Decoder Attention, MaxProb Assembly, and Agreement-Gated TTA for Interactive 3D Biomedical Segmentation

Boyan Lv[1[0009−0003−1075−4770]], Zhou Jiang[1[0000−0001−8146−2028]], Changkai Ji[1[0009−0007−7090−7360]], and Lisheng Wang[1[0000−0003−3234−7511]]*

School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, China
lswang@sjtu.edu.cn

Abstract. Interactive 3D biomedical image segmentation iteratively refines a prediction from user prompts, typically bounding boxes and corrective clicks. nnInteractive is a strong 3D interactive segmentation baseline that predicts each target from its own prompts. However, this per-object inference strategy leaves multi-target consistency to post-hoc assembly, exposing two failure modes in multi-object settings. First, independently predicted masks can overlap; last-wins merging resolves these shared voxels by object order, allowing later labels to suppress earlier ones. Second, before assembly, click-only mask prediction can already be weak: sparse clicks lack the coarse object-extent cue of a bounding box, so small or low-contrast structures may be under-localized or spatially imprecise; such predictions are more likely to overlap with nearby targets, making post-hoc label assignment less reliable. We address these failures with three lightweight extensions, two of which are training-free. **MaxProb assembly** replaces last-wins merging by assigning each contested voxel to the object with the highest foreground probability, with a safety net against empty masks; it accounts for the main broad-set gain (+0.041 DSC AUC). **Decoder attention** with LoRA introduces a decoder-side adapter for click-only localization. For each target, it converts both current-target clicks and other-target clicks into sparse tokens that interact with decoder features; the other-target tokens are intended as auxiliary cues about competing structures and relative layout. Controlled ablations show that removing these other-target tokens preserves nearly all of the gain, so the observed improvement mainly comes from the current target's own click tokens. **Round-aware test-time augmentation (TTA)** applies agreement-gated flip augmentation only at the final interaction round. On a broad 176-case validation set spanning five modalities, the full deployed system reaches 3.161 DSC AUC (+0.051 over nnInteractive), with MaxProb providing the reliable general improvement. On a constructed 90-case no-box subset, decoder attention adds +0.050 DSC AUC and +0.062 normalized surface Dice (NSD) AUC over MaxProb.

* Corresponding author.

Keywords: Interactive segmentation · 3D medical imaging · Multi-object assembly · Test-time augmentation

1 Introduction

Background. Accurate segmentation of 3D biomedical images underpins many clinical and research workflows. Although automatic segmentation has advanced rapidly [13,9], robustly handling the breadth of anatomical structures, imaging modalities, and pathologies remains difficult. Interactive segmentation addresses this gap by incorporating user prompts and iteratively refining the prediction [4]. The CVPR 2026 FMV Interactive Track [8] evaluates methods over iterative prompting trajectories: boxed targets receive an initial bounding box followed by five corrective clicks placed in the largest remaining error region, whereas click-only targets must be localized from corrective clicks alone.

Related work. The Segment Anything Model (SAM) [5] and SAM 2 [10] popularized promptable segmentation as a general-purpose paradigm, but they operate on 2D images or video frames rather than directly on biomedical volumes. This mismatch has motivated medical promptable segmentation methods that either adapt the paradigm to 2D medical images, such as MedSAM [6], or extend it to volumetric inputs, such as SAM-Med3D [11], SegVol [1], and VISTA3D [2]. We build on nnInteractive [4], a strong 3D interactive segmentation model for the FMV Interactive Track that encodes prompts as additional input channels to a 3D Residual Encoder U-Net trained at scale on more than 120 datasets. This early-fusion design is efficient and effective, but because targets are predicted one at a time, multi-object consistency must be recovered after inference by assembling independent per-object predictions.

Diagnosis. nnInteractive predicts targets one at a time: each target is inferred from its own prompt history, with no interaction state carried across targets, and the final multi-class mask is assembled from independent per-object binary predictions. This design is effective for many single-target or box-guided cases, but it exposes two failure modes in multi-object settings, especially when localization relies on clicks alone rather than on a bounding box. *First*, independently predicted object masks can overlap, leaving post-hoc assembly to decide which label should occupy each shared voxel. A last-wins rule resolves this ambiguity by object order, so later predictions can overwrite earlier labels in contested regions. In our profiling subset, multiple object masks claim ∼19% of predicted object voxels (Fig. 1); on hard BraTS cases, replacing last-wins with a ground-truth-guided oracle merge raises round-1 multi-class Dice from 0.15 to 0.46, indicating substantial assembly headroom. *Second*, click-only prompting weakens the per-object predictions before assembly. Unlike a bounding box, sparse clicks do not directly encode object extent. For small, low-contrast, or weakly bounded structures, the resulting masks can be incomplete or spatially imprecise. Such imprecise predictions are more likely to overlap with nearby targets,

forcing post-hoc assembly to decide label ownership from already ambiguous per-object masks and making the merge less reliable. These two failure modes motivate complementary changes: MaxProb addresses how contested voxels are assigned after per-object inference, while decoder attention targets the click-only localization weakness before assembly. In the deployed adapter, we also include clicks from other targets in the same case as auxiliary tokens, motivated by the possibility that they may provide context about competing structures or relative layout. Section 4.2 removes the attention update and the other-target tokens to identify what actually drives the gain.

