DISC: Latent Diffusion Models with Self-Distillation from Separated Conditions for Prostate Cancer Grading (Supplementary Document)

Man M. Ho[1] Elham Ghelichkhan [1] Yosep Chong[2,4] Yufei Zhou[3]
Beatrice Knudsen[4,5] Tolga Tasdizen[1,6]
[1] Scientific Computing and Imaging Institute, University of Utah, Utah, USA
[2] The Catholic University of Korea College of Medicine, Seoul, Korea
[3] Case Western Reserve University, Ohio, USA
[4] Departmant of Pathology, University of Utah, Utah, USA
[5] Huntsman Cancer Institute, University of Utah Health, Utah, USA
[6] Department of Electrical and Computer Engineering, University of Utah, Utah, USA

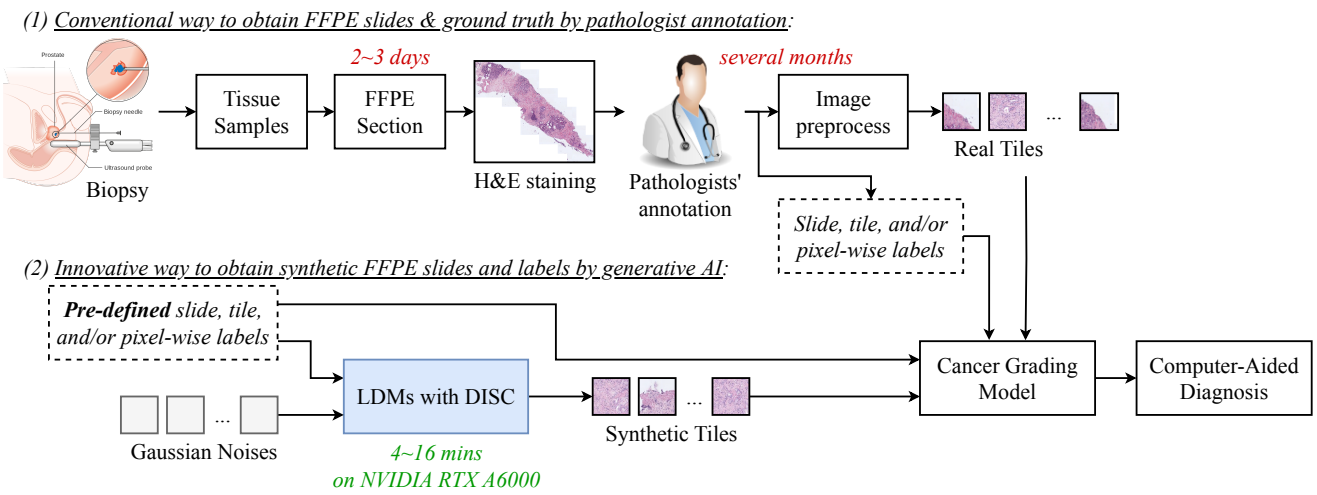


Figure 1. Application Overview on Our Training Whole Slide Images (WSIs) Augmentation. Pathologists commonly rely on Formalin-Fixed Paraffin-Embedded (FFPE) slides for crucial diagnostic, prognostic, and treatment decisions. To alleviate the pathologist's workload, innovative machine learning-based cancer grading techniques have emerged. These methods leverage tiles extracted from slides to offer supplementary insights, aiding pathologists or directly informing prognosis and treatment strategies. However, it is costly to prepare FFPE slides (typically 2-3 days to prepare a single FFPE slide) and substantial human involvement required to annotate these slides (which may cost several months) for these cancer grading models. In this study, we present Latent Diffusion Models (LDMs) with Self-**Di**stillation from **S**eparated **C**onditions (DISC). This novel methodology enables the generation of tiles based on provided slide-level labels, improving the efficiency and accuracy in cancer diagnostics.

1. Introduction of Training Whole Slide Images Preparation for Computer-Aided Prostate Cancer Diagnosis and Our Proposed Generative Models for Data Augmentation

We provide an overview of how real Whole Slide Images (WSIs) and our synthetic WSIs are generated and their applications for prostate cancer diagnosis, prognosis, and treatment in Figure 1.

2. Pre-defining Tile Shape Masks and Mask Sampling

See Figure 2.

3. Ablation Study on Random Weights

To further prove the effectiveness of our sampling technique where Random Weights (probability of being chosen for a label) are applied to make sure the generated primary and secondary Gleason Grades (GGs) are similar to pre-defined GGs, we train two models with our empirically-set Random Weights as [40%, 25%, 5%, 30%] for choosing labels [Primary GG, Secondary GG, Tertiary GG, Non-Cancer], respectively, and without the Random Weights, where all labels (excepting Tertiary GG) have the same chance of being chosen, as shown in Figure 3.

4. Ablation Study on LDMs' Conditions

We provide additional results for ablation study on LDMs' conditions including Tile Labels, Tile+Slide Labels, and Pixel-wise Labels in Figures 4 and 5.

5. Generated tiles by LayoutDiffusion

We convert the sampled annotated masks to COCO-like layouts [3] by finding contours and bounding boxes for all pixel-wise labels to train LayoutDiffusion [9]. We observe that LayoutDiffusion [9] finds difficulty dealing with overlapped bounding boxes as they are weak labels and not specific enough to blend semantic information. Hence, it tends to ignore small bounding boxes and generate the patterns for the larger ones, as shown in Figure 6.

6. Additional Results

Qualitative Comparison. We provide an additional qualitative comparison between the baseline Stable Diffusion (SD) [6] and our ablation models SD with **S**eparated **C**onditions (SD-SC), SD with Self-**Di**stillation from **S**eparated **C**onditions (SD-DISC), and SD-DISC also fine-tuned with real tiles (SD-DISC-CoTrain) in Figures 7, 8, and 9.

