Towards Integrating Uncertainty for Domain-Agnostic Segmentation

Jesse Brouwers **Xiaoyan Xing** **Alexander Timans**
UvA-Bosch Delta Lab, University of Amsterdam

Abstract

Foundation models for segmentation such as the Segment Anything Model (SAM) family exhibit strong zero-shot performance, but remain vulnerable in shifted or limited-knowledge domains. This work investigates whether uncertainty quantification can mitigate such challenges and enhance model generalisability in a domain-agnostic manner. To this end, we (1) curate `UncertSAM`, a benchmark comprising eight datasets designed to stress-test SAM under challenging segmentation conditions including shadows, transparency, and camouflage; (2) evaluate a suite of lightweight, *post-hoc* uncertainty estimation methods; and (3) assess a preliminary uncertainty-guided prediction refinement step. Among evaluated approaches, a last-layer Laplace approximation yields uncertainty estimates that correlate well with segmentation errors, indicating a meaningful signal. While refinement benefits are preliminary, our findings underscore the potential of incorporating uncertainty into segmentation models to support robust, domain-agnostic performance. Our benchmark and code are made available at `https://github.com/JesseBrouw/UncertSAM`.

1 Introduction

As in other domains, large-scale foundation models have transformed vision-based tasks and enabled effective generalisation to novel tasks through zero- or few-shot prompting [Bommasani et al., 2021]. For segmentation tasks, the Segment Anything Model family (SAM) stands out through its high-quality segmentation masks leveraging minimal user input, such as point clicks or bounding boxes [Kirillov et al., 2023]. Trained on the SA-1B dataset of over *one billion* masks, the second SAM iteration [Ravi et al., 2025] demonstrates exceptional generalisation, yet struggles with fine-grained structures, precise object boundaries, and sensitivity to common real-world degradations such as motion blur, noise, shadows, and transparency [Kirillov et al., 2023, Ji et al., 2024, Wang et al., 2024, Chen et al., 2024]. Uncertainty quantification (UQ) offers a potential avenue to enhance SAM's robustness to such cases. By supplementing predictions with an added measure of uncertainty or confidence, UQ can help raise the trustworthiness of segmentation outputs and notify the model when it risks being wrong [Gawlikowski et al., 2023]. However, current efforts to incorporate uncertainty into SAM are limited in scope. Most existing studies focus on narrow task settings or estimate uncertainty purely via heuristics such as boundary cues, rather than eliciting it more fundamentally from the model [Zhang et al., 2023, Zhou et al., 2025, Liu et al., 2024, Kaiser et al., 2025, Xie et al., 2024]. Thus, it remains somewhat unclear whether meaningful spatial uncertainty estimates can be obtained, and whether these can serve as a signal to refine predictions in a general, domain-agnostic fashion (see Fig. 1 for a visual example).

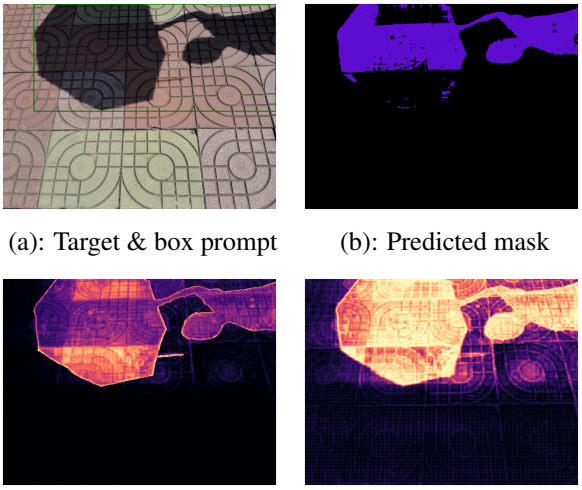
In this work, we address these questions by comparing four approaches to quantify pixel-level segmentation uncertainty via controlled perturbations to the input, prompts, model parameters, or an additional variance network (Fig. 2). Our approaches are driven by the desire for a lightweight,

2nd Edition of the Workshop on Frontiers in Probabilistic Inference (NeurIPS 2025)

post-hoc uncertainty integration amenable to work with a pretrained and *frozen* SAM encoder[1]. While all methods yield reasonable uncertainties, we find that a last-layer Laplace approximation most strongly correlates with segmentation errors, highlighting its potential to guide prediction refinement.

We take a first step towards a refinement strategy via a dense embedding that fuses uncertainty maps into SAM's encoder representations (App. D), but minor gains over control baselines suggest it falls short of fully exploiting the uncertainty signal. We posit this originates from limitations of our purely *post-hoc* approach, and stipulate a deeper integration of uncertainty into model architecture to improve performance benefits. In summary, we contribute:

- `UncertSAM`, a curated multi-domain benchmark featuring challenging segmentation cases and enabling domain-agnostic evaluation of methods;
- a systematic comparison of four pixel-level uncertainty methods for SAM, showing strong alignment between uncertainty and segmentation errors;
- a simple prediction refinement strategy leveraging uncertainty estimates, in part achieving small gains while necessitating no domain-specific fine-tuning.



(a): Target & box prompt (b): Predicted mask

(c): Prob. error map (d): Uncertainty map

Figure 1: **An example of SAM's failure in shadow detection.** The predicted mask misses shadow regions despite an accurate bounding box prompt. In contrast, a Laplace-based uncertainty map correctly recovers the full shadow (from the ISTD dataset [Wang et al., 2018]).