Contributions. Guided by this diagnosis, we introduce three lightweight extensions to nnInteractive: two training-free inference-time changes and one small trained adapter. The first improves post-hoc multi-object assembly, the second introduces a decoder-side attention adapter for click-only multi-target cases, and the third adds a final-round robustness step:

1. **MaxProb assembly.** We replace last-wins merging with probability-based arbitration: each contested voxel is assigned to the object with the highest foreground probability, with a minimum-voxel safety net against degenerate empty masks. This component is training-free and acts only where per-object predictions overlap.
2. **Decoder attention with LoRA.** We add a lightweight adapter at decoder stage 1 (24^3 resolution), together with LoRA on decoder stages 2–3. For the target being predicted, the adapter converts both its own clicks and clicks from other targets in the same case into sparse tokens. These tokens attend to decoder features, allowing the prediction to use the target's direct prompts together with auxiliary cues about competing structures and relative layout. The module is activated only for click-only multi-target cases and bypassed otherwise. Section 4.2 ablates the attention update and removes other-target tokens to identify what drives the gain.
3. **Round-aware agreement-gated TTA.** We apply flip augmentation only at the final interaction round, when no subsequent clicks depend on the prediction, and average the original and flipped logits only when their binarized predictions agree with Dice > 0.90.

2 Method

Our method builds on nnInteractive [4], which predicts one target at a time with a 3D Residual Encoder U-Net that takes the image and seven interaction channels as input. Fig. 2 summarizes our deployed inference pipeline. For each target, the backbone produces a foreground probability map; our additions refine click-only predictions with decoder attention when applicable, apply agreement-gated TTA only at the final interaction round, and assemble all per-object masks with MaxProb rather than last-wins.

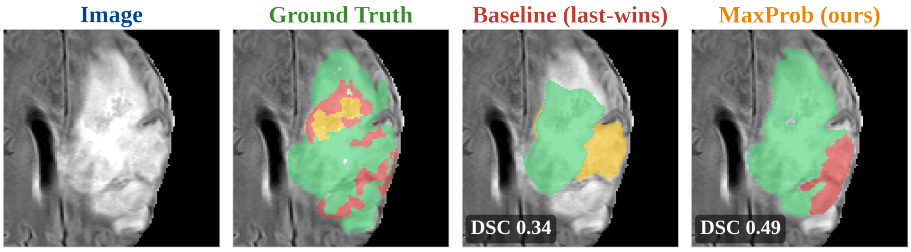

Fig. 1. Diagnostic example of the assembly failure targeted by MaxProb (Case #BraTS-T2f_bratsgli_0385, slice #33). In this BraTS glioma MRI case with three tumor subregions, independently predicted per-object masks overlap. Last-wins assembly assigns overlapping voxels according to object order, suppressing earlier labels and reducing multi-class DSC to 0.34. MaxProb changes only contested voxels, assigning them by foreground probability, and improves DSC to 0.49.

2.1 MaxProb Assembly with Safety Net

nnInteractive assembles multi-object predictions with a last-wins rule, so labels predicted later can overwrite labels predicted earlier wherever their binary masks overlap. We replace this order-dependent merge with order-independent probability arbitration at contested voxels. Let $p_k(v) = \sigma(z_k(v))$ denote the foreground probability predicted for object k at voxel v, and let

$$\mathcal{C}(v) = \{k \mid p_k(v) > 0.5\}$$

be the set of objects that claim v. Voxels with $|\mathcal{C}(v)| = 0$ remain background, and voxels with $|\mathcal{C}(v)| = 1$ keep their only claimed label. For contested voxels, $|\mathcal{C}(v)| > 1$, MaxProb assigns

$$\hat{y}(v) = \arg\max_{k \in \mathcal{C}(v)} p_k(v).$$

Thus, MaxProb leaves all uncontested predictions unchanged while removing the dependence on object ordering.

As a safeguard, if arbitration would remove an object entirely, we restore up to $m = 5$ voxels from that object's pre-arbitration claim, choosing the voxels with the highest foreground probability for that object. This safety net is triggered only in degenerate empty-mask cases; $m = 5$ is a small validity safeguard rather than a tuned performance hyperparameter. In our 116-case overlap analysis, using foreground sigmoid probability, foreground softmax probability, or logit margin as the arbitration score changed DSC AUC by less than 0.001.

2.2 Decoder Attention with LoRA

The second failure mode is weak localization from sparse clicks when no bounding box is available. We address this with a lightweight decoder-side adapter that

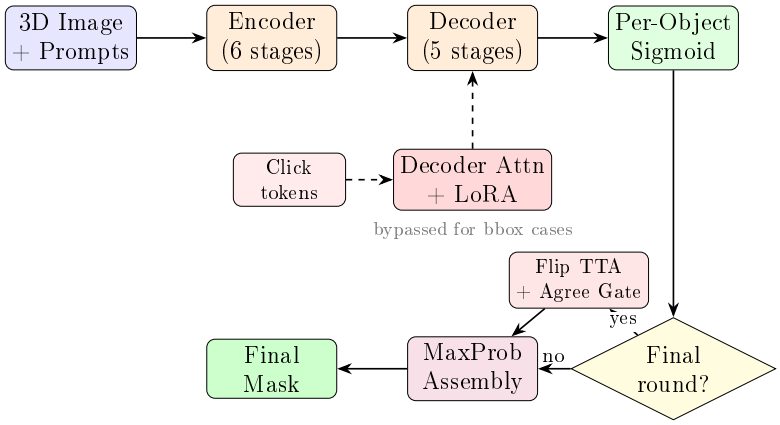

Fig. 2. Pipeline overview. nnInteractive predicts each target independently from the image and interaction prompts. Our three additions are shown as colored blocks: decoder attention with LoRA refines click-only multi-target predictions and is bypassed for bounding-box cases; agreement-gated flip TTA is applied only at the final interaction round; and MaxProb assembly replaces last-wins merging when per-object masks overlap.

conditions intermediate decoder features on sparse click tokens while keeping the pretrained nnInteractive backbone frozen (Fig. 3). The adapter is activated only for click-only multi-target cases and bypassed for box-guided inference. When predicting a target, the deployed adapter builds its token sequence from all available clicks in the case: clicks on the current target are marked as *self*, and clicks on other targets are marked as *other*. This exposes the adapter to both the current target's own prompts and prompts from other targets in the same image. Section 4.2 later ablates the attention update and removes the other-target tokens to determine which part of the deployed adapter drives the gain.