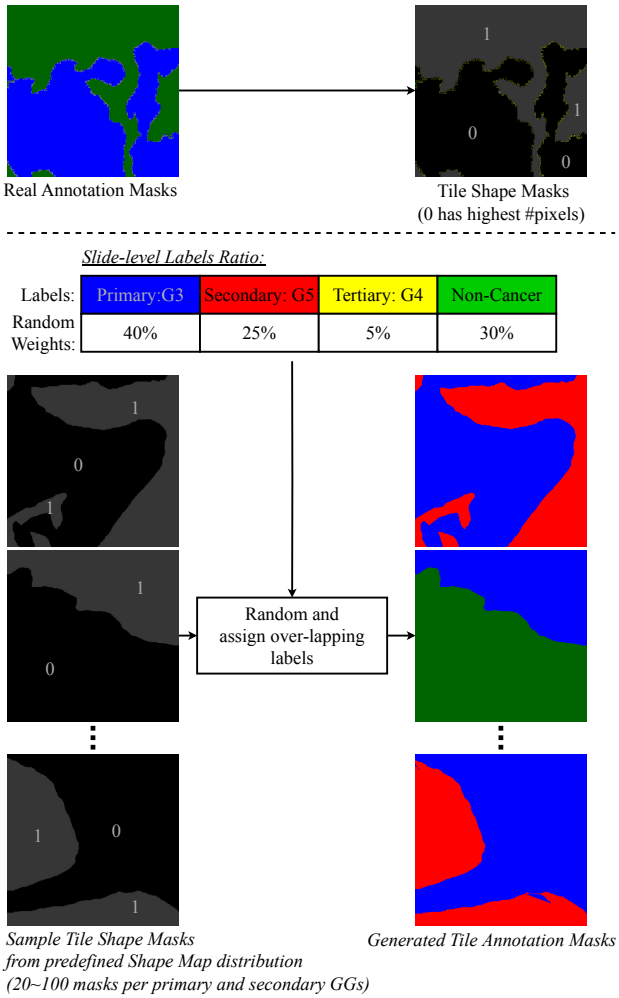
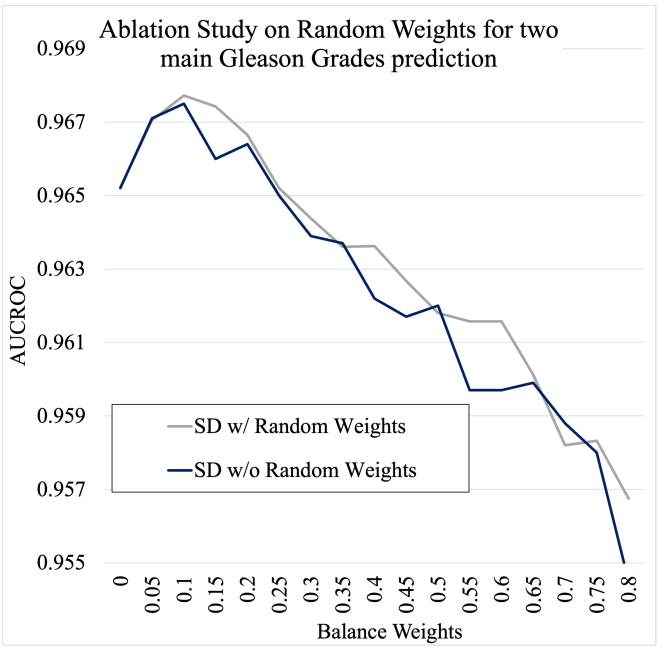


Figure 2. We eliminate semantic information from human-annotated masks in SICAPv2 [8], converting them to tile shape masks with labels mapped to their frequency distribution. To generate a tile set for slide-level classification via Multiple Instance Learning, we randomly select 20-100 annotation shape masks per designated primary and secondary GGs, with non-overlapping cancer grading labels distributed according to random weights. When the primary GG is Non-Cancer, the Non-Cancer label is exclusively applied.

Quantitative Comparison on Pixel-Level Classification. On the in-distribution SICAPv2 dataset, Carcino-Net with SD-DISC achieves the best pixel-level precision of **0.8061±0.0599**, an increase of **+0.027** from the baseline, thanks to the diversified patterns generated by our generative models. On out-of-distribution LAPC, Carcino-Net trained with SD obtained the best precision as **0.7863±0.0547**, **+0.1202** increased from Carcino-Net itself, while SD-DISC-CoTrain provided more stable performance across classes, with **0.7431±0.1452**, as shown in Table 1.



(a)

SD w/ Random Weights:

Primary	Secondary	Tertiary	Non-Cancer
40%	25%	5%	30%

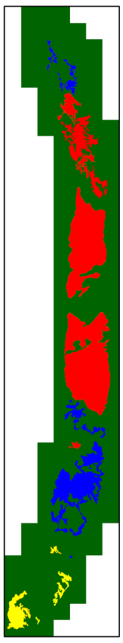
SD w/o Random Weights:

Primary	Secondary	Tertiary	Non-Cancer
33%	33%	0%	33%

Accuracy of **generated primary and secondary GGs** estimated by their #pixels in all sampled masks compared to their **pre-defined primary and secondary GGs**:

Method	Accuracy
SD w/ Random Weights	96.75%
SD w/o Random Weights	67.75%

(b)



(c)

Figure 3. Ablation Study on Random Weights. These weights are empirically established based on the typical label distribution in actual slides (e.g., depicted in (*c*)). Slides generated by Stable Diffusion (SD) [6] with and without the Random Weights can fulfill TransMIL's [7] objective of predicting two main Gleason Grades (GGs), regardless of their primary or secondary nature. However, in slides produced by SD without Random Weights, where label selection (except for Tertiary GG) occurs with equal probability, the label distribution significantly deviates from real-world norms, slightly reducing TransMIL's performance, as shown in (*a*). Moreover, as shown in (*b*), only 67.75% slides generated w/o Random Weights have primary and secondary GGs estimated from their annotated masks similar to their pre-defined primary and secondary GGs, resulting weaker identification of primary and secondary GGs. This issue also affects patient prognosis and treatment, especially when cancer grading models are trained on these inaccurately assigned primary and secondary GGs.

