ICLR 2026 Oral

TRACE:

Your Diffusion Model is Secretly an Instance Edge Detector

Sanghyun Jo[1*], Ziseok Lee[2*], Wooyeol Lee[2], Jonghyun Choi[2], Jaesik Park[2†], Kyungsu Kim[2†]

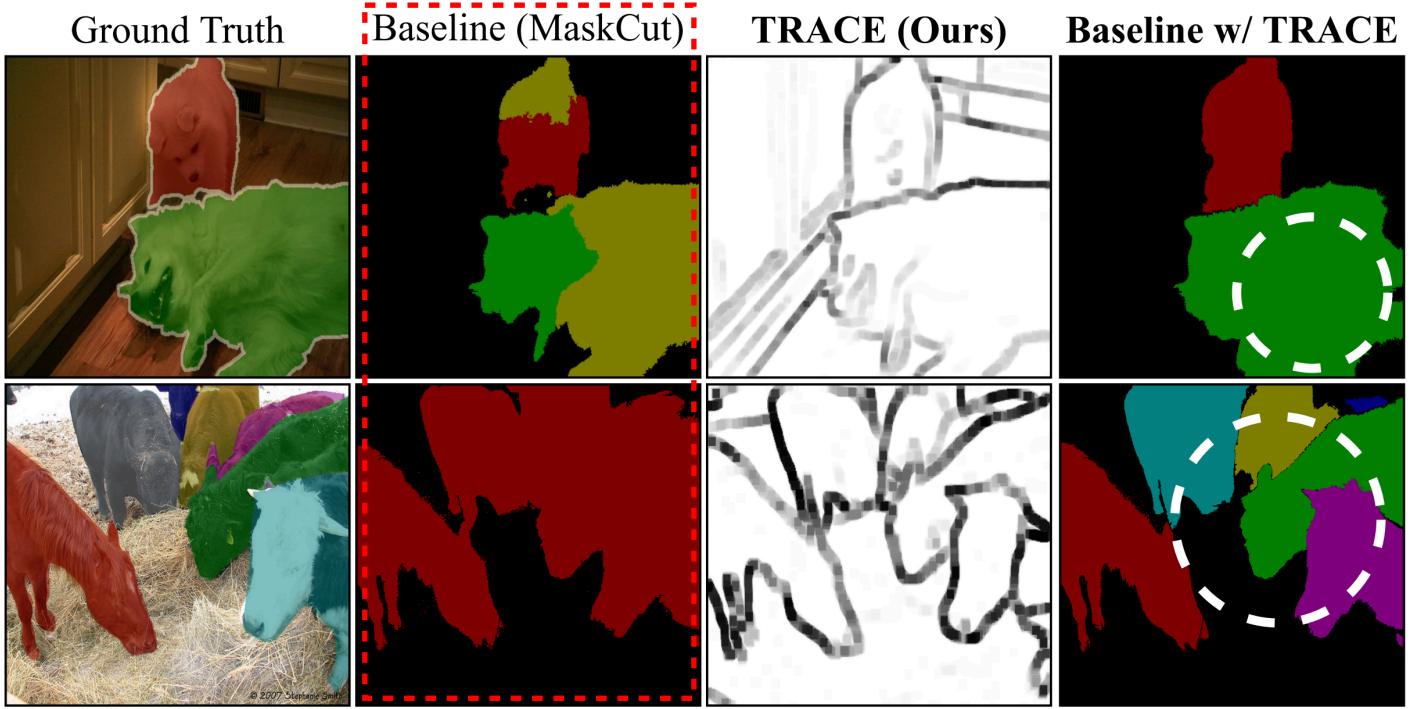
[1]OGQ, Seoul, Korea [2]Seoul National University, Seoul, Korea

*Equal contribution. †Corresponding authors

Problem

Why Instance Boundaries Are Hard?

