The other vision-based model we propose is based on the Segment Anything Model (SAM). SAM is a foundation model for image and videos segmentation. Developed by Meta's FAIR (Fundamental AI Research) lab, SAM represents a significant advancement in computer vision, designed to serve as a versatile framework for various segmentation tasks. His architecture can be summarized in three parts: a Vision Transformer image encoder, a prompt encoder, and a modified mask Transformer decoder block (see Figure 1).

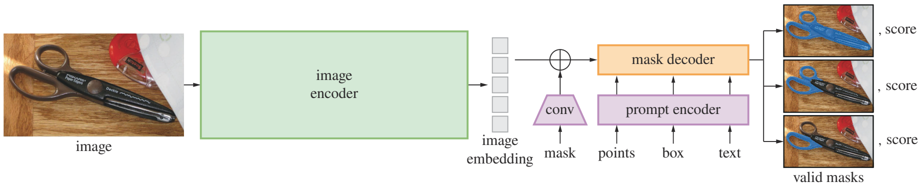

Figure 1: Segment Anything Model (SAM) overview

Now, to address the task, we first produce a high-quality image using the GeoJSON-to-Image pipeline presented earlier. Then, we use SAM to infer masks on the image as segmentation of all the rooms. Finally, the contours of the masks are extracted, and after rescaling the contour coordinates, we generate the corresponding GeoJSON file. This process essentially reverses the GeoJSON-to-Image pipeline, completing the cycle from GeoJSON to image and back to GeoJSON. In Table 1, we find the SAM model checkpoint used.

Checkpoint	`sam_vit_h_4b8939.pth`
Backbone Size	Huge (ViT-H)
Model Parameters	636 million
File Size	2.4 GB
GPU Memory Requirement	At least 8 GB
Use Case	Tasks requiring high-quality segmentations

Table 1: Model Checkpoint Information