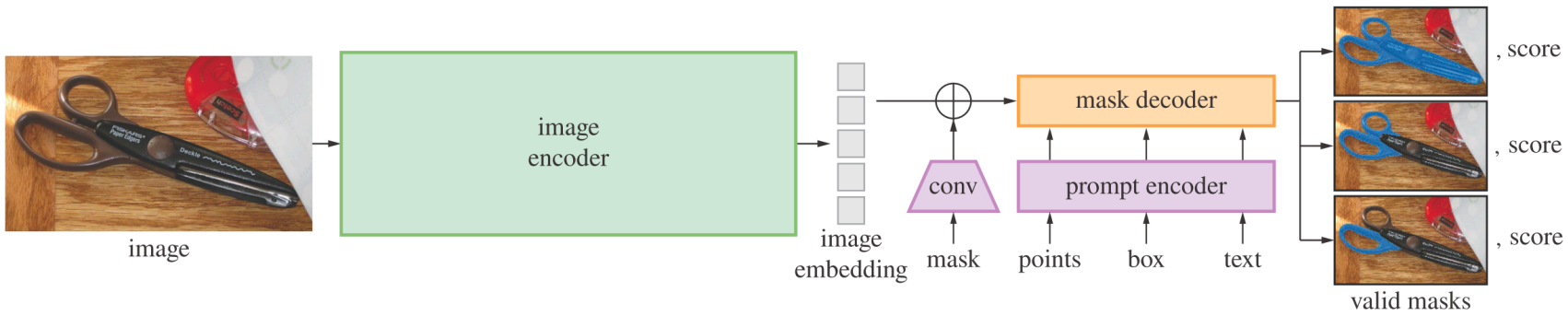

Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

image encoder：MAE based ViT model, run once per image (prior to prompting models).

Prompt encoder:

1. Prompts set: point, box, mask and free text
2. Mask + conv (sum with image embedding)
3. Present Point/box with positional encoding+learning embedding
4. Free text->CLIP feature

Segment anything Data Engine

Assisted-manual stage.

1. First round: SAM powered by common segmentation dataset, classic interactive annotation using SAM. Annotation time is 30 seconds per image.
2. Retrain SAM with the dataset of step 1. Retrain the model 6 time, including image encoder from ViT-B to ViT-H. Annotation time down to 14 sec. (6.5 * Faster labeled than COCO, but with average masks from 20 to 44, Dataset size generate 4.3m masks from 120k images)

Semi-automatic stage

Aim to increase diversities.

1. Detected confidence masks (based on bounding box detector on all first stage masks labeled with objects)
2. presented annotators with images prefilled with these masks and asked them to annotate any additional unannotated objects.
3. Train 5 round, time down to 34 sec, mask numbers increase to 72 per image

Fully automatic stage

Ambiguity-aware

1. select confident and stable masks
2. NMS + multiple overlapping zoomed-in image crops
3. 1.1 B masks on 1.1 m images.