Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Deterministic and Gaussian Decoder LoRA for Point-Prompted Abdominal CT Segmentation

Aritrik Ghosh

aritrikghosh012@gmail.com

swami vivekananda university https://orcid.org/0009-0003-7016-3315

Research Article

Keywords: abdominal CT, low-rank adaptation, point prompting, SAM2, uncertainty estimation

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11759900/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Deterministic and Gaussian Decoder LoRA for Point-Prompted Abdominal CT Segmentation

Aritrik Ghosh
Department of Computer Science and Engineering
Swami Vivekananda University
Barrackpore, West Bengal, India

Sourav Saha
Department of Computer Science and Engineering
Swami Vivekananda University
Barrackpore, West Bengal, India

***Abstract*—This study examines deterministic low-rank adaptation (LoRA) and Gaussian decoder LoRA with heuristic rank selection for point-prompted abdominal CT segmentation using SAM2. Both approaches were evaluated on a single BTCV split comprising 19 training, five validation, and six test volumes, with one training seed and 2,466 matched positive-point prompts. Across 78 volume–organ pairs, mean Dice scores were 0.5123 for frozen SAM2, 0.8049 for deterministic LoRA, and 0.7982 for Gaussian decoder LoRA. The deterministic model achieved a 0.667-percentage-point advantage under equal volume–organ weighting (descriptive 95% t interval: -0.214 to $+1.549$), although the difference became -0.019 points under equal prompt weighting. We also investigated whether decoder uncertainty provides additional information for identifying segmentation failures. Adding full-image mutual information (MI) to a failure-ranking model containing organ identity, predicted area, and entropy produced a negligible change in mean AUROC (-0.0001). Retrospective inference-only analyses showed similarly limited effects from replacing learned decoder standard deviations with fixed values. Adding dropout MI to organ, area, and entropy on 189 deterministic-model failures changed AUROC from 0.8222 to 0.8238. The Gaussian decoder also required substantially more inference time than its deterministic counterpart. Both adaptation strategies improved overlap relative to frozen SAM2, but the available evidence does not establish a consistent advantage for Gaussian decoder adaptation or its uncertainty estimates beyond simpler alternatives.**

***Index Terms*—abdominal CT, low-rank adaptation, point prompting, SAM2, uncertainty estimation**

I. Introduction

Accurate segmentation of abdominal organs is important for medical image analysis, particularly when organ boundaries are difficult to distinguish from surrounding tissue. Computed tomography (CT) presents several challenges, including overlapping tissue intensities, variation in organ appearance, and anatomical structures that may occupy disconnected regions within individual slices.

Prompt-based segmentation models offer a flexible way to identify structures of interest. Instead of requiring a separate model for each organ, a user can indicate the target region through a point or another spatial prompt. However, a positive point identifies a location within the desired structure without defining its complete boundary. The model must therefore infer the remaining spatial extent from the image.

The Segment Anything Model 2 (SAM2) provides a foundation for prompt-driven segmentation [1]. Although such models offer broad segmentation capabilities, adapting their representations to medical imaging remains important. Low-rank adaptation (LoRA) introduces trainable low-rank weight updates while retaining the pretrained model weights [2].

Existing medical adaptations differ in their training datasets, prompt types, network modifications, and evaluation protocols [3]–[9]. Consequently, results reported under different conditions cannot establish whether stochastic decoder adaptation offers an advantage over deterministic LoRA when both are evaluated using identical point prompts and the same pretrained backbone.

A related question concerns the usefulness of predictive uncertainty. Sampling model parameters can produce multiple segmentation predictions, but additional variability does not necessarily improve segmentation accuracy or the identification of unreliable outputs. It is therefore useful to test whether uncertainty measures provide information beyond simpler predictors, including entropy, organ identity, and predicted mask size.

We compare frozen SAM2, deterministic rank-16 LoRA, and Gaussian decoder LoRA with heuristic rank selection. The Gaussian configuration learns distributions over decoder adaptation factors while retaining deterministic encoder adapters. Its rank-selection controller combines uncertainty-related attribution with a sparsity-regularized gate to choose among predefined ranks. We assess (i) segmentation overlap under matched positive-point prompts and (ii) the additional value and computational cost of stochastic decoder uncertainty for failure ranking. The experiments involve one completed BTCV split and one training seed. Because the test cohort was inspected before later retrospective analyses, the findings are descriptive rather than confirmatory estimates of generalization.

II. Related Work

SAM can return several plausible masks for an ambiguous prompt [3]. Medical adaptations vary in their training data, interaction policy, and dimensionality. MedSAM and MedSAM2 target medical images and sequences [4], [5]. SAMed adapts SAM with low-rank updates, while SAM-Med2D and Medical SAM Adapter use medical-image adaptation strategies [6]–[8]. nnU-Net provides a non-prompted segmentation reference [9]. Their published scores use different protocols from the present benchmark.

AdaLoRA studies importance-driven rank allocation [10]. Gaussian factors follow the variational weight-inference tradition [11]. Laplace-LoRA and BLoB provide distinct Bayesian low-rank approaches [12], [13]. The historical implementation identifier BALoRA v1 refers here to our Gaussian decoder configuration. BaLoRA is also the name of a separate method [14], so we use a descriptive name throughout. MC dropout and independently trained ensembles offer additional uncertainty baselines [15], [16]. Probability calibration and failure ranking address different questions [17]. Temperature scaling is a validation-fit comparator for calibration [18].