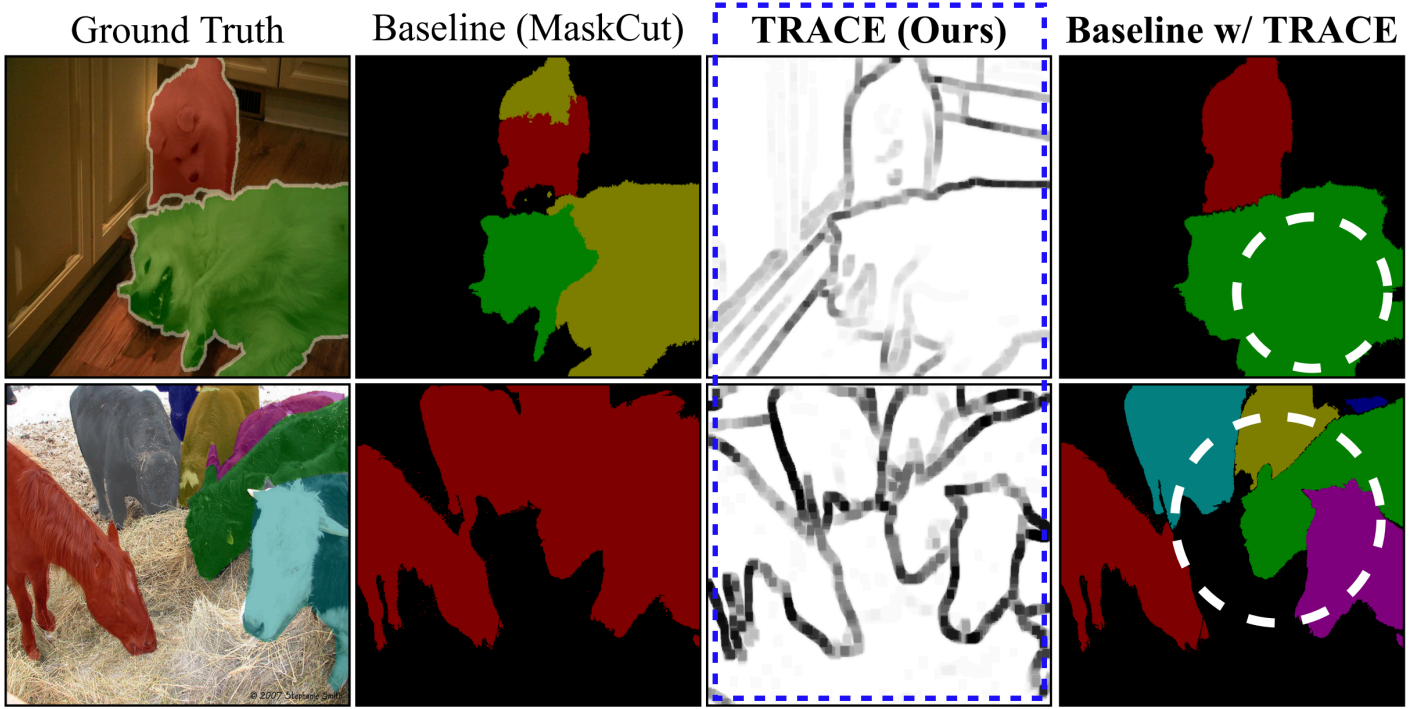
- Vision encoders (e.g., DINO/CLIP) group by class, blending adjacent objects into one blob.
- Forced clustering with fixed object counts shatters single instances into pieces.



Solution

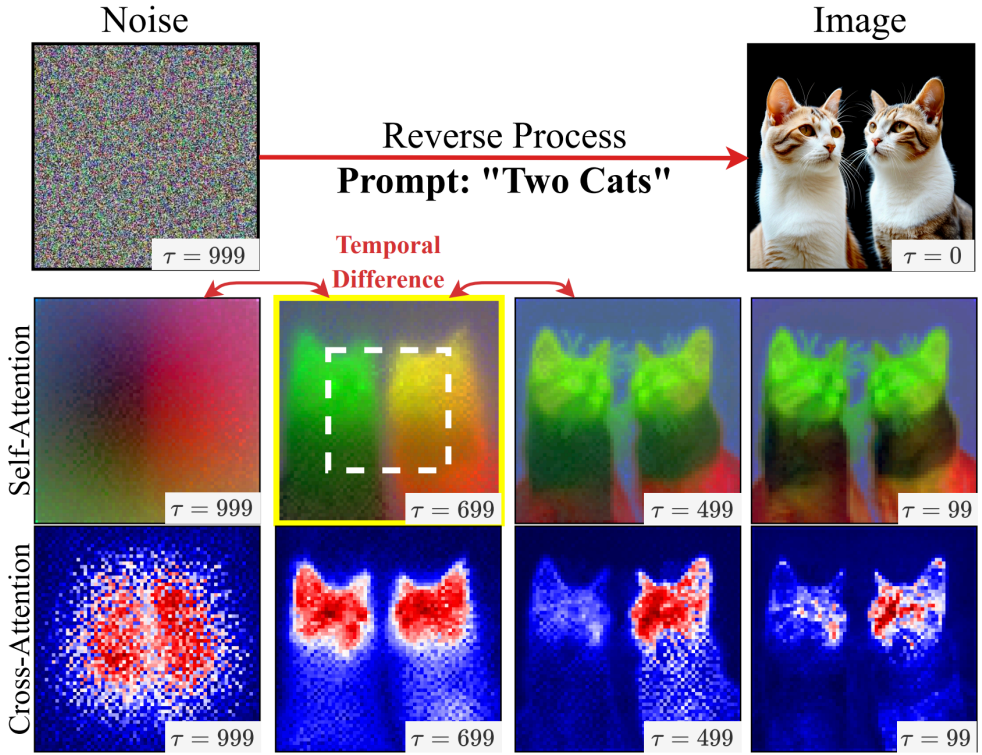
TRACE = Boundary-Centric Separation

- Extracts pure, instance-aware boundaries instead of relying on semantic clustering.
- Uses these edges as barriers to separate adjacent objects for merging and fragmentation.



