VOXELPROMPT: A VISION AGENT FOR END-TO-END MEDICAL IMAGE ANALYSIS

Andrew Hoopes [1,2], **Neel Dey** [1,2,3], **Victor Butoi** [1], **John Guttag** [1], **Adrian V. Dalca** [1,2,3]
1. Massachusetts Institute of Technology, **2.** Massachusetts General Hospital, **3.** Harvard Medical School

ABSTRACT

We present VoxelPrompt, an end-to-end image analysis agent that tackles free-form radiological tasks. Given any number of volumetric medical images and a natural language prompt, VoxelPrompt integrates a language model that generates executable code to invoke a jointly-trained, adaptable vision network. This code further carries out analytical steps to address practical quantitative aims, such as measuring the growth of a tumor across visits. The pipelines generated by VoxelPrompt automate analyses that currently require practitioners to painstakingly combine multiple specialized vision and statistical tools. We evaluate VoxelPrompt using diverse neuroimaging tasks and show that it can delineate hundreds of anatomical and pathological features, measure complex morphological properties, and perform open-language analysis of lesion characteristics. VoxelPrompt performs these objectives with an accuracy similar to that of specialist single-task models for image analysis, while facilitating a broad range of compositional biomedical workflows.

1 INTRODUCTION

Clinicians and scientists routinely pose complex questions involving specific targets in medical imaging that extend well beyond simple segmentation or classification tasks. These questions involve multi-step efforts to track the evolution of a particular pathology over many scans, quantify subtle asymmetries of a specific anatomy, or integrate information from multiple acquisitions.

As a detailed example, consider tracking the growth of a specific lesion over time in a patient with multiple abnormalities. After image pre-processing, the first challenge is segmenting *only* the specific lesion of interest. Available tools rarely generalize to diverse, real-world lesion types, and even those that do offer no way to identify a specific lesion using natural language descriptors (e.g., by anatomical location, size, or intensity). Additionally, current tools do not typically accommodate a flexible number of acquisitions from a scan session. As a result, the user must choose a single suitable scan, develop a custom pipeline to programmatically select the target lesion, repeat the process for later scans, and then compute the desired downstream metrics to track changes.

The example above illustrates a fundamental barrier in integrating AI in real imaging workflows. While existing tools perform well on specific segmentation or classification targets (Billot et al., 2023; Isensee et al., 2025), they are specialized to their intended use cases and cannot be used directly to perform broader, integrated analyses that require executing multiple steps. This task-level specialization limits the adoption of AI in radiology, leading to practitioners with complex radiological questions needing to manually chain together multiple fragile, task-specific models and develop extensive post-processing and metric-extraction workflows for each new study.

VoxelPrompt is fundamentally different in functionality and design from existing medical image analysis systems. In VoxelPrompt, we jointly train a language model agent and vision network from scratch to generate and execute *end-to-end* image analysis workflows. Given a task described in natural language, the agent iteratively orchestrates a sequence of instructions as executable code. The dynamically evaluated instructions generate spatial features (e.g., segmentations) using the vision network, incorporate natural language responses, and access a library of functions to compute and provide quantitative outputs. Through diverse output modalities, VoxelPrompt can segment and

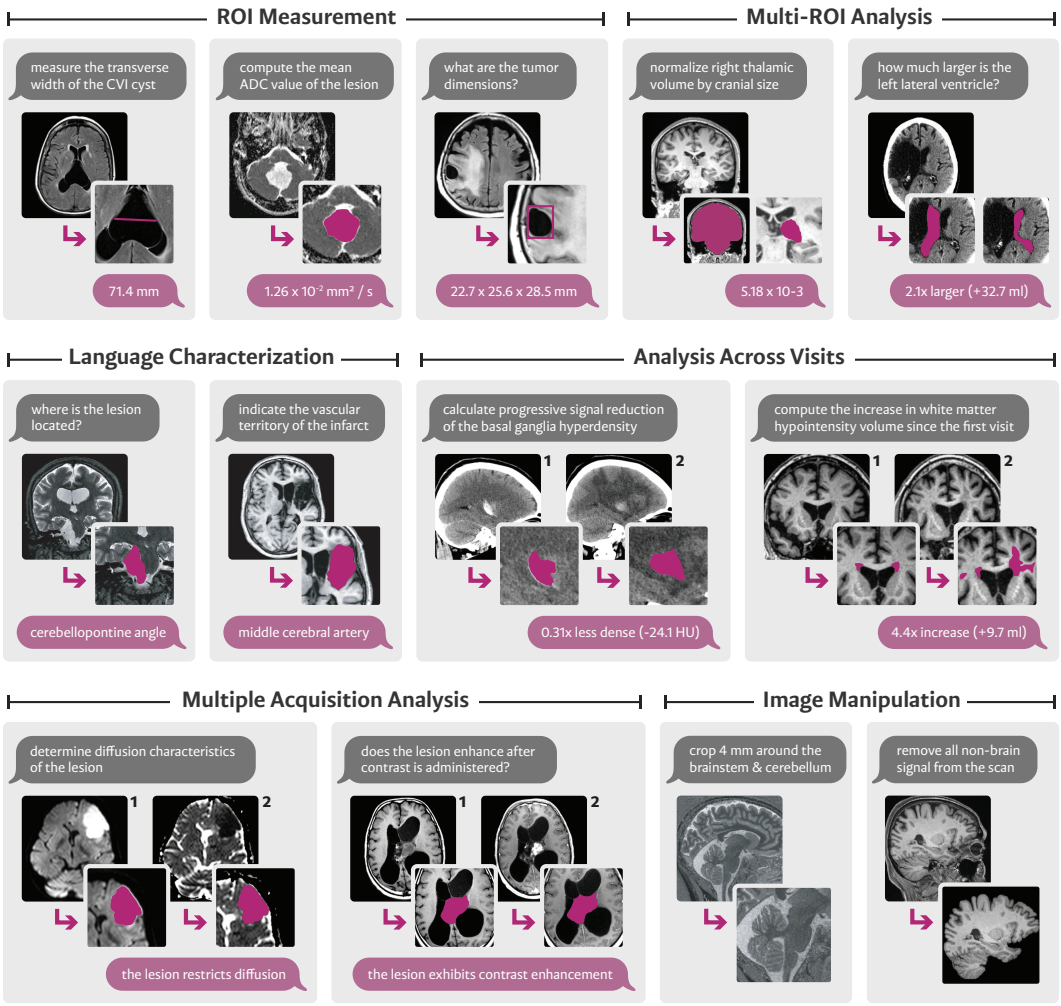

Figure 1: Illustrative examples of VoxelPrompt capabilities, each showing the input prompt (gray) and volumes with VoxelPrompt's predicted annotations and language responses (purple).

localize user-specified anatomical and pathological regions of interest, calculate measurements that relate multiple scans to one another, and perform biomedical characterization (Figure 1).

We make several technical contributions to realize VoxelPrompt's capabilities for real-world medical imaging aims. Our convolutional vision network enables fine-grained, language-controlled visual analysis by integrating jointly-trained language model embeddings as conditioning. To also support multi-acquisition and longitudinal studies (Reuter et al., 2012), the vision network uses attention to process volumetric features across sequences of any length. Further, unlike typical models tied to fixed channels and voxel spacing (Zhang et al., 2024), VoxelPrompt operates on variable-sized inputs at native resolution. This native processing yields substantial memory and runtime efficiency, enabling the joint training of vision and language components on large neuroimaging volumes on standard GPUs without prior resampling. Lastly, we facilitate robustness to acquisition type as well as anatomical and pathological variation by creating and training on a large neurological dataset combining public cohorts, new annotations of unlabeled pathological volumes, and simulated lesions.

We focus on brain imaging and show that VoxelPrompt enables end-to-end analysis on nuanced and diverse tasks covering a wide range of MRI and CT acquisitions, anatomies, and diseases. Quantitatively, we show that a single VoxelPrompt model captures, and often exceeds, the individual accuracy and capabilities of many single-task specialist neuroimaging baselines, while retaining unique language prompted flexibility. These results highlight VoxelPrompt's promise as a foundation for tackling diverse and complex radiology workflows.

2 Related Work

Brain Region Analysis. Widely-used neuroimage analysis pipelines typically delineate regions and quantify their size, shape, composition, and change over time (Fischl, 2012; Jenkinson et al., 2012). Modern approaches train networks to segment anatomical and pathological structures, including cerebral subregions (Billot et al., 2023; Henschel et al., 2020), vessels (Hilbert et al., 2020; Livne et al., 2019), and lesions (Hssayeni et al., 2020; Liu et al., 2021). While performant, these networks generally work for fixed segmentation targets and require significant human involvement for analyzing data and deriving downstream ROI measures. VoxelPrompt aims to match or outperform these methods in segmentation accuracy, while tackling a wider range of targets, enabling flexible specification of target regions, and facilitating end-to-end workflows.

Learning Across Medical Imaging Tasks. Recent medical imaging methods aim to improve performance by exploiting shared representations across diverse segmentation, classification, registration, and statistical modeling objectives in a single framework (Elmahdy et al., 2021; Graham et al., 2023; Tellez et al., 2020; Liu et al., 2025; Czolbe & Dalca, 2023). Broad, segmentation-focused tools, like interactive or in-context segmentation models, can adapt to specific biomedical targets, prompted by partial image annotations (Cheng et al., 2023; Luo et al., 2021; Ma et al., 2024; Wong et al., 2023) or example image-segmentation pairs (Min et al., 2021; Xie et al., 2021; Butoi et al., 2023; Ouyang et al., 2022; Rakic et al., 2024; Roy et al., 2020). However, these multi-task models do not aim to address a complete analytical pipeline and can require finetuning in real scenarios. In contrast, VoxelPrompt integrates supervision from many tasks to create computational workflows, where multiple components interact to carry out requested analyses.

