Appendix

This appendix includes the following five parts: *(i)* We provide detailed descriptions of the datasets used in our experiments in Sec. A. *(ii)* The introduction and implementation details of our CoPS method and the SOTA comparison methods are presented in Sec. B. *(iii)* Additional experimental results, including comparative experiments, ablation studies, and further analysis, are reported in Sec. C. *(iv)* More detailed quantitative results are presented in Sec. D. *(v)* More detailed qualitative results are provided in Sec. E.

A. Detailed Dataset Descriptions

We evaluate the ZSAD performance of our proposed method on 13 publicly available datasets from industrial and medical domains. As shown in Tab. S1, we employ five industrial and eight medical datasets commonly used in prior studies.

MVTec-AD [2] is one of the most challenging datasets in the industrial domain. This dataset contains 15 high-resolution industrial product categories divided into texture and object groups, including over 70 types of defects. In this work, we only use its labeled test set, which contains 467 normal and 1258 anomalous samples.

VisA [47] is one of the largest datasets for industrial anomaly detection, including 10821 images across 12 categories of colored industrial parts. The dataset covers diverse appearance defects under varying illumination and background conditions. In this work, we only use its labeled test set, which contains 962 normal samples and 1200 anomalous samples.

BTAD [29] is a real-world industrial anomaly detection dataset composed of 3 high-resolution categories. It includes 451 normal and 290 anomalous samples with pixel-level annotations. Similar to MVTec-AD, BTAD captures both structural and surface-level defects in practical inspection scenarios.

MPDD [18] is a real-world industrial dataset focused on defect detection in metal parts. The dataset captures fine-grained structural anomalies commonly encountered in industrial manufacturing. In this work, we only use its labeled test set, which contains 176 normal samples and 282 anomalous samples.

DTD-Synthetic [1] is a synthetic industrial dataset containing 12 texture categories with 357 normal and 947 anomalous images. Despite being synthetically generated, it provides pixel-level anomaly annotations, enabling both image-level and pixel-level evaluation.

HeadCT [34] is a medical anomaly detection dataset comprising head CT scans across 1 anatomical category with 100 normal and 100 anomalous images. The dataset covers diverse pathological conditions and is widely used for evaluating anomaly detection methods in medical imaging. Since HeadCT offers only image-level labels and lacks pixel-level annotations, it is primarily used for image-level evaluation.

BrainMRI [21] is a medical anomaly detection dataset consisting of brain MRI scans within a single anatomical class, containing 98 normal and 155 anomalous images. The collection spans varied neuropathologies and serves as a standard benchmark for assessing medical anomaly detection methods. As BrainMRI provides only image-level labels without pixel-level annotations, it is primarily used for image-level evaluation.

Br35H [14] is a medical anomaly detection dataset of brain MRI scans within a single anatomical class, comprising 1500 normal and 1500 anomalous images. The images encompass a variety of brain pathologies and are widely adopted for benchmarking medical anomaly detection methods. Also, Br35H provides only image-level labels without pixel-level annotations, so it is primarily used for image-level evaluation.

ISIC [10] is a medical anomaly detection dataset of dermoscopic skin images within a single anatomical class, comprising 379 anomalous images and no normal images. Each image is supplied with pixel-level lesion masks, making the dataset a benchmark for evaluating pixel-level anomaly segmentation rather than image-level anomaly classification.

CVC-ColonDB [35] is a colonoscopy anomaly dataset containing 380 anomalous images and no normal images. Every image is annotated with a pixel-level polyp mask, establishing the dataset as a standard benchmark for evaluating pixel-level anomaly segmentation rather than image-level anomaly classification.

CVC-ClinicDB [3] is a colonoscopy anomaly dataset containing 612 anomalous images and no normal images, similar to CVC-ColonDB. Each image includes a pixel-level polyp mask, making the dataset a standard benchmark for evaluating pixel-level anomaly segmentation rather than image-level anomaly classification.

Kvasir [19] is a colonoscopy anomaly dataset containing 1000 anomalous images and no normal images, similar to CVC-ColonDB. Every image is annotated with a pixel-level polyp mask, establishing the dataset as a standard benchmark for evaluating pixel-level anomaly segmentation rather than image-level anomaly classification.

Endo [15] is a colonoscopy anomaly dataset containing 200 anomalous images and no normal images, similar to CVC-ColonDB. Every image is annotated with a pixel-level polyp mask, establishing the dataset as a standard benchmark for evaluating pixel-level anomaly segmentation rather than image-level anomaly classification.

B. Additional Implementation Details

This section provides additional implementation details of our method CoPS, as well as descriptions and reproduc-

Table S1. Key statistics of the 13 industrial and medical datasets used in our experiments. '#' denotes the number of instances.

Domain	Dataset	#Class	#Normal Image	#Anomaly Image	Data Type	Pixel Label	Real-world
Industrial	MVTec-AD	15	467	1258	object & texture	✔	✔
	VisA	12	962	1200	object	✔	✔
	BTAD	3	451	290	object & texture	✔	✔
	MPDD	6	176	282	object	✔	✔
	DTD-Synthetic	12	357	947	texture	✔	×
Medical	HeadCT	1	100	100	brain	×	✔
	BrainMRI	1	98	155	brain	×	✔
	Br35H	1	1500	1500	brain	×	✔
	ISIC	1	0	379	skin	✔	✔
	CVC-ColonDB	1	0	380	colon	✔	✔
	CVC-ClinicDB	1	0	612	colon	✔	✔
	Kvasir	1	0	1000	colon	✔	✔
	Endo	1	0	200	colon	✔	✔

tion settings of other SOTA comparison methods. For a fair comparison, all methods are evaluated under the same CLIP backbone, input resolution, and evaluation protocol.

CoPS is our proposed method, which dynamically synthesizes visually conditioned prompts to adapt CLIP for zero-shot anomaly detection, achieving SOTA performance. Following previous works [4, 45, 46], we adopt the publicly available CLIP (ViT-L/14@336px) pre-trained by OpenAI [32]. Input images are resized to 518×518. Visual and textual embeddings are extracted from the final layers of the vision and text encoders, with a dimensionality of $C = 768$. In the 2nd to 9th layers of the text encoder, eight sets of four learnable tokens replace the input prefix to refine the textual representation. All layers of the vision encoder employ both Q-KV and V-VV branches in parallel. For ESTS, the context token length K, state token length M, and class token length N are set to 6, 6, and 2, respectively. The prototype extractor $\mathcal{P}_\theta$ is configured with 12 attention heads and a two-layer feed-forward network whose hidden layer has the same dimensionality as the input. For ICTS, the sampling count R is set to 10. The VAE employs two-layer MLPs for both the encoder $q_{\psi'}$ and decoder $p_{\psi''}$, with hidden layers matching the input dimensionality. For SAGA, the distance coefficient α and glocal coefficient β are set to 0.3 and 0.9, respectively. The default temperature hyperparameter τ is 0.07. CoPS is trained using the Adam optimizer for 10 epochs with an initial learning rate of 0.001 and a batch size of 8. During inference, a Gaussian filter with $\sigma = 4$ is applied to smooth the anomaly map. The results are reported with the random seed fixed to 0 for reproducibility. All experiments are conducted on a system equipped with a single NVIDIA GeForce RTX 3090 GPU and an Intel Xeon Gold 6226R CPU.

