HYPER-FSAD: TRAINING-FREE AND LANGUAGE-FREE FEW-SHOT ANOMALY DETECTION VIA SPARSE HYPER MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot anomaly detection (FSAD) is particularly valuable when only a few normal images are available in a new target domain, while anomalous cases are rare, diverse, and difficult to enumerate in advance. However, existing methods often still require task-specific fitting or language prompts, and their patch-level retrieval commonly relies on brittle nearest-neighbor or fixed Top-p rules. We propose **Hyper-FSAD**, a training-free and language-free framework that performs support-only inference with a frozen visual encoder. We formulate FSAD as support-only scoring in frozen feature space and establish the selective stability of sparse retrieval and sufficient conditions for normal–anomaly separation. Guided by this analysis, **Sparse Hyper Matching** uses *sparsemax* to adaptively select support patches for each query patch, exactly suppressing below-threshold distractors without a manually specified retrieval hyperparameter. **Dual-Branch Image Scoring** further combines local reconstruction discrepancies with support-conditioned global <CLS> deviation. Across four industrial and two medical benchmarks, Hyper-FSAD achieves the best overall performance across the 1/2/4-shot settings, while requiring only 52.6 ms per image and 0.89 GB GPU memory. The code will be released.

1 INTRODUCTION

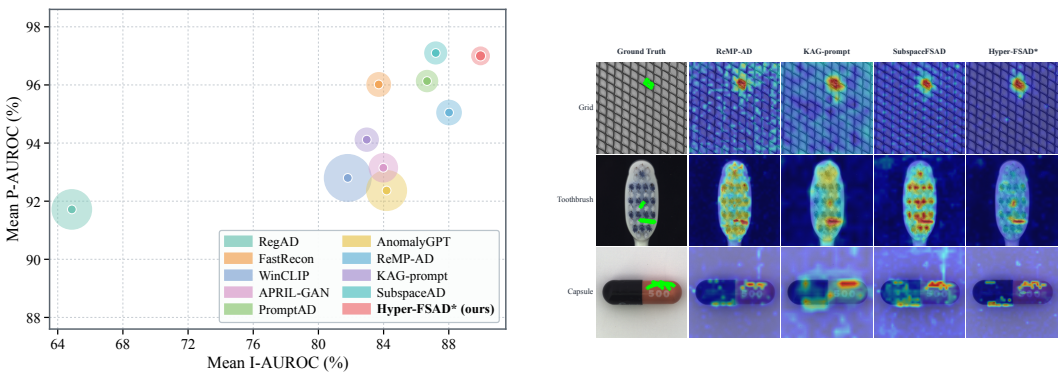
Few-shot anomaly detection (FSAD) aims to detect anomalies in images and identify potentially anomalous regions with only a few normal samples available for support. It is particularly important in real-world applications where representative target-domain data are scarce and anomalous cases are rare, diverse, and difficult to collect or enumerate in advance, such as medical image analysis Mahapatra et al. (2021); Menze et al. (2014) and industrial defect detection Qu et al. (2023); Chen et al. (2025).

Existing FSAD methods generally follow two paradigms: source-domain training and target-domain training. Source-domain training typically trains a model on a dataset's training set, potentially augmented with synthetic anomalies, and then transfers it to new datasets for testing Tao et al. (2025); Ma et al. (2025); Huang et al. (2022); Zhou et al. (2024); Chen et al. (2025). In target-domain training, a prompt or adapter is learned on the target dataset using only normal samples before testing Li et al. (2024); Fang et al. (2023); Jeong et al. (2023). However, both paradigms remain constrained in real-world deployment because they rely on training or adaptation under specific data distributions.

In practice, target-domain distributions may evolve continuously and are difficult to model explicitly. For example, lighting fluctuations or material batch differences can introduce implicit distribution shifts Guo et al. (2024); Goodarzi et al. (2025); Wang et al. (2025), causing trained models to degrade during deployment. Moreover, data privacy and cross-domain sharing restrictions can make repeated retraining or domain adaptation difficult to implement. These challenges limit the robustness and deployability of training-dependent FSAD paradigms.

To improve generalization under limited target-domain data, recent FSAD methods have increasingly leveraged the language-derived semantic priors of large-scale vision-language models, particularly CLIP Radford et al. (2021), demonstrating impressive performance across various domains

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: **(a)** One-shot accuracy–efficiency comparison of Hyper-FSAD with representative FSAD methods across six datasets. Each point reports the mean image- and pixel-level AUROC, while bubble area is log-scaled by per-image runtime (smaller is faster). Hyper-FSAD* denotes our DINOv3 ViT-B/16 instantiation. **(b)** Qualitative comparison of 1-shot anomaly maps.

Jeong et al. (2023); Cao et al. (2024); Gu et al. (2024); Qu et al. (2024); Zhou et al. (2024); Li et al. (2024); Tao et al. (2025); Ma et al. (2025). However, these methods often rely on complex prompt engineering or additional training strategies, which still limit their flexibility and deployability. Moreover, their dependence on textual descriptions restricts their applicability to purely visual encoders such as DINOv3 Siméoni et al. (2025). This raises an important question: Do we really need language supervision for FSAD?

Beyond their reliance on training and language supervision, existing FSAD methods also face limitations in patch-level visual matching. Most methods match each query patch to its nearest normal support patch and convert the similarity into an anomaly score Jeong et al. (2023); Li et al. (2024); Tao et al. (2025). However, under limited normal references, such one-to-one matching is easily affected by background clutter, noise, and local appearance variations, leading to accidental matches or mismatches. This issue is especially evident in medical and industrial images, where anomalous regions may resemble normal tissues or materials, while normal regions can exhibit large intra-class variations. Some methods therefore adopt Top-p matching to aggregate multiple support patches Ma et al. (2025). Yet, its performance depends heavily on the manually specified p: a small p provides insufficient normal evidence, whereas a large p may introduce irrelevant patches and dilute anomaly cues. These limitations motivate a data-adaptive matching mechanism that selects relevant normal evidence without a fixed size.

To address these challenges, we take a theory-guided approach. We formulate FSAD as a generic support-only scoring problem in frozen feature space and derive sufficient normal-coverage and anomaly-margin conditions for separation, an exponential few-shot coverage bound, and selective-stability guarantees for retrieval as the memory grows. These results lead to **Hyper-FSAD**, a training-free and language-free framework that requires no target-task optimization, text input, or prompting, and is compatible with frozen visual backbones, including the CLIP image encoder Radford et al. (2021), DINOv2 Oquab et al. (2023), and DINOv3 Siméoni et al. (2025). Here, *Hyper* denotes a one-to-set matching structure: instead of pairing a query patch with a single support patch, each query patch and an adaptively selected set of normal support patches jointly form a higher-order matching unit. We instantiate this view as **Sparse Hyper Matching**, where *sparsemax* Martins & Astudillo (2016); Peters et al. (2019) assigns nonzero weights only to a sparse, similarity-dependent subset of relevant support patches—without fixing its size—and reconstructs the query patch from this subset; all below-threshold support patches receive exactly zero weight. Our analysis proves that adding any number of such below-threshold distractors leaves the reconstruction unchanged and that the retrieval weights are non-expansive, while the coverage–margin conditions provide a sufficient guarantee for training-free normal–anomaly separation. Finally, **Dual-Branch Image Scoring** combines a spatial branch that summarizes patch reconstruction discrepancies with a support-aware <CLS> branch that captures global semantic deviation, producing the final anomaly score using only visual evidence.

Our main contributions are summarized as follows:

- We formulate FSAD as a generic support-only scoring problem in frozen feature space. We prove sufficient normal-coverage and anomaly-margin conditions for training-free visual separation, and quantify few-shot global coverage as a function of K.
- The analysis identifies selective, simplex-valued retrieval as a key requirement. We prove exact below-threshold distractor rejection and non-expansive weights for Euclidean simplex projection, which directly motivates our **Sparse Hyper Matching**.
- Guided by the two additive separation margins, we propose **Hyper-FSAD**, combining sparse spatial reconstruction with support-conditioned global <CLS> evidence in a training-free and language-free framework.

2 Related work

FSAD has evolved rapidly, driven by the need to handle open-world defects with minimal supervision. Early approaches like RegAD Huang et al. (2022) adopted registration-based paradigms to align test images with a normal reference, while others, such as FastRecon Fang et al. (2023), focused on rapid feature reconstruction to identify deviations. Recently, the success of Vision-Language Models (VLMs) has spurred a wave of methods like WinCLIP Jeong et al. (2023), AnomalyCLIP Zhou et al. (2024), PromptAD Li et al. (2024), and KAG-prompt Tao et al. (2025), which leverage CLIP's zero-shot capabilities by designing text prompts to distinguish normal from anomalous samples. ReMP-AD Ma et al. (2025) further enhances this by fusing multi-modal prompts with retrieval mechanisms. Despite their progress, these methods largely rely on either source-domain meta-training Huang et al. (2022) or target-domain training Li et al. (2024), and thus inherit the practical constraints of training in deployment, such as distribution drift, open-set defect emergence, limited compute budgets, and privacy restrictions that hinder frequent retraining or domain adaptation. Moreover, their dependence on text prompts often introduces a semantic gap when describing complex industrial defects. This issue is especially pronounced when defects are subtle, texture-like, or category-specific, where precise linguistic specification is difficult even for human experts. More recently, SubspaceAD Lendering et al. (2026) fits a per-category PCA subspace over frozen DINOv2 patch features and scores anomalies via the reconstruction residual. RAD Zhang et al. (2026) instead uses fixed-memory nearest-neighbor retrieval and analyzes the stability of its distance-to-memory score. Complementarily, Hyper-FSAD studies data-adaptive sparse reconstruction and dual-granularity scoring: it constructs matches directly from visual embeddings, without target-category parameter/subspace fitting or prompt engineering, and our analysis characterizes the resulting adaptive support and few-shot separation.

3 Theory-Guided Formulation

Before specifying Hyper-FSAD, we ask which properties are sufficient for FSAD using only frozen visual features and normal supports. Our analysis first treats retrieval as an unspecified support-only operator; the resulting guarantees then determine the concrete design of Hyper-FSAD.

3.1 A Generic Support-Only Scoring Family

Let a frozen encoder produce unit-normalized patch tokens $\mathbf{q}^l_{X,t}$ and a unit-normalized global token $\mathbf{c}^l_X$ for image X, patch t, and layer $l \in L$. The K normal support images induce patch and global memories $\mathcal{P}^l = \{\mathbf{p}^l_i\}_{i=1}^M$ and $\mathcal{C}^l = \{\mathbf{c}^l_k\}_{k=1}^K$, where $M = KN_p$. For similarities $z^l_{X,t,i} = \langle \mathbf{q}^l_{X,t}, \mathbf{p}^l_i \rangle$, consider any training-free rule $\rho^l : \mathbb{R}^M \to \Delta_M$, where $\Delta_M = \{\mathbf{w} \geq \mathbf{0} : \mathbf{1}^\top \mathbf{w} = 1\}$. It yields weights $\mathbf{w}^l_{X,t} = \rho^l(\mathbf{z}^l_{X,t})$ and normal evidence $\boldsymbol{\varepsilon}^l_{X,t} = \sum_i w^l_{X,t,i} \mathbf{p}^l_i$.

Dense weighting accumulates irrelevant evidence, whereas hard nearest-neighbor and Top-p rules predefine the retained neighborhood by a fixed cardinality or proportion and can switch discontinuously at rank boundaries. This motivates the following regularized simplex rule, which remains generic at this stage:

$$\rho_{\mathrm{proj}}(\mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{w} \in \Delta_M} \tfrac{1}{2} \|\mathbf{w} - \mathbf{z}\|_2^2. \tag{1}$$

Proposition 1 (Adaptive support and selective stability)**.** *For a nonempty $R \subseteq \{1, \dots, M\}$, let $\tau_R = (\sum_{i \in R} z_i - 1)/|R|$. If*

$$\max_{j \notin R} z_j \leq \tau_R < \min_{i \in R} z_i, \tag{2}$$

then Eq. (1) *selects exactly R, with $w_i = z_i - \tau_R$ on R. Appending any number of candidates with similarity at most τ_R leaves the original weights and reconstruction unchanged. Moreover, $\|\rho_{\mathrm{proj}}(\mathbf{z}) - \rho_{\mathrm{proj}}(\mathbf{z}')\|_2 \leq \|\mathbf{z} - \mathbf{z}'\|_2$.*

Proof sketch. The simplex KKT conditions give $w_i = [z_i - \tau]_+$ and $\sum_i [z_i - \tau]_+ = 1$. Under Eq. (2), τ_R activates exactly R; it remains a valid threshold after any below-threshold additions. Non-expansiveness follows because Euclidean projection onto a closed convex set is non-expansive. The full proof is in Appendix A.