Medical Vision-Language Models. Vision-language models (VLMs) trained on large-scale biomedical image-caption datasets (Johnson et al., 2019; Lin et al., 2023; Zhang et al., 2023a) can facilitate biomedical visual question-answering (Chen et al., 2023a;b; Zhang et al., 2023a;b) and clinical report generation (Bannur et al., 2024; Wang et al., 2023c;b). However, current biomedical VLMs remain largely limited to narrow-domain, text generation tasks, and do not capture the quantitative metrics required in real-world clinical imaging workflows. In contrast to current vision-language models that produce text outputs in a black-box manner, VoxelPrompt explicitly produces code for all relevant intermediate outputs and a traceable sequence of operations. This provides analytical transparency for high-stakes applications. Also, unlike existing models, the VoxelPrompt operations involve explicit vision operations to compute and present images depicting the essential intermediate features. Finally, aside from few recent works (Chen et al., 2023a;b; Liu et al., 2023; Wu et al., 2025; Zhou et al., 2024), most models are trained exclusively on two-dimensional image slices, often X-rays, making them inappropriate for MR and CT imaging. VoxelPrompt is instead trained directly at native acquisition resolution, enabling it to process 3D volumes.

Language Models as Agents. Recent efforts extend large language models beyond plain text prediction into agents capable of planning and executing actions for computational tasks. Often, these generate code (Gupta & Kembhavi, 2023; Ke et al., 2025) that call external APIs for mathematical computation (Ruan et al., 2023; Gou et al., 2023), image analysis (Subramanian et al., 2023; Surís et al., 2023; Yang et al., 2023), scientific discovery (Bran et al., 2023; Boiko et al., 2023), and more. Adaptive, feedback-driven agents address complex and dynamic problems by iteratively planning, executing, and interpreting intermediate outcomes rather than predicting entire action sequences at once (Huang et al., 2022; Rana et al., 2023; Wang et al., 2023d;a; Yao et al., 2022; Zhu et al., 2023). Building on this idea, VoxelPrompt trains an adaptive agent that interacts with a library of processing functions. Unlike other methods, VoxelPrompt jointly trains an adaptable vision network to guide image processing. Recent work in medical imaging (Li et al., 2024) trains an agent to select from a set of pretrained, task-specific tools, but it does not execute downstream operations or leverage flexible language prompting to distinguish ROIs with specific characteristics as in VoxelPrompt.

3 Methods

3.1 Modeling Details

VoxelPrompt processes volumes $\mathcal{V}$ in response to a text prompt p. A language model agent α translates the prompt into Python code executed in a persistent environment Ω, invoking actions involving

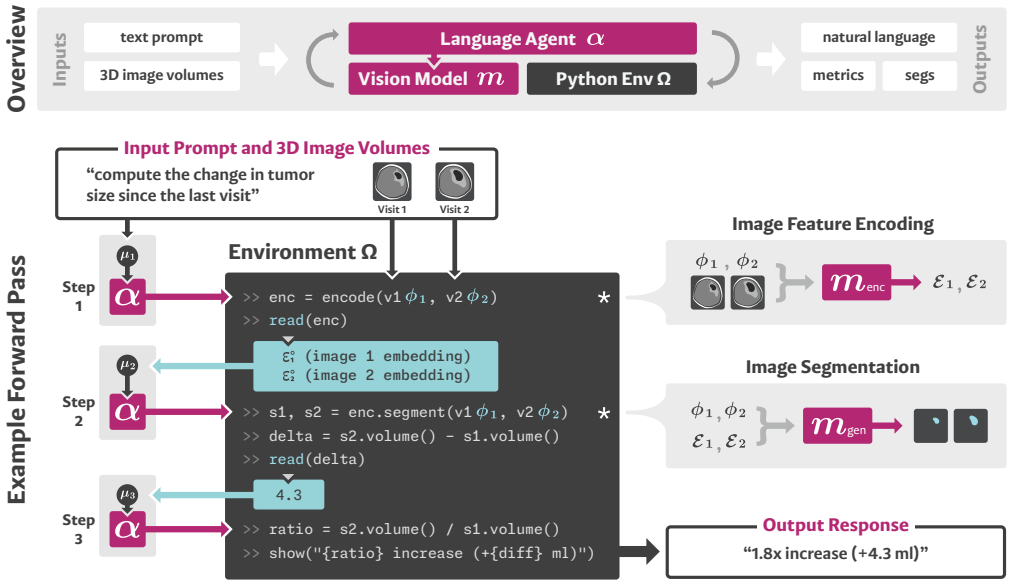

Figure 2: **Top:** VoxelPrompt takes a text prompt and volumes as input to a trainable agent α. The agent iteratively produces executable code in a Python environment Ω, which controls a jointly-trained vision model m. **Bottom:** To solve an example language-prompted task, the agent α interprets execution outcomes z (blue) to guide subsequent instruction prediction across multiple steps. To perform vision operations, such as volume encoding or generation, α employs vision networks m_{enc} and m_{gen}, which are manipulated by image-specific latent instruction embeddings ϕ.

mathematical computation, morphological operations, and interface interaction. A core function set runs a jointly trained vision network directed by the agent to perform vision operations. Figure 2 summarizes this framework and outlines an example use-case. We provide low-level implementation details in Appendix A.

Agent. The agent α produces code iteratively, with each step building on outcomes of prior actions. At step i, it generates executable code $c_i = \alpha(\mu_i)$ based on a state representation $\mu_i \in \mathbb{R}^{\ell,d}$ with sequence length ℓ and embedding dimension d. The code runs in environment Ω, which preserves variables across steps. Intermediate results from Ω can be read and embedded into a representation z_i and incorporated into the next state $\mu_{i+1} = \mu_i \shortparallel z_i$, where $\shortparallel$ denotes sequence concatenation. The initial state μ_1 encodes the prompt p and acquisition date metadata for each volume $v \in \mathcal{V}$. This loop of code generation, execution, and feedback continues until a stopping code signals completion.

Vision Network. Several functions in environment Ω invoke a shared convolutional vision network, consisting of an input feature encoder $\mathcal{E} = m_{\text{enc}}(\mathcal{V}, \phi)$ and a volume generator $\mathcal{W} = m_{\text{gen}}(\mathcal{E}, \phi)$. The latent embeddings ϕ are produced by the agent, passed as arguments to the functions, and condition the vision network for a specific goal. For example, they might direct the vision network to segment the edema around the lesion in the frontal lobe. Vision network outputs $\mathcal{E}$, $\mathcal{W}$ can then be further processed by downstream actions to execute the user prompt.

Volume Interaction. The vision subnetworks share information across an arbitrary number of input volumes using an attention mechanism. Each input volume v is processed by m_{enc} (or its encoding $\mathcal{E}_v$ processed by m_{gen}), producing individual *streams* of intermediate activations from each volume that interact with each other at each layer. Specifically, for voxel features $a_s \in \mathbb{R}^c$ in volume stream s, we concatenate a_s with stream-specific ϕ_s, and use a fully-connected layer to yield $a'_s \in \mathbb{R}^c$. We then stack corresponding voxel representations a'_s across S streams to construct $A' \in \mathbb{R}^{S,c}$. We interact streams in A' using attention with dimension b: $B = f(\text{softmax}(QK^T b^{-1/2})V) + A$, where $Q, K, V \in \mathbb{R}^{S,b}$ are learnable linear transformations of A, and fully-connected layer f projects the output to $\mathbb{R}^{S,c}$. We then separate B into stream-specific voxel features for each volume.

Native Space Processing. Volumetric image formats define a world-coordinate transform specifying in-plane voxel spacing x and inter-slice spacing y, with anisotropy ratio $r = y/x$. Standard tools

often resample images to isotropic resolution, which greatly inflates data size for thick-slice acquisitions. Instead, we implement a vision network that operates in native voxel resolutions by tracking and updating spacings throughout the multi-scale hierarchy. In the downsampling arm, following resolution level n, the target in-plane spacing is set to $x_{n+1} = 2x_n$, while the target slice spacing is updated to $y_{n+1} = x_{n+1}$ only when $r_n \leq 2$. In the upsampling arm, voxel spacings are inferred from the skip connections. During stream interaction, volume features are resampled to a common geometry, then returned to their previous space.

Supervised Training. We jointly train the language model α and vision network from scratch on a curated, diverse task set $\mathcal{T}$ (Section 3.2). Each task $\tau \in \mathcal{T}$ is paired with target (ground-truth) code c^* that carries out the task objective, as illustrated in Figure 2. At each training step, we sample $\tau \sim \mathcal{T}$, generate a prompt p, and sample input volumes $\mathcal{V}$ with ground-truth outputs $\mathcal{W}^*$. The training loss is $\mathcal{L}_{ce}\big(P(c), c^*\big) + \lambda \sum_{j=1}^{|\mathcal{W}|} \mathcal{L}_{img}\big(\mathcal{W}_j, \mathcal{W}_j^*\big)$, where $P(c)$ is the language model output, volumes $\mathcal{W}$ are generated by the vision networks while executing c^*, $\mathcal{L}_{ce}$ is cross-entropy, and $\mathcal{L}_{img}$ compares predicted and target volumes (using soft Dice loss for segmentation).

