Beyond Visible Boundaries: Benchmarking Foundation Models for Overlapping Cell Segmentation in Microscopic Imaging

Vo Minh Khoi Nguyen[1*] Khanh Nguyen Vo Ngoc[2*] Vi Vu[3] Thien Nguyen[3]
Thanh-Huy Nguyen[3] Min Xu[3]

[1]Vietnamese-German University [2]University of Information Technology, Vietnam National University
[3]Carnegie Mellon University, USA

Abstract

Overlapping cell segmentation remains a critical bottleneck in computational pathology, yet existing benchmarks evaluate models exclusively on clean, non-overlapping structures, and standard metrics conflate visible and hidden anatomy into a single score, masking fundamental amodal completion failures. In this work, we present an algorithm for synthesising overlapping cell occlusion that is applicable across diverse cell imaging datasets. Building on this, we construct a controlled benchmark for stress-testing SAM-family foundation models under overlapping cell occlusion, generating images from the ISBI-2014 (cervical cytology) and SegPC-2021 (plasma-cell microscopy) datasets across three severity levels: Light, Medium, and Heavy. We further propose a decomposed evaluation protocol that partitions each cell mask into three regions: the full mask, the non-overlap sub-region, and the overlap sub-region. Standard metrics such as Dice and Precision are inappropriate in this setting due to mis-penalization across sub-regions, leading to systematic bias. To address this, we propose a fixed-prior recall-weighted F-measure that computes the weighted harmonic mean between sub-region recall and full-mask precision. Experiments with SAM, SAM 2, MedSAM, and MedSAM2 show that medical SAM-based models are more effective at recovering the overlap sub-region while remaining competitive on the non-overlap sub-region.

1. Introduction

Automated cell segmentation in microscopic imaging underpins computer-aided diagnosis (CAD) in cytology and histopathology, providing quantitative measurements of cell morphology, nuclear-to-cytoplasmic ratio, and chromatin texture for early detection of precancerous lesions [9, 19]. Yet clinical microscopy images are rarely clean: adjacent cells routinely occlude one another, and even modern monolayer preparations such as ThinPrep cannot eliminate cellular overlap [9]. This challenge spans domains - from overlapping red blood cell clusters in blood-smear analysis [14] to plasma-cell clumps in bone-marrow pathology [3]. Recovering the hidden portion of a partially occluded cell - *amodal completion* - is therefore a prerequisite for reliable clinical deployment, yet remains largely uncharacterised for modern segmentation foundations.

Recent foundation models have dramatically raised the ceiling for general segmentation. SAM [6] introduced prompt-driven zero-shot segmentation; SAM 2 [16] extended it to spatiotemporal settings. To close the domain gap, MedSAM [11] fine-tuned SAM across ten-plus imaging modalities, while MedSAM2 [23] reframes 2D/3D segmentation as video object tracking via a self-sorting memory bank. Together, they represent the current state of the art in general and medical image segmentation.

Despite their promise, a critical blind spot remains: *How robust are these models to overlapping cell occlusion?* Existing benchmarks rely on clean, well-separated structures and do not systematically probe this failure mode [4, 10]. Although SAMEO [17] explored amodal segmentation in natural scenes, no benchmark targets microscopic imaging or quantifies robustness across clinically relevant overlap severities. Moreover, standard metrics like Dice and Precision make this worse: when applied to either the overlap or non-overlap sub-region, they incorrectly penalize pixels that are correctly predicted in the complementary sub-region.

To address these problems, we design a controlled evaluation framework to stress-test SAM, SAM2, MedSAM, and MedSAM2 on overlapping cell occlusion, exploring the reasons and factors that make SAM-based models succeed or fail under this condition. Our contributions can be summarized as follows:

- We introduce a **synthesis algorithm** for generating synthetic overlapping cell images with explicit control over

*Equal contribution.

