Appendix: Fine-grained Abnormality Prompt Learning for Zero-shot Anomaly Detection

Jiawen Zhu[1] **Yew-Soon Ong**[2] **Chunhua Shen**[3] **Guansong Pang**[1*]
[1]Singapore Management University, Singapore
[2]Nanyang Technological University, Singapore
[3]Zhejiang University, China

Table 1. Data statistics of MVTec AD and VisA.

Dataset	Subset	Type	Original Training	Original Test	
			Normal	Normal	Anomalous
MVTec AD	Carpet	Texture	280	28	89
	Grid	Texture	264	21	57
	Leather	Texture	245	32	92
	Tile	Texture	230	33	83
	Wood	Texture	247	19	60
	Bottle	Object	209	20	63
	Capsule	Object	219	23	109
	Pill	Object	267	26	141
	Transistor	Object	213	60	40
	Zipper	Object	240	32	119
	Cable	Object	224	58	92
	Hazelnut	Object	391	40	70
	Metal_nut	Object	220	22	93
	Screw	Object	320	41	119
	Toothbrush	Object	60	12	30
VisA	candle	Object	900	100	100
	capsules	Object	542	60	100
	cashew	Object	450	50	100
	chewinggum	Object	453	50	100
	fryum	Object	450	50	100
	macaroni1	Object	900	100	100
	macaroni2	Object	900	100	100
	pcb1	Object	904	100	100
	pcb2	Object	901	100	100
	pcb3	Object	905	101	100
	pcb4	Object	904	101	100
	pipe_fryum	Object	450	50	100

Table 2. Data statistics of the other 17 AD datasets. They are used for ZSAD inference only.

Data type	Dataset	Modalities	C	Normal	Anomalous
Object	SDD	Photography	1	286	54
	BTAD	Photography	3	451	290
	MPDD	Photography	6	176	282
Textual	AITEX	Photography	12	564	183
	DAGM	Photography	10	6996	1054
	DTD-Synthetic	Photography	12	357	947
	ELPV	Electroluminescence	2	377	715
Brain	BrainMRI	Radiology (MRI)	1	98	155
	HeadCT	Radiology (CT)	1	100	100
	Br35H	Radiology (MRI)	1	1500	1500
Fundus	LAG	Fundus Photography	1	786	1711
Colon	CVC-ColonDB	Endoscopy	1	0	380
	CVC-ClinicDB	Endoscopy	1	0	612
	Kvasir	Endoscopy	1	0	1000
	Endo	Endoscopy	1	0	200
Skin	ISIC	Photography	1	0	379
Thyroid	TN3K	Radiology (Utralsound)	1	0	614

A. Dataset Details

A.1. Data Statistics of Training and Testing

We conduct extensive experiments on 19 real-world Anomaly Detection (AD) datasets, including nine industrial defect inspection datasets (MVTecAD [2], VisA [33], DAGM [28], DTD-Synthetic [1], AITEX [24], SDD [25], BTAD [21], MPDD [15], ELPV[5]) and ten medical anomaly detection datasets (BrainMRI [23], HeadCT [23], LAG [20], Br35H [11], CVC-ColonDB [26], CVC-ClinicDB [3], Kvasir [16], Endo [12], ISIC [10], TN3K [8]).

To assess the ZSAD performance, the test set of MVTec AD is used as the auxiliary training data, on which AD models are trained, and they are subsequently evaluated on the test set of the other 18 datasets without any further training. We train the model on the test set of VisA when evaluating the performance on MVTec AD. Table 1 provides the data statistics of MVTec AD and VisA, while Table 2 shows the test set statistics of the other 17 datasets.

*Corresponding author: Guansong Pang (gspang@smu.edu.sg)

B. Implementation Details

B.1. Details of Model Configuration.

Following previous works [4, 7, 32], `FAPrompt` adopts a modified version of CLIP –OpenCLIP [13] and its publicly available pre-trained backbone `VIT-L/14@336px`– as the VLM backbone to enhance the model's attention to local features while preserving its original structure. The parameters of both visual and text encoders in CLIP are kept frozen. Following [32], we replace the original Q-K self-attention mechanism in the visual encoder with a V-V self-attention mechanism during patch feature extraction, starting from the 6th layer of the visual encoder. The parameters of both the visual and text encoders in CLIP are frozen throughout the experiments.

Inspired by previous works [17, 18, 32], We use text prompt tuning to refine the original textual space of CLIP by adding additional learnable token embeddings into its text encoder. By default, the learnable token embeddings are attached to the first 9 layers of the text encoder to refine the textual space, with a token length of four for each layer. The lengths of the learnable normal prompt and abnormal tokens in CAP are set to five and two, respectively. The number of fine-grained abnormality prompts (K) and selected patch tokens (M) in DAP are both set to 10. To align with the dimension of `VIT-L/14@336px`, the abnormality prior network $\psi(\cdot)$ is configured with the input and output dimensions of $768 \times M$ and 768, respectively, and includes a hidden layer of size $(768 \times M)/16$ with ReLU activation.

We utilize the Adam optimizer with an initial learning rate of 1e-3 to update the model parameters. The input images are resized to 518×518 with a batch size of eight. This resizing is also applied to other baseline models for a fair comparison, while preserving their original data preprocessing methods, if applicable. The training is conducted for seven epochs across all experiments. During the inference stage, a Gaussian filter with $\sigma = 10$ is applied to smooth the anomaly score map. We follow the same random seed (111) as previous methods for fair comparison. All experiments are conducted using PyTorch on a single GPU (NVIDIA GeForce RTX 3090).

B.2. Implementation of Comparison Methods

To evaluate the efficiency of `FAPrompt`, we compare its performance against ten state-of-the-art (SotA) baselines. The results for CLIP [13], CLIP-AC [13], WinCLIP [14], APRIL-GAN [4], CoOp [31], and AnomalyCLIP [32] are sourced from AnomalyCLIP, except the newly added datasets (SDD, AITEX, ELPV, LAG). For fair comparison, these implementations follow the setup of AnomalyCLIP. We use the official implementations of AnoVL [7], CoCoOp [30], FiLO [9] and BLIP (ViT-B/16) [19] on all our datasets. To adapt CoCoOp for ZSAD, we replace its learnable text prompt templates with normality and abnormality text prompt templates, which is consistent with the implementation of CoOp in existing ZSAD studies. We obtain the results of BLIP by only changing the backbone. All other parameters remain consistent with those specified in their original papers.

B.2.1. The Algorithm of `FAPrompt`

To better illustrate the interactions between the CAP and DAP, we summarize the step-by-step procedure of Fine-grained Abnormality Learning (`FAPrompt`) in Algorithm 1.

Algorithm 1 Fine-grained Abnormality Learning (`FAPrompt`)

Input: Dataset $\mathcal{D} = \{x, y, \mathbf{G}\}$, visual encoder $f_v(\cdot)$, text encoder $f_t(\cdot)$, abnormality prior network $\psi(\cdot)$, normal learnable tokens $\{V_1, V_2, \ldots, V_E\}$, abnormal learnable tokens $\{A_1^i, A_2^i, \ldots, A_{E'}^i\}_{i=1}^K$

Output: Text encoder $f_t(\cdot)$, abnormality prior network $\psi(\cdot)$, normal learnable tokens $\{V_1, V_2, \ldots, V_E\}$, abnormal learnable tokens $\{A_1^i, A_2^i, \ldots, A_{E'}^i\}_{i=1}^K$

for *epoch* $= 1$ to N **do**
  // Compound Abnormality Prompt Learning
  Construct initial normal prompt $\mathcal{P}^n$ and abnormal prompts $\mathcal{P}^a = \{\mathcal{P}^{a_1}, \ldots, \mathcal{P}^{a_K}\}$ based on normal and abnormal learnable tokens using Eq. (1).
  Encode prompts: $\mathbf{F}_n = f_t(\mathcal{P}^n)$, $\mathbf{F}_{a_i} = \{f_t(\mathcal{P}^{a_i})\}_{i=1}^K$
  Compute orthogonal constraint loss $\mathcal{L}_{oc}$ (Eq. 2)
  Compute abnormal prompt prototype: $\mathbf{F}_a = \frac{1}{K}\sum_{i=1}^K \mathbf{F}_{a_i}$
  Generate segmentation maps $\mathcal{M}^n$, $\mathcal{M}^a$
  // Data-dependent Abnormality Prior Learning
  Select top-M patch tokens $\mathbf{p}_x = \{p_1, p_2, \ldots, p_M\}$ that most similar to $\mathbf{F}_a$ (Eq. 3)
  Compute sample-wise abnormality prior: $\Omega_x = \psi(\mathbf{p}_x)$
  Refine abnormal prompts $\hat{\mathcal{P}}^a$ based on Ω_x (Eq. 4)
  Compute refined prototype: $\hat{\mathbf{F}}_a = \frac{1}{|\hat{\mathcal{P}}^a|}\sum_{\hat{\mathcal{P}}^{a_i} \in \hat{\mathcal{P}}^a} f_t(\hat{\mathcal{P}}^{a_i})$
  Compute prior loss $\mathcal{L}_{prior}$ (Eq. 5)
  Generate refined segmentation maps $\hat{\mathcal{M}}^n$, $\hat{\mathcal{M}}^a$
  Compute image-level anomaly score $s(x)$ (Eq. 8, 9)
  Compute pixel-level anomaly map $\mathcal{M}_x$ (Eq. 11)
  Compute pixel-level loss $\mathcal{L}_{local}$ (Eq. 7)
  Compute image-level loss $\mathcal{L}_{global}$ (Eq. 10)
  Update parameters of learnable tokens $\{V_1, V_2, \ldots, V_E\}$, $\{A_1^i, A_2^i, \ldots, A_{E'}^i\}_{i=1}^K$, text encoder $f_t(\cdot)$, and abnormality prior network $\psi(\cdot)$
end for

C. Additional Results

C.1. Model Complexity of `FAPrompt` vs. SotA Methods

We compare the model complexity of `FAPrompt` with SotA methods in Table 3, evaluating the number of parameters, per-batch training time, and per-image inference time. The batch size for all approaches is set to eight for fair comparison, excluding training-free methods WinCLIP and AnoVL. While `FAPrompt` introduces additional trainable parameters, leading to a slightly longer training time, this minor computational overhead results in substantial performance improvements over competing methods. Addition-

Table 3. Number of parameters, per-batch training time (ms) and per-image inference time (ms) in comparison with competing methods.

