One Click per Cell Type Suffices: Training-free Group Interaction for Cell Instance Segmentation

Sanghyun Jo[1,2], Seo Jin Lee[2], Seohyung Hong[2], Yoorim Gang[2], Hyeongsub Kim[2,3], Hyungseok Seo[2†], and Kyungsu Kim[2†]

[1]OGQ, Korea [2]Seoul National University, Korea [3]LG CNS, Korea

Abstract. Cell instance segmentation models trained on cell-specific datasets suffer severe performance drops on out-of-distribution cell types, while interactive foundation models overcome this through per-instance prompting at a cost that is prohibitively expensive for histopathology images containing hundreds to thousands of densely packed instances. We introduce *Group Prompting*, a new paradigm that shifts interactive segmentation from per-instance $O(N)$ to per-type $O(T)$, where a single click per cell type suffices to segment all instances of that type. Our key observation is that the frozen image encoder of the Segment Anything Model (SAM) already clusters same-type cells in its feature space before any prompt is given. Exploiting this property, we propose **Chain-of-Prompts (CoP)**, a training-free framework that recursively expands a single user click by (1) identifying reliable same-type locations through non-parametric gating of multi-scale encoder features, and (2) selecting the most spatially distant reliable point as the next prompt to maximize coverage. On three cell-type-annotated benchmarks, CoP with one click per type retains over 90% of per-instance performance and surpasses fully-supervised methods without any additional training. On four morphologically homogeneous benchmarks, a single click retains over 99%. **Project Page:** `shjo-april.github.io/Chain-of-Prompts`

Keywords: Cell Instance Segmentation · Interactive Segmentation

1 Introduction

Cell instance segmentation is essential for quantitative analysis in computational pathology, yet existing cell-specific methods [20,3] remain fundamentally constrained by their training data. Whether unsupervised [11], weakly-supervised [8], or fully-supervised [7], these approaches learn cell representations tied to specific tissue types and cell morphologies encountered during training, leading to severe performance degradation on out-of-distribution (OOD) cell types (see Fig. 1). Recent interactive foundation models such as SAM3 [2] offer an alternative by accepting per-instance point prompts, enabling segmentation of arbitrary cell types without task-specific training. However, unlike natural images [4,14] where object numbers are in the tens, histopathology images [18,5]

† Corresponding authors: `kyskim@snu.ac.kr`, `h.seo@snu.ac.kr`

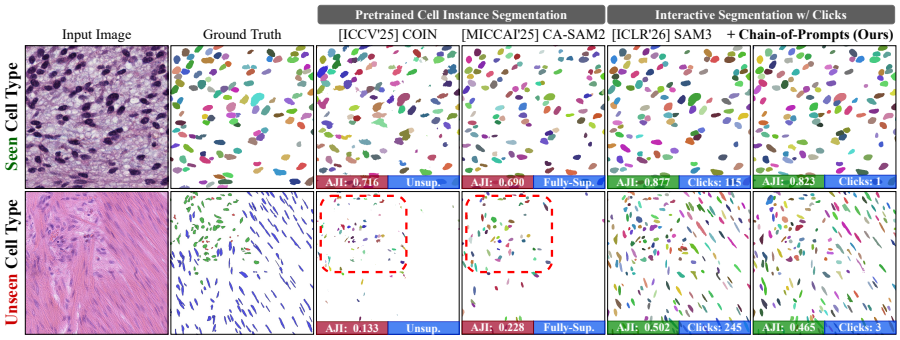

Fig. 1. One Click per Cell Type is All You Need. Pretrained models fail to identify unseen types and their performance is limited to a specific cell type (red dashed boxes). While SAM3 [2] generalizes, it requires per-instance clicks (*e.g.*, 245). Our CoP achieves 92.7% of the upper bound performance [2] with only 3 clicks.

contain hundreds to thousands of densely packed cell instances, making per-instance prompting prohibitively expensive in practice. This contrast motivates a paradigm shift from per-instance prompting, which scales as $O(N)$ with the number of cells, to per-type group prompting at $O(T)$, where a single click per cell type suffices to segment all instances of that type.

