Segmentar cualquier objeto

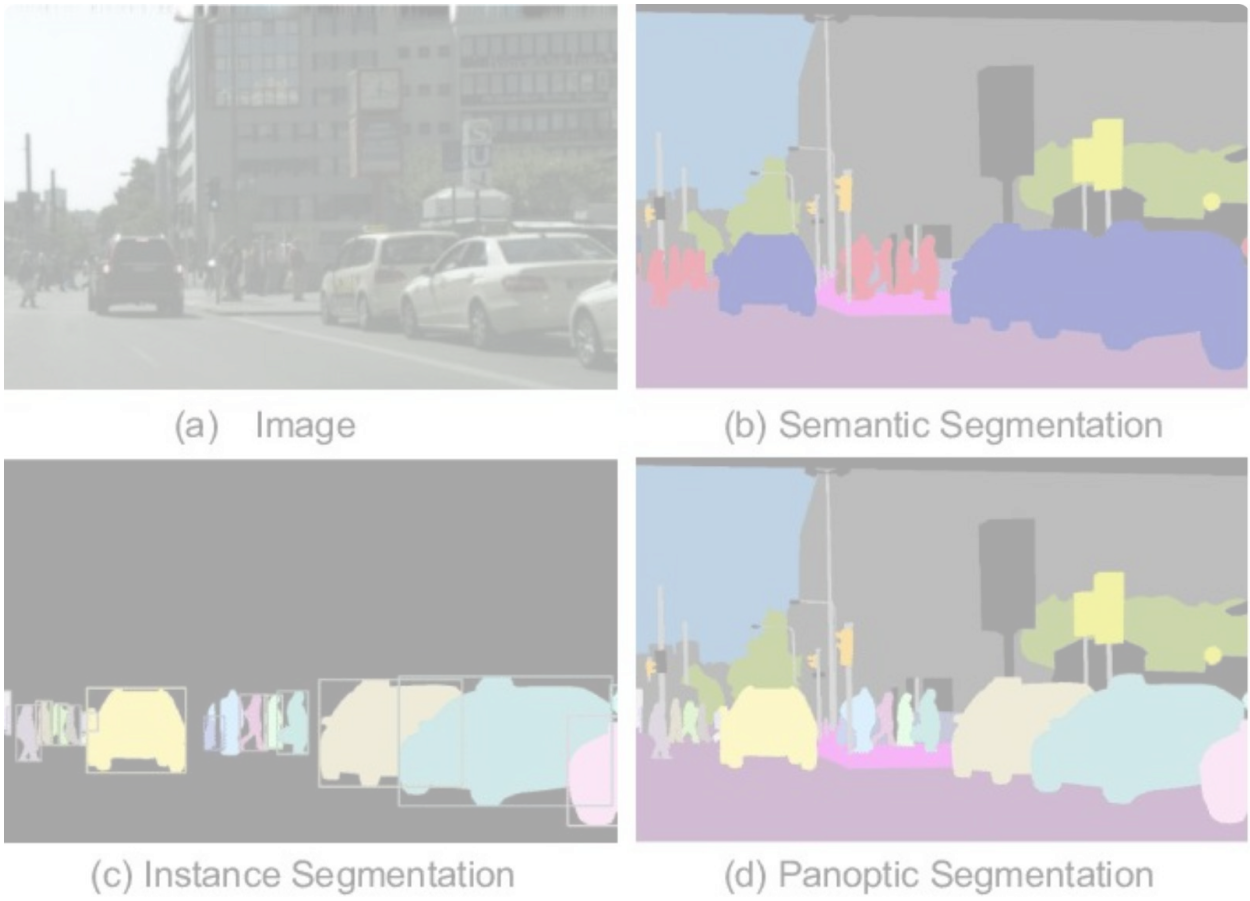
La familia SAM convierte una indicación visual o textual en máscaras precisas.

SAM 1 trabaja con imagen; SAM 2 añade vídeo; SAM 3 incorpora conceptos abiertos.

Segment Anything es una familia de modelos promptables. Las capacidades dependen de la versión: no todos los checkpoints aceptan texto ni mantienen memoria temporal.

Una decisión por píxel

La máscara delimita la forma del objeto, mientras una caja solo aproxima su posición.



