2026-06-03

Vision AI Already Knows:

Understanding the World Without Labels

Sanghyun Jo

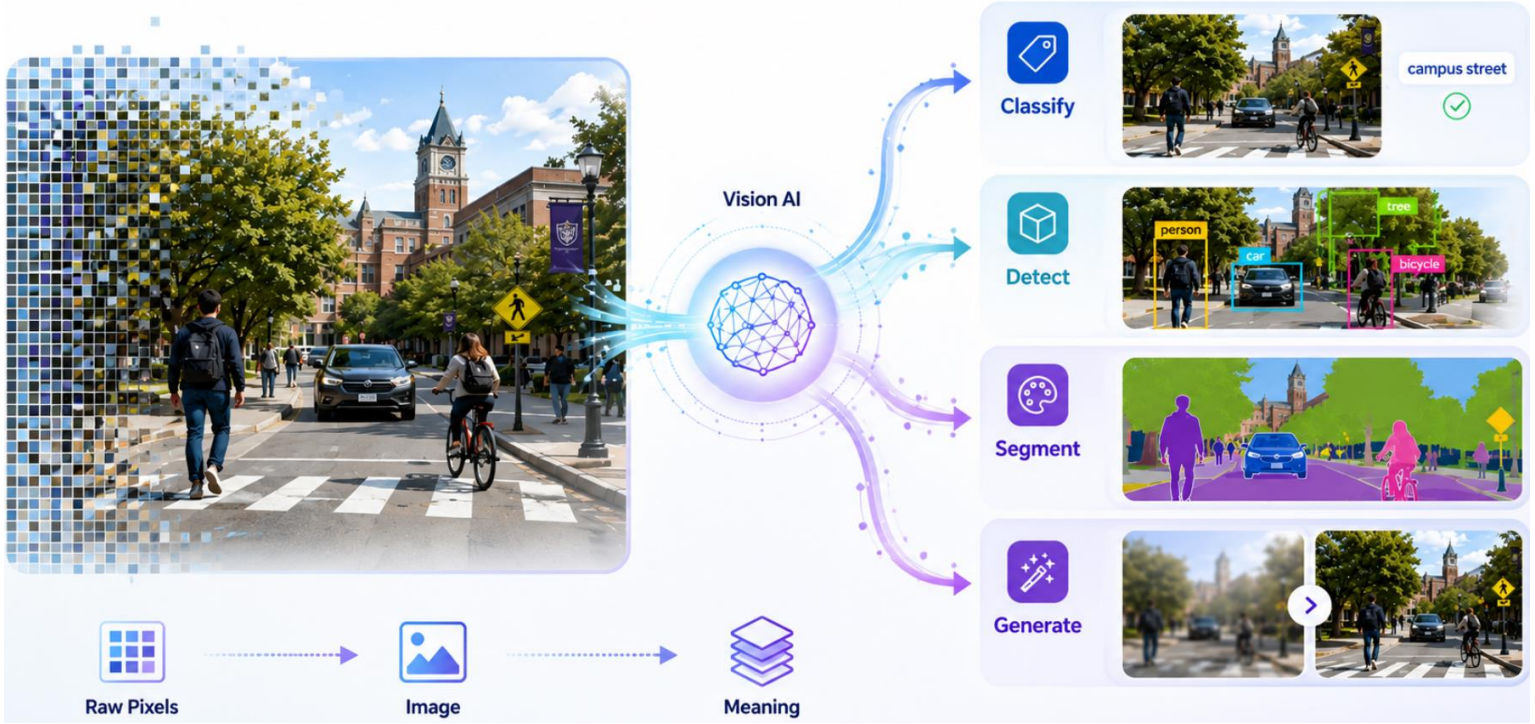
Principal AI Researcher, OGQ

Affiliated Researcher, SNU AIBL Lab

Homepage

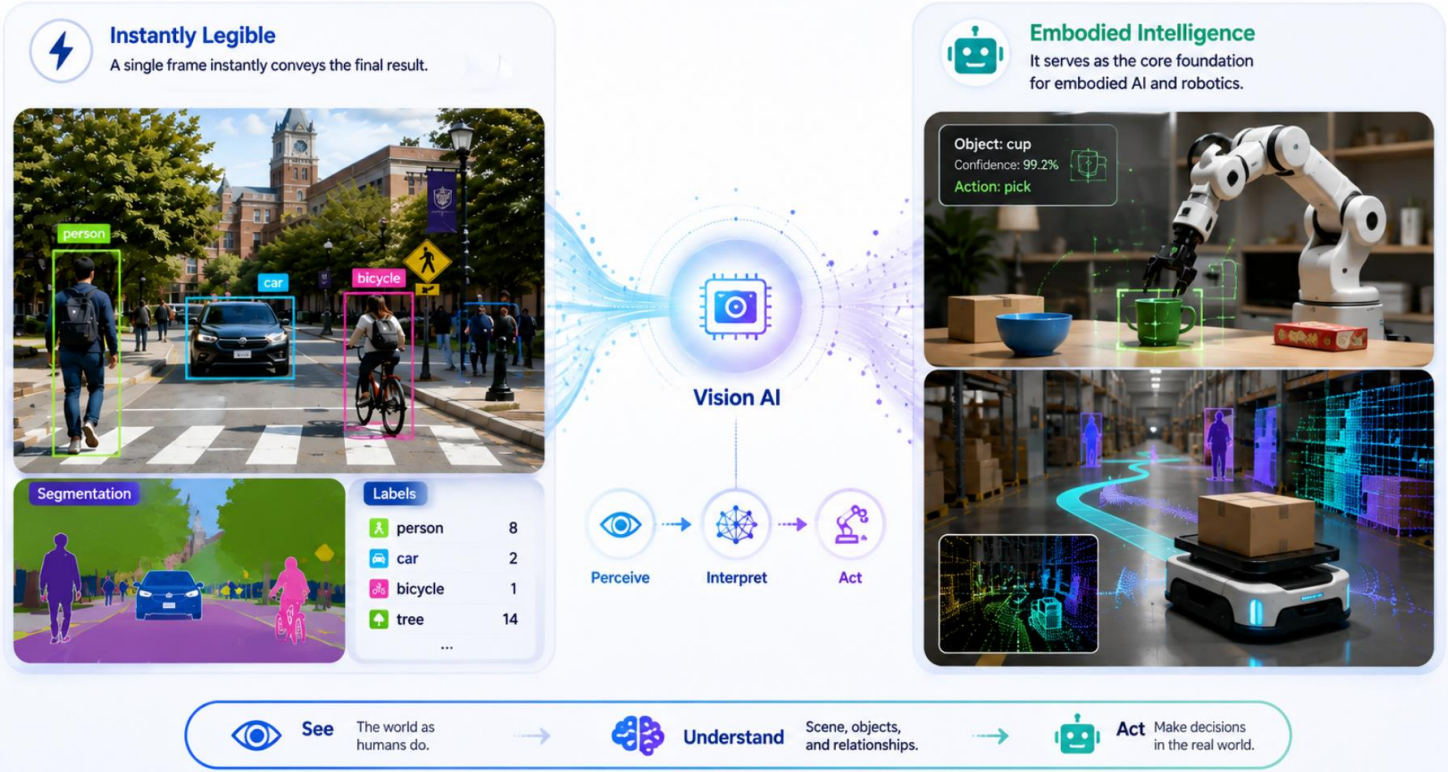
Decoding the Visual World

- **Translating Raw Data:** Vision AI extracts structured meaning from raw pixels through classification, detection, segmentation, and generation.
- **Intuitive Intelligence:** The before/after images speak for themselves.



Seeing is Believing

- **Immediate Impact:** A single frame instantly conveys the result.
- **Long Term Vision:** It serves as the core foundation for embodied AI and robotics.

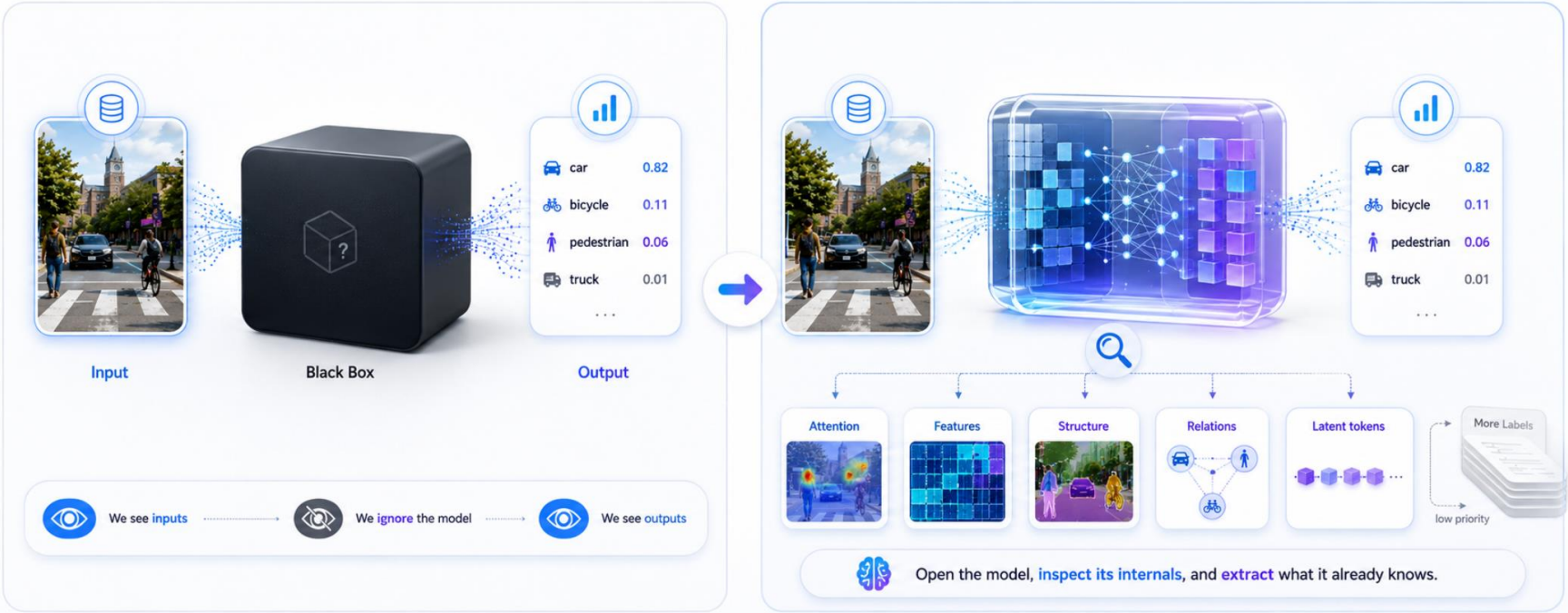


The 80 Percent Trap: The Hidden Tax of AI

- **The Hidden Cost:** Data cleansing and labeling consume 80 percent of the entire AI workflow.
- **The Annotation Penalty:** Pixel-level masking takes ~247 seconds/image, while tags only take seconds.

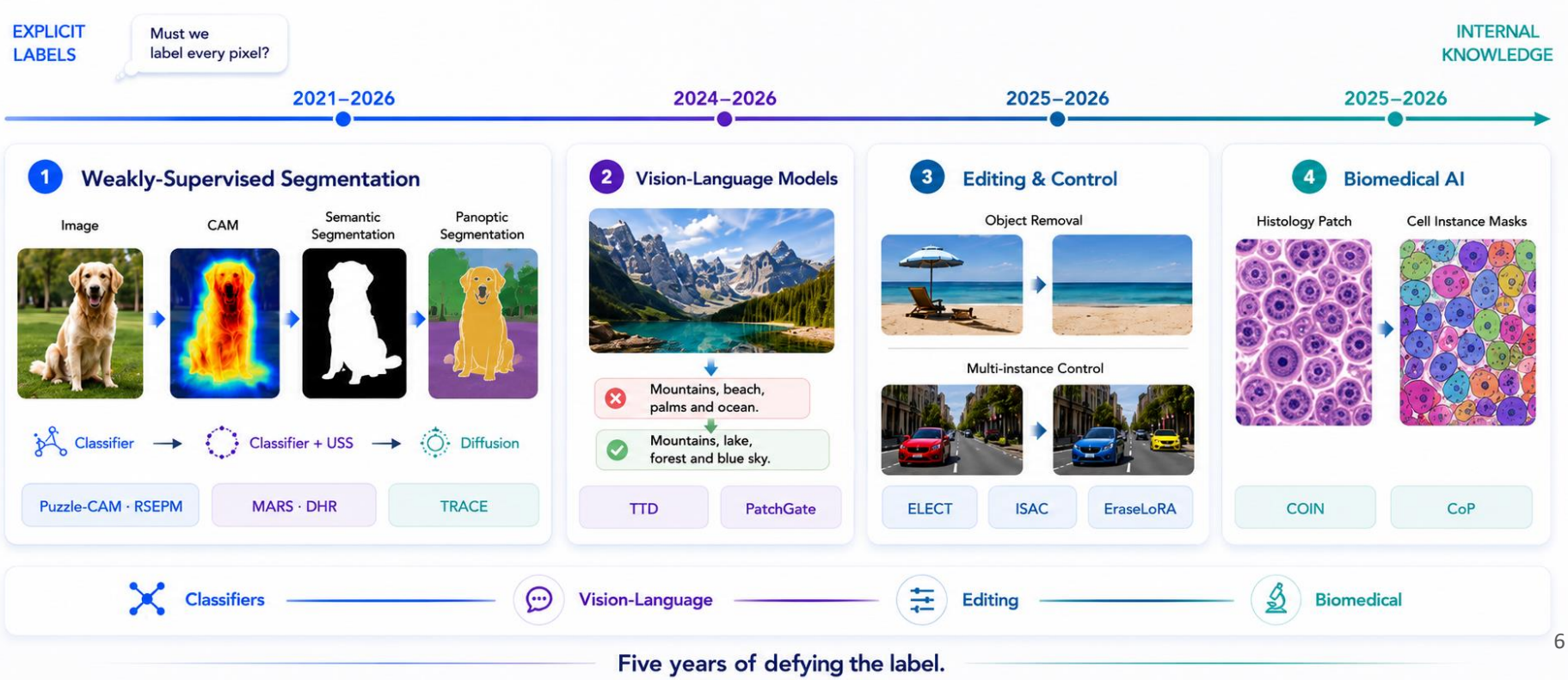
Shattering the Black Box

- **The Status Quo:** The industry focuses on inputs and outputs while ignoring internal mechanisms.
- **The Paradigm Shift:** Instead of blindly collecting more labels, we must uncover what the model already knows.



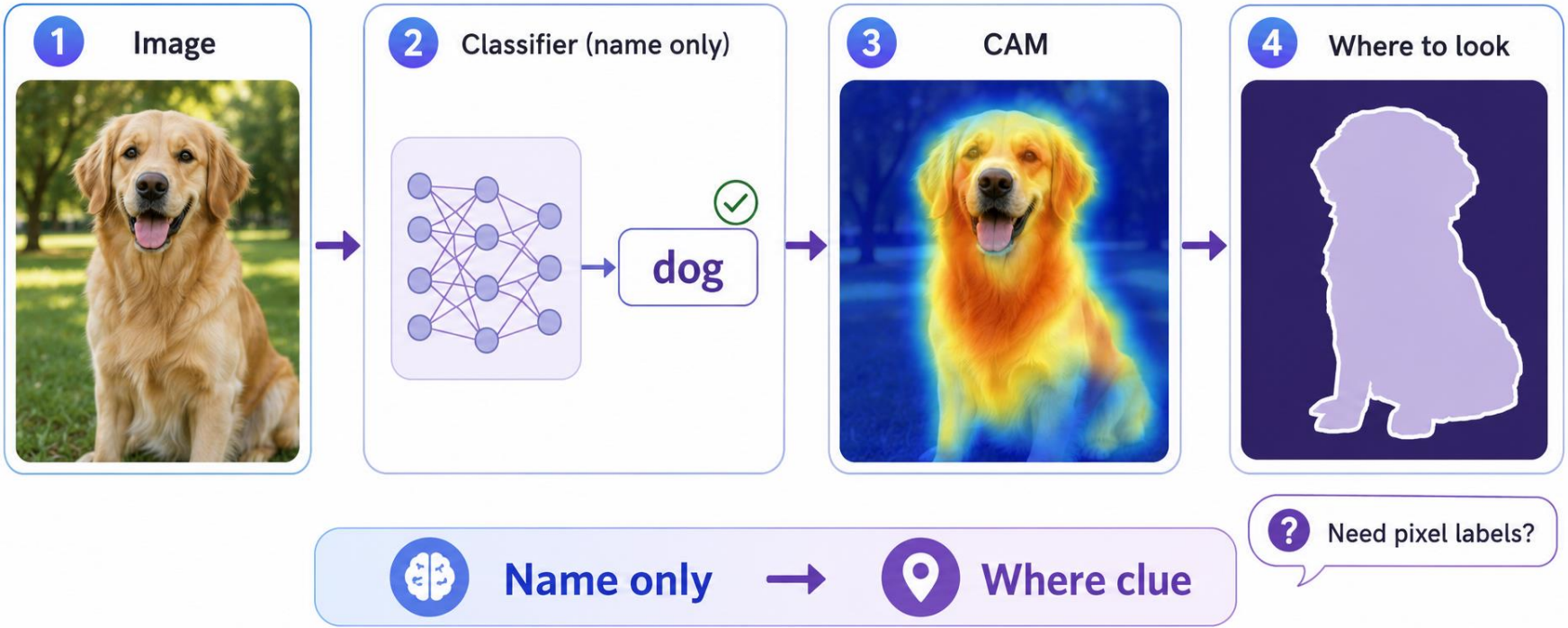
A Five Year Journey: Defying the Label

- **A Simple Question:** Must we label every pixel? This has driven my research from 2021 to today.
- **Evolving Targets:** Expanding from image classifiers to multimodal and diffusion models.



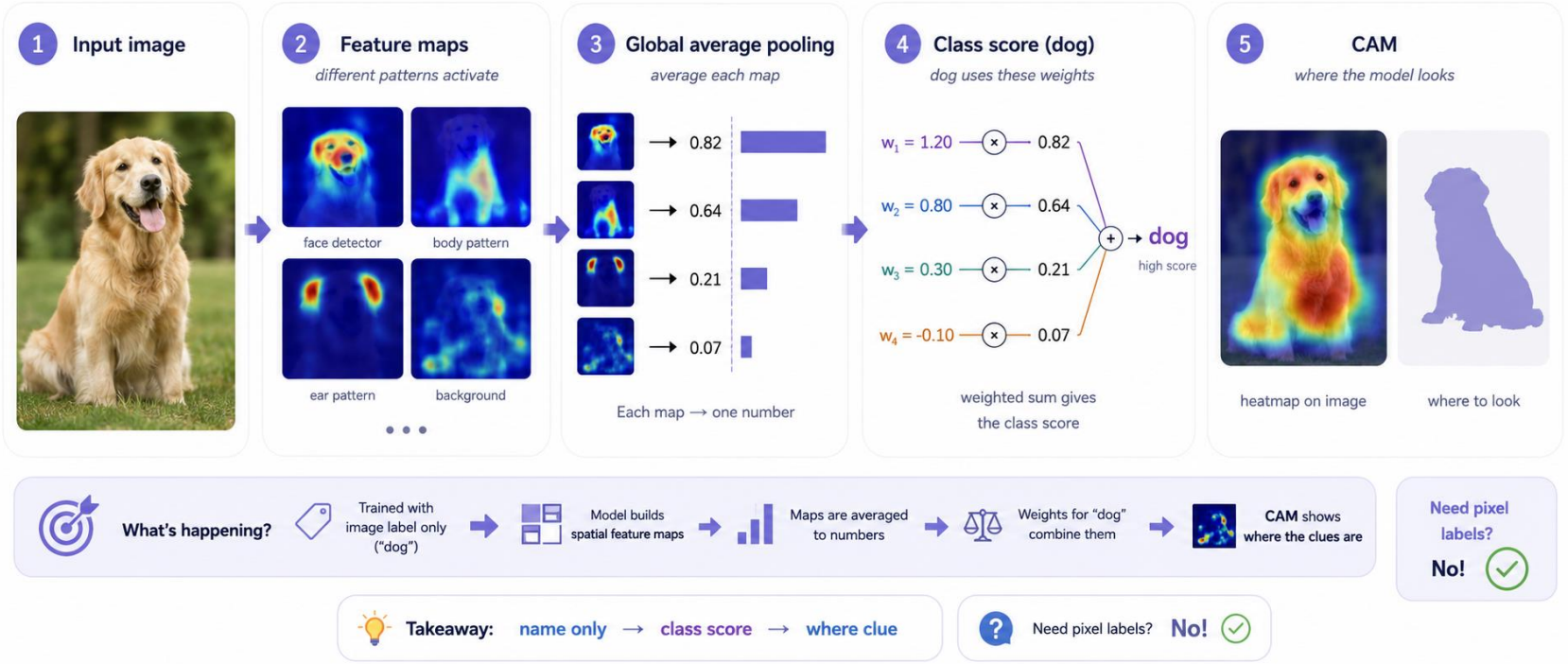
The Classifier Already Knows Where to Look

- A classifier trained only to name an image reveals where the object is through its Class Activation Map.
- The seed of everything. If the model already points at the object, do we still need pixel labels?