A common strategy to reduce per-instance cost is to generate pseudo prompts (*e.g.*, points) automatically using external open-vocabulary or cell-specific detection models [15,24,10]. However, these detectors are trained on specific cell and tissue types and therefore inherit the same OOD limitation (see Fig. 1). In this work, we bypass external detectors by leveraging a key intrinsic property of SAM [12,2,1]. Because SAM's architecture dictates that the image encoder must embed all instance information before receiving user prompts at the decoding stage, its frozen feature space inherently performs instance-aware encoding. When combined with shared morphological traits (*e.g.*, size, shape, staining pattern), this naturally gives rise to cell-type-aware clustering without any supervision. As a result, computing similarity from a cell's feature reliably activates other cells of the same cell type across the image.

While this intrinsic property provides the theoretical foundation for propagating a single click to all instances of the same cell type, directly exploiting it presents two challenges. First, SAM's multi-scale features dictate a strict trade-off between spatial precision and type selectivity: high-resolution features localize densely but activate background regions with similar texture, whereas low-resolution features accurately isolate cell types but blur adjacent instances due to limited resolution. Second, naive one-shot propagation is highly sensitive to similarity thresholds, yielding either excessive false positives or missed cells.

To address these challenges, we propose **Chain-of-Prompts (CoP)**, a training-free framework that recursively leverages newly discovered cells as prompts for subsequent propagation. CoP consists of two complementary components. First,

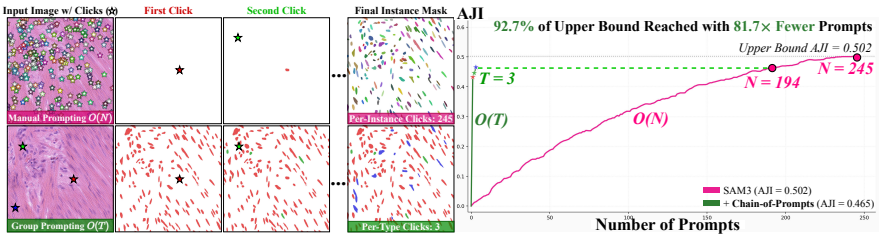

Fig. 2. From 245 Clicks to 3: Group Prompting. Manual prompting requires one click per instance; our group prompting propagates each click to all same-type instances, reaching **92.7%** of the upper bound with **81.7**× fewer prompts.

Hierarchical Similarity Gating (HSG) combines SAM's multi-scale features to non-parametrically identify reliable cell points recursively, achieving precision above 96% without any learnable parameters. Second, Farthest Prompt Recursion (FPR) ensures comprehensive tissue coverage by selecting the next prompt farthest from all prior clicks, maximizing spatial diversity by uncovering cells in unexplored regions. By iterating these two steps, our CoP expands from a single click to segment most of the same-type cells. On three benchmarks [5,6,21], CoP uses only $O(T)$ per-type clicks and retains over 90% of $O(N)$ per-instance performance of SAM3 [2] with 97% reduction in annotation cost, while outperforming fully-supervised models [9,7,22] (see Fig. 2). Our contributions are as follows:

- We introduce **Group Prompting**, shifting interactive segmentation from per-instance $O(N)$ to per-type $O(T)$ interaction, thereby reducing annotation cost from the number of cells to the number of cell types while remaining robust to out-of-distribution cell types without cell-specific training.
- We propose **Chain-of-Prompts (CoP)**, a training-free framework that recursively expands prompt coverage while maintaining high precision ($\geq$96%) at each iteration.
- On seven benchmarks, CoP retains over 90% of per-instance performance on cell-type-annotated datasets [5,6,21] and over 99% on morphologically homogeneous datasets [13,18,16,23], outperforming fully-supervised methods [9,7,22] that require complete mask annotations for training.