2 Methodology

We next detail our three-step approach on benchmarking, uncertainty estimation and refinement.

Benchmark curation. We establish the `UncertSAM` benchmark by collecting and standardising eight datasets spanning a range of challenging visual conditions and environments for segmentation. These include fine-grained salient objects (BIG [Cheng et al., 2020], COIFT [Liew et al., 2021]), camouflaged objects (COD [Fan et al., 2022]), medical CT scans (MSD Spleen [Antonelli et al., 2022]), shadows (ISTD [Wang et al., 2018], SBU [Vicente et al., 2016]), lighting artifacts (Flare [Dai et al., 2022]), and transparent objects (Trans [Xie et al., 2021]). This results in a collection of over 23,000 images and 44,000 annotated masks across different domains and edge cases, which can be used to evaluate baseline segmentation performance, UQ methods, and uncertainty-guided refinement. To ensure domain-agnostic analysis, any uncertainty fitting (for the Laplace) or training (for the variance network) is done on a representative subset of SAM's original training set (SA-1B [Kirillov et al., 2023]). Further dataset details are provided in App. B.

Uncertainty quantification for SAM. We consider four UQ strategies which target different components of the model design, thus providing complementary views on arising uncertainty (see Fig. 2). Since we refrain from modifying the parameter-heavy SAM image encoder, our approaches are generally *post-hoc* and target the lightweight SAM prompt and decoder modules. We refer to Ravi et al. [2025] for an architectural overview of SAM and its three key components. Consider the input combination of an image $X \in \mathbb{R}^{H \times W \times C}$ and prompt configuration q, such as user-provided bounding box coordinates $q \in \mathbb{R}^4$ around the target object to segment. We generically define f_θ as the parametrised SAM model with weights θ, and subsequently $f_\theta(X, q) \in \mathbb{R}^{H \times W}$ as the model's pixel-wise output logits for the input tuple (X, q). As SAM targets binary foreground/background segmentation, a sigmoid function σ can be applied pixel-wise to obtain a final probability map $P \in [0, 1]^{H \times W}$. With this notation in hand, we employ the following four uncertainty strategies:

[1]We refer to SAM more broadly, but in practice work with SAM-2 [Ravi et al., 2025].

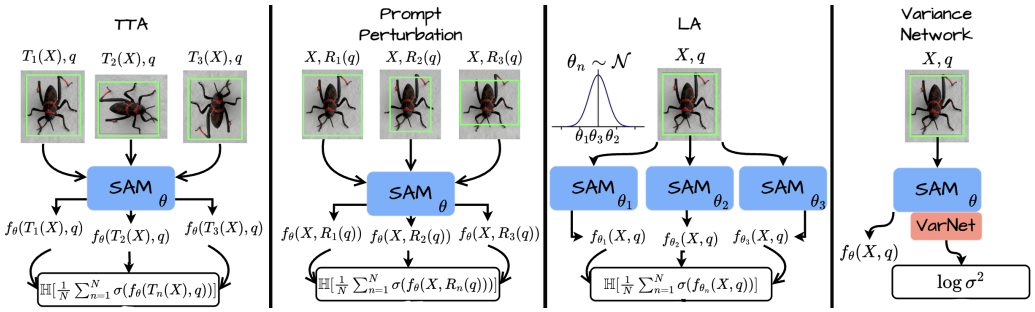


Figure 2: **Overview of our four uncertainty strategies.** *From left to right:* We quantify pixel-level uncertainty by targeting the input image (via test-time augmentations, TTA), prompts, model parameters (via the Laplace approximation, LA), and an additional variance prediction head. The first three methods leverage stochasticity to generate ensembles, whereas the latter is deterministic.

(i) Test-Time Augmentations (TTA). We perturb the input image by sampling and applying a stochastic augmentation $T_n \sim \mathcal{T}$, resulting in the probability map $P_n = \sigma(f_\theta(T_n(X), q))$. We consider augmentations used during SAM's original training (*e.g.* flips, resizes, jitter; see Table 4) as well as additional hue shifts. Sampling N times yields a set of maps $\{P_1, \ldots, P_N\}$, which can then be used to obtain a pixel-wise mean prediction map $\bar{P} = \frac{1}{N}\sum_{n=1}^{N} P_n$ and uncertainty map U. Among different options to measure uncertainty we consider the pixel-wise *predictive entropy* $\mathbb{H}[\cdot]$, given for the binary case as $U = \mathbb{H}[\bar{P}] = -\bar{P}\log\bar{P} - (1-\bar{P})\log(1-\bar{P})$. We refer to this as our *predictive* spatial uncertainty map, rather than claiming distinct origins [Hüllermeier and Waegeman, 2021].

(ii) Prompt Perturbations. Instead of augmenting the input image, we may also perturb the input prompt by sampling bounding box coordinate perturbations $R_n \sim \mathcal{R}$ and obtaining the probability map $P_n = \sigma(f_\theta(X, R_n(q)))$. We leverage the existing prompt perturbation schedule used during SAM's training [Kirillov et al., 2023, Ravi et al., 2025], and generate an ensemble mean $\bar{P}$ and uncertainty map U as above.