III. Materials and Methods

A. Data, prompts, and comparators

We used 30 labeled BTCV CT volumes [19] with volume-disjoint train, validation, and test sets of 19, five, and six. This is the single split identified as F1 in the archived run. Patient linkage was unavailable. Thirteen organs were evaluated as individual binary targets. We clipped CT intensities to $[-175, 250]$ HU, scaled them to $[0, 1]$, and resized slices along the stored third NIfTI axis to 1024×1024. Images used bilinear interpolation and masks used nearest-neighbor interpolation. We repeated grayscale in three channels and applied ImageNet channel normalization. No physical-spacing resampling was applied.

Training, validation, and test contained 1,347, 371, and 460 organ-positive slices. One positive point per present organ was sampled uniformly inside the resized reference mask. The validation and test prompt generators used base seeds 90,043 and 700,043. Each test model used the same coordinates for all 2,466 slice–organ targets. These oracle prompts fix target presence and localization. We requested one output mask from SAM2.1 Hiera Base+. The official builder's stability-based fallback can internally select an alternate mask at evaluation [20]. The frozen comparator used the same points and decoding path without fitting. The two adapted models used fresh adapters on the frozen backbone and prompt encoder.

B. Adapters and rank-selection heuristic

We inserted LoRA adapters into 96 image-encoder linear transforms and 32 mask-decoder transforms of SAM2.1 Hiera Base+. The pretrained backbone and prompt encoder remained frozen. At layer ℓ, a rank-r update is

$$z_\ell = W_\ell x + b_\ell + \frac{16}{r} B_{\ell,r} A_{\ell,r} D(x), \tag{1}$$

where $A_{\ell,r} \in \mathbb{R}^{r \times d_{\text{in}}}$, $B_{\ell,r} \in \mathbb{R}^{d_{\text{out}} \times r}$, and D applies dropout with probability 0.05 during training. Both models use deterministic rank-16 encoder factors. Deterministic LoRA also maintains fixed rank-16 decoder factors; Gaussian decoder LoRA stores rank-16 decoder factors but selects active prefixes at ranks 4, 8, or 16 (Fig. 1).

Each stored Gaussian decoder factor entry has an independent learned posterior $\mathcal{N}(\mu_j, \sigma_j^2)$ and a unit Gaussian prior. The A means use Kaiming-uniform initialization, while the B means start at zero. The initial standard deviation (SD) is 0.05, and log SD is clamped to $[-12, 6]$ before exponentiation. For pixel i, let $p_i^{(k)}$ be the predicted foreground probability from draw k. We compute

$$\bar{p}_i = K^{-1} \sum_{k=1}^{K} p_i^{(k)}, \tag{2}$$

$$u_i = \left[H(\bar{p}_i) - K^{-1} \sum_{k=1}^{K} H(p_i^{(k)}) \right]_+, \tag{3}$$

where H is binary entropy in nats and $[\cdot]_+$ clips negative numerical residuals to zero.

Two full-rank probe predictions provide a mean MI objective J. For each decoder layer, the detached attribution g_ℓ is the mean magnitude of $t_\ell \odot \partial J / \partial t_\ell$, where t_ℓ is the unscaled activation retained from the second probe. The rank-selection score is

$$s_\ell = 0.8 \, \text{sigmoid}\left(4 \left(\frac{g_\ell}{\max_j g_j} - 0.5 \right)\right) + 0.2 \, \text{sigmoid}(\eta_\ell). \tag{4}$$

The attribution term is set to zero if $\max_j g_j \le 10^{-12}$. The bank index is $\text{round}(2 s_\ell)$, corresponding to the three available rank choices. Fresh samples at the selected ranks provide two final draws during training and four at inference.

Learned parameters versus heuristic decisions: decoder-factor means and SDs are fitted through optimization. Gate parameters η_ℓ receive the sparsity gradient and AdamW weight decay. In contrast, the attribution values are detached, and integer rank indexing does not propagate task gradients through the rank choice. Thus, the controller is a heuristic selection rule operating on learned factors, not an independently trained end-to-end differentiable rank policy. Changing rank also changes the multiplier $16/r$ in Eq. (1), coupling active prefix length to update magnitude.

C. Training and decisions

For C present organs on a training slice, we optimize

$$\mathcal{L} = \frac{1}{C} \sum_{c=1}^{C} (\mathcal{L}_{\text{Dice},c} + \mathcal{L}_{\text{BCE},c} + 0.1 \mathcal{L}_{\text{cal},c}) + 10^{-4} \frac{\text{KL}(q \| p)}{2391} + \frac{0.01}{32} \sum_{\ell=1}^{32} \text{sigmoid}(\eta_\ell). \tag{5}$$

The code uses the prefilter dataset length of 2,391 to normalize the KL sum over stored decoder-factor entries. Optimization uses 1,347 organ-positive slices. The coefficient is 4.18×10^{-8} per nat, or 10^{-4} times an inverse-raw-slice coefficient of $1/2391$. This denominator is an inherited implementation convention. The hybrid objective uses Dice, binary cross-entropy (BCE) on the mean probability, and additional penalties. Dice and BCE use the mean probability clamped to $[10^{-6}, 1 - 10^{-6}]$. Soft Dice uses smoothing 10^{-6}. The calibration term weights confidence errors by soft bin mass over 15 triangular bins. Deterministic LoRA uses one probability prediction, the same calibration term, and no KL or rank penalty.