Adapter placement. We insert the adapter at decoder stage 1, where the feature map has resolution 24^3 and 256 channels. Rather than encoding multi-target prompt context as additional input channels, which would require full retraining and could disrupt the input-channel semantics learned during large-scale pretraining, we attach the adapter to the frozen decoder path. This choice keeps the 102.4 M-parameter backbone fixed and makes training feasible under our available compute budget. The bottleneck feature map (12^3), a common site for such attention [12,14], is spatially too coarse for small structures, whereas higher-resolution decoder stages (48^3, 96^3, and 192^3) make attention substantially more expensive. Stage 1 is therefore the coarsest decoder stage that still retains enough spatial detail for click-based localization.

Click-token construction. For the target currently being predicted, we convert the click history from all targets in the case into sparse tokens. Clicks from

the current target are tagged as *self*, and clicks from other targets in the same case are tagged as *other*; foreground and background clicks are distinguished in the role embedding. Each click token combines three components: a Fourier positional encoding of the click coordinate, projected to dimension d=128; a role embedding with five categories (*self*/*other* $\times$ foreground/background, plus a learned background-token category); and an interaction-round embedding. One learnable background token is appended, and all tokens are projected by a two-layer MLP.

Two-way decoder attention. Let $F \in \mathbb{R}^{256\times24\times24\times24}$ denote the stage-1 decoder feature map, and let T denote the sparse click-token sequence. Following the two-way attention of the SAM mask decoder [5,10], we flatten F into spatial tokens and apply two attention layers with four heads each. Each layer first performs token self-attention over the token sequence; in the deployed adapter, this sequence contains both self-click tokens and other-target tokens. The tokens then attend to the spatial decoder features to gather image context, and the spatial features attend back to the tokens to receive click-conditioned updates. The attention blocks use pre-LayerNorm, L2-normalized queries and keys with a learnable temperature, and a 2×-expansion feed-forward network. The resulting feature update is modulated by a per-voxel gate and passed through a zero-initialized 1×1×1 Conv3d before residual addition, so the adapter initially leaves the frozen decoder path unchanged.

LoRA adaptation and activation rule. Because the adapter perturbs the stage-1 decoder features, we add LoRA [3] with rank 4 to decoder stages 2–3 (48^3 and 96^3), allowing the downstream decoder to adapt while keeping the pretrained weights frozen. The module is used only in click-only multi-target cases. For bounding-box cases, where localization is already anchored by the box prompt, the adapter is bypassed and inference follows the original nnInteractive path.

Training. Only the decoder-attention adapter and LoRA weights are trained. The trainable component has ∼2.7 M parameters and is trained on 1,065 multi-object files from 29 datasets and five modalities, balanced approximately 1:1 between spatially overlapping and spatially separated multi-object cases, with no case overlap with the evaluation splits. Training uses no initial bounding boxes: for each sampled multi-object case, interaction is simulated as corrective clicks only, matching the no-box evaluation protocol used in Section 4.2. We use AdamW with learning rate 3×10^{-4}, polynomial learning-rate decay, Dice + CE loss with five-level deep supervision, and the nnInteractive augmentation pipeline. Training takes ∼29 h on a single RTX 3090 over three epochs.

2.3 Round-Aware Agreement-Gated TTA

Test-time augmentation can improve individual predictions, but in an interactive protocol it should avoid altering the prompting trajectory. Although applying flip TTA at every round can improve individual rounds, it approximately doubles

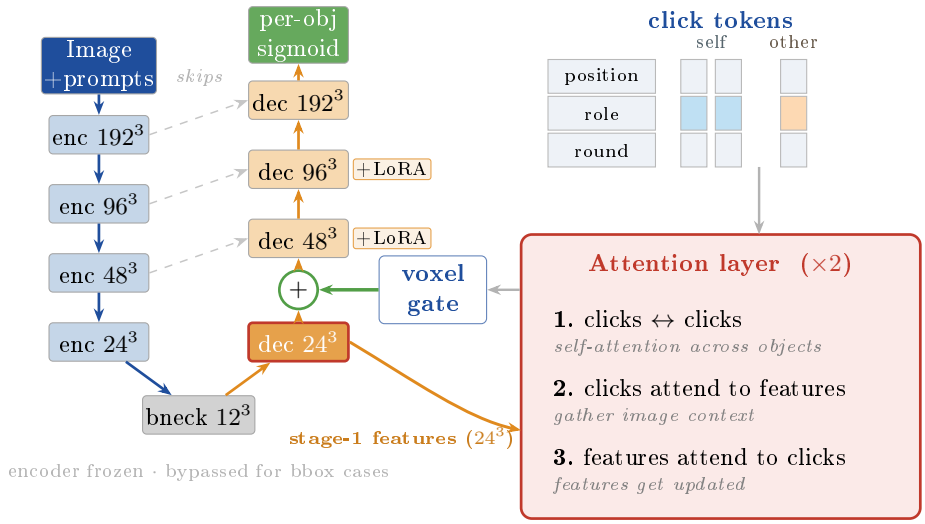

Fig. 3. Decoder-attention adapter. **(left)** The adapter is inserted into the frozen nnInteractive U-Net at decoder stage 1 (24^3). Its gated residual update is added back to the decoder stream ($\oplus$), while rank-4 LoRA adapts decoder stages 2–3. **(right)** For the object currently being predicted, each click is converted into a sparse token carrying position, role (*self* or *other*, foreground or background), and interaction round. Two attention layers first allow tokens to interact, then let click tokens gather image context from stage-1 features, and finally write click-conditioned information back to the spatial features. A zero-initialized per-voxel gate scales the residual update. The adapter is bypassed for bounding-box cases; Section 4.2 ablates the attention update and removes other-target tokens to identify which part of the deployed adapter drives the gain.

inference cost and can exceed the 90 s/class budget. Applying it before the final round is also undesirable, because the augmented prediction changes the error map from which the next corrective click is generated. We therefore apply axis-2 flip TTA only after the final click, when no subsequent prompts depend on the prediction.

