.UBAfiuba
FACULTAD DE INGENIERÍA

Vision Transformers

Docentes:
Esp. Abraham Rodriguez - FIUBA
Mg. Oksana Bokhonok - FIUBA

Programa de la materia

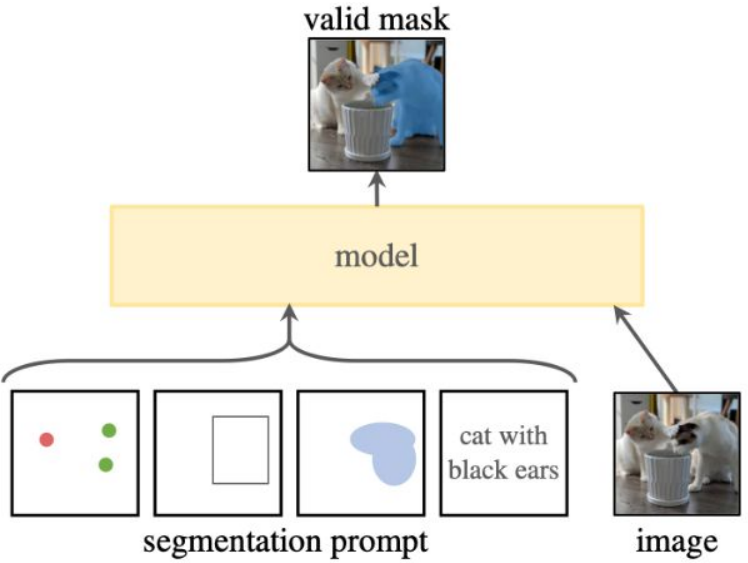
1. Arquitectura de Transformers e imágenes como secuencias.
2. Arquitecturas de ViT y el mecanismo de Attention.
3. Ecosistema actual, Huggingface y modelos pre entrenados.
4. GPT en NLP e ImageGPT.
5. Modelos multimodales: combinación de visión y lenguaje.
6. Segmentación con SAM y herramientas de auto etiquetado multimodales.
7. OCR y detección con modelos multimodales.
8. Presentación de proyectos.

Segment Anything

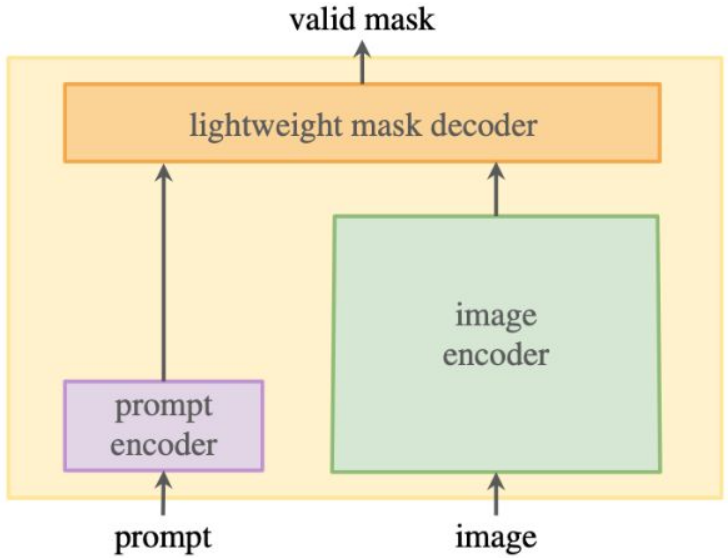
Segment Anything

Presentado en el paper "Segment Anything", consiste en el desarrollo de un modelo para segmentación mediante tres componentes y premisas:

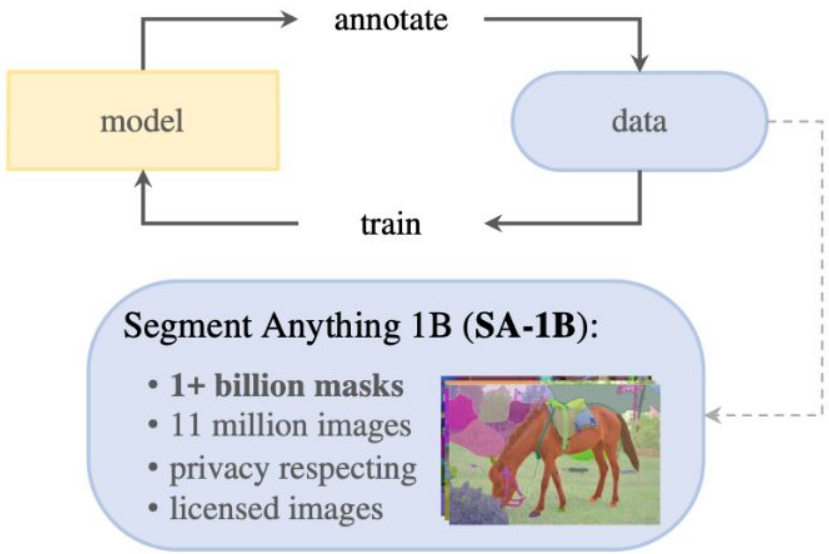
Componente	Premisa
Una tarea de segmentation.	Que tarea permite la generalization mediante zero-shot?
SAM, para la anotación de datos con capacidad de zero-shot mediante prompt engineering.	Cual es la arquitectura del modelo correspondiente?
Un motor de datos para recolectar un dataset con más de 1B máscaras.	Qué datos pueden satisfacer al modelo y la tarea?



