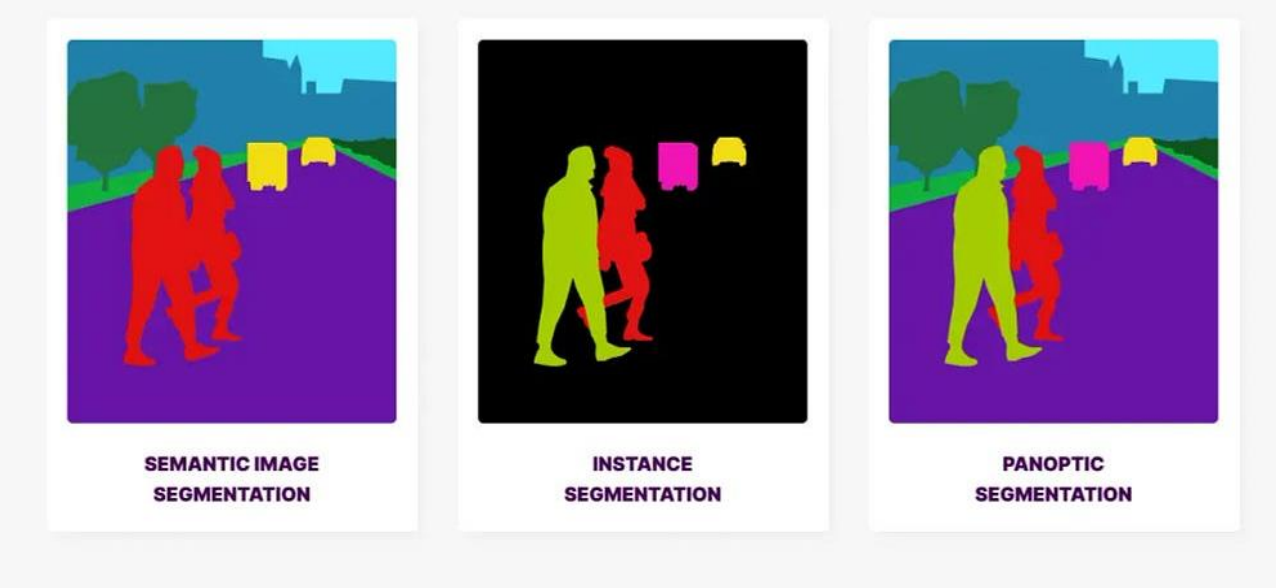
Segmentation

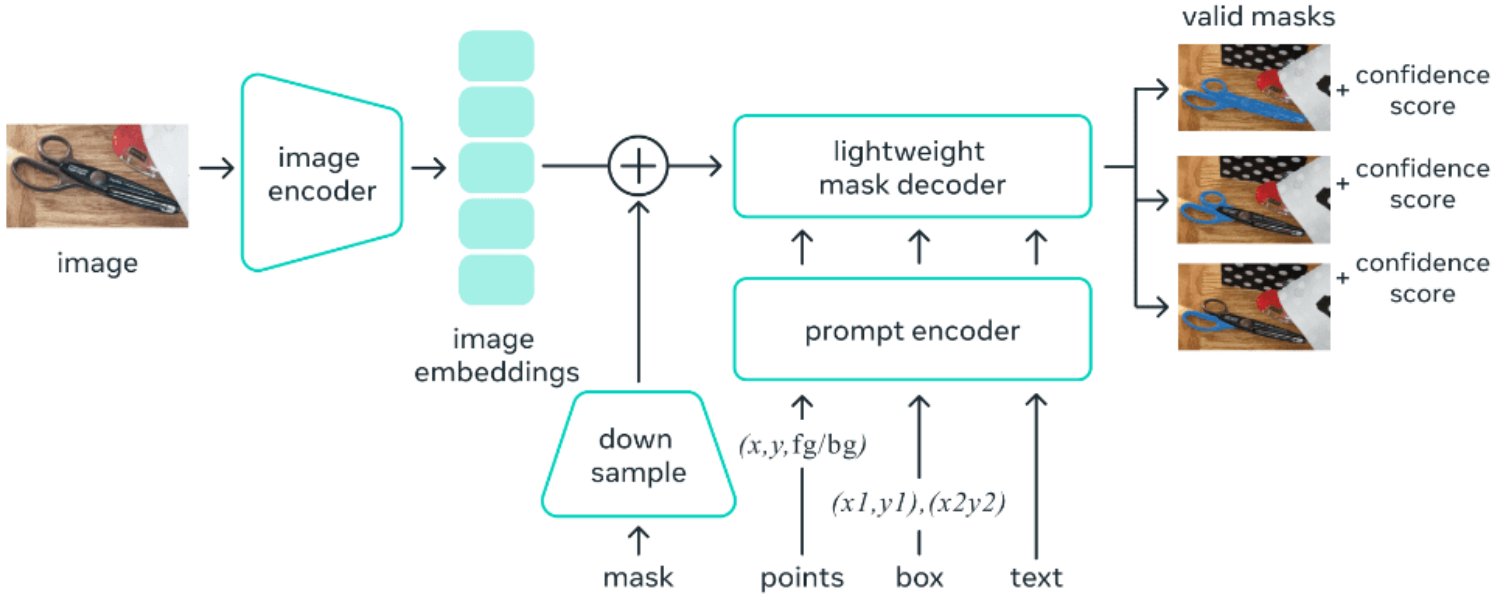
Tipos de Segmentation

- Semantic Segmentation:
  - todo lo que es del mismo tipo con el mismo color
- Instance Segmentation
  - Ahora quiero separar cada objeto individual
- Panoptic Segmentation
  - Combinemos las dos ideas



SAM en 3 ideas

- **Promptable**: acepta **puntos**, **cajas**, **máscaras** (y texto en SAM-2).
- **Generalista**: entrenado a gran escala
- **Interactividad**: devuelve **varias máscaras** + **scores** (maneja ambigüedad).



¿Qué es Zero-Shot Learning?

- Un modelo *zero-shot* no necesita ejemplos de entrenamiento del nuevo dominio o clase.
- Aprende **conceptos generales** y luego **razona por analogía** cuando enfrenta algo nuevo.
- **Ejemplo intuitivo:**
  - Si aprendiste qué es “un perro” y “un gato”, podés reconocer “un zorro” aunque nadie te lo haya mostrado: comparás formas, colores y contexto.

Cómo lo logra

- Aprende **representaciones generales** (por ejemplo, con millones de imágenes).
- Usa **embeddings** que capturan similitudes semánticas (“esto se parece a aquello”).
- Luego, ante algo nuevo, **transfiere** ese conocimiento sin volver a entrenar.
- SAM fue entrenado con **billones de máscaras** y aprendió representaciones visuales muy generales.
- En modo *zero-shot*, **no lo reentrenás**: solo le das una **pista (prompt) y SAM intenta separar los píxeles que corresponden a ese objeto.**

