Segment Anything with Concepts (SAM 3)

1. Introducción

La segmentación visual se ha considerado uno de los principales componentes en lo que respecta a visión computacional contemporáneo y su gran importancia radica en su evolución de modelos como HAAR hasta modelos avanzados de redes neuronales y arquitecturas mucho más avanzadas en la actualidad las necesidades de soluciones en computer vision siguen creciendo en contextos más complejos y dinámicos hoy por hoy tenemos a la robótica autónoma la realidad aumentada y los volúmenes gigantescos de información visual que se requiere en procesar.

En todos estos escenarios se recalca la importancia de identificar localizar y de limitar objetos de manera flexible y precisa por eso la indispensable de contar con algoritmos o estructuras de redes neuronales a la altura.

El proceso de delimitar objetos de manera flexible y precisa resulta indispensable para construir sistemas visuales robustos y adaptable en ese contexto la familia de modelos *Segment Anything Model* (SAM) Representan un punto de avance significativo y moderno al introducir el concepto **segmentación promptable** el cual puede orientar guiar activamente en el proceso de segmentación a través de diferentes interacciones simples como por ejemplo texto y visuales como puntos cajas y máscara.

Este nuevo enfoque desarrollado por Facebook da las pautas a lo que es los nuevos modelos inteligentes que hoy por hoy se integran en el ecosistema de soluciones avanzados en lo que es segmentación detección y rastreo en el campo de computer vision.

Sin embargo a pesar de los múltiples avances que hemos venido siguiendo desde las versiones SAM y SAM 2 presentaban grandes limitaciones estructurales importantes entre ellas su diseño estaba orientado a la segmentación de una única instancia por cada prompt, lo que limitaba su eficacia en escenarios donde resulta necesario la identificación de varios patrones. Esta limitación conceptual de diversas instancias asociadas a un mismo concepto dentro de una imagen o de un vídeo fueron resueltas por *"SAM 3: Segment Anything with Concepts" El cual propone una extensión conceptual y metodológica mediante la Concepción de* **Promptable Concept Segmentation (PCS)**, un enfoque que amplía la segmentación promptable hacia una gran colección de vocabulario o tokens Capaces de operar de manera consistente tanto en imágenes como en secuencias de vídeo

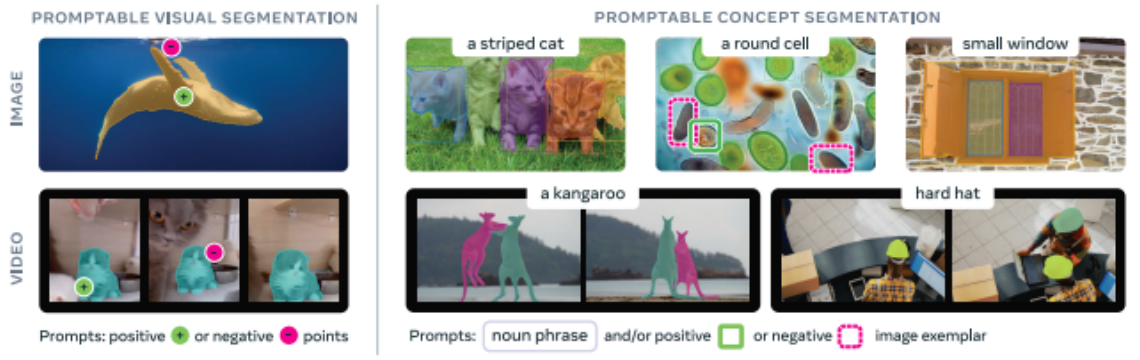

Figure 1 SAM 3 improves over SAM 2 on promptable visual segmentation with clicks (left) and introduces the new promptable concept segmentation capability (right). Users can segment all instances of a visual concept specified by a short noun phrase, image exemplars (positive or negative), or a combination of both.