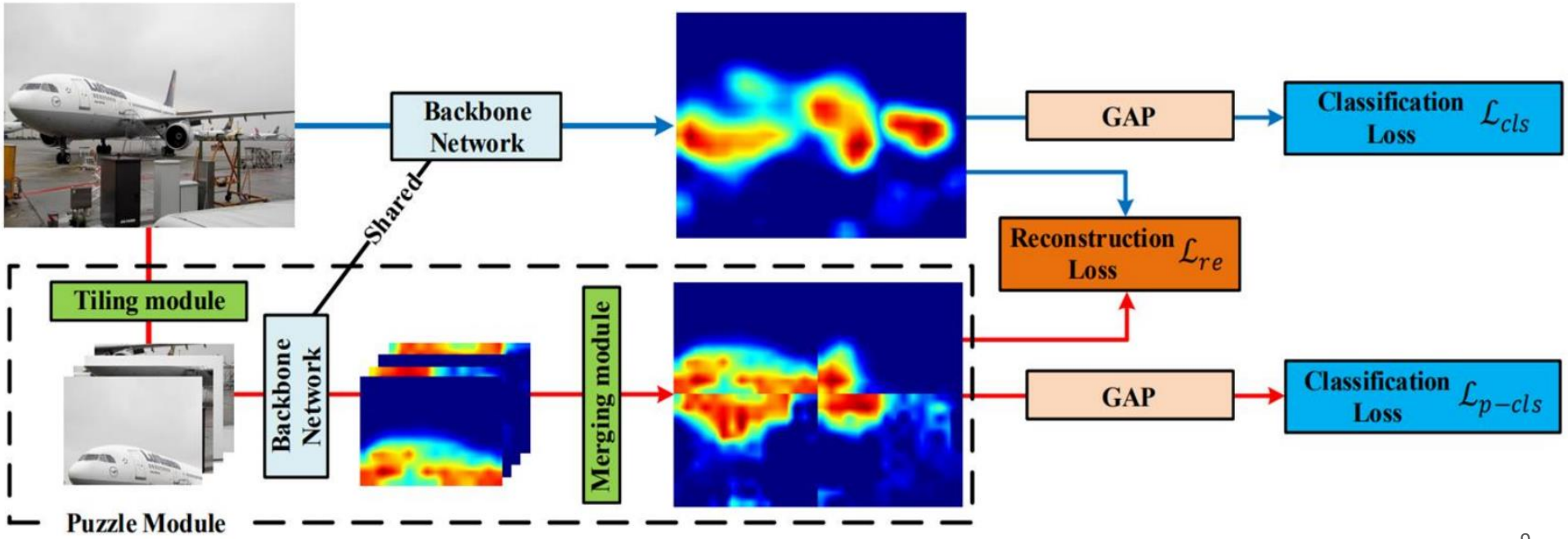
The Classifier Already Knows Where to Look

- A classifier trained only to name an image reveals where the object is through its Class Activation Map.
- The seed of everything. If the model already points at the object, do we still need pixel labels?



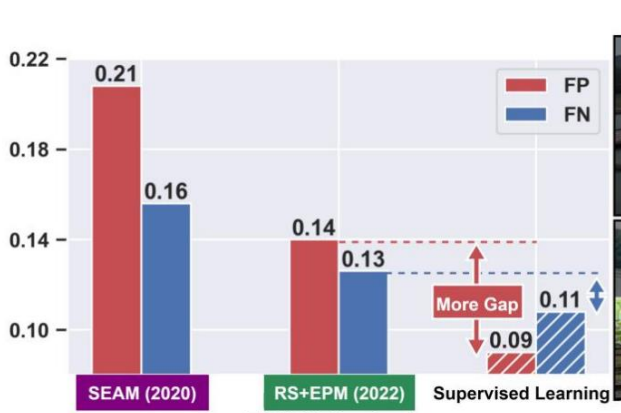
[ICIP 2021] Puzzle-CAM: Seeing the Whole, Not Just the Part

- Tiling the image and matching partial against full features expands coverage beyond the most discriminative part.
- Achieves dense localization from image level tags alone.

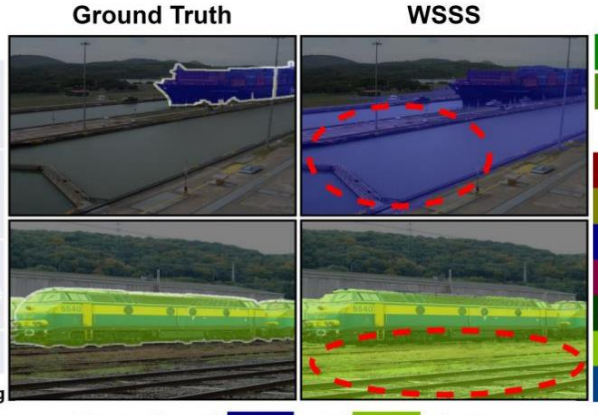


[ICCV 2023] MARS: Removing Hidden Bias Automatically

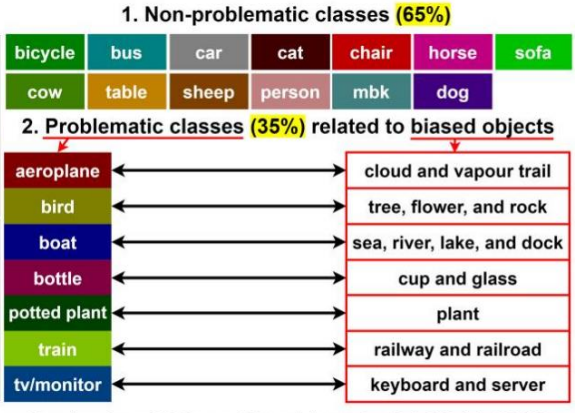
- Classifiers learn false shortcuts, such as associating a boat with water, affecting ~35 percent of VOC.



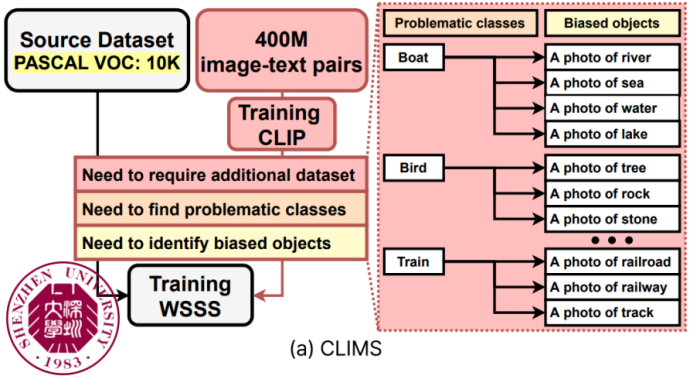
FP/FN Analysis



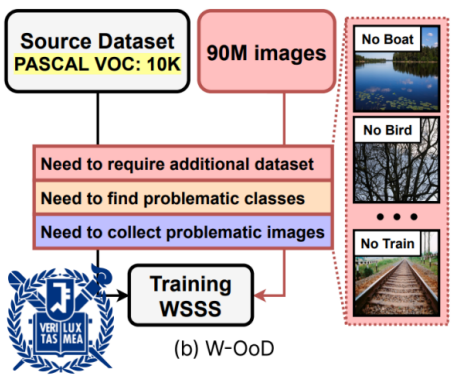
Examples of boat and train classes



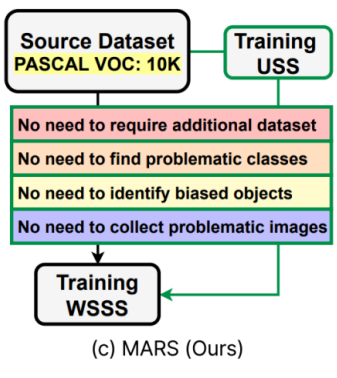
Analysis of Biased Problem in PASCAL VOC



(a) CLIMS



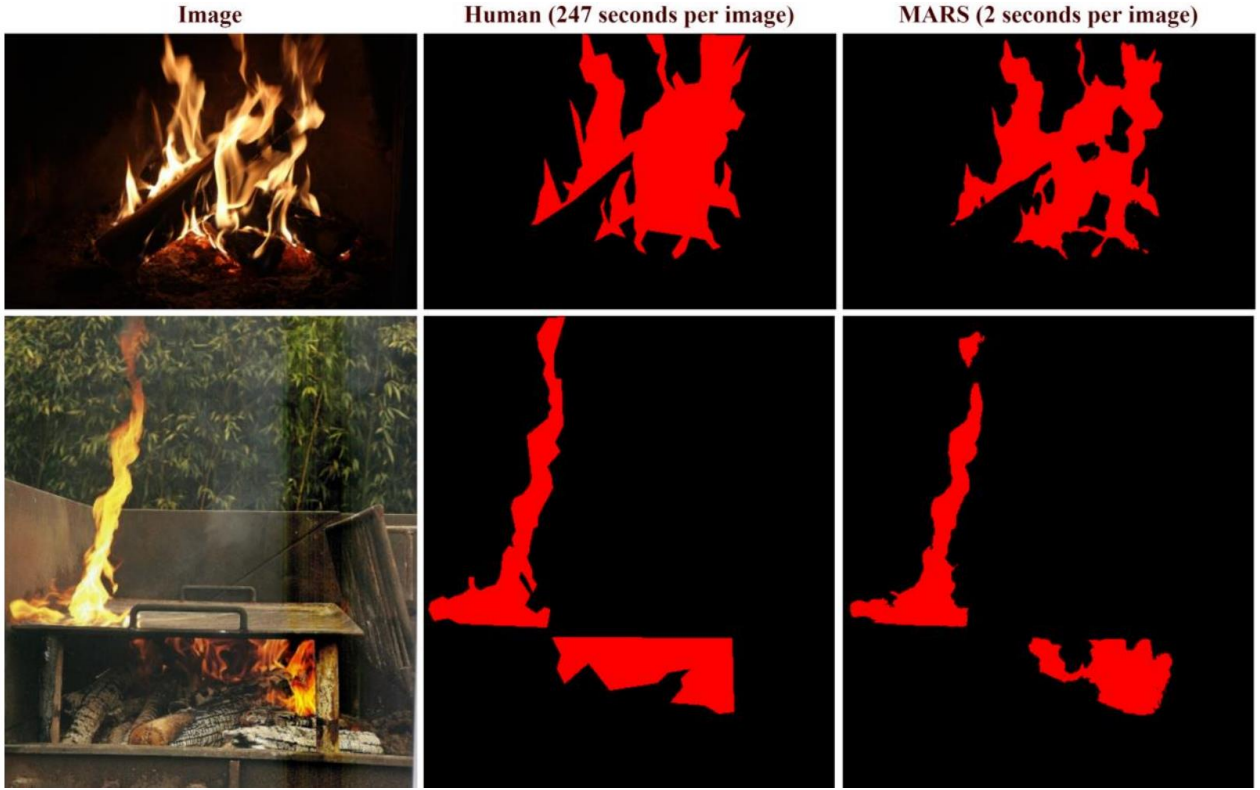
(b) W-OoD



(c) MARS (Ours)

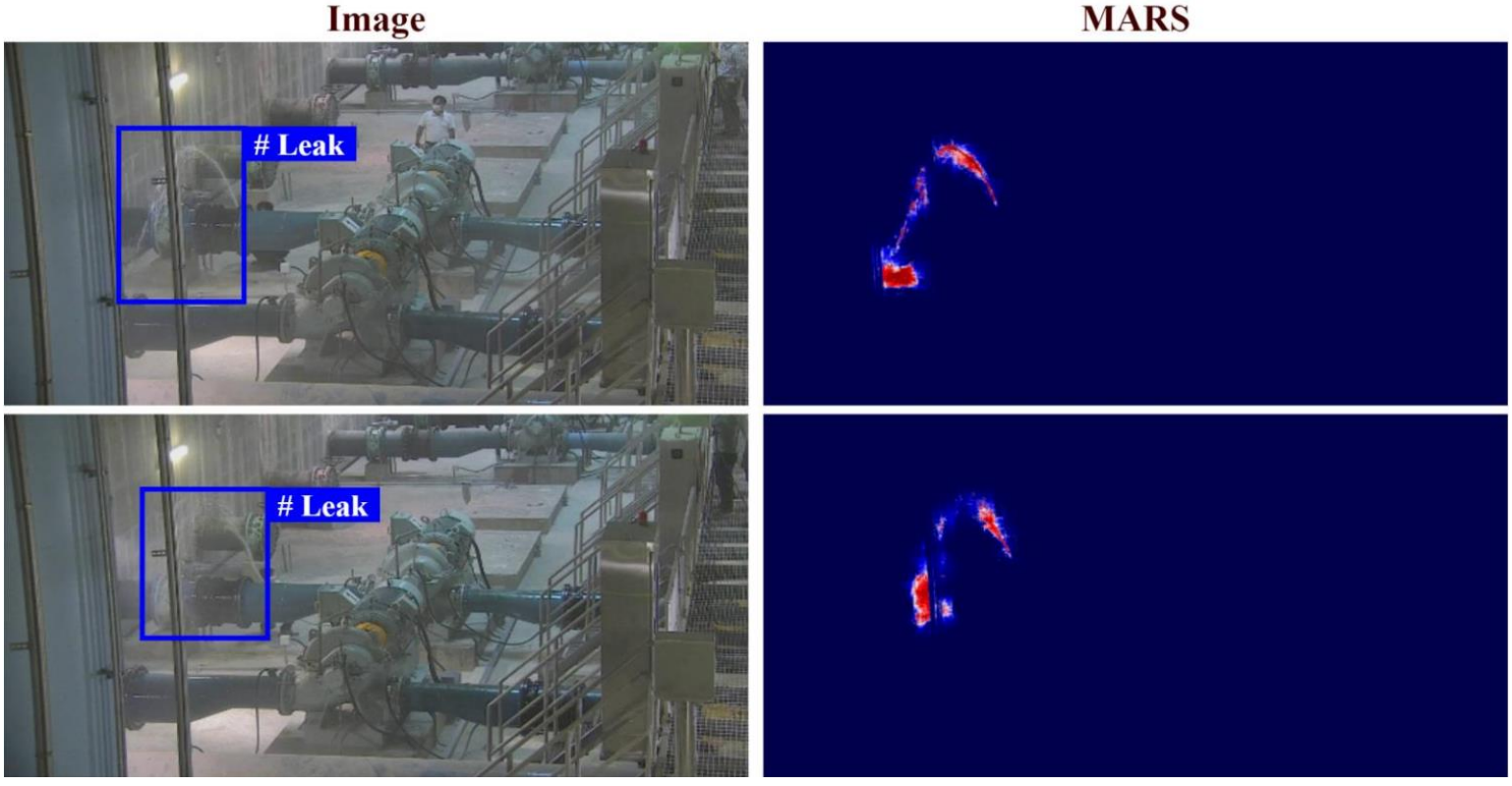
[ICCV 2023] MARS: Removing Hidden Bias Automatically

- Removed with zero extra supervision, reaching SOTA and cutting labeling time from 247 seconds to 2 seconds per image.



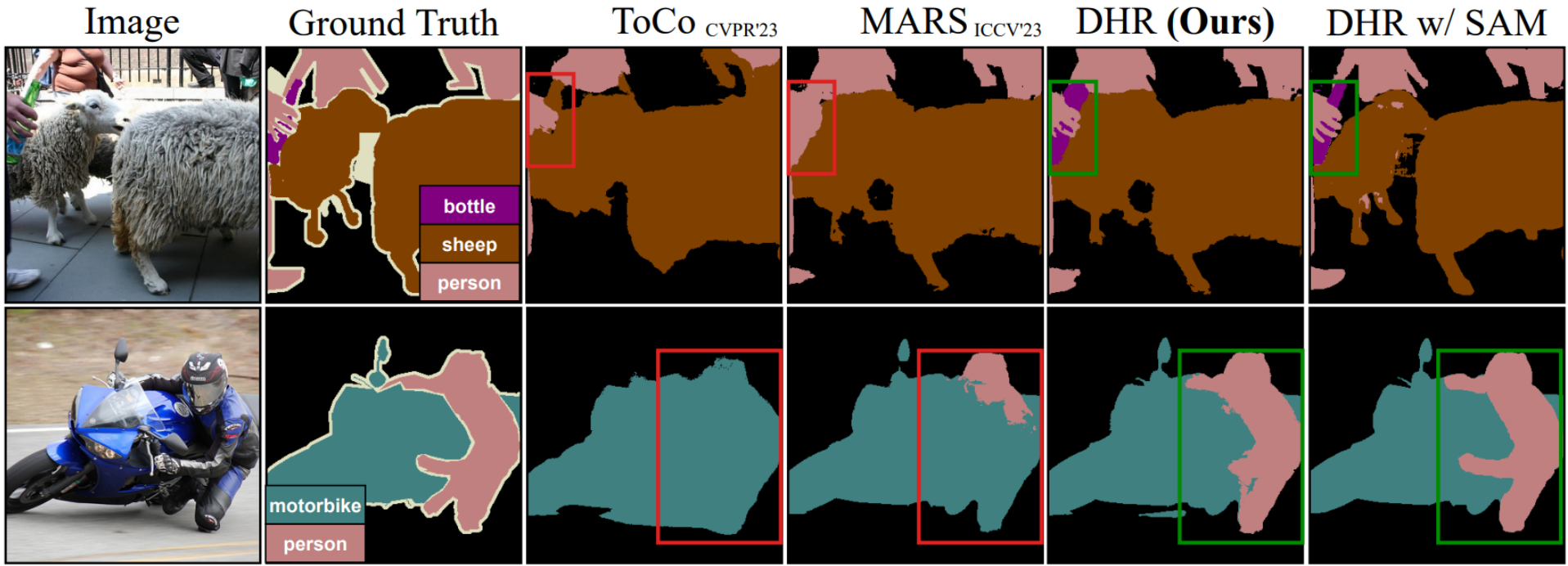
[ICCV 2023] MARS: Removing Hidden Bias Automatically

- Removed with zero extra supervision, reaching SOTA and cutting labeling time from 247 seconds to 2 seconds per image.



