DO WORDS MATTER? WHAT TEXT CONTRIBUTES TO FEW AND FULL-SHOT VISION–LANGUAGE ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language anomaly detectors describe normal and abnormal appearance with text, and their gains are commonly credited to the semantic content of these descriptions. This attribution is plausible in the zero-shot setting, but few and full-shot detectors interpose a trained visual adapter, and in that setting it has not been isolated. Changing a description changes both its meaning and its numerical geometry, so accuracy comparisons cannot separate the two. We ask what a fixed pair of text embeddings, or *anchors*, contributes to such a detector: semantic content, effective scale, or joint anchor geometry. For a single anchor pair, the answer is effective scale alone. The score reads the anchors only through their difference, and a trained adapter can rotate to follow any replacement direction, leaving that direction, and with it the linguistic content, unidentifiable. We prove that replacement anchors with the same difference-vector length permit exactly the same probability maps and attainable losses. Across seven medical datasets, three vision backbones, both supervision regimes, and three adapter families, scale-matched random and orthogonal anchors track correct prompts; breaking the scale match opens a clear gap that restoring the scale closes without text. The same pattern holds empirically in MVFA's residual, multi-level detector and when MadCLIP's learned prompts are replaced by a text-free learned scale. When one adapter serves many categories, the answer expands to their joint arrangement. Rotating all category axes together does not change anything, as predicted. For the tested CLIP prompt banks, these axes are nearly collinear, and meaning-free rotations that spread them apart improve performance on both industrial benchmarks. Even under frozen zero-shot transfer, a random anchor frame transfers as well as the prompt frame: coordinate consistency, not meaning. What these detectors read from their prompts is scale and arrangement, not semantic content.

1 INTRODUCTION

Anomaly detection asks a model to identify appearances it has rarely or never seen. Because anomalies are rare, visually diverse, and seldom labeled, the field has largely trained detectors on normal data alone, scoring a test image by how far it departs from the resulting model of normality (Roth et al., 2022; Deng & Li, 2022; Gudovskiy et al., 2022; Salehi et al., 2021); supervising anomalies directly only shifts the scarcity, since labeled anomalous examples are rarer still (Ding et al., 2022; Yao et al., 2023). Such detectors carry no explicit representation of what *constitutes* abnormality, and must define it negatively, as departure from a reference distribution that has to be estimated exhaustively from images. The cost is a per-category fit: every new object or organ needs its own corpus of normal data, and accuracy degrades when only a handful of examples are available (Huang et al., 2022). Vision–language models supply the missing half. Contrastive pretraining on web-scale image–text pairs places visual and textual concepts in a shared space (Radford et al., 2021), so the reference for *abnormal* can be written in language instead of estimated from target-domain images. This both enables low-sample operation and lets a single category-agnostic detector be re-targeted by changing the noun in a prompt.

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.

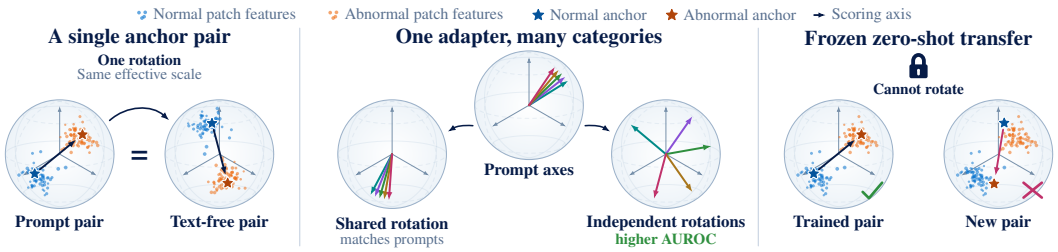

Figure 1: **What the text anchors contribute to a trained fixed-anchor detector.** The score reads the two anchors only through their difference, the scoring axis. **Left:** a separately trained adapter rotates to follow any replacement axis of the same effective scale, so its direction, and with it the semantic content, is unidentifiable. **Middle:** with one adapter serving many categories, a shared rotation preserves the arrangement of their axes; independent rotations keep each scale but change the arrangement. **Right:** a frozen adapter cannot rotate, so it must keep the anchor frame it was trained with, which need not be linguistic. Embeddings are schematic.

These detectors perform well in practice. Handcrafted prompts give competitive zero-shot results on industrial benchmarks (Jeong et al., 2023); trained visual projections or learned prompts improve zero- and few-shot performance (Chen et al., 2023; Zhou et al., 2024); and the approach extends to medical imaging, where labeled anomalies are scarcest (Huang et al., 2024; Shiri et al., 2025). Language-conditioned detection is thus common in the low-label regime. The trained fixed-anchor subclass studied here uses two prescribed embeddings, or *anchors*, for its normal–abnormal readout.

The field typically attributes these gains to semantic factors, including anomaly semantics, language-compatible features, or vision–language alignment (Li et al., 2024; Zuo et al., 2024; Huang et al., 2024; Zhang et al., 2024; Shiri et al., 2025). In a training-free detector, the anchors directly define the scoring rule: no learned visual map can compensate when their coordinates change. The fixed-anchor few- and full-shot setting differs: real or synthetic anomaly targets train a visual adapter against the anchors, letting it absorb changes in their direction. Because changing a description alters both its meaning and its numerical geometry, ordinary prompt comparisons cannot determine which anchor properties remain observable after adaptation. Curated prompt banks cost clinical expertise for each new dataset and finding; if anchors contribute only geometry, pursuing better descriptions and text encoders targets the wrong variable. This motivates our central question: *For trained fixed-anchor vision–language anomaly detectors, what information from the text anchors actually affects representational capacity and observed performance: semantic content, effective scale, or joint anchor geometry?* We answer this by identifying what the score can read, then testing whether training respects it. The binary score depends on the anchors only through their difference: a direction, which a trained adapter can rotate its output to follow (Figure 1, left), and a magnitude, which it cannot. We prove the direction is therefore unidentifiable, and probe the claim with five anchor conditions (correct medical prompts, a mismatched-dataset bank, and three text-free pairs that preserve or break the surviving magnitude) across seven medical datasets, two supervision regimes, three vision backbones and three adapter families. We then find the same scale-controlled pattern in MVFA's residual multi-level detector and in MadCLIP after replacing its learned prompt path with a text-free learned scale. Two further settings test the boundaries: one adapter serving many categories, where the invariant becomes the joint geometry of their difference vectors (Figure 1, middle), and frozen zero-shot transfer, where the adapter can no longer rotate.

Our contributions are:

Single-pair identification. Only the effective scale is identifiable. Scale-matched random and orthogonal anchors track correct medical prompts with mean gaps of 0.16%–0.26% image AUROC; breaking the scale opens a clear gap that restoring it closes with no text.

Shared multi-category geometry. The invariant expands to the Gram matrix of the category difference vectors. The tested CLIP prompt axes are nearly collinear, and spreading them apart, while preserving every category's scale and adding no meaning, improves both industrial benchmarks, by 5.6% image and 4.4% pixel AUROC on VisA and up to 8.8% and 8.3% across vision backbones.

Frozen zero-shot transfer boundary. With the adapter frozen, its anchor frame is load-bearing but need not be linguistic: a retained random frame averages within 0.5% AUROC of prompts.

2 Related Work

Vision–language anomaly detectors differ in where adaptation occurs. WinCLIP scores handcrafted prompts with frozen encoders (Jeong et al., 2023); APRIL-GAN learns visual projections into a fixed text space (Chen et al., 2023); PALADIN aligns DINOv3 patch features to the CLIP text space for prompt-guided localization (Basaran, 2026); and AnomalyCLIP learns object-agnostic prompt tokens (Zhou et al., 2024). Medical systems likewise combine prescribed prompt ensembles with residual visual adapters (Huang et al., 2024), or jointly learn prompt tokens and visual adaptation (Zhang et al., 2024; Shiri et al., 2025), while AnomalyDINO removes text entirely (Damm et al., 2025). These distinctions determine whether a changed text readout can be compensated by a trainable visual branch. Prompt-based detectors commonly credit anomaly semantics or vision–language alignment (Li et al., 2024; Zuo et al., 2024; Huang et al., 2024; Zhang et al., 2024; Shiri et al., 2025), yet strong visual-only systems show that accuracy alone cannot identify a contribution from language (Roth et al., 2022; Deng & Li, 2022; Gudovskiy et al., 2022; Salehi et al., 2021; Liu et al., 2023; Li et al., 2021; Schlüter et al., 2022; Huang et al., 2022). Learned-prompt methods further blur attribution because their text representations are fitted rather than prescribed (Zhou et al., 2022b;a; Khattak et al., 2023). Evidence on textual content is mixed even with frozen visual features: semantic descriptors can improve classification and make its criteria editable (Menon & Vondrick, 2023), whereas random words can match curated descriptors (Roth et al., 2023). Neither setting studies fixed text anchors with a separately trainable visual branch. Identifiability analyses more generally characterize representations only up to transformations that leave the observable model unchanged (Roeder et al., 2021). The remaining question is which properties of fixed text embeddings stay identifiable after visual adaptation.

3 Preliminaries

This section fixes notation and describes the detector family the paper analyzes: how frozen image patches, a trainable adapter, and a prescribed pair of text embeddings produce an anomaly map.