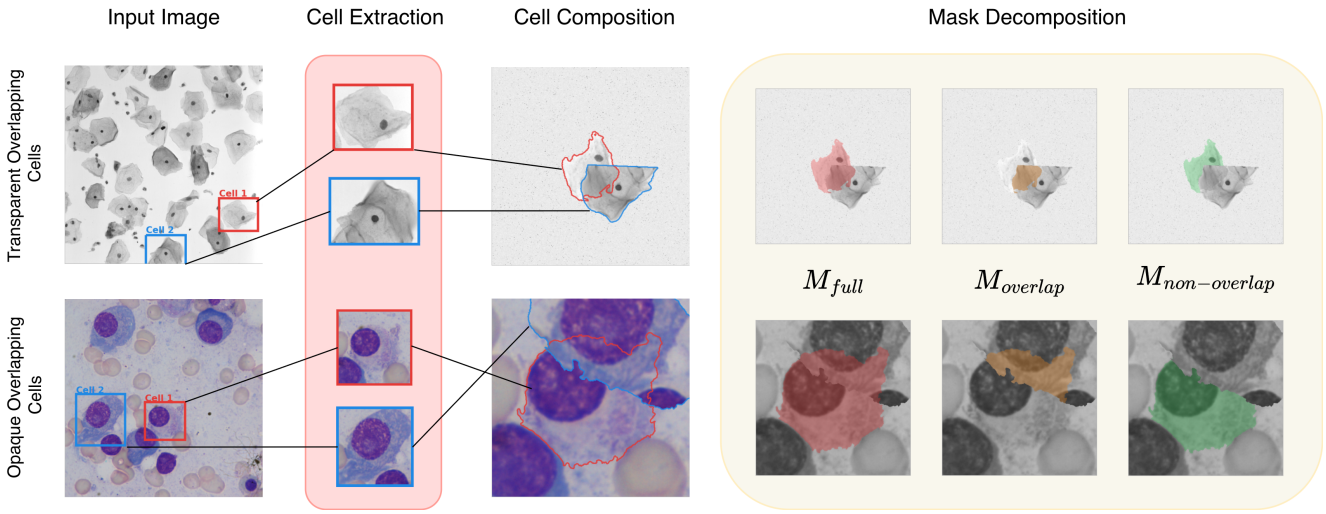


Figure 1. **Overlapping cell synthesis pipeline.** Cell pairs are extracted from input images and composited into unified bounding regions. Ground-truth full mask M_{full} is decomposed into overlap (M_{overlap}) and non-overlap ($M_{\text{non-overlap}}$) components for targeted evaluation, shown for opaque cells from SegPC 2021 (top) and transparent cells from ISBI 2014 (bottom).

cell's transparency, the overlapping ratio range and the number of cells per image. The algorithm supports both transparent and opaque cell variants, making it applicable across diverse cytology datasets and imaging modalities.

- We propose a **decomposed evaluation protocol** that partitions each cell mask into overlap, non-overlap, and full components, enabling independent assessment of *amodal cell completion* and *non-occluded boundary delineation*-two distinct challenges conflated by full-mask metrics.
- We introduce F_2^{sub}, a **fixed-prior recall-weighted F-measure** for sub-region evaluation that resolves the cross-region contamination inherent in standard metrics such as Dice and Precision. By using subregion recall and full-mask precision in the harmonic mean, F_2^{sub} penalises over-segmentation without penalising correctly predicted pixels in the complementary sub-region.
- Based on our experiments across three overlap severity levels, we observed that MedSAM and MedSAM2 are more effective at recovering the overlap sub-region while remaining competitive in the non-overlap sub-region, suggesting that, within the SAM family, amodal cell completion rather than visible boundary delineation is the dominant challenge.

2. Related Work

2.1. Overlapping Cell Segmentation

Classical and early deep-learning approaches to overlapping cell segmentation can be grouped by how they model the overlap region itself. Watershed-based methods partition intensity gradients and use marker control or multi-pass refinement to resolve nucleus and cytoplasm boundaries in clumps [10, 15, 18], but remain prone to over-segmentation wherever strong and weak edges coexist. Level-set methods afford finer boundary control by coupling repulsive contour penalties [15] or joint pairwise overlap-area constraints [9] across all cells in a clump; geometry-driven alternatives detect concave boundary points and fit ellipses to convex arc segments [14]. All three families degrade under severe deformation or deep multi-cell stacking, and none scales gracefully to the high cell counts encountered in dense smears.

Deep learning shifted the paradigm from hand-crafted geometry to learned representations. Boundary-weighted semantic models with CRF post-processing [19] and instance-based frameworks exploiting inter-instance relations [21, 22] raised segmentation quality on the ISBI 2014 benchmark but lacked explicit modelling of the overlap region as a distinct geometric entity. DoNet [4] addressed this directly with a decompose-and-recombine architecture that separates intersection from complement sub-regions, setting the prior state of the art on ISBI 2014 and the CPS cytology dataset. Despite these advances, few prior methods have systematically analyzed how performance degrades as the overlap coefficient increases, and large-scale foundation models have not been evaluated under controlled overlapping cell occlusion.

2.2. Foundation Models in Medical Imaging

SAM [6] established a prompt-driven segmentation framework trained on over one billion masks with strong zero-shot generalisation; SAM 2 [16] extended it to video and volumetric settings via a streaming memory architecture. MedSAM [11] fine-tuned SAM on a large-scale multi-

modal medical corpus for robust cross-modality transfer, while MedSAM2 [23] reframed both 2D and 3D medical image segmentation as video object tracking with a self-sorting memory bank that selects relevant embeddings across slices or temporal frames. None of these models has been evaluated under controlled overlapping cell occlusion; their benchmarks focus on anatomical structures that are clearly delineated and spatially separated.

