



BÁO CÁO BÀI TẬP LỚN
XỬ LÝ ẢNH

**Chủ đề : Phương pháp
Segment Anything Model**

gee



Nội dung

- ★ I. Giới thiệu và lý do chọn bài báo
- ★ II. Nội dung bài báo
- ★ III. Ưu điểm, hạn chế và ứng dụng
- ★ IV. Demo

I.Giới thiệu và lý do chọn bài báo

Segment Anything

1

Giới thiệu

2

*Lý do chọn
bài báo*

1. Giới thiệu

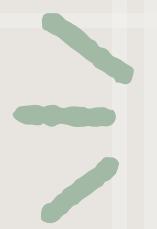
- Tên bài báo: Segment Anything
- Ngày công bố: 05/04/2023(arXiv: 2304.02643)
 - chính thức tại ICCV 2023
- Tác giả: 43 tác giả thuộc Meta FAIR với Alexander Kirillov là trưởng nhóm
- Giá trị học thuật:
 - + 30.000 trích dẫn (Google Scholar – 11/2025)
 - + Thuộc Top 3 paper được trích dẫn nhiều nhất thế giới năm 2023–2024
 - + Được trao Best Paper Award – ICCV 2023



2. Lý do chọn bài báo

- Đưa ra khái niệm hoàn toàn mới: “Foundation Model cho Image Segmentation”.
- Mã nguồn, trọng số và dataset SA-1B được công bố 100% miễn phí → cực kỳ thuận lợi cho sinh viên.
- Được cộng đồng sử dụng rộng rãi (hàng nghìn dự án phái sinh: SAM 2, MobileSAM, FastSAM, EfficientSAM, SAM-Med2D...).
- Có tiềm năng ứng dụng thực tế rất cao tại Việt Nam trong y tế, nông nghiệp, giao thông.





II. Nội dung bài báo

1

*Bài toán, động cơ
nghiên cứu, mục tiêu*

2

*Kiến trúc mô hình
SAM*

Segment Anything

3

*Dataset SA-1B và
Quy Trình Huấn
Luyện*

4

Kết quả

5

*So sánh với các
phương pháp khác*

1. Mục tiêu, bài toán và động cơ nghiên cứu

a. Bài toán

Vấn đề	Nội dung	
1	Mỗi dataset phải có một mô hình riêng	<ul style="list-style-type: none">• COCO, ADE20K, ... đều cần kiến thức hoặc mô hình được huấn luyện riêng• Không tồn tại “một mô hình chung cho tất cả”
2	Mỗi lớp đối tượng phải có nhãn trong bộ dữ liệu training	<ul style="list-style-type: none">• Nếu trong dataset không có lớp cat,motorbike,... thì mô hình không thể segment chính xác những lớp đó
3	Không có mô hình nào có thể “segment bất kỳ thứ gì” chỉ bằng một thao tác đơn giản	Các mô hình truyền thông không hỗ trợ việc segment qua prompt (ví dụ text, click,box)
4	Chi phí gắn nhãn quá tốn kém	<ul style="list-style-type: none">• Pixel-level annotation mất rất nhiều thời gian• Ngay cả COCO- một trong những dataset lớn nhất- cũng chỉ có:<ul style="list-style-type: none">+ 80 lớp+ 118k ảnh+ 820k mask



=> **Đây là bài toán căn bản:** *Làm sao tạo ra một mô hình segmentation có khả năng khái quát hóa cực mạnh và hoạt động được với bất kỳ loại đối tượng nào trong ảnh?*

1. Mục tiêu, bài toán và động cơ nghiên cứu

b. Động cơ nghiên cứu

Những hạn chế ở trên tạo ra nhu cầu cấp bách:

- Cần một mô hình duy nhất, không phải huấn luyện theo từng dataset.
- Cần khả năng segment cả những đối tượng chưa từng thấy (zero-shot).
- Cần một cơ chế tương tác linh hoạt, tương tự như prompt trong NLP.
- Cần một nguồn dữ liệu cực lớn để mô hình có tính tổng quát mạnh.
- Cần giảm chi phí gắn nhãn pixel-level bằng một cách tạo mask bán tự động.

=> Tất cả dẫn đến động lực phát triển một “foundation model” dành cho hình ảnh, giống như ChatGPT là foundation model cho ngôn ngữ.



1. Mục tiêu, bài toán và động cơ nghiên cứu

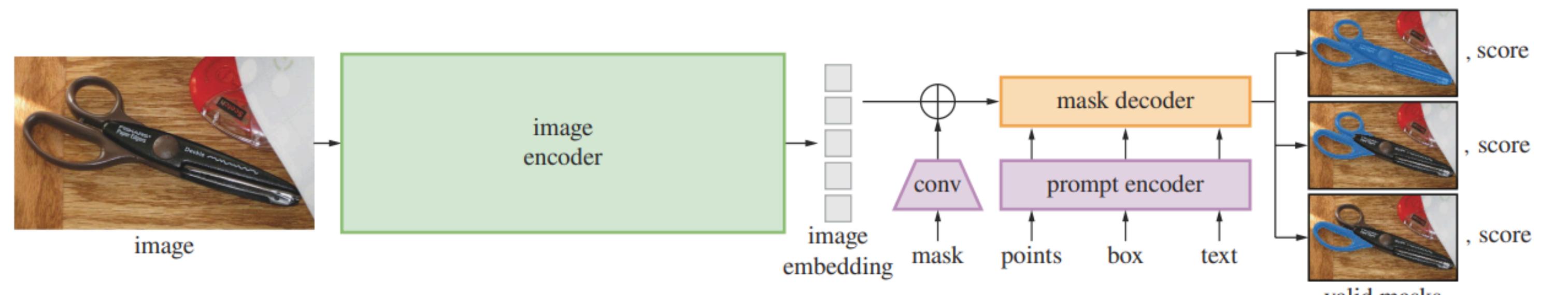
c. Mục tiêu

Với những động cơ đó, Meta đã đặt mục tiêu táo bạo khi phát triển SAM (2023):

“Xây dựng một mô hình duy nhất, một kiểu nhiệm vụ duy nhất và một cách tương tác duy nhất có thể giải quyết gần như mọi bài toán segmentation.”

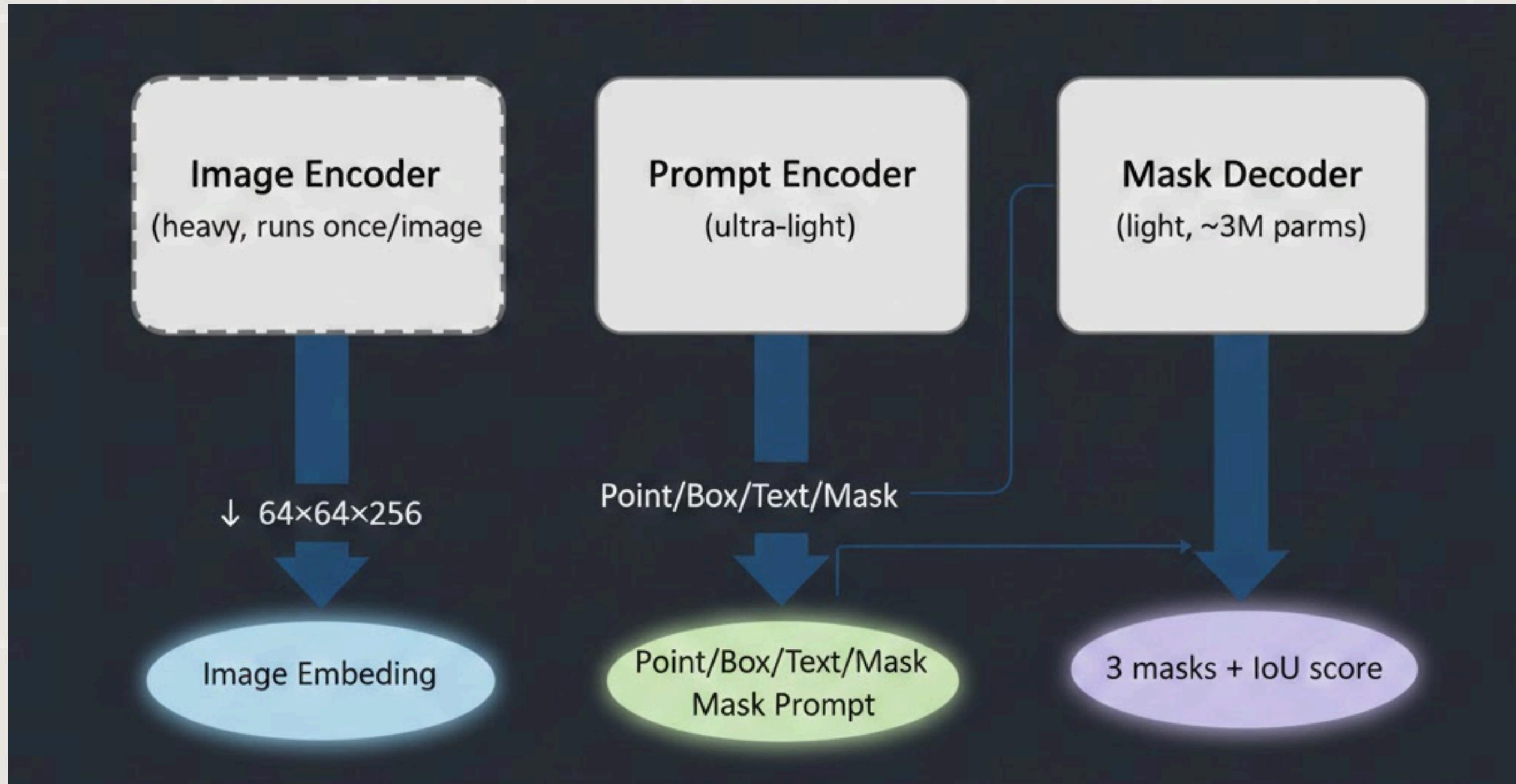


Pipeline tóm quan



2. Kiến trúc SAM

SAM có 3 thành phần chính – thiết kế cực kỳ thông minh để vừa mạnh vừa nhanh:



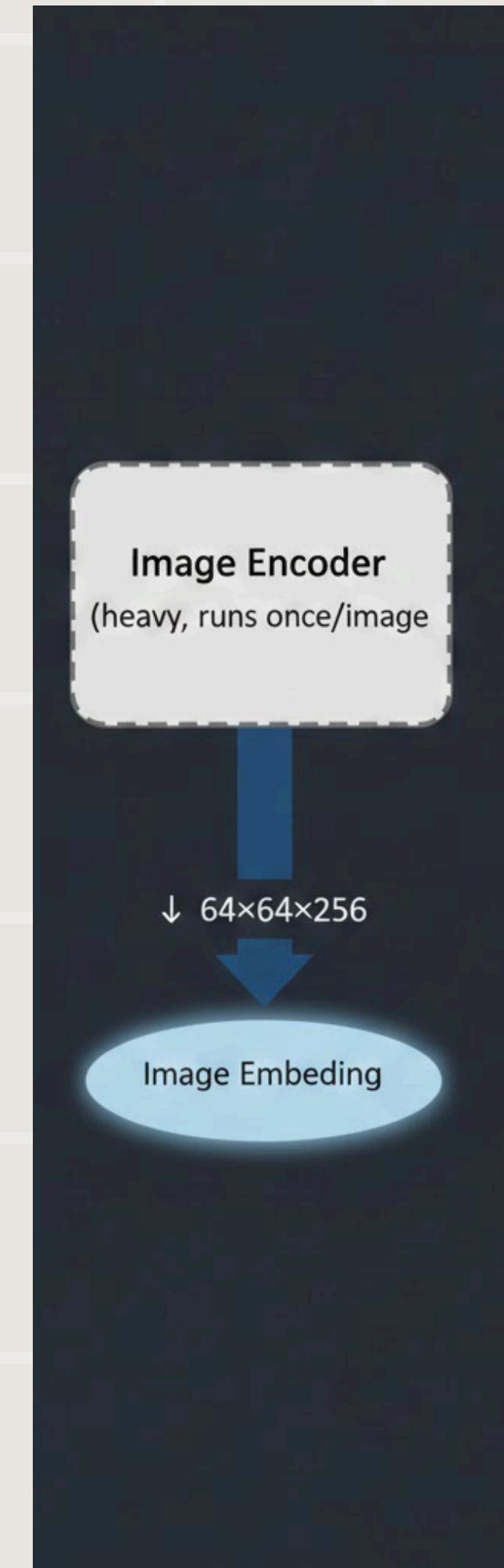
2. Kiến trúc SAM

a. *Image Encoder*

Image Encoder là *thành phần quan trọng nhất* của mô hình Segment Anything. Nó nhận ảnh đầu vào và biến đổi thành các image embeddings – dạng biểu diễn đặc trưng có thể sử dụng cho mọi loại prompts (point, box, text,...).

