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.

Selective Foundation-Model Intervention for Vertebral CT Segmentationunder Predictive Uncertainty

Xinyu Liu
Sichuan University
Xinlu Lv
Sichuan University
Yihan Liu
Sichuan University
Bingran Wang
Sichuan University
Wenchao Du
wenchaodu.cs@scu.edu.cn
Sichuan University

Article

Keywords: medical image segmentation, foundation models, selective inference, predictive uncertainty, Segment Anything Model, vertebral CT segmentation

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11113234/v1

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

Additional Declarations: No competing interests reported.

Selective Foundation-Model Intervention for Vertebral CT Segmentation under Predictive Uncertainty

Xinyu Liu[1], Xinlu Lv[1], Yihan Liu[1], Bingran Wang[2], Wenchao Du[1*]

[1*] College of Computer Science, Sichuan University, Chengdu, Sichuan, China .

[2] West China School of Stomatology, Sichuan University, Chengdu, Sichuan, China .

*Corresponding author(s). E-mail(s): wenchaodu.cs@scu.edu.cn;

Abstract

Foundation segmentation models offer strong generalization capabilities, but their intervention is not necessarily beneficial when a task-specific model already provides reliable predictions. We propose an uncertainty-guided selective collaboration framework (UGSCF) for vertebral CT segmentation, in which the Segment Anything Model (SAM) acts as an on-demand correction proposer rather than a mandatory second-stage segmenter. UGSCF follows a route–propose–verify–correct strategy: a task-specific UNet first generates the default segmentation and predictive uncertainty; uncertainty determines where SAM intervention should be allocated; SAM-derived proposals are restricted to uncertain disagreement regions; and a candidate-value verifier determines whether a proposed correction should modify the original prediction.

Experiments on VerSe19 showed that unconditional SAM intervention was frequently harmful, with direct SAM output degrading 91.66% of eligible validation slices. Predictive uncertainty nevertheless provided useful allocation information, with the fraction of beneficial interventions increasing from 2.04% in the lowest-uncertainty quartile to 30.49% in the highest. Under nested subject-level validation, the best deployable candidate-value policy achieved a mean volumetric Dice of 0.814658, compared with 0.814139 for UNet alone. SAM was invoked for 19.53% of foreground-positive validation slices, while only 9.49% received an accepted correction, of which 65.62% were beneficial. The validation gain was small and not statistically supported (95% CI: $\mathbf{-0.000312}$ to $\mathbf{0.001493}$; $\boldsymbol{p = 0.307821}$). In a post-hoc 40-subject test-set evaluation, UGSCF

achieved a mean volumetric Dice of 0.825360 compared with 0.825776 for UNet ($\mathbf{\Delta = -0.000416}$; $\boldsymbol{p = 0.265373}$).
These results indicate that collaboration between task-specific and foundation segmentation models is better formulated as a selective decision process than as unconditional model fusion. Rather than establishing a consistent improvement in aggregate segmentation accuracy, UGSCF provides a hierarchical mechanism that separates where additional foundation-model computation is allocated from whether the resulting proposal is trusted.

Keywords: medical image segmentation, foundation models, selective inference, predictive uncertainty, Segment Anything Model, vertebral CT segmentation

1 Introduction

Accurate vertebral segmentation from computed tomography (CT) images is an important component of quantitative spine analysis, supporting tasks such as vertebral localization, morphological assessment, and the analysis of spinal abnormalities. The release of large-scale benchmarks such as VerSe has facilitated the development and systematic evaluation of automated vertebral segmentation methods across heterogeneous multi-detector CT scans [1]. Deep convolutional architectures, particularly UNet and its derivatives, have become a standard paradigm for biomedical image segmentation [2]. More recently, highly optimized task-specific frameworks such as nnU-Net have further demonstrated that carefully configured supervised segmentation models can achieve strong performance across diverse biomedical imaging tasks [3].

In parallel with the development of task-specific segmentation models, large-scale vision foundation models have introduced a different paradigm for image segmentation. The Segment Anything Model (SAM), trained on a large collection of natural images and masks, provides a promptable segmentation interface capable of transferring to previously unseen objects and image distributions [4]. Its flexible use of point and bounding-box prompts has motivated extensive interest in medical image analysis, where foundation-model priors may complement models trained for individual anatomical structures or imaging modalities. Medical adaptations such as MedSAM have further demonstrated the potential of large-scale promptable segmentation models across multiple medical imaging tasks [5].

However, the availability of a powerful foundation model does not imply that its prediction should always replace or refine that of a task-specific model. Unlike natural images, medical images frequently contain small structures, weak boundaries, heterogeneous acquisition conditions, and substantial domain-specific appearance variation. Experimental evaluations of SAM across medical imaging datasets have shown considerable task-dependent variability, with performance strongly affected by anatomical target, imaging modality, and prompt design [6]. Consequently, when a task-specific model already produces a reliable segmentation, an additional SAM prediction may

either correct an existing error or introduce a new one. This creates a decision problem that is distinct from conventional model fusion: *when should a foundation model be invoked, and when should its proposed correction be trusted?*

Existing efforts to adapt foundation segmentation models to medical images have primarily focused on improving their medical-domain adaptation, prompt generation, or computational efficiency [5, 7]. These directions improve the capability of the foundation model itself, but a complementary question arises when the foundation model is used together with an already strong task-specific segmenter. In this setting, unconditional intervention can be unnecessary or even harmful, while invoking the foundation model on every image also introduces additional computational cost. A practical collaborative system therefore requires not only a mechanism for generating alternative segmentation proposals, but also explicit control over the allocation and acceptance of such interventions.

