

Disentangling Prompt Dependence to Evaluate Segmentation Reliability in Gynecological MRI

Elodie Germani¹[0000–0002–5786–9538], Krystel Nyangoh-Timoh^{1,2}, Pierre Jannin¹, and John S.H. Baxter¹

¹ Laboratoire Traitement du Signal et de l’Image (LTSI, INSERM UMR 1099),
Université de Rennes, Rennes, France

² Centre Hospitalier Universitaire de Rennes (CHU Rennes), Rennes, France
elodie.germani@univ-rennes.fr

Abstract. Promptable segmentation models (*e.g.*, the Segment Anything Models) enable generalizable, zero-shot segmentation across diverse domains. Although predictions are deterministic for a fixed image–prompt pair, the robustness of these models to variations in user prompts, referred to as *prompt dependence*, remains underexplored. In safety-critical workflows with substantial inter-user variability, interpretable and informative frameworks are needed to evaluate prompt dependence. Here, we assess the reliability of promptable segmentation by analyzing and measuring its sensitivity to prompt variability. We introduce the first formulation of prompt dependence that explicitly disentangles prompt ambiguity (inter-user variability) from local sensitivity (interaction imprecision), offering an interpretable view of segmentation robustness. Experiments on two female pelvic MRI datasets for uterus and bladder segmentation reveal a strong negative correlation between both metrics and segmentation performance, highlighting the value of our framework for assessing robustness. The two metrics have low correlation and mutual information, supporting the disentangled design of our formulation. By providing meaningful indicators of prompt-related failure modes, our framework helps identify, at the image level, which samples are at risk and how to mitigate those risks.

Keywords: Promptable Segmentation · Robustness · Pelvic MRI · Prompt Dependence · Inter-user Variability · Measurement noise.

1 Introduction

Foundation Models [1] have recently reshaped the field of semantic segmentation by separating visual understanding from user input, enabling a single model to generalize across datasets and tasks. The Segment Anything Model (SAM) [2] exemplifies this paradigm: given an image and a lightweight prompt (*e.g.*, points or bounding boxes), the model outputs a segmentation mask without task- or domain-specific fine-tuning. In medical imaging, this paradigm is of high interest given the scarcity of annotations and the heterogeneity of acquisition conditions.

Recent work, such as MedSAM [3], shows that effectively deploying SAM-based models in clinical settings still requires addressing important challenges.

One challenge is that promptable segmentation models’ outputs can be highly dependent on the prompt: small changes in a bounding box can lead to large changes in the predicted mask [4]. This effect is exacerbated in medical imaging, where low contrast, artifacts, and diffuse boundaries can induce inter-observer variability even among experts [5, 6]. Moreover, inter-observer variability varies across image modalities and structures: organs with sharp boundaries (*e.g.*, bladder) show higher agreement than organs with smooth or infiltrative boundaries (*e.g.*, uterus) [7]. In this work, we aim to improve the reliability of promptable segmentation models by analyzing their dependence on prompt variability.

Prompt dependence can be decomposed into two distinct components reflecting different sources of variability. First, *prompt ambiguity* refers to situations where multiple prompts are clinically plausible for a given image and target. This image property captures inter-observer variability in prompt definition due to diffuse boundaries or subjective interpretation. Second, *measurement noise* reflects imprecision: when an observer intends a specific prompt, discretization and drawing variability introduce small perturbations. This property is also intrinsic to the model: even when the target is unambiguous, a model may yield different outputs for minor variations in the prompt.

Previous studies on prompt dependence in segmentation models typically address these components together. To mitigate prompt dependence and improve robustness, prompt refinement methods [8, 9] use detection networks to infer an optimal prompt for a given image-label pair. Similarly, [10] proposes iteratively generating prompts and selecting those that maximize a heuristic or an uncertainty-guided criterion, which can improve results but at the expense of increased computational cost. On the contrary, [11] treats the prompt as fixed to compute uncertainty at the mask level via ensembles or stochastic inference. However, both approaches conflate local sensitivity with prompt uncertainty, leading to prompt refinement strategies that collapse multiple plausible hypotheses into a single prompt or favor consistent yet incorrect solutions, and mask-level uncertainties that appear overconfident by ignoring prompt ambiguity.