Model	Number of Para.	Training Time	Inference Time
WinCLIP	0	0	227.5±0.7
AnoVL	0	0	171.4±0.5
APRIL-GAN	3148800	368.7±0.5	47.9±0.1
CoOp	9216	643.8±1.1	89.9±0.7
CoCoOp	83760	737.4±3.6	93.8±0.7
AnomalyCLIP	5555200	914.1±0.9	124.2±0.9
FAPrompt	9612256	1354.1±1.7	214.7±0.8

ally, since training is performed offline, this training computational overhead is generally negligible in real-world applications. In terms of inference time, our approach remains reasonably efficient and responsive.

C.2. Comparison with SOTA Full-shot Methods and Prompt Tuning Methods

We conduct experiments on five of the most commonly used datasets to examine the performance gap between FAPrompt and two SotA full-shot methods, PatchCore [22] and RD4AD [6]. Note that it is not a fair comparison as PatchCore and RD4AD utilize the full training data of each testing dataset in its detection while ZSAD methods like FAPrompt does not use any of such training data. The results presented in Table 4 are only for analyzing the possible upper bound performance of ZSAD. Despite the unfair utilization of the dataset-specific training data in PatchCore and RD4AD, FAPrompt obtains rather impressive detection performance, further reducing the performance gap between ZSAD and full-shot methods.

We also compare FAPrompt with SotA prompt tuning approach TCP [29] to further verify the effectiveness of fine-grained abnormality prompt in Table 5. Sine TCP is not originally designed for anomaly detection and its contextual information relies heavily on handcrafted text prompts, we adapted TCP for the ZSAD by testing two types of AD-oriented text prompts, resulting in two variants of TCP for ZSAD, **TCP_V1** and **TCP_V2**:

- **TCP_V1**, where we use a straightforward prompt design: the normal prompt is in the form of "This is a photo of [cls]." while the abnormal prompt is in the form of "This is a photo of damaged [cls]."
- **TCP_V2**, where we adopt the complete set of the prompt templates from WinCLIP.

For a fair comparison, we maintained the original model designs of TCP throughout the experiments. As shown in Table 5, both TCP variants largely underperform AnomalyCLIP and FAPrompt in the ZSAD task. This is primarily due to the fact that TCP is not designed for ZSAD and also has strong reliance on handcrafted text prompts.

In contrast, FAPrompt is specifically designed for the ZSAD task, leveraging data-dependent abnormality prior of the query images to learn complementary abnormality prompts. This adaptive approach enables FAPrompt to more effectively capture a wide variety of anomalies, resulting in promising performance in both image-level and pixel-level ZSAD tasks.

Table 4. Comparison of ZSAD performance between FAPrompt and two SotA full-shot methods. The best and second-best results are respectively highlighted in red and blue.

Dataset	AnomalyCLIP	FAPrompt	PatchCore	RD4AD
Image-level (AUROC, AP)				
MVTecAD	(91.5, 96.2)	(91.9, 95.7)	(99.0, 99.7)	(98.7, 99.4)
VisA	(82.1, 85.4)	(84.5, 86.8)	(94.6, 95.9)	(95.3, 95.7)
BTAD	(88.3, 87.3)	(92.2, 92.5)	(93.2, 98.6)	(93.8, 96.8)
MPDD	(77.0, 82.0)	(80.1, 83.9)	(94.1, 96.3)	(91.6, 93.8)
DAGM	(97.5, 92.3)	(98.8, 95.3)	(92.7, 81.3)	(92.9, 79.1)
Pixel-level (AUROC, PRO)				
MVTecAD	(91.1, 81.4)	(90.6, 83.3)	(98.1, 92.8)	(97.8, 93.6)
VisA	(95.5, 87.0)	(95.9, 87.5)	(98.5, 92.2)	(98.4, 91.2)
BTAD	(94.2, 74.8)	(95.6, 75.1)	(97.4, 74.4)	(97.5, 75.1)
MPDD	(96.5, 87.0)	(96.5, 87.9)	(98.8, 94.9)	(98.4, 95.2)
DAGM	(95.6, 91.0)	(98.2, 95.0)	(95.9, 87.9)	(96.8, 91.9)

Table 5. Comparison with TCP.

Model	Industrial		Medical	
	image-level	pixel-level	image-level	pixel-level
AnomalyCLIP	(85.0, 83.6)	(94.4, 84.8)	(87.7, 90.6)	(83.2, 62.9)
TCP_V1	(61.3, 55.9)	(87.2, 66.6)	(56.4, 61.7)	(80.2, 60.9)
TCP_V2	(64.9, 59.1)	(88.5, 71.5)	(53.3, 60.3)	(76.8, 52.9)
FAPrompt	(**88.5**, **87.5**)	(**95.0**, **85.6**)	(**91.0**, **93.0**)	(**85.7**, **66.2**)

C.3. t-SNE Visualization of Prompt-wise Anomaly Score Map

To explore the complementarity of abnormality prompts in FAPrompt, we provide two-dimensional t-SNE visualization of the anomaly score map S_x^a and quantitative results of 'AnomalyCLIP', prompt ensemble method 'AnomalyCLIP Ensemble*' for their comparison with FAPrompt on the three datasets. The results are shown in Fig. 1. Note that the difference between AnomalyCLIP and FAPrompt/AnomalyCLIP Ensemble* in the figure is because AnomalyCLIP learns one single abnormality prompt only while the FAPrompt/AnomalyCLIP Ensemble* learns 10 abnormality prompts.

FAPrompt vs. AnomalyCLIP. It is clear that compared to AnomalyCLIP, FAPrompt learns a set of effective complementary abnormal patterns captured by the 10 abnormality prompts, resulting in better detection performance on datasets with complex anomaly cases.

For example, on the datasets BTAD(01) and VisA (pcb4), several anomalies, which are distributed very closely to, or overlapped with part of the normal im-

ages, are difficult to detect using single abnormality prompt in AnomalyCLIP, indicating that its single abnormality prompt is not discriminative w.r.t. these anomalies. FAPrompt alleviates this situation with the abnormality prompts that show visually different, discriminative power.

For datasets with simpler patterns like VisA (chewing-gum), single abnormality prompt is sufficient, while having multiple abnormality prompts in FAPrompt do not have adverse effect. This demonstrates the performance of FAPrompt in achieving stable, effective detection across simple and complex datasets.

FAPrompt vs. the prompt ensemble method 'AnomalyCLIP Ensemble*'. Despite also learning multiple abnormality prompts, it is clear from the visualization that the abnormality prompts in AnomalyCLIP Ensemble* tend to be clustered closely, while that in FAPrompt is much more disperse, *e.g.*, two clustered patterns on BTAD(01) and one clustered pattern on VisA (pcb4) learned by AnomalyCLIP Ensemble* vs. four disperse patterns on both datasets learned by FAPrompt. Importantly, the more disperse abnormal patterns from FAPrompt provides complementary discriminative power to each other, substantiated by the enhanced AUROC/AP performance compared to AnomalyCLIP Ensemble*.

C.4. Hyperparameter Sensitivity Analysis

C.4.1. Complete Sensitivity Analysis for K and M

We present the complete image-level and pixel-level results for the sensitivity w.r.t. the number of abnormality prompts (K) in CAP and the number of selected patch tokens (M) in DAP across Industrial and Medical datasets in Fig. 2. The trend of the results is consistent with our analysis in the main text.

C.4.2. Sensitivity Analysis for Length of Learnable Normal and Abnormal Tokens.

We also evaluate the sensitivity of the length of learnable normal and abnormal tokens $\{E, E'\}$ in **CAP** module. The Image-level and pixel-level ZSAD results are shown in Fig. 3. Overall, the setting of (5, 2) works best for both industrial and medical AD, yielding strong ZSAD performance. Longer prompt lengths, such as (10, 4), can introduce more complexity without clear performance improvement, particularly in pixel-level performance. Using shorter prompt lengths, *e.g.*, the setting of (2, 1), lacks sufficient capacity to support the ZSAD task, leading to consistently weaker performance.

C.4.3. Sensitivity Analysis for Learnable Tokens.

To evaluate the sensitivity of the learnable tokens, we also conduct ablation studies on the number of layers with learnable tokens and the length of the tokens. As shown by the results in Table 6, the performance generally gets improved with an increasing number of layers, reaching optimal performance at 9 layers. Beyond 9 layers, it tends to over-generalization, leading to a decrease in the detection performance. A similar pattern was observed with the token length, where FAPrompt achieves the best overall performance with a token length of 4 and 6.