2 Method

The proposed **Chain-of-Prompts (CoP)** is a training-free framework that discovers all same-type cells from a single user click and produces their instance masks. CoP operates exclusively on the frozen features of a pretrained SAM image encoder (*e.g.*, SAM3 [2]), which extracts a high-resolution feature map $F_h \in \mathbb{R}^{D\times H/4\times W/4}$ and a low-resolution feature map $F_l \in \mathbb{R}^{D\times H/16\times W/16}$ from an input image I. As illustrated in Fig. 3, CoP comprises two components. First, Hierarchical Similarity Gating (Sec. 2.1) leverages the complementary strengths of F_h and F_l to identify a high-precision set of reliable points $\mathcal{R}^{(0)}$ from the initial

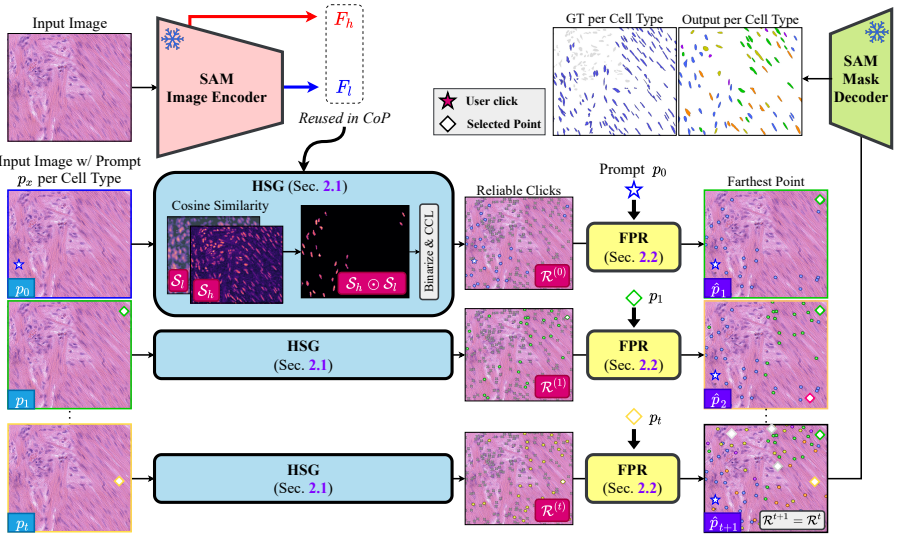

Fig. 3. Overview of Chain-of-Prompts (CoP). A frozen SAM encoder extracts F_h and F_l once per image. For each user click p_x (☆), HSG (Sec. 2.1) produces initial reliable points $\mathcal{R}^{(0)}$ via hierarchical similarity and connected-component labeling (CCL). FPR (Sec. 2.2) then expands $\mathcal{R}^{(0)}$ by iteratively prompting the farthest uncovered point (◆) until no new points are found. All propagated points per cell type are finally decoded into instance masks.

prompt. Second, Farthest Prompt Recursion (Sec. 2.2) then iteratively selects new prompts from this reliable set to expand spatial coverage until convergence ($\mathcal{R}^{(t+1)}$=$\mathcal{R}^{(t)}$). The resulting point set is decoded into instance masks via SAM's decoder.

2.1 Hierarchical Similarity Gating

A single feature scale cannot simultaneously achieve spatial precision and type selectivity. F_h localizes individual cells even among tightly packed neighbors, but also activates tissue regions with similar texture beyond the target cell type. Conversely, F_l selectively responds to the target type, but its coarse resolution causes neighboring instances to merge. HSG addresses this trade-off by combining both scales via element-wise gating to obtain a reliable point set $\mathcal{R}$ with high precision.

Given a point prompt p per cell type, we interpolate F_l to match the spatial resolution of F_h and compute two cosine similarity maps: $S_h(x) = \cos(F_h(x), F_h(p))$ and $S_l(x) = \cos(F_l(x), F_l(p))$. The element-wise product $S_h \odot S_l$ suppresses false activations in S_h that fall outside the target cell type according to S_l, while preserving spatially precise responses (Fig. 3, HSG). We then binarize the gated map with a non-parametric threshold $\tau = \mu(S_h \odot S_l) + \sigma(S_h \odot S_l)$, inspired by COIN [11], and apply connected-component labeling (CCL) to extract the

similarity-weighted centroid from each connected region, as it provides a simple and deterministic way to convert dense activations into discrete point prompts without additional hyperparameters. The resulting centroids form the reliable set $\mathcal{R}^{(0)} = \{c_1, \ldots, c_K\}$, which typically covers cells near the initial prompt but misses spatially distant instances.