(a) **Task**: promptable segmentation



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

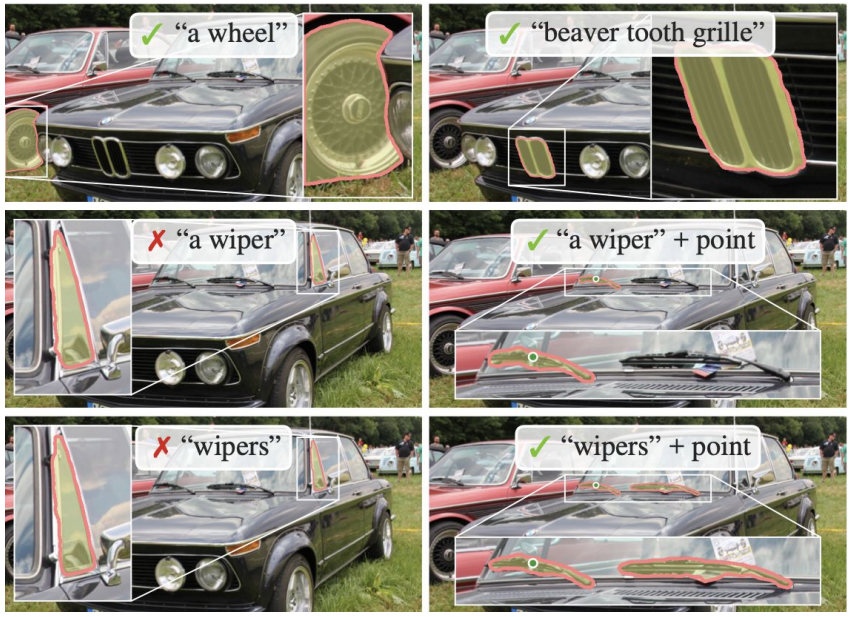


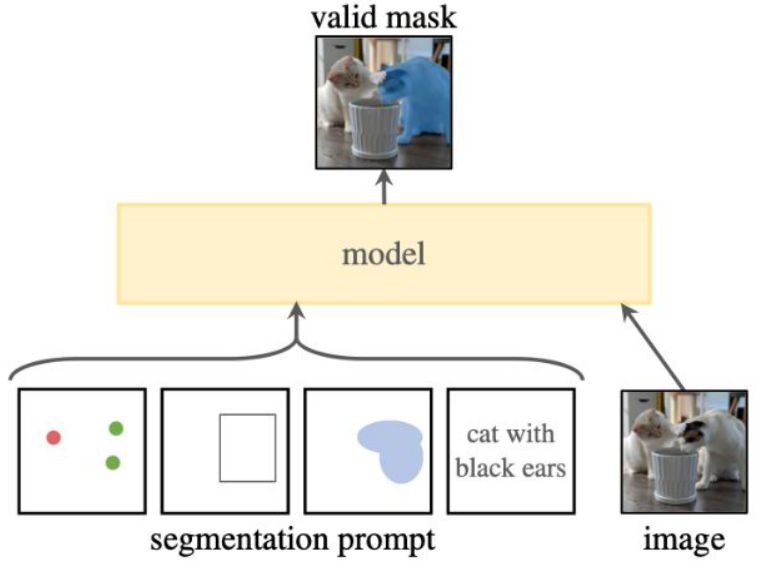
(c) **Data**: data engine (top) & dataset (bottom)

Segment Anything Task

En NLP y Computer Vision, los modelos pueden realizar zero-shot y few-shot learning mediante prompting para nuevos conjuntos de datos y tareas. **Segment Anything** propone utilizar prompts, donde el objetivo es **devolver una máscara de segmentación válida** dado cualquier prompt de segmentación.

Un prompt especifica qué objetos segmentar en una imagen, puede incluir información espacial o textual que identifique un objeto. La salida debe ser una máscara razonable para al menos uno de esos objetos.





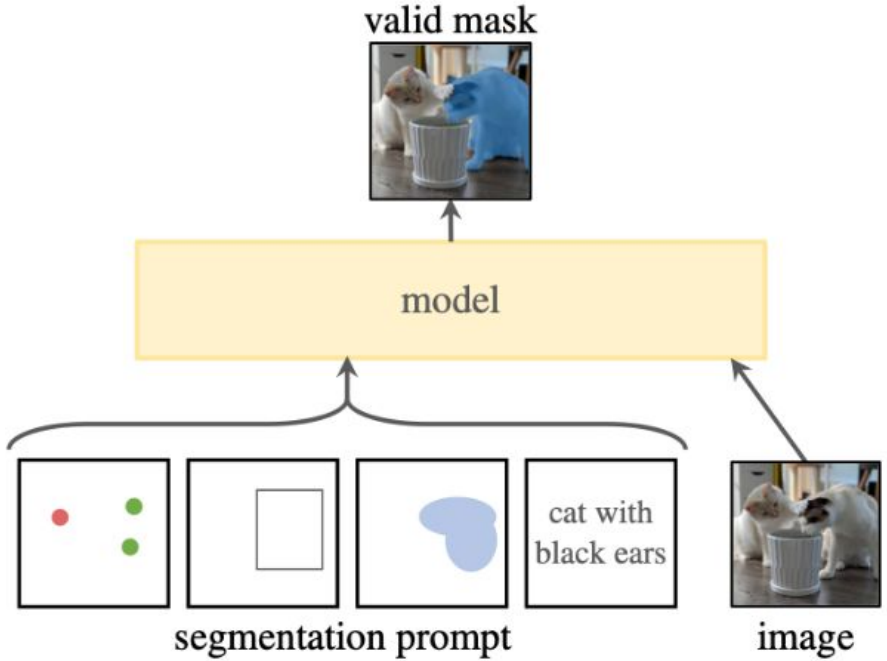
(a) **Task**: promptable segmentation

Segment Anything Task

Un prompt puede ser un conjunto de puntos, una aproximación mediante bounding box o máscara, texto. Básicamente cualquier información descriptiva del objeto a segmentar.

La máscara retornada debe de ser “válida” dado cualquier tipo de prompt.

