AdaptCLIP: Adapting CLIP for Universal Visual Anomaly Detection

Association for the Advancement of Artificial Intelligence

Bin-Bin Gao*, Yue Zhou*, Jiangtao Yan, Yuezhi Cai,
Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, Chengjie Wang

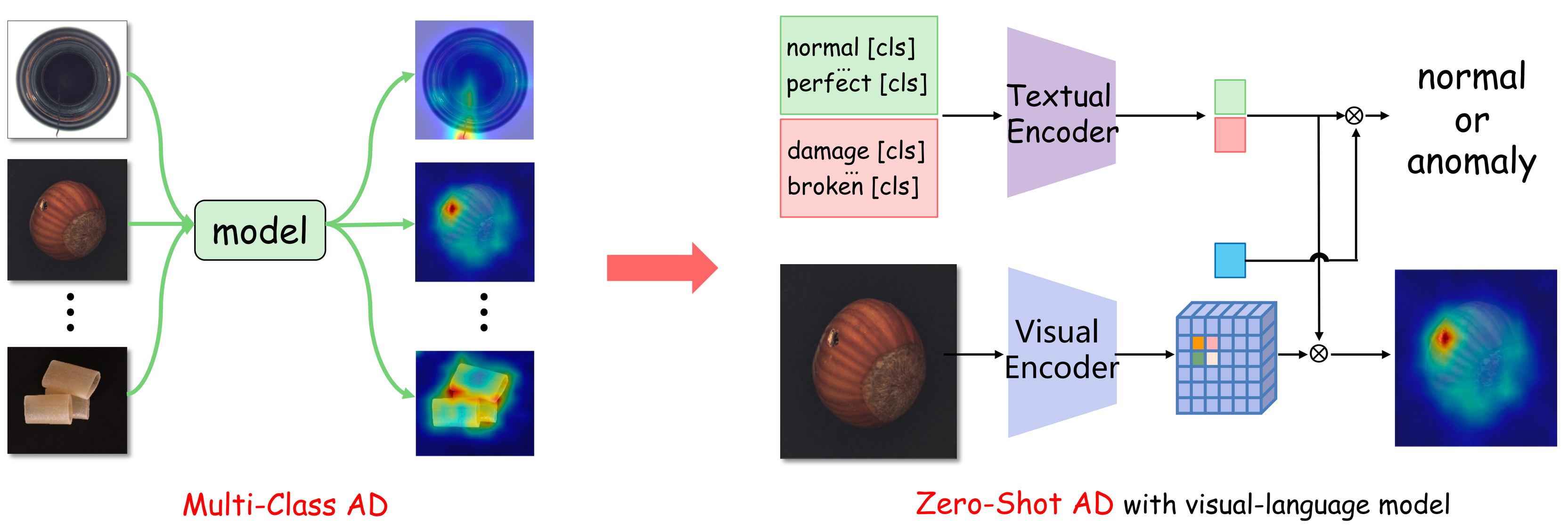
Tencent YouTu Lab

SIEMENS

AdaptCLIP

1. Motivation

Visual Anomaly Detection: from multi-class to any-class



Challenges:

✓restricted to zero-shot or few-shot
only zero-shot: AnomalyCLIP and AdaCLIP
only few-shot: InCtrl, PromptAD and MetaUAS

✓ hurt original ability of CLIP
concatenate learnable tokens to intermediate layers of CLIP, such as AnomalyCLIP, AdaCLIP

