StructSAM: Structure- and Spectrum-Preserving Token Merging for Segment Anything Models

Duy M. H. Nguyen [1 2 3] **Tuan A. Tran** [2 4] **Duong Nguyen** [2] **Siwei Xie** [1] **Trung Q. Nguyen** [2] **Mai T. N. Truong** [2] **Daniel Palenicek** [5] **An T. Le** [5 6 7] **Michael Barz** [2] **Eric Hannus** [8] **TrungTin Nguyen** [9] **Tuan Dam** [10] **Tran Le** [11] **Ngan Le** [12] **Minh Vu** [6 7] **Khoa Doan** [7] **Vien Ngo** [6 7] **Pengtao Xie** [13] **James Zou** [14] **Daniel Sonntag** [2 4] **Jan Peters** [2 5] **Mathias Niepert** [1 3]

Abstract

SAM achieves strong segmentation but at high inference cost dominated by its ViT image encoder. Token merging accelerates ViTs without retraining, yet directly applying it to SAM is nontrivial: SAM mixes windowed and global attention and requires dense, prompt-conditioned features for precise boundary prediction. We systematically evaluate representative token-merging methods on the SAM family in a strict off-the-shelf setting and find that existing destination-selection heuristics erode boundaries and leak prompt information as merge rates increase. We propose **StructSAM**, a resolution-preserving framework that computes lightweight token-energy scores from first-order feature gradients, protects boundary and prompt regions via grid-based flatness screening, and merges flat-region tokens toward low-energy targets with explicit recovery. We further present a spectral graph coarsening analysis showing that score-guided merging yields bounded Laplacian spectral distortion relative to random or window-restricted baselines. Across five natural and medical benchmarks, StructSAM reduces encoder FLOPs by 25–30% (up to 40%+ with prompt-aware merging) with minor drops in mIoU/Dice, consistently outperforming recent merging techniques at the same compute.

[1]University of Stuttgart, Germany [2]German Research Center for Artificial Intelligence (DFKI) [3]International Max Planck Research School for Intelligent Systems, Germany [4]Carl von Ossietzky University of Oldenburg, Germany [5]Technical University of Darmstadt, Germany [6]VinRobotics, Vietnam [7]VinUniversity, Vietnam [8]Aalto University, Finland [9]Queensland University of Technology, Australia [10]Hanoi University of Science and Technology (HUST), Vietnam [11]Technical University of Denmark, Denmark [12]University of Arkansas, USA [13]University of California San Diego, USA [14]Stanford University, USA. Correspondence to: Mathias Niepert <mathias.niepert@ki.uni-stuttgart.de>.

International Conference on Machine Learning, AdaptFM: Resource-Adaptive Foundation Model Inference, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

SAM (Kirillov et al., 2023) has emerged as a foundation model for segmentation, combining a powerful ViT image encoder (Dosovitskiy, 2020) with prompt-conditioned mask decoding to enable flexible, interactive segmentation across diverse domains, including medical imaging (Ma et al., 2024), robot surgery (Wang et al., 2023a), and embodied AI (Li et al., 2025; Noh et al., 2025). However, its practical deployment is severely constrained by computational cost (Zhang et al., 2023): in large variants such as ViT-L and ViT-H, the image encoder alone accounts for over 98% of the total parameters and FLOPs.

Recent efforts to improve SAM's efficiency largely rely on model compression, including knowledge distillation (Zhang et al., 2023; Zhou et al., 2025), lightweight backbone re-design (Xiong et al., 2024), and post-training quantization (Lv et al., 2024; Zhang et al., 2025). While effective, these approaches typically require retraining or task-specific calibration, limiting their applicability when SAM is used off-the-shelf or domain-specific fine-tuning is undesirable.

Since SAM's pre-trained representations are already highly expressive, this **motivates approaches that retain the original weights while reducing inference cost**. Token merging (Bolya et al., 2022; Tran et al., 2024) has recently emerged as a promising strategy for accelerating ViTs by dynamically reducing tokens processed by self-attention, with substantial gains in classification (Bolya et al., 2022) and segmentation (Norouzi et al., 2024), often without retraining. *However, directly applying existing token merging to SAM is nontrivial*: (i) SAM's encoder *interleaves windowed and global attention* and preserves fine-grained spatial details crucial for mask prediction, and (ii) segmentation *requires dense, structured outputs*, making aggressive token reduction (Bolya & Hoffman, 2023; Kim et al., 2024) incompatible without careful unmerging.

Motivated by these challenges, we systematically study representative token-merging techniques developed for ViTs (Bolya et al., 2023; Tran et al., 2024; Li et al., 2024) and dense segmentation (Norouzi et al., 2024; Bolya &

Hoffman, 2023), and adapt them to the SAM family (SAM and MedSAM) on boundary-sensitive natural-image and cross-domain medical benchmarks in a *strict off-the-shelf setting*. We find that prior approaches - which rely on random, global, or window-restricted destination selection - struggle to preserve object boundaries and prompt-relevant regions, leading to noticeable degradation as the merge rate increases.

To address this, we propose **StructSAM**, a structure- and spectrum-preserving token-merging framework tailored to SAM-style architectures. StructSAM (i) identifies boundary-critical tokens using a lightweight energy score computed from first-order finite differences on the encoder feature map (Ziou & Tabbone, 1998; Forsyth & Ponce, 2002), inspired by spectral graph energy (Balakrishnan, 2004; Gutman & Zhou, 2006); (ii) groups tokens into grid-based cells and ranks cells by flatness to select mergeable regions with spatial coherence; and (iii) merges tokens within selected cells toward low-energy destinations while explicitly unmerging to recover the original token resolution required by SAM's mask decoder. When box prompts are available, a prompt-aware variant restricts aggressive merging to tokens outside the prompted region. We further show that our merging admits a spectral graph-theoretic interpretation (Jin et al., 2020; Tran et al., 2024) that provably preserves intrinsic spectral properties under mild conditions.

Contributions.

- We present the first systematic evaluation of *inference-time* token merging for the SAM family in a strict off-the-shelf setting, revealing why existing strategies fail under boundary- and prompt-sensitive segmentation.
- We propose StructSAM, a boundary- and prompt-aware merging strategy that leverages gradient-based token energy and cell flatness, reducing FLOPs by 25–30% (up to 40%+ with prompt-aware merging) while maintaining segmentation quality across natural and medical benchmarks.
- We provide a spectral graph-theoretic analysis showing that score-guided merging yields a provable bound on spectral distortion, explaining its robustness over random or similarity-only merging.

2. Related Work

Compression, Distillation, and Quantization for SAM. Most existing methods focus on *backbone replacement or structured compression*: MobileSAM (Zhang et al., 2023), FastSAM (Zhao et al., 2023), EdgeSAM (Zhou et al., 2025), and EfficientSAM (Xiong et al., 2024) replace the original ViT-H encoder with lightweight architectures (e.g., TinyViT (Wu et al., 2022), EfficientViT (Zhang et al., 2024)), typically requiring training from scratch with high data and compute costs. SlimSAM (Chen et al., 2024) compresses the original SAM via structured pruning and distillation, better preserving pre-trained knowledge but introducing additional optimization complexity. In contrast, StructSAM investigates *off-the-shelf acceleration* via inference-time token merging, without modifying weights or requiring retraining.

Additionally, although prior work has investigated quantization techniques (Liu et al., 2024a; Lv et al., 2024; Xiao et al., 2023) to lower SAM's memory usage and bit-width requirements, our merging approach is complementary to these methods. Integrating the two enables synergistic improvements, further reducing FLOPs and speeding up inference even for already quantized models (Fig. 5).

Token Pruning and Merging in Transformers. Token pruning has been explored in NLP (Goyal et al., 2020; Zhong et al., 2023) and ViTs (Yin et al., 2022; Wang et al., 2023b), but typically requires training and yields input-dependent token counts that complicate batching. Token merging methods, led by ToMe (Bolya et al., 2023) and follow-ups (Chen et al., 2023; Shi et al., 2024), merge similar tokens via lightweight bipartite matching, achieving better efficiency–accuracy trade-offs but remaining sensitive to token partitioning. More principled clustering or graph-based methods (Loukas & Vandergheynst, 2018; Tran et al., 2024) offer stronger guarantees but incur substantial overhead and typically reduce tokens *progressively across layers*, which is ill-suited to SAM's architecture and dense segmentation setting.

Token Reduction for Semantic Segmentation. Token halting approaches pause high-confidence tokens but limit information flow (Tang et al., 2023; Liu et al., 2024b). Clustering strategies such as ELViT (Liang et al., 2022) and AiluRus (Li et al., 2023) merge neighboring tokens within a single layer, offering limited efficiency gains. Content-aware token sharing (Lu et al., 2023) uses a policy network but incurs additional cost and addresses only local redundancies. ALGM (Norouzi et al., 2024) performs adaptive local-then-global merging based on cosine similarity. In comparison, our *structure-aware* strategy uses gradient-based energy scores to identify protected boundary tokens and mergeable regions, avoiding the random or purely local window-based decisions of prior work.

3. Method

3.1. SAM architecture

SAM (Kirillov et al., 2023) uses a transformer-based image encoder that embeds an image into visual tokens and processes them through a hierarchical ViT encoder to produce multi-scale features for the mask decoder.

To balance efficiency and global context modeling, SAM interleaves *local* and *global* attention. Given image tokens

$\mathcal{X} = \{x_1, \ldots, x_N\}$, tokens are partitioned into disjoint spatial windows $\{\mathcal{P}_k\}_{k=1}^{K}$ such that

$$\bigcup_{k=1}^{K} \mathcal{P}_k = \mathcal{X}, \qquad \mathcal{P}_i \cap \mathcal{P}_j = \emptyset \ \ (i \neq j).$$

Most layers apply self-attention independently within each window:

$$\mathrm{Attn}_{\mathrm{local}}(x_i) = \sum_{x_j \in \mathcal{P}(i)} \mathrm{softmax}_j\left(\frac{\mathbf{q}_i^\top \mathbf{k}_j}{\sqrt{d}}\right) \mathbf{v}_j.$$

To enable long-range interactions, SAM periodically applies global attention over locally updated tokens $\tilde{x}_i = \mathrm{Attn}_{\mathrm{local}}(x_i)$:

$$\mathrm{Attn}_{\mathrm{global}}(\tilde{x}_i) = \sum_{j=1}^{N} \mathrm{softmax}_j\left(\frac{\tilde{\mathbf{q}}_i^\top \tilde{\mathbf{k}}_j}{\sqrt{d}}\right) \tilde{\mathbf{v}}_j.$$

This hybrid design captures global context while reducing the cost of full self-attention. Unlike standard ViTs, SAM requires a dense 2D feature grid for mask decoding. We therefore use a *merge–compute–unmerge* scheme that reduces attention cost while restoring full-resolution tokens afterward.

3.2. Resolution-preserving merge–unmerge interface

To reduce SAM's self-attention cost, we introduce a token merging framework for its mixed local–global attention. At encoder layer ℓ, let $\mathcal{X}^{(\ell)} = \{x_1^{(\ell)}, \ldots, x_{N_\ell}^{(\ell)}\}$ be the full-resolution token set. We define a merging operator

$$f_\ell : \mathcal{X}^{(\ell)} \to \widetilde{\mathcal{X}}^{(\ell)}, \qquad |\widetilde{\mathcal{X}}^{(\ell)}| < |\mathcal{X}^{(\ell)}|,$$

which merges spatially redundant tokens while preserving important semantics.

Attention is computed on merged tokens and restored to the original resolution via an unmerging operator f_ℓ^{-1}:

$$\widetilde{\mathcal{X}}^{(\ell+1)} = \mathrm{Attn}\left(\widetilde{\mathcal{X}}^{(\ell)}\right), \qquad \mathcal{X}^{(\ell+1)} = f_\ell^{-1}\left(\widetilde{\mathcal{X}}^{(\ell+1)}\right),$$

where $\mathrm{Attn}(\cdot)$ denotes local or global attention. This merge–compute–unmerge design reduces attention cost while preserving SAM segmentation quality.

Plug-in baselines. Existing token merging methods (e.g., ToMe (Bolya et al., 2023), PI-ToMe (Tran et al., 2024), ToMe-SD (Bolya & Hoffman, 2023), VidToMe (Li et al., 2024), and ALGM (Norouzi et al., 2024)) can be viewed as different choices of f_ℓ. We compare them under the same interface across multiple SAM encoder sizes and datasets.

3.3. StructSAM: gradient-guided structure-aware token merging

We overview our method in Figure 1. At transformer layer ℓ of the SAM image encoder, the image tokens

$$\mathcal{X}^{(\ell)} = \{x_1^{(\ell)}, \ldots, x_{N_\ell}^{(\ell)}\}$$

are reshaped into a feature map

$$\mathbf{I}^{(\ell)} \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell}, \qquad |\mathcal{X}^{(\ell)}| = H_\ell W_\ell.$$

Each token $x_i^{(\ell)}$ corresponds to a spatial position $p(i) = (h(i), w(i))$, aligned with SAM's relative positional embeddings. Our goal is to reduce self-attention cost by merging spatially redundant tokens while preserving tokens important for object boundaries and prompt-conditioned segmentation.

Feature gradient–based energy estimation. We interpret token features as a discrete feature field

$$\mathbf{I}^{(\ell)} : (h, w) \mapsto \mathbf{f}_{h,w}^{(\ell)} \in \mathbb{R}^{C_\ell},$$

and estimate local gradients using finite differences:

$$\nabla_x \mathbf{I}^{(\ell)}(h, w) \approx \mathbf{f}_{h,w+1}^{(\ell)} - \mathbf{f}_{h,w-1}^{(\ell)},$$

$$\nabla_y \mathbf{I}^{(\ell)}(h, w) \approx \mathbf{f}_{h+1,w}^{(\ell)} - \mathbf{f}_{h-1,w}^{(\ell)}.$$

The gradient magnitude

$$\mathbf{G}^{(\ell)}(h, w) = \sqrt{\|\nabla_x \mathbf{I}^{(\ell)}(h, w)\|_2^2 + \|\nabla_y \mathbf{I}^{(\ell)}(h, w)\|_2^2}$$

serves as a lightweight energy score, where high values typically indicate strong local feature variations (e.g., boundaries) and are thus preserved during merging. Compared with graph-based energy methods (Tran et al., 2024), this first-order approximation achieves similar effectiveness with substantially lower overhead (65–75% FLOPs reduction).