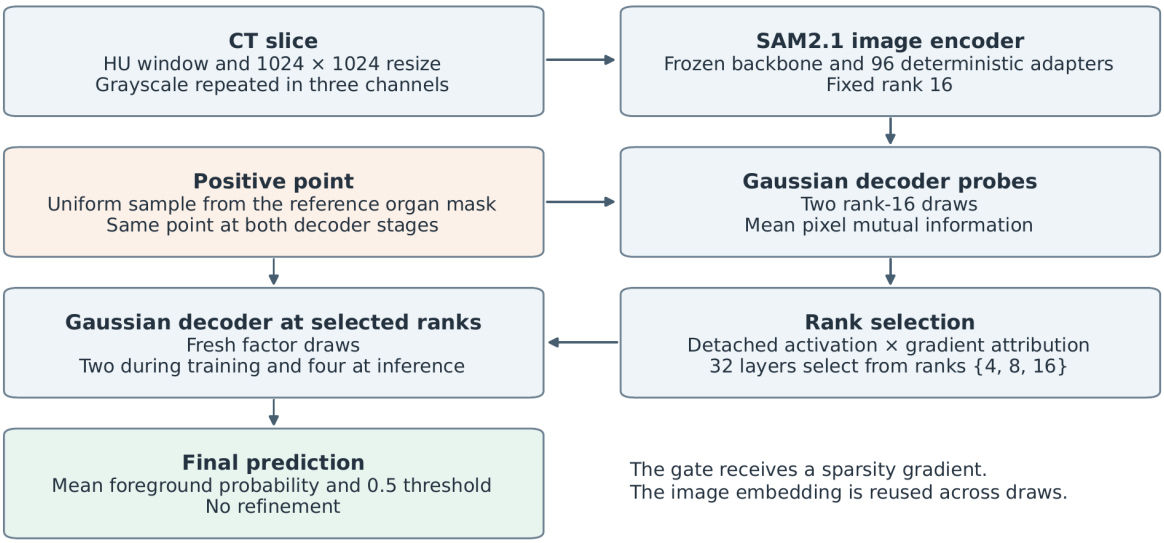


Fig. 1. Gaussian decoder LoRA pipeline. The deterministic comparison uses the same oracle point and a fixed rank-16 decoder.

Both adaptations used seed 42, AdamW with initial learning rate 10^{-4} and weight decay 10^{-5}, gradient norm clipping at 1, slice batch size one, and a 20-epoch cosine schedule. Encoder autocast used FP16 with gradient scaling; decoder and loss computations used FP32. The Gaussian fit used per-block encoder gradient checkpointing. Strict improvements in validation volume–organ mean Dice selected checkpoints. Training used at least eight epochs, at most 20, and patience of six validation evaluations. The primary test policy used a threshold of 0.5 and no refinement. A separate validation audit required guarded refinement to improve paired macro-organ Dice by more than 10^{-6} before selection; it retained no refinement.

D. Endpoints and retrospective analyses

Primary overlap averages Dice across slices within each volume–organ pair, then gives all 78 pairs equal weight. The paired contrast averages 13 organ means per volume, yielding six differences. We report a two-sided paired t interval descriptively, with exhaustive percentile volume bootstrap and exact sign-flip calculations as sensitivity analyses. All 6^6 ordered bootstrap resamples and 2^6 sign assignments were enumerated. These analyses are conditional on the fitted models and six observed volumes. Equal-prompt weighting is also reported. No equivalence or noninferiority margin was specified before the results were observed.

We report IoU, HD95, average symmetric surface distance (ASSD), 15-bin expected calibration error (ECE), Brier score, and binary negative log-likelihood (NLL). Surface distances use resized 2D pixels. The implementation returns infinite distance when only one mask is empty and zero when both are empty. All evaluated reference targets were present, and every prediction was nonempty. The archived reference-foreground diagnostic conditions on positives. The reference-boundary band contains both labels and uses the reference mask. A further shared-region calibration check uses a dilated union of reference and predicted masks for 18 anatomy-selected targets with saved arrays.

Mask failure was defined retrospectively as Dice below 0.5 for a single slice–organ prediction. We computed AUROC within volumes containing both outcomes, then averaged over defined volumes. Mean predictive entropy and Gaussian-decoder MI each average all 1024×1024 pixels; this combines uncertainty intensity with spatial extent and may depend on mask size and boundary length. For each model's own failures, fixed L_2 logistic models ($C = 1$) included organ identity and $\log(1 + \text{predicted area})$, followed by entropy. Gaussian models additionally included MI with and without entropy. Predicted area was recovered from archived confusion statistics and checked against saved masks. Encoding, scaling, and fitting used five volumes for each held-out-volume prediction. Paired AUROC increments were summarized over six volumes by means and descriptive t intervals; overlapping retrospective training sets restrict interval interpretation.