✓ require fine-tuning or heavy computation
such as PromptAD and WinCLIP

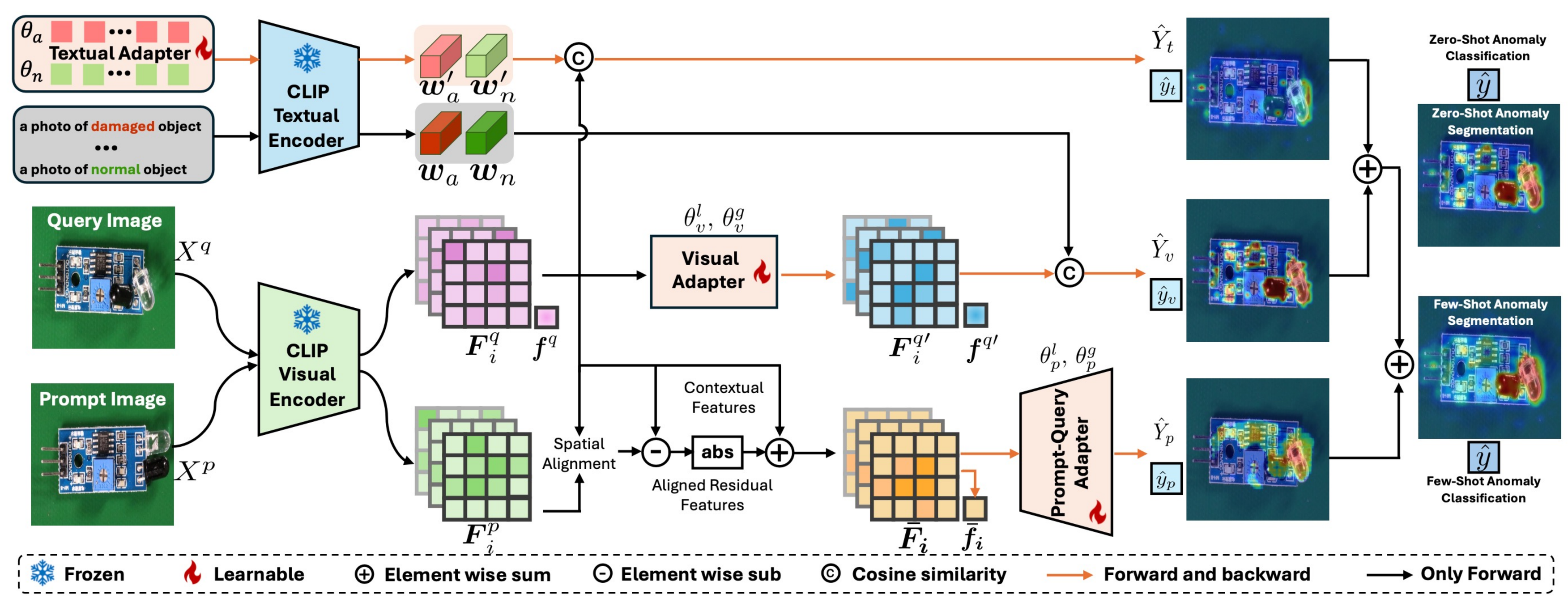
State of the art methods

Methods	ZS	FS	OA	w/o FT
WinCLIP	✓	✓	✓	✓
AdaCLIP	✓	✗	✗	✓
InCtrl	✗	✓	✓	✓
AnomalyCLIP	✓	✗	✗	✓
PromptAD	✗	✓	✓	✗
MetaUAS	✗	✓	✓	✓
AdaptCLIP	✓	✓	✓	✓

We want to explore a universal AD (both zero-shot and few-shot) model, aiming to detect any anomalies without any dataset-specific fine-tuning

2. Our Method

2.1 AdaptCLIP Framework



The philosophy of AdaptCLIP is that "less and simpler could be better", and it contains two key insights based on three simple adapters.

Our Contributions:

- ✓ We propose a simple but effective universal visual anomaly detection framework (AdaptCLIP) based on visual-language CLIP, which is capable of detecting any visual anomalies with a training-free manner on target domains.
- ✓ We find that adaptive visual and textual representations should be learned alternately rather than jointly, using separate visual and text adapters.
- ✓ We also find prompt-query comparative learning should incorporate contextual and aligned residual features rather than relying solely on residual features.
- ✓ AdaptCLIP outperforms zero- and few-shot AD methods on 8 industrial and 4 medical benchmarks. Meanwhile, AdaptCLIP possesses simpler adapters, fewer parameters, and competitive efficiency.

2.2 Insight 1: Alternating Learning (zero-shot)

✓ fixing two-class textual embeddings and learning visual tokens with a visual adapter.