3.2 Training Tasks and Data Design

We construct and curate a dataset $\mathcal{T}$ of brain imaging tasks, designing new task formulations and labels across a wide range of image acquisitions, segmentation protocols, and annotation types. We use this dataset to both train and evaluate VoxelPrompt in the joint prediction of analytical instructions, spatial delineations, and natural language descriptions. We include a set of clinically-oriented objectives, which are broadly categorized as either ROI processing or pathology description tasks. For each task, we create ground-truth code c^*, used in training and evaluation. Additional training data details are described in Appendix B.

Training Code for Quantitative ROI Processing. Quantitative processing tasks involve image feature segmentation, optionally followed by downstream steps to compute ROI measures. We include a core segmentation task for all structures and pathology classes in our dataset. Downstream processing tasks use predicted segmentations, sometimes in conjunction with the input volumes. For example, some tasks involve removing, extracting, or cropping the field of view (FOV) around a segmented region. Others use segmentations to compute ROI-specific statistics of image signal intensities (e.g., mean intensity). Morphological tasks analyze ROI shape and compute total volume, bounding box dimensions, or the maximum height, width, and depth of a segmented structure. We also include tasks that compute and compare such metrics across multiple segmentations. For example, longitudinal tasks measure change in ROI properties across a series of scan sessions, and multi-region tasks compare metrics from different ROIs in a single scan session. To support these applications, ground-truth code c^* specifies a sequence of functions that predict the required segmentations, compute relevant metrics, and format the results into an output message (Figure 2).

Training Code for Question Answering. We also train on question-answering tasks, where VoxelPrompt produces natural language responses from a combinatorially large set of possible answers. For example, some tasks involve classifying lesion signal intensity as hyperintense, hypointense, or isointense relative to surrounding tissue, while others require identifying anatomical location. Certain tasks integrate information across multiple images, for example, detecting restricted diffusion from paired DWI and ADC maps, or assessing post-contrast lesion enhancement. We handcraft a target language response for all possible answers. For each task, we construct the ground-truth instruction code c^* with functions to (1) encode the input volumes, (2) read the encoded volume features, and (3) output a text response with the correct natural language answer.

Training Prompt Synthesis. We synthesize a combinatorially diverse set of prompts for training. For each task τ, we define a set of prompt templates $\mathcal{P}_\tau$ containing placeholders to accommodate multiple words, terminologies, and phrases with similar meanings. We compile a list of interchangeable text $\mathcal{C}_k$ for each placeholder k. To generate a prompt p for task τ, we sample a template from $\mathcal{P}_\tau$, then fill each placeholder k with text sampled from $\mathcal{C}_k$. Placeholders may themselves contain other placeholders, making the process recursive. This yields a diverse distribution of prompts spanning clinical and imaging terminology, as well as variations in tense, syntax, and word choice.

Training Images and Segmentations. We assemble and annotate a collection of 6,925 3D brain MRI and CT scans from 15 public datasets, comprising 185 bilateral anatomical structures and 14

pathology classes, focusing on a breadth of imaging types, regions of interest, and tasks. The MRI sequences span T1w, T2w, FLAIR, PD, GRE, and DWI with various scan resolutions. The subjects are split into 4,852 training, 213 validation, and 1,860 test volumes. Anatomical segmentations are derived from established pipelines (Fischl, 2012; Greve et al., 2021; Hoopes et al., 2022), atlas annotations (Adil et al., 2021; Pauli et al., 2018), manual corrections, and manual labeling of additional structures in a small set of images, yielding high-quality whole-brain labels across multiple cohorts.

To capture diverse pathologies, we integrate expert-annotated lesions from BraTS, ISLES, ATLAS, and WMH (Baid et al., 2021; Hernandez Petzsche et al., 2022; Liew et al., 2022; Kuijf et al., 2019), covering gliomas, edema, infarcts, and white matter hyperintensities. We further compile rare cases from *Radiopaedia* and manually delineate infarcts, arachnoid and epidermoid cysts, papillomas, and many others. These new annotations also include sub-components like edema, enhancing tissue, and heterogeneous intra-lesion features. Finally, we augment the dataset with a conditional synthesis procedure that generates diverse lesions in healthy brains, broadening the distribution of pathological presentations (Appendix B.5). To support analysis of lesion characteristics, we annotate each lesion with its anatomical location, intensity profile, size, and position relative to surrounding structures, and, when applicable, indicators of diffusion restriction or post-contrast enhancement.

4 Experiments

VoxelPrompt addresses non-standard, open-ended workflows rather than a single fixed task. Its evaluation, therefore, requires a diverse set of complementary experiments. We first present experiments that evaluate VoxelPrompt's ability to generate and execute accurate end-to-end brain analyses across several representative practitioner use-cases. We then present analyses and ablations of modeling decisions. We provide further experimental and test data details in Appendix C and include additional results demonstrating disease characterization performance in Appendix D.

4.1 Brain Image Analysis

***Ad hoc* Neuroimaging Workflow Generation.** Figure 1 shows that a single VoxelPrompt model can execute a wide range of workflows on held out test data, including localizing brain anatomy and pathology regions, extracting intensity metrics and morphology measures within user-specified ROIs, and masking or cropping tissues for focused visualization. The model can compute and compare metrics across ROIs, such as hippocampal asymmetry, normalized subcortical volumes, and acute versus chronic hemorrhage components, as well as track longitudinal changes such as tumor size across scans. By integrating multiple acquisitions, VoxelPrompt can further characterize lesion locations and tissue properties, such as diffusion restriction or post-contrast enhancement. Figure 3A shows that VoxelPrompt facilitates fine-grained specificity by supporting flexible, language-guided analysis, such as isolation or differentiation of lesions in multifocal disease based on signal intensity, size, relative position, or anatomical context (e.g., hemisphere, lobe, etc.). Examples in Figure 1 reflect common practitioner use-cases, but many are qualitative as no benchmark dataset currently exists to evaluate free-form workflow generation outcomes for complex pathology analyses.

Text-prompted Zero-shot Lesion Segmentation. Zero-shot brain lesion segmentation enables medical practitioners and researchers to rapidly localize and quantify pathologies without requiring a disease-specific model. We evaluate VoxelPrompt's zero-shot segmentation capabilities on entirely unseen abnormality datasets using a dataset-specific prompt: "segment the $\langle ROI \rangle$," where $\langle ROI \rangle$ is the target lesion type or informative description. We benchmark our approach against multi-dataset foundation models that include brain pathology segmentation as training tasks. These include volumetric BiomedParse v2 (Zhao et al., 2024; 2025a) and SAT (Zhao et al., 2025b), both of which use text prompts to target abnormalities. Most other existing text-prompted vision models, or vision–language models (VLMs), are limited to question-answering tasks, and do not perform quantitative analyses such as segmentation, thereby precluding them as baselines. We also use MoME (Zhang et al., 2024), a recent generalizable brain abnormality segmentation model. Our evaluation suite spans diverse targets: 30 meningiomas from BraTS-MEN (LaBella et al., 2024), 9 pediatric MRIs of multiple sclerosis lesions from PediMS (Popa et al., 2025), 35 resection cavities from EPISURG (Pérez-García et al., 2020), and 36 hemorrhages from BHSD (Wu et al., 2023). VoxelPrompt has not been trained on these datasets, allowing for zero-shot performance assessment.

