

Efficient SAM3 Adaptation for Multi-Class Semantic Segmentation via Dense Competitive Representations

Wenbin Liao, Hao Zhu, Yike Ma, Hao Jiang, Feng Dai*

Institute of Computing Technology, Chinese Academy of Sciences

School of Computer Science and Technology, University of Chinese Academy of Sciences

33×

Inference Speedup vs. SAM3

68.7

mIoU on ADE20K (SOTA)

8

Benchmarks Evaluated

Two Bottlenecks When Extending SAM3 to Multi-Class Segmentation

What SAM3 Does Well

- Segment Anything Model 3 (SAM3) achieves breakthroughs in **Promptable Concept Segmentation (PCS)**
- Processes short noun phrases (e.g., "a red billboard") to generate unique instance masks
- Strong cross-modal attention enables precise visual condition
- Robust visual-linguistic representations generalize across domains

Two Critical Bottlenecks

① Semantic Misclassification

Independent inference lacks inter-class competition. Similar concepts (tree vs. plant) trigger "double-high" activations, causing overlapping masks.

② Prohibitive $O(n)$ Inference Latency

Each concept requires separate cross-modal attention. On ADE20K (150 classes), inference takes **8038 ms per image**.

Our Goal: Adapt SAM3 to multi-class semantic segmentation by decoupling cross-modal interactions and enforcing explicit inter-class competition — achieving both high accuracy and low latency.

Independent Binary Masks Lack Inter-Class Competition

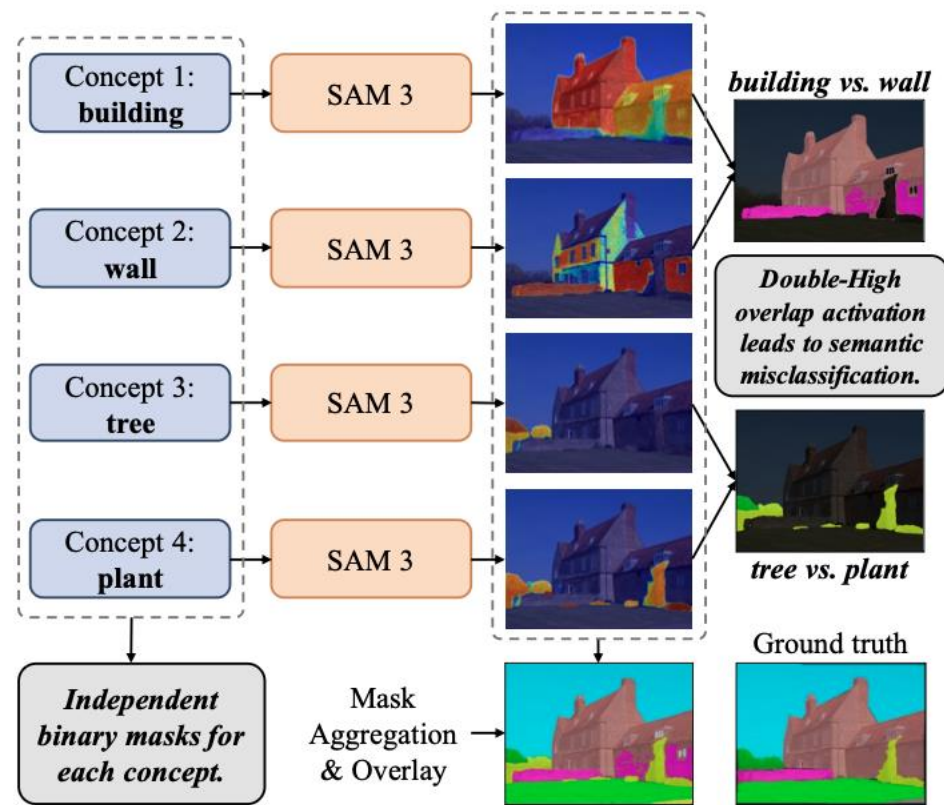


Figure 1: Direct mask aggregation in native SAM3

Why It Fails

- Each concept (building, wall, tree, plant) is treated as an **isolated binary segmentation task**
- No competition mechanism exists across categories
- Similar semantic concepts produce "**double-high**" overlapping activations

Consequence

Overlapping masks and severe misclassification between conceptually similar categories (e.g., tree vs. plant, building vs. wall). The final aggregated mask is unreliable for dense multi-class prediction.