3.2 Sufficient Conditions for Frozen-Feature Separation

For a nonzero reconstruction, define a generic local score, global score, and their convex fusion:

$$S_{\mathrm{map}}(X) = \max_t \frac{1}{|L|} \sum_{l \in L} \left(1 - \cos(\mathbf{q}^l_{X,t}, \boldsymbol{\varepsilon}^l_{X,t})\right), \tag{3}$$

$$S_{\mathrm{cls}}(X) = \frac{1}{|L|} \sum_{l \in L} \left(1 - \max_k \langle \mathbf{c}^l_X, \mathbf{c}^l_k \rangle\right), \tag{4}$$

$$S_{\mathrm{image}}(X) = \lambda S_{\mathrm{map}}(X) + (1 - \lambda) S_{\mathrm{cls}}(X). \tag{5}$$

Theorem 1 (Sufficient conditions for training-free visual separation)**.** *Consider a normal image N and an anomalous image A. At every $l \in L$, suppose: (i) $w^l_{N,t,i} > 0$ implies $\|\mathbf{q}^l_{N,t} - \mathbf{p}^l_i\|_2 \leq \epsilon_l < \sqrt{2}$; (ii) one patch $t^\star$ of A has a nonzero reconstruction and cosine similarity at most $1 - \gamma_l$ to every nonzero $\mathbf{v} \in \mathrm{conv}(\mathcal{P}^l)$; and (iii) $\min_k \|\mathbf{c}^l_N - \mathbf{c}^l_k\|_2 \leq r_l$ and $\min_k \|\mathbf{c}^l_A - \mathbf{c}^l_k\|_2 \geq \delta_l$. Then*

$$S_{\mathrm{image}}(A) - S_{\mathrm{image}}(N) \geq \lambda \Delta_{\mathrm{map}} + (1 - \lambda) \Delta_{\mathrm{cls}}, \tag{6}$$

where $\Delta_{\mathrm{map}} = |L|^{-1} \sum_l (\gamma_l - \epsilon_l^2/2)$ and $\Delta_{\mathrm{cls}} = (2|L|)^{-1} \sum_l (\delta_l^2 - r_l^2)$. Consequently, a positive right-hand side guarantees separation without target-task fitting or language input.

Proof sketch. Simplex weights make each reconstruction a convex combination. For a normal patch, unit normalization and (i) give cosine similarity at least $1 - \epsilon_l^2/2$; (ii) gives anomaly score at least γ_l. For global tokens, $1 - \langle \mathbf{c}, \mathbf{c}_k \rangle = \|\mathbf{c} - \mathbf{c}_k\|_2^2/2$. Averaging over layers, max pooling, and convex fusion yield Eq. (6); Appendix A gives the full proof.

The global coverage condition also exposes the role of shots. For a fixed normal token $\mathbf{c}^l_N$, if each of the K i.i.d. supports falls within radius r_l with probability at least α_l, then

$$\Pr\left[\min_k \|\mathbf{c}^l_N - \mathbf{c}^l_k\|_2 > r_l\right] \leq (1 - \alpha_l)^K \leq e^{-K\alpha_l}. \tag{7}$$

These are sufficient rather than distribution-free conditions; the analysis does not assume that every frozen encoder satisfies the required margins.

From theory to design. The analysis yields three requirements. First, local normal evidence should be a convex reconstruction from compatible support patches. Second, increasing the patch memory calls for exact rejection of irrelevant entries; Proposition 1 shows that the projection in Eq. (1), known as *sparsemax* Martins & Astudillo (2016), is an adaptive and stable choice. Third, the additive margins in Theorem 1 motivate combining local spatial evidence with support-conditioned global evidence. These conclusions lead directly to Hyper-FSAD.

4 Method

4.1 Framework Overview

Guided by Proposition 1 and Theorem 1, we instantiate the generic support-only family as **Hyper-FSAD**, illustrated in Fig. 2. Given K normal support images and a query image, a frozen visual

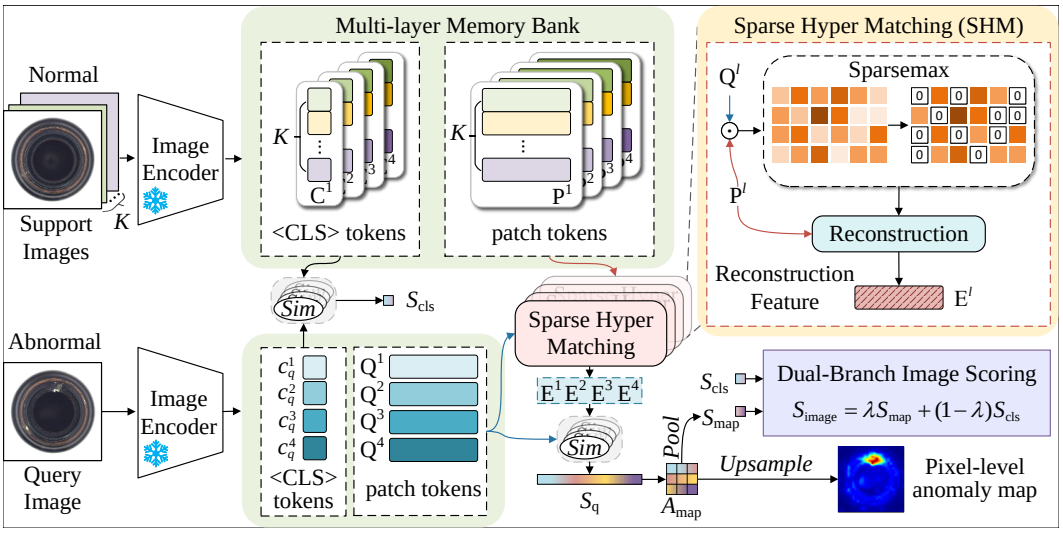


Figure 2: Overview of **Hyper-FSAD**. Given K normal support images and a query image, a frozen visual encoder extracts multi-layer patch tokens and <CLS> tokens to construct a *Multi-layer Memory Bank* of normal features. For each query patch at layer l, we perform *Sparse Hyper Matching*: similarities to all support patches in $\mathbf{P}^l$ are converted into sparse fusion weights via *sparsemax*, selecting a few informative neighbors to form a sparse reconstruction. The pixel-level anomaly map is obtained by the cosine distance between each query patch and its sparse reconstruction, aggregated across layers. *Dual-Branch Image Scoring* further combines a pooled map-based score S_{map} from the *Spatial Evidence Branch* and a <CLS>-based score S_{cls} from the *Global Semantic Branch* into the final image-level anomaly score S_{image}.

encoder extracts multi-layer patch and <CLS> tokens. We store both in a normal memory, instantiate ρ_{proj} as **Sparse Hyper Matching**, and implement the two margin terms in Theorem 1 through **Dual-Branch Image Scoring**. No parameters are fitted.

Every token entering a similarity computation is ℓ_2-normalized, so inner products are cosine similarities. Here, *language-free* means that target-task adaptation and inference require no text input, text encoder, prompt, or language supervision; the DINOv2/v3 instantiations additionally use visually pre-trained backbones.

4.1.1 Multi-layer Memory Bank.

We construct a multi-layer memory bank to store multi-layer features of normal samples, leveraging the strong representations of modern frozen visual encoders. Instead of relying on a single layer, which may miss multi-scale cues and thereby limit robustness, we extract features from four selected layers of the backbone, $L = \{l_1, l_2, l_3, l_4\}$, whose patch-token grid size and dimension are identical across layers. For each support image in the few-shot support set $\mathcal{S} = \{I_s^1, I_s^2, \ldots, I_s^K\}$, where K denotes the number of normal support images, we collect two distinct types of tokens: patch tokens, which preserve fine-grained local spatial information essential for pixel-level anomaly detection, and the class token <CLS>, which captures global semantic information suitable for image-level anomaly detection. These tokens are then stored in layer-specific memory banks to retain complementary cues across depths. Specifically, for each selected layer l, all patch tokens from the support images are concatenated to form $\mathbf{P}^l \in \mathbb{R}^{(K \cdot N_p) \times D}$, where N_p denotes the total number of patch tokens of the feature map at layer l and D is the token dimension. Meanwhile, we store the <CLS> tokens as $\mathbf{C}^l \in \mathbb{R}^{K \times D}$ to represent global normality. This structure allows our model to perform precise matching at both spatial and global granularities without the need for additional training or labor-intensive hand-crafted text prompts, thus improving practicality in deployment.

4.2 SPARSE HYPER MATCHING

Proposition 1 motivates instantiating the generic rule ρ^l with the projection ρ_{proj}, *i.e.*, *sparsemax*. For query patches $\mathbf{Q}^l \in \mathbb{R}^{N_p \times D}$, each row $\mathbf{q}^l$ is compared with the layer memory $\mathbf{P}^l$ to obtain $\mathbf{z}^l = \mathbf{q}^l (\mathbf{P}^l)^\top \in \mathbb{R}^{1 \times M}$. For completeness, the retrieval alternatives evaluated in our ablation are

$$\hat{\mathbf{w}}^l = \begin{cases} \mathrm{softmax}(\mathbf{z}^l) & \text{Dense Lookup,} \\ \mathrm{softmax}\big((\mathbf{z}^l)_{\text{top-}p}\big) & \text{Top-}p \text{ Lookup,} \\ \mathrm{sparsemax}(\mathbf{z}^l) & \text{Sparse Lookup.} \end{cases} \tag{8}$$

Dense Lookup weights every patch, whereas Top-p retains only a fixed percentage of patches. Sparse Hyper Matching instead uses the closed-form threshold $\hat{w}_u^l = [z_u^l - \tau]_+$ implied by Eq. (1), so its active set is data-dependent and below-threshold patches receive exact zero weight. The computation is detailed in Appendix A.5.

Finally, $\boldsymbol{\varepsilon}^l = \hat{\mathbf{w}}^l \mathbf{P}^l$ is the sparse normal reconstruction. Applying this rule to every query patch gives $\mathbf{E}^l \in \mathbb{R}^{N_p \times D}$.

4.3 DUAL-BRANCH IMAGE SCORING

The two additive margins in Theorem 1 motivate complementary local and global evidence. We implement them as a *spatial evidence branch* based on sparse patch reconstruction and a *global semantic branch* based on nearest-support <CLS> matching.

Spatial Evidence Branch. Given a query image I_q, we extract its patch tokens $\mathbf{Q}^l \in \mathbb{R}^{N_p \times D}$ and the corresponding sparse reconstructions $\mathbf{E}^l \in \mathbb{R}^{N_p \times D}$ at each selected layer $l \in L$ using Sparse Hyper Matching. We compute the patch-wise anomaly score vector by averaging the cosine distance across layers:

$$\mathbf{S}_q = \frac{1}{|L|} \sum_{l \in L} \big(\mathbf{1} - \cos\big(\mathbf{Q}^l, \mathbf{E}^l\big)\big) \in \mathbb{R}^{N_p}, \tag{9}$$

where $\cos(\mathbf{Q}^l, \mathbf{E}^l)$ denotes row-wise cosine similarity between corresponding rows of $\mathbf{Q}^l$ and $\mathbf{E}^l$. We reshape $\mathbf{S}_q$ into a patch-grid anomaly map $A_{map} \in \mathbb{R}^{H_p \times W_p}$ with $H_p W_p = N_p$.

We further reduce A_{map} to a scalar score through a pooling operator $f_{\mathrm{pool}}(\cdot)$, which can be instantiated as max, Top-n, or Top-p pooling; unless otherwise stated, we use max pooling in all experiments (see Appendix D), yielding map-based anomaly score $S_{\mathrm{map}} = f_{\mathrm{pool}}(A_{map})$.

Global Semantic Branch. For each layer, we match the query <CLS> token $\mathbf{c}_q^l$ to the K normal tokens in $\mathbf{C}^l$ and average their nearest-neighbor cosine distances, exactly implementing Eq. (4). Finally, Eq. (5) fuses S_{map} and S_{cls}; we use $\lambda = 0.5$ throughout.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

5.1.1 DATASETS.

Experiments are conducted on six real-world datasets spanning industrial inspection and medical imaging. We adopt four industrial anomaly detection benchmarks: MVTecAD Bergmann et al. (2019), VisA Zou et al. (2022), MPDD Jezek et al. (2021), and BTAD Mishra et al. (2021), together with two medical datasets, RESC Hu et al. (2019) for retinal lesion detection and BraTS Menze et al. (2014) for brain tumor analysis. Hyper-FSAD follows a strict *no-fitting* protocol: for each test category, we use K normal images as the support set and directly infer anomaly maps and image-level anomaly scores via memory-based matching, without auxiliary data, prompt/adaptor learning, or target-category parameter/subspace estimation. For fair comparison, we reproduce prior methods under their *original training settings* (source-domain or target-domain training, including prompt/adaptor learning when applicable), following the configurations reported in their papers.

Table 1: Comparison with existing state-of-the-art methods across six datasets under 1-shot/2-shot/4-shot settings. The best results are in **bold** and the second-best are underlined. **I** denotes I-AUROC (%) and **P** denotes P-AUROC (%). †, ‡, and ∗ denote our Hyper-FSAD using CLIP ViT-L/14@336px, DINOv2 ViT-B/14, and DINOv3 ViT-B/16 as the backbone, respectively. The ‡ on SubspaceAD indicates that it is likewise based on the DINOv2 ViT-B/14 backbone. Results for Hyper-FSAD (ours) are reported as mean ± standard deviation over 5 runs.