2.2 Farthest Prompt Recursion

While HSG identifies highly reliable cells in the local vicinity of the prompt, feature similarity naturally decays across distant, morphologically diverse tissue regions. Consequently, a single prompt yields lower precision for distant cells. FPR addresses this by automatically selecting the point in $\mathcal{R}^{(t)}$ that is farthest from all previously used prompts $\mathcal{Q}^{(t)} = \{p_0, \ldots, p_t\}$ at each iteration t:

$$p_{t+1} = \arg\max_{c \in \mathcal{R}^{(t)}} \min_{q \in \mathcal{Q}^{(t)}} \|c - q\|_2. \tag{1}$$

By computing distance in image coordinates rather than feature space, we ensure each new prompt explores spatially uncovered tissue regions without feature drift. The selected prompt p_{t+1} is then fed back into HSG as a new prompt. Newly discovered points from the next round of HSG are merged into the reliable set: $\mathcal{R}^{(t+1)} = \mathcal{R}^{(t)} \cup \mathrm{HSG}(p_{t+1}, F_h, F_l)$. This cycle repeats until no new points are discovered ($\mathcal{R}^{(t+1)} = \mathcal{R}^{(t)}$), indicating that every target cell that shares feature similarity with the initial click instance has been identified. Finally, each point $r \in \mathcal{R}$ is decoded into an instance mask via SAM's decoder, where overlapping predictions are resolved through non-maximum suppression at IoU > 0.5.

3 Experiments

3.1 Implementation Details

All compared methods use their official code and pretrained weights. Open-vocabulary methods [24,10] use "cell" as the text prompt; for visual prompting, we provide a cropped cell patch as the reference image. Interactive baselines [12,19,2,1] receive N foreground clicks simulated by computing the centroid of each GT instance mask. Fully-supervised methods [9,7,22] are evaluated using their publicly released models trained on their respective datasets. CoP requires only T clicks (one per cell type present) for cell-type-annotated datasets and a single click for datasets without type labels, where most instances are morphologically similar and thus behave as a single cell type.

All experiments run on a single NVIDIA RTX A6000. On a 1000×1000 input, SAM3 image encoding takes ∼2 s as a one-time cost; each subsequent CoP click (HSG propagation + FPR until convergence) completes in ∼4 s on average, with individual FPR iterations at ∼170 ms. A typed image with T=3 cell types thus finishes in under 15 s excluding the encoder forward pass. Since CoP operates entirely in feature space without backpropagation, it adds negligible memory overhead beyond the frozen encoder.

Table 1. Quantitative comparison on cell-type-annotated benchmarks. $\mathcal{T}$: text prompt (*i.e.*, "cell"), $\mathcal{V}$: visual prompt (reference image patch), $\mathcal{M}$: pixel-level supervision for training, $\mathcal{P}_N$: one point per instance, $\mathcal{P}_T$: one point per cell type.

Method	Prompt	CoNIC		CoNSeP		GlaS	
		AJI ↑	Dice ↑	AJI ↑	Dice ↑	AJI ↑	Dice ↑
YOLOE [24] ICCV'25	$\mathcal{T}$	0.000	0.057	0.000	0.000	0.000	0.000
SAM3 [2] ICLR'26	$\mathcal{T}$	0.450	0.696	0.000	0.001	0.000	0.002
Rex-Omni [10] CVPR'26	$\mathcal{T}$	0.002	0.056	0.151	0.316	0.004	0.154
YOLOE [24] ICCV'25	$\mathcal{V}$	0.028	0.110	0.000	0.000	0.001	0.303
SAM3 [2] ICLR'26	$\mathcal{V}$	0.390	0.620	0.000	0.001	0.000	0.003
Rex-Omni [10] CVPR'26	$\mathcal{V}$	0.045	0.269	0.126	0.348	0.016	0.125
CellViT [9] MedIA'23	$\mathcal{M}$	0.371	0.670	0.495	0.802	0.265	0.643
CA-SAM2 [7] MICCAI'25	$\mathcal{M}$	0.269	0.561	0.382	0.700	0.213	0.593
CellPose3 [22] Nat. Methods'25	$\mathcal{M}$	0.173	0.333	0.158	0.354	0.033	0.080
μSAM [1] Nat. Methods'25	$\mathcal{P}_N$	0.705	0.834	0.316	0.470	0.279	0.529
+ **CoP (Ours)**	$\mathcal{P}_T$	**0.652**	**0.791**	**0.286**	**0.431**	**0.252**	**0.491**
SAM3 [2] ICLR'26	$\mathcal{P}_N$	0.641	0.795	0.411	0.760	0.327	0.704
+ **CoP (Ours)**	$\mathcal{P}_T$	**0.579**	**0.759**	**0.374**	**0.729**	**0.292**	**0.673**