La figura uno es una evidencia clara de lo que se concibe como la introducción de SAM 3 con respecto a sus predecesores en el gráfico de la izquierda se puede

observar la operación de SAM 2 la de segmentación visual promptable mediante clics, limitado a la segmentación de instancias individuales en comparación con la imagen de la derecha que introduce el concepto de **segmentación conceptual promptable mediante la cual SAM3** es capaz de identificar y segmentar todas las instancias relacionadas como el mismo concepto visual dicho de otra manera a través de frases nominales breves o conceptos como también se suma la selección visual de verdaderos positivos con falsos positivos o la combinación de ellos.

Este gran cambio representa mucho más que una mejora incremental en el rendimiento de modelo sino es en residencia redefine el modo en que los sistemas de computer vision pueden interactuar con el lenguaje natural y transferir una lógica centrada en instancias aisladas con un enfoque **conceptual, exhaustivo y semánticamente guiado** abriendo la posibilidad de operar directamente sobre conceptos visuales como “red Apple” o “striped cat” que introduce un nivel de abstracción que se aproxima la percepción artificial de las formas humanas de interpretación visual

Desde una perspectiva crítica en el ámbito de *computer vision*, SAM 3 no se limita a proponer una arquitectura mejorada, sino que establece un marco de trabajo más amplio que se materializa en tres aportes fundamentales. En primer lugar, formaliza un **nuevo task**, denominado *Promptable Concept Segmentation (PCS)*, que extiende la segmentación promptable hacia escenarios de vocabulario abierto. En segundo lugar, introduce el **benchmark SA-Co**, cuya diversidad conceptual supera en varios órdenes de magnitud a los datasets tradicionales, permitiendo evaluaciones más realistas y exigentes. Finalmente, propone una **arquitectura híbrida detector–tracker**, diseñada para desacoplar explícitamente los procesos de reconocimiento semántico, localización espacial y seguimiento temporal, evitando los conflictos habituales entre estas tareas.

Bajo este marco, el presente informe aborda el análisis del trabajo desde una perspectiva técnica, comparativa y crítica, examinando tanto sus contribuciones

metodológicas y experimentales como las limitaciones que aún persisten en este enfoque.

2. Trabajos relacionados

2.1 Segmentación promptable e interactiva

En el presente trabajo, se basa directamente en la línea de investigación inaugurada por **SAM (Kirillov et al., 2023)** y posteriormente extendida por **SAM 2 (Ravi et al., 2024)**. Estas propuestas introdujeron una mejora radical en el concepto de segmentación visual guiada por prompts y su gran posibilidad de refinar los resultados de manera interactiva. **Este nuevo punto de vista** hace que el usuario participe activamente en el proceso de segmentación mediante indicaciones simples, como pueden ser los puntos, las cajas o las máscaras, lo que representa **un avance significativo metodológico** frente a las versiones anteriores que estaban basadas en clases cerradas y **anotaciones rígidas**. No obstante, a pesar de su innovación, ambas versiones mantienen dentro del paradigma de **segmentación de instancia única**, en el que cada **prompt** está asociado a un único objeto. Esta restricción limita su gran utilidad en los escenarios en donde se requiere una detección exhaustiva, es decir, la de identificar **simultáneamente todas las instancias que pertenecen a un mismo concepto visual** dentro de una imagen o una secuencia de **video**. En contraste con este enfoque, **SAM 3** hace mejoras radicales, extendiendo el paradigma de la **segmentación promptable** hacia una **segmentación conceptual exhaustiva**, permitiendo funcionar directamente sobre conceptos visuales definidos **de forma abierta**. Esta gran diferencia no solo se manifiesta a nivel conceptual, sino

también **se refleja de manera empírica en los resultados experimentales**. Tal como se observa en la **Figura 2 del paper**, SAM 3 supera **de forma consistente** a los métodos previos como **OWLv2** en escenarios de **vocabulario abierto**, dando como evidencia una mejora sustantiva en la capacidad del modelo para **generalizar y segmentar conceptos no restringidos a un conjunto cerrado de clases**.