For spatial sensitivity analyses, we reused archived Gaussian uncertainty means on a two-pixel reference-mask boundary band. We repeated the fits using those means and added boundary MI to a full-image entropy model. We also repeated cross-fitting after excluding each volume from all fitting and evaluation. These analyses assess sensitivity to region and cohort composition; the reference band is an oracle region. On Tesla T4 hardware, CUDA-synchronized model-call timing includes one encoder call per slice and one decoder call per target, summed per slice. It excludes image transfers, CPU metrics, export, and other workflow overhead. Original timings include warm-up and represent one recorded run per method. Failure thresholds, weighting sensitivity, and risk-model reanalysis had no prospective registration; checkpoint

TABLE I
SINGLE-SPLIT TEST OVERLAP

Method	Pair Dice	Prompt Dice	Pair IoU
Frozen SAM2	.5123	.62370	.4471
Deterministic LoRA	.8049	.84657	.7216
Gaussian decoder LoRA	.7982	.84676	.7162

TABLE II
PAIR-WEIGHTED SURFACE AND PROBABILITY METRICS

Metric	Frozen	Deterministic	Gaussian
HD95 (px)	180.22	34.30	35.67
ASSD (px)	100.77	11.74	12.34
All-pixel ECE	.11227	.00424	.00449
All-pixel Brier	.09554	.00412	.00403
All-pixel NLL	.37233	.02708	.02686
Reference-boundary ECE	.34532	.32902	.31106

selection and the refinement audit used validation data.

E. Inference-only replay

We replayed selected seed-42 checkpoints on the same prompts and test volumes without refitting segmentation models. For frozen SAM2, the original single-output path was compared with three candidate masks selected by predicted IoU or by highest reference Dice; another oracle also considered the original single output. The oracles used references only for candidate scoring and selection. For deterministic LoRA, we enabled decoder-adapter dropout at $p = 0.05$ and averaged four predictive draws. For Gaussian LoRA, we replaced learned decoder SDs with 0.05 and reran rank selection and four final draws. We separately logged stability-fallback substitutions for rank probes and final draws.

Dropout MI was evaluated on deterministic single-output failure labels. Full-image entropy and dropout MI were compared as standalone scores, then dropout MI was added to a leave-one-volume-out model with organ identity, deterministic single-output area, and deterministic full-image entropy. We also ranked learned-SD failures using MI from both Gaussian replays. Predicted-band analyses used a two-pixel dilation–erosion band around each predicted mask and added log-transformed perimeter to the failure models. A matched encoder benchmark alternated model order after three warm-up calls per model, using two slices per volume and three timed repeats per slice.

IV. RESULTS

A. Checkpoint selection and overlap

Both fits completed 13 epochs and selected epoch 7. Validation Dice for deterministic LoRA was 0.7808 at epoch 4 and 0.8134 at epoch 7 (Fig. 2). The 3.26-point span exceeds the test difference, illustrating checkpoint sensitivity rather than variability across independent refits. The archived total training objectives contain different terms, so Fig. 2 compares validation Dice only.

TABLE III
LOW-DICE TARGETS BY VOLUME AND METHOD

Volume	Targets	Frozen	Deterministic	Gaussian
img0009	715	232	56	65
img0010	493	174	45	43
img0026	337	144	28	33
img0028	295	89	17	17
img0034	293	76	16	16
img0038	333	122	27	26
Total	2466	837	189	200

Table I shows improvements of 29.25 and 28.59 Dice points over frozen SAM2 for deterministic and Gaussian decoder LoRA. Deterministic LoRA had a 0.667-point advantage under equal pair weighting. The descriptive t interval was $[-0.214, 1.549]$ points, the percentile bootstrap interval was $[0.075, 1.306]$, and the two-sided exact sign-flip result was $6/64 = 0.09375$. The disagreement among interval methods matters with six volumes. Leaving out one volume gave differences from 0.396 to 0.905 points. Equal weighting of the 2,466 prompts gave a -0.019-point contrast because organs and volumes contributed different numbers of targets. These findings do not identify equivalence or a stable ranking across training runs.

Deterministic LoRA had higher Dice in five test volumes. Gaussian decoder LoRA led on five organ means, including gallbladder ($+2.91$ points) and inferior vena cava ($+2.14$). The largest deficit was for right adrenal gland (-5.37 points). Each organ comparison is based on six volumes and is hypothesis-generating.

B. Calibration, surface distances, and refinement

All-pixel ECE and surface distances were slightly lower for deterministic LoRA, whereas Gaussian decoder LoRA had slightly lower all-pixel Brier score and NLL (Table II). All-pixel summaries include extensive background. The boundary-band ECE conditions on reference boundaries. On 18 anatomy-selected targets, shared-region ECE was 0.0496 for deterministic LoRA and 0.0580 for Gaussian decoder LoRA.

The archived refinement comparison used five validation volumes, 371 slices, and 1,993 targets. Macro-organ Dice was 0.8034 without refinement, 0.6183 with legacy refinement, and 0.7968 with guarded refinement. These comparisons used the same validation weights. The primary test policy remained the recorded no-refinement configuration.

C. Failure ranking

Each method–volume combination included both failure outcomes (Table III). Each mean AUROC uses six volumes; Gaussian LoRA had 16–65 failures per volume. For deterministic-model failures, entropy alone reached 0.6983, organ and area reached 0.7769, and adding entropy reached 0.8222. Frozen-model entropy reached 0.8515 for its own failure labels. Failure labels differ by model.