We evaluate on seven cell instance segmentation benchmarks using their official test splits. Three provide cell-type annotations: CoNIC [6] (6 types), CoNSeP [5] (4 types), and GlaS [21]. Four contain instance masks without type labels: MoNuSeg [13], TNBC [18], CryoNuSeg [16], and CPM-17 [23]. Following prior studies [6,7], we report AJI (instance-level overlap with false-positive penalty) and Dice (pixel-level foreground overlap).

3.2 Comparison with State-of-the-art Approaches

Evaluating on three cell-type-annotated benchmarks (Tab. 1), interactive models (*e.g.*, [2]) and open-vocabulary detectors (*e.g.*, [24,10]) use text/visual ($\mathcal{T}/\mathcal{V}$) prompts to avoid per-instance interaction. However, they fail to generalize: SAM3 [2] yields predictions only on CoNIC [6], whereas Rex-Omni [10] is restricted to CoNSeP [5]. This is because text/visual prompt pathways depend on domain-specific alignment learned during training, whereas point prompts bypass this alignment and directly query the frozen image encoder, whose features already separate cell instances regardless of domain (Fig. 5). Fully-supervised methods [9,7,22] similarly suffer out-of-distribution degradation: on CoNIC [6], the strongest baseline CellViT [9] achieves an AJI of only 0.371, well below zero-shot point-prompted models. Qualitatively (Fig. 4), supervised baselines miss entire cell populations, whereas CoP discovers them via recursive feature propagation from user clicks. Using only ∼3 clicks per image (one per cell type), CoP with SAM3 reduces prompt costs by >97% versus per-instance annotation ($\mathcal{P}_N$), retaining ≥90% of $\mathcal{P}_N$ performance across all three benchmarks.

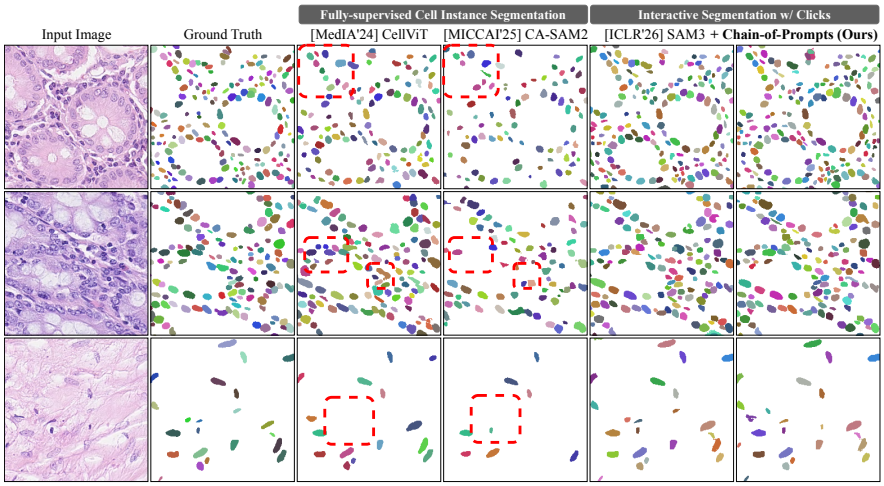

Fig. 4. Qualitative comparison on CoNIC [6]. Fully-supervised methods miss cell populations absent from their training set (red dashed boxes), whereas CoP discovers them from a single click per type.

