TIPS Over Tricks: Simple Prompts For Effective Zero-Shot Anomaly Detection

A. Salehi[1], E. Karami[1], Se. Noey[3], Sa. Noey[3], M. Yamada[2], R. Hosseini[1], M. Sabokrou[2*]

[1] University of Tehran, [2] Okinawa Institute of Science and Technology, [3] Amirkabir University of Technology

***mohammad.sabokrou@oist.jp**

PROBLEM & MOTIVATION

- **Anomaly detection:** industrial inspection and medical imaging.
- Traditional methods require **normal** data, which is often not available (Privacy) or is expensive
- **Zero-shot anomaly detection (ZSAD)** solves this by using vision-language models (VLMs) like **CLIP**.

⚠ CLIP LIMITATION:

🎯 Poor spatial alignment

🔍 Limited anomaly sensitivity

🧩 Requires complex add-ons (adapters, attention tricks, …)

✓TIPS Solution:

✓ Better backbone

✓ Spatial alignment

✓ Simple prompts

KEY CONTRIBUTIONS

🧠 **Rethink the Backbone**
Use TIPS instead of relying on CLIP fixes.

💬 **Decoupled Prompting**
Separate prompts for detection and localization.

🔗 **Global–Local Fusion**
Combine image-level and pixel-level evidence.

⚡ **Simple Zero-Shot Pipeline**
Lightweight design with no complex add-ons.

QUANTITATIVE RESULTS

Zero-shot performance

Setting	Metrics	Ours	Gain vs. Best Prior
Industrial Image-level	AUROC / AP / F1-max	**92.5 / 93.1 / 90.7**	**+2.3 / +3.9 / +1.1**
Industrial Pixel-level	AUROC / AUPRO / F1-max	**96.5 / 90.2 / 51.7**	**+2.0 / +6.9 / +1.5**
Medical Pixel-level	AUROC / AUPRO / F1-max	**87.2 / 71.4 / 54.8**	**+3.2 / +4.4 / +5.3**

📈 **Strong Performance Gains**
Both anomaly detection and localization

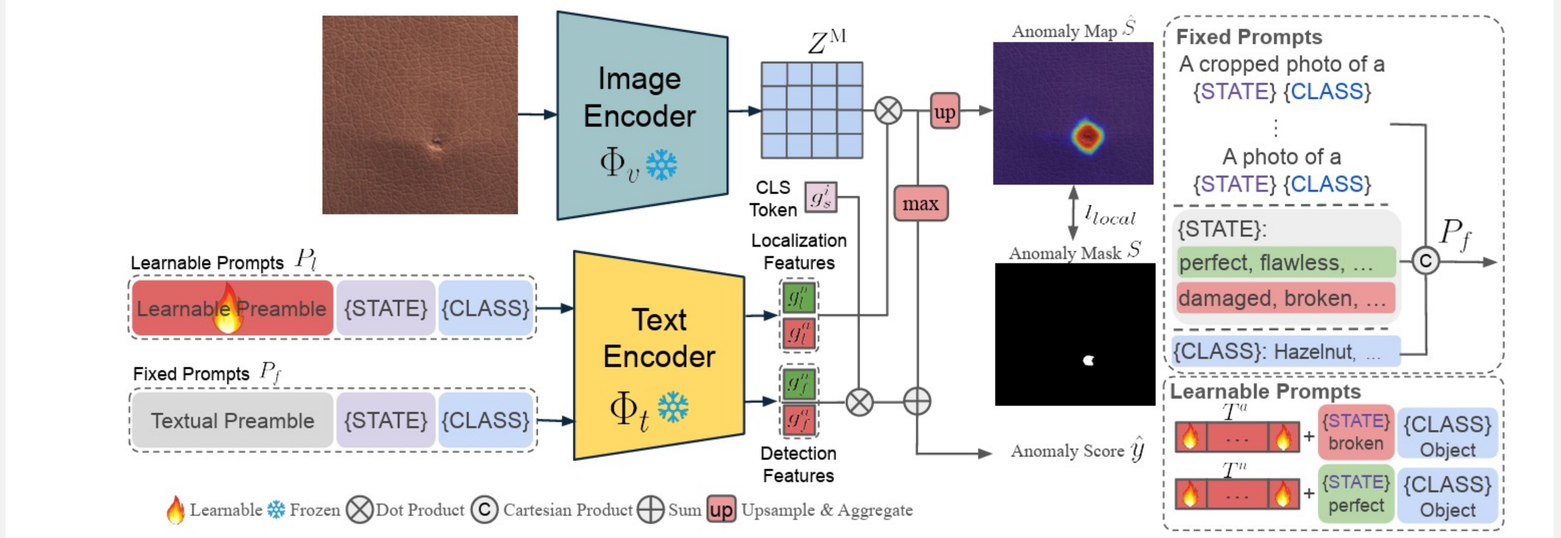
Average improvement over CLIP-based baseline