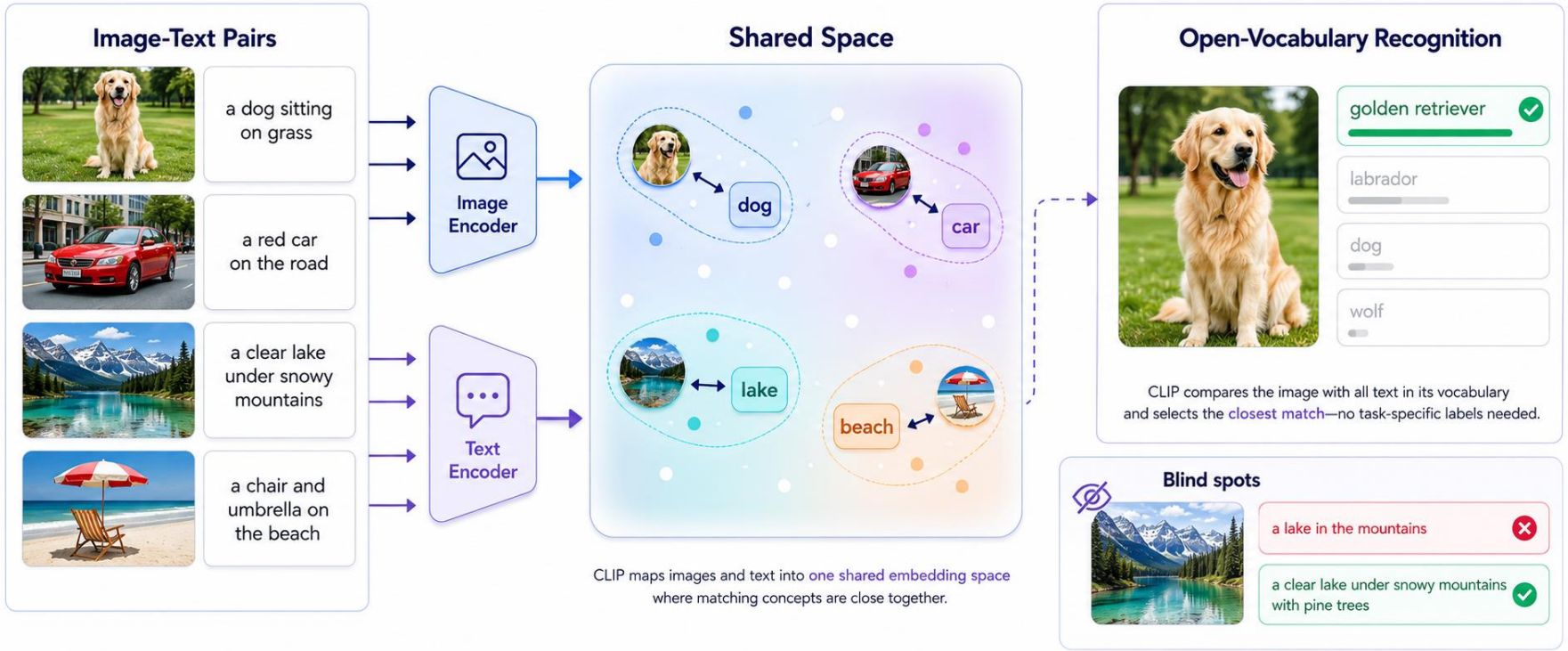
[ECCV 2024] DHR: Balancing What the Model Overlooks

- Combining two feature sources rebalances underrepresented inter- and intra-class regions.
- Rare and small regions stop disappearing, significantly lifting segmentation quality.



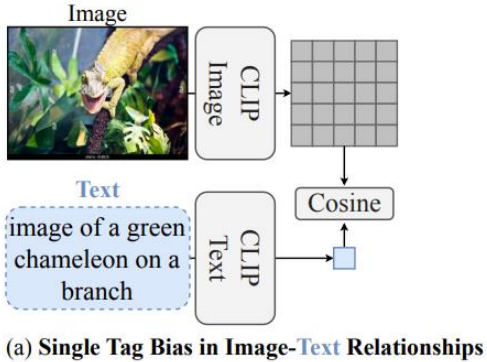
What Is a Vision-Language Model?

- CLIP learns from millions of image-text pairs, placing pictures and words into one shared space.
- This enables zero-shot classification w/o task-specific labels, though it inherits blind spots that require fixing.

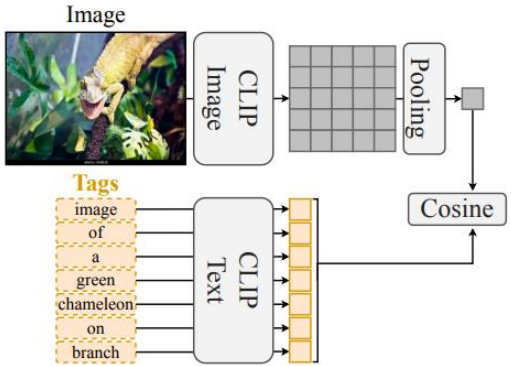


[ECCV 2024] TTD: Fixing the Single-Tag Blind Spot

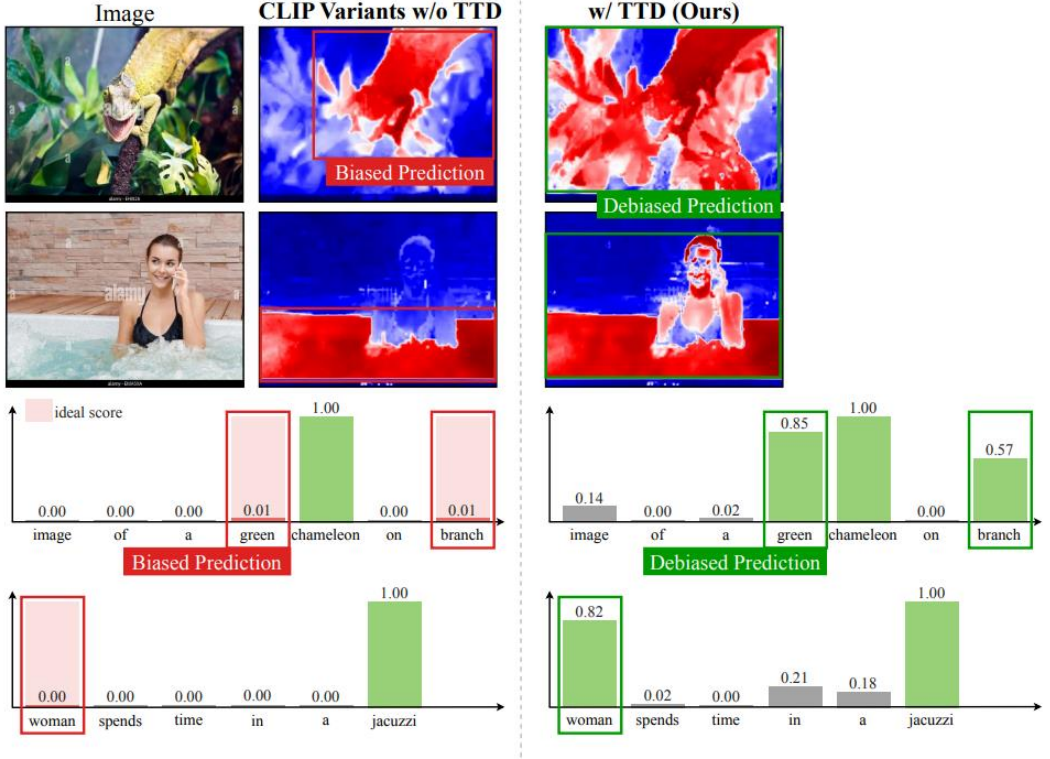
- CLIP tends to latch onto one dominant tag and ignore the rest.
- TTD self-distills text tags to restore full image-text alignment, improving multi-label understanding.



(a) **Single Tag Bias in Image-Text Relationships**

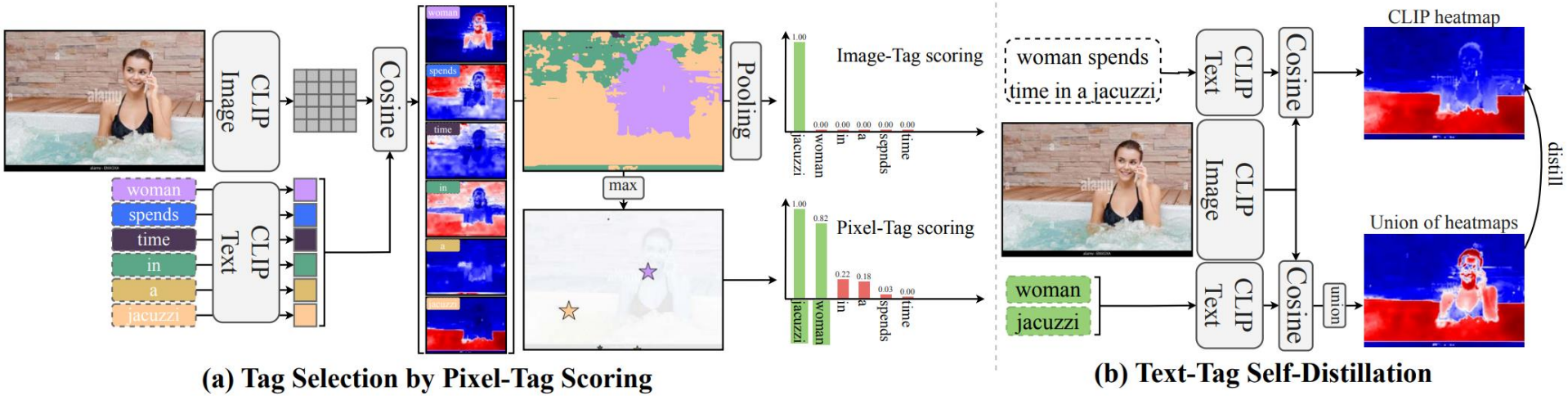


(b) **Single Tag Bias in Image-Tags Relationships**



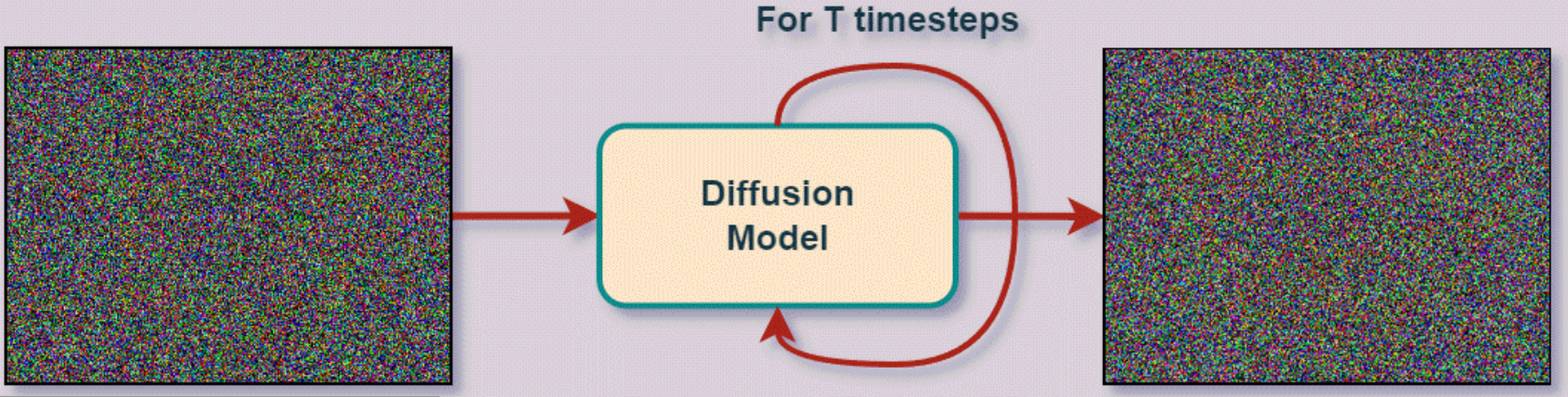
[ECCV 2024] TTD: Fixing the Single-Tag Blind Spot

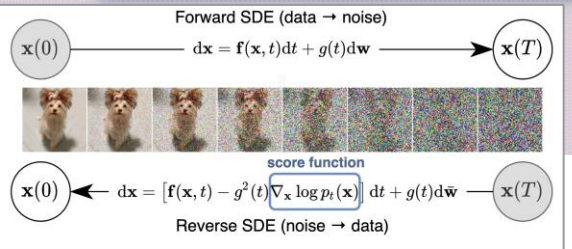
- CLIP tends to latch onto one dominant tag and ignore the rest.
- TTD self-distills text tags to restore full image-text alignment, improving multi-label understanding.



What Is a Diffusion Model?

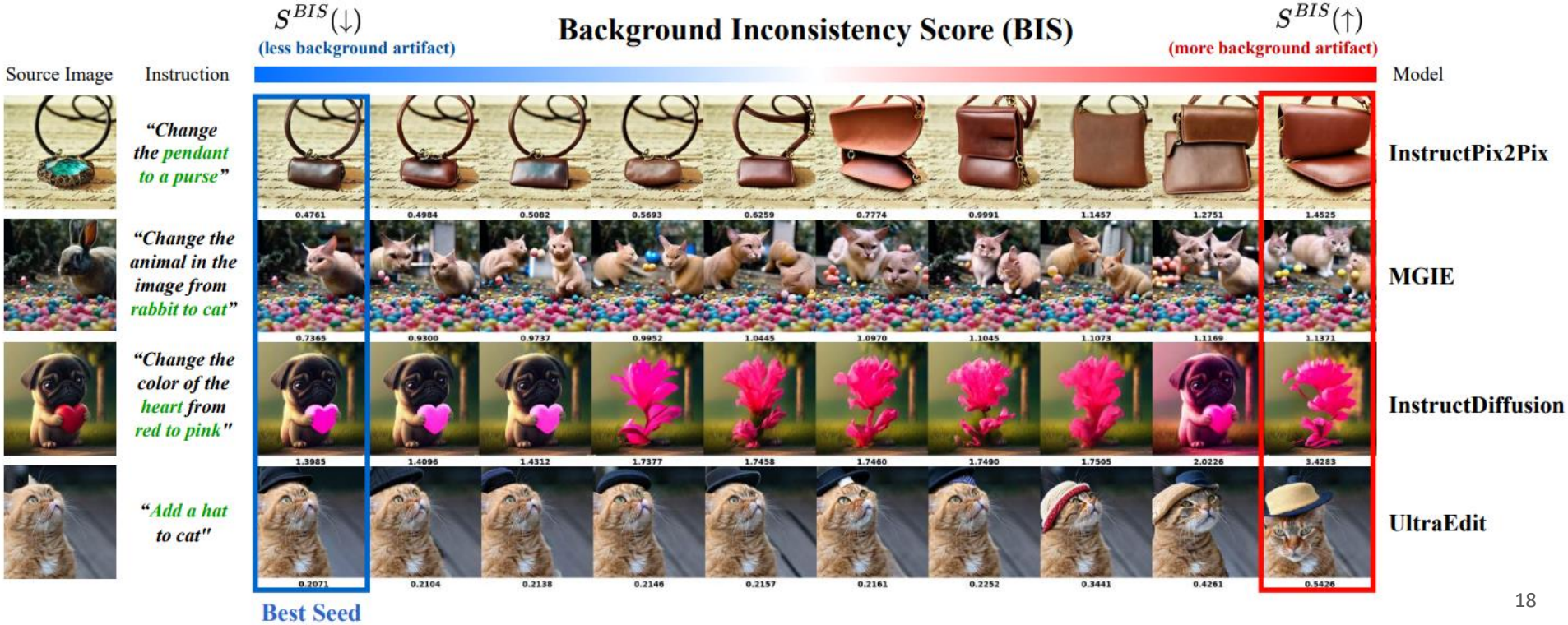
- A diffusion model turns pure noise into an image by removing noise step-by-step.
- The reverse process is not just for drawing. Its intermediate steps hold structures we can read & steer.





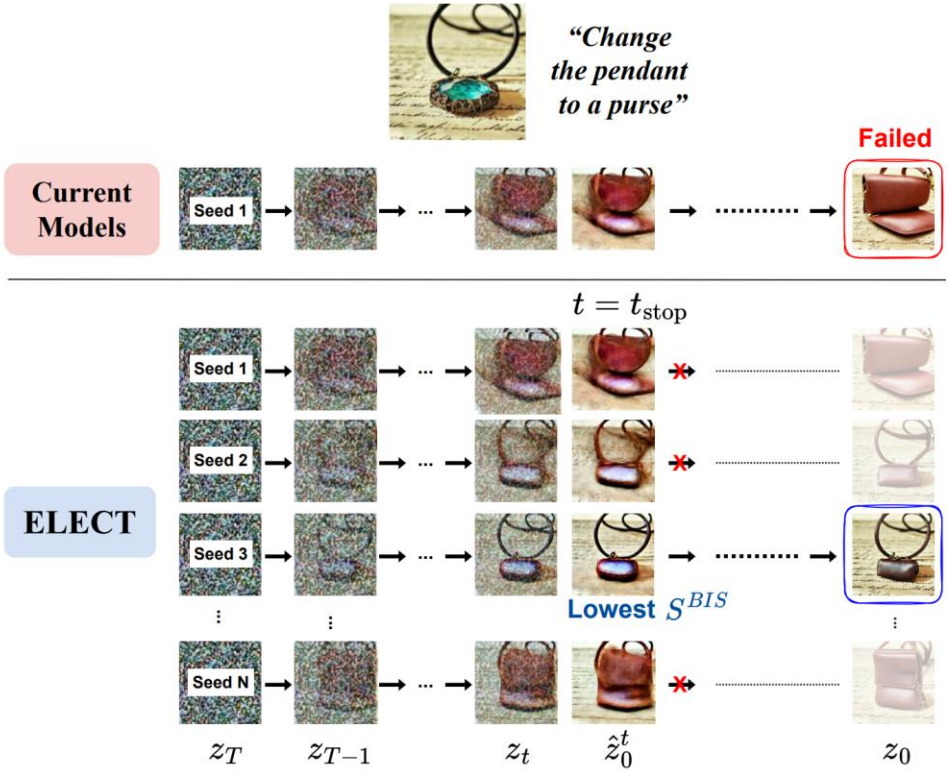
[ICCV 2025] ELECT: Choosing the Right Edit Early

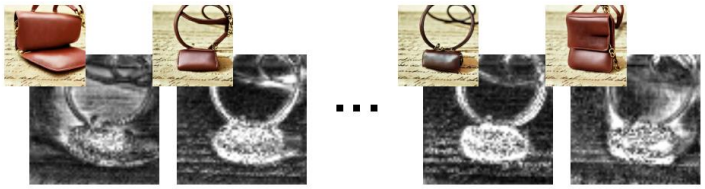
- For instruction-guided editing, ELECT picks the most promising candidate at an early denoising step.
- Zero-shot and no training required, saving significant compute on dead-end edits.