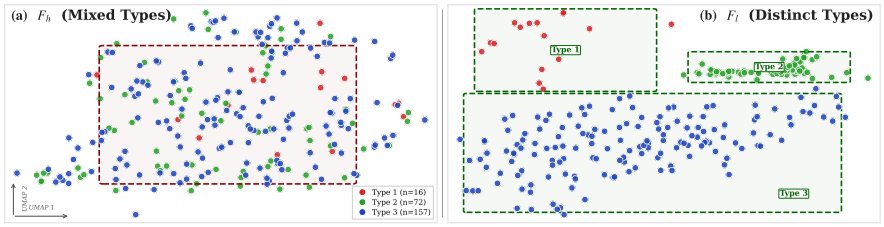

Fig. 5. UMAP [17] of SAM's frozen image encoder features at GT instance centroids. The UMAP embeddings are extracted from the input image used in Fig. 2. (a) F_h mixes cell types; (b) F_l groups same-type cells without any training.

On four benchmarks without cell-type annotations (Tab. 2), instances within each image are morphologically homogeneous, forming a single cell type. CoP therefore operates from one click per image, propagating it to all instances via iterative FPR. Under this setting, CoP retains over 99% of the per-instance prompting performance for both μSAM [1] and SAM3 [2], while consistently outperforming fully-supervised methods [22,7].

3.3 Ablation Study

We ablate the core design choices of CoP on CoNIC [6]. With all proposed components enabled, CoP achieves AJI 0.579 (90% of the per-instance upper bound 0.641, Tab. 1). Each component is critical: recursive propagation accounts for 65% relative AJI gain, multi-scale gating contributes 20–39% (depending on

Table 2. Quantitative results on benchmarks without cell-type annotations. Most cells share similar morphology within each image, allowing CoP to segment from one click.

Method	MoNuSeg		TNBC		CryoNuSeg		CPM-17	
	AJI	Dice	AJI	Dice	AJI	Dice	AJI	Dice
CellViT [9] MedIA'23	0.676	0.832	0.669	0.817	0.492	0.800	0.703	0.853
CA-SAM2 [7] MICCAI'25	0.645	0.807	0.644	0.799	0.468	0.781	0.657	0.819
CellPose3 [22] Nat. Methods'25	0.316	0.499	0.224	0.365	0.140	0.303	0.326	0.479
μSAM [1] Nat. Methods'25	0.732	0.849	0.809	0.893	0.567	0.746	0.763	0.753
+ **CoP (Ours)**	**0.729**	**0.845**	**0.805**	**0.884**	**0.563**	**0.745**	**0.754**	**0.749**
SAM3 [2] ICLR'26	0.712	0.838	0.752	0.863	0.518	0.706	0.766	0.874
+ **CoP (Ours)**	**0.706**	**0.837**	**0.750**	**0.864**	**0.503**	**0.696**	**0.761**	**0.871**

which scale is removed), and performance is robust to initial-click choice (±0.003 std). We isolate each contribution below.

- **Effect of recursive propagation.** Without FPR (Sec. 2.2), HSG (Sec. 2.1) alone produces reliable points only near the initial click, reaching AJI 0.203 (−65%, ↓0.376). Adding FPR (Sec. 2.2) restores AJI to 0.579, confirming that recursive expansion is essential for whole-image coverage.
- **Selection strategy within FPR (Sec. 2.2).** Farthest-point (0.579) outperforms closest-point (0.492, −15%, ↓0.087) and midpoint (0.515, −11%, ↓0.064), both of which tend to revisit covered areas. Thus, FPR resolves the spatial coverage bottleneck of HSG (Sec. 2.1) by maximizing prompt-to-prompt distance.
- **Multi-scale similarity gating of HSG (Sec. 2.1).** Replacing $S_h \odot S_l$ with S_h alone degrades AJI to 0.463 (↓0.116), as precision drops below 0.60 by t=15 due to tissue-level false positives propagating through each recursion. Using S_l alone yields 0.351 (↓0.228), as its coarse resolution causes poor prompt localization. $S_h \odot S_l$ maintains precision above 0.96 throughout all iterations at comparable recall. This is because F_l, extracted from deeper layers with a larger receptive field, encodes overall morphology and naturally clusters cells by their semantic identity (Fig. 5), while F_h precisely locates cell centers but with high semantic uncertainty. By gating them together, HSG filters out the spatial noise of F_l and the semantic uncertainty of F_h.
- **Initial Click Sensitivity.** We repeat all CoNIC experiments (Tab. 1) with 30 random seeds. CoP achieves a mean AJI of 0.579 ± 0.003, indicating that performance is robust to the choice of initial prompt location.
- **Representative Failure Modes.** CoP inherits the base model's limitations: instances that SAM3 [2] cannot segment from a correct point prompt will also be missed by CoP. CoP further assumes that same-type cells share coherent appearance in feature space, which may not hold under extreme morphological heterogeneity within a single cell type.

