VoxelPrompt: A Vision-Language Agent for Grounded Medical Image Analysis

Andrew Hoopes [1,2], **Victor Ion Butoi** [1], **John V. Guttag** [1], **Adrian V. Dalca** [1,2,3]

1. Massachusetts Institute of Technology, **2.** Massachusetts General Hospital, **3.** Harvard Medical School

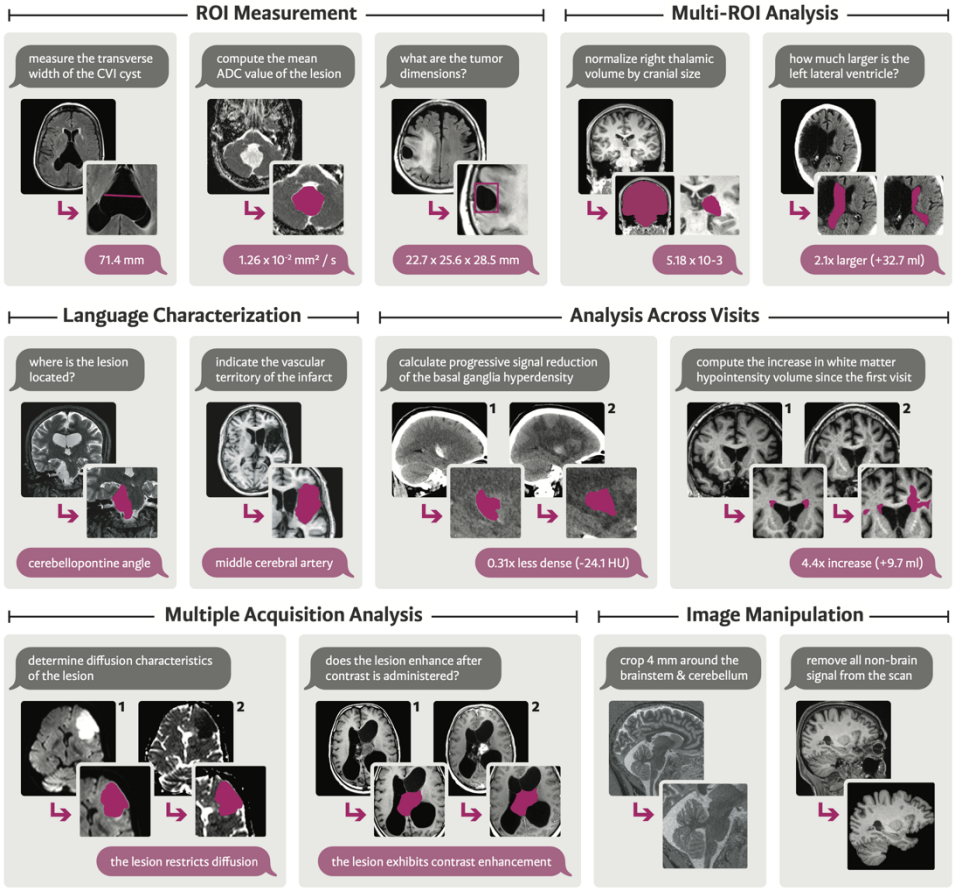

Figure 1. Examples from the diverse set of tasks supported by the VoxelPrompt framework. For each example, we show the input prompt (gray) above the input images(s). VoxelPrompt annotates the images and generates language responses (shown in purple). These scans are processed entirely in 3D, but here we show only a single extracted slice.

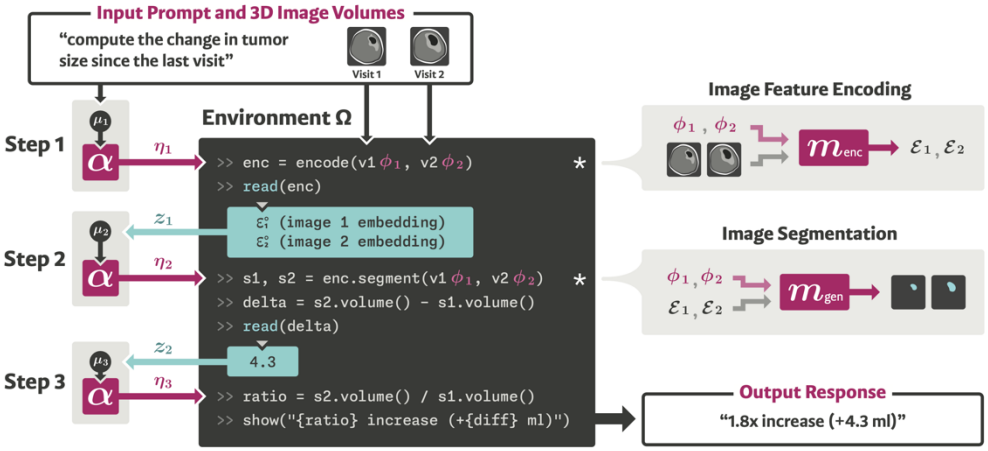

Figure 2. To solve a language-prompted task, the adaptive agent model α outputs instructions η as code to run in a persistent execution environment Ω. Across multiple steps, the agent interprets execution outcomes z (blue) to guide subsequent instruction prediction. To perform vision operations, such as volume encoding or generation, α can instruct the execution of vision networks m_{enc} and m_{gen}, which are manipulated by image-specific latent instruction embeddings ϕ.