Cell partitioning aligned to attention windows. We partition the token grid into non-overlapping $s \times s$ cells (Fig. 1). For local-attention layers, partitioning is applied within each attention window $\mathcal{P}_k$; for global-attention layers, the full grid is treated as one window. Let

$$\mathcal{C}^{(\ell)} = \{\mathcal{C}_1^{(\ell)}, \ldots, \mathcal{C}_M^{(\ell)}\}, \ \bigcup_{m=1}^{M} \mathcal{C}_m^{(\ell)} = \mathbf{I}^{(\ell)}, \quad \mathcal{C}_i^{(\ell)} \cap \mathcal{C}_j^{(\ell)} = \emptyset.$$

Cell flatness and protected set. We define the cell flatness score as

$$\phi(\mathcal{C}_m^{(\ell)}) = -\max_{(h,w) \in \mathcal{C}_m^{(\ell)}} \mathbf{G}^{(\ell)}(h, w),$$

where higher values indicate smoother regions. Given a merge rate $r \in [0, 1)$, cells are sorted by ϕ, and the first M_{merge} cells are selected as mergeable such that the final token count matches $(1 - r)H_\ell W_\ell$. Since each $s \times s$ cell reduces s^2 tokens to one, each selected cell removes $s^2 - 1$ tokens. Remaining cells form the protected set whose tokens are preserved.

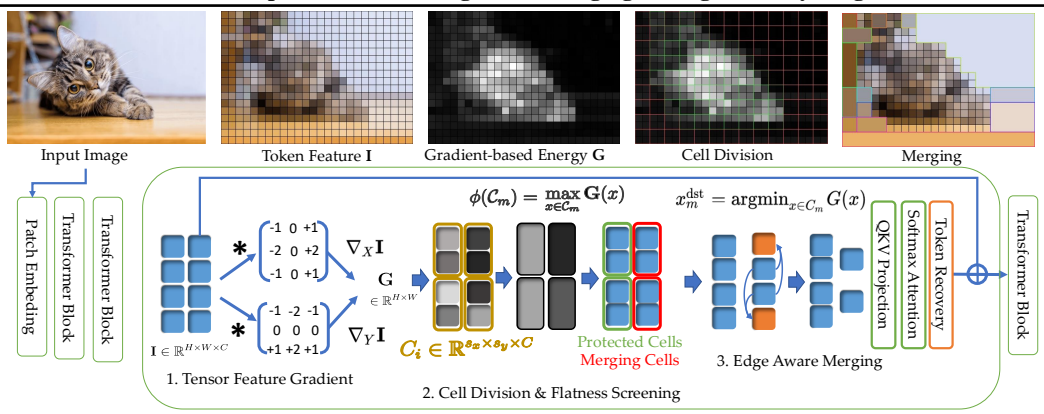


Figure 1. **StructSAM overview.** Feature-gradient energy identifies structurally important tokens, forming a **protected set** that is kept at full resolution. Visually flat regions are selectively merged (one representative per mergeable cell) and followed by lightweight token recovery (unmerging), resulting in SAM's mask decoder still receiving a dense feature grid.

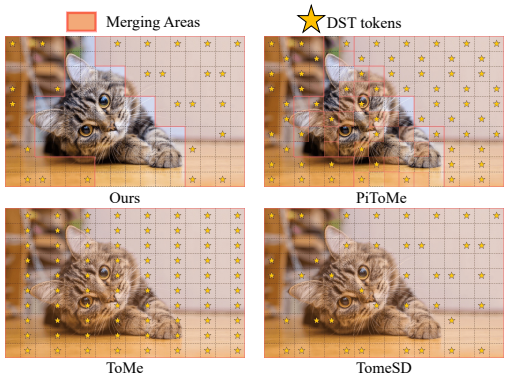


Figure 2. Illustration of token merging strategies. ToMe and ToMeSD treat all tokens as mergeable, while PiToMe introduces a protected set that is effective only at low merge rates. In contrast, our method preserves structurally important tokens while supporting aggressive token reduction.

Destination and source token selection. For each mergeable cell $\mathcal{C}_m^{(\ell)}$, the destination token is chosen as

$$x_{m,\mathrm{dst}}^{(\ell)} = \arg \min_{x_i^{(\ell)} \in \mathcal{C}_m^{(\ell)}} \mathbf{G}^{(\ell)}(h(i), w(i)),$$

and the remaining tokens form the source set

$$\mathcal{S}_m^{(\ell)} = \mathcal{C}_m^{(\ell)} \setminus \{x_{m,\mathrm{dst}}^{(\ell)}\}.$$

Cell-wise token merging. Source tokens are merged into the destination token via bipartite soft matching followed by averaging:

$$\widetilde{\mathbf{f}}_{x_{m,\mathrm{dst}}^{(\ell)}}^{(\ell)} = \frac{1}{|\widetilde{\mathcal{C}}_m^{(\ell)}|} \sum_{x_i^{(\ell)} \in \widetilde{\mathcal{C}}_m^{(\ell)}} \mathbf{f}_i^{(\ell)},$$

where $\widetilde{\mathcal{C}}_m^{(\ell)}$ denotes the matched token set. The merged token set consists of all tokens in protected cells and the destination tokens of mergeable cells.

Token recovery (unmerging). After attention on the merged tokens, the updated destination feature is copied back to all tokens in $\mathcal{S}_m^{(\ell)} \cup \{x_{m,\mathrm{dst}}^{(\ell)}\}$, restoring the original spatial layout required for dense mask prediction.

Prompt-aware variant. When box prompts are available, we apply a lower merge rate inside the prompted region and a higher rate outside, preserving prompt-relevant details while further reducing computation.

4. Graph Coarsening View and Spectral Stability in StructSAM

We analyze *StructSAM* merging via *spectral graph theory*. Tokens within attention windows define weighted graphs; StructSAM's merge–unmerge procedure induces *graph coarsening* and a canonical *lifting* back to the original resolution. Using the feature-gradient energy (Figure 1), mergeable cells act as coarse nodes while protected regions remain at full resolution.

At layer ℓ, tokens in window $\mathcal{P}_k$ form a graph $\mathcal{G}_{\ell,k}$ with normalized Laplacian $\mathcal{L}_{\ell,k}$. After lifting, we obtain $\mathcal{G}_{\ell,k,l}$. We quantify distortion via the *spectral discrepancy*: $\mathrm{SD}_\ell \triangleq \sum_{k=1}^{K_\ell} \|\boldsymbol{\lambda}_{\ell,k} - \boldsymbol{\lambda}_{\ell,k,l}\|_1$, where $\boldsymbol{\lambda}$ are the sorted eigenvalues of the original and lifted Laplacians.

Our analysis relies on three core assumptions (Appendix H): Assumptions 1 and 2 posit within-region concentration and margin separation, ensuring coherent feature structure as seen in Figure 6, while Assumption 3 assumes gradient-separation, where high-gradient boundaries are distinguished from low-gradient interiors to allow reliable flatness screening.

Theorem 1 (Informal: Layerwise Spectrum Stability)**.** *Under Assumptions 1 to 3, let $SD_\ell(\mathrm{SG})$ be the discrepancy of StructSAM and $SD_\ell(\mathrm{Base})$ be a non-score-guided baseline (e.g., ToMe-SD):*

1. ***StructSAM (Vanishing Drift):*** *Flatness screening protects boundaries, forcing merges within coherent regions; thus $\mathbb{E}[SD_\ell(\mathrm{SG})] \to 0$.*

2. ***Baselines (Irreducible Drift):*** *Heuristic selection may merge across boundaries with non-negligible probabil-*

Table 1. Performance of merging methods on DIS5K, ThinObject5K-TE, COIFT, and HRSOD datasets at the merging rate $r = 55\%$. Note that TOME and PITOME can not work with $r > 50\%$.

Model	Method	GFlops	Mem (GB)	DIS5K		ThinObject5K-TE		COIFT		HRSOD	
				mIoU	b-mIoU	mIoU	b-mIoU	mIoU	b-mIoU	mIoU	b-mIoU
ViT-B	Baseline	486.4	3.53	55.30	46.97	63.28	52.65	89.14	82.54	86.64	76.56
	TomeSD	362.3 ↓25.5%	**2.68**	51.39	43.16	58.09	48.24	87.29	80.39	85.34	75.66
	ViDTome	399.8 ↓17.8%	3.43	43.80	35.77	49.74	37.21	80.52	69.55	75.89	62.20
	ALGM	381.1 ↓21.6%	2.93	51.26	42.30	50.24	40.97	49.10	39.48	47.33	37.04
	StructSAM	**347.8**↓28.5%	**2.68**	**54.61**	**45.57**	**63.30**	**52.06**	**87.74**	**80.85**	**86.67**	**76.84**
ViT-L	Base Model	1493.8	5.97	62.27	53.94	75.50	64.71	92.65	87.40	89.67	82.65
	TomeSD	1188.2 ↓20.4%	**4.84**	60.32	50.81	73.68	61.98	90.51	**84.67**	**88.58**	**80.99**
	ViDTome	1274.8 ↓14.7%	5.83	39.44	30.36	47.51	31.05	61.73	46.16	60.79	46.56
	ALGM	1249.6 ↓16.3%	5.17	56.93	44.44	54.42	40.88	52.82	38.17	49.93	33.64
	StructSAM	**1167.1**↓21.9%	**4.84**	**61.01**	**51.36**	**75.80**	**63.81**	**90.73**	84.26	88.39	80.46

ity, implying $\liminf \mathbb{E}[SD_\ell(\text{Base})] > 0$.

Theorem 1 explains why StructSAM maintains segmentation quality at high merge rates where heuristic baselines degrade.

5. Experiments

Q1. Boundary and Thin Structure Preservation. We evaluate StructSAM's ability to preserve fine details and boundary precision across ViT-B and ViT-L architectures. Our experiments utilize four boundary-sensitive benchmarks: DIS5K for pixel-accurate annotations; THINOBJECT-5K and COIFT for globally thin or intricate structures; and HRSOD for precise delineation in complex high-resolution scenes.

Token Merging Comparison. We compare against five representative token merging baselines. TOME (Bolya et al., 2023) uses bipartite soft matching to merge similar tokens, while ToMeSD (Bolya & Hoffman, 2023) extends this with timestep-aware merging for diffusion models in image generation. PiToMe (Tran et al., 2024) improves selection via pivot-based strategies to preserve important tokens, and VidToMe (Li et al., 2024) leverages temporal redundancy in video transformers. ALGM (Norouzi et al., 2024) introduces adaptive local-to-global merging to the ViT architecture for dense segmentation.

Observations. We assess segmentation quality via mIoU and boundary mIoU, and efficiency via GFLOPs, memory, and throughput. As shown in Table 1 and Figure 4, StructSAM reduces FLOPs by ∼28.5% (ViT-B) and ∼21.9% (ViT-L) with lower memory usage and comparable performance to the baseline. It effectively preserves structural details on boundary-sensitive benchmarks, with only minor degradation (e.g., COIFT).

Compared to existing token merging methods, StructSAM maintains superior stability and accuracy at high merge rates (35%–65%), particularly for precise boundaries. While not always the fastest in throughput, StructSAM offers a superior balance by prioritizing structural fidelity and segmentation quality over aggressive speedups (Fig 3b).

Q2. Generalization to SAM Variants. We evaluate StructSAM across (I) MEDICAL SAM (MedSAM) and (II) EFFICIENT-SAM. For MedSAM, we evaluate the INBREAST mammography dataset in both prompt-based and prompt-free settings to assess robustness in high-precision medical scenarios. For EFFICIENT-SAM, the speedup from token merging enables extension from segmentation to video tracking, supporting its use in VLA models (Sec. F Appendix). **Observations.** On MedSAM, StructSAM reduces GFLOPs by up to 28.5% (41.8% with prompt-aware merging) with marginal Dice score impact, consistently outperforming other merging methods at high rates (Fig. 3a). Integrated into Efficient-SAM for VLA tracking, it matches SAM-2's task success rates with a ∼45% speedup, proving its efficiency in both medical and robotic applications.

Table 2. Ablation study on SAM-B. StructSAM (Full) denotes the complete model.

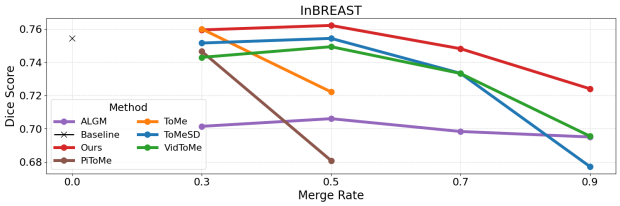
Dataset	Method	$r = 0.35$		$r = 0.55$	
		mIoU	B-IoU	mIoU	B-IoU
DIS5K	StructSAM (Full)	**54.7**	**46.0**	**54.6**	**45.6**
	Central-Diff	53.8	45.6	53.3	44.6
	Mean-Flatness	54.5	45.8	54.4	45.4
	No-Cell	53.9	44.5	54.2	43.9
	Rand-Cell	53.8	43.9	53.5	43.2
	Max-Dst	54.5	45.9	53.9	45.2
	Rand-Dst	**54.7**	**46.0**	54.3	45.4

Table 3. FLOP counts analysis between StructSAM energy scoring and graph-based methods such as PiToMe (Tran et al., 2024).

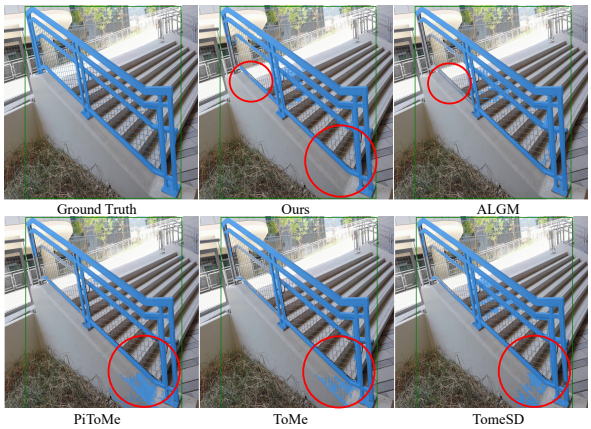
Attention	Method	(10^9) FLOPs/im↓
Global	PiToMe	1.0737
	StructSAM (Central Diff)	0.2684 ↓75.00%
	StructSAM (Sobel)	0.2732 ↓74.56%
Window	PiToMe	0.0615
	StructSAM (Central Diff)	0.0154 ↓74.96%
	StructSAM (Sobel)	0.0210 ↓65.85%

Q3. Compatibility with Quantization. We investigate whether StructSAM is complementary to model quantization and can be applied on top of quantized SAM variants. Among existing quantization approaches for SAM (Liu et al., 2024a; Lv et al., 2024; Xiao et al., 2023), we adopt *SmoothQuant* (Xiao et al., 2023) due to its strong compatibility with GPU kernels and efficient post-training deployment. Our results in Figure 5 show that StructSAM can be

Method	GFLOPs	Dice Score
Base Model	486.4	75.43
TomeSD	362.3 $_{\downarrow 25.5\%}$	73.33
ViDTome	399.8 $_{\downarrow 17.8\%}$	73.32
ALGM	381.1 $_{\downarrow 21.6\%}$	69.83
StructSAM	347.8 $_{\downarrow 28.5\%}$	**74.81**
StructSAM + **prompt-aware**	**283.0** $_{\downarrow 41.8\%}$	74.72



(a)



(b)

Figure 3. (a) *Top*: results on INbreast with **MedSAM**, including a prompt-aware StructSAM variant that restricts token processing for targeted efficiency. *Bottom*: performance across varying merge rates. (b) Qualitative comparison showing that StructSAM better preserves fine structures and detailed regions, while other methods often miss boundaries or over-merge objects into the background.