Key Insight | The Hidden Instance Prior

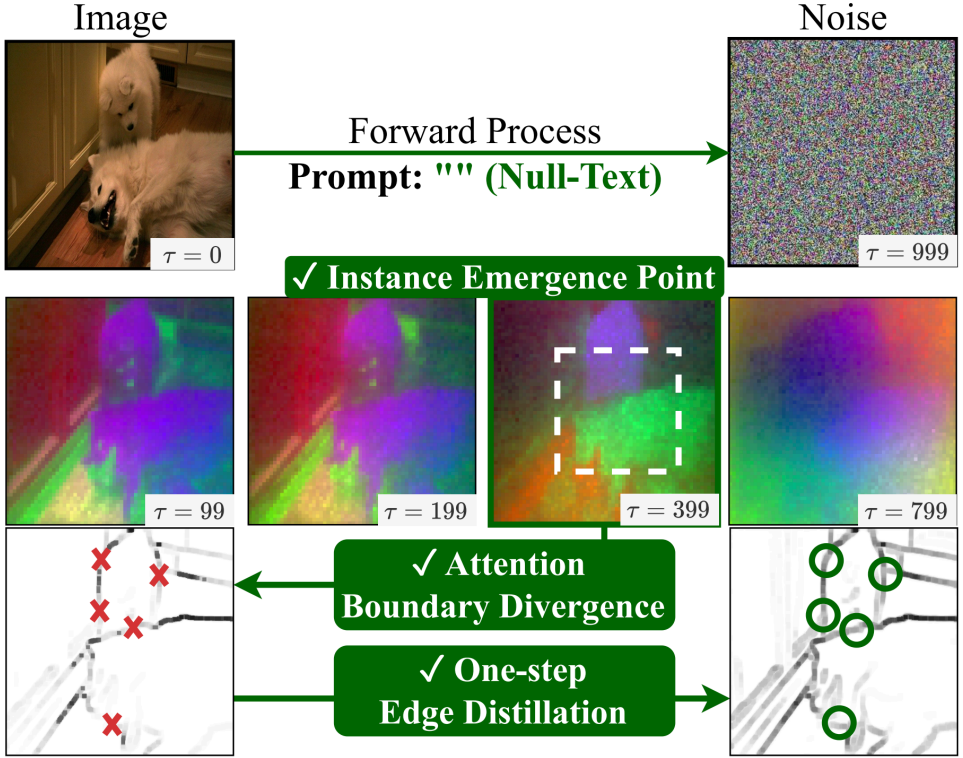
- **Cross-Attention**: Remains semantic, merging adjacent objects even with explicit text prompts.
- **Self-Attention**: Secretly reveals distinct, instance-level structures at early denoising steps.



Summary

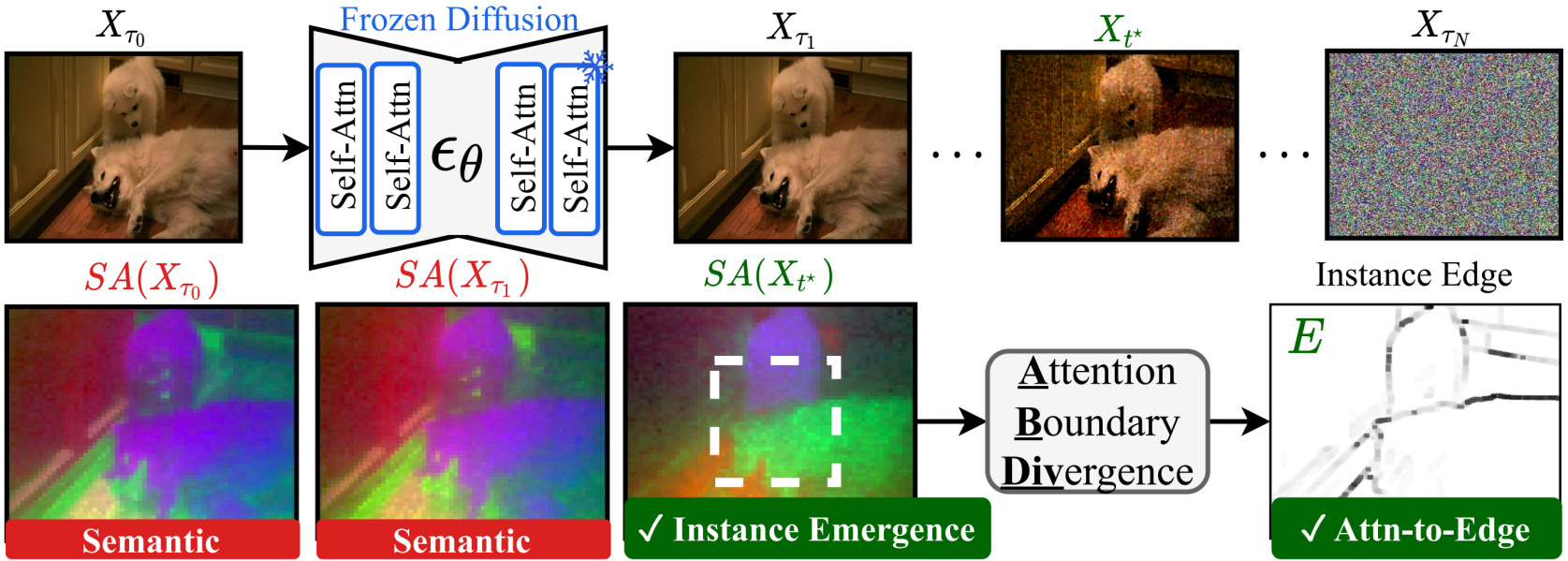
From Hidden Priors to Edges: The TRACE

- **Stage 1.** Instance Emergence Point (IEP) → Attention Boundary Divergence (ABDiv)
- **Stage 2.** One-step Edge Distillation for Real-Time Edge Prediction



Stage 1 | Extracting Hidden Priors (IEP & ABDiv)

- **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.