	Methods	Venue/Year	MVTecAD		VisA		MPDD		BTAD		RESC		BraTS	
			I	P	I	P	I	P	I	P	I	P	I	P
1-shot	RegAD	ECCV'22	73.3	92.3	69.3	93.3	47.9	93.2	84.4	95.6	55.9	84.6	58.4	91.3
	FastRecon	ICCV'23	92.0	95.1	81.0	96.6	76.5	96.2	93.7	96.8	82.8	96.0	76.2	95.4
	WinCLIP	CVPR'23	92.6	91.6	84.8	95.3	79.9	95.6	89.5	88.9	57.4	92.3	86.6	93.1
	APRIL-GAN	CVPR'23	91.1	91.2	87.1	95.9	75.1	94.9	86.5	93.0	77.3	93.0	86.8	90.9
	PromptAD	CVPR'24	93.0	95.2	85.2	97.2	79.3	96.0	93.4	96.1	87.4	**96.4**	81.7	95.9
	AnomalyGPT	AAAI'24	92.8	95.3	86.4	87.4	72.4	95.6	93.6	95.7	86.8	86.0	73.1	94.2
	ReMP-AD	ICCV'25	94.7	95.6	91.3	97.6	78.6	96.0	94.8	96.5	81.6	91.4	87.1	93.2
	KAG-prompt	AAAI'25	95.8	96.2	91.6	97.0	68.3	96.0	89.4	96.0	69.2	83.0	83.5	96.5
	SubspaceAD‡	CVPR'26	93.3	97.2	93.2	**98.0**	77.8	**97.7**	94.9	97.7	73.3	94.3	**90.7**	**97.7**
	Hyper-FSAD†	-	$93.1_{\pm 0.4}$	$93.4_{\pm 0.1}$	$92.4_{\pm 0.7}$	$95.0_{\pm 0.1}$	$82.8_{\pm 0.7}$	$96.5_{\pm 0.1}$	$87.6_{\pm 1.5}$	$96.9_{\pm 0.1}$	$89.5_{\pm 1.6}$	$90.8_{\pm 0.4}$	$89.9_{\pm 1.2}$	$89.3_{\pm 0.3}$
	Hyper-FSAD‡	-	$95.9_{\pm 0.1}$	$97.4_{\pm 0.0}$	$93.5_{\pm 0.1}$	$97.3_{\pm 0.1}$	$\mathbf{83.0}_{\pm \mathbf{0.9}}$	$97.2_{\pm 0.2}$	$\mathbf{95.4}_{\pm \mathbf{1.1}}$	$97.8_{\pm 0.0}$	$84.7_{\pm 0.1}$	$94.4_{\pm 0.2}$	$82.0_{\pm 1.7}$	$95.5_{\pm 0.2}$
	Hyper-FSAD∗	-	$\mathbf{96.3}_{\pm \mathbf{0.2}}$	$\mathbf{97.7}_{\pm \mathbf{0.0}}$	$\mathbf{94.0}_{\pm \mathbf{0.1}}$	$97.7_{\pm 0.1}$	$78.8_{\pm 1.5}$	$96.4_{\pm 0.1}$	$\mathbf{95.4}_{\pm \mathbf{1.2}}$	$\mathbf{97.9}_{\pm \mathbf{0.0}}$	$\mathbf{93.0}_{\pm \mathbf{1.4}}$	$96.3_{\pm 0.0}$	$82.2_{\pm 1.7}$	$96.0_{\pm 0.2}$
2-shot	RegAD	ECCV'22	76.6	94.5	70.4	94.3	52.5	94.0	88.9	96.9	59.4	85.9	57.4	92.7
	FastRecon	ICCV'23	94.2	95.5	81.1	96.6	81.8	96.5	93.8	97.2	87.6	96.2	75.8	95.2
	WinCLIP	CVPR'23	93.8	91.9	83.5	95.7	81.5	96.5	90.7	89.6	60.3	92.7	87.0	93.0
	APRIL-GAN	CVPR'23	90.1	91.6	86.6	96.1	75.1	95.1	86.1	93.2	78.3	93.4	87.5	90.9
	PromptAD	CVPR'24	95.4	95.6	85.1	97.7	83.3	96.8	92.7	96.4	89.2	**96.7**	83.0	95.8
	AnomalyGPT	AAAI'24	94.4	95.9	87.2	87.7	79.7	96.3	93.4	96.0	87.8	86.3	74.9	94.1
	ReMP-AD	ICCV'25	95.9	96.2	92.9	97.8	83.1	96.5	95.4	96.9	85.2	93.2	88.1	93.8
	KAG-prompt	AAAI'25	96.6	96.5	92.7	97.4	69.4	96.1	89.7	96.0	77.4	82.9	85.1	96.6
	SubspaceAD‡	CVPR'26	94.7	97.5	93.7	**98.3**	79.4	**97.7**	**96.9**	98.0	74.9	95.0	89.1	**98.0**
	Hyper-FSAD†	-	$93.6_{\pm 0.3}$	$94.2_{\pm 0.1}$	$94.5_{\pm 0.4}$	$95.7_{\pm 0.1}$	$\mathbf{83.9}_{\pm \mathbf{0.6}}$	$96.6_{\pm 0.1}$	$90.5_{\pm 0.7}$	$97.1_{\pm 0.0}$	$88.1_{\pm 0.7}$	$90.9_{\pm 0.2}$	$\mathbf{90.3}_{\pm \mathbf{1.4}}$	$91.2_{\pm 0.6}$
	Hyper-FSAD‡	-	$95.8_{\pm 0.2}$	$97.6_{\pm 0.1}$	$\mathbf{94.9}_{\pm \mathbf{0.1}}$	$97.8_{\pm 0.1}$	$82.2_{\pm 0.6}$	$97.1_{\pm 0.1}$	$96.4_{\pm 0.5}$	$\mathbf{98.1}_{\pm \mathbf{0.1}}$	$84.7_{\pm 0.1}$	$95.0_{\pm 0.2}$	$80.5_{\pm 1.7}$	$96.5_{\pm 0.4}$
	Hyper-FSAD∗	-	$\mathbf{96.8}_{\pm \mathbf{0.0}}$	$\mathbf{97.8}_{\pm \mathbf{0.1}}$	$94.8_{\pm 0.3}$	$98.0_{\pm 0.1}$	$80.3_{\pm 0.5}$	$96.4_{\pm 0.1}$	$95.8_{\pm 0.7}$	$98.0_{\pm 0.0}$	$\mathbf{91.5}_{\pm \mathbf{1.7}}$	$96.5_{\pm 0.2}$	$83.2_{\pm 1.5}$	$96.9_{\pm 0.3}$
4-shot	RegAD	ECCV'22	83.4	95.7	72.0	94.7	61.1	94.9	91.3	97.3	64.2	87.9	63.3	93.8
	FastRecon	ICCV'23	96.2	96.3	84.4	97.0	81.9	96.9	94.4	97.4	87.5	95.8	78.6	96.1
	WinCLIP	CVPR'23	95.5	92.4	85.7	96.0	84.0	97.0	91.7	90.3	63.8	93.1	87.0	93.1
	APRIL-GAN	CVPR'23	91.0	92.2	87.2	96.2	76.5	95.3	86.1	93.3	78.3	93.7	88.0	91.3
	PromptAD	CVPR'24	95.9	96.0	87.5	97.9	84.0	97.3	92.6	96.6	**90.2**	**96.8**	86.4	96.6
	AnomalyGPT	AAAI'24	97.0	96.4	91.4	96.5	**85.9**	96.7	93.5	96.2	88.5	86.7	79.4	95.4
	ReMP-AD	ICCV'25	96.8	96.6	94.1	97.8	80.0	97.0	96.0	97.0	86.6	93.6	88.2	94.7
	KAG-prompt	AAAI'25	**97.1**	96.7	93.3	97.7	71.6	96.6	88.4	96.0	82.0	84.0	87.5	96.7
	SubspaceAD‡	CVPR'26	95.5	97.7	94.9	**98.4**	79.8	**97.8**	**97.2**	98.1	82.4	95.7	88.1	**97.4**
	Hyper-FSAD†	-	$94.0_{\pm 0.2}$	$95.0_{\pm 0.1}$	$\mathbf{94.9}_{\pm \mathbf{0.2}}$	$96.0_{\pm 0.0}$	$84.2_{\pm 0.8}$	$96.9_{\pm 0.1}$	$91.0_{\pm 0.4}$	$97.5_{\pm 0.1}$	$88.7_{\pm 0.7}$	$91.2_{\pm 0.2}$	$\mathbf{90.1}_{\pm \mathbf{0.8}}$	$91.6_{\pm 0.6}$
	Hyper-FSAD‡	-	$96.1_{\pm 0.2}$	$97.8_{\pm 0.0}$	$\mathbf{95.5}_{\pm \mathbf{0.1}}$	$98.1_{\pm 0.0}$	$81.7_{\pm 0.5}$	$97.4_{\pm 0.1}$	$96.8_{\pm 0.5}$	$\mathbf{98.2}_{\pm \mathbf{0.1}}$	$84.6_{\pm 0.8}$	$95.0_{\pm 0.3}$	$78.5_{\pm 0.9}$	$96.0_{\pm 0.4}$
	Hyper-FSAD∗	-	$96.9_{\pm 0.1}$	$\mathbf{98.0}_{\pm \mathbf{0.0}}$	$94.8_{\pm 0.2}$	$98.3_{\pm 0.0}$	$80.3_{\pm 0.7}$	$96.5_{\pm 0.1}$	$96.2_{\pm 0.6}$	$\mathbf{98.2}_{\pm \mathbf{0.0}}$	$89.1_{\pm 1.2}$	$95.9_{\pm 0.2}$	$80.5_{\pm 1.1}$	$96.5_{\pm 0.6}$

5.1.2 Evaluation metrics.

Industrial visual anomaly detection is evaluated using the area under the receiver operating characteristic (AUROC), the F1-score at the optimal threshold (F1), average precision (AP), and per-region overlap (PRO). For image-level anomaly classification, we report I-AUROC, I-F1, and I-AP, while pixel-level anomaly segmentation is assessed using P-AUROC, PRO, and AP. Due to space constraints, we present only I-AUROC and P-AUROC in the main paper, and report results for the remaining metrics in Appendix B. We report performance with K set to 1, 2, and 4 normal support images.

5.1.3 Implementation details.

Since Hyper-FSAD is agnostic to the choice of frozen visual encoder, we instantiate it with three representative pre-trained backbones to demonstrate its generality: the visual encoder of CLIP ViT-L/14@336px Radford et al. (2021), DINOv2 ViT-B/14 Oquab et al. (2023), and DINOv3 ViT-B/16 Siméoni et al. (2025). Unless otherwise specified, we adopt dinov3_vitb16 as the default backbone. For the default backbone, all input images are resized to 448×448, and, following our multi-layer design, we extract features from transformer layers $L = \{3, 6, 9, 12\}$; the other two backbones follow the same four-layer multi-scale design at their standard input resolutions. For every backbone, we then construct layer-wise memory banks by storing support-set patch tokens and <CLS> tokens for subsequent matching and scoring. Our method is strictly training-free: no auxiliary data, prompt/adaptor learning, or parameter optimization is performed on either source or target datasets; inference is carried out directly using the proposed Sparse Hyper Matching and Dual-Branch Image Scoring. All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24 GB memory, using a test batch size of 8.

5.2 Comparison with State-of-the-art Methods

5.2.1 Competing Methods.

We compare our proposed **Hyper-FSAD** with nine recent baselines, including RegAD Huang et al. (2022), FastRecon Fang et al. (2023), WinCLIP Jeong et al. (2023), APRIL-GAN Chen et al. (2023), PromptAD Li et al. (2024), AnomalyGPT Gu et al. (2024), ReMP-AD Ma et al. (2025), KAG-prompt Tao et al. (2025), and SubspaceAD Lendering et al. (2026). For a fair comparison, all methods are evaluated under identical 1-, 2-, and 4-shot settings on six datasets (MVTecAD Bergmann et al. (2019), VisA Zou et al. (2022), MPDD Jezek et al. (2021), BTAD Mishra et al. (2021), RESC Hu et al. (2019), and BraTS Menze et al. (2014)), using the same sampled support sets, with each experiment repeated five times.

5.2.2 Quantitative Comparison.

As summarized in Tab. 1, we report image-level AUROC (I-AUROC) and pixel-level AUROC (P-AUROC) under 1/2/4-shot settings for Hyper-FSAD instantiated with three frozen backbones († CLIP ViT-L/14@336px, ‡ DINOv2 ViT-B/14, and $*$ DINOv3 ViT-B/16). Overall, Hyper-FSAD is highly competitive across all three backbones and consistently delivers strong *average* results. With the default DINOv3 ViT-B/16, Hyper-FSAD reaches average pixel-level AUROC values of 97.00%/97.27%/97.23% for 1/2/4-shot, surpassing *PromptAD* by **+0.87**, **+0.77**, and **+0.36** points, respectively. For image-level detection, Hyper-FSAD also achieves the best average I-AUROC at every shot, reaching 89.95%/90.40% with DINOv3 under 1/2-shot and 90.48% with CLIP under 4-shot, compared with 88.02%/90.10%/90.28% for the strongest baseline (*ReMP-AD*). The aggregate trade-off in Fig. 1a further shows that the default Hyper-FSAD combines 89.95%/97.00% mean 1-shot I/P-AUROC with a 1-shot runtime of 52.6 ms, lower than all compared baselines in Tab. 4. Crucially, all three instantiations remain on par with or superior to prior training-based, language-guided methods, showing that the improvements stem from our Sparse Hyper Matching and Dual-Branch Image Scoring rather than from a particular backbone. It is worth noting that some competing methods further require additional training or per-task adaptation when applied to new categories or datasets. In contrast, Hyper-FSAD performs inference in a *training-free* manner, enabling a simpler deployment pipeline while maintaining superior accuracy. Additional support images generally improve normal coverage, although the gains need not be monotonic for every dataset or metric.