Table 6. Hyperparameter analysis of the number of layers with learnable tokens and the length of the tokens.

Model	Industrial Datasets		Medical Datasets	
	Image-level	Pixel-level	Image-level	Pixel-level
Length of learnable token				
2	(88.4, 87.4)	(95.0, 84.8)	(90.7, 91.7)	(84.9, 65.1)
4	(88.5, 87.5)	(95.0, 85.6)	(91.0, 93.0)	(85.7, 66.2)
6	(90.0, 87.7)	(94.8, 85.3)	(91.2, 93.5)	(85.0, 65.2)
8	(87.8, 86.6)	(94.9, 84.3)	(90.6, 92.3)	(85.0, 65.1)
Layers having learnable tokens				
5	(88.0, 87.3)	(94.2, 85.5)	(91.2, 93.0)	(84.6, 65.0)
7	(88.0, 86.9)	(94.6, 84.3)	(91.0, 93.3)	(85.3, 65.2)
9	(88.5, 87.5)	(95.0, 85.6)	(91.0, 93.0)	(85.7, 66.2)
11	(88.1, 87.2)	(94.9, 84.5)	(90.5, 92.7)	(84.5, 63.5)

C.5. Qualitative Results of FAPrompt

C.5.1. Comparison with SOTA ZSAD methods

We compare the anomaly maps generated by FAPrompt with those produced by other ZSAD models across various datasets, as shown in Fig. 4. APRIL-GAN and AnomalyCLIP are selected as representatives of handcrafted and learnable text prompt competitors, respectively. The visualization results show that FAPrompt demonstrates significantly more accurate segmentation compared to the other two methods across both industrial and medical domains. In particular, despite not accessing any additional information or training from medical data, FAPrompt effectively localizes abnormal lesion/tumor regions, which highlight the cross-dataset generalization superiority of the fine-grained abnormality semantics learned by FAPrompt.

C.5.2. Visualization on Samples with Multiple Anomalous Types

To assess the performance on samples containing multiple anomalous types within a single image, we also provide visualization of pixel-level detection results on such samples from three MVTecAD categories (zipper, pill and wood) and AITEX. The results shown in Fig. 5 demonstrate that despite using a single abnormality prompt prototype, FAPrompt can still effectively detect multiple anomaly types in a single image.

C.5.3. Visualization on Diverse Datasets

In addition, we also provide pixel-level anomaly score maps on diverse datasets to further showcase the strong

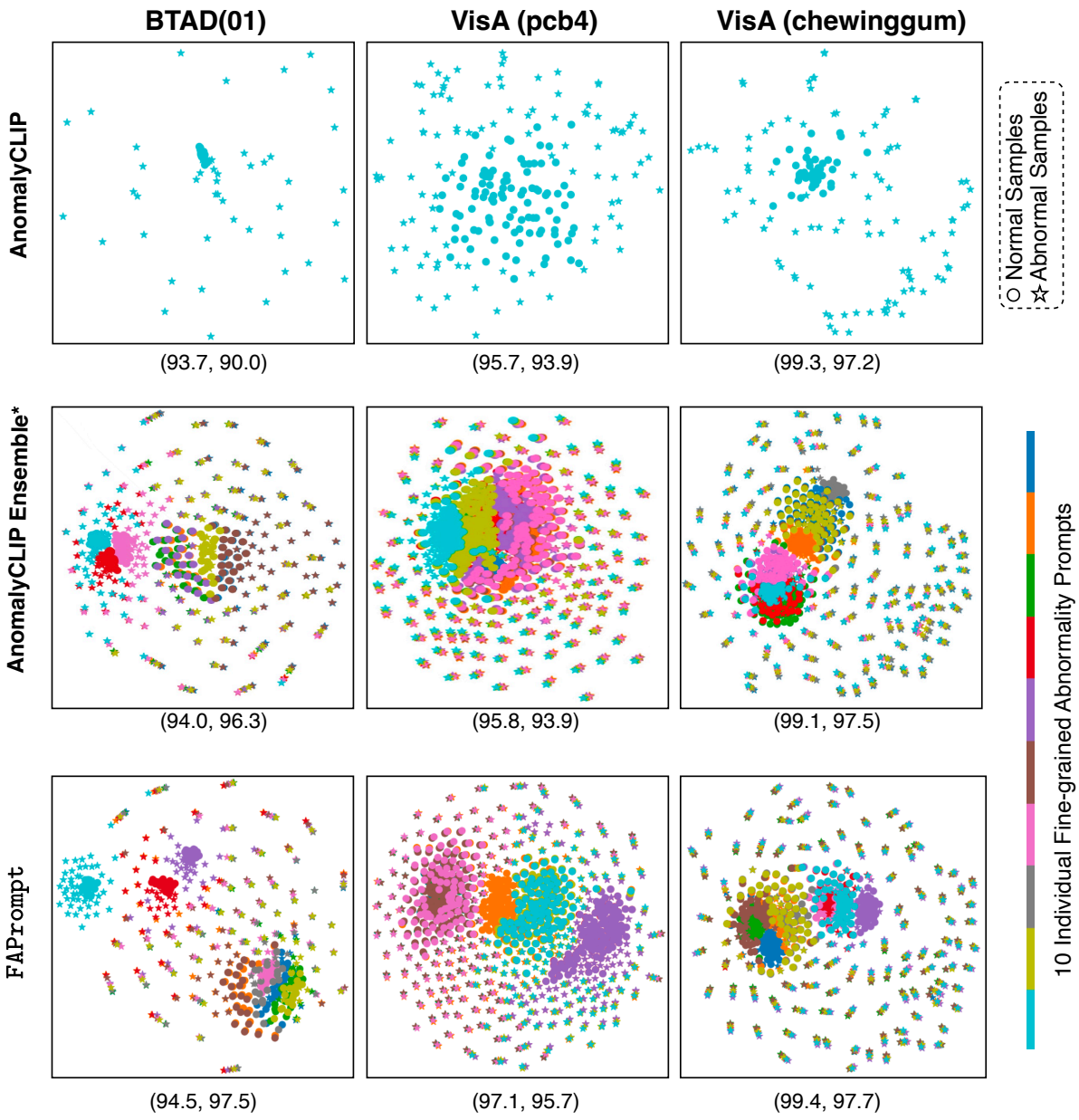

Figure 1. 2-D t-SNE visualizations and quantitative results (Image-level AUROC, Pixel-level AUROC) of FAPrompt, AnomalyCLIP and its ensemble method AnomalyCLIP Ensemble*.

segmentation capability of FAPrompt in Figs. 6 to 15. Specifically, for the industrial AD datasets, we select three object categories (capsule, pipe_fryum in VisA and metal_plate in MPDD) and three texture categories (grid, tile in MVTecAD and AITEX) for visualization. For the medical AD datasets, we visualize the pixel-level anomaly detection performance for the brain, colon, skin, and thyroid anomalies.

C.6. Failure Cases and Limitations

While the proposed FAPrompt demonstrates promising detection results across various categories without any dataset-specific references, it may fail in certain cases. Fig. 16 illustrates some of these failure cases. Some cases

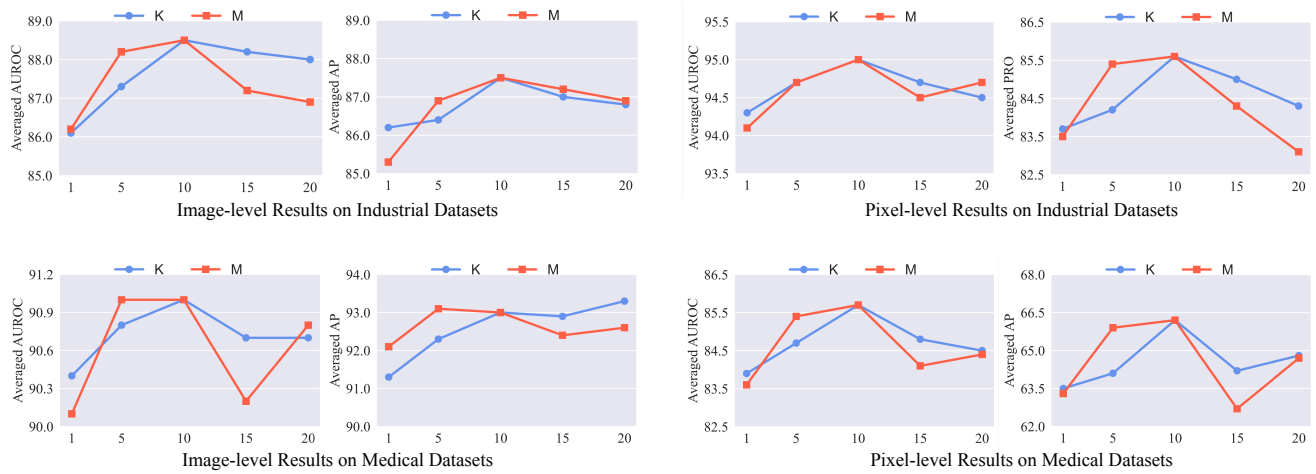

Figure 2. Averaged results with varying K and M.