Notation and the studied detector. A frozen vision backbone b exposes P patch features $\boldsymbol{z}_{1:P}^{(\ell)}$, one per token of the patch grid, at each of L retained layers indexed by ℓ. Each layer has its own trainable adapter $g^{(\ell)}$ into the d-dimensional anchor space, and its normalized adapted patch is

$$\boldsymbol{z}_{1:P}^{(\ell)} = b^{(\ell)}(x), \qquad \hat{\boldsymbol{u}}_p^{(\ell)} = \mathrm{normalize}\Big(g^{(\ell)}(\boldsymbol{z}_p^{(\ell)})\Big) \in \mathbb{R}^d. \tag{1}$$

The scoring head contains two prescribed unit vectors, $\boldsymbol{a}_N$ for normal and $\boldsymbol{a}_A$ for abnormal, which we call *anchors*. For text conditions, each is the normalized mean of a fixed prompt ensemble encoded by the frozen text encoder; random-coordinate controls construct the two vectors directly. In either case the vectors remain fixed throughout a run. Patch p at layer ℓ receives the two cosine logits

$$s_{p,N}^{(\ell)} = \tau \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \boldsymbol{a}_N \right\rangle, \qquad s_{p,A}^{(\ell)} = \tau \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \boldsymbol{a}_A \right\rangle. \tag{2}$$

Here $\tau > 0$ is the multiplicative similarity, or logit, scale; under the conventional softmax parameterization it is an inverse temperature rather than a temperature.

From logits to anomaly maps. The patch-grid logits are reshaped and bilinearly upsampled to image resolution before the two-class softmax. Its abnormal channel is

$$p_A^{(\ell)} = \frac{\exp s_A^{(\ell)}}{\exp s_N^{(\ell)} + \exp s_A^{(\ell)}} = \sigma\Big(s_A^{(\ell)} - s_N^{(\ell)}\Big). \tag{3}$$

Training objective and evaluation. Focal and Dice losses (Lin et al., 2017; Milletari et al., 2016) train every layer map. At inference, the L abnormal maps are averaged; pixel AUROC reads this averaged map, while the image score is the mean of its top 1% pixels. Appendix A.1 verifies the interpolation, loss, and evaluation steps exactly.

4 THEORETICAL ANALYSIS

The goal of this section is to identify which properties of the prompt embeddings actually determine the detector's capabilities, and thereby separate the contribution of linguistic meaning from that of embedding geometry. We ask when text-derived anchors can be replaced by text-free vectors without changing the representable abnormal-probability maps. We first answer this for one normal–abnormal pair, then for several pairs sharing an adapter.

4.1 ONE ANCHOR PAIR: SEPARATING DIRECTION FROM SCALE

What does the detector read from its anchors? By equation 3, only the abnormal logit minus the normal logit determines the abnormal probability. The anchors therefore enter through their difference, $\boldsymbol{\delta} = \boldsymbol{a}_A - \boldsymbol{a}_N$, the *scoring axis*. With anchor cosine $\rho = \langle \boldsymbol{a}_N, \boldsymbol{a}_A \rangle$, we separate its length from its direction:

$$s_{p,A}^{(\ell)} - s_{p,N}^{(\ell)} = \tau \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \boldsymbol{\delta} \right\rangle = \underbrace{\tau\sqrt{2-2\rho}}_{\lambda_{\text{eff}}} \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \hat{\boldsymbol{\delta}} \right\rangle, \qquad \hat{\boldsymbol{\delta}} = \frac{\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2}. \tag{4}$$

The inner product measures an adapted patch's alignment with the scoring axis; positive values favor abnormality. The factor λ_{eff}, the *effective logit scale* (hereafter effective scale), sets how strongly this alignment changes the probability. Closer anchors give a smaller effective scale.

Suppose we rotate the anchors and then retrain the adapter. Training now encourages higher cosine alignment with the new scoring axis for abnormal patches and lower alignment for normal patches. An adapter that can rotate its outputs with that axis can preserve the original cosine values (Figure 1, left). Normalization fixes the output length, however, so rescaling the adapter cannot absorb a change in axis length. This motivates replacing linguistic coordinates while preserving effective scale.

An adapter family $\mathcal{G}$ is *rotation-closed* if every rotation of an adapter's outputs can be achieved by changing its parameters. One matrix $\boldsymbol{Q} \in O(d)$ rotates the outputs of every layer; $O(d)$ is the set of d-dimensional orthogonal matrices, which preserve lengths and angles. A freely trained final linear map absorbs $\boldsymbol{Q}$, so our linear, MLP, and Transformer adapters all qualify (Appendix A.8).

This gives a direct test of prompt meaning: if replacement anchors preserve the effective scale, can a retrained adapter reproduce every original abnormal-probability map? Theorem 1 below answers this question for representable probability maps.

Theorem 1 (Single-pair equivalence and effective scale)**.** *Consider unit anchor pairs $(\boldsymbol{a}_N, \boldsymbol{a}_A)$ and $(\boldsymbol{a}'_N, \boldsymbol{a}'_A)$, the original and replacement, used with positive similarity scales τ and τ'. Hold each pair fixed within its run and train a separate adapter from a rotation-closed family. If*

$$\tau \|\boldsymbol{a}_A - \boldsymbol{a}_N\|_2 = \tau' \|\boldsymbol{a}'_A - \boldsymbol{a}'_N\|_2, \tag{5}$$

then the two choices permit exactly the same category-agnostic abnormal-probability maps. Any training objective or evaluation statistic computed only from those maps has the same attainable values under both choices. Its infimum and supremum coincide, as does its optimum whenever attained. At fixed τ, matching the anchor cosine ρ is sufficient; more generally the surviving scalar is the effective scale $\lambda_{\text{eff}} = \tau\sqrt{2-2\rho}$.

Thus, scale-matched text-free anchors can reproduce every map achievable with prompts, and vice versa. Equal representational capabilities need not yield identical training outcomes: stochasticity and finite optimization can lead to different solutions. Our experiments test whether the two anchor choices achieve comparable performance. Appendices A.3 and A.9 give the proof and explain possible residuals.

4.2 SEVERAL ANCHOR PAIRS: WHY THEIR ARRANGEMENT MATTERS

On industrial anomaly detection datasets such as MVTec-AD, one adapter serves object categories such as bottles and cables, each with its own anchor pair. Rotating all scoring axes together still allows one compensating adapter rotation. Rotating them independently can change their relative angles, which no common rotation can undo (Figure 1, middle). We therefore need to track their joint arrangement, as well as each scale. Category t selects only its own unit anchor pair $(\boldsymbol{a}_t^N, \boldsymbol{a}_t^A)$,

with cosine ρ_t; all T categories share the adapter. Stack their scoring axes $\boldsymbol{\delta}_t = \boldsymbol{a}_t^A - \boldsymbol{a}_t^N$ as rows of $\boldsymbol{D}$. Their pairwise inner products form a *Gram matrix*, recording lengths and alignments:

$$\boldsymbol{M} = \boldsymbol{D}\boldsymbol{D}^\top, \qquad M_{ts} = \langle \boldsymbol{\delta}_t, \boldsymbol{\delta}_s \rangle, \qquad M_{tt} = 2 - 2\rho_t. \tag{6}$$

The diagonal fixes each category's scale at a given τ; the off-diagonals record how axes align. If two banks share $\boldsymbol{M}$, one rotation maps every original scoring axis to its replacement. A shared adapter can follow that one rotation. The corollary states what this implies for the probability maps of all categories.

Corollary 1 (Shared multi-category extension)**.** *Let two category-conditioned anchor banks have difference matrices $\boldsymbol{D}$ and $\boldsymbol{D}'$. If $\boldsymbol{D}\boldsymbol{D}^\top = \boldsymbol{D}'\boldsymbol{D}'^\top$, then separately trained rotation-closed adapters permit exactly the same category-conditioned abnormal-probability maps under the two banks at the same similarity scale.*

Our experiments therefore use shared rotations to test equivalence at fixed $\boldsymbol{M}$, and independent rotations to test the effect of changing the arrangement at fixed category scales. The corollary predicts no direction for that effect. Here $\boldsymbol{M}$ characterizes the axes up to a common orthogonal transformation; matching it is sufficient, but not universally necessary, for equal capabilities (Appendix A.4).

When the argument stops applying. Both results require fixed anchors, losses and metrics computed from cosine-based binary probability maps, and an adapter able to follow changed anchors. The next proposition collects cases where the compensating rotation is unavailable. Learned prompts fall outside the fixed-anchor premise; these conclusions concern the studied detector family.

Proposition 1 (Proof in Appendix A.10)**.** *The equivalence guarantees above no longer follow if the adapter family cannot absorb the required output rotation or if the adapter is frozen. In shared multi-category scoring, they also do not follow when the replacement changes $\boldsymbol{D}\boldsymbol{D}^\top$, because no single rotation then maps all scoring axes to their replacements. None of these conditions guarantees different performance or a benefit from correct text; each only removes the equivalence guarantee.*

5 Method: Anchor Controls

Suppose we train an adapter with prompts and then replace their anchors with scale-matched, text-free vectors. Theorem 1 says we *could* rotate the trained adapter's outputs to keep every abnormal-probability map unchanged. In our experiments, however, we do not apply that rotation: we train a new adapter from random weights for a fixed schedule. Its initialization is tied to coordinate axes, and Adam rescales updates coordinate by coordinate, so the new run need not follow the corresponding rotated path. Stochastic batches and limited numerical precision can add further variation (Appendix A.9). We therefore measure whether ordinary training reaches similar held-out performance without words, then change scale, joint geometry, and the ability to retrain the adapter to test the limits of this result.