2.3. Amodal Segmentation

Amodal segmentation, inferring the complete extent of a partially visible object, has been studied extensively in natural scenes [7]. Occlusion R-CNN [2] adds three mask prediction heads to R-CNN (visible, amodal, and occlusion) to jointly infer both the observable and hidden extents of each instance; Xiao *et al.* [20] use shape-prior codebooks to infer hidden regions; BCNet [5] directly models the occluder-occludee relationship; and SAMEO [17] adapts SAM for amodal mask decoding in general occluded scenes. Direct transfer to cytological microscopy faces two fundamental obstacles: occlusions arise between densely clustered instances of the *same class*, producing entangled boundaries unlike the distinct occluder-occludee pairs in natural scenes.

3. Methodology

To begin with, cytology datasets differ in the optical properties of their cells depending on the imaging modality and staining protocol used. In Pap smear datasets such as ISBI 2014 [9], CPS [22], the cytoplasm exhibits characteristic translucency, such that the underlying cell remains partially visible beneath an occluding cell; we classify such datasets as **transparent**. In contrast, datasets that employ specific staining protocols such as SegPC 2021 [3], which images plasma cells produce cells that are optically opaque, fully occluding any underlying cell content at the overlap region; we classify these as **opaque**. This distinction directly governs the compositing step of the synthesis procedure, as described in Section 3.1.

3.1. Synthetic Overlapping Cell Generation

Transparent Overlapping Cell Generation. Inspired by Lu et al. [9], who proposed an algorithm for generating synthetic overlapping cells with alpha-blended transparency at overlap regions, and further extended by Wan et al. [19], we adapt this approach into a general purpose method for synthesizing overlapping cell images with transparent properties, applicable across diverse cytology datasets and imaging modalities.

Source material. Individual free-lying cells are extracted from the input images, where each cell is accompanied by its corresponding segmentation mask.

Synthesis procedure. The synthesis procedure, reproduced as Algorithm 1 generates synthetic images of size 512×512 where at least one cell has an overlapping ratio within a specified range and a certain amount of cells N through the following steps: (1) **Background generation**: Background is formed using background pixels randomly selected from any input image, with mirror transformations applied to smooth the transitions between pixels. (2) **Pick and transform cell**: pick one of the cells, apply a random rigid geometric transform (translation, rotation from $(0, 2\pi)$ and scale from (0.8, 1.2)) and random linear brightness transform and place it on the synthetic image. Using a random value between 0.88 and 0.99 for the alpha channel to simulate the partial transparency effect. (3) **Adding N-1 cell** for the first cell of N-1 cells, we would repeatedly transform (like in step 2) and then randomly place it on the target cell until it has the overlapping ratio within the specified range. From the second cell of N-1 cells, each of these new cells would be transformed and placed at random locations, where each of these new cells overlaps with at least one of the cells present in the image.

Opaque Overlapping Cell Generation. For datasets in which cells are opaque, we first crop the cell from the image with the crop of size 512×512 and then use another cell from the same image to iteratively transform and place on top of the target cell until the overlapping ratio of the target cell falls within the specified range. The transformation would be the same as in step 2 of Algorithm 1 but alpha is set to 1.

Overlap severity levels. To systematically evaluate model robustness under varying degrees of occlusion, we define three controlled overlap configurations based on the proportion of the target cell area that is occluded by other cells. **Light** overlap corresponds to 10-20% occlusion, **Medium** overlap to 30-40%, and **Heavy** overlap to 50-60%.

These ranges are enforced during the synthesis process by constraining the pairwise overlap ratio between the target cell and at least one occluding cell (see Algorithm 1). The overlap ratio is computed as the fraction of the target cell mask that lies within the intersection region with an occluding cell. This design ensures that each severity level reflects a distinct level of geometric complexity and occlusion difficulty. Increasing overlap severity not only enlarges the hidden (amodal) region M_{overlap}, but also reduces the directly observable boundary in $M_{\text{non-overlap}}$, thereby creating a progressively more challenging setting for both boundary delineation and amodal completion. Reporting results across these levels provides a more granular view of model behavior under occlusion than aggregate evaluation.

Ground truth decomposition. For each generated sample, three complementary masks are stored: the **full mask** M_{full} representing the complete extent of the target cell, the **non-overlap mask** $M_{\text{non-overlap}} = M_{\text{full}} \setminus M_{\text{overlap}}$ captur-

Algorithm 1 Cell Synthesis Procedure

Require: Cell image set $\mathcal{C}$, input image set $\mathcal{I}$, number of cells N, overlapping ratio range $[r_{\min}, r_{\max}]$
Ensure: Synthetic image $\mathbf{S}$ of size 512×512