Predictive uncertainty provides a natural signal for the first of these decisions. Regions or images associated with greater uncertainty indicate where the task-specific model is less confident and where additional model capacity may be more informative. Nevertheless, uncertainty alone cannot determine whether a particular alternative segmentation is actually superior. A highly uncertain prediction may still be preferable to the proposal generated by a foundation model. We therefore distinguish between two complementary decisions: *intervention routing*, which determines where foundation-model computation should be allocated, and *proposal verification*, which determines whether the resulting candidate should modify the original task-specific prediction.

Based on this distinction, we propose an uncertainty-guided selective collaboration framework, termed UGSCF, for vertebral CT segmentation. Rather than treating SAM as a mandatory second-stage segmenter, UGSCF follows a hierarchical *route–propose–verify–correct* procedure. A task-specific UNet first produces the default segmentation and its predictive uncertainty. Uncertainty is then used to route a subset of inputs for additional SAM intervention. SAM acts as an on-demand correction proposer, and its outputs are restricted to localized uncertain regions in which the two models disagree. Finally, a candidate-value verifier estimates whether the proposed correction is likely to improve the original prediction and rejects interventions whose estimated value is insufficient.

The main contributions of this study are threefold. First, we systematically characterize the failure of unconditional SAM intervention when applied to predictions from a strong task-specific vertebral CT segmenter, demonstrating that additional foundation-model capacity does not necessarily translate into improved segmentation. Second, we introduce a hierarchical selective-intervention formulation that explicitly separates uncertainty-guided computational allocation from candidate-level correction acceptance. Third, we develop and evaluate a candidate-value verification mechanism that preserves the task-specific prediction when a SAM-derived correction is not expected to be beneficial, providing a practical framework for selective collaboration between task-specific and foundation segmentation models.

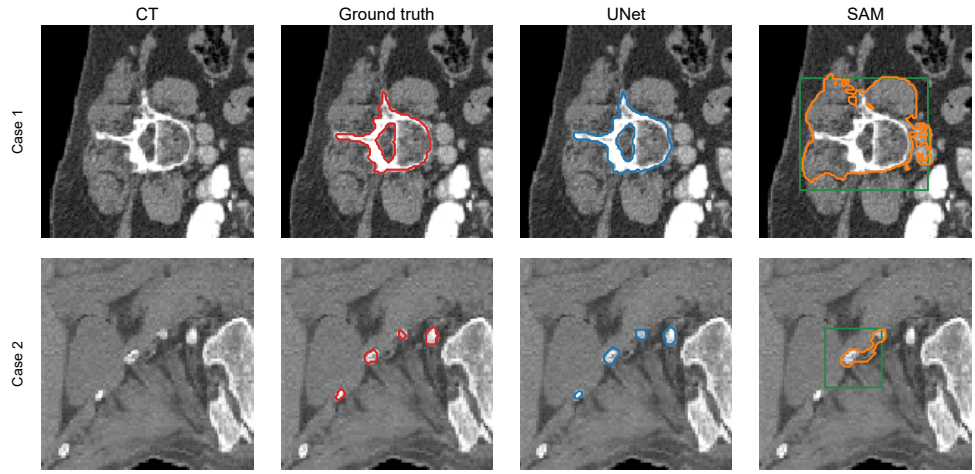


Fig. 1 Representative failure cases of SAM intervention on the internal validation set. Red, blue, and orange contours denote the ground truth, UNet prediction, and SAM prediction, respectively; green boxes indicate the automatically generated SAM prompts.

2 Results

2.1 Task-specific segmentation establishes a strong baseline

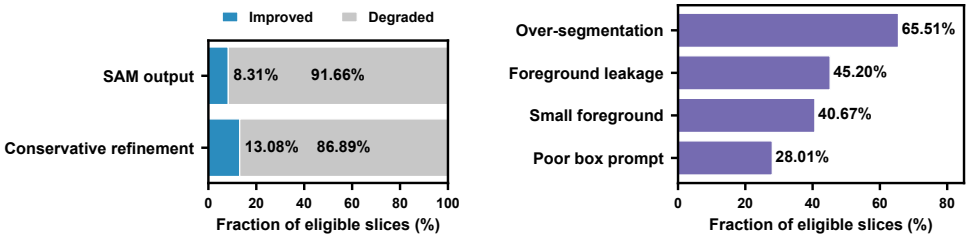
The VerSe19 dataset was partitioned at the subject level into 65 training, 15 internal-validation, and 40 test subjects. The test subjects were excluded from UNet training and checkpoint selection, while the internal-validation subjects were used to develop the uncertainty-routing and candidate-value verification components.

The ResNet18-encoder UNet achieved a mean volumetric Dice of 0.8258 ± 0.0685, an HD95 of 16.02 ± 28.91, and an ASD of 3.28 ± 7.13 on the 40-subject test set. Metrics were computed after reconstructing slice predictions into subject-level three-dimensional volumes. This strong task-specific baseline motivated treating SAM as a selective correction proposer rather than a mandatory post-processing component.

2.2 Unconditional SAM intervention frequently degrades segmentation

We first examined whether SAM could reliably refine the UNet predictions. Among 3,339 foreground-positive internal-validation slices, 3,334 were eligible for SAM intervention after automatic prompt generation. Direct SAM output improved the corresponding UNet prediction on only 8.31% of eligible slices and degraded it on 91.66%. An earlier conservative strategy restricting SAM modification to a narrow region around the UNet boundary improved 13.08% of slices and degraded 86.89%. Thus, reducing the spatial extent of SAM modification alone was insufficient to make intervention consistently beneficial.