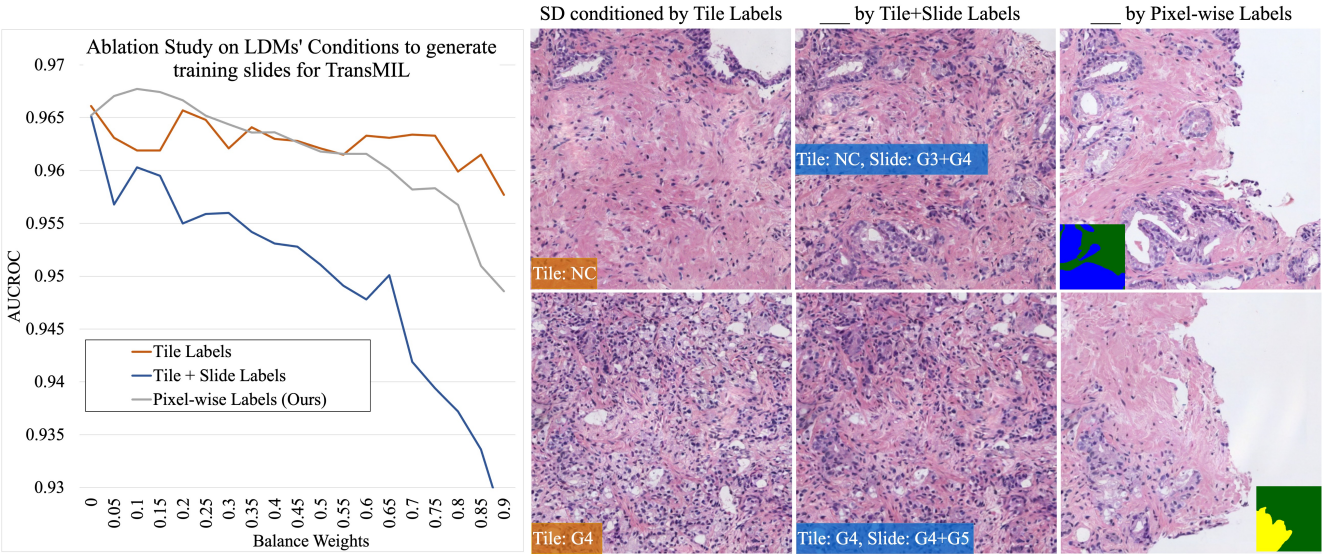


Figure 4. Additional Ablation Study investigating the impact of Latent Diffusion Models' (LDMs) conditions on enhancing TransMIL's performance (*left*) in AUCROC and qualitative evaluation (*right*). TransMIL trained on tiles generated with the unreasonable combination of Tile+Slide Labels provides the worst cancer grading performance compared to other conditioning types.

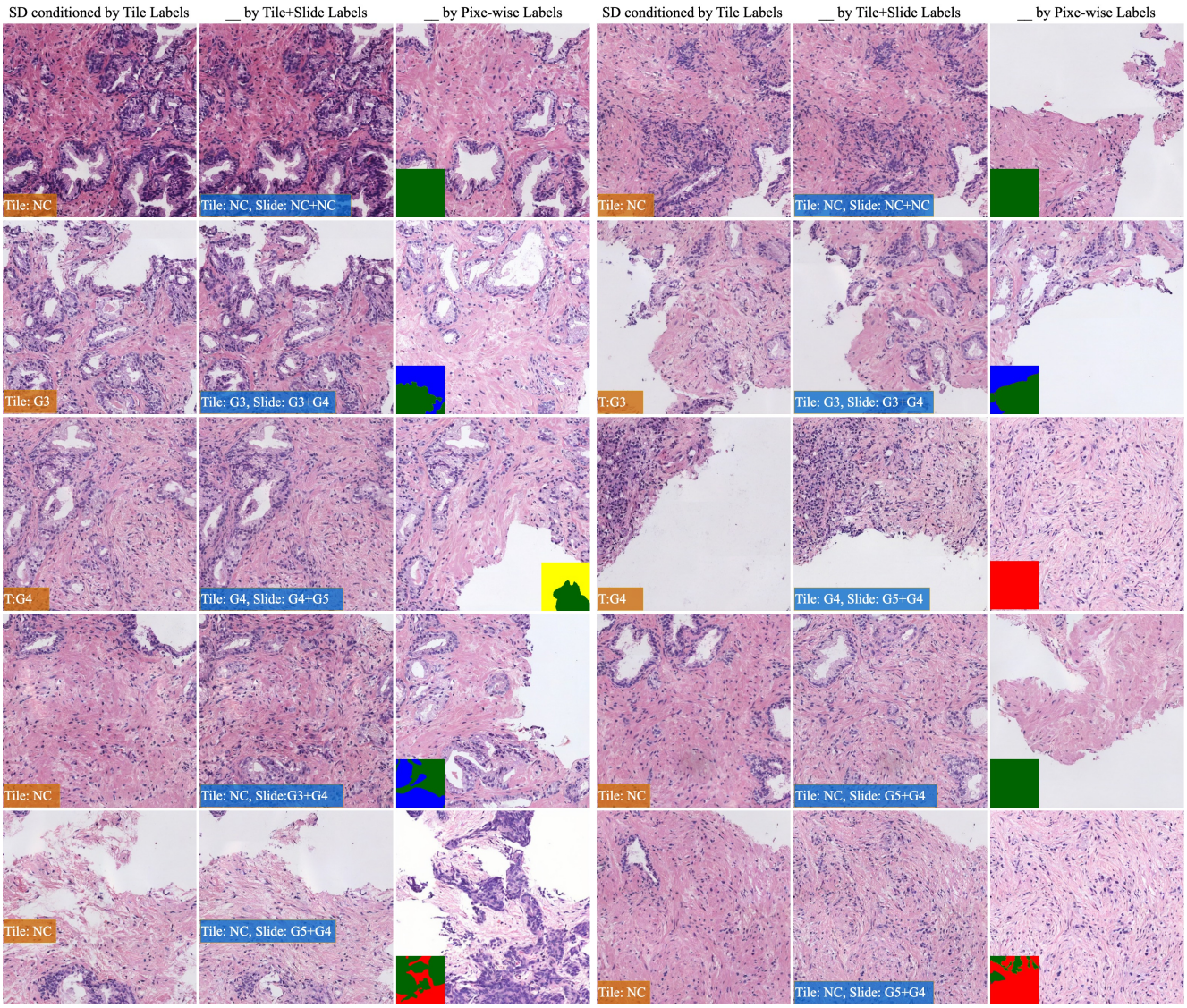


Figure 5. Tiles generated with three different LDMs' Conditions.

