

# AI를 활용한 이미지 영역 나눔을 위한 Segment Anything

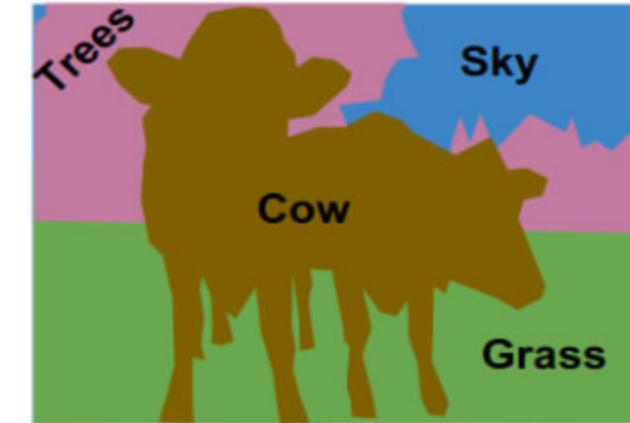
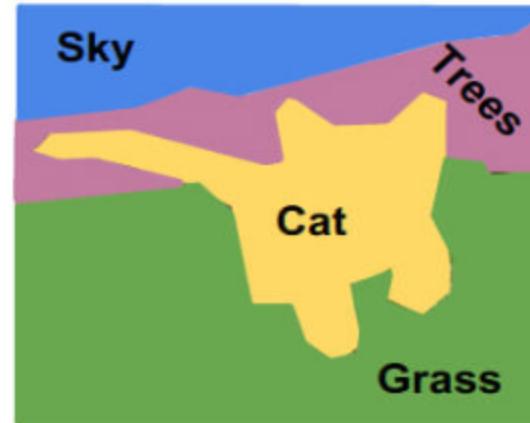
기술 논문 리뷰

통계학전공 4학년 최지우

Image Segmentation?

## Image Segmentation?

이미지 안에서 물체의 영역을 마스킹하는 컴퓨터 비전 테스크  
픽셀 단위로 어느 클래스에 속하는지를 분류해야하는 난이도가 매우 높은 테스크



### Segment Anything

Task : 어떤 task로 모델을 학습시켜야 GPT처럼 general vision 모델을 만들 수 있을까?

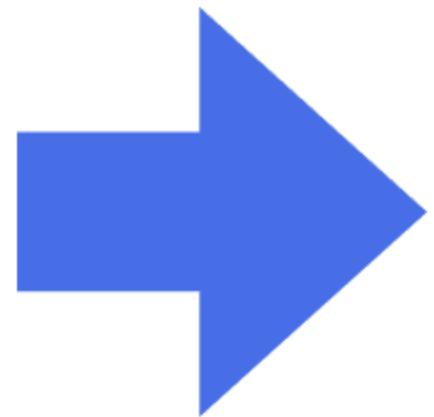
Data : 이 모델을 학습시키려면 어떤 데이터들이 필요할까?

Model : 이 task를 잘 수행하면서도 general하려면 어떤 모델 구조여야 할까?

---

01 | TASK

## TASK



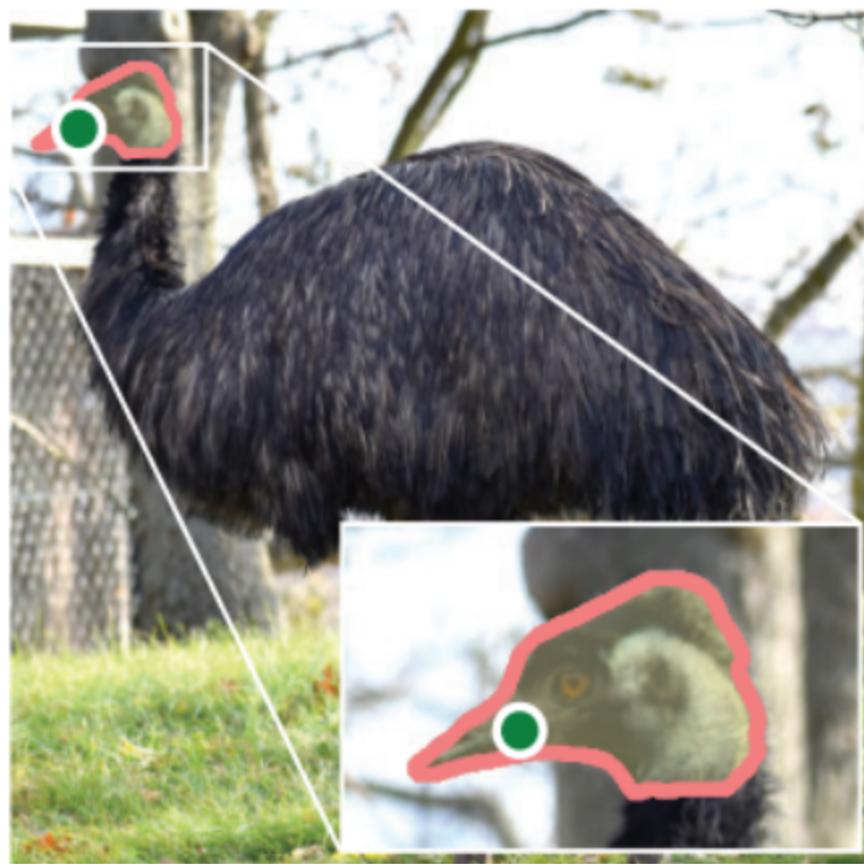
# Computer Vision?