Figure 1 illustrates two representative failure patterns. In the upper example, the UNet prediction already closely follows the vertebral target, whereas SAM expands



(a) Intervention outcomes (b) Failure modes

Fig. 2 Quantitative analysis of SAM intervention failures on the internal validation set. (a) Fractions of improved and degraded predictions following direct SAM output and conservative boundary refinement. (b) Frequencies of the major observed failure modes; categories are not mutually exclusive.

beyond the relevant anatomical region, producing substantial over-segmentation and foreground leakage. In the lower example, the relatively small foreground region leads to a less informative automatically generated box prompt, and the resulting SAM prediction again introduces additional segmentation error. These examples illustrate why a plausible SAM output cannot be assumed to constitute a beneficial correction to an already strong task-specific prediction.

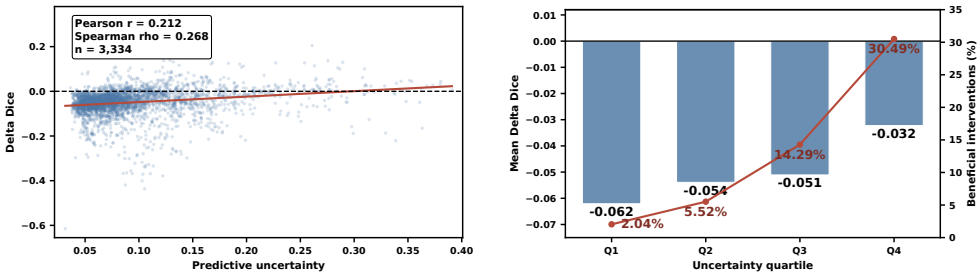
Across the eligible validation slices, failure analysis identified four recurring characteristics: over-segmentation, foreground leakage, small foreground regions, and poor automatic box prompts, occurring in 65.51%, 45.20%, 40.67%, and 28.01% of cases, respectively. These categories were not mutually exclusive.

As summarized in Fig. 2(a), restricting SAM modification to a local boundary region reduced but did not eliminate harmful interventions. Figure 2(b) further shows that these failures arose from multiple overlapping mechanisms rather than a single dominant cause. Together, the qualitative and quantitative analyses indicate that additional foundation-model capacity alone is insufficient to determine whether a proposed correction should be applied, motivating selective intervention based on predictive uncertainty.

2.3 Predictive uncertainty identifies more promising interventions

We next examined whether predictive uncertainty could identify slices for which SAM intervention was relatively more promising. Across the 3,334 SAM-eligible validation slices, uncertainty was positively but modestly associated with the Dice change following conservative SAM refinement. Pearson correlation was $r = 0.2122$ (95% bootstrap CI: 0.1748–0.2504), and Spearman correlation was $\rho = 0.2678$ (95% bootstrap CI: 0.2347–0.3026).

The fraction of beneficial interventions increased from 2.04% in the lowest-uncertainty quartile to 30.49% in the highest, while the mean Dice change improved from -0.0621 to -0.0322 (Fig. 3). Because the mean effect remained negative in every quartile, uncertainty was used as an allocation signal rather than an acceptance criterion.



(a) Uncertainty vs. intervention value (b) Uncertainty-stratified outcomes

Fig. 3 Predictive uncertainty is associated with the value of SAM intervention. (a) Relationship between predictive uncertainty and the change in Dice following conservative SAM refinement across 3,334 SAM-eligible foreground-positive validation slices. The horizontal reference line denotes no change relative to the original UNet prediction. (b) Mean change in Dice and fraction of beneficial interventions across uncertainty quartiles. Although the mean intervention effect remained negative in all quartiles, the fraction of beneficial interventions increased from 2.04% in the lowest-uncertainty quartile to 30.49% in the highest-uncertainty quartile.

2.4 Candidate-value verification enables selective SAM correction

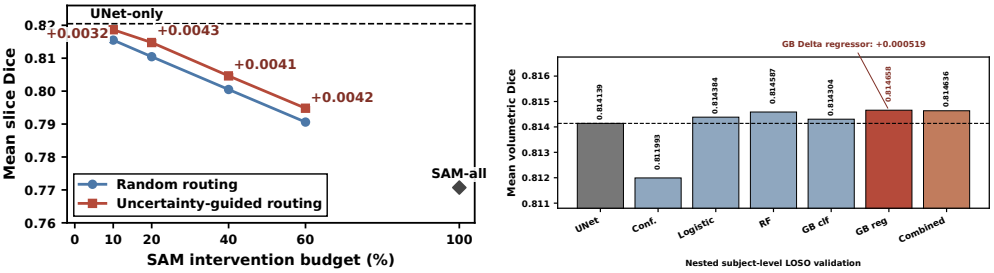
Although uncertainty-guided routing reduced unnecessary SAM usage, routing alone could not determine whether a particular SAM proposal would improve the corresponding UNet prediction. We therefore evaluated a candidate-value verification stage that estimated the expected Dice improvement of each localized correction candidate before it was accepted. All verifier development and model selection were performed exclusively on the 15 internal-validation subjects using nested subject-level leave-one-subject-out (LOSO) evaluation.

After removing exact duplicates from the cached SAM proposals, 3,384 unique candidates remained for verifier development. For each candidate, image-, uncertainty-, disagreement-, and mask-derived features were extracted and used to predict the candidate-specific Dice change relative to the original UNet prediction. We compared a confidence-only policy, logistic regression, random forest, gradient boosting classification, gradient boosting regression, and a combined verification strategy. The UNet prediction was retained whenever no candidate satisfied the acceptance criterion.