Method	In-distribution SICAPv2 Dataset				Out-of-distribution LAPC Dataset		
	GG3	GG4	GG5	Avg	Low-Grade (GG3)	High-Grade (GG4+GG5)	Avg
Carcino-Net [5]	0.7595 ± 0.054	0.8415 ± 0.027	0.738 ± 0.156	0.7796 ± 0.0652	0.6426 ± 0.2136	0.6895 ± 0.2508	0.6661 ± 0.0902
w/ SD	0.7557 ± 0.0518	0.8426 ± 0.0239	0.5901 ± 0.1851	0.7295 ± 0.0691	0.6271 ± 0.104	0.9456 ± 0.0441	**0.7863 ± 0.0547**
w/ SD-SC	0.7942 ± 0.0534	0.8211 ± 0.0438	0.7212 ± 0.351	0.7788 ± 0.1298	0.6212 ± 0.1982	0.4259 ± 0.5043	0.5236 ± 0.2583
w/ SD-DISC	0.7295 ± 0.1431	0.7767 ± 0.0745	0.9121 ± 0.1061	**0.8061 ± 0.0599**	0.728 ± 0.1755	0.1823 ± 0.3317	0.4551 ± 0.2179
w/ SD-DISC-CoTrain	0.7634 ± 0.0538	0.7793 ± 0.043	0.638 ± 0.3175	0.7269 ± 0.12	0.7429 ± 0.2421	0.7433 ± 0.3351	0.7431 ± 0.1452

Table 1. Quantitative comparison between Carcino-Net [5] and itself trained with our techniques on SICAPv2 [8] and LAPC [4] using pixel-level precision while ignoring Non-Cancer label.

Quantitative Comparison on Slide-Level Classification. We provide all class-wise quantitative comparisons between TransMIL [7], Mixed Supervision [1] - an extended version of TransMIL, and our ablation models based on TransMIL including TransMIL jointly trained with generated tiles from SD (TransMIL+SD), SD-SC (TransMIL+SD-SC), SD-DISC (TransMIL+SD-SC), and SD-DISC-CoTrain (TransMIL+SD-SC-CoTrain) on in-distribution SICAPv2 [8] in the top part of Figures 10, 11, 12, 13, and 14. To prove the generalization of our generated tiles, which consistently improves the slide-level prostate cancer grading performance of TransMIL, we also evaluate these models, which are trained on SICAPv2, on out-of-distribution PANDA [2] in the bottom part of Figures 10, 11, 12, 13, and 14.

References

[1] Hao Bian, Zhuchen Shao, Yang Chen, Yifeng Wang, Haoqian Wang, Jian Zhang, and Yongbing Zhang. Multiple instance learning with mixed supervision in gleason grading. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 204–213. Springer, 2022. 5

[2] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022. 5, 10, 11, 12, 13, 14

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 6

[4] Wenyuan Li, Jiayun Li, Karthik V Sarma, King Chung Ho, Shiwen Shen, Beatrice S Knudsen, Arkadiusz Gertych, and Corey W Arnold. Path r-cnn for prostate cancer diagnosis and gleason grading of histological images. *IEEE transactions on medical imaging*, 38(4):945–954, 2018. 4

[5] Avinash Lokhande, Saikiran Bonthu, and Nitin Singhal. Carcino-net: A deep learning framework for automated gleason grading of prostate biopsies. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1380–1383. IEEE, 2020. 4

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[7] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 3, 5

[8] Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195: 105637, 2020. 2, 4, 5, 10, 11, 12, 13, 14