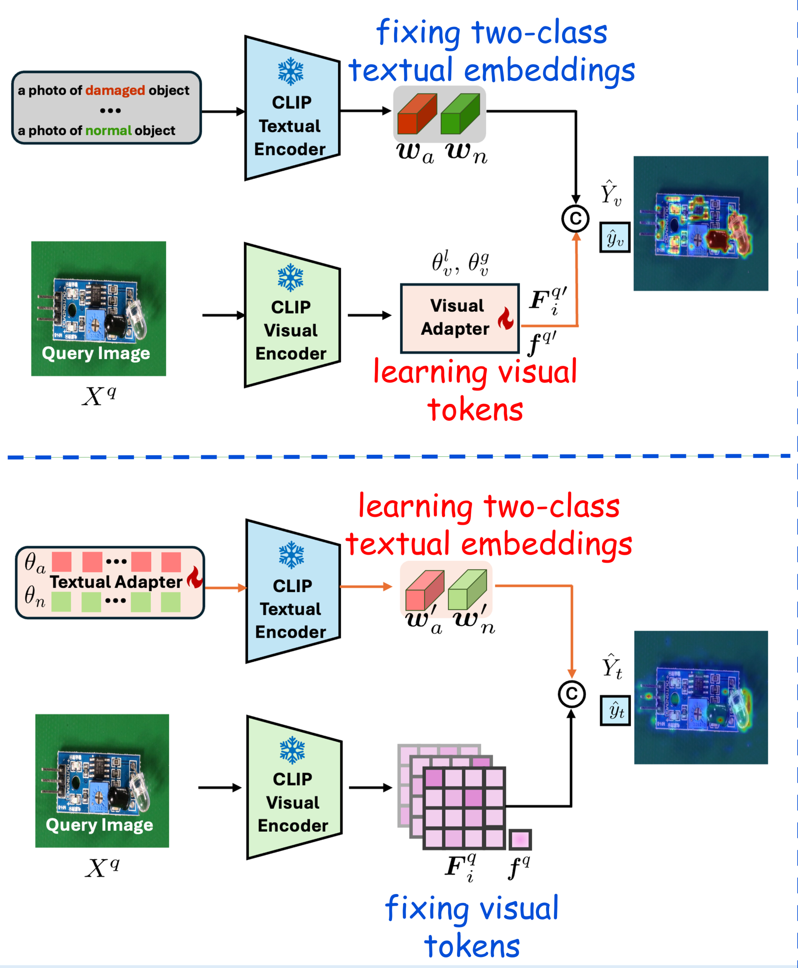
$$F_i^{q'} = F_i^q + \mathrm{MLP}(F_i^q;\theta_v^l);\ f^{q'} = f^q + \mathrm{MLP}(f^q;\theta_v^g)$$

simple residual multi-layer perceptron

✓ fixing visual tokens and learning two-class textual prompt embeddings with a textual adapter.

$$w'_a = \mathcal{T}(\theta_a), w'_n = \mathcal{T}(\theta_n)$$

two-class prompts embeddings



2.3 Insight 2: Comparative Learning (few-shot)

It is intuitive to use a normal image as a visual prompt for anomaly detection.

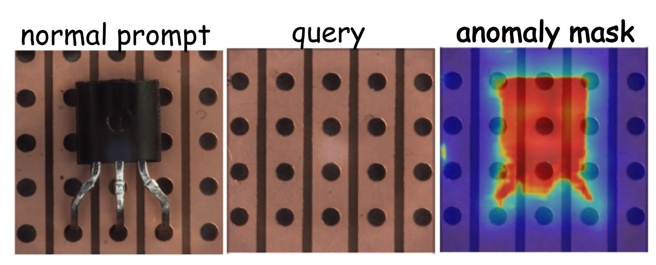
Step1: spatial alignment;