For Gaussian-model failures, organ and area reached AUROC 0.7783; adding MI increased it to 0.8133. The paired mean

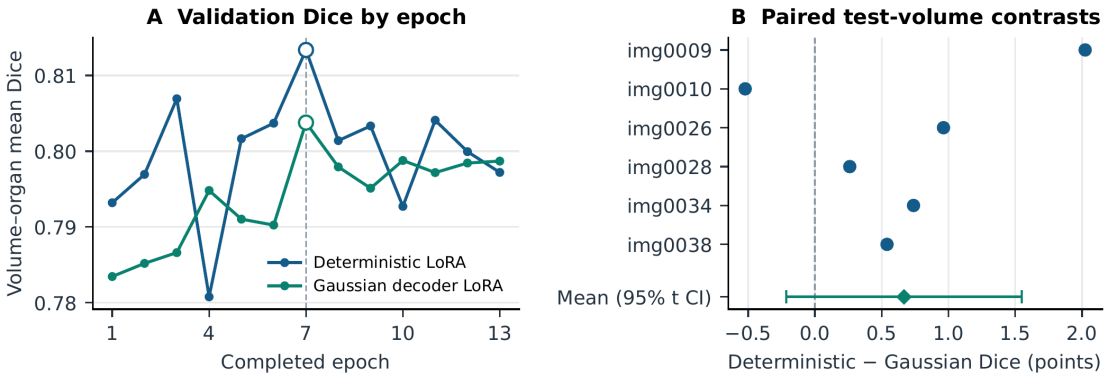


Fig. 2. Verified validation histories and test-volume Dice contrasts. Open circles mark the selected epoch. The diamond and interval summarize six paired volume contrasts; positive differences favor deterministic LoRA.

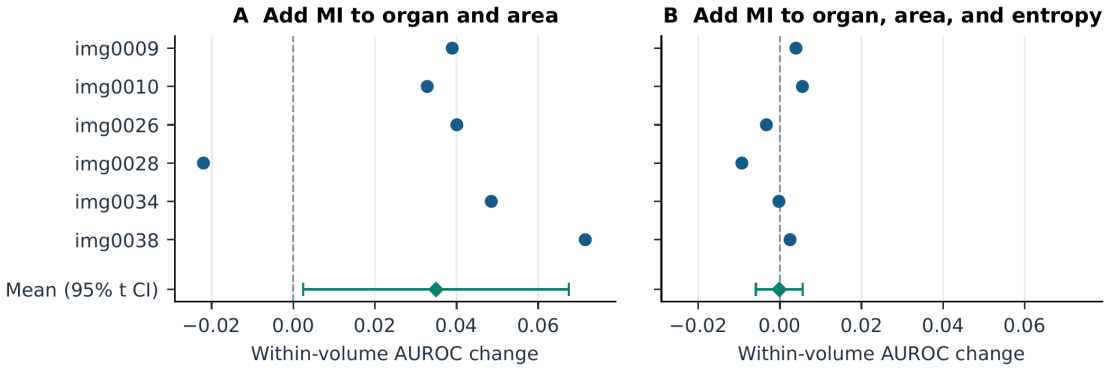


Fig. 3. Paired AUROC increments after adding full-image mean MI to cross-fitted failure models. Both panels use the same 200 Gaussian-model failures. Dots denote held-out volumes; diamonds and bars show means and descriptive 95% t intervals over six volumes with overlapping fit sets.

increment was 0.0350 (descriptive 95% t interval, 0.0024 to 0.0675), with volume differences from -0.0220 to $+0.0715$. Organ, area, and entropy reached 0.82589, and adding MI yielded 0.82575. The increment from unrounded values was -0.00014 (descriptive interval, -0.00589 to $+0.00561$). Excluding one volume and refitting gave mean increments from -0.00576 to $+0.00282$. Small failure counts and overlapping training sets restrict inference (Fig. 3). MI alone reached 0.7974 versus 0.7464 for entropy alone, a mean difference of 0.0510 (0.0261 to 0.0759).

Adding reference-boundary MI to reference-boundary entropy increased AUROC from 0.7736 to 0.8372; adding it to the stronger full-image entropy model changed AUROC from 0.8259 to 0.8229. Archived pixel-error AUROCs for Gaussian entropy and MI were 0.8576 and 0.7863 across all pixels, and 0.5987 and 0.5780 in the reference-boundary band. These histogram estimates used 256 equal-width bins on $[0, \log 2]$ nats, introducing score ties. Four draws produce a finite-sample MI estimate; sensitivity to draw count was not measured.

D. Rank, factor scales, and timing

Across 78,912 decoder-layer decisions, ranks 4, 8, and 16 accounted for 50.14%, 46.14%, and 3.72%, respectively. Mean active rank was 6.29; every prompt used rank 16 in at least one layer (mean 1.19 layers). Within-prompt maximum normalization discards absolute attribution scale. The controller redistributes relative ranks, changes the $16/r$ update multiplier as well as the active prefix, and stores both factor tensors at rank 16.