Few-shot medical controls. We first test the single-pair result with a newly trained adapter for every anchor choice. Correct medical prompts provide the reference: for each dataset, complete normal and abnormal description banks specify its anatomy, imaging modality, state definitions, and templates. We encode each bank, average and normalize its embeddings, and obtain the reference cosine ρ_{correct} and effective scale λ_{eff}. All anchor pairs stay fixed throughout training; learning them would also change ρ. We repeat the comparisons across supervision regimes, vision backbones, and adapter families to check finite-training behavior. The decisive test removes language while keeping exactly what the theorem says matters. A seeded random unit-vector pair with cosine ρ_{correct} bypasses the text encoder but preserves λ_{eff}. These *Gram-matched random anchors* permit the same probability maps as correct prompts. Any performance gap after separately training the adapters is therefore a finite-training residual, not a loss of representational capability.

Next we ask whether scale, rather than words, explains a gap when anchors are replaced. *Unmatched orthogonal anchors* are seeded, perpendicular unit vectors with zero cosine. They also contain no text, but change λ_{eff}. For *matched orthogonal anchors*, we reuse those same vectors and multiply the similarity scale during training and evaluation by $\sqrt{1 - \rho_{\text{correct}}}$. This restores the original λ_{eff} without restoring text. The orthogonal pair is thus a scale-changing control, and its matched version is a second exact probability-map-level control: recovery after scale matching implicates scale rather than linguistic coordinates.

Finally, a *mismatched-dataset bank* tests a more recognizable semantic error. We replace the target's complete description banks with those of a fixed donor, changing anatomy, modality, and state descriptions together; for example, Kaggle BrainMRI is scored with LiverCT banks. This is a stress test, not an exact test of meaning, because the donor need not match ρ_{correct}. A difference cannot isolate semantics, and a null result cannot show that every word is ignored. We fix donor mappings and full-bank construction before training. The two scale-matched text-free conditions carry the semantic-necessity test; the mismatched-dataset bank provides complementary evidence. The audit follows controlled identifiability and capacity-matched controls (Locatello et al., 2019; Hewitt & Liang, 2019) to test a possible geometric shortcut (Geirhos et al., 2020). Section 6.1 gives the protocol; Appendices A.5 and A.7 document the donor banks and geometry.

Architectural and model extensions. The exact single-pair result assumes fixed anchors and a visual adapter that can absorb output rotations. We test how far its observed scale dependence carries as the detector becomes more complex. First, approximately parameter-matched linear, MLP, and contextual Transformer adapters retain rotation closure but differ in nonlinear capacity and patch interaction. This tests whether the result depends on our linear head. We then retain the published visual and scoring pipelines of MVFA (Huang et al., 2024) and MadCLIP (Shiri et al., 2025). MVFA inserts residual adapters at multiple levels of one CLIP image encoder and uses separate detection and segmentation readouts; its prescribed anchors permit the same scale controls. MadCLIP instead has separate normal and abnormal adaptation branches within one CLIP image encoder and learns its prompts, so its anchor cosine is fitted. There we replace the prompt path with a text-free learned scale on a random frame. These studies test empirical reach where the fixed-anchor theorem alone does not guarantee equivalence. Appendix C specifies the arms, optimization, scoring, and checkpoint rules.

Unified industrial controls. When one adapter serves many industrial categories, matching each pair's scale is not enough to guarantee equivalence: Corollary 1 also requires the arrangement of their scoring axes to stay fixed. We therefore compare two geometric interventions. A *shared rotation* applies one orthogonal transformation to the whole bank, preserving $\boldsymbol{M}$. A single adapter rotation can compensate, so this is the exact control for finite-training invariance. *Independent rotations* reorient categories separately. They preserve every diagonal M_{tt}, hence every category's effective scale, but generally change the off-diagonal alignments. Their performance tests whether joint geometry matters when individual scales are held fixed.

Frozen zero-shot medical transfer. Theorem 1 requires the adapter to follow a new anchor frame. To test this boundary, we train an adapter on a medical source dataset, freeze it, and score a target without target-side training. We either retain the source pair or substitute the target pair. Comparing those swaps for correct and random anchors asks whether transfer depends on linguistic content or consistency with the adapter's trained coordinates (Section 6.5).

6 Experiments

6.1 Setup and reading rule

We report results in the order of the controls: text-free anchors after retraining, changes in anchor geometry, and adapters that cannot retrain. We separated datasets by their source of anomaly supervision. The real-anomaly-support regime used K normal and K observed abnormal samples, whereas the normal-only regime used K normal samples and generated synthetic anomalies online during training. We used the former for BraTS2021 (Baid et al., 2021), LiverCT (Bao et al., 2024), RESC (Hu et al., 2019), and OCT2017 (Kermany et al., 2018), with $K \in \{2, 4, 8, 16\}$; we used the latter for BUSI (Al-Dhabyani et al., 2020), CheXpert (Irvin et al., 2019), and Kaggle BrainMRI (Zhang et al., 2024; Chakrabarty, 2019), with $K \in \{4, 8, 16, 32\}$. Because the supervision mechanisms and support grids differ, we did not pool effect magnitudes, and only $K \in \{4, 8, 16\}$ supports a statement across both regimes. We used image AUROC (AC) as the primary metric and reported pixel AUROC (AS) only where valid masks exist. Within each table block, conditions shared the dataset, support size, vision backbone, adapter, and scoring protocol. We compared their three-seed means descriptively without treating closeness as a formal equivalence test. For each control, we report the mean and largest absolute correct-minus-control image-AUROC difference within each supervision regime; signed differences only describe direction (positive favors correct

Table 1: CLIP ViT-L/14 anchor substitution at $K{=}16$ in both supervision regimes. **Left:** real-anomaly support (16 normal + 16 observed abnormal samples). **Right:** normal-only (16 normal samples with online synthetic anomalies). Cells are image-AUROC (AC) mean±sd over three seeds (%). Section 5 defines the five anchor conditions. Pixel AUROC, the remaining support sizes, and the complete factorials for all three vision backbones are in Appendix Tables 27 and 28.

	Real-anomaly support					Normal-only			
Anchor condition	**OCT2017**	**BraTS2021**	**LiverCT**	**RESC**	**Avg**	**Kaggle BrainMRI**	**CheXpert**	**BUSI**	**Avg**
Correct prompts	99.6±0.1	96.9±0.2	91.6±1.1	97.2±0.9	96.3	92.9±0.8	69.3±1.6	88.2±0.6	83.4
Mismatched-dataset bank	99.7±0.1	96.9±0.2	90.8±1.3	97.2±0.9	96.1	92.9±0.7	69.0±2.3	87.7±0.9	83.2
Gram-matched random	99.6±0.1	96.9±0.2	91.5±1.2	97.2±0.9	96.3	92.9±0.2	69.5±1.3	88.4±0.6	83.6
Unmatched orthogonal	99.2±0.0	97.0±0.1	89.0±0.3	97.1±1.0	95.6	92.5±0.4	69.8±1.3	87.8±0.6	83.4
Matched orthogonal	99.6±0.1	96.9±0.2	91.6±1.4	97.2±0.9	96.3	92.9±0.7	69.4±1.7	87.8±0.8	83.4

prompts). The main tables show CLIP ViT-L/14 (Radford et al., 2021), the one configuration whose visual features and text anchors share a pretrained joint space and in which semantic anchors are therefore most likely to help; Appendix D provides the complete mean±sd factorials for all three vision backbones while Appendix B gives the training, checkpoint, and evaluation protocol.

For each industrial benchmark, we trained one full-shot CLIP ViT-L/14 detector using all normal training images and synthetic anomalies. The 15 MVTec-AD (Bergmann et al., 2019) or 12 VisA (Zou et al., 2022) category-specific normal–abnormal pairs shared its category-agnostic linear adapters. We scored each sample only against its own pair and macro-averaged pixel and image AUROC across categories. Appendices A.6 and B.1 specify the exact industrial prompt banks and synthetic-anomaly generator.

6.2 Few-shot medical anchor controls: effective scale, not prompt content

If prompt meaning helps after retraining, correct prompts should outperform the two exact text-free controls despite their matched effective scale. Pooled over three vision backbones, their mean absolute image-AUROC gaps were only 0.16% and 0.18% with real-anomaly support (maxima 1.2% and 1.0%), and 0.22% and 0.26% with normal-only training (maxima 0.9% and 0.7%), for Gram-matched random and matched orthogonal anchors respectively. These small absolute gaps, without a consistent advantage for correct prompts, show that finite training closely tracked the theorem's attainable equivalence in the studied settings. Table 1 reports CLIP ViT-L/14; Appendix Tables 27 and 28 report all three vision backbones.

The next comparison asks what happens when we change scale while keeping the anchors text-free. Without scale matching, the mean absolute gap increased to 1.46% with real-anomaly support and 0.93% with normal-only training (maxima 11.8% and 3.4%); the 11.8% maximum occurred on DINOv3-L/16 (Siméoni et al., 2025) OCT2017 at $K = 2$ (93.6% versus 81.8%). Restoring scale with the same orthogonal vectors restored the near-zero gaps above, identifying scale rather than text removal as the source of this deterioration. The mismatched-dataset bank changed the complete descriptions yet differed by at most 1.7% and 1.3%.