2.2 Detección y segmentación open-vocabulary

Diversos trabajos recientes han abordado el problema de la detección **Open Vocabulary** mediante el uso de representaciones conjuntas de visión y lenguaje. Métodos como **OWLv2**, **GroundingDINO**, **GLIP** y **MDETR** han demostrado que es posible extender la detección de objetos más allá de un conjunto cerrado de clases, apoyándose en **descripciones textuales** para guiar el reconocimiento visual. Estos enfoques han contribuido de manera positiva a ampliar la capacidad semántica de los sistemas de visión por computadora. Adicionalmente, pese a sus avances, dichos métodos presentan **limitaciones importantes** cuando se evalúa desde una perspectiva más amplia de segmentación y uso interactivo. En la mayoría de los casos, **su salida principal** se restringe a cajas delimitadoras, sin otorgar **máscaras de segmentación de alta calidad** que permitan una delimitación más precisa a nivel de cada píxel. Asimismo, podemos indicar que estos modelos no han sido concebidos para procesos de **refinamiento interactivo**, lo que hace es limitar la participación del usuario en la corrección o el ajuste en los resultados. A ello se adiciona la ausencia, en muchos casos de estos enfoques, de una **integración nativa con mecanismos de seguimiento en video**, dificultando su aplicación en escenarios dinámicos y temporales. Aparte, en el contexto de **SAM 3**, retomamos elementos fundamentales de arquitecturas basadas en **DETR** y **MDETR**, pero introduce un cambio en su arquitectura, la cual podemos indicar como la **separación explícita entre el reconocimiento**

semántico del concepto y su localización espacial. Esta disociación permite abordar de manera más práctica las exigencias propias de la segmentación conceptual exhaustiva, al evitar que una misma cabeza del modelo deba resolver simultáneamente qué es un objeto y dónde se encuentra. Dicha elección, ausente o poco explorada en gran parte de los trabajos previos, constituye uno de los factores claves que explican los grandes avances observados en escenarios de vocabulario abierto.

2.3 Segmentación y tracking en video

En las tareas de segmentación y seguimiento de objetos en video, la literatura especializada ha tenido que agrupar en torno a estos dos enfoques principales. Por un lado, los métodos de **tracking-by-detection** como **SORT** y **ByteTrack**, realizan la detección de objetos de manera independiente en cada fotograma y posteriormente asocian dichas detecciones a lo largo del tiempo mediante criterios geométricos y de apariencia. Por otro lado, los métodos como **end-to-end**, entre los que se incluyen propuestas como **TrackFormer** y **MOTR**, buscan integrar de manera única una arquitectura donde los procesos de detección están asociados a la variable temporal. Si bien es cierto estos enfoques han mostrado resultados competitivos, el trabajo analizado señala que, de manera acertada, una limitación estructural compartida: **la existencia de conflictos inherentes entre la detección semántica y la preservación de la identidad de los objetos**. En particular, **mientras la detección requiere priorizar la identificación correcta de las categorías visuales**, el seguimiento demanda separar y mantener identidades consistentes, incluso en situaciones de solapamiento, oclusión o cambios drásticos en la apariencia. Frente a este problema, **SAM 3** adopta una estrategia diferente orientada en una **arquitectura modular**, en la cual las responsabilidades se encuentran claramente delimitadas. El detector opera de manera **agnóstica a la identidad**, concentrándose exclusivamente en ubicar las instancias que corresponden al concepto dado, mientras que el módulo **tracking**

asume de manera explícita la tarea de seguimiento temporal. La separación funcional permite reducir los conflictos entre ambas.

3. Metodología

3.1 Definición formal de PCS

La **Promptable Concept Segmentation (PCS)** se define como la tarea orientada a **detectar, segmentar y rastrear de manera exhaustiva todas las instancias asociadas a un mismo concepto visual**, el cual es especificado mediante una **frase nominal simple**, ejemplos visuales, o una combinación de ambos. Esta tarea se plantea tanto para imágenes individuales como para secuencias de video de corta duración, incorporando explícitamente la dimensión temporal cuando corresponde.

