SAM 3: Segment Anything with Concepts

Nicolas Carion[*], **Laura Gustafson**[*], **Yuan-Ting Hu**[*], **Shoubhik Debnath**[*], **Ronghang Hu**[*], **Didac Suris**[*], **Chaitanya Ryali**[*], **Kalyan Vasudev Alwala**[*], **Haitham Khedr**[*], **Andrew Huang**, **Jie Lei**, **Tengyu Ma**, **Baishan Guo**, **Arpit Kalla**, **Markus Marks**, **Joseph Greer**, **Meng Wang**, **Peize Sun**, **Roman Rädle**, **Triantafyllos Afouras**, **Effrosyni Mavroudi**, **Katherine Xu**[°], **Tsung-Han Wu**[°], **Yu Zhou**[°], **Liliane Momeni**[°], **Rishi Hazra**[°], **Shuangrui Ding**[°], **Sagar Vaze**[°], **Francois Porcher**[°], **Feng Li**[°], **Siyuan Li**[°], **Aishwarya Kamath**[°], **Ho Kei Cheng**[°], **Piotr Dollár**[†], **Nikhila Ravi**[†], **Kate Saenko**[†], **Pengchuan Zhang**[†], **Christoph Feichtenhofer**[†]

Meta Superintelligence Labs
[*]core contributor, [°]intern, [†]project lead, order is random within groups

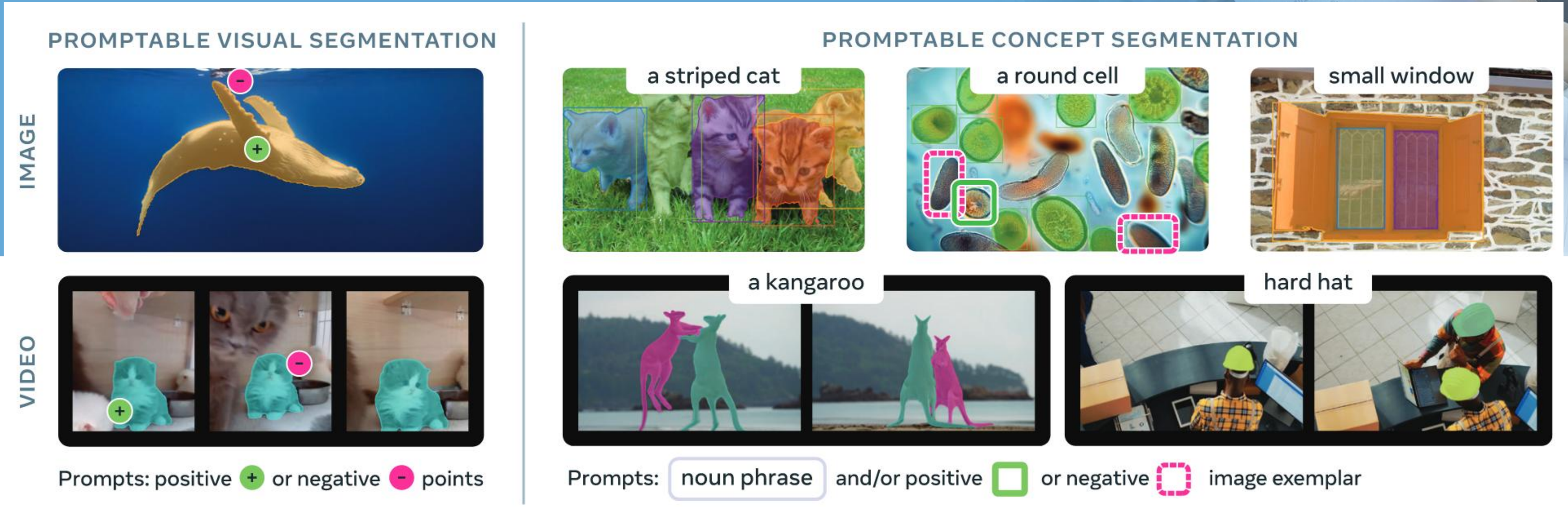
汇报人：蒋明忠

时间：2026.03

Background

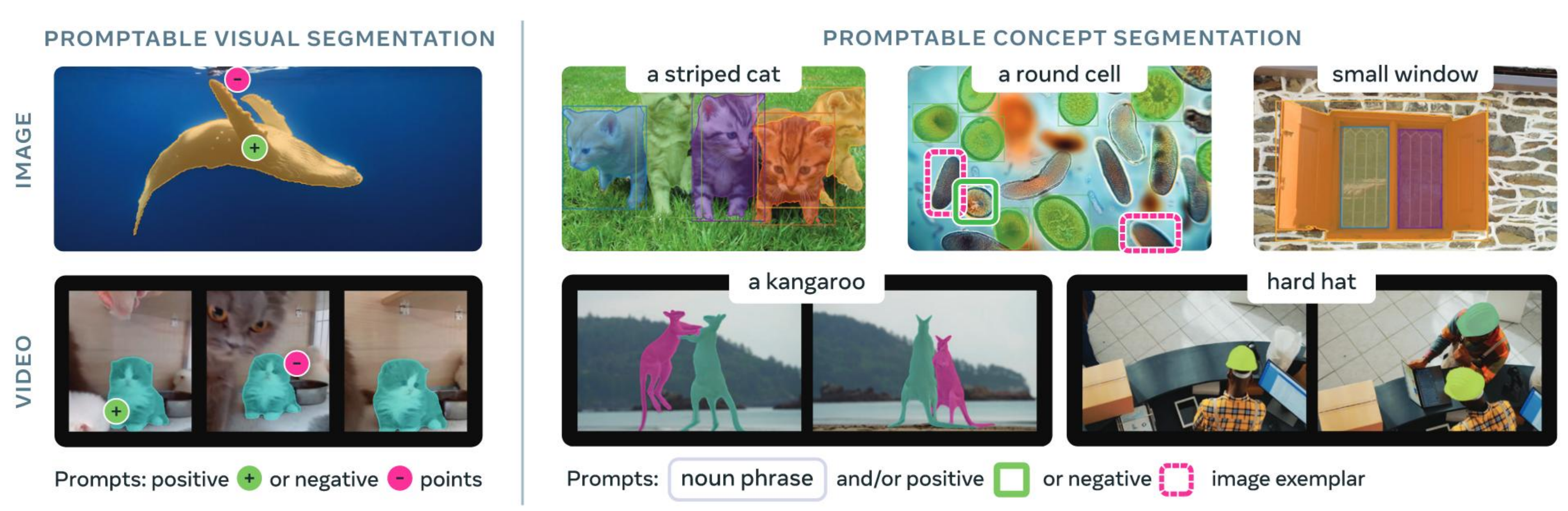
在视觉场景中找到并分割任何事物的能力是多模态人工智能的基础，为机器人、内容创作、增强现实、数据 标注和更广泛的科学领域的应用提供动力。SAM系列提出了图像和视频的可提示分割任务，专注于可提示视觉分割(PVS)，通过点、框或掩码对每个提示进行单个对象分割。虽然这些方法取得了突破，但它们并没有解决在输入中任何地方发现和分割所有概念实例的通用任务。