(iii) Last-layer Laplace Approximation (LA). In order to generate an ensemble from model parameters, we consider a scalable Bayesian treatment via the Laplace approximation over the final linear decoder layer [MacKay, 1992, Daxberger et al., 2021]. A Gaussian posterior approximation is centred over the layer's pretrained model weights—interpretable as the *maximum a posteriori* estimates $\hat{\theta}_{\text{MAP}}$—with variance dictated by the local curvature of the (diagonal) Hessian $\hat{H}$, that is $p(\theta \mid \mathcal{D}_{\text{fit}}) \approx \mathcal{N}(\theta \mid \hat{\theta}_{\text{MAP}}, \hat{H}^{-1})$. This offers a relatively crude, but scalable and *post-hoc* Bayesian model treatment, and model weights can now simply be sampled as $\theta_n \sim p(\theta \mid \mathcal{D}_{\text{fit}})$ to produce the probability map $P_n = \sigma(f_{\theta_n}(X, q))$ and obtain $\bar{P}$ and U as before.

(iv) Learnable variance network. Finally, a more distinct approach inspired by Kendall and Gal [2017] sees training an auxiliary uncertainty prediction head on top of SAM's decoder features. The head learns to predict a pixel-wise log variance term, and is interpretable as a Gaussian 'spread' given the trained likelihood objective (see Kendall and Gal [2017] and App. C for more details). Both the decoder and mean prediction are kept frozen, and thus only a notion of uncertainty is learned in a strictly *post-hoc* way. In contrast to above, the returned uncertainty map U is not based on ensembling but comes from a deterministic variance prediction given the inputs (X, q).

Uncertainty-guided prediction refinement. How can the obtained map U subsequently be used to improve predictions and help mitigate failures such as Fig. 1? In this work, we consider a simple approach dubbed *Dense Embedding Fusion*, which encodes both prediction and uncertainty into a dense embedding and applies a 1×1 convolution to produce an uncertainty-aware fused feature map. Repeating a second forward pass through SAM, we expect this map to supplement the decoder's internal spatial features with uncertainty information to aid correct initial mistakes. See App. D for a schematic and details. We stress that this is a *preliminary first step* towards more elaborate strategies.

3 Experimental Results

We assess UQ methods by their correlation with prediction error in Fig. 3, and then test their utility for downstream refinement in Table 1. All experiments make use of bounding box prompts only.

	BIG	COIFT	COD	MSD	ISTD	SBU	Flare	Trans
Mean IoU (↑)								
Ground truth mask	*93.96*	*95.65*	*89.20*	*91.35*	*92.52*	*87.09*	*82.15*	*94.92*
SAM (No Refinement)	**89.95**	**94.77**	**82.82**	87.86	65.71	66.11	42.82	84.87
Dense Fusion w/ SAM (Ones Map)	89.82	94.73	82.72	**87.91**	66.53	66.27	43.12	85.05
Dense Fusion w/ LA (Ones Map)	89.34	94.74	82.54	87.25	**67.89**	**67.37**	**46.61**	**86.54**
Dense Fusion w/ LA	88.58	94.73	82.41	87.49	67.70	67.26	46.19	86.49
Mean Boundary IoU (↑)								
Ground truth mask	*86.50*	*91.25*	*83.20*	*87.76*	*76.92*	*80.34*	*72.36*	*85.15*
SAM (No Refinement)	**83.81**	**89.74**	**75.01**	82.69	47.71	57.49	35.47	69.82
Dense Fusion w/ SAM (Ones Map)	83.34	**89.74**	74.86	**82.71**	48.17	57.62	35.32	69.75
Dense Fusion w/ LA (Ones Map)	83.40	89.58	74.47	81.84	**48.92**	**58.04**	**36.71**	**72.31**
Dense Fusion w/ LA	82.40	89.60	74.46	82.21	**48.92**	57.98	36.48	72.28

Table 1: **Comparison of uncertainty-guided prediction refinement** across the `UncertSAM` benchmark. *Ground truth mask* represents an empirical upper bound using the ground truth mask for refinement. *SAM (No Refinement)* refers to the baseline SAM prediction without refinement step. *Dense Fusion w/ SAM* applies dense embedding fusion with baseline SAM prediction maps, whereas *Dense Fusion w/ LA* uses Laplace-based prediction and uncertainty maps. Variants marked *(Ones Map)* explicitly fuse a constant map of ones instead of uncertainty maps. **Best values**, second best.

Uncertainty alignment with error. We measure alignment via the *Pearson correlation coefficient* $\rho(U, E)$, where $E = |P - M|$ denotes the probabilistic model error, *i.e.* the gap between probability map P and ground-truth foreground/background segmentation mask $M \in [0, 1]^{H \times W}$. A larger positive value indicates stronger linear correlation between U and E [Benesty et al., 2009]. Our results in Fig. 3, averaged across pixels and samples for each dataset, find that correlation is generally high across methods ($\rho > 0.5$), indicating good correspondence. The variance network records lowest values, whereas the Laplace Approximation gives strongest alignment. Thus we focus particularly on this approach for subsequent refinement. Similar analysis via the *Brier score* metric [Gneiting and Raftery, 2007] indicated comparable probabilistic accuracy across methods (Fig. 6), and more visuals are given in App. F.