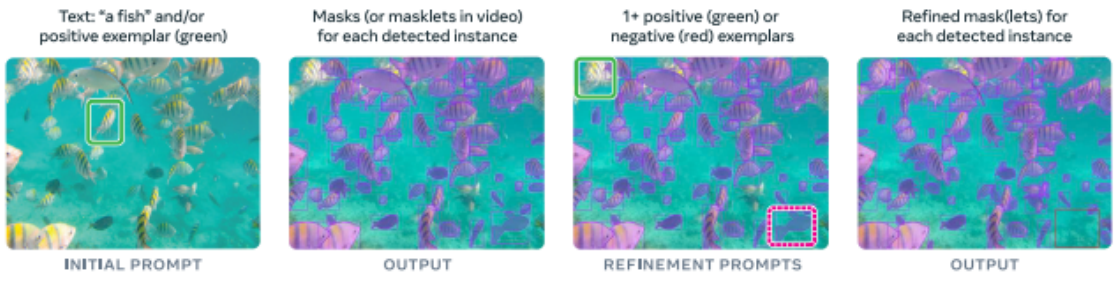

Figure 3 Illustration of supported initial and optional interactive refinement prompts in the PCS task.

La **Figura 3 del paper** ilustra los distintos tipos de *prompts* que el modelo es capaz de procesar, incluyendo descripciones textuales de alcance global, ejemplos visuales positivos y negativos, así como mecanismos de refinamiento interactivo. Desde una perspectiva metodológica, esta formulación resulta especialmente relevante, ya que **delimita de manera explícita el espacio lingüístico de entrada** al restringirlo a frases nominales simples. De este modo, se evita abordar el problema abierto del razonamiento lingüístico complejo, permitiendo que el modelo se concentre en la correcta asociación entre conceptos visuales y sus manifestaciones en la escena.

Esta decisión de diseño no solo simplifica el proceso de anotación y evaluación, sino que también contribuye a una mayor estabilidad y reproducibilidad experimental, al reducir ambigüedades semánticas inherentes al lenguaje natural más elaborado. En consecuencia, la definición formal de PCS establece una base metodológica clara sobre la cual se construyen tanto la arquitectura del modelo como los procedimientos de entrenamiento y evaluación descritos en el trabajo.

3.2 Arquitectura del modelo

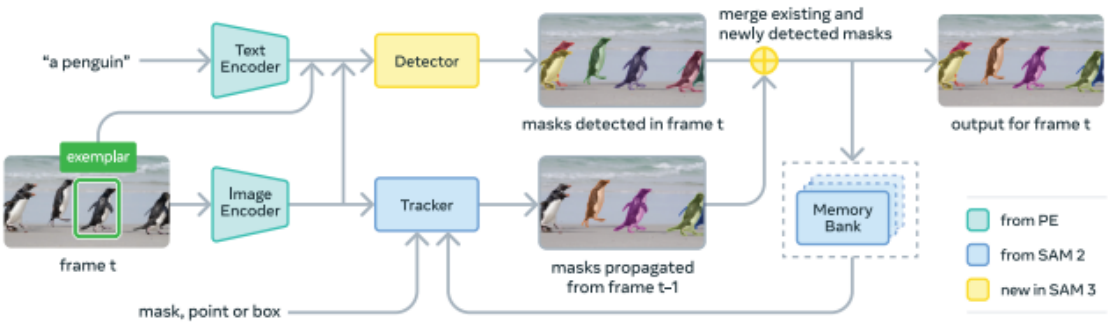

Figure 4 SAM 3 architecture overview

La **Figura 4** presenta la arquitectura general de SAM 3, compuesta por:

- Un **Perception Encoder** compartido visión–lenguaje.

- Un **detector basado en DETR**, condicionado por prompts.
- Un **tracker con memoria**, heredado de SAM 2.

Un aporte clave es el **presence token**, descrito en detalle la importancia y diferencias clave entre SAM2 y SAM3 esto permite reevaluar la coincidencia de las instancias propuestas con el requirimiento del prompt conceptual. Este concepto de token desacopla provee:

- La probabilidad de que el concepto esté presente en la imagen.
- La probabilidad de que una propuesta específica corresponda al concepto.