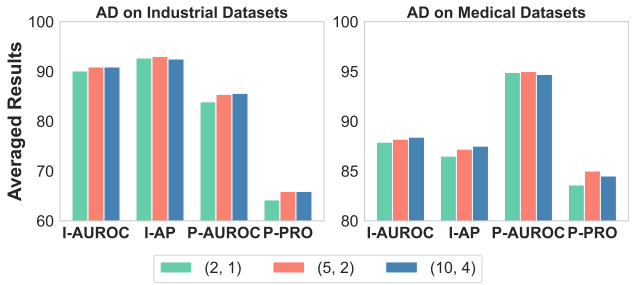

Figure 3. Averaged results of FAPrompt with varying prompt sizes of (E, E').

can be attributed to annotation errors. For example, images that contain multiple types of anomalies but are only partially labeled may lead to segmentation errors due to labeling inconsistencies, as can be seen in the stain defect in Fig. 16 (1). Additionally, instrument artifacts in some medical datasets are often misinterpreted as anomalies, leading to incorrect detection, *e.g.*, Fig. 16 (2). In other cases, FAPrompt may fail in challenging cases like the ones illustrated in Fig. 16 (3)-(6), where the anomalous regions may be too small, subtle, or overshadowed by other suspicious areas (according to FAPrompt's interpretation). Nevertheless, as demonstrated in this figure and Figs. 6 to 15, FAPrompt consistently strives to identify the most likely abnormal regions, without relying on any reference from the target datasets. Moving forward, incorporating more prior knowledge, *e.g.*, from in-context examples, knowledge graphs, or Large Language Models (LLMs), would be helpful for providing more discriminative information for achieving more accurate anomaly detection.

In Addition, for the auxiliary training data, following previous works, we only consider the commonly used MVTec AD and VisA datasets. We believe incorporating more recent large-scale datasets, such as Real-IAD [27], further enhance the generalizability of this research direction.

D. Detailed Empirical Results

D.1. Breakdown Results on VisA and MVTec AD

Tables 7 to 14 present detailed downbreak ZSAD results of FAPrompt against eight SotA methods across each category of the MVTecAD and VisA datasets.

D.2. Dataset-specific Results on Ablation Study

In this section, we present the dataset-specific image-level and pixel-level ZSAD results for module ablation in Table 15 and Table 16, respectively.

References

[1] Toshimichi Aota, Lloyd Teh Tzer Tong, and Takayuki Okatani. Zero-shot versus many-shot: Unsupervised texture anomaly detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5564–5572, 2023. 1

[2] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9592–9600, 2019. 1

[3] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 1

[4] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation

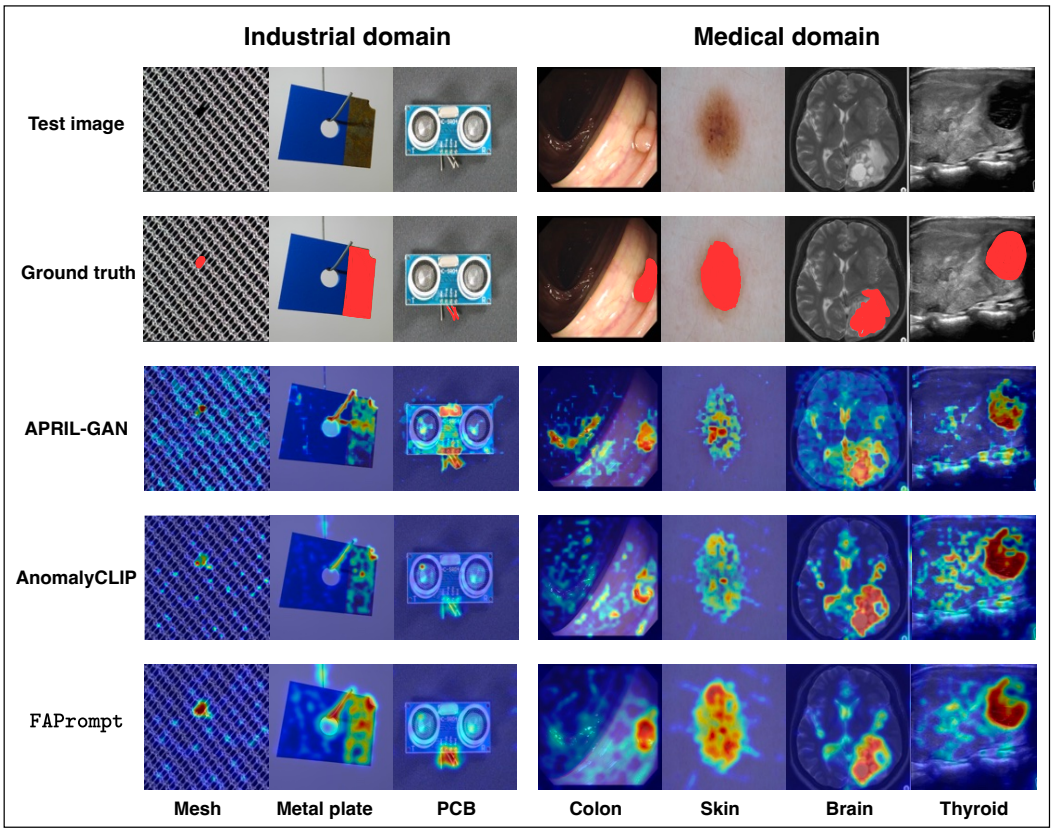

Figure 4. Visualization of anomaly maps generated by different ZSAD methods.

Table 7. Breakdown AUROC results of image-level ZSAD performance comparison on MVTecAD.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
Carpet	96.0	93.1	100.0	99.5	-	99.9	98.7	100.0	100.0
Grid	72.5	63.7	98.8	86.3	-	94.7	87.7	97.0	97.9
Leather	99.4	99.5	100.0	99.7	-	99.9	98.5	99.8	99.9
Tile	88.5	89.0	100.0	99.9	-	99.7	99.4	100.0	99.7
Wood	94.0	94.9	99.4	99.0	-	97.7	44.4	96.8	98.0
Bottle	45.9	46.1	99.2	92.0	-	87.7	80.2	89.3	89.8
Capsule	71.4	68.8	72.9	79.9	-	81.1	84.2	89.9	92.4
Pill	73.6	73.8	79.1	80.5	-	78.6	83.3	81.8	89.6
Transistor	48.8	51.2	88.0	80.8	-	92.2	77.3	92.8	81.7
Zipper	60.1	36.1	91.5	89.6	-	98.8	54.5	98.5	98.4
Cable	58.1	46.6	86.5	88.4	-	56.7	29.6	69.8	74.7
Hazelnut	88.7	91.1	93.9	89.6	-	93.5	11	97.2	96.5
Metal_nut	62.8	63.4	97.1	68.4	-	85.3	81.3	93.6	89.7
Screw	78.2	66.7	83.3	84.9	-	88.9	59	81.1	85.0
Toothbrush	73.3	89.2	88.0	53.8	-	77.5	88.6	84.7	85.6
MEAN	**74.1**	**71.5**	**91.8**	**86.2**	**92.5**	**88.8**	**71.8**	**91.5**	**91.9**

method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 1, 2

[5] Sergiu Deitsch, Vincent Christlein, Stephan Berger, Claudia Buerhop-Lutz, Andreas Maier, Florian Gallwitz, and Christian Riess. Automatic classification of defective photovoltaic

Table 8. Breakdown AP results of image-level ZSAD performance comparison on MVTecAD.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
Carpet	98.8	97.8	100.0	99.8	-	100.0	99.6	100.0	100.0
Grid	87.1	83.9	99.6	94.9	-	98.1	95.8	99.1	99.3
Leather	99.8	99.8	100.0	99.9	-	100.0	99.3	99.9	100.0
Tile	95.9	96.2	100.0	100.0	-	99.9	99.8	100.0	99.9
Wood	97.9	98.3	99.8	99.7	-	99.4	68.2	99.2	99.4
Bottle	78.9	79.8	99.8	97.7	-	96.4	93.1	97.0	96.7
Capsule	92.1	90.9	91.5	95.5	-	95.7	96.5	97.9	98.4
Pill	93.4	93.6	95.7	96.0	-	94.2	96.2	95.4	97.9
Transistor	48.1	49.9	87.1	77.5	-	90.2	71.1	90.6	78.9
Zipper	87.4	73.9	97.5	97.1	-	99.7	86.7	99.6	99.5
Cable	70.8	64.3	91.2	93.1	-	69.4	50.8	81.4	82.9
Hazelnut	94.6	95.9	96.9	94.8	-	96.7	45.9	98.6	98.1
Metal_nut	87.7	89.2	99.3	91.9	-	96.3	93.6	98.5	97.5
Screw	91.4	86.6	93.1	93.6	-	96.2	81.2	92.5	93.6
Toothbrush	90.7	96.0	95.6	71.5	-	90.4	95.1	93.7	93.8
MEAN	**87.6**	**86.4**	**96.5**	**93.5**	**96.7**	**94.8**	**84.9**	**96.2**	**95.7**