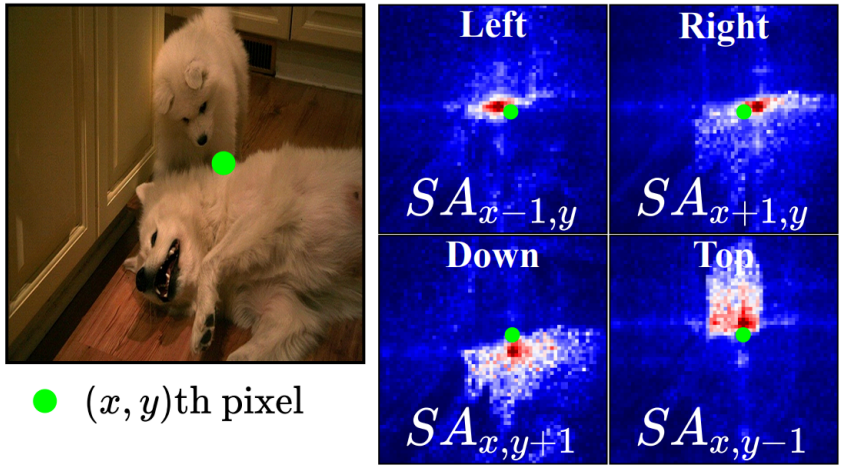


Details of IEP and ABDiv

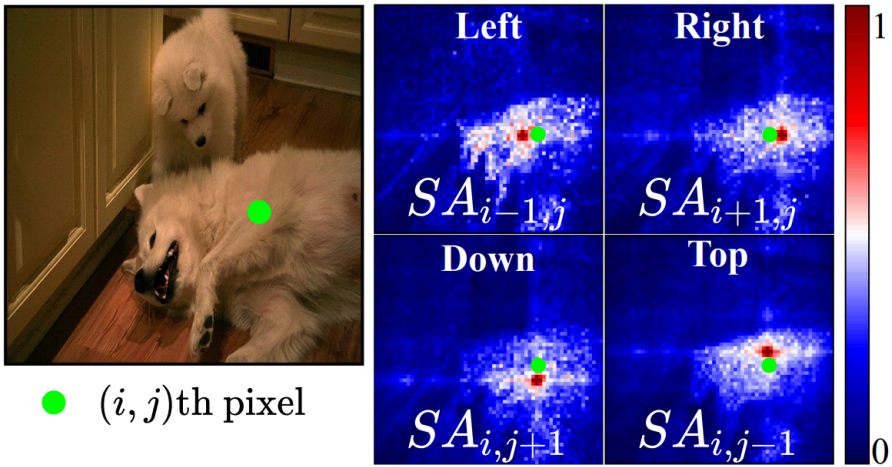
- **Instance Emergence Point:** Locates the exact timestep where temporal KL divergence peaks.

$$t^\star = \underset{t \in \{\tau_1, \ldots, \tau_N\}}{\operatorname{argmax}} \; D_{\mathrm{KL}}(SA(X_{t_{\mathrm{prev}}}) \,\|\, SA(X_t))$$

- **Attention Boundary Divergence:** Scores spatial divergence between opposite neighbors.