다음에 올 단어를 예측하는 방식으로 학습시킨 GPT가  
한번도 학습한 적 없는 테스크들을 잘 수행하면서 NLP에 혁신을 가져옴

어떤 단일 task로 모델을 학습시키면 GPT처럼 여러  
테스크들을 잘하는 모델을 만들 수 있을까?

## TASK

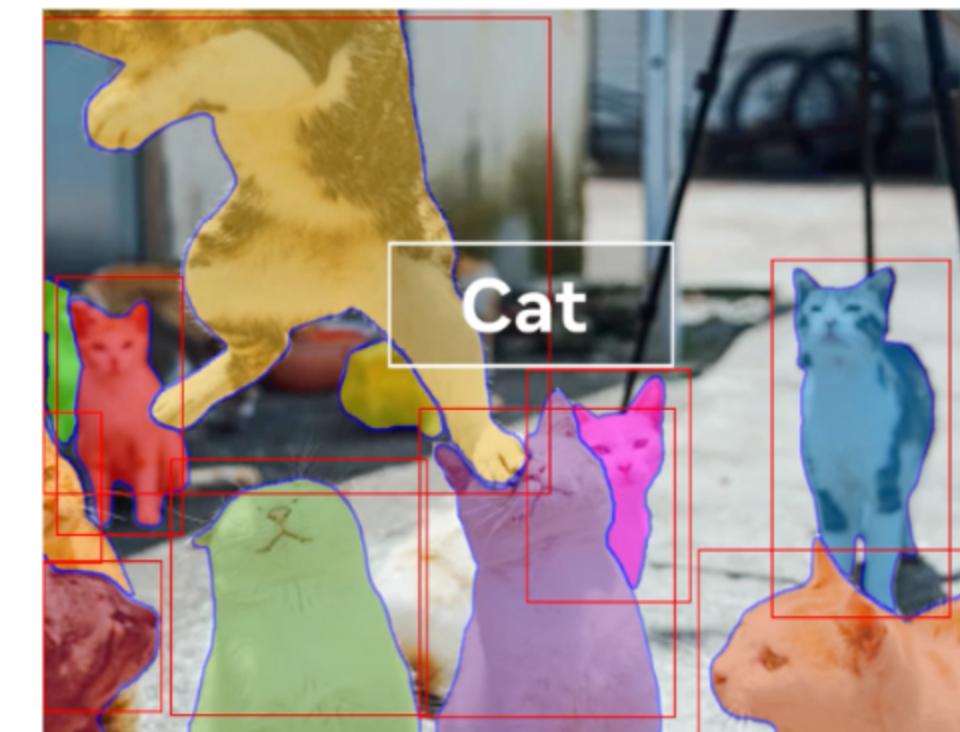
**Promptable Segmentation은  
그 어떤 모호한 정보가 Prompt로 입력되더라도 유효한 마스크를 반환하는 것을  
목표로 한다.**



**point**



**box**



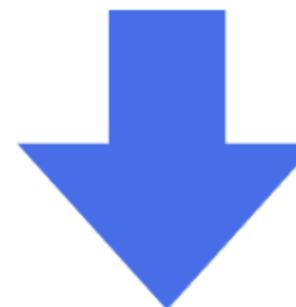
**text**

---

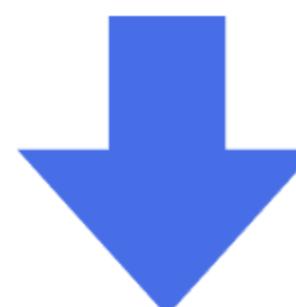
# 02 | DATA

## DATA

유연하게 동작하는 general한 AI -> 상상을 초월하는 데이터가 필요  
GPT의 경우, 웹 상의 텍스트들을 크롤링한 Common Crawl을 주요 데이터 셋으로 사용



segmentation 모델을 학습시키기 위해선 마스크 라벨이 붙어있는 데이터가 필요  
이는 단순 크롤링으로 해결이 안 될뿐더러 제작에 엄청난 수고가 들어감



data engine을 만들어서 전례 없는 규모의 데이터 셋을 직접 만듦

## DATA

### 1. Assisted-manual

기존 데이터셋으로 SAM 모델 학습, AI가 먼저 segmentation을 수행 -> 사람이 이를 수정 및 추가  
430만개의 마스크를 라벨링, 라벨링 하는 와중에도 데이터 쌓이면 모델 재학습

### 2. Semi-automatic

이전 단계에서 모은 데이터셋으로 SAM 학습  
AI가 먼저 segmentation -> 사람이 빠진 것들만 채워넣음  
마스크 590만개 추가 라벨링 -> 총 1020만개

### 3. Fully automatic

1, 2 단계에서 모은 마스크 1020만개를 가지고 SAM 모델을 학습  
이미지 1100만장에 대해 11억개의 마스크 라벨을 생성함 -> SA-1B 데이터셋