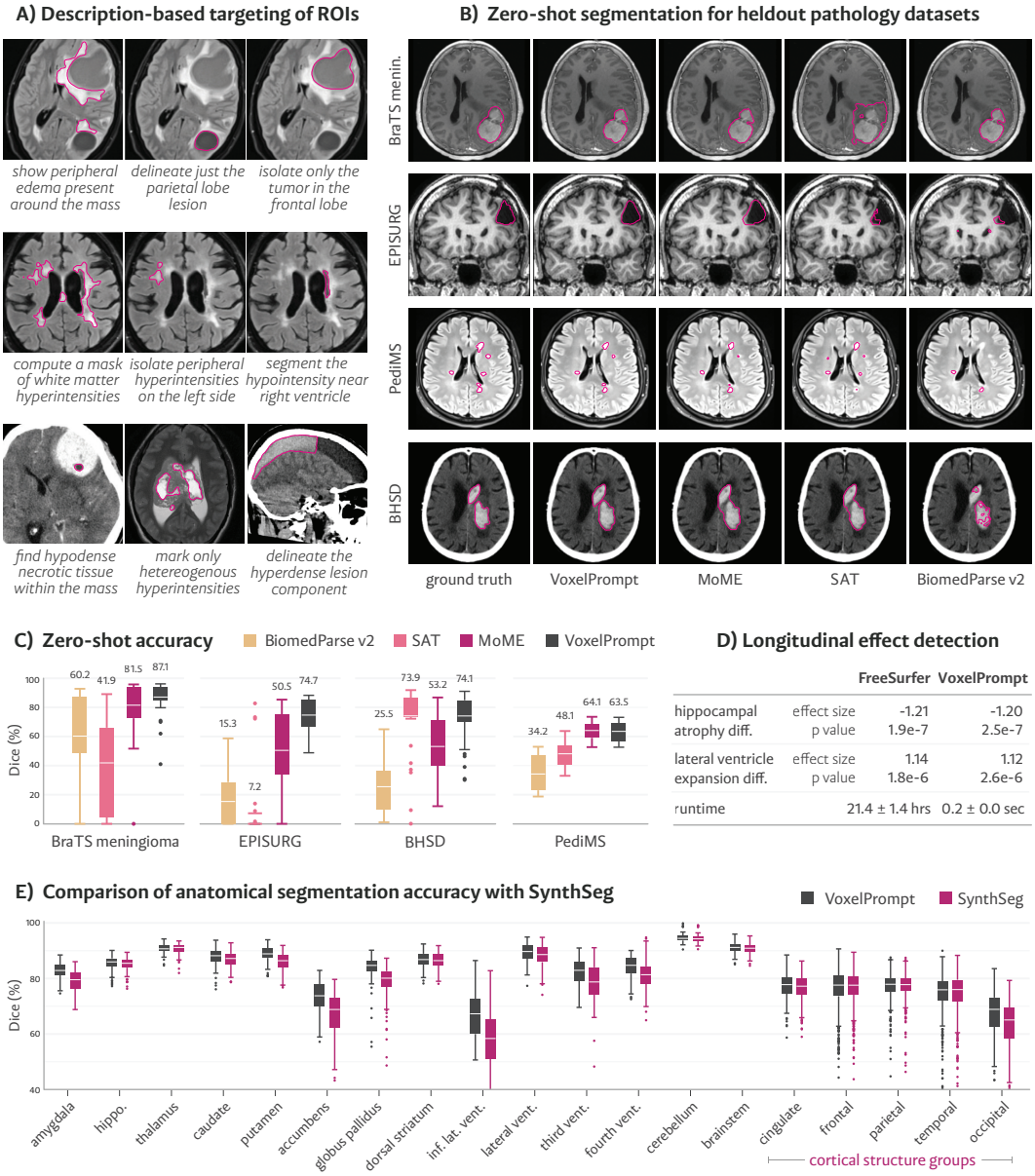

Figure 3: **VoxelPrompt performance. (A)** Free-form text prompts, shown below each image, guide VoxelPrompt to perform targeted analysis and delineation of nuanced, context-specific image regions, even in scans with multiple lesions. **(B, C)** On unseen datasets with diverse brain abnormalities, VoxelPrompt is the only method achieving consistently high-quality results both qualitatively and quantitatively. **(D)** Compared to longitudinal FreeSurfer, VoxelPrompt achieves the same effect size in distinguishing Alzheimer's disease from controls with a $10^5\times$ faster runtime. **(E)** VoxelPrompt outperforms the state-of-the-art specialist model (SynthSeg) on whole brain segmentation.

Figures 3B and C demonstrate that VoxelPrompt is the only method that achieves consistently high performance across all lesion target types, and on average achieves the highest Dice score. We find that no baseline achieves generalization across abnormalities, and the second-best method varies from dataset to dataset. Quantitatively, VoxelPrompt achieves a mean 12.53 Dice points higher than MoME, the overall second-best method. Appendix Figure 6 shows per-subject performances.

Whole Brain Anatomical Segmentation. We evaluate VoxelPrompt's ability to segment diverse neuroanatomical targets. Since many brain structures are bilateral, we prompt VoxelPrompt to "seg-

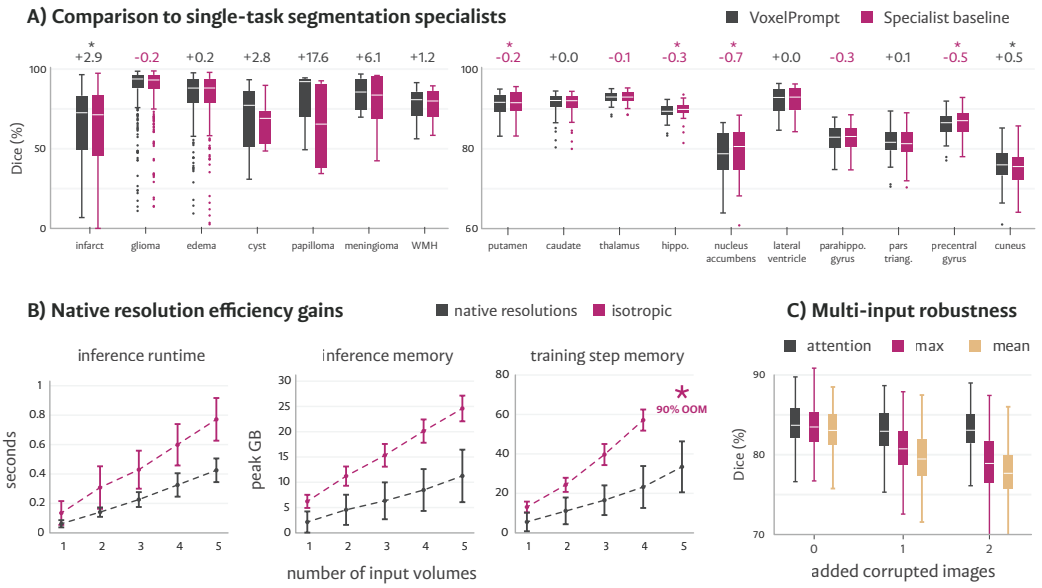

Figure 4: **Ablations and analyses**. **(A)** A single VoxelPrompt model trained jointly on all tasks matches or exceeds task-specific models for both lesions (left) and anatomy (right). Asterisks indicate statistically significant differences. **(B)** Our proposed native-resolution convolutions are more efficient in runtime and memory than isotropic resampling. **(C)** Our attention mechanism for multi-input volume interaction is more robust to image corruptions compared to max and mean reductions.

ment the left and right $\langle ROI \rangle$" to generate joint segmentations, where applicable. We compare against the widely used state-of-the-art SynthSeg v2 (Billot et al., 2023) method for multi-class anatomical segmentation, which generalizes across the diverse acquisition contrasts exhibited in our image dataset. We use a structural MRI test set of 108 unseen volumes, which span various tissue contrasts and contain ground-truth segmentations for the 45 structures predicted by SynthSeg.

Figure 3E shows that VoxelPrompt significantly outperforms SynthSeg ($p < 0.05$) on 23/45 ROIs, with a mean Dice improvement of $+1.1 \pm 2.3\%$ over all structures. While VoxelPrompt achieves a modest improvement, we emphasize that our main goal is not to outperform established tools for segmentation sub-tasks, but rather to provide reliable anatomical segmentations while retaining VoxelPrompt's unique flexibility of natural language prompting for workflows.

Longitudinal Analyses. Figure 1 shows qualitatively that VoxelPrompt performs volumetric analyses across time for various pathologies. Here, we quantify performance for longitudinal analyses of *anatomical* structures, a core component of large neuroimaging studies. We aggregate 100 subjects with two MRI sessions separated by two years from the ADNI dataset (Weber et al., 2021), equally split between controls and Alzheimer's disease (AD) subjects. We assess VoxelPrompt's off-the-shelf ability to measure the change in AD-affected structures over time, and use that to distinguish controls from AD subjects. Specifically, we compare effect sizes and runtime against longitudinal FreeSurfer (Reuter et al., 2012), the widely used standard for multi-session analysis. As demonstrated in Figure 3D, VoxelPrompt can detect well-established AD-related effects, such as increased hippocampal atrophy and increased ventricle volume expansion over time, with a similar effect size to FreeSurfer, while offering a dramatic 3.8×10^5 speedup in runtime.

4.2 Ablations and Analyses

Multi-Task Training. We test whether the proposed *single* VoxelPrompt model trained jointly on multiple tasks can match the performance of single-task specialist networks. We optimize individual, label-specific segmentation networks with the same architecture as the VoxelPrompt vision network, for a subset of distinct segmentation tasks, using a soft Dice loss. Since optimizing a specialized baseline for each ROI in our training dataset is computationally prohibitive, we select a subset of

10 anatomical and 7 pathology targets spanning diverse shapes and locations. In total, the resulting evaluation subset encompasses 638 held-out subjects.

Figure 4A shows that VoxelPrompt performance is on par with ($p > 0.05$) or exceeds ($p < 0.05$) the performance of $13/17$ single-task specialists. The mean Dice difference relative to the specialists is $+4.3 \pm 5.7\%$ for pathology targets and $-0.1 \pm 0.3\%$ for anatomical structures. This shows that multi-task training in VoxelPrompt rivals specialist models, while offering substantial improvement for brain abnormality segmentation, especially for variable lesions and limited data.

Native Resolution Efficiency. We evaluate the efficiency gains of the proposed native-resolution vision network by comparing the processing of images at their native resolution to the standard approach of conforming all inputs to a 1 mm^3 isotropic geometry (Billot et al., 2023). For both approaches, we measure inference runtime, peak GPU memory during inference, and training memory incurred during the backward pass of a single optimization step using the VoxelPrompt model. Image geometries are drawn from a distribution reflecting those encountered in a single clinical scan session (Appendix C.3), and results are averaged over 500 samples. To assess scalability, we repeat the experiment with increasing numbers of input images per sample.

Figure 4B shows that, averaged across all numbers of input volumes, native-resolution processing achieves a $2\times$ reduction in inference runtime and a $2.4\times$ memory reduction compared to isotropic resolution conformation. Isotropic resampling in training incurs $2.2\times$ higher memory costs, rendering it challenging to train a multi-modal model that accepts only isotropic inputs on standard hardware. For example, with 5 input volumes, frequently present in a longitudinal MRI series, isotropic inputs cause out-of-memory errors on 90% of sampled batches on an 80 GB GPU.