5.2.3 Qualitative Comparison.

Fig. 1b visualizes anomaly localization results under the 1-shot setting on both industrial and medical datasets. Overall, **Hyper-FSAD** consistently produces noticeably more accurate anomaly maps with clearer boundaries and fewer false positives in normal regions, while preserving fine-grained abnormal structures. Across these representative examples, Hyper-FSAD concentrates responses tightly within annotated anomalies while effectively suppressing background activations. Additional Hyper-FSAD*–GT comparisons in Appendix E further demonstrate reliable pixel-level localization across diverse anomaly appearances under limited normal references.

5.3 Ablation Study

Unless otherwise specified, all ablation experiments in this subsection are conducted on MVTecAD Bergmann et al. (2019) and VisA Zou et al. (2022) under the 1/2/4-shot settings, using DINOv3 ViT-B/16 with 448×448 inputs. The input-resolution ablation instead varies the resolution from 224×224 to 672×672 while retaining the same backbone and few-shot protocol.

5.3.1 The effects of Lookup Strategies.

Tab. 2 compares sparse weighting variants on MVTecAD and VisA; results on MPDD, BTAD, RESC, and BraTS are provided in Appendix B.1. Sparsemax (our *Sparse Hyper Matching*) achieves the strongest pixel-level AUROC across both datasets and all shots, together with the best or tied-best image-level results on MVTecAD and competitive results on VisA. Its rows are taken directly from the default DINOv3 ViT-B/16, 448×448 configuration in Tab. 1. Softmax degrades sharply because dense weights mix abundant distractors into reconstruction; entmax provides only partial sparsity,

Table 2: Patch-weighting ablation on MVTecAD (M) and VisA (V), using DINOv3 ViT-B/16 at 448×448. I/P denote I/P-AUROC (%); sparsemax matches Tab. 1.

	Top-1	Top-p (p=1%)	Softmax	Entmax (α=1.5)	Sparsemax (ours)
M1	93.4/95.7	92.8/94.3	83.1/86.5	**96.3**/96.1	**96.3/97.7**
M2	94.6/95.9	93.6/95.1	82.9/86.4	96.6/96.6	**96.8/97.8**
M4	95.1/96.4	94.1/95.4	82.7/86.4	96.7/97.0	**96.9/98.0**
V1	92.2/96.7	90.1/96.6	84.8/93.2	**94.6**/96.3	94.0/**97.7**
V2	92.5/96.9	91.7/97.3	84.8/93.2	**95.2**/97.1	94.8/**98.0**
V4	92.8/96.9	91.8/97.8	84.6/93.1	**95.5**/97.5	**94.8/98.3**

Table 3: DINOv3 layer-selection analysis on MVTecAD (M) Bergmann et al. (2019) and VisA (V) Zou et al. (2022) at 448×448. I/P denote I/P-AUROC (%); standard deviations are omitted.

	$\{6\}$	$\{12\}$	$\{6, 12\}$	$\{3, 6, 9, 12\}$
M1	93.5/96.2	92.6/95.3	94.4/97.4	**96.3/97.7**
M2	94.0/96.6	93.3/95.6	95.2/97.5	**96.8/97.8**
M4	94.4/97.0	94.2/95.8	95.5/97.7	**96.9/98.0**
V1	89.7/96.2	84.3/93.7	91.2/97.3	**94.0/97.7**
V2	91.1/96.7	87.3/95.3	92.9/97.9	**94.8/98.0**
V4	91.2/97.1	88.6/95.7	92.9/98.1	**94.8/98.3**

while Top-p remains sensitive to a manually chosen retention ratio. Sparsemax instead forms an adaptive active set with exact zeros, yielding robust localization without a sparsity hyperparameter.

5.3.2 The effects of DINOv3 layer selection.

Tab. 3 reports the layer-selection analysis on MVTecAD and VisA. A single intermediate or final layer is consistently inferior to combining multiple depths, while $\{6, 12\}$ already recovers much of the performance. The default $\{3, 6, 9, 12\}$ configuration gives the strongest overall localization and competitive image-level detection. This behavior is consistent with the complementary-layer ranking guarantee in Theorem 2, proved in Appendix A.3: when layer-wise margins are positive on average and their fluctuations are not perfectly correlated, multi-layer aggregation yields a strictly tighter pairwise-misranking, and hence AUROC, certificate than a single layer. Results on MPDD, BTAD, RESC, and BraTS appear in Appendix B.2.

5.4 Efficiency Analysis

We compare Hyper-FSAD with representative FSAD baselines in inference efficiency. Tab. 4 reports GPU memory and per-image inference time under the 1-shot setting on a single NVIDIA RTX 3090 GPU with 24 GB memory. Although several baselines use lower input resolutions, Hyper-FSAD achieves the lowest latency with low memory consumption, requiring only 0.89 GB and 52.6 ms per image. Despite its higher-resolution visual multi-layer pipeline, this efficiency follows from training-free feed-forward inference, which avoids costly prompt optimization, adapter learning, and iterative reconstruction.

Table 4: Comparison of 1-shot inference efficiency on an NVIDIA RTX 3090.

Method	Backbone	Resolution	GPU Cost (GB)	Time (ms)
RegAD	ResNet18	224×224	3.82	1517.1
FastRecon	Wide-ResNet50	224×224	**0.55**	106.6
WinCLIP	ViT-L/14-336	336×336	15.70	8033.5
APRIL-GAN	ViT-L/14-336	336×336	2.05	204.6
PromptAD	ViT-L/14-336	336×336	2.10	80.1
AnomalyGPT	ImageBind-Huge	224×224	19.17	1653.2
ReMP-AD	ViT-L/14-336	336×336	1.48	118.8
KAG-prompt	ViT-L/14-336	336×336	3.37	100.6
SubspaceAD	DINOv2 ViT-B/14	448×448	1.01	86.9
Hyper-FSAD	DINOv3 ViT-B/16	448×448	0.89	**52.6**

6 Conclusion

We presented **Hyper-FSAD**, a training- and language-free FSAD framework using a frozen visual encoder. Sparse Hyper Matching rejects distractors without fixed top-p, while Dual-Branch Image Scoring fuses spatial and global visual evidence. We prove exact below-threshold rejection and conditional few-shot separation. Across six industrial and medical benchmarks and three backbones, Hyper-FSAD delivers consistently strong results without task-specific parameter learning or prompts.

AI USE STATEMENT

We used generative AI tools to polish the wording of selected passages of this manuscript and to assist in implementing parts of the experimental code. The authors take responsibility for the final manuscript, implementation, and reported results.

REFERENCES

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 9592–9600, 2019.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *Eur. Conf. Comput. Vis.*, pp. 55–72, 2024.

Qiyu Chen, Huiyuan Luo, Haiming Yao, Wei Luo, Zhen Qu, Chengkan Lv, and Zhengtao Zhang. Center-aware residual anomaly synthesis for multiclass industrial anomaly detection. *IEEE Trans. Ind. Informatics*, 2025.

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

Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Int. Conf. Comput. Vis.*, pp. 17481–17490, 2023.

P. Goodarzi, A. Schütze, and T. Schneider. Domain shifts in industrial condition monitoring: a comparative analysis of automated machine learning models. *Journal of Sensors and Sensor Systems*, 14(2):119–132, 2025.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI*, pp. 1932–1940, 2024.

Brian Guo, Darui Lu, Gregory Szumel, Rongze Gui, Tingyu Wang, Nicholas Konz, and Maciej A Mazurowski. The impact of scanner domain shift on deep learning performance in medical imaging: an experimental study. *arXiv preprint arXiv:2409.04368*, 2024.

Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical image analysis*, 55:216–227, 2019.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *Eur. Conf. Comput. Vis.*, pp. 303–319, 2022.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 19606–19616, 2023.

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)*, pp. 66–71. IEEE, 2021.

Camile Lendering, Erkut Akdag, and Egor Bondarev. Subspacead: Training-free few-shot anomaly detection via subspace modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 28557–28566, 2026.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 16838–16848, 2024.

Hongchi Ma, Guanglei Yang, Debin Zhao, Yanli Ji, and Wangmeng Zuo. Remp-ad: Retrieval-enhanced multi-modal prompt fusion for few-shot industrial visual anomaly detection. In *Int. Conf. Comput. Vis.*, pp. 20425–20434, 2025.

Dwarikanath Mahapatra, Behzad Bozorgtabar, and Zongyuan Ge. Medical image classification using generalized zero shot learning. In *Int. Conf. Comput. Vis.*, pp. 3344–3353, 2021.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Int. Conf. Mach. Learn.*, pp. 1614–1623, 2016.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging*, 34(10):1993–2024, 2014.

P Mishra, R Verk, D Fornasier, C Piciarelli, GL Foresti, et al. Vt-adl: A vision transformer network for image anomaly detection and localization. In *IEEE International Symposium on Industrial Electronics*, pp. 01–06, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ben Peters, Vlad Niculae, and André FT Martins. Sparse sequence-to-sequence models. In *Assoc. Comput. Linguistics*, pp. 1504–1519, 2019.

Zhen Qu, Xian Tao, Fei Shen, Zhengtao Zhang, and Tao Li. Investigating shift equivalence of convolutional neural networks in industrial defect segmentation. *IEEE Trans. Instrumentation and Measurement*, 72:1–17, 2023.

Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *Eur. Conf. Comput. Vis.*, pp. 301–317, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pp. 8748–8763, 2021.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Fenfang Tao, Guo-Sen Xie, Fang Zhao, and Xiangbo Shu. Kernel-aware graph prompt learning for few-shot anomaly detection. In *AAAI*, pp. 7347–7355, 2025.

Fangjun Wang, Jianhao Wu, Zhouwang Yang, and Yanzhi Song. Industrial vision inspection using digital twins: bridging cad models and realistic scenarios. *Journal of Intelligent Manufacturing*, 36(7):4963–4975, 2025.

Xingwu Zhang, Guanxuan Li, Paul Henderson, Gerardo Aragon-Camarasa, and Zijun Long. Is task-specific training necessary for anomaly detection? *arXiv preprint arXiv:2601.22763*, 2026.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *Int. Conf. Learn. Represent.*, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *Eur. Conf. Comput. Vis.*, pp. 392–408, 2022.

APPENDIX OVERVIEW

This appendix provides supplementary details supporting the main manuscript, organized as follows:

- **Sec. A** provides proofs and additional theoretical details.
- **Sec. B** reports additional metrics and cross-dataset patch-weighting and layer-selection ablations.
- **Sec. C** studies the effect of the fusion strategy.
- **Sec. D** studies the effect of the pooling operator.
- **Sec. E** provides additional qualitative results.
- **Sec. F** reports the input-resolution ablation.

A PROOFS AND ADDITIONAL THEORETICAL DETAILS

A.1 PROOF OF PROPOSITION 1

For clarity, we use column vectors in the proofs and omit the layer index. Let $\Delta_M = \{\mathbf{w} \in \mathbb{R}^M : \mathbf{w} \geq \mathbf{0}, \mathbf{1}^\top \mathbf{w} = 1\}$ and Π_{Δ_M} denote Euclidean projection onto this simplex.

Proof. The KKT conditions for $\min_{\mathbf{w}\in\Delta_M} \frac{1}{2}\|\mathbf{w} - \mathbf{z}\|_2^2$ give the unique solution

$$w_i = [z_i - \tau]_+, \qquad \sum_{i=1}^{M} [z_i - \tau]_+ = 1, \tag{10}$$

where $[a]_+ = \max(a, 0)$. Under Eq. (2), setting $\tau = \tau_R$ makes the active coordinates exactly R, and $\sum_{i\in R}(z_i - \tau_R) = 1$ by definition. This proves the support and weight formula.

Now append similarities $\tilde{z}_j \leq \tau_R$. The same threshold satisfies

$$\sum_{i=1}^{M} [z_i - \tau_R]_+ + \sum_j [\tilde{z}_j - \tau_R]_+ = 1 + 0 = 1. \tag{11}$$

Hence Eq. (10) gives the extended solution $[\mathbf{w}; \mathbf{0}]$. Since the simplex projection is unique, neither the original weights nor the reconstruction $\mathbf{P}^\top \mathbf{w}$ changes.