// Step 1: Background Generation
Sample background pixels randomly from $I \sim \mathcal{I}$
Apply mirror transformations to smooth pixel transitions
Initialize $\mathbf{S}$ with the generated background
// Step 2: Pick and Transform Target Cell
Pick a random cell $c_0 \sim \mathcal{C}$
Apply rigid transform: translation, $\theta \sim \mathcal{U}(0,\, 2\pi)$, $s \sim \mathcal{U}(0.8,\, 1.2)$
Apply random linear brightness transform to c_0
Sample $\alpha \sim \mathcal{U}(0.88,\, 0.99)$
Composite c_0 onto $\mathbf{S}$ using alpha channel α
// Step 3: Add Remaining $N-1$ Cells
Pick a random cell $c_1 \sim \mathcal{C}$
repeat
    Apply random rigid geometric and brightness transforms to c_1
    Place c_1 at a random location on $\mathbf{S}$
until overlap$(c_1,\, c_0) \in [r_{\min},\, r_{\max}]$
Composite c_1 onto $\mathbf{S}$ using $\alpha \sim \mathcal{U}(0.88,\, 0.99)$
for $i \leftarrow 2$ **to** $N-1$ **do**
    Pick a random cell $c_i \sim \mathcal{C}$
    repeat
        Apply random rigid geometric and brightness transforms to c_i
        Place c_i at a random location on $\mathbf{S}$
    until c_i overlaps with at least one cell present in $\mathbf{S}$
    Composite c_i onto $\mathbf{S}$ using $\alpha \sim \mathcal{U}(0.88,\, 0.99)$
end for
return $\mathbf{S}$

ing only the non-overlapped region, and the **overlap mask** $M_{\text{overlap}} = M_{\text{full}} \cap M_{\text{occluding cell}}$ covering the hidden region beneath the overlapping cell.

3.2. Evaluation Process

We evaluate each model under three mask settings: the full mask, the overlap sub-region, and the non-overlap sub-region. The **overlap sub-region** $M_{\text{overlap}} = M_{\text{full}} \cap M_{\text{occluding cell}}$ covers the portion of the target cell hidden beneath an occluding cell, while the **non-overlap sub-region** $M_{\text{non-overlap}} = M_{\text{full}} \setminus M_{\text{overlap}}$ covers the directly visible cell boundary. For the full-mask setting, we report the Dice Similarity Coefficient (DSC), Precision, and Recall between the predicted mask $\hat{M}$ and the ground-truth mask M_{full}:

$$\text{Precision} = \frac{|\hat{M} \cap M_{\text{full}}|}{|\hat{M}|}, \tag{1}$$

$$\text{Recall} = \frac{|\hat{M} \cap M_{\text{full}}|}{|M_{\text{full}}|}, \tag{2}$$

$$\text{DSC} = \frac{2|\hat{M} \cap M_{\text{full}}|}{|\hat{M}| + |M_{\text{full}}|}. \tag{3}$$

Why standard metrics do not apply to sub-regions. Precision and DSC cannot be directly applied to either sub-region (M_{overlap} or $M_{\text{non-overlap}}$). Because their denominators include the full predicted mask $|\hat{M}|$, any correctly predicted pixel *outside* the evaluated sub-region is treated as a false positive. Concretely, restricting precision to a sub-region $M_{\text{sub}} \in \{M_{\text{overlap}},\, M_{\text{non-overlap}}\}$ gives:

$$\text{Precision}_{\text{sub}} = \frac{|\hat{M} \cap M_{\text{sub}}|}{|\hat{M}|} \;\leq\; \frac{|M_{\text{sub}}|}{|\hat{M}|}, \tag{4}$$

which systematically penalises models that correctly predict the full cell extent, since those correct predictions lie outside M_{sub}. DSC inherits the same bias through its $|\hat{M}|$ term. This cross-region contamination is an artefact of the evaluation design, not a genuine indicator of segmentation quality.

Sub-region recall. Recall avoids this problem because its denominator is anchored entirely to the ground-truth sub-region:

$$\text{Recall}_{\text{sub}} = \frac{|\hat{M} \cap M_{\text{sub}}|}{|M_{\text{sub}}|}, \quad M_{\text{sub}} \in \{M_{\text{overlap}},\, M_{\text{non-overlap}}\}. \tag{5}$$

Specifically, $\text{Recall}_{\text{overlap}}$ measures a model's ability to recover hidden cell anatomy beneath an occluding cell, while $\text{Recall}_{\text{non-overlap}}$ measures delineation of the directly observable cell boundary.