Table 9. Breakdown AUROC results of pixel-level ZSAD performance comparison on MVTecAD.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
Carpet	11.5	10.7	95.4	98.4	-	6.7	96.7	98.8	99.0
Grid	8.7	11.9	82.2	95.8	-	7.8	89.8	97.3	96.9
Leather	9.9	5.6	96.7	99.1	-	11.7	98.5	98.6	98.5
Tile	49.9	39.1	77.6	92.7	-	41.7	87.4	94.6	95.7
Wood	45.7	42.4	93.4	95.8	-	31.4	94.5	96.5	96.4
Bottle	17.5	23.3	89.5	83.4	-	23.1	89.7	90.4	90.3
Capsule	50.9	49.1	86.9	92.0	-	35.5	80.1	95.8	95.2
Pill	55.8	60.8	80.0	76.2	-	46.5	78.7	92.0	90.5
Transistor	51.1	48.5	74.7	62.4	-	50.1	66.2	71.0	69.8
Zipper	51.5	44.7	91.6	91.1	-	33.4	92.0	91.4	91.8
Cable	37.4	37.5	77.0	72.3	-	49.7	73.3	78.9	79.5
Hazelnut	25.2	34.0	94.3	96.1	-	30.2	95.9	97.1	97.5
Metal_nut	43.9	53.6	61.0	65.4	-	49.3	71.0	74.4	71.4
Screw	80.1	76.4	89.6	97.8	-	17.0	98.3	97.5	97.4
Toothbrush	36.3	35.0	86.9	95.8	-	64.9	89.1	91.9	89.7
MEAN	**38.4**	**38.2**	**85.1**	**87.6**	**89.8**	**33.3**	**86.7**	**91.1**	**90.6**

module cells in electroluminescence images. *Solar Energy*, 185:455–468, 2019. 1

[6] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 3

[7] Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2023. 1, 2

[8] Haifan Gong, Guanqi Chen, Ranran Wang, Xiang Xie, Mingzhi Mao, Yizhou Yu, Fei Chen, and Guanbin Li. Multi-task learning for thyroid nodule segmentation with thyroid region prior. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 257–261. IEEE, 2021. 1

[9] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024. 2

[10] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 1

Table 10. Breakdown PRO results of pixel-level ZSAD performance comparison on MVTecAD.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
Carpet	2.9	1.9	84.1	48.5	-	0.5	94.1	90.1	94.1
Grid	0.9	2.4	57.0	31.6	-	1.0	74.5	75.6	81.6
Leather	0.2	0.0	91.1	72.4	-	1.8	97.9	92.2	95.7
Tile	21.5	16.3	51.2	26.7	-	10.1	76.9	87.6	89.3
Wood	13.7	10.3	74.1	31.1	-	5.1	93.1	91.2	92.3
Bottle	1.4	4.9	76.4	45.6	-	4.5	79.4	80.9	81.0
Capsule	13.2	14.9	62.1	51.3	-	5.7	82.8	87.2	83.9
Pill	6.0	8.2	65.0	65.4	-	3.2	84.4	88.2	87.6
Transistor	15.3	11.2	43.4	21.3	-	9.3	51.5	58.1	59.0
Zipper	17.7	15.2	71.7	10.7	-	11.6	78.3	65.3	75.1
Cable	7.3	6.9	42.9	25.7	-	12.2	55.5	64.4	68.2
Hazelnut	2.8	9.4	81.6	70.3	-	4.7	89.2	92.4	93.3
Metal_nut	2.9	10.3	31.8	38.4	-	7.0	71.5	71.0	70.9
Screw	57.8	56.2	68.5	67.1	-	6.4	93.8	88.0	89.7
Toothbrush	5.8	5.2	67.7	54.5	-	16.6	71.6	88.5	87.3
MEAN	**11.3**	**11.6**	**64.6**	**44.0**	**76.2**	**6.6**	**79.6**	**81.4**	**83.3**

Table 11. Breakdown AUCROC results of image-level ZSAD performance comparison on VisA.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
candle	37.9	33.0	95.7	83.8	-	46.2	63.7	79.3	87.0
capsules	69.7	75.3	85.0	61.2	-	77.2	69.8	81.5	92.0
cashew	69.1	72.7	92.2	87.3	-	75.7	93.3	76.3	90.7
chewinggum	77.5	76.9	95.3	96.4	-	84.9	96.5	97.4	97.7
fryum	67.2	60.9	75.3	94.3	-	80.0	76.6	93.0	96.1
macaroni1	64.4	67.4	77.8	71.6	-	53.6	68.0	87.2	81.4
macaroni2	65.0	65.7	66.7	64.6	-	66.5	75.4	73.4	71.6
pcb1	54.9	43.9	79.8	53.4	-	24.7	81.5	85.4	70.6
pcb2	62.6	59.5	52.6	71.8	-	44.6	61.6	62.2	66.5
pcb3	52.2	49.0	70.2	66.8	-	54.4	66.4	62.7	68.6
pcb4	87.7	89.0	84.5	95.0	-	66.0	93.8	93.9	95.7
pipe_fryum	88.8	86.4	69.4	89.9	-	80.1	91.0	92.4	97.5
MEAN	**66.4**	**65.0**	**78.7**	**78.0**	**79.2**	**62.8**	**78.1**	**82.1**	**84.6**

[11] Ahmed Hamada. Br35h:: Brain tumor detection. kaggle (2020). 2020. 1

[12] Steven A Hicks, Debesh Jha, Vajira Thambawita, Pål Halvorsen, Hugo L Hammer, and Michael A Riegler. The endotect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII*, pages 263–274. Springer, 2021. 1

[13] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, et al. Openclip. *Zenodo*, 4:5, 2021. 1, 2

[14] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 2

[15] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops (ICUMT)*, pages 66–71. IEEE, 2021. 1

[16] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26*, pages 451–462. Springer, 2020. 1

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie,

Table 12. Breakdown AP results of image-level ZSAD performance comparison on VisA.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
candle	42.9	40.0	96.1	86.9	-	52.9	67.7	81.1	89.3
capsules	81.0	84.3	91.0	74.3	-	85.3	81.9	88.7	96.3
cashew	83.4	86.1	96.5	94.1	-	87.1	96.8	89.4	96.0
chewinggum	90.4	90.2	97.9	98.4	-	93.1	98.6	98.9	99.1
fryum	82.0	76.6	88.1	97.2	-	90.2	89.6	96.8	98.3
macaroni1	56.8	58.7	77.7	70.9	-	52.3	73.0	86.0	81.3
macaroni2	65.0	65.8	63.3	63.2	-	62.2	72.2	72.1	67.7
pcb1	56.9	48.4	81.8	57.2	-	36.0	82.4	87.0	74.8
pcb2	63.2	59.8	50.4	73.8	-	47.3	64.6	64.3	68.1
pcb3	53.0	47.6	70.4	70.7	-	54.8	71.1	70.0	75.9
pcb4	88.0	90.6	81.5	95.1	-	66.3	94.0	94.4	95.9
pipe_fryum	94.6	93.7	82.1	94.8	-	89.7	95.1	96.3	98.7
MEAN	**71.4**	**70.2**	**81.4**	**81.4**	**81.7**	**68.1**	**82.3**	**85.4**	**86.8**

Table 13. Breakdown AUROC results of pixel-level ZSAD performance comparison on VisA.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
candle	33.6	50.0	88.9	97.8	-	16.3	97.9	98.8	98.9
capsules	56.8	61.5	81.6	97.5	-	47.5	89.7	95.0	96.3
cashew	64.5	62.5	84.7	86.0	-	32.5	85.8	93.8	95.3
chewinggum	43.0	56.5	93.3	99.5	-	3.4	98.5	99.3	99.3
fryum	45.6	62.7	88.5	92.0	-	21.7	93.3	94.6	94.4
macaroni1	20.3	22.9	70.9	98.8	-	36.8	98.6	98.3	98.2
macaroni2	37.7	28.8	59.3	97.8	-	27.5	99.0	97.6	96.8
pcb1	57.8	51.6	61.2	92.7	-	19.8	90.4	94.1	96.0
pcb2	34.7	38.4	71.6	89.7	-	22.9	89.3	92.4	92.7
pcb3	54.6	44.6	85.3	88.4	-	18.0	91.3	88.4	88.2
pcb4	52.1	49.9	94.4	94.6	-	14.0	93.6	95.7	97.1
pipe_fryum	58.7	44.7	75.4	96.0	-	29.2	96.1	98.2	98.1
MEAN	**46.6**	**47.8**	**79.6**	**94.2**	**89.9**	**24.1**	**93.6**	**95.5**	**95.9**

Table 14. Breakdown PRO results of pixel-level ZSAD performance comparison on VisA.

Data Subset	Handcrafted Text Prompting					Learnable Text Prompting			
	CLIP	CLIP-AC	WinCLIP	APRIL-GAN	AnoVL	CoOp	CoCoOp	AnomalyCLIP	FAPrompt
candle	3.6	6.0	83.5	92.5	-	1.1	92.4	96.2	96.7
capsules	15.8	22.4	35.3	86.7	-	18.4	72.8	78.5	84.6
cashew	9.6	10.9	76.4	91.7	-	1.7	93.6	91.6	91.8
chewinggum	17.8	30.2	70.4	87.3	-	0.1	86.1	91.2	93.2
fryum	12.1	29.3	77.4	89.7	-	2.6	91.3	86.8	88.1
macaroni1	8.1	13.4	34.3	93.2	-	18.1	93.9	89.8	91.1
macaroni2	20.9	18.4	21.4	82.3	-	2.7	89.5	84.2	80.9
pcb1	11.7	12.5	26.3	87.5	-	0.1	82.1	81.7	85.3
pcb2	12.8	13.9	37.2	75.6	-	0.7	72.9	78.9	73.7
pcb3	31.7	23.6	56.1	77.8	-	0.0	84.6	77.1	78.4
pcb4	17.1	20.3	80.4	86.8	-	0.0	84.8	91.3	91.3
pipe_fryum	16.7	6.0	82.3	90.9	-	0.6	96.2	96.8	96.8
MEAN	**14.8**	**17.2**	**56.8**	**86.8**	**71.2**	**3.8**	**86.7**	**87.0**	**87.7**

Table 15. Dataset-specific image-level ZSAD results (AUROC, AP) of our ablation study.