Our **contributions** are threefold: (i) we introduce the first framework for prompt dependence that disentangles prompt ambiguity (inter-user variability) from local sensitivity (interaction imprecision), providing an interpretable perspective on segmentation reliability ; (ii) we model the image-conditioned prompt distribution using a Mixture Density Network (MDN), allowing us to quantify image-level prompt ambiguity and propagate it to pixel-level uncertainty ; (iii) we propose a novel stability margin formulation to estimate local sensitivity under measurement noise, using realistic prompt perturbations to predict the minimal perturbation level required to induce mask changes. We perform an extensive evaluation of two promptable segmentation models in female pelvic MRI, assessing the potential of our framework for reliability and robustness estimation.

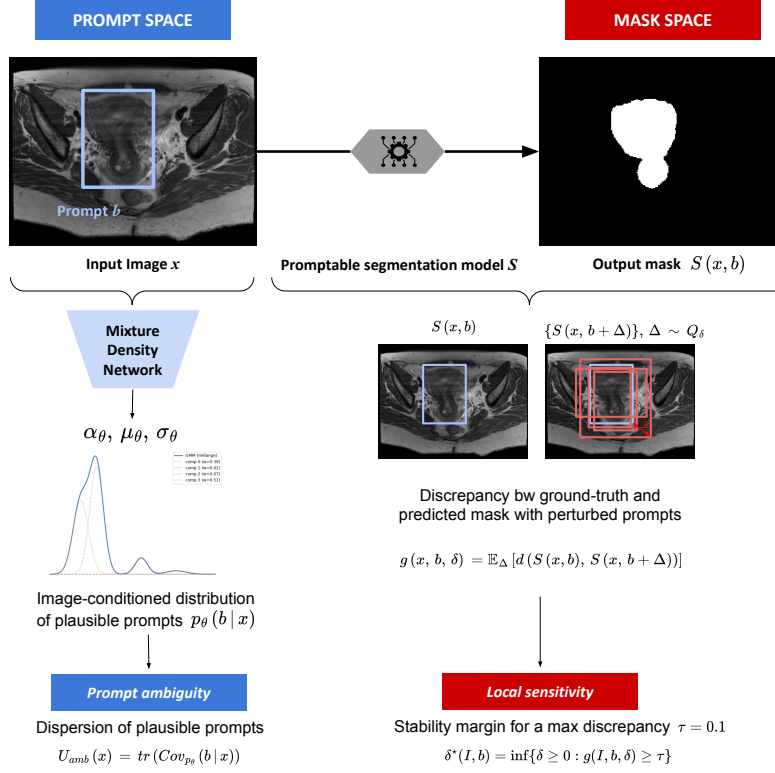


Fig. 1. Overview of the prompt dependence evaluation framework. In prompt space (*left*), a conditional mixture model estimates the distribution of plausible prompts $P_{\text{valid}}(b|I)$. The dispersion of this distribution, quantified by the trace of its covariance, defines **prompt ambiguity**. By adding controlled perturbations around reference prompts and measuring their impact in mask space, we train a stability-margin predictor to estimate the minimum change needed to produce a meaningful discrepancy between masks, thereby providing a formulation of **local sensitivity**.

2 Methods

2.1 Theoretical framework

Let $S : \mathcal{I} \times \mathcal{B} \rightarrow [0, 1]^{H \times W}$ be a frozen promptable segmentation model mapping an image $I \in \mathcal{I}$ and a prompt $b \in \mathcal{B} \subset \mathbb{R}^4$ to a probabilistic mask $\hat{y} = S(I, b)$. Here, both the image I and the model S are fixed, and we analyze variability in probabilistic masks $\hat{y}$ for a segmentation task t with respect to a distribution of prompts B . We explore two levels of variability: between masks generated from different plausible prompts, and local variability around a reference prompt.

Variance decomposition. For a given image I and task t , we assume a distribution of clinically plausible prompts $B \sim P_{\text{valid}}(b|I, t)$ capturing geometric

ambiguity. The prompt-marginalized prediction $\hat{Y} = \mathcal{S}(I, B)$ is then a random variable conditioned on (I, t) , whose variability is quantified by

$$\text{Var}(\hat{Y} \mid I, t) = \text{Var}_{B \sim P_{\text{valid}}(\cdot \mid I, t)}[\mathcal{S}(I, B)]$$

To characterize local prompt sensitivity, we further introduce a perturbation variable $\Delta \sim Q_\delta(\cdot \mid B)$ representing small deviations around a reference prompt B , and define the perturbed prediction $\hat{Y}' = \mathcal{S}(I, B + \Delta)$.