Desde una perspectiva crítica, esta decisión arquitectónica es elegante porque aborda un problema clásico de DETR: forzar a una misma cabeza a resolver simultáneamente *qué* y *dónde*.

3.3 Interactividad y ejemplos visuales

	Instance Segmentation						Box Detection								Semantic Segmentation		
	LVIS		SA-Co				LVIS		COCO		SA-Co				ADE-847	PC-59	Cityscapes
Model	cgF_1	AP	Gold cgF_1	Silver cgF_1	Bronze cgF_1	Bio pmF_1	cgF_1	AP	AP	AP_o	Gold cgF_1	Silver cgF_1	Bronze cgF_1	Bio pmF_1	mIoU	mIoU	mIoU
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†	21.3δ	–	–	–	–	52.4†	56.0†	–	22.5δ	–	–	–	–	–	–
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**

Table 1 Evaluation on image concept segmentation with text. APo corresponds to COCO-O accuracy, ⋆: partially trained on LVIS, †: from original papers, δ: from DINO-X API. Gray numbers indicate usage of respective closed set training data (LVIS/COCO).

Model	ODinW13		RF-100VL	
	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**

Table 2 Zero-shot and 10-shot transfer on in-the-wild datasets

Model	COCO				LVIS				ODinW13			
	AP T	AP^+ T	AP^+ I	AP^+ T+I	AP T	AP^+ T	AP^+ I	AP^+ T+I	AP T	AP^+ T	AP^+ I	AP^+ 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**

Table 3 Prompting with 1 exemplar on COCO, LVIS and ODinW13. Evaluation per prompt type: T (text-only), I (image-only), and T+I (combined text and image). AP+ is evaluated only on positives examples

A diferencia de los enfoques basados exclusivamente en descripciones textuales, **SAM 3** incorpora de manera explícita la posibilidad de utilizar **ejemplos visuales tanto positivos como negativos** como parte del proceso de especificación del concepto. Esta capacidad amplía significativamente la expresividad de los *prompts*, permitiendo al modelo capturar matices visuales que pueden resultar difíciles de describir únicamente mediante lenguaje natural.

Los resultados experimentales presentados en la **Tabla 3 del paper** evidencian que la **combinación de información textual y visual (T+I)** conduce de forma consistente a un mejor desempeño en comparación con los enfoques que emplean únicamente texto o únicamente imágenes. Esta ventaja se vuelve particularmente notable en **datasets de alta complejidad y gran diversidad semántica**, como **LVIS**, donde la ambigüedad conceptual y la variabilidad visual representan un desafío significativo. En estos escenarios, la integración conjunta de ambas modalidades permite una delimitación más precisa del concepto objetivo, reforzando la robustez y la capacidad de generalización del modelo.

3.4 Motor de datos (Data Engine)

Uno de los aportes más innovadores del trabajo es el diseño de un **motor de datos** (*data engine*), ilustrado en la **Figura 5**, cuya función principal es sostener el entrenamiento de SAM 3 a gran escala mediante un proceso de generación y validación continua de datos. Este sistema integra de manera coordinada la participación de **anotadores humanos**, **modelos de inteligencia artificial especializados en verificación** en particular para la evaluación de calidad de máscaras (*Mask Verification*) y exhaustividad de anotación (*Exhaustivity Verification*), así como la **producción masiva de datos sintéticos**.

A diferencia de los enfoques tradicionales, basados en **datasets estáticos y previamente cerrados**, este motor de datos adopta una lógica **iterativa y auto-mejorable**, en la que el propio modelo, junto con los verificadores automáticos, contribuye activamente a identificar errores, ambigüedades y casos difíciles. Este esquema permite dirigir el esfuerzo humano hacia las situaciones más complejas, mientras que los casos rutinarios son gestionados de forma automática, incrementando de manera significativa la eficiencia del proceso de anotación.