Inference Time Grows Linearly with the Number of Categories

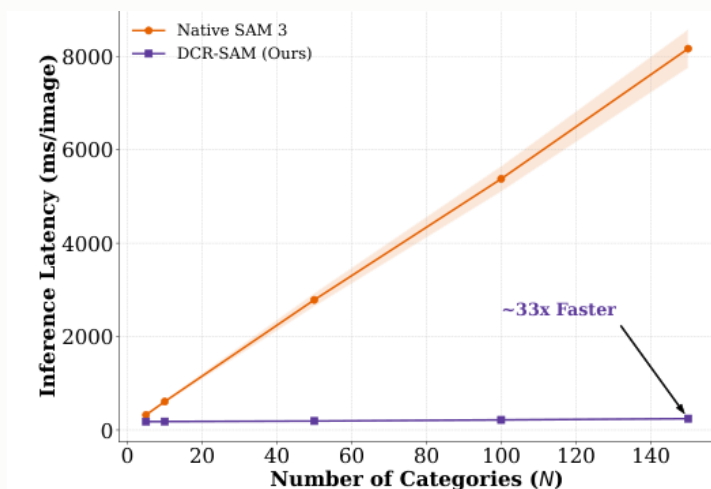


Figure 2: Latency vs. number of categories (ADE20K)

Root Cause

Native SAM3 runs separate **cross-modal attention** and feature decoding for each concept. Complexity is $O(n)$ where n is the number of target categories.

8038 ms

per image on ADE20K (150 classes) —
too slow for real-world deployment

Our Solution Preview

By decoupling cross-modal fusion into a **single forward pass**, DCR-SAM reduces latency to **244 ms** — a **33× speedup**.

DCR-SAM: From Independent Decoding to Dense Competition

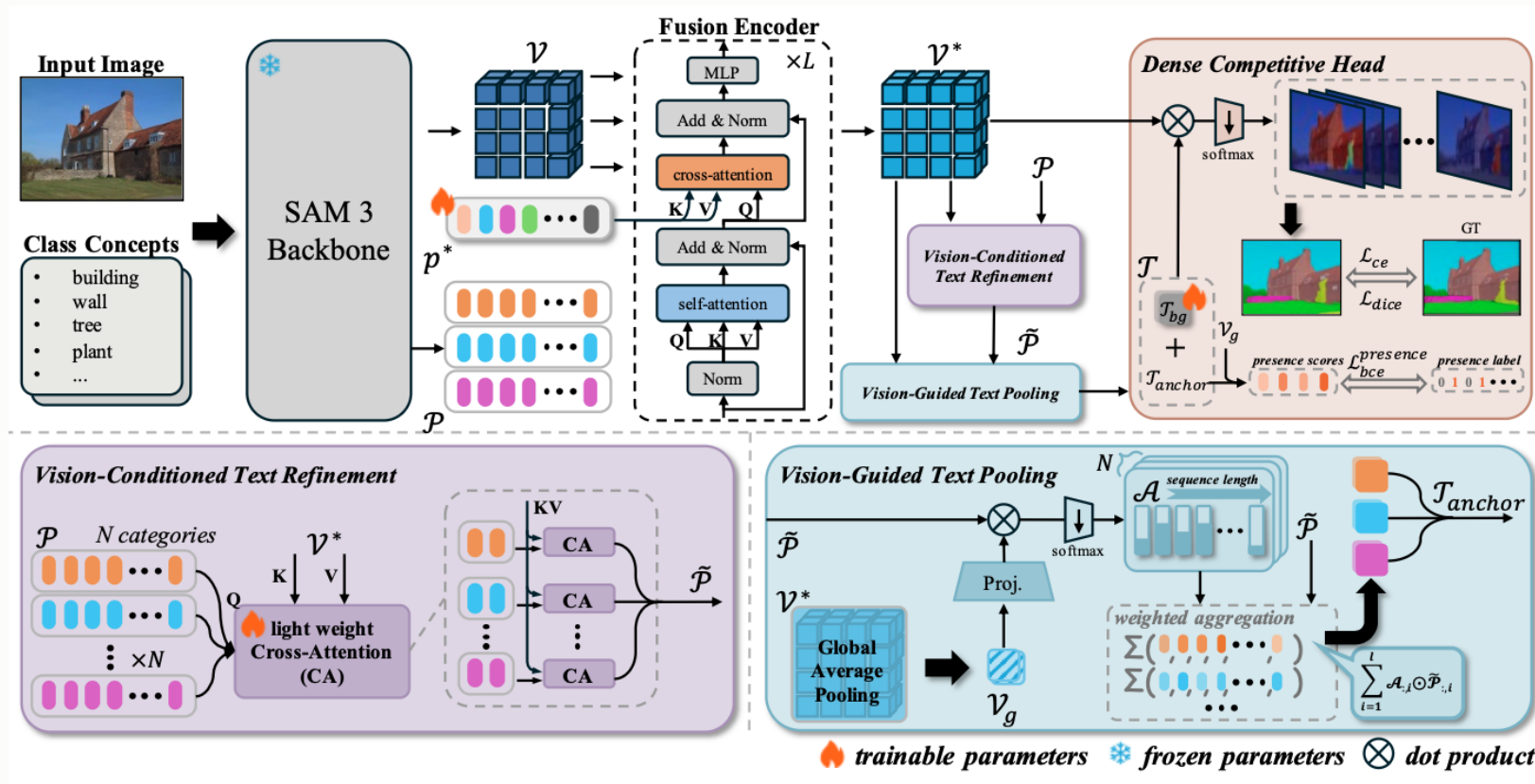


Figure 3: Overview of the DCR-SAM framework

- ① **Decouple**
One-pass cross-modal fusion
- ② **Align**
VCTR + VGTP text-visual fusion
- ③ **Compete**
Dense Competitive Head

Four Design Choices Solve Both Latency and Misclassification

① Decoupled Visual Conditioning

Replace per-concept text queries with a unified learnable prompt p^* . Cross-modal fusion happens in **one forward pass**, eliminating $O(n)$ overhead.

