IVAAN: Instance-Level Vision-Language Alignment via Attribute-Guided Text Prompt Generation for Nuclei Analysis (Supplementary Material)

Jaehoon Jeong[1,2] Yi Hu[2] Soopil Kim[1] Jongseong Jang[2] Soonyoung Lee[2] Sang Hyun Park[1]

[1]DGIST, Daegu, Republic of Korea [2]LG AI Research, Seoul, Republic of Korea

{j.hoon, soopilkim, shpark13135}@dgist.ac.kr, {j.jang, soonyoung.lee}@lgresearch.ai

The supplementary material is structured as follows:

1. We demonstrate generalization ability of the proposed model in Section A.
2. We present the details on attribute definition and selection B.
3. We provide additional qualitative results of the proposed method in Section C.

A. Generalization ability

Test Group	mPQ		PQ		AJI	
	PromptNucSeg	Ours	PromptNucSeg	Ours	PromptNucSeg	Ours
Group 1	0.276	0.317	0.647	0.649	0.662	0.688
Group 2	0.233	0.293	0.577	0.610	0.581	0.612
Group 3	0.315	0.335	0.651	0.664	0.665	0.682
Group 4	0.326	0.343	0.625	0.649	0.638	0.671
Mean	0.288	**0.322**	0.625	**0.643**	0.637	**0.663**
std	0.037	**0.019**	0.029	**0.020**	0.034	**0.030**

Table 1. 4-fold cross validation results across organ groups. Each group is used for testing while training on the remaining three groups. mPQ, PQ and AJI are reported. Our model consistently outperforms the previous SOTA (PromptNucSeg [1]) across all groups. All experiments were conducted with ViT-L

To evaluate the model's cross-organ generalization ability, we conducted leave-one-out cross validation across the 19 organs in the PanNuke dataset. However, the dead class is extremely imbalanced among organs. Only 11 of the 19 organs contain any dead instances, and among those, 5 organs have fewer than 50 dead nuclei, while only 4 organs contain more than 500. Such imbalance can cause unstable, inaccurate, and unreliable results when each organ is evaluated individually, as organs with very few dead samples yield fluctuating scores.

To address this issue, we grouped the organs into four subsets so that each group contains a more balanced number of dead instances. Specifically:

- Group 1: lung, bladder, kidney, prostate
- Group 2: colon, skin, pancreatic, liver, testis
- Group 3: uterus, esophagus, ovarian, thyroid, bile duct
- Group 4: stomach, head & neck, cervix, breast, adrenal gland

Each group contains approximately 5,000, 1,500, 1,000, and 1,000 dead instances, respectively. We then performed 4-fold cross validation by training on three groups and testing on the remaining one in rotation.

As summarized in Table 1, our method outperforms the previous SOTA (PromptNucSeg) across all groups and metrics (mPQ, PQ, and AJI). Here, mPQ denotes the mean of class-wise PQ scores, whereas PQ represents the image-wise mean PQ. Notably, our model not only achieves higher mean performance but also shows substantially lower variance across groups. In

particular, the standard deviation of mPQ decreases from 0.037 in PromptNucSeg to 0.019 in ours—nearly a 50% reduction. Because each group includes nuclei from distinct organs, even the same semantic class may present strong visual variation between groups, often leading to large fluctuations in class-wise PQ. The reduced variance therefore demonstrates our model's superior cross-organ generalization, maintaining consistent segmentation and classification across visually diverse domains.

B. Details on Attribute Definition and Selection

To facilitate reproducibility, this section details the equations and hyperparameters used to construct the text prompts.

B.1 Definition of Per-Nucleus Attributes

We compute a set of morphological and appearance attributes for each nucleus from the ground-truth (GT) instance masks and the corresponding RGB image. These attributes are used both for our GT analysis and as the basis for text prompts. For the attribute analysis, we discard nuclei that are truncated by the image boundary and very small instances whose area is less than 10 pixels, since their size and shape cannot be reliably estimated.

Let $M \subset \mathbb{Z}^2$ denote the binary mask of a single nucleus in an image $I : \mathbb{Z}^2 \to \mathbb{R}^3$. We denote by A the area (number of pixels) of the mask, by P its perimeter, and by A_{cvx} the area of its convex hull (all obtained via `skimage`).