Vai trò:

- Được xem như “*bộ não chính*” của SAM.
- Chỉ cần *chạy 1 lần duy nhất* cho mỗi ảnh, sau đó mọi lần người dùng đưa prompt (điểm, hộp, nét vẽ...) đều chỉ chạy phần Mask Decoder rất nhanh.
- Giúp SAM trở thành *real-time interactive segmentation*.

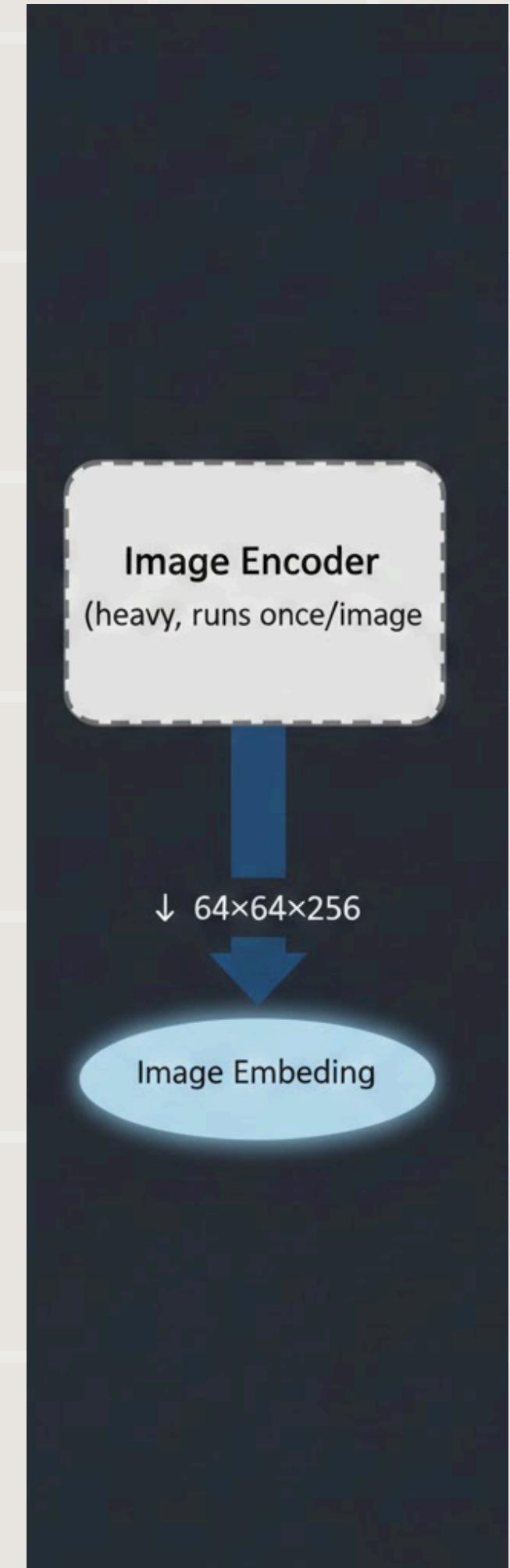


2. Kiến trúc SAM

a. Image Encoder

Kiến trúc: SAM dùng một backbone cực lớn: ViT-H
(Vision Transformer – Huge, 632M parameters)

- Patch size: 16×16
- 32 attention blocks
- 14×14 window attention
- Cấu trúc transformer giúp học được quan hệ toàn cục trong ảnh (global context).



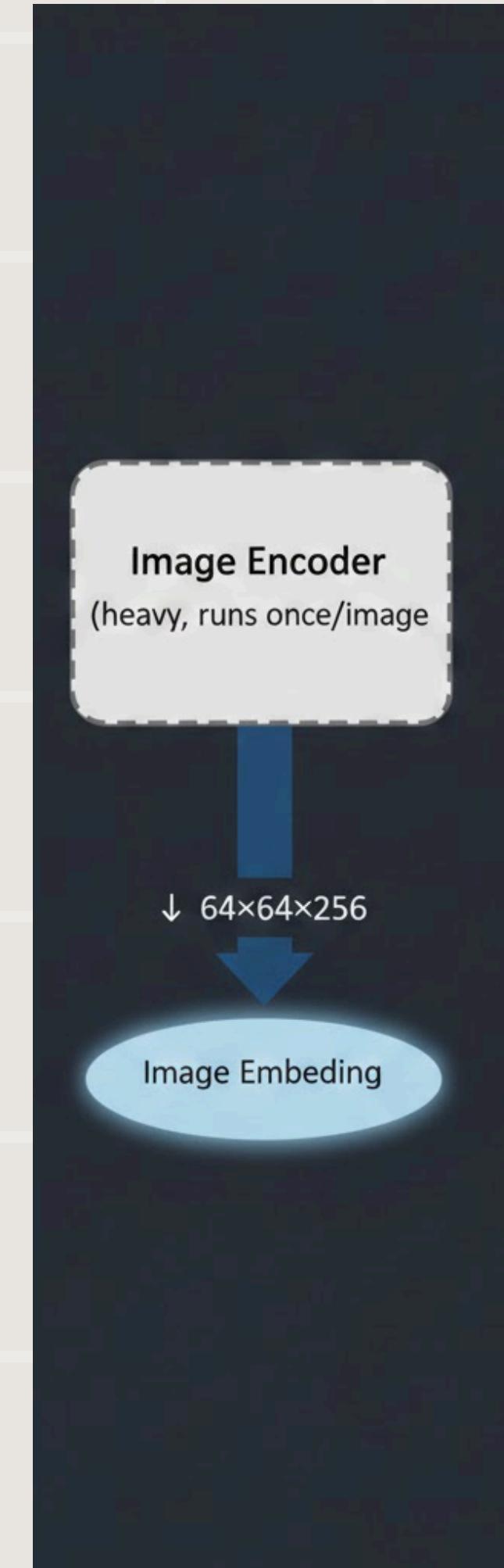
2. Kiến trúc SAM

a. *Image Encoder*

Output của Image Encoder:

Encoder không dự đoán mask. Nó chỉ tạo ra:

- Image embedding kích thước $256 \times 64 \times 64$
- Bao gồm thông tin:
 - Màu sắc
 - Kết cấu
 - Biên dạng
 - Ngữ cảnh không gian
 - Feature đa cấp độ
- Embeddings này được lưu lại → các bước segment sau lặp lại cực nhanh (vì không cần chạy encoder).

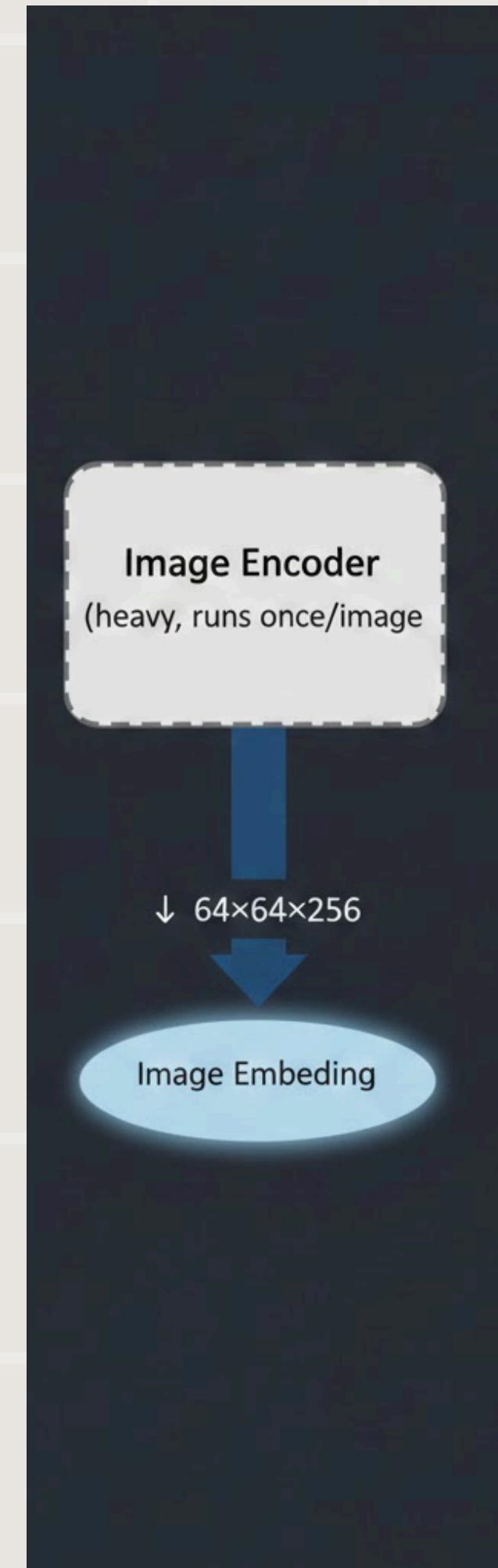


2. Kiến trúc SAM

a. *Image Encoder*

Ưu điểm lớn:

- Tách rời image encoder và prompt decoder
→ SAM trở thành foundation model, dùng 1 embedding cho vô hạn prompts.
- Tương tác real-time
 - + Image Encoder: chạy ~400ms
 - + Prompt + Mask Decoder: 5–15ms
- Ổn định với mọi loại dữ liệu (zero-shot)
- Nhờ training trên SA-1B (11 triệu ảnh + 1 tỷ masks).



2. Kiến trúc SAM

b. *Prompt Encoder*

+ Point Prompt(điểm nhập)

Point Embedding = Positional Encoding + Point-Type Embedding

- Positional Encoding của tọa độ (x,y)

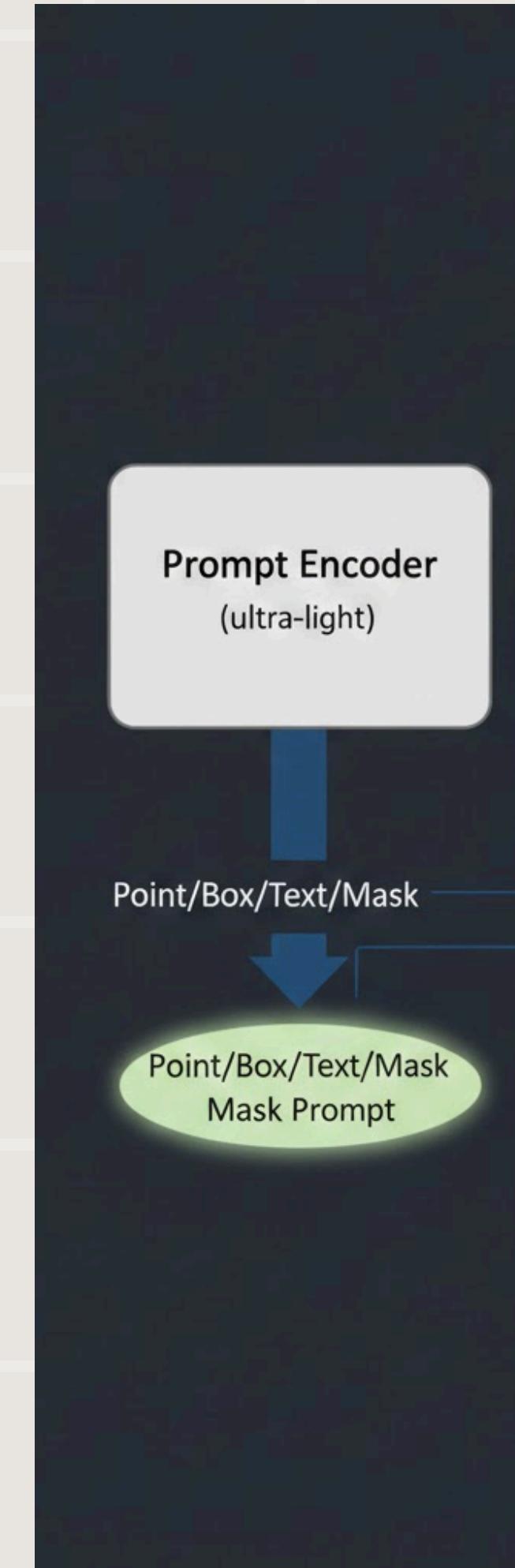
SAM sẽ lấy tọa độ chuẩn hóa: $(x,y) \rightarrow PE(x,y)$ rồi tạo sinusoidal positional encoding giống Transformers (tức là biến vị trí(pos) thành một vector gồm nhiều sóng sin/cos với nhiều tần số)

- Embedding học được cho 2 loại điểm(foreground/background)

SAM dùng 2 vector embedding riêng biệt để biểu diễn:

E_{fg} (điểm foreground ý nói “vật nằm ở đây”)