Mechanisms of Stream Interaction. We compare our attention-based interaction module with existing non-parametric alternatives (Butoi et al., 2023) that interact features across image inputs using mean or max reductions. To isolate the effect of interaction, we train vision-only models using three mechanisms: attention, mean, and max reduction (Appendix C.5). We train models using synthetic multi-contrast brain images generated from 360 MRIs using a domain randomization pipeline (Gopinath et al., 2024). Synthetic inputs enable us to explicitly test accuracy gains from multi-contrast integration. We also measure robustness to groups of images of variable quality by synthetically corrupting a subset of images in a group. We evaluate using 500 synthetic brains with arbitrary contrasts, measuring average Dice over 35 anatomical labels. We find that all three mechanisms achieve similar overall accuracy improvements as the number of uncorrupted inputs is increased (see Appendix Table 2). However, as shown in Figure 4C, attention interaction is markedly more robust to image quality: when including corrupted inputs, Dice degrades by only $0.6 \pm 3.4\%$, compared to $4.6 \pm 4.5\%$ and $5.4 \pm 4.0\%$ for max- and mean-reduction, respectively.

5 Discussion

Limitations and Future Work. The VoxelPrompt vision and language networks are trained from scratch on simulated user prompts generated from templates, which limits their utility when given entirely unseen prompts. This limitation can be addressed by constructing more diverse datasets containing a broader range of tasks, either through simulations or language instructions from real users. While VoxelPrompt was trained on a combination of public brain imaging datasets and images we aggregated and annotated from the open Internet, its training set can be extended further by inferring tasks from clinical images and their associated reports, which might better cover the complex edge-case pathologies seen in practice. Additionally, instead of training the language model from scratch, a lightweight pre-trained language model with broad programming and natural language knowledge could be finetuned to support generalization to new tasks. We believe that our training strategy is generic and could be productively applied to radiology beyond neuroimaging.

Conclusions. We introduced VoxelPrompt, a vision-language system that can address radiological aims not possible with existing methods as well as tasks that today require a multitude of specialized models and extensive manual user work. We demonstrated that agent-based VoxelPrompt accurately solves a broad spectrum of neuroimaging tasks involving end-to-end image analysis. Moreover, it provides transparent execution steps that can provide users with confidence in its results. We anticipate VoxelPrompt's use in projects, *ad hoc* studies, and clinical pipelines, empowering biomedical users to adopt AI into their medical imaging workflows.

ACKNOWLEDGEMENTS

This work is supported in part by the National Institute of Biomedical Imaging and Bioengineering (R01 EB033773, T32 EB001680), the Harvard MIT Neuroimaging Training Program, the National Science Foundation Graduate Research Fellowships Program, Quanta Computer Incorporated, and Felicis Ventures.

REFERENCES

Syed M Adil, Evan Calabrese, Lefko T Charalambous, James J Cook, Shervin Rahimpour, Ahmet F Atik, Gary P Cofer, Beth A Parente, G Allan Johnson, Shivanand P Lad, et al. A high-resolution interactive atlas of the human brainstem using magnetic resonance imaging. *Neuroimage*, 237: 118135, 2021.

Anahit Babayan, Miray Erbey, Deniz Kumral, Janis D Reinelt, Andrea MF Reiter, Josefin Röbbig, H Lina Schaare, Marie Uhlig, Alfred Anwander, Pierre-Louis Bazin, et al. A mind-brain-body dataset of mri, eeg, cognition, emotion, and peripheral physiology in young and old adults. *Scientific data*, 6(1):1–21, 2019.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*, 2023.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023.

Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. *arXiv preprint arXiv:2305.10799*, 2023a.

Yinda Chen, Che Liu, Wei Huang, Sibo Cheng, Rossella Arcucci, and Zhiwei Xiong. Generative text-guided 3d vision-language pretraining for unified medical image segmentation. *arXiv preprint arXiv:2306.04811*, 2023b.

Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

Steffen Czolbe and Adrian V Dalca. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6217–6230, 2023.

Neel Dey, Benjamin Billot, Hallee E Wong, Clinton J Wang, Mengwei Ren, P Ellen Grant, Adrian V Dalca, and Polina Golland. Learning general-purpose biomedical volume representations using randomized synthesis. *arXiv preprint arXiv:2411.02372*, 2024.

Mohamed S Elmahdy, Laurens Beljaards, Sahar Yousefi, Hessam Sokooti, Fons Verbeek, Uulke A Van Der Heide, and Marius Staring. Joint registration and segmentation via multi-task learning for adaptive radiotherapy of prostate cancer. *IEEE Access*, 9:95551–95568, 2021.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

Karthik Gopinath, Andrew Hoopes, Daniel C Alexander, Steven E Arnold, Yael Balbastre, Benjamin Billot, Adrià Casamitjana, You Cheng, Russ Yue Zhi Chua, Brian L Edlow, et al. Synthetic data in generalizable, learning-based neuroimaging. *Imaging Neuroscience*, 2:1–22, 2024.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujiu Yang, Minlie Huang, Nan Duan, Weizhu Chen, et al. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*, 2023.

Simon Graham, Quoc Dang Vu, Mostafa Jahanifar, Shan E Ahmed Raza, Fayyaz Minhas, David Snead, and Nasir Rajpoot. One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification. *Medical Image Analysis*, 83:102685, 2023.

Douglas N Greve, Benjamin Billot, Devani Cordero, Andrew Hoopes, Malte Hoffmann, Adrian V Dalca, Bruce Fischl, Juan Eugenio Iglesias, and Jean C Augustinack. A deep learning toolbox for automatic segmentation of subcortical limbic structures from mri images. *Neuroimage*, 244: 118610, 2021.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023.

Michael Hanke, Florian J Baumgartner, Pierre Ibe, Falko R Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler. A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie. *Scientific data*, 1(1):1–18, 2014.

Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219: 117012, 2020.

Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9 (1):762, 2022.

Adam Hilbert, Vince I Madai, Ela M Akay, Orhun U Aydin, Jonas Behland, Jan Sobesky, Ivana Galinovic, Ahmed A Khalil, Abdel A Taha, Jens Wuerfel, et al. Brave-net: fully automated arterial brain vessel segmentation in patients with cerebrovascular disease. *Frontiers in artificial intelligence*, 3:552258, 2020.

Malte Hoffmann, Andrew Hoopes, Douglas N. Greve, Bruce Fischl, and Adrian V. Dalca. Anatomy-aware and acquisition-agnostic joint registration with SynthMorph. *Imaging Neuroscience*, 2: 1–33, 06 2024. ISSN 2837-6056. doi: 10.1162/imag_a_00197.

Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: Skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022.

Murtadha D Hssayeni, Muayad S Croock, Aymen D Salman, Hassan Falah Al-Khafaji, Zakaria A Yahya, and Behnaz Ghoraani. Intracranial hemorrhage segmentation using a deep convolutional model. *Data*, 5(1):14, 2020.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al. nninteractive: Redefining 3d promptable segmentation. *arXiv preprint arXiv:2503.08373*, 2025.

Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

Fucai Ke, Joy Hsu, Zhixi Cai, Zixian Ma, Xin Zheng, Xindi Wu, Sukai Huang, Weiqing Wang, Pari Delir Haghighi, Gholamreza Haffari, et al. Explain before you answer: A survey on compositional visual reasoning. *arXiv preprint arXiv:2508.17298*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019.

Dominic LaBella, Omaditya Khanna, Shan McBurney-Lin, Ryan Mclean, Pierre Nedelec, Arif S Rashid, Nourel Hoda Tahon, Talissa Altes, Ujjwal Baid, Radhika Bhalerao, et al. A multi-institutional meningioma mri dataset for automated multi-sequence image segmentation. *Scientific data*, 11(1):496, 2024.

Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, pp. 2019–12, 2019.

Binxu Li, Tiankai Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*, 2024.

Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1):320, 2022.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 525–536. Springer, 2023.

Che Liu, Cheng Ouyang, Yinda Chen, Cesar César Quilodrán-Casas, Lei Ma, Jie Fu, Yike Guo, Anand Shah, Wenjia Bai, and Rossella Arcucci. T3d: Towards 3d medical image understanding through vision-language pre-training. *arXiv preprint arXiv:2312.01529*, 2023.

Chin-Fu Liu, Johnny T. C. Hsu, Xin Xu, Sandhya Ramachandran, Victor Wang, Michael I. Miller, Argye Elizabeth Hillis, Andreia Vasconcellos Faria, Max Steven J. Gregory W. Stephen M. James C. Werner Do Wintermark Warach Albers Davis Grotta Hacke Kang K, Max Wintermark, Steven J. Warach, Gregory W. Albers, Stephen M. Davis, James Charles Grotta, Werner Hacke, Dong-Wha Kang, Chelsea Kidwell, Walter J. Koroshetz, Kennedy R. Lees, Michael H. Lev, David S. Liebeskind, A. Gregory Sorensen, Vincent N. Thijs, Götz Thomalla, Joanna Marguerite Wardlaw, and Marie Luby. Deep learning-based detection and segmentation of diffusion abnormalities in acute ischemic stroke. *Communications Medicine*, 1, 2021.