ARCHITECTURE

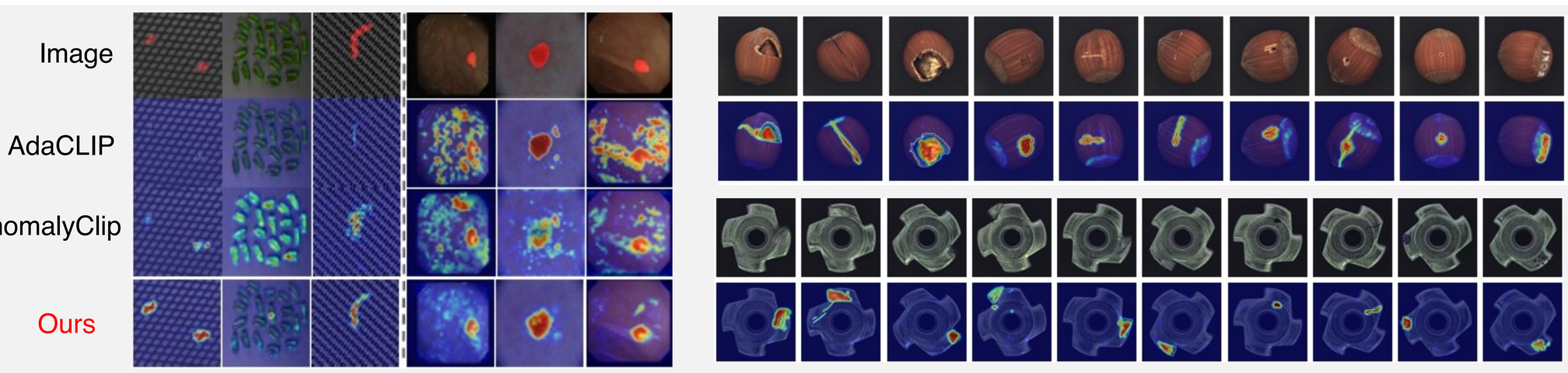
1. **Inputs**:
   - Image → ViT → global (g_s^i , g_o^i) + patch **features** (Z^M)
   - Prompts → Transformer → textual **prototypes**
2. **Scoring**:
   - Fixed prototypes (g_f) → image-level scoring
   - Learnable prototypes (g_l) → pixel-level anomaly map
   - Visual embeddings (e) → $e \in \{g_s^i , g_o^i , Z^M\}$

$$p(e, G_{(\cdot)}) = \mathrm{softmax}\left(\frac{G_{(\cdot)}^{\top} e}{\tau}\right), \quad p_a = [p(e, G_{(\cdot)})]_a$$

3. **Combine**: global score + max(local anomaly)
4. **Output**:
   - anomaly score + anomaly map



QUALITATIVE RESULTS



☞**TIPS produces sharper localization and cleaner anomaly maps across industrial and medical domains.**

CONCLUSION

- Using a better **backbone (TIPS)** is more effective than patching CLIP.
- Identified a **key issue**:
  - ⚠ Global–local feature gap
- Solved with:
  - **Decoupled prompting**
  - **Local-global score fusion**

A simple, **spatially-aware** design can outperform complex CLIP-based adaptations in zero-shot anomaly detection.

PAPER & CODE

arXiv Paper

GitHub Repo