4 Conclusion

In this paper, we present **Chain-of-Prompts (CoP)**, a training-free framework that discovers and segments all same-type cells from a single user click by recursively propagating prompts through frozen SAM features. Our key finding is that SAM's frozen image encoder already clusters same-type cells in its multi-scale feature space before any prompt is given, and CoP exploits this intrinsic property through non-parametric gating without additional training. Across seven diverse benchmarks, CoP retains over 90% of per-instance performance while requiring up to 97% fewer clicks, generalizes to unseen cell types without adaptation, and even surpasses fully-supervised methods. By reducing hundreds of manual annotations to a single click per cell type, CoP demonstrates that interactive foundation models can be leveraged far more efficiently than the current paradigm assumes, establishing group prompting as a practical and scalable alternative for clinical workflows.

Acknowledgements. This work was partly supported by the KHIDI grant funded by the Korean government (MOHW) [No.RS-2025-02307233], the NRF or IITP grants funded by the Korean government (MSIT) [No.RS-2026-25472075, No.RS-2026-25483206, No.RS-2025-02305581, No.RS-2025-25442338 (AI Star Fellowship-SNU), and No.RS-2021II211343 (SNU AI)], the ITIP grant funded by the Korean government (MOTIR) [No.RS-2026-25549946], the Advanced GPU Utilization and AI Computing Infrastructure Enhancement User Support Programs funded by the Korean government (MSIT) [No.05-26-04-0094], the Research grant from SNU, and the Strategic Hub grant for International Research Collaboration of SNU.

Kyungsu Kim is affiliated with the School of Transdisciplinary Innovations, Department of Biomedical Science, Interdisciplinary Program in Artificial Intelligence (IPAI), Medical Research Center, and AI Institute at SNU.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Archit, A., Freckmann, L., Nair, S., Pape, C., et al.: Segment anything for microscopy. Nature Methods **22**(3), 579–591 (2025). https://doi.org/10.1038/s41592-024-02580-4
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. In: ICLR (2026)
3. Chen, P., Zhu, C., Shui, Z., Cai, J., Zheng, S., Zhang, S., Yang, L.: Exploring unsupervised cell recognition with prior self-activation maps. In: MICCAI. pp. 559–568. Springer Nature Switzerland, Cham (2023)
4. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**(2), 303–338 (2010)