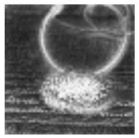
[ICCV 2025] ELECT: Choosing the Right Edit Early

- For instruction-guided editing, ELECT picks the most promising candidate at an early denoising step.
- Zero-shot and no training required, saving significant compute on dead-end edits.





$$M_t^i = |\epsilon_{cond} - \epsilon_{uncond}|$$
$$M_t^{mean} = \frac{1}{|\mathbb{S}|} \sum_{i \in \mathbb{S}} M_t^i$$



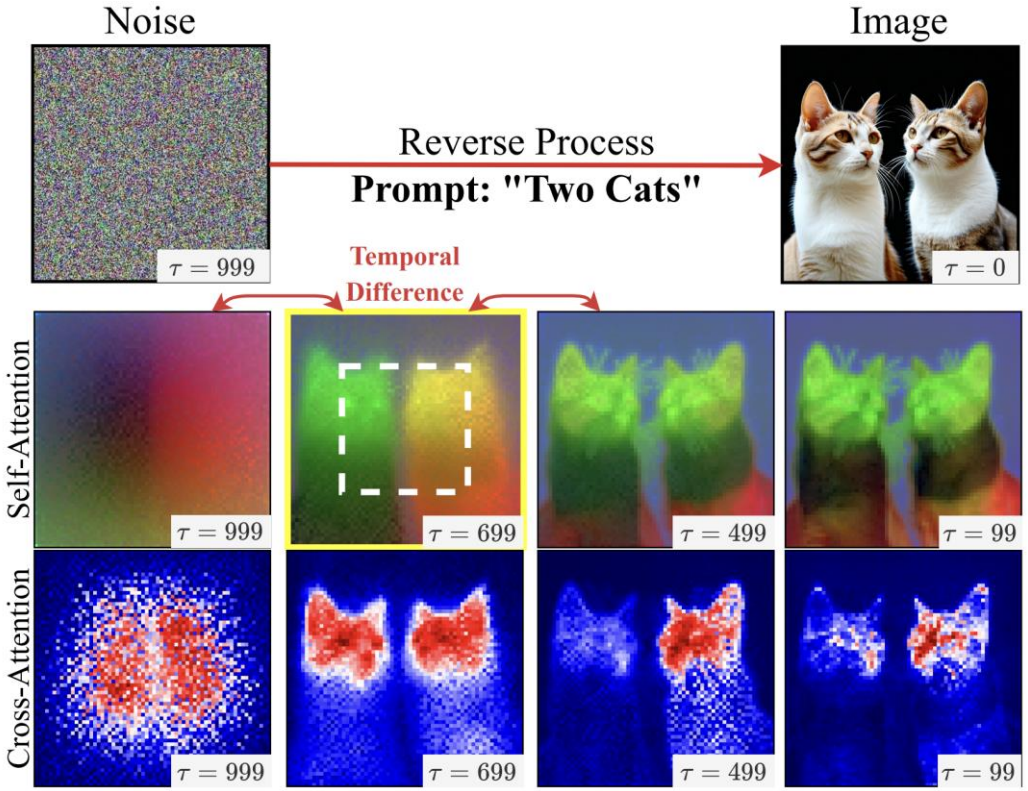
M_t^{mean}

$$S^{\text{BIS}}(i,t) := (1 - (M_t^{\text{mean}})^2) \odot |\hat{z}_0^i - \mathcal{E}(I)|$$

[ICLR 2026 Oral] TRACE: The Secret Life of Diffusion Models

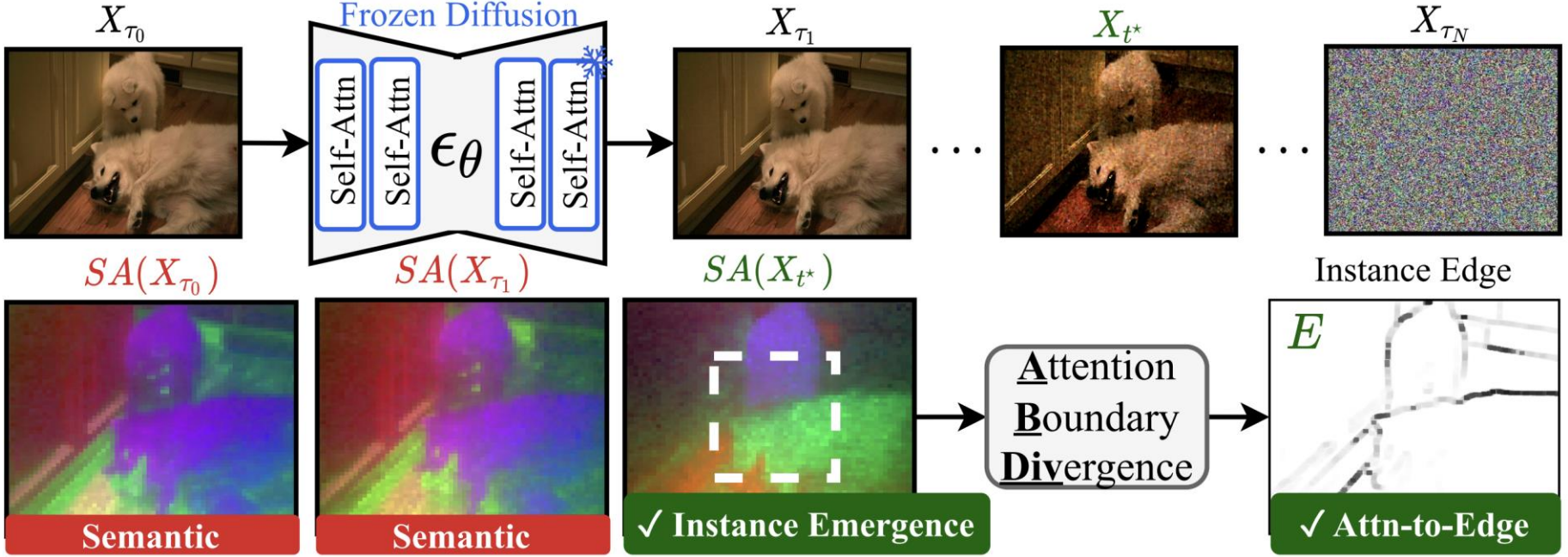
- Our hidden prior reveals that diffusion models naturally separate objects on their own during denoising steps.
- Boundary-centric separation extracts precise instance edges with zero human annotations.

ICLR 2026 Oral



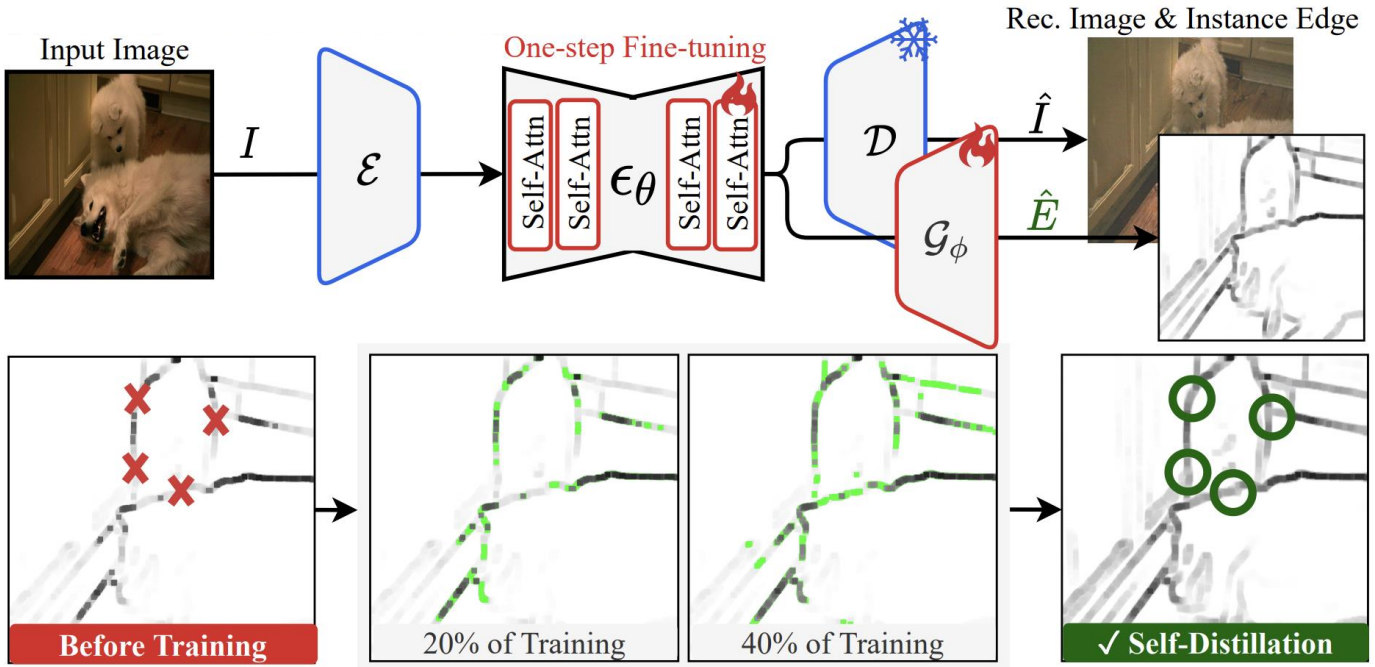
[ICLR 2026 Oral] TRACE: The Secret Life of Diffusion Models

- **IEP:** Tracks the forward process to pinpoint t*, the exact timestep where self-attention shifts from semantic blobs to sharp instance structures.
- **ABDiv:** Transforms this attention map into an initial edge map w/o requiring any annotations.



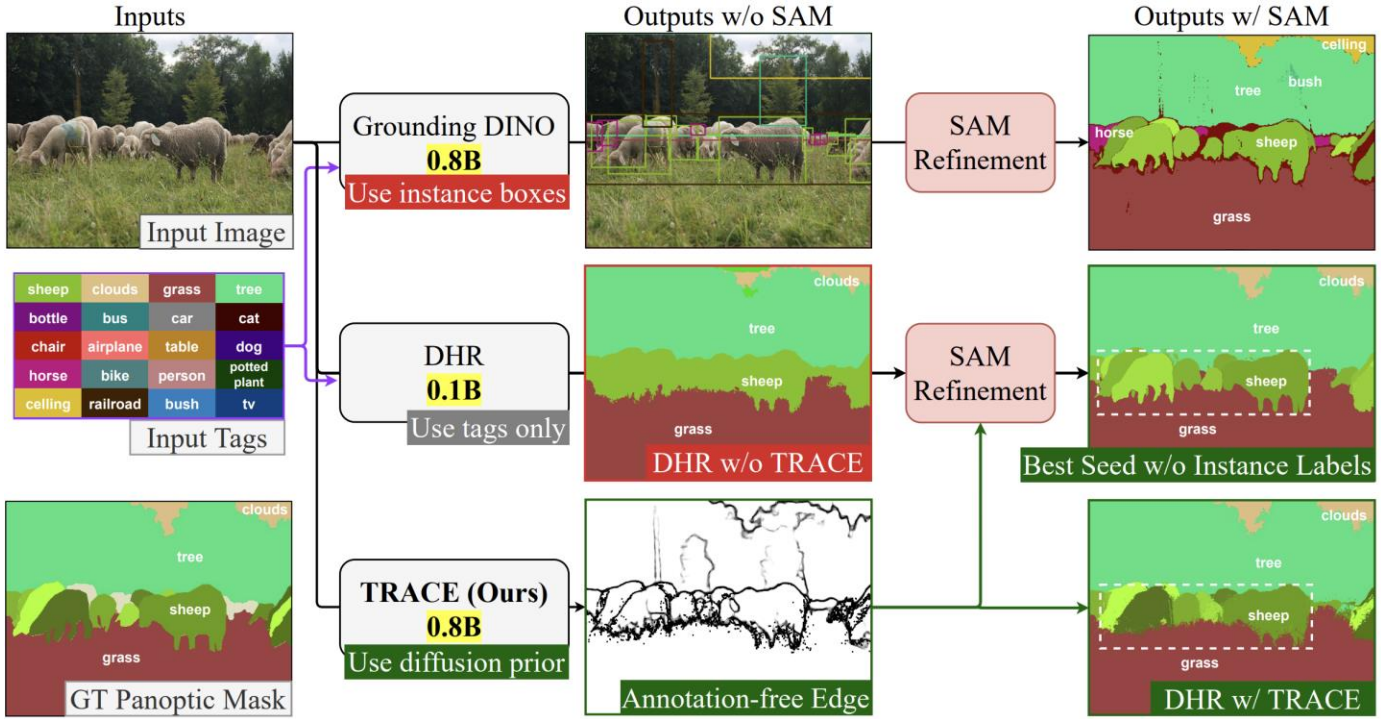
[ICLR 2026 Oral] TRACE: The Secret Life of Diffusion Models

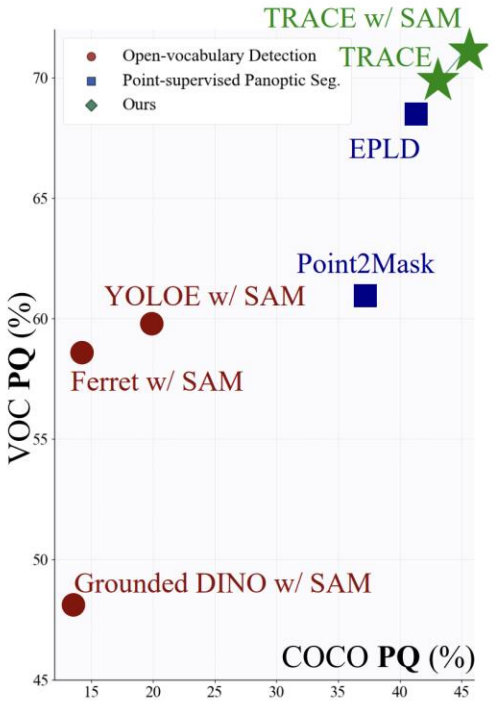
- **Self-Distillation:** Fine-tunes the diffusion backbone (via LoRA) alongside a lightweight edge decoder, using Stage 1 pseudo-edges as training targets.
- **Real-Time Inference:** Eliminates the iterative IEP search, slashing inference latency from over 3 seconds to just 45ms per image (an 81x speedup).



[ICLR 2026 Oral] TRACE: The Secret Life of Diffusion Models

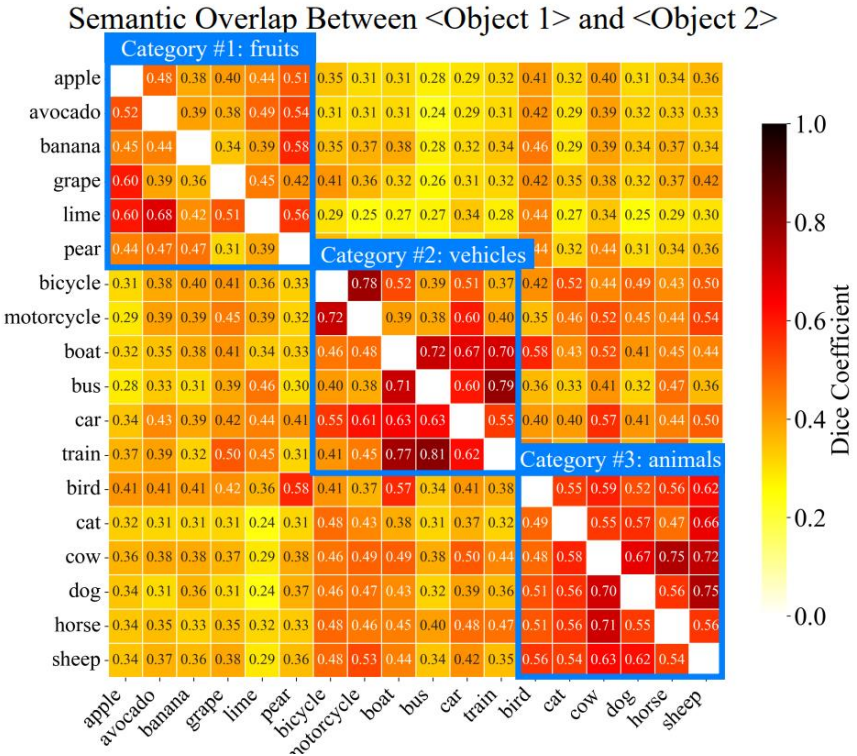
- **Annotation-Free Seeds:** Converts tag-supervised masks into panoptic masks.
- **Surpassing SOTA:** Outperforms point- and box-supervised methods w/o instance labels

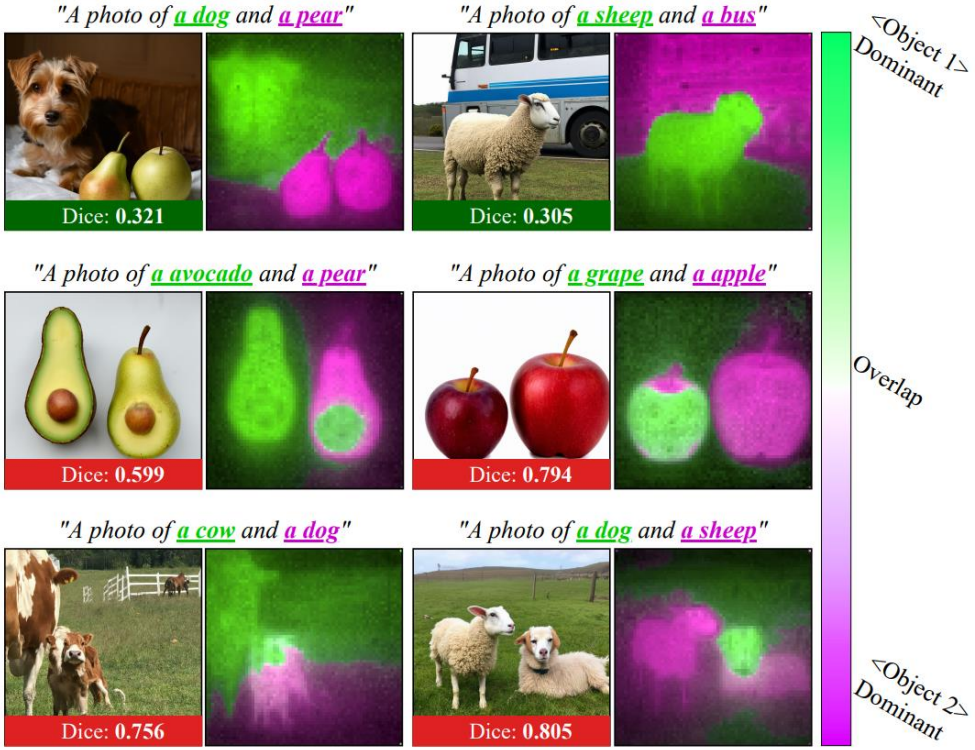




[Under Review] ISAC: Making Diffusion Count Correctly

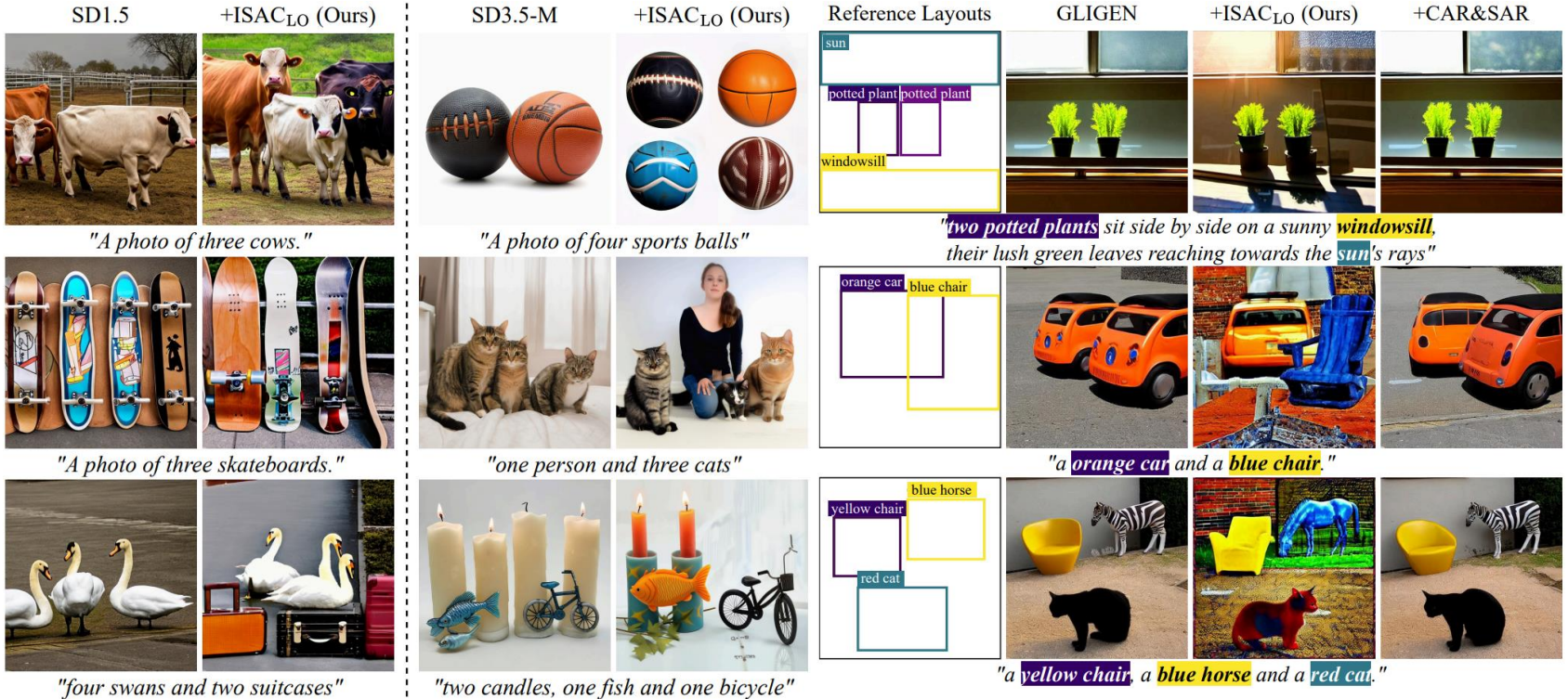
- Text-to-image diffusion often blurs multiple instances together.