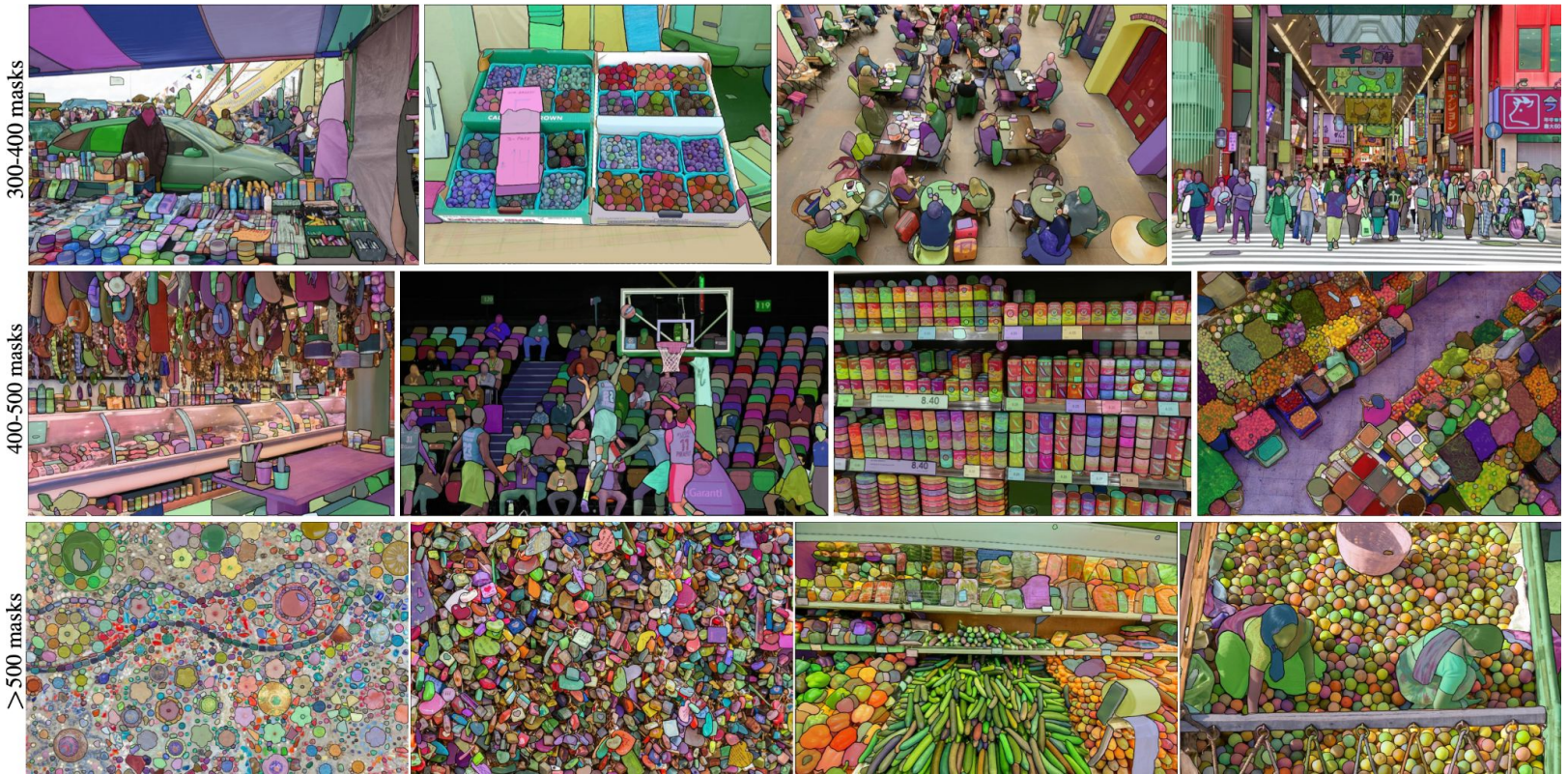
La “validez” significa que inclusive cuando el prompt es ambiguo y pueda referirse a multiples objetos, la salida debe de ser una máscara razonable **para al menos uno** de los objetos.

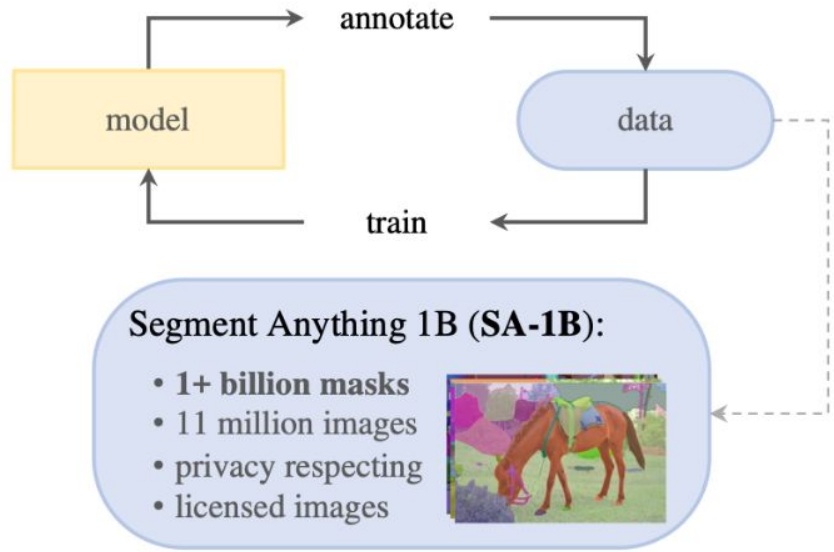


Segment Anything Data Engine

Para que SAM logre generalizar, fue entrenado un un conjunto amplo y diverso de masks, ya que la naturaleza del dataset es poco común, se construyó el Data Engine. Consiste en tres etapas:

I. SAM, asiste a anotadores en generar masks de anotación.
II. SAM, puede generar masks automáticamente mediante prompting.
III. Mediante una grilla de puntos, SAM puede generar máscaras.