El mismo píxel puede recibir categoría e identidad según la tarea.

La segmentación permite medir contornos, áreas y solapes. Una máscara no implica que el modelo conozca la clase correcta: localización y reconocimiento son problemas relacionados pero distintos.

Tres preguntas diferentes

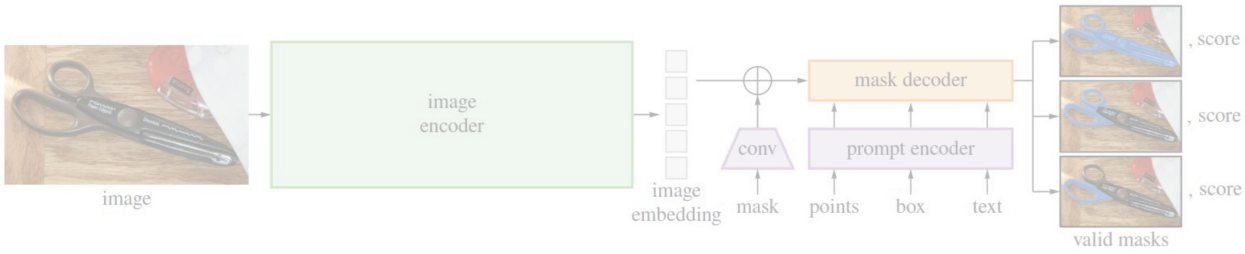
Categoría, identidad y cobertura completa definen el tipo de segmentación.

Semántica	Instancias	Panóptica
¿Qué clase ocupa cada píxel?	¿Qué objeto individual ocupa cada píxel?	¿Qué clase e instancia cubren toda la escena?

La semántica une objetos de una misma clase; la segmentación por instancias los separa; la panóptica combina ambas y asigna también las regiones de fondo etiquetables.

Codificar una vez, preguntar muchas

SAM 1 separa el encoder de imagen del prompt y del decoder de máscaras.

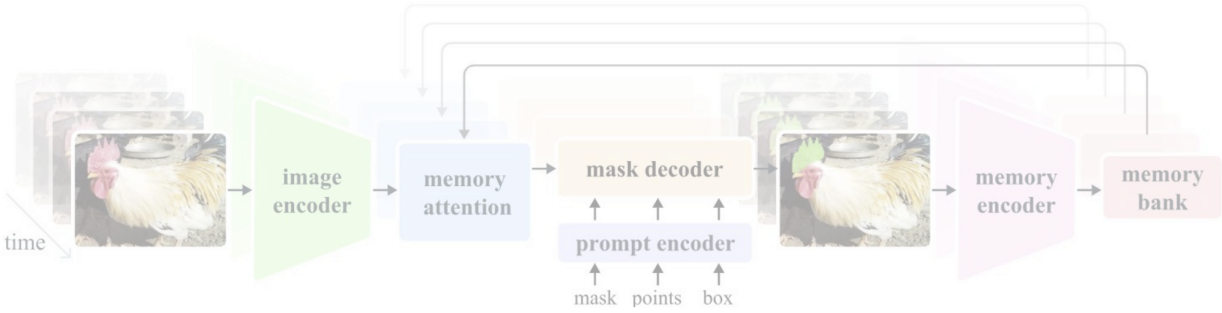


El embedding de imagen se reutiliza durante una interacción con varios prompts.

Un Vision Transformer produce el embedding costoso. El prompt encoder representa puntos, cajas o una máscara previa. Un decoder ligero cruza ambas fuentes y puede ofrecer varias máscaras cuando la indicación es ambigua.

La memoria conecta fotogramas

SAM 2 extiende la interacción a vídeo mediante un banco de memoria temporal.