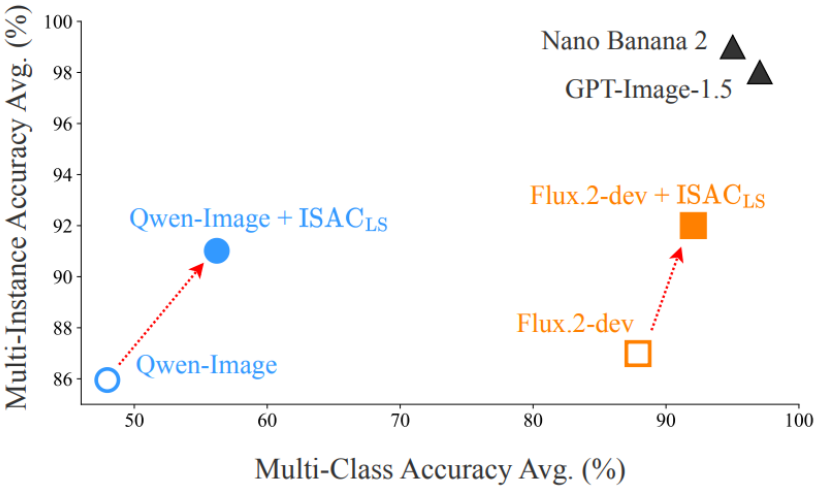
[Under Review] ISAC: Making Diffusion Count Correctly

- ISAC steers the model's internal attention from instances to semantics so each object renders distinctly.
- Training-free control for reliable multi-instance generation.

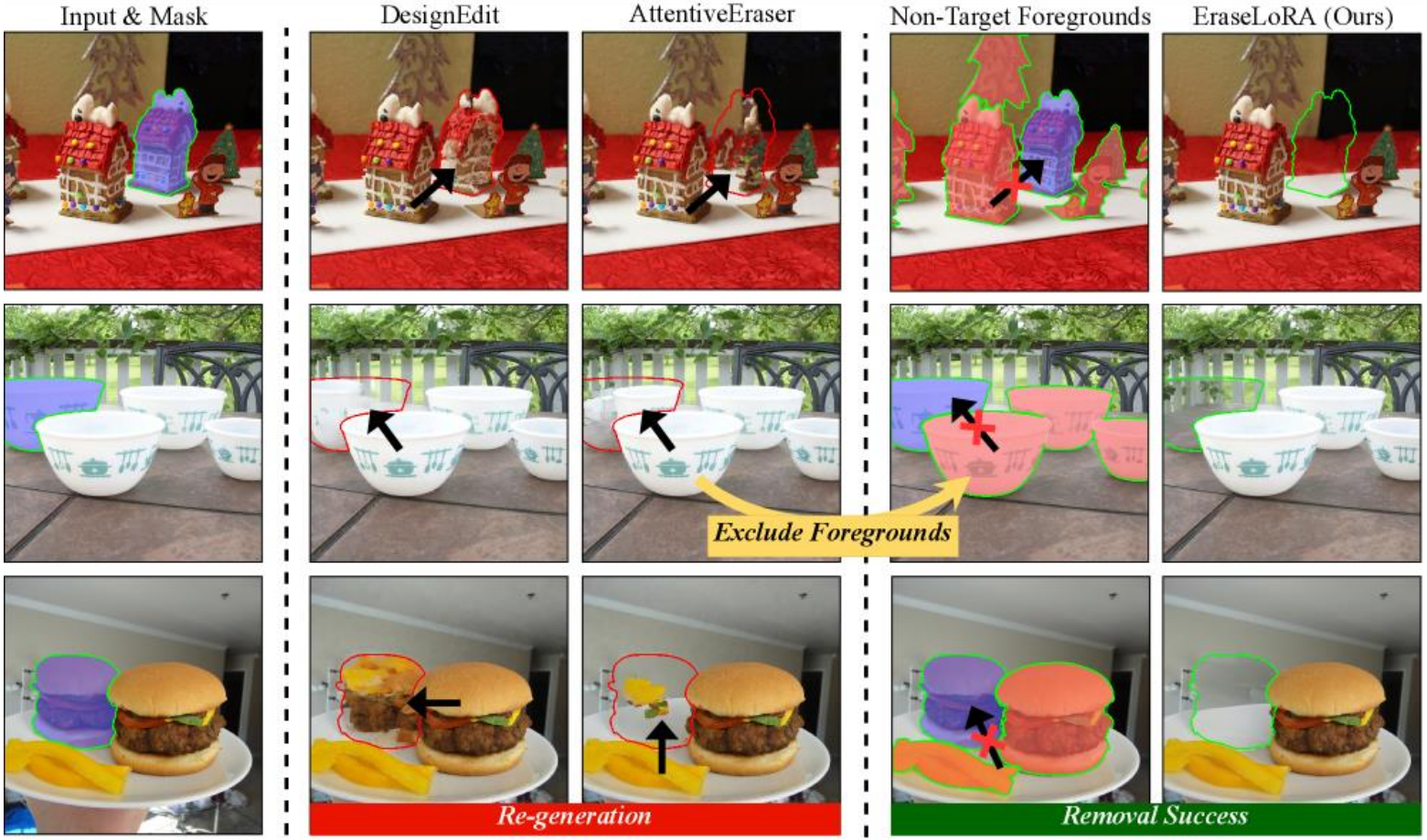


[Under Review] ISAC: Making Diffusion Count Correctly

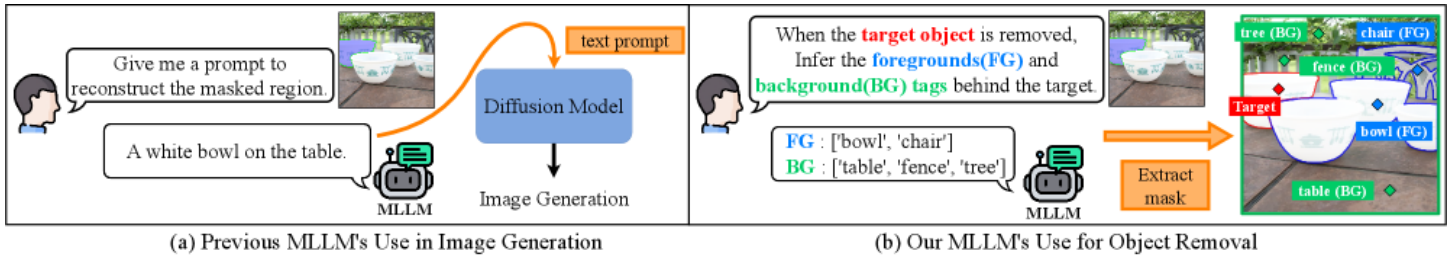
Method	Multi-Class Accuracy (↑)					Multi-Instance Accuracy (↑)					Efficiency (↓)	
	#2	#3	#4	#5	Avg.	#2	#3	#4	#5	Avg.	Latency	VRAM
Qwen-Image [78]	91%	45%	33%	10%	48%	98%	92%	84%	70%	86%	**140s**	**60.1GB**
+ **ISAC$_{LS}$ (Ours)**	**99%**	**58%**	**42%**	**25%**	**56%**	**99%**	**96%**	**89%**	**78%**	**91%**	210s	65.3GB
Flux.2-dev [42]	97%	95%	84%	78%	88%	**100%**	93%	81%	75%	87%	**205s**	**74.2GB**
+ **ISAC$_{LS}$ (Ours)**	**99%**	**98%**	**89%**	**83%**	**92%**	**100%**	**98%**	**88%**	**81%**	**92%**	305s	79.8GB
GPT-Image-1.5 [56]	99%	99%	98%	95%	97%	100%	100%	99%	94%	98%	N/A	N/A
Nano Banana 2 [23]	99%	97%	93%	92%	95%	100%	100%	100%	95%	99%	N/A	N/A



[Under Review] EraseLoRA: Erasing Objects Without a Dataset

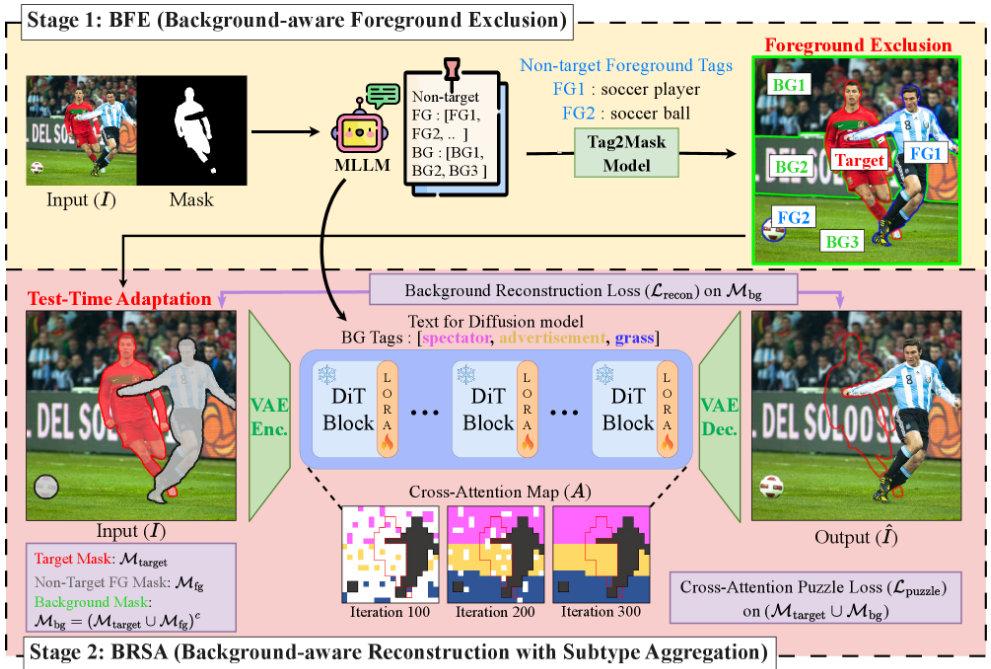


[Under Review] EraseLoRA: Erasing Objects Without a Dataset



(a) Previous MLLM's Use in Image Generation

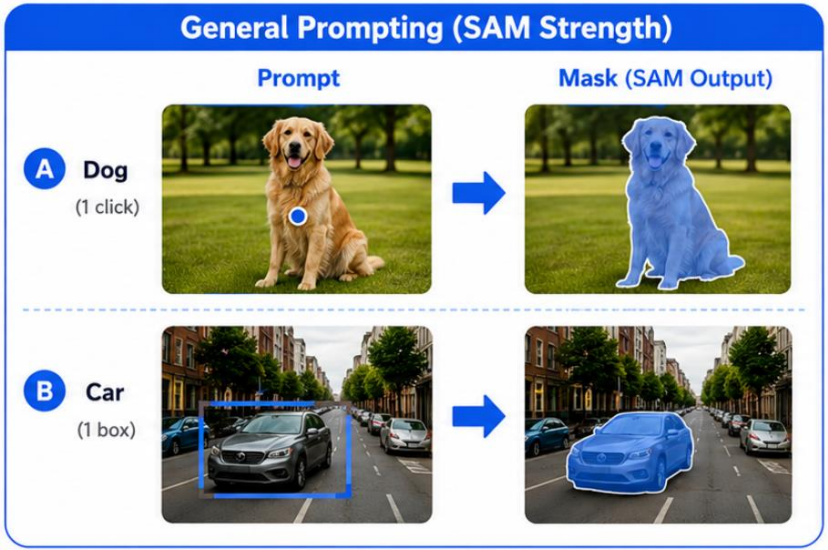
(b) Our MLLM's Use for Object Removal

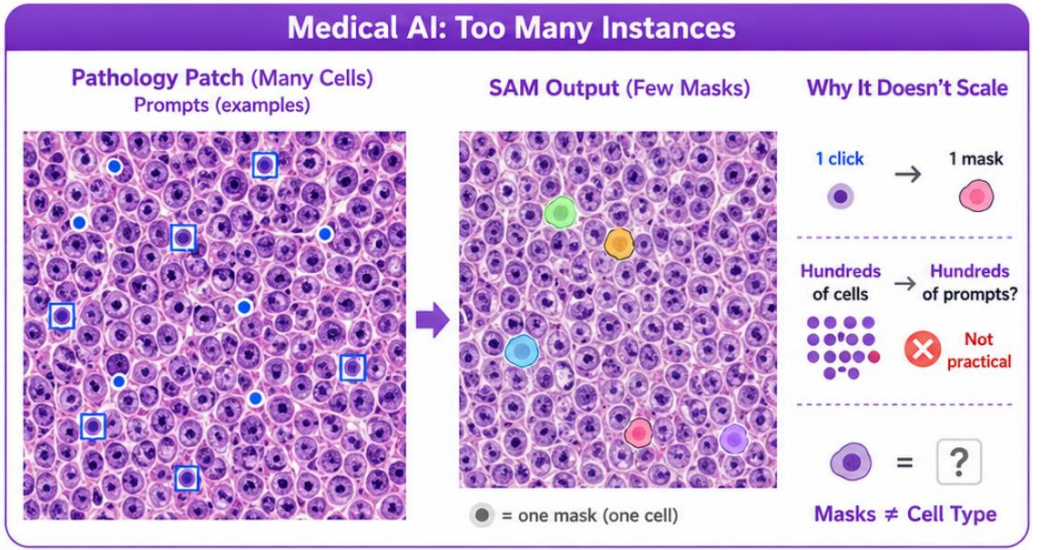


[Under Review] D2R / SOTA Object Composition

What Is SAM?

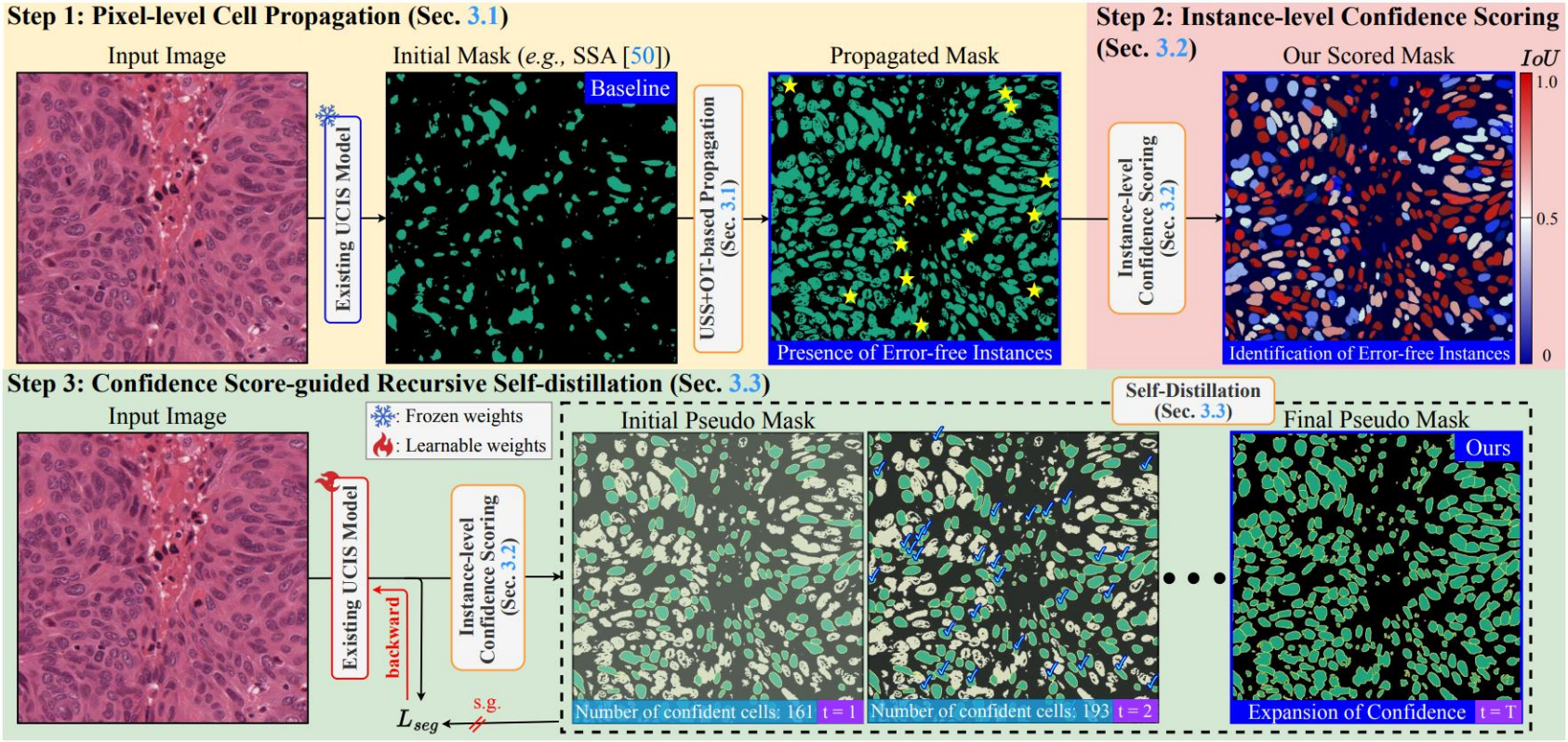
- Segment Anything Model segments almost any object from a simple prompt like a click or a box.





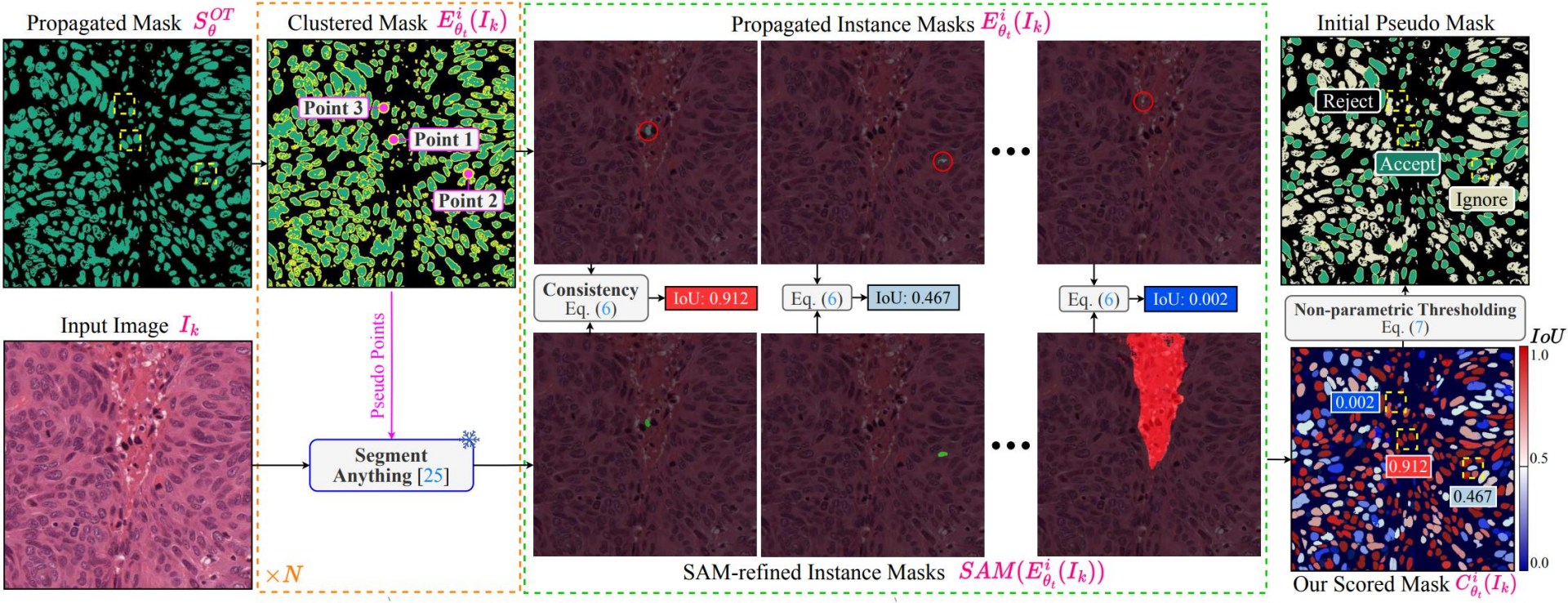
[ICCV 2025] COIN: Segmenting Cells Without Annotations

- COIN teaches a model to segment cells with no human labels.
- The foundation for bringing label-efficient vision into pathology.



[ICCV 2025] COIN: Segmenting Cells Without Annotations

- COIN teaches a model to segment cells with no human labels.
- The foundation for bringing label-efficient vision into pathology.