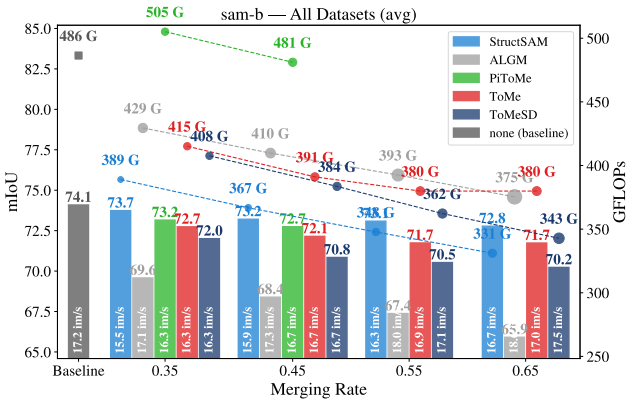


Figure 4. Comparison across merging methods, showing mIoU, GFLOPs, and throughput (img/s) at **different merge rates** on SAM-B. Results for other architectures are in Appendix.

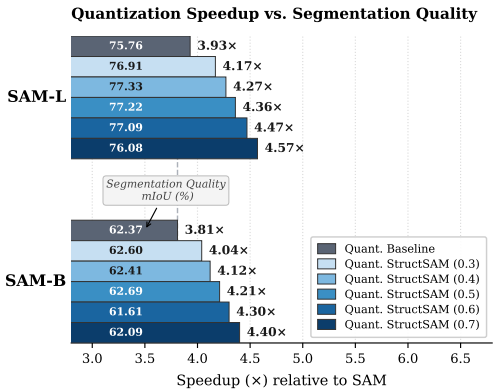


Figure 5. Speedup comparison of quantized baseline and StructSAM (relative to the unquantized baseline) on ThinObject5K across merge rates (30–70%). Numbers indicate mIoU.

seamlessly integrated with a quantized model, providing additional gains in inference speed and memory efficiency while maintaining accuracy, demonstrating its complementary and orthogonal nature to quantization methods.

Q4. Analysis Other Choices. Table 2 reports an ablation study on SAM-B evaluated on DIS5K, using both *mIoU* and *boundary IoU (B-IoU)* under two mask ratios. Across all settings, our **full method** achieves the best or tied-best performance, demonstrating the effectiveness of the proposed design. The ablation for **COIFT** is presented in the Appendix.

- Effect of gradient estimation. Replacing the Sobel operator with a simple central difference (*Central-Diff*) in the token energy score degrades performance, especially in B-IoU (Table 2), highlighting the importance of accurate gradient estimation for preserving boundaries during merging. Our energy score effectively captures fine structures while remaining computationally efficient (Table 3). Additional discussion of failure cases is included in the Appendix.

- Other results. We provide in the Appendix evaluations with (i) point-based prompts instead of box prompts. We also (ii) visualize the heatmap difference between token compression across layers, comparing merged and original tokens to better understand the model's behavior.

6. Discussion and Limitations

Token merging methods for classification ViTs often fail to preserve structure in prompt-conditioned dense tasks. To address this, we introduce a lightweight gradient-based energy that identifies boundary-critical tokens for structure-aware merging. Our method generalizes across natural, medical, and robotic tracking applications. However, its effectiveness depends on representation quality; noisy or low-texture features can reduce the informativeness of gradient cues. Future robustness could be improved by jointly retraining the backbone with merging. Future directions include extending StructSAM to 3D models (Chen et al., 2025) and optimizing ranking operations for better GPU parallelism (Liu et al., 2026).

Impact Statement

This work enables efficient deployment of foundation segmentation models in low-resource environments by substantially reducing inference cost without retraining or architectural modification. By preserving fine-grained structural details under aggressive token reduction, our method broadens access to high-quality segmentation for applications with limited compute, memory, or energy budgets, including medical imaging and embedded vision systems.

References

Balakrishnan, R. The energy of a graph. *Linear Algebra and its Applications*, 387:287–295, 2004.

Bolya, D. and Hoffman, J. Token merging for fast stable diffusion. *CVPR Workshop on Efficient Deep Learning for Computer Vision*, 2023.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your ViT but faster. In *International Conference on Learning Representations*, 2023.

Chen, M., Shao, W., Xu, P., Lin, M., Zhang, K., Chao, F., Ji, R., Qiao, Y., and Luo, P. Diffrate : Differentiable compression rate for efficient vision transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17118–17128, 2023. doi: 10.1109/ICCV51070.2023.01574.

Chen, X., Chu, F.-J., Gleize, P., Liang, K. J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Chen, Z., Fang, G., Ma, X., and Wang, X. Slimsam: 0.1% data makes segment anything slim. *Advances in Neural Information Processing Systems*, 37:39434–39461, 2024.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Forsyth, D. A. and Ponce, J. *Computer vision: a modern approach*. prentice hall professional technical reference, 2002.

Goyal, S., Choudhury, A. R., Raje, S., Chakaravarthy, V., Sabharwal, Y., and Verma, A. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *International Conference on Machine Learning*, pp. 3690–3699. PMLR, 2020.

Gutman, I. and Zhou, B. Laplacian energy of a graph. *Linear Algebra and its applications*, 414(1):29–37, 2006.

Hannus, E., Malin, M., Le, T. N., and Kyrki, V. Ia-vla: Input augmentation for vision-language-action models in settings with semantically complex tasks, 2025. URL https://arxiv.org/abs/2509.24768.

Jin, Y., Loukas, A., and JaJa, J. Graph coarsening with preserved spectral properties. In *International Conference on Artificial Intelligence and Statistics*, pp. 4452–4462. PMLR, 2020.

Kim, M., Gao, S., Hsu, Y.-C., Shen, Y., and Jin, H. Token fusion: Bridging the gap between token pruning and token merging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1383–1392, 2024.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Li, J., Wang, Y., Zhang, X., Shi, B., Jiang, D., Li, C., Dai, W., Xiong, H., and Tian, Q. Ailurus: a scalable vit framework for dense prediction. *Advances in Neural Information Processing Systems*, 36:30979–30996, 2023.

Li, P., Wu, Y., Xi, Z., Li, W., Huang, Y., Zhang, Z., Chen, Y., Wang, J., Zhu, S.-C., Liu, T., et al. Controlvla: Few-shot object-centric adaptation for pre-trained vision-language-action models. *arXiv preprint arXiv:2506.16211*, 2025.

Li, X., Ma, C., Yang, X., and Yang, M.-H. Vidtome: Video token merging for zero-shot video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7486–7495, 2024.

Liang, W., Yuan, Y., Ding, H., Luo, X., Lin, W., Jia, D., Zhang, Z., Zhang, C., and Hu, H. Expediting large-scale vision transformer for dense prediction without fine-tuning. *Advances in Neural Information Processing Systems*, 35:35462–35477, 2022.

Liu, X., Ding, X., Yu, L., Xi, Y., Li, W., Tu, Z., Hu, J., Chen, H., Yin, B., and Xiong, Z. Pq-sam: Post-training quantization for segment anything model. In *European Conference on Computer Vision*, pp. 420–437. Springer, 2024a.

Liu, Y., Zhou, Q., Wang, J., Wang, Z., Wang, F., Wang, J., and Zhang, W. Dynamic token-pass transformers for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1827–1836, 2024b.

Liu, Y., Chen, X., Tian, A., Li, H., Li, Q., Zhang, X., Zhou, A., Zhang, C. J., Li, Q., and Chen, L. Gpu-accelerated algorithms for graph vector search: Taxonomy, empirical study, and research directions. *arXiv preprint arXiv:2602.16719*, 2026.

Loukas, A. and Vandergheynst, P. Spectrally approximating large graphs with smaller graphs. In *International conference on machine learning*, pp. 3237–3246. PMLR, 2018.

Lu, C., De Geus, D., and Dubbelman, G. Content-aware token sharing for efficient semantic segmentation with vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23631–23640, 2023.

Lv, C., Chen, H., Guo, J., Ding, Y., and Liu, X. Ptq4sam: Post-training quantization for segment anything. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 15941–15951, 2024.

Ma, J., He, Y., Li, F., Han, L., You, C., and Wang, B. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Noh, S., Kim, J., Nam, D., Back, S., Kang, R., and Lee, K. Graspsam: When segment anything model meets grasp detection. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14023–14029. IEEE, 2025.

Norouzi, N., Orlova, S., De Geus, D., and Dubbelman, G. Algm: Adaptive local-then-global token merging for efficient semantic segmentation with plain vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15773–15782, 2024.

Shi, D., Tao, C., Rao, A., Yang, Z., Yuan, C., and Wang, J. Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers. *International Conference on Machine Learning*, 2024.

Tang, Q., Zhang, B., Liu, J., Liu, F., and Liu, Y. Dynamic token pruning in plain vision transformers for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 777–786, 2023.

Tran, C., MH Nguyen, D., Nguyen, M.-D., Nguyen, T., Le, N., Xie, P., Sonntag, D., Zou, J. Y., Nguyen, B., and Niepert, M. Accelerating transformers with spectrum-preserving token merging. *Advances in Neural Information Processing Systems*, 37:30772–30810, 2024.

Wang, A., Islam, M., Xu, M., Zhang, Y., and Ren, H. Sam meets robotic surgery: an empirical study on generalization, robustness and adaptation. In *International conference on medical image computing and computer-assisted intervention*, pp. 234–244. Springer, 2023a.

Wang, H., Dedhia, B., and Jha, N. K. Zero-tprune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. *arXiv preprint arXiv:2305.17328*, 2023b.

Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., and Yuan, L. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*, pp. 68–85. Springer, 2022.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16111–16121, 2024.

Yin, H., Vahdat, A., Alvarez, J. M., Mallya, A., Kautz, J., and Molchanov, P. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10809–10818, 2022.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., and Hong, C. S. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

Zhang, W., Zhong, Y., Ando, S., and Yoshioka, K. Ahcptq: Accurate and hardware-compatible post-training quantization for segment anything model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22383–22392, 2025.

Zhang, Z., Cai, H., and Han, S. Efficientvit-sam: Accelerated segment anything model without performance loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7859–7863, 2024.

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., and Wang, J. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

Zhong, Q., Ding, L., Liu, J., Liu, X., Zhang, M., Du, B., and Tao, D. Revisiting token dropping strategy in efficient bert pretraining. *arXiv preprint arXiv:2305.15273*, 2023.

Zhou, C., Li, X., Loy, C. C., and Dai, B. Edgesam: Prompt-in-the-loop distillation for sam. *International Journal of Computer Vision*, 133(12):8452–8468, 2025.

Ziou, D. and Tabbone, S. Edge detection techniques-an overview. *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications*, 8(4): 537–559, 1998.

SUPPLEMENTARY MATERIAL FOR "STRUCTSAM: STRUCTURE- AND SPECTRUM-PRESERVING TOKEN MERGING FOR SEGMENT ANYTHING MODELS"

Contents

A. Additional Discussion on StructSAM Limitations

Despite these advantages, StructSAM has several limitations. First, our current formulation relies on fixed, hand-crafted gradient operators (e.g., Sobel or central differences), which may be suboptimal for highly textured regions or domain-specific imagery. Second, the use of predefined cell partitions introduces an additional design choice that may require tuning for different input resolutions or architectures. Finally, while our method is evaluated primarily in the context of SAM-based segmentation, its effectiveness for other vision tasks or transformer architectures remains to be fully explored.

These limitations suggest several promising directions for future work. An interesting extension would be to learn adaptive or task-specific gradient operators that retain the efficiency of first-order information while improving robustness. In addition, dynamically adjusting cell structures or sampling strategies based on content or model depth could further enhance flexibility. More broadly, we believe that StructSAM opens up new opportunities for rethinking token merging as a *local, structure-driven* problem rather than a global graph optimization task, and we hope this perspective will inspire more efficient designs for scalable vision transformers.

B. Implementation Details

B.1. Pseudo Code

We provide pseudocode for StructSAM's token partitioning procedure in Algorithm 1.

Algorithm 1 StructSAM token partitioning and merge map (cell-wise)

Input: Token tensor $T \in \mathbb{R}^{H\times W\times C}$; cell size s_x, s_y; merge rate $r \in [0, 1)$
Procedure: Source tokens $\mathbf{A}$ (to be merged); kept tokens $\mathbf{B}$; assignment map π for unmerging

Initialize empty sets $\mathbf{A}, \mathbf{B}$ and empty map π
Partition T into non-overlapping cells $\{C_i \in \mathbb{R}^{s_x\times s_y\times C}\}$
for each cell C_i **do**
    Compute gradient magnitudes $G(t)$ for tokens $t \in C_i$
    Compute cell flatness $\phi_i \leftarrow -\max_{t\in C_i} G(t)$
end for
Sort cells by ϕ_i in decreasing order ▷ Higher ϕ_i = flatter (more mergeable)
$M_{\text{merge}} \leftarrow \left\lceil \dfrac{r\cdot H\cdot W}{s_x s_y - 1} \right\rceil$ ▷ #cells needed to remove rHW tokens
$\mathcal{M} \leftarrow$ first M_{merge} cells in the sorted list ▷ mergeable cells
$\mathcal{P} \leftarrow$ remaining cells ▷ protected cells
for each protected cell $C_i \in \mathcal{P}$ **do**
    $\mathbf{B} \leftarrow \mathbf{B} \cup C_i$ ▷ keep all tokens
end for
for each mergeable cell $C_i \in \mathcal{M}$ **do**
    Compute $G(t)$ for tokens $t \in C_i$
    $t_{\text{dst}} \leftarrow \arg\min_{t\in C_i} G(t)$ ▷ stable destination token
    $\mathbf{B} \leftarrow \mathbf{B} \cup \{t_{\text{dst}}\}$
    for each token $t \in C_i \setminus \{t_{\text{dst}}\}$ **do**
        $\mathbf{A} \leftarrow \mathbf{A} \cup \{t\}$
        $\pi(t) \leftarrow t_{\text{dst}}$ ▷ unmerge target
    end for
end for
return $\mathbf{A}, \mathbf{B}, \pi$

Bounding Box Generation Following the standard SAM evaluation protocol, we use bounding box prompts derived from ground truth segmentation masks. For each ground truth mask, we compute the tight axis-aligned bounding box by extracting the minimum and maximum coordinates of foreground pixels (threshold ¿ 128). The resulting boxes are provided to SAM in the format of top-left and bottom-right corners as spatial prompts. This deterministic box generation ensures reproducible evaluation while simulating realistic user-provided region annotations. For our prompt-aware token merging

strategy, pixel-space boxes are converted to token-space coordinates by dividing by the patch size (16 pixels), enabling differentiated merging policies for tokens inside versus outside the prompted region.