The confidence-only strategy did not provide sufficient protection against harmful corrections, reducing the mean subject-level volumetric Dice from 0.814139 to 0.811993. In contrast, learned candidate-value verification substantially reduced unnecessary interventions. Logistic regression, random forest, gradient boosting classification, gradient boosting regression, and the combined verifier achieved mean Dice scores of 0.814384, 0.814587, 0.814304, 0.814658, and 0.814636, respectively (Table 1).

The two levels of selective control are summarized in Fig. 4: uncertainty-guided routing determines where SAM computation is allocated, whereas candidate-value verification determines whether the resulting proposal is accepted.

Among the deployable policies, the gradient boosting Δ-regressor achieved the highest mean validation Dice of 0.814658, corresponding to a numerical improvement of 0.000519 over UNet. The paired bootstrap 95% confidence interval for this difference



(a) Intervention allocation (b) Proposal verification

Fig. 4 Selective control of SAM intervention through uncertainty-guided allocation and candidate-value verification. (a) Mean slice Dice under different SAM intervention budgets for random routing and uncertainty-guided routing. Uncertainty-guided routing consistently achieved higher Dice than random routing at matched intervention budgets, while unconditional SAM intervention produced substantially lower performance. (b) Mean subject-level volumetric Dice of different candidate-value verification policies under nested subject-level leave-one-subject-out validation. The gradient boosting Δ-regressor achieved the highest mean Dice among the evaluated deployable policies.

was $[-0.000312, 0.001493]$, and the Wilcoxon signed-rank test yielded $p = 0.307821$. At the subject level, 8 subjects improved, 4 degraded, and 3 were tied. Thus, the validation result represents a small numerical improvement rather than evidence of a statistically supported aggregate gain.

The selectivity of the final policy was more pronounced than its aggregate Dice difference. SAM was invoked for 652 of the 3,339 foreground-positive validation slices (19.53%), whereas corrections were accepted for only 317 slices (9.49%). Thus, approximately half of the routed slices were ultimately rejected by the candidate-value verifier. Among accepted corrections, 65.62% were beneficial relative to the corresponding UNet predictions. These results indicate that the verifier primarily acted as a safeguard against indiscriminate foundation-model intervention.

To examine the frozen framework beyond internal validation, we subsequently performed a post-hoc evaluation on the 40-subject test set. Because the test set had been evaluated at an earlier project stage, this analysis was not treated as an untouched confirmatory evaluation. Importantly, development and selection of the final candidate-value verifier remained restricted to the internal-validation subjects.

On the post-hoc test evaluation, the original UNet achieved a mean subject-level volumetric Dice of 0.825776 ± 0.068484, whereas the final UGSCF framework achieved 0.825360 ± 0.069728, corresponding to a difference of -0.000416 ($p = 0.265373$). Fifteen subjects improved, 20 degraded, and 5 were tied. The final framework invoked SAM for 45.01% of evaluated slices and accepted corrections for 16.53%. Therefore, the small positive effect observed during internal validation did not translate into an aggregate Dice improvement on the post-hoc test evaluation.

As an additional task-specific control, we evaluated a simple validation-selected largest-3D-component post-processing strategy. This deterministic baseline achieved a mean test Dice of 0.833313 ± 0.056149, corresponding to an improvement of $+0.007537$ over the original UNet ($p = 0.000031$), with 34 of the 40 subjects improved. This result provides an important reference: for this dataset, a simple anatomy-oriented

Table 1 Quantitative evaluation of the proposed selective-intervention framework. Post-hoc test-set performance is reported over 40 subjects using subject-level three-dimensional volumetric evaluation. Candidate-value verification is evaluated on 15 internal-validation subjects using nested leave-one-subject-out validation.

Post-hoc test-set evaluation ($n = 40$)				
Method	**Dice ↑**	**Δ Dice**	**HD95 ↓**	**ASD ↓**
UNet	0.8258 ± 0.0685	–	16.02 ± 28.91	3.28 ± 7.13
UNet + largest 3D component	$\mathbf{0.8333 \pm 0.0561}$	$\mathbf{+0.0075}$	–	–
Final UGSCF	0.8254 ± 0.0697	-0.0004	16.75 ± 29.34	3.42 ± 7.60

Internal-validation verifier analysis ($n = 15$)				
Method	**Dice ↑**	**Δ Dice**	**Accepted**	**Beneficial**
UNet	0.8141 ± 0.0665	–	–	–
Confidence only	0.8120 ± 0.0656	-0.0021	17.82%	53.28%
Logistic regression	0.8144 ± 0.0659	$+0.0002$	9.46%	65.82%
Random forest	0.8146 ± 0.0656	$+0.0004$	9.55%	68.65%
GB classifier	0.8143 ± 0.0657	$+0.0002$	7.91%	64.02%
GB Δ regressor (UGSCF)	$\mathbf{0.8147 \pm 0.0654}$	$\mathbf{+0.0005}$	**9.49%**	**65.62%**
Combined verifier	0.8146 ± 0.0656	$+0.0005$	8.81%	67.01%
Oracle selection	0.8191 ± 0.0600	$+0.0049$	–	–

Note: Dice, HD95, and ASD are subject-level three-dimensional volumetric metrics. HD95 and ASD were not recomputed for the largest-component baseline. Test and internal-validation results correspond to different evaluation cohorts and should not be compared directly. During internal validation, SAM was invoked for 19.53% of foreground-positive slices. "Accepted" denotes the fraction of foreground-positive slices for which a correction was applied, whereas "Beneficial" denotes the fraction of accepted corrections that improved Dice relative to the corresponding UNet prediction. Oracle selection uses ground-truth candidate quality and is non-deployable.

post-processing operation could produce a larger aggregate Dice change than selective SAM intervention. Accordingly, the principal contribution of UGSCF should not be interpreted as uniformly improving segmentation accuracy, but as a mechanism for deciding when a foundation-model proposal should be considered, verified, and selectively incorporated.