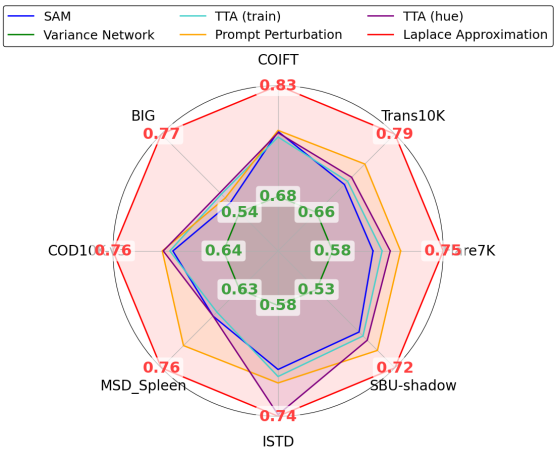


Figure 3: **Pearson correlation coefficient** ρ between error maps and uncertainty maps for each method, averaged across samples per dataset. Higher positive values indicate better correlation, with the LA giving strongest error alignment.

Utility for prediction refinement. We measure predictive performance via two standard segmentation metrics, *mean intersection-over-union* (mIoU) and its boundary-only version (mBIoU) which focuses on pixels exclusively on the mask contours, following Ke et al. [2023]. We benchmark our uncertainty-guided refinement step (*Dense Fusion w/ LA*) against multiple controls to isolate sources of potential gains, detailed further in App. A. We also evaluate SAM directly, and report an additional empirical *performance upper bound* leveraging the ground truth mask for refinement. We observe mixed results across the `UncertSAM` benchmark in Table 1. We find that leveraging the LA's ensemble prediction improves over SAM's baseline predictions, in particular for more challenging domain-shifted datasets (ISTD, SBU, Flare and Trans). However, our dense fusion approach does not improve meaningfully over controls (Ones Map), suggesting our explicit uncertainty handling in its current preliminary form does not contribute consistently to refine predictions.

Conclusion. We observe the recovery of meaningful uncertainty estimates that can benefit predictions, but a simple fusion approach fails to fully leverage this uncertainty signal. In part, this may be due to our *post-hoc* driven strategies which freeze SAM's image encoder, containing most of its expressivity. A deeper fusion of uncertainty into the model architecture, or even the learning process itself, should yield better leverage for subsequent refinement. Nonetheless, we believe such a more holistic perspective on using uncertainty to guide predictions, as opposed to pure per-domain fine-tuning, can offer a more principled path to robust and reliable domain-agnostic segmentation.

References

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 2022.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise Reduction in Speech Processing*, 2009.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *Radiology: Artificial Intelligence*, 2021.

Wei-Ting Chen, Yu-Jiet Vong, Sy-Yen Kuo, Sizhou Ma, and Jian Wang. Robustsam: segment anything robustly on degraded images. *Conference on Computer Vision and Pattern Recognition*, 2024.

Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. *Conference on Computer Vision and Pattern Recognition*, 2020.

Yuekun Dai, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Flare7k: A phenomenological nighttime flare removal dataset. *Advances in Neural Information Processing Systems*, 2022.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 2021.

Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *Advances in Neural Information Processing Systems*, 2025.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 2023.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 2021.

Wei Ji, Jingjing Li, Qi Bi, Tingwei Liu, Wenbo Li, and Li Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. *Machine Intelligence Research*, 2024.

Timo Kaiser, Thomas Norrenbrock, and Bodo Rosenhahn. Uncertainsam: Fast and efficient uncertainty quantification of the segment anything model. *International Conference on Machine Learning*, 2025.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 2023.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *International Conference on Computer Vision*, 2023.

Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. *Winter Conference on Applications of Computer Vision*, 2021.

Kangning Liu, Brian Price, Jason Kuen, Yifei Fan, Zijun Wei, Luis Figueroa, Krzysztof Geras, and Carlos Fernandez-Granda. Uncertainty-aware fine-tuning of segmentation foundation models. *Advances in Neural Information Processing Systems*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019.

David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 1992.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *International Conference on Learning Representations*, 2025.

Tomás F Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. *European Conference on Computer Vision*, 2016.

Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. *Conference on Computer Vision and Pattern Recognition*, 2018.

Yuqing Wang, Yun Zhao, and Linda Petzold. An empirical study on the robustness of the segment anything model (sam). *Pattern Recognition*, 2024.

Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. *International Joint Conferences on Artificial Intelligence*, 2021.

Zhaozhi Xie, Bochen Guan, Weihao Jiang, Muyang Yi, Yue Ding, Hongtao Lu, and Lei Zhang. Pa-sam: Prompt adapter sam for high-quality image segmentation. *International Conference on Multimedia and Expo*, 2024.

Yichi Zhang, Shiyao Hu, Chen Jiang, Yuan Cheng, and Yuan Qi. Segment anything model with uncertainty rectification for auto-prompting medical image segmentation. *arXiv Preprint (arXiv:2311.10529)*, 2023.

Nan Zhou, Ke Zou, Kai Ren, Mengting Luo, Linchao He, Meng Wang, Yidi Chen, Yi Zhang, Hu Chen, and Huazhu Fu. Medsam-u: Uncertainty-guided auto multi-prompt adaptation for reliable medsam. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Towards Integrating Uncertainty for Domain-Agnostic Segmentation
— Supplementary Material —

A Further Experimental Design & Discussion

We measure predictive performance via two standard segmentation metrics, *mean intersection-over-union* (mIoU) and its boundary-only version (mBIoU) which focuses on pixels exclusively on the mask contours. Following Ke et al. [2023] the boundary distance is set dynamically based on image size. All experiments are conducted on NVIDIA H100 GPUs (80GB VRAM) with CUDA 12.6.0.