For each object k, let z_k be the original final-round logit map and let $\tilde{z}_k$ be the axis-2-flipped prediction inverted back to the original coordinates. We compute agreement on the binarized predictions,

$$a_k = \mathrm{Dice}(\mathbf{1}\{\sigma(z_k) > 0.5\},\ \mathbf{1}\{\sigma(\tilde{z}_k) > 0.5\})\,.$$

If $a_k > 0.90$, we average logits,

$$\bar{z}_k = \tfrac{1}{2}(z_k + \tilde{z}_k);$$

otherwise, we discard the augmented pass and set $\bar{z}_k = z_k$. The gate is applied independently for each object, and the resulting logits are passed to the usual binarization and MaxProb assembly.

Table 1. Inference efficiency on the evaluation platform (NVIDIA Quadro RTX 5000, 16 GB). Runtime is wall-clock per case for a complete interaction trajectory; variation mainly reflects case size, number of targets, and whether attention/TTA are activated. Both configurations meet the 90 s/class budget.

Configuration	Runtime/case (s)	Peak GPU mem. (GB)	Within 90 s budget
nnInteractive baseline	∼31	∼10†	Yes
Full deployed pipeline	∼30–50	∼10†	Yes

†Peak VRAM was not logged separately on the RTX 5000. The ∼10 GB estimate is based on our development-hardware benchmark (peak 9.89 GB); all evaluation runs completed within the 16 GB limit.

Additional implementation details. In click-only cases, there is no Round 0 box prediction, so decoder attention is evaluated over the five corrective-click rounds. LoRA uses α=1 at rank 4 (scale α/r=0.25), no dropout, and the standard zero-initialized B matrix, so the adapted decoder initially matches the frozen decoder.

2.4 Inference Efficiency

The deployed pipeline uses the standard nnInteractive inference code with lightweight wrappers for decoder attention, MaxProb assembly, and final-round TTA. We do not use JIT tracing or CUDA Graphs; in our implementation, these graph-level optimizations were not compatible with the wrapped decoder and were unnecessary for meeting the challenge budget. On the evaluation platform (NVIDIA Quadro RTX 5000, 16 GB), a complete interaction trajectory runs in ∼30–50 s per case for the full deployed pipeline, within the 90 s/class budget. The baseline runtime is ∼31 s; the added cost is case-dependent because decoder attention is activated only for click-only multi-target cases, and agreement-gated TTA is evaluated only at the final interaction round. Table 1 summarizes the observed runtime and memory footprint.

3 Experiments

3.1 Dataset and evaluation metrics

The development data extend the CVPR 2024 MedSAM Challenge [7] and span five modalities: CT, MRI, PET, Ultrasound, and Microscopy. We use the publicly released pretrained nnInteractive model as the frozen backbone and train only the decoder-attention/LoRA adapter; full retraining under the coreset or all-data track was outside our computational budget. Metrics are DSC/NSD AUC and final-round DSC/NSD. Following the official metric, AUC is the trapezoidal area over the five corrective-click rounds (the initial bounding-box prediction is excluded): $\mathrm{AUC} = \sum_{i=1}^{5} s_i - (s_1 + s_5)/2$, where s_i is the metric after click i, giving a range of 0–4. Click-only cases have the same five rounds. All methods are evaluated under the 90 s/class runtime budget.

Table 2. Data splits and their roles. The broad set reflects the realistic, mostly bounding-box challenge distribution. The no-box set is the 90-case click-only evaluation used to measure decoder attention where it is active. The adapter-training split is separate from the evaluation cases and uses click-only prompt simulation.

Split	n	Prompt protocol	Role
Broad eval	176	144 bbox / 32 no-box	Realistic full-pipeline evaluation; MaxProb/TTA analysis; attention active only on no-box cases
No-box eval	90	click-only	Dedicated attention evaluation; case-disjoint from adapter training
Adapter train	1,065	click-only	Multi-object adapter training; corrective-click simulation only

Evaluation sets. We use two evaluation sets, summarized in Table 2. The **broad set** contains 176 cases from 38 datasets, including 144 bounding-box cases and 32 no-box cases. It reflects the realistic, mostly bounding-box challenge distribution and is used for the full deployed pipeline and for analyzing the training-free components, MaxProb assembly and TTA. Because the decoder-attention adapter is bypassed for bounding-box cases, its effect on this broad set is expected to be small and is not the primary attention evaluation.

To evaluate decoder attention where it is active, we construct a dedicated **no-box set** of 90 multi-object cases from nine CT/MRI datasets. Because the official validation data contain almost no click-only multi-object cases beyond BraTS, most cases in this set are originally bounding-box cases that we re-run under a click-only protocol: no initial bounding box is provided, and the model receives only corrective clicks. This no-box set spans a wide range of baseline strength and is used for the main decoder-attention evaluation.

Adapter training and leakage control. The decoder-attention/LoRA adapter is trained on a separate 1,065-file multi-object subset from 29 datasets and five modalities. Training uses the same no-box/click-only prompt setting as the no-box evaluation: no initial bounding box is given, and interaction is simulated through corrective clicks. All 90 no-box evaluation cases are case-disjoint from the adapter-training files. The datasets differ in training exposure: the lesion datasets (ISLES2022, CrossMoDA, and AbdTumor) share a family or subset with training, so they test case-level generalization, whereas EMIDEC, LCTSC, HVSMR, CURVAS, HaN-Seg, and CHAOS are absent from adapter training and therefore test dataset-level generalization. Since MaxProb assembly and TTA are training-free, dataset-family overlap with the adapter training data does not create parameter-fitting leakage for those components.