Peirong Liu, Oula Puonti, Xiaoling Hu, Karthik Gopinath, Annabel Sorby-Adams, Daniel C Alexander, W Taylor Kimberly, and Juan E Iglesias. A modality-agnostic multi-task foundation model for human brain imaging. *arXiv preprint arXiv:2509.00549*, 2025.

Michelle Livne, Jana Rieger, Orhun Utku Aydin, Abdel Aziz Taha, Ela Marie Akay, Tabea Kossen, Jan Sobesky, John D Kelleher, Kristian Hildebrand, Dietmar Frey, et al. A u-net deep learning framework for high performance vessel segmentation in patients with cerebrovascular disease. *Frontiers in neuroscience*, 13:97, 2019.

Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical image analysis*, 72:102102, 2021.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

Inés Mérida, Julien Jung, Sandrine Bouvard, Didier Le Bars, Sophie Lancelot, Franck Lavenne, Caroline Bouillot, Jérôme Redouté, Alexander Hammers, and Nicolas Costes. Cermep-idb-mrxfdg: A database of 37 normal adult human brain [18f] fdg pet, t1 and flair mri, and ct images available for research. *EJNMMI research*, 11(1):1–10, 2021.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Wolfgang M Pauli, Amanda N Nili, and J Michael Tyszka. A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei. *Scientific data*, 5(1):1–13, 2018.

Fernando Pérez-García, Roman Rodionov, Ali Alim-Marvasti, Rachel Sparks, John S Duncan, and Sébastien Ourselin. Simulation of brain resection for cavity segmentation using self-supervised and semi-supervised learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 115–125. Springer, 2020.

Ana Luísa Pinho, Alexis Amadon, Torsten Ruest, Murielle Fabre, Elvis Dohmatob, Isabelle Denghien, Chantal Ginisty, Séverine Becuwe-Desmidt, Séverine Roger, Laurence Laurier, et al. Individual brain charting, a high-resolution fmri dataset for cognitive mapping. *Scientific data*, 5 (1):1–15, 2018.

Maria Popa, Gabriela Adriana Vișa, and Ciprian Radu Șofariu. Pedims: A pediatric multiple sclerosis lesion segmentation dataset. *Scientific Data*, 12(1):1184, 2025.

Marianne Rakic, Hallee E Wong, Jose Javier Gonzalez Ortiz, Beth Cimini, John Guttag, and Adrian V Dalca. Tyche: Stochastic in-context learning for medical image segmentation. *arXiv preprint arXiv:2401.13650*, 2024.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. *arXiv preprint arXiv:2307.06135*, 2023.

Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020.

Jingqing Ruan, Yihong Chen, Bin Zhang, Zhiwei Xu, Tianpeng Bao, Hangyu Mao, Ziyue Li, Xingyu Zeng, Rui Zhao, et al. Tptu: Task planning and tool usage of large language model-based ai agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

Sanjay Subramanian, Medhini Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. Modular visual question answering via code generation. *arXiv preprint arXiv:2306.05392*, 2023.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

David Tellez, Diederik Höppener, Cornelis Verhoef, Dirk Grünhagen, Pieter Nierop, Michal Drozdzal, Jeroen Laak, and Francesco Ciompi. Extending unsupervised neural image compression with supervised multitask learning. In *Medical Imaging with Deep Learning*, pp. 770–783. PMLR, 2020.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30. IEEE, 2017.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11558–11567, 2023b.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023c.

Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023d.

Christopher J Weber, Maria C Carrillo, William Jagust, Clifford R Jack Jr, Leslie M Shaw, John Q Trojanowski, Andrew J Saykin, Laurel A Beckett, Cyrille Sur, Naren P Rao, et al. The worldwide alzheimer's disease neuroimaging initiative: Adni-3 updates and global perspectives. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 7(1):e12226, 2021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any medical image. *arXiv preprint arXiv:2312.07381*, 2023.

Biao Wu, Yutong Xie, Zeyu Zhang, Jinchao Ge, Kaspar Yaxley, Suzan Bahadir, Qi Wu, Yifan Liu, and Minh-Son To. Bhsd: A 3d multi-class brain hemorrhage segmentation dataset. In *International workshop on machine learning in medical imaging*, pp. 147–156. Springer, 2023.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023b.

Xinru Zhang, Ni Ou, Berke Doga Basaran, Marco Visentin, Mengyun Qiao, Renyang Gu, Cheng Ouyang, Yaou Liu, Paul M Matthews, Chuyang Ye, et al. A foundation model for brain lesion segmentation with mixture of modality experts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 379–389. Springer, 2024.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024.

Theodore Zhao, Ho Hin Lee, Alberto Santamaria-Pang, Noel C Codella, Sid Kiblawi, Yu Gu, Yu Fang, Wenxuan Teng, Naiteek Sangani, Ivan Tarapov, Matthew P. Lungren, Matthias Blondeel, Tristan Naumann, Naoto Usuyama, Sheng Wang, Paul Vozila, Hoifung Poon, and Mu Wei. Biomedparse-v : Scaling foundation model for universal text-guided volumetric biomedical image segmentation. In *Foundation Models for 3D Biomedical Image Segmentation Workshop at CVPR*, 2025a.

Ziheng Zhao, Yao Zhang, Chaoyi Wu, Xiaoman Zhang, Xiao Zhou, Ya Zhang, Yanfeng Wang, and Weidi Xie. Large-vocabulary segmentation for medical images with text prompts. *npj Digital Medicine*, 8(1):566, 2025b.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.

Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, et al. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*, 2023.

A Model Implementation Details

We implement VoxelPrompt with PyTorch (Paszke et al., 2019) and use Python as the programming language of the code c and persistent programming environment Ω. To support the wide range of imaging operations required by $\mathcal{T}$, we develop and use a PyTorch library of volumetric medical image utilities, called *Voxel*, available at github.com/dalcalab/voxel.

A.1 Language Agent

We implement the agent model α as a decoder-only transformer stack, using a randomly initialized LLaMA architecture (Touvron et al., 2023; Wolf et al., 2019) with 16 transformer blocks, a hidden representation of size $d = 512$, a linear representation of size 2048, and 32 attention heads. We convert text into an embedding space by splitting character groups into tokens (from a vocabulary of size γ) and mapping them to a sequence of $\mathbb{R}^d$ features via an embedding matrix in $\mathbb{R}^{\gamma,d}$. We use the pre-computed tokenizer released with LLaMA 2, with $\gamma = 32{,}000$.

The language model auto-regressively generates instruction embeddings $\varphi = \varphi^c \shortparallel \varphi^\phi$ based on the input μ. We pass φ^c through a fully-connected layer to obtain text token probabilities $P(c)$, and decode into code c by choosing the maximum probability token at each sequence position. We pass φ^ϕ through a fully-connected layer to compute the vision network modulators ϕ.

To split φ^ϕ and φ^c, we first transform φ embeddings into a sequence of max-probability tokens. We extract φ^ϕ from all sequence positions that immediately follow special token `<MOD>`, and we extract φ^c from all remaining positions. The agent α predicts `<MOD>` and subsequent φ^ϕ features after each volume encoding and generation function argument. We project φ^ϕ embeddings to ϕ using a fully-connected layer with 32 output channels and SiLU activation.

A.2 Persistent Environment

In environment Ω, we predefine a variable corresponding to each volume v. As the code c_i is executed, new variables are defined and retained in Ω, persisting across steps. To guide the next instruction step, c_i can include *read* operations, which extract the value of a variable in Ω and embed it in a representation z_i as feedback in the next state $\mu_{i+1} = \mu_i \shortparallel \varphi_i \shortparallel z_i$.

For each volume v passed through m_{enc}, we reduce the spatial dimensions of the deepest layer output using a global max operator. We pass these pooled features through a fully-connected layer to compute $\varepsilon_v^\circ \in \mathbb{R}^d$. When a *read* instruction is executed on a set of volume encodings $\mathcal{E}$ defined in Ω, we concatenate each $\varepsilon_v^\circ \in \mathcal{E}$ into the feedback embeddings z.

A.3 Volume Encoder and Generator Subnetworks

We implement m_{enc} and m_{gen} as the respective down-sampling and up-sampling arms of a six-level UNet-like model (Ronneberger et al., 2015). Each level consists of a 3D convolutional layer followed by a latent feature ϕ mixing layer and stream interaction layer with $b = 32$, as defined in Section 3.1. All layers use SiLU activations. The spatial outputs at each level are channel-normalized with a group size of four, then max-pooled (m_{enc}) or trilinearly upsampled (m_{gen}) by a factor of two. Convolution kernels have size 3^3, with 32 output channels at the top resolution level and 96 output channels at all other levels.

For all input volume streams, we populate $\mathcal{E}$ with spatial features output at each level in m_{enc}. We use these latent features as skip-connections to corresponding level inputs in the generator m_{gen}, which predicts volumes through a convolutional layer with one output channel. Lastly, we apply the sigmoid activation to generate binary segmentation maps. If multiple volumes from a single scan session are passed as input to VoxelPrompt, we compute a merged, session-specific segmentation by extracting the max values across outputs corresponding to each session.

A.4 Optimization

We train VoxelPrompt using the Adam optimizer (Kingma & Ba, 2014) with an initial learning rate of 10^{-4}, a batch size of one, and 10 gradient accumulation steps on an NVIDIA A100 GPU. We

halve the learning rate after 10^5 steps with no improvement in validation accuracy, stopping training after four sets of learning rate updates. We set the volume loss weight $\lambda = 0.1$.