Recall-weighted F-measure (F_2^{sub}). Recall alone, however, admits a trivial maximiser: a model that predicts the entire image as foreground achieves $\text{Recall}_{\text{sub}} = 1.0$ regardless of quality. To address this, we replace the ill-defined $\text{Precision}_{\text{sub}}$ with the full-mask precision P_{full} (Eq. (1)) as a cross-region-free proxy that directly penalises over-segmentation, and combine it with $\text{Recall}_{\text{sub}}$ into a weighted F-measure:

$$F_2^{\text{sub}} = \frac{5\; P_{\text{full}}\; \text{Recall}_{\text{sub}}}{4\; P_{\text{full}} + \text{Recall}_{\text{sub}}}, \tag{6}$$

where M_{sub} is instantiated as either M_{overlap} or $M_{\text{non-overlap}}$ depending on the region of interest. We set $\beta = 2$ to reflect the clinical priority of recovering the specific sub-region over avoiding false positives. This formulation eliminates

the trivial maximiser, as a model predicting the entire image as foreground collapses $P_{\text{full}} \to 0$, driving $F_2^{\text{sub}} \to 0$ even at perfect recall. We report $\text{Recall}_{\text{sub}}$ and F_2^{sub} for both sub-regions side by side throughout Section 4.

3.3. Prompt Type Selection

All experiments use bounding-box prompts derived from the tight axis-aligned bounding box $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ over ground-truth foreground pixels. Box prompts are consistently superior to point prompts across SAM and its medical variants: Mazurowski *et al*. [13] demonstrated this across 19 medical datasets; Mattjie *et al*. [12] corroborated it across eight prompt strategies; Liu *et al*. [8] confirmed the same finding in MedSAM-specific evaluations; and Dai *et al*. [1] showed that point prompts for general SAM require explicit augmentation strategies to approach box-level performance.

4. Experimental Results

4.1. Dataset Settings

Transparent Cell Dataset. For the transparent cell setting, we use the ISBI 2014 Overlapping Cervical Cytology Segmentation Challenge dataset [9], a standard benchmark for overlapping cell segmentation in Pap smear cytology, in which cells exhibit characteristic cytoplasmic translucency at overlap regions. Applying Algorithm 1 yields 1,216 synthetic samples distributed across three overlapping severity ranges.

Opaque Cell Dataset. For the opaque cell setting, we use SegPC 2021 [3], a dataset of plasma cells stained with Jenner-Giemsa stain, in which cells are optically opaque and fully occlude any underlying cell content at overlap regions. Applying the opaque variant of Algorithm 1 yields 4,635 synthetic samples distributed across three overlapping severity ranges.

4.2. Full-Mask Evaluation

Per-dataset leaders. As shown in Table 1, on SegPC 2021, MedSAM is the strongest full-mask model, attaining DSC of 0.862, 0.821, and 0.830 across Light, Medium, and Heavy overlap - the only model whose DSC does not decrease monotonically with severity. On ISBI 2014, MedSAM 2 leads at Medium and Heavy overlap (0.912, 0.900), while SAM 2 edges it at Light overlap (0.939 vs. 0.938); MedSAM remains competitive at 0.903, 0.884, and 0.860 across all levels.

Precision-recall asymmetry between model families. General-domain models attain the highest precision in most settings (up to 0.979 and 0.962 on SegPC 2021) but pay a steep recall cost: SAM 2 drops to 0.409 recall at Heavy overlap on SegPC 2021. Medically adapted models maintain a more balanced profile, which explains their stronger DSC under severe occlusion. On ISBI 2014, MedSAM 2 achieves the highest recall at all levels (0.897, 0.883, 0.891) while the general-domain models retain higher precision.

Robustness to increasing overlap severity. The DSC gap between model families widens substantially as overlap increases. On SegPC 2021, general-domain models drop by 0.221 (SAM) and 0.232 (SAM 2) from Light to Heavy, versus only 0.032 (MedSAM) and 0.054 (MedSAM 2) for medically adapted models. On ISBI 2014, the corresponding drops are 0.099 and 0.115 versus 0.043 and 0.038, confirming that medically adapted models are far more robust and that MedSAM is the strongest cross-dataset choice overall.

4.3. Overlap-Only Evaluation

The general models (SAM, SAM 2) essentially fail at recovering occluded anatomy on SegPC 2021. In Table 2, at Light severity, their Overlap Recall is just 0.016 and 0.013, respectively near zero, indicating they make no attempt to predict the cell region hidden beneath an occluding cell. Their F_2^{sub} scores (0.020 and 0.016) confirm this is not a precision recall trade-off artefact but a genuine inability to recover the covered anatomy.

Overlap recall increases with occlusion severity for all models, which is an important structural property: a heavier overlap creates a larger hidden region, so models that implicitly predict the full cell extent score progressively higher on this metric as severity increases. MedSAM's Overlap Recall leaps from $0.291 \to 0.696$ (a gain of 0.405) on SegPC 2021, while MedSAM2 gains 0.277, both reflecting a capacity to infer occluded structure.