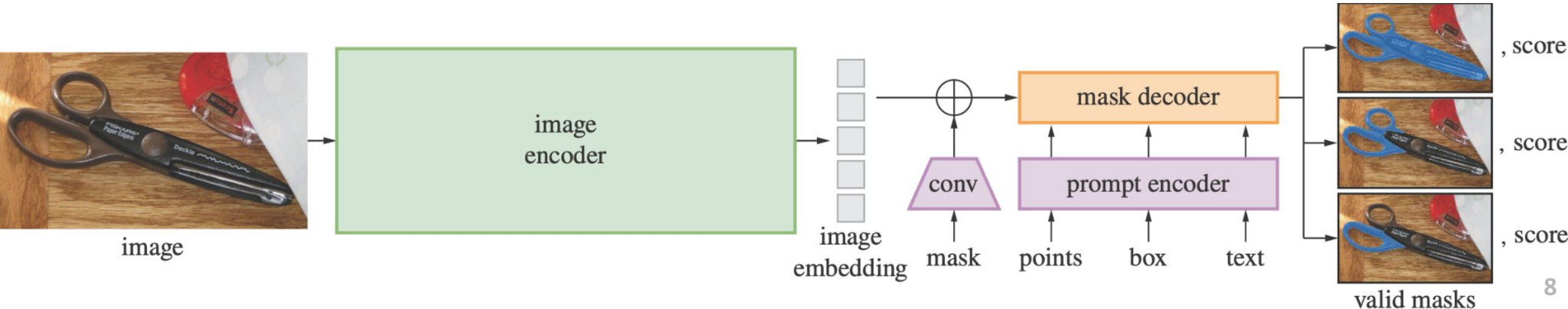
(c) **Data**: data engine (top) & dataset (bottom)

Segment Anything Model (SAM)

La naturaleza de la tarea a realizar en **tiempo real**, representa una restricción para el modelo, donde debe ser flexible ante los prompts y debe de ser “consciente de ambigüedad”. Un diseño simple satisface lo anterior, mediante un image encoder, un prompt encoder y un mask decoder.

Dado un image embedding **precomputado**, los prompt e image decoders predice una mask en approx. 50ms, para que SAM sea capaz de manejar la ambigüedad, **se predicen múltiples masks** para un prompt.

Website



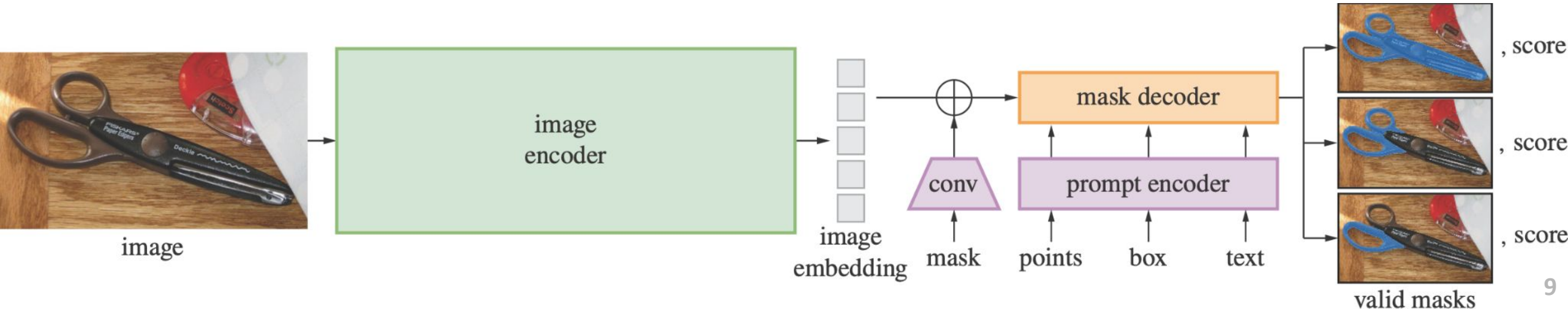
Segment Anything Model (SAM)

El Image encoder es un Masked Autoencoder (MAE) ViT, adaptado para procesar imágenes de alta resolucion.

Para el prompt encoder se consideran varios tipos de prompts siendo: esparsos(puntos, boxes y texto) y denso(masks).

- Los puntos y boxes son representados mediante **positional encoding.**
- Para texto se utiliza el mismo Transformer utilizado en **CLIP**.
- Los masks son embebidos mediante **convolución** y la operación de suma sobre image embeddings.