(例如，视频中的所有“猫”)


PROMPTABLE VISUAL SEGMENTATION
IMAGE
VIDEO
Prompts: positive + or negative − points
PROMPTABLE CONCEPT SEGMENTATION
a striped cat
a round cell
small window
a kangaroo
hard hat
Prompts: noun phrase and/or positive 🟩 or negative ▨ image exemplar

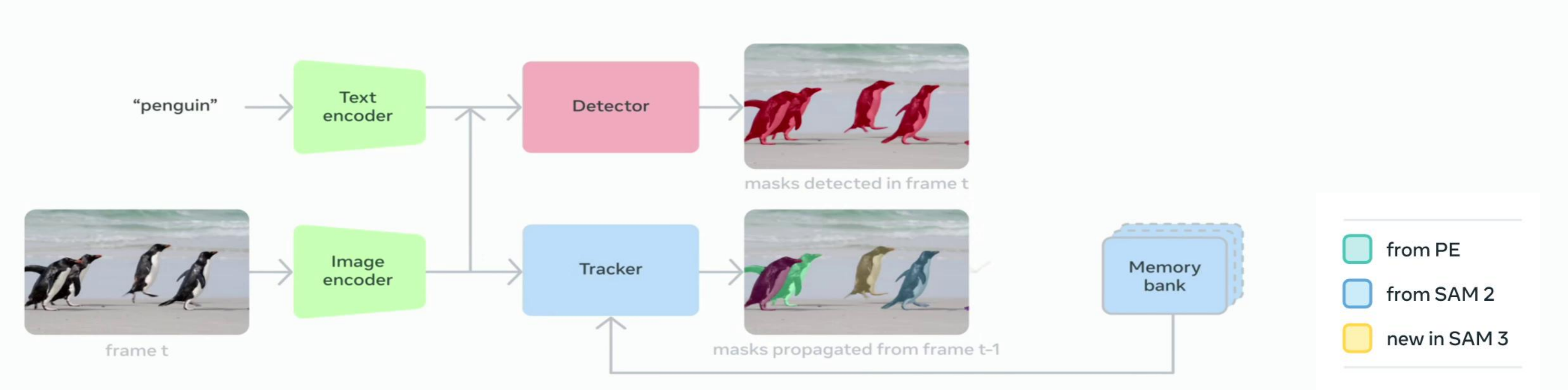
Method

为填补这一空白，我们提出了SAM 3模型，该模型在图像和视频的可提示分割方面实现了质的飞跃，相较于 SAM 2提升了PVS，并为可提示概念分割(PCS)设定了新标准。我们将PCS任务形式化为：以文本和/或图像样本作为输入，预测与概念匹配的每个对象的实例和语义掩码，同时保持视频帧间对象身份的一致性。


PROMPTABLE VISUAL SEGMENTATION
IMAGE
VIDEO
Prompts: positive + or negative − points
PROMPTABLE CONCEPT SEGMENTATION
a striped cat
a round cell
small window
a kangaroo
hard hat
Prompts: noun phrase and/or positive or negative image exemplar

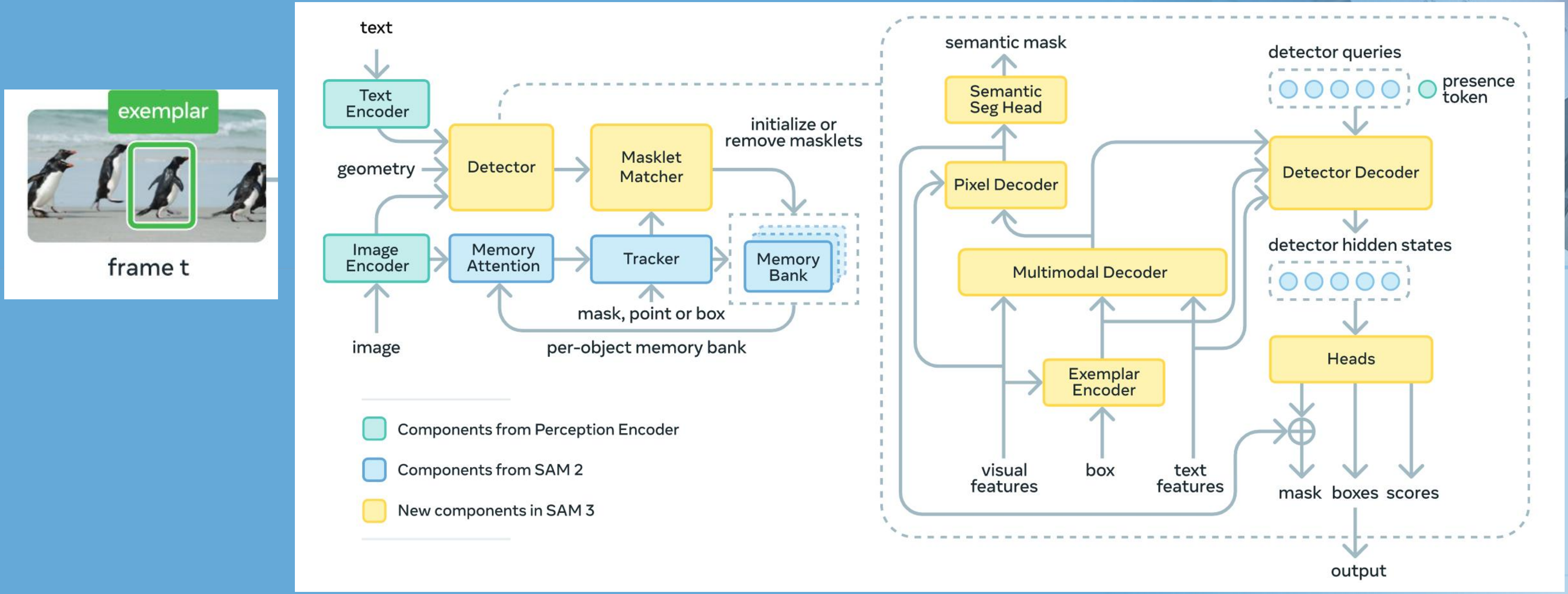
Method

我们的模型由一个检测器和一个跟踪器组成，它们共享一个视觉编码器（PE - Perception Encoder (Meta, 2025)）。该检测器基于DETR架构，基于文本、几何和图像样本输入的模型。为解决开放词汇概念检测的挑战，我们引入了一个独立的存在头来解耦识别和定位，这在使用具有挑战性的否定短语进行训练时特别有效。该追踪器继承了SAM 2 Transformer encoder-decoder编架构，支持视频分割与交互式精修。检测与跟踪的解耦设计避免了任务冲突，因为检测器需要具备身份无关性，而跟踪器的主要目标是分离视频中的身份。