Applying the law of total variance to $\hat{Y}'$ with conditioning variable (B) yields

$$\text{Var}(\hat{Y}' \mid I, t) = \mathbb{E}_B [\text{Var}_\Delta(\mathcal{S}(I, B + \Delta) \mid B)] + \text{Var}_B (\mathbb{E}_\Delta[\mathcal{S}(I, B + \Delta) \mid B])$$

The first term $\text{Var}_\Delta(\mathcal{S}(I, B + \Delta) \mid B)$ quantifies the variability of the model $\mathcal{S}$ under prompt perturbations Δ around a reference prompt B . The second term $\text{Var}_B (\mathbb{E}_\Delta[\mathcal{S}(I, B + \Delta) \mid B])$, can be approximated for sufficiently small perturbations Δ by $\mathbb{E}_\Delta[\mathcal{S}(I, B + \Delta) \mid B] \approx \mathcal{S}(I, B)$. Overall, this leads to the following decomposition of the variance of prompt-marginalized predictions:

$$\text{Var}(\hat{Y}') \approx \underbrace{\mathbb{E}_B [\text{Var}_\Delta \mathcal{S}(I, B + \Delta)]}_{\text{Local sensitivity}} + \underbrace{\text{Var}_B [\mathcal{S}(I, B)]}_{\text{Prompt ambiguity}}.$$

By introducing a local perturbation variable and applying the law of total variance, we decompose the variance of prompt-marginalized predictions into (i) the expected local variance induced by prompt perturbations (*local sensitivity*) and (ii) the variance across reference prompts (*prompt ambiguity*).

Interpretation of local sensitivity. Rather than estimating local sensitivity via infinitesimal perturbations, we define it in terms of a stability margin. For an image I and a reference prompt b , we set $d(\cdot, \cdot)$ to be a mask discrepancy measure. For perturbations $\Delta \sim Q_\delta(\cdot \mid b)$ of amplitude δ , we define

$$g(I, b, \delta) = \mathbb{E}_\Delta [d(\mathcal{S}(I, b), \mathcal{S}(I, b + \Delta))].$$

Given a tolerance threshold τ representing the maximum acceptable discrepancy between masks, we define the stability margin as $\delta^*(I, b) = \inf\{\delta \geq 0 : g(I, b, \delta) \geq \tau\}$. This quantity represents the smallest perturbation level required to induce a non-acceptable change in the predicted mask. Larger values of δ^* indicate robust segmentation with respect to prompt perturbations.

Interpretation of prompt ambiguity. For a fixed image I , let $B \sim P_{\text{valid}}(b \mid I)$ denote a random plausible prompt and define $f(b) = \mathcal{S}(I, b)$. The prompt-marginalized prediction $f(B)$ is then a random variable with variability $\text{Var}_B[f(B)]$. Assuming f is locally differentiable in a neighborhood of $\mu_B = \mathbb{E}[B]$ and that B has covariance Σ_B , a first-order expansion gives

$$f(B) \approx f(\mu_B) + \nabla_b f(\mu_B)^\top (B - \mu_B) \Rightarrow \text{Var}_B[f(B)] \approx \nabla_b f(\mu_B)^\top \Sigma_B \nabla_b f(\mu_B).$$

This approximation makes explicit that prompt-induced variability increases with the dispersion of the valid-prompt distribution in prompt space (captured by Σ_B) and with the local prompt sensitivity (captured by $\nabla_b f(\mu_B)$).

2.2 Empirical estimation

Prompt ambiguity. In practice, the true distribution of valid prompts $P_{\text{valid}}(b | I)$ is unknown. We approximate it with a conditional density model $p_{\theta}(b | x)$, $x = \text{pool}(E(I))$, where E is the frozen image encoder and p_{θ} is implemented as a mixture density network (MDN) [12, 13] (see Fig. 1, left). The MDN is trained by conditional maximum likelihood on pairs (x_i, b_i) . This parametric approximation of the image-conditioned prompt distribution enables estimation of prompt ambiguity via the predictive covariance of the learned mixture, which captures both intra-mode dispersion and inter-mode separation $U_{\text{amb}}(I) = \text{tr}(\text{Cov}_{p_{\theta}}(b | x))$. Here, U_{amb} only captures the prompt-space dispersion Σ_B and deliberately omits the Jacobian term $\nabla_b f(\mu_B)$, making it, by design, a model-agnostic property of the image. Thus, U_{amb} is a faithful proxy of mask-space prompt ambiguity insofar as $\nabla_b f$ is non-degenerate and approximately comparable across the images.