On ISBI 2014, the overlap gap between model families narrows, but medical models still lead. SAM and SAM 2 achieve moderate recall (0.491–0.602), unlike their near-zero scores on SegPC 2021, because ISBI 2014 cells are transparent: the occluding cell does not fully block the underlying cell's texture and intensity, providing a residual visual signal that even general-purpose models can exploit to partially recover the hidden boundary. By contrast, SegPC 2021 cells are opaque, leaving no such cue and forcing models to rely entirely on shape priors—a capacity that medical pretraining provides but general-purpose training does not. Critically, MedSAM2 overtakes MedSAM at Medium (0.816 vs. 0.783) and Heavy (0.890 vs. 0.815) on ISBI 2014, while MedSAM leads only at Light (0.666 vs. 0.625), and the MedSAM vs. SAM F_2^{sub} gap at Heavy on SegPC 2021 remains 0.447 (0.736 vs. 0.289), underscoring that on fully opaque datasets, medical-domain pretraining is the *only* available substitute for the absent visual signal.

4.4. Non-overlap-Only Evaluation

Complementary trend: general-domain models lead on non-overlap sub-regions. Table 3 shows the inverse of

Table 1. **Full-mask evaluation under overlapping cell occlusion on SegPC 2021 and ISBI 2014, box prompts.** Metrics include DSC, Precision, and Recall evaluated at Light, Medium, and Heavy severity levels of cell overlap. **Best** and second best highlighted per column per dataset.

Model	SegPC 2021									ISBI 2014								
	DSC			Precision			Recall			DSC			Precision			Recall		
	Light	Med.	Heavy	Light	Med.	Heavy	Light	Med.	Heavy	Light	Med.	Heavy	Light	Med.	Heavy	Light	Med.	Heavy
SAM	.804	.666	.583	**.979**	.970	.950	.695	.519	.446	.930	.876	.831	**.997**	**.993**	.983	.874	.793	.743
SAM 2	.785	.670	.553	.976	**.972**	**.962**	.675	.524	.409	**.939**	.879	.824	.996	.992	**.985**	.890	.800	.734
MedSAM	**.862**	**.821**	**.830**	.971	.961	.953	**.780**	**.726**	**.748**	.903	.884	.860	.982	.977	.965	.840	.812	.786
MedSAM2	.764	.720	.710	.953	.936	.922	.671	.620	.617	.938	**.912**	**.900**	.986	.953	.921	**.897**	**.883**	**.891**

Table 2. **Overlap Sub-region Evaluation under overlapping cell occlusion, box prompts.** Evaluated on SegPC 2021 and ISBI 2014 datasets across Light, Medium, and Heavy severity levels. Each severity shows Overlap Recall (Rec) and the fixed-prior recall-weighted F-measure (F_2^{sub}). **Best** and second best highlighted per column per dataset.

Model	SegPC 2021						ISBI 2014					
	Light		Medium		Heavy		Light		Medium		Heavy	
	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}
SAM	.016	.020	.078	.096	.246	.289	.491	.546	.532	.586	.602	.653
SAM 2	.013	.016	.039	.048	.127	.154	.495	.550	.528	.582	.573	.625
MedSAM	**.291**	**.338**	**.505**	**.558**	**.696**	**.736**	**.666**	**.712**	.783	.815	.815	.841
MedSAM2	.274	.320	.441	.493	.551	.599	.625	.674	**.816**	**.840**	**.890**	**.896**

the overlap sub-region pattern. On ISBI 2014, SAM 2 is the strongest model across all severity levels (0.957, 0.943, 0.930), with SAM close behind; MedSAM is consistently the weakest (0.869, 0.827, 0.749), likely due to domain shift. On SegPC 2021, the ordering reverses: MedSAM leads at all levels (0.867, 0.844, 0.811) while MedSAM 2 yields the lowest recall (0.742, 0.716, 0.698). Crucially, inter-model differences here are far smaller than in the overlap sub-region setting, confirming that the primary challenge in overlapping-cell segmentation is amodal cell completion, not non-overlap sub-region delineation.

Combined view: non-overlap sub-region delineation vs. amodal overlap recovery. Jointly considering Tables 2 and 3 and Figure 4, a clear division emerges: general-domain models excel at delineating the non-overlap sub-region but fail on the overlap sub-region, while medically adapted models maintain competitive non-overlap sub-region performance *and* substantially better amodal cell completion. MedSAM is the most practical cross-dataset choice when both the overlap and non-overlap sub-regions must be segmented reliably.

F_2^{sub} validates non-overlap sub-region performance. Across both datasets, F_2^{sub} for the non-overlap sub-region systematically exceeds raw recall, confirming that strong performance is not driven by over-segmentation but reflects a balanced trade-off between coverage and precision. The exception is MedSAM on ISBI 2014 at Heavy overlap (Rec=0.749, F_2^{sub}=0.784), which signals genuine under-recovery of the non-overlap sub-region, likely due to domain shift rather than conservative bias or excessive false positives. On SegPC 2021, MedSAM's F_2^{sub} advantage over general-domain models (e.g., 0.886 vs. 0.845 and 0.825 at Light overlap) confirms that its non-overlap sub-region superiority reflects real segmentation improvement rather than metric artefacts.