WinCLIP [17] is the first work to employ frozen CLIP for ZSAD. It leverages window-based patch sampling and computes text-image similarity at the region level to localize anomalies. Anomaly scores are derived by aggregating the dissimilarity between visual patches and the textual description of normality. This method does not require additional training data or fine-tuning, making it a training-free solution for ZSAD. As the official implementation of WinCLIP is unavailable, we adopt the reproduced code from [45].

APRIL-GAN [8] builds on a frozen CLIP backbone and adds lightweight trainable linear layers to project patch features into the shared image-text space for finer alignment with compositional prompts. It further maintains class-specific memory banks of normal references whose features are contrasted with test features to refine anomaly maps during inference. These designs allow APRIL-GAN to perform zero-/few-shot anomaly classification and segmentation without task-specific retraining. As APRIL-GAN adopts the same backbone (ViT-L/14@336px) and input resolution (518×518) as ours, we evaluate it directly using the official implementation and pre-trained weights.

CLIP-AD [9] builds on APRIL-GAN's lightweight linear adapters and further integrates representative vector selection and multi-scale feature fusion to produce both image-/pixel-level anomaly scores. Since CLIP-AD originally uses an uncommon backbone (ViT-B/16@240px) and input resolution (240×240), we retrain it with the official code under our backbone (ViT-L/14@336px) and input resolution (518×518).

AdaCLIP [4] adapts CLIP for ZSAD by jointly fine-tuning both vision and text encoders while learning hybrid prompts that combine globally optimized static tokens with per-image dynamic tokens. The hybrid prompts guide the dual encoders to disentangle normal and abnormal semantics. Since AdaCLIP adopts a non-standard evaluation protocol with multiple auxiliary datasets, we retrain it with the official code under the same single-auxiliary setting as ours.

AnomalyCLIP [45] builds on CoOp [44] by learning context tokens for dual prompts that represent "normal" and "anomalous" states with frozen vision encoder and trainable text encoder. Additionally, it employs consistent self-

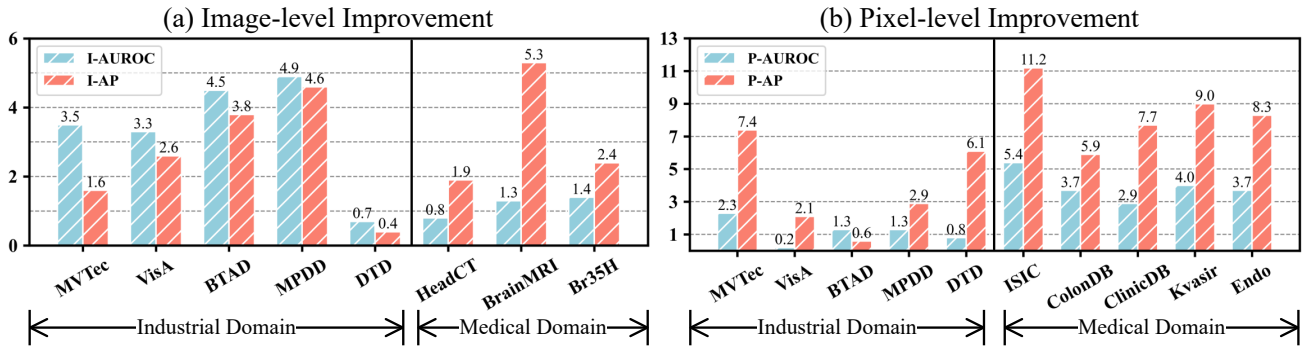

Figure S1. Performance improvements of CoPS over the baseline AnomalyCLIP across all 13 industrial and medical datasets.

attention (i.e., V-VV) across the visual encoder layers to emphasize diagonally prominent local features and improve fine-grained anomaly localization. AnomalyCLIP serves as the baseline for our proposed method. As AnomalyCLIP adopts the same backbone (ViT-L/14@336px) and input resolution (518×518) as ours, we evaluate it directly using the official implementation and pre-trained weights.

FAPrompt [46] introduces a fine-grained prompt tuning method for ZSAD which is built upon CoCoOp [43]. It synthesizes compound abnormality prompts by leveraging the mapped top-K abnormality-prior patch features. This approach allows for more flexible and adaptive prompt design, improving the model's ability to capture diverse anomalies. Although FAPrompt provides official code, no pre-trained weights are available. Thus, we retrain it with the official implementation under the same evaluation protocol.

C. Extended Experimental Analysis

We provide additional comparative and ablation experiments to further validate the effectiveness of CoPS.

Performance improvement. Fig. S1 illustrates the relative improvements of CoPS over the prompt-tuning baseline, AnomalyCLIP. CoPS achieves consistent gains over AnomalyCLIP on all 13 datasets, particularly showing larger improvements in image-level performance (Fig. S1(a)) for industrial datasets and in pixel-level performance (Fig. S1(b)) for medical datasets. These results demonstrate the effectiveness of CoPS's prototype extraction, class sampling, and glocal alignment components.

Computational efficiency. Tab. S2 compares the accuracy and computational efficiency of various methods in terms of image-level and pixel-level performance, model size, memory consumption, and inference speed. CoPS achieves the best performance across all four evaluation metrics (I-AUROC, I-AP, P-AUROC, and P-AP), outperforming all prior methods. In terms of efficiency, CoPS maintains a competitive model size (19 MB) and moderate memory usage (7.1 GB for training and 2.7 GB for testing),

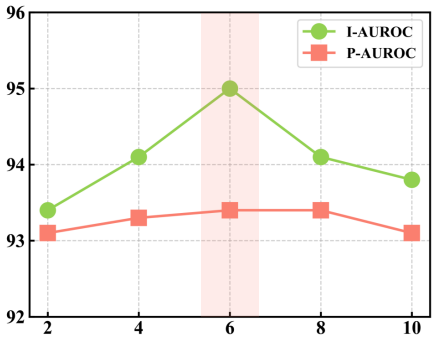

Figure S2. Performance ablation of context token length K.

while achieving an inference speed of 168 ms per frame. Although WinCLIP requires no fine-tuning, it suffers from high inference latency. APRIL-GAN and CLIP-AD offer relatively efficient inference, but their accuracy remains notably limited. Furthermore, CoPS outperforms AdaCLIP in both accuracy and computational efficiency. While its inference time is marginally higher than that of AnomalyCLIP, CoPS yields significantly improved results on both image-level and pixel-level metrics. Finally, CoPS also outperforms the SOTA method FAPrompt, achieving consistent improvements across all evaluation metrics. These results demonstrate that CoPS offers a favorable balance between accuracy and efficiency, making it a practical and scalable solution for real-world ZSAD tasks.

Influence of pre-trained backbone. As shown in Tab. S3, we analyze the impact of different pre-trained CLIP backbones and input image resolutions on model performance. The results indicate a consistent performance improvement with larger backbones and higher input resolutions. However, when the resolution increases to 700×700, performance declines due to semantic misalignment caused by overly small patch sizes. The best performance is achieved using ViT-L/14@336px with an input size of 518×518, which is also adopted as the default configuration in our

Table S2. Comparison of accuracy and efficiency across various methods on MVTec-AD dataset. The best results are highlighted in bold.