Method

我们的模型由一个检测器和一个跟踪器组成，它们共享一个视觉编码器（PE - Perception Encoder (Meta, 2025)）。该检测器基于DETR架构，基于文本、几何和图像样本输入的模型。为解决开放词汇概念检测的挑战，我们引入了一个独立的存在头来解耦识别和定位，这在使用具有挑战性的否定短语进行训练时特别有效。该追踪器继承了SAM 2 Transformer encoder-decoder编架构，支持视频分割与交互式精修。检测与跟踪的解耦设计避免了任务冲突，因为检测器需要具备身份无关性，而跟踪器的主要目标是分离视频中的身份。



🟩 from PE

🟦 from SAM 2

🟨 new in SAM 3

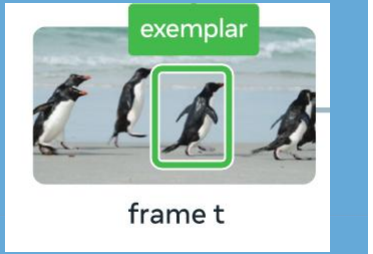
Method

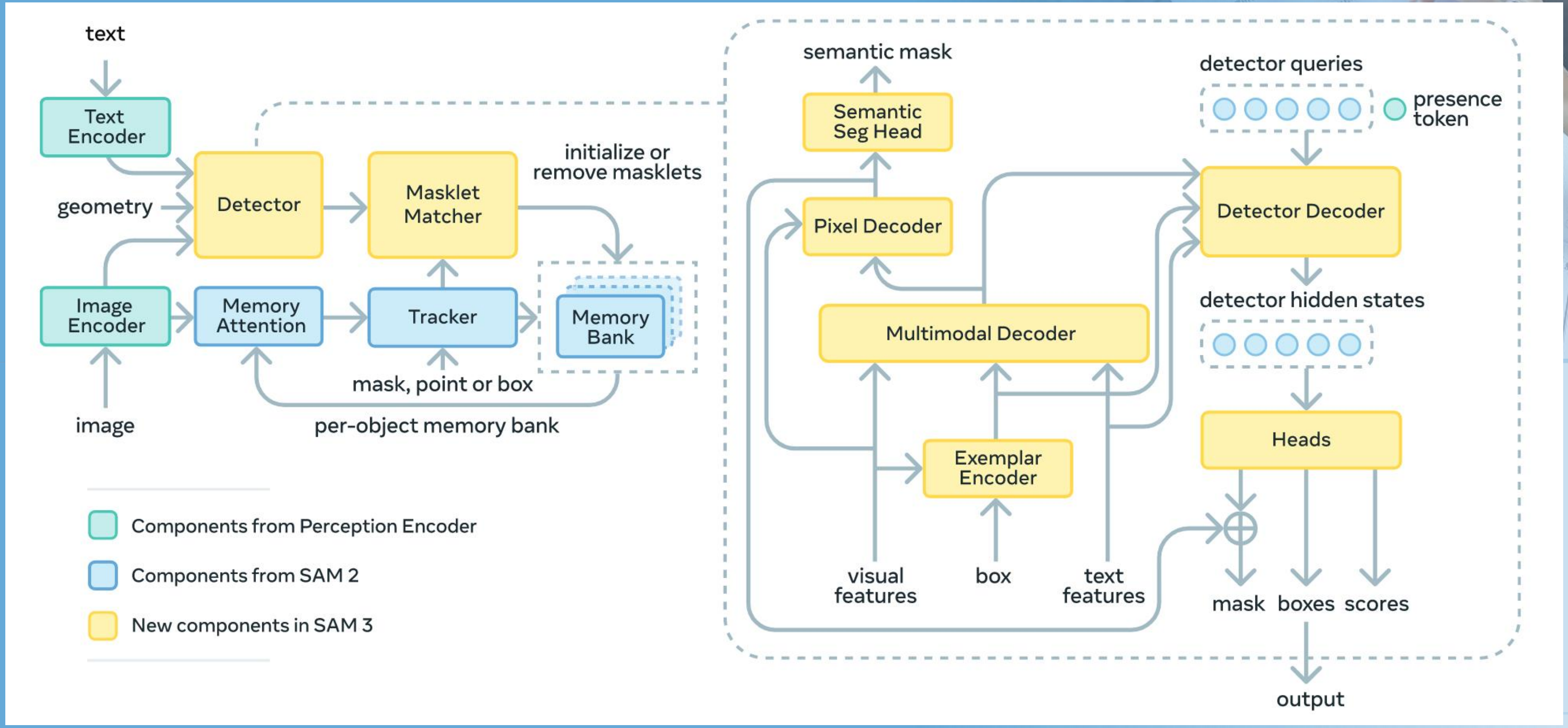
$$\hat{\mathcal{M}}_t = \text{propagate}\left(\mathcal{M}_{t-1}\right), \qquad \mathcal{O}_t = \text{detect}\left(I_t, P\right), \qquad \mathcal{M}_t = \text{match\_and\_update}\left(\hat{\mathcal{M}}_t, \mathcal{O}_t\right).$$

$$\Delta_i(\tau) = \begin{cases} +1, & \text{if } \exists\, d \in \mathcal{D}_\tau \text{ s.t. } \text{IoU}(d, \hat{\mathcal{M}}^i_\tau) > \text{iou\_threshold} \\ -1, & \text{otherwise,} \end{cases}$$