Local sensitivity. We estimate the stability margin $\delta^*(I, b)$ for a given image around a representative plausible prompt b sampled from $p_{\theta}(b | x)$. We evaluate $g(I, b, \delta)$ over a discrete set of perturbation levels and select the smallest δ such that the average mask difference exceeds $\tau = 0.1$. This empirical procedure requires no additional ground-truth masks and relies solely on the model’s response to synthetic, clinically relevant perturbations. To enable efficient inference, we use a pretrained stability-margin predictor r_{ψ} , that provides a fast approximation of local sensitivity without iterative sampling (see Fig. 1, right): $\delta^*(I, b) = r_{\psi}(I, b, \tau)$ with $\tau = 0.1$.

The proposed framework provides theory-motivated operational proxies of the two theoretical components of prompt-induced variability: (i) ambiguity of valid prompts, modeled explicitly via a conditional mixture density, and (ii) local sensitivity under measurement noise, captured by the stability margin. Their interpretation relies on assumptions: $p_{\theta}(b | x)$ captures dominant modes of plausible prompts; the prompt-to-mask mapping is locally regular within these prompts; and perturbations reflect realistic user imprecision. Under these assumptions, U_{amb} captures image-driven ambiguity in prompt definition, while δ^* captures local model robustness around a plausible prompt.

3 Experiments and Results

Datasets. We evaluate the proposed framework on two datasets of T2-weighted pelvic MRI acquired in the axial plane: UT-EndoMRI [14] and MOGaMBO [15]. UT-EndoMRI contains $N = 81$ female patients with endometriosis and manual uterus segmentations from 1 to 3 raters, while MOGaMBO includes $N = 94$ female patients with locally advanced cervical cancer and bladder segmentations. We apply a standard preprocessing pipeline: 3D volumes and labels converted into 2D axial slices, resampled to a spacing of 1.0×1.0 mm, and reoriented to a common coordinate space. Pixel intensities are min-max normalized per slice, and only slices containing organ annotations are retained. Each selected slice is

resized to 1024×1024 and exported as a 3-channel image. We derive a ground-truth bounding box $b = x_1, y_1, x_2, y_2$ from each manual mask and reparameterize it for stable learning as $u = \text{logit}(x_c), \text{logit}(y_c), \log H, \log W$, where (x_c, y_c) are normalized box-center coordinates and (H, W) its height and width. Datasets are split at the patient level into 70%/10%/20% train/val/test, with respectively $\{773, 501, 255\}$ and $\{195, 128, 62\}$ slices for MOGaMBO and UT-EndoMRI.

Implementation. We apply our framework for prompt dependence assessment on two frozen SAM-based architectures: MobileSAM [16] and MedSAM [3]. Image embeddings x are extracted for each 2D slice using each image encoder (with $\dim(x) = 512$), and ground-truth bounding boxes are computed from manual segmentation masks. When multiple masks are available for a slice, each mask yields a separate pair (x_i, b_i) , so that $p_\theta(b | x)$ is fit on real inter-rater mask variability. For prompt ambiguity, we train a conditional mixture density network (MDN) implemented as a two-layer multi-layer perceptron (MLP) with ReLU activations, taking as input the mean-max pooled image embedding and predicting the parameters of a Gaussian mixture distribution $p_\theta(b | x)$ with $K = 8$ components, including mixture weights, means, and diagonal covariances. The number of components has been chosen by fitting standard GMM with various K values and the value giving the lowest BIC score on the validation set was chosen. The MDN is trained with a negative log-likelihood (nll) loss, the Adam optimizer with a learning rate of 1×10^{-3} , and a batch size of 8 for 100 epochs. The parameters giving the lowest nll on the validation sets were selected. The stability margin predictor is also implemented as a two-layer MLP that takes as input the concatenation of the image embedding, bounding box, and the tolerance threshold τ . Oracle stability margins are computed offline using 32 samples per perturbation type (translation or scale) and level over a discrete grid of δ values ($\delta \in \{0.005, 0.01, 0.05\}$). The model learns to predict stability margins δ^* using an ℓ_1 loss, an Adam optimizer with a learning rate of 1×10^{-3} , and a batch size of 64 for 100 epochs. All models are implemented in PyTorch v2.4.1.