C. PCA Visualization

We visualize the feature space using PCA projections across different layers of SAM-B in Fig. 6 and Fig. 8. Even with 65% of the tokens merged, the features remain faithful and preserve fine-grained forground object details.

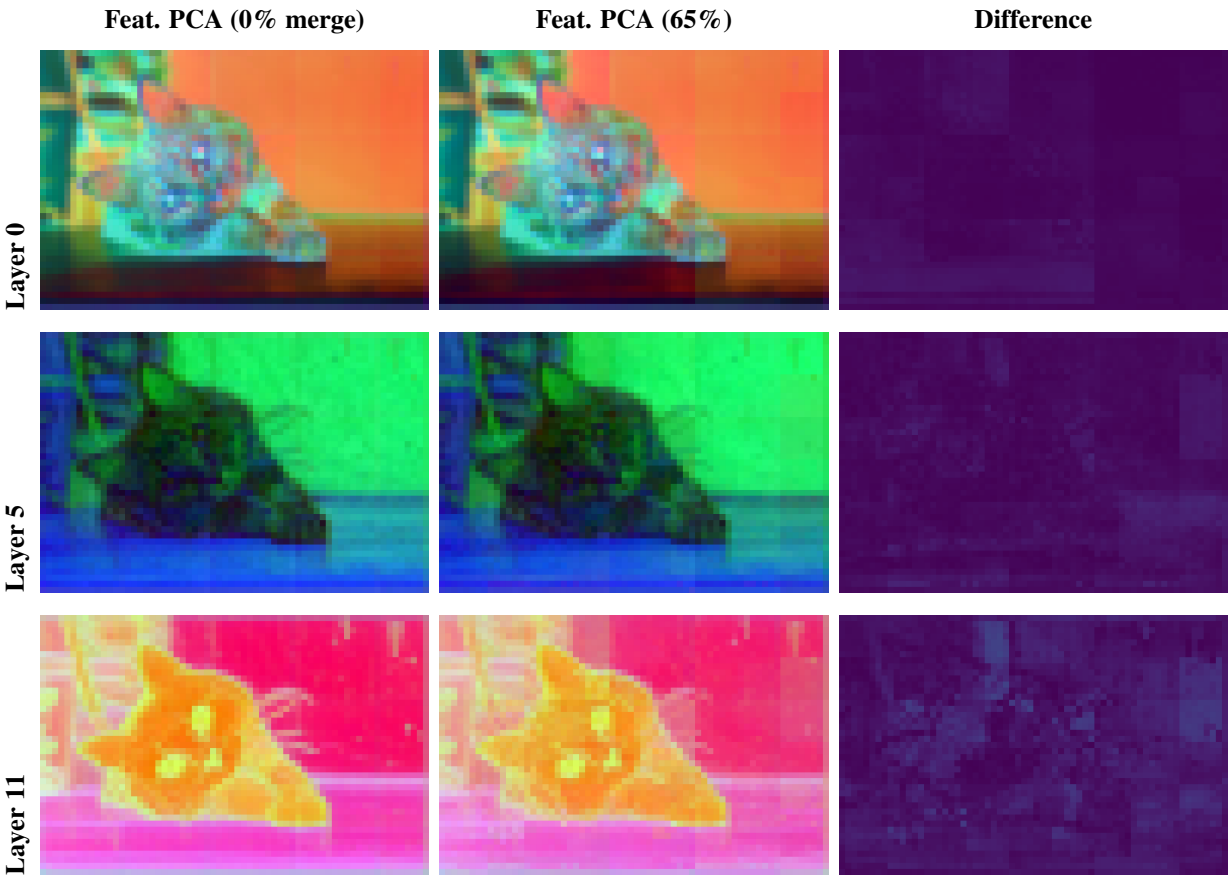


Figure 6. PCA visualizations across image encoder layers.

D. When the StructSAM energy will fail

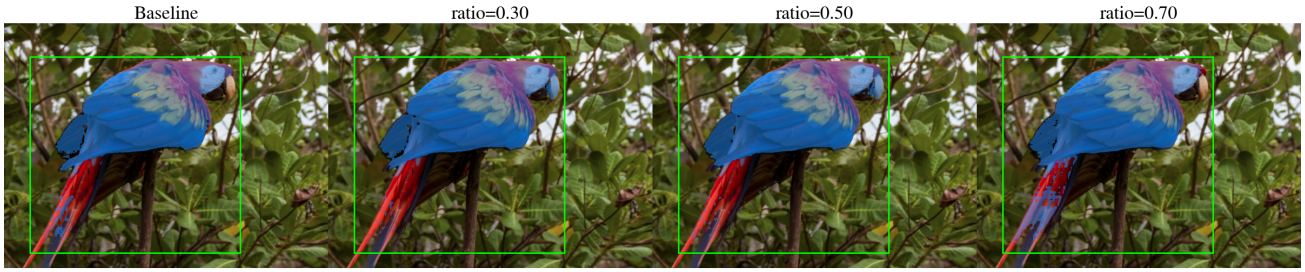
StructSAM becomes more unstable when processing regions where the original model exhibits low confidence or high ambiguity induced by the prompt. The top two rows of Fig. 7 illustrate this failure mode: the tail of the parrot is an ambiguous region, where even slight changes in the bounding box can lead to large variations in the original SAM output. In such cases, StructSAM shows increased instability across different merging rates.

Another failure case arises from StructSAM's tendency to prioritize merging background features while preserving highly distinctive and salient foreground objects. Although this behavior can improve segmentation quality for foreground regions, it makes background segmentation more challenging, as shown in the last row of Fig. 7 and the last row of Fig. 8.

E. Additional Ablation Studies

Cell Size : Table 5 presents an ablation study on cell size. To ensure compatibility, we select cell sizes that are divisible by the SAM model's window size. Our results indicate that 2×2 cells yield the highest quality, as they offer superior spatial preservation compared to larger configurations.

(a) The baseline models exhibit low confidence around the parrot's tail region; consequently, token merging becomes inconsistent across different merging rates.



(b) The baseline models exhibit low confidence around the parrot's tail region; even a slight adjustment of the bounding box leads to significant changes in the segmentation output..

(c) Leaf segmentation with token merging. Background regions are more easily merged due to weaker feature importance.

Figure 7. Segmentation results of SAM under different token merging rates. The parrot examples show that token merging becomes unstable in regions where the model has low confidence, particularly around challenging areas such as the tail. In contrast, the leaf example illustrates that non-foreground regions are more aggressively merged due to their lower feature saliency, which can lead to segmentation failures.

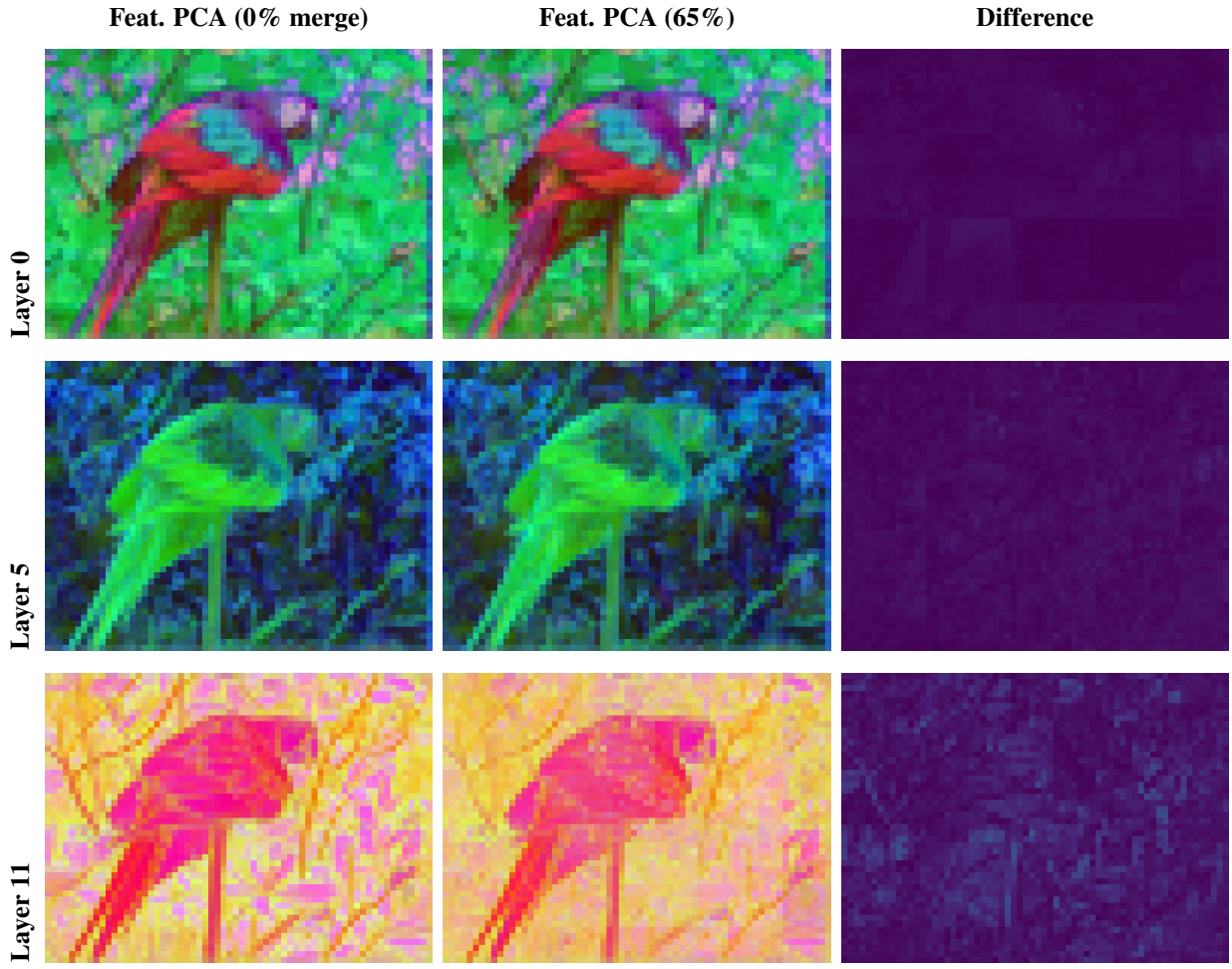


Figure 8. PCA visualizations across image encoder layers.

Additional ablation on the COIFT dataset: We further extend the ablation study presented in the main paper to the COIFT dataset in Table 4, providing additional evidence to support the design choices adopted in the final model.

Table 4. Ablation study on SAM-B. mIoU and boundary IoU (B-IoU) are reported in %. GradCell (Full) denotes the complete model using Sobel-based flatness and structured cell sampling.

		$r = 0.35$		$r = 0.55$	
Dataset	Method	mIoU	B-IoU	mIoU	B-IoU
COIFT	GradCell (Full)	**64.0**	**53.5**	**63.3**	**52.1**
	Central-Diff	60.5	50.4	59.5	49.1
	Mean-Flatness	63.8	53.3	62.9	52.1
	No-Cell	62.7	50.6	59.2	47.7
	Rand-Cell	63.9	52.5	63.0	51.1
	Max-Dst	63.3	52.7	62.0	51.4
	Rand-Dst	63.7	53.2	62.8	51.9

Table 5. Effect of cell size on INbreast dataset segmentation performance. The image encoder processes tokens on a 64×64 grid for global attention layers and a 14×14 grid for window attention layers, so the cell sizes must be divisors of their respective grid dimensions.

Window Cell Size	**Global Cell Size**	**Dice Score**	**GFLOPs**
2×2	2×2	0.7551	348.3
7×7	8×8	0.7481	347.8
14×14	16×16	0.7364	347.8

F. Extension StructSAM to Efficient-SAM for Video Tracking

We extend StructSAM to Efficient-SAM to enable lightweight video tracking by leveraging token merging for improved efficiency. While merging may introduce minor accuracy degradation, the resulting segmentation quality remains sufficient for tracking scenarios that rely on coarse prompts such as bounding boxes. This design prioritizes speed and scalability, making it suitable for real-time applications. Our goal is to evaluate whether a StructSAM-enhanced Efficient-SAM can serve as a practical alternative to more powerful but computationally intensive models such as SAM-2 for video tracking.

We illustrate the tracking quality of EfficientSAM in Fig. 9. Despite its simplicity, the method with a 70% merging rate performs on par with SAM2 while requiring significantly less memory and computation. Fig. 10 shows the gains in throughput and memory consumption when applying StructSAM to EfficientSAM; at a 70% merging rate, the system achieves real-time performance at 30 frames per second. Fig. 11 demonstrates that replacing SAM2 with EfficientSAM (70% merging rate) in our tracking pipeline maintains comparable performance on the robot stacking task with IA-VLA, while substantially improving efficiency and speed.

Extension for efficient video object tracking StructSAM extends to video co-tracking and segmentation by propagating masks across frames while restricting computation to relevant regions. For a video sequence I_i, the bounding box ∂_i derived from the previous mask M_{i-1} is used as a prompt to segment the current frame, yielding $M_i = \text{SAM}(I_i, \partial_i)$. Computation is focused within ∂_i, while tokens outside are merged or skipped for efficiency. The next bounding box is updated via $\partial_{i+1} = \mathcal{B}(M_i)$, enabling iterative mask propagation. This region-focused strategy allows efficient and accurate segmentation over time by leveraging temporal consistency between consecutive frames.

