

Not Quite Anything: Overcoming SAM’s Limitations for 3D Medical Imaging

Keith E. Moore

Stanford University, Stanford, California, U.S.A.

kem1@stanford.edu

Abstract

Foundation segmentation models (such as SAM and SAM-2) perform well on natural images but struggle with brain MRIs where structures like the caudate and thalamus lack sharp boundaries and have poor contrast. Rather than fine-tune these models (e.g., MedSAM), we propose a compositional alternative where we treat the foundation model’s output as an additional input channel (like an extra color channel) and pass it alongside the MRI to highlight regions of interest.

We generate SAM-2 segmentation prompts (e.g., a bounding box or positive/negative points) using a lightweight 3D U-Net that was previously trained on MRI segmentation. However, the U-Net might have been trained on a different dataset. As such its guesses for prompts are often inaccurate but often in the right region. The edges of the resulting foundation segmentation *guesses* are then smoothed to allow better alignment with the MRI. We also test prompt-less segmentation using DINO attention maps within the same framework.

This “has-a” architecture avoids modifying foundation weights and adapts to domain shift without retraining the foundation model. It achieves 96% volume accuracy on basal ganglia segmentation, which is sufficient for our study of longitudinal volume change. Our approach is faster, more label-efficient, and robust to out-of-distribution scans. We apply it to study inflammation-linked changes in sudden-onset pediatric OCD.

1 Introduction

Despite significant progress in segmenting natural images, foundation models like SAM [5], SAM-2 [10], and DINO [4] fail to properly segment brain Magnetic Resonance Images (MRIs). The issue is that regions in the brain (particularly the subcortical regions) are difficult to distinguish due to limited contrast or clear visual boundaries. Retraining or fine-tuning the model on medical images helps (e.g., MedSAM [10]) but the resulting segmentation is inaccurate even with extensive prompting (see Figure 1 where we are trying to isolate the caudate with SAM).

We have had success using supervised models (such as 3D U-Net and UNETR) for brain segmentation [8], but these models are sensitive to dataset distribution shift. We thought SAM might be doing poorly because it lacked 3D spatial information. In SAM-2, we tried encoding the Z dimension along the time dimension (i.e., treating the slices like frames of a movie), but

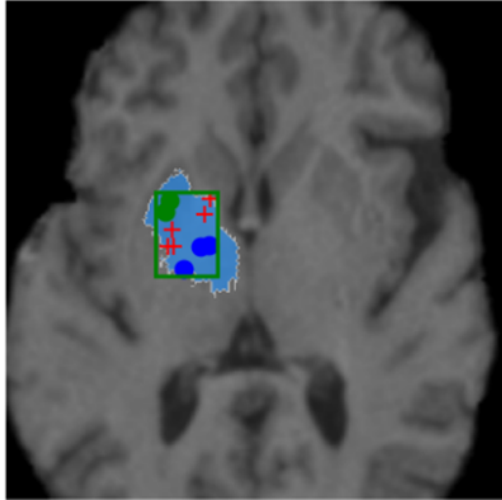


Figure 1: Poor isolation of caudate despite point prompts and bounding box (red-positive, blue-negative, and green-”don’t care”).

this had no improvement to segmentation accuracy. The other option of re-architecting SAM-2 to add the Z-axis and retrain on MRI images was impractical. This led us to explore a different strategy.

Instead of fine-tuning the foundation model, we use the existing inference engine to generate a 2D segmentation guess of each MRI slice as a second ”channel” or color. This output is passed alongside the original MRI slice-by-slice into a multi-channel 3D U-Net to reconstruct the segmented volume. Our hypothesis was that the foundation model might help the U-Net find the intended structures.

In this architecture, the likely-inaccurate ”guesses” by the foundation model act like an alternate imaging sequence (similar to T2 or FLAIR MRI imaging modalities), drawing attention to features in the T1 image. This ”has-a” construction leverages the foundation model for landmark discovery, adapts to distribution shift, and remains fast to train with minimal supervision.

1.1 Clinical Context and Motivation

Our clinical objective is to analyze thousands of pediatric MRIs to investigate immune-related inflammation in neuropsychiatric disorders. We are particularly interested in the caudate and thalamus (small subcortical structures comprising approximately 20,000 voxels within an 8.6M voxel 3D image (240x240x155)). The effect we aim to measure is roughly a 10% change in volume across two timepoints requiring a segmentation accuracy exceeding 94% to distinguish signal from noise (see Appendix A.9).

Unlike segmentation tasks involving large lesions or tumors, our goal is to detect small volume changes in pre-defined structures. Even small boundary errors can obscure clinically meaningful differences, requiring models with higher edge precision than typical MRI segmentation tools. Real-world scans vary in quality, resolution, and orientation across sites. While improved instrumentation may help, clinical research must often work with heterogeneous and imperfect data. We sought methods that were robust to distribution shift, do not require

retraining, and can operate at scale without extensive manual labeling.