2.5 Ablation analysis of selective intervention components

We further performed a component-wise ablation analysis to determine how the successive control mechanisms contributed to the behavior of UGSCF. All variants were evaluated on the same 15 internal-validation subjects using subject-level three-dimensional volumetric Dice. To avoid ground-truth-dependent candidate selection, ablations without the candidate-value verifier used the deterministic box-only SAM proposal (P0). Uncertainty-routing thresholds were estimated independently within each outer LOSO fold using only the corresponding development subjects.

As shown in Table 2, applying SAM proposals without selective control substantially degraded segmentation, reducing mean Dice from 0.814139 for UNet alone to 0.625612. Introducing uncertainty-guided routing increased Dice to 0.712070, indicating that allocating SAM intervention to higher-uncertainty slices reduced the damage

Table 2 Ablation analysis of the main selective-intervention components on the 15-subject internal-validation cohort. All Dice values are computed from complete reconstructed three-dimensional subject volumes.

Variant	Route	Localize	Verify	Dice	Δ Dice
A0: UNet only	–	–	–	0.814139	–
A1: Unselective SAM	–	–	–	0.625612	−0.188527
A2: Uncertainty routing	✓	–	–	0.712070	−0.102069
A3: + Local correction	✓	✓	–	0.811803	−0.002336
A4: Full UGSCF	✓	✓	✓	**0.814658**	**+0.000519**

Note: A1–A3 use the deterministic box-only SAM proposal (P0) to avoid ground-truth-dependent candidate selection. A2–A4 use fold-specific uncertainty routing estimated without the held-out subject. A4 uses the complete candidate pool and the nested-LOSO candidate-value verifier. Δ Dice is calculated relative to A0.

caused by indiscriminate intervention but was insufficient by itself to preserve the task-specific baseline.

Restricting SAM modifications to uncertainty-restricted disagreement regions produced the largest recovery among the successive ablations, increasing mean Dice from 0.712070 to 0.811803. This result indicates that spatial localization was important for preventing a SAM proposal from overwriting reliable regions of the original UNet prediction. Nevertheless, routing and localization alone remained slightly below the UNet baseline.

Finally, adding candidate-value verification increased mean Dice from 0.811803 to 0.814658. The verifier reduced the fraction of foreground-positive slices receiving a SAM-derived modification from 19.53% to 9.49%, rejecting proposals whose predicted intervention value was insufficient. The complete UGSCF therefore recovered the performance lost under uncontrolled SAM intervention and produced a small numerical gain of 0.000519 over the original UNet. Taken together, the ablation results show complementary roles for the three control mechanisms: uncertainty routing limits where SAM is queried, localized correction limits where a proposal may alter the prediction, and candidate-value verification determines whether the resulting correction is ultimately accepted.

2.6 Oracle analysis reveals remaining correction headroom

The limited aggregate improvement of the deployable verifier raises an important question: whether SAM failed to generate useful correction candidates, or whether the main difficulty was identifying those useful candidates reliably. To separate these two factors, we performed a retrospective oracle analysis using the same candidate pool. Ground truth was used only in this non-deployable analysis to identify candidates according to their true Dice improvement.

Oracle candidate selection achieved a mean subject-level volumetric Dice of 0.819059 ± 0.059980, compared with 0.814139 ± 0.066463 for the original UNet (Table 1). This corresponds to an oracle improvement of +0.004920. In comparison, the deployable gradient boosting Δ-regressor achieved an improvement of +0.000519, recovering approximately 10.6% of the available oracle headroom.

This gap indicates that useful SAM-derived corrections were present within the generated candidate pool, but most of their potential could not yet be recovered by a deployable selection policy. The principal remaining bottleneck was therefore not simply candidate generation, but the reliable estimation of candidate value without access to ground-truth segmentation. This observation also explains why unrestricted SAM replacement performed poorly despite the existence of a smaller subset of genuinely beneficial proposals.

Taken together, the routing, verification, and oracle analyses support a consistent interpretation of UGSCF. Predictive uncertainty provides a useful signal for determining where intervention may be valuable, while candidate-value verification is necessary to suppress harmful foundation-model proposals. However, uncertainty alone is not a sufficient acceptance criterion, and the remaining gap between deployable and oracle performance indicates substantial room for improved candidate-value estimation.

3 Discussion

This study investigated how a general-purpose foundation segmentation model can collaborate with a task-specific model when the latter already provides a strong segmentation baseline. The experiments reveal that the main challenge is not simply whether SAM can generate plausible vertebral masks, but whether its intervention is beneficial relative to an already competent UNet prediction. Direct SAM intervention frequently degraded segmentation, indicating that foundation-model predictions should not be assumed to provide uniformly beneficial refinement in this setting.

Predictive uncertainty provided a useful first level of control. Higher-uncertainty slices were more likely to contain relatively favorable opportunities for SAM intervention, and uncertainty-guided routing consistently outperformed random routing under matched intervention budgets. Nevertheless, routing alone remained insufficient: even among slices selected for intervention, a SAM proposal could still be inferior to the original UNet prediction. This distinction suggests that identifying *where to intervene* and determining *whether to accept the resulting intervention* are separate problems.