Quantitative analysis of the effect of StructSAM on EfficientSAM In this part, we assess the change in performance of EfficientSAM when it is extended by StructSAM. We illustrate the tracking performance in a robotic task in Figure 12. For evaluation, we compare the bounding box and segmentation results of EfficientSAM tracking with those of EfficientSAM + StructSAM tracking (using the algorithm described above), in terms of mIoU over a 15s video, as shown in Fig. 12. Table 6 demonstrate the quantitative results. Even with the high merging ratio, StructSAM still shows comparative results to

Tracking using EffecicientSAM with our algorithm at 0.7 merging rate

Tracking with SAM 2

Figure 9. **Results.** We compare TrackSAM (with a merge rate of 0.70) against the base model EfficientSAM and a SAM2. While the overall performance is similar, our method enables real-time processing with lower memory consumption.

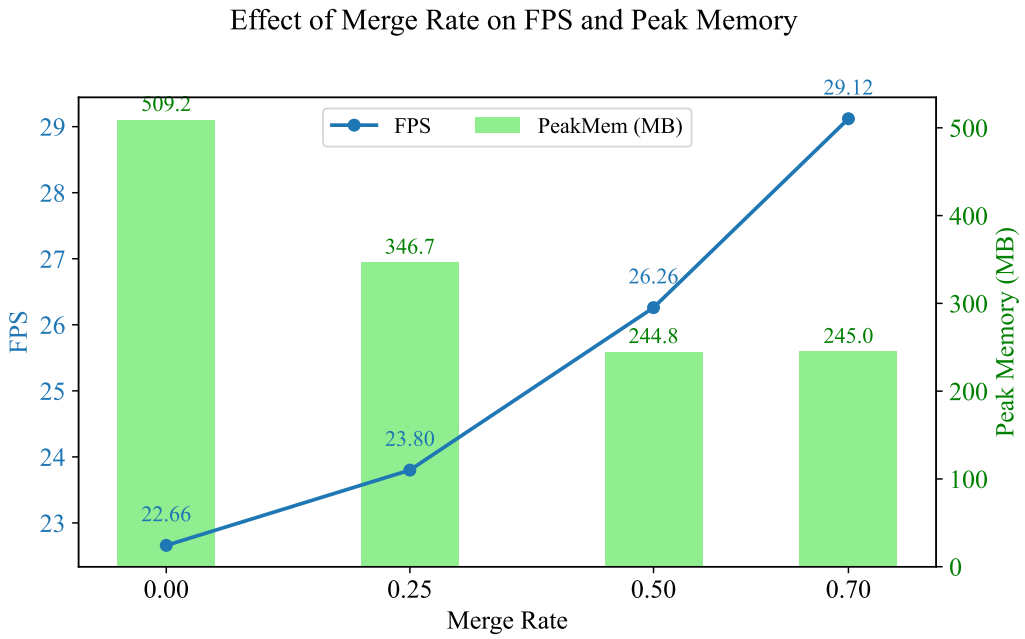


Figure 10. At 70% merging rate, we boost up robot execution speed to 3x while still maintaining success rate

original EfficientSAM, with Fig. 12 showing the success of StructSAM at merging ratio of 0.8. Even though the task is simple, which explains the mIoU and bbox IoU even at high merging ratio, it shows that in many cases StructSAM can be used to speed up an already efficient algorithm significantly while still maintaining decent performance.

Table 6. Segmentation and bounding box quality of applying StructSAM on EfficientSAM

Merging rate	mIoU	Bbox IoU
$R = 0.0$ (baseline)	0.920	0.916
$R = 0.6$	0.929	0.942
$R = 0.7$	0.928	0.942
$R = 0.8$	0.931	0.947
$R = 0.9$	0.922	0.932

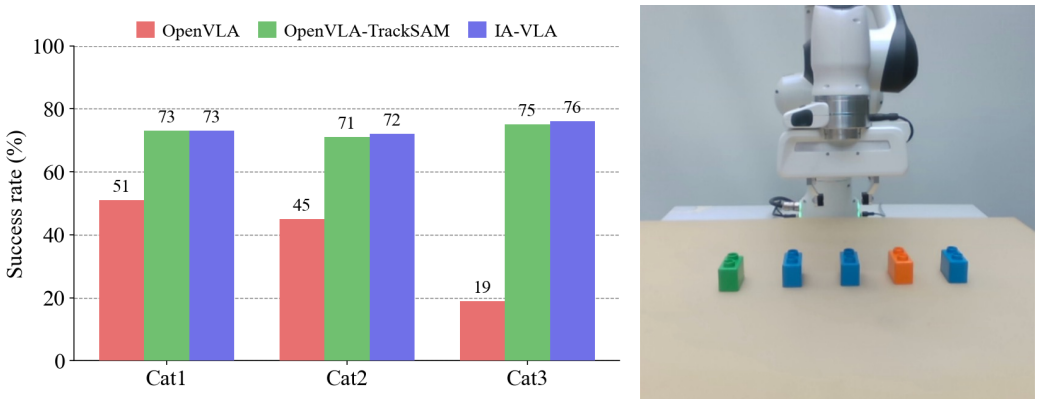
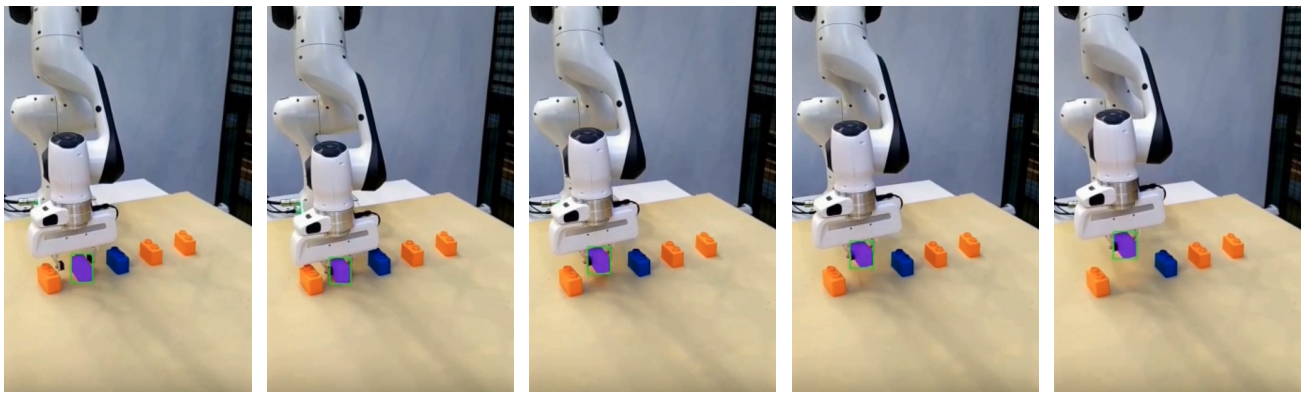


Figure 11. At 70% merging rate, our method (Green) boost up robot execution speed to 1.4x while keeping the same performance as IA-VLA (Blue).



Frame 1 Frame 30 Frame 60 Frame 90 Frame 120

Figure 12. Segmentation and tracking across video frames, merging rate at 0.8

Robot Setup for VLA experiments To assess the benefits of StructSAM in downstream robotic applications, we integrate it into Efficient-SAM for object tracking within vision-language-action (VLA) pipelines. We then evaluate its performance through real-world experiments designed to test robustness against distractor objects. In these experiments, a Franka Research 3 robot is instructed to pick up the correct Lego block based on a human prompt (see Figure 13), requiring both semantic and visual reasoning to accurately identify the target.

The language instructions follow structured patterns such as *"lift the leftmost / rightmost orange / green / blue block"* and *"lift the second / third / fourth / fifth orange / green / blue block from the left / right."*

We collect 120 demonstrations across 12 distinct language instructions covering a subset of the instructions above. Each scene is defined by the number, color, and spatial arrangement of the blocks, with semantic concepts grounded in color and positional references.

To evaluate generalization of the VLA models, we divide tasks into three difficulty categories. Category 1 includes instructions seen during training. Category 2 combines familiar positional references with novel color assignments. Category 3 introduces previously unseen ways of referring to object positions.

We use OpenVLA with raw robot observations as a baseline. To mitigate the impact of distractors, inspired by (Hannus et al., 2025), we augment the input with a highlighted mask of the target object, generated using StructSAM-enhanced Efficient-SAM and SAM2. We then compare baseline performance against OpenVLA augmented with SAM2 and with StructSAM-enhanced Efficient-SAM. The results show that input augmentation consistently improves performance across all task categories. In addition, StructSAM integrated into Efficient-SAM achieves task success rates comparable to SAM2 while delivering a 45% speedup, highlighting an effective balance between efficiency and performance in both robotic and

medical imaging contexts.

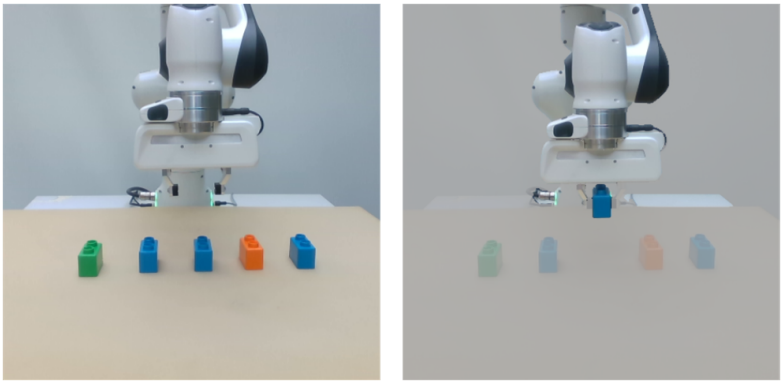


Figure 13. Task: *"lift the second blue block from the right"*. The raw robot observation is on the left, and the augmented observation with propagated masks at the end of the task is at the right.

G. Additional Results

We present in Figure 14 a visualization of token-merging outputs from different algorithms on a representative image from the INbreast dataset. Additionally, we provide a detailed comparison between StructSAM and baseline methods across different merging ratios on four high-quality datasets and two SAM variants (ViT-B and ViT-L backbones), as shown in Figs. 15, 16, and 17. We show additional analysis for peak GPU memory consumptions from Figs. 18 and 19.

In Table 7, we additionally report segmentation performance under point-prompt evaluation. We follow the original SAM evaluation protocol, where points are sampled from the ground-truth mask and used sequentially to progressively refine the predicted segmentation.

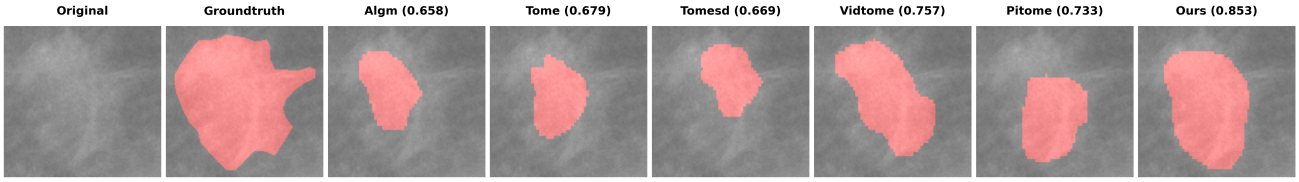


Figure 14. Segmentation results on INbreast dataset using MedSAM

Table 7. Point-prompt segmentation results across all datasets (model: sam-b, merge ratio: 0.55). **Green** = best, blue = second-best per column.

Dataset	Method	1 Point		2 Points		3 Points	
		mIoU	b-mIoU	mIoU	b-mIoU	mIoU	b-mIoU
DIS5K	Baseline (No Merging)	15.98	13.58	34.54	**29.10**	**47.86**	**39.94**
	ToMeSD	16.72	13.81	34.21	28.08	46.43	37.99
	ALGM	13.09	10.11	30.76	23.94	43.14	33.68
	vidtome	11.52	9.59	26.61	21.74	39.75	31.90
	StructSAM (Ours)	**17.03**	**13.96**	**34.65**	28.35	47.57	38.89
ThinObject5K	Baseline (No Merging)	46.70	39.02	69.04	58.54	**78.43**	67.88
	ToMeSD	50.04	42.05	70.84	59.89	77.69	67.50
	ALGM	48.60	40.32	69.97	59.87	77.54	66.83
	vidtome	40.92	31.80	60.63	47.10	70.75	56.45
	StructSAM (Ours)	**52.54**	**44.40**	**71.54**	**60.79**	78.43	**67.95**
HRSOD	Baseline (No Merging)	47.33	39.67	66.18	56.37	75.68	64.88
	ToMeSD	49.92	42.08	66.60	56.99	75.77	65.13
	ALGM	47.46	38.79	64.82	54.13	73.96	62.51
	vidtome	36.96	29.02	54.68	43.30	68.00	53.84
	StructSAM (Ours)	**52.07**	**44.10**	**68.62**	**58.74**	**76.34**	**65.90**
COIFT	Baseline (No Merging)	37.51	29.52	58.08	45.57	**69.73**	**55.15**
	ToMeSD	40.29	31.14	58.67	45.62	69.05	54.31
	ALGM	38.81	28.56	56.23	41.54	66.16	49.18
	vidtome	27.73	19.29	46.02	31.97	57.80	40.46
	StructSAM (Ours)	**41.42**	**31.97**	**59.32**	**46.26**	69.64	54.92

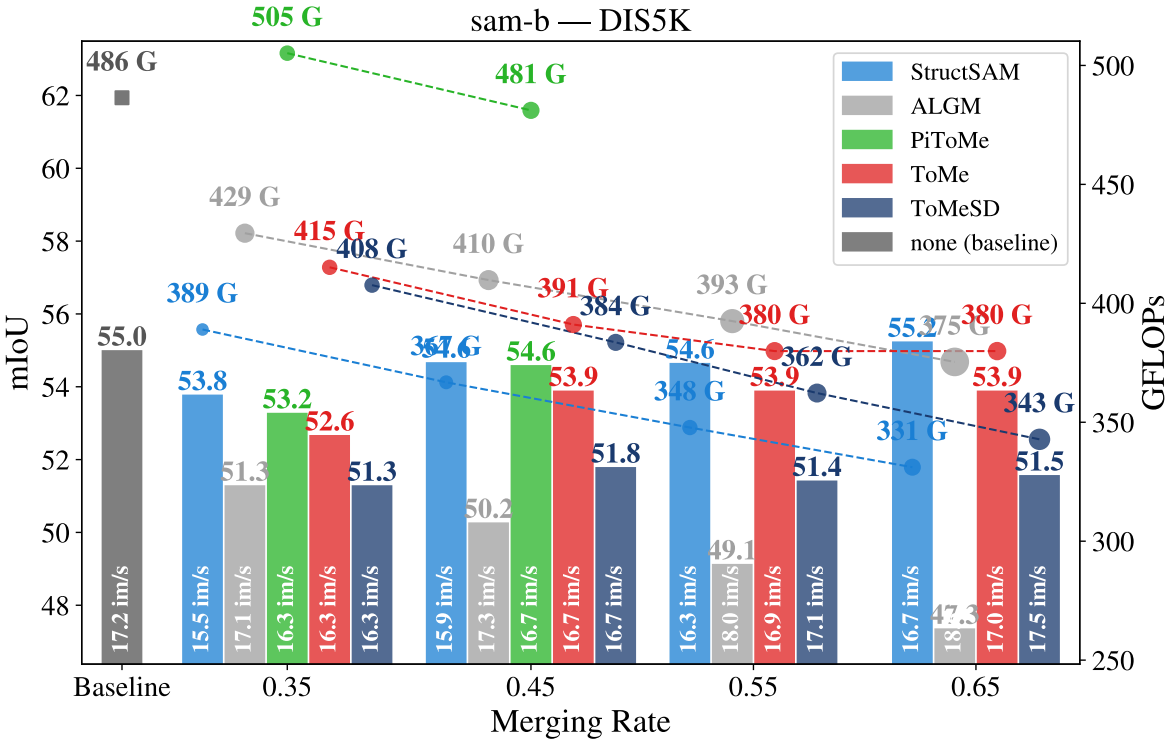


Figure 15. Results on DIS5K dataset, SAM with ViT-B backbone.