1.2 Prior Related Work

Segmentation is often considered a **solved problem** in computer vision and relies on edge or texture cues to isolate structures. Brain MRI presents harder challenges where the relevant structures lack clear visual boundaries and exhibit low contrast with surrounding tissue (see Appendix A.9).

Finding the exact border is difficult even for a skilled radiologist. In tasks such as tumor segmentation (e.g., BRATS21 [7]), broader margins are often acceptable because precise boundaries are resolved histologically. Our use case, however, requires high precision in edges as a 10% error is on average only 12 voxels per slice (i.e., 20,000/155).

Traditional neuroimaging pipelines such as ASEG [15] and SAMSEG [2] generate smooth, topologically consistent segmentations using atlas-based or probabilistic models, but they are computationally expensive (often requiring over an hour per scan).

CNN/Deep learning: FastSurfer ([3]) improves runtime performance using a 2.5D CNN-based architecture, but remains single-modal and can be sensitive to imaging artifacts and image distribution shift. Similarly, 3D U-Net [11] and its transformer cousin, UNETR [13], improve accuracy through improved spatial information but have a similar issue in sensitivity to distribution shifts despite data augmentation[9].

Foundation models (such as SAM [5], SAM-2 [10] and DINO [4]) have poor performance on medical images (see Table 2). Others have fine-tuned these models on 2D medical images (e.g., MedSAM[12] and SAM-UNet[13]); however, these have worked best in situations where there is clear boundaries between regions (i.e., MRIs of the abdomen).

Our approach: These limitations led us to reframe the problem. Rather than treat segmentation as an "is-a" specialization of a foundation model (requiring full retraining), we adopt a compositional "has-a" strategy. We treat the foundation model as a frozen module that emits noisy guesses (used as an additional input channel) allowing a lightweight 3D U-Net to learn robust, high-accuracy segmentations from weak supervision.

1.3 From Fine-Tuning to Composition

We hypothesized that even flawed foundation model outputs could aid segmentation if we treated the segmentation "guess" as an additional input channel. This mirrors radiology practice, where multiple views (T1, T2, FLAIR) are integrated as channels to help resolve ambiguity. We treat a coarse segmentation from SAM or DINO as a synthetic modality and pair it with the raw MRI in a 3D multi-channel U-Net student model. The student learns to integrate this weak signal into an accurate 3D segmentation.

Functionally, the segmentation guess acts like a crude attention map. Architecturally, we shift from an *is-a* inheritance model (where the foundation model is the solution) to a *has-a* design (where the segmentation model *has-a* foundation model as one of its components).

2 System Architecture

Figure 2 outlines the system pipeline. At its core is a **foundation model** (SAM, SAM-2, MedSAM, or DINO) that produces 2D segmentation guesses. For prompt-based models, bounding boxes and point prompts are generated by a prior U-Net iteration (called the **prompt**

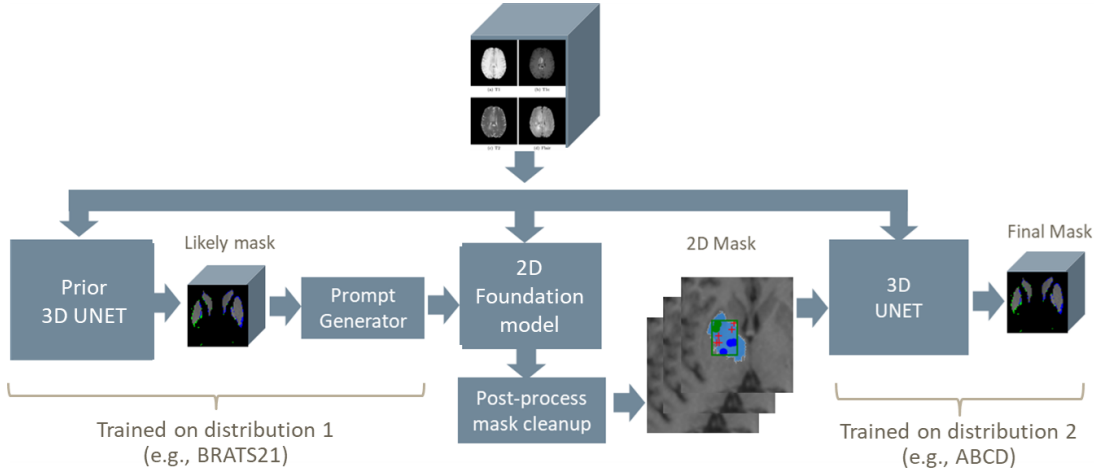


Figure 2: System architecture. A prior U-Net prompts a 2D foundation model. Its mask guess is merged with a raw MRI and passed to a lightweight 3D U-Net student.

generator). These prompts need not be precise. Rough bounding boxes are sufficient to trigger meaningful guesses.