Control baselines. In order to thoroughly benchmark our uncertainty-guided refinement step (Dense Fusion w/ LA), we compare against two controls to isolate sources of potential gains. To verify if explicit fusion of the uncertainty map is useful, we compare to fusion with an uninformative constant map instead (Dense Fusion w/ LA, Ones Map). To verify if benefits are gained from using the LA's ensemble predictions, we compare to a fusion using constant maps *and* SAM's baseline prediction (Dense Fusion w/ SAM, Ones Map). We also evaluate SAM directly, and report an additional empirical *performance upper bound* leveraging the ground truth mask for refinement.

Results analysis. Our results across the `UncertSAM` benchmark in Table 1 draw mixed conclusions. On one hand, fusing the uncertainty map *does not* improve over its constant control (Ones Map), suggesting that the explicit uncertainty handling using our approach contributes little to refine predictions. On the other hand, leveraging the LA's ensemble prediction *does* improve over SAM's baseline predictions, in particular for more challenging domain-shifted datasets (ISTD, SBU, Flare and Trans)[2]. However, the gap to the ground-truth upper bound across these datasets remains fairly large across the board. In contrast, highly in-domain (but fine-grained) datasets such as BIG and COIFT see strong performance even from baseline SAM. Overall, we observe that meaningful uncertainty estimates are recovered and can certainly benefit predictions, but our simple fusion approach fails to fully leverage this uncertainty signal for explicit prediction refinement.

Discussion. Uncertainty estimates that correlate well with model error are meaningful, and should help the model refine its predictions especially in high-error regions (Fig. 1). Yet, our control experiments suggest that prediction benefits originate primarily from improved ensemble predictions, rather than explicit uncertainty information passed on through dense embedding fusion. Despite modest gains, our prediction refinement step is hindered by two major limitations.

Firstly, our *post-hoc* driven strategies operate by freezing SAM's image encoder, which contains most of its expressivity. Thus, the model may lack capacity to adapt necessary internal representations in response to uncertainty signals. A deeper fusion of uncertainty into the model architecture, or even the learning process itself, should yield better leverage for subsequent refinement. Secondly, our design prioritised domain-agnostic generalisation, and as such any uncertainty fitting or training is done on in-domain training data (from SA-1B [Kirillov et al., 2023]) containing predominantly high-confidence examples. This will have constrained the refinement module's exposure to different lower-confidence uncertainty patterns arising in domain-shifted settings. A more balanced regime, or cross-domain fine-tuning exposure could help the model develop a richer representation of uncertainty, benefitting downstream generalisation.

Moving forward, these directions can be explored to shift from a strictly *post-hoc* perspective toward uncertainty-aware model integration and learning.

B UncertSAM benchmark

Table 2 contains an overview of the datasets used in the `UncertSAM` benchmark. Dataset licenses, where available, are included. Most licenses restrict usage to research purposes only. The SA-1B dataset [Kirillov et al., 2023] has a more restrictive license, and therefore it is not included in our

[2] In part, this may also stem from the introduced fusion layer's ability to re-weigh latent features in a way that benefits generalisation.

Table 2: Datasets overview of the UncertSAM benchmark, including licensing information and pre-processing configurations. The columns indicate pre-processing steps on connected component analysis (CCA), colour-coding (CC), and specific medical CT scan pre-processing.

Dataset	Reference	License	CCA	CC	CT
BIG	Cheng et al. [2020]	Research Only	✓		
COIFT	Liew et al. [2021]	Attribution-NonCommercial 4.0			
COD10K-v3	Fan et al. [2022]	Research Only	✓	✓	
MSD Spleen	Antonelli et al. [2022]	Attribution-ShareAlike 4.0			✓
ISTD	Wang et al. [2018]	Research Only	✓		
SBU	Vicente et al. [2016]	Unknown	✓		
Flare7K	Dai et al. [2022]	S-Lab License 1.0	✓	✓	
Trans10K	Xie et al. [2021]	Research Only	✓		
SA-1B Subset	Kirillov et al. [2023]	SA-1B V1.0			

publicly available dataset. However, original filenames from the randomly sampled subset used in this study are retained by the author. For reproducibility, this subset is available upon request. The preprocessing steps indicated in the table columns are described below:

- **Connected Component Analysis (CCA).** For datasets containing images with multiple disconnected surfaces potentially representing valid masks according to SAM's broad entity-part strategy, we apply CCA to separate masks. We use the OpenCV library to perform the following steps:
  1. Apply morphological closing twice to the initial mask using a 3×3 kernel to reduce the likelihood of generating semantically irrelevant masks when small parts are disconnected but belong to a larger coherent target.
  2. Extract connected components from the closed mask.
  3. Perform a binary AND operation between the resulting mask and the initial mask to remove regions introduced by morphological closing.
  4. Retain connected components larger than 1,000 pixels to eliminate small artifacts.
- **Colour Coded (CC).** When masks are colour coded, we extract unique RGB values and split the mask accordingly into multiple targets.
- **CT scan processing.** The CT images in the MSD Spleen dataset are 3D CT volumes in nii.gz format, sliced along the axial plane to produce 2D images. Only slices containing foreground labels are retained. Normalisation follows the method described in Du et al. [2025]:
  1. Filter the volume to keep only foreground voxels.
  2. Apply Z-score normalisation: $\frac{x-\mu}{\sigma}$.
  3. Clip voxel intensities to the [0.05, 99.95] percentile range.