Metric → Method ↓	I-AUROC (%)	I-AP (%)	P-AUROC (%)	P-AP (%)	Model Size (MB)	Train Mem. (GB)	Test Mem. (GB)	1/FPS (ms)
WinCLIP	91.8	95.1	85.1	18.0	-	-	**2.0**	840
APRIL-GAN	86.1	93.5	87.6	40.8	12	**5.5**	3.3	**105**
CLIP-AD	89.8	95.3	89.8	40.0	**8.7**	6.7	3.4	115
AdaCLIP	92.0	96.4	86.8	38.1	41	10	3.3	183
AnomalyCLIP	91.5	96.2	91.1	34.5	22	6.9	2.7	131
FAPrompt	91.9	95.9	90.4	34.6	39	11	2.7	174
CoPS	**95.0**	**97.8**	**93.4**	**41.9**	19	7.1	2.7	168

Table S3. Performance ablation of different backbones and input image sizes, as measured by I-AUROC%, I-AP%, P-AUROC%, and P-AP%. The best results are highlighted in bold, and the second-best results are underlined.

Backbone	Image Size	I-AUROC	I-AP	P-AUROC	P-AP
ViT-B/16@224px	336×336	85.4	93.4	88.6	31.2
ViT-L/14@224px	336×336	89.9	95.1	91.5	34.8
ViT-L/14@224px	518×518	90.0	95.4	91.8	40.6
ViT-L/14@336px	336×336	92.2	96.4	<u>92.6</u>	37.1
ViT-L/14@336px	518×518	**95.0**	**97.8**	**93.4**	**41.9**
ViT-L/14@336px	700×700	<u>92.4</u>	<u>97.2</u>	92.5	<u>41.8</u>

experiments. These settings are widely adopted in most existing methods [4, 8, 45].

Influence of context token length. As illustrated in Fig. S2, increasing the context token length K initially improves both I-AUROC and P-AUROC, reaching peak performance at $K = 6$. Further increasing K beyond this point leads to performance degradation, likely due to overfitting in the prompt space. Therefore, we adopt $K = 6$ as the default setting in all experiments.

D. More Quantitative Results

As the industrial datasets contain multiple categories, we report the detailed performance of CoPS across all categories and further compare it with all SOTA methods on MVTec-AD and VisA. Specifically, Tabs. S7 and S8 present the image-level and pixel-level performance of SOTA methods on MVTec-AD across four metrics for each category. These metrics include image-level AUROC (I-AUROC) and image-level AP (I-AP) for classification, and pixel-level AUROC (P-AUROC) and pixel-level AP (P-AP) for segmentation. Tabs. S9 and S10 report the image-level and pixel-level performance of SOTA methods on VisA, also evaluated using four metrics per category. Tabs. S4–S6 show the per-category results of CoPS on DTD-Synthetic, BTAD, and MPDD, respectively.

E. More Qualitative Results

In this section, we present the visualization results of CoPS across all categories on 13 industrial and medical datasets. Specifically, Figs S3–S7 show the visualization results of CoPS on all 15 categories of the MVTec-AD dataset. Figs S8–S11 present results on all 12 categories of the VisA dataset. Fig. S12 displays the visualizations for all 3 categories in BTAD. Figs S13–S14 illustrate results on all 6 categories of MPDD. Figs S15–S18 cover all 12 categories of DTD-Synthetic. Figs. S19–S26 show comprehensive visualizations for all categories included in the HeadCT, BrainMRI, Br35H, ISIC, CVC-ColonDB, CVC-ClinicDB, Kvasir, and Endo datasets. It is important to note that HeadCT, BrainMRI, and Br35H (Figs. S19–S21) do not provide pixel-level annotations. As a result, only the input images can be displayed, and no ground-truth contours are available for qualitative comparison.

Table S4. Performance of CoPS on each category of the DTD-Synthetic dataset under the ZSAD setting.

Category	I-AUROC	I-AP	P-AUROC	P-AP
Woven_001	100	100	99.5	63.3
Woven_127	92.5	94.2	96.0	42.9
Woven_104	99.1	99.8	97.5	60.4
Stratified_154	98.9	99.8	99.7	79.4
Blotchy_099	99.2	99.8	99.2	66.8
Woven_068	96.4	98.0	98.3	37.6
Woven_125	100	100	99.4	66.8
Marbled_078	98.6	99.7	99.1	61.1
Perforated_037	93.1	98.3	97.1	55.9
Mesh_114	85.0	93.9	97.5	49.9
Fibrous_183	99.4	99.9	99.2	67.1
Matted_069	80.3	94.4	98.5	50.8
Average	95.2	98.1	98.4	58.5

Table S5. Performance of CoPS on each category of the BTAD dataset under the ZSAD setting.

Category	I-AUROC	I-AP	P-AUROC	P-AP
01	95.8	98.4	93.7	46.5
02	87.3	97.9	95.2	61.7
03	97.7	88.2	95.0	19.5
Average	93.6	94.9	94.6	42.6

Table S6. Performance of CoPS on each category of the MPDD dataset under the ZSAD setting.

Category	I-AUROC	I-AP	P-AUROC	P-AP
bracket_black	63.4	75.1	97.3	6.70
bracket_brown	60.6	76.4	95.8	12.4
bracket_white	77.9	66.9	99.0	6.92
connector	80.0	72.1	97.1	21.1
metal_plate	92.7	97.4	98.3	87.6
tubes	97.2	98.7	97.6	50.9
Average	78.6	81.1	97.5	30.9

Table S7. Performance comparison of various SOTA methods on each category of the MVTec-AD dataset under ZSAD setting, as measured by I-AUROC% / I-AP%. The best results are highlighted in bold, and the second-best results are underlined.

Method →	Prompt Design				Prompt Learning		
Category ↓	WinCLIP	APRIL-GAN	CLIP-AD	AdaCLIP	AnomalyCLIP	FAPrompt	**CoPS**
bottle	**99.2** / 98.3	92.0 / 97.7	96.4 / **98.8**	95.6 / 98.6	88.7 / 96.8	89.3 / 96.4	92.5 / 97.8
cable	86.5 / 86.2	88.2 / 92.9	80.4 / 88.9	79.0 / 87.3	70.3 / 81.7	73.3 / 81.8	**89.6 / 94.1**
capsule	72.9 / 93.4	79.8 / 95.4	82.8 / 96.4	89.3 / 97.8	89.5 / 97.8	93.0 / 98.6	**95.5 / 99.1**
carpet	**100** / 99.9	99.4 / 99.8	99.5 / 99.8	**100 / 100**	99.9 / 99.9	**100 / 100**	**100 / 100**
grid	98.8 / **99.8**	86.2 / 94.9	94.1 / 97.9	99.2 / 99.7	97.8 / 99.3	98.2 / 99.3	**99.3 / 99.8**
hazelnut	93.9 / 96.3	89.4 / 94.6	98.0 / 99.0	95.5 / 97.5	97.2 / 98.5	96.8 / 98.2	**98.8 / 99.3**
leather	**100 / 100**	99.7 / 99.9	**100 / 100**	**100 / 100**	99.8 / 99.9	99.9 / **100**	**100 / 100**
metal_nut	**97.1** / 97.9	68.2 / 91.8	75.1 / 94.4	79.9 / 95.6	92.4 / **98.1**	87.6 / 97.1	89.6 / 97.5
pill	79.1 / 96.5	80.8 / 96.1	87.7 / 97.6	**92.6 / 98.6**	81.1 / 95.3	89.4 / 97.9	92.0 / 98.2
screw	83.3 / 88.4	85.1 / 93.6	**89.1 / 96.2**	83.9 / 93.0	82.1 / 92.9	86.9 / 94.7	84.2 / 94.2
tile	**100** / 99.9	99.8 / 99.9	99.6 / 99.8	99.7 / 99.9	**100 / 100**	99.8 / 99.9	99.9 / **100**
toothbrush	87.5 / 96.7	53.2 / 71.9	76.1 / 90.2	95.2 / 97.9	85.3 / 93.9	83.9 / 92.7	**96.1 / 98.8**
transistor	88.0 / 74.9	80.9 / 77.6	79.3 / 73.7	82.0 / 83.8	**93.9 / 92.1**	85.1 / 83.1	90.9 / 88.9
wood	**99.4** / 98.8	98.9 / **99.6**	98.9 / **99.6**	98.5 / 99.5	96.9 / 99.2	97.8 / 99.4	98.2 / 99.5
zipper	91.5 / 98.9	89.4 / 97.1	88.6 / 96.9	89.4 / 97.1	**98.4 / 99.5**	97.4 / 99.3	97.8 / 99.4
Average	91.8 / 95.1	86.1 / 93.5	89.8 / 95.3	92.0 / 96.4	91.5 / 96.2	91.9 / 95.9	**95.0 / 97.8**