Morphological attributes

Equivalent diameter (size). We define the nucleus size as the diameter of a circle with the same area as the mask:

$$d_{\text{eq}} = \sqrt{\frac{4A}{\pi}}.$$

Solidity. Solidity measures how convex the nucleus is:

$$\text{solidity} = \frac{A}{\max(A_{\text{cvx}}, 10^{-6})}.$$

Boundary complexity (perimeter ratio). To quantify contour irregularity, we normalize the perimeter by the circumference of a circle with diameter d_{eq}:

$$\text{perim\_ratio} = \frac{P}{2\pi(d_{\text{eq}}/2)}.$$

A value close to 1 corresponds to a nearly circular contour, while larger values indicate increasingly complex/irregular boundaries. This is used as our boundary-complexity attribute.

Eccentricity (shape). From the best-fitting ellipse of the mask, we obtain the eccentricity $\text{eccentricity} \in [0, 1]$, where 0 corresponds to a circle and values closer to 1 correspond to elongated shapes. This is used as our shape attribute.

Extent and aspect ratio. We also compute

$$\text{extent} = \frac{A}{A_{\text{bbox}}}, \quad \text{aspect\_ratio} = \frac{\text{major\_axis\_length}}{\text{minor\_axis\_length}},$$

for additional morphological analysis.

Color and texture attributes

We measure stain intensity and chromatin texture from the hematoxylin channel $H(x, y)$. For each image, we first compute a stain-normalized hematoxylin channel. All intensity-based attributes below are computed on this channel.

For robustness, we slightly erode the mask before measuring intensity, in order to avoid boundary noise. Let M' denote the eroded mask: $M' = \text{erode}\big(M, \text{rect}(3 \times 3)\big)$, when the mask area is sufficiently large; otherwise we use $M' = M$.

For each nucleus, we compute raw color and texture values:

$$c_{\text{raw}} = \frac{1}{|M'|} \sum_{(x,y)\in M'} H(x, y), \quad t_{\text{raw}} = \text{Var}\big(H(x, y) : (x, y) \in M'\big).$$

To make the attributes comparable across images, we express them as per-image z-scores. For each image, we collect c_{raw} and t_{raw} over all nuclei in that image, and compute

$$\mu_c = \text{mean}(c_{\text{raw}}), \qquad \sigma_c = \text{std}(c_{\text{raw}}) + 10^{-6},$$
$$\mu_t = \text{mean}(t_{\text{raw}}), \qquad \sigma_t = \text{std}(t_{\text{raw}}) + 10^{-6}.$$

Stain intensity (color_z). The color attribute is the per-image z-score of the mean hematoxylin intensity:

$$\mathrm{color\_z} = \frac{c_{\mathrm{raw}} - \mu_c}{\sigma_c}.$$

Here, larger values correspond to nuclei that are darker (more strongly stained) than the image-specific average, and smaller values correspond to lighter nuclei.

Texture (texture_z). The chromatin texture attribute is the per-image z-score of the hematoxylin variance:

$$\mathrm{texture\_z} = \frac{t_{\mathrm{raw}} - \mu_t}{\sigma_t}.$$

Larger values indicate more heterogeneous/"coarse" chromatin, whereas smaller values indicate more homogeneous/"smooth" nuclei.

Core–rim contrast attribute

To capture core–rim intensity differences inside a nucleus, we split the mask into an inner core and an outer rim via morphological erosion.

Given the equivalent diameter d_{eq}, we choose the erosion size $k = \mathrm{clip}\Big(\mathrm{round}(0.08\, d_{\mathrm{eq}}), 2, 6\Big)$ and use a $k \times k$ rectangular structuring element. The core and rim masks are defined as $M_{\mathrm{core}} = \mathrm{erode}\big(M, \mathrm{rect}(k \times k)\big),\ M_{\mathrm{rim}} = M \setminus M_{\mathrm{core}}$.

If either region is empty, we omit this attribute.