6.3 Extensions across adapter families, residual detectors, and learned prompts

Rotation-closed adapter families. We next ask whether the result depends on adapter architecture. The approximately parameter-matched linear, MLP, and contextual Transformer adapters differ in nonlinear capacity and patch interaction but remain rotation-closed (Appendix A.8). In regime averages across two vision backbones, correct-prompt versus Gram-matched random gaps stayed within 0.5% image AUROC, with no consistent sign. Unmatched orthogonal anchors instead reduced contextual-Transformer performance by 5.4% on CLIP ViT-L/14 and 18.7% on DINOv3-L/16 with real-anomaly support; matching their scale brought both losses back to within 0.2%. Thus the exact controls agree across three implementations that still satisfy the theorem's architectural premise. Appendix Tables 25 and 26 give the per-dataset results and parameter counts.

MVFA's prescribed prompts support three text-free controls: Gram-matched random anchors preserve their geometry, unmatched orthogonal anchors change effective scale, and scale matching re-

Table 2: External-detector anchor controls at $K{=}16$ (AUROC %, mean±sd over three seeds). AC is image and AS pixel AUROC; OCT2017 pixel AUROC is not reported, so it has no AS column and Avg AS averages BraTS2021, LiverCT, and RESC. Checkpoints follow each released method's test-selection rule; Appendix C gives the complete grids and provenance.

Method	Anchor condition	OCT2017	BraTS2021		LiverCT		RESC		Avg	
		AC	AC	AS	AC	AS	AC	AS	AC	AS
MVFA	Correct prompts	99.9±0.0	95.0±0.6	97.5±0.9	86.1±1.5	98.8±1.4	98.2±0.4	99.4±0.0	94.8	98.6
	Gram-matched random	99.9±0.0	94.1±1.5	97.6±0.1	83.7±2.5	99.1±0.9	98.0±0.2	99.4±0.1	93.9	98.7
	Unmatched orthogonal	92.2±2.1	66.8±4.0	93.1±1.1	63.0±5.2	96.5±0.3	72.9±9.1	89.9±2.4	73.7	93.2
	Matched orthogonal	99.9±0.1	94.5±1.1	97.5±0.6	85.5±1.8	99.0±1.0	98.0±0.3	99.3±0.1	94.5	98.6
MadCLIP	Learned prompts	99.6±0.2	96.0±0.5	97.6±0.4	90.5±0.9	99.5±0.5	98.6±0.2	99.4±0.1	96.2	98.8
	Learned scale (no text)	99.7±0.1	95.8±0.6	97.7±0.6	91.1±0.8	99.1±0.7	98.6±0.2	99.4±0.1	96.3	98.7
	Prompt-scale (no text)	99.6±0.1	96.1±0.5	98.1±0.1	87.7±3.4	99.1±1.0	97.7±0.2	99.2±0.1	95.3	98.8

stores it. Because MadCLIP learns its prompt pair, we compare its original implementation (learned prompts) with two text-free arms on a frozen random orthonormal frame: one learns the cosine (learned scale), the other fixes it at the prompt-like value $+0.986$ (prompt-scale).

MVFA: residual, multi-level visual adaptation. MVFA (Huang et al., 2024) inserts multiple residual adapter branches into one CLIP image encoder and uses separate readouts for detection and segmentation. Its prescribed anchors admit the same controls, although the residual architecture itself is outside the exact closure result. At $K{=}16$, mean image AUROC was 94.8% with prompts and 93.9% with Gram-matched random anchors. Unmatched orthogonal anchors reduced it to 73.7%; restoring only their scale recovered 94.5% (Table 2). This provides the clearest external drop-and-recovery example, as empirical evidence beyond the theorem's guaranteed scope.

MadCLIP: learned prompts. MadCLIP (Shiri et al., 2025) uses separate normal and abnormal adaptation branches in one CLIP image encoder and learns the prompt pair, making its cosine an outcome of training. A fixed-anchor replacement therefore cannot reproduce its internal protocol. At $K{=}16$, learned prompts and the text-free learned scale obtained 96.2% and 96.3% mean image AUROC. Across the full grid, their mean learned cosines were -0.984 and -0.994, respectively. Both paths approach antipodal anchors, the maximum effective scale: given freedom to learn its prompts, MadCLIP finds what a single scalar finds. The prompt-scale arm reached 95.3% at $K{=}16$ but was less stable at small K; Appendix Tables 20 and 21 report all image and pixel results.

6.4 Unified multi-category detection: joint anchor geometry

The industrial comparison asks whether a shared adapter needs more than each category's scale. The corollary points to the relative directions of the category scoring axes, but a table of pairwise angles would be hard to read. We summarize their spread with the participation-ratio effective rank. It is near 1 when the axes mostly lie along one direction and rises as they occupy more directions. For the nonnegative eigenvalues $\mu_1, \dots, \mu_T$ of $\boldsymbol{M}$, it is

$$r_{\text{eff}}(\boldsymbol{M}) = \frac{\left(\sum_{i=1}^{T} \mu_i\right)^2}{\sum_{i=1}^{T} \mu_i^2}, \qquad 1 \le r_{\text{eff}}(\boldsymbol{M}) \le \operatorname{rank}(\boldsymbol{M}). \tag{7}$$

The prompt axes were concentrated: $r_{\text{eff}} = 1.79$ on MVTec-AD and 1.78 on VisA (Appendix Table 16). Independent rotations raised these values to 14.28–14.29 of 15 and about 11.10 of 12, while preserving every category-specific scale. The shared rotation stayed within 0.1% of correct prompts on all four AUROC endpoints, consistent with its exact equivalence. Independent rotations instead gained 5.6% image and 4.4% pixel AUROC on VisA and 1.2% image and 3.0% pixel AUROC on MVTec-AD (Table 3). Thus, individual scales did not settle performance: in these benchmarks the spread-out bank worked better for the shared adapter. DINOv2-L/14 (Oquab et al., 2024) and DINOv3-L/16 reproduced the result on both benchmarks, with gains up to 8.8% image and 8.3% pixel AUROC on VisA; Appendix Table 22 reports all three vision backbones.

One plausible mechanism is reduced cross-category interference. Because the detector scores category t through $\langle \hat{\boldsymbol{u}}, \boldsymbol{\delta}_t \rangle$, concentrating difference vectors in shared directions may make category updates compete, whereas spreading them provides distinct readouts. We did not measure gradient alignment, so this mechanism remains unverified.

Table 3: Full-shot industrial results (%), mean±sd over three training seeds (and geometry seeds where applicable). AC and AS are image and pixel AUROC, category-macro means over the T categories; Appendix Table 22 adds PRO, AUPR, and F1max.

Anchor condition	VisA		MVTec-AD	
	AC	AS	AC	AS
Correct prompts	86.2±0.1	91.7±0.1	96.2±0.1	93.6±0.1
Shared rotation	86.3±0.1	91.6±0.0	96.2±0.1	93.6±0.1
Independent rotations	91.8±0.1	96.1±0.1	97.4±0.1	96.6±0.0

6.5 Frozen zero-shot transfer: the boundary of anchor equivalence

The final test asks what happens when an adapter cannot follow changed anchors. At $K = 16$, we trained each CLIP ViT-L/14 adapter on a medical source dataset, froze it, and evaluated a target without target-side training. We scored the same target images with either the source pair that trained the adapter or the target dataset's pair. Their AUROC difference (source minus target) is the anchor-swap cost, separate from the domain shift common to both scores (Appendix Table 23).

Retaining a text-free random source pair preserved transfer, whereas redrawing a random target pair without changing the frozen adapter was costly. Relative to correct prompts, retained random pairs differed by only 0.1% and 0.3% image AUROC on CLIP ViT-L/14 and 0.2% and 0.5% on DINOv3-L/16 in the real-anomaly-support and normal-only regimes. Random-pair swaps cost 9.5% and 19.1% on CLIP and 19.6% and 13.2% on DINOv3; the corresponding correct-prompt swaps cost 0.6% and 0.8% on CLIP, and -1.3% and -0.3% on DINOv3. Correct target prompts shared a text-encoder frame, whereas independently drawn random target pairs did not (Appendix Table 24).

Reusing the random source pair preserves the trained readout without linguistic coordinates. A frozen adapter cannot make the compensating change used in Theorem 1; swapping pairs can change direction and scale together, so these results establish the importance of keeping the adapter and its anchor pair coupled, not which geometric component causes the swap cost.

7 Conclusion

Summary. We asked which information from the text anchors of a trained fixed-anchor vision–language anomaly detector affects its representational capacity and observed performance: semantic content, effective scale, or joint anchor geometry. The answer is geometry, not semantic content. For a single anchor pair, a rotation-closed adapter reparameterizes the difference-vector direction away, so scale-matched random or orthogonal anchors permit exactly the maps of correct prompts and matched them after training, while changing only the scale opened gaps that restoring it closed. The same pattern held in MVFA and in MadCLIP with a text-free learned scale. When one adapter serves many categories, the joint geometry of their difference vectors matters: spreading the near-collinear prompt axes apart, without adding meaning, improved both industrial benchmarks. Under frozen transfer, a retained random frame transferred as well as prompts, because an adapter and its anchors form a coupled coordinate system.

Practical implications. Within this family, curating expert prompt banks or searching for better wording and text encoders cannot improve accuracy through meaning: a text-free anchor pair with matched or learned effective scale suffices, and a unified detector can improve by spreading its category axes apart. A frozen adapter should be deployed with the anchors it was trained with, and claims that prompt semantics drive performance should control for effective scale.

Limitations. Our guarantees require prescribed fixed anchors, rotation-closed adapters, and cosine-based binary maps (Proposition 1). MVFA and MadCLIP lie outside this premise, yet text-free anchors matched their prompts there too, suggesting the absence of a semantic benefit extends beyond the theorem; this evidence is empirical, not guaranteed. All anchors come from one frozen text encoder, so the near-collinearity of category axes may be specific to it. We compared only the reference effective scale with the larger one of orthogonal anchors, without sweeping it, so the best scale remains unknown; we also did not measure the proposed interference mechanism. Closeness between conditions is descriptive over three seeds, not a formal equivalence test.