Data type	Dataset	Base	CAP	CAP w\o $\mathcal{L}_{oc}$	DAP	DAP w\o $\mathcal{L}_{prior}$	FAPrompt
Object	VisA	(82.1, 85.4)	(83.8, 86.7)	(83.8, 86.7)	(82.7, 85.0)	(81.0, 83.3)	(84.6, 86.8)
	BTAD	(88.3, 87.3)	(91.5, 92.4)	(90.8, 91.1)	(90.7, 90.7)	(91.0, 89.3)	(92.2, 92.5)
	MPDD	(77.0, 82.0)	(78.7, 81.3)	(77.9, 81.3)	(74.6, 78.3)	(73.4, 77.8)	(80.1, 83.9)
	SDD	(98.1, 93.4)	(98.6, 96.1)	(98.0, 95.8)	(98.1, 95.5)	(98.3, 95.3)	(98.4, 95.6)
Textual	AITEX	(62.2, 40.4)	(72.8, 55.8)	(72.7, 75.4)	(73.6, 54.1)	(75.9, 57.8)	(74.1, 55.5)
	DAGM	(97.5, 92.3)	(97.9, 93.0)	(97.9, 93.0)	(96.5, 88.2)	(95.7, 89.6)	(98.8, 95.3)
	DTD-Synthetic	(93.5, 97.0)	(96.3, 98.5)	(95.7, 93.9)	(96.0, 98.0)	(96.3, 98.1)	(96.2, 98.1)
	ELPV	(81.5, 91.3)	(84.8, 92.6)	(80.8, 90.7)	(83.0, 91.6)	(80.6, 89.9)	(83.7, 92.1)
Medical	BrainMRI	(90.3, 92.2)	(95.2, 95.2)	(95.0, 94.6)	(95.9, 96.0)	(95.9, 96.5)	(95.8, 96.2)
	HeadCT	(93.4, 91.6)	(94.7, 94.6)	(93.7, 90.4)	(92.3, 90.4)	(92.0, 91.0)	(94.0, 92.4)
	LAG	(74.3, 84.9)	(75.2, 85.4)	(75.2, 85.4)	(75.2, 85.5)	(74.5, 84.6)	(76.6, 86.1)
	Br35H	(94.6, 94.7)	(97.4, 97.1)	(97.1, 96.8)	(97.3, 97.1)	(97.0, 96.9)	(97.6, 97.1)

Table 16. Dataset-specific pixel-level ZSAD results (AUROC, PRO) of our ablation study.

Data type	Dataset	Base	CAP	CAP w\o $\mathcal{L}_{oc}$	DAP	DAP w\o $\mathcal{L}_{prior}$	FAPrompt
Object	VisA	(95.5, 87.0)	(95.1, 85.1)	(95.1, 85.0)	(95.8, 86.1)	(95.6, 85.1)	(95.9, 87.5)
	BTAD	(94.2, 74.8)	(94.4, 70.5)	(94.4, 70.5)	(95.4, 73.7)	(95.5, 75.2)	(95.6, 75.1)
	MPDD	(96.5, 87.0)	(95.9, 86.2)	(95.9, 86.2)	(95.8, 86.4)	(95.5, 85.4)	(96.5, 87.9)
	SDD	(98.1, 95.2)	(98.3, 93.8)	(98.3, 93.2)	(97.9, 95.6)	(97.7, 92.5)	(98.3, 94.1)
Textual	AITEX	(83.0, 66.5)	(82.3, 64.5)	(81.3, 61.9)	(82.4, 65.2)	(82.0, 62.1)	(82.0, 66.2)
	DAGM	(95.6, 91.0)	(98.1, 95.2)	(97.5, 95.2)	(98.5, 96.0)	(98.2, 94.4)	(98.2, 95.0)
	DTD-Synthetic	(97.9, 92.3)	(97.9, 92.3)	(97.9, 92.3)	(98.1, 91.4)	(98.1, 91.3)	(98.3, 93.3)
Medical	CVC-ColonDB	(81.9, 71.3)	(83.7, 72.8)	(82.9, 68.1)	(83.8, 73.9)	(84.0, 73.0)	(85.0, 73.3)
	CVC-ClinicDB	(82.9, 67.8)	(83.2, 67.8)	(83.4, 72.9)	(83.6, 68.4)	(83.3, 68.3)	(84.7, 70.1)
	Kvasir	(78.9, 45.6)	(78.8, 48.1)	(78.5, 48.0)	(79.3, 45.5)	(79.0, 45.3)	(82.1, 49.9)
	Endo	(84.1, 63.6)	(84.3, 63.4)	(84.1, 63.4)	(84.7, 63.8)	(84.8, 64.2)	(86.8, 67.6)
	ISIC	(89.7, 78.4)	(88.7, 78.0)	(88.1, 76.8)	(91.0, 80.9)	(91.4, 81.3)	(91.1, 81.6)
	TN3K	(81.5, 50.4)	(84.2, 52.7)	(84.5, 53.4)	(84.9, 56.0)	(84.2, 53.5)	(84.7, 54.6)

Serge J. Belongie, Bharath Hariharan, and Ser Nam Lim. Visual prompt tuning. *ArXiv*, abs/2203.12119, 2022. 2

[18] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19113–19122, 2022. 2

[19] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[20] Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. Attention based glaucoma detection: A large-scale database and cnn model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2019. 1

[21] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 1

[22] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 3

[23] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14902–14912, 2021. 1

[24] Javier Silvestre-Blanes, Teresa Albero-Albero, Ignacio Miralles, Rubén Pérez-Llorens, and Jorge Moreno. A public fabric database for defect detection methods and results. *Autex Research Journal*, 19(4):363–374, 2019. 1

[25] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel

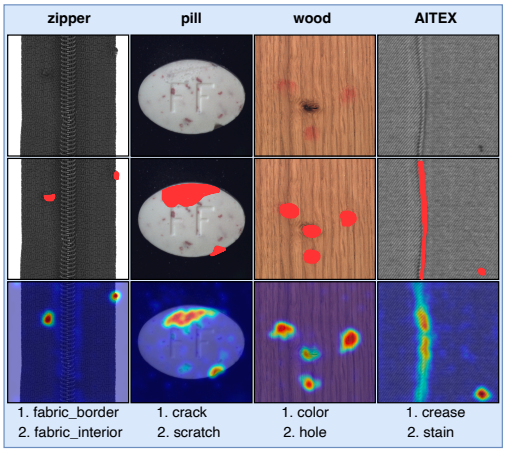

Figure 5. Visualization of anomaly maps of `FAPrompt` on samples containing multiple anomalous types in a single image.

Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, 2020. 1

[26] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 1

[27] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22883–22892, 2024. 6

[28] Matthias Wieler and Tobias Hahn. Weakly supervised learning for industrial optical inspection. In *DAGM symposium in*, page 11, 2007. 1

[29] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 3

[30] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2

[31] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[32] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[33] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408, 2022. 1

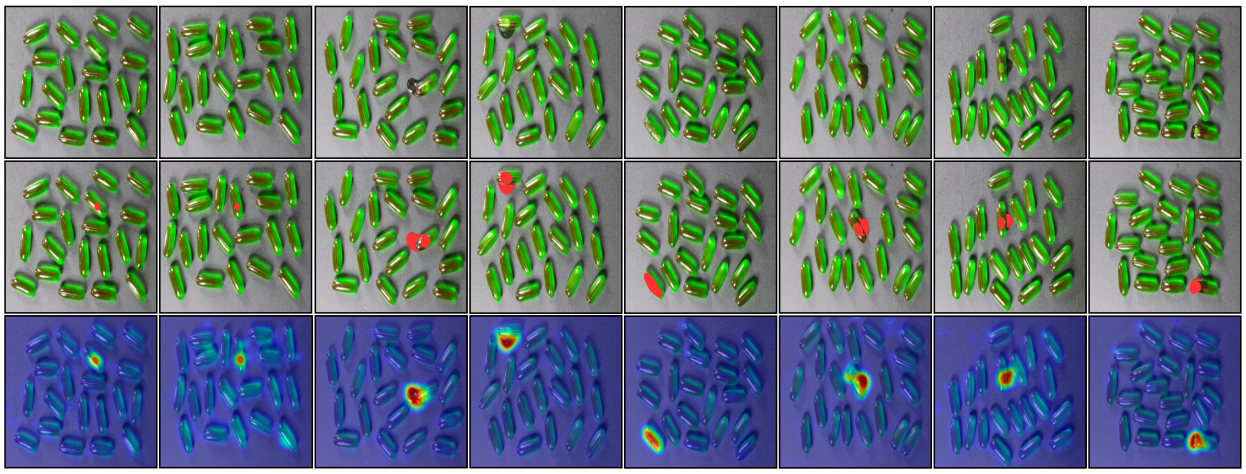

Figure 6. Anomaly maps generated by FAPrompt for the capsules category in VisA. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