---

# 03 | Model

SAM은 크게 **Image Encoder, Prompt Encoder, Mask Decoder**로 구성됨

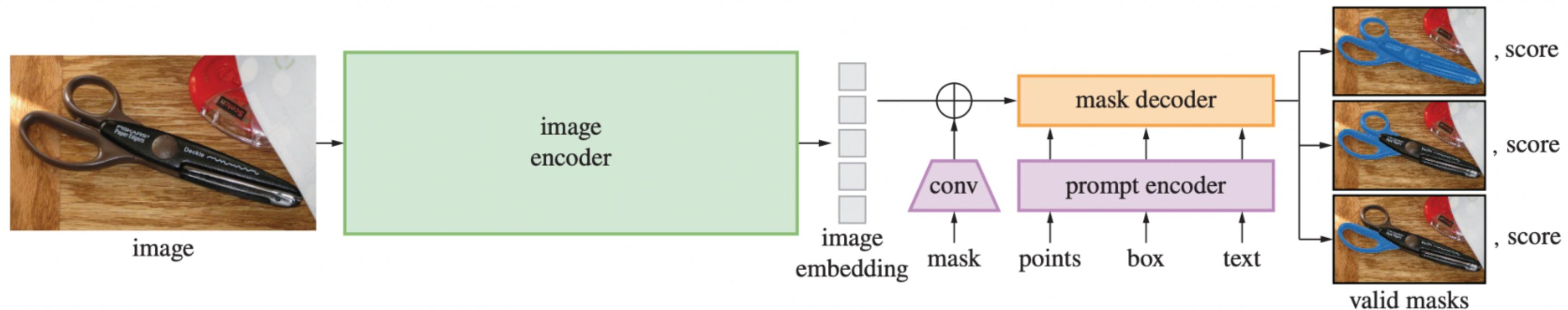


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

## Image Encoder

### Image Encoder

- 고해상도의 이미지를 처리하기 위해 **MAE**로 Pre-training을 한 **ViT**기반의 구조를 사용한다
- Image encoder의 출력은 입력 이미지 크기를 기준으로 16배 다운스케일된 Embedding이다.
- 1024x1024의 이미지 입력을 사용했으며 좌우/위아래 중 짧은 부분에는 padding을 한다.
- 즉, Embedding의 크기는 1024 / 16에 해당하는  $64 \times 64$ 가 된다.
- 또한  $1 \times 1$  Convolution을 사용해서 256 채널의 출력을 얻으며 이어서 256채널의  $3 \times 3$  Convolution을 적용한다.
- 각 Convolution 연산 뒤에는 Layer Normalization을 적용한다.