E_{bg} (điểm background ý nói “vật không nằm ở đây”)



2. Kiến trúc SAM

b. *Prompt Encoder*

+ Box Prompt(hộp):

Box bao gồm: Góc top_left (x1,y1) và góc bottom_right(x2,y2) =>

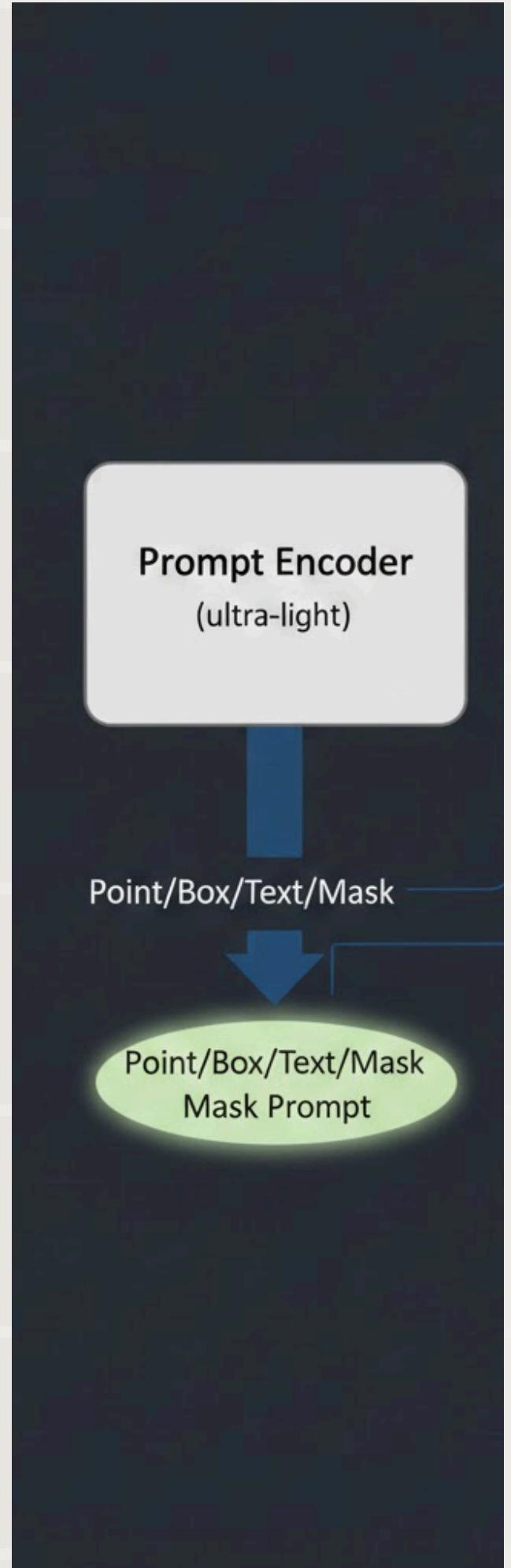
SAM tạo 2 vector embedding: BoxEmbedding = PE + Box_type

PE cho từng điểm, mỗi góc sẽ thành PE(x1,y1) và PE(x2,y2)

Embedding học được 2 loại góc hộp : E_top_left và E_bottom_right

=> Như vậy sẽ tạo ra 2 vector 256dim đưa vào mask decoder

Ý nghĩa: Mô hình cho biết bạn đang cung cấp phạm vi bao phủ của vật thể



2. Kiến trúc SAM

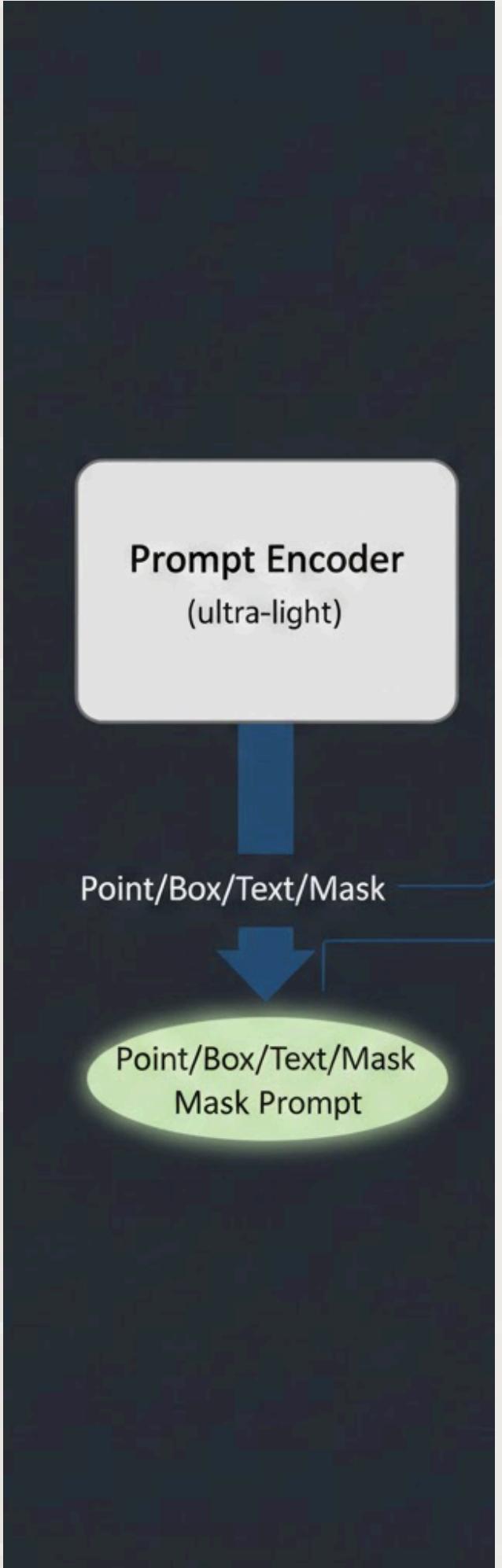
b. *Prompt Encoder*

+ Mask prompt:

Là bộ “dịch mask thành thông tin mà SAM hiểu được”.

Bạn đưa mask thô → SAM biến nó thành dạng gọn nhẹ → trộn vào ảnh
→ giúp SAM khoanh đúng hơn.

- Mask (ảnh đen-trắng) được đưa vào một bộ xử lý nhỏ gồm vài lớp lọc ảnh (convolution) để thu gọn kích thước và rút ra thông tin quan trọng.
- Kết quả là SAM tạo ra một “bản tóm tắt” của mask với 256 kênh thông tin.
- Bản tóm tắt này được ghép trực tiếp vào thông tin đặc trưng của ảnh, ở đúng vị trí pixel tương ứng.
- Nếu bạn không đưa mask, SAM dùng một mask mặc định đã được học sẵn.
- Phần xử lý mask rất nhỏ, chỉ có vài nghìn – vài chục nghìn tham số, không đáng kể so với mô hình chính

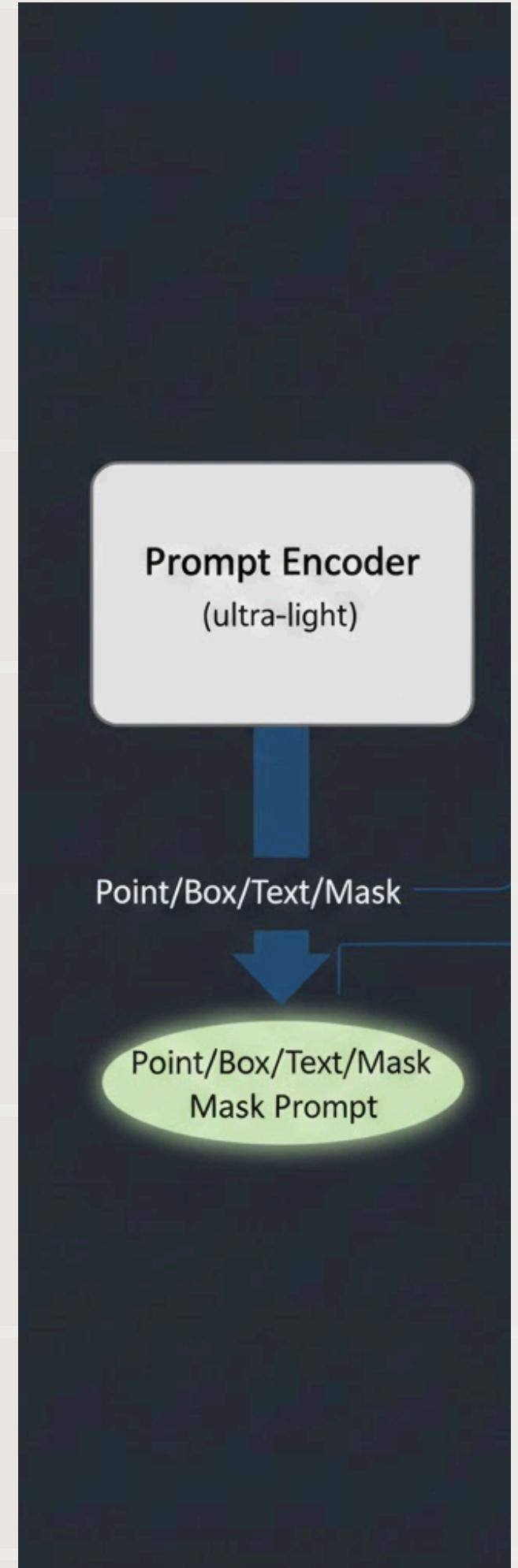


2. Kiến trúc SAM

b. *Prompt Encoder*

Tóm lại:

Loại Prompt	Cách mã hoá (rất đơn giản)
Point	Positional encoding + foreground/background token
Box	2 góc (top-left+ bottom-right)-> 2 vector 256-dim
Mask (mask input)	Convolution-> embed thành 1 vector 256- dim



2. Kiến trúc SAM

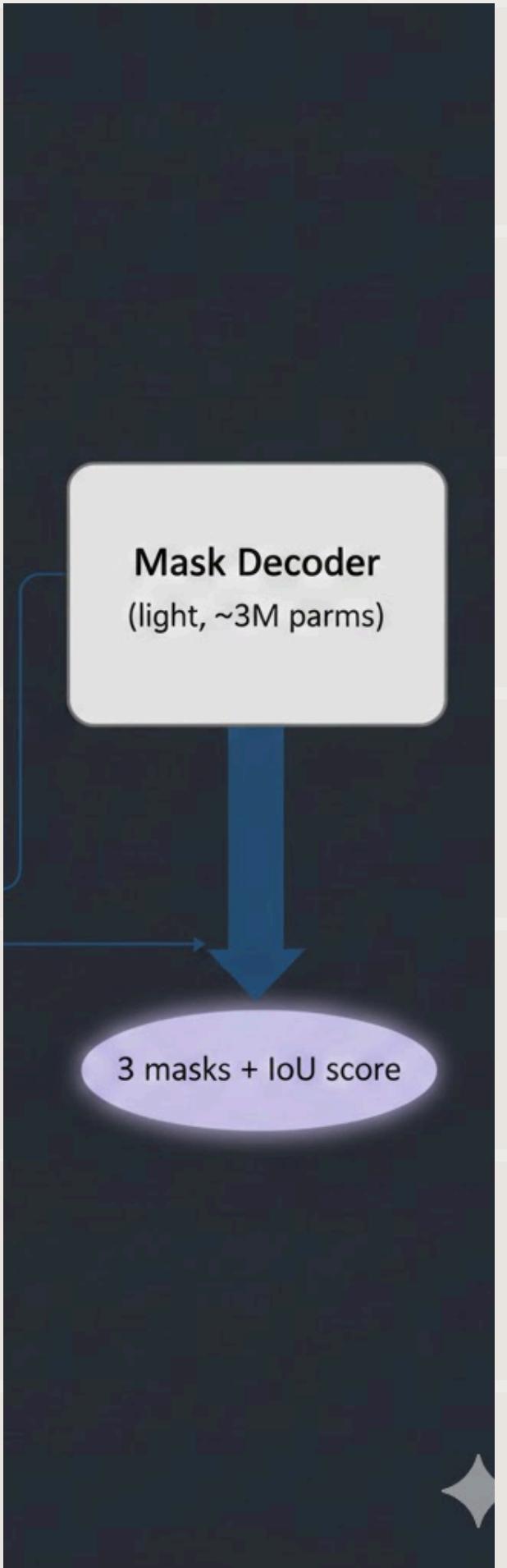
c. Mask Decoder

Có thể nói đây là “linh hồn” của SAM

Mask Decoder là nơi SAM quyết định khoanh vùng cái gì.

Dù rất nhỏ (~3 triệu tham số) nhưng lại là phần quan trọng nhất.