[9] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 2, 6

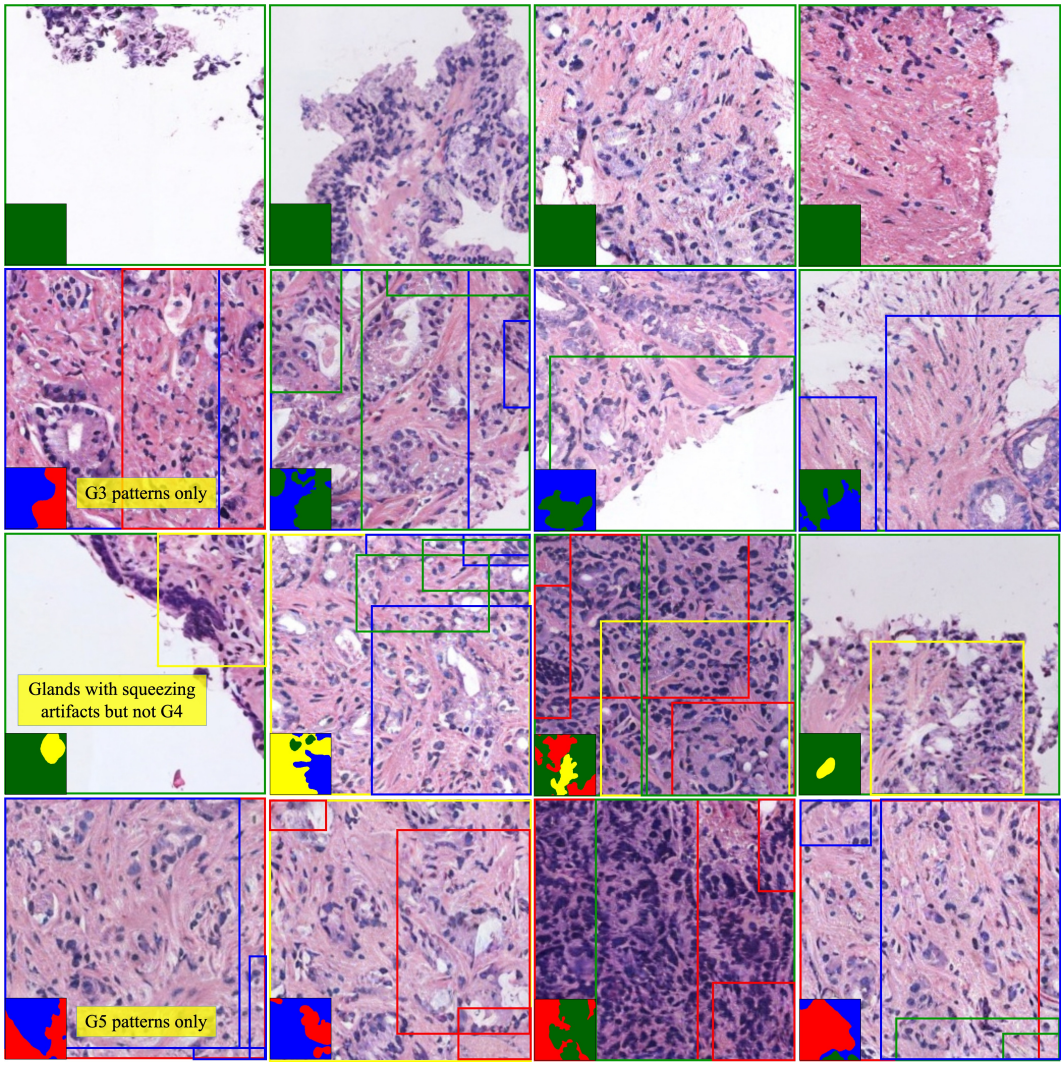


Figure 6. Layouts to Tiles with LayoutDiffusion [9]. We convert the sampled annotated masks to COCO-like layouts [3] by finding contours and bounding boxes for all pixel-wise labels to train LayoutDiffusion [9]. We observe that LayoutDiffusion [9] finds difficulty dealing with overlapped bounding boxes as they are weak labels and not specific enough to blend semantic information. Hence, it tends to ignore small bounding boxes and generate the patterns for the larger ones.