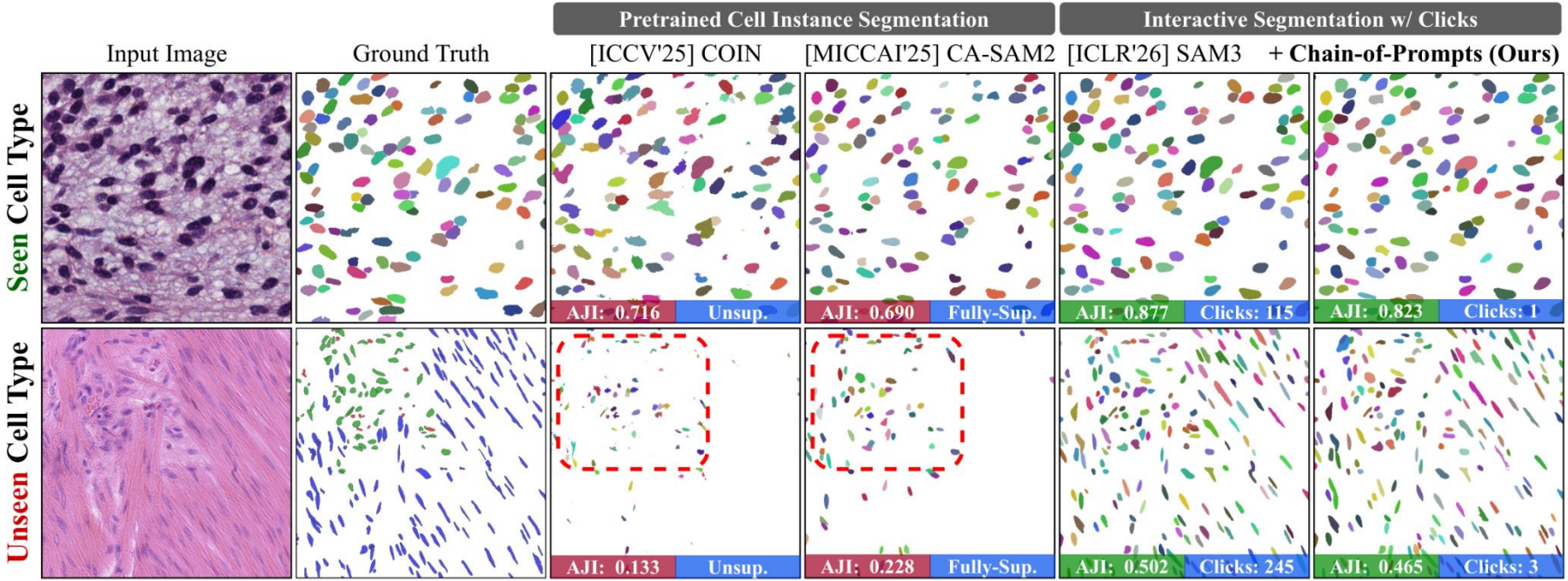
[ICCV 2025] COIN: Segmenting Cells Without Annotations

- COIN teaches a model to segment cells with no human labels.
- The foundation for bringing label-efficient vision into pathology.

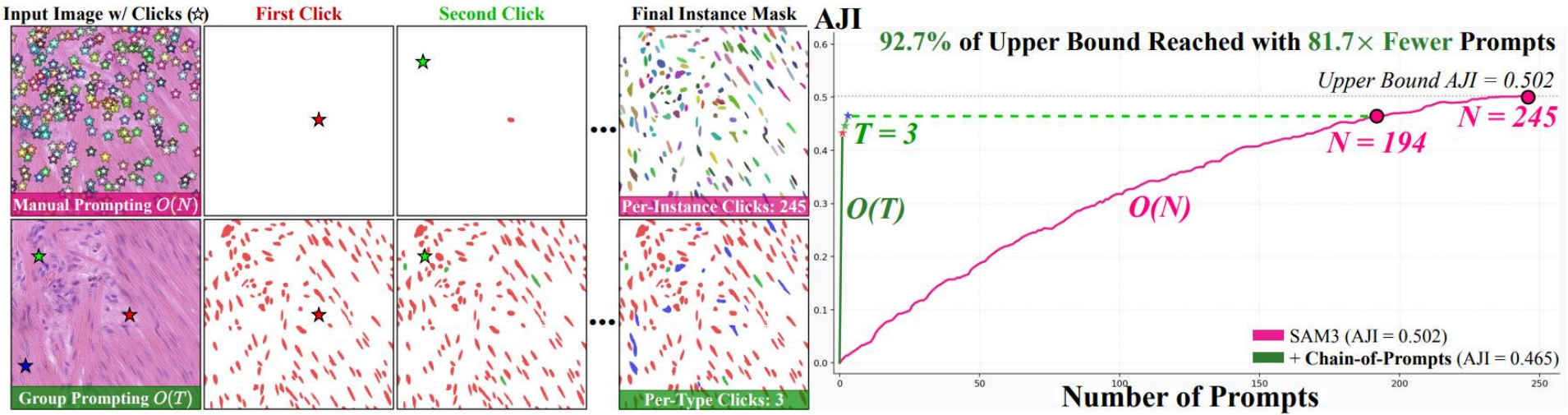
Method	Cell Supervision	MoNuSeg				TNBC			
		AJI (↑)	PQ (↑)	IoU (↑)	Dice (↑)	AJI (↑)	PQ (↑)	IoU (↑)	Dice (↑)
Annotation-free Instance Segmentation									
MaskCut [59] CVPR'23	✗	0.001*	0.000*	0.053*	0.089*	0.006*	0.000*	0.048*	0.088*
CutLER [59] CVPR'23	✗	0.002*	0.000*	0.143*	0.244*	0.003*	0.000*	0.082*	0.146*
ProMerge [32] ECCV'24	✗	0.000*	0.000*	0.013*	0.024*	0.004*	0.000*	0.046*	0.076*
Annotation-free Cell Instance Segmentation									
CellProfiler [4] Genome Biology'06	✗	0.123	-	-	0.404	0.208	-	-	0.415
Fiji [51] Nature Methods'12	✗	0.273	-	-	0.665	-	-	-	-
Hou *et al.* [18] CVPR'19	✗	0.498	-	-	0.750	-	-	-	-
SSA [50] MICCAI'20	✗	0.259*	0.185*	0.618*	0.575*	0.273*	0.253*	0.647*	0.538*
SSA + COIN (Ours)	✗	**0.580**	**0.536**	**0.776**	**0.794**	**0.568**	**0.540**	**0.797**	**0.774**
PSM [5] MICCAI'23	✗	0.471*	0.355*	0.689*	0.682*	-	-	-	-
PSM + COIN (Ours)	✗	**0.579**	**0.539**	**0.777**	**0.797**	-	-	-	-
Weakly-supervised Cell Instance Segmentation									
Qu *et al.* [46] MIDL'19	Point	0.496	-	-	0.702	-	-	-	-
C2FNet [53] MICCAI'20	Point	0.493	-	0.624	-	-	-	-	-
Mixed Anno [47] ISBI'20	Point & Mask	0.516	-	-	0.733	-	-	-	-
BB-WSIS [58] MICCAI'21	Box	-	-	-	0.728	-	-	-	0.703
Liu *et al.* [35] ISBI'22	Point	0.534	-	-	0.740	-	-	-	-
SPPNet [63] MLMI'23	Point	0.497*	0.392*	0.709*	0.719*	-	-	-	-
All-in-SAM [8] IOPscience'23	Box	0.502	-	-	0.738	-	-	-	-
PROnet [38] MICCAI'23	Point	0.555	-	-	0.750	-	-	-	-
InstaSAM [39] MICCAI'24	Point	0.574	-	-	0.772	-	-	-	-
Semi-supervised Cell Instance Segmentation									
CDCL [61] CVPR'22	Mask	-	-	-	0.782	-	-	-	-
TextDiff [11] MICCAI'24	Mask & Text	0.510*	0.410*	0.726*	0.726*	0.464*	0.358*	0.728*	0.666*

[MICCAI 2026 Early Accept] One Click per Cell Type Suffices

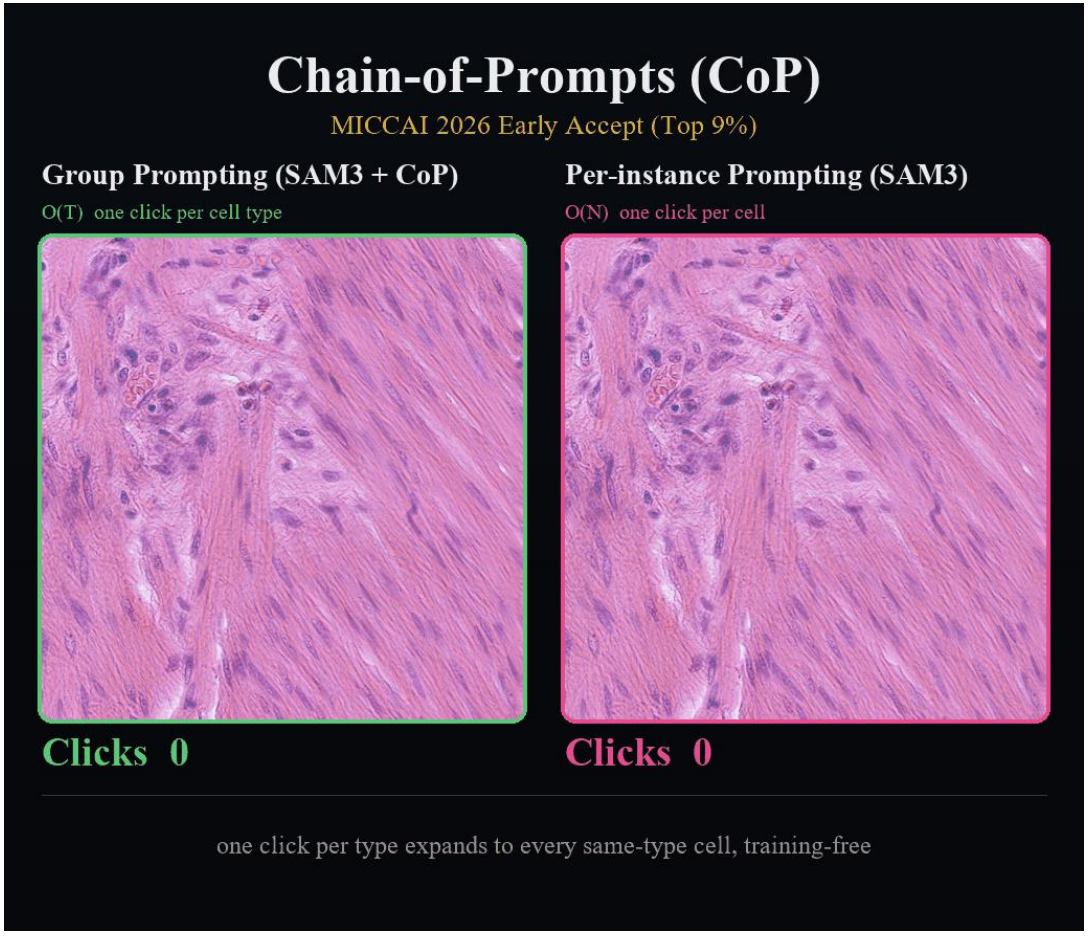
- A pathologist gives a single click per cell type. Training-free group interaction turns it into dense cell instance masks.
- The label-efficient idea completed. We evolved from tags to zero labels and finally to one click.



[MICCAI 2026 Early Accept] One Click per Cell Type Suffices



[MICCAI 2026 Early Accept] One Click per Cell Type Suffices

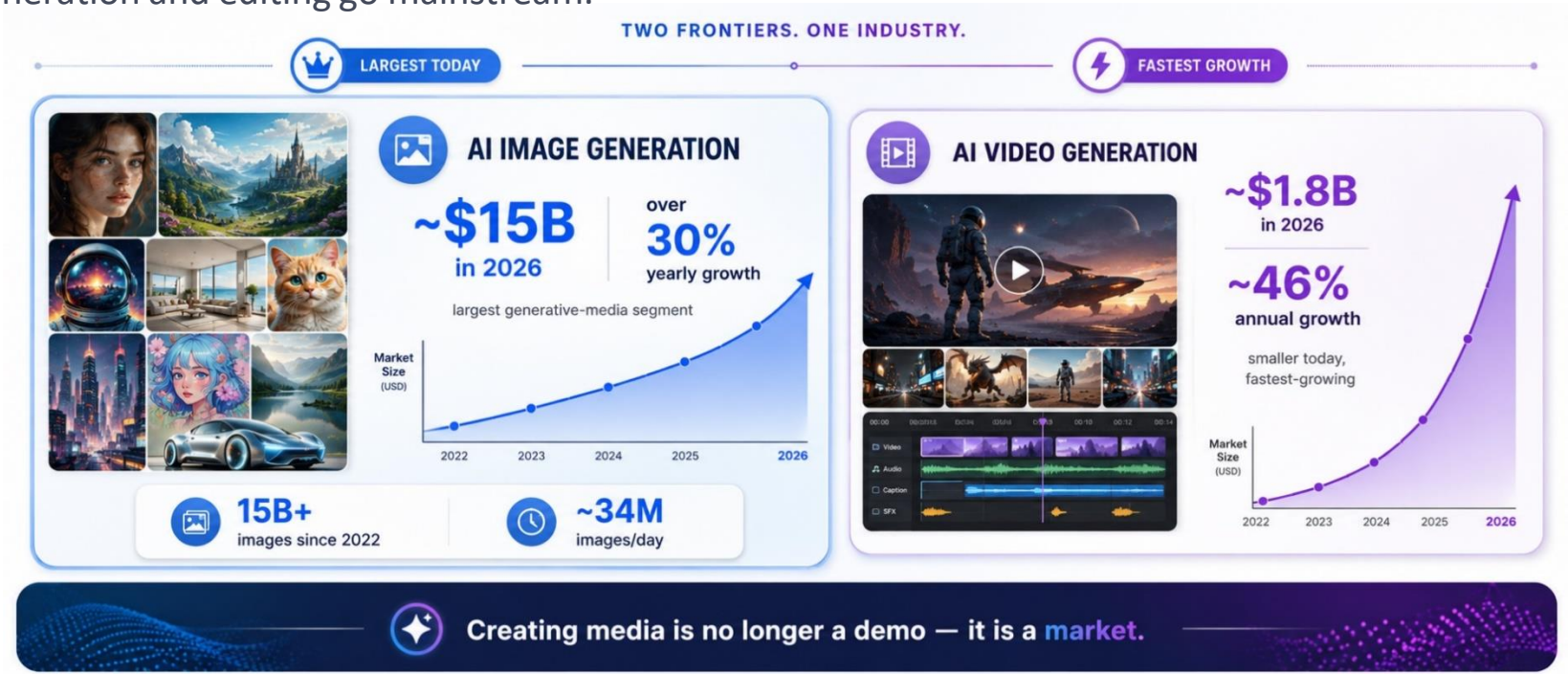


The Generator Steps Outside the Lab

- Having just explored the hidden workings of generative models, we now see these same models already redrawing entire industries out in the real world.
- Vision AI is changing the world along two key frontiers: creating media and modeling reality itself.

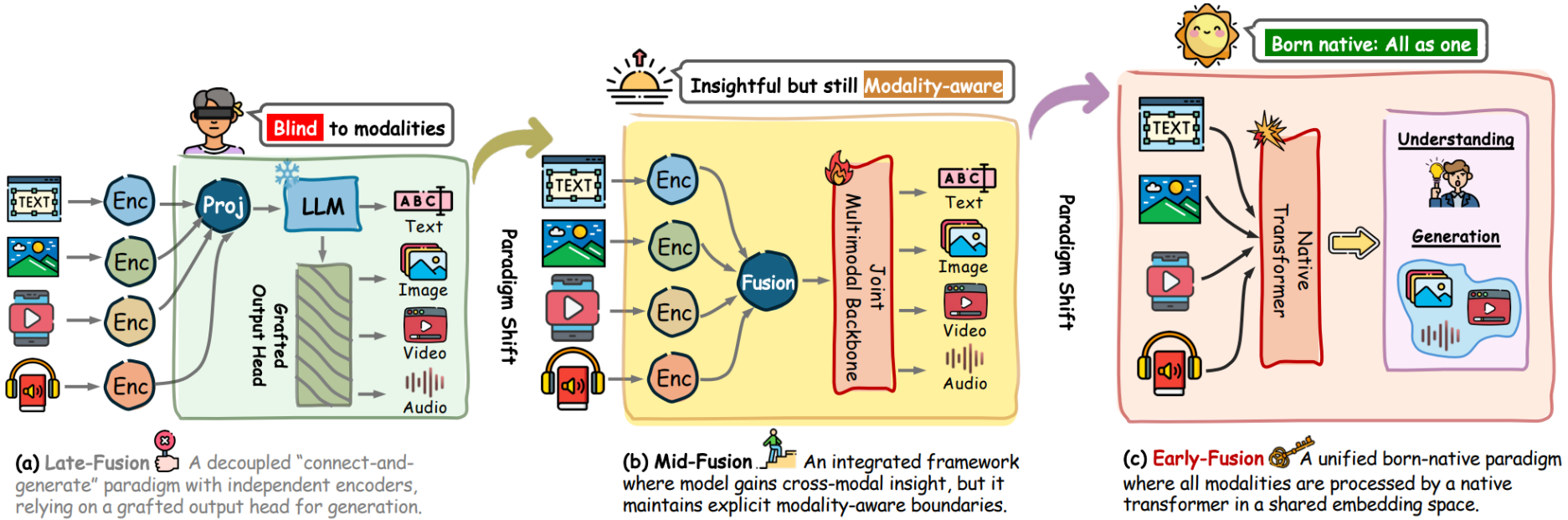
Generating Reality Is Now an Industry

- AI image generation is already the largest generative-media segment, around USD 15 billion in 2026 and growing over 30% a year.
- AI video is smaller but fastest-growing, multiplying several times this decade as photorealistic generation and editing go mainstream.



How One Model Does It All

- The old way bolted a vision encoder onto a language model; the backbone stayed blind to raw pixels.
- A single transformer handles every modality in one shared space from the start, a born-native model.



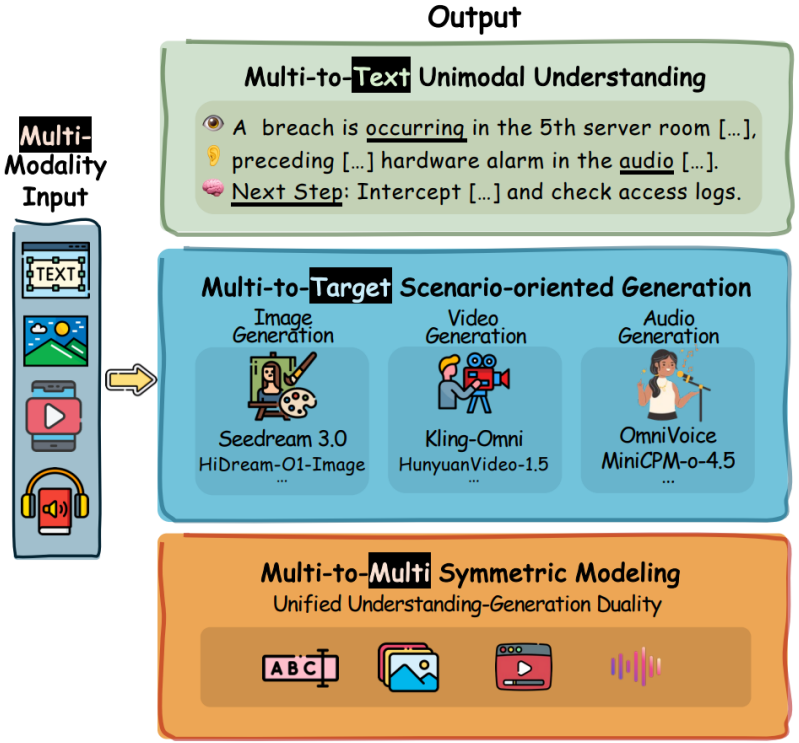
(a) Late-Fusion A decoupled "connect-and-generate" paradigm with independent encoders, relying on a grafted output head for generation.

(b) Mid-Fusion An integrated framework where model gains cross-modal insight, but it maintains explicit modality-aware boundaries.

(c) Early-Fusion A unified born-native paradigm where all modalities are processed by a native transformer in a shared embedding space.