Cơ chế hoạt động chính gồm 4 bước



2. Kiến trúc SAM

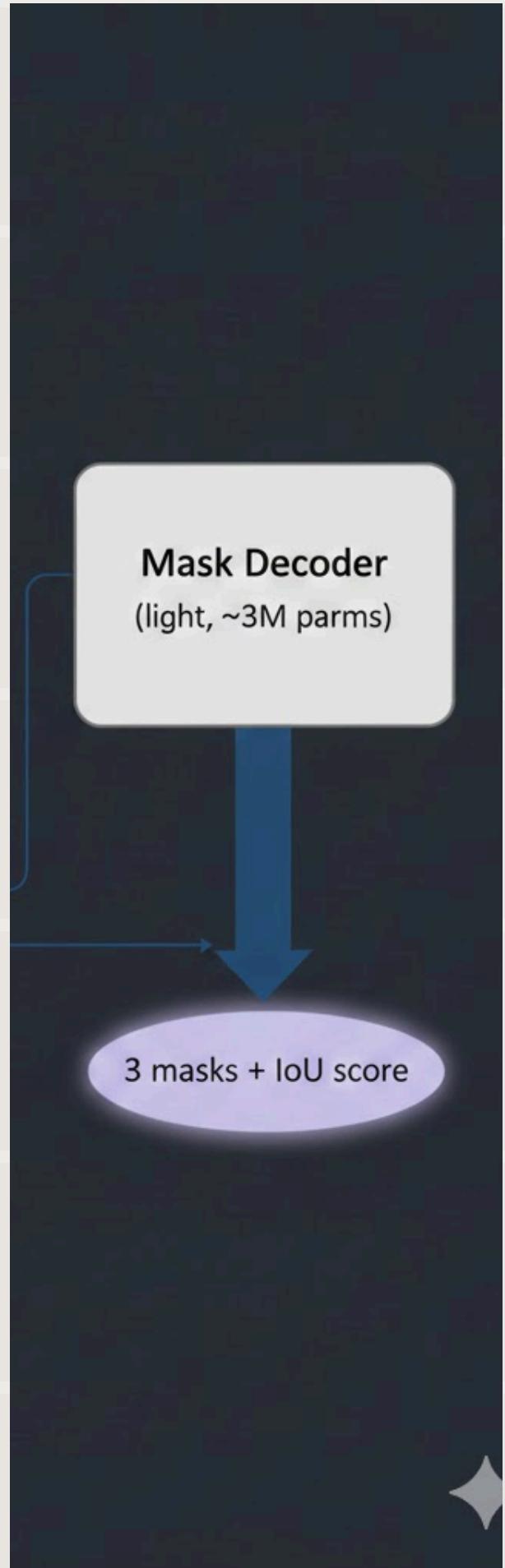
c. Mask Decoder

Bước 1: Nhận các loại “gợi ý” (prompt)

Tất cả những gì người dùng đưa cho SAM (click điểm, vẽ box, đưa mask cũ) đều được đổi thành một dạng vector mà mô hình hiểu được:

- Điểm (Point) → vị trí + loại điểm (chọn / loại bỏ)
- Hộp (Box) → hai góc của hộp
- Mask cũ → nén lại thành 1 gói thông tin nhỏ

→ Hiểu đơn giản: SAM biến các thao tác của người dùng thành “ngôn ngữ nội bộ”.



2. Kiến trúc SAM

c. Mask Decoder

Bước 2: Hai vòng xử lý (giống như trao đổi thông tin)

Trong mỗi vòng:

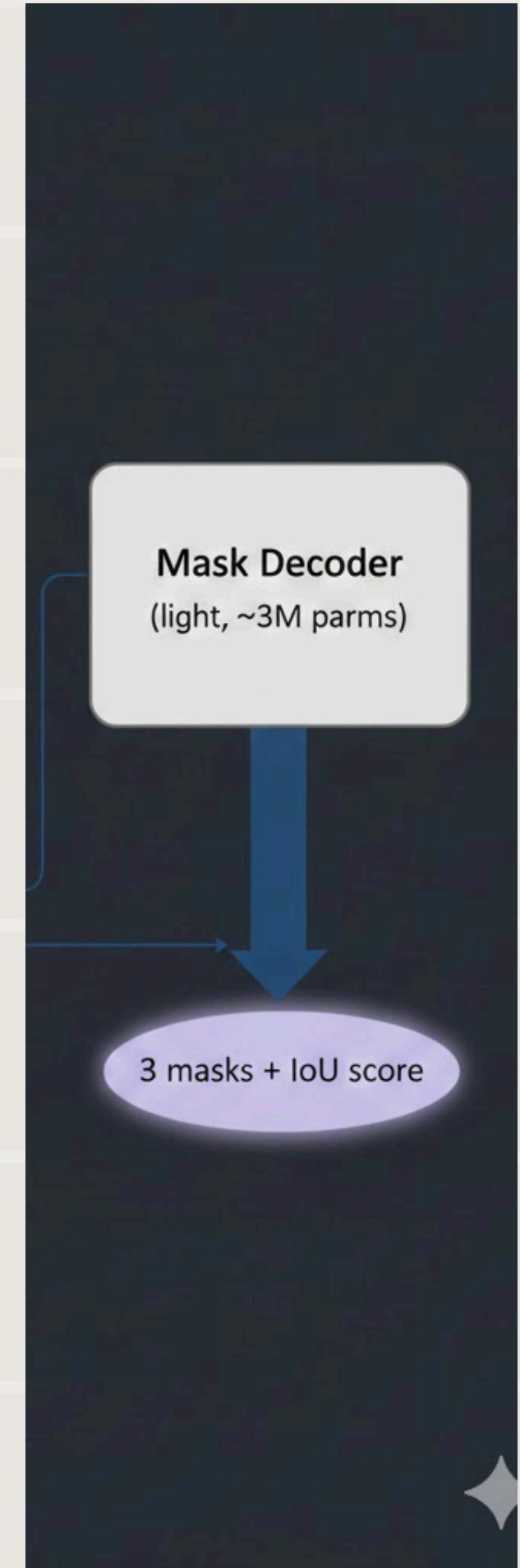
(a) Prompt tự trao đổi với nhau

- Điểm nói chuyện với điểm, box, mask.
- → Giúp SAM hiểu: người dùng đang chọn vùng nào, bỏ vùng nào.

(b) Prompt trao đổi với Ảnh & Ảnh trao đổi lại

- Prompt tìm xem vùng ảnh nào liên quan.
- Ảnh gửi ngược lại thông tin chi tiết.

→ Đây là bước SAM hiểu ý định của người dùng.



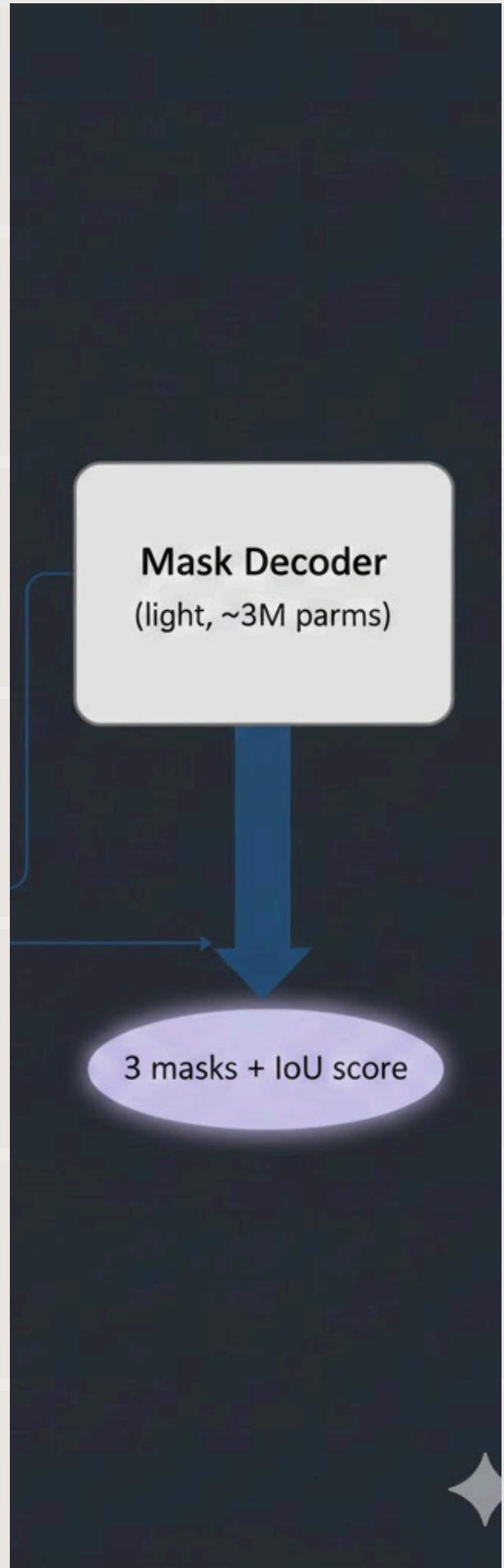
2. Kiến trúc SAM

c. Mask Decoder

Bước 3: Tạo mask

Sau khi hiểu đủ:

- Thông tin ảnh được làm rõ nét lại
- Một token “đặc biệt” tạo ra bộ lọc
- Bộ lọc này được áp vào ảnh và mask được sinh ra
→ Giống như “vẽ lại vùng mà người dùng đang muốn chọn”.