The final framework therefore assigns different roles to the task-specific and foundation models. UNet remains the primary segmentation model, while SAM acts as an on-demand correction proposer. Predictive uncertainty controls when the additional foundation-model computation is considered, and candidate-value verification determines whether the resulting proposal is allowed to modify the original prediction. This route–propose–verify–correct formulation differs from unconditional replacement or fusion by preserving the task-specific prediction whenever the expected value of SAM intervention is insufficient.

The validation results support this distinction, although the magnitude of the observed improvement was limited. Under nested subject-level leave-one-subject-out evaluation, the best candidate-value policy increased mean volumetric Dice from 0.814139 to 0.814658. The paired confidence interval included zero, however, and the improvement should therefore be interpreted as numerical rather than statistically supported. Importantly, a confidence-only acceptance strategy remained below the UNet baseline, whereas multiple learned candidate-value policies produced

numerically higher Dice values. This suggests that SAM confidence alone is not an adequate indicator of whether a proposed correction will improve a strong task-specific segmentation.

The selective behavior of the framework also has computational implications. SAM was invoked for approximately 19.5% of foreground-positive validation slices in the best validation policy, while only approximately 9.5% ultimately received an accepted correction. The framework therefore separates the decision to spend foundation-model computation from the decision to alter the final segmentation. Such selective use may be particularly relevant in computationally constrained medical imaging settings, where invoking a comparatively expensive foundation model for every input may be unnecessary when a lightweight task-specific model already performs reliably on most cases.

Several limitations remain. First, the numerical improvement over the UNet baseline was small and was not statistically supported by the subject-level confidence interval. Second, the experiments were conducted on a single vertebral CT dataset, and the generalizability of the proposed decision mechanism to other anatomical structures, modalities, and acquisition conditions remains to be established. Third, the effectiveness of the framework depends on the quality and diversity of the available SAM correction proposals. The oracle analysis indicated that useful corrections remained available within the candidate set, but the deployable verifier recovered only a limited fraction of this potential headroom. Finally, the present implementation used a frozen SAM model and lightweight automatic prompting; medical adaptation or stronger prompt-generation strategies may produce more informative correction candidates, although these alternatives require separate evaluation.

The post-hoc test-set analysis provided a complementary perspective. After freezing the final verifier configuration, UGSCF achieved a mean volumetric Dice of 0.825360 compared with 0.825776 for UNet alone, corresponding to a small numerical difference of -0.000416. This difference was not statistically significant ($p = 0.265373$). Thus, the small positive change observed during nested internal validation did not translate into an aggregate Dice improvement on the post-hoc test evaluation. Taken together, these findings indicate that the current framework should not be interpreted as consistently improving segmentation accuracy over the task-specific baseline. Rather, its main contribution is a selective decision mechanism for limiting when foundation-model proposals are generated and when they are allowed to modify an already strong task-specific prediction.

Overall, the results suggest that collaboration between task-specific and foundation segmentation models should not be viewed as unconditional model fusion. In the studied setting, the more appropriate role of the foundation model was that of a selectively queried correction proposer, with uncertainty controlling resource allocation and proposal verification controlling prediction modification. This perspective provides a practical basis for further investigation of selective foundation-model intervention in medical image segmentation.

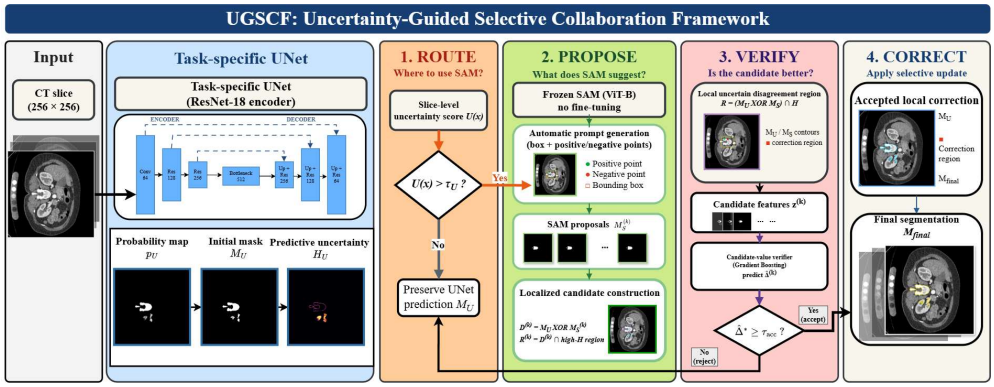


Fig. 5 Overview of the proposed uncertainty-guided selective collaboration framework (UGSCF). A task-specific UNet first produces the foreground probability map p_U, initial segmentation M_U, and predictive uncertainty map H_U. The slice-level uncertainty score $U(x)$ determines whether the input is routed to the frozen SAM. For routed slices, automatic box and point prompts generate multiple SAM proposals $M_S^{(k)}$, from which localized correction candidates are constructed within uncertainty-restricted disagreement regions. A Gradient Boosting candidate-value verifier predicts the expected Dice improvement $\hat{\Delta}^{(k)}$ of each candidate. The best candidate is accepted only when its predicted improvement exceeds the acceptance threshold τ_{acc}; otherwise, the original UNet prediction is preserved. The resulting output is the final segmentation M_{final}.

4 Methods

4.1 Overview of the UGSCF framework

The proposed UGSCF follows a route–propose–verify–correct strategy, in which a task-specific UNet remains the default segmentation model and SAM is invoked only for selected high-uncertainty slices. Rather than directly replacing the UNet prediction with the SAM output, UGSCF treats SAM predictions as candidate corrections whose expected utility is explicitly verified before application. The overall framework is illustrated in Fig. 5.