The output mask is then post-processed and merged with the raw T1 MRI as a second channel, forming a 3D volume passed into a **3D U-Net student model**. This model refines the segmentation using both raw structure and foundation cues. The student model is small (3M parameters) and fast to train (under 30 minutes for 100 MRIs). Earlier U-Nets can be re-used as prompt engines to bootstrap later training rounds.

2.1 Segmentation Guess Construction

Our pipeline produces a segmentation “guess” that serves as a second input channel to the 3D U-Net. This guess is derived either via prompted foundation models (e.g., SAM, SAM-2, MedSAM) or prompt-less attention models (e.g., DINO).

Prompt-based generation: Using a prior U-Net iteration, we extracted bounding boxes and five labeled points per slice (foreground, background, or “don’t care”) to serve as prompts. Boxes were jittered for robustness. These were passed into the respective API for SAM, SAM-2, or MedSAM.¹ Appendix A.5 (Figure 6) shows multiple segmentation masks returned by SAM-2 for one prompt.

Prompt-less generation: DINO attention maps were used directly as segmentation guesses, bypassing the prompt engine. These maps often resembled T2-weighted MRIs (Appendix A.4, Figure 5), supporting our multi-modal framing of foundation outputs as synthetic MRI contrasts.

Post-processing: For SAM-based models, we selected the highest-confidence mask per slice. Inspired by MedSAM, we tested contour-based filtering and found that selecting the largest contour within the bounding box was generally effective. To smooth boundaries, we applied a signed distance map using a clipped sigmoid blur (10-pixel range). This improved volume alignment despite minor DICE degradation.

¹MedSAM from [medsam-vit-base](#); SAM-2 from [sam2.1_hiera_large.pt](#); DINO from [dino_vits8](#).

Volume reconstruction and augmentation: Segmentation guesses were reassembled into 3D Nifti volumes stored alongside the raw MRIs. At training time, the T1 image and the segmentation guess were loaded as separate channels. Dataloaders supported optional substitution of the second channel (e.g., blank or T2 for baseline comparisons). Each subject volume was augmented 4x using MONAI transforms (RandomGamma, GaussianNoise, affine perturbations). The 3D U-Net trained on these dual-channel inputs.

2.2 Datasets

We used BRATS21 [6] for initial training and validation. It includes 1,940 multi-modal MRIs (T1, T2, T1ce, FLAIR) from 140 patients, totaling 168,000 slices. Ground truth labels were generated in prior work using SAMSEG [15], though accuracy may degrade in tumor-distorted regions. A subset of ABCD [1] was used to test distribution shift.

Dataset	Description	Subjects	MRIs	% Labels
BRATS21 [6]	Tumor surgery	140	1,940	100%
ABCD [1]	Pediatric longitudinal cohort	11,857	45,678	20%

Table 1: Datasets used in this study; results focus on BRATS21 (Phase 1).

3 Experiments Overview

We organized our experiments into four stages to isolate the effects of architectural dimensionality, generalization, foundation model input, and signed distance smoothing:

- 2D/3D U-Net Baselines (Supervised):** Includes a naive 2D U-Net (T1+T2), a 2D SAM run using U-Net-derived prompts (no refinement), and a full 3D U-Net (T1+T2) to test the benefit of volumetric context.
- 3D UNETR Baselines (Supervised):** UNETR trained on BRATS21 (in-distribution) and then applied to ABCD scans (out-of-distribution) without fine-tuning to test brittleness under distribution shift.
- 3D U-Net + Foundation Guess (No Distance Map):** Dual-channel input using T1 plus segmentation guesses from MedSAM, SAM-2, or DINO; used to assess weak priors without labels or retraining.
- 3D U-Net + Guess + Signed Distance Map:** Same input as above, but the guess is softened at the edge by using a signed distance map[14] and a sigmoid falloff to smooth gradients and improve alignment despite noise.

All U-Net variants used the same architecture and training pipeline, enabling clean comparison of structure, prior quality, and refinement.

4 Results

To establish labels, we used generated labels from SAMSEG and Freesurfer (two well established tools). As we wanted to compare T1+T2 vs T1+SAM-2 guesses, we used SAMSEG as a common labeling tool.

4.1 Baseline Performance

Table 2 reports performance from models without foundation model input. These baselines underscore the value of 3D context and expose the limits of both 2D segmentation and unrefined foundation models:

- **2D U-Net (T1+T2)** generalized poorly from slices to volumes despite high training DICE, underscoring the limitations of 2D architectures.
- **2D SAM (Prompted)** outperformed the 2D U-Net on test DICE (~ 0.46), though this likely reflects prompt quality more than SAM’s segmentation skill.
- **3D UNETR** trained on BRATS21 performed well in-distribution (DICE 0.76, Vol Acc 0.93) but degraded substantially on ABCD scans, highlighting brittleness under domain shift.