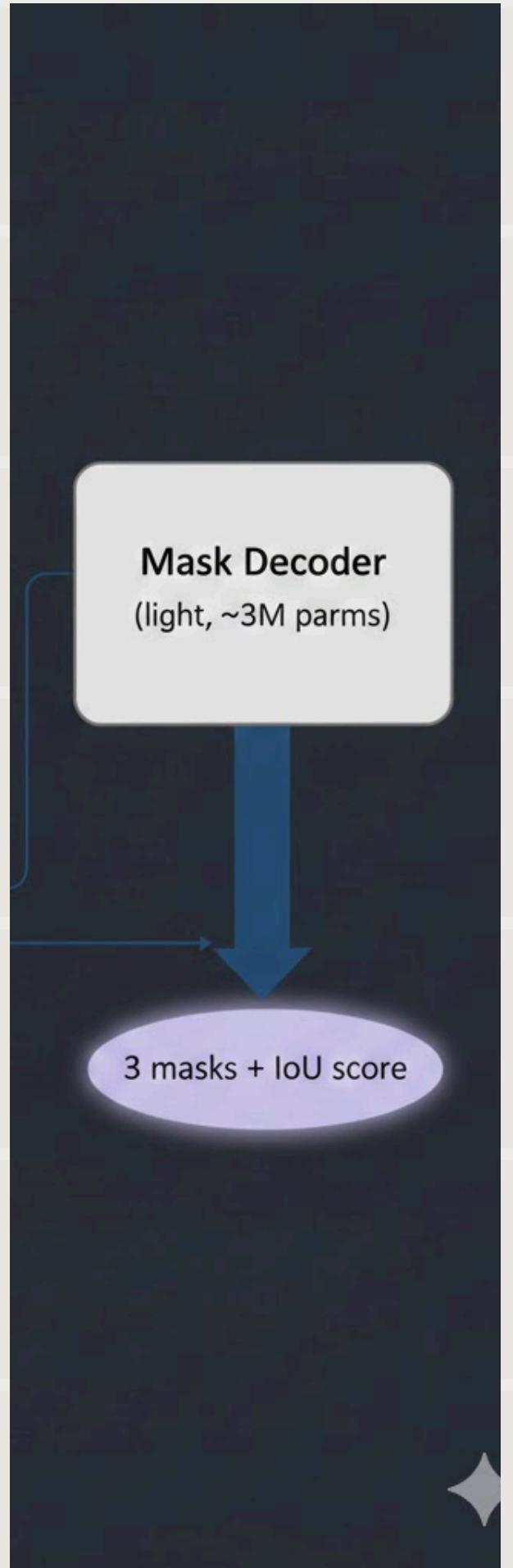
2. Kiến trúc SAM

c. Mask Decoder

Bước 4: Xuất kết quả cuối cùng

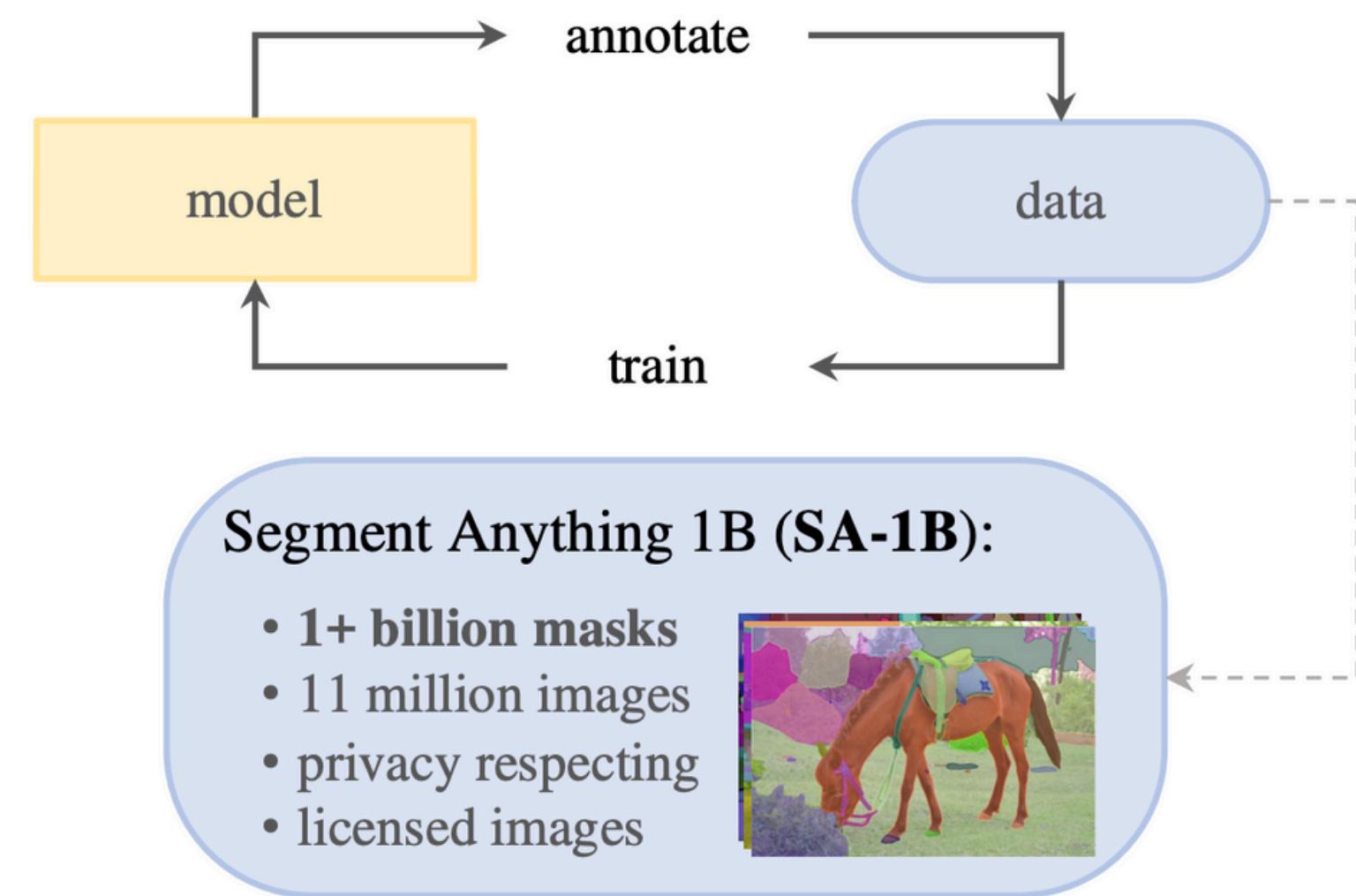
Mask Decoder trả về:

- 3 mask khác nhau (vùng lớn, vùng nhỏ, vùng tinh)
 - 3 điểm tin cậy
- SAM chọn mask tốt nhất trả lại cho bạn.



3. Dataset SA-1B và Quy Trình Huấn Luyện

- ✓ Dataset lớn nhất lịch sử: 1.1 TỶ MASK
- ✓ Data Engine: 3 giai đoạn tự động hóa
- ✓ 3 giai đoạn huấn luyện chuyên sâu



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

Dataset SA-1B - Bộ Dữ Liệu Segmentation Lớn Nhất Lịch Sử

SA-1B: SEGMENT ANYTHING 1-BILLION

11 TRIỆU ẢNH | 1.1 TỶ MASK

- Licensed & Privacy-respecting
- Độ phân giải: 3300×4950px
- ~100 mask/ảnh trung bình



Tiêu chí	SA-1B	COCO [64]	Open Images [58]	LVIS [43]	Tỷ lệ so với lớn nhất trước đó
Số ảnh	11 triệu	330k	9 triệu	100k	×1.2 (so Open Images)
Số mask	1.1 tỷ	2.5 triệu	2.8 triệu	2.2 triệu	×400 (so Open Images)
Độ phân giải ảnh trung bình	3300×4950	480×640	Thấp	Trung bình	Cao hơn đáng kể
Số mask trung bình/ảnh	~100	~7	~0.3	~22	Cao hơn nhiều
Độ đa dạng đối tượng	Hàng nghìn lớp tự nhiên (things + stuff)	91 lớp	500 lớp	1.2k lớp	Rất cao, bao phủ góc cạnh ảnh

ĐA DẠNG & CHẤT LƯỢNG SA-1B

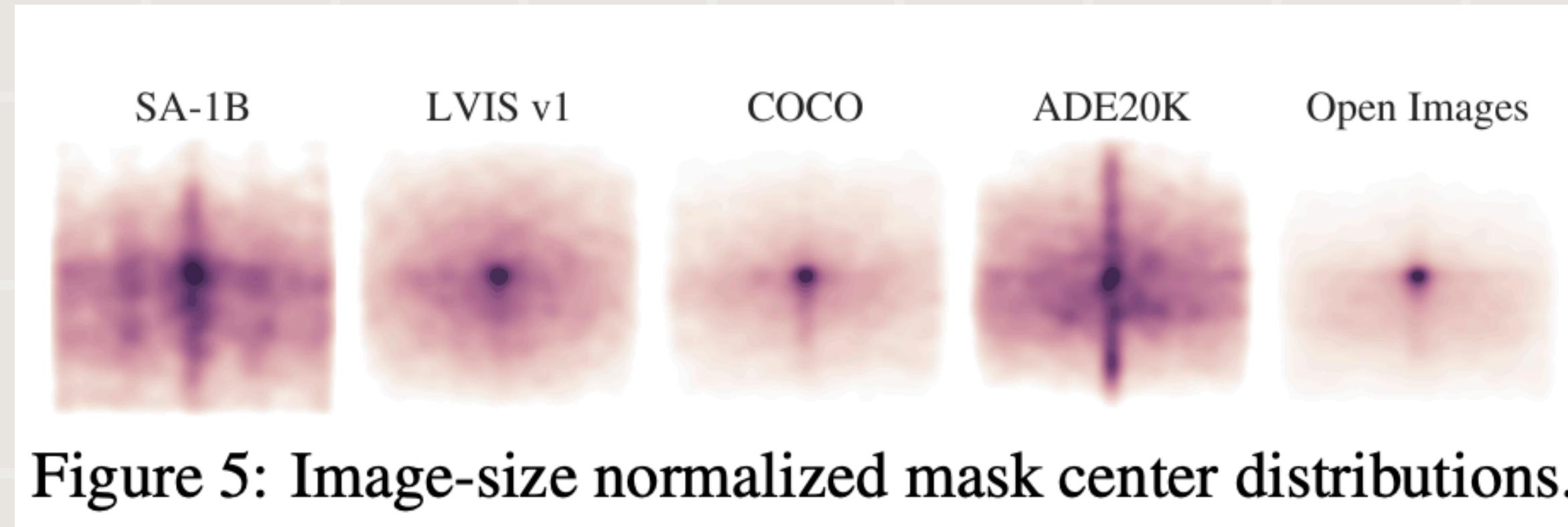


Figure 5: Image-size normalized mask center distributions.

PHÂN BỐ MASK CENTERS

SA-1B: ĐỀU TRỌNG TÂM

→ GÓC ẢNH COCO/LVIS: THIÊN VỀ GIỮA ẢNH

CHẤT LƯỢNG: 94% mask auto có IoU > 90% so với pro annotation

≈ Độ nhất quán giữa annotator con người (85-91%)

DATA ENGINE: VÒNG LẶP MÔ HÌNH-DỮ LIỆU

Giai đoạn 1	Giai đoạn 2	Giai đoạn 3
Assisted-Manual	Semi-Automatic	Fully Automatic
120K ảnh 4.3M mask 120 annotators	180K ảnh 5.9M mask +Auto detect	11M ảnh 1.1B mask 99.1% auto

↻ RETRAIN SAM: 11 LẦN TỔNG CỘNG

⌚ 2 NĂM | 120 ANNOTATORS FULL-TIME

GIAI ĐOẠN 1: HỖ TRỢ THỦ CÔNG

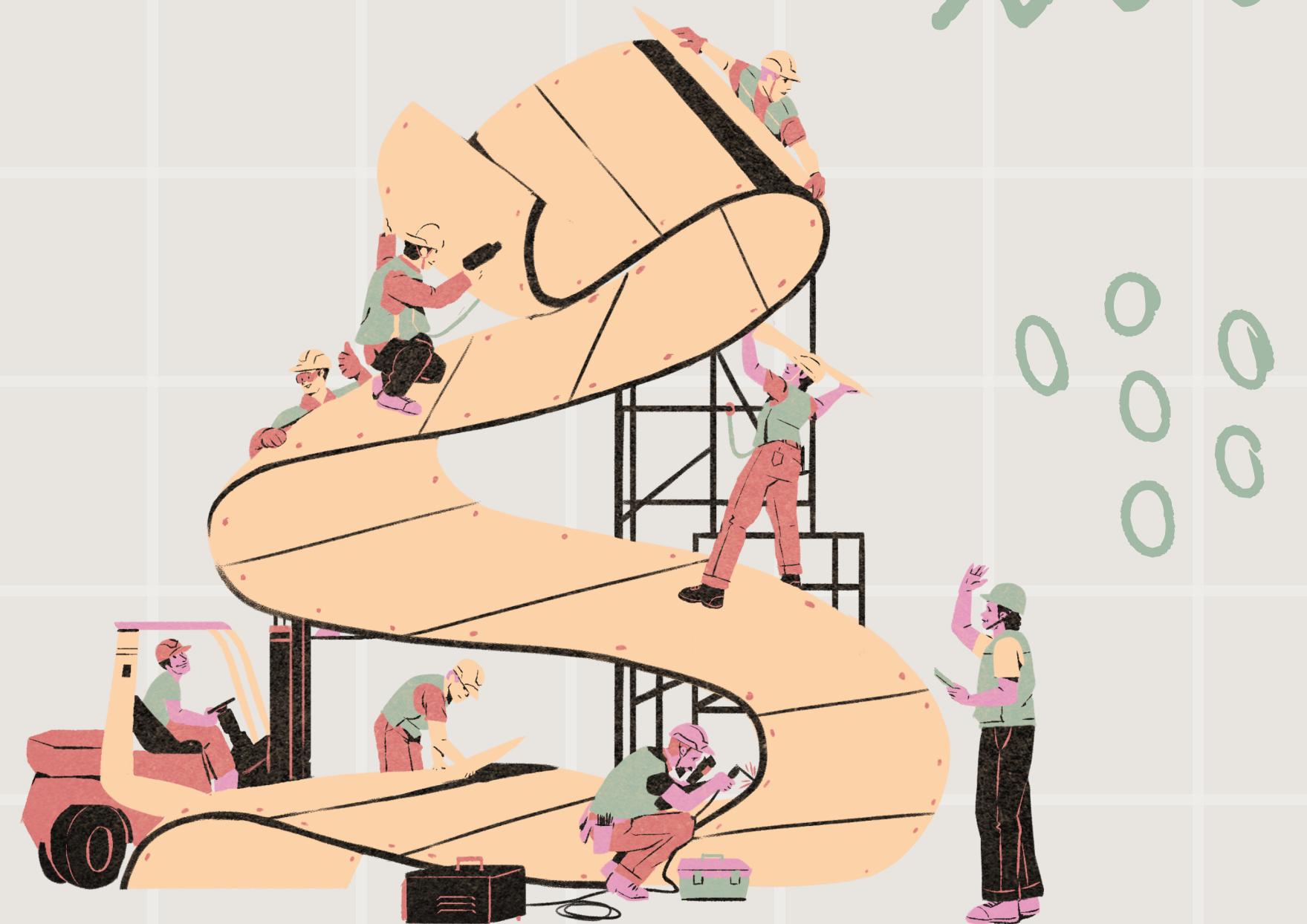
🖱️ Interactive Segmentation Tool

- Click FG/BG points
- Brush/Eraser pixel-precise
- Real-time (~50ms/prompt)

📊 KẾT QUẢ:

- 120K ảnh → 4.3M mask
- Thời gian/mask: 34s → 14s
- Mask/ảnh: 20 → 44
- RETRAIN: 6 lần