The framework consists of four sequential stages: uncertainty-guided routing, SAM proposal generation, candidate-value verification, and selective correction.

4.2 Dataset and task-specific segmentation

Experiments were conducted on the VerSe19 vertebral CT dataset. Subjects were partitioned at the patient level into 65 training, 15 internal-validation, and 40 test subjects, corresponding to 14,657, 4,980, and 7,521 preprocessed axial slices, respectively. All partitions were defined at the subject level to prevent slices from the same CT volume from appearing in different subsets.

Images and segmentation masks were preprocessed to 256×256 pixels. Training augmentation included random horizontal flipping, rotation uniformly sampled from $[-10°, 10°]$, and scaling from $[0.95, 1.05]$. Foreground-aware sampling used a foreground-to-background ratio of 75:25.

The task-specific segmenter was a UNet with a ResNet-18 encoder, single-channel input, and two-class output. The encoder was initialized using ImageNet-pretrained

weights. Training used combined cross-entropy and Dice losses with AdamW (learning rate 1×10^{-4}; weight decay 1×10^{-4}). Foreground probabilities were thresholded at 0.5. After checkpoint selection, the UNet was frozen and its prediction served as the default segmentation throughout all subsequent experiments.

4.3 Predictive uncertainty and intervention routing

Let p_i denote the frozen UNet foreground probability at pixel i. After clipping p_i to $[\epsilon, 1-\epsilon]$, with $\epsilon = 10^{-8}$, predictive entropy was computed as

$$H_i = -p_i \log(p_i + \epsilon) - (1 - p_i) \log(1 - p_i + \epsilon). \tag{1}$$

Slice uncertainty was concentrated around the predicted anatomical boundary. The one-pixel inner boundary of the thresholded UNet mask was dilated by five pixels to obtain $\mathcal{B}_5(M_U)$, and the slice-level uncertainty score was

$$U(x) = \frac{1}{|\mathcal{B}_5(M_U)|} \sum_{i \in \mathcal{B}_5(M_U)} H_i. \tag{2}$$

For an empty foreground mask, full-image mean entropy was used as a fallback. For intervention-budget experiments, eligible slices were ranked in descending order of $U(x)$. During nested validation of the final verifier, the routing threshold was estimated exclusively from development subjects in each outer fold as the 80th percentile of SAM-eligible slice uncertainty. Across the outer folds, this procedure routed 652 of 3,339 foreground-positive validation slices (19.53%).

4.4 SAM proposal generation and local correction

A frozen SAM ViT-B model generated correction proposals only for routed slices. An automatic bounding-box prompt was constructed from the largest connected component of the thresholded UNet foreground mask, padded by 8 pixels in each direction, and clipped to the image boundaries. Slices with an empty UNet foreground mask were not eligible for SAM prompting.

Three prompt configurations were considered. P0 used the bounding box only; P1 added up to one positive and one negative uncertainty-guided point; and P2 added up to two positive and two negative points. Candidate points were selected deterministically from high-entropy pixels within a dilated UNet-boundary neighborhood intersected with the prompt box. Positive points ($p_i \geq 0.5$) were ranked by $H_i p_i$, whereas negative points ($p_i < 0.5$) were ranked by $H_i(1 - p_i)$. A 10-pixel Euclidean non-maximum-suppression radius reduced spatial redundancy. Point-prompt configurations without valid points fell back to P0. Candidate prompt regions were evaluated using entropy thresholds corresponding to the 70th, 80th, and 90th percentiles. When SAM returned multiple masks, the output with the highest SAM-predicted confidence was retained.

SAM predictions were converted into localized correction candidates rather than replacing the UNet output globally. For UNet mask M_U and SAM mask M_S, their disagreement region was

$$D = M_U \oplus M_S, \tag{3}$$

where $\oplus$ denotes pixel-wise exclusive OR. For non-empty D, the candidate-specific entropy threshold was

$$\tau_D = Q_{0.90}\left(\{H_i \mid i \in D\}\right), \tag{4}$$

and the local correction region was

$$R = D \cap \{i \mid H_i \geq \tau_D\}. \tag{5}$$

The resulting candidate correction was

$$M_C(i) = \begin{cases} M_S(i), & i \in R, \\ M_U(i), & i \notin R. \end{cases} \tag{6}$$

Thus, SAM could modify only the highest-entropy portion of its candidate-specific disagreement with UNet. Fold-specific generation produced repeated equivalent records, which were removed before verifier learning. The resulting candidate pool contained 3,384 unique interventions from 802 unique physical slices: 802 P0, 1,110 P1, and 1,472 P2 candidates.

4.5 Candidate-value verification

For each candidate correction M_C, the verifier target was its slice-level Dice change relative to the original UNet prediction,

$$\Delta D_C = D(M_C, Y) - D(M_U, Y), \tag{7}$$

where Y denotes the ground-truth mask. Ground truth was used only to construct training targets and evaluate candidate quality and was not included in any inference-time verifier feature.

The verifier used features describing four complementary aspects of an intervention: UNet uncertainty and segmentation geometry; SAM confidence and segmentation geometry; UNet–SAM agreement, disagreement, and area change; and prompt/proposal characteristics. These included slice-level and boundary uncertainty, foreground area and connected components, SAM confidence and area, UNet–SAM Dice/IoU, disagreement fractions and boundary discrepancy, point counts, prompt type, and proposal-stability measures.

Logistic regression, random forest, gradient-boosting classification, and gradient-boosting regression were evaluated. The final verifier was a gradient-boosting regressor with 80 estimators, maximum tree depth 2, and learning rate 0.05. Numeric features were median-imputed, and prompt type was one-hot encoded.