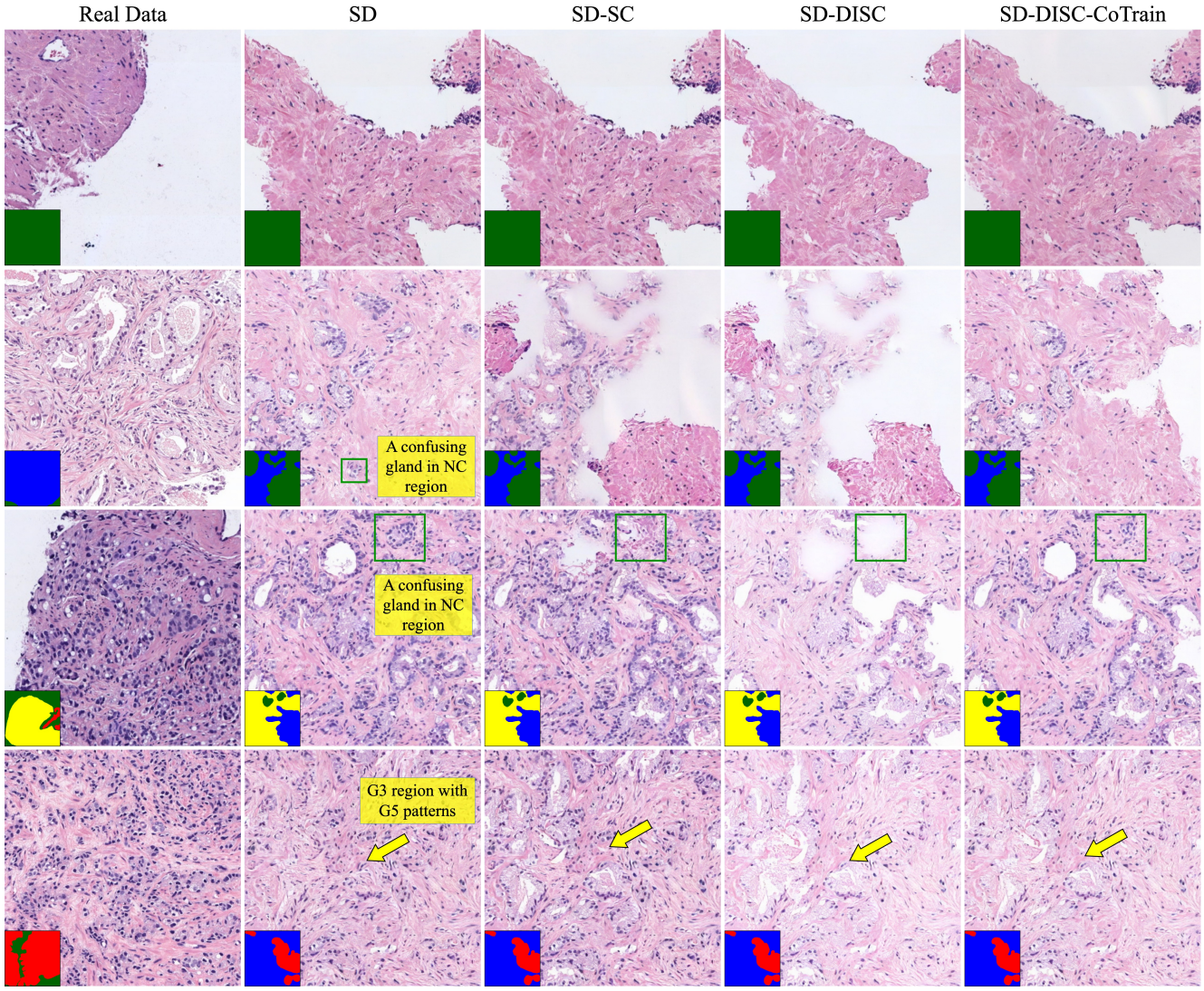


Figure 7. An additional qualitative comparison

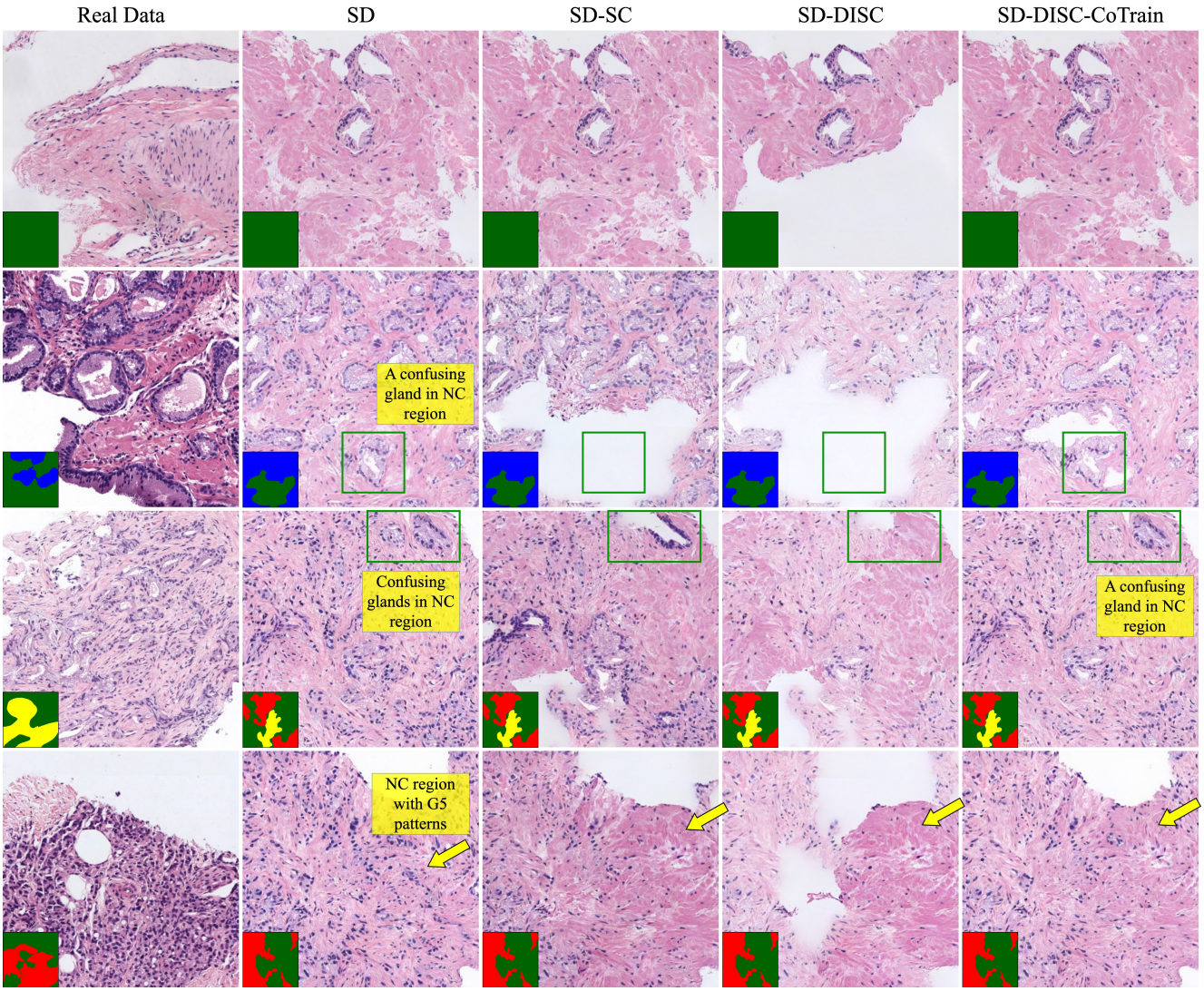


Figure 8. An additional qualitative comparison

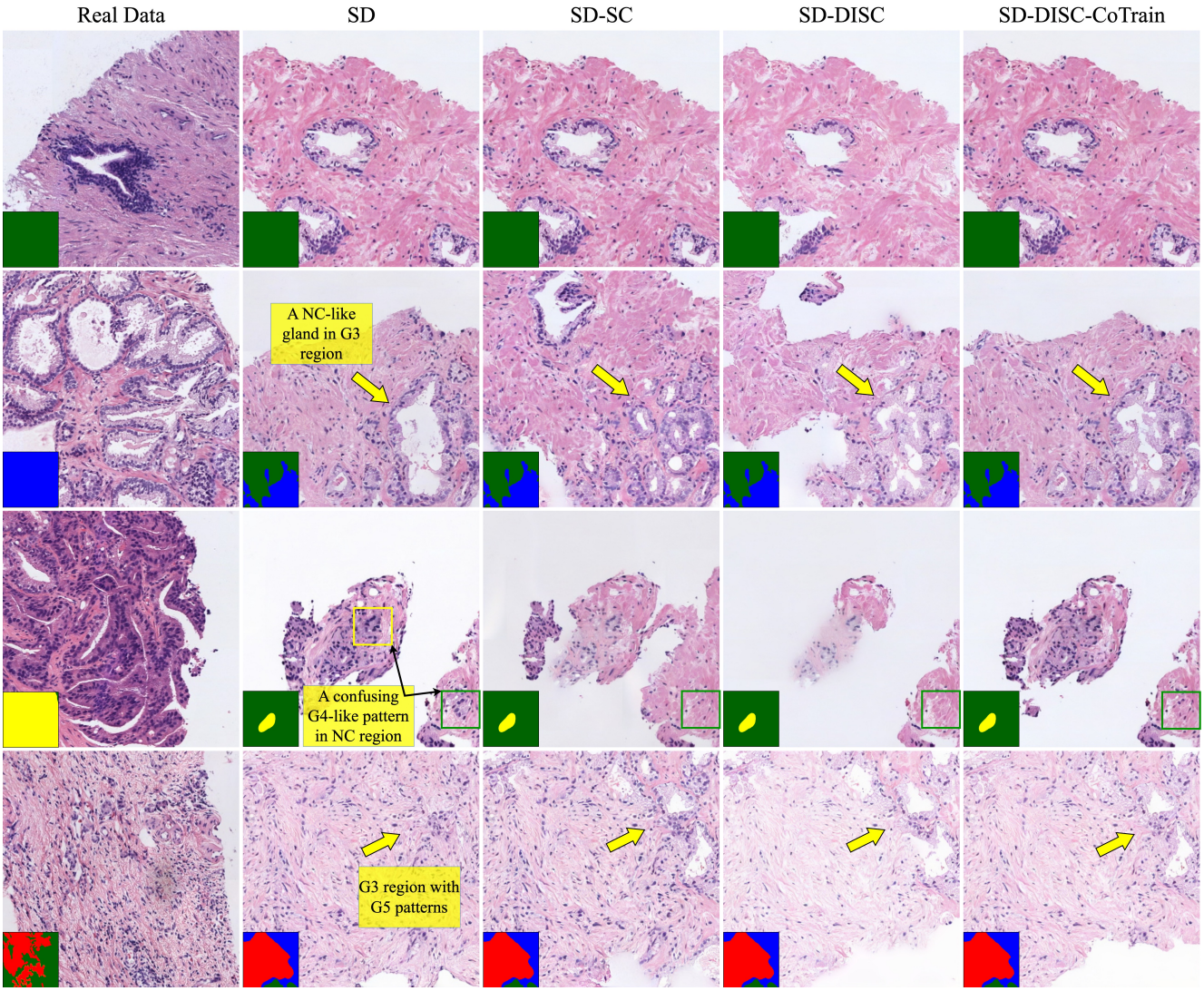