The selected checkpoint stores 335,872 scale entries across 32 decoder layers. The 5th percentile, median, and 95th percentile were 0.0494, 0.0739, and 0.0990, respectively (initialization: 0.05). Root-mean-square SD divided by root-mean-square factor mean ranged from 1.67 to 10.37 across layers, with median 5.02. Factorwise medians were 3.34 for A and 11.78 for B; these ratios depend on factorization. Checkpoint KL was approximately 7.27×10^5 nats before weighting and 0.0304 afterward.

Table IV reports mean encode, organ-prompt decode, and total model-call time per slice. Gaussian decoder time was 8.70 times that of deterministic LoRA, and total per-slice model-call

TABLE IV
OBSERVED MODEL-CALL TIMES

Method	Enc. (ms)	Dec. (ms)	Sum (s/slice)
Frozen SAM2	115.89	15.23	0.198
Deterministic LoRA	156.22	23.40	0.282
Gaussian decoder LoRA	142.78	203.53	1.234

time was 4.38 times as high across 460 slices (mean 5.36 targets per slice). Gaussian decoding includes two probes, attribution, and four final draws. The deterministic path uses live adapters. In the matched 12-slice replay, mean per-slice encoder medians were 145.88 ms for deterministic and 145.92 ms for Gaussian LoRA (paired difference -0.038 ms), so the original encoder gap did not recur.

E. Inference-only comparisons

The replay reproduced all 7,398 archived single-output and learned-SD Dice scores exactly. Predicted-IoU selection among frozen candidates gave Dice 0.5099, versus 0.5123 for the single-output path. Reference selection among the same three candidates reached 0.6246, and the combined oracle including the original output reached 0.6277. Stability fallback substituted 1,130/2,466 frozen outputs (45.8%), 48/2,466 deterministic outputs (1.9%), and 242/9,864 Gaussian learned-SD final draws (2.5%).

Decoder MC dropout gave Dice 0.8053, versus 0.8049 for deterministic single output (paired difference $+0.043$ percentage points; descriptive 95% interval, -0.038 to $+0.123$). On 189 deterministic single-output failures, dropout MI alone reached AUROC 0.8037 and entropy alone reached 0.6983 (paired increment $+0.1054$, interval $+0.0339$ to $+0.1769$). The organ, area, and entropy model reached 0.8222. Adding dropout MI increased AUROC to 0.8238 (increment $+0.0016$; descriptive 95% t interval, -0.0060 to $+0.0092$).

Fixed-SD Dice was 0.7985, versus 0.7982 with learned SDs (paired difference $+0.033$ percentage points; interval -0.256 to $+0.322$). On the same 200 learned-SD failures, AUROC was 0.7953 for fixed-SD MI versus 0.7974 for learned-SD MI (increment -0.0021; interval -0.0203 to $+0.0161$). Rank-4 assignments rose from 39,566 to 51,978 of 78,912 decisions in the fixed-SD replay.

On learned-SD predictions, the organ, area, perimeter, and predicted-band entropy model reached AUROC 0.8785. Adding predicted-band MI changed this to 0.8764 (increment -0.0021). With full-image entropy and perimeter, the MI increment was less than 0.0001. All analyses used six held-out volumes from the same test cohort.

V. DISCUSSION

Both deterministic and Gaussian decoder LoRA substantially improved segmentation overlap relative to frozen SAM2 under identical reference-mask-derived positive points. Mean volume–organ Dice increased from 0.5123 to 0.8049 and 0.7982, respectively. This supports the value of adapting the pretrained model under the evaluated prompting protocol, but does not establish generalization to unrestricted clinical prompting.

The difference between the two adapted models was much smaller. Deterministic LoRA performed slightly better under equal volume–organ weighting, but the paired descriptive interval included zero, and equal-prompt weighting yielded a small contrast in the opposite direction. The ranking is therefore sensitive to the analysis unit and unresolved across training runs. Moreover, the Gaussian approach changes factor sampling, adaptive rank selection, KL regularization, sparsity penalty, and rank-dependent scaling together. The present comparison cannot isolate those components. In particular, selection changes both the active factor prefix and update magnitude.

The Gaussian controller selected low active ranks on average but did not reduce observed inference cost. Recorded decoder calls were approximately 8.70 times as expensive as deterministic calls, reflecting additional uncertainty probes, attribution computation, and final draws. A matched encoder-timing replay did not reproduce the original encoder difference; the strongest efficiency conclusion concerns decoder workload rather than a systematic encoder advantage.

The uncertainty results are similarly qualified. Gaussian MI improved failure ranking beyond organ identity and predicted area alone, but its incremental benefit was negligible once entropy was included. Deterministic dropout MI outperformed entropy as a standalone failure score, yet its contribution was small after adding organ, area, and entropy features. Thus, the value of MI depends on what information the failure-ranking model already contains. A direct comparison of Gaussian and dropout MI would require common failure labels and matched baseline features.

Calibration conclusions depend on the evaluation region. All-pixel metrics include extensive background; reference-boundary assessments provide complementary information but require reference labels and are not deployment-time procedures. Both adaptations shared a calibration-related loss, while validation-fit temperature scaling remains untested. Reference-assisted selection among frozen SAM2 candidates also provides an oracle bound, not a practical mask-selection policy. Replacing Gaussian decoder SDs with fixed values altered inference-time rank selection; identifying the effect of fixed-SD *training* would require refitting.