消歧策略： （1）删除未确认的 masklet

（2） 删除重复的 masklet

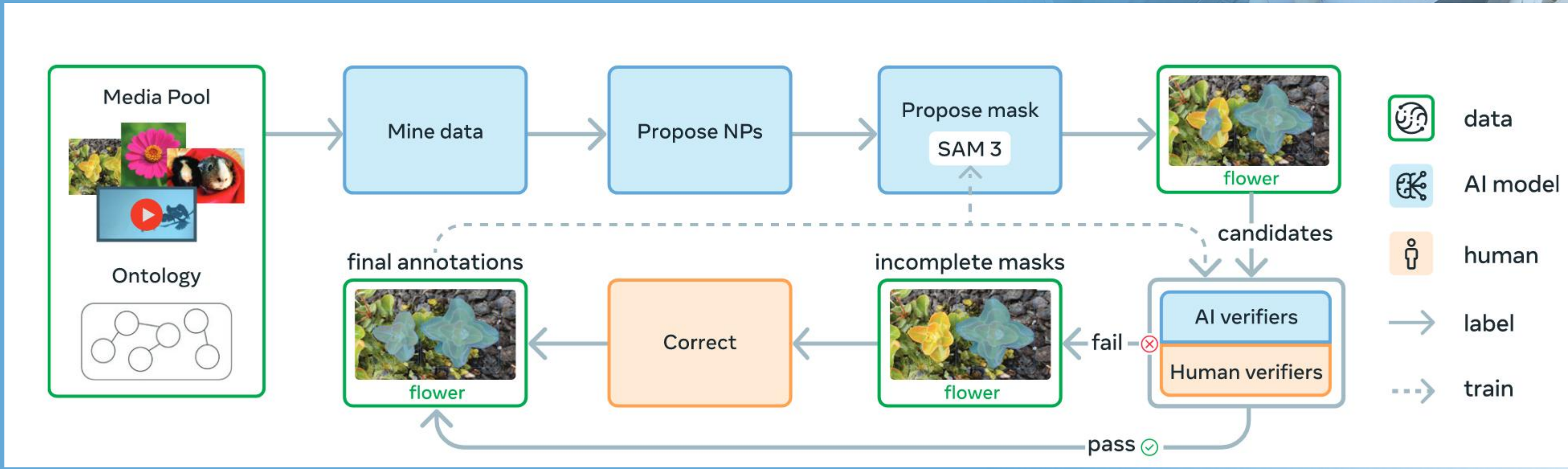




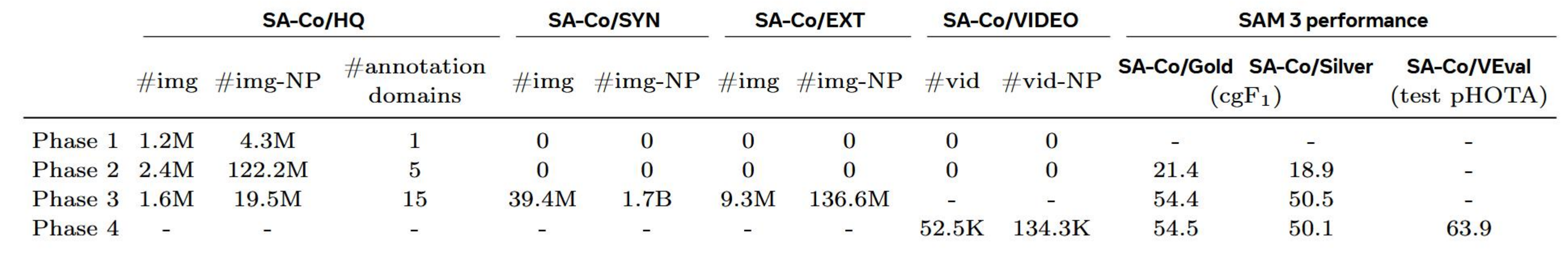
Method

通过SAM 3实现PCS的阶跃式提升需要对大量多样化概念和视觉域进行训练，这超出了现有数据集的范畴。我们构建了一个高效的数据引擎，通过与SAM 3、人工标注者和AI标注者的反馈循环迭代生成标注数据，通过主动挖掘当前SAM 3版本无法生成高质量训练数据的媒体短语对，以进一步优化模型性能。 通过将特定任务委托给AI标注员(其模型准确度与人类相当或更优)，相较于纯人工标注流程，我们的处理效率可提升逾一倍。

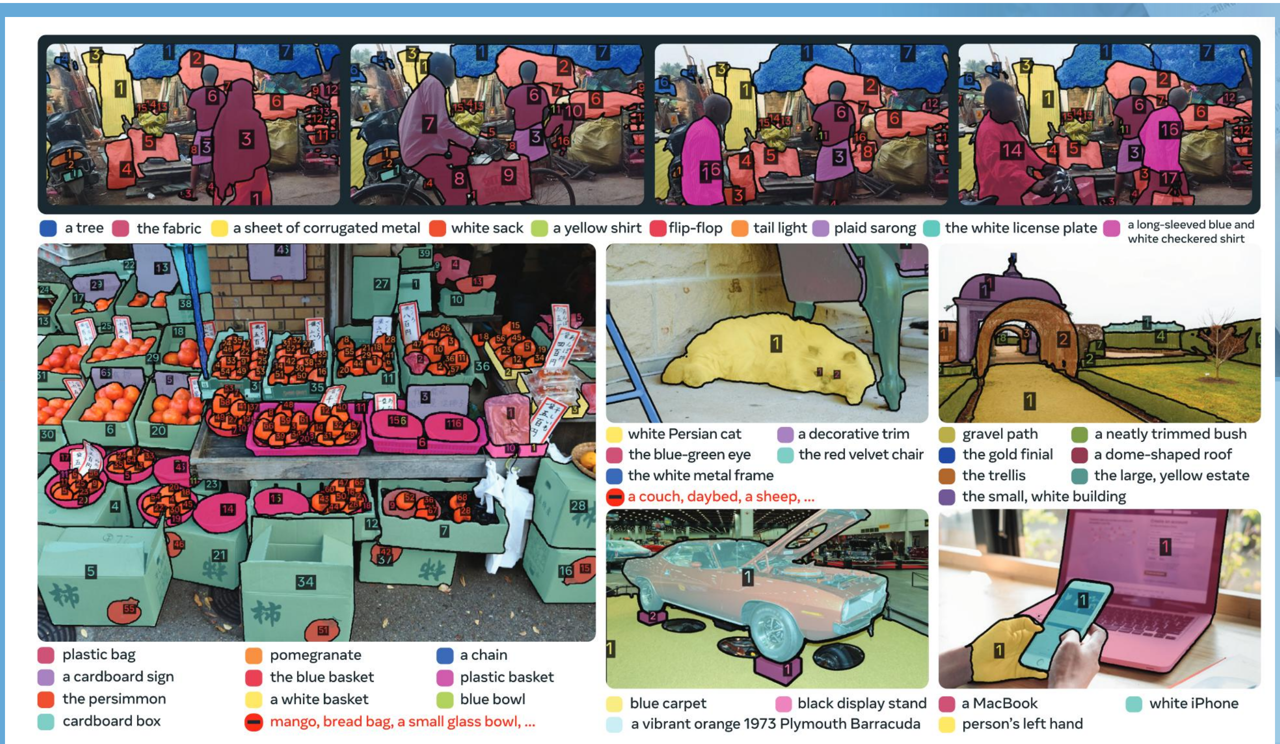
第一阶段：纯人工验证（冷启动） 4.3M image-NP 第二阶段：人机混合验证（效率飞跃）122M image-NP 第三阶段：规模化与领域扩展（长尾覆盖）19.5M image-NP 第四阶段：视频标注（时空连贯）52.5K videos and 467K masklets.