We then compute the mean hematoxylin intensity in each region:

$$c_{\mathrm{core}} = \frac{1}{|M_{\mathrm{core}}|} \sum_{(x,y)\in M_{\mathrm{core}}} H(x,y), \quad c_{\mathrm{rim}} = \frac{1}{|M_{\mathrm{rim}}|} \sum_{(x,y)\in M_{\mathrm{rim}}} H(x,y),$$

and their difference $\Delta_{\mathrm{core}} = c_{\mathrm{core}} - c_{\mathrm{rim}}$.

To express this in the same scale as the stain intensity attribute, we normalize by the per-image standard deviation σ_c used for $\mathrm{color\_z}$:

$$\mathrm{core\_contrast\_z} = \frac{\Delta_{\mathrm{core}}}{\sigma_c}.$$

Positive values indicate nuclei whose inner core is darker than their rim (strong core–rim contrast), whereas negative values indicate the opposite.

Boundary sharpness attribute

We quantify boundary sharpness by averaging the gradient magnitude along a thin band around the nucleus contour. For a given nucleus, we extract a scalar channel $C(x, y)$, using the hematoxylin channel H when available, and otherwise grayscale intensity.

We construct a boundary band via a morphological gradient:

$$M_{\mathrm{dil}} = \mathrm{dilate}\big(M, \mathrm{rect}(3\times 3)\big), \quad M_{\mathrm{ero}} = \mathrm{erode}\big(M, \mathrm{rect}(3\times 3)\big), \quad B = M_{\mathrm{dil}} \oplus M_{\mathrm{ero}}.$$

We then compute the Sobel gradient magnitude on C, $G = |\nabla C|_{\mathrm{Sobel}}$, and define

$$\mathrm{boundary\_grad\_mean} = \frac{1}{|B|} \sum_{(x,y)\in B} G(x,y).$$

Higher values correspond to sharper, more pronounced boundaries.

Nearest-neighbor distance attribute

For some analyses we also measure local nucleus density. For each nucleus we compute its centroid (x_i, y_i). The nearest-neighbor distance is defined as

$$\mathrm{nn\_dist\_px}(i) = \min_{j\neq i} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

Smaller distances indicate crowded regions, whereas large distances correspond to more isolated nuclei.

B.2 Attribute selection based on class separability

We consider a pool of candidate per-nucleus metrics $\mathcal{M} = \{m_1, \dots, m_K\}$ (e.g., size, eccentricity, ...). Let $\mathcal{C} = \{1, \dots, C\}$ denote the set of semantic classes (e.g., neoplastic, inflammatory, ...).

For each metric $m \in \mathcal{M}$ and each class $c \in \mathcal{C}$, we collect all per-nucleus values

$$X_c^{(m)} = \{x_{c,1}^{(m)}, x_{c,2}^{(m)}, \dots, x_{c,n_c}^{(m)}\}.$$

Pairwise separability (Cohen's d)

For every pair of distinct classes $i, j \in \mathcal{C}$, we quantify how well metric m separates them using Cohen's d effect size:

$$d_{ij}^{(m)} = \frac{\mu_i^{(m)} - \mu_j^{(m)}}{s_p^{(m)}},$$

where $\mu_i^{(m)}$ and $\mu_j^{(m)}$ are the sample means of $X_i^{(m)}$ and $X_j^{(m)}$, and

$$s_p^{(m)} = \sqrt{\frac{(n_i - 1)(s_i^{(m)})^2 + (n_j - 1)(s_j^{(m)})^2}{n_i + n_j - 2}}$$

is the pooled standard deviation based on the unbiased sample variances $(s_i^{(m)})^2$ and $(s_j^{(m)})^2$. We use the absolute value $|d_{ij}^{(m)}|$ as a scale-invariant separability measure (we are interested in how strongly the distributions differ, not in which class has the larger mean).

Global and per-class separability scores

To summarize per-metric separability across all classes, we define:

Global separability of metric m: median pairwise effect size over all class pairs,

$$S_{\text{global}}(m) = \text{median}_{1 \le i < j \le C}\big(|d_{ij}^{(m)}|\big).$$

This captures how well m separates typical class pairs across the whole label space.