REPRODUCIBILITY STATEMENT

All datasets are publicly available (Section 6.1). Appendix B and Table 19 give the complete training, checkpoint, and evaluation protocol, including seeds, and Appendix B.1 specifies both synthetic-anomaly generators. Appendices A.5 and A.6 list every prompt sentence, Appendix A.7 reports the resulting anchor cosines and effective scales, and Section 5 and Appendix C define the random and orthogonal anchor constructions. Appendix C also documents every change to MVFA and MadCLIP. Complete proofs are in Appendix A, and complete result tables in Appendix D. Code will be released upon acceptance.

AI USE STATEMENT

Generative AI tools were used to assist with mathematical derivations and proof checking, experimental methodology, research-code implementation and verification, interpretation of results, and the drafting and polishing of the manuscript text. The authors independently reviewed and verified all AI-assisted content, executed the reported experiments, and take full responsibility for the final manuscript and its claims.

REFERENCES

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. BMAD: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4042–4053, 2024.

Arda Basaran. PALADIN: Prompt-aligned localization and anomaly detection with DINOv3. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 7693–7701, 2026.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD – a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (LiTS). *Medical Image Analysis*, 84:102680, 2023. doi: 10.1016/j.media.2022.102680.

Navoneel Chakrabarty. Brain MRI images for brain tumor detection. Kaggle dataset, 2019. URL `https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection`.

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. *arXiv preprint arXiv:2305.17382*, 2023.

Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. AnomalyDINO: Boosting patch-based few-shot anomaly detection with DINOv2. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1319–1329, 2025.

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

Choubo Ding, Guansong Pang, and Chunhua Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7388–7398, 2022.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Denis A. Gudovskiy, Shun Ishizaka, and Kazuki Kozuka. CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 98–107, 2022.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019.

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. doi: 10.1016/j.media.2019.05.002.

Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yanfeng Wang. Registration based few-shot anomaly detection. In *Proceedings of the European Conference on Computer Vision*, pp. 303–319, 2022.

Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11375–11385, 2024.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 590–597, 2019.

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

Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C. S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalene Y. L. Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A. N. Huu, Cindy Wen, Edward D. Zhang, Chao L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Bennett Landman, Zhoubing Xu, Juan Eugenio Igelsias, Martin Styner, Thomas Robin Langerak, and Arno Klein. MICCAI multi-atlas labeling beyond the cranial vault – workshop and challenge. In *Proceedings of the MICCAI Multi-Atlas Labeling Beyond the Cranial Vault Workshop and Challenge*, 2015. doi: 10.7303/syn3193805.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. CutPaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664–9674, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16838–16848, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20402–20411, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning*, 2019.

Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations*, 2023.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *Proceedings of the International Conference on 3D Vision*, pp. 565–571, 2016.

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. *Transactions on Machine Learning Research*, 2024.

Ken Perlin. An image synthesizer. *ACM SIGGRAPH Computer Graphics*, 19(3):287–296, 1985.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

Geoffrey Roeder, Luke Metz, and Diederik P Kingma. On linear identifiability of learned representations. In *Proceedings of the International Conference on Machine Learning*, 2021.

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

Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15700–15711, 2023.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad Hossein Rohban, and Hamid R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14902–14912, 2021.

Hannah M. Schlüter, Jeremy Tan, Benjamin Hou, and Bernhard Kainz. Natural synthetic anomalies for self-supervised anomaly detection and localization. In *Proceedings of the European Conference on Computer Vision*, pp. 474–489, 2022.

Mahshid Shiri, Cigdem Beyan, and Vittorio Murino. MadCLIP: Few-shot medical anomaly detection with CLIP. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 416–426, 2025.

Oriane Siméoni et al. DINOv3, 2025.

Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, and Chongyang Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24490–24499, 2023.

Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. MediCLIP: Adapting CLIP for few-shot medical image anomaly detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 458–468, 2024.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022a.

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

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. AnomalyCLIP: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, 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 *Proceedings of the European Conference on Computer Vision*, pp. 392–408, 2022.

Zuo Zuo, Yao Wu, Baoqiang Li, Jiahao Dong, You Zhou, Lei Zhou, Yanyun Qu, and Zongze Wu. CLIP-FSAC: Boosting CLIP for few-shot anomaly classification with synthetic anomalies. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 1834–1842, 2024. doi: 10.24963/ijcai.2024/203.

A Theoretical Details and Proofs

This appendix first verifies that the implemented loss and metrics factor through abnormal probability maps. It then gives the Gram-to-rotation lemma, the binary proof, and the shared multi-category proof. Prompt banks, anchor geometry, rotation closure, finite-training residuals, and the proof of Proposition 1 follow.

A.1 From the implemented detector to abnormal probability maps

Theorem 1 concerns the binary probability maps actually consumed by training and evaluation. We therefore first verify that the complete implementation factors through the normal–abnormal logit difference.

Layer-wise scorer. The vision backbone retains $L = 4$ intermediate layers. Layer ℓ has its own adapter $g^{(\ell)}$, but every layer uses the same fixed unit anchor pair. With $\hat{\boldsymbol{u}}_p^{(\ell)} = \mathrm{normalize}(g^{(\ell)}(\boldsymbol{z}_p^{(\ell)}))$,

$$s_{p,c}^{(\ell)} = \tau \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \boldsymbol{a}_c \right\rangle, \qquad c \in \{N, A\}, \tag{8}$$

where $\tau = 100$ in training and $\tau = 20$ at evaluation, each multiplied by the configured similarity-scale multiplier. Define

$$\Delta s_p^{(\ell)} = s_{p,A}^{(\ell)} - s_{p,N}^{(\ell)}. \tag{9}$$

Upsampling. The two logit maps are bilinearly resampled from the patch grid to the image grid before softmax. If w_{qp} denotes the class-independent interpolation weight from patch p to pixel q, linearity gives

$$\widetilde{\Delta s}_q^{(\ell)} = \tilde{s}_{q,A}^{(\ell)} - \tilde{s}_{q,N}^{(\ell)} = \sum_{p=1}^{P} w_{qp} \Delta s_p^{(\ell)}. \tag{10}$$

Binary probability. A two-class softmax is a logistic function of this difference:

$$p_q^{(\ell)} = \frac{e^{\tilde{s}_{q,A}^{(\ell)}}}{e^{\tilde{s}_{q,N}^{(\ell)}} + e^{\tilde{s}_{q,A}^{(\ell)}}} = \sigma\left(\widetilde{\Delta s}_q^{(\ell)}\right), \qquad p_{q,N}^{(\ell)} = 1 - p_q^{(\ell)}. \tag{11}$$

Thus the normal channel carries no information beyond the abnormal map.

Training loss. For target mask y, the focal term uses the probability assigned to the target,

$$\pi_q^{(\ell)} = p_q^{(\ell)} y_q + \big(1 - p_q^{(\ell)}\big)(1 - y_q), \qquad \mathcal{L}_{\mathrm{foc}} = \frac{1}{BR^2} \sum_q -\big(1 - \pi_q^{(\ell)}\big)^{\gamma} \log \pi_q^{(\ell)}, \tag{12}$$

with $\gamma = 2$ and the implemented numerical clamp. The Dice term reads the abnormal map directly:

$$\mathcal{L}_{\mathrm{dice}} = 1 - \frac{1}{B} \sum_{b=1}^{B} \frac{2 \sum_q p_{b,q}^{(\ell)} y_{b,q} + 1}{\sum_q p_{b,q}^{(\ell)} + \sum_q y_{b,q} + 1}. \tag{13}$$

Consequently,

$$\mathcal{L} = \sum_{\ell=1}^{L} (\mathcal{L}_{\mathrm{foc}} + \mathcal{L}_{\mathrm{dice}}) = \Phi\Big(\Delta s^{(1)}, \ldots, \Delta s^{(L)}; y\Big) \tag{14}$$

for a fixed functional Φ. Neither loss term observes the two logits separately.

Evaluation. The layer maps are averaged and the image score is the top-fraction mean,

$$\mathrm{score}(x) = \frac{1}{k} \sum \mathrm{TopK}_q \left(\frac{1}{L} \sum_{\ell=1}^{L} p_q^{(\ell)} \right), \qquad k = \max\big(1, \lfloor \alpha R^2 \rfloor\big). \tag{15}$$

Pixel AUROC depends on the averaged maps and masks; image AUROC depends on these scores and image labels. Both therefore factor through the same abnormal probability maps as the training objective.