3.2 Implementation details

Input images follow the challenge NPZ format used by MedSAM-style datasets [6], with intensities stored in $[0, 255]$. For CT, we use modality-specific intensity

windows; for non-CT modalities, intensities are clipped to the 0.5th–99.5th percentiles. At inference, we keep nnInteractive's per-image Z-score normalization. We do not use partial labels or unlabeled images.

Data augmentation follows the nnInteractive pipeline: random flips, rotations ($\pm 30°$), transposition (p=0.5), scaling in $[0.5, 2.0]$, and intensity augmentations. Interaction channels are generated after spatial augmentation, so simulated prompts remain aligned with the augmented image. Corrective-click simulation uses EDT center-biased sampling (α=8) with decay factor 0.98.

Table 3 reports the development environment used to train and validate the decoder-attention/LoRA adapter. Runtime benchmarking on the official evaluation hardware is reported separately in Table 1. Table 4 summarizes the adapter training protocol.

Table 3. Development environment used for adapter training and validation. Runtime benchmarking on the official evaluation platform is reported separately in Table 1.

System	Ubuntu 18.04.5 LTS
CPU	Intel(R) Core(TM) i9-10920X @ 3.50 GHz
RAM	128 GB
GPU	2× NVIDIA GeForce RTX 3090 (24 GB each)
CUDA version	11.8
Programming language	Python 3.10
Deep learning framework	PyTorch 2.6.0, torchvision 0.21.0
Key dependencies	nnInteractive 1.1.3, nnU-Net v2 2.6.2

4 Results and discussion

4.1 Quantitative results on 176 validation cases

Table 5 reports the broad-set ablation on 176 validation cases from 38 datasets, including 144 bounding-box cases and 32 no-box cases. This setting reflects the realistic challenge distribution and is therefore most informative for the full deployed pipeline, but it is dominated by box-guided cases where the decoder-attention adapter is bypassed. The clearest improvement on this set comes from MaxProb assembly: replacing last-wins merging with probability-based arbitration increases DSC AUC from 3.110 to 3.151, a gain of +0.041.

Per-case logs are available for 171 of the 176 validation cases. On these cases, 47 improve, 104 are unchanged, and 20 worsen under MaxProb, giving a paired Wilcoxon $p < 10^{-4}$ with 95% CI $[+0.019, +0.067]$. This behavior matches the design of MaxProb: it is identical to last-wins wherever no voxel is contested, and can change the prediction only on voxels claimed by multiple objects. On a dedicated 116-case multi-object overlap analysis, MaxProb improves 92 cases, ties 13, and worsens 11 ($p < 10^{-15}$), confirming that the broad-set gain is driven by the intended overlap regime.

Table 4. Training protocol for the decoder-attention/LoRA adapter. The pretrained nnInteractive backbone remains frozen; only the adapter and LoRA weights are optimized.

Pretrained model	nnInteractive v1.0 `fold_all` (102.4 M parameters, frozen)
Trainable modules	Decoder-attention adapter + rank-4 LoRA on decoder stages 2–3 (∼2.7 M parameters)
Training data	1,065 multi-object files from 29 datasets and five modalities
Batch / patch / objects	Batch size 1; patch size 192^3; at most 8 objects per case
Epochs / optimizer / LR	3 epochs; AdamW with weight decay 3×10^{-5}; learning rate 3×10^{-4} with polynomial decay
Loss function	Dice + CE with five-level deep supervision
Training time	∼29 hours on a single RTX 3090
Model parameters	102.4 M frozen backbone + ∼2.7 M trainable adapter
FLOPs	≈2472 G per forward pass for an 8-channel 192^3 input, estimated with fvcore; the frozen backbone dominates the cost

Adding MaxProb alone increases DSC AUC to 3.151 (+0.041). Adding the decoder-attention adapter on top of MaxProb yields 3.147 (+0.037); the slight decrease from 3.151 reflects the adapter's near-neutral effect on the 144 bounding-box cases where it is bypassed, combined with minor variance on the 32 no-box cases. The full deployed system, which further adds final-round agreement-gated TTA, reaches 3.161 DSC AUC (+0.051 over the baseline). Because this row is an end-to-end deployed-system result rather than a same-trajectory single-component ablation, we do not decompose the small residual beyond MaxProb into individual components on this bounding-box-dominated set. MaxProb is therefore the reliable general-purpose improvement here, while decoder attention is evaluated in its intended click-only setting in Section 4.2.

4.2 Decoder attention on no-box cases

To isolate the attention module's contribution, we evaluate every case under the click-only protocol (no bounding box) on the single held-out **no-box set** of Table 2: **90 cases** from nine CT/MRI datasets (most originally bounding-box cases re-evaluated click-only), spanning a wide range of baseline strength (aggregate last-wins AUC 2.58; per-dataset from 1.98 to 3.61). Leakage is controlled at the *case* level: all 90 cases are disjoint from the training files. The datasets differ in training exposure: the lesion datasets (ISLES2022, CrossMoDA, AbdTumor) each share a family or subset with training, whereas the others (EMIDEC, LCTSC, HVSMR, CURVAS, HaN-Seg, CHAOS) are absent from it, so the set probes both case-level and, on the unseen datasets, dataset-level generalization. See Table 6.

Table 5. Incremental ablation on the 176-case broad validation set. Each row is an independent end-to-end run of the full interactive trajectory; because corrective clicks are regenerated per run, the small Final-DSC differences among the improved variants (within ±0.002) reflect run-to-run trajectory variance rather than a controlled single-component effect, and DSC AUC is the primary ranking metric. Decoder attention is bypassed for the 144 bounding-box cases; its click-only effect is isolated in Table 6. NSD was computed only for the baseline on this set; NSD for the no-box set is in Table 6.