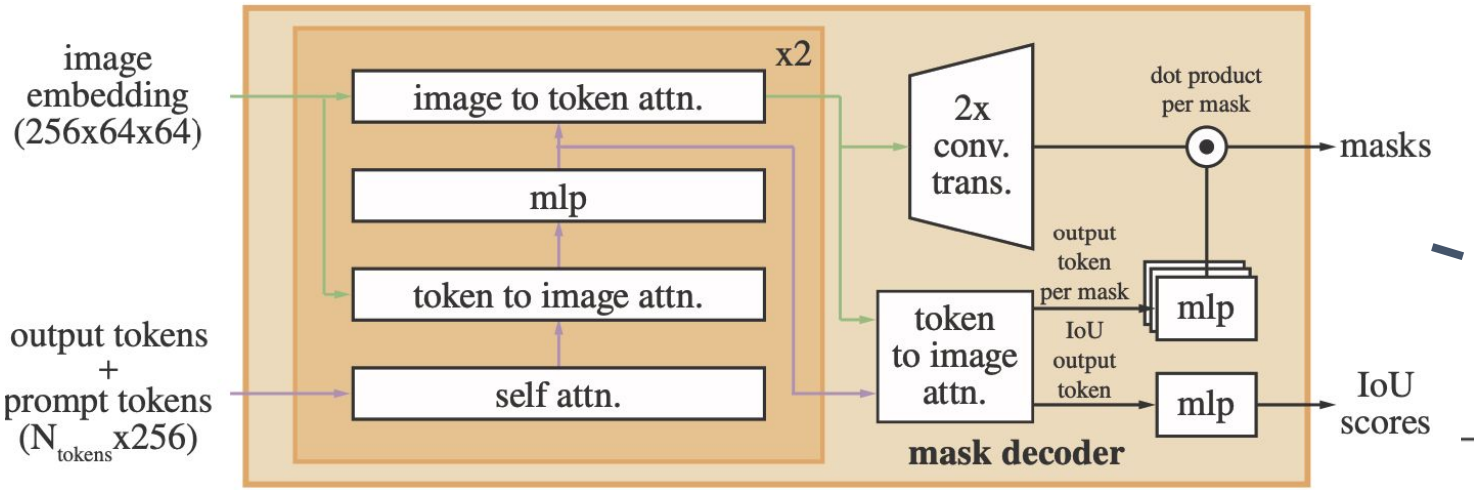
Website



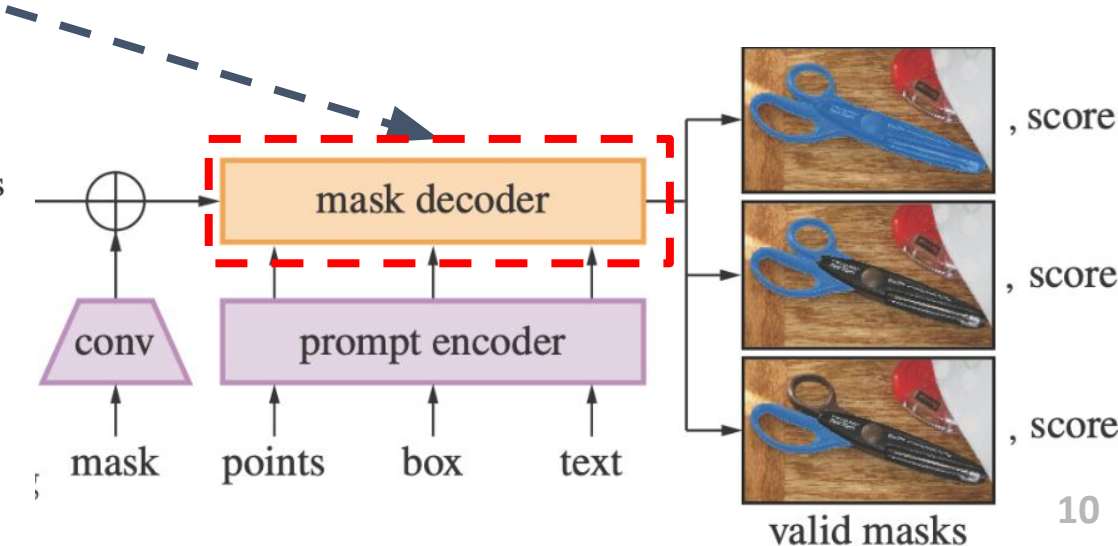
Segment Anything Model (SAM)

El mask decoder consiste en dos bloques de Transformer decoder inspirado de DETR con un **head MLP de predicción de masks** que genera un token representando una máscara.

Github



Website



Segment Anything Model (SAM)

SAM es capaz de realizar distintas tareas las cuales no fue entrenado mediante Zero-shot, como detección de bordes.

Website

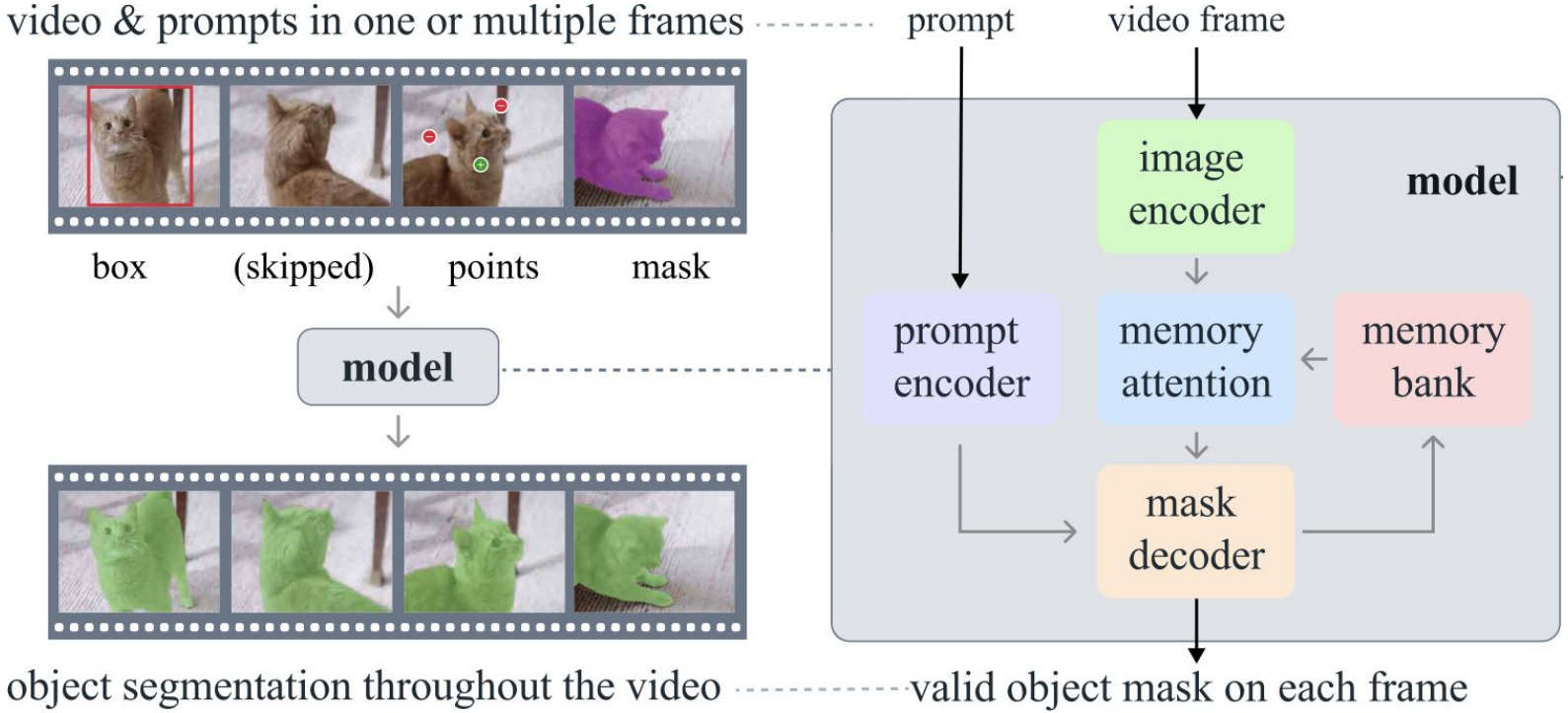
Figure 10: Zero-shot edge prediction on BSDS500. SAM was not trained to predict edge maps nor did it have access to BSDS images or annotations during training.