Table S8. Performance comparison of various SOTA methods on each category of the MVTec-AD dataset under ZSAD setting, as measured by P-AUROC% / P-AP%. The best results are highlighted in bold, and the second-best results are underlined.

Method →	Prompt Design				Prompt Learning		
Category ↓	WinCLIP	APRIL-GAN	CLIP-AD	AdaCLIP	AnomalyCLIP	FAPrompt	**CoPS**
bottle	89.5 / 49.8	83.5 / 53.0	91.2 / 56.8	83.8 / 49.8	90.4 / 55.3	90.5 / 57.4	**92.8 / 61.8**
cable	77.0 / 6.20	72.2 / **18.2**	76.2 / 17.3	**85.6** / 16.5	78.9 / 12.3	78.0 / 11.9	79.0 / 15.1
capsule	86.9 / 8.60	92.0 / 29.6	95.1 / 27.2	86.2 / 24.8	95.8 / 27.7	95.4 / 26.0	**97.4 / 30.6**
carpet	95.4 / 25.9	98.4 / 67.5	99.1 / 65.4	94.8 / 63.5	98.8 / 56.6	99.0 / 60.7	**99.3 / 74.5**
grid	82.2 / 5.70	95.8 / **36.5**	96.3 / 30.7	90.6 / 27.8	97.3 / 24.1	97.2 / 23.9	**97.8** / 26.9
hazelnut	94.3 / 33.3	96.1 / 49.7	97.2 / 59.2	**98.7 / 69.5**	97.2 / 43.4	97.5 / 45.9	97.4 / 51.1
leather	96.7 / 20.4	99.1 / 52.3	**99.3** / 50.5	97.8 / **53.6**	98.6 / 22.7	98.6 / 25.1	99.2 / 36.0
metal_nut	61.0 / 10.8	65.5 / 25.9	58.9 / 21.2	55.4 / 19.9	74.6 / 26.4	68.0 / 23.3	**88.5 / 39.0**
pill	80.0 / 7.00	76.2 / 23.6	83.7 / 26.1	77.5 / 25.8	91.8 / 34.1	90.2 / 31.2	**92.3 / 35.0**
screw	89.6 / 5.40	97.8 / 33.7	98.7 / 39.1	**99.2 / 41.6**	97.5 / 27.5	97.3 / 24.7	98.2 / 22.4
tile	77.6 / 21.2	92.7 / 66.3	94.5 / 65.2	83.9 / 48.8	94.7 / 61.7	96.2 / 64.6	**97.9 / 79.2**
toothbrush	86.9 / 5.50	**95.8 / 43.2**	92.7 / 29.9	93.4 / 24.7	91.9 / 19.3	89.8 / 15.9	94.9 / 25.0
transistor	74.7 / **20.2**	62.4 / 11.7	**75.5** / 14.2	71.4 / 11.9	70.8 / 15.6	70.8 / 16.1	73.6 / 17.0
wood	93.4 / 32.9	95.8 / 61.8	96.9 / 59.4	91.2 / 56.6	96.4 / 52.6	96.9 / 55.2	**97.5 / 68.4**
zipper	91.6 / 19.4	91.1 / 38.7	92.8 / 38.5	91.8 / 36.0	91.2 / 38.7	90.7 / 37.7	**95.2 / 46.3**
Average	85.1 / 18.0	87.6 / 40.8	89.8 / 40.0	86.8 / 38.1	91.1 / 34.5	90.4 / 34.6	**93.4 / 41.9**

Table S9. Performance comparison of various SOTA methods on each category of the VisA dataset under ZSAD setting, as measured by I-AUROC% / I-AP%. The best results are highlighted in bold, and the second-best results are underlined.

Method →	Prompt Design				Prompt Learning		
Category ↓	WinCLIP	APRIL-GAN	CLIP-AD	AdaCLIP	AnomalyCLIP	FAPrompt	**CoPS**
candle	95.4 / 95.6	82.5 / 85.9	89.4 / 91.6	**95.9 / 96.4**	80.9 / 82.6	84.5 / 87.1	87.8 / 91.0
capsules	85.0 / 80.9	62.3 / 74.6	75.2 / 86.6	81.1 / 86.7	82.7 / 89.4	**91.1 / 95.8**	88.9 / 93.4
cashew	**92.1** / 95.2	86.7 / 93.9	83.7 / 92.4	89.6 / **95.4**	76.0 / 89.3	89.2 / **95.4**	87.1 / 94.6
chewinggum	96.5 / 98.8	96.5 / 98.4	95.6 / 98.1	**98.5 / 99.4**	97.2 / 98.8	97.4 / 99.0	98.1 / 99.2
fryum	80.3 / 92.5	93.8 / 97.0	78.7 / 90.4	89.5 / 95.1	92.7 / 96.6	**96.0 / 98.2**	93.8 / 97.4
macaroni1	76.2 / 64.5	69.5 / 67.5	80.0 / 81.1	86.3 / 85.0	**86.7** / 85.5	81.2 / 81.5	84.1 / **85.7**
macaroni2	63.7 / 65.2	65.7 / 64.9	67.0 / 65.3	56.7 / 54.3	72.2 / 70.8	**72.5 / 72.6**	70.5 / 69.3
pcb1	73.6 / 74.6	50.6 / 54.6	68.6 / 72.5	74.0 / 73.5	85.2 / 86.7	69.4 / 74.1	**86.6 / 89.1**
pcb2	51.2 / 44.2	**71.6 / 73.8**	69.7 / 71.4	71.1 / 71.6	62.0 / 64.4	66.2 / 67.8	67.1 / 69.1
pcb3	73.4 / 66.2	66.9 / 70.5	67.3 / 71.9	**75.2 / 77.9**	61.7 / 69.4	68.2 / 75.5	66.4 / 71.3
pcb4	79.6 / 70.1	94.6 / 94.8	96.2 / 96.0	89.6 / 89.8	93.9 / 94.3	95.3 / 95.1	**97.7 / 97.3**
pipe_fryum	69.7 / 82.1	89.4 / 94.6	86.5 / 93.7	88.8 / 93.9	92.3 / 96.3	**97.2** / 98.6	97.1 / **98.7**
Average	78.1 / 77.5	78.0 / 81.4	79.8 / 84.3	83.0 / 84.9	82.1 / 85.4	84.0 / 86.7	**85.4 / 88.0**

Table S10. Performance comparison of various SOTA methods on each category of the VisA dataset under ZSAD setting, as measured by P-AUROC% / P-AP%. The best results are highlighted in bold, and the second-best results are underlined.