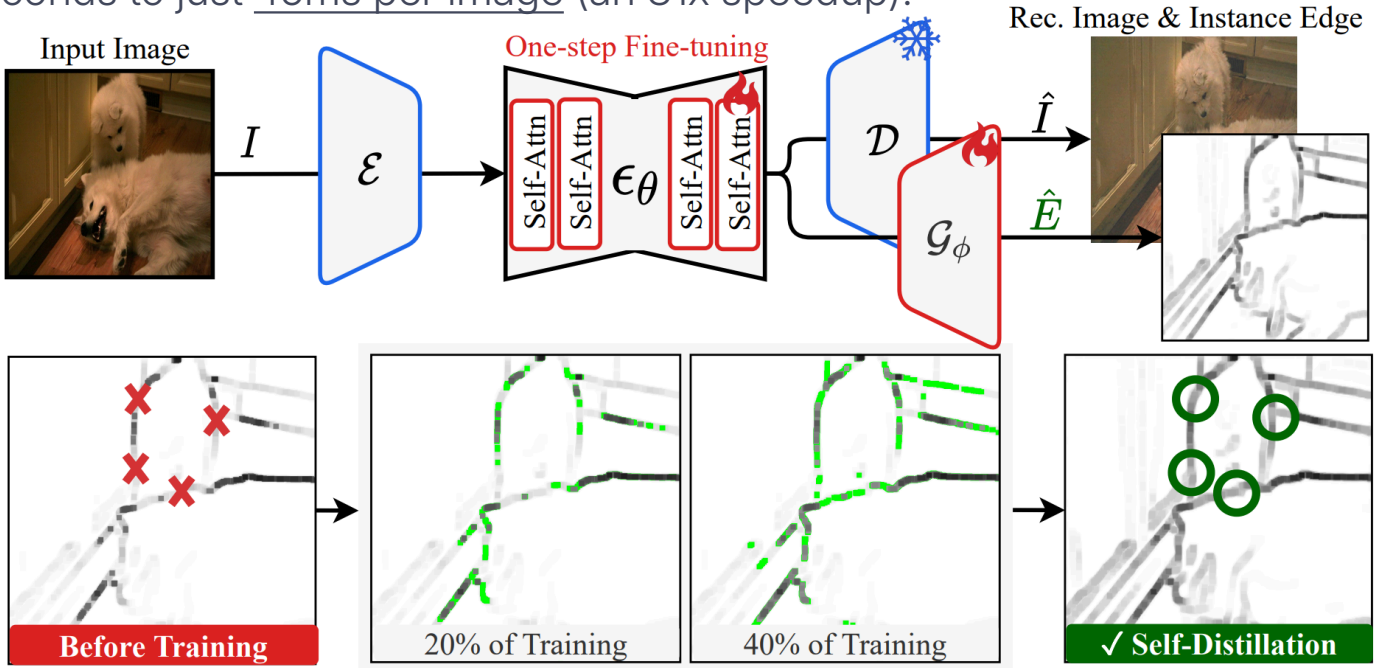
(a) Instance Boundary (ABDiv: 0.114)



(b) Instance Interior (ABDiv: 0.027)

Stage 2 | One-step Edge Distillation

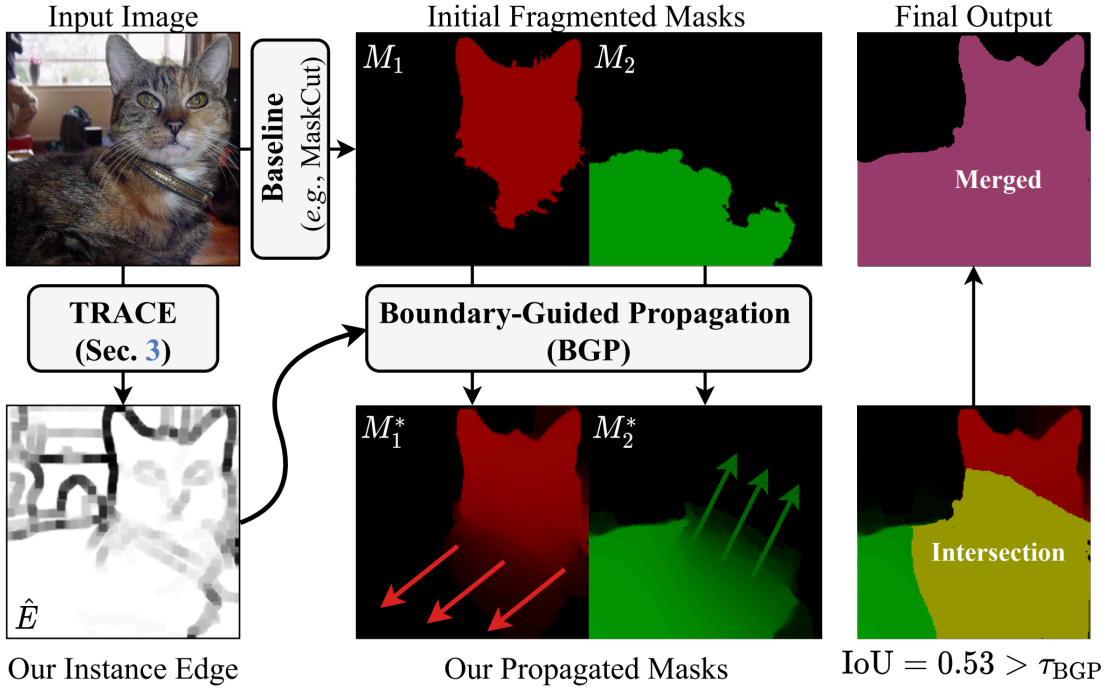
- **Self-Distillation:** Fine-tunes the diffusion backbone (via LoRA) alongside a lightweight edge decoder, using Stage 1 pseudo-edges as training targets.
- **Real-Time Inference:** Eliminates the iterative IEP search, slashing inference latency from over 3 seconds to just 45ms per image (an 81x speedup).