5. Graham, S., Vu, Q.D., Jahanifar, M., Abraham, A., Durr, N.J., Rajpoot, N., Raza, S.E.A.: A dataset for prostate cancer semantic segmentation and gland detection from whole slide images. IEEE Transactions on Medical Imaging **40**(12), 3923–3933 (2021). https://doi.org/10.1109/TMI.2021.3113172
6. Graham, S., Vu, Q.D., Raza, S.E.A., Rajpoot, N., et al.: Conic challenge: Pushing the frontiers of nuclear detection, segmentation, classification and counting. MedIA **91**, 103049 (2024). https://doi.org/10.1016/j.media.2023.103049
7. Huang, H., He, H., Xu, L., Zhu, X., Feng, S., Fu, G.: Ca-sam2: Sam2-based context-aware network with auto-prompting for nuclei instance segmentation. In: MICCAI. pp. 86–95. Springer Nature Switzerland (2025)
8. Huang, L., Liang, Y., Liu, J.: DES-SAM: Distillation-Enhanced Semantic SAM for Cervical Nuclear Segmentation with Box Annotation . In: MICCAI. vol. LNCS 15009. Springer Nature Switzerland (October 2024)
9. Hörst, F., Rempe, M., Heine, L., Seibold, C., Keyl, J., Baldini, G., Ugurel, S., Siveke, J., Grünwald, B., Egger, J., Kleesiek, J.: Cellvit: Vision transformers for precise cell segmentation and classification. MedIA **94**, 103143 (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103143
10. Jiang, Q., Huo, J., Chen, X., Xiong, Y., Zeng, Z., Chen, Y., Ren, T., Yu, J., Zhang, L.: Detect anything via next point prediction. In: CVPR (2026)
11. Jo, S., Lee, S.J., Lee, S., Hong, S., Seo, H., Kim, K.: Coin: Confidence score-guided distillation for annotation-free cell segmentation. In: ICCV. pp. 20324–20335 (2025)
12. 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: ICCV. pp. 4015–4026 (2023)
13. Kumar, N., Verma, R., Anand, D., Zhou, Y., Onder, O.F., Tsougenis, E., Chen, H., Heng, P.A., Li, J., Hu, Z., Wang, Y., Koohbanani, N.A., Jahanifar, M., Tajeddin, N.Z., Gooya, A., Rajpoot, N., Ren, X., Zhou, S., Wang, Q., Shen, D., Yang, C.K., Weng, C.H., Yu, W.H., Yeh, C.Y., Yang, S., Xu, S., Yeung, P.H., Sun, P., Mahbod, A., Schaefer, G., Ellinger, I., Ecker, R., Smedby, O., Wang, C., Chidester, B., Ton, T.V., Tran, M.T., Ma, J., Do, M.N., Graham, S., Vu, Q.D., Kwak, J.T., Gunda, A., Chunduri, R., Hu, C., Zhou, X., Lotfi, D., Safdari, R., Kascenas, A., O'Neil, A., Eschweiler, D., Stegmaier, J., Cui, Y., Yin, B., Chen, K., Tian, X., Gruening, P., Barth, E., Arbel, E., Remer, I., Ben-Dor, A., Sirazitdinova, E., Kohl, M., Braunewell, S., Li, Y., Xie, X., Shen, L., Ma, J., Baksi, K.D., Khan, M.A., Choo, J., Colomer, A., Naranjo, V., Pei, L., Iftekharuddin, K.M., Roy, K., Bhattacharjee, D., Pedraza, A., Bueno, M.G., Devanathan, S., Radhakrishnan, S., Koduganty, P., Wu, Z., Cai, G., Liu, X., Wang, Y., Sethi, A.: A multi-organ nucleus segmentation challenge. TMI **39**(5), 1380–1391 (2020). https://doi.org/10.1109/TMI.2019.2947628
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
15. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: ECCV. pp. 38–55. Springer (2024)
16. Mahbod, A., Schaefer, G., Bancher, B., Löw, C., Dorffner, G., Ecker, R., Ellinger, I.: Cryonuseg: A dataset for nuclei instance segmentation of cryosectioned h&e-stained histological images. Computers in Biology and Medicine **132**, 104349 (2021). https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104349, https://www.sciencedirect.com/science/article/pii/S0010482521001438
17. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)

18. Naylor, P., Laé, M., Reyal, F., Walter, T.: Segmentation of nuclei in histopathology images by deep regression of the distance map. IEEE Transactions on Medical Imaging **38**(2), 448–459 (2018)
19. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: ICLR (2025)
20. Sahasrabudhe, M., Christodoulidis, S., Salgado, R., Michiels, S., Loi, S., André, F., Paragios, N., Vakalopoulou, M.: Self-supervised nuclei segmentation in histopathological images using attention. In: MICCAI. pp. 393–402. Springer (2020)
21. Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al.: Gland segmentation in colon histology images: The GlaS challenge contest. MedIA **35**, 489–502 (2017)
22. Stringer, C., Pachitariu, M.: Cellpose3: one-click image restoration for improved cellular segmentation. Nature Methods **22**(3), 592–599 (2025). https://doi.org/10.1038/s41592-025-02595-5
23. Vu, Q.D., Graham, S., Kurc, T., To, M.N.N., Shaban, M., Qaiser, T., Koohbanani, N.A., Khurram, S.A., Kalpathy-Cramer, J., Zhao, T., Gupta, R., Kwak, J.T., Rajpoot, N., Saltz, J., Farahani, K.: Methods for segmentation and classification of digital microscopy tissue images. Frontiers in Bioengineering and Biotechnology **Volume 7 - 2019** (2019). https://doi.org/10.3389/fbioe.2019.00053, https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2019.00053
24. Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., Ding, G.: Yoloe: Real-time seeing anything. In: ICCV. pp. 24591–24602 (2025)