method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Segment Anything 2

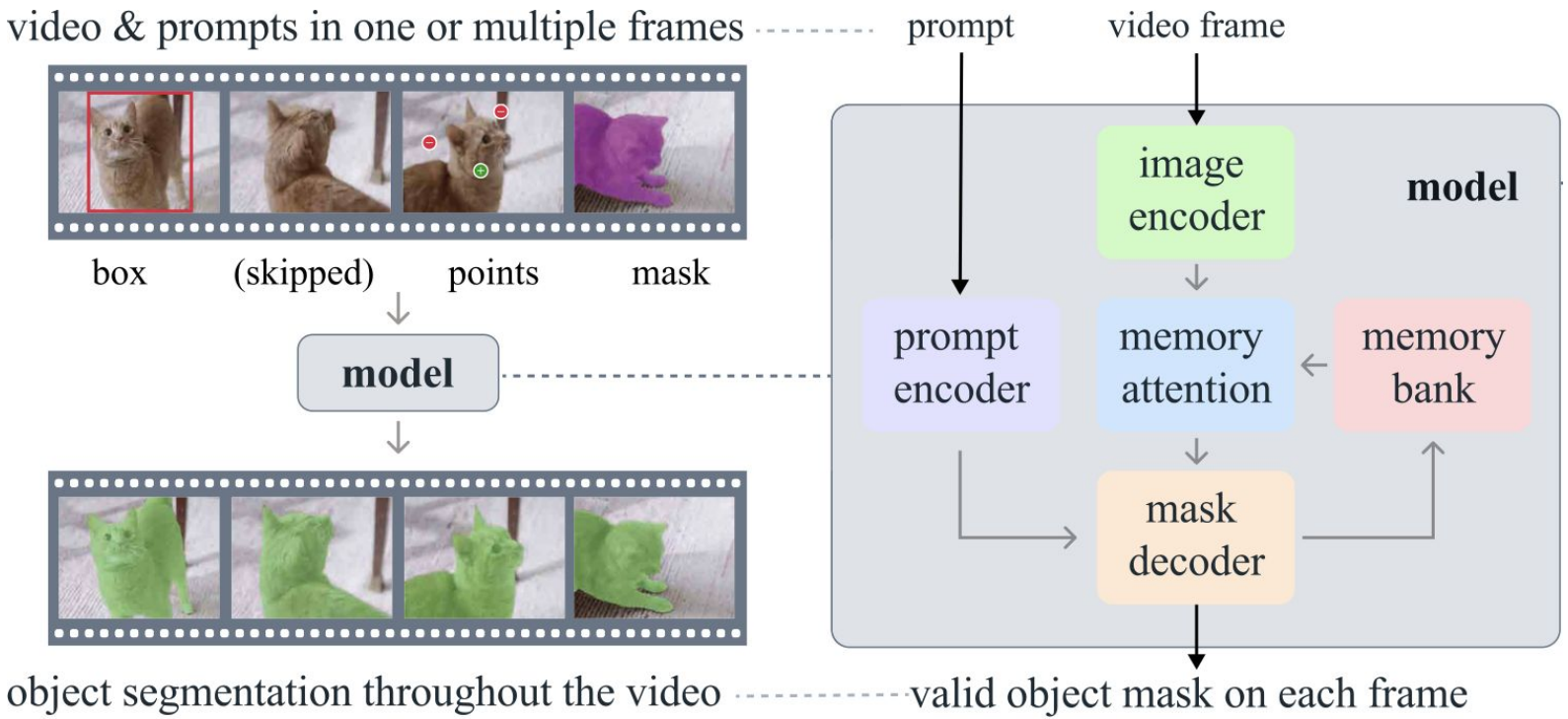
En Julio 2024, se presentó el paper "SAM 2: Segment Anything in Images and Videos", que expande a SAM con capacidades de segmentación en **videos** e imágenes, es 6x más rápido que SAM en imágenes.