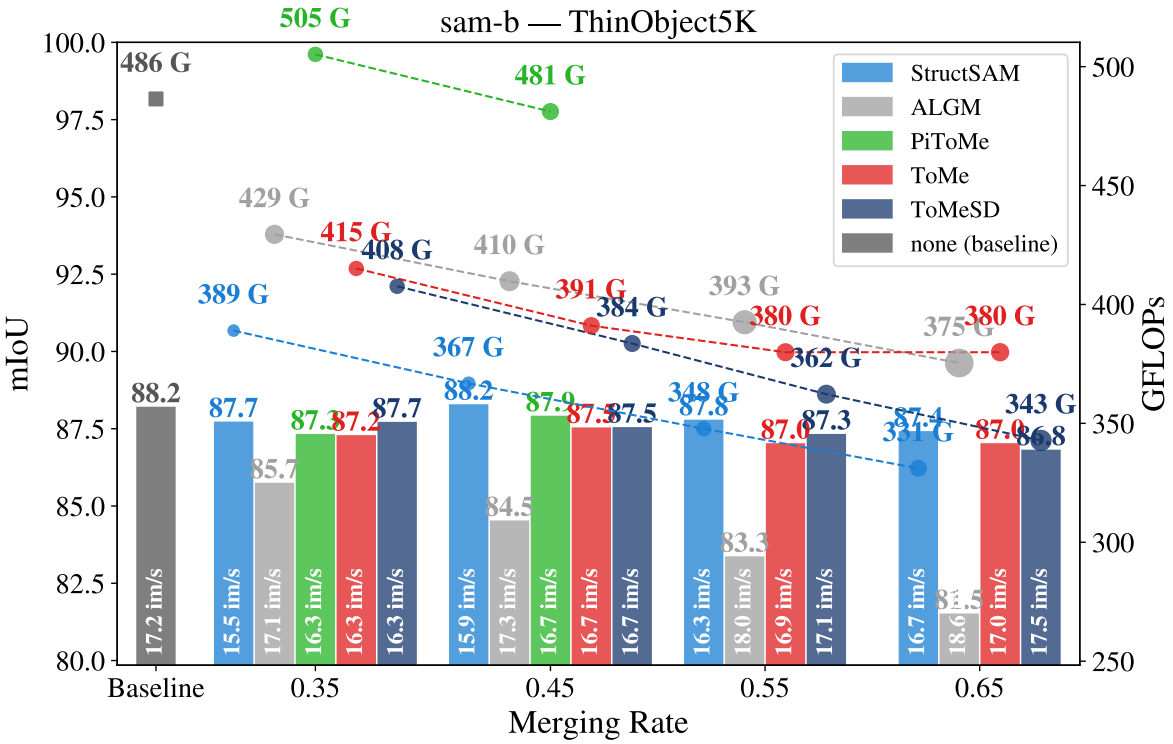


Figure 16. Results on ThinObject5K dataset, SAM with ViT-B backbone.

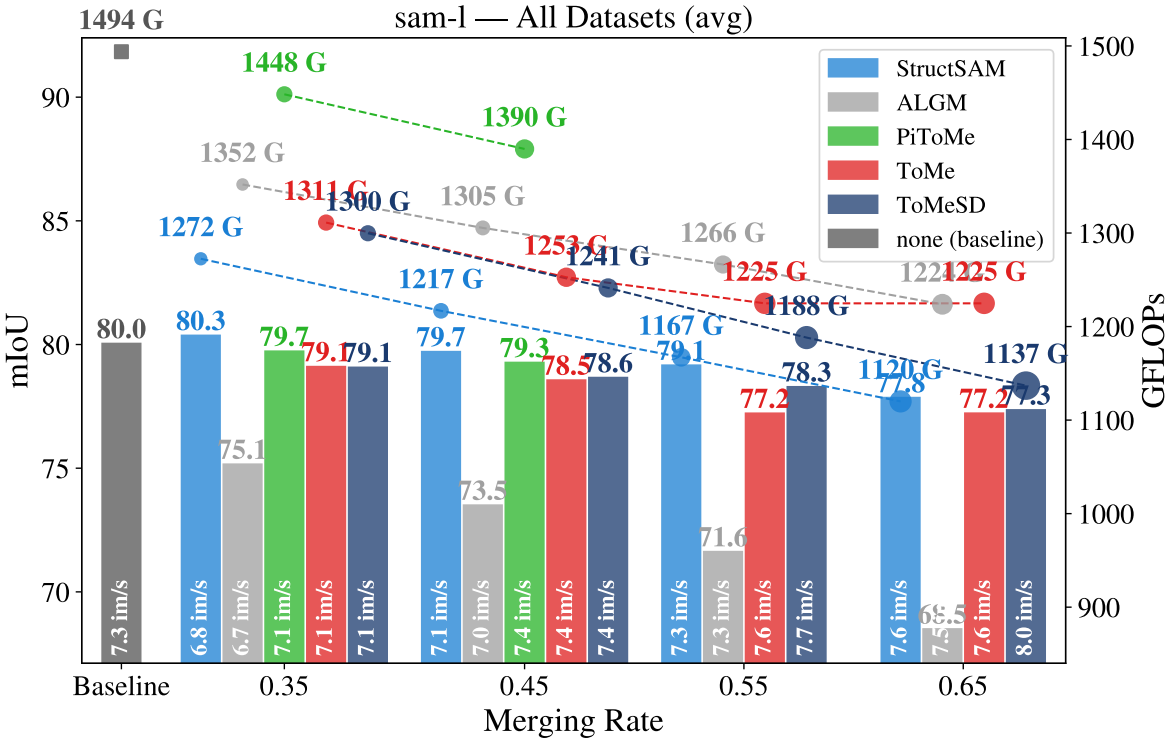


Figure 17. Average results on all 4 datasets, SAM with ViT-L backbone.

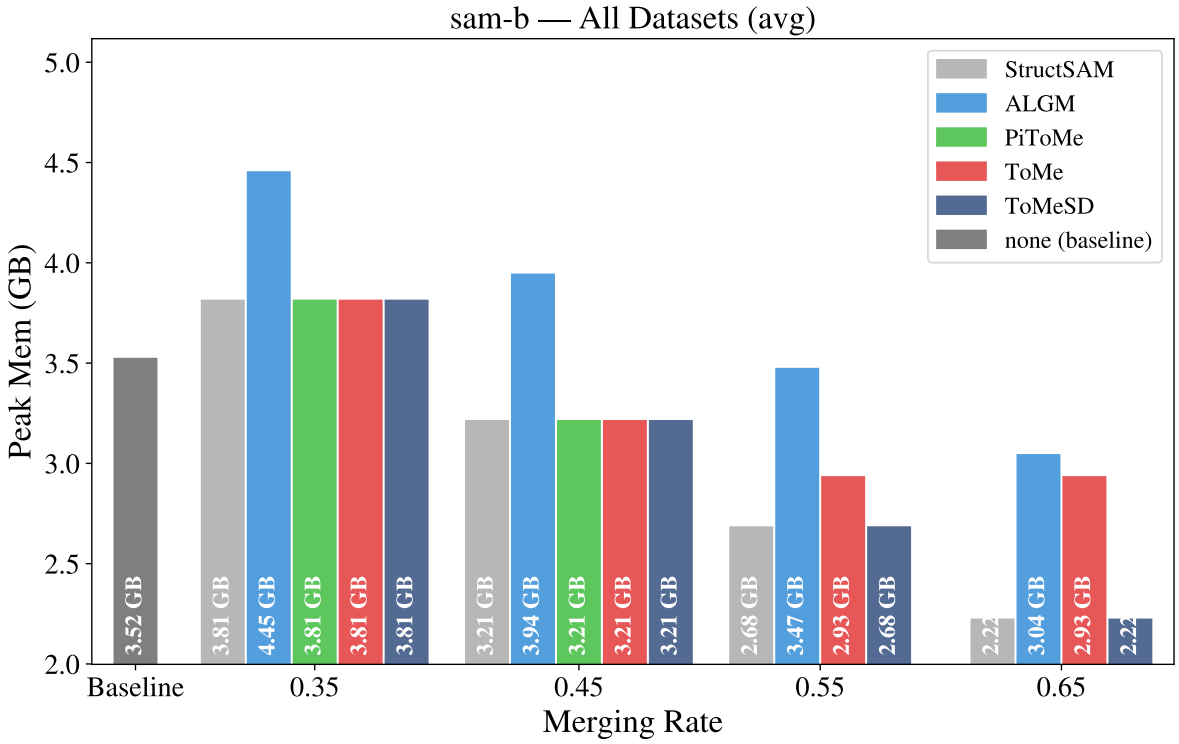


Figure 18. Average peak memory results on all 4 datasets, SAM with ViT-b backbone. Note that ToMe does not applicable at above 50% merging rate, we note the memory consumption of ToMe at 50% for 0.55 and 0.65

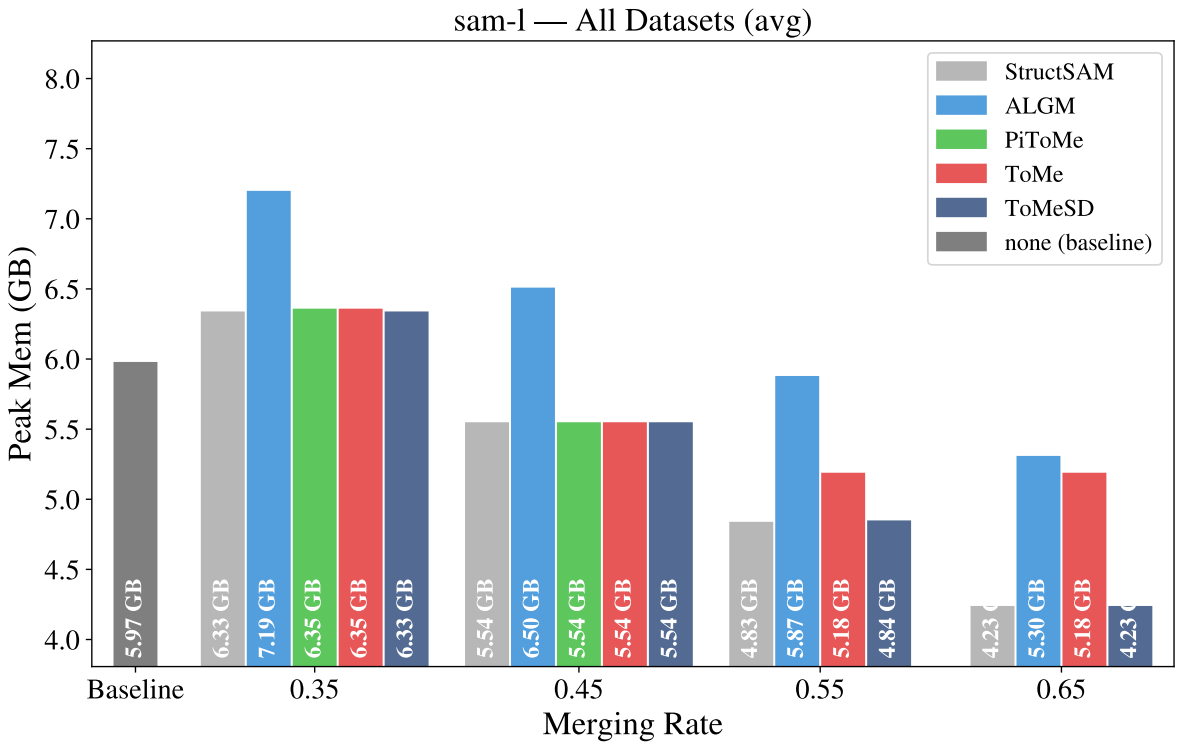


Figure 19. Average peak memory results on all 4 datasets, SAM with ViT-l backbone. Note that ToMe does not applicable at above 50% merging rate, we note the memory consumption of ToMe at 50% for 0.55 and 0.65

H. Formal Proofs and Statements of Theoretical Guarantee for Section 4

This appendix provides formal statements and proofs supporting the informal Theorem 1 in the main paper. We analyse StructSAM's token merging inside an attention block through spectral graph theory: merging induces a graph coarsening map on a token graph, while the unmerging step corresponds to a canonical lifting that restores the original token resolution required by dense mask prediction.

Layerwise spectral discrepancy (main paper). At encoder layer ℓ, tokens in each attention window $\mathcal{P}_{\ell,k}$ define a weighted graph $\mathcal{G}_{\ell,k}$ with normalized Laplacian $\mathcal{L}_{\ell,k}$. After merging and lifting, we obtain a lifted graph $\mathcal{G}_{\ell,k,l}$ and its Laplacian $\mathcal{L}_{\ell,k,l}$. We measure structural distortion at layer ℓ via

$$\mathrm{SD}_\ell \triangleq \sum_{k=1}^{K_\ell} \left\| \boldsymbol{\lambda}_{\ell,k} - \boldsymbol{\lambda}_{\ell,k,l} \right\|_1,$$

where $\boldsymbol{\lambda}_{\ell,k}$ and $\boldsymbol{\lambda}_{\ell,k,l}$ are eigenvalues of the original and lifted normalized Laplacians, respectively.