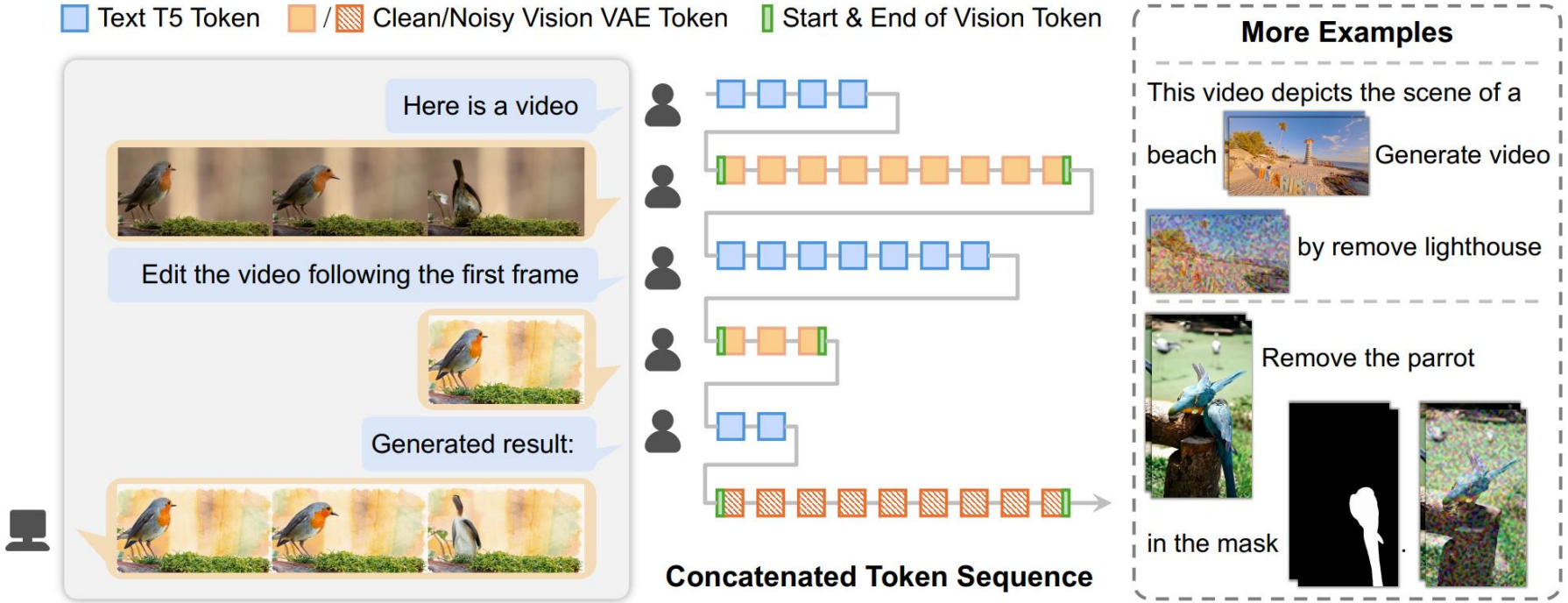
Understanding and Generation as One

- **Three regimes:** read inputs into text (M2T), generate a target like image or video (M2G), or do both symmetrically (M2M).
- The endpoint of M2M is a single network where understanding and generation coexist.



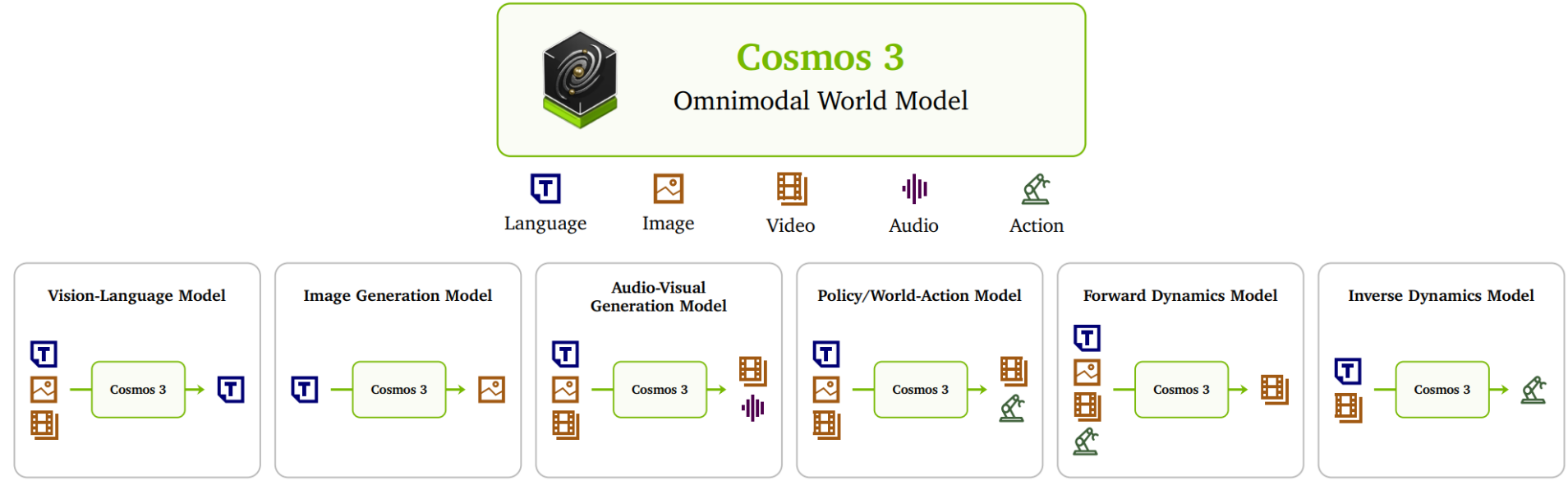
Understanding and Generation as One

- **[ICLR 2026 Oral]** EditVerse: Unifying Image and Video Editing and Generation with In-Context Learning

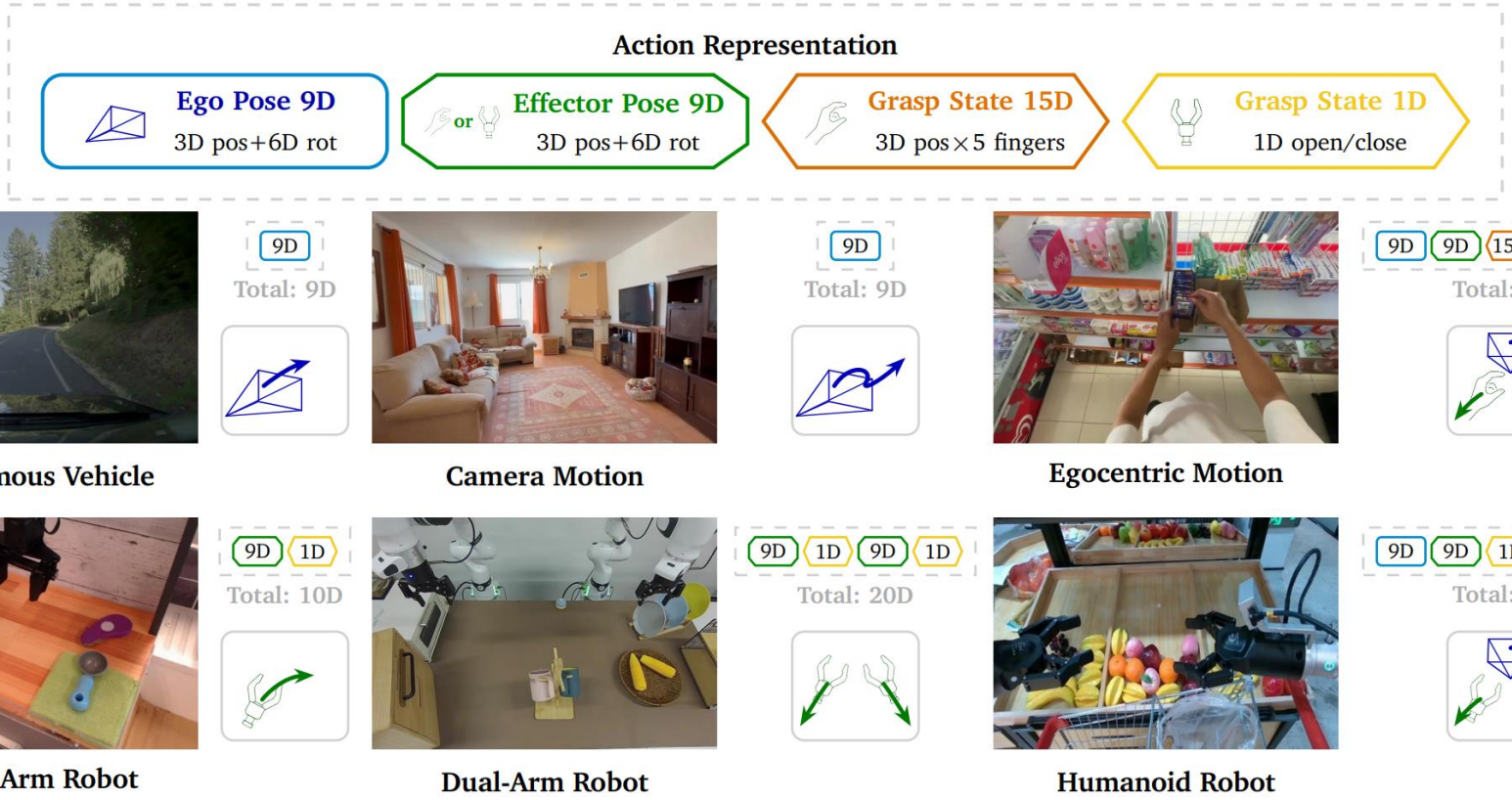


From Images to Worlds

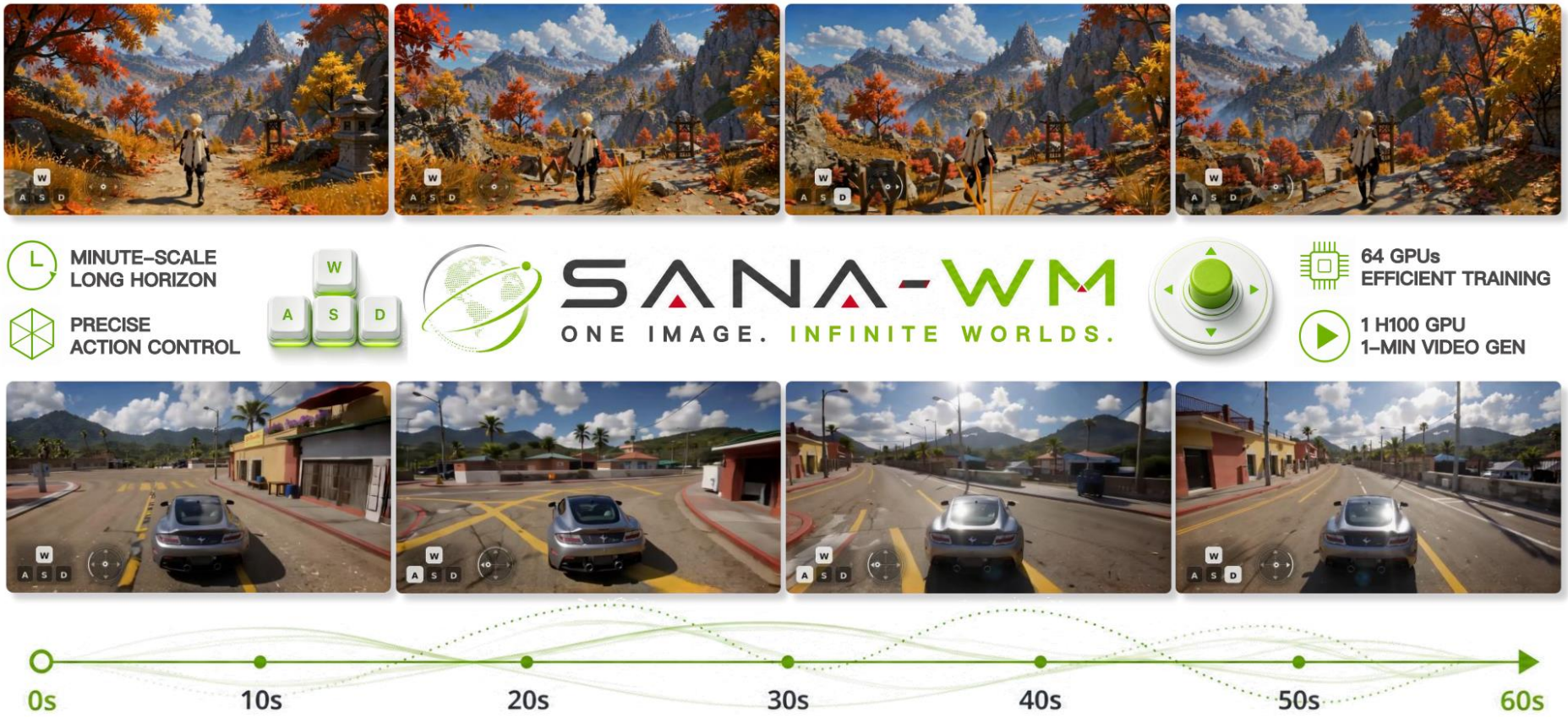
- Beyond single images and clips, world models generate physically plausible worlds and predict what happens next.
- They train robots and self-driving cars in simulation, cutting development from months to days.



From Images to Worlds

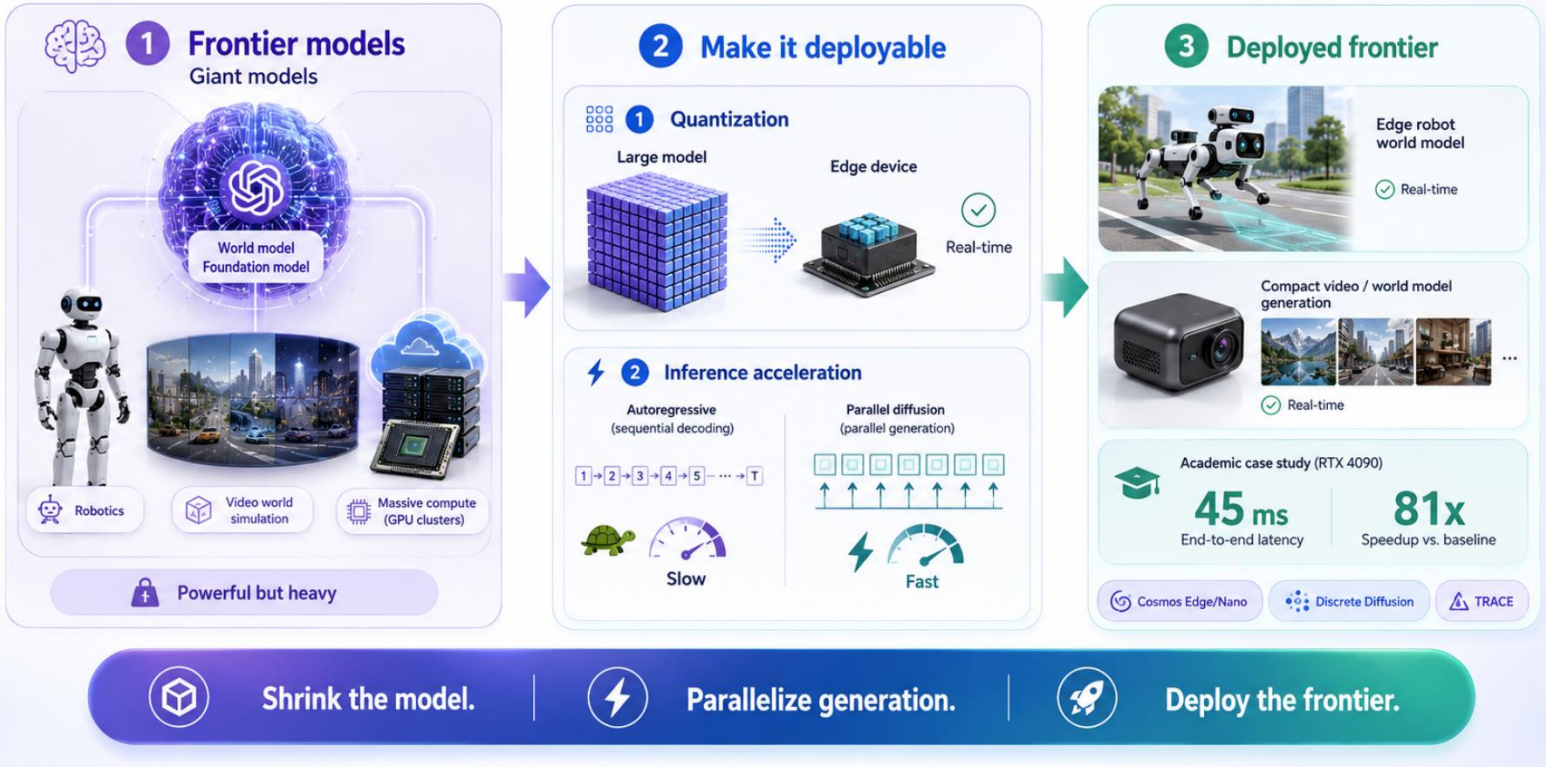


From Images to Worlds



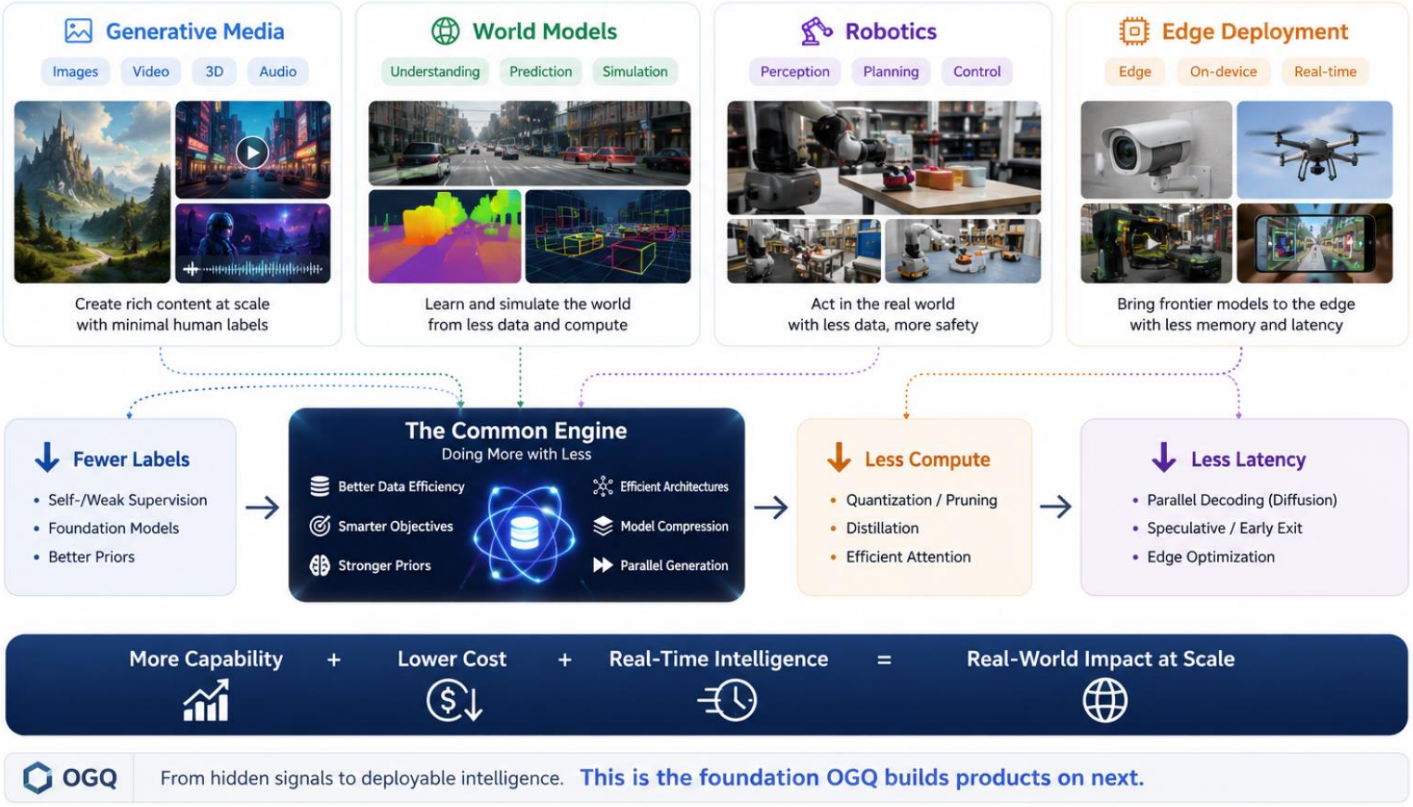
Making the Frontier Deployable

- Quantization shrinks giant models for edge devices, bringing world models to real-time inference.
- Moving generation from slow autoregressive decoding toward parallel diffusion for large speedups.