TRACE as a Universal Plug-in

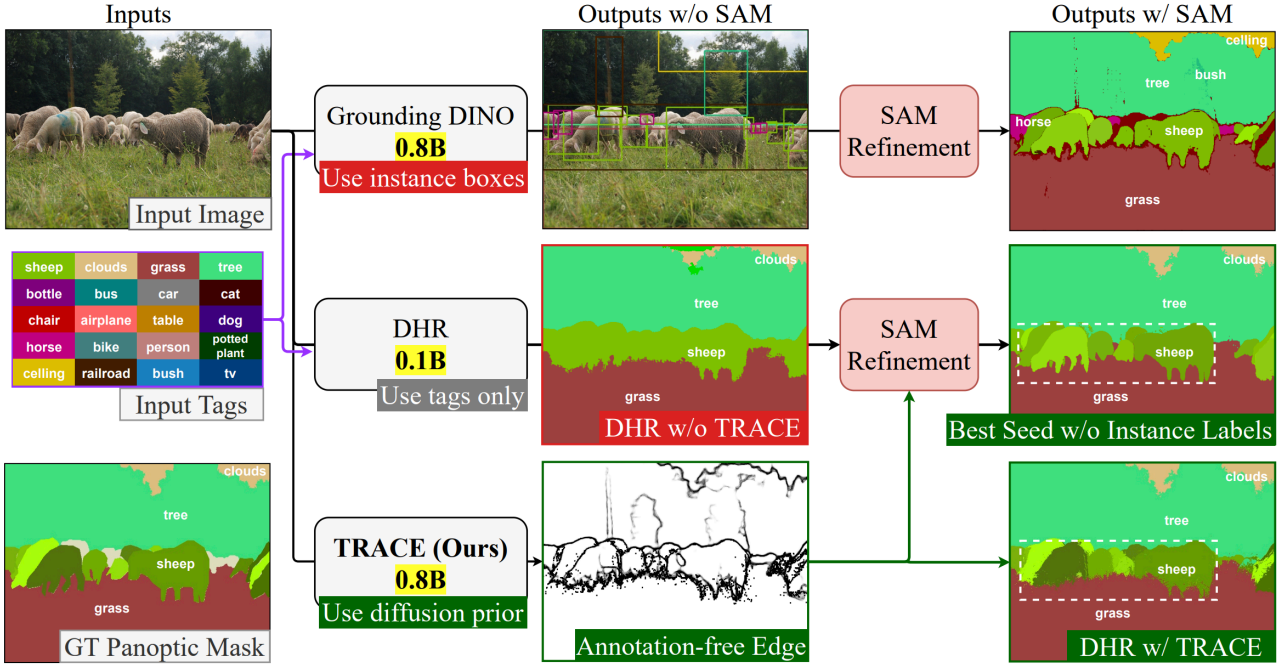
Boundary-Guided Propagation

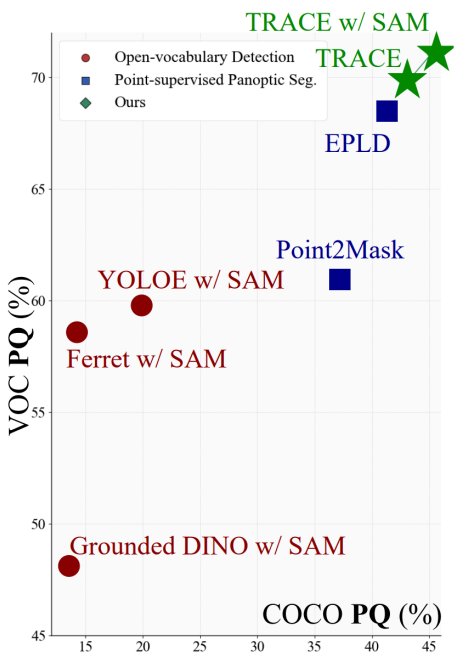
- TRACE edges strictly contain mask expansion, preventing merging while filling internal gaps.
- Overlapping propagated masks are iteratively merged to restore whole objects.



Application | TRACE as a Universal Plug-in

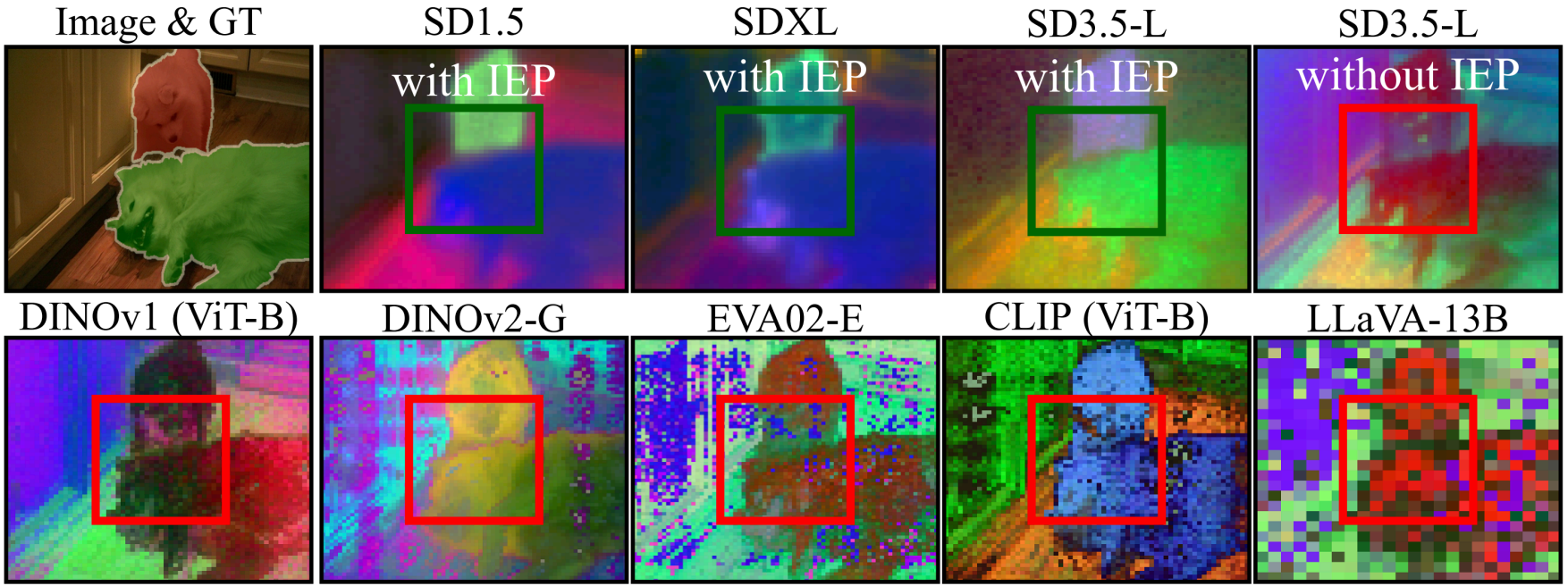
- **Annotation-Free Seeds:** Converts tag-supervised masks into panoptic masks & provides clean instance seeds for SAM.
- **Surpassing SOTA:** Outperforms point- and box-supervised methods w/o instance labels





