SAM 2: Segment Anything in Images and Videos

DISCLAIMER: Summarized by AI

Problem they are trying to solve / Purpose of method

The SAM 2 model aims to advance **promptable visual segmentation (PVS)** by extending the capabilities of the original SAM (Segment Anything Model) from static image segmentation to dynamic video segmentation.

Previous problems:

- Original SAM handled static images only, ignoring the temporal complexity of videos.
- Video segmentation is more challenging due to occlusions, motion blur, deformations, and lighting changes.
- Existing video segmentation models lacked generalization, struggled with sparse prompts, and had inefficient refinement mechanisms.

Why SAM 2 is needed:

- There is a growing need for a **unified segmentation system** that works seamlessly across images and videos.
- Applications in AR/VR, robotics, video editing, and autonomous systems require accurate and efficient **spatio-temporal segmentation**.
- Existing datasets and tools were limited in scale and diversity, hindering progress in segmenting "anything" in videos.

How does it differ from other methods?

- **Unified model**: SAM 2 treats an image as a single-frame video and works on both images and videos using the same architecture.
- **Streaming memory architecture**: Unlike tracker-based methods, SAM 2 includes a memory bank that allows it to maintain object context across frames.
- **Interactive refinement**: Users can provide additional prompts (points, boxes, masks) on any frame to refine results without restarting the segmentation.
- **Efficient annotation**: Paired with a new data engine and the large-scale SA-V dataset, annotation becomes 8.4× faster than frame-by-frame annotation.
- **Superior performance**:
  - Outperforms SAM in image segmentation (6× faster, more accurate).
  - Outperforms prior video segmentation methods with 3× fewer interactions.
  - Achieves state-of-the-art results on multiple video and image benchmarks.

How the method works

Simple overview:

SAM 2 extends SAM's image segmentation capabilities to video by adding:

- A **memory mechanism** to track object appearances across time.
- A **promptable architecture** that accepts user inputs (clicks, boxes, masks) at any point.
- A **data engine** to build a large-scale, diverse segmentation dataset (SA-V) for training and evaluation.

Detailed explanation:

1. **Input & Task**:
   - Takes a video (or image) and a prompt on any frame.
   - Predicts a "masklet": the segmented object across frames.
   - Supports iterative refinement through new prompts.
2. **Architecture Components**:
   - **Image Encoder**: Uses a MAE-pretrained Hiera encoder for multi-scale frame features.
   - **Prompt Encoder**: Converts user prompts into embeddings.
   - **Mask Decoder**: Predicts masks using frame features and prompts.
   - **Memory Encoder & Bank**:
     - Stores spatio-temporal information from previous frames.
     - Enables tracking objects through occlusion or motion.
   - **Memory Attention Module**: Allows the model to condition current frame predictions on past memory and prompts.
   - **Occlusion Detection**: Predicts whether an object is visible in a frame.
3. **Data Engine & SA-V Dataset**:
   - 3 annotation phases:
     1. Manual per-frame using SAM.
     2. Assisted propagation using SAM + SAM2Mask.
     3. Interactive propagation using full SAM 2.
   - Results in **50.9K videos and 35.5M masks**, including auto-generated annotations.
   - Significantly larger and more diverse than previous datasets (e.g., DAVIS, YouTube-VOS).
4. **Training**:
   - Joint training on both image (SA-1B) and video (SA-V, Internal, VOS datasets).
   - Simulated interactive training: random prompts, corrections sampled over 8-frame sequences.
5. **Performance**:
   - Outperforms state-of-the-art in zero-shot and semi-supervised settings.
   - Works efficiently in real-time with high accuracy.