Vision-Language Models

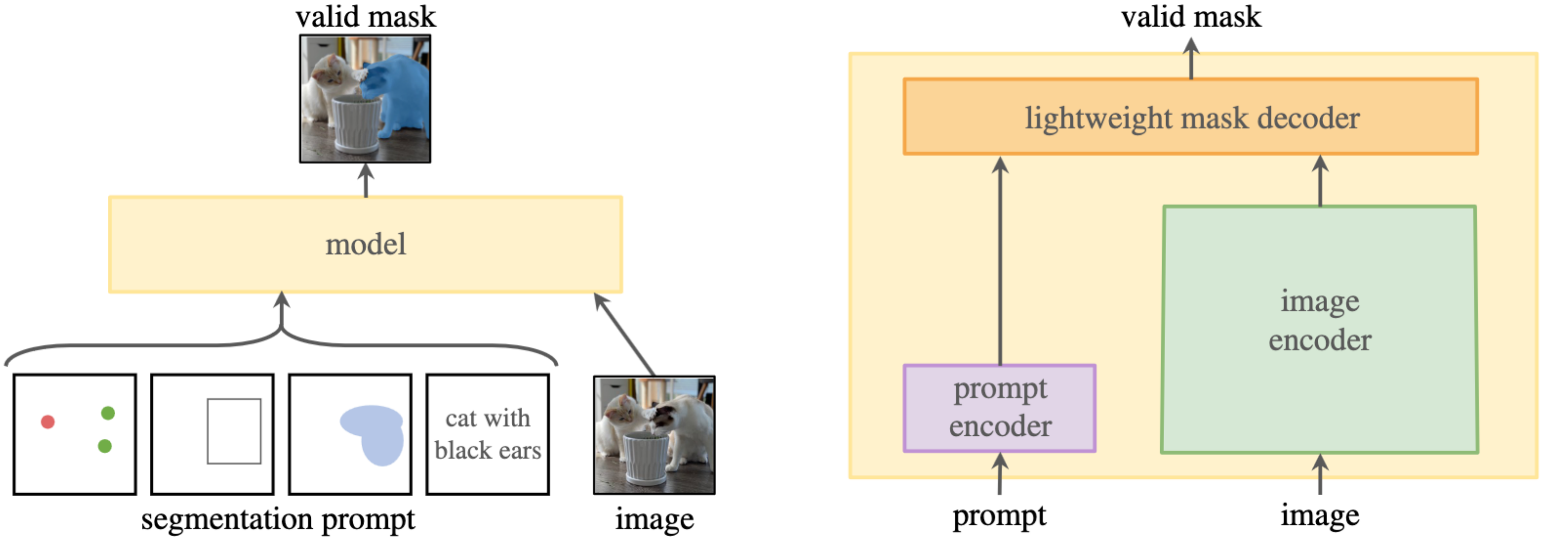
Part II:

VLMs using LLMs

SAM: Transformers for segmentation in image and video

Segment Anything

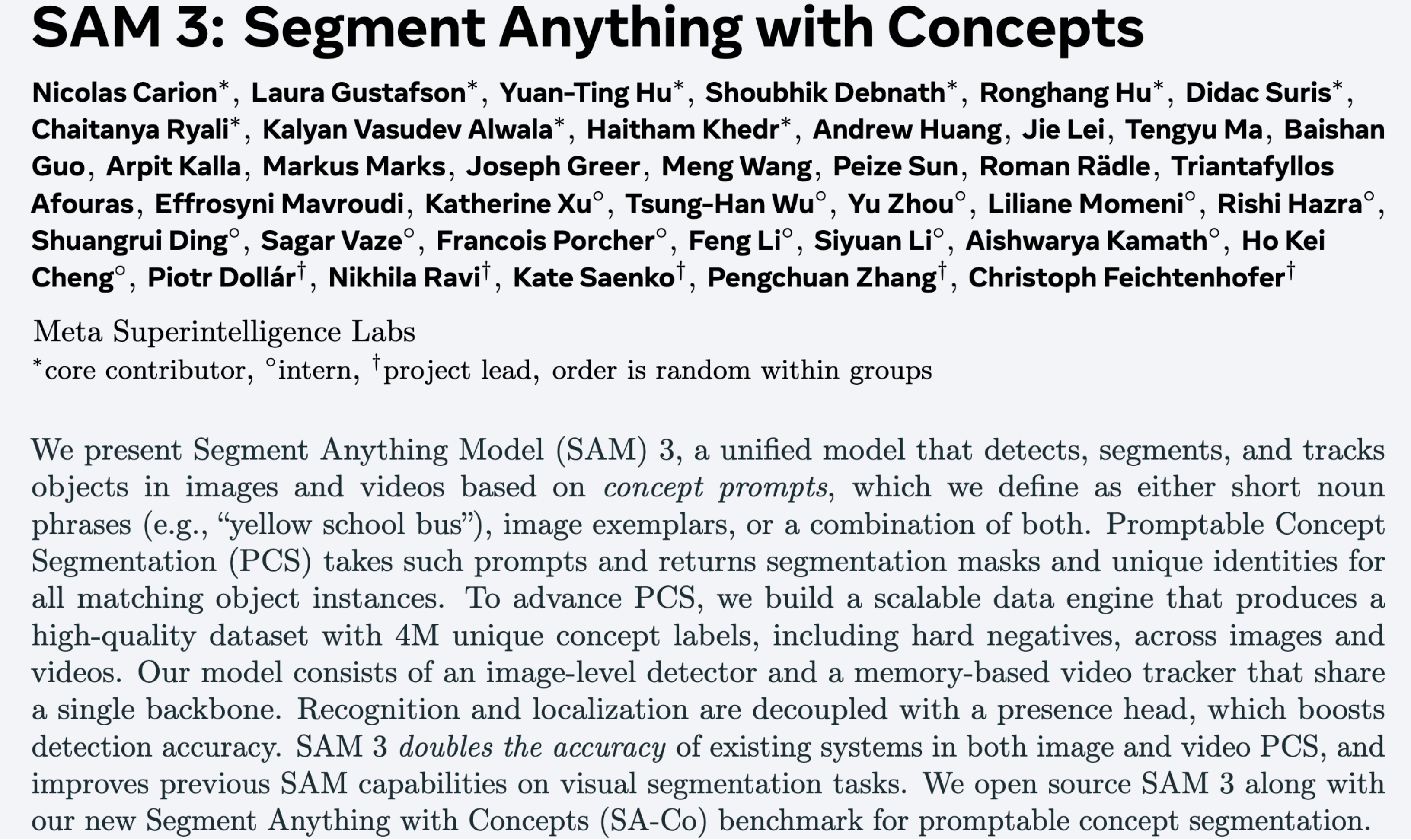
Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]



(a) **Task**: promptable segmentation

(b) **Model**: Segment Anything Model (**SAM**)

SAM: Transformers for segmentation in image and video



SAM 3: Segment Anything with Concepts

Nicolas Carion[*], Laura Gustafson[*], Yuan-Ting Hu[*], Shoubhik Debnath[*], Ronghang Hu[*], Didac Suris[*], Chaitanya Ryali[*], Kalyan Vasudev Alwala[*], Haitham Khedr[*], Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu[°], Tsung-Han Wu[°], Yu Zhou[°], Liliane Momeni[°], Rishi Hazra[°], Shuangrui Ding[°], Sagar Vaze[°], Francois Porcher[°], Feng Li[°], Siyuan Li[°], Aishwarya Kamath[°], Ho Kei Cheng[°], Piotr Dollár[†], Nikhila Ravi[†], Kate Saenko[†], Pengchuan Zhang[†], Christoph Feichtenhofer[†]

Meta Superintelligence Labs
[*]core contributor, [°]intern, [†]project lead, order is random within groups

We present Segment Anything Model (SAM) 3, a unified model that detects, segments, and tracks objects in images and videos based on *concept prompts*, which we define as either short noun phrases (e.g., "yellow school bus"), image exemplars, or a combination of both. Promptable Concept Segmentation (PCS) takes such prompts and returns segmentation masks and unique identities for all matching object instances. To advance PCS, we build a scalable data engine that produces a high-quality dataset with 4M unique concept labels, including hard negatives, across images and videos. Our model consists of an image-level detector and a memory-based video tracker that share a single backbone. Recognition and localization are decoupled with a presence head, which boosts detection accuracy. SAM 3 *doubles the accuracy* of existing systems in both image and video PCS, and improves previous SAM capabilities on visual segmentation tasks. We open source SAM 3 along with our new Segment Anything with Concepts (SA-Co) benchmark for promptable concept segmentation.