Why Diffusion? | Diffusion vs. Non-Diffusion

- **Semantic Collapse:** Non-diffusion models (DINO, CLIP, MLLMs) only group by semantics, failing to separate instances.



Why Diffusion? | Diffusion vs. Non-Diffusion

- **Generative Superiority:** Even a lightweight 0.6B diffusion model outperforms a massive 72B MLLM (Qwen2.5-VL) in instance edge extraction.

Method	Backbone	Params	AP^{mk}	AR^{mk}_{100}
ProMerge	–	–	3.1	7.6
Non-diffusion backbones (ABDiv only)				
+ TRACE	DINOv2-G	1.1B	2.6	7.7
+ TRACE	EVA02-E	5.0B	3.2	7.9
+ TRACE	DINOv3	7.0B	4.3	8.9
+ TRACE	LLaVA	13B	3.8	8.4
+ TRACE	Qwen2.5-VL	72B	4.1	8.5
Diffusion backbones (IEP + ABDiv)				
+ TRACE	SD1.5	0.8B	6.8	11.2
+ TRACE	PixArt-α	**0.6B**	7.1	11.8
+ TRACE	SDXL	2.5B	7.4	12.3
+ TRACE	SD3.5-L	8.1B	8.2	13.1
+ TRACE	FLUX.1	12B	**8.3**	**13.4**

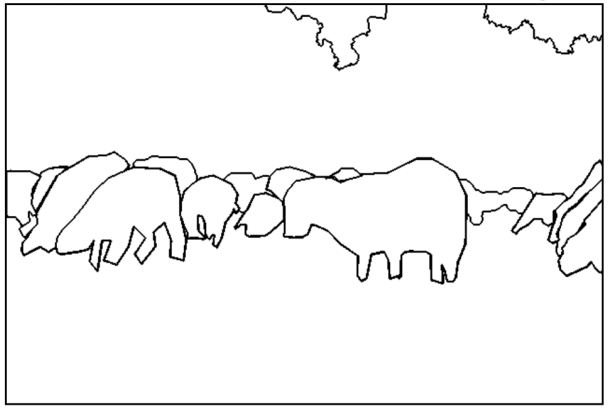
Defining Instance-Aware Edge Ground Truth

- **The Problem:** Standard edge benchmarks evaluate low-level textures and gradients, not instance separation.
- **Our Benchmark:** Extracted pure instance boundaries directly from COCO Panoptic masks to properly evaluate topological connectivity and precision.