It remains to prove stability. A projection onto any nonempty closed convex set is non-expansive. In particular, the variational inequalities for $\mathbf{w} = \Pi_{\Delta_M}(\mathbf{z})$ and $\mathbf{w}' = \Pi_{\Delta_M}(\mathbf{z}')$ imply

$$\|\mathbf{w} - \mathbf{w}'\|_2^2 \leq \langle \mathbf{w} - \mathbf{w}', \mathbf{z} - \mathbf{z}' \rangle \leq \|\mathbf{w} - \mathbf{w}'\|_2 \|\mathbf{z} - \mathbf{z}'\|_2. \tag{12}$$

Canceling the common factor proves the non-expansiveness claim. Since $\mathbf{z} = \mathbf{P}\mathbf{q}$ and $\boldsymbol{\varepsilon}(\mathbf{q}) = \mathbf{P}^\top \Pi_{\Delta_M}(\mathbf{P}\mathbf{q})$, submultiplicativity further gives the reconstruction bound

$$\|\boldsymbol{\varepsilon}(\mathbf{q}) - \boldsymbol{\varepsilon}(\mathbf{q}')\|_2 \leq \|\mathbf{P}\|_2 \|\Pi_{\Delta_M}(\mathbf{P}\mathbf{q}) - \Pi_{\Delta_M}(\mathbf{P}\mathbf{q}')\|_2 \tag{13}$$

$$\leq \|\mathbf{P}\|_2^2 \|\mathbf{q} - \mathbf{q}'\|_2, \tag{14}$$

which holds even when the active set changes. □

The invariance is specific to sufficiently low-scoring additions. For comparison, if J new patches all have similarity a, dense softmax assigns them total mass

$$\frac{Je^a}{\sum_{i=1}^{M} e^{z_i} + Je^a} \longrightarrow 1 \quad \text{as } J \to \infty, \tag{15}$$

whereas sparsemax remains exactly unchanged whenever $a \leq \tau_R$.

Local form on a fixed active set. When the active set R does not change, let $\mathbf{H}_R = \mathbf{I} - |R|^{-1}\mathbf{1}\mathbf{1}^\top$ and let $\mathbf{P}_R$ contain its selected support rows. Direct substitution of the threshold gives

$$\boldsymbol{\varepsilon}(\mathbf{q}) = \bar{\mathbf{p}}_R + \mathbf{G}_R\mathbf{q}, \tag{16}$$

$$\mathbf{G}_R = \mathbf{P}_R^\top \mathbf{H}_R \mathbf{P}_R \tag{17}$$

$$= \sum_{i\in R} (\mathbf{p}_i - \bar{\mathbf{p}}_R)(\mathbf{p}_i - \bar{\mathbf{p}}_R)^\top, \tag{18}$$

where $\bar{\mathbf{p}}_R = |R|^{-1}\sum_{i\in R}\mathbf{p}_i$. Thus the local reconstruction gain is governed by the dispersion $\|\mathbf{G}_R\|_2$ of the *selected* normal evidence, rather than directly by the total memory size.

A.2 Proof of Theorem 1

Proof. All tokens are unit-normalized. For a normal query patch and every active support patch,

$$\langle \mathbf{q}_N^l, \mathbf{p}_i^l \rangle = 1 - \tfrac{1}{2}\|\mathbf{q}_N^l - \mathbf{p}_i^l\|_2^2 \geq 1 - \tfrac{1}{2}\epsilon_l^2. \tag{19}$$

Because the generic retrieval weights lie in Δ_M, the reconstruction is a convex combination: $\boldsymbol{\varepsilon}_N^l = \sum_i w_i^l \mathbf{p}_i^l$. Therefore,

$$\langle \mathbf{q}_N^l, \boldsymbol{\varepsilon}_N^l \rangle \geq 1 - \tfrac{1}{2}\epsilon_l^2, \qquad \|\boldsymbol{\varepsilon}_N^l\|_2 \leq 1. \tag{20}$$

The first quantity is positive because $\epsilon_l < \sqrt{2}$, and hence $\cos(\mathbf{q}_N^l, \boldsymbol{\varepsilon}_N^l) \geq 1 - \epsilon_l^2/2$. Every patch of N consequently has layer-averaged score at most $|L|^{-1}\sum_l \epsilon_l^2/2$, so max pooling preserves this upper bound.

For the specified anomalous patch, each reconstruction lies in $\mathrm{conv}(\mathcal{P}^l)$. Assumption (ii) of Theorem 1 therefore gives a layer score of at least γ_l. Averaging across layers and then taking the maximum over patches gives

$$S_{\mathrm{map}}(A) - S_{\mathrm{map}}(N) \geq \frac{1}{|L|}\sum_{l\in L}\left(\gamma_l - \frac{\epsilon_l^2}{2}\right) = \Delta_{\mathrm{map}}. \tag{21}$$

For unit <CLS> tokens,

$$1 - \langle \mathbf{c}, \mathbf{c}_k \rangle = \tfrac{1}{2}\|\mathbf{c} - \mathbf{c}_k\|_2^2. \tag{22}$$

The two <CLS> distance assumptions thus imply

$$S_{\mathrm{cls}}(A) - S_{\mathrm{cls}}(N) \geq \frac{1}{2|L|}\sum_{l\in L}(\delta_l^2 - r_l^2) = \Delta_{\mathrm{cls}}. \tag{23}$$

Applying the convex fusion in Eq. (5) proves Eq. (6). If its right-hand side is positive, the anomalous score is strictly larger than the normal score, so a threshold between them separates the pair. If the assumptions hold uniformly for two populations, the same threshold separates those populations. □

A.3 Complementarity of Multi-Layer Evidence

For compactness, define the layer-wise patch and global discrepancies

$$d_{X,t}^l = 1 - \cos(\mathbf{q}_{X,t}^l, \boldsymbol{\varepsilon}_{X,t}^l), \qquad g_X^l = 1 - \max_k \langle \mathbf{c}_X^l, \mathbf{c}_k^l \rangle. \tag{24}$$

We write S_{image}^L for the score in Eq. (5) computed with layer set L.

Theorem 2 (Complementary-layer ranking guarantee)**.** *Let $m = |L| > 1$ and consider a normal–anomalous pair (N, A) with a common anomalous patch $t^\star$ across the selected feature grids. Define the layer-wise ranking certificate*

$$\eta_l = \lambda\Big(d_{A,t^\star}^l - \max_t d_{N,t}^l\Big) + (1-\lambda)(g_A^l - g_N^l). \tag{25}$$

Then

$$S_{\mathrm{image}}^L(A) - S_{\mathrm{image}}^L(N) \geq \bar{\eta}_L := \frac{1}{m}\sum_{l\in L}\eta_l. \tag{26}$$

Under the assumptions of Theorem 1, each certificate further satisfies

$$\eta_l \geq b_l := \lambda\left(\gamma_l - \frac{\epsilon_l^2}{2}\right) + \frac{1-\lambda}{2}(\delta_l^2 - r_l^2). \tag{27}$$

Thus $\sum_l b_l > 0$ certifies multi-layer separation, even if some individual b_l are non-positive.

Now let the support set and normal–anomalous pair be random. Suppose that, for every $l \in L$, $\mathbb{E}[\eta_l] \geq \mu > 0$, $\mathrm{Var}(\eta_l) \leq \sigma^2$, and $\mathrm{Cov}(\eta_l, \eta_r) \leq \rho\sigma^2$ for $l \neq r$, where $0 \leq \rho < 1$. Then

$$\Pr\big[S^L_{\mathrm{image}}(A) \leq S^L_{\mathrm{image}}(N)\big] \leq \frac{\sigma^2 c_m}{\sigma^2 c_m + \mu^2}, \tag{28}$$

where $c_m = [1 + (m-1)\rho]/m < 1$. For $\sigma > 0$, this is strictly tighter than the corresponding uniform single-layer certificate $\sigma^2/(\sigma^2 + \mu^2)$.

Proof. Using Eq. (3), the order of averaging and max pooling gives

$$S^L_{\mathrm{map}}(A) \geq \frac{1}{m}\sum_{l\in L} d^l_{A,t^\star}, \tag{29}$$

$$S^L_{\mathrm{map}}(N) = \max_t \frac{1}{m}\sum_{l\in L} d^l_{N,t} \leq \frac{1}{m}\sum_{l\in L}\max_t d^l_{N,t}. \tag{30}$$

Moreover, Eq. (4) gives $S^L_{\mathrm{cls}}(A) - S^L_{\mathrm{cls}}(N) = m^{-1}\sum_l (g^l_A - g^l_N)$. Applying the convex fusion in Eq. (5) proves Eq. (26). The patch and global bounds established in the proof of Theorem 1 give

$$\begin{aligned} d^l_{A,t^\star} - \max_t d^l_{N,t} &\geq \gamma_l - \epsilon_l^2/2, \\ g^l_A - g^l_N &\geq (\delta_l^2 - r_l^2)/2, \end{aligned} \tag{31}$$

which proves Eq. (27).

It remains to establish the ranking guarantee. Let $\bar{\mu}_L = \mathbb{E}[\bar{\eta}_L]$ and $v_L = \mathrm{Var}(\bar{\eta}_L)$. Because the actual score gap is lower-bounded by $\bar{\eta}_L$, a misranking implies $\bar{\eta}_L \leq 0$. Cantelli's inequality therefore yields

$$\Pr(\bar{\eta}_L \leq 0) = \Pr(\bar{\eta}_L - \bar{\mu}_L \leq -\bar{\mu}_L) \leq \frac{v_L}{v_L + \bar{\mu}_L^2}. \tag{32}$$

Under the stated covariance conditions,

$$v_L = \frac{1}{m^2}\left(\sum_l \mathrm{Var}(\eta_l) + 2\sum_{l<r}\mathrm{Cov}(\eta_l, \eta_r)\right) \tag{33}$$

$$\leq \frac{\sigma^2}{m}\big(1 + (m-1)\rho\big) = \sigma^2 c_m. \tag{34}$$

Together with $\bar{\mu}_L \geq \mu$, this proves Eq. (28). Since $m > 1$ and $\rho < 1$, we have $c_m < 1$, so its right-hand side is strictly smaller than the uniform single-layer bound whenever $\sigma > 0$. □

Under the standard pairwise-ranking interpretation of AUROC, Eq. (32) also gives $\mathrm{AUROC}_L \geq 1 - v_L/(v_L + \bar{\mu}_L^2)$ when ties are counted pessimistically. For pre-interpolation pixel scores, the same derivation applies to $\eta^{\mathrm{patch}}_l = d^l_{A,t_A} - d^l_{N,t_N}$. The result is conditional rather than a claim that adding an arbitrary layer must help: a strongly negative mean margin can reduce $\bar{\mu}_L$, while perfectly correlated fluctuations ($\rho = 1$) provide no concentration gain. It therefore attributes the improvement to complementary evidence across depths, rather than to layer count alone.

A.4 Few-Shot Coverage and Stability of the Global Branch

For a fixed normal token $\mathbf{c}^l_N$, let $\mathbf{c}^l_{\mathrm{sup}}$ denote one independently sampled normal support token and define $p_l = \Pr(\|\mathbf{c}^l_{\mathrm{sup}} - \mathbf{c}^l_N\|_2 \leq r_l \mid \mathbf{c}^l_N) \geq \alpha_l$. Independence of the K sampled normal supports gives

$$\Pr\left[\min_k \|\mathbf{c}^l_N - \mathbf{c}^l_k\|_2 > r_l\right] \tag{35}$$

$$= (1-p_l)^K \leq (1-\alpha_l)^K \leq e^{-K\alpha_l}, \tag{36}$$

which proves Eq. (7). A union bound shows that all selected layers satisfy their respective coverage conditions with probability at least $1 - \sum_{l \in L} e^{-K\alpha_l}$; no independence across layers is required for this final step.

The global score also has a direct stability property. For a fixed unit-normalized memory $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{K}$, define $s_{\mathcal{C}}(\mathbf{c}) = 1 - \max_k \langle \mathbf{c}, \mathbf{c}_k \rangle$. It is zero on every stored token, and for any unit $\mathbf{u}, \mathbf{v}$,

$$|s_{\mathcal{C}}(\mathbf{u}) - s_{\mathcal{C}}(\mathbf{v})| = \left| \max_k \langle \mathbf{v}, \mathbf{c}_k \rangle - \max_k \langle \mathbf{u}, \mathbf{c}_k \rangle \right| \tag{37}$$

$$\leq \max_k |\langle \mathbf{u} - \mathbf{v}, \mathbf{c}_k \rangle| \leq \|\mathbf{u} - \mathbf{v}\|_2. \tag{38}$$

Thus each layer's implemented cosine score is 1-Lipschitz on the unit sphere. By Eq. (22), $\sqrt{2 s_{\mathcal{C}}(\mathbf{c})} = \min_k \|\mathbf{c} - \mathbf{c}_k\|_2$ is exactly the distance to the fixed normal <CLS> memory. The maximal stable-score result for distance-to-memory Zhang et al. (2026) therefore applies to this square-root distance, but not to the sparse patch-reconstruction score.

A.5 Sparsemax Computation

Algorithm 1 gives the closed-form computation used by Sparse Hyper Matching. The permutation is explicitly inverted before returning the weights so that they remain aligned with the original memory-bank rows.