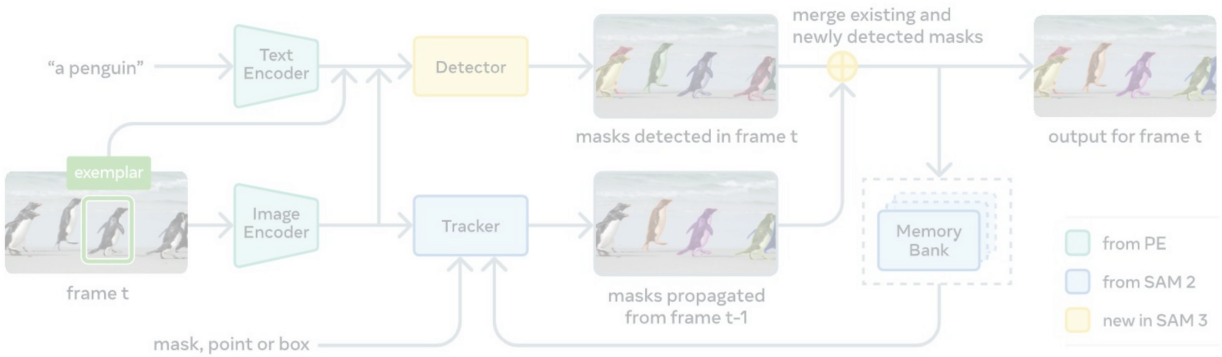


La memoria conserva información del objeto para propagarlo en el vídeo.

El backbone jerárquico Hiera extrae características. Memory attention consulta fotogramas previos y prompts para mantener identidad frente a movimiento u oclusión. Las correcciones del usuario pueden propagarse a otros fotogramas.

Primero presencia, después posición

Un token estima si el concepto existe; detector y tracker localizan y mantienen sus instancias.

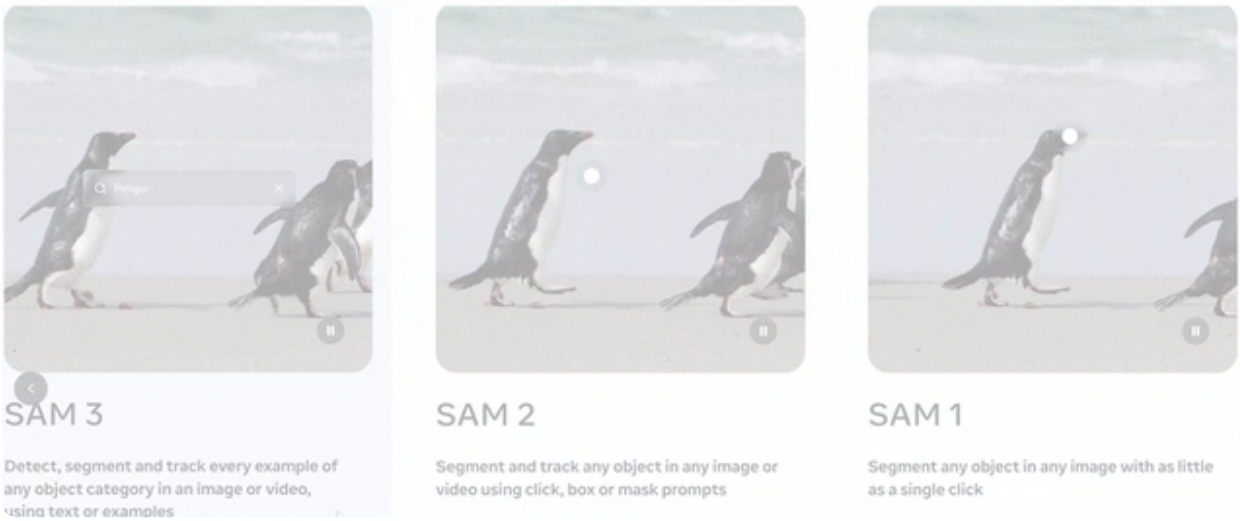


Separar reconocimiento y localización ayuda a reducir falsos positivos de vocabulario abierto.

SAM 3 condiciona la percepción con texto o ejemplares visuales. El Presence Token estima de forma global si el concepto solicitado aparece en la imagen; las object queries se concentran en localizar sus instancias. Un detector basado en DETR encuentra objetos nuevos y el tracker heredado de SAM 2 conserva identidades mediante memoria, pudiendo reiniciarse cuando reaparece una instancia.

Imagen, vídeo y conceptos

Cada generación amplía el tipo de entrada y la persistencia de la máscara.