(a) Input Image

(b) Ground-Truth Panoptic Mask

(c) Ground-Truth Instance Edge

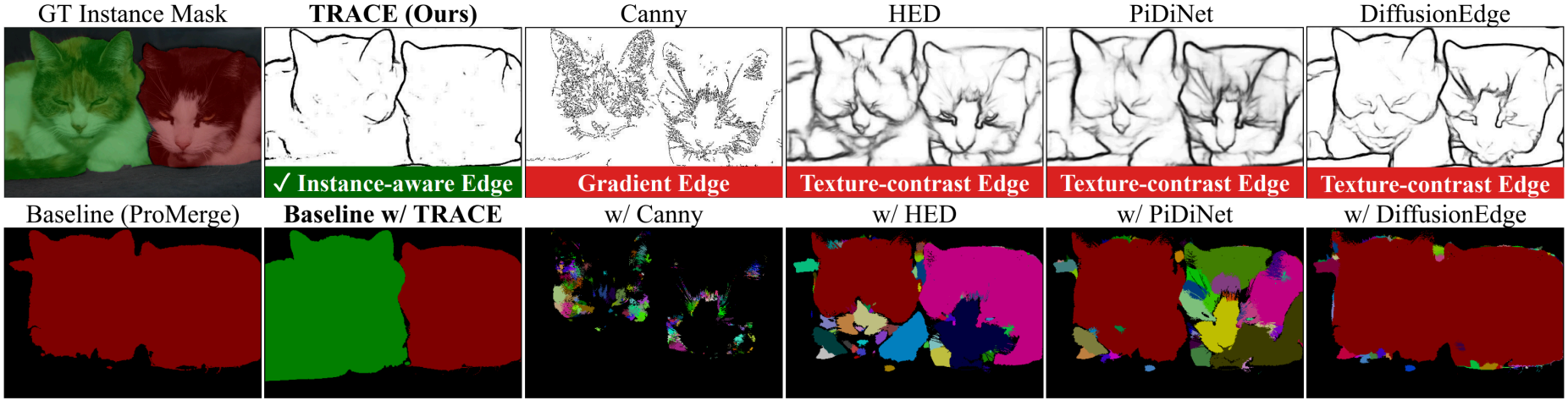


Why Not Traditional Edges?

TRACE vs. Existing

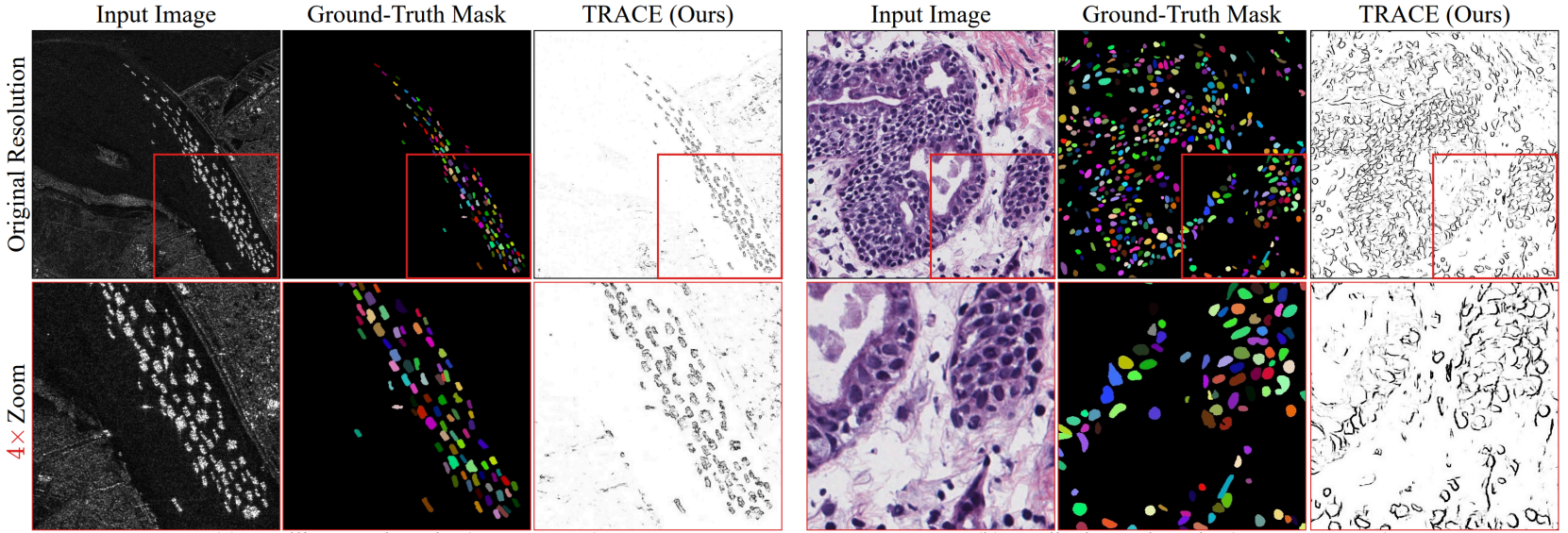
- **Texture Bias:** Conventional detectors (Canny, HED, PiDiNet) capture internal textures and illumination, ruining instance masks.

Method	ODS	OIS	clDice
Canny	0.129	0.202	0.134
HED	0.347	0.443	0.446
PiDiNet	0.362	0.450	0.574
DiffusionEdge	0.428	0.485	0.576
TRACE (Ours)	**0.889**	**0.899**	**0.826**



Discussion | Limitations & Future Directions

- **Tiny Instances (Satellite):** VAE latent compression in diffusion models severely compresses very small objects, blending their boundaries.
- **Out-of-Distribution Domains (Medical):** Natural-image diffusion priors misalign with non-photographic structures like histopathology cells.



(a) Satellite Benchmarks (*e.g.*, HRSID) (b) Medical Benchmarks (*e.g.*, MoNuSeg)

Conclusion

Uncovering Hidden Priors	Diffusion models inherently encode precise instance structures, acting as powerful perception engines.
Pure Boundary Extraction	TRACE decodes these signals to extract true instance edges, overcoming the texture bias of traditional detectors.
Annotation-Free SOTA	Achieves state-of-the-art performance in both unsupervised and weakly-supervised segmentation using zero instance labels.
Real-Time Efficiency	One-step edge distillation accelerates the process by 81x, taking only 45ms per image for practical deployment.