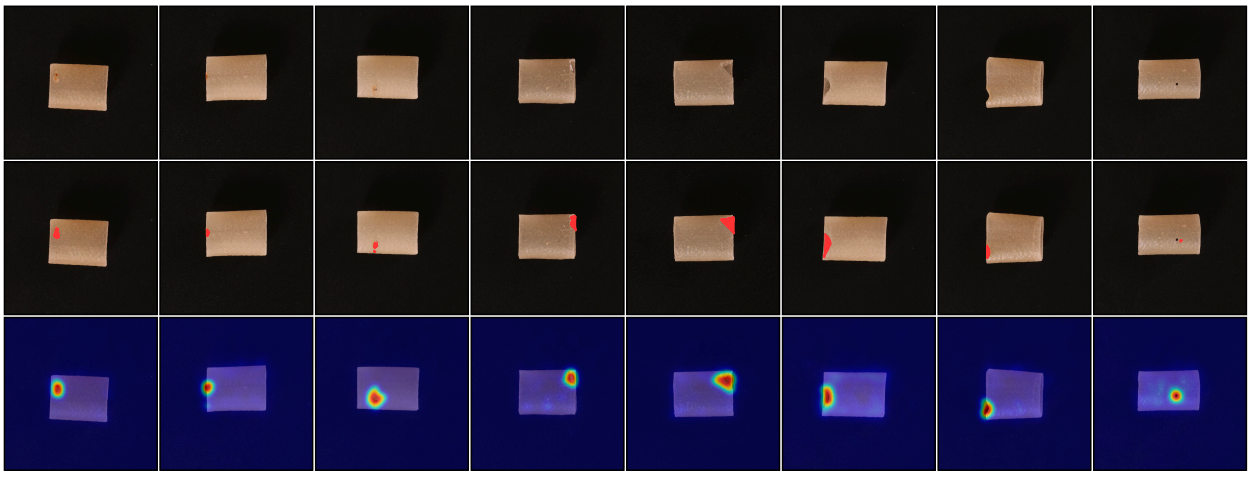

Figure 7. Anomaly maps generated by FAPrompt for the pipe_fryum category in VisA. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

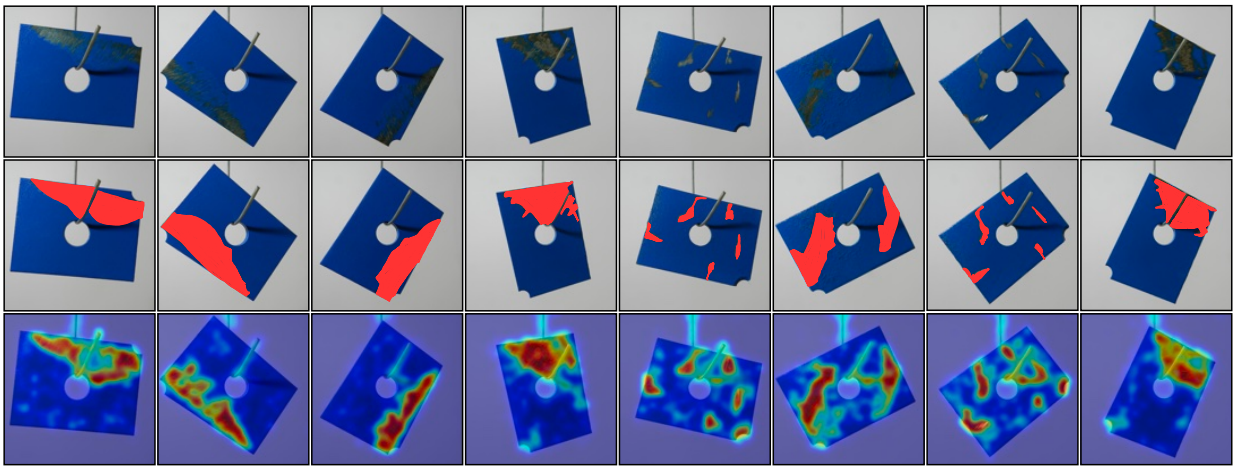

Figure 8. Anomaly maps generated by FAPrompt for the metal_plate category in MPDD. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

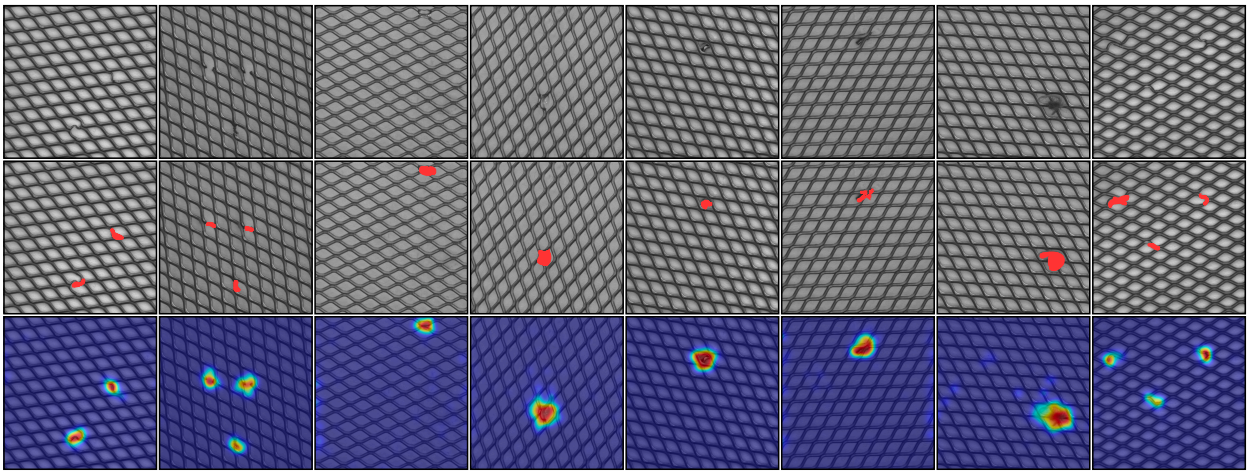

Figure 9. Anomaly maps generated by FAPrompt for grid category in MVTecAD. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

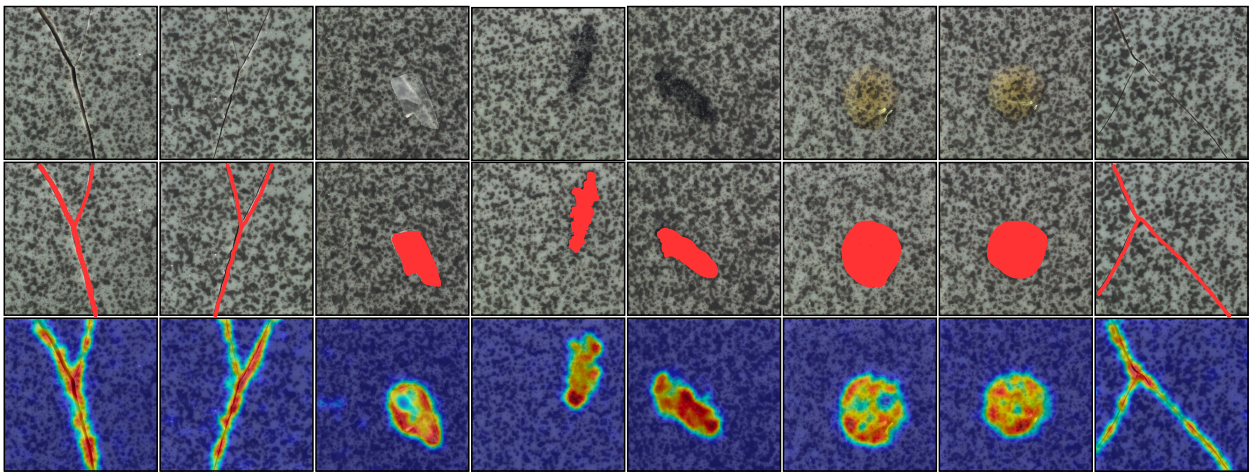

Figure 10. Anomaly maps generated by FAPrompt for tile category in MVTecAD. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

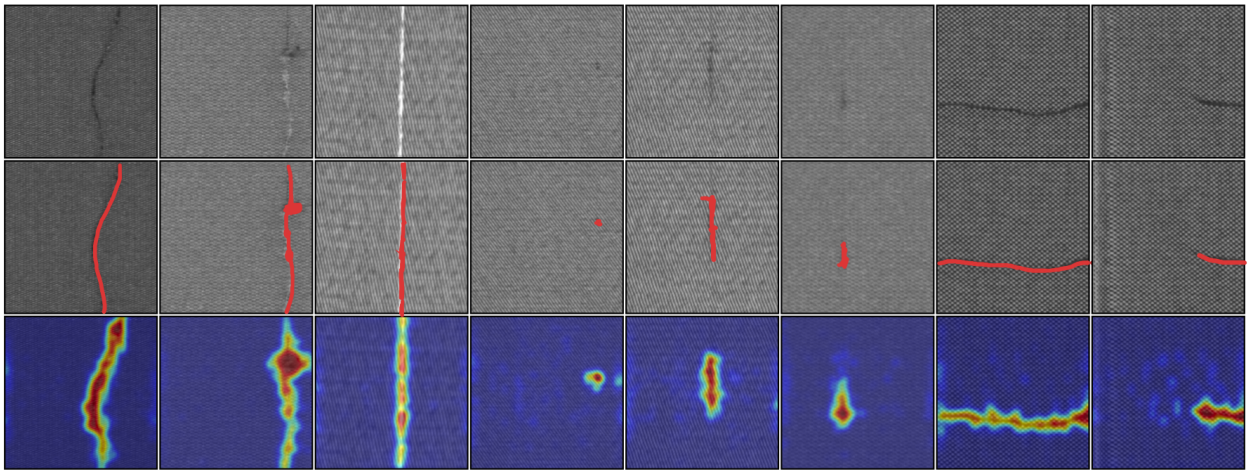

Figure 11. Anomaly maps generated by FAPrompt for AITEX. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