🎯 Nguyên tắc: "Segment bất kỳ thứ gì có thể mô tả" (không cần semantic)



GIAI ĐOẠN 2: BÁN TỰ ĐỘNG

🤖 AUTO + HUMAN

1. Train Box Detector (Detectron2)

trên 4.3M mask giai đoạn 1

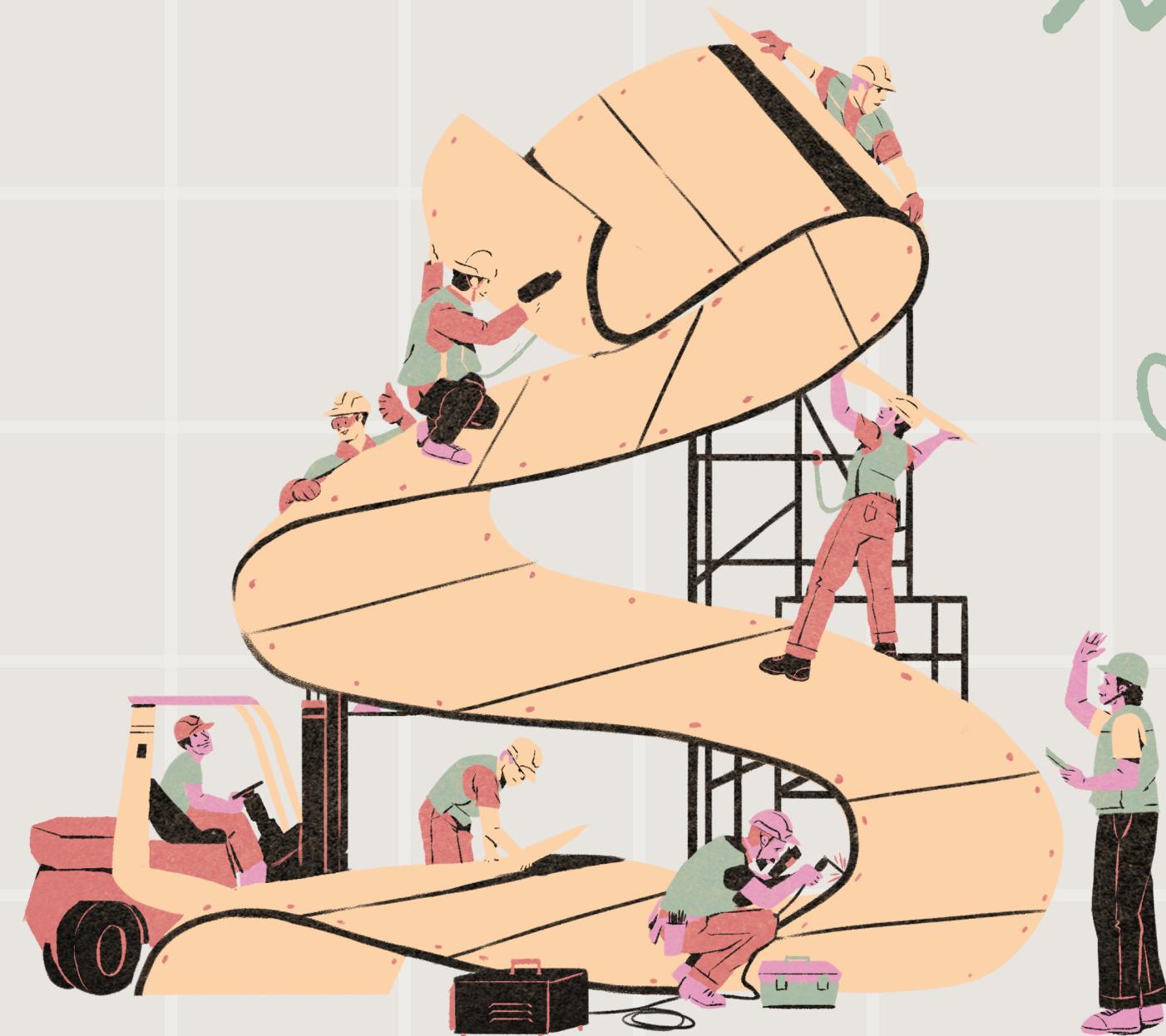
2. SAM predict confident masks

3. Annotator: THÊM mask ít nổi bật

- Merge/Split overlap
- Object khó

📊 KẾT QUẢ:

- 180K ảnh → 5.9M mask
- TỔNG: 10.2M mask
- Mask/ảnh: 72
- RETRAIN: 5 lần



GIAI ĐOẠN 3: HOÀN TOÀN TỰ ĐỘNG

🚀 SAM TỰ GENERATE 1.1B MASKS

QUY TRÌNH:

1. 32×32 FG point grid
2. Predict 3 masks/point (subpart, part, whole)

3. FILTER:

- Confidence score > threshold
- Stable masks ($\text{IoU}@0.5 \pm \delta$)
- NMS loại trùng lặp

4. Zoomed-in crops cho mask nhỏ

- ✓ 11M ảnh \rightarrow 1.1B mask
- ✓ 99.1% automatic



QUY TRÌNH HUẤN LUYỆN 3 GIAI ĐOẠN

Giai đoạn	Mục tiêu	Data
1.MAE Pretrain	Image Encoder (ViT-H 632M)	11M ảnh (no mask)
2.Mask Prediction	Mask Prediction +IoU	1.1B triplets
3.Ambiguity Handling	All prompts (point/box/..)	+Edge cases



TASK: Promptable Segmentation

- 11 interactive rounds/mask
- Predict VALID mask (even ambiguous)

GIAI ĐOẠN 1: MAE PRE-TRAINING

🎯 MỤC TIÊU: Robust Image Features

KIẾN TRÚC:

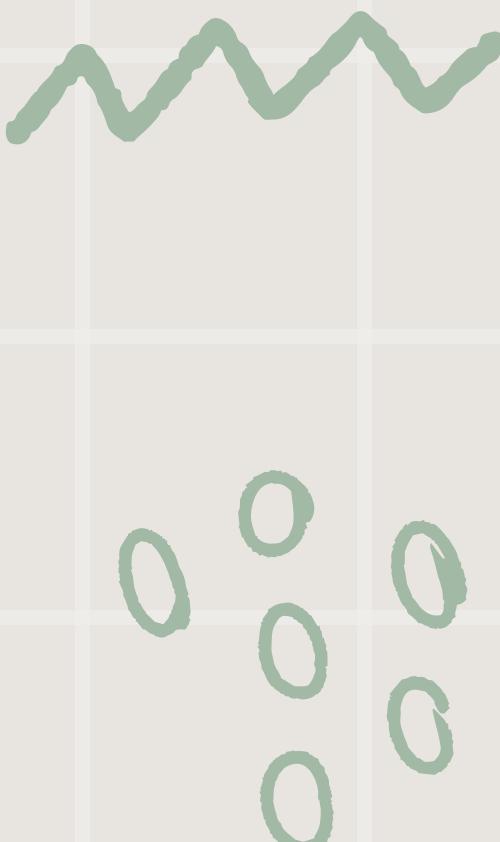
ViT-Huge (632M params)

75% patches MASKED → Reconstruction

HYPERPARAMS:

Epochs	Batch	LR	Optimizer
800	4096	1.5e-4	AdamW

✓ KẾT QUẢ: SOTA feature quality
 (= DINOv2 + CLIP)



GIAI ĐOẠN 2: MASK PREDICTION

INPUT: Image embedding ($64 \times 64 \times 256$) + Prompt embedding

MASK DECODER (3M params):

2 Transformer layers

➡ Cross-attention (prompt \leftrightarrow image)

→ 3 masks + IoU scores

LOSS: Focal + Dice

DATA: 1.1B triplets Prompts: 50% point | 30% box | 20% mask

HYPERPARAMS:

Epochs	Batch	LR
12	2048	6e-5



GIAI ĐOẠN 3: FULL PROMPT FLEXIBILITY

🎯 XỬ LÝ AMBIGUITY & MQI PROMPT

PROMPT TYPES:

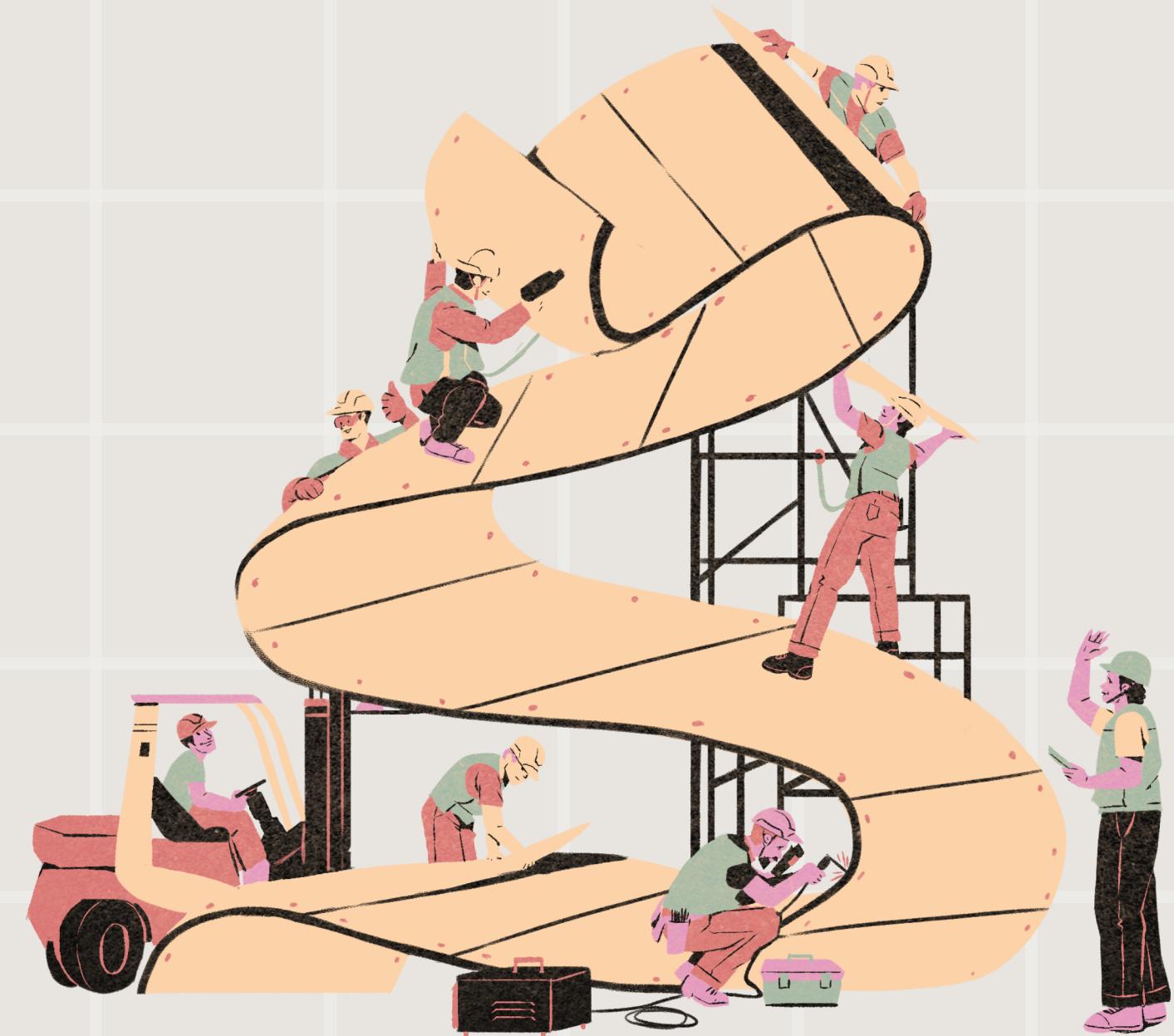
- Point (single/multi + BG)
- Box
- Mask
- Everything (no prompt)
- Combined

AMBIGUITY-AWARE:

- Predict 3 masks/prompt
- Backprop MIN loss
- IoU score để RANK

LOSS: Focal + Dice + $0.2 \times$ Quality

EPOCHS: 8 | LR: 3e-5

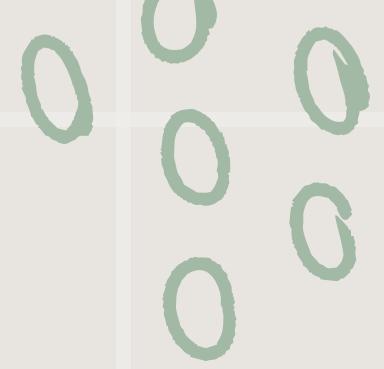


KẾT QUẢ SA-1B + 3 GIAI ĐOẠN TRAINING

🔥 ĐỘT PHÁ:

1. SCALE: 1.1B masks → Generalization
2. PROMPT ENGINEERING: All types
3. EFFICIENCY: Decoder chỉ 3% params
4. DATA ENGINE: Tự động hóa hoàn toàn

SAM = ĐẦU TIÊN "SEGMENT ANYTHING"!