Method →	Prompt Design				Prompt Learning		
Category ↓	WinCLIP	APRIL-GAN	CLIP-AD	AdaCLIP	AnomalyCLIP	FAPrompt	**CoPS**
candle	88.9 / 2.40	97.8 / 29.9	98.7 / 36.6	98.6 / **45.3**	**98.8** / 25.6	**98.8** / 25.4	98.2 / 25.9
capsules	81.6 / 1.40	**97.5 / 40.0**	97.4 / 38.5	96.1 / 18.2	94.9 / 29.3	96.4 / 30.9	95.6 / 31.3
cashew	84.7 / 4.80	86.0 / 15.1	91.4 / 24.1	**97.2 / 44.8**	93.7 / 19.6	93.8 / 17.6	95.5 / 25.1
chewinggum	93.3 / 24.0	**99.5** / 83.6	99.2 / 83.4	99.2 / **87.6**	99.2 / 56.3	99.4 / 61.3	**99.5** / 65.5
fryum	88.5 / 11.1	92.0 / 22.1	93.0 / 22.4	93.6 / 24.0	94.6 / 22.6	94.1 / 21.4	**94.7 / 26.1**
macaroni1	70.9 / 0.03	**98.8** / 24.8	98.7 / 23.2	**98.8 / 27.1**	98.3 / 14.9	98.1 / 12.9	98.5 / 12.8
macaroni2	59.3 / 0.02	97.8 / **6.80**	97.6 / 2.30	**98.2** / 3.00	97.6 / 1.50	96.5 / 0.88	96.7 / 1.69
pcb1	61.2 / 0.40	92.7 / 8.40	92.6 / 7.20	90.7 / 7.80	94.0 / 8.60	**95.6** / 9.72	93.7 / **9.76**
pcb2	71.6 / 0.40	89.8 / 15.4	91.0 / 8.20	91.3 / **17.5**	92.4 / 9.10	92.2 / 8.62	**92.7** / 8.18
pcb3	85.3 / 0.70	88.4 / 14.1	87.5 / 11.7	87.7 / **16.1**	88.3 / 4.30	88.0 / 3.66	**89.8** / 5.71
pcb4	94.4 / 15.5	94.6 / 24.9	95.9 / 31.2	94.6 / 34.2	95.7 / 30.6	**97.2 / 38.3**	95.9 / 35.1
pipe_fryum	75.4 / 4.40	96.0 / 23.6	96.9 / 27.2	95.7 / 24.4	**98.2** / 33.2	97.8 / 26.3	**98.2 / 33.3**
Average	79.6 / 5.00	94.2 / 25.7	95.0 / 26.3	95.1 / **29.2**	95.5 / 21.3	**95.7** / 21.4	**95.7** / 23.4

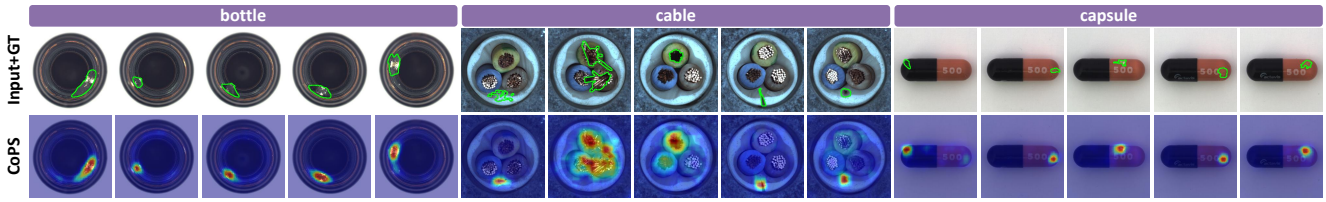

Figure S3. Qualitative segmentation results for the bottle, cable, and capsule categories from the MVTec-AD dataset.

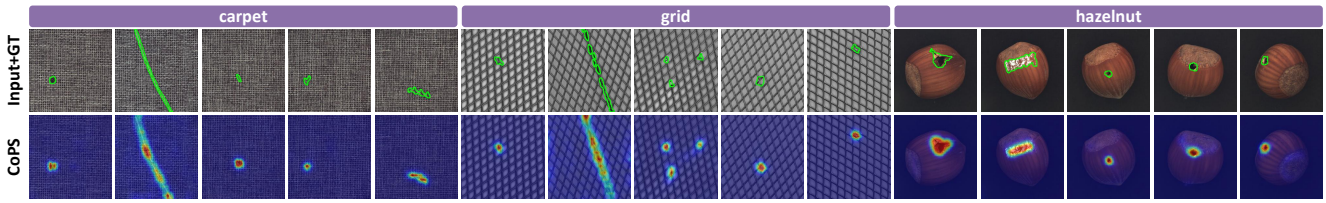

Figure S4. Qualitative segmentation results for the carpet, grid, and hazelnut categories from the MVTec-AD dataset.

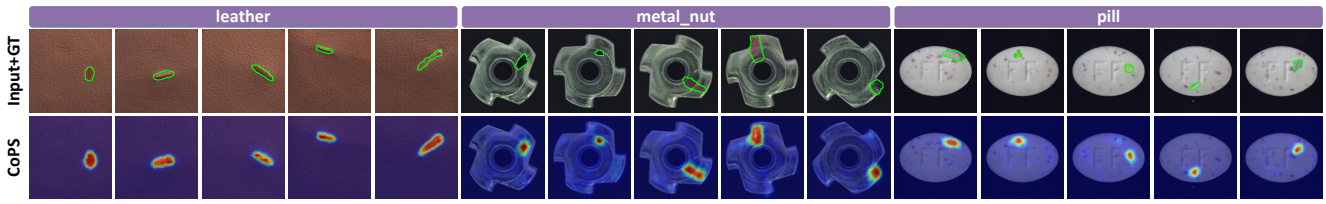

Figure S5. Qualitative segmentation results for the leather, metal nut, and pill categories from the MVTec-AD dataset.

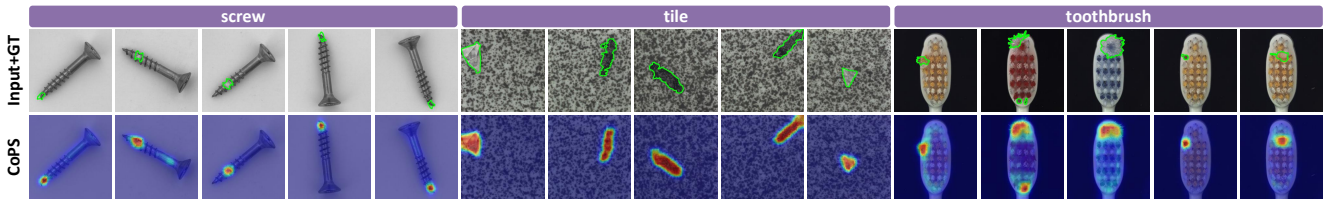

Figure S6. Qualitative segmentation results for the screw, tile, and toothbrush categories from the MVTec-AD dataset.