Desde una perspectiva más amplia, este cambio de paradigma tiene implicancias profundas para la **escalabilidad futura de los modelos de visión por computadora**. Al desacoplar el crecimiento del dataset de una dependencia estricta de anotación manual, el *data engine* propuesto sienta las bases para sistemas de aprendizaje más sostenibles, capaces de adaptarse progresivamente a nuevos dominios visuales y conceptuales sin incurrir en costos prohibitivos.

4. Resultados

4.1 Segmentación de conceptos en imágenes

La **Tabla 1** constituye uno de los resultados más relevantes del trabajo, ya que sintetiza el desempeño de SAM 3 en tareas de segmentación conceptual sobre imágenes. En el benchmark **SA-Co/Gold**, el modelo logra **duplicar el valor de la métrica cgF1** en comparación con OWLv2 y alcanza aproximadamente un **74 % del desempeño humano estimado**. Este resultado adquiere especial importancia si se considera que SA-Co/Gold es un benchmark de **vocabulario abierto** que abarca alrededor de **207 000 conceptos únicos**, una escala sin precedentes en trabajos previos de segmentación y detección visual.

Desde una perspectiva crítica, estos resultados respaldan de manera sólida la hipótesis central del trabajo: la segmentación conceptual exhaustiva no puede resolverse únicamente mediante el uso de **representaciones visión-lenguaje genéricas**, como los *embeddings* derivados de CLIP. Por el contrario, el desempeño observado sugiere que factores como la **calidad y diversidad del dataset**, así como la **separación explícita entre reconocimiento semántico y localización espacial**, desempeñan un papel determinante en la capacidad del modelo para generalizar en escenarios de vocabulario abierto.

4.2 Generalización y few-shot learning

Los resultados presentados en la **Tabla 2** muestran que SAM 3 alcanza el **estado del arte** tanto en configuraciones *zero-shot* como *10-shot* sobre datasets *in-the-wild*. Este comportamiento pone de manifiesto la capacidad del modelo para adaptarse eficazmente a nuevos conceptos con una cantidad mínima o incluso nula de datos de entrenamiento específicos.

En conjunto, estos resultados refuerzan la idea de que la **Promptable Concept Segmentation (PCS)** constituye una representación más **transferible y robusta** que los enfoques tradicionales de detección basados en clases cerradas. Al operar directamente sobre conceptos y no sobre etiquetas predefinidas, SAM 3 demuestra una mayor flexibilidad frente a variaciones semánticas y visuales, lo

que resulta especialmente valioso en escenarios reales donde la diversidad y la imprevisibilidad de los objetos son la norma.

4.3 Video y seguimiento

	SA-Co/VEval benchmark test split						Public benchmarks			
	SA-V		YT-Temporal-1B		SmartGlasses		LVVIS	BURST	YTVIS21	OVIS
	(2.0K NPs)		(1.7K NPs)		(2.4K NPs)		(1.2K NPs)	(482 NPs)	(40 NPs)	(25 NPs)
Model	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**

Table 5 Video PCS from a text prompt (open-vocabulary video instance segmentation) on SA-Co/VEval and public benchmarks. SAM 3 shows strong performance, especially on benchmarks with a large number of NPs. †: GLEE and LLMDet do not perform well zero-shot on SA-Co/VEval.

En el contexto de secuencias de video, los resultados presentados en la **Tabla 5** evidencian que **SAM 3 supera de manera consistente** a baselines relevantes, como **GLEE** y diversas configuraciones híbridas evaluadas en el trabajo. En particular, el desempeño alcanzado en términos de la métrica **pHOTA** pone de manifiesto la capacidad del modelo para **preservar la coherencia temporal y semántica de las instancias segmentadas**, incluso en escenarios caracterizados por una alta diversidad conceptual y un número elevado de conceptos distintos.

4.4 Ablaciones y análisis crítico

	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.

Table 9 Selected model and data ablations on SA-Co/Gold. Numbers across tables are not directly comparable.