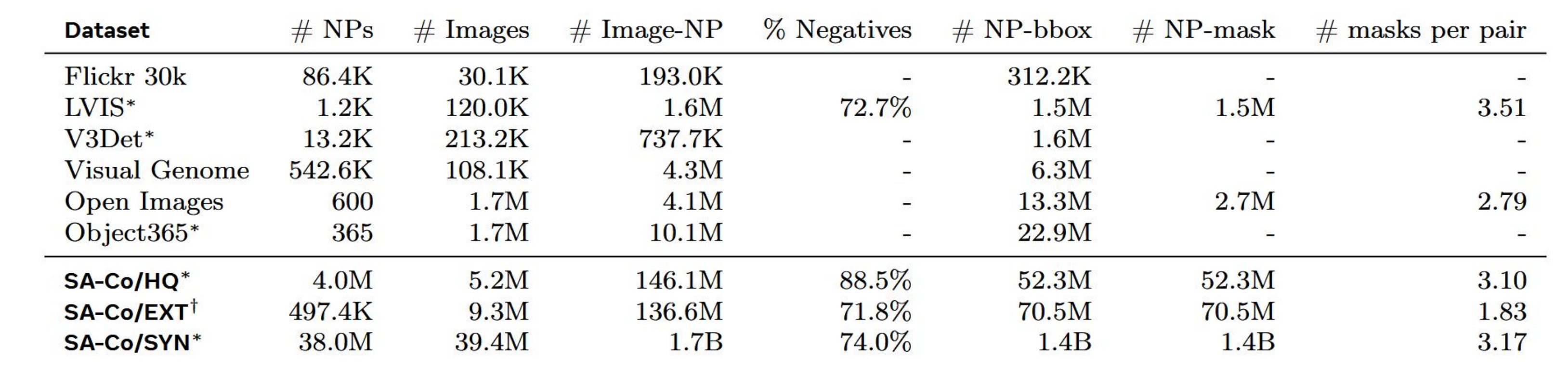
Method



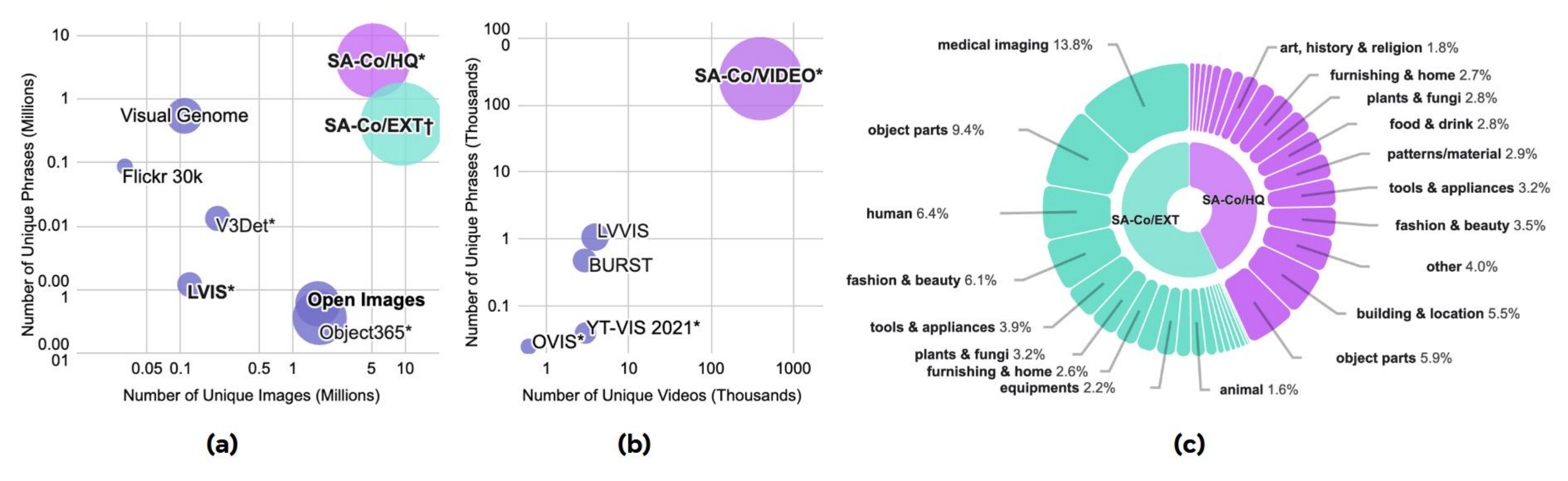
	SA-Co/HQ			SA-Co/SYN		SA-Co/EXT		SA-Co/VIDEO		SAM 3 performance		
	#img	#img-NP	#annotation domains	#img	#img-NP	#img	#img-NP	#vid	#vid-NP	SA-Co/Gold	SA-Co/Silver	SA-Co/VEval
										(cgF$_1$)		(test pHOTA)
Phase 1	1.2M	4.3M	1	0	0	0	0	0	0	-	-	-
Phase 2	2.4M	122.2M	5	0	0	0	0	0	0	21.4	18.9	-
Phase 3	1.6M	19.5M	15	39.4M	1.7B	9.3M	136.6M	-	-	54.4	50.5	-
Phase 4	-	-	-	-	-	-	-	52.5K	134.3K	54.5	50.1	63.9



Method



Dataset	# NPs	# Images	# Image-NP	% Negatives	# NP-bbox	# NP-mask	# masks per pair
Flickr 30k	86.4K	30.1K	193.0K	-	312.2K	-	-
LVIS*	1.2K	120.0K	1.6M	72.7%	1.5M	1.5M	3.51
V3Det*	13.2K	213.2K	737.7K	-	1.6M	-	-
Visual Genome	542.6K	108.1K	4.3M	-	6.3M	-	-
Open Images	600	1.7M	4.1M	-	13.3M	2.7M	2.79
Object365*	365	1.7M	10.1M	-	22.9M	-	-
SA-Co/HQ*	4.0M	5.2M	146.1M	88.5%	52.3M	52.3M	3.10
SA-Co/EXT†	497.4K	9.3M	136.6M	71.8%	70.5M	70.5M	1.83
SA-Co/SYN*	38.0M	39.4M	1.7B	74.0%	1.4B	1.4B	3.17