Per-class separability of metric m for class c:

$$S_{\text{class}}(m, c) = \text{median}_{j \in \mathcal{C},\, j \neq c}\big(|d_{cj}^{(m)}|\big).$$

This measures how well metric m separates a specific class c from all other classes. In practice, we prioritize metrics with high global separability when selecting candidate prompt attributes, while also checking the per-class scores to ensure that important classes are well separated.

We optionally aggregate these into a single overall score per metric:

$$S_{\text{overall}}(m) = \alpha\, S_{\text{global}}(m) + (1 - \alpha)\, \frac{1}{C} \sum_{c=1}^{C} S_{\text{class}}(m, c),$$

with $\alpha \in [0, 1]$ (we use $\alpha = 0.5$ in our experiments), which balances "global" discriminative power and per-class coverage.

Redundancy penalty between metrics

Different metrics may capture highly overlapping information (e.g., size vs. area, or perimeter ratio vs. solidity). To avoid selecting redundant attributes, we penalize metrics that are strongly correlated with already selected ones.

For any pair of metrics $m, m' \in \mathcal{M}$, we compute the Pearson correlation coefficient over all nuclei:

$$\rho(m, m') = \text{corr}\big(X^{(m)}, X^{(m')}\big),$$

where $X^{(m)}$ denotes the concatenation of all class-specific samples $X_c^{(m)}$.

Given a current set of selected attributes $A \subset \mathcal{M}$, the redundancy of a candidate metric $m \notin A$ is quantified as

$$R(m \mid A) = \begin{cases} 0, & \text{if } A = \varnothing, \\ \dfrac{1}{|A|} \displaystyle\sum_{m' \in A} \big|\rho(m, m')\big|, & \text{otherwise.} \end{cases}$$

High $R(m \mid A)$ indicates that m is largely redundant with the already chosen attributes.

Greedy selection of four attributes

We select four attributes in a greedy manner by maximizing a separability–redundancy trade-off. Let $\lambda \geq 0$ control the strength of redundancy penalization.

Initialization. Choose the first attribute as the metric with the highest overall separability:

$$m_1 = \arg\max_{m \in \mathcal{M}} S_{\text{overall}}(m),$$

and set $A = \{m_1\}$.

Greedy expansion. For $k = 2, 3, 4$, we select

$$m_k = \arg\max_{m \in \mathcal{M} \setminus A} J(m \mid A),$$

where the objective is

$$J(m \mid A) = S_{\text{overall}}(m) - \lambda\, R(m \mid A).$$

We then update $A \leftarrow A \cup \{m_k\}$.

In practice, we restrict $\mathcal{M}$ to semantically interpretable candidates (size, shape, stain intensity, boundary complexity, density, etc.), and use a small λ (e.g., $\lambda \in [0.3, 0.5]$). This encourages the selected attributes to be (i) strongly discriminative across classes and (ii) complementary to each other rather than redundant.

With this procedure, the four attributes consistently selected in our setting are: size: equivalent diameter (d_{eq}), shape: eccentricity, stain intensity: color_z, boundary complexity: perim_ratio, which together span distinct morphological and appearance factors while providing high class separability in the GT analysis.

C. Additional Qualitative Results

We present qualitative results of the proposed method on the PanNuke datasets, as shown in Fig. S1. To demonstrate the robustness of our model, we visualize diverse samples of nuclei from each of the 19 organs. As illustrated in Fig. S1, our method successfully classified and segmented the nuclei.As can be seen in the results for bile duct, skin, and testis, crowded and mixed connective and inflammatory nuclei are still well classified even though they look very similar. This is particularly challenging for the nuclei classification models, as connective and inflammatory nuclei often occupy the similar feature space.