En la vida real un sistema de segmentación **universal** debe manejar imágenes como videos con **movimientos complejos**. La segmentación en videos implica identificar entidades en un contexto espacio-temporal, considerando deformaciones, oclusiones, cambios en iluminación, etc. Además, los videos suelen tener menor calidad que imágenes debido al movimiento, desenfoque y baja resolución.



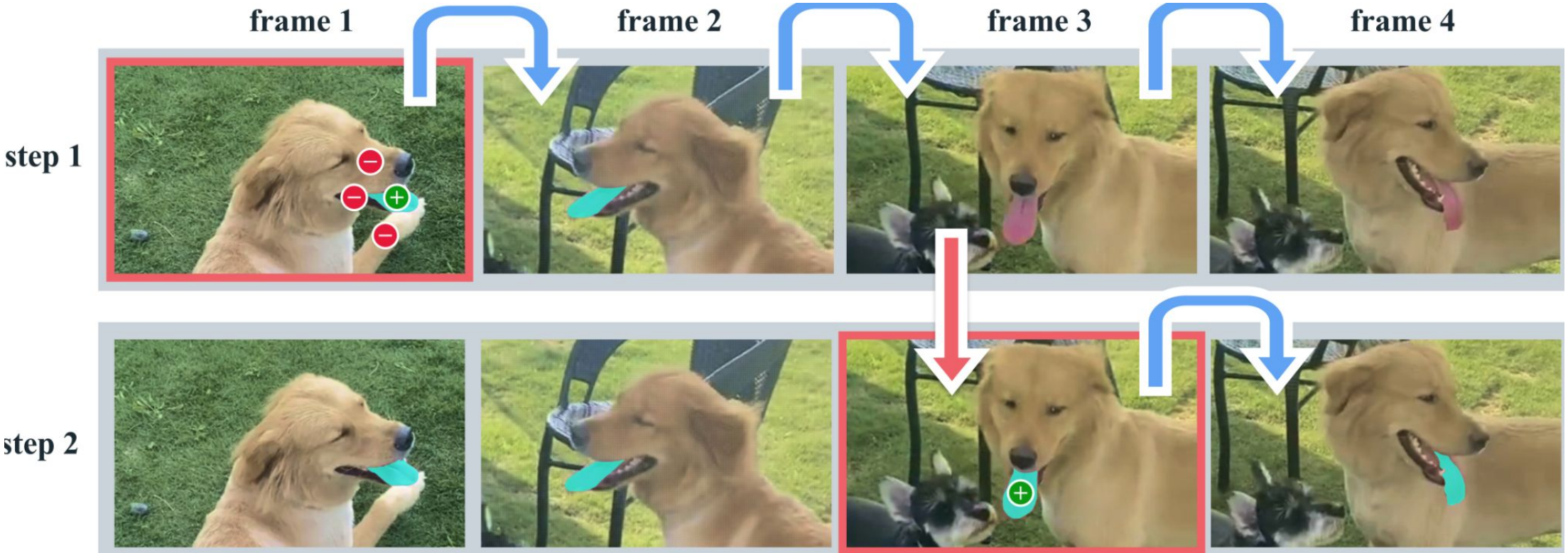
Segment Anything 2

SAM 2, produce masks de un objeto de interés, está equipado con **memoria** donde almacena información sobre el objeto e **interacciones previas**, lo cual le permite predecir masklets a través del video, además de **corregirlos** mediante la memoria almacenada de frames anteriormente observados.



Segment Anything 2 Task

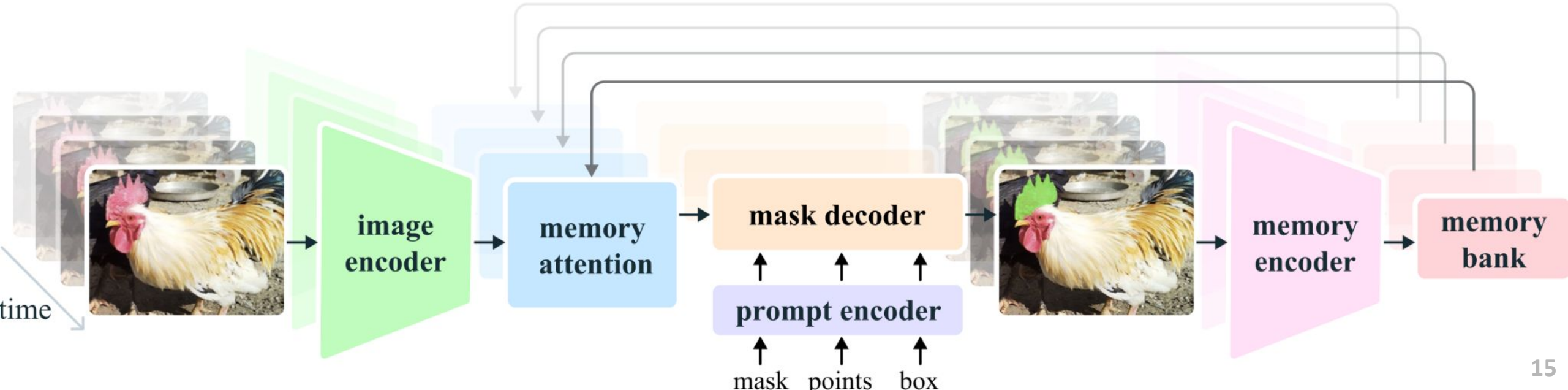
La tarea implica proveer prompts (clicks, boxes, masks) para definir un objeto a segmentar en cualquier frame del video, sin embargo SAM 2 puede perder el objeto durante transicion. para ello la memoria y un prompt puede recuperar la segmentación.



Segment Anything 2 Model (SAM 2)

SAM 2, es una generalización de SAM, con el agregado de tener un memory encoder. Los frames son procesados mediante streaming, uno por uno, la segmentación es condicionada en el prompt actual y/o en la memoria acumulada. En su mayoría l**a arquitectura se mantiene igual que SAM**.