Algorithm 1: Sparsemax with Original-Index Recovery

Require: Similarities $\mathbf{z}^l \in \mathbb{R}^M$, where $M = KN_p$
Ensure: Sparse weights $\hat{\mathbf{w}}^l \in \mathbb{R}^M$

```
Sort and record π: z^l_{π(1)} ≥ ··· ≥ z^l_{π(M)}
m ← max{j : 1 + j z^l_{π(j)} > Σ_{u=1}^{j} z^l_{π(u)}}
τ ← ( Σ_{u=1}^{m} z^l_{π(u)} − 1 ) / m
for j = 1, ..., M do
    ŵ^l_{π(j)} ← max(0, z^l_{π(j)} − τ)
end for
return ŵ^l
```

B Detailed Results on Additional Metrics

In this section, we report detailed quantitative results on additional evaluation metrics beyond the AUROC results presented in the main paper. Specifically, Tab. 5 reports image-level F1-max (I-F1) and average precision (I-AP), while Tab. 6 reports pixel-level PRO (P-PRO) and average precision (P-AP), across six datasets under 1-shot, 2-shot, and 4-shot settings. For consistency with the main paper, all methods are evaluated on the same support sets and the results are averaged over five runs. The best and second-best results are marked in **bold** and underlined, respectively. As shown in Tabs. 5 and 6, the superiority of Hyper-FSAD extends consistently beyond AUROC to threshold-dependent and region-aware metrics. At the image level, our method achieves the strongest overall performance in both I-F1 and I-AP across different shot settings, indicating that the proposed scoring mechanism not only ranks anomalous samples effectively, but also produces more reliable decision boundaries.

At the pixel level, Hyper-FSAD also delivers highly competitive or best results on P-PRO and P-AP for most datasets and shot settings, demonstrating that Sparse Hyper Matching yields anomaly maps with improved localization quality and region consistency. These observations further support our claim that the combination of sparse reconstruction and dual-branch scoring provides robust benefits for both classification and segmentation in a training-free and language-free setting.

B.1 Cross-Dataset Patch-Weighting Ablation

We further report the patch-weighting comparison on MPDD, BTAD, RESC, and BraTS, complementing the MVTecAD and VisA results in Tab. 2. All configurations use the same default DINOv3

Table 5: Image-level F1-max (**I-F1**) and average precision (**I-AP**) across six datasets and 1/2/4-shot settings. Symbols, highlighting, backbone variants, and reporting protocol follow Tab. 1.

	Methods	Venue/Year	MVTecAD		VisA		MPDD		BTAD		RESC		BraTS	
			I-F1	I-AP	I-F1	I-AP	I-F1	I-AP	I-F1	I-AP	I-F1	I-AP	I-F1	I-AP
1-shot	RegAD	ECCV'22	87.1	87.2	76.2	72.2	72.9	61.5	78.2	80.5	60.4	46.6	83.0	73.2
	FastRecon	ICCV'23	93.4	95.6	81.4	82.3	80.3	76.1	92.0	95.0	71.3	80.4	86.5	85.1
	WinCLIP	CVPR'23	92.0	96.1	82.8	87.0	80.9	82.5	81.8	86.3	60.7	48.1	87.4	92.5
	APRIL-GAN	CVPR'23	90.9	95.6	83.1	90.5	80.1	80.8	84.0	88.5	69.5	69.7	88.9	92.5
	PromptAD	CVPR'24	93.7	96.6	83.3	86.8	81.6	83.5	90.4	94.4	78.2	84.3	87.5	88.5
	AnomalyGPT	AAAI'24	94.3	96.1	84.4	87.4	79.3	75.9	89.7	94.6	76.2	83.4	85.8	82.0
	ReMP-AD	ICCV'25	94.3	97.3	86.9	93.1	82.9	84.0	91.5	95.4	72.6	78.2	87.5	93.8
	KAG-prompt	AAAI'25	94.9	97.3	86.9	92.1	79.4	70.8	90.0	95.5	64.4	62.6	86.5	89.9
	SubspaceAD‡	CVPR'26	95.4	96.6	89.6	93.5	82.6	81.4	94.4	97.7	71.2	61.4	**89.5**	**95.5**
	Hyper-FSAD†	-	$93.2_{\pm 0.2}$	$96.7_{\pm 0.1}$	$89.1_{\pm 0.5}$	$93.2_{\pm 0.6}$	$85.2_{\pm 0.4}$	$84.0_{\pm 2.0}$	$88.4_{\pm 1.4}$	$94.0_{\pm 1.2}$	$80.7_{\pm 1.9}$	$85.8_{\pm 1.6}$	$89.4_{\pm 0.8}$	$94.9_{\pm 0.5}$
	Hyper-FSAD‡	-	$\mathbf{96.8}_{\pm \mathbf{0.1}}$	$\mathbf{98.1}_{\pm \mathbf{0.0}}$	$89.9_{\pm 0.1}$	$93.8_{\pm 0.2}$	$\mathbf{85.9}_{\pm \mathbf{0.6}}$	$\mathbf{85.8}_{\pm \mathbf{1.0}}$	$93.6_{\pm 1.9}$	$98.1_{\pm 1.2}$	$77.8_{\pm 0.2}$	$71.2_{\pm 0.3}$	$86.9_{\pm 1.6}$	$87.8_{\pm 1.8}$
	Hyper-FSAD*	-	$96.6_{\pm 0.3}$	$\mathbf{98.1}_{\pm \mathbf{0.1}}$	$\mathbf{90.6}_{\pm \mathbf{0.1}}$	$\mathbf{94.0}_{\pm \mathbf{0.1}}$	$82.5_{\pm 0.4}$	$81.6_{\pm 2.1}$	$\mathbf{95.1}_{\pm \mathbf{0.2}}$	$\mathbf{98.7}_{\pm \mathbf{0.4}}$	$\mathbf{83.1}_{\pm \mathbf{1.4}}$	$\mathbf{90.4}_{\pm \mathbf{2.3}}$	$87.7_{\pm 1.0}$	$89.6_{\pm 1.1}$
2-shot	RegAD	ECCV'22	88.8	88.9	75.8	73.6	73.6	62.2	89.2	92.1	62.0	48.0	83.5	72.0
	FastRecon	ICCV'23	94.5	96.5	81.8	81.3	83.0	81.4	90.3	94.9	75.7	84.9	87.2	83.6
	WinCLIP	CVPR'23	93.0	96.6	81.3	85.9	81.0	83.3	84.2	87.6	61.0	50.6	88.0	93.4
	APRIL-GAN	CVPR'23	91.0	95.5	82.6	90.4	79.4	80.2	84.2	88.5	70.8	71.0	89.2	93.0
	PromptAD	CVPR'24	95.1	97.7	83.0	87.0	83.6	**88.2**	89.0	94.4	79.6	85.9	88.1	89.3
	AnomalyGPT	AAAI'24	95.0	97.0	84.1	88.8	82.1	81.1	89.9	95.0	78.2	83.5	86.6	83.3
	ReMP-AD	ICCV'25	94.9	97.8	89.0	94.7	85.5	86.5	92.1	96.6	76.0	81.7	88.6	94.3
	KAG-prompt	AAAI'25	96.2	**98.4**	88.6	93.8	79.9	71.1	90.1	95.1	67.3	75.3	86.5	92.3
	SubspaceAD‡	CVPR'26	96.3	97.3	90.2	93.4	85.0	84.2	**95.5**	98.5	71.8	58.3	89.0	93.6
	Hyper-FSAD†	-	$93.7_{\pm 0.3}$	$96.8_{\pm 0.4}$	$90.4_{\pm 0.4}$	$\mathbf{95.1}_{\pm \mathbf{0.4}}$	$\mathbf{85.9}_{\pm \mathbf{0.4}}$	$85.9_{\pm 0.9}$	$89.8_{\pm 1.1}$	$95.1_{\pm 0.6}$	$79.3_{\pm 1.1}$	$84.3_{\pm 0.9}$	$\mathbf{89.9}_{\pm \mathbf{1.6}}$	$\mathbf{95.2}_{\pm \mathbf{1.2}}$
	Hyper-FSAD‡	-	$\mathbf{97.1}_{\pm \mathbf{0.2}}$	$98.0_{\pm 0.1}$	$91.1_{\pm 0.2}$	$\mathbf{95.1}_{\pm \mathbf{0.2}}$	$85.1_{\pm 0.7}$	$84.6_{\pm 0.7}$	$93.9_{\pm 2.0}$	$98.0_{\pm 0.9}$	$78.9_{\pm 0.2}$	$70.0_{\pm 0.3}$	$86.8_{\pm 1.9}$	$86.2_{\pm 1.0}$
	Hyper-FSAD*	-	$96.9_{\pm 0.2}$	$98.2_{\pm 0.0}$	$\mathbf{91.7}_{\pm \mathbf{0.2}}$	$94.5_{\pm 0.3}$	$82.5_{\pm 0.1}$	$83.6_{\pm 0.4}$	$95.3_{\pm 0.7}$	$\mathbf{98.6}_{\pm \mathbf{0.3}}$	$\mathbf{81.9}_{\pm \mathbf{2.2}}$	$\mathbf{87.3}_{\pm \mathbf{1.2}}$	$89.0_{\pm 2.1}$	$89.2_{\pm 1.7}$
4-shot	RegAD	ECCV'22	89.8	91.7	77.1	73.9	75.9	66.9	91.2	94.5	63.7	51.0	83.9	75.5
	FastRecon	ICCV'23	95.3	97.2	82.6	85.1	82.8	81.0	91.8	96.2	76.4	84.6	87.6	86.4
	WinCLIP	CVPR'23	94.0	97.3	82.8	87.8	83.1	86.1	84.3	88.0	62.6	54.0	88.0	93.4
	APRIL-GAN	CVPR'23	91.6	95.9	83.3	91.1	80.7	81.6	83.9	88.4	70.7	71.2	89.1	93.5
	PromptAD	CVPR'24	95.2	97.5	83.9	89.2	87.6	**92.6**	91.0	94.4	**81.0**	**87.3**	88.2	92.4
	AnomalyGPT	AAAI'24	95.9	98.0	87.2	92.6	**88.5**	89.0	91.0	95.9	78.8	85.4	86.2	87.8
	ReMP-AD	ICCV'25	95.4	98.2	89.5	95.3	83.7	84.7	94.3	98.0	77.4	83.0	88.6	94.3
	KAG-prompt	AAAI'25	96.6	**98.3**	88.9	93.9	81.7	75.3	91.0	95.2	72.4	79.1	87.3	94.1
	SubspaceAD‡	CVPR'26	96.9	97.6	91.2	94.9	84.7	84.3	95.4	98.7	74.1	72.3	89.3	91.4
	Hyper-FSAD†	-	$94.7_{\pm 0.2}$	$97.6_{\pm 0.2}$	$90.8_{\pm 0.1}$	$95.3_{\pm 0.4}$	$86.4_{\pm 0.9}$	$86.7_{\pm 1.7}$	$91.3_{\pm 1.7}$	$96.4_{\pm 0.7}$	$80.7_{\pm 0.5}$	$84.3_{\pm 1.4}$	$\mathbf{89.9}_{\pm \mathbf{0.4}}$	$\mathbf{94.9}_{\pm \mathbf{0.4}}$
	Hyper-FSAD‡	-	$\mathbf{97.4}_{\pm \mathbf{0.1}}$	$98.1_{\pm 0.1}$	$\mathbf{91.7}_{\pm \mathbf{0.2}}$	$\mathbf{95.6}_{\pm \mathbf{0.2}}$	$85.4_{\pm 0.8}$	$84.6_{\pm 0.8}$	$95.7_{\pm 0.5}$	$98.5_{\pm 0.1}$	$78.9_{\pm 0.4}$	$70.2_{\pm 0.5}$	$86.9_{\pm 0.3}$	$84.7_{\pm 0.8}$
	Hyper-FSAD*	-	$97.1_{\pm 0.0}$	$\mathbf{98.3}_{\pm \mathbf{0.0}}$	$\mathbf{91.7}_{\pm \mathbf{0.1}}$	$94.7_{\pm 0.3}$	$82.6_{\pm 1.0}$	$83.2_{\pm 0.8}$	$\mathbf{96.2}_{\pm \mathbf{0.1}}$	$\mathbf{98.9}_{\pm \mathbf{0.1}}$	$79.6_{\pm 1.6}$	$84.9_{\pm 0.9}$	$88.6_{\pm 0.4}$	$86.8_{\pm 0.7}$

Table 6: Pixel-level PRO (**P-PRO**) and average precision (**P-AP**) across six datasets and 1/2/4-shot settings. Symbols, highlighting, backbone variants, and reporting protocol follow Tab. 1.