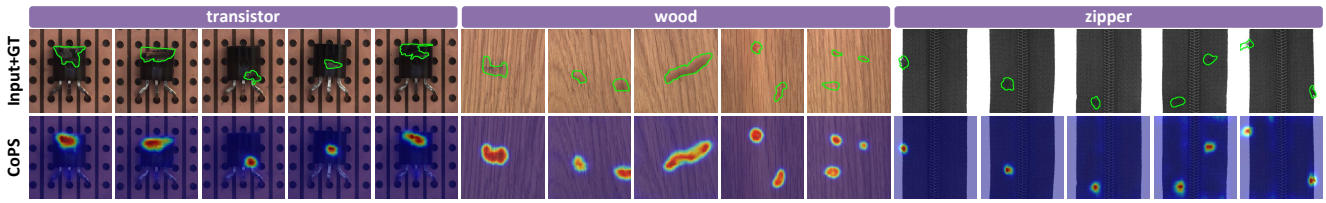

Figure S7. Qualitative segmentation results for the transistor, wood, and zipper categories from the MVTec-AD dataset.

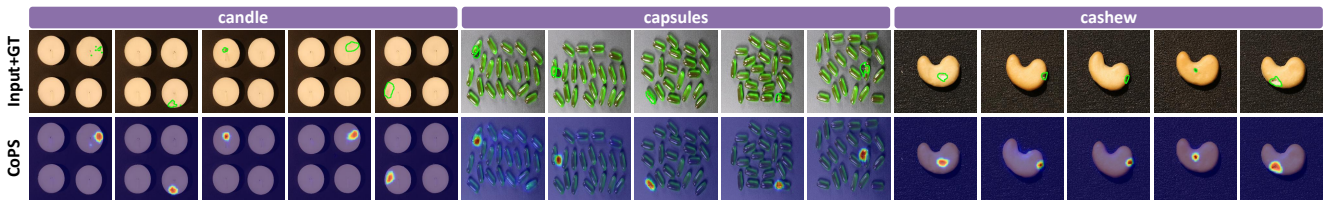

Figure S8. Qualitative segmentation results for the candle, capsules, and cashew categories from the VisA dataset.

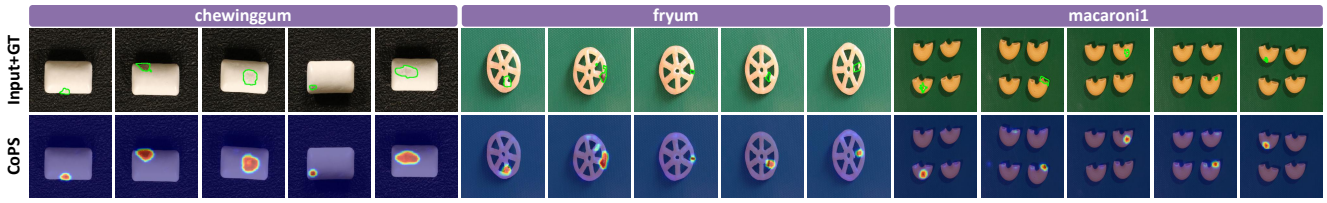

Figure S9. Qualitative segmentation results for the chewinggum, fryum, and macaroni1 categories from the VisA dataset.

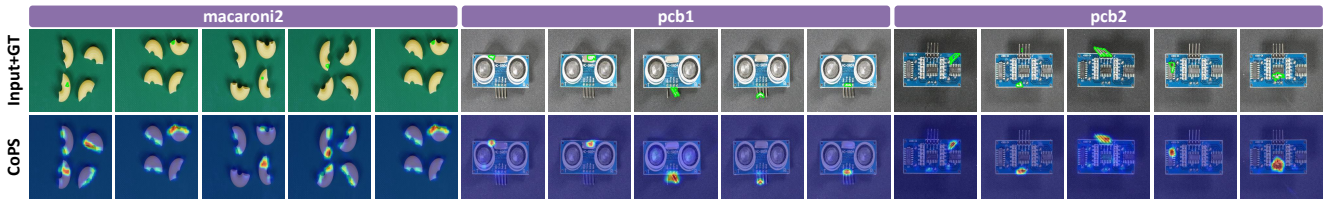

Figure S10. Qualitative segmentation results for the macaroni2, pcb1, and pcb2 categories from the VisA dataset.

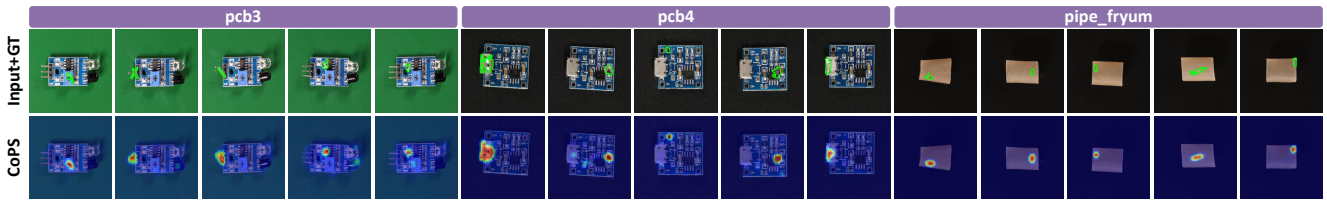

Figure S11. Qualitative segmentation results for the pcb3, pcb4, and pipe fryum categories from the VisA dataset.

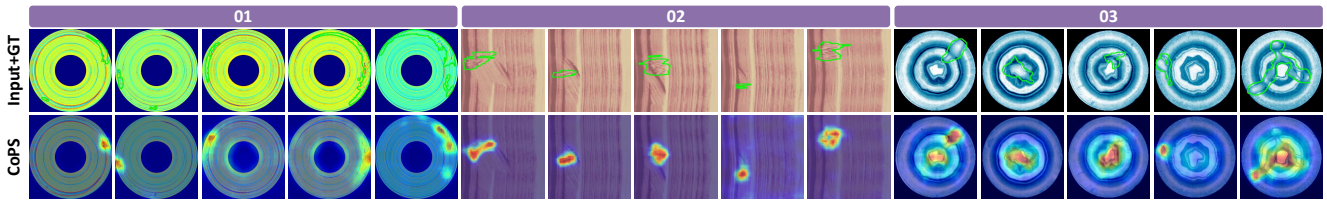

Figure S12. Qualitative segmentation results for the 01, 02, and 03 categories from the BTAD dataset.