The study has important limitations: six test volumes, one split, one seed, oracle point locations, and retrospective uncertainty analyses on an inspected cohort. Patient linkage was unavailable. Paired refits across seeds and volume-disjoint splits, an untouched external cohort, and ablations separating Gaussian sampling, rank selection, KL regularization, and $16/r$ scaling are needed. Matched comparisons with medical segmentation baselines would further clarify practical value.

VI. CONCLUSION

Deterministic and Gaussian decoder LoRA improved point-prompted abdominal CT segmentation overlap relative to frozen SAM2 in this single-split BTCV study. Deterministic LoRA had slightly higher mean volume–organ Dice, but the difference

was small, sensitive to weighting, and not established across seeds or cohorts. Gaussian decoder LoRA incurred greater inference cost, and its mutual-information estimates added little to failure-ranking models already containing entropy and mask-derived features. Retrospective fixed-SD inference also had little effect on mean Dice. These results do not rule out a benefit from Gaussian parameter uncertainty; rather, they identify questions that require independent replication and component-level ablations.

Data, Code, and Declarations

Dataset, ethics, and permissions. This study is a secondary analysis of 30 previously collected, de-identified abdominal CT volumes and reference masks from the public BTCV challenge dataset (Synapse project syn3193805) [19], used under its data-use terms. No new patient data were collected and no human participants were recruited, so institutional ethics approval was not required. CT-derived images in this paper are reproduced under the dataset license.

Reproducibility. The companion package contains an archive inventory, scalar predictions, split manifests, analysis scripts, factor-scale summaries, and original figures. The 54 saved dense records cover 18 qualitative targets for three methods. The SAM2 source commit 2b90b9f5ceec907a1c18123530e92e794ad901a4 was recorded in the run configuration.

Funding and competing interests. The authors received no specific funding for this work and declare no competing interests.

Author contributions. A.G. implemented the models, ran the experiments and analyses, and wrote the first draft. S.S. supervised the work and reviewed the manuscript.

Release. The companion package is available from the authors on request. Fitted adapter weights are not released.

References

[1] N. Ravi *et al.*, "SAM 2: Segment anything in images and videos," arXiv:2408.00714, 2024.

[2] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[3] A. Kirillov *et al.*, "Segment anything," in *Proc. ICCV*, 2023, pp. 4015–4026.

[4] J. Ma *et al.*, "Segment anything in medical images," *Nature Commun.*, vol. 15, Art. no. 654, 2024, doi: 10.1038/s41467-024-44824-z.

[5] J. Ma *et al.*, "MedSAM2: Segment anything in 3D medical images and videos," arXiv:2504.03600, 2025.

[6] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," arXiv:2304.13785, 2023.

[7] J. Cheng *et al.*, "SAM-Med2D," arXiv:2308.16184, 2023.

[8] J. Wu *et al.*, "Medical SAM Adapter: Adapting Segment Anything Model for medical image segmentation," arXiv:2304.12620, 2023.

[9] F. Isensee *et al.*, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2021, doi: 10.1038/s41592-020-01008-z.

[10] Q. Zhang *et al.*, "Adaptive budget allocation for parameter-efficient fine-tuning," in *Proc. ICLR*, 2023.

[11] C. Blundell *et al.*, "Weight uncertainty in neural network," in *Proc. ICML*, 2015, pp. 1613–1622.

[12] A. X. Yang, M. Robeyns, X. Wang, and L. Aitchison, "Bayesian low-rank adaptation for large language models," in *Proc. ICLR*, 2024.

[13] Y. Wang *et al.*, "BLoB: Bayesian low-rank adaptation by back-propagation for large language models," in *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 67758–67794, 2024.

[14] D. Coscia, S. Löwe, and M. Welling, "BaLoRA: Bayesian Low-Rank Adaptation of Large Scale Models," arXiv:2605.08110, 2026. [Online]. Available: https://arxiv.org/abs/2605.08110.

[15] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. ICML*, 2016, pp. 1050–1059.

[16] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[17] A. Mehrtash *et al.*, "Confidence calibration and predictive uncertainty estimation for deep medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 3868–3878, 2020, doi: 10.1109/TMI.2020.3006437.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. ICML*, 2017, pp. 1321–1330.

[19] "Multi-Atlas Labeling Beyond the Cranial Vault: Workshop and Challenge," MICCAI, 2015. Synapse, project syn3193805. [Online]. Available: https://www.synapse.org/Synapse:syn3193805/wiki/89480.

[20] Meta AI, "SAM 2 official implementation and checkpoints." [Online]. Available: https://github.com/facebookresearch/sam2.

SUPPLEMENTARY QUALITATIVE CT COMPARISONS

Four selected BTCV test targets from the archived BALoRA CT Case Explorer.
Mint: reference contour; amber: predicted contour; pale yellow cross: shared positive prompt.

Fig. S1. Liver (img0009, slice 116; prompt (607, 558)).

Reference	Frozen SAM2	Deterministic LoRA	Gaussian Decoder LoRA
			
Ground-truth mask	Dice 0.6723 | IoU 0.5064	Dice 0.9653 | IoU 0.9329	Dice 0.9719 | IoU 0.9453

Fig. S2. Pancreas (img0009, slice 63; prompt (600, 617)).