3.1 Prompt ambiguity and local sensitivity in organ segmentation

Both the conditional MDN and the margin predictor were trained successfully, achieving $\text{nll} \approx -2.5$ and $\ell_1 \approx 0.001$ on validation sets. Fig. 2 (top panel) displays representative samples from each dataset, their corresponding ground-truth bounding boxes, and prompt dependency metrics from MobileSAM experiments. Bounding box locations and sizes vary across samples and organs, but do not appear to be associated with ambiguity or sensitivity. Organ shapes and boundaries, however, appear irregular and blurry in samples with high U_{amb} . Regarding local sensitivity, samples with low stability margins δ^* tend to have smaller bounding boxes that are close to adjacent structures, which could explain the low perturbation level required to induce mask changes at $\tau = 0.1$. Similar observations were made for experiments with MedSAM. Average metrics values for each dataset and model are shown in Fig. 2 (bottom panel). Across both datasets and models, prompt ambiguity and local sensitivity exhibit low

UT-EndoMRI

MOGaMBO

$U_{amb}=0.129$
 $\delta^*=0.0095$

$U_{amb}=0.241$
 $\delta^*=0.00475$

$U_{amb}=0.00922$
 $\delta^*=0.00987$

$U_{amb}=0.299$
 $\delta^*=0.00631$

Dataset	$Dice$ (SEM)	U_{amb} (min, max)	δ^* (min, max)	$r(U_{amb}, \delta^*)$	$r(U_{amb}, Dice)$	$r(\delta^*, Dice)$
MobileSAM [16]						
MOGaMBO	0.81 (0.004)	0.14 (0.02-0.34)	0.01 (0.003-0.05)	-0.07	-0.45*	0.36*
UT-EndoMRI	0.86 (0.004)	0.19 (0.12-0.23)	0.01 (0.003-0.02)	-0.001	-0.30*	0.44*
MedSAM [3]						
MOGaMBO	0.33 (0.07)	0.07 (0.002-0.37)	0.003 (0.0002-0.02)	-0.04	-0.30*	0.21*
UT-EndoMRI	0.80 (0.005)	0.21 (0.18-0.27)	0.0001 ($1e^{-5}$ -0.0004)	0.05	-0.37*	0.46*

Fig. 2. Qualitative examples (top) and dataset-level summary statistics (bottom) of prompt dependence metrics, U_{amb} and δ^* , for UT-EndoMRI and MOGaMBO. For each dataset, two representative slices are shown with the corresponding U_{amb} and δ^* values; the blue boxes highlight the ground-truth bounding box. The table reports the mean Dice ($\pm$ SEM), the ranges of U_{amb} and δ^* , and Pearson correlations on the test set. Stars $\star$ indicate significant correlations at $p < 0.001$.

correlations ($r_{MOG} = -0.07$ and $r_{UT} = 0.001$ with MobileSAM), supporting the complementary yet distinct aspects of prompt dependence.

3.2 Prompt dependence as an indicator of segmentation quality

We use the median values of U_{amb} and δ^* within each dataset (Fig. 3, top panel) to define four classes of prompt dependence. Overall, segmentation quality is lower in images with both high prompt ambiguity and high local sensitivity (+/+) than in the other classes ($\overline{Dice}_{+/+} = 0.29$, $\overline{Dice}_{-/-} = 0.73$ for MOGaMBO with MobileSAM). The association between prompt dependence metrics and segmentation quality is further supported by the inverse correlations of each metric with Dice scores (e.g., $r(Dice, U_{amb}) = -0.45$ for MOGaMBO and $r(Dice, U_{amb}) = -0.30$ for UT-EndoMRI with MobileSAM). However, the level of correlation varies across datasets, models, and metrics: in MOGaMBO, Dice is more strongly correlated with U_{amb} ($r = -0.45$) than with δ^* ($r = 0.36$); conversely, in UT-EndoMRI, Dice is more strongly correlated with δ^* ($r = 0.44$) than with U_{amb} ($r = -0.30$). Similar, though less pronounced, results are observed in experiments with MedSAM (see Fig. 3, top panel). These results suggest that the main driver of prompt dependence is dataset-specific and differently encoded by image encoders, reflecting complementary failure modes and differences in image quality or organ appearance.