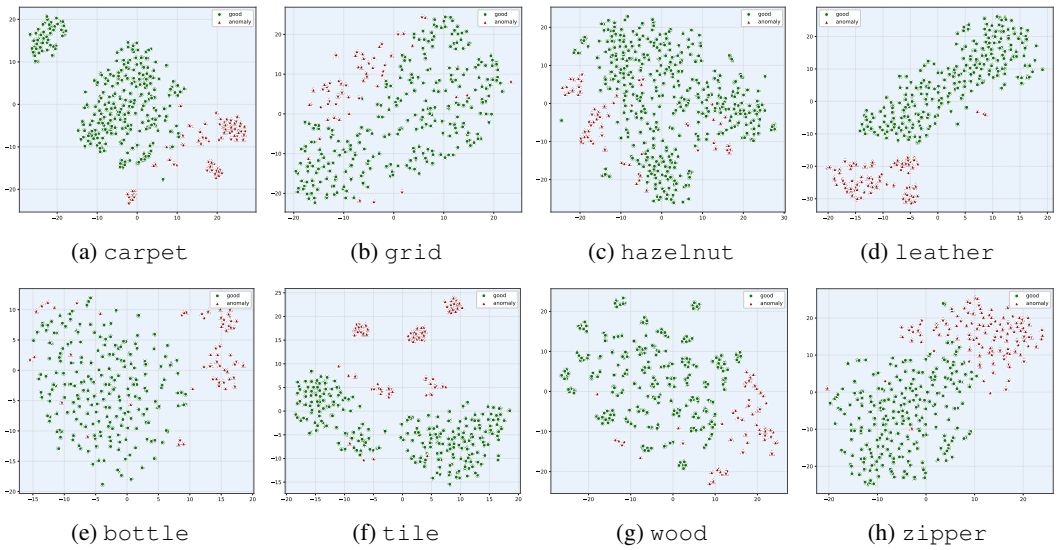
	Methods	Venue/Year	MVTecAD		VisA		MPDD		BTAD		RESC		BraTS	
			P-PRO	P-AP	P-PRO	P-AP	P-PRO	P-AP	P-PRO	P-AP	P-PRO	P-AP	P-PRO	P-AP
1-shot	RegAD	ECCV'22	76.8	36.1	68.7	17.9	74.6	8.4	68.9	33.1	53.2	14.6	62.6	17.5
	FastRecon	ICCV'23	90.8	50.5	85.2	26.0	89.3	26.0	80.6	60.0	83.6	66.5	71.4	39.0
	WinCLIP	CVPR'23	82.0	35.5	85.2	19.2	86.7	23.6	61.7	26.0	73.3	33.4	64.2	33.2
	APRIL-GAN	CVPR'23	84.5	43.8	87.0	29.3	85.1	28.3	73.4	50.4	74.9	54.1	62.7	38.7
	PromptAD	CVPR'24	90.9	53.3	88.4	29.1	90.4	30.1	79.4	61.3	85.8	**68.2**	74.8	46.0
	AnomalyGPT	AAAI'24	89.0	48.8	65.3	16.8	89.9	31.3	71.7	49.9	58.5	27.4	69.9	30.1
	ReMP-AD	ICCV'25	90.7	56.1	91.5	37.2	92.2	34.5	**82.8**	58.1	71.2	43.9	64.3	41.3
	KAG-prompt	AAAI'25	90.5	55.0	83.5	37.0	88.0	33.4	72.1	52.0	51.2	20.3	77.8	55.3
	SubspaceAD‡	CVPR'26	**92.9**	47.7	**93.6**	24.9	**92.7**	26.4	81.4	35.9	77.7	32.7	**79.7**	39.0
	Hyper-FSAD†	-	$87.0_{\pm 0.2}$	$62.2_{\pm 0.1}$	$87.0_{\pm 0.1}$	$38.0_{\pm 0.6}$	$89.9_{\pm 0.7}$	$\mathbf{43.4}_{\pm \mathbf{1.1}}$	$79.4_{\pm 0.4}$	$63.0_{\pm 0.4}$	$71.7_{\pm 1.4}$	$59.9_{\pm 0.7}$	$59.5_{\pm 0.1}$	$54.9_{\pm 1.8}$
	Hyper-FSAD‡	-	$\mathbf{92.9}_{\pm \mathbf{0.1}}$	$65.5_{\pm 0.2}$	$91.0_{\pm 0.1}$	$37.2_{\pm 0.3}$	$91.7_{\pm 0.9}$	$39.8_{\pm 0.7}$	$82.2_{\pm 0.2}$	$\mathbf{72.3}_{\pm \mathbf{0.1}}$	$79.0_{\pm 0.7}$	$45.5_{\pm 1.1}$	$74.6_{\pm 0.6}$	$57.0_{\pm 0.9}$
	Hyper-FSAD*	-	$92.5_{\pm 0.1}$	$\mathbf{70.0}_{\pm \mathbf{0.1}}$	$90.4_{\pm 0.1}$	$\mathbf{40.7}_{\pm \mathbf{0.5}}$	$89.2_{\pm 0.8}$	$35.2_{\pm 0.4}$	$80.9_{\pm 0.4}$	$70.6_{\pm 0.8}$	$\mathbf{86.1}_{\pm \mathbf{0.5}}$	$62.7_{\pm 0.3}$	$76.3_{\pm 0.6}$	$\mathbf{62.7}_{\pm \mathbf{1.3}}$
2-shot	RegAD	ECCV'22	82.7	42.1	70.2	21.6	79.3	13.1	74.1	42.3	54.5	15.1	66.0	20.6
	FastRecon	ICCV'23	91.5	51.9	85.2	30.6	92.7	35.7	80.5	61.6	84.7	68.4	71.7	35.3
	WinCLIP	CVPR'23	82.7	37.4	85.9	23.6	89.4	26.8	63.4	27.5	74.6	35.7	63.6	32.9
	APRIL-GAN	CVPR'23	85.5	45.1	86.8	30.1	86.6	30.2	73.2	50.8	76.5	56.0	63.1	38.8
	PromptAD	CVPR'24	91.5	54.8	89.4	34.4	92.6	34.5	79.6	62.3	**86.6**	**69.9**	75.2	45.7
	AnomalyGPT	AAAI'24	90.2	50.7	65.0	19.7	91.8	34.5	72.4	50.6	59.0	27.9	70.2	29.7
	ReMP-AD	ICCV'25	91.7	58.6	92.0	38.8	**93.2**	37.4	83.2	61.0	75.7	50.0	66.9	45.1
	KAG-prompt	AAAI'25	91.1	56.7	85.2	37.5	88.1	34.2	71.8	52.5	50.4	21.8	78.1	58.0
	SubspaceAD‡	CVPR'26	**93.6**	48.6	**94.4**	26.4	91.0	28.4	82.6	37.1	79.9	33.9	**82.3**	42.1
	Hyper-FSAD†	-	$87.7_{\pm 0.3}$	$63.6_{\pm 0.4}$	$88.3_{\pm 0.2}$	$40.8_{\pm 0.3}$	$89.3_{\pm 0.4}$	$\mathbf{43.7}_{\pm \mathbf{0.8}}$	$79.7_{\pm 0.3}$	$64.9_{\pm 0.8}$	$71.3_{\pm 0.5}$	$61.3_{\pm 0.9}$	$63.4_{\pm 1.9}$	$61.8_{\pm 2.0}$
	Hyper-FSAD‡	-	$93.1_{\pm 0.2}$	$66.0_{\pm 0.2}$	$92.2_{\pm 0.1}$	$39.8_{\pm 0.6}$	$90.5_{\pm 0.7}$	$40.1_{\pm 0.9}$	$\mathbf{83.3}_{\pm \mathbf{0.6}}$	$\mathbf{73.8}_{\pm \mathbf{0.6}}$	$81.3_{\pm 0.3}$	$48.2_{\pm 1.2}$	$77.3_{\pm 1.4}$	$59.6_{\pm 1.8}$
	Hyper-FSAD*	-	$92.9_{\pm 0.2}$	$\mathbf{70.8}_{\pm \mathbf{0.1}}$	$91.5_{\pm 0.1}$	$\mathbf{42.7}_{\pm \mathbf{0.2}}$	$88.6_{\pm 0.3}$	$37.9_{\pm 0.9}$	$82.1_{\pm 0.2}$	$72.1_{\pm 0.7}$	$86.3_{\pm 1.4}$	$64.8_{\pm 1.1}$	$79.2_{\pm 1.5}$	$\mathbf{67.9}_{\pm \mathbf{1.0}}$
4-shot	RegAD	ECCV'22	86.0	46.5	72.8	21.4	83.3	16.4	75.5	44.1	60.0	18.1	70.2	24.8
	FastRecon	ICCV'23	92.2	53.9	86.2	32.5	93.1	37.8	80.8	62.2	82.8	68.5	73.8	43.9
	WinCLIP	CVPR'23	83.8	39.2	86.5	25.7	90.7	29.3	64.7	28.5	75.7	38.4	63.0	40.0
	APRIL-GAN	CVPR'23	86.6	46.6	86.6	30.6	86.9	31.4	74.6	50.9	77.6	57.3	63.0	40.0
	PromptAD	CVPR'24	92.4	57.5	89.5	37.5	94.0	40.5	80.1	62.5	**86.8**	**71.3**	77.0	54.4
	AnomalyGPT	AAAI'24	91.2	52.9	65.4	20.8	93.2	40.8	73.5	50.6	60.0	28.5	73.6	41.8
	ReMP-AD	ICCV'25	92.6	61.4	92.3	41.6	**94.2**	39.7	83.2	61.6	77.5	53.7	69.5	48.2
	KAG-prompt	AAAI'25	91.4	57.4	85.5	38.8	90.2	35.4	70.9	52.5	53.9	24.1	78.6	59.5
	SubspaceAD‡	CVPR'26	**93.9**	50.4	**94.9**	27.0	91.7	28.6	**83.4**	36.6	81.8	36.4	**79.9**	45.4
	Hyper-FSAD†	-	$88.4_{\pm 0.2}$	$64.9_{\pm 0.3}$	$89.3_{\pm 0.2}$	$41.8_{\pm 0.5}$	$90.3_{\pm 0.4}$	$\mathbf{44.8}_{\pm \mathbf{1.0}}$	$80.8_{\pm 0.3}$	$66.5_{\pm 0.6}$	$72.1_{\pm 0.6}$	$61.7_{\pm 0.8}$	$64.0_{\pm 1.1}$	$\mathbf{62.0}_{\pm \mathbf{2.3}}$
	Hyper-FSAD‡	-	$93.5_{\pm 0.1}$	$66.2_{\pm 0.1}$	$93.2_{\pm 0.1}$	$41.2_{\pm 0.2}$	$91.4_{\pm 0.6}$	$41.4_{\pm 0.9}$	$\mathbf{83.4}_{\pm \mathbf{0.6}}$	$\mathbf{74.2}_{\pm \mathbf{0.4}}$	$80.9_{\pm 1.1}$	$48.3_{\pm 1.1}$	$77.4_{\pm 0.8}$	$52.3_{\pm 1.6}$
	Hyper-FSAD*	-	$93.5_{\pm 0.1}$	$\mathbf{71.5}_{\pm \mathbf{0.1}}$	$92.5_{\pm 0.1}$	$\mathbf{43.2}_{\pm \mathbf{0.2}}$	$89.3_{\pm 0.3}$	$40.3_{\pm 1.6}$	$82.2_{\pm 0.4}$	$73.0_{\pm 0.4}$	$83.2_{\pm 1.0}$	$64.1_{\pm 2.1}$	$77.4_{\pm 1.4}$	$61.5_{\pm 1.7}$

Table 7: Cross-dataset patch-weighting ablation across 1/2/4-shot settings. I/P denote I-AUROC/P-AUROC (%); sparsemax values match Tab. 1, and standard deviations are omitted.

Dataset	Metric	1-shot				2-shot				4-shot			
		Softmax	Entmax	Top-p	Sparsemax	Softmax	Entmax	Top-p	Sparsemax	Softmax	Entmax	Top-p	Sparsemax
MPDD	I	67.1	78.6	74.6	**78.8**	68.0	80.1	76.8	**80.3**	67.7	80.1	77.0	**80.3**
	P	89.0	95.3	94.6	**96.4**	88.6	95.7	95.0	**96.4**	88.4	95.9	95.3	**96.5**
BTAD	I	83.7	95.2	91.2	**95.4**	83.5	95.6	92.3	**95.8**	83.6	96.0	92.9	**96.2**
	P	90.5	96.8	96.1	**97.9**	90.2	97.3	96.6	**98.0**	90.1	97.6	97.0	**98.2**
RESC	I	81.3	92.8	88.8	**93.0**	79.2	91.3	88.0	**91.5**	76.5	88.9	85.8	**89.1**
	P	88.9	95.2	94.5	**96.3**	88.7	95.8	95.1	**96.5**	87.8	95.3	94.7	**95.9**
BraTS	I	70.5	82.0	78.0	**82.2**	70.9	83.0	79.7	**83.2**	67.9	80.3	77.2	**80.5**
	P	88.6	94.9	94.2	**96.0**	89.1	96.2	95.5	**96.9**	88.4	95.9	95.3	**96.5**

Table 8: Cross-dataset DINOv3 layer-selection analysis across 1/2/4-shot settings. I/P denote I-AUROC/P-AUROC (%); standard deviations are omitted.