C Variance Network Architecture

The auxiliary variance network deterministically predicts log variance by adding two components to the SAM mask decoder: a variance prediction head, identical to the existing mask prediction heads, and a lightweight CNN that upsamples spatial embeddings from the image-to-token attention. To reduce checkerboard artifacts in the variance outputs, we replace transpose convolutions in the CNN with bilinear upsampling followed by 2D convolution. Fig. 4 shows the modified mask decoder.

D Dense Embedding Fusion Architecture

We extend SAM's prompt encoder with a parallel uncertainty embedding module that processes a spatial uncertainty map to produce dense features. These are concatenated with the mask prompt embeddings and fused via a 1×1 convolution to retain the original channel dimensionality. This fused representation replaces the original mask embedding, allowing uncertainty to be integrated directly into the internal spatial features used by the decoder. An overview of the architecture is shown in Fig. 5.

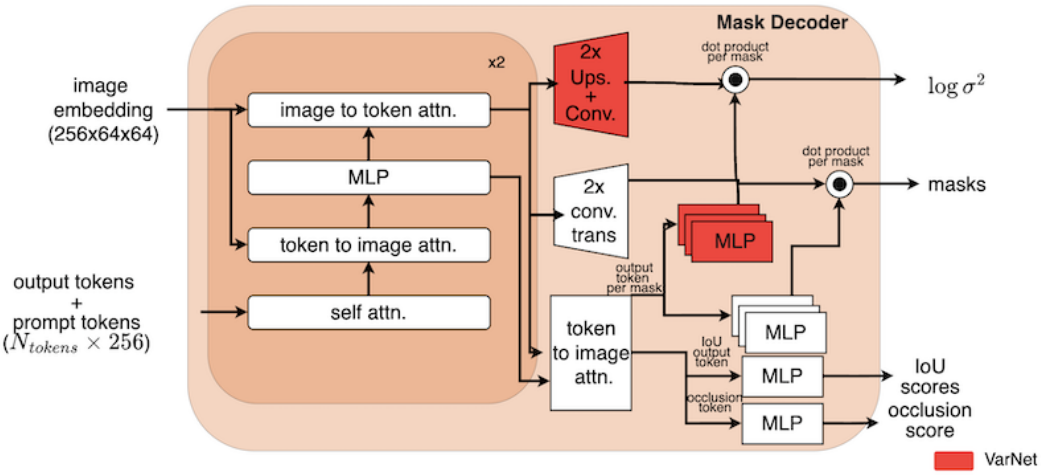


Figure 4: Modified SAM mask decoder architecture with an auxiliary variance network.

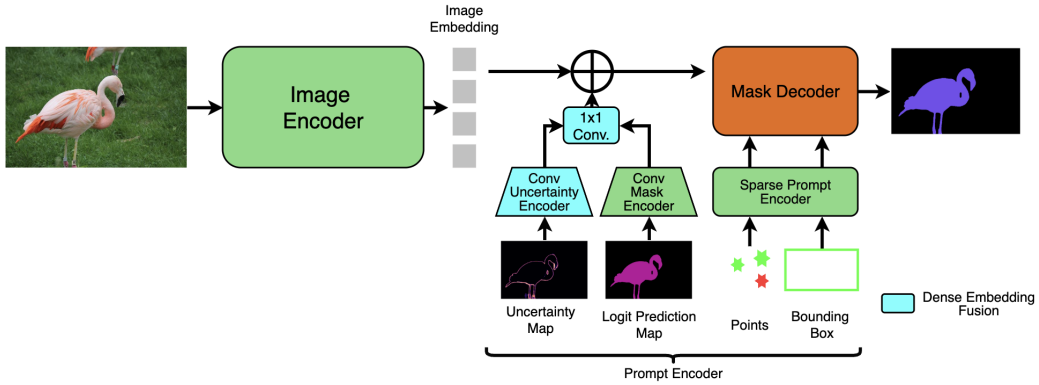


Figure 5: Architecture of the Dense Embedding Fusion Network. The prompt encoder is extended with a CNN-based uncertainty encoder, operating in parallel to the mask prompt encoder used in SAM models. The resulting dense spatial embeddings are concatenated along the channel dimension and fused via a 1×1 convolution layer.

E Training and Hyperparameter Configurations

All experiments are conducted on NVIDIA H100 GPUs (80GB VRAM) with CUDA 12.6.0. All post-hoc fitting and training uses the SA-1B subset to enable a domain-agnostic analysis.

Multi-step training schedule. We follow an 8-step schedule similar to SAM [Ravi et al., 2025]. The first prompt is sampled as a bounding box or a single foreground point with equal probability. Each subsequent step adds one point, sampled uniformly from foreground or background. Unlike the setup of SAM, which places new points in error regions to simulate interactive correction, our uniform sampling targets broad prompt diversity to better characterise uncertainty rather than maximise iterative correction quality.

Fitting the Laplace Approximation. We use the SAM model with frozen weights and fit a diagonal Hessian approximation over the final layer. We fit the LA for one epoch with a batch size of 1. Due to memory constraints, we sample one of the eight prompts per optimisation step to maintain diversity while keeping memory usage low. The loss is computed on predictions and masks downsampled to 128×128 because of memory constraints.