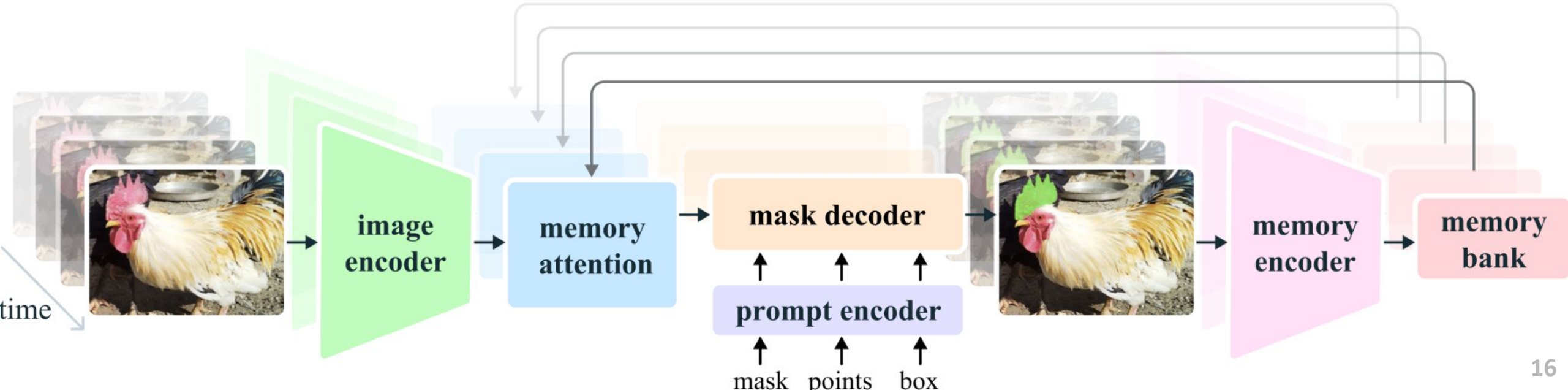
El rol de **memory attention** es condicionar los frames actuales, pasados y predichos. Consiste en un stack de Transformer blocks, donde el primero recibe como input al encoded image. Cada bloque realiza self-attention seguido de cross attention con los frames y prompts guardados en el memory bank, seguido de un MLP.



Segment Anything 2 Model (SAM 2)

El **memory encoder**, genera una memoria realizando un downsample mediante convolución.

El **memory bank**, contiene una queue FIFO que retiene predicciones pasadas del objeto de interés hasta N frames recientes. También se retiene una lista de object pointers de vectores que contienen la información semántica del objeto segmentado, basado en el mask decoder.

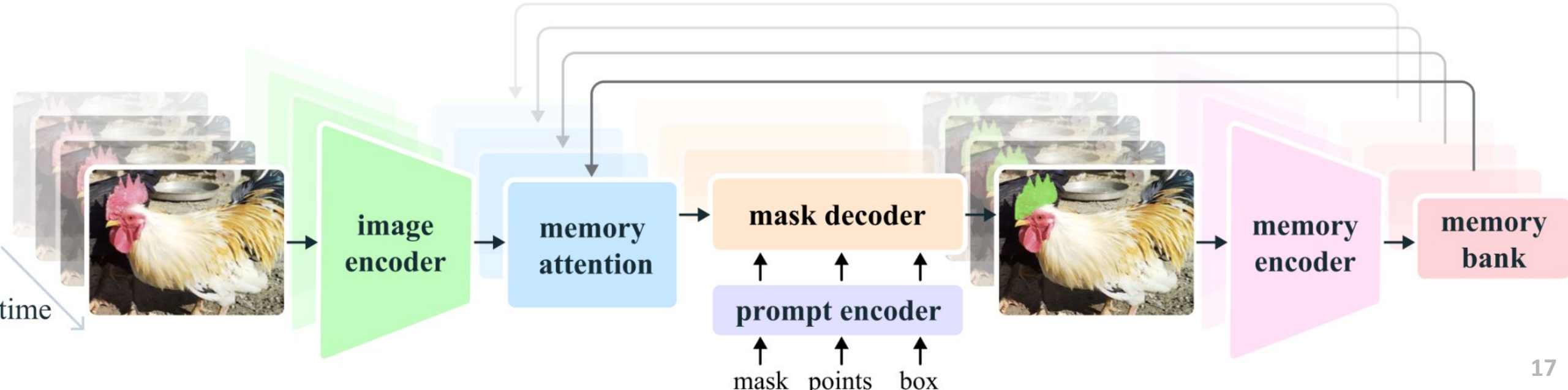


Segment Anything 2 Model (SAM 2)

[Roboflow: What is Segment Anything 2?]

[Ultralytics SAM 2]

[Website]



Auto Labeling Tools

Label Studio

Label Studio es una plataforma open-source para data labeling para LLMs, Vision por Computadora, Audio, Texto, Series temporales, OCR y Video.

- Es facil de configurar e integrar a pipelines existentes mediante SDKs y APIs.
- Permite agregar labeling asistido por modelos.
- Integracion con AWS y GCP.
- Permite el manejo del datasets.
- Soporta múltiples proyectos y usuarios.

Label Studio

Integración con ML pipelines.

Ejemplos y tutoriales de integración con modelos.

Integration directory

Label Studio Enterprise

Github

Roboflow Auto Label

Parte de los servicios de Roboflow, Auto Label utiliza modelos para realizar tareas de computer vision de manera automática.

SAM 2 Auto label

Examples

Otras herramientas

CVAT

Roboflow Autodistill

Matlab

SuperAnnotate

Encord

Práctica

Preguntas?