Dataset	Metric	1-shot				2-shot				4-shot			
		$\{6\}$	$\{12\}$	$\{6,12\}$	$\{3,6,9,12\}$	$\{6\}$	$\{12\}$	$\{6,12\}$	$\{3,6,9,12\}$	$\{6\}$	$\{12\}$	$\{6,12\}$	$\{3,6,9,12\}$
MPDD	I	75.4	72.3	76.6	**78.8**	77.3	75.1	78.8	**80.3**	77.5	76.1	78.9	**80.3**
	P	94.9	93.2	96.1	**96.4**	95.2	94.0	96.2	**96.4**	95.4	94.1	96.3	**96.5**
BTAD	I	92.0	88.9	93.2	**95.4**	92.8	90.6	94.3	**95.8**	93.4	92.0	94.8	**96.2**
	P	96.4	94.7	97.6	**97.9**	96.8	95.6	97.8	**98.0**	97.1	95.8	98.0	**98.2**
RESC	I	89.6	86.5	90.8	**93.0**	88.5	86.3	90.0	**91.5**	86.3	84.9	87.7	**89.1**
	P	94.8	93.1	96.0	**96.3**	95.3	94.1	96.3	**96.5**	94.8	93.5	95.7	**95.9**
BraTS	I	78.8	75.7	80.0	**82.2**	80.2	78.0	81.7	**83.2**	77.7	76.3	79.1	**80.5**
	P	94.5	92.8	95.7	**96.0**	95.7	94.5	96.7	**96.9**	95.4	94.1	96.3	**96.5**



(a) `carpet` (b) `grid` (c) `hazelnut` (d) `leather`

(e) `bottle` (f) `tile` (g) `wood` (h) `zipper`

Figure 3: **t-SNE visualization of selected-layer dinov3_vitb16 <CLS> embeddings** for normal vs. anomalous samples across eight MVTecAD categories.

ViT-B/16 backbone and 448×448 input resolution. For consistency, every sparsemax entry is identical to the corresponding DINOv3 result in Tab. 1, and standard deviations are omitted.

B.2 Cross-Dataset DINOv3 Layer Selection

We further report the layer-selection comparison on MPDD, BTAD, RESC, and BraTS, complementing the MVTecAD and VisA results in Tab. 3. All configurations use DINOv3 ViT-B/16 with 448×448 inputs. The $\{3, 6, 9, 12\}$ entries are aligned with the corresponding DINOv3 results in Tab. 1, and standard deviations are omitted.

Table 9: The effects of λ across different shots on MVTecAD Bergmann et al. (2019), evaluated by **I-AUROC** (%), **I-F1-max** (%), and **I-AP** (%).

λ	1-shot			2-shot			4-shot		
	I-AUROC	I-F1	I-AP	I-AUROC	I-F1	I-AP	I-AUROC	I-F1	I-AP
0	90.3	93.2	94.3	91.2	94.0	94.7	91.5	94.3	95.0
0.5	**96.3**	**97.6**	**98.4**	**96.8**	**97.7**	**98.5**	**96.9**	**97.9**	**98.7**
1	95.7	97.1	98.1	96.0	97.3	98.3	96.3	97.6	98.5

Table 10: The effects of different pooling operators across shots on MVTecAD Bergmann et al. (2019), evaluated by **I-AUROC** (%), **I-F1** (%), and **I-AP** (%).

Pooling operator	1-shot			2-shot			4-shot		
	I-AUROC	I-F1	I-AP	I-AUROC	I-F1	I-AP	I-AUROC	I-F1	I-AP
Top 10	95.3	95.9	97.5	95.8	96.6	97.8	95.9	97.0	97.9
Top 1%	**96.5**	**97.7**	**98.4**	**96.9**	97.6	**98.5**	96.8	**97.9**	98.5
max	96.3	97.6	98.4	96.8	**97.7**	**98.5**	**96.9**	**97.9**	**98.7**

C Effect of the Fusion Strategy

Tab. 9 studies the fusion weight λ in our Dual-Branch Image Scoring, where λ=0 uses only the Global Semantic Branch (S_{cls}), λ=1 uses only the Spatial Evidence Branch (S_{map}), and λ=0.5 fuses both. When relying solely on global <CLS> semantics (λ=0), the method remains functional but clearly suboptimal, suggesting that support-to-query <CLS> cosine matching encodes a meaningful normality prior, yet lacks the spatial evidence needed for robust anomaly assessment. This observation is consistent with Fig. 3, where normal and anomalous samples are well separated in the <CLS> embedding space across multiple categories, indicating discriminative global representations while not guaranteeing accurate localization-aware scoring. Conversely, using only spatial cues from the anomaly map (λ=1) performs strongly (1-shot: 95.7%/97.1%/98.1%) but is still inferior to the fusion, as isolated high responses or local clutter may dominate map pooling. Notably, the balanced fusion λ=0.5 consistently achieves the best results across all shots, reaching 96.3%/97.6%/98.4% in 1-shot and 96.9%/97.9%/98.7% in 4-shot, confirming that global <CLS>-level semantics and spatial localization cues are complementary: the former stabilizes image-level decisions under distributional variations, while the latter injects fine-grained spatial evidence to improve reliability in challenging cases.

D Effect of the Pooling Operator

Tab. 10 evaluates different pooling operators for reducing the anomaly map A_{map} into the map-based image score S_{map}. Overall, **max** pooling and Top-1% pooling yield very similar performance across shots. In the 1-shot setting, Top-1% is slightly better in I-AUROC (96.5% vs. 96.3%) and I-F1 (97.7% vs. 97.6%), while tying I-AP at 98.4%. This suggests that the anomaly map produced by our Sparse Hyper Matching is already highly concentrated on truly anomalous regions and thus benefits only marginally from aggregating multiple high-response locations. In contrast, Top-10 pooling is consistently inferior across all shots, indicating that fixed-cardinality pooling is less reliable in our setting, likely because the number of informative high-score patches varies with anomaly scale and background complexity, so a fixed n can either include spurious responses (hurting specificity) or under-summarize extended anomalies (hurting sensitivity). Despite the small gain of Top-1% over max pooling on MVTecAD, we adopt **max** pooling as the default to improve robustness in practice: percentile-based pooling introduces an extra ratio hyperparameter whose optimal value can vary with image resolution, category, and anomaly size, whereas max pooling is hyperparameter-free and thus avoids dataset-specific tuning while remaining highly competitive (*e.g.*, 4-shot: 96.9%/97.9%/98.7% for I-AUROC/I-F1-max/I-AP).

E More Qualitative Results

Figs. 4 to 8 provide additional 1-shot comparisons between ground-truth (GT) annotations and anomaly maps predicted by Hyper-FSAD*. Across industrial and medical images, Hyper-FSAD* consistently concentrates high anomaly responses within the annotated regions, covering compact, elongated, multiple, and diffuse anomalies. Together, they demonstrate robust localization across diverse anomalies without target-task training or language supervision.

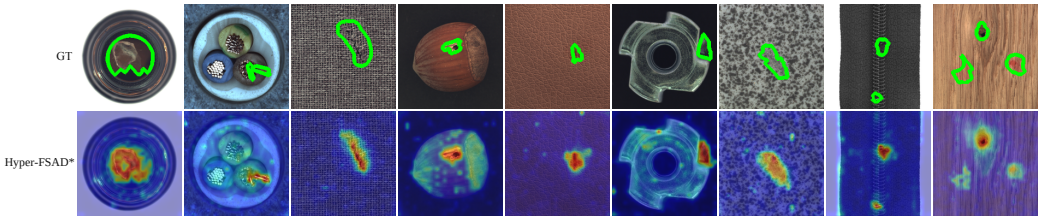


Figure 4: Additional 1-shot comparison between GT annotations (top) and Hyper-FSAD* anomaly maps (bottom), set 1. Hyper-FSAD* denotes our DINOv3 ViT-B/16 instantiation.

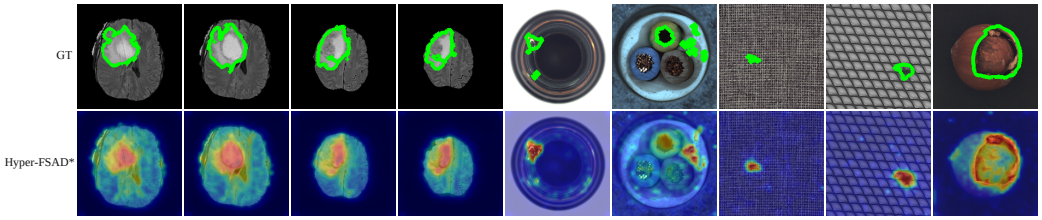


Figure 5: Additional 1-shot comparison between GT annotations (top) and Hyper-FSAD* anomaly maps (bottom), set 2.

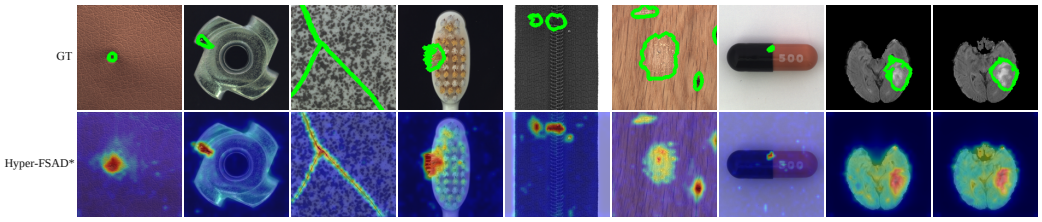


Figure 6: Additional 1-shot comparison between GT annotations (top) and Hyper-FSAD* anomaly maps (bottom), set 3.

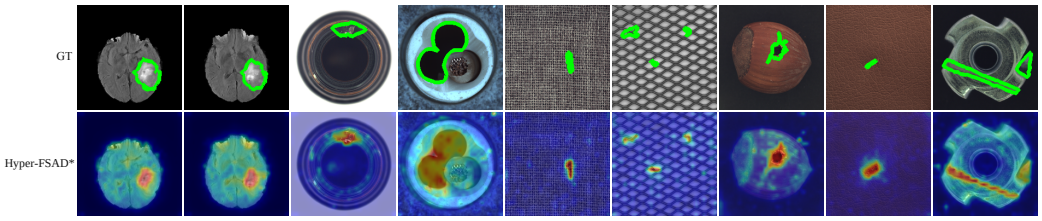


Figure 7: Additional 1-shot comparison between GT annotations (top) and Hyper-FSAD* anomaly maps (bottom), set 4.

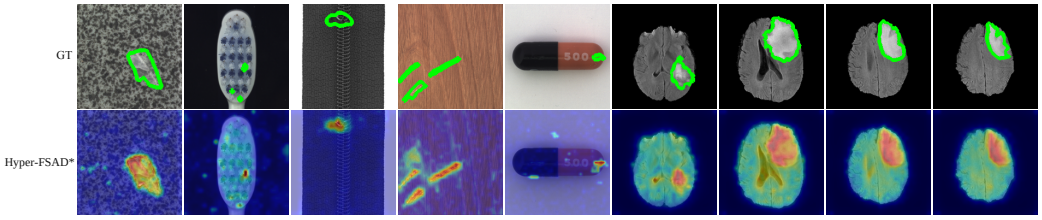


Figure 8: Additional 1-shot comparison between GT annotations (top) and Hyper-FSAD* anomaly maps (bottom), set 5.

F Input Resolution Analysis

F.1 The effects of input resolution.

We study the performance–efficiency trade-off across five input resolutions from 224×224 to 672×672 under the 1/2/4-shot settings. As shown in Tab. 11, higher resolutions generally benefit both datasets, especially in P-AUROC, because finer spatial details are retained for patch-level matching. The improvement is not strictly monotonic, however, and largely saturates beyond 448×448. Under the 4-shot setting, increasing the resolution from 448×448 to 672×672 changes MVTecAD I/P-AUROC from 96.9%/98.0% to 96.7%/98.1% and VisA I/P-AUROC from 94.8%/98.3% to 94.8%/98.5%. Thus, the higher resolution yields only 0.1 and 0.2 percentage-point gains in P-AUROC, respectively, while MVTecAD I-AUROC decreases by 0.2 points and VisA I-AUROC remains unchanged. Meanwhile, runtime increases more than fourfold, from 59.3 ms to 269.2 ms. Similar saturation is observed in the 1- and 2-shot settings. Because runtime is determined by the shared backbone, resolution, and inference pipeline, the same values apply to both datasets. We therefore adopt 448×448 as the default, which preserves strong image- and pixel-level performance while avoiding the substantial overhead of higher resolutions.

Table 11: Input-resolution ablation on MVTecAD (M) and VisA (V). I/P: I/P-AUROC (%); T: shared runtime (ms). Best/second-best values are bolded/underlined.

Shot	Resolution	M (I/P)	V (I/P)	T
1-shot	224×224	95.4/97.1	93.1/97.4	**38.0**
	336×336	96.0/97.4	93.5/97.7	<u>44.2</u>
	448×448	96.3/**97.7**	94.0/97.7	52.6
	560×560	<u>96.6</u>/97.5	<u>94.3</u>/<u>98.0</u>	88.6
	672×672	**96.7**/<u>97.6</u>	**94.5**/**98.2**	165.8
2-shot	224×224	95.8/97.4	94.0/97.8	**38.2**
	336×336	<u>96.3</u>/97.7	94.4/97.8	<u>44.3</u>
	448×448	**96.8**/<u>97.8</u>	<u>94.8</u>/98.0	53.1
	560×560	**96.8**/<u>97.8</u>	<u>94.8</u>/<u>98.2</u>	91.4
	672×672	**96.8**/**97.9**	**94.9**/**98.3**	195.1
4-shot	224×224	96.0/97.6	93.9/97.9	**38.4**
	336×336	96.4/<u>98.0</u>	94.6/98.1	<u>44.6</u>
	448×448	**96.9**/<u>98.0</u>	<u>94.8</u>/<u>98.3</u>	59.3
	560×560	<u>96.7</u>/**98.1**	**94.9**/**98.5**	106.4
	672×672	<u>96.7</u>/**98.1**	<u>94.8</u>/**98.5**	269.2