H.1. Preliminaries: coarsening, lifting, and eigenvalue inclusion

We reuse *Graph Coarsening* and *Graph Lifting* (Definitions 1 and 2) window-wise to interpret token merging as a graph coarsening operation and to define the lifted proxy used in our spectral analysis.

Definition 1 (Graph Coarsening)**.** *Let $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathbf{W})$ be a weighted graph with $|\mathcal{V}| = N$ and adjacency matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$. Let $\mathcal{P} = \{\mathcal{V}_i\}_{i \in [n]}$ be a partition of $\mathcal{V}$ into n disjoint subsets. The* coarsened graph *of $\mathcal{G}$ with respect to $\mathcal{P}$ is the weighted graph $\mathcal{G}_c(\mathcal{V}_c, \mathcal{E}_c, \mathbf{W}_c)$, where each subset $\mathcal{V}_i$ is collapsed into a single node $\nu_i \in \mathcal{V}_c$. Its adjacency entries are defined by block-averaging:*

$$\mathbf{W}_c[i,j] \;=\; \frac{1}{|\mathcal{V}_i|\,|\mathcal{V}_j|} \sum_{u \in \mathcal{V}_i} \sum_{v \in \mathcal{V}_j} \mathbf{W}[u,v], \qquad i,j \in [n].$$

Let $\mathbf{D}$ *be the degree matrix of* $\mathcal{G}$ *with* $\mathbf{D}[p,p] = d_p := \sum_{q=1}^{N} \mathbf{W}[p,q]$*, and define the combinatorial and normalized Laplacians:*

$$\mathbf{L} = \mathbf{D} - \mathbf{W}, \qquad \mathcal{L} = \mathbf{I}_N - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}.$$

Similarly define $\mathbf{D}_c$*,* $\mathbf{L}_c = \mathbf{D}_c - \mathbf{W}_c$*, and* $\mathcal{L}_c = \mathbf{I}_n - \mathbf{D}_c^{-1/2}\mathbf{W}_c\mathbf{D}_c^{-1/2}$ *for* $\mathcal{G}_c$*. We denote eigenvalues and eigenvectors of* $\mathcal{L}$ *by* $(\boldsymbol{\lambda}, \mathbf{u})$ *and those of* $\mathcal{L}_c$ *by* $(\boldsymbol{\lambda}_c, \mathbf{u}_c)$*.*

Definition 2 (Graph Lifting)**.** *Given a coarsened graph* $\mathcal{G}_c(\mathcal{V}_c, \mathcal{E}_c, \mathbf{W}_c)$ *induced by partition* $\mathcal{P} = \{\mathcal{V}_i\}_{i\in[n]}$*, the* lifted graph $\mathcal{G}_l(\mathcal{V}, \mathcal{E}_l, \mathbf{W}_l)$ *is defined on the original node set* $\mathcal{V}$ *by*

$$\mathbf{W}_l[u,v] \;=\; \mathbf{W}_c[i,j], \qquad \forall\, u \in \mathcal{V}_i,\ v \in \mathcal{V}_j,\ i,j \in [n].$$

Let $\mathbf{D}_l$ *be the degree matrix of* $\mathcal{G}_l$ *with* $\mathbf{D}_l[p,p] = d_{lp} := \sum_{q=1}^{N} \mathbf{W}_l[p,q]$*, and define*

$$\mathbf{L}_l = \mathbf{D}_l - \mathbf{W}_l, \qquad \mathcal{L}_l = \mathbf{I}_N - \mathbf{D}_l^{-1/2}\mathbf{W}_l\mathbf{D}_l^{-1/2}.$$

We denote eigenvalues and eigenvectors of $\mathcal{L}_l$ *by* $(\boldsymbol{\lambda}_l, \mathbf{u}_l)$*.*

Lemma 1 (Eigenvalue inclusion under lifting)**.** *Let* $\mathcal{G}_c$ *be coarsened from* $\mathcal{G}$ *by a partition* $\mathcal{P}$*, and let* $\mathcal{G}_l$ *be the lifted graph. Then the eigenvalues of* $\mathcal{L}_l$ *contain all eigenvalues of* $\mathcal{L}_c$*, and the remaining* $(N-n)$ *eigenvalues equal* 1*.*

Proof. A standard lifting argument shows that $\mathcal{L}_l$ has an invariant subspace of vectors constant on each cluster, on which $\mathcal{L}_l$ is similar to $\mathcal{L}_c$. The orthogonal complement contributes eigenvalue 1 with multiplicity $N - n$. □

H.2. Token graphs for windowed and global attention

Fix an encoder layer ℓ with window partition $\{\mathcal{P}_{\ell,k}\}_{k=1}^{K_\ell}$, where $K_\ell \geq 1$. For local-attention layers, $\mathcal{P}_{\ell,k}$ are spatial windows; for global attention layers, $K_\ell = 1$ and $\mathcal{P}_{\ell,1} = \mathcal{X}$. For each window (ℓ, k), define a weighted token graph $\mathcal{G}_{\ell,k}(\mathcal{V}_{\ell,k}, \mathcal{E}_{\ell,k}, \mathbf{W}_{\ell,k})$ on the tokens in $\mathcal{P}_{\ell,k}$. Let its normalized Laplacian be

$$\mathcal{L}_{\ell,k} \;=\; \mathbf{I}_{N_{\ell,k}} - \mathbf{D}_{\ell,k}^{-1/2}\mathbf{W}_{\ell,k}\mathbf{D}_{\ell,k}^{-1/2}, \qquad N_{\ell,k} = |\mathcal{V}_{\ell,k}|.$$

StructSAM performs merging within each window (equivalently, within grid cells aligned to the window), producing a coarsened graph $\mathcal{G}_{\ell,k,c}$; lifting yields $\mathcal{G}_{\ell,k,l}$ on $N_{\ell,k}$ nodes. Let $\boldsymbol{\lambda}_{\ell,k}$ and $\boldsymbol{\lambda}_{\ell,k,l}$ be eigenvalues of $\mathcal{L}_{\ell,k}$ and $\mathcal{L}_{\ell,k,l}$, respectively.

H.3. Merge correctness event

Fix (ℓ, k). Let $\mathcal{P}_{0,\ell,k}^{(s)} = \{\mathcal{V}_{0,\ell,k,1}^{(s)}, \ldots, \mathcal{V}_{0,\ell,k,s_{\ell,k}}^{(s)}\}$ be an (unknown) semantic partition of the window at merge step s (e.g., local regions separated by edges). At step s, the algorithm merges a pair $(v_{a_{\ell,k,s}}, v_{b_{\ell,k,s}})$ inside that window. Define the within-region merge event

$$\mathcal{E}_{\ell,k,s} \;\triangleq\; \left\{\exists i \in [s_{\ell,k}] \text{ s.t. } v_{a_{\ell,k,s}}, v_{b_{\ell,k,s}} \in \mathcal{V}_{0,\ell,k,i}^{(s)}\right\}. \tag{1}$$

H.4. Assumptions

We keep Assumptions 1 and 2 identical in spirit to PiToME's Theorem 1 and establish a gradient-separation Assumption 3 that matches StructSAM's flatness screening rule and implies $\delta_{\ell,k,s} \to 0$.

Assumption 1 (Within-region concentration)**.** *For each merge step* s *and each true part* $\mathcal{V}_{0,\ell,k,i}^{(s)}$*,*

$$\mathbb{E}\big[\cos(v_u, v_v)\big] \to 1, \qquad \forall\, v_u, v_v \in \mathcal{V}_{0,\ell,k,i}^{(s)}.$$

Assumption 2 (Margin across regions)**.** *There exists a margin* $m \in (0,1)$ *such that for all* $i \neq j$*,*

$$\cos(v_u, v_v) \;\geq\; m \;>\; \cos(v_u, v_w), \qquad \forall\, v_u, v_v \in \mathcal{V}_{0,\ell,k,i}^{(s)},\ \forall\, v_w \in \mathcal{V}_{0,\ell,k,j}^{(s)}.$$

Assumption 3 (Gradient separation for flatness screening)**.** *Fix a layer ℓ and an attention window $\mathcal{P}_{\ell,k}$. Partition $\mathcal{P}_{\ell,k}$ into disjoint grid cells $\{\mathcal{C}_{\ell,k,m}\}_{m=1}^{M_{\ell,k}}$ aligned to window geometry, each of size $s \times s$. Let $S_\ell(x) = \mathbf{G}^{(\ell)}(x)$ be the feature gradient–based energy score. Define the cell flatness score*

$$\phi_{\ell,k}(\mathcal{C}_{\ell,k,m}) \;:=\; -\max_{x \in \mathcal{C}_{\ell,k,m}} S_\ell(x).$$

StructSAM selects the mergeable-cell set $\mathcal{M}^{\mathrm{mer}}_{\ell,k}$ as the $\rho M_{\ell,k}$ cells with largest $\phi_{\ell,k}$, and defines the protected set $\mathcal{M}^{\mathrm{pro}}_{\ell,k} = [M_{\ell,k}] \setminus \mathcal{M}^{\mathrm{mer}}_{\ell,k}$.

Assume there exist constants $0 < \tau_{\mathrm{in}} < \tau_{\mathrm{bd}}$ and a function $\delta_{\ell,k}(s) \in (0,1)$ such that for each merge step s:

1. ***Boundary-gradient lower bound.*** *If a cell $\mathcal{C}_{\ell,k,m}$ intersects two or more true parts of $\mathcal{P}^{(s)}_{0,\ell,k}$, then*

$$\max_{x \in \mathcal{C}_{\ell,k,m}} S_\ell(x) \;\geq\; \tau_{\mathrm{bd}} \quad \textit{with probability at least } 1 - \delta_{\ell,k}(s).$$

2. ***Interior-gradient upper bound.*** *If a cell $\mathcal{C}_{\ell,k,m}$ is fully contained in a single true part of $\mathcal{P}^{(s)}_{0,\ell,k}$, then*

$$\max_{x \in \mathcal{C}_{\ell,k,m}} S_\ell(x) \;\leq\; \tau_{\mathrm{in}} \quad \textit{with probability at least } 1 - \delta_{\ell,k}(s).$$

3. ***Mergeable-cell budget.*** *The merge rate ρ satisfies*

$$\rho \;\leq\; \frac{\#\{\textit{interior cells in } (\ell,k)\}}{M_{\ell,k}},$$

so that mergeable cells can be chosen exclusively among interior cells whenever (A3a)–(A3b) hold.

Moreover, $\delta_{\ell,k}(s) \to 0$ as token resolution increases (equivalently, as cell size decreases at fixed image resolution).

Lemma 2 (Gradient separation implies boundary protection)**.** *Under Assumption 3, with probability at least $1 - 2\,\delta_{\ell,k}(s)$, every mergeable cell selected by StructSAM is an interior cell (i.e., it does not intersect a boundary). In particular, defining*

$$\delta_{\ell,k,s} \;:=\; 2\,\delta_{\ell,k}(s),$$

we have $\delta_{\ell,k,s} \to 0$.

Proof. On the event that (A3a) and (A3b) hold, every boundary cell has $\max S_\ell \geq \tau_{\mathrm{bd}}$ while every interior cell has $\max S_\ell \leq \tau_{\mathrm{in}}$, with $\tau_{\mathrm{in}} < \tau_{\mathrm{bd}}$. Hence all cells with smallest values of $\max S_\ell$ are interior. By (A3c), StructSAM can select its mergeable-cell budget from these interior cells. A union bound over the two failure events yields probability at least $1 - 2\,\delta_{\ell,k}(s)$. □

Lemma 3 (From gradient screening to merge correctness)**.** *Under Assumptions 2 and 3, for each merge step s,*

$$\mathbb{P}(\mathcal{E}_{\ell,k,s}) \;\geq\; 1 - \delta_{\ell,k,s}, \qquad \textit{where } \delta_{\ell,k,s} = 2\,\delta_{\ell,k}(s) \to 0.$$

Proof. By Lemma 2, with probability at least $1 - \delta_{\ell,k,s}$, all mergeable cells are interior (contained in a single true part). Within an interior cell, StructSAM chooses a destination token inside that same true part. By the margin assumption Assumption 2, BSM assigns each source token to a destination in the same true part, hence no cross-part merges occur and $\mathcal{E}_{\ell,k,s}$ holds. □

Proposition 1 (Failure of Assumption 3 under coarse cell partitions)**.** *There exist token graphs satisfying Assumptions 1 and 2 for which Assumption 3 fails solely due to the choice of cell size, even when the gradient score $S_\ell = \mathbf{G}^{(\ell)}$ perfectly separates boundary tokens from interior tokens.*

In particular, fix a window $\mathcal{P}_{\ell,k}$ whose tokens lie on a $H \times W$ grid, and consider a latent partition into two true parts separated by a boundary curve that intersects every grid cell of a given cell partition $\{\mathcal{C}_{\ell,k,m}\}_{m=1}^{M_{\ell,k}}$. Then $\#\{\textit{interior cells}\} = 0$, and thus the mergeable-cell budget condition (A3c) fails for any $\rho > 0$. Consequently, StructSAM must select mergeable cells that are boundary/mixed cells, so the boundary-exclusion conclusion of Assumption 3 cannot hold.

Proof. Construct embeddings as follows. Let all tokens in true part 1 share a unit vector u, and all tokens in true part 2 share a unit vector v such that $u^\top v = \gamma < m$. Then within-part cosine similarity equals 1, so Assumption 1 holds (trivially), and cross-part similarity equals $\gamma < m$, so Assumption 2 holds.

Now choose a cell partition with no interior cells, i.e. every cell intersects both true parts (for example, take a single cell covering the entire window, or take cells so large that each cell straddles the boundary). Then $\#\{\text{interior cells}\} = 0$, and the condition (A3c), $\rho \le \#\{\text{interior cells}\}/M_{\ell,k}$, fails for any $\rho > 0$. Hence StructSAM must choose at least one mergeable cell that is mixed. Even if the gradient score perfectly separates boundary tokens from interior tokens, there are no interior cells to select, so the boundary-exclusion mechanism cannot be satisfied. □

H.5. Practicality of the assumptions in Section H.4

Why Assumptions 1 and 2 are plausible in foundation encoders. Assumptions 1 and 2 posit that token embeddings concentrate within coherent regions and admit a margin across different regions. In SAM-style foundation encoders, this behaviour is empirically supported by the strong spatial organisation visible in PCA projections of token features. In Figure 6, the *Feat. PCA (0% merge)* panels exhibit piecewise-smooth colour structure that aligns with salient objects and backgrounds across early (Layer 0) through deeper layers (Layer 11), suggesting that tokens within a region occupy a compact neighbourhood in feature space, while tokens across different regions remain separated. This is consistent with within-region similarity concentration and cross-region margin, motivating Assumptions 1 and 2 in practice.