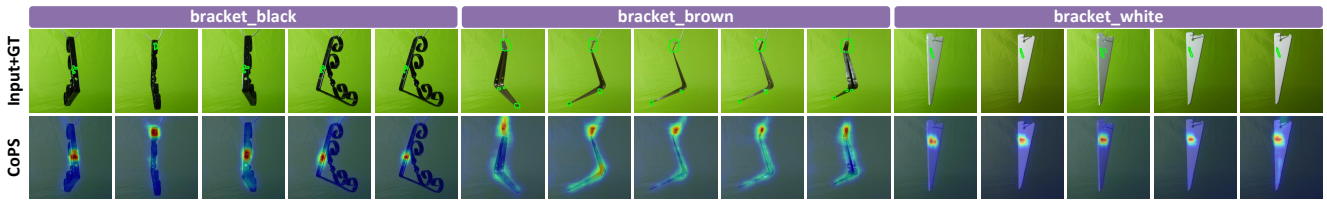

Figure S13. Qualitative segmentation results for the bracket black, brown, and white categories from the MPDD dataset.

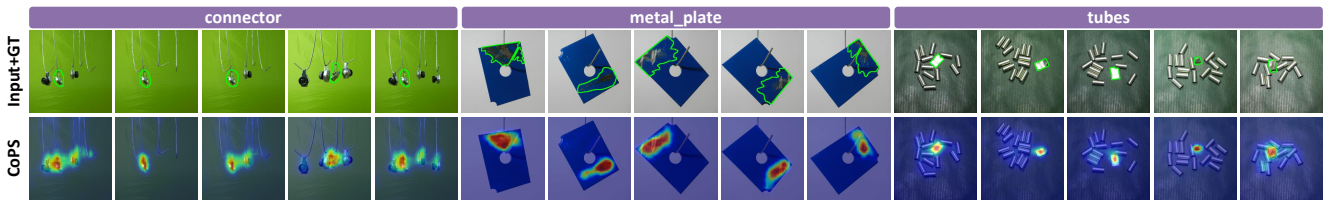

Figure S14. Qualitative segmentation results for the connector, metal plate, and tubes categories from the MPDD dataset.

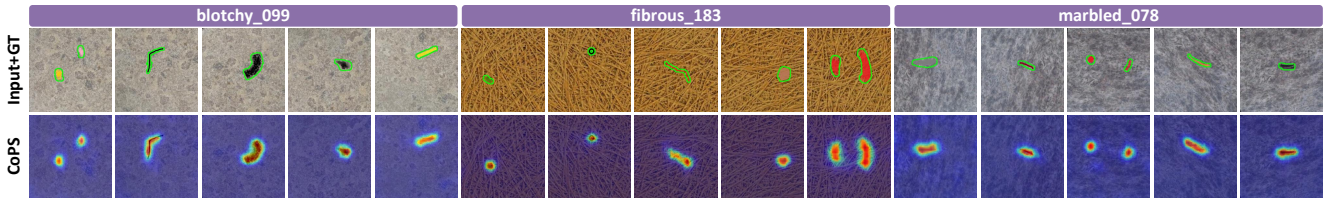

Figure S15. Qualitative segmentation results for the blotchy, fibrous, and marbled categories from the DTD-Synthetic dataset.

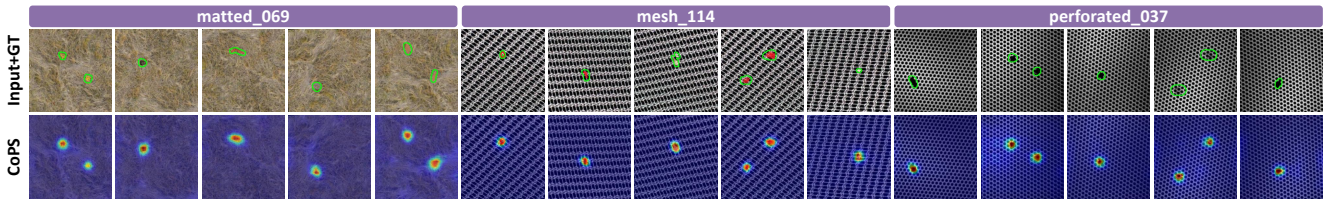

Figure S16. Qualitative segmentation results for the matted, mesh, and perforated categories from the DTD-Synthetic dataset.

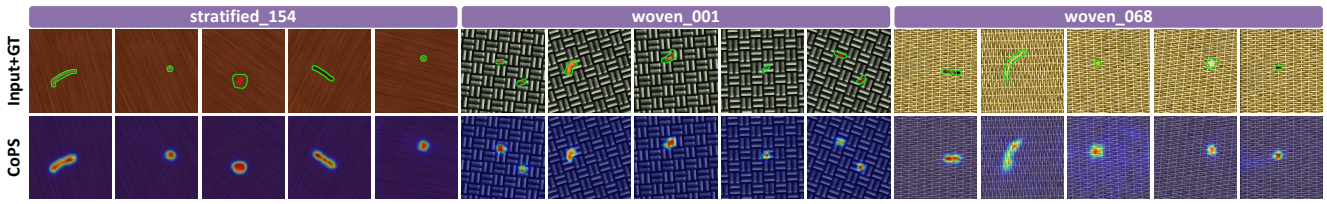

Figure S17. Qualitative segmentation results for the stratified, woven1, and woven2 categories from the DTD-Synthetic dataset.

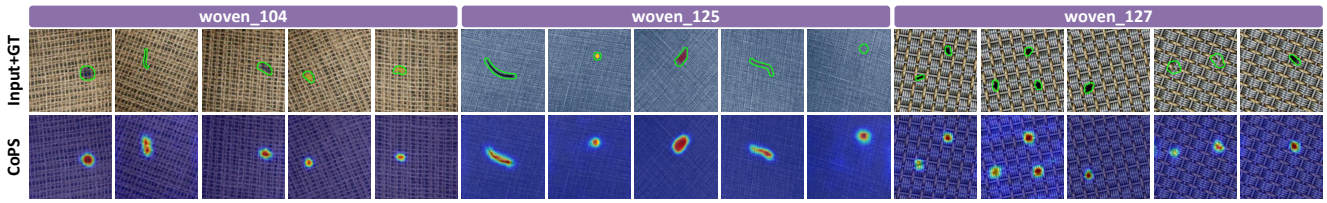

Figure S18. Qualitative segmentation results for the woven3, woven4, and woven5 categories from the DTD-Synthetic dataset.

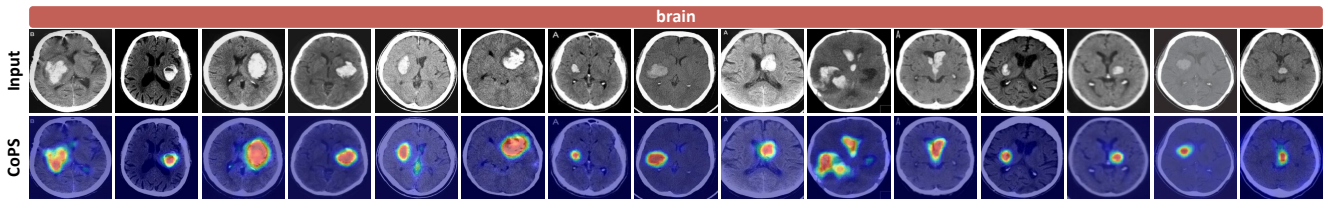

Figure S19. Qualitative segmentation results for the brain category from the HeadCT dataset.