SAM: Transformers for segmentation in image and video

Promptable Concept Segmentation (PCS): given an image or short video (≤30 secs), detect, segment and track all instances of a visual concept specified by a short text phrase, image exemplars, or a combination of both.

Concepts restricted to those defined by simple noun phrases (NPs) consisting of a noun and optional modifiers.

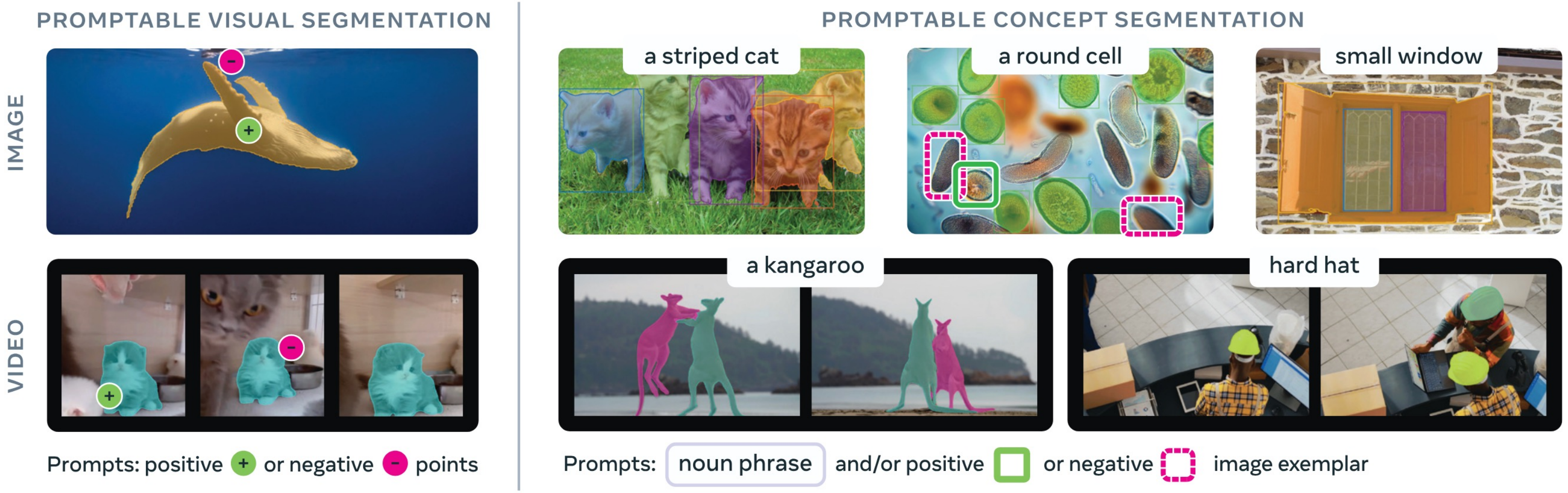
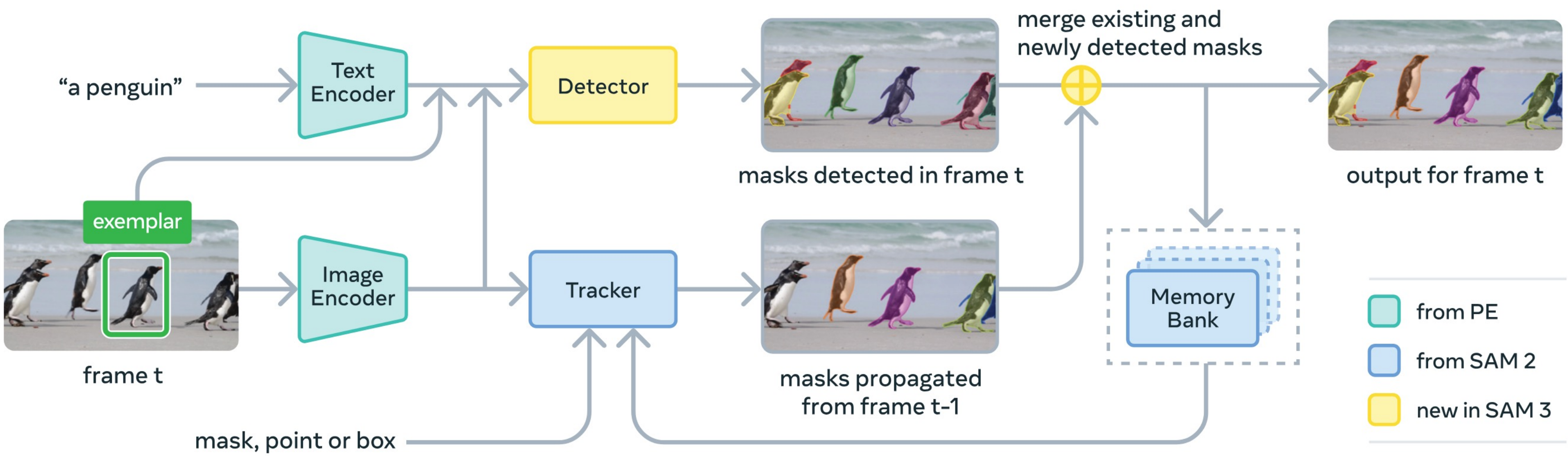