Variant	DSC AUC	NSD AUC	Final DSC	Δ AUC
Baseline (last-wins)	3.110	3.247	0.806	—
+ MaxProb	3.151	—	**0.814**	+0.041
+ Attention	3.147	—	0.813	+0.037
+ TTA (deployed)	**3.161**	—	0.812	**+0.051**

Table 6. Per-dataset AUC on the 90 no-box cases (click-only evaluation), ordered by baseline DSC AUC. BL = baseline MaxProb; Δ = decoder-attention gain over baseline MaxProb. †-marked datasets share a family or subset with training (case-disjoint); others are unseen in training.

		DSC AUC			NSD AUC		
Dataset	n	BL	+Attn	Δ_{DSC}	BL	+Attn	Δ_{NSD}
ISLES2022†	20	1.993	**2.113**	**+.121**	2.744	**2.949**	**+.205**
AbdTumor†	24	2.517	2.549	+.032	2.955	2.972	+.016
EMIDEC	4	2.571	**2.684**	**+.113**	2.967	**3.130**	**+.163**
HaN-Seg	4	2.700	2.709	+.009	3.285	3.311	+.026
CrossMoDA†	20	2.671	**2.728**	**+.057**	3.638	**3.693**	**+.055**
HVSMR	6	3.389	3.368	−.021	3.305	3.265	−.040
CURVAS	4	3.503	3.505	+.002	3.395	3.381	−.014
LCTSC	7	3.558	3.540	−.017	3.464	3.424	−.040
CHAOS-T1	1	3.610	3.557	−.053	3.589	3.428	−.161
Total	**90**	**2.640**	**2.691**	**+.050**	**3.165**	**3.227**	**+.062**

† Dataset sharing a family/subset with training (case-disjoint); others unseen.

Table 6 reports the per-dataset breakdown over the 90 no-box cases. On aggregate, the module adds **+0.050 DSC AUC** over MaxProb. The gain varies substantially across datasets: it is concentrated on a subset of cases and is negligible or slightly negative on others, as Table 6 shows.

A per-case Wilcoxon signed-rank test over the full 90-case no-box set confirms the DSC AUC gain is statistically significant (48 improve, 38 worsen, 4 tie; $p = 0.023$, 95% CI $[+0.019, +0.084]$). The improvement also carries over to the surface metric: NSD AUC rises from 3.165 (baseline MaxProb) to 3.227 with attention (+0.062), so it is not limited to region overlap. Because the challenge ranks teams per case and averages the ranks, a per-case improvement can affect the final rank even though the module's effect on the bounding-box-heavy full-set average is small (Table 5).

What drives the gain? The deployed adapter uses two sources of prompt information: clicks from the current target and clicks from other targets in the same case. To determine which source drives the gain, we retrain two ablation controls on the same data and evaluate all variants on the same 90 no-box cases under the identical protocol (Table 7). Removing the attention update while keeping only LoRA does not improve performance (-0.022 DSC AUC, $p = 0.02$), ruling out a simple fine-tuning or extra-capacity explanation. Removing the other-target tokens, i.e., using a self-only attention control, recovers almost the entire gain (+0.049, $p = 0.010$). The deployed adapter, which uses both self-click and other-target tokens, reaches +0.050, only +0.001 above the self-only control; this difference is not statistically significant ($p = 0.76$, 95% CI $[-0.017, +0.020]$).

This ablation shows that the measured gain is driven by attention-based refinement of the current target's own clicks against intermediate decoder features. The other-target tokens included in the deployed adapter do not provide a significant additional benefit under the current design and evaluation. We interpret this as a limitation of where multi-target prompt context is injected, not as evidence that such context is useless in general. Our adapter operates at a single coarse decoder stage (24^3, close to the bottleneck); at this resolution, the current target's own clicks may already provide enough localization signal, leaving little room for other-target tokens to contribute. The self-only control therefore suggests that future multi-object modeling should be injected earlier, at the input or full-resolution prompt-encoding level, rather than only through token exchange at one coarse decoder stage.

4.3 Qualitative results

Fig. 1 (Introduction) illustrates the MaxProb assembly effect on a BraTS glioma case with three tumor subregions. Under last-wins assembly, the later-predicted enhancing tumor (yellow) overwrites much of the edema (green), and the necrotic core (red) nearly disappears. MaxProb restores all three subregions by assigning each contested voxel to the label with the highest foreground probability: the edema reappears around the periphery and the necrotic core is recovered, raising

Table 7. Controlled ablation of the decoder-attention gain on the 90 no-box cases. All controls are retrained on the same data and evaluated under the identical protocol. Because each training run takes approximately 29 hours on one RTX 3090, each variant uses one training seed; statistical uncertainty is therefore reported over paired evaluation cases rather than over seeds. Removing the attention update leaves only LoRA and does not improve performance. Removing other-target tokens gives the self-only control, which recovers almost all of the deployed adapter's gain.

Variant	DSC AUC	Δ_{DSC}	NSD AUC	Δ_{NSD}
Baseline (MaxProb)	2.640	—	3.165	—
+ LoRA only (no attention)	2.618	−0.022	3.136	−0.029
+ Attn, self-click tokens only	2.689	+0.049	3.229	+0.063
+ Attn, self + other (deployed)	**2.691**	**+0.050**	**3.227**	**+0.062**

DSC from 0.34 to 0.49. Fig. 4 shows the complementary effect of the decoder-attention adapter on three click-only no-box cases across different datasets and anatomies. The difference is most visible in the center column (EMIDEC, cardiac case with four subregions): the adapter sharpens the boundaries between the myocardial wall (red), cavity (purple/orange), and surrounding tissue (green), producing a segmentation closer to the ground truth. In the two CrossMoDA columns, the vestibular schwannoma and cochlea are small structures where the visual change on any single slice is subtle. Per-slice analysis shows that the adapter primarily reduces false-positive over-segmentation of the cochlea and fills under-segmented tumor regions on other slices; the DSC improvements (+0.07 and +0.06) accumulate from these small boundary corrections distributed across the 3D volume. These examples are selected from the cases with the largest final-mask differences between baseline and attention; on many other cases, the adapter's improvement manifests primarily in earlier interaction rounds, raising DSC AUC through better intermediate predictions rather than through visible changes in the final-round mask. A remaining failure mode shared by both components is probability misassignment in ambiguous overlaps: when two structures have similar appearance or intensity profiles, the highest foreground probability may still select the wrong object.