A.2 GRAM-TO-ROTATION LEMMA

Lemma 1 (Equal Gram matrices give one orthogonal map)**.** *Let $\boldsymbol{X}, \boldsymbol{X}' \in \mathbb{R}^{m\times d}$ have corresponding rows and satisfy $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{X}'\boldsymbol{X}'^\top$. Then some $\boldsymbol{Q} \in O(d)$ maps every row vector of $\boldsymbol{X}$ to the corresponding row vector of $\boldsymbol{X}'$.*

Proof. The row correspondence defines a linear map from $\mathrm{span}\{\boldsymbol{x}_i\}$ to $\mathrm{span}\{\boldsymbol{x}'_i\}$. Equality of the Gram matrices makes it well-defined and inner-product preserving: every two linear combinations have the same inner product before and after the map. It is therefore an isometry between the two row spans. Extend orthonormal bases of those spans to bases of $\mathbb{R}^d$; mapping the completed bases gives an orthogonal $\boldsymbol{Q}$. □

A.3 COMPLETE PROOF OF THEOREM 1

Let $\boldsymbol{\delta} = \boldsymbol{a}_A - \boldsymbol{a}_N$, $\boldsymbol{\delta}' = \boldsymbol{a}'_A - \boldsymbol{a}'_N$, and assume first that both are nonzero. The score definition in equation 8 gives

$$\Delta s_p^{(\ell)} = \tau \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \boldsymbol{\delta} \right\rangle = \tau \left\|\boldsymbol{\delta}\right\|_2 \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \hat{\boldsymbol{\delta}} \right\rangle. \tag{16}$$

Apply Lemma 1 to the one-row matrices $\hat{\boldsymbol{\delta}}^\top$ and $\widehat{\boldsymbol{\delta}'}^\top$. It supplies $\boldsymbol{Q} \in O(d)$ with

$$\boldsymbol{Q}\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}'. \tag{17}$$

For every layer set $g'^{(\ell)} = \boldsymbol{Q} \circ g^{(\ell)}$. Rotation closure makes all transformed adapters admissible, and

$$\tau' \left\|\boldsymbol{\delta}'\right\|_2 \left\langle \boldsymbol{Q}\hat{\boldsymbol{u}}_p^{(\ell)}, \hat{\boldsymbol{\delta}}' \right\rangle = \tau \left\|\boldsymbol{\delta}\right\|_2 \left\langle \hat{\boldsymbol{u}}_p^{(\ell)}, \hat{\boldsymbol{\delta}} \right\rangle \tag{18}$$

by equation 5. Thus corresponding adapters produce identical logit-difference fields. Equations 10–15 then give identical probability maps, losses, and evaluation scores. Applying $\boldsymbol{Q}^\top$ proves the reverse inclusion of attainable maps. If both difference vectors vanish, both binary probabilities are identically $1/2$; equality in equation 5 excludes the case in which only one vanishes.

For unit anchors,

$$\left\|\boldsymbol{\delta}\right\|_2^2 = \left\|\boldsymbol{a}_A\right\|_2^2 + \left\|\boldsymbol{a}_N\right\|_2^2 - 2\left\langle \boldsymbol{a}_A, \boldsymbol{a}_N \right\rangle = 2 - 2\rho, \tag{19}$$

which yields $\lambda_{\text{eff}} = \tau\sqrt{2-2\rho}$. An orthogonal pair scaled by a multiplier c matches the correct pair when

$$c\tau\sqrt{2} = \tau\sqrt{2-2\rho_{\text{correct}}} \qquad \Longleftrightarrow \qquad c = \sqrt{1-\rho_{\text{correct}}}. \tag{20}$$

The implementation applies c in training and evaluation alike. Normalizing the adapter output is essential: it prevents the adapter from absorbing $\left\|\boldsymbol{\delta}\right\|_2$ by rescaling itself.

A.4 UNIFIED MULTI-CATEGORY SPECIALIZATION

Category t has unit anchors $(\boldsymbol{a}_t^N, \boldsymbol{a}_t^A)$, difference vector $\boldsymbol{\delta}_t = \boldsymbol{a}_t^A - \boldsymbol{a}_t^N$, and known category label. The shared detector selects only that pair:

$$p_{t,A}(\boldsymbol{z}; g) = \sigma(\tau \left\langle \mathrm{normalize}(g(\boldsymbol{z})), \boldsymbol{\delta}_t \right\rangle). \tag{21}$$

The category selects a readout pair, not a different adapter. Let $\boldsymbol{D}$ stack the $\boldsymbol{\delta}_t^\top$ rows and let $\boldsymbol{M} = \boldsymbol{D}\boldsymbol{D}^\top$.

Proof of Corollary 1. If $\boldsymbol{D}\boldsymbol{D}^\top = \boldsymbol{D}'\boldsymbol{D}'^\top$, Lemma 1 supplies one $\boldsymbol{Q} \in O(d)$ with $\boldsymbol{\delta}'_t = \boldsymbol{Q}\boldsymbol{\delta}_t$ for every category. Compose the same $\boldsymbol{Q}$ into every layer-wise adapter. For a category-t sample,

$$\left\langle \boldsymbol{Q}\,\mathrm{normalize}(g(\boldsymbol{z})), \boldsymbol{\delta}'_t \right\rangle = \left\langle \boldsymbol{Q}\,\mathrm{normalize}(g(\boldsymbol{z})), \boldsymbol{Q}\boldsymbol{\delta}_t \right\rangle = \left\langle \mathrm{normalize}(g(\boldsymbol{z})), \boldsymbol{\delta}_t \right\rangle. \tag{22}$$

Thus every selected binary probability is preserved. The loss and evaluation factorization in Appendix A.1 completes the proof. Equality of $\boldsymbol{M}$ is also necessary and sufficient for the two finite difference-vector configurations to lie in the same global orthogonal orbit. We make no converse claim about equality of hypothesis classes for an arbitrary adapter family and input domain.

Relation to the full anchor Gram. Stack the anchors as rows

$$\boldsymbol{A} = [\boldsymbol{a}_1^N; \boldsymbol{a}_1^A; \ldots; \boldsymbol{a}_T^N; \boldsymbol{a}_T^A], \qquad \boldsymbol{G} = \boldsymbol{A}\boldsymbol{A}^\top,$$

and let $\boldsymbol{R} = \boldsymbol{I}_T \otimes [-1 \;\; 1]$. Then

$$\boldsymbol{D} = \boldsymbol{R}\boldsymbol{A}, \qquad \boldsymbol{M} = \boldsymbol{R}\boldsymbol{G}\boldsymbol{R}^\top. \tag{23}$$

Hence an exact full-Gram match implies an exact difference-Gram match, but the converse fails. Writing

$$\boldsymbol{a}_t^N = \boldsymbol{m}_t - \boldsymbol{\delta}_t/2, \qquad \boldsymbol{a}_t^A = \boldsymbol{m}_t + \boldsymbol{\delta}_t/2, \qquad \langle \boldsymbol{m}_t, \boldsymbol{\delta}_t \rangle = 0,$$

shows why: rotating a midpoint $\boldsymbol{m}_t$, while preserving its norm, inside the orthogonal complement of $\boldsymbol{\delta}_t$ changes entries of $\boldsymbol{G}$ while leaving every $\boldsymbol{\delta}_t$, probability, loss, and metric unchanged. The full $2T \times 2T$ anchor Gram is therefore sufficient but not necessary. The exact statement used in this paper is that $\boldsymbol{M}$ completely characterizes the T difference vectors up to one global orthogonal rotation.

Shared-rotation control. The shared-rotation condition applies one seeded $\boldsymbol{Q}$ to every correct anchor. It consequently rotates every $\boldsymbol{\delta}_t$ by the same $\boldsymbol{Q}$ and preserves all of $\boldsymbol{M}$. Correct prompts versus this shared rotation measures only the finite-training residual allowed by Appendix A.9.

Independent-rotation control. For each category, the independent-rotation condition constructs a seeded random pair with the same $\rho_t = \langle \boldsymbol{a}_t^N, \boldsymbol{a}_t^A \rangle$. Equivalently, some category-specific $\boldsymbol{Q}_t$ maps the correct pair to the replacement pair. It preserves

$$M_{tt} = \|\boldsymbol{\delta}_t\|_2^2 = 2 - 2\rho_t$$

but generally changes M_{ts} for $t \neq s$. Shared rotation versus independent rotations therefore changes the joint geometry while retaining every category's effective scale. Neither randomized condition is independent of text-derived geometry: the shared condition inherits the full reference joint geometry, and the independent condition inherits each ρ_t.

Seed scope of the current result. Every industrial condition uses three training seeds. In each randomized condition, the anchor seed equals the training seed, so the runs sample three shared rotations or three independently rotated sets of difference vectors. Correct-prompt anchors are deterministic and vary only through training.

A.5 Clinical prompt banks and mismatched-dataset intervention

All seven medical datasets use the same fixed prompt protocol: 14 normal and 14 abnormal complete sentences, with the same ordered grammatical scaffolds for every dataset. Only the anatomy, modality, and dataset-verified binary finding change. Each bank encodes the slice- or image-level binary task of its own dataset, following the benchmark constructions and their underlying datasets (Bao et al., 2024; Zhang et al., 2024; Baid et al., 2021; Landman et al., 2015; Bilic et al., 2023; Hu et al., 2019; Kermany et al., 2018; Al-Dhabyani et al., 2020; Irvin et al., 2019; Chakrabarty, 2019). LiverCT follows the BMAD construction, which combines BTCV and LiTS slices (Bao et al., 2024; Landman et al., 2015; Bilic et al., 2023). In particular, a normal slice is defined by the absence of the evaluated annotation on that image; we do not infer that it came from a healthy patient. Abnormal subtypes receive equal prompt weight, not dataset-frequency weight.

The medical semantic stress test replaces the complete prompt specification of target dataset t by the prompt specification of a fixed donor dataset $w(t)$. Let $\mathcal{P}_t^N$ and $\mathcal{P}_t^A$ denote all normal and abnormal descriptions used for the correct target condition. The mismatched-dataset condition uses $\mathcal{P}_{w(t)}^N$ and $\mathcal{P}_{w(t)}^A$ without editing them to mention the target. It therefore changes the organ, imaging modality, normality definition, abnormality definition, state descriptions, and any dataset-specific templates together. It is not implemented by replacing one anatomy noun inside the target templates.