Model	Description	Validation DICE	Test DICE	Vol Acc
2D baseline (Supervised)				
2D U-Net (T1 + T2)	Slice-wise supervised baseline	.72	.31	.83
2D SAM (Prompted)	SAM mask compared to ground truth	N/A	.46	.48
3D Baseline (Supervised)				
3D UNETR (BRATS21)	Used for prompt generation; trained on BRATS21 (in distribution)	.92	.76	.93
3D UNETR (ABCD)	Tested on ABCD data without fine-tuning (out-of-distribution)	.92	.69	.82

Table 2: Baseline performance of 2D and 3D models without foundation model inputs

Note: The strong BRATS21 results were expected, but the distribution shift to ABCD proved substantial. This is likely due to anatomical differences between adult and pediatric brains.²

4.2 Performance with Foundation Model Guesses

Table 3 shows results using segmentation guesses from MedSAM, SAM-2, and DINO, each merged as a second channel with the T1 image. We evaluate both raw guesses and versions processed through a Signed Distance Map (SDM).

- **Without SDM:** Results varied. SAM-2 improved test DICE to 0.75; DINO was less effective.
- **With SDM:** Volume accuracy improved across all models. SAM-2 + SDM achieved the best overall performance (DICE 0.76, Vol Acc 0.96).

All models were trained with identical hyperparameters, random seeds, and 4x augmentation.

²In retrospect, BRATS tumors distort anatomy, but pediatric scans from ABCD showed unexpected structural variation.

Model	Description	Validation DICE	Test DICE	Vol Acc
3D U-Net Without Signed Distance Map				
3D U-Net (T1 only)	No auxiliary input	.44	.46	.40
3D U-Net (T1 + T2)	T1 and T2 as channels	.64	.66	.77
3D U-Net (T1 + MedSAM)	MedSAM as second channel	.70	.69	.80
3D U-Net (T1 + SAM-2)	SAM-2 as second channel	.75	.75	.89
3D U-Net (T1 + DINO)	DINO attention map as 2nd channel	.57	.56	.51
3D U-Net With Signed Distance Map				
3D U-Net (T1 + T2 + SDM)	T2 smoothed with SDM	.58	.62	.65
3D U-Net (T1 + MedSAM + SDM)	MedSAM guess smoothed with SDM	.71	.71	.91
3D U-Net (T1 + SAM-2 + SDM)	Best model (guess + SDM)	.76	.76	.96
3D U-Net (T1 + DINO + SDM)	DINO guess smoothed with SDM	.61	.62	.65

Table 3: Segmentation performance using foundation model guesses (with/without SDM)

5 Discussion

The central result is that segmentation guesses from a prompted SAM-2 model, when merged with the T1 image, significantly improve performance (especially when smoothed via a signed distance map (SDM)).³

5.1 Impact of the Signed Distance Map

The form of the SDM was critical. Initial tests used a broad soft gradient across the entire mask, which degraded performance (likely due to misalignment with DICE scores). Restricting the blur to a 10-voxel band around the boundary yielded substantially better results.

We adjusted our loss function as a dynamic loss (i.e., it started by prioritizing a DICE loss, then once the regions mostly overlap, the loss shifts to prioritizing the boundary loss).⁴ Even with imperfect alignment, SDM provides a consistent optimization signal, leading to improved volume estimation.

5.2 Edge Uncertainty Remains

While SDM improves volume accuracy, precise edge alignment remains imperfect (Appendix A.8, Figure 9). Most errors for SAM-2 and DINO occur within 1–2 pixels of the boundary. This is not surprising given the low contrast of regions like the caudate.

Improving edge precision would likely require higher weighting on edge losses or significantly more training. Still, this may not be necessary: volume is cubic in radius, while surface area is

³We had some experience with signed distance maps[14] in a prior project [9] but were surprised at how effective it was for aligning edges between channels.

⁴Conceptually this is a $loss = \sigma(loss_1 - .5) * (loss_1) + (1 - \sigma(loss_1 - .5)) * loss_2$.

quadratic. Small edge errors tend to have limited impact on volume estimates, which are our primary focus.

5.3 Being Close is Good Enough

A surprising result was how tolerant the student model was to noisy segmentation guesses. As long as the foundation model’s output overlapped the general region, the 3D U-Net learned the correct structure. The guess appears to act as a weak “attention” signal, guiding spatial focus even though the architecture lacks explicit attention.

Appendix A.6 (Figure 7) shows an example where the SAM-2 mask overlaps the target region but includes irrelevant areas. Despite this, the U-Net correctly localizes the structure, likely by learning spatial correlations across slices.

5.4 Fine-Tuning vs Adaptation

The architecture resembles a teacher-student setup: the teacher (e.g., UNETR + SAM-2) provides noisy inference-only supervision, and the student (3D U-Net) is trained on these outputs. No gradients are passed to the teacher, avoiding the cost of fine-tuning.

This adaptation strategy lets us wrap foundation model outputs with domain-specific processing (e.g., SDM) without unlocking layers or training new weights. It is faster, simpler, and more flexible.