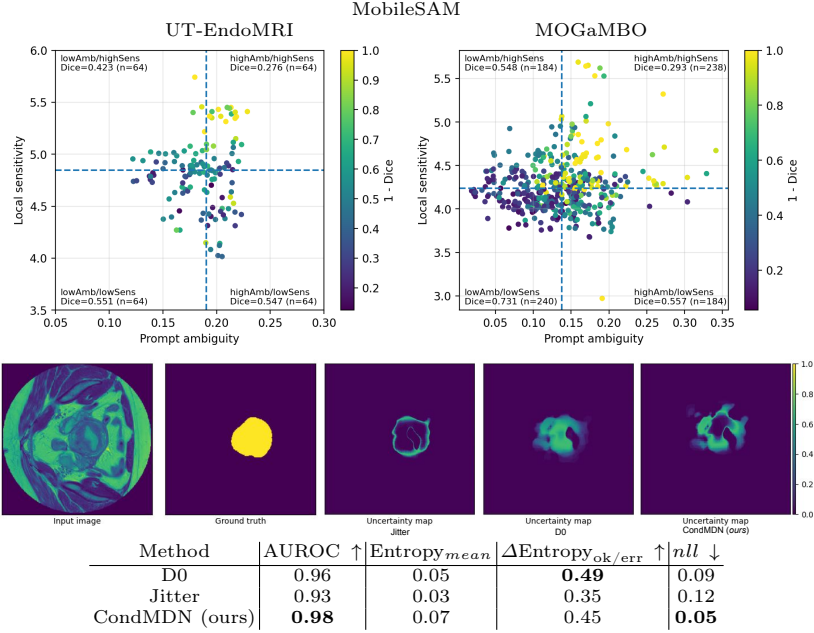


Fig. 3. (*Top*) Distribution of prompt ambiguity (x-axis) and local sensitivity (y-axis, encoded as $\log(\frac{1}{1-\delta^*})$), color-coded by $1 - \text{Dice}$. Dashed lines denote the median values used to define the four classes. (*Bottom*) Qualitative example of uncertainty maps produced by Jitter, DO, and CondMDN with the input image and ground truth mask on UT-EndoMRI, with averaged evaluation metrics of uncertainty maps.

3.3 From prompt ambiguity to pixel-level uncertainty

Beyond prompt-ambiguity scores, $p_\theta(b \mid x)$ enables a wide range of analyses. Here, we propose using this distribution to derive pixel-wise uncertainty maps that represent uncertainty due to inter-user variability. For each image, we sample $N = 1000$ prompts, decode the corresponding masks, and compute the per-pixel entropy across the resulting mask ensemble. Fig. 3 (bottom) illustrates uncertainty maps obtained with three strategies: (i) jittering a fixed, user-provided prompt, (ii) sampling prompts from a non-conditional distribution D_0 , and (iii) sampling from the conditional model $p_\theta(b \mid x)$. Unlike standard perturbation-based approaches, which primarily capture local prompt sensitivity, conditional sampling marginalizes over the space of clinically plausible prompts, yielding spatially coherent uncertainty concentrated along ambiguous anatomical boundaries. Compared to the aleatoric prompt-entropy baseline D_0 , our model captures image-specific ambiguity rather than generic prompt variability. Quantitatively, uncertainty maps derived from $p_\theta(b \mid x)$ show high spatial calibration, with high entropy consistently aligned with regions of segmentation errors (on average, $\Delta\text{Entropy}_{\text{ok/err}} = 0.45$ in UT-EndoMRI, and 0.38 in MOGaMBO for uncertainty maps with MobileSAM).

4 Conclusion

In this work, we evaluate the reliability of medical image segmentation by focusing on prompt variability, characterized by inter-user variability in prompt definitions, which increases in ambiguous cases, and by measurement imprecision due to local sensitivity in image-model settings. We propose theory-motivated operational proxies that correlate significantly with segmentation quality and enable interpretable evaluation of failure modes in promptable segmentation. Our study has several limitations. First, our empirical evaluation does not verify whether the tested SAM-based models satisfy the local-regularity assumption underlying our proxies, particularly in the high-ambiguity regime that the framework aims to characterize; assessing the smoothness of the prompt-to-mask mapping in this regime remains an open question. Second, U_{amb} is interpreted as reflecting inter-user variability; while $p_{\theta}(b | x)$ is fit on multi-rater masks, our current evaluation establishes an association with segmentation difficulty rather than a direct validation against prompts explicitly elicited from multiple annotators across both datasets. Third, further experiments are needed to better understand the impact of hyperparameters, such as the number of components used to model $p_{\theta}(b | x)$ and the stability margin threshold τ . Future work will address these points and explore how best to integrate this evaluation framework into clinical practice, to improve target delineation in radiotherapy and to extract robust organ and lesion biomarkers from images. Our code is publicly available at: <https://github.com/elodiegermani/PromptDependence>.