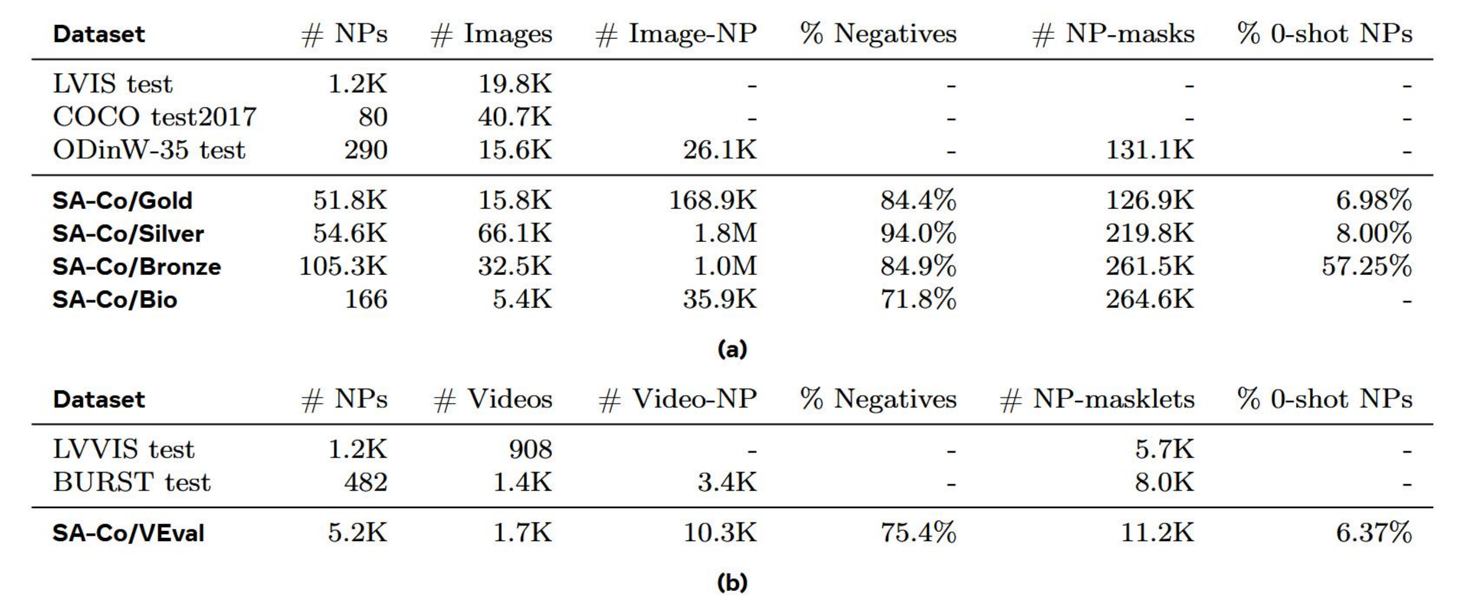


(a)

(b)

(c)

Method



Dataset	# NPs	# Images	# Image-NP	% Negatives	# NP-masks	% 0-shot NPs
LVIS test	1.2K	19.8K	-	-	-	-
COCO test2017	80	40.7K	-	-	-	-
ODinW-35 test	290	15.6K	26.1K	-	131.1K	-
SA-Co/Gold	51.8K	15.8K	168.9K	84.4%	126.9K	6.98%
SA-Co/Silver	54.6K	66.1K	1.8M	94.0%	219.8K	8.00%
SA-Co/Bronze	105.3K	32.5K	1.0M	84.9%	261.5K	57.25%
SA-Co/Bio	166	5.4K	35.9K	71.8%	264.6K	-

(a)

Dataset	# NPs	# Videos	# Video-NP	% Negatives	# NP-masklets	% 0-shot NPs
LVVIS test	1.2K	908	-	-	5.7K	-
BURST test	482	1.4K	3.4K	-	8.0K	-
SA-Co/VEval	5.2K	1.7K	10.3K	75.4%	11.2K	6.37%

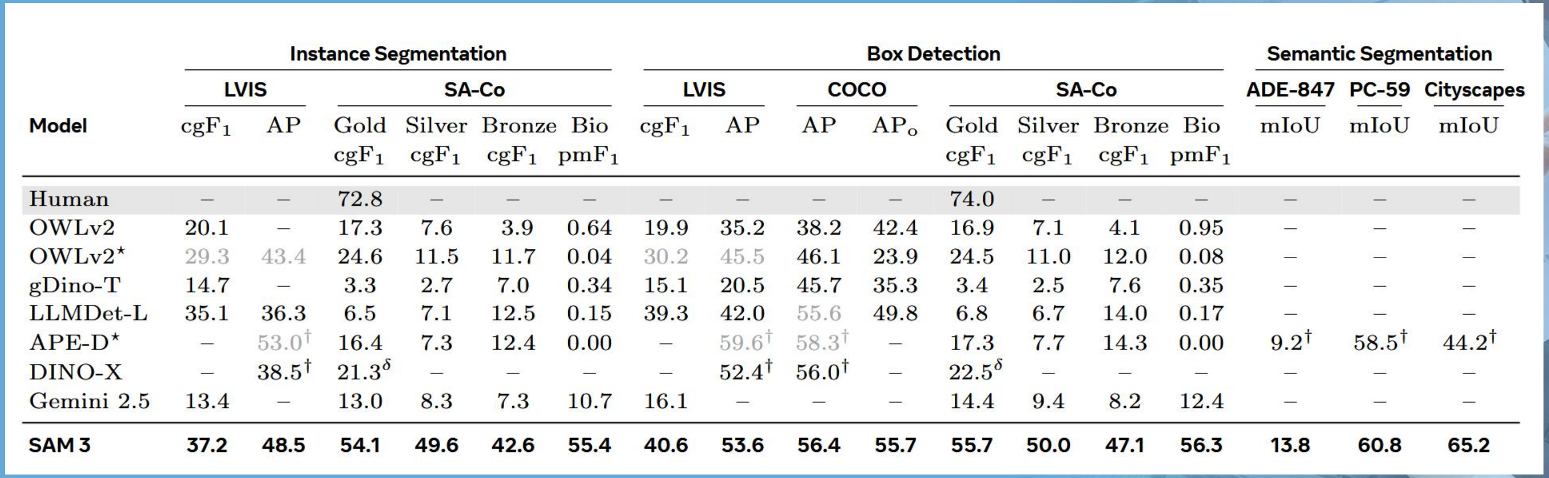
(b)

$$\mathrm{pmF}_1^\tau = \frac{2\mathrm{TP}^\tau_{total}}{2\mathrm{TP}^\tau_{total} + \mathrm{FP}^\tau_{total} + \mathrm{FN}^\tau_{total}}, \quad pmF_1 = \frac{1}{10} \sum_{\tau \in \mathrm{np.linspace}(0.5,0.95,10)} \mathrm{pmF}_1^\tau .$$

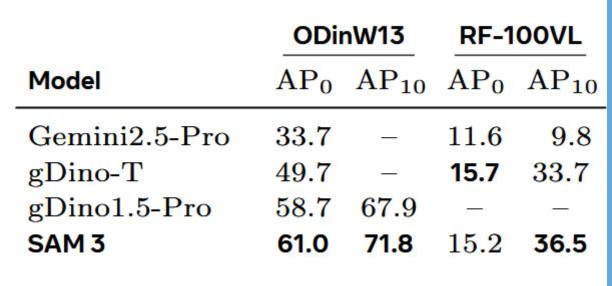
$$\mathrm{IL\_MCC} = \frac{\mathrm{IL\_TP} \cdot \mathrm{IL\_TN} - \mathrm{IL\_FP} \cdot \mathrm{IL\_FN}}{\sqrt{(\mathrm{IL\_TP} + \mathrm{IL\_FP}) \cdot (\mathrm{IL\_TP} + \mathrm{IL\_FN}) \cdot (\mathrm{IL\_TN} + \mathrm{IL\_FP}) \cdot (\mathrm{IL\_TN} + \mathrm{IL\_FN})}}.$$