Training was efficient, and the student generalized well (even with noisy guesses) by learning distribution shift. A small number of labeled samples served as anchors, while the second channel guided spatial attention.

This hybrid setup, combining frozen foundation models with structural smoothing and a lightweight student, is promising for low-supervision segmentation tasks.

6 Conclusion and Next Steps

This project showed that even inaccurate guesses from foundation models can support accurate 3D segmentation when properly adapted. Fine-tuning MedSAM added little; adapting outputs worked better (perhaps due to recovering spatial information). Prompt generation from a prior student model was effective, and DINO worked as a prompt-free option (although not as well). Using a clipped signed distance map consistently improved volume accuracy, especially for weaker inputs.

The key insight is that **adapting foundation model outputs** (rather than modifying foundation model weights) can deliver robust segmentation. Whether this generalizes remains open, and next steps includes ablation studies, but early results on pediatric ABCD data are promising.

Our goal was to enable volumetric tracking of basal ganglia in adolescents to study inflammation-linked OCD phenotypes. The next step is to apply this model to real clinical scans and assess its ability to detect volume shifts linked to inflammation.

7 Acknowledgements

The author would like to recognize the guidance of Prof. Serena Yeung-Levy, Xiaohan Wang and James Burgess of the MARVL lab at Stanford for their help and guidance on the project.

References

- [1] B. J. Casey, T. Cannonier, M. I. Conley, et al. The adolescent brain cognitive development (abcd) study: Imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, 2018.
- [2] John Cho, Sanghyun Ahn, Byungsoo Kim, et al. Label-efficient 3d brain segmentation via complementary 2d diffusion models with orthogonal views. arXiv preprint arXiv:2407.12329, 2024.
- [3] Lukas Henschel, Sailesh Conjeti, Salvador Estrada, Katharina Diers, Bruce Fischl, and Martin Reuter. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, 2020.
- [4] Shubham Juneja, Povilas Daniušis, and Virginijus Marcinkevičius. Dino pre-training for vision-based end-to-end autonomous driving. arXiv preprint arXiv:2407.10803, 2024. Preprint; PDF available.
- [5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.
- [6] Hongbin Li, Giovanni M. Conte, Sajid M. Anwar, et al. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn). arXiv preprint arXiv:2305.09011v5, 2023.
- [7] Hongbin Li, Giovanni M. Conte, Sajid Mohammed Anwar, et al. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn). arXiv preprint arXiv:2305.09011v5, 2023.
- [8] Keith Moore. Measuring longitudinal basal ganglia volume changes using u-net. https://drive.google.com/file/d/1_tFXev6CTiuJyH7bG_YuANOMTlm_IfrH/view?usp=drive_link, 2022. Accessed Sept 20, 2025.
- [9] Keith Moore. Is ocd correlated with volume change of the basal ganglia? https://drive.google.com/file/d/14CuVFN-0hHQmbBuChNmLx4G4z2TZWIIn/view?usp=drive_link, 2023. Accessed Sept 20, 2025.
- [10] Nikhila Ravi, Valentine Gabeur, Yetunde T. Hu, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024.
- [11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [12] Jie Wu, Wei Ji, Yi Liu, et al. Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620, 2023.
- [13] Shuo Yang, Huan Bi, Hao Zhang, and Jian Sun. Sam-unet: Enhancing zero-shot segmentation of sam for universal medical images. arXiv preprint arXiv:2408.09886, 2024.
- [14] Chenyu You, Yujun Zhou, Rui Zhao, Lawrence H. Staib, and James S. Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(9):2228–2237, 2022.

- [15] Liang Zhao, Xi Chen, Ethan Z. Chen, Yi Liu, Tian Chen, Sheng Sun, et al. Retrieval-augmented few-shot medical image segmentation with foundation models. arXiv preprint arXiv:2408.08813, 2024.

A Supplemental Material

A.1 Issues in intensity

The figure below shows the complexity of determining the label based on intensity. The intensities overlap for nearby regions. Brain Atlases tend to do better here as they use the probabilities that a pixel is part of one region or another by looking at the adjacent pixels. Contrast dyes are used in MRIs to try to help separate regions; however, the contrast dyes have some risk and even when used these regions are difficult to discern. The reality is that ground truth is exceptionally difficult with high variance by radiologists on where the exact edge is located in an MRI (making volume measurements quite difficult).