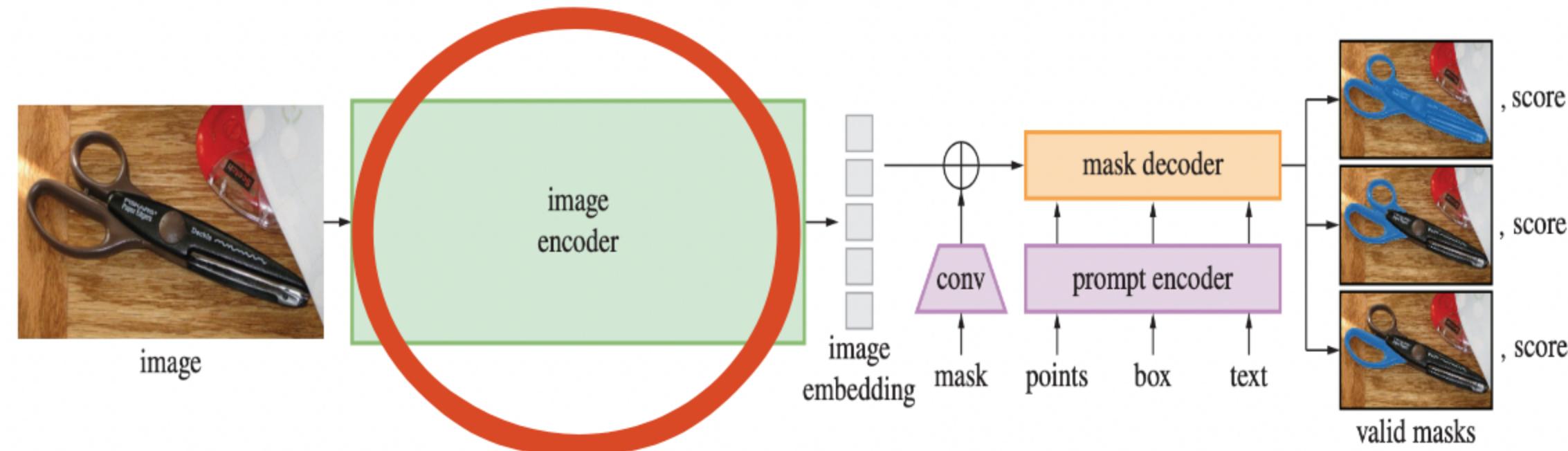
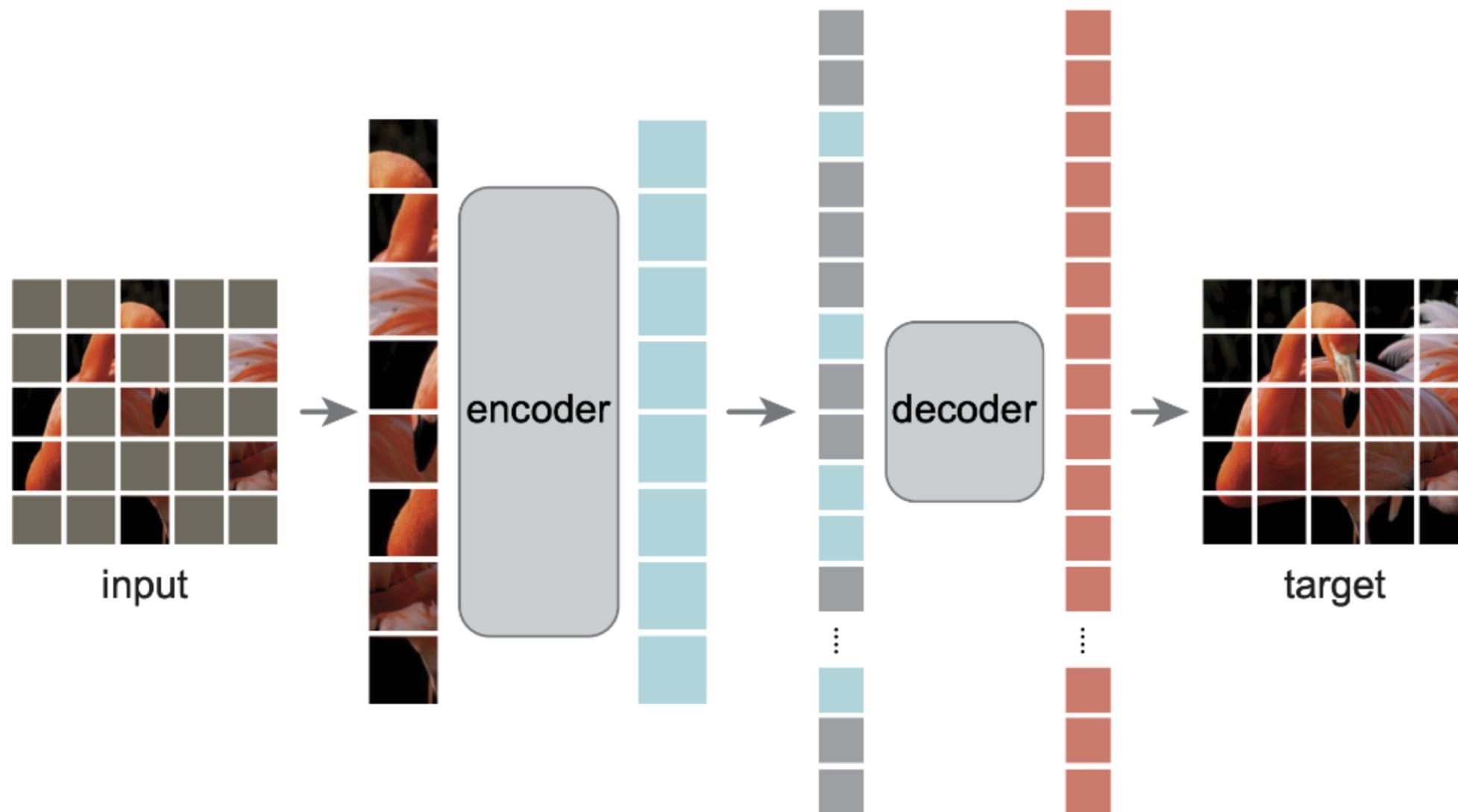