Figure 9. An additional qualitative comparison

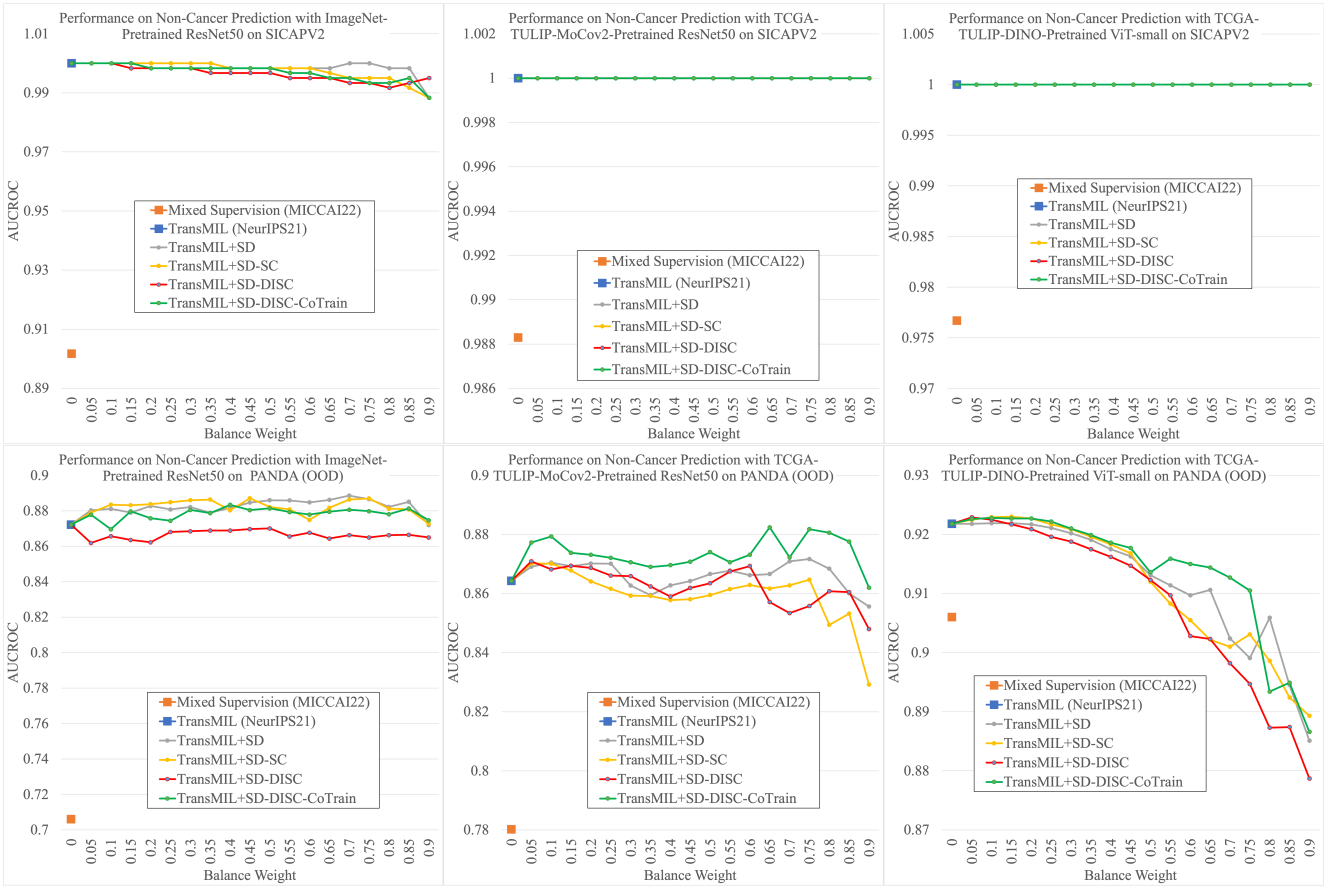


Figure 10. Slide-Level Prostate Cancer Grading Performance on Non-Cancer on in-distribution SICAPv2 [8] (top) and out-of-distribution PANDA [2] (bottom).