PROMPTABLE VISUAL SEGMENTATION
IMAGE
VIDEO
Prompts: positive + or negative − points
PROMPTABLE CONCEPT SEGMENTATION
a striped cat
a round cell
small window
a kangaroo
hard hat
Prompts: noun phrase and/or positive or negative image exemplar

Figure 1 SAM 3 improves over SAM 2 on promptable *visual* segmentation with clicks (left) and introduces the new promptable *concept* segmentation capability (right). Users can segment all instances of a visual concept specified by a short noun phrase, image exemplars (positive or negative), or a combination of both.

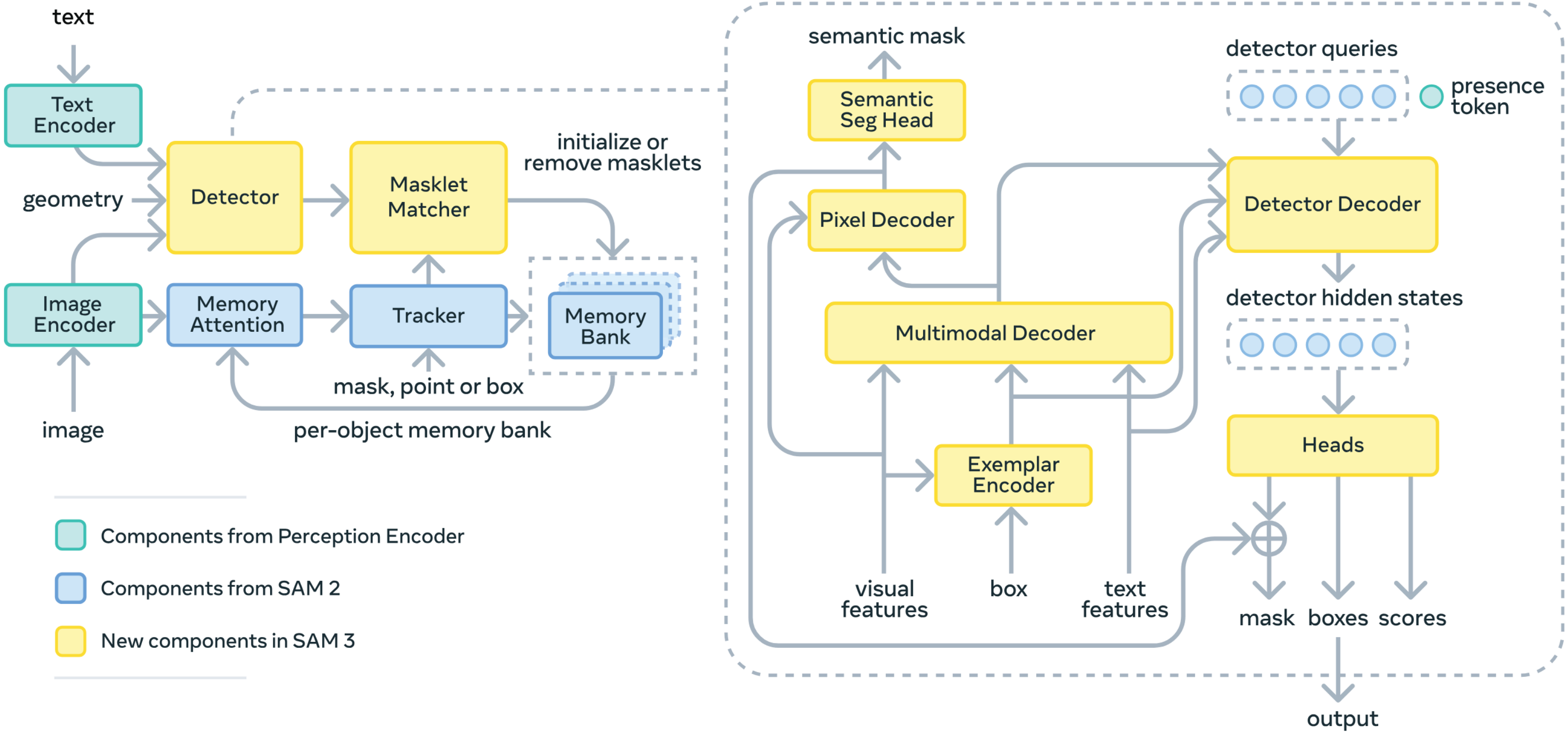
SAM: Transformers for segmentation in image and video