4. Kết quả

a. Đánh giá Zero-Shot từ Một điểm đơn lẻ

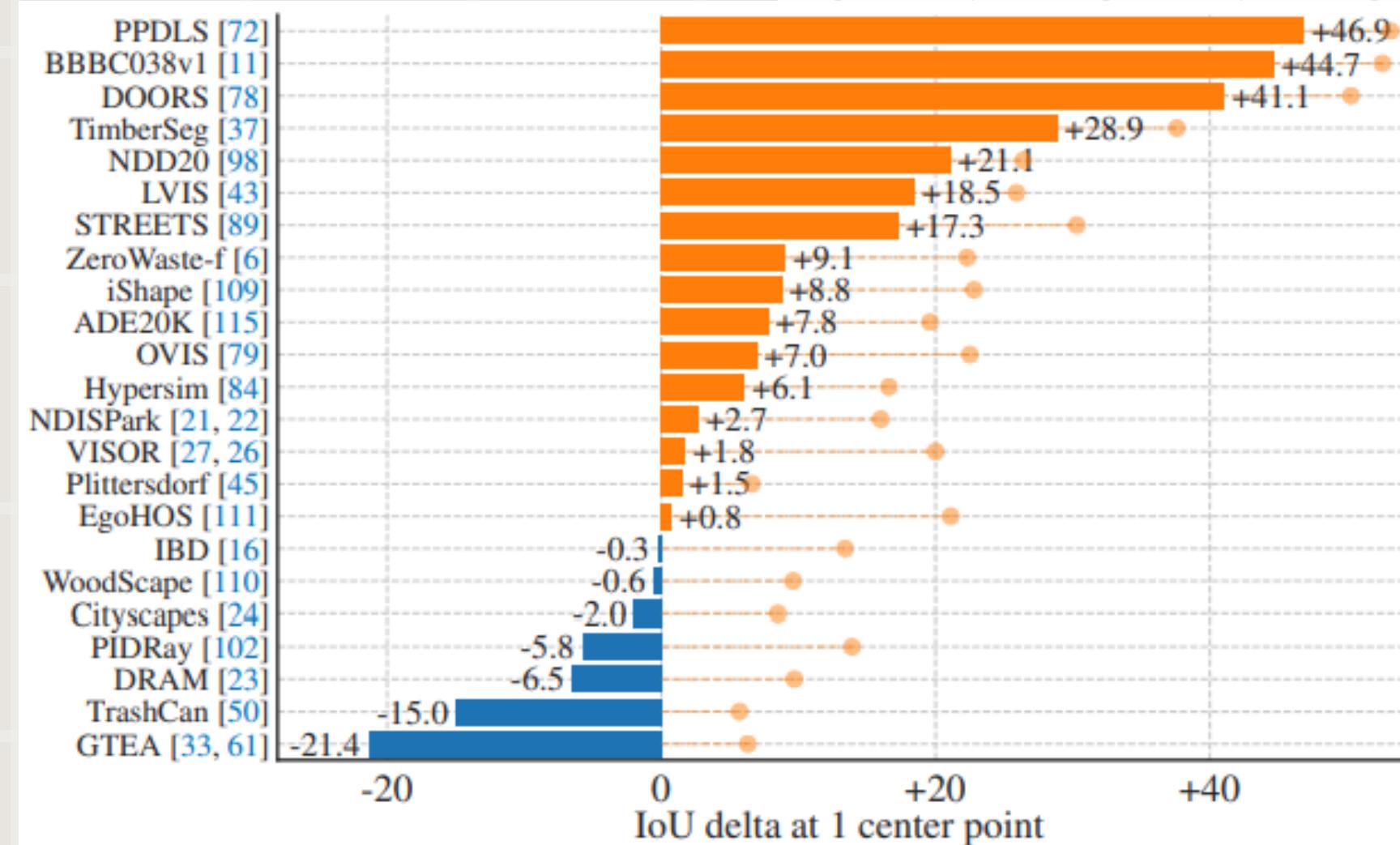
Tác vụ cốt lõi mà SAM được đánh giá là tạo ra một mask phân đoạn hợp lệ từ chỉ một điểm tiền cảnh duy nhất làm prompt. Đây là một tác vụ khó vì một điểm có thể mơ hồ và đề cập đến nhiều đối tượng.

4. Kết quả

0

a. Đánh giá Zero-Shot từ Một điểm đơn lẻ

SAM được so sánh với RITM - phương pháp phân đoạn tương tác:



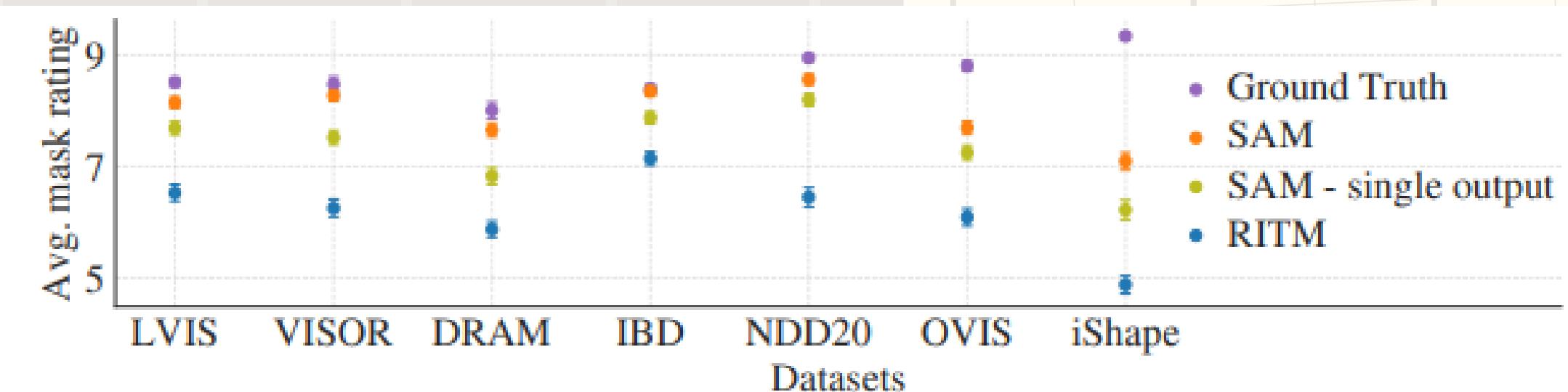
(b) SAM vs. RITM [90] on 23 datasets

4. Kết quả

a. Đánh giá Zero-Shot từ Một điểm đơn lẻ

Do sự mơ hồ khiến các chỉ số tự động như mIoU không hoàn toàn đáng tin cậy nên một số nghiên cứu của con người đã được thực hiện:

- Annotators đánh giá chất lượng mask của SAM cao hơn đáng kể so với RITM một cách nhất quán.



(c) Mask quality ratings by human annotators

4. Kết quả

a. Đánh giá Zero-Shot từ Một điểm đơn lẻ

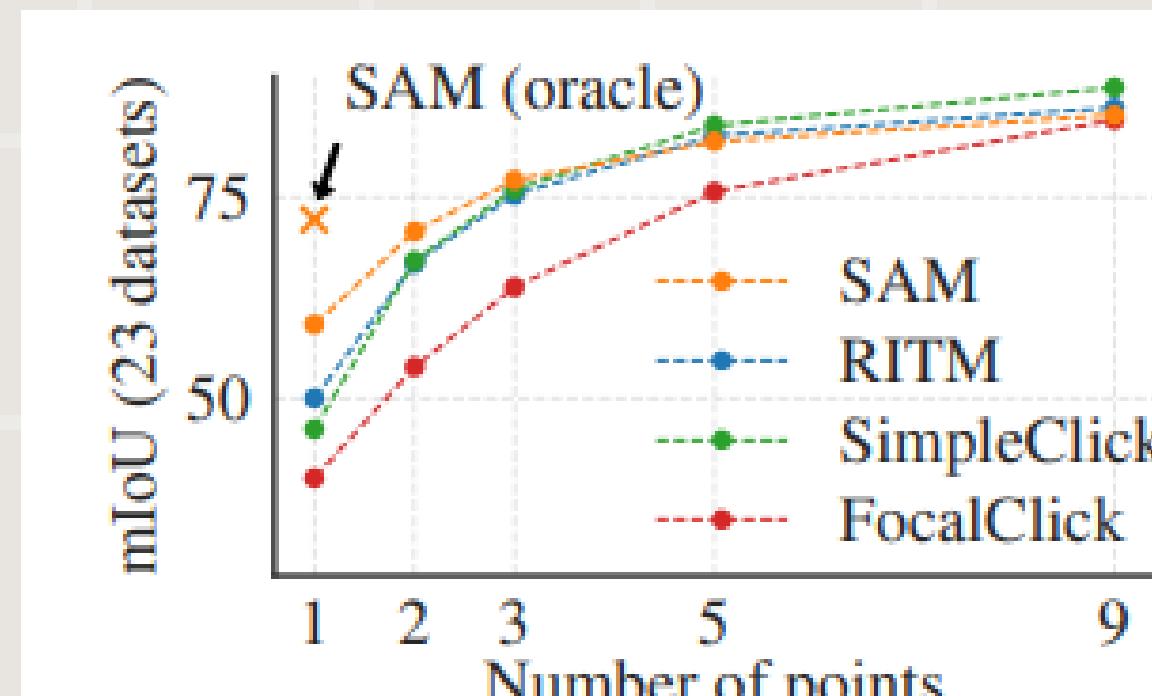
Do sự mơ hồ khiến các chỉ số tự động như mIoU không hoàn toàn đáng tin cậy nên một số nghiên cứu của con người đã được thực hiện:

- Chất lượng Mask cao : Xếp hạng trung bình của SAM nằm trong khoảng 7-9.
- Xử lý mơ hồ hiệu quả : Phiên bản SAM không nhận biết sự mơ hồ có xếp hạng thấp hơn một cách nhất quán. Ngay cả trên các tập dữ liệu mà SAM có mIoU tự động thấp hơn RITM, SAM vẫn nhận được xếp hạng chất lượng cao hơn từ con người.

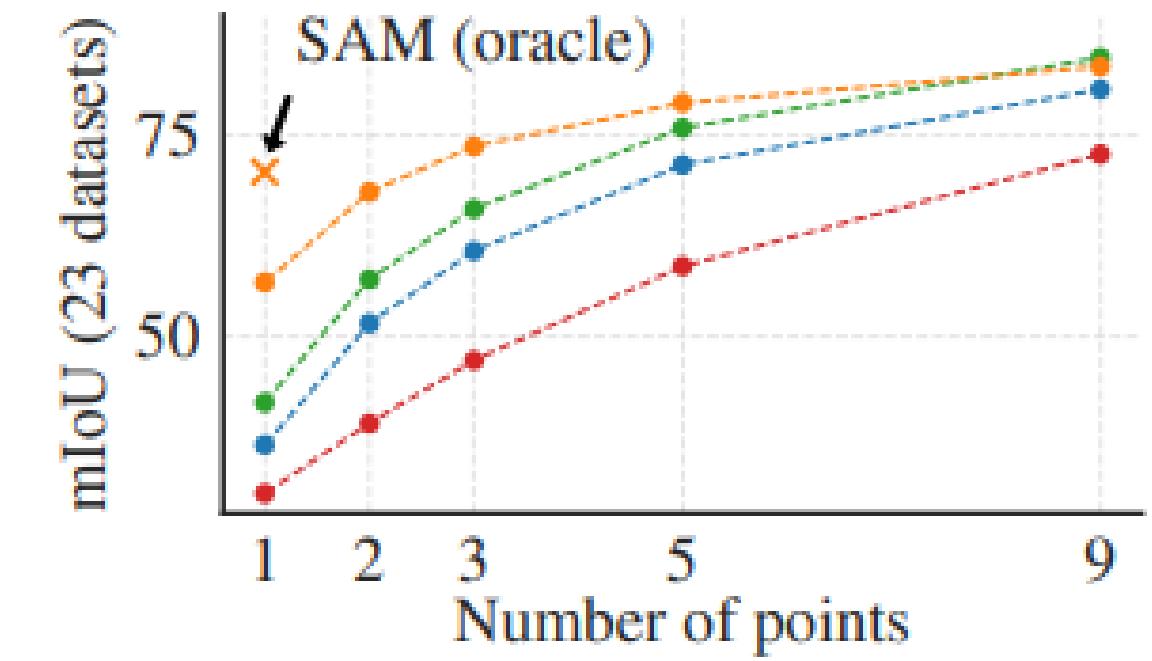
4. Kết quả

a. Đánh giá Zero-Shot từ Một điểm đơn lẻ

Khi số lượng điểm Prompt tăng từ 1 lên 9, khoảng cách hiệu suất mIoU giữa SAM và các phương pháp tương tác khác như RITM, SimpleClick, FocalClick giảm dần.