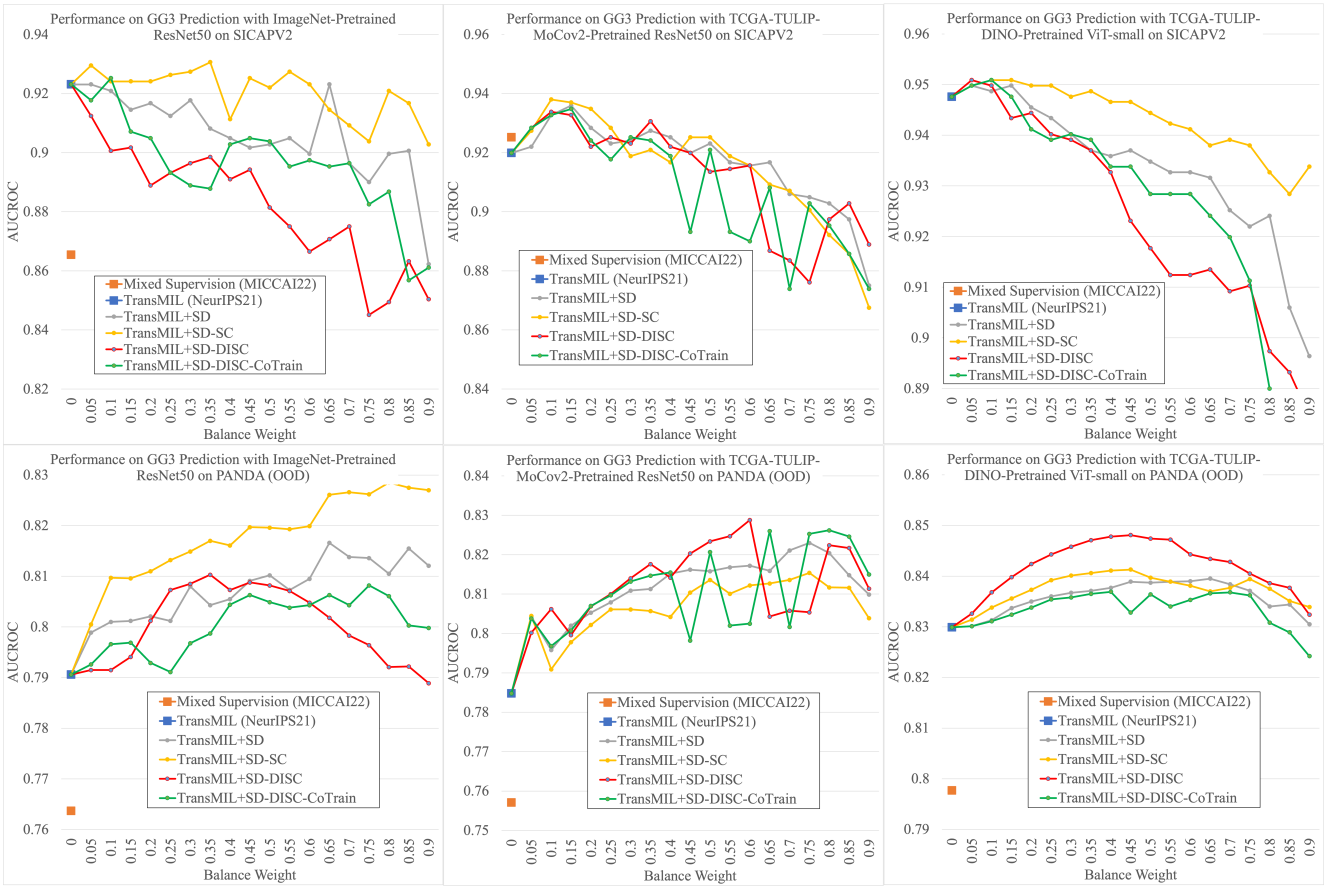


Figure 11. Slide-Level Prostate Cancer Grading Performance on GG3 on in-distribution SICAPv2 [8] (top) and out-of-distribution PANDA [2] (bottom).

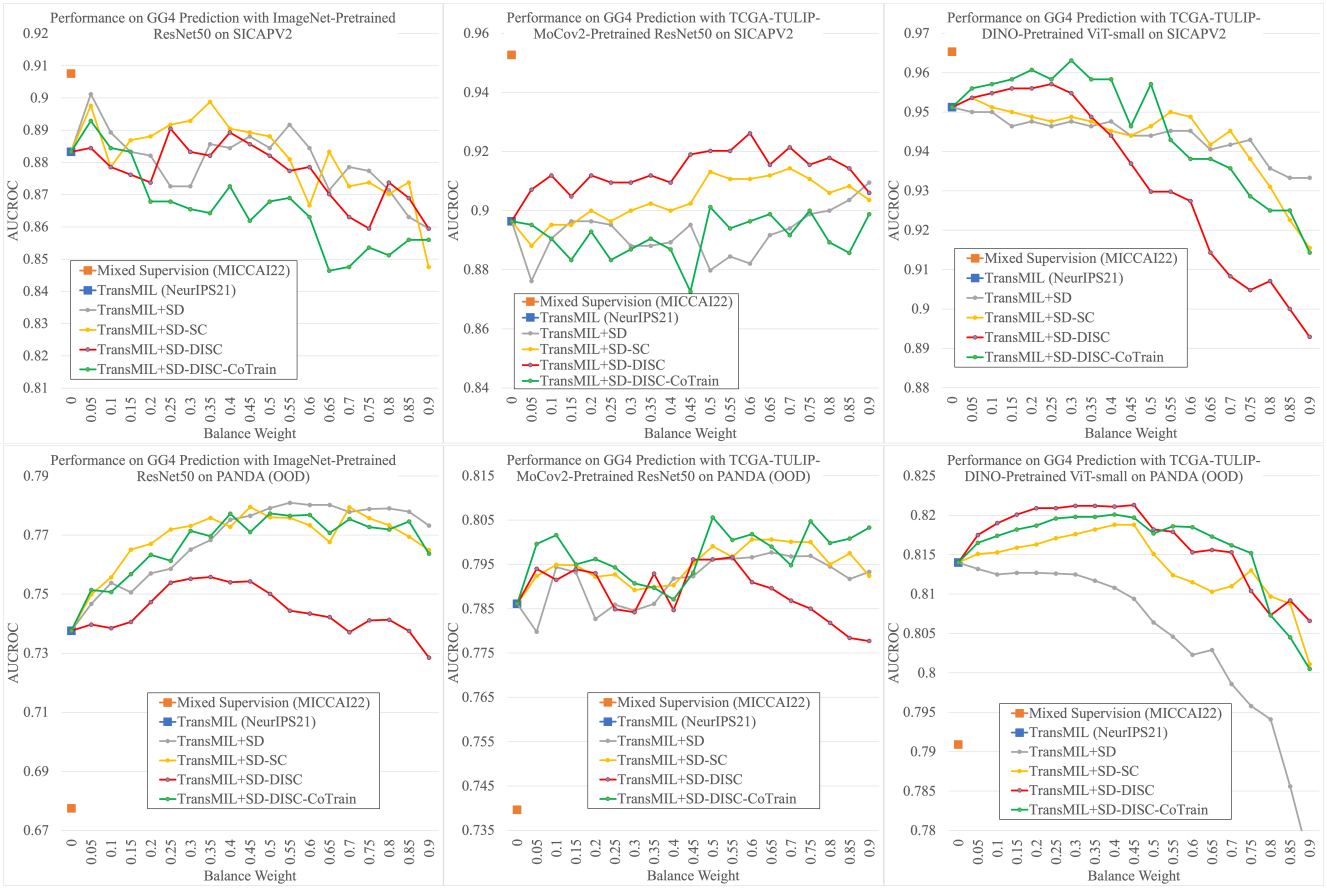


Figure 12. Slide-Level Prostate Cancer Grading Performance on GG4 on in-distribution SICAPv2 [8] (top) and out-of-distribution PANDA [2] (bottom).

Figure 13. Slide-Level Prostate Cancer Grading Performance on GG5 on in-distribution SICAPv2 [8] (top) and out-of-distribution PANDA [2] (bottom).

Figure 14. Slide-Level Prostate Cancer Grading Performance on All-Class on in-distribution SICAPv2 [8] (top) and out-of-distribution PANDA [2] (bottom).