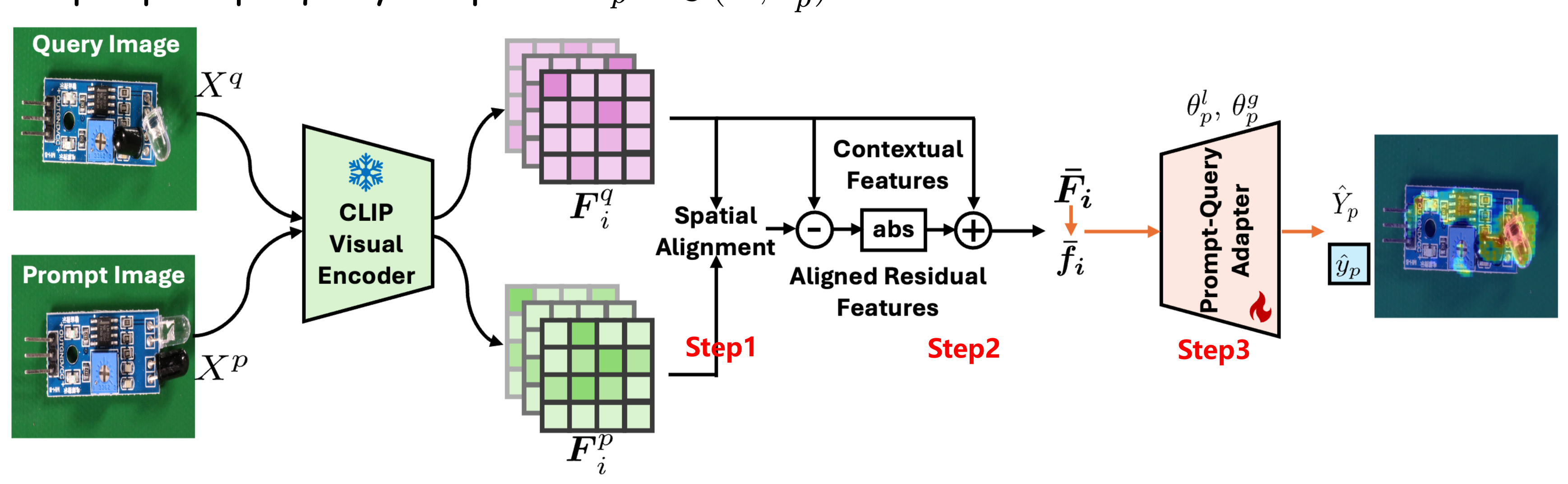
$$F_i^{p'} = F_{k}^p, k = \arg\min_j \|F_i^q - F_j^p\|_2.$$