③ Learnable Dynamic Background Token

A learnable token T_{bg} adaptively absorbs non-target objects, avoiding hand-crafted background prompts and unknown distractors.

② VCTR + VGTP

Vision-Conditioned Text Refinement dynamically aligns text with visual cues; Vision-Guided Text Pooling aggregates discriminative class anchors.

④ Dense Competitive Head

Temperature-scaled dot-product between class anchors and visual features enforces **explicit inter-class competition**, suppressing overlapping predictions.

Only the lightweight refinement modules and competitive head are trained; the SAM3 backbone remains frozen.

SOTA on Both Natural Scene and Remote Sensing Benchmarks

Method	ADE20K	Cityscapes	LoveDA	Potsdam	ISAID
SAM3 [†] (Binary)	46.5	72.6	52.8	61.3	64.3
SAM3 [‡] (Multi-Class)	60.3	75.4	55.8	62.4	61.1
SegEarth-OV3	36.6	—	47.1	—	—
EfficientSAM3	43.2	—	50.3	—	—
DCR-SAM (Ours)	68.7	89.1	66.5	89.1	75.4

mIoU (%). Best results in bold green. [†] Native-Binary paradigm, [‡] Native-Multi-Class paradigm. Data from Table 1 of the paper.

+2.3% ~ 10.7%

improvement over SOTA across natural scene benchmarks

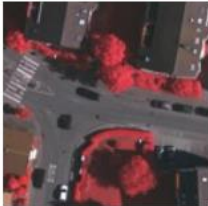
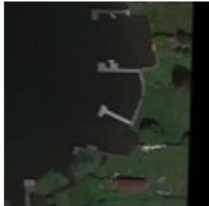
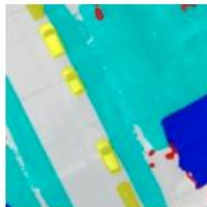
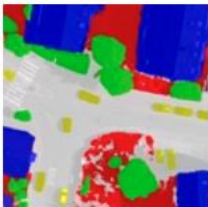
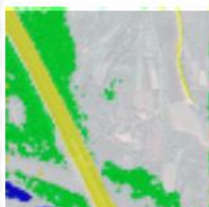
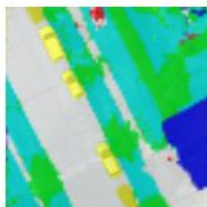
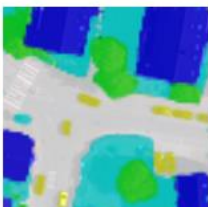
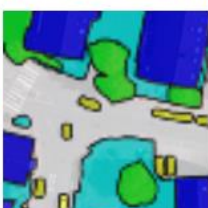
+2.4% ~ 10.7%

improvement on remote sensing benchmarks (stronger domain adaptation)

21.5% / 13.6%

gain over strongest SAM3 baselines on ADE20K / LoveDA

DCR-SAM Preserves Sharp Boundaries and Eliminates Overlaps

	ADE20K	Cityscapes	COCO-Stuff	Context59	LoveDA	Potsdam	Vaihingen	iSAID
Image								
SAM 3								
DCR-SAM								
GT								

SAM3 shows semantic fragmentation and noisy artifacts; DCR-SAM produces clean boundaries with correct category discrimination.

Each Component Contributes to Both Speed and Accuracy

33× speedup

8038 ms → 244 ms on ADE20K

Decoupled Visual Conditioning alone reduces inference time from 8038 ms to 224 ms. Adding VCTR brings it to 244 ms with only negligible overhead, while gaining +2.1 mIoU.

Component Ablation (ADE20K mIoU)

Configuration	mIoU	Time (ms)
Baseline (SAM3+)	46.5	8038
+ Decoupled VC	61.4	224
+ VCTR	68.1	244
Full DCR-SAM (+VGTP, BG)	68.7	244

VCTR: +6.7 mIoU with only +20 ms overhead — most impactful module

VGTP: outperforms first/last/mean pooling, yielding more discriminative anchors

Dynamic BG: beats generic and hard-coded prompts, no manual tuning

DCR-SAM: Efficient, Competitive, Generalizable

- ① **Problem:** SAM3's independent inference causes semantic misclassification and $O(n)$ latency.
 - ② **Method:** Decoupled Visual Conditioning + VCTR/VGTP + Dynamic Background + Dense Competitive Head.
 - ③ **Results:** SOTA on 8 benchmarks with **33× speedup** over native SAM3.
 - ④ **Future:** Further efficiency and generalization for open-vocabulary segmentation.
-