(d) Center points (default)



(e) Random points

4. Kết quả

b. Hiệu suất trong tác vụ hạ nguồn

Kết quả cho thấy SAM có thể được sử dụng out-of-the-box với prompt engineering để giải quyết nhiều tác vụ hạ nguồn :

- Phát hiện cạnh (Edge Detection).
 - Tạo đề xuất đối tượng (Object Proposal Generation).
 - Phân đoạn thể hiện (Instance Segmentation) : bằng cách kết hợp SAM với một bộ phát hiện đối tượng.
- SAM là một mô hình có khả năng tổng quát hóa mạnh mẽ và có thể hoạt động như một thành phần đáng tin cậy trong hệ thống lớn hơn.

4. Kết quả

c. Quy mô và chất lượng dữ liệu (SA-1B)

Thành công của SAM có được nhờ việc tạo ra tập dữ liệu SA-1B, lớn hơn bất kỳ tập dữ liệu segmentation hiện có.

- Quy mô vô song : SA-1B chứa 1,1 tỷ mask từ 11 triệu hình ảnh.
- SA-1B có số lượng mask gấp 400 lần so với tập dữ liệu segmentation lớn nhất hiện có(Open Images).
- Chất lượng mask cao : Việc so sánh mask tự động với mask được chỉnh sửa chuyên nghiệp cho thấy 94% các cặp có IoU lớn hơn 90% . → Khẳng định mask tự động của SAM có chất lượng cao và hiệu quả cho việc huấn luyện mô hình.

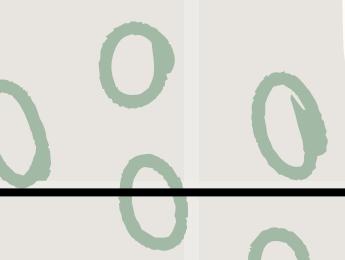
5. So sánh với các phương pháp khác

Ưu điểm	SAM	Phương pháp khác
Tính tổng quát Zero-Shot	Được thiết kế để chuyển giao zero-shot sang các miền và tác vụ mới. Khả năng này bắt nguồn từ việc huấn luyện trên 1 tập dữ liệu quy mô lớn và đa dạng	Các mô hình segmentation truyền thống thường yêu cầu huấn luyện lại (fine-tuning) để thích ứng với dữ liệu mới. SAM có thể sử dụng out-of-the-box.

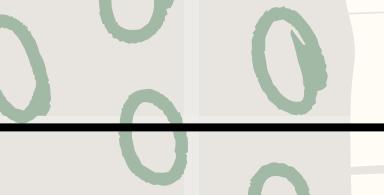
5. So sánh với các phương pháp khác



Ưu điểm	SAM	Phương pháp khác
Xử lý mơ hồ	SAM được thiết kế để dự đoán nhiều mask hợp lệ cho một prompt mơ hồ. Mô hình có thể trả về các mask cho nhiều cấp độ đối tượng (tổng thể, bộ phận, chi tiết).	Các mô hình tương tác khác (RITM) là mô hình đơn mask, có xu hướng tính trung bình các mask hợp lệ khi prompt mơ hồ, dẫn đến mask chất lượng kém hơn.

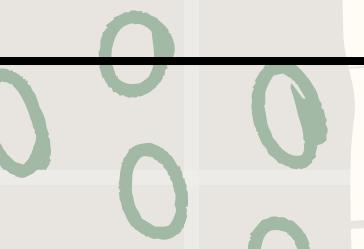


5. So sánh với các phương pháp khác

Ưu điểm	SAM	Phương pháp khác
Hiệu quả Tốc độ (Amortized Real-Time)	<p>Thiết kế chia mô hình thành Image Encoder nặng (chạy một lần) và Prompt Encoder/Mask Decoder nhẹ (chạy nhiều lần). Sau khi Image Embedding được tính, quá trình dự đoán mask chỉ mất ~50ms trên CPU, cho phép tương tác liền mạch, thời gian thực.</p> 	<p>Các mô hình tương tác cố định (như RITM hoặc FocalClick) thường chạy toàn bộ mô hình trong mỗi lần tương tác, làm cho chúng chậm hơn.</p> 

5. So sánh với các phương pháp khác

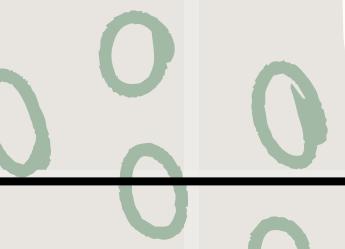
Ưu điểm	SAM	Phương pháp khác
Tính Khả dụng (Compositionality)	SAM hoạt động như một thành phần có giao diện đáng tin cậy. Nó có thể được kết hợp với các mô-đun khác (như bộ phát hiện, mô hình ngôn ngữ) để thực hiện các tác vụ phức tạp hơn	Các hệ thống đa tác vụ (multi-task systems) trước đây thực hiện một tập hợp các tác vụ cố định đã được huấn luyện.



5. So sánh với các phương pháp khác



Ưu điểm	SAM	Phương pháp khác
Quy mô Dữ liệu (SA-1B)	SAM được huấn luyện trên SA-1B, tập dữ liệu segmentation lớn nhất từ trước đến nay, với 1.1 tỷ mask và 11 triệu hình ảnh.	SA-1B có số lượng mask gấp 400 lần so với tập dữ liệu segmentation lớn nhất hiện có khác (Open Images). Quy mô này là chìa khóa cho khả năng tổng quát hóa của SAM.



III. Ưu điểm, hạn chế và ứng dụng

Segment Anything

1

Ưu điểm và tính mới

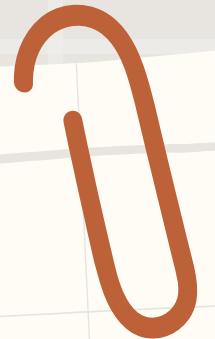
2

*Hạn chế
còn tồn tại*

3

*Ứng dụng và góc nhìn
riêng*

1. Ưu điểm và tính mới



SAM không chỉ là một mô hình tốt hơn mà còn là một sự thay đổi mô hình trong lĩnh vực phân đoạn, lấy cảm hứng từ NLP.

- Mô hình Nền tảng cho Phân đoạn: SAM là nỗ lực đầu tiên nhằm xây dựng một "mô hình nền tảng" cho phân đoạn hình ảnh, nhằm tổng quát hóa sang các tác vụ và phân phối dữ liệu mới.
- Thiết kế Tác vụ Mới : Khác với các tác vụ cố định (Semantic/Instance segmentation), SAM giới thiệu tác vụ "phân đoạn có thể nhắc nhở được" (Promptable Segmentation). Tác vụ này hoạt động như một mục tiêu huấn luyện mạnh mẽ và là cơ chế chuyển giao zero-shot.

1. Ưu điểm và tính mới



SAM không chỉ là một mô hình tốt hơn mà còn là một sự thay đổi mô hình trong lĩnh vực phân đoạn, lấy cảm hứng từ NLP.

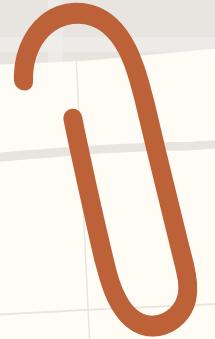
- Cấu trúc Khả dụng (Compositionality): SAM được thiết kế để trở thành một thành phần đáng tin cậy trong một hệ thống lớn hơn, có thể kết hợp với các mô-đun khác (detector, text encoder). Điều này cho phép mở rộng ứng dụng ngoài những gì được hình dung ban đầu.
- Bộ dữ liệu SA-1B Không lồ: Việc xây dựng SA-1B với 1.1 tỷ mask trên 11 triệu ảnh là một thành tựu đáng kể, tạo ra nguồn dữ liệu chưa từng có cho nghiên cứu về mô hình nền tảng trong thị giác máy tính.

2. Hạn chế còn tồn tại

a. Giới hạn kỹ thuật và hiệu suất

- Độ trễ Xử lý Tổng thể (Latency): Mặc dù Prompt Encoder và Mask Decoder cực nhanh, hiệu suất tổng thể của SAM không đạt tốc độ thời gian thực (real-time) do Image Encoder nặng. Điều này là gánh nặng khi xử lý hàng loạt hoặc cần phản hồi tức thì trên phần cứng giới hạn.

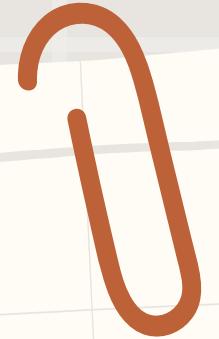
2. Hạn chế còn tồn tại



a. Giới hạn kỹ thuật và hiệu suất

- Độ chính xác Ranh giới (Boundary Sharpness): SAM có thể bỏ sót các cấu trúc tinh vi (fine structures) và không tạo ra các đường ranh giới sắc nét như các phương pháp chuyên biệt có tính toán chuyên sâu hơn (ví dụ: các phương pháp sử dụng kỹ thuật "zoom-in" như FocalClick).
- Hallucination: Mô hình đôi khi dự đoán sai (hallucinates) các thành phần nhỏ bị ngắt kết nối (small disconnected components).

2. Hạn chế còn tồn tại



b. Giới hạn về phạm vi tác vụ

- Không tối ưu cho IoU rất cao: SAM được thiết kế cho tính tổng quát, và các mô hình tương tác chuyên dụng được kỳ vọng sẽ vượt trội hơn SAM khi người dùng cung cấp nhiều điểm prompt.
- Tác vụ Phân đoạn Cấp cao: Vẫn chưa rõ cách thiết kế các prompt đơn giản để thực hiện các tác vụ phân đoạn cấp cao như Semantic Segmentation và Panoptic Segmentation. SAM cần được kết hợp với các mô hình phân loại hoặc các công cụ khác cho các tác vụ này.

3. Ứng dụng và góc nhìn riêng

a. Ứng dụng

Y tế - xử lý ảnh chuẩn đoán :

- SAM có thể được tích hợp vào các hệ thống chẩn đoán hình ảnh để phân đoạn tổn thương/u bướu (sau khi được phát hiện bằng Detector) . Lợi ích chính là tăng tốc quá trình tạo mask cho các nghiên cứu lâm sàng hoặc các bệnh lý hiếm gặp, nơi dữ liệu huấn luyện (fine-tuning) chuyên biệt bị giới hạn.

3. Ứng dụng và góc nhìn riêng

a. Ứng dụng

Nông nghiệp Thông minh - Phân tích Ảnh UAV/Flycam :

- SAM có thể được sử dụng để phân đoạn tức thì các khu vực bị nhiễm bệnh, cỏ dại, hoặc cây trồng bị stress trong ảnh chụp từ máy bay không người lái. Điều này giúp đẩy nhanh quá trình tạo dữ liệu Ground Truth và hỗ trợ nông nghiệp chính xác (precision agriculture).

3. Ứng dụng và góc nhìn riêng

a. Ứng dụng

Thương mại Điện tử và Marketing :

- Ứng dụng phổ biến nhất là tự động xóa phông và phân đoạn sản phẩm chính xác, giúp tiết kiệm thời gian và chi phí xử lý hình ảnh hàng loạt cho các doanh nghiệp.

3. Ứng dụng và góc nhìn riêng

b. Góc nhìn riêng

Thách thức về Tài nguyên Điện toán :

- Dù tương tác nhanh, gánh nặng của Image Encoder nặng vẫn là một rào cản về hiệu suất và chi phí đối với các ứng dụng yêu cầu xử lý trên phần cứng biên (edge devices) không mạnh hoặc các giải pháp cần độ trễ cực thấp.

3. Ứng dụng và góc nhìn riêng

b. Góc nhìn riêng

Vai trò Chiến lược (Composable Tool) :

- SAM không phải là một giải pháp "All-in-one" thay thế hoàn toàn các mô hình phân đoạn chuyên dụng. Vai trò mạnh mẽ nhất của nó là một thành phần cấu thành (composable component). Các nhà phát triển tại Việt Nam có thể tận dụng SAM như một giao diện segmentation đáng tin cậy bằng cách kết hợp nó với các mô hình khác (ví dụ: mô hình phát hiện đối tượng bản địa) để xây dựng các ứng dụng phức tạp mà không cần phải lo lắng về việc gán nhãn hàng triệu mask mới.

IV. Demo

Segment Anything