Development and evaluation followed nested subject-level leave-one-subject-out (LOSO) cross-validation over the 15 internal-validation subjects. In each outer fold, one subject was held out for evaluation and all routing, verifier fitting, and acceptance-rule selection used only the remaining development subjects. Inner LOSO predictions on the development subjects were used to select the regression acceptance threshold from

$$\mathcal{T} = \{0,\ 0.005,\ 0.010,\ 0.020\} \tag{8}$$

by maximizing development-subject mean volumetric Dice.

For a routed slice x, the candidate with the highest predicted intervention value was selected,

$$C^*(x) = \arg \max_{C \in \mathcal{C}(x)} \hat{\Delta}(C), \tag{9}$$

and accepted only when

$$\hat{\Delta}(C^*) \geq \tau_{\text{acc}}. \tag{10}$$

Otherwise, the original UNet prediction was preserved. Across the outer folds, 317 of 3,339 foreground-positive validation slices received an accepted correction (9.49%).

4.6 Overall inference procedure

For an input CT slice x, the frozen UNet produces the default mask M_U and predictive uncertainty. Slices below the uncertainty-routing threshold return M_U directly. For routed slices, frozen SAM generates automatically prompted proposals, which are converted into localized correction candidates as described above. The candidate-value verifier selects the candidate with the largest predicted Dice improvement and applies it only when its predicted value exceeds the development-derived acceptance threshold. The final output is therefore

$$M_{\text{final}}(x) = \begin{cases} M_{C^*}, & r(x) = 1 \text{ and } \hat{\Delta}(C^*) \geq \tau_{\text{acc}}, \\ M_U, & \text{otherwise}, \end{cases} \tag{11}$$

where $r(x)$ denotes the uncertainty-based routing decision. Thus, UNet remains the default prediction both when SAM is not invoked and when a generated proposal fails candidate-value verification.

4.7 Evaluation and statistical analysis

Primary segmentation performance was evaluated at the subject level after reconstructing axial predictions into complete three-dimensional volumes. Dice similarity coefficient, 95th-percentile Hausdorff distance (HD95), and average surface distance (ASD) were computed for each subject and summarized across the evaluation cohort. Slice-level Dice was used only for diagnostic analyses of SAM intervention, uncertainty–intervention relationships, and intervention-budget experiments.

For selective intervention, we additionally reported the SAM invocation rate, accepted-correction rate, and the proportion of accepted corrections producing a positive Dice change relative to the corresponding UNet prediction.

Paired subject-level differences between UGSCF and UNet were summarized using the mean Dice change and a 95% paired bootstrap confidence interval and were assessed using a two-sided Wilcoxon signed-rank test with $\alpha = 0.05$. Pearson and Spearman correlations with bootstrap confidence intervals were used for uncertainty–intervention analyses. All verifier development used subject-level nested

LOSO, ensuring that candidate interventions originating from the same subject did not cross development and evaluation partitions. Slice-level diagnostic analyses were not treated as independent subject-level statistical tests.

4.8 Computational environment

The project environment specified Python 3.10.21, PyTorch 2.14.0, segmentation-models-pytorch 0.5.0, and scikit-learn 1.7.2. The foundation-model component used the public Segment Anything implementation with the frozen ViT-B checkpoint `sam_vit_b_01ec64.pth`. SAM proposals were cached after generation and reused during candidate construction and verifier evaluation to avoid repeated foundation-model inference and to ensure consistent candidate predictions across comparative policies.

The audited project environment used a CPU build of PyTorch. Because the available execution records do not establish the hardware used to generate all previously cached SAM outputs, no specific GPU model is reported. Model parameters and data partitions were fixed throughout the selective-intervention experiments, and random seeds and evaluation protocols were kept fixed where supported by the corresponding procedures.

Data availability

The data used in this study were obtained from the publicly available VerSe19 vertebral segmentation dataset. The dataset is available through the official VerSe challenge repository and associated data distribution platform. No additional private or patient-specific datasets were used in this study.

Code availability

The source code for the proposed UGSCF framework is publicly available at https://github.com/XinyuLiu-SCU/UGSCF. The repository includes the implementation of uncertainty-guided routing, SAM-based correction proposal generation, candidate-value verification, selective correction, and the complete evaluation pipeline used in this study.

References

[1] Sekuboyina, A., Husseini, M. E., Bayat, A. *et al.* Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical Image Analysis* **73**, 102166 (2021).

[2] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention* **9351**, 234–241 (2015).

[3] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods* **18**, 203–211 (2021).

[4] Kirillov, A. *et al.* Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision* 4015–4026 (2023).

[5] Ma, J. *et al.* Segment anything in medical images. *Nature Communications* **15**, 654 (2024).

[6] Mazurowski, M. A. *et al.* Segment anything model for medical image analysis: An experimental study. *Medical Image Analysis* **89**, 102918 (2023).

[7] Dong, G. *et al.* An efficient segment anything model for the segmentation of medical images. *Scientific Reports* **14**, 19425 (2024).

Author contributions

Xinyu Liu conceived the study, developed the methodology, implemented the proposed framework, conducted the experiments, analyzed the results, and wrote the original manuscript. Xinlu Lv, Yihan Liu, and Bingran Wang participated in project discussions and assisted with the review and organization of the study materials. Wenchao Du supervised the project, provided methodological guidance, and reviewed and revised the manuscript. All authors reviewed the final manuscript.

Funding

This work was supported by the Undergraduate Innovation and Entrepreneurship Training Program of Sichuan University (Project No. X2026106100606).

Competing interests

The authors declare no competing interests.