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

## Image Encoder

### MAE(Masked auto-encoder)

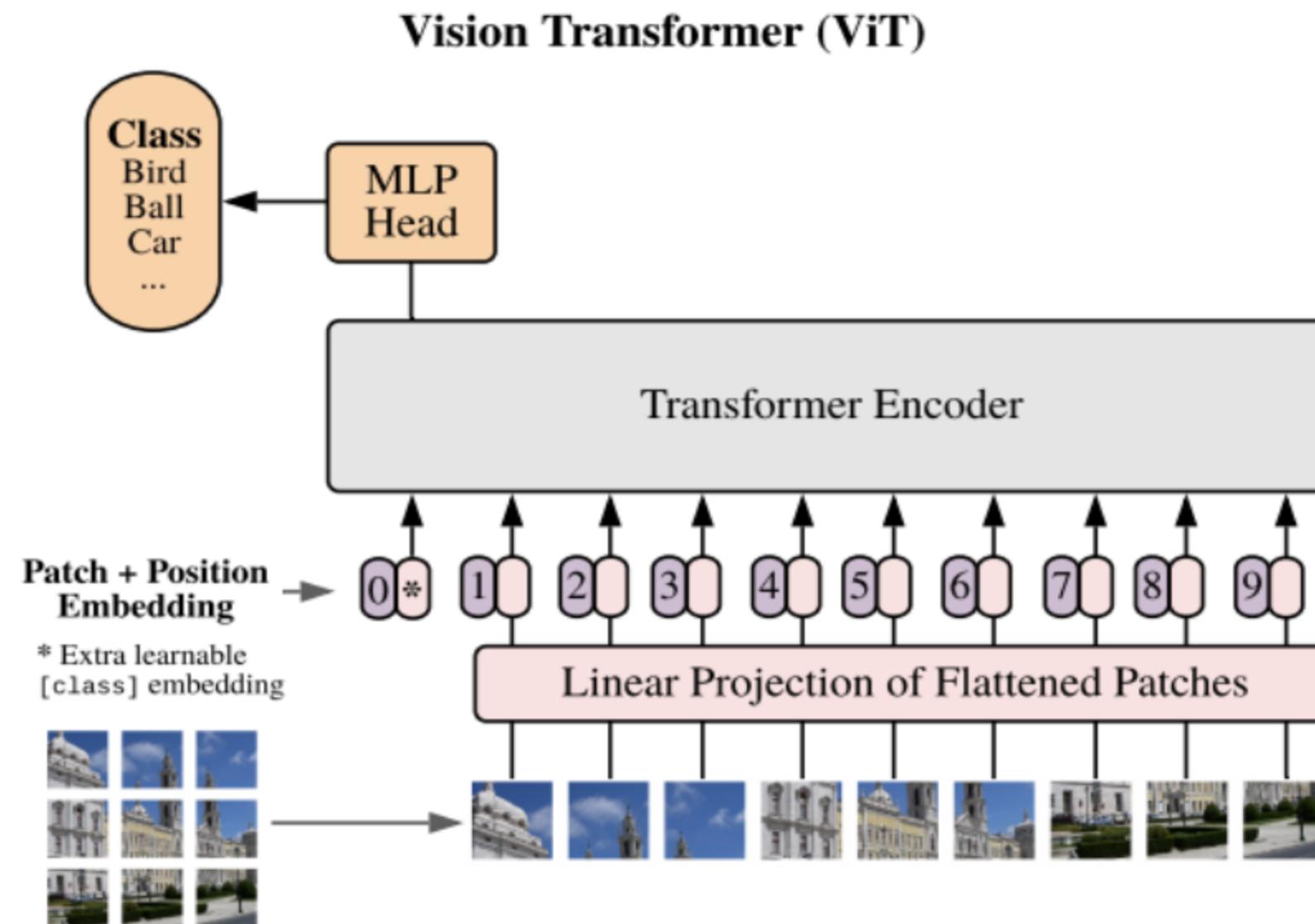
- MAE는 이미지를 일정한 크기의 그리드로 나누고 랜덤하게 가린 뒤, 복원하도록 모델을 학습시키는 기법
- decoder는 모델 학습시에만 사용한 뒤 버리고, encoder만 사용



## Image Encoder

### ViT(Vision transformer)

- Vision transformer(ViT)는 이미지를 일정한 크기의 패치로 쪼갬
- 자연어에서 토큰처럼 사용하는 트랜스포머 모델로 많은 computer vision 테스크에서 기본 블록으로 사용됨



## Prompt Encoder

### Prompt Encoder

- 각 prompt 타입에 맞는 인코딩 방식을 적용
- **Mask** : Convolution으로 차원을 맞춘고 Image embedding에 pixel wise sum
- **Points, Bounding Box** : Positional encodings summed with learned embeddings
- **Text** : CLIP 모델의 text encoder를 가져와 embedding