4.4 Results on final testing set

Table 8 reports the organizer-provided final leaderboard scores for the Interactive Segmentation track. On the hidden final set, our deployed system matches the nnInteractive baseline on the rounded AUC metrics reported by the leaderboard, with DSC AUC 3.30 and NSD AUC 3.45 for both methods. It slightly improves the final-round scores, increasing final DSC from 0.854 to 0.856 and final NSD from 0.902 to 0.904.

We interpret this result as a deployment check rather than a component-level ablation or evidence of aggregate hidden-set improvement. The final leaderboard provides only aggregate rounded scores and no per-case logs, so it cannot determine which component contributes to the small final-round difference. All

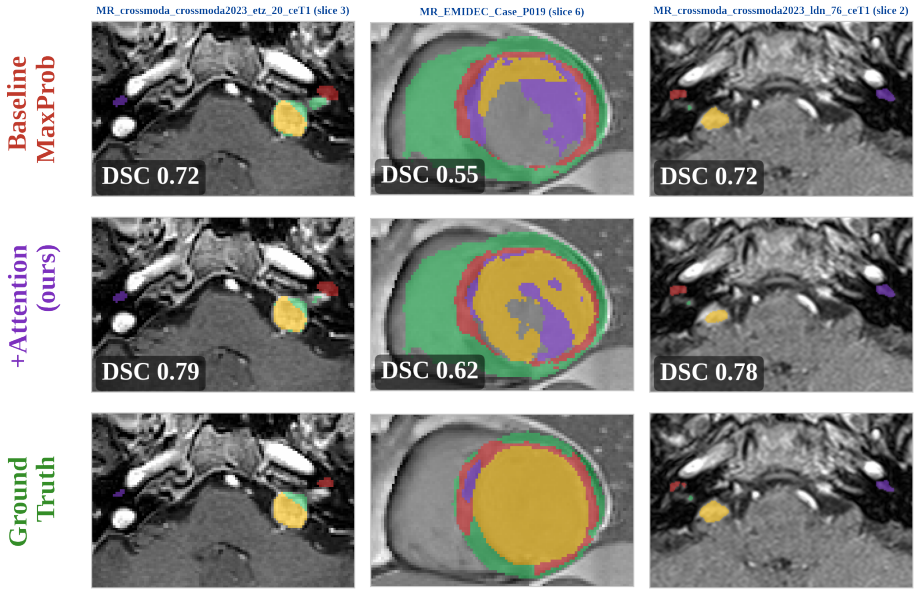

Fig. 4. Qualitative effect of the decoder-attention adapter on three click-only no-box cases (Case #crossmoda_crossmoda2023_etz_20_ceT1, slice #3; Case #EMIDEC_Case_P019, slice #6; Case #crossmoda_crossmoda2023_ldn_76_ceT1, slice #2). Each column shows one case; rows show baseline MaxProb assembly, MaxProb with the deployed attention adapter, and ground truth. The adapter improves click-only localization, with final-round DSC increasing from 0.72→0.79, 0.55→0.62, and 0.72→0.78 across two vestibular schwannoma cases (CrossMoDA) and a cardiac case with four subregions (EMIDEC).

component-level claims in this paper are therefore based on the controlled validation and no-box evaluations in Sections 3.1–4.2. At the reported precision, the hidden leaderboard shows that our deployed system matches nnInteractive on DSC/NSD AUC and slightly improves final-round DSC/NSD by 0.002, indicating no observable deployment-time degradation relative to the strong baseline.

4.5 Limitations and future work

This study has three main limitations. First, the decoder-attention adapter is intentionally lightweight and is inserted only at decoder stage 1 (24^3), a relatively coarse resolution close to the bottleneck. This design keeps the pretrained nnInteractive backbone frozen and satisfies the runtime budget, but it limits where multi-target prompt context can enter the model. Our controlled ablation shows that the current adapter's gain comes from conditioning on the current target's own clicks against decoder features, while the other-target tokens add no significant benefit. We therefore view this negative result as a limitation of the

Table 8. Final leaderboard results for the Interactive Segmentation track. Scores are reported as displayed by the organizer leaderboard. Our deployed system matches nnInteractive on rounded AUC metrics and slightly improves the final-round DSC/NSD.

Method	DSC AUC	NSD AUC	Final DSC	Final NSD
nnInteractive baseline	3.30	3.45	0.854	0.902
Ours (sjtu426439)	**3.30**	**3.45**	**0.856**	**0.904**

injection point rather than evidence that multi-target prompt context is useless in general. A more principled future direction is to encode multi-object prompt context at the input or prompt-encoding level, where nnInteractive already represents user prompts at full resolution, and to train the model end-to-end on larger multi-object data.

Second, MaxProb assembly is a post-hoc arbitration rule. It removes the object-order dependence of last-wins merging and is effective when multiple objects claim the same voxel, but it cannot recover objects that were never predicted, correct poorly calibrated probabilities, or fully close the oracle headroom observed in the diagnosis. A learned multi-object output head, overlap resolver, or training-time consistency loss could address these failures more directly by moving multi-object reasoning into the model rather than correcting conflicts after independent per-object inference.

Third, computational constraints limited both training and evaluation scope. The adapter was trained on a 1,065-file subset with the backbone frozen, rather than end-to-end on the full dataset; training on larger and more diverse multi-object data with a jointly optimized backbone could yield stronger results. On the evaluation side, NSD was computed for all variants on the 90-case no-box set but only for the baseline on the 176-case broad set; the remaining broad-set NSD evaluations were not completed within our budget. Each retrained ablation variant also uses a single training seed because each control takes approximately 29 hours on one RTX 3090. We mitigate these limitations by evaluating all variants on the same cases and reporting paired statistical tests with per-case confidence intervals, but future work should train on larger data, complete the full NSD evaluation, and quantify training-seed variability.