$$\mathrm{cgF}_1 = 100 \cdot \mathrm{pmF}_1 \cdot \mathrm{IL\_MCC}.$$

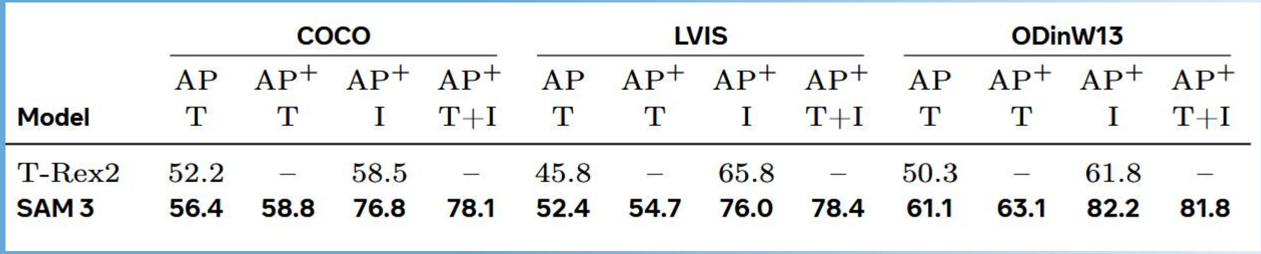
Experiments



	Instance Segmentation						Box Detection								Semantic Segmentation		
	LVIS		SA-Co				LVIS		COCO		SA-Co				ADE-847	PC-59	Cityscapes
Model	cgF$_1$	AP	Gold	Silver	Bronze	Bio	cgF$_1$	AP	AP	AP$_o$	Gold	Silver	Bronze	Bio	mIoU	mIoU	mIoU
			cgF$_1$	cgF$_1$	cgF$_1$	pmF$_1$					cgF$_1$	cgF$_1$	cgF$_1$	pmF$_1$			
Human	–	–	72.8	–	–	–	–	–	–	–	74.0	–	–	–	–	–	–
OWLv2	20.1	–	17.3	7.6	3.9	0.64	19.9	35.2	38.2	42.4	16.9	7.1	4.1	0.95	–	–	–
OWLv2*	29.3	43.4	24.6	11.5	11.7	0.04	30.2	45.5	46.1	23.9	24.5	11.0	12.0	0.08	–	–	–
gDino-T	14.7	–	3.3	2.7	7.0	0.34	15.1	20.5	45.7	35.3	3.4	2.5	7.6	0.35	–	–	–
LLMDet-L	35.1	36.3	6.5	7.1	12.5	0.15	39.3	42.0	55.6	49.8	6.8	6.7	14.0	0.17	–	–	–
APE-D*	–	53.0†	16.4	7.3	12.4	0.00	–	59.6†	58.3†	–	17.3	7.7	14.3	0.00	9.2†	58.5†	44.2†
DINO-X	–	38.5†	21.3$^\delta$	–	–	–	–	52.4†	56.0†	–	22.5$^\delta$	–	–	–	–	–	–
Gemini 2.5	13.4	–	13.0	8.3	7.3	10.7	16.1	–	–	–	14.4	9.4	8.2	12.4	–	–	–
SAM 3	**37.2**	**48.5**	**54.1**	**49.6**	**42.6**	**55.4**	**40.6**	**53.6**	**56.4**	**55.7**	**55.7**	**50.0**	**47.1**	**56.3**	**13.8**	**60.8**	**65.2**

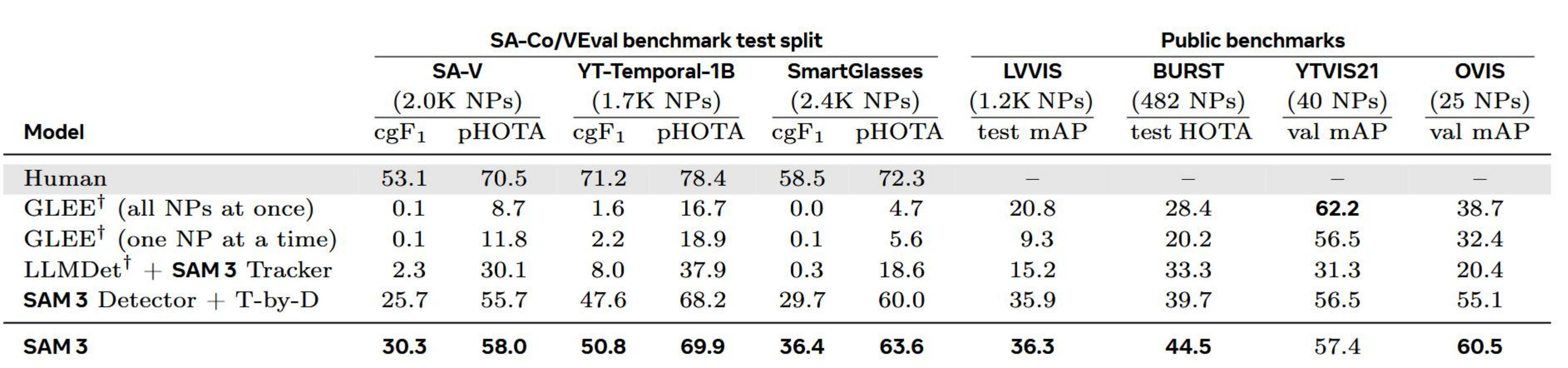


	ODinW13		RF-100VL	
Model	AP$_0$	AP$_{10}$	AP$_0$	AP$_{10}$
Gemini2.5-Pro	33.7	–	11.6	9.8
gDino-T	49.7	–	**15.7**	33.7
gDino1.5-Pro	58.7	67.9	–	–
SAM 3	**61.0**	**71.8**	15.2	**36.5**

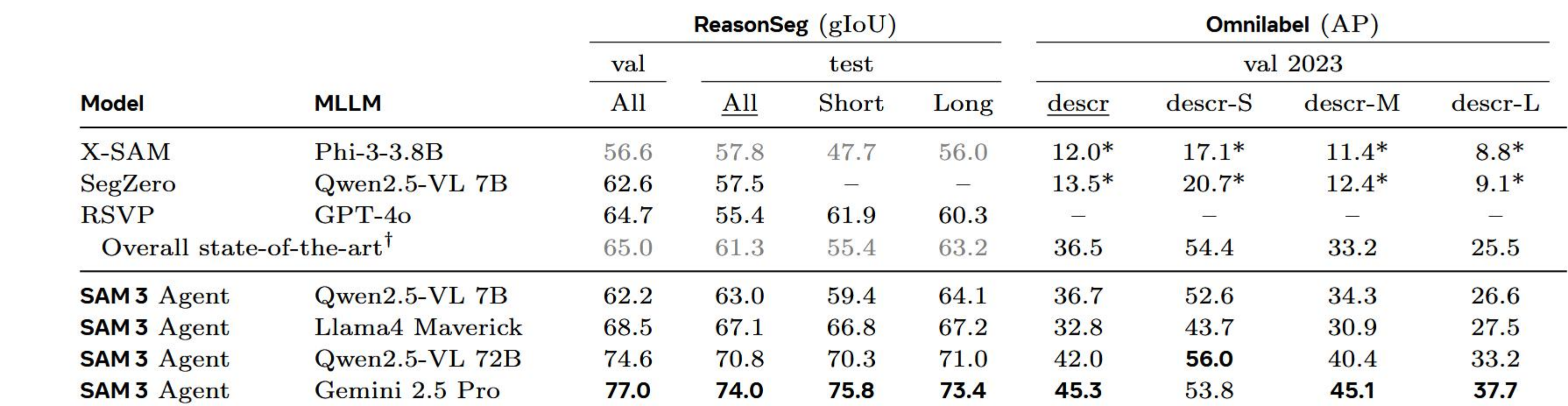


	COCO				LVIS				ODinW13			
	AP	AP$^+$	AP$^+$	AP$^+$	AP	AP$^+$	AP$^+$	AP$^+$	AP	AP$^+$	AP$^+$	AP$^+$
Model	T	T	I	T+I	T	T	I	T+I	T	T	I	T+I
T-Rex2	52.2	–	58.5	–	45.8	–	65.8	–	50.3	–	61.8	–
SAM 3	**56.4**	**58.8**	**76.8**	**78.1**	**52.4**	**54.7**	**76.0**	**78.4**	**61.1**	**63.1**	**82.2**	**81.8**