Both donor banks are encoded by the same frozen text encoder used for the correct condition. The normal embeddings are averaged and normalized to form $\boldsymbol{a}_N^{\mathrm{mis}}$, and the abnormal embeddings are averaged and normalized to form $\boldsymbol{a}_A^{\mathrm{mis}}$. The anchors remain fixed throughout the run, and the visual adapter is trained on the target dataset exactly as in the other anchor conditions. Thus only the fixed prompt-derived anchor pair changes; images, supervision, support owners, optimization settings, and evaluation remain target-specific.

Table 4 specifies the donor mapping a priori. It is a derangement, uses each dataset once as a donor, and pairs every target with a different anatomy and imaging modality. For example, Kaggle

BrainMRI uses the complete LiverCT prompt banks: its anchors describe normal and abnormal liver CT appearances rather than normal and abnormal brain MR appearances.

Table 4: Fixed target-to-donor mapping for the mismatched-dataset prompt-bank intervention. The donor contributes its complete normal and abnormal description banks; no donor prompt is rewritten to mention the target dataset.

Target dataset	Target domain	Donor prompt bank
BraTS2021	brain MRI	BUSI (breast ultrasound)
LiverCT	liver CT	OCT2017 (retinal OCT)
RESC	retinal OCT	CheXpert (chest radiograph)
OCT2017	retinal OCT	Kaggle BrainMRI (brain MRI)
BUSI	breast ultrasound	RESC (retinal OCT)
CheXpert	chest radiograph	BraTS2021 (brain MRI)
Kaggle BrainMRI	brain MRI	LiverCT (liver CT)

This condition deliberately does not match $\rho = \langle \boldsymbol{a}_N^{\mathrm{mis}}, \boldsymbol{a}_A^{\mathrm{mis}} \rangle$ to the correct pair. It is therefore not an exact probability-map control and cannot by itself separate semantics from effective scale. Its role is narrower: it tests whether training with a grossly incorrect, internally coherent textual specification produces a repeatable performance change. The exact semantic-necessity test continues to be correct prompts versus the text-free controls matched in effective scale.

The following tables list the exact expanded banks used by both training and evaluation.

Table 5: Exact prompts for BraTS2021.

Normal	**Abnormal**
An axial T2-FLAIR brain MRI slice of the brain showing absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	An axial T2-FLAIR brain MRI slice of the brain showing annotated brain tumor tissue.
A clinical axial T2-FLAIR brain MRI slice of the brain showing absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	A clinical axial T2-FLAIR brain MRI slice of the brain showing a tumor-bearing axial brain slice.
A diagnostic axial T2-FLAIR brain MRI slice of the brain demonstrating absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	A diagnostic axial T2-FLAIR brain MRI slice of the brain demonstrating glioma tissue.
This axial T2-FLAIR brain MRI slice of the brain shows absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	This axial T2-FLAIR brain MRI slice of the brain shows a whole-tumor region.
This axial T2-FLAIR brain MRI slice of the brain demonstrates absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	This axial T2-FLAIR brain MRI slice of the brain demonstrates peritumoral edema.
The brain on this axial T2-FLAIR brain MRI slice shows absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	The brain on this axial T2-FLAIR brain MRI slice shows edema-associated T2-FLAIR hyperintensity.
The appearance of the brain on this axial T2-FLAIR brain MRI slice is consistent with absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	The appearance of the brain on this axial T2-FLAIR brain MRI slice is consistent with a necrotic tumor core.

Continued on next page

Table 5 – continued

Normal	Abnormal
The primary finding on this axial T2-FLAIR brain MRI slice is absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	The primary finding on this axial T2-FLAIR brain MRI slice is a non-enhancing tumor core.
The evaluated region of the brain on this axial T2-FLAIR brain MRI slice shows absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	The evaluated region of the brain on this axial T2-FLAIR brain MRI slice shows an enhancing tumor region.
Image-level assessment of this axial T2-FLAIR brain MRI slice indicates absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	Image-level assessment of this axial T2-FLAIR brain MRI slice indicates tumor core with surrounding edema.
The observable pattern on this axial T2-FLAIR brain MRI slice is consistent with absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	The observable pattern on this axial T2-FLAIR brain MRI slice is consistent with enhancing and non-enhancing tumor tissue.
The scan appearance demonstrates absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	The scan appearance demonstrates one or more BraTS tumor subregions.
An example of axial T2-FLAIR brain MRI slice showing absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor.	An example of axial T2-FLAIR brain MRI slice showing a focal glioma-related abnormality.
This image is interpreted as showing absence of any annotated glioma region, including edema, necrotic or non-enhancing tumor core, and enhancing tumor on axial T2-FLAIR brain MRI slice.	This image is interpreted as showing a glioma lesion comprising edema, core, or enhancing tumor on axial T2-FLAIR brain MRI slice.

Table 6: Exact prompts for LiverCT.

Normal	Abnormal
An axial liver CT slice of the liver showing absence of any annotated primary or metastatic liver tumor on the slice.	An axial liver CT slice of the liver showing an annotated liver tumor.
A clinical axial liver CT slice of the liver showing absence of any annotated primary or metastatic liver tumor on the slice.	A clinical axial liver CT slice of the liver showing a neoplastic liver lesion.
A diagnostic axial liver CT slice of the liver demonstrating absence of any annotated primary or metastatic liver tumor on the slice.	A diagnostic axial liver CT slice of the liver demonstrating a primary liver tumor.
This axial liver CT slice of the liver shows absence of any annotated primary or metastatic liver tumor on the slice.	This axial liver CT slice of the liver shows hepatocellular carcinoma.
This axial liver CT slice of the liver demonstrates absence of any annotated primary or metastatic liver tumor on the slice.	This axial liver CT slice of the liver demonstrates cholangiocarcinoma.
The liver on this axial liver CT slice shows absence of any annotated primary or metastatic liver tumor on the slice.	The liver on this axial liver CT slice shows a metastatic liver lesion.
The appearance of the liver on this axial liver CT slice is consistent with absence of any annotated primary or metastatic liver tumor on the slice.	The appearance of the liver on this axial liver CT slice is consistent with a colorectal cancer metastasis in the liver.
The primary finding on this axial liver CT slice is absence of any annotated primary or metastatic liver tumor on the slice.	The primary finding on this axial liver CT slice is a breast cancer metastasis in the liver.

Continued on next page

Table 6 – continued

Normal	Abnormal
The evaluated region of the liver on this axial liver CT slice shows absence of any annotated primary or metastatic liver tumor on the slice.	The evaluated region of the liver on this axial liver CT slice shows a lung cancer metastasis in the liver.
Image-level assessment of this axial liver CT slice indicates absence of any annotated primary or metastatic liver tumor on the slice.	Image-level assessment of this axial liver CT slice indicates a hypodense liver tumor.
The observable pattern on this axial liver CT slice is consistent with absence of any annotated primary or metastatic liver tumor on the slice.	The observable pattern on this axial liver CT slice is consistent with a hyperdense liver tumor.
The scan appearance demonstrates absence of any annotated primary or metastatic liver tumor on the slice.	The scan appearance demonstrates a small focal liver tumor.
An example of axial liver CT slice showing absence of any annotated primary or metastatic liver tumor on the slice.	An example of axial liver CT slice showing multiple liver tumors.
This image is interpreted as showing absence of any annotated primary or metastatic liver tumor on the slice on axial liver CT slice.	This image is interpreted as showing one or more liver tumors with variable size and appearance on axial liver CT slice.

Table 7: Exact prompts for RESC.

Normal	Abnormal
A retinal OCT B-scan of the retina showing absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	A retinal OCT B-scan of the retina showing an annotated retinal-edema lesion.
A clinical retinal OCT B-scan of the retina showing absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	A clinical retinal OCT B-scan of the retina showing retinal edema.
A diagnostic retinal OCT B-scan of the retina demonstrating absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	A diagnostic retinal OCT B-scan of the retina demonstrating an edematous retinal region.
This retinal OCT B-scan of the retina shows absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	This retinal OCT B-scan of the retina shows edema-associated retinal thickening.
This retinal OCT B-scan of the retina demonstrates absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	This retinal OCT B-scan of the retina demonstrates subretinal fluid.
The retina on this retinal OCT B-scan shows absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	The retina on this retinal OCT B-scan shows a localized pocket of subretinal fluid.
The appearance of the retina on this retinal OCT B-scan is consistent with absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	The appearance of the retina on this retinal OCT B-scan is consistent with fluid accumulation beneath the retina.
The primary finding on this retinal OCT B-scan is absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	The primary finding on this retinal OCT B-scan is pigment epithelial detachment.
The evaluated region of the retina on this retinal OCT B-scan shows absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	The evaluated region of the retina on this retinal OCT B-scan shows a localized pigment epithelial detachment.

Continued on next page

Table 7 – continued

Normal	Abnormal
Image-level assessment of this retinal OCT B-scan indicates absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	Image-level assessment of this retinal OCT B-scan indicates elevation of the retinal pigment epithelium.
The observable pattern on this retinal OCT B-scan is consistent with absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	The observable pattern on this retinal OCT B-scan is consistent with subretinal fluid with pigment epithelial detachment.
The scan appearance demonstrates absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	The scan appearance demonstrates one or more edema-related lesions.
An example of retinal OCT B-scan showing absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment.	An example of retinal OCT B-scan showing retinal-layer distortion caused by edema.
This image is interpreted as showing absence of any annotated retinal-edema lesion, including subretinal fluid and pigment epithelial detachment on retinal OCT B-scan.	This image is interpreted as showing a fluid-related abnormality disrupting normal retinal structure on retinal OCT B-scan.