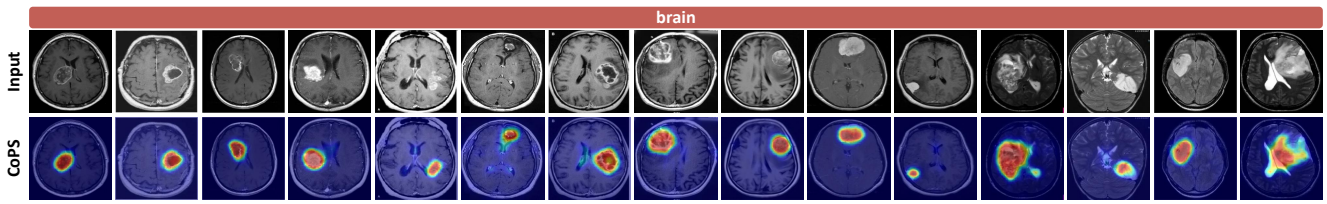

Figure S20. Qualitative segmentation results for the brain category from the BrainMRI dataset.

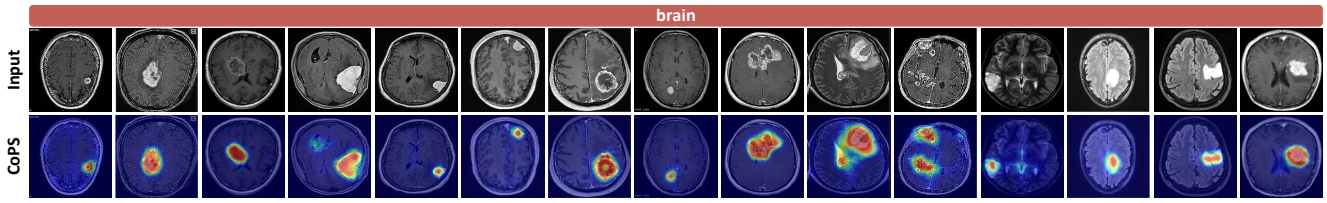

Figure S21. Qualitative segmentation results for the brain category from the Br35H dataset.

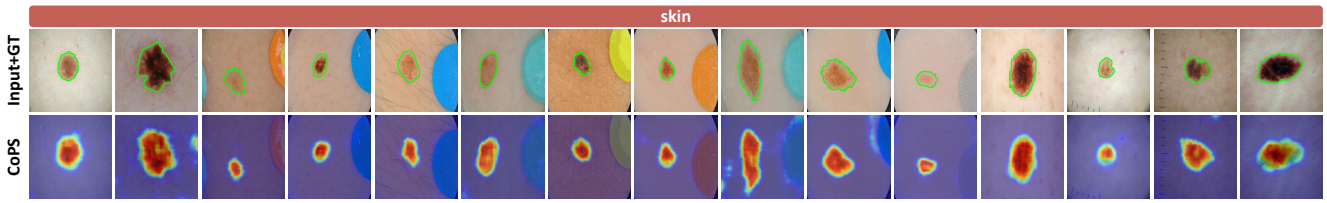

Figure S22. Qualitative segmentation results for the skin category from the ISIC dataset.

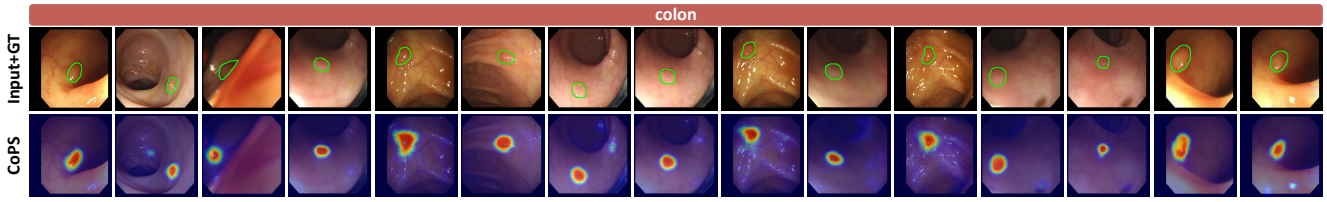

Figure S23. Qualitative segmentation results for the colon category from the CVC-ColonDB dataset.

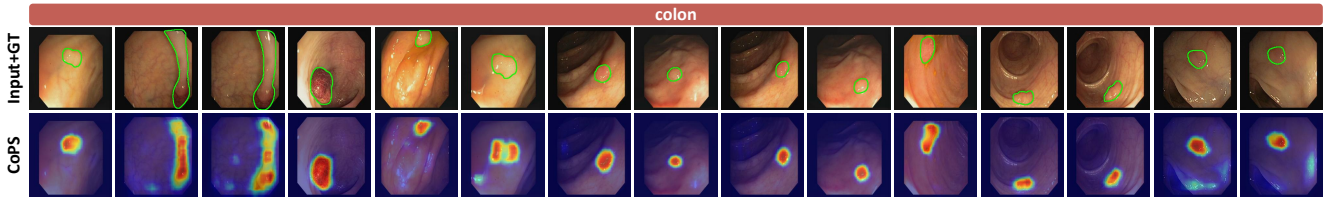

Figure S24. Qualitative segmentation results for the colon category from the CVC-ClinicDB dataset.

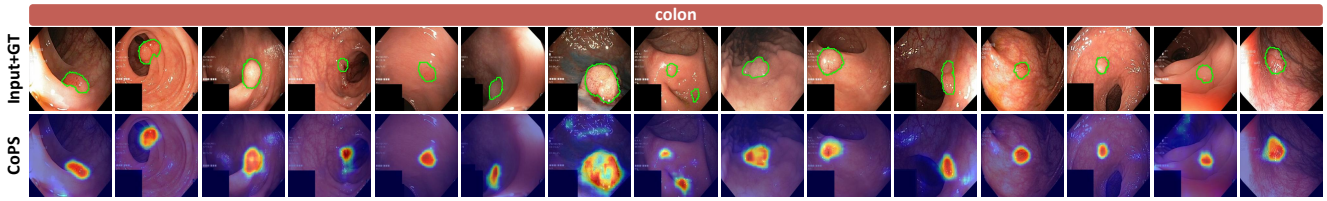

Figure S25. Qualitative segmentation results for the colon category from the Kvasir dataset.

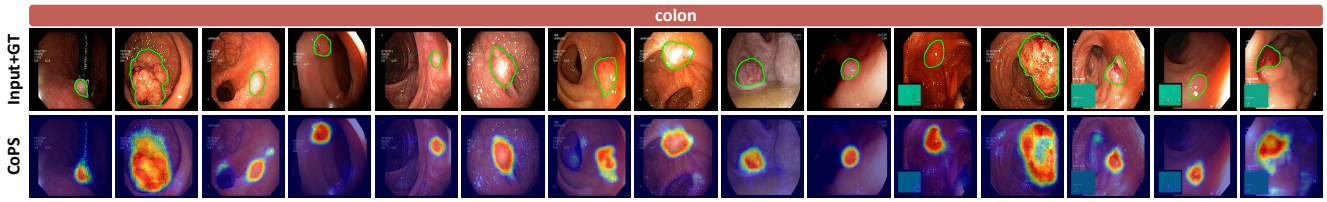

Figure S26. Qualitative segmentation results for the colon category from the Endo dataset.