	Supervise mask scores only when concept present	Supervise total score	Sup. total score, detach presence	SA-Co/Gold		
				$\mathrm{cgF_1}$	IL_MCC	$\mathrm{pmF_1}$
a.	✓	×	×	54.0	0.82	65.5
b.	×	×	×	52.2	0.81	64.2
c.	✓	✓	×	**54.9**	**0.83**	**66.0**
d.	✓	×	✓	53.6	**0.83**	64.9

Table 10 Supervision strategy for object/mask scores for a model with a presence token. We find the best supervision strategy is to supervise mask scores only for positive concepts and to supervise the presence and mask scores separately, although their product is used as the total object score during inference.

Este comportamiento resulta especialmente significativo en tareas de segmentación conceptual en video, donde la consistencia a lo largo del tiempo constituye un desafío crítico debido a fenómenos como oclusiones, cambios de apariencia y movimientos complejos. La robustez observada sugiere que la arquitectura propuesta logra integrar de manera efectiva la detección conceptual con los mecanismos de seguimiento temporal.

Por otro lado, los estudios de ablación resumidos en las **Tablas 9 y 10** confirman de forma empírica la relevancia de varios componentes clave del sistema, entre ellos el **presence token**, la incorporación de **hard negatives** durante el entrenamiento y el uso de **datos sintéticos verificados**. Un aspecto particularmente destacable es que los **verificadores basados en inteligencia artificial** logran cerrar aproximadamente **la mitad de la brecha existente entre el desempeño del modelo y el desempeño humano**, lo que apunta hacia una evolución progresiva de los procesos de anotación. En este sentido, los resultados sugieren una tendencia clara hacia la construcción de **datasets cada vez menos dependientes de anotación manual exhaustiva**, con implicancias directas en la escalabilidad y sostenibilidad de futuros sistemas de visión por computadora.

5. Conclusiones

SAM3 constituye un avance significativo en la evolución reciente de la segmentación visual, no solamente por aportar **mejoras cuantitativas** que introduce al modelo con respecto a versiones anteriores como **SAM y SAM2**, sino sobre todo por su **alineamiento natural con un nuevo paradigma de sistemas de inteligencia artificial basados en la orquestación de modelos**, donde modelos como este aportan una metodología especializada en el campo de **computer vision** y de **segmentación, ubicación y trazabilidad de instancias**. Este nuevo enfoque emergente, donde los modelos dejan de concebirse como soluciones monolíticas y autosuficientes para ser entendidas como **componentes especializados, reutilizables y combinables**, de forma análoga a **piezas modulares** que se integran dentro de una arquitectura de mayor alcance. Dentro de esa perspectiva, **SAM3** puede interpretarse como un componente de gran valor en el ecosistema de **visión por computadora**, específicamente diseñado para abordar de manera efectiva el problema de la **segmentación conceptual open-vocabulary**, tanto en imágenes estáticas como en secuencias de video. Su objetivo no solamente es sustituir otros enfoques existentes, sino **complementarlos**, aportando capacidades particulares dentro de la **segmentación exhaustiva basada en conceptos**, como la **interactividad con el usuario** y el **seguimiento temporal consistente**, que pueden ser activadas de manera selectiva según el contexto y los requerimientos de la tarea existente.

Este énfasis en el contexto resulta fundamental, al igual que ocurre con otros algoritmos relevantes en **computer vision**, como los modelos de **Human Action Recognition (HAR)**, donde la detección de eventos, la estimación de pose o el reconocimiento de escenas responden a patrones de reconocimiento útiles en situaciones específicas. La aplicabilidad óptima de **SAM3** depende del problema visual que se desea resolver, del dominio considerado y de las restricciones operativas del sistema, pudiendo maximizar la **productividad**, la **eficiencia computacional**, la **precisión semántica** o la **robustez temporal** según sea el caso.

En ese sentido, tanto la arquitectura como la filosofía de diseño de **SAM3** lo hacen particularmente adecuado para su integración en **sistemas de orquestación de modelos**, como los implementados en frameworks del tipo **LangGraph**, donde múltiples redes neuronales y algoritmos cooperan bajo un flujo de control explícito. En estos sistemas, **SAM3** puede desempeñar distintos roles funcionales, ya sea como **módulo de percepción visual especializada**, como **proveedor de máscaras semánticas** que alimentan procesos de razonamiento posteriores, o como **herramienta visual invocada dinámicamente por agentes de mayor nivel**, incluidos modelos de lenguaje multimodal.