Step2: joint contextual and aligned residual feature

$$\bar{F}_i = F_i^q + |F_i^q - F_i^{p'}|.$$

Step3: prompt-query adapter $\hat{Y}_p = \mathcal{G}(\bar{F};\theta_p^l)$





3. Experiments

3.1 Comparisons with Zero-/Few-Shot Methods

Anomaly classification

Shots	Methods	Industrial MVTec	VisA	BTAD	MVTec3D	DTD	KSDD	MPDD	Real-IAD	AVG	Medical Br35H	Covid	AVG
0	WinCLIP (Jeong et al. 2023)	90.4	75.5	68.2	69.4	95.1	92.9	61.5	67.0	77.5	80.5	66.4	73.5
	AdaCLIP† (Cao et al. 2024)	90.7	81.7	89.9	76.2	92.7	96.6	64.0	73.3	83.1	96.7	69.4	83.0
	AnomalyCLIP (Zhou et al. 2024)	91.6	82.0	88.3	73.9	93.9	97.8	77.5	69.5	84.3	94.2	77.7	86.0
	AdaptCLIP-Zero	93.5	84.8	91.0	78.6	96.0	98.1	73.6	74.2	86.2	94.8	86.5	90.7
1	WinCLIP+ (Jeong et al. 2023)	93.6±0.4	80.0±2.4	84.4±1.5	74.1±0.4	97.9±0.2	93.8±0.4	69.3±2.9	74.7±0.2	83.4	80.1±2.1	90.1±3.6	85.1
	InCtrl (Zhu and Pang 2024)	91.3±0.4	83.2±2.4	88.5±0.4	75.3±1.3	97.9±0.3	92.0±0.9	73.0±2.7	76.6±0.0	84.7	83.9±6.4	89.2±5.3	86.6
	AnomalyCLIP+ (Zhou et al. 2024)	95.2±0.2	86.1±0.7	88.5±0.8	76.7±2.1	98.0±0.2	97.5±0.3	83.4±2.6	78.2±0.0	88.0	90.8±5.1	87.3±2.6	89.1
	AdaptCLIP	94.5±0.5	90.5±1.2	93.4±0.0	81.7±1.5	98.0±0.0	96.9±0.3	83.8±2.2	81.8±0.3	90.1	93.7±2.4	91.8±2.5	92.8
2	WinCLIP+ (Jeong et al. 2023)	94.5±1.0	82.7±1.0	85.8±1.8	74.3±0.3	98.1±0.2	93.8±0.2	69.3±2.3	76.1±0.1	84.3	81.6±0.6	91.8±2.5	86.7
	InCtrl (Zhu and Pang 2024)	91.8±0.9	86.3±1.4	86.2±2.0	75.4±0.5	98.3±0.2	91.6±0.9	74.2±1.8	78.5±0.0	85.3	86.1±1.7	89.7±5.1	87.9
	AnomalyCLIP+ (Zhou et al. 2024)	95.4±0.1	87.8±0.5	89.2±1.1	78.3±1.3	98.2±0.1	97.9±0.2	83.4±1.5	78.3±0.0	88.6	91.5±4.0	89.3±2.7	90.4
	AdaptCLIP	95.7±0.6	92.2±0.8	93.4±0.2	82.9±1.1	98.3±0.0	97.2±0.0	84.4±0.7	82.9±0.2	90.8	94.0±1.7	94.9±0.9	94.5
4	WinCLIP+ (Jeong et al. 2023)	95.3±0.1	84.3±0.6	87.8±0.8	75.7±0.3	98.2±0.0	94.0±0.2	71.2±1.6	77.0±0.0	85.4	82.3±0.4	92.9±2.1	87.6
	InCtrl (Zhu and Pang 2024)	93.1±0.7	87.8±0.2	67.5±2.4	78.1±1.1	97.7±0.1	91.6±0.9	78.6±2.3	81.8±0.0	84.5	89.1±1.2	91.4±4.1	90.3
	AnomalyCLIP+ (Zhou et al. 2024)	96.1±0.1	88.8±0.5	90.5±1.2	79.2±1.3	98.4±0.1	97.8±0.1	86.3±1.8	78.4±0.0	89.4	91.1±4.4	91.4±3.0	91.3
	AdaptCLIP	96.6±0.3	93.1±0.2	93.3±0.3	84.2±0.6	98.5±0.1	97.0±0.2	86.8±1.1	83.9±0.2	91.7	93.7±2.0	95.8±0.9	94.8

Anomaly segmentation

Shots	Methods	Industrial MVTec	VisA	BTAD	MVTec3D	DTD	KSDD	MPDD	Real-IAD	AVG	Medical Kvasir	Endo	AVG
0	WinCLIP (Jeong et al. 2023)	18.2	5.4	12.9	5.3	9.8	7.1	14.1	3.3	9.5	27.8	23.8	25.8
	AdaCLIP† (Cao et al. 2024)	39.1	31.0	42.9	37.5	75.2	48.2	25.9	30.5	41.3	36.6	43.7	40.1
	AnomalyCLIP (Zhou et al. 2024)	34.5	21.3	45.5	30.5	62.6	51.9	28.9	26.7	37.7	39.6	46.6	43.1
	AdaptCLIP-Zero	38.3	26.1	41.8	31.4	68.7	58.3	25.3	28.2	39.7	45.3	52.0	48.7
1	WinCLIP+ (Jeong et al. 2023)	38.3±0.8	15.8±0.2	41.3±2.6	18.4±1.1	47.8±0.9	19.2±0.3	29.8±2.0	13.9±0.2	28.1	27.6±2.9	23.6±0.1	25.6
	InCtrl (Zhu and Pang 2024)	47.8±1.1	17.7±0.6	44.1±1.4	18.7±0.5	64.3±0.5	26.7±0.7	27.9±2.2	19.1±0.0	33.3	22.1±1.7	20.3±3.7	21.2
	AnomalyCLIP+ (Zhou et al. 2024)	40.8±0.1	24.8±0.9	41.3±1.1	30.6±1.1	67.4±0.4	47.5±0.5	34.2±0.8	27.9±0.0	39.3	46.9±3.9	47.8±4.9	47.4
	AdaptCLIP	53.7±0.9	38.9±0.3	60.6±1.0	40.7±0.6	76.9±0.1	57.8±1.2	33.5±2.5	36.6±0.1	49.8	49.2±4.7	52.4±4.7	50.8
2	WinCLIP+ (Jeong et al. 2023)	39.5±0.6	17.2±0.8	42.8±1.3	19.1±0.8	48.2±0.9	19.0±0.5	30.7±1.1	14.8±0.1	28.9	29.1±0.2	27.6±2.3	28.4
	InCtrl (Zhu and Pang 2024)	49.2±0.7	18.5±0.2	44.2±0.8	20.3±0.6	64.4±0.4	26.4±2.5	29.2±1.3	20.1±0.0	34.0	24.9±1.9	24.5±7.5	24.7
	AnomalyCLIP+ (Zhou et al. 2024)	41.5±0.1	26.2±0.7	41.9±0.6	32.4±1.5	68.1±0.2	47.6±0.4	35.3±1.1	28.1±0.0	40.1	47.3±2.9	49.6±4.8	48.5
	AdaptCLIP	55.1±0.5	40.7±0.6	61.0±0.6	42.3±1.1	77.4±0.2	57.5±1.1	35.0±0.7	37.8±0.1	50.9	49.0±4.1	53.1±4.2	51.1
4	WinCLIP+ (Jeong et al. 2023)	41.2±0.9	18.1±1.3	44.0±0.4	19.9±0.6	49.3±0.1	19.1±0.7	32.0±0.2	15.4±0.2	29.9	29.6±0.8	27.7±0.5	28.7
	InCtrl (Zhu and Pang 2024)	50.9±0.3	19.2±0.6	44.0±0.2	22.2±1.2	64.9±0.3	26.0±1.4	31.4±0.8	21.0±0.0	35.0	24.7±1.6	22.3±1.0	23.5
	AnomalyCLIP+ (Zhou et al. 2024)	42.4±0.0	27.5±1.1	45.8±3.0	33.4±1.3	68.5±0.2	46.4±0.7	36.8±1.0	28.2±0.0	41.1	45.9±1.5	49.2±3.4	47.6
	AdaptCLIP	57.2±0.8	41.8±0.6	62.3±0.3	44.5±0.3	78.2±0.2	56.4±1.4	37.4±1.1	39.1±0.3	52.1	47.5±2.7	52.2±3.1	49.9