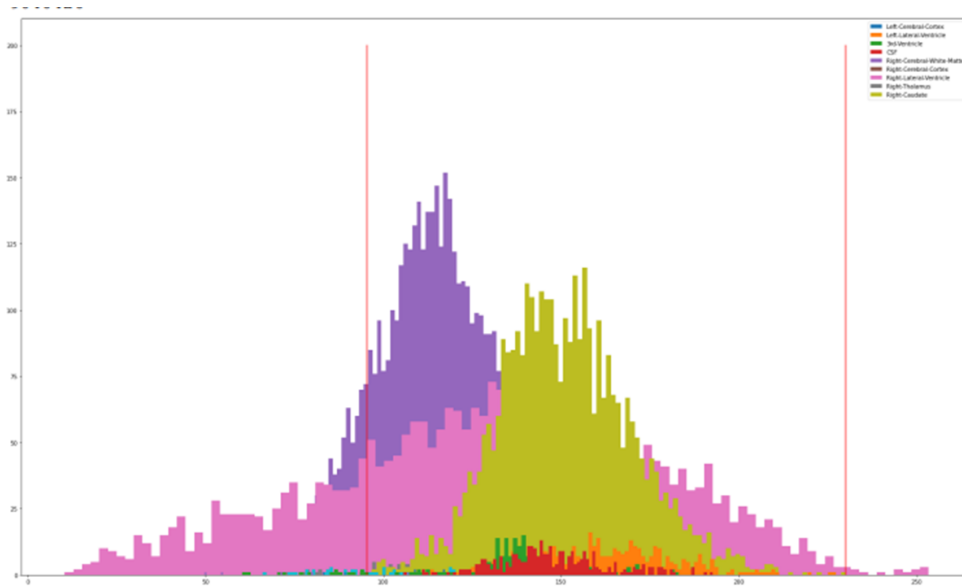


Figure 3: MRI pixel intensities overlap for brain regions making segmentation difficult

A.2 DICE score improved through selective classification

There are outliers in the BRATS21 dataset (as can be seen in supplementary figure 8). Here we note that three of the volumes have very strong differences in labels from the "ground truth". Analyzing the data, we see that in one of the MRIs a tumor has strongly distorted the region of interest making it unrecognizable as being the basal ganglia. The other issues were due to some blurring in the image acquisition. This is common and can, to some degree, be corrected.

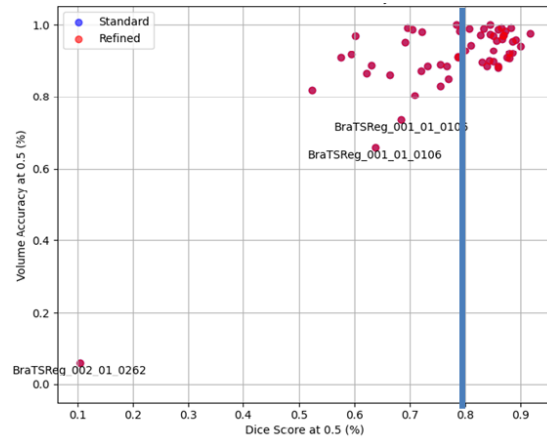


Figure 4: Outliers in the DICE distribution for BRATS21

A.3 Comparison of 2D, 3D U-Net and UNETR volume accuracy

Overall the performance of the UNETR did best with a 93% volume accuracy when fine-tuned on the BRATS21 distribution. The 3D U-Net also had good scores having only 3M parameters. Note, the study below is when using 4 radiology modes that have been aligned (T1, T2, T1ce and FLAIR)

A.4 Qualitative foundation outputs

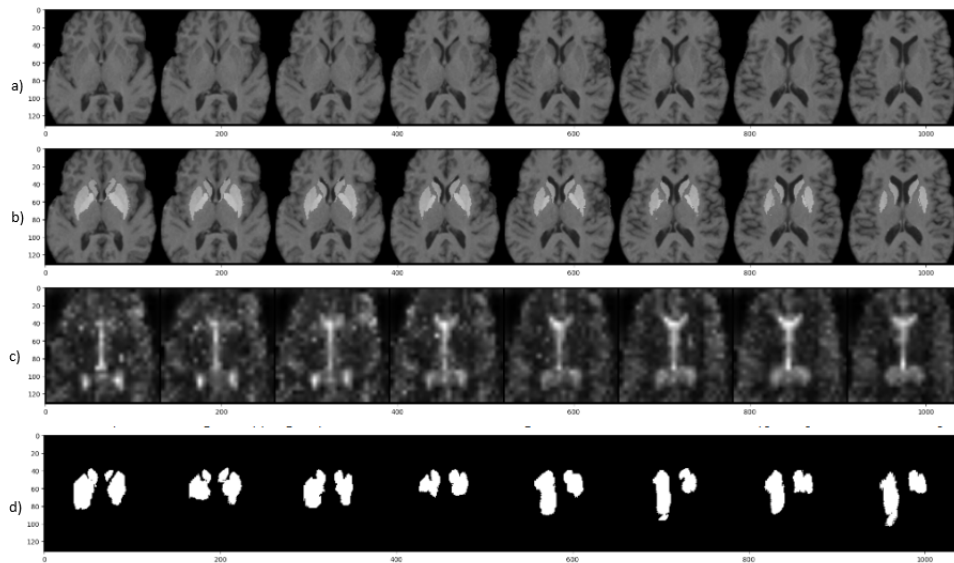


Figure 5: a) T1 image, b) ground truth, c) DINO attention, d) SAM-2 mask.