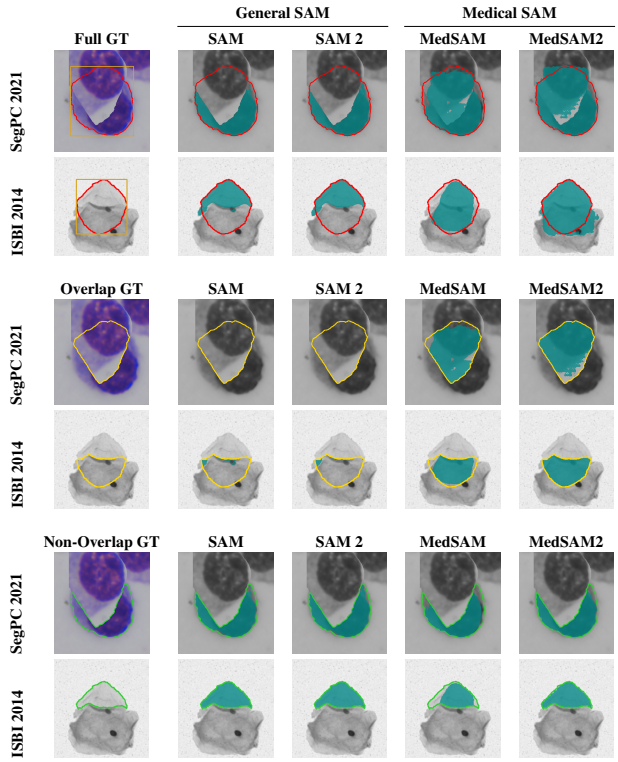
4.5. Qualitative Results

Qualitative comparison of cell segmentation performance across overlapping and non-overlapping conditions on SegPC 2021 (top of each pair) and ISBI 2014 (bottom of each pair). Rows 1-2: full-mask segmentation under increasing overlap severity (Light, Medium, Heavy), where medical models (MedSAM, MedSAM2) recover more complete cell boundaries than generalist SAM and SAM 2, aligning with quantitative trends in Table 1. Rows 3-4: overlap-mask predictions with Overlap Recall (Ovr-Rec) scores, where under heavy occlusion SAM and SAM 2 severely under-segment intersecting territories, while MedSAM and MedSAM2 robustly reconstruct hidden boundaries in dense overlap regions. Rows 5-6: non-overlap mask predictions with Non-Overlap Recall (NonOvr-Rec) scores; SAM and SAM 2 excel at isolated cells, while MedSAM and MedSAM2 maintain competitive clean-boundary accuracy.

Table 3. **Non-overlap Sub-region Evaluation under overlapping cell occlusion, box prompts.** Evaluated on SegPC 2021 and ISBI 2014 datasets across Light, Medium, and Heavy severity levels. Each severity shows Non-overlap Recall (Rec) and the fixed-prior recall-weighted F-measure (F_2^{sub}). **Best** and second best highlighted per column per dataset.

Model	SegPC 2021						ISBI 2014					
	Light		Medium		Heavy		Light		Medium		Heavy	
	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}	Rec	F_2^{sub}
SAM	.817	.845	.756	.791	.687	.727	.939	.950	.930	.942	.916	.929
SAM 2	.794	.825	.785	.816	.751	.785	**.957**	**.965**	**.943**	**.952**	**.930**	**.941**
MedSAM	**.867**	**.886**	**.844**	**.865**	**.811**	**.836**	.869	.889	.827	.853	.749	.784
MedSAM2	.742	.776	.716	.751	.698	.734	.942	.950	.918	.925	.891	.897

Table 4. Qualitative comparison of SAM, SAM 2, MedSAM, and MedSAM2 on SegPC 2021 and ISBI 2014, showing full-mask, overlap-mask, and non-overlap-mask cell segmentation under varying overlap severity, with medical models yielding more complete and robust boundaries, heavily occluded regions.



5. Conclusion and Discussion

We proposed an algorithm for synthesising overlapping cell occlusion with specific control over overlapping ratio, number of cells, and cell transparency, and used it to construct a controlled benchmark for stress-testing SAM-family foundation models under overlapping cell occlusion across three severity levels Light, Medium, and Heavy on ISBI-2014 (cervical cytology) and SegPC-2021 (plasma-cell microscopy). We further proposed a decomposed mask protocol into full-mask, overlap sub-region, and non-overlap sub-region, alongside F_2^{sub}, a fixed-prior recall-weighted F-measure that evaluates models on sub-regions without the problems of cross-region contamination and trivial recall maximisation.