B Training Data Details

B.1 Image Preprocessing and Augmentation

For each image volume, we normalize intensities within the range $[0, 1]$, conform the data layout to a right-anterior-superior (RAS) orientation, and crop the field of view to a 20 mm margin around the cranial cavity. We co-register all images acquired from each subject using Hoffmann et al. (2024).

In training, we randomly sample up to 8 (or max available) images corresponding to a scan session. We augment images by applying random affine transformations, spatial intensity distortions (bias field simulations, spatial smoothing, k-space corruptions), exponential scaling, lateral anatomical flipping, cropping, anatomical masking, and voxel resizing. We take advantage of our resolution-agnostic vision network and randomly sparsify training data to reduce voxel throughput and significantly reduce total train time. This random sparsification is performed by sampling slice separations from the range [1, 6] mm or by cropping the field of view. We ensure that the target ROI, if applicable, is not removed during this process. Volume sparsification is performed with 50% probability for each sample or when total input voxels exceeds a preset threshold to prevent device memory errors.

B.2 Anatomical Dataset Details

In addition to the pathology datasets outlined in 3.2, we generate segmentations for whole-brain anatomical structures on images from the FSM (Greve et al., 2021), OASIS (Marcus et al., 2007; LaMontagne et al., 2019), Mind Brain Body (Babayan et al., 2019), IBC (Pinho et al., 2018), CERMEP (Mérida et al., 2021), and Forrest Gump (Hanke et al., 2014) cohorts. We select high-quality acquisitions and thoroughly inspect and correct errors in the label maps. Additionally, we make use of multiple image atlases with precomputed segmentations (Adil et al., 2021; Pauli et al., 2018).

B.3 Anatomical Structures

We use segmentations of various anatomical classes, listed below. Bilateral brain structures are defined by two distinct hemisphere-specific labels.

Global tissue classes include the brain, dura, skull cavity, cerebrum, cerebral white matter, cerebral cortex, brainstem, cerebellum, ventricular system, and cerebral spinal fluid (CSF).

Brain sub-structure labels include the amygdala, nucleus accumbens, hippocampus, thalamus, caudate, putamen, dorsal striatum, globus pallidus (externus and internus), basal ganglia, hypothalamus, fornix (body, crus, and column), mammillary body, septal nucleus, subthalamic nucleus, habenula, ventral pallidum, extended amygdala, red nucleus, anterior and posterior commissures, pars compacta, pars reticulata, parabrachial pigmented nucleus, ventral tegmental area, fimbria, septum pellucidum, tectum, pineal gland, superior and inferior colliculus, cerebral peduncle, medullary pyramid, medial lemniscus, cerebellar peduncle (superior, middle, inferior), cerebellar gray matter, and cerebellar white matter.

Ventricular sub-structure labels include the lateral ventricle, inferior lateral ventricle, posterior lateral ventricle, anterior lateral ventricle, atrium, third ventricle, fourth ventricle, interventricular foramen, and cerebral aqueduct.

Cortical sub-region labels include the frontal lobe, parietal lobe, temporal lobe, occipital lobe, cingulate cortex, insular cortex, anterior cingulate cortex, caudal anterior cingulate cortex, rostral anterior cingulate cortex, posterior cingulate cortex, isthmus cingulate cortex, frontal pole, middle frontal gyrus, caudal middle frontal gyrus, rostral middle frontal gyrus, superior frontal gyrus, inferior frontal gyrus, pars opercularis, pars orbitalis, pars triangularis, lateral orbitofrontal cortex, medial orbitofrontal cortex, precentral gyrus, paracentral lobule, inferior parietal lobule, superior parietal lobule, supramarginal gyrus, precuneus, postcentral gyrus, entorhinal cortex, fusiform gyrus, parahippocampal gyrus, temporal pole, inferior temporal gyrus, middle temporal gyrus, superior temporal gyrus, transverse temporal gyrus, cuneus, lingual gyrus, and pericalcarine cortex.

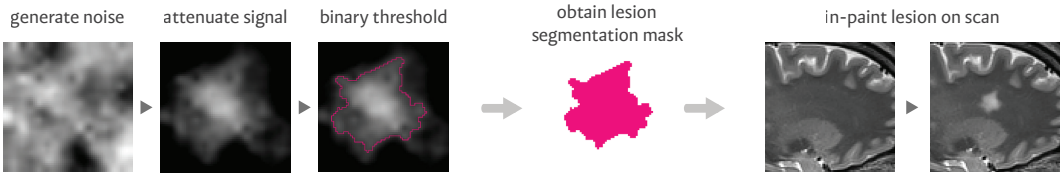

Figure 5: Schematic of the lesion synthesis procedure. A lesion shape is first generated by attenuating and thresholding Brownian noise. The resulting segmentation map is resampled into the target image space, with size and position determined based on anatomical priors. The lesion is in-painted by pasting the tissue mask into the image with procedurally-generated texture and mean signal intensity based on randomly selected relative tissue characteristics.

B.4 Radiopaedia Data

We download and annotate 101 patient cases from Radiopaedia, a radiology reference at *https://radiopaedia.org*. Each case includes text-based notes and scans in the form of 2D image slices. We reconstruct volumetric data by stacking these slices and estimating an affine matrix to map voxel coordinates in world space. We compute this mapping by registering the image to an average template.

B.5 Lesion Synthesis

To extend the range of pathological features observed during training, we synthesize brain lesions with variable characteristics using a model-based domain randomization technique (Gopinath et al., 2024). Images spanning diverse acquisition types from 26 subjects in the OASIS, Mind Brain Body, and CERMEP datasets are paired with whole-brain anatomical segmentation maps as the basis for this process.

For each synthetic lesion, we first sample parameters describing anatomical location, dimensions, intensity relative to surrounding tissue, and texture. As illustrated in Figure 5, volumetric multi-scale Brownian noise is generated with a signal fall-off matched to the sampled lesion dimensions, then thresholded to produce lobulated structures. These shapes define lesion boundaries, which are further constrained by anatomical maps. For example, parenchymal lesions are restricted to white and gray matter, while ventricular lesions are restricted to cerebrospinal fluid spaces. Lesion interiors are inpainted into the native image using randomly generated textures derived from Perlin noise. The mean intensity of the inpainted lesion is determined by the underlying healthy image signal and sampled relative intensity shift.

We also synthesize multiple lesions per subject with varying properties, providing negative examples for VoxelPrompt to learn to differentiate abnormalities based on descriptive features. In addition, we simulate heterogeneous lesions by superimposing secondary Brownian noise–derived masks within an existing lesion, producing intra-lesional components or heterogeneous signal profiles.

C Experimental Details and Data

C.1 Segmentation Evaluation Data

In the table below, we summarize the number of unique images from held-out subjects used when comparing VoxelPrompt segmentation accuracy to SynthSeg (Billot et al., 2023) and individual specialist models. These evaluations use the same set of anatomical test images (Appendix B.2), so we group them together below.

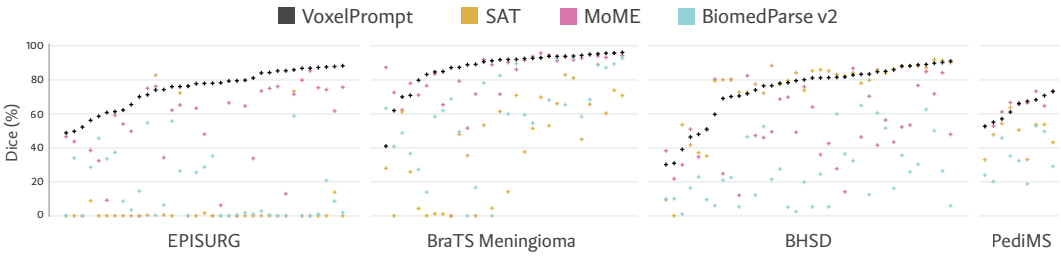

Figure 6: Per-subject accuracy for zero-shot lesion segmentation. Subject indices are sorted along the x-axis in ascending order of VoxelPrompt prediction accuracy. We find that VoxelPrompt consistently achieves high-quality segmentation across all evaluation datasets and lesion types considered.

Table 1: Numbers of held-out test images and subjects corresponding to the whole brain anatomical segmentation experiment in Section 4.1 and the multi-task training ablation in Section 4.2.

Segmentation Target	Images	Subjects
infarct	206	206
glioma	1376	344
edema	1386	347
cyst	24	11
papilloma	16	6
meningioma	21	8
white matter hyperintensities	40	20
anatomical structures	108	40

C.2 Zero-Shot Lesion Segmentation Baselines

Baseline settings. For the pathology segmentation experiments, we evaluate baseline performance across diverse input configurations to ensure fairness and avoid bias. For both BiomedParse v2 (Zhao et al., 2025a) and SAT (Zhao et al., 2025b), we explore a range of prompting strategies on each dataset, following the formats recommended in the original repositories or used during their training. We test multiple levels of pathology classification terminology and adopt the phrasing that yields the best performance. BiomedParse v2 does not specify a preferred anatomical orientation, so we evaluate across all possible orientations and report the best-performing one. We also find that BiomedParse v2 does not benefit from resampling inputs to isotropic resolution. The MoME (Zhang et al., 2024) baseline requires skull-stripping and affine alignment to the MNI template prior to prediction. We follow this preprocessing and report MoME's results in the original coordinate system.

The dataset-specific prompts used as input to language-conditioned methods are described below.