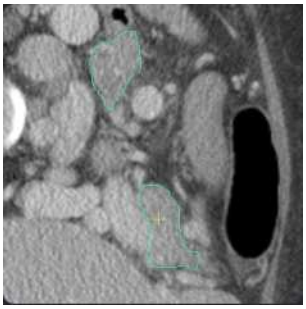
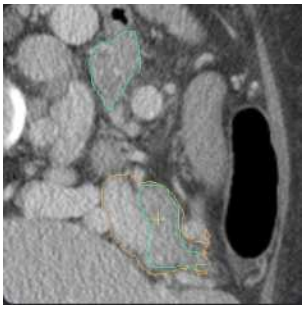
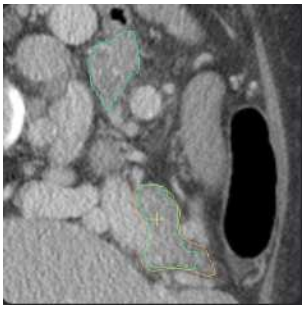
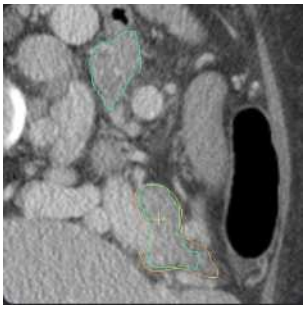
Reference	Frozen SAM2	Deterministic LoRA	Gaussian Decoder LoRA
			
Ground-truth mask	Dice 0.4802 | IoU 0.3160	Dice 0.6167 | IoU 0.4458	Dice 0.6094 | IoU 0.4382

Fig. S3. Right adrenal gland (img0009, slice 97; prompt (338, 380)).

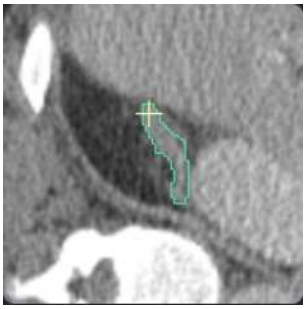
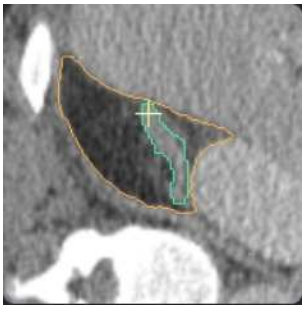
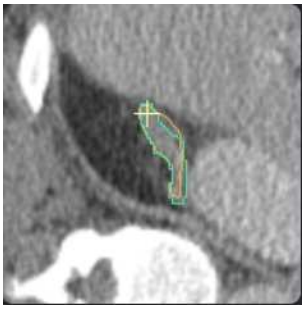
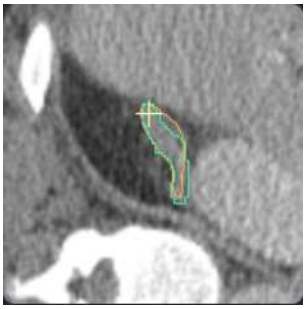
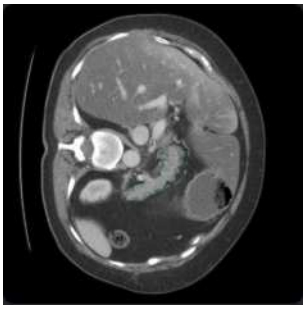
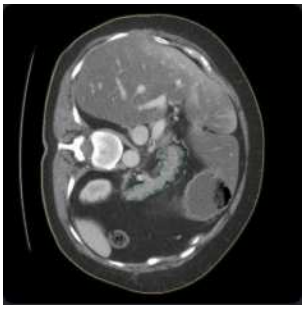
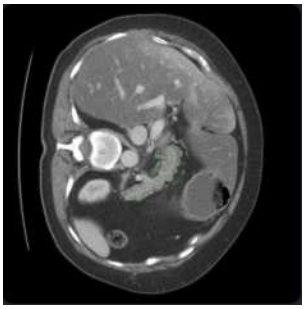
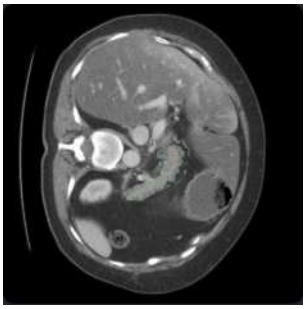
Reference	Frozen SAM2	Deterministic LoRA	Gaussian Decoder LoRA
			
Ground-truth mask	Dice 0.2662 | IoU 0.1535	Dice 0.8199 | IoU 0.6948	Dice 0.8149 | IoU 0.6876

Fig. S4. Pancreas (img0028, slice 64; prompt (419, 641)).

Reference	Frozen SAM2	Deterministic LoRA	Gaussian Decoder LoRA
			
Ground-truth mask	Dice 0.0438 | IoU 0.0224	Dice 0.8783 | IoU 0.7829	Dice 0.8778 | IoU 0.7821

Selected illustrative slice-organ cases, not independent experiments or additional test cohorts.
Results are per-prompt Dice/IoU, distinct from volume-organ means in Table I of the paper.