Variance network training. This approach adds a variance head on top of the SAM backbone, which remains frozen. The model is trained using the heteroscedastic loss proposed by Kendall and

Gal [2017], given as

$$\mathcal{L} = \frac{1}{2HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \exp\left(-\log \sigma_{i,j}^2\right) \|M_{i,j} - \sigma(f_\theta(X, q))_{i,j}\|^2 + \log \sigma_{i,j}^2,$$

where $\log \sigma_{i,j}^2$ is the predicted log variance at spatial position (i, j), M is the ground-truth mask, $\sigma(\cdot)$ is the sigmoid function applied element-wise to each spatial element, and $f_\theta(X, q)$ is the output of SAM. Unlike LA, this method uses a single backward pass that combines all eight steps.

Hyperparameter settings. Table 3 summarises the hyperparameters used during training of the modules in this study. Table 4 details the hyperparameters of the data augmentation settings. All training procedures follow the 8-step schedule. Data augmentation is limited to random horizontal flips, mirroring the pre-training setup used in SAM [Ravi et al., 2025]. We also adopt their bounding box perturbation strategy, adding noise up to 10% of box dimensions (max 20 pixels). Optimisation uses AdamW [Loshchilov and Hutter, 2019] with constant learning rate scheduling and precision tailored to stability: bfloat16 for the auxiliary variance network and float32 for Laplace-based setups to avoid overflow/underflow during sampling.

Parameter	**Value**
Multi-Step Training Procedure	
Num. Steps	8
Foreground Probability	0.5
Step 1: point/bbox probability	0.5
Sampling Strategy	Uniform
Augmentations	Rand. HFlip
Max Number of Objects	32
bbox noise level	10% Box dim., <= 20
Optimisation	
Optimiser	AdamW
Learning Rate	1e-4
Schedule	Constant
Weight Decay	1e-4 (VarNet), 0.01
Gradient Clip Norm	0.1
Batch Size (images)	1
Steps	17,500
Warmup steps	195
Cooldown steps	972

Table 3: Overview of training hyperparameters.

Training Augmentations	
Horizontal Flip (hflip)	p: 0.5
Resize	1024 (square)
TTA (Train)	
Color Jitter (1)	b: 0.1, c: 0.03 s: 0.03, h: null
Greyscale	p: 0.05
Color Jitter (2)	b: 0.1, c: 0.05 s: 0.05, h: null
Resize	1024 (square)
TTA (Hue)	
Color Jitter	b: null, c: null s: null, h: 0.5
Resize	1024 (square)

Table 4: Data augmentation sets and corresponding hyperparameters.

F Additional Results

In addition to the correlation analysis in the main text, Fig. 6 provides a comparison of the *Brier score* [Gneiting and Raftery, 2007], a proper scoring rule capturing both calibration and prediction properties. We observe similarly low values across UQ methods, indicating comparable probabilistic accuracy across methods. In addition, Fig. 7 and Fig. 8 provide two qualitative comparisons along the lines of Fig. 1 for the uncertainty maps generated by each of our four considered uncertainty estimation methods.

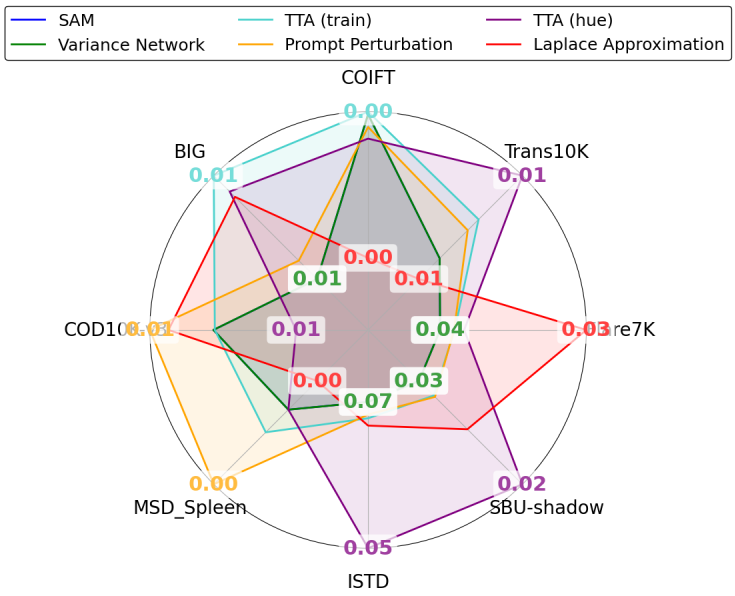


Figure 6: Radar plot illustrating Brier scores of the UQ methods, averaged across samples per dataset. Lower scores indicate better probabilistic performance.