SAM 1 (2023) introdujo la segmentación promptable en imagen. SAM 2 (2024) unificó imagen y vídeo con memoria. SAM 3 (2025) añadió segmentación abierta por conceptos mediante texto y ejemplares, además de detección y tracking.

El modelo acelera sus etiquetas

El motor de datos alterna anotación humana, asistencia y generación automática.

Manual

Personas dibujan las primeras máscaras.

Asistida

SAM propone y una persona corrige.

Automática

Se conservan propuestas de alta calidad.

Este ciclo permitió construir SA-1B a gran escala: aproximadamente 11 millones de imágenes y 1.1 mil millones de máscaras. En la fase final, la gran mayoría de máscaras se generó automáticamente y se filtró por criterios de calidad.

Indicar sin reentrenar

Un prompt convierte la intención del usuario en una condición espacial o semántica.

Puntos	Caja	Máscara
Incluyen o excluyen zonas concretas.	Acota el objeto con rapidez.	Refina una propuesta anterior.

Un prompt no modifica los pesos. SAM 1 y SAM 2 trabajan de forma nativa con indicaciones visuales; el texto suele requerir un modelo externo como Grounding DINO. SAM 3 incorpora texto y ejemplares como prompts nativos para conceptos.

Un clic elige el objeto

El punto positivo señala una región que debe pertenecer a la máscara.

Positivo dentro del objeto.

Negativo en una zona que sobra.

Varios para resolver ambigüedad.

Quieres segmentar una manzana entre otras frutas. ¿Qué prompt sirve para empezar?

- A Un punto positivo dentro de la manzana.
- B El nombre de la clase sin imagen.
- C Una métrica de precisión.

Respuesta A. Exacto: indica qué objeto quieres incluir.

La primera indicación puede admitir varias interpretaciones —objeto completo, parte o subparte—. Por eso algunas versiones devuelven varias máscaras candidatas con puntuaciones.

Corregir sin volver a entrenar

Los nuevos clics añaden información sobre qué incluir y qué excluir.

Observa dónde falla el contorno.

Añade el prompt en esa región.

Repite hasta alcanzar el criterio.

La máscara incluye parte del fondo. ¿Qué harías?

- A Añadir un punto negativo sobre el fondo.
- B Subir la temperatura del modelo.
- C Cambiar el nombre del objeto.

Respuesta A. Exacto: el punto negativo señala qué región excluir.

En vídeo con SAM 2, una corrección puede actualizar la memoria y propagarse. La interacción debe evaluarse por calidad final y por número de clics o tiempo requerido.

Texto y ejemplos cambian la búsqueda

SAM 3 puede segmentar todas las instancias que coinciden con un concepto.

Texto	Ejemplar
Una frase nominal como “paraguas rojo a rayas”.	Un recorte muestra visualmente qué instancias buscar.

Los prompts de concepto son útiles para encontrar, contar o seguir varias instancias. Una consulta lógica compleja puede requerir un VLM que razone y use SAM como herramienta.

La máscara no nombra la especie

SAM delimita; otra señal debe reconocer, decidir o justificar.

Úsalo con objetos nuevos y anotación interactiva.

Combínalo con detector, clasificador o VLM.

Compara modelos ligeros en entornos fijos.

Necesitas saber qué especie aparece y segmentarla.

- A Combinar una señal de categoría con SAM.
- B Usar solo una máscara sin revisar.
- C Eliminar los prompts.

Respuesta A. Exacto: SAM aporta la máscara, otra parte debe aportar la categoría.

SAM resulta valioso con poca anotación, interacción humana o vídeo. Un modelo supervisado pequeño puede ser más rápido y barato para clases estables y fondos controlados. Siempre hay que evaluar bordes, objetos pequeños, oclusiones y consumo de memoria.

Segmenta y corrige

Prueba cómo cambian las máscaras con puntos positivos y negativos.

1. Guarda una copia en Drive.
2. Segmenta con un punto.
3. Refina y compara máscaras.

Abrir en Colab ↗

Empieza con un punto positivo y añade un punto negativo donde la máscara invada el fondo. Compara número de interacciones, contorno y objetos pequeños. Comprueba qué versión carga el cuaderno: texto, memoria temporal y ejemplares no están disponibles en todos los modelos SAM.