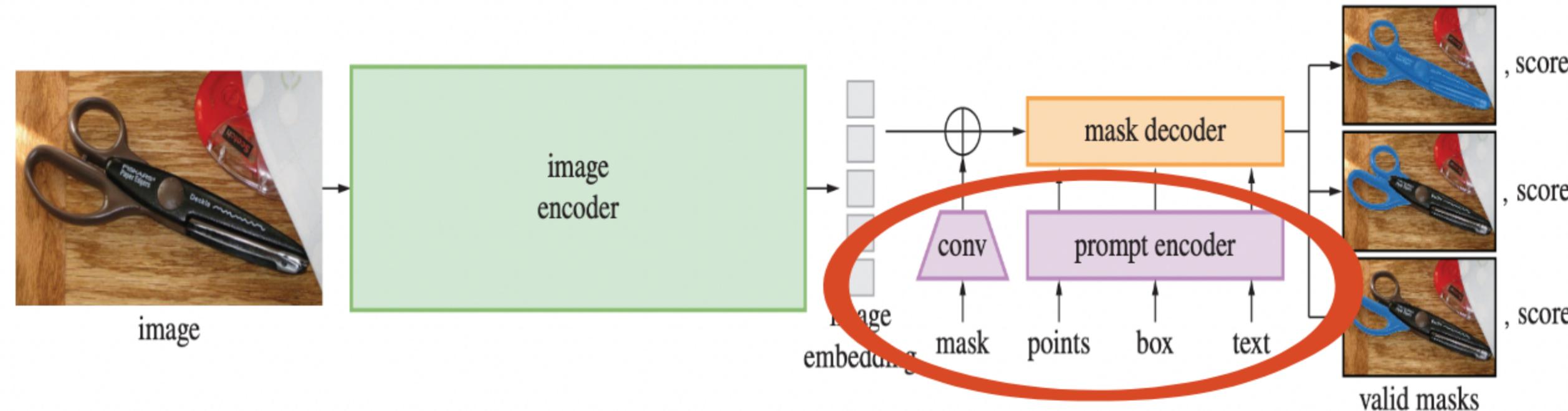
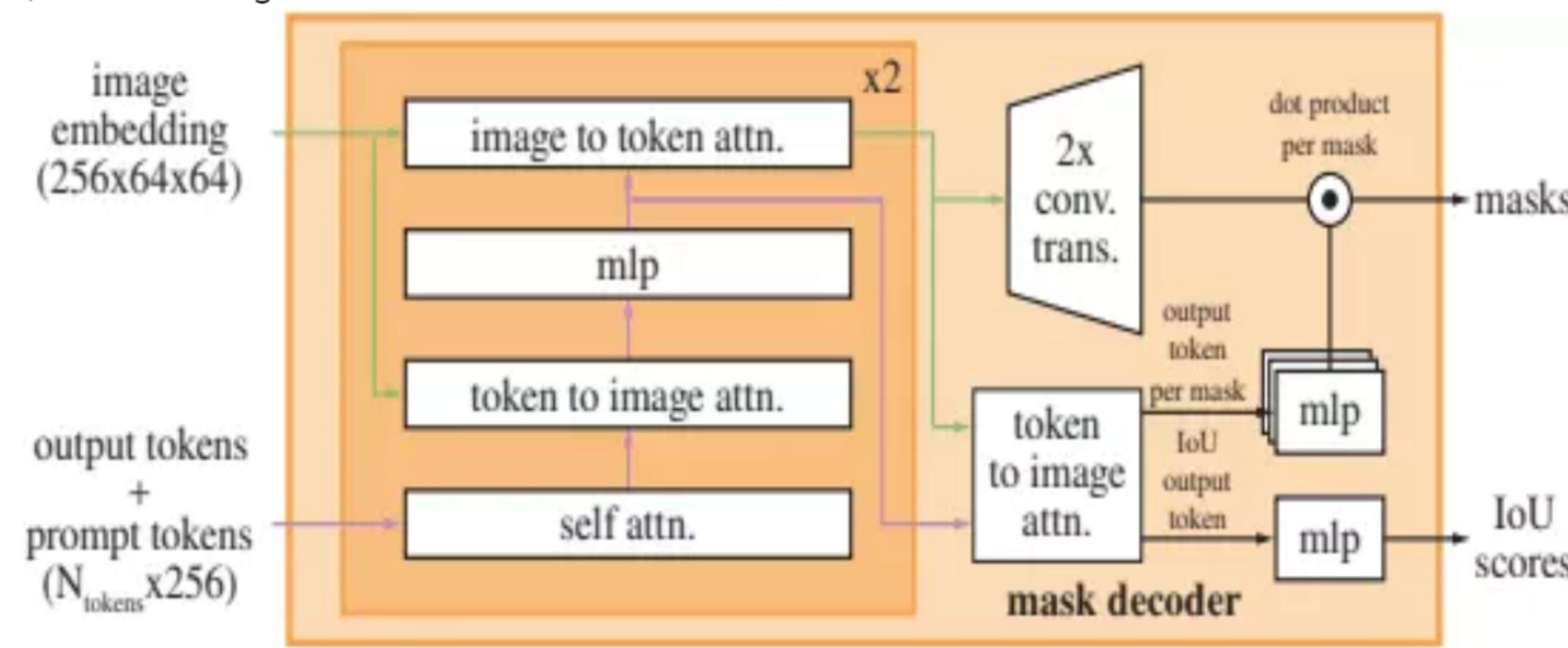


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.

## Mask Decoder

### Mask Decoder

- Image embedding과 prompt embedding을 받아 마스크를 예측하는 부분
- Prompt Embedding과 Image Embedding을 빠르게 Output Mask에 매핑하는 역할
- 두 개의 입력을 합치는 방법으로는 Transformer Segmentation Models를 참고함
- 이미지 임베딩과 프롬프트 임베딩 간의 cross attention 메커니즘을 적용해준 뒤, 마스크와 IoU를 리턴함
- image to token attn, token to image attn -> cross attention 메커니즘