5 Conclusion

We presented a failure-driven extension of nnInteractive for interactive 3D biomedical segmentation. The main broad-set improvement comes from MaxProb assembly, a training-free and order-independent replacement for last-wins merging that arbitrates contested voxels using foreground probabilities. On the 176-case broad validation set, the full deployed system reaches 3.161 DSC AUC (+0.051 over nnInteractive), with MaxProb providing the verified general gain of +0.041.

We also introduced a lightweight decoder-attention adapter with LoRA for click-only multi-object cases. On the 90-case no-box set, it adds +0.050 DSC AUC and +0.062 NSD AUC over MaxProb. Controlled ablations show that

this gain is not explained by LoRA alone and is almost entirely recovered by the current target's own click tokens; the other-target tokens do not provide a significant additional benefit in the current design. This negative result is informative: it suggests that future multi-object modeling should inject multi-target prompt context earlier, at the input or prompt-encoding level, rather than only through token exchange at a coarse decoder stage.

Finally, round-aware agreement-gated TTA provides a small final-round robustness step while preserving the interactive prompting trajectory and remaining within the 90 s/class runtime budget. On the hidden final leaderboard, the deployed system matches nnInteractive on rounded DSC/NSD AUC and slightly improves final-round DSC/NSD, indicating no deployment-time degradation relative to the strong baseline. Together, the results support a simple conclusion: MaxProb is a reliable post-hoc correction for multi-object assembly, decoder attention improves click-only localization, and the next step is to move multi-object reasoning from post-processing into the model itself.

Acknowledgements We thank the challenge organizers, all data owners, Codabench [15] for hosting the evaluation platform, and the nnInteractive team [4] for releasing the pretrained model and inference code.

Disclosure of Interests. The authors have no competing interests.

References

1. Du, Y., Bai, F., Huang, T., Zhao, B.: SegVol: Universal and interactive volumetric medical image segmentation. In: Advances in Neural Information Processing Systems. vol. 37, pp. 110746–110783 (2024) 2
2. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D., Li, W.: VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (2024) 2
3. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022) 6
4. Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., Deissler, J., Floca, R., Maier-Hein, K.: nnInteractive: Redefining 3D promptable segmentation. arXiv preprint arXiv:2503.08373 (2025) 2, 3, 17
5. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the International Conference on Computer Vision. pp. 4015–4026 (2023) 2, 6
6. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 2, 9

7. Ma, J., Li, F., Kim, S., Asakereh, R., Le, B.H., Nguyen-Vu, D.K., Pfefferle, A., Wei, M., Gao, R., Lyu, D., Yang, S., Purucker, L., Marinov, Z., Staring, M., Lu, H., Dao, T.T., Ye, X., Li, Z., Brugnara, G., Vollmuth, P., Foltyn-Dumitru, M., Cho, J., Mahmutoglu, M.A., Bendszus, M., Pflüger, I., Rastogi, A., Ni, D., Yang, X., Zhou, G.Q., Wang, K., Heller, N., Papanikolopoulos, N., Weight, C., Tong, Y., Udupa, J.K., Patrick, C.J., Wang, Y., Zhang, Y., Contijoch, F., McVeigh, E., Ye, X., He, S., Haase, R., Pinetz, T., Radbruch, A., Krause, I., Kobler, E., He, J., Tang, Y., Yang, H., Huo, Y., Luo, G., Kushibar, K., Amankulov, J., Toleshbayev, D., Mukhamejan, A., Egger, J., Pepe, A., Gsaxner, C., Luijten, G., Fujita, S., Kikuchi, T., Wiestler, B., Kirschke, J.S., de la Rosa, E., Bolelli, F., Lumetti, L., Grana, C., Xie, K., Wu, G., Puladi, B., Martín-Isla, C., Lekadir, K., Campello, V.M., Shao, W., Brisbane, W., Jiang, H., Wei, H., Yuan, W., Li, S., Zhou, Y., Wang, B.: Efficient MedSAMs: Segment anything in medical images on laptop. arXiv:2412.16085 (2024) 8
8. Ma, J., Patel, V.M., Zhou, Y., Wang, B.: Interactive 3D Biomedical Image Segmentation Challenge, CVPR 2026 Workshop on Foundation Models for Medical Vision (FMV). https://fmv-cvpr26workshop.github.io/ (2026), codabench competition 5263 2
9. Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., Cao, S., Zhang, Q., Liu, S., Wang, Y., Li, Y., He, J., Yang, X.: Abdomenct-1k: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6695–6714 (2022) 2
10. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations (2025) 2, 6
11. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2024) 2
12. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J.: TransBTS: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 109–119 (2021) 5
13. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023) 2
14. Xie, Y., Zhang, J., Shen, C., Xia, Y.: CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 171–180 (2021) 5
15. Xu, Z., Escalera, S., Pavão, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 17

Table 9. Checklist Table.

Requirements	Answer
A meaningful title	Yes
The number of authors ($\leq$6)	4
Author affiliations, Email, and ORCID	Yes
Corresponding author is marked	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Fig. 2
Pre-processing	Page 9
Strategies to use the partial label	Not used (N/A)
Strategies to use the unlabeled images	Not used (N/A)
Strategies to improve model inference	Section 2.4
Post-processing	Page 4 (MaxProb + safety net)
Dataset and evaluation metric section is presented	Section 3.1
Environment setting table is provided	Table 3
Training protocol table is provided	Table 4
Ablation study	Tables 5, 6
Efficiency evaluation results are provided	Table 1
Visualized segmentation example is provided	Figs. 1, 4
Limitation and future work are presented	Yes
Reference format is consistent	Yes