Through our experiments, MedSAM and MedSAM2 are more effective at recovering the overlap sub-region while remaining competitive on the non-overlap sub-region. SAM and SAM 2 nearly fail to recover occluded anatomy on opaque cells, whereas MedSAM and MedSAM2 demonstrate a strong capacity to infer hidden structure - a gap that narrows on transparent cells, where residual visual signals beneath the occluder provide partial recovery cues even for general-domain models. In the non-overlap sub-region, inter-model differences are relatively small, indicating that, within the SAM family, amodal cell completion rather than visible boundary delineation is the dominant challenge.

Limitations and future work. Our evaluation covers SAM-family models, single-occluder scenarios, and bounding-box prompts. Extending the benchmark to task-specific architectures such as DoNet [4] and to multi-occluder settings is left for future work.

References

[1] Haixing Dai, Chong Ma, Zhiling Yan, Zhengliang Liu, Enze Shi, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, et al. Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*, 2023. 5

[2] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1328–1336. IEEE, 2019. 3

[3] Anubha Gupta, Shiv Gehlot, Shubham Goswami, Sachin Motwani, Ritu Gupta, Álvaro García Faura, Dejan Štepec, Tomaž Martinčič, Reza Azad, Dorit Merhof, et al. Segpc-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images. *Medical Image Analysis*, 83:102677, 2023. 1, 3, 5

[4] Hao Jiang, Rushan Zhang, Yanning Zhou, Yumeng Wang, and Hao Chen. Donet: Deep de-overlapping network

for cytology instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15641–15650, 2023. 1, 2, 7

[5] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4019–4028, 2021. 3

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 1, 2

[7] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016. 3

[8] Xiaofeng Liu, Jonghye Woo, Chao Ma, Jinsong Ouyang, and Georges El Fakhri. Point-supervised brain tumor segmentation with box-prompted medsam. *ArXiv*, pages arXiv–2408, 2024. 5

[9] Zhi Lu, Gustavo Carneiro, and Andrew P Bradley. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. *IEEE transactions on image processing*, 24(4):1261–1272, 2015. 1, 2, 3, 5

[10] Zhi Lu, Gustavo Carneiro, Andrew P Bradley, Daniela Ushizima, Masoud S Nosrati, Andrea GC Bianchi, Claudia M Carneiro, and Ghassan Hamarneh. Evaluation of three algorithms for the segmentation of overlapping cervical cells. *IEEE journal of biomedical and health informatics*, 21(2):441–450, 2016. 1, 2

[11] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024. 1, 2

[12] Christian Mattjie, Luis Vinicius De Moura, Rafaela Ravazio, Lucas Kupssinskü, Otávio Parraga, Marcelo Mussi Delucis, and Rodrigo C Barros. Zero-shot performance of the segment anything model (sam) in 2d medical imaging: A comprehensive evaluation and practical guidelines. In *2023 IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 108–112. IEEE, 2023. 5

[13] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. 5

[14] Korranat Naruenatthanaset, Thanarat H Chalidabhongse, Duangdao Palasuwan, Nantheera Anantrasirichai, and Attakorn Palasuwan. Red blood cell segmentation with overlapping cell separation and classification on imbalanced dataset. *arXiv preprint arXiv:2012.01321*, 2020. 1, 2

[15] Xin Qi, Fuyong Xing, David J Foran, and Lin Yang. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. *IEEE Transactions on Biomedical Engineering*, 59(3):754–765, 2011. 2

[16] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2

[17] Wei-En Tai, Yu-Lin Shih, Cheng Sun, Yu-Chiang Frank Wang, and Hwann-Tzong Chen. Segment anything, even occluded. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29385–29394, 2025. 1, 3

[18] Afaf Tareef, Yang Song, Heng Huang, Dagan Feng, Mei Chen, Yue Wang, and Weidong Cai. Multi-pass fast watershed for accurate segmentation of overlapping cervical cells. *IEEE transactions on medical imaging*, 37(9):2044–2059, 2018. 2

[19] Tao Wan, Shusong Xu, Chen Sang, Yulan Jin, and Zengchang Qin. Accurate segmentation of overlapping cells in cervical cytology with deep convolutional neural networks. *Neurocomputing*, 365:157–170, 2019. 1, 2, 3

[20] Yuting Xiao, Yanyu Xu, Ziming Zhong, Weixin Luo, Jiawei Li, and Shenghua Gao. Amodal segmentation based on visible region segmentation and shape prior. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2995–3003, 2021. 3

[21] Yanning Zhou, Hao Chen, Jiaqi Xu, Qi Dou, and Pheng-Ann Heng. Irnet: Instance relation network for overlapping cervical cell segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 640–648. Springer, 2019. 2

[22] Yanning Zhou, Hao Chen, Huangjing Lin, and Pheng-Ann Heng. Deep semi-supervised knowledge distillation for overlapping cervical cell instance segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 521–531. Springer, 2020. 2, 3

[23] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024. 1, 3