	VoxelPrompt	BiomedParse V2	SAT
BraTS menin.	segment the hyperintense mass	MRI imaging of a brain tumor	meningioma
EPISURG	segment the hypointense lesion	MRI imaging of a brain lesion	stroke
BHSD	segment the hyperdense lesions	CT imaging of brain lesions	intracranial hemorrhage
PediMS	segment hyperintensities	MRI imaging of lesions	white matter hyperintensities

Additional results. In Figure 6, we show per-subject performance plots for all methods on zero-shot lesion segmentation. VoxelPrompt consistently outperforms the baseline foundation models for brain pathology segmentation.

C.3 Scan Geometry Sampling

To evaluate the efficiency improvements provided by the native-resolution vision network, we sample test images with spatial geometries drawn from distributions representative of clinical MR and

CT brain acquisitions. Image dimensions are sampled uniformly around a $155 \times 190 \times 165$ mm field of view, with a max deviation of ± 15 mm in each dimension. This range is derived from the 10^{th} and 90^{th} percentile values of image sizes in our preprocessed datasets.

To reflect real-world voxel resolutions, we consider imaging modalities most frequently collected in a standard clinical brain imaging session. For each acquisition type, we define uniform sampling ranges for in-plane resolution and slice separation, based on standard protocol guidelines and empirical resolution distributions observed in our dataset. Voxel spacings are clamped to a minimum of 0.8 mm. During each experimental sample, we randomly select a field of view, acquisition type, and resolution from these distributions, and randomly populate the volumes with Gaussian noise. This distribution, outlined below, is not exhaustive, but is designed to provide a representative coverage of common acquisitions sufficient to evaluate the benefits of native-resolution processing.

	in-plane spacing (mm)	slice separation (mm)
T1-weighted (isotropic)	0.8 – 1.2	iso.
T2-weighted	0.8 – 1.0	3.0 – 5.0
FLAIR	0.8 – 1.0	3.0 – 5.0
diffusion-weighted (DWI)	1.5 – 2.5	2.0 – 3.5
gradient-echo (GRE)	0.8 – 1.2	4.0 – 6.0
perfusion MRI	1.5 – 2.5	4.0 – 6.0
susceptibility-weighted (SWI)	0.8 – 1.0	1.5 – 3.0
CT (isotropic)	0.8 – 1.0	iso.
CT (thick slice)	0.8 – 1.0	3.0 – 6.0

C.4 Image Synthesis for Evaluating Stream Interaction

Brain image synthesis techniques are used to train neuroimaging models that are robust to acquisition variability and anatomical differences (Gopinath et al., 2024). These approaches employ domain randomization methods (Tobin et al., 2017), in which whole-brain anatomical segmentations are mapped to randomized tissue intensities, warped by spatial transformations, and augmented with simulated artifacts. The resulting synthetic images extend beyond the realistic range of clinical scans, enabling models to generalize across diverse real-world data and tasks (Dey et al., 2024).

We adopt this strategy as a controlled framework for evaluating VoxelPrompt's ability to integrate complementary information across arbitrary numbers of input volumes. By generating multiple synthetic images from a single anatomical label map, we test whether segmentation accuracy improves as additional inputs are provided.

For stream-interaction evaluation, we employ a standard image synthesis protocol widely used in brain imaging (Gopinath et al., 2024). Briefly, for each evaluation sample, we sample a whole-brain anatomical segmentation map from a set of OASIS subjects. To generate an individual image from this segmentation, each label is assigned an intensity distribution defined by Gaussian parameters sampled uniformly, following a classical Bayesian segmentation formulation. Voxel labels are then recoded into grayscale values drawn from their respective label distributions, producing synthetic images. Finally, we apply random artifact simulations, including spatial blurring, additive noise, and bias-field distortion.

To generate corrupted images, we synthesize random label maps from multi-scale Brownian noise. Between 10 and 20 noise maps are generated, and voxels are assigned to the index of the maximum-valued map. This synthetic label map is then converted to an image using the same label-to-intensity procedure described above.

C.5 Stream Interaction Variants

We compare our attention-based stream interaction module with two reduction-based variants commonly used in multi-input medical image analysis (Butoi et al., 2023). In these variants, features are aggregated across input streams by mean or max pooling along the stream dimension. The resulting global feature representation is then concatenated channel-wise with the original stream-specific features. This combined feature map is passed through a linear projection layer whose output di-

mensionality matches that of the block input, ensuring compatibility with the downstream network. We test both mean and max-reduction as alternative aggregation operations to our proposed attention mechanism.

To compare interaction mechanisms, we implement vision models matching the VoxelPrompt architecture but without language-conditioning blocks, varying only the stream interaction module. Models are trained on synthetic data (Appendix C.4) using multi-class Soft Dice loss to segment 35 anatomical brain structures. At each training step, between one and three images corresponding to a single segmentation target are sampled, with a 10% probability of including a corrupted image (described in Appendix C.4).

For evaluation, we generate 500 synthetic image sets, each containing three images derived from a held-out test set of 50 subjects and a predefined subset of corrupted images. We evaluate model performance by varying the number of images provided as input from each set and measuring Dice overlap between predicted and reference segmentations.

Table 2: On *uncorrupted* input image sets, attention- and reduction-based stream interaction methods result in similar model segmentation accuracy (Dice), which improves for all methods as image inputs are added for a single forward pass.

Method	1 image	2 images	3 images
attention	83.7 ± 2.8	85.8 ± 1.8	86.9 ± 1.4
max-reduction	83.5 ± 2.6	85.8 ± 1.8	87.1 ± 1.5
mean-reduction	83.0 ± 2.7	85.7 ± 1.8	86.7 ± 1.6

D Pathology Characterization Experiment

We evaluate the ability of VoxelPrompt to produce a natural language characterization of image features. We focus on five pathology-based visual question-answering tasks (also used during training). These involve classifying lesions based on (1) signal intensity, (2) broad cerebral location, (3) stroke-affected vascular territory, (4) diffusion restriction, and (5) post-contrast enhancement.

Data. For each of these tasks, we curate a subset of held-out subjects with relevant features, while ensuring equal representation of possible classification categories in each subset. In total, the evaluation set consists of 102 cases, with per-task breakdowns detailed below.

Classification Task	Images	Subjects
characterize lesion signal intensity	26	26
identify lesion cerebral location	112	30
identify infarct vascular territory	16	12
detect diffusion restriction	28	14
detect post-contrast enhancement	40	20

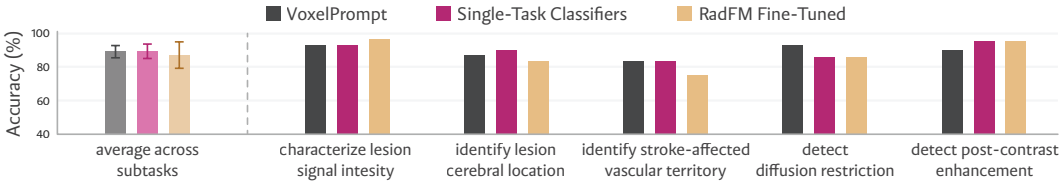

Figure 7: Accuracy of pathology characterization using natural language for five separate classification subtasks. Average subtask accuracy is shown on the left. VoxelPrompt (black) parallels the performance of individually-trained, single-task classifiers (purple) and a fine-tuned RadFM model (yellow) – a state-of-the-art method for 3D visual question-answering.

Evaluation. During evaluation, we consider a prediction as correct if the output natural language response exactly matches the expected characterization. Using a paired *t*-test, we compare the VoxelPrompt per-task classification accuracy to that of multiple baselines.

Baselines. We compare VoxelPrompt to a set of classifier benchmarks, each trained for one of the five pathology characterization tasks in $\mathcal{T}$. As opposed to using language, the single-task benchmark models directly predict label probabilities for a fixed set of task-specific characterizations. We implement these models using the architecture of m_{enc}, with ϕ mixing layers replaced. We reduce the spatial dimensions of the deepest encoder layer output using a global max operator, then compute the maximum value over all input volume streams. To compute classification probabilities for n possible descriptions, we pass the stream-pooled features to a fully-connected layer with n output channels and softmax activation. During benchmark optimization, we use the categorical cross-entropy loss on these predicted probabilities.

We also compare VoxelPrompt to RadFM (Wu et al., 2025), a publicly released, state-of-the-art architecture for medical visual question answering that can process multiple 3D images simultaneously. In our preliminary experiments, we find that the pretrained RadFM cannot generalize to the neuroimaging tasks used in this experiment. Therefore, we *fine-tune* RadFM on our subset of pathology characterization tasks, using the training code released with their pretrained model weights. To fit the optimization within 80 GB of GPU memory, we keep only the first eight hidden transformer layers of the language model and do not modify any other model components. As required by the vision transformer, we resize all input volume spatial dimensions to the nearest multiple of $32 \times 32 \times 4$. During fine-tuning, we use only the expected language response (without code) as the target text.

Results. Figure 7 shows that VoxelPrompt achieves an average classification accuracy of $89.0 \pm 3.6\%$ over all tasks, matching the performance of the single-task benchmarks ($89.3 \pm 4.2\%$) as well as the fine-tuned RadFM model ($87.1 \pm 7.9\%$). These results demonstrate that VoxelPrompt can achieve language-based image characterization with comparable performance to specialized classification and medical vision-language architectures, while also able to perform the wide variety of tasks described in the main experiments.