La contribución más profunda de este trabajo, por tanto, no se limita a la mejora del estado del arte en segmentación visual, sino que reside en demostrar cómo una red neuronal cuidadosamente diseñada, con un **task claramente delimitado como la Promptable Concept Segmentation (PCS)**, puede integrarse de forma armónica dentro de un ecosistema más amplio de **inteligencia artificial modular y orquestada**. Este enfoque favorece el desarrollo de sistemas más **escalables**, **interpretables** y **adaptables**, en los que cada algoritmo aporta valor desde su especialización sin imponer rigidez estructural al conjunto.

En síntesis, **SAM3** debe entenderse como un bloque fundamental dentro del nuevo paradigma de **inteligencia artificial compuesta**, en el cual la **sinergia entre modelos**, más que la supremacía de uno solo, constituye el verdadero motor del progreso de la visión por computadora y de los sistemas inteligentes de la próxima generación.

Por último, se reconoce como debilidad del modelo que su entrenamiento haya sido realizado **exclusivamente en idioma inglés**, lo que introduce una dependencia lingüística como limitante en su operatividad. No obstante, se considera que esta limitación es de carácter temporal y que, al integrarse en **modelos de orquestación de agentes con estructura de structured output**, como los proporcionados por **LangGraph**, dicho inconveniente puede superarse sin mayores dificultades

Referencias

(Formato APA 7 – Informe de Tesis)

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). *End-to-end object detection with transformers*. En **Proceedings of the European Conference on Computer Vision (ECCV)** (pp. 213–229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

Cheng, B., Schwing, A. G., & Kirillov, A. (2021). *Per-pixel classification is not all you need for semantic segmentation*. En **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)** (pp. 1–11). https://doi.org/10.1109/CVPR46437.2021.00958

Gupta, A., Dollár, P., & Girshick, R. (2019). *LVIS: A dataset for large vocabulary instance segmentation*. En **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)** (pp. 5356–5364). https://doi.org/10.1109/CVPR.2019.00549

Hu, Y.-T., Debnath, S., & Feichtenhofer, C. (2023). *DAC-DETR: Dual assignment consistency for end-to-end object detection*. En **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment anything*. En **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)** (pp. 4015–4026). https://doi.org/10.1109/ICCV51070.2023.00374

Li, J., Li, D., Xiong, Y., Xiong, Y., Zhang, S., & Zhang, X. (2023). *DINOv: Towards universal visual detection*. En **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**.

Minderer, M., Gritsenko, A., Houlsby, N., & Zhai, X. (2024). *Scaling open-vocabulary object detection*. En **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2020). *Phrase grounding: Aligning textual phrases with image regions*. **IEEE Transactions on Pattern Analysis and Machine Intelligence, 42**(9), 2201–2214. https://doi.org/10.1109/TPAMI.2019.2896989

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning transferable visual models from natural language supervision*. En **Proceedings of the International Conference on Machine Learning (ICML)** (pp. 8748–8763).

Ravi, N., Mintun, E., Kirillov, A., et al. (2024). *SAM 2: Segment anything in images and videos*. En **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**.

Wu, J., Li, X., Zhang, S., & Zhao, H. (2024). *GLEE: Generalist language enhanced entity segmentation*. En **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**.

Zhang, Y., Sun, P., Jiang, Y., & Luo, P. (2022). *ByteTrack: Multi-object tracking by associating every detection box*. En **Proceedings of the European Conference on Computer Vision (ECCV)**. https://doi.org/10.1007/978-3-031-19830-4_1

Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Suris, D., Ryali, C., … Feichtenhofer, C. (2025). *SAM 3: Segment anything with concepts*. Meta Superintelligence Labs. https://ai.meta.com/sam3