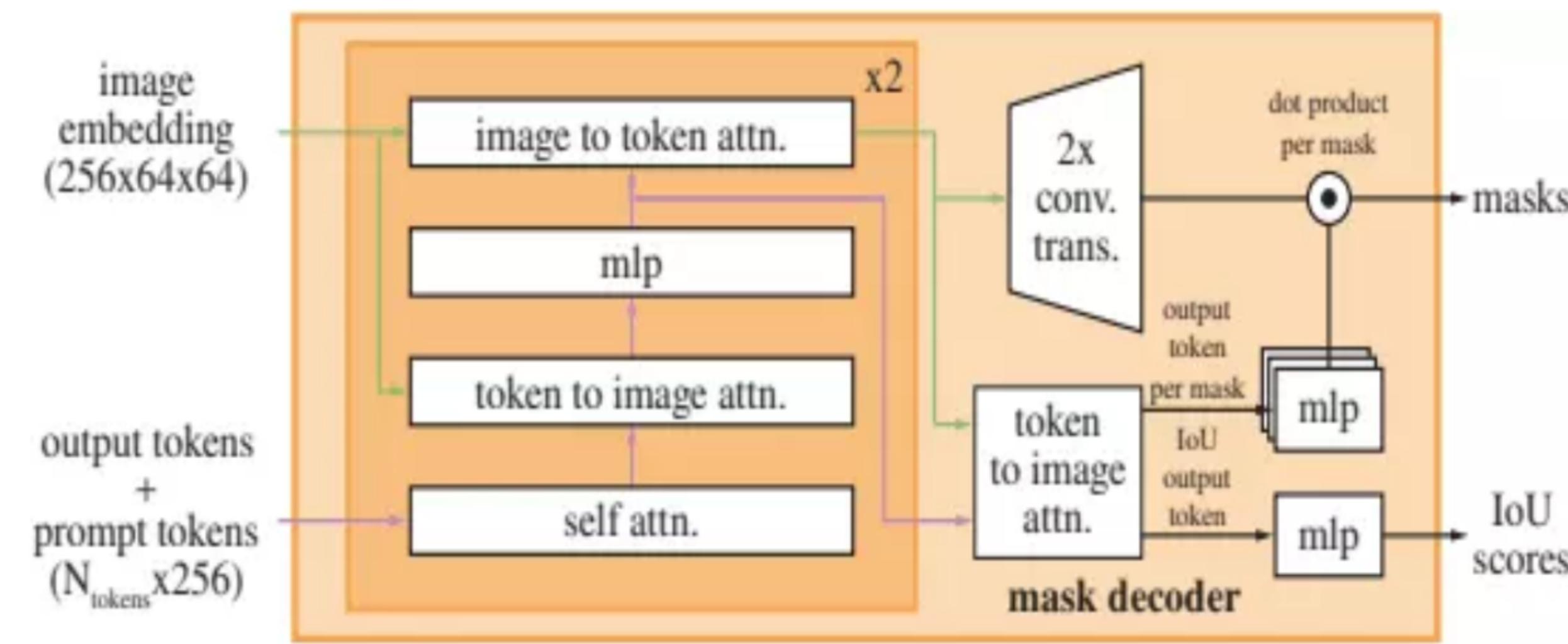


Mask Decoder 구조

## Mask Decoder

### Mask Decoder

1. Token에 대한 Self-attention
2. Token을 Query로 하여 Image Embedding에 Cross-Attention
3. Point-wise MLP로 각 Token을 업데이트
4. Image Embedding을 Query로 하여 Token에 Cross-Attention



Mask Decoder 구조

# Ambiguity

## Resolving Ambiguity

- 하나의 prompt라도 유저의 의도에 따라서 **여러 마스크가 정답**일 수 있음
- Ambiguity에 대응하기 위해서 SAM은 하나의 **prompt에 3개의 마스크를 생성하고 loss 계산**
- 3개의 마스크 중 가장 작은 loss만 역전파 진행