A.5 SAM-2 prompt options

SAM-2 returned multiple segmentation masks as shown in Figure 6.

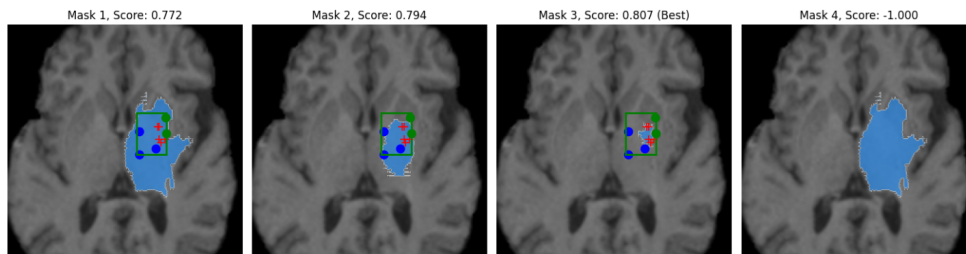


Figure 6: Multiple SAM-2 segmentations for a given prompt

A.6 Being close seems good enough

Surprisingly, the poor guess by SAM-2 shown in Figure 7 is "good enough" to help the U-Net focus on relevant spatial information. We think this is due to the SAM-2 mask finding other relevant anchor/landmark points.

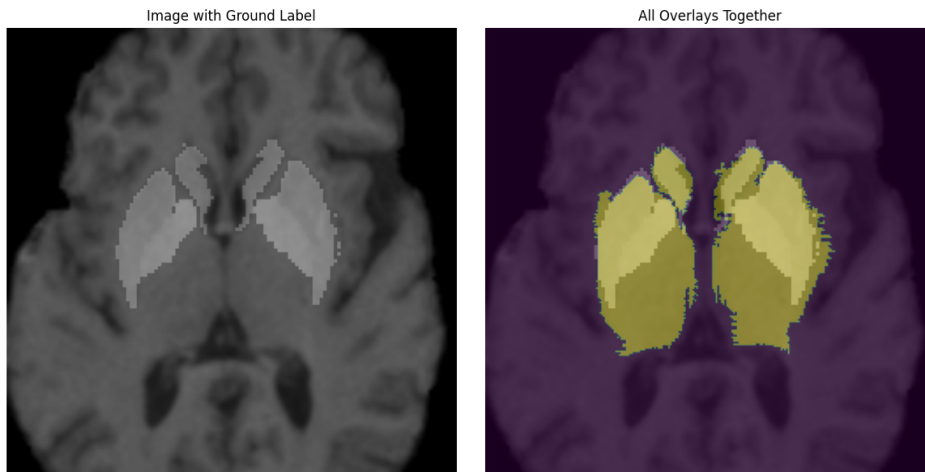


Figure 7: SAM-2 guess overlaps with target region but is not precise.

A.7 DINO does pretty well

Figure 8 shows a comparison of the DINO initial guess and the U-Net refinement. The final row shows the error to the target labels. This is frankly an amazing result (i.e., it is very, very close) and it is possible that this result would improve with a stronger focus in the loss function or a lower learning rate. Next step would be to conduct those experiments. This figure is moved here as it wasn't the main focus of the paper.

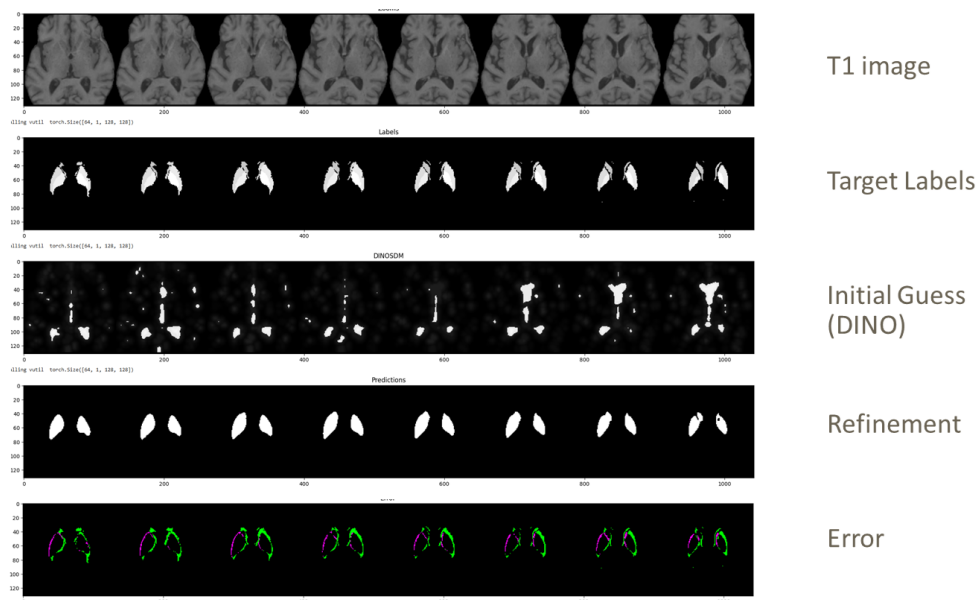


Figure 8: Result of refinement after DINO’s initial guess.