3.2 Complexity and Efficiency Comparisons

Shots	Methods	CLIP Models	Input Size	# Params (M)	Inf.Time (ms)
0	WinCLIP (Jeong et al. 2023)	ViT-B-16+240	240×240	208.4 + 0.0	201.3
		ViT-B-16+240	512×512	208.4 + 0.0	3912.6
	AdaCLIP (Cao et al. 2024)	ViT-L/14@336px	518×518	428.8 + 10.7	212.0
	AnomalyCLIP (Zhou et al. 2024)	ViT-L/14@336px	518×518	427.9 + 5.6	154.9
	AdaptCLIP-Zero	ViT-B-16+240	512×512	208.4 + 0.4	49.9
		ViT-L/14@336px	518×518	427.9 + 0.6	162.2
1	WinCLIP+ (Jeong et al. 2023)	ViT-B-16+240	240×240	208.4 + 0.0	339.5
		ViT-B-16+240	512×512	208.4 + 0.0	7434.9
	InCtrl (Zhu and Pang 2024)	ViT-B-16+240	240×240	208.4 + 0.3	337.0
	AnomalyCLIP+ (Zhou et al. 2024)	ViT-L/14@336px	518×518	427.9 + 5.6	158.6
	AdaptCLIP	ViT-B-16+240	512×512	208.4 + 1.4	54.0
		ViT-L/14@336px	518×518	427.9 + 1.8	168.2

✓ strong generalization;
from industrial to medial

✓ fewer learnable parameters;
zero-shot: 0.6M, few-shot: 1.8M

✓ support zero-shot and few-shot;

✓ training-free on target domains;