References

[1] Zhongyi Shui, Yunlong Zhang, Kai Yao, Chenglu Zhu, Sunyi Zheng, Jingxiong Li, Honglin Li, Yuxuan Sun, Ruizhe Guo, and Lin Yang. Unleashing the power of prompt-driven nucleus instance segmentation. In *European conference on computer vision*, pages 288–304. Springer, 2024. 1

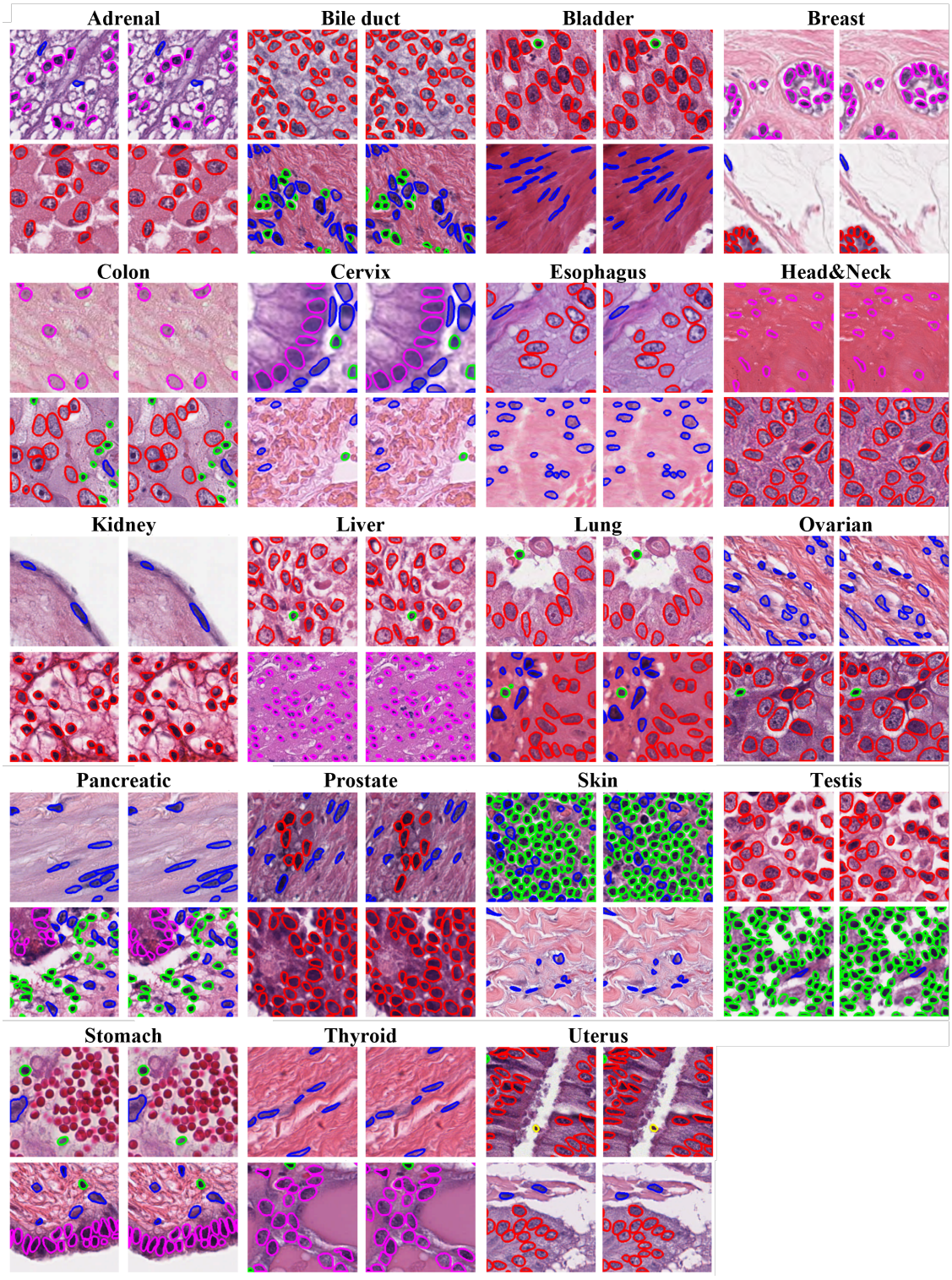

Figure S1. Qualitative results on the PanNuke dataset. We show two example images for each of the 19 organs. Left: ground-truth annotations. Right: predictions from our model. Instance boundaries are color-coded as follows: red (neoplastic), blue (connective), green (inflammatory), yellow (dead), and magenta (epithelial).