A.8 Dealing with edges

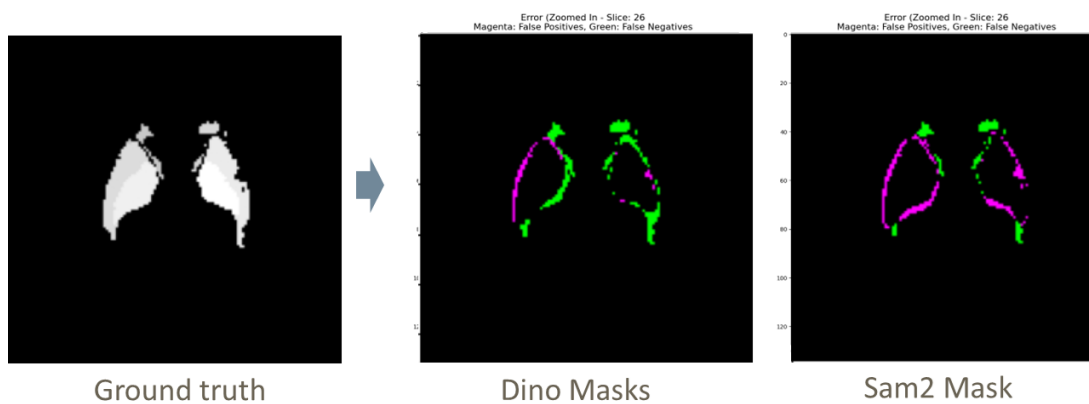


Figure 9: Edge discrepancies: magenta = false positives, green = false negatives

A.9 Per-scan precision needed to resolve 10% change

To estimate the per-scan volume measurement precision needed to resolve a 10% longitudinal effect over two timepoints, we model independent measurement noise at each timepoint and derive a bound on relative error.

Key Variables

- V_{avg} : Typical volume of the region (e.g., caudate).
- ΔV_{diff} : Target effect size in volume change between timepoints (here $0.10 V_{\text{avg}}$).
- ϵ : Relative (fractional) *one-sigma* error of a single volume measurement at one timepoint, so $V_{\text{meas}} = V_{\text{true}} + \eta$ with $\text{SD}(\eta) = \epsilon V_{\text{avg}}$.
- Z : Normal critical value for a one-sided confidence bound (e.g., $Z = 1.282$ for 90% one-sided; $Z = 1.645$ for 95% one-sided).

Error in a two-timepoint volume change Let the measured change be $\widehat{\Delta V} = (V_2 + \eta_2) - (V_1 + \eta_1)$ where η_1, η_2 are independent with $\text{SD}(\eta_t) = \epsilon V_{\text{avg}}$. Then

$$\text{Var}(\widehat{\Delta V} - \Delta V) = \text{Var}(\eta_2 - \eta_1) = \text{Var}(\eta_2) + \text{Var}(\eta_1) = 2(\epsilon V_{\text{avg}})^2,$$

so the one-sigma error in the measured change is

$$\text{SD}(\widehat{\Delta V} - \Delta V) = \sqrt{2} \epsilon V_{\text{avg}}.$$

Condition to resolve a 10% effect above measurement noise To have the target effect exceed measurement noise at a one-sided confidence level Z , requires

$$\Delta V_{\text{diff}} \geq Z \cdot \text{SD}(\widehat{\Delta V} - \Delta V) = Z \cdot \sqrt{2} \epsilon V_{\text{avg}}.$$

Substitute $\Delta V_{\text{diff}} = 0.10 V_{\text{avg}}$ and cancel V_{avg} :

$$0.10 \geq Z\sqrt{2}\epsilon \quad \Rightarrow \quad \epsilon \leq \frac{0.10}{Z\sqrt{2}}.$$

Example thresholds

- 90% one-sided ($Z = 1.282$): $\epsilon \leq \frac{0.10}{1.282\sqrt{2}} \approx 0.0552$ (about 5.5% one-sigma relative error).
- 95% one-sided ($Z = 1.645$): $\epsilon \leq \frac{0.10}{1.645\sqrt{2}} \approx 0.0430$ (about 4.3% one-sigma relative error).

Conclusion Under an independent measurement-noise model, resolving a 10% longitudinal volume effect at roughly 90% one-sided confidence suggests a per-timepoint relative error scale of about 5.5% (and about 4.3% for 95% one-sided). This corresponds to roughly 94–96% per-timepoint volume fidelity when volume accuracy is reported as $1 - \frac{|V_{\text{meas}} - V_{\text{true}}|}{V_{\text{true}}}$.