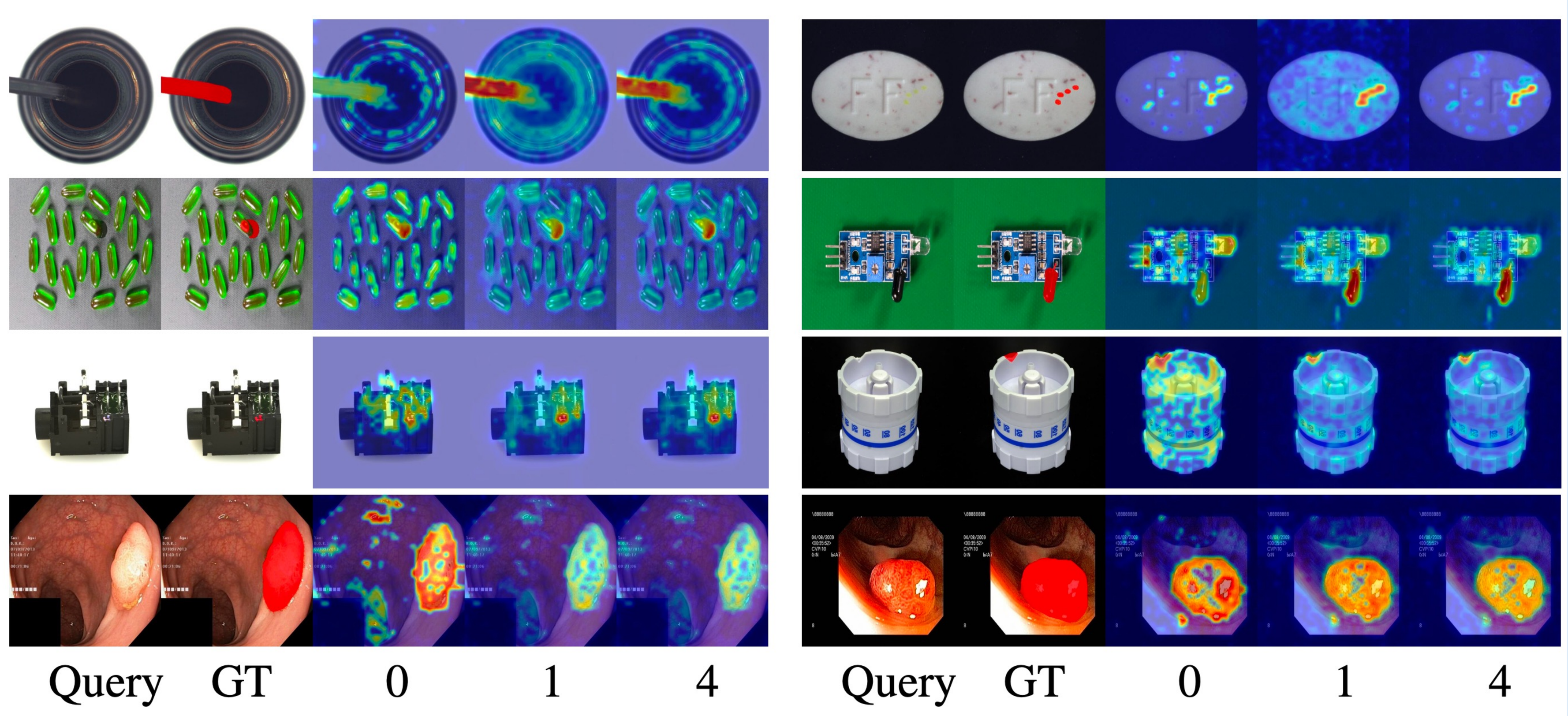
3.3 Ablation Studies

No.	Methods	Shots	TA	VA	PQA	MVTec	VisA
0	baselines	0	✗	✗	✗	91.1 / 33.0	82.1 / 18.0
1		0	✓	✗	✗	92.2 / 31.4	82.9 / 19.7
2		0	✗	✓	✗	90.5 / 39.4	81.0 / 22.1
3	joint	0	✓	✓	✗	89.3 / 36.2	81.6 / 21.5
4	alternating	0	✓	✓	✗	93.5 / 38.3	84.8 / 26.1
5	w/o context	1	✗	✗	✓	62.6 / 7.0	85.3 / 28.7
6	w context	1	✗	✗	✓	88.1 / 50.2	88.9 / 38.1
7	AdaptCLIP	1	✓	✓	✓	94.2 / 52.5	92.0 / 38.8

✓ Alternating learning helps us fully utilize the prior knowledge of CLIP and thus improve cross-domain generalization (Line 4 vs. 3)

✓ The joint of contextual and aligned residual features performs better than residual features alone (Line 6 vs. Line 5);

3.4 Qualitative Comparisons



Generally, few-shot normal image prompts help AdaptCLIP segment anomalies more accurately and produce fewer false positives than in a zero-shot manner.