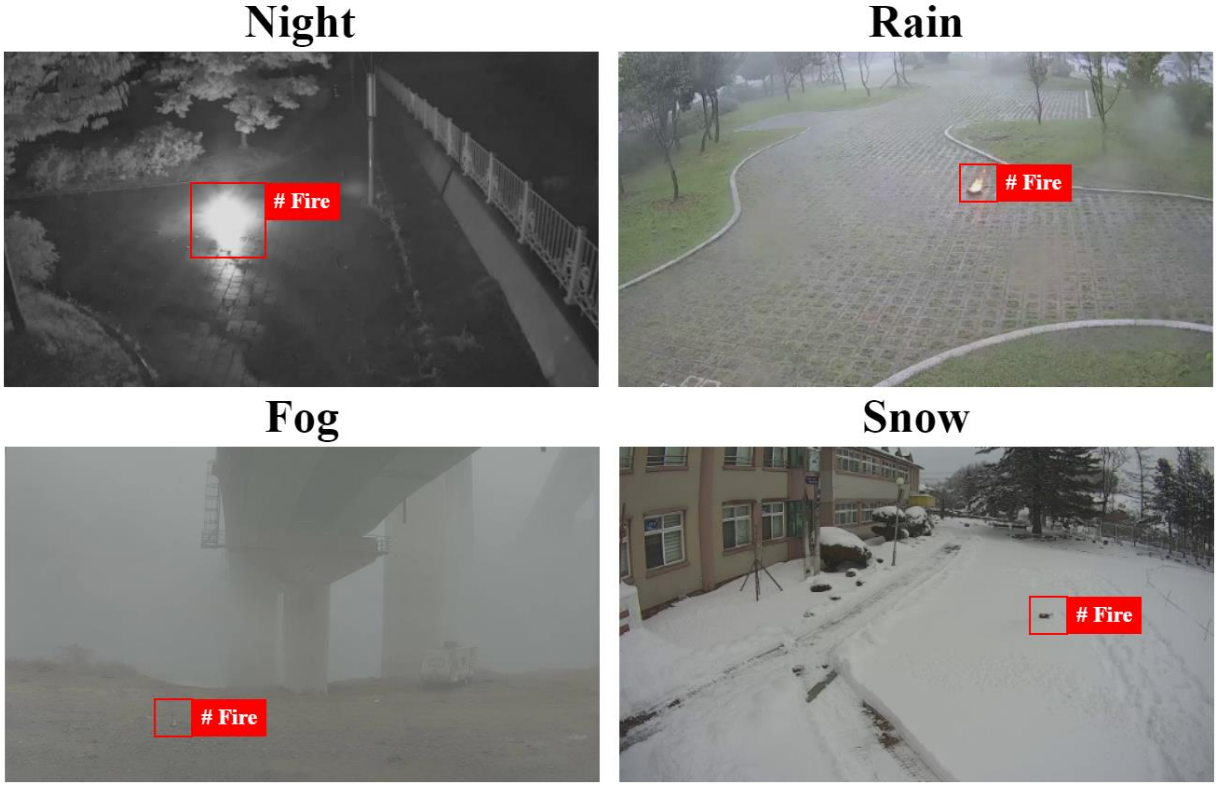
The Common Engine: Doing More with Less

- Media, world models, robots, edge: all advance by needing fewer labels, less compute, less latency.
- Now redrawing the frontier, and the foundation OGQ builds products on next!



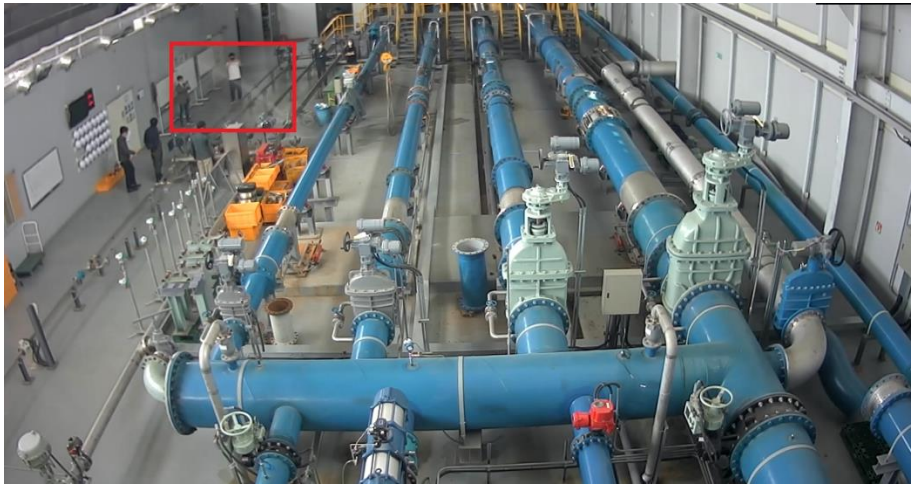
Private & On-Device: Vision for Public Safety

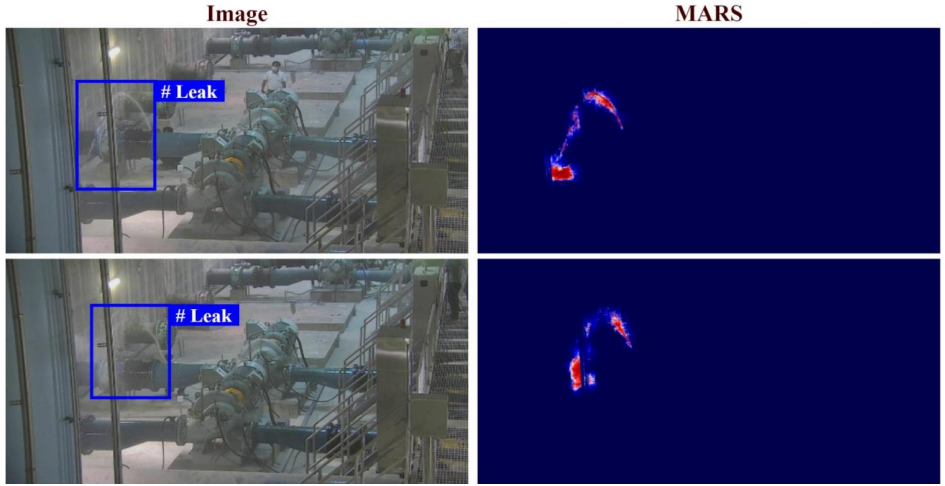
- **Robust Field Detection:** Validated on real-world data to accurately detect leaks, fires, and structural collapses across extreme conditions, including night, rain, fog, and snow.



Private & On-Device: Vision for Public Safety

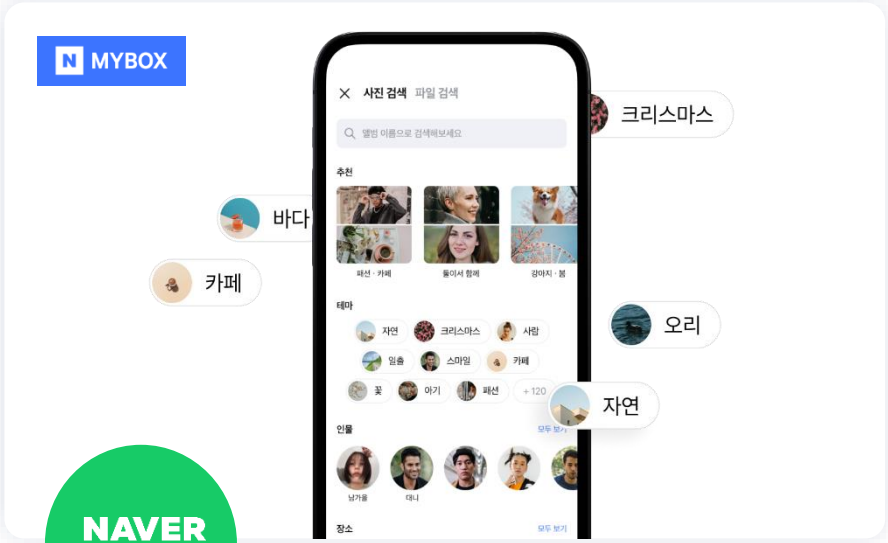
- **Privacy-Preserving & On-Device Deployment:** Rolled out with public partners (K-water, IITP, and police drone programs). Powered by our latest TRACE research to ensure high performance without compromising data privacy.

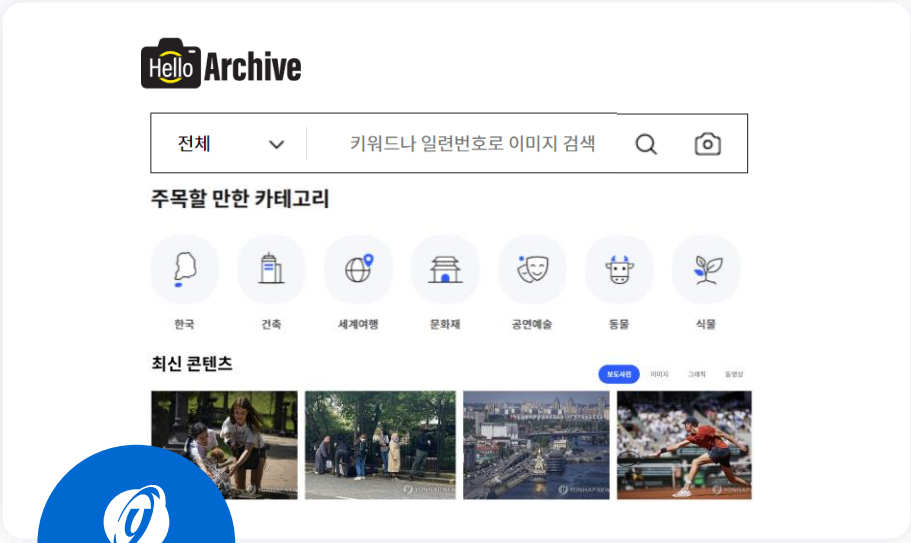




Content Search: Making Assets Findable

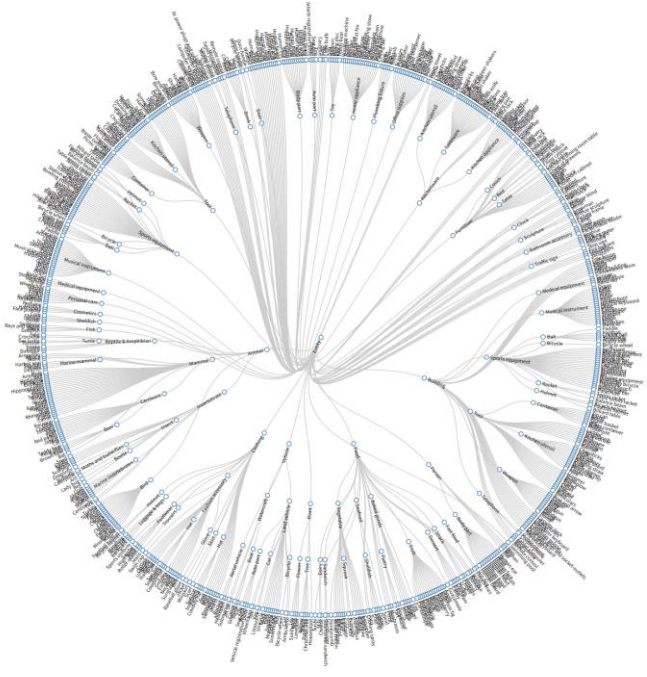
- **Semantic Auto-Tagging:** Transforming massive asset catalogs to be searchable by their underlying meaning, moving beyond traditional filename searches.
- **Powering Partner Ecosystems:** Driving the core content search engine for major platforms like Naver MYBOX, directly leveraging our foundational recognition research.





Content Search: Making Assets Findable

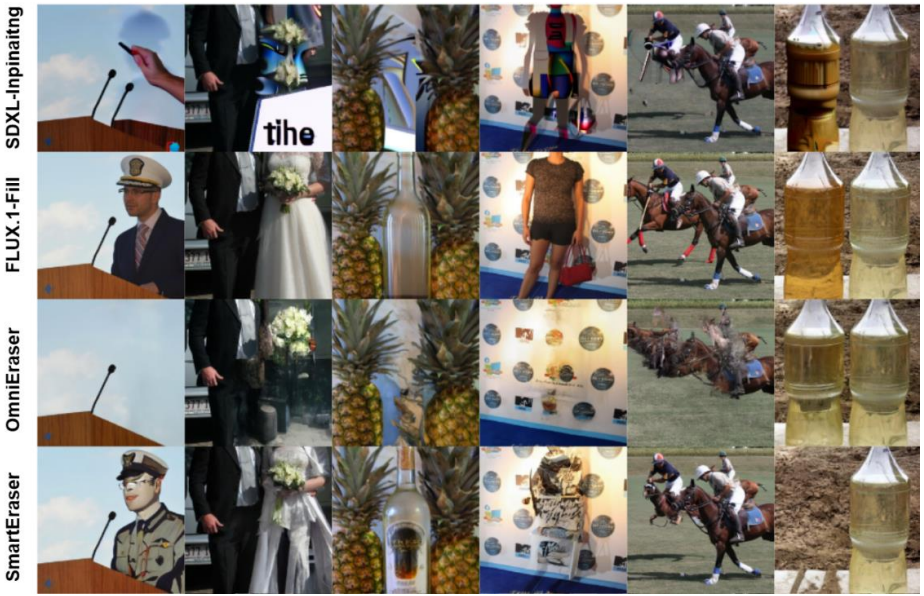
- **Semantic Auto-Tagging:** Transforming massive asset catalogs to be searchable by their underlying meaning, moving beyond traditional filename searches.
- **Powering Partner Ecosystems:** Driving the core content search engine for major platforms like Naver MYBOX, directly leveraging our foundational recognition research.

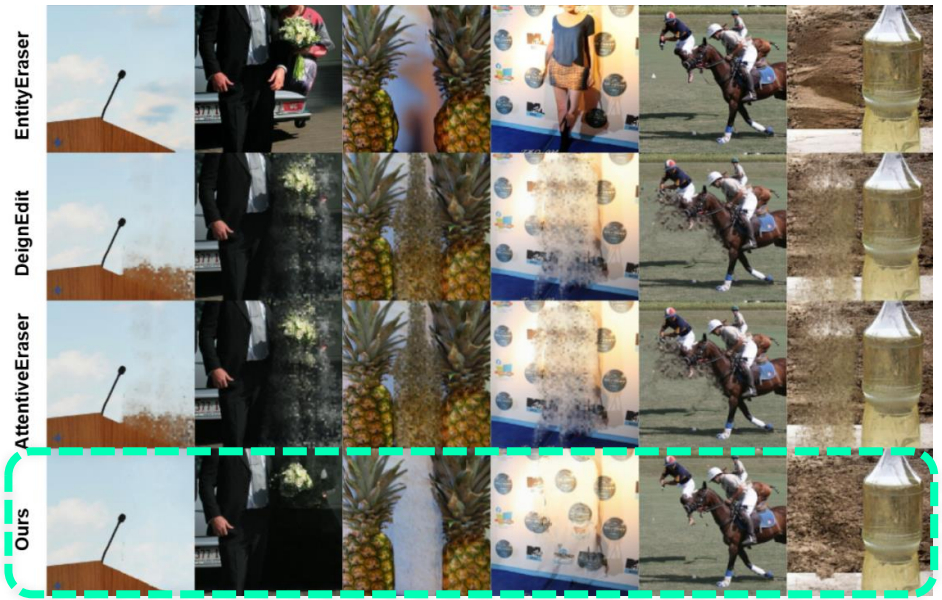


tag_id	en	checl	테마명	ko1	ko2
1	plant			식물	공장
2	person	t	사람	사람	
3	sky	t	하늘	하늘	
4	flower	t	꽃	꽃	
5	food	t	음식	음식	
6	tree	t	나무	나무	
7	font	s		폰트	
8	illustration	s		삽화	
9	nature	t	자연	자연	
10	natural landscape	s		자연 경관	
11	cuisine	s		요리법	요리
12	flowering plant	s		꽃식물	꽃 피는 식물
13	dish	t	접시	접시	
14	ingredient			재료	
15	architecture			건축	건축학
16	photography	s		사진술	사진
17	logo	s		로고	
18	graphics	s		삽화	
19	text	t	텍스트	본문	텍스트
20	building	t	건물	건축물	
21	pink	t	분홍색	분홍	분홍색
22	product	t	제품	제품	

Secondary Creation: Editing for Creators

- **Empowering Derivative Works:** Equipping creators with intuitive prompt-based editing and intelligent object removal tools to accelerate secondary creation.
- **Research-Backed Innovation:** Delivered through national R&D initiatives (NIPA) and built directly upon our core Diffusion and EraseLoRA.

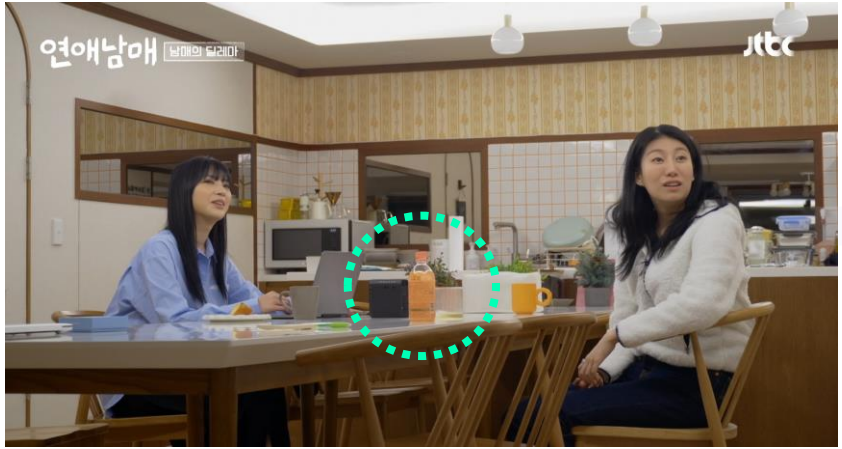


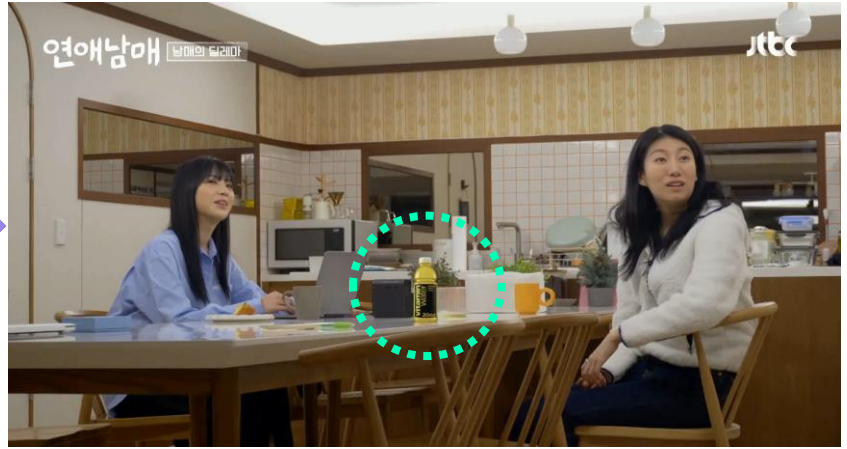


Secondary Creation: Editing for Creators

w/o AI PPL

w/ OGQ AI PPL





Secondary Creation: Editing for Creators

w/o Removal

w/ Removal

Secondary Creation: Editing for Creators

- **ONAIR:** Broadcast footage with overlaid subtitles, captions, and visual effects.
- **CLEAN:** Original footage before post-production overlay.

Input Video

Output Video

The Academia ↔ Industry Flywheel

- **Real-World Constraints as Catalysts:** Industrial challenges, such as labeling costs, deployment speeds, and on-device limits, directly fuel our research agenda.
- **The Continuous Innovation Loop:** These practical challenges inspire top-tier academic papers like TRACE and CoP, which immediately flow back to upgrade our core products.