Disclosure of Interests. The authors have no competing interests to declare relevant to the content of this article.

References

1. Moor, M., Banerjee, O., Abad, Z., Krumholz, H., Leskovec, J., Topol, E. & Rajpurkar, P. Foundation models for generalist medical artificial intelligence. *Nature*. **616**, 259-265 (2023).
2. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A., Lo, W., Dollar, P. & Girshick, R. Segment Anything. *Proceedings Of The IEEE/CVF International Conference On Computer Vision (ICCV)*. pp. 4015-4026 (2023,10).
3. Ma, J., He, Y., Li, F., Han, L., You, C. & Wang, B. Segment anything in medical images. *Nature Communications*. **15**, 654 (2024,1).
4. Rahman, M., Munir, M., Jha, D., Bagci, U. & Marculescu, R. PP-SAM: Perturbed Prompts for Robust Adaption of Segment Anything Model for Polyp Segmentation. 4989-4995, (2024), *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*
5. Andreadis, G., Visser-Groot, W., Boer, S., Bosman, P. & Alderliesten, T. Inter-observer variability of corresponding anatomical landmark placement in pelvic CT and MRI scans. *Radiotherapy And Oncology*. pp. 111430 (2026,2).
6. Zheng, W., Yang, X., Cheng, Z., Lian, J., Li, E., Mo, S., Liu, Y. & Huang, S. Interobserver and sequence variability in the delineation of pelvic organs at risk on magnetic resonance images. *Radiology And Oncology*. **59**, 139-146, (2025).

7. Zuo, J., Feng, F., Wang, Z., Ashton-Miller, J., DeLancey, J. & Luo, J. Semi-supervised deep learning for uterus and bladder segmentation on female pelvic floor magnetic resonance imaging with limited labeled data. *American Journal Of Obstetrics And Gynecology*. **234**, 461-470 (2026,2).
8. Chen, Y., Son, M., Hua, C. & Kim, J. AoP-SAM: automation of prompts for efficient segmentation. *Proceedings Of The Thirty-Ninth AAAI Conference On Artificial Intelligence And Thirty-Seventh Conference On Innovative Applications Of Artificial Intelligence And Fifteenth Symposium On Educational Advances In Artificial Intelligence*. **39** pp. 2284-2292 (2025,2).
9. Zheng, X., Zhang, Y., Zhang, H., Liang, H., Bao, X., Jiang, Z. & Lao, Q. Curriculum Prompting Foundation Models for Medical Image Segmentation. *Medical Image Computing And Computer Assisted Intervention – MICCAI 2024*. pp. 487-497 (2024)
10. Zhou, N., Zou, K., Ren, K., Luo, M., He, L., Wang, M., Chen, Y., Zhang, Y., Chen, H. & Fu, H. MedSAM-U: Uncertainty-Guided Auto Multi-Prompt Adaptation for Reliable MedSAM. *IEEE Transactions On Circuits And Systems For Video Technology*. pp. 1-1 (2025).
11. Kaiser, T., Norrenbrock, T. & Rosenhahn, B. UncertainSAM: Fast and Efficient Uncertainty Quantification of the Segment Anything Model. (2025,6).
12. Bishop, C. Mixture density networks. (1994).
13. McLachlan, G. & Basford, K. Mixture models : inference and applications to clustering. (Marcel Dekker,1988,1).
14. Liang, X., Alpuing Radilla, L., Khalaj, K., Dawoodally, H., Mokashi, C., Guan, X., Roberts, K., Sheth, S., Tammisetti, V. & Giancardo, L. A Multi-Modal Pelvic MRI Dataset for Deep Learning-Based Pelvic Organ Segmentation in Endometriosis. *Scientific Data*. **12**, 1292 (2025,7).
15. Manna, S., Das, S., Ghosh, S. & Bhattacharya, S. Federated Self-Supervised Learning for One-Shot Cross-Modal and Cross-Imaging Technique Segmentation. (arXiv,2025).
16. Zhang, C., Han, D., Qiao, Y., Kim, J., Bae, S., Lee, S. & Hong, C. Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. (arXiv,2023)