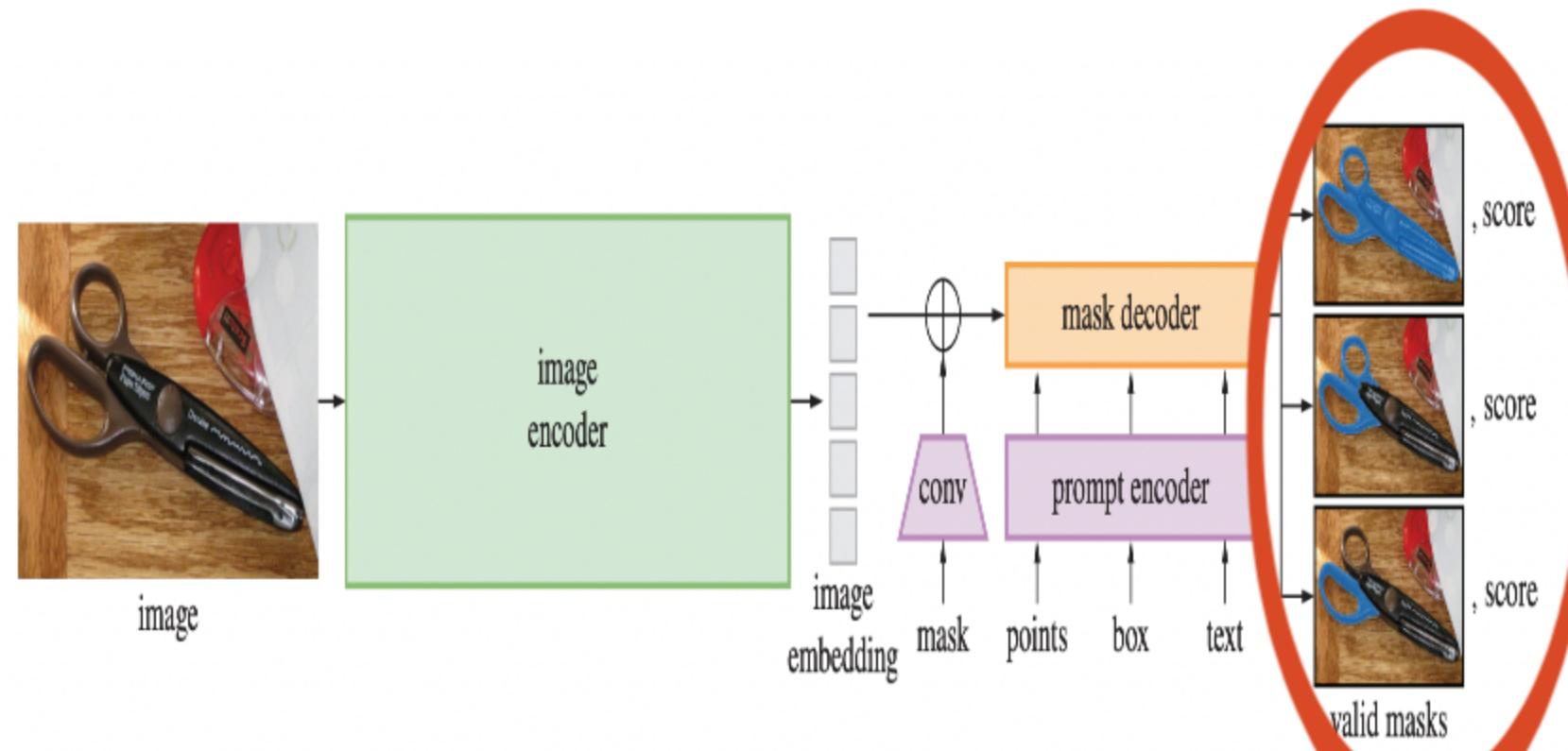
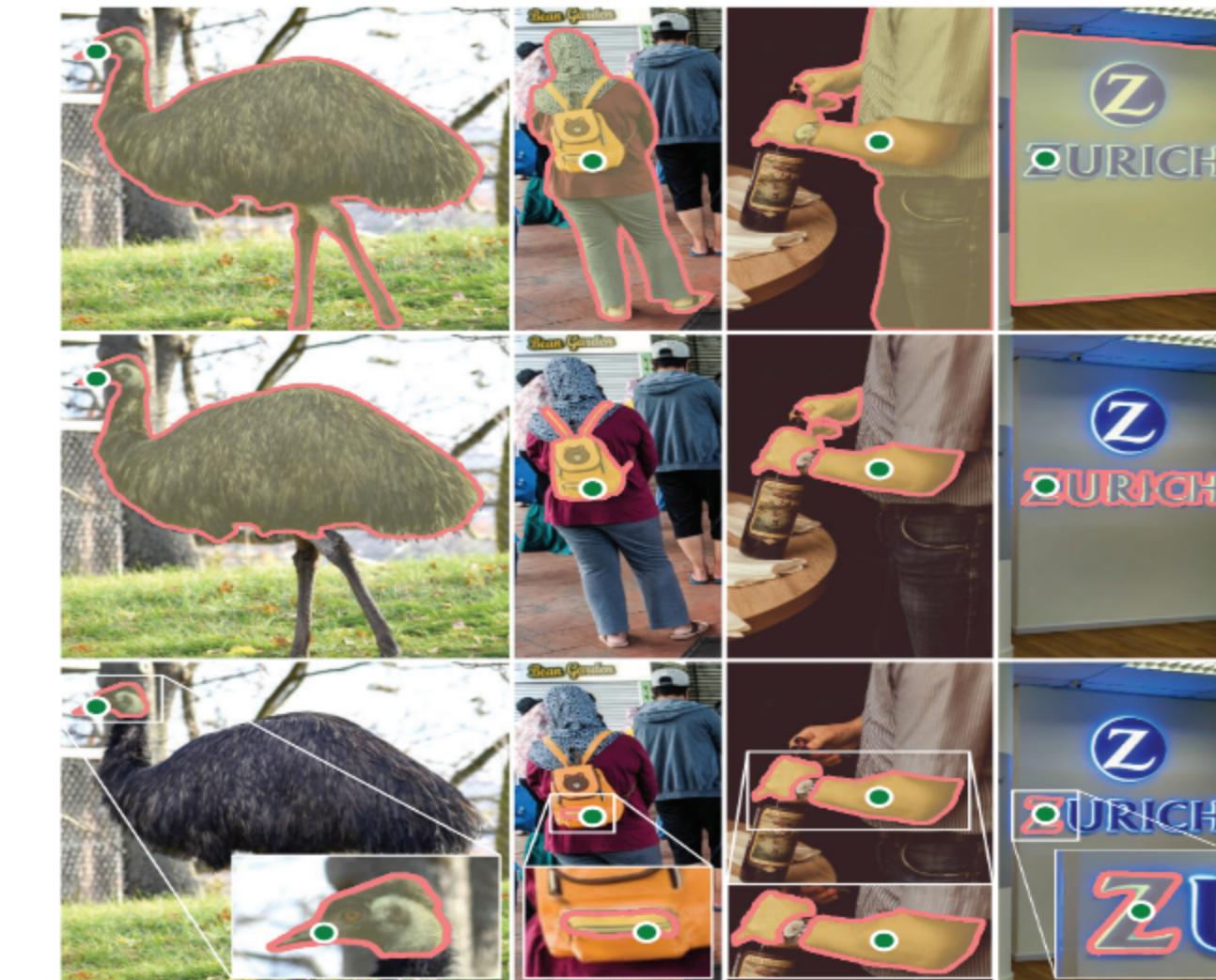


Figure 4: Segment Anything Model (SAM) overview. A heavyweight image encoder outputs an image embedding that can then be efficiently queried by a variety of input prompts to produce object masks at amortized real-time speed. For ambiguous prompts corresponding to more than one object, SAM can output multiple valid masks and associated confidence scores.



---

# 04

## Experimental Results

## Zeroshot Experiments

### Zeroshot Experiments

원래 의도처럼 단일 테스크로 학습시킨 SAM이 GPT처럼 학습하지 않았던 테스크들을  
잘 수행하는지 테스트



Zero-Shot Single Point Valid Mask Evaluation

Zero-Shot Edge Detection

Zero-Shot Object Proposals

Zero-Shot Instance Segmentation

Zero-Shot Text-to-Mask



좋은 성과를 보임

---

감사합니다