¿Qué es un prompt?

- Un **prompt** es la **instrucción visual** que le dice a SAM *qué* parte de la imagen debe segmentar.
  - Podés pensar que es como señalarle con el dedo “mirá acá”.
- Tipos de Prompts en SAM
  - **Point Prompt**
    - Le das **coordenadas (x, y)**
    - **Le decís:** *“Segmentá este punto y lo que lo rodea.”*
  - **Box Prompt**
    - Le das un **rectángulo [x_1, y_1, x_2, y_2]**
    - **Le decís:** *“Segmentá lo que hay dentro de esta zona.”*
  - **Mask Prompt**
    - Le das una **máscara previa**
    - **Le pedís:** *“Refiná este resultado.”*
  - *(SAM-2)* **Text Prompt**
    - Le das una **frase** (ej. “todos los autos”)
    - Usa embeddings de **CLIP (contrastive language-image pre trained model)** para entender lenguaje y segmentar por texto.

¿Cuándo alcanza zero-shot? ¿Cuándo fine-tunear?

- **No fine-tune**:
  - imágenes naturales, recursos limitados, <100 GT.
- **Sí fine-tune**:
  - **dominio específico** (médico/industrial/aéreo)
  - **objetos finos/pequeños** (cracks, cables)
  - **shift fuerte** (IR/termal).

Nueva métrica: Dice

- ¿Qué mide?
  - Mide **cuánto se parecen** dos máscaras:
  - la **predicha por el modelo**
  - y la **real (ground truth)**
  - Cuanto más se solapan, **mayor es el valor de Dice**.

$$Dice = \frac{2 \times |A \cap B|}{|A| + |B|}$$

- Donde:
  - A = píxeles predichos como objeto
  - B = píxeles verdaderos del objeto
  - A∩B = píxeles donde ambas coinciden