Experiments

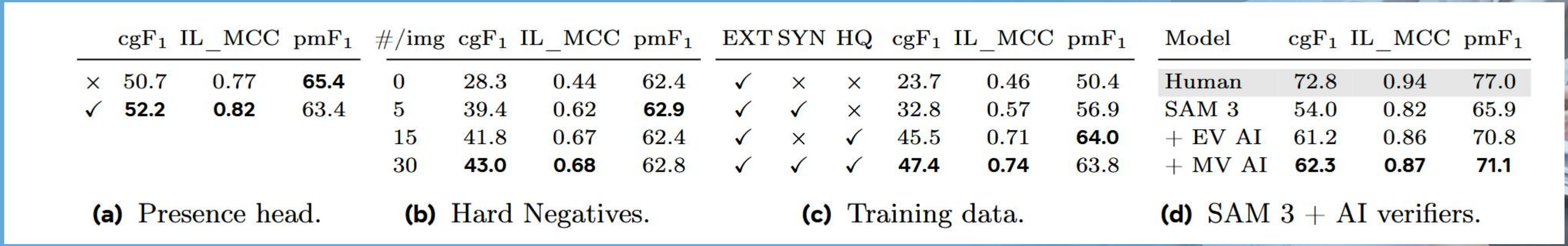


Model	SA-Co/VEval benchmark test split						Public benchmarks			
	SA-V (2.0K NPs)		YT-Temporal-1B (1.7K NPs)		SmartGlasses (2.4K NPs)		LVVIS (1.2K NPs)	BURST (482 NPs)	YTVIS21 (40 NPs)	OVIS (25 NPs)
	cgF$_1$	pHOTA	cgF$_1$	pHOTA	cgF$_1$	pHOTA	test mAP	test HOTA	val mAP	val mAP
Human	53.1	70.5	71.2	78.4	58.5	72.3	–	–	–	–
GLEE† (all NPs at once)	0.1	8.7	1.6	16.7	0.0	4.7	20.8	28.4	**62.2**	38.7
GLEE† (one NP at a time)	0.1	11.8	2.2	18.9	0.1	5.6	9.3	20.2	56.5	32.4
LLMDet† + **SAM 3** Tracker	2.3	30.1	8.0	37.9	0.3	18.6	15.2	33.3	31.3	20.4
SAM 3 Detector + T-by-D	25.7	55.7	47.6	68.2	29.7	60.0	35.9	39.7	56.5	55.1
SAM 3	**30.3**	**58.0**	**50.8**	**69.9**	**36.4**	**63.6**	**36.3**	**44.5**	57.4	**60.5**



Model	MLLM	ReasonSeg (gIoU)				Omnilabel (AP)			
		val	test			val 2023			
		All	All	Short	Long	descr	descr-S	descr-M	descr-L
X-SAM	Phi-3-3.8B	56.6	57.8	47.7	56.0	12.0*	17.1*	11.4*	8.8*
SegZero	Qwen2.5-VL 7B	62.6	57.5	–	–	13.5*	20.7*	12.4*	9.1*
RSVP	GPT-4o	64.7	55.4	61.9	60.3	–	–	–	–
Overall state-of-the-art†		65.0	61.3	55.4	63.2	36.5	54.4	33.2	25.5
SAM 3 Agent	Qwen2.5-VL 7B	62.2	63.0	59.4	64.1	36.7	52.6	34.3	26.6
SAM 3 Agent	Llama4 Maverick	68.5	67.1	66.8	67.2	32.8	43.7	30.9	27.5
SAM 3 Agent	Qwen2.5-VL 72B	74.6	70.8	70.3	71.0	42.0	**56.0**	40.4	33.2
SAM 3 Agent	Gemini 2.5 Pro	**77.0**	**74.0**	**75.8**	**73.4**	**45.3**	53.8	**45.1**	**37.7**

Experiments



	cgF$_1$	IL_MCC	pmF$_1$
×	50.7	0.77	**65.4**
✓	**52.2**	**0.82**	63.4

(a) Presence head.

#/img	cgF$_1$	IL_MCC	pmF$_1$
0	28.3	0.44	62.4
5	39.4	0.62	**62.9**
15	41.8	0.67	62.4
30	**43.0**	**0.68**	62.8

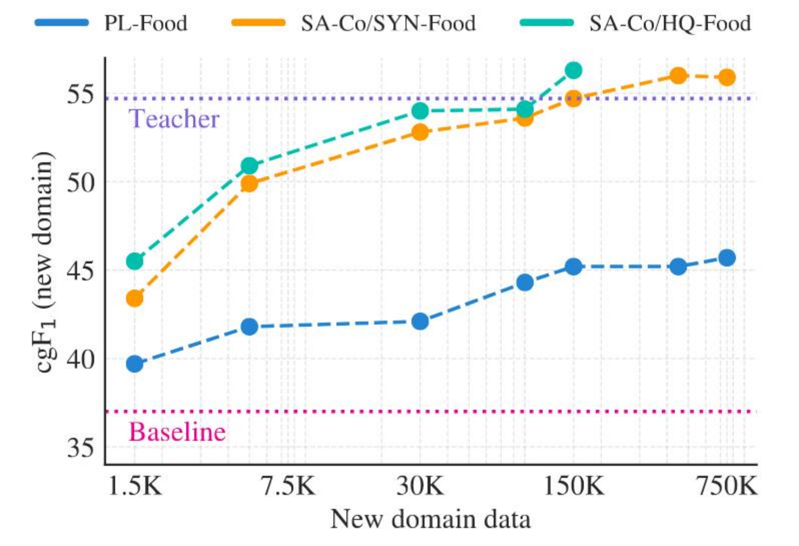
(b) Hard Negatives.

EXT	SYN	HQ	cgF$_1$	IL_MCC	pmF$_1$
✓	×	×	23.7	0.46	50.4
✓	✓	×	32.8	0.57	56.9
✓	×	✓	45.5	0.71	**64.0**
✓	✓	✓	**47.4**	**0.74**	63.8

(c) Training data.

Model	cgF$_1$	IL_MCC	pmF$_1$
Human	72.8	0.94	77.0
SAM 3	54.0	0.82	65.9
+ EV AI	61.2	0.86	70.8
+ MV AI	**62.3**	**0.87**	**71.1**

(d) SAM 3 + AI verifiers.



Thanks