The detector and tracker ingest vision-language inputs from an aligned Perception Encoder (PE) backbone



SAM: Transformers for segmentation in image and video

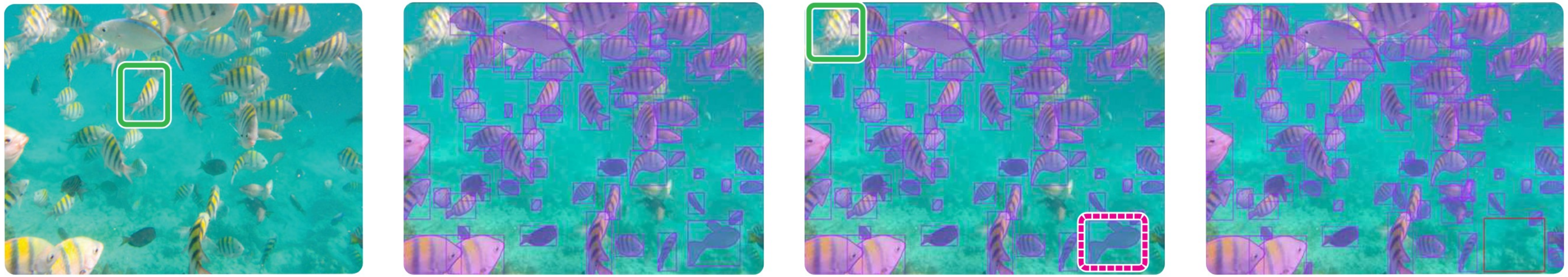
The fusion encoder accepts the unconditioned embeddings from the image encoder and conditions them by cross-attending to the prompt tokens.
The fusion is followed by a CAIT-like Perceiver-like DETR-like decoder, where learned object queries cross-attend to the conditioned image embeddings from the fusion encoder.



SAM: Transformers for segmentation in image and video

Noun-phrase prompts (when provided) are global to all frames of the image/video

Image exemplars can be provided on individual frames as positive or negative bounding boxes to iteratively refine the target masks



Text: “a fish” and/or positive exemplar (green)

INITIAL PROMPT

Masks (or masklets in video) for each detected instance

OUTPUT

1+ positive (green) or negative (red) exemplars

REFINEMENT PROMPTS

Refined mask(lets) for each detected instance

OUTPUT