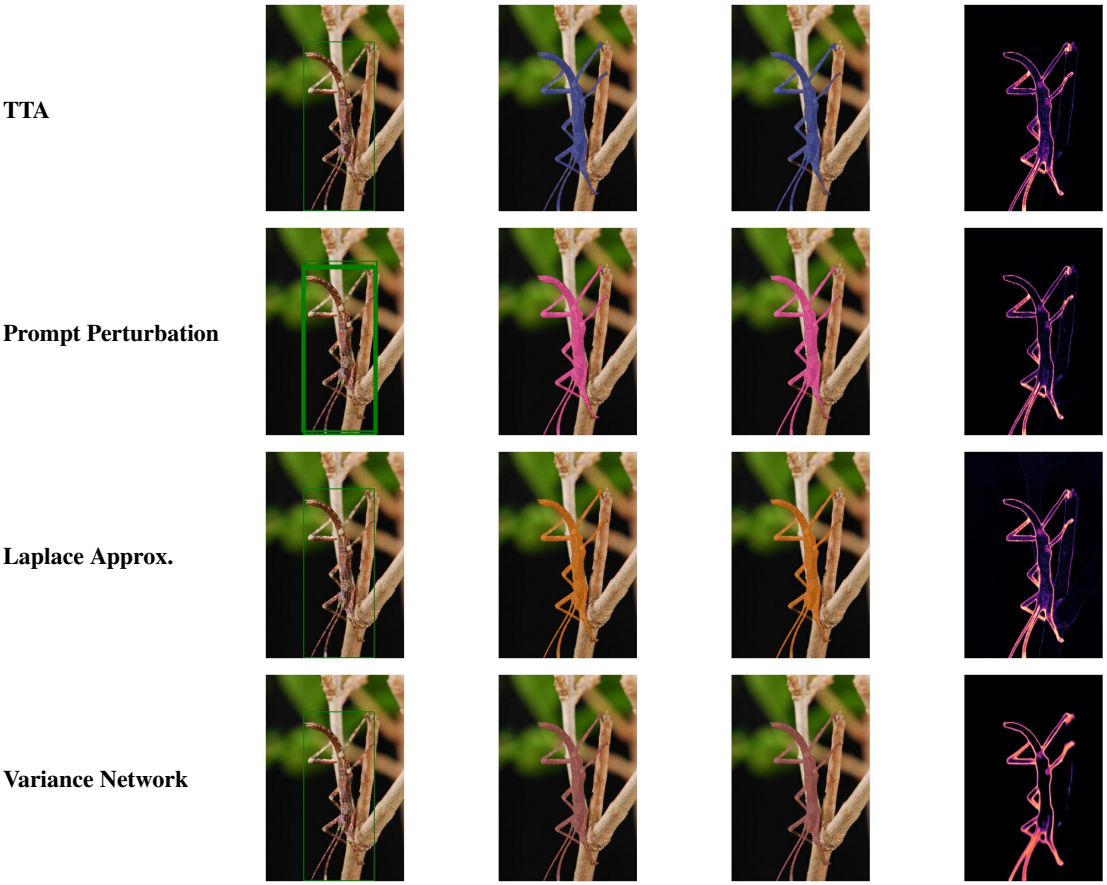


Figure 7: A comparative example from the COD camouflage dataset of the four UQ methods. *From left to right in each row*: (1) Input image with bounding box prompt, (2) Ground truth segmentation mask, (3) Predicted segmentation mask, and (4) Uncertainty estimation mask.

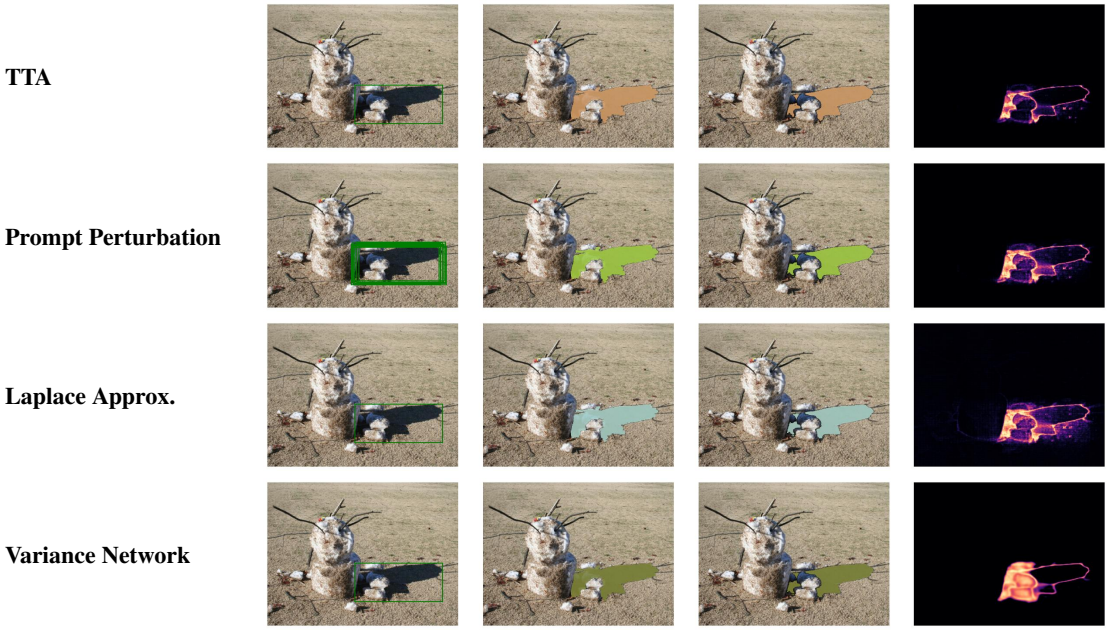


Figure 8: A comparative example from the ISTD shadow dataset of the four UQ methods. *From left to right in each row*: (1) Input image with bounding box prompt, (2) Ground truth segmentation mask, (3) Predicted segmentation mask, and (4) Uncertainty estimation mask.