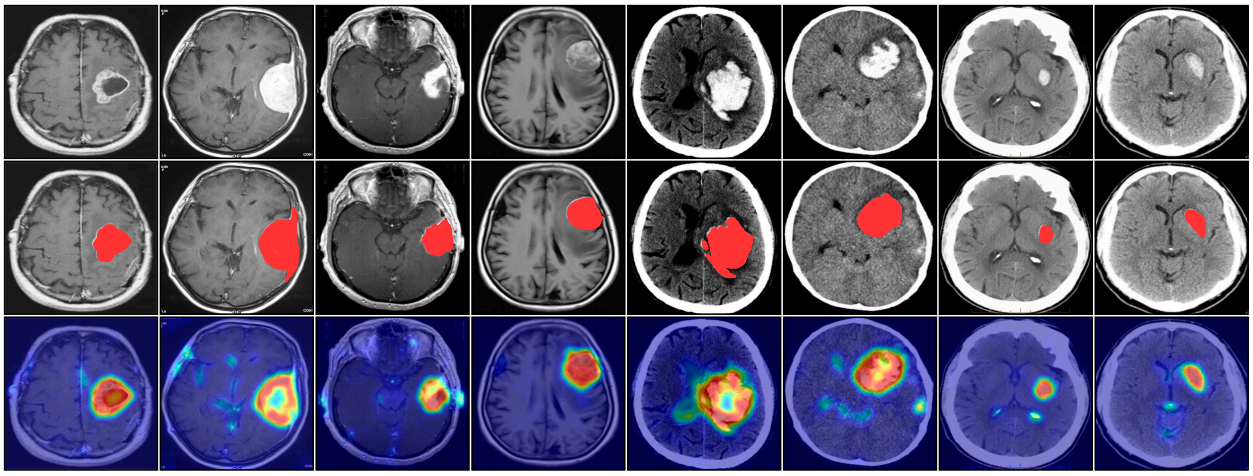

Figure 12. Anomaly maps generated by FAPrompt for brain-related anomalies. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

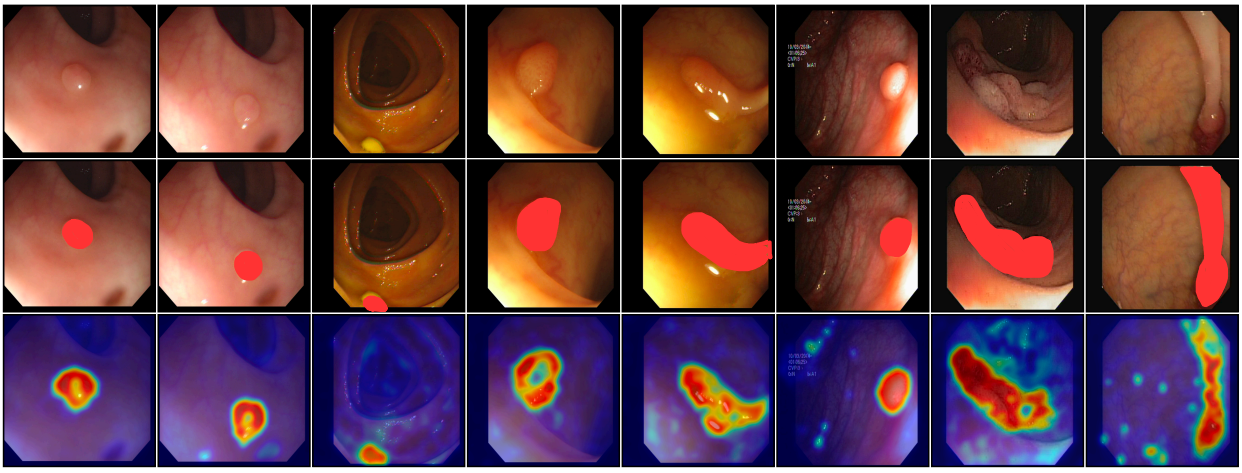

Figure 13. Anomaly maps generated by FAPrompt for colon-related anomalies. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

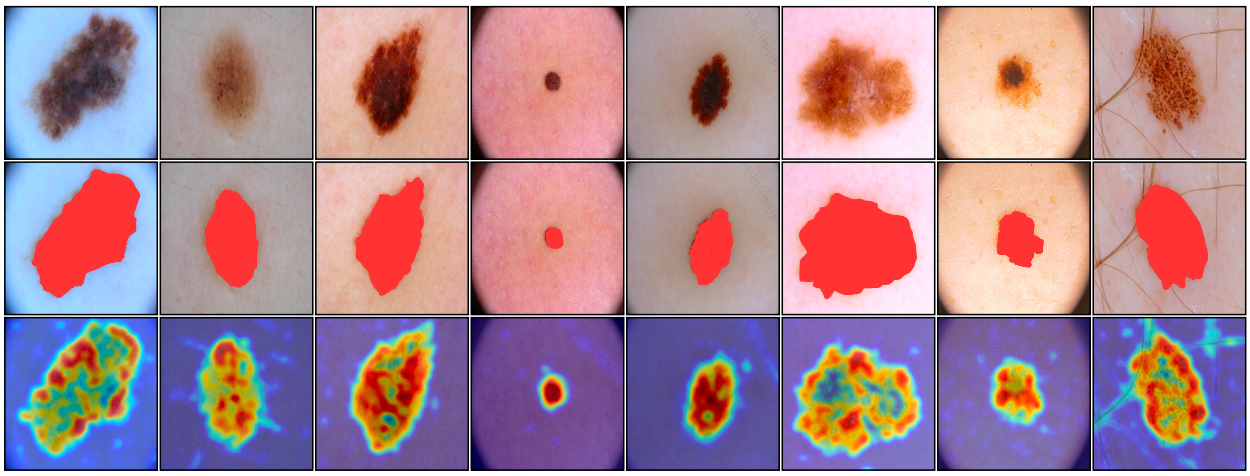

Figure 14. Anomaly maps generated by FAPrompt for skin-related anomalies. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

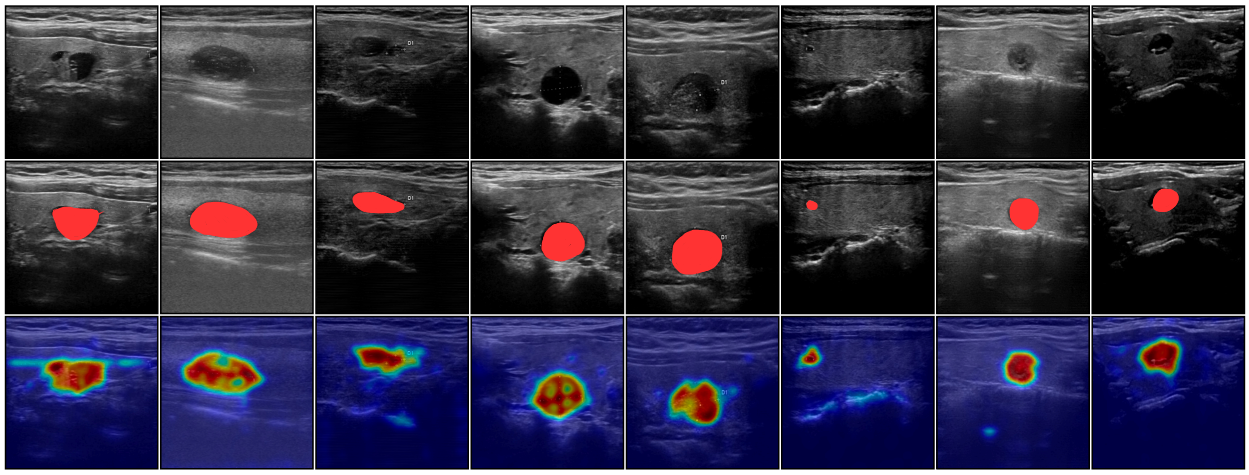

Figure 15. Anomaly maps generated by FAPrompt for thyroid-related anomalies. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.

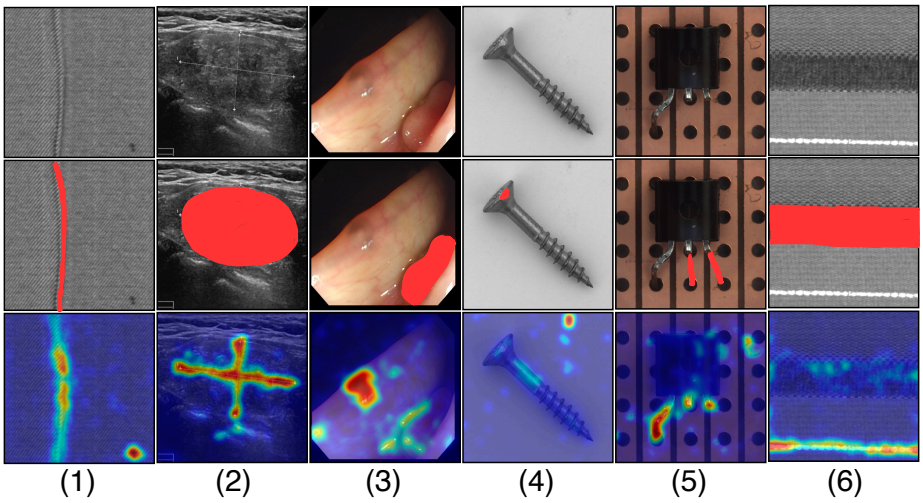

Figure 16. Failure cases of FAPrompt. The first row represents the input images, while the second row displays the ground truth of anomalous regions. The bottom row illustrates the segmentation results from FAPrompt.