Decomposing Assumption 3: score separation vs. geometric budget. Assumption 3 decomposes into two ingredients.

Score separation (A3a)–(A3b). These conditions require that the cell-wise statistic $\max_{x\in\mathcal{C}} \mathbf{G}^{(\ell)}(x)$ separates boundary/mixed cells from interior cells. This is a standard edge-detection heuristic: a boundary-crossing cell typically contains at least one high-gradient token, while interior cells remain low-gradient. The PCA diagnostics in Figure 6 support this separation across layers even under aggressive merging. Specifically, comparing *Feat. PCA (0% merge)* and *Feat. PCA (65%)* shows that the dominant spatial structure is largely preserved, while the *Difference* maps remain sparse and localised rather than diffuse. This indicates that the merge–unmerge perturbation concentrates on a small subset of locations and does not globally scramble the feature geometry, making it plausible that gradient-energy remains a stable signal for identifying boundary-sensitive cells across layers.

Geometric/budget condition (A3c). Condition (A3c) is structural: the mergeable-cell ratio ρ (equivalently, the cell size s) must be chosen so that sufficiently many *interior cells* exist to populate the mergeable set. If s is too large (or ρ is too aggressive), interior cells may be absent or too few, forcing the selection of boundary/mixed cells and invalidating (A3c), as formalised by Proposition 1. In practice, StructSAM's use of window-aligned cells and moderate cell sizes typically yields many interior cells per window, so (A3c) is satisfied except in extreme high-merge regimes or for very thin structures whose width is comparable to the chosen cell size.

Takeaway. Together, Figure 6 and the sparse difference patterns across layers provide empirical support that the representation geometry of SAM's encoder makes Assumptions 1 and 2 and the score-separation component (A3a)–(A3b) easy to satisfy, while (A3c) highlights the practical trade-off between merge rate and the availability of interior cells under window-aligned partitioning.

H.6. Formal theorem matching the main-paper informal statement in Theorem 1

Theorem 2 (Formal: Layerwise spectrum stability of score-guided merging)**.** *Fix an encoder layer ℓ with windows $\{\mathcal{P}_{\ell,k}\}_{k=1}^{K_\ell}$. Let $SD_\ell(\mathrm{SG})$ denote the spectral discrepancy induced by StructSAM's score-guided merging, and $SD_\ell(\mathrm{Base})$ that of a non–score-guided baseline (e.g., random or stride-based dst selection). Assume bounded degrees and bounded weights in each window: there exist constants $0 < d_{\min} \le d_{\max} < \infty$ and $0 < w_{\max} < \infty$ such that, for all merge steps,*

$$d_{\min} \;\le\; d^{(s)}_{\ell,k}(i) \;\le\; d_{\max}, \qquad 0 \le \mathbf{W}^{(s)}_{\ell,k}[i,j] \le w_{\max}.$$

Then:

1. *Under Assumptions 1 to 3, we have $\mathbb{E}[SD_\ell(\mathrm{SG})] \to 0$.*

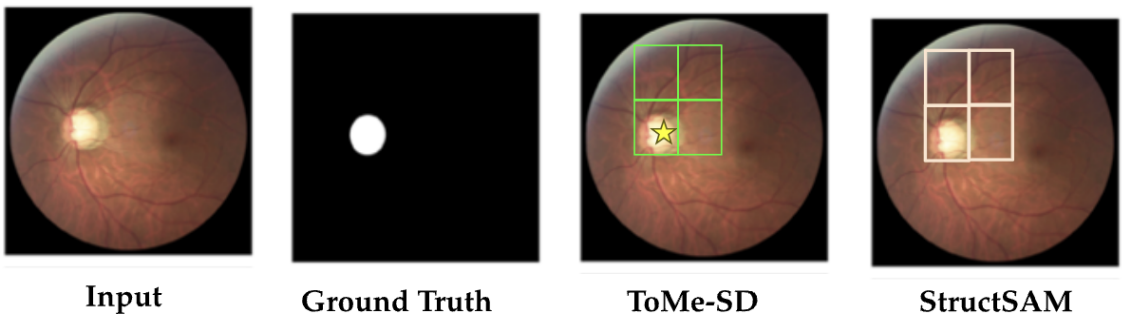


Figure 20. **Illustrative failure case of ToMeSD**. Stride-based destination selection may place a destination token inside a boundary-crossing (mixed) cell, forcing cross-region token merges that degrade dense segmentation quality. In the example shown for ToMe-SD, the destination token is sampled from the ground-truth cell, causing the algorithm to merge structurally inconsistent neighboring tokens and produce erroneous segmentation. In contrast, StructSAM uses flatness-based screening to protect mixed cells and preserve structural boundaries.

2. *If the baseline has a non-vanishing probability of cross-region merging, i.e., there exists* $\delta > 0$ *such that*

$$\inf_{k,s} \mathbb{P}(\mathcal{E}_{\ell,k,s}) \;\le\; 1 - \delta,$$

then $\liminf \mathbb{E}[SD_\ell(\mathrm{Base})] \ge c_0\,\delta > 0$ *for a constant* c_0 *depending only on* $(m, d_{\min}, d_{\max}, w_{\max})$.

H.7. A baseline counterexample (ToMeSD-style dst selection)

Proposition 2 (ToMeSD admits non-vanishing cross-region merges)**.** *Assume Assumptions 1 and 2. Consider a window* $\mathcal{P}_{\ell,k}$ *that contains two true parts* $\mathcal{V}^{(s)}_{0,\ell,k,1}$ *and* $\mathcal{V}^{(s)}_{0,\ell,k,2}$ *separated by a boundary that intersects at least one grid cell* $\mathcal{C}$ *(a* mixed cell*). Suppose the baseline destination-selection rule chooses a destination token in* $\mathcal{C}$ *with probability at least* $p_0 > 0$ *and, conditional on selecting* $\mathcal{C}$*, picks a destination from the non-dominant part in* $\mathcal{C}$ *with probability at least* $q_0 > 0$*. Then for infinitely many merge steps* s*,*

$$\mathbb{P}(\mathcal{E}^c_{\ell,k,s}) \;\ge\; p_0 q_0,$$

and hence $\inf_{k,s} \mathbb{P}(\mathcal{E}_{\ell,k,s}) \le 1 - p_0 q_0$.

Proof. On the event that a destination is chosen in the mixed cell $\mathcal{C}$ from the non-dominant true part, there exist source tokens in the dominant true part within $\mathcal{C}$. Because the baseline does not enforce boundary protection, at least one such source token must be assigned to a destination in the other true part, and therefore a cross-region merge occurs, i.e. $\mathcal{E}^c_{\ell,k,s}$ holds. The claim follows from the lower bounds p_0, q_0. □

Figure 20 visualises the mixed-cell event in Proposition 2: ToMeSD may select a destination in a boundary-crossing cell, forcing cross-region merges, whereas StructSAM's flatness screening avoids selecting such cells.

H.8. Proof of Theorem 2

We follow the same high-level route as in PiToME's Appendix E: (i) control a one-step adjacency discrepancy, (ii) translate it to a Laplacian perturbation bound, (iii) convert Laplacian perturbation to eigenvalue drift via Hoffman–Wielandt, (iv) telescope over merge steps and sum over windows.

One-step discrepancy. At merge step s in window (ℓ, k), StructSAM merges indices $(a_{\ell,k,s}, b_{\ell,k,s})$. Define the one-step (row/column) discrepancy

$$\Delta_{\ell,k,s} \;\triangleq\; \big\|\mathbf{W}^{(s)}_{\ell,k}[a_{\ell,k,s},:] - \mathbf{W}^{(s)}_{\ell,k}[b_{\ell,k,s},:]\big\|_1 + \big\|\mathbf{W}^{(s)}_{\ell,k}[:,a_{\ell,k,s}] - \mathbf{W}^{(s)}_{\ell,k}[:,b_{\ell,k,s}]\big\|_1. \tag{2}$$

For symmetric affinities, the two terms coincide; we keep both for generality.

Proposition 3 (Row/column drift under correct vs. incorrect merges)**.** *There exist constants $c_{\mathrm{row}} > 0$ and $c_0 > 0$, depending only on the affinity construction and bounds, such that:*

1. *On $\mathcal{E}_{\ell,k,s}$,*
$$\Delta_{\ell,k,s} \;\le\; c_{\mathrm{row}} \sqrt{2\big(1 - \cos(v_{a_{\ell,k,s}}, v_{b_{\ell,k,s}})\big)}.$$
2. *On $\mathcal{E}^c_{\ell,k,s}$, under Assumption 2 we have*
$$\Delta_{\ell,k,s} \;\ge\; c_0(1-m).$$

Proof. (1) For cosine-type affinities $\mathbf{W}[i,j] = \psi(\cos(v_i, v_j))$ with ψ Lipschitz, $|\mathbf{W}[a,j] - \mathbf{W}[b,j]| \le L_\psi |\cos(v_a, v_j) - \cos(v_b, v_j)| \le L_\psi \|v_a - v_b\|_2$. Summing over j inside the window yields $\|\mathbf{W}[a,:] - \mathbf{W}[b,:]\|_1 \le c' \|v_a - v_b\|_2$, and similarly for the column term. Using $\|v_a - v_b\|_2^2 = 2(1 - \cos(v_a, v_b))$ gives the claim.

(2) If $\mathcal{E}^c_{\ell,k,s}$ occurs, the merged pair lies in different true parts. By Assumption 2, within-part similarities are at least m while cross-part similarities are strictly below m, which implies a nontrivial mismatch in adjacency rows/columns to tokens in at least one of the parts. This yields an ℓ_1 discrepancy bounded below by a constant multiple of $1 - m$. □

Proposition 4 (Adjacency-to-Laplacian perturbation)**.** *Let $\mathcal{L}^{(s)}_{\ell,k}$ be the normalized Laplacian before the merge at step s, and let $\mathcal{L}^{(s-1)}_{\ell,k,l}$ be the lifted normalized Laplacian after that merge. Under the boundedness condition in Theorem 2, there exists $c_{\mathrm{lap}} > 0$ such that*
$$\big\|\mathcal{L}^{(s)}_{\ell,k} - \mathcal{L}^{(s-1)}_{\ell,k,l}\big\|_F \;\le\; c_{\mathrm{lap}}\, \Delta_{\ell,k,s}.$$

Proof. Write $\mathcal{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$. A single merge changes $\mathbf{W}$ only through the merged pair's rows/columns under coarsening and lifting, hence $\|\mathbf{W} - \mathbf{W}_l\|_F$ is controlled by $\Delta_{\ell,k,s}$. Degree matrices differ by row sums of $\mathbf{W}$, so $\|\mathbf{D} - \mathbf{D}_l\|_F$ is also controlled by $\Delta_{\ell,k,s}$. Using degree bounds, $\|\mathbf{D}^{-1/2}\|_2$ and $\|\mathbf{D}_l^{-1/2}\|_2$ are uniformly bounded. A triangle inequality expansion then yields the stated Frobenius bound. □

Lemma 4 (Hoffman–Wielandt)**.** *For symmetric matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N\times N}$ with eigenvalues $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$,*
$$\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_1 \le \sqrt{N}\, \|\mathbf{A} - \mathbf{B}\|_F.$$

Proof. By Hoffman–Wielandt, $\|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2 \le \|\mathbf{A} - \mathbf{B}\|_F$. Apply Cauchy–Schwarz to obtain the ℓ_1 bound. □

Proposition 5 (One-step eigenvalue drift)**.** *There exists $c_{\mathrm{sp}} > 0$ such that for each merge step s,*
$$\big\|\boldsymbol{\lambda}^{(s)}_{\ell,k} - \boldsymbol{\lambda}^{(s-1)}_{\ell,k,l}\big\|_1 \;\le\; c_{\mathrm{sp}}\, \Delta_{\ell,k,s}.$$

Proof. Apply Lemma 4 with $\mathbf{A} = \mathcal{L}^{(s)}_{\ell,k}$ and $\mathbf{B} = \mathcal{L}^{(s-1)}_{\ell,k,l}$, then use Proposition 4. □

Proof of Theorem 2. Step 1 (telescoping). For fixed (ℓ, k), telescope over merge steps:
$$\|\boldsymbol{\lambda}_{\ell,k} - \boldsymbol{\lambda}_{\ell,k,l}\|_1 \le \sum_{s=n_{\ell,k}+1}^{N_{\ell,k}} \|\boldsymbol{\lambda}^{(s)}_{\ell,k} - \boldsymbol{\lambda}^{(s-1)}_{\ell,k,l}\|_1 \le c_{\mathrm{sp}} \sum_{s=n_{\ell,k}+1}^{N_{\ell,k}} \Delta_{\ell,k,s}.$$
Summing over $k = 1, \dots, K_\ell$ yields
$$\mathrm{SD}_\ell \le c_{\mathrm{sp}} \sum_{k=1}^{K_\ell} \sum_{s=n_{\ell,k}+1}^{N_{\ell,k}} \Delta_{\ell,k,s}.$$

Step 2 (score-guided vanishing). By Lemma 3, $\mathbb{P}(\mathcal{E}^c_{\ell,k,s}) \le \delta_{\ell,k,s}$ with $\delta_{\ell,k,s} \to 0$. Using Proposition 3 and the law of total expectation,
$$\mathbb{E}[\Delta_{\ell,k,s}] \le c_{\mathrm{row}} \mathbb{E}\Big[\sqrt{2\big(1 - \cos(v_{a_{\ell,k,s}}, v_{b_{\ell,k,s}})\big)}\Big] + c_0(1-m)\, \delta_{\ell,k,s}.$$

Under Assumption 1, the cosine term converges to 1 on correct merges, hence the first term vanishes. Since $\delta_{\ell,k,s} \to 0$ by construction, we obtain $\mathbb{E}[\Delta_{\ell,k,s}] \to 0$, and therefore $\mathbb{E}[\mathrm{SD}_\ell(\mathrm{SG})] \to 0$.

Step 3 (baseline non-vanishing). By Proposition 2, ToMeSD-style destination selection can yield a non-vanishing lower bound $\mathbb{P}(\mathcal{E}^c_{\ell,k,s}) \geq \delta$ for some $\delta > 0$ on infinitely many merge steps. By the lower bound in Proposition 3, this implies $\mathbb{E}[\Delta_{\ell,k,s}] \geq c_0(1-m)\delta$, which yields $\liminf \mathbb{E}[\mathrm{SD}_\ell(\mathrm{Base})] \geq c_0\delta > 0$. □