Table 8: Exact prompts for OCT2017.

Normal	Abnormal
A retinal OCT B-scan of the retina showing a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	A retinal OCT B-scan of the retina showing choroidal neovascularization.
A clinical retinal OCT B-scan of the retina showing a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	A clinical retinal OCT B-scan of the retina showing CNV-related neovascular change.
A diagnostic retinal OCT B-scan of the retina demonstrating a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	A diagnostic retinal OCT B-scan of the retina demonstrating a choroidal neovascular membrane.
This retinal OCT B-scan of the retina shows a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	This retinal OCT B-scan of the retina shows diabetic macular edema.
This retinal OCT B-scan of the retina demonstrates a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	This retinal OCT B-scan of the retina demonstrates DME-related retinal thickening.
The retina on this retinal OCT B-scan shows a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	The retina on this retinal OCT B-scan shows diabetic macular edema with intraretinal fluid.
The appearance of the retina on this retinal OCT B-scan is consistent with a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	The appearance of the retina on this retinal OCT B-scan is consistent with drusen deposits.
The primary finding on this retinal OCT B-scan is a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	The primary finding on this retinal OCT B-scan is multiple drusen beneath the retinal pigment epithelium.

Continued on next page

Table 8 – continued

Normal	Abnormal
The evaluated region of the retina on this retinal OCT B-scan shows a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	The evaluated region of the retina on this retinal OCT B-scan shows an early age-related macular degeneration appearance with drusen.
Image-level assessment of this retinal OCT B-scan indicates a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	Image-level assessment of this retinal OCT B-scan indicates a retinal disease pattern from CNV, DME, or drusen.
The observable pattern on this retinal OCT B-scan is consistent with a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	The observable pattern on this retinal OCT B-scan is consistent with a diseased retinal OCT appearance.
The scan appearance demonstrates a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	The scan appearance demonstrates retinal fluid or edema associated with DME.
An example of retinal OCT B-scan showing a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen.	An example of retinal OCT B-scan showing abnormal retinal structure associated with CNV, DME, or drusen.
This image is interpreted as showing a disease-free retinal appearance with no choroidal neovascularization, diabetic macular edema, or drusen on retinal OCT B-scan.	This image is interpreted as showing one of the OCT2017 CNV, DME, or drusen disease classes on retinal OCT B-scan.

Table 9: Exact prompts for BUSI.

Normal	Abnormal
A grayscale breast ultrasound image of breast tissue showing absence of any focal breast mass labeled benign or malignant.	A grayscale breast ultrasound image of breast tissue showing a focal breast mass.
A clinical grayscale breast ultrasound image of breast tissue showing absence of any focal breast mass labeled benign or malignant.	A clinical grayscale breast ultrasound image of breast tissue showing an annotated breast lesion.
A diagnostic grayscale breast ultrasound image of breast tissue demonstrating absence of any focal breast mass labeled benign or malignant.	A diagnostic grayscale breast ultrasound image of breast tissue demonstrating a localized breast nodule.
This grayscale breast ultrasound image of breast tissue shows absence of any focal breast mass labeled benign or malignant.	This grayscale breast ultrasound image of breast tissue shows a benign breast mass.
This grayscale breast ultrasound image of breast tissue demonstrates absence of any focal breast mass labeled benign or malignant.	This grayscale breast ultrasound image of breast tissue demonstrates a benign breast lesion.
Breast tissue on this grayscale breast ultrasound image shows absence of any focal breast mass labeled benign or malignant.	Breast tissue on this grayscale breast ultrasound image shows a benign breast nodule.
The appearance of breast tissue on this grayscale breast ultrasound image is consistent with absence of any focal breast mass labeled benign or malignant.	The appearance of breast tissue on this grayscale breast ultrasound image is consistent with a localized benign tumor region.
The primary finding on this grayscale breast ultrasound image is absence of any focal breast mass labeled benign or malignant.	The primary finding on this grayscale breast ultrasound image is a malignant breast mass.
The evaluated region of breast tissue on this grayscale breast ultrasound image shows absence of any focal breast mass labeled benign or malignant.	The evaluated region of breast tissue on this grayscale breast ultrasound image shows a malignant breast lesion.

Continued on next page

Table 9 – continued

Normal	Abnormal
Image-level assessment of this grayscale breast ultrasound image indicates absence of any focal breast mass labeled benign or malignant.	Image-level assessment of this grayscale breast ultrasound image indicates a malignant breast tumor.
The observable pattern on this grayscale breast ultrasound image is consistent with absence of any focal breast mass labeled benign or malignant.	The observable pattern on this grayscale breast ultrasound image is consistent with a localized malignant tumor region.
The scan appearance demonstrates absence of any focal breast mass labeled benign or malignant.	The scan appearance demonstrates a breast mass labeled benign or malignant.
An example of grayscale breast ultrasound image showing absence of any focal breast mass labeled benign or malignant.	An example of grayscale breast ultrasound image showing a tumor-bearing breast ultrasound appearance.
This image is interpreted as showing absence of any focal breast mass labeled benign or malignant on grayscale breast ultrasound image.	This image is interpreted as showing one or more focal breast lesions on grayscale breast ultrasound image.

Table 10: Exact prompts for CheXpert.

Normal	Abnormal
A chest radiograph of the chest showing the No Finding state with none of the twelve evaluated disease observations.	A chest radiograph of the chest showing enlarged cardiomediastinum.
A clinical chest radiograph of the chest showing the No Finding state with none of the twelve evaluated disease observations.	A clinical chest radiograph of the chest showing cardiomegaly.
A diagnostic chest radiograph of the chest demonstrating the No Finding state with none of the twelve evaluated disease observations.	A diagnostic chest radiograph of the chest demonstrating lung opacity.
This chest radiograph of the chest shows the No Finding state with none of the twelve evaluated disease observations.	This chest radiograph of the chest shows a lung lesion.
This chest radiograph of the chest demonstrates the No Finding state with none of the twelve evaluated disease observations.	This chest radiograph of the chest demonstrates pulmonary edema.
The chest on this chest radiograph shows the No Finding state with none of the twelve evaluated disease observations.	The chest on this chest radiograph shows pulmonary consolidation.
The appearance of the chest on this chest radiograph is consistent with the No Finding state with none of the twelve evaluated disease observations.	The appearance of the chest on this chest radiograph is consistent with pneumonia.
The primary finding on this chest radiograph is the No Finding state with none of the twelve evaluated disease observations.	The primary finding on this chest radiograph is atelectasis.
The evaluated region of the chest on this chest radiograph shows the No Finding state with none of the twelve evaluated disease observations.	The evaluated region of the chest on this chest radiograph shows pneumothorax.
Image-level assessment of this chest radiograph indicates the No Finding state with none of the twelve evaluated disease observations.	Image-level assessment of this chest radiograph indicates pleural effusion.
The observable pattern on this chest radiograph is consistent with the No Finding state with none of the twelve evaluated disease observations.	The observable pattern on this chest radiograph is consistent with another pleural abnormality.
The scan appearance demonstrates the No Finding state with none of the twelve evaluated disease observations.	The scan appearance demonstrates a fracture.

Continued on next page

Table 10 – continued

Normal	Abnormal
An example of chest radiograph showing the No Finding state with none of the twelve evaluated disease observations.	An example of chest radiograph showing one or more of the twelve evaluated disease observations.
This image is interpreted as showing the No Finding state with none of the twelve evaluated disease observations on chest radiograph.	This image is interpreted as showing an abnormal chest radiographic finding from the evaluated disease set on chest radiograph.

Table 11: Exact prompts for Kaggle BrainMRI.

Normal	Abnormal
A 2D brain MRI image of the brain showing absence of any visible brain tumor or intracranial mass.	A 2D brain MRI image of the brain showing a visible brain tumor.
A clinical 2D brain MRI image of the brain showing absence of any visible brain tumor or intracranial mass.	A clinical 2D brain MRI image of the brain showing an intracranial tumor.
A diagnostic 2D brain MRI image of the brain demonstrating absence of any visible brain tumor or intracranial mass.	A diagnostic 2D brain MRI image of the brain demonstrating a tumor-bearing brain MRI appearance.
This 2D brain MRI image of the brain shows absence of any visible brain tumor or intracranial mass.	This 2D brain MRI image of the brain shows a focal intracranial mass.
This 2D brain MRI image of the brain demonstrates absence of any visible brain tumor or intracranial mass.	This 2D brain MRI image of the brain demonstrates a neoplastic brain lesion.
The brain on this 2D brain MRI image shows absence of any visible brain tumor or intracranial mass.	The brain on this 2D brain MRI image shows a space-occupying brain lesion.
The appearance of the brain on this 2D brain MRI image is consistent with absence of any visible brain tumor or intracranial mass.	The appearance of the brain on this 2D brain MRI image is consistent with a localized brain tumor region.
The primary finding on this 2D brain MRI image is absence of any visible brain tumor or intracranial mass.	The primary finding on this 2D brain MRI image is abnormal brain tissue consistent with a tumor.
The evaluated region of the brain on this 2D brain MRI image shows absence of any visible brain tumor or intracranial mass.	The evaluated region of the brain on this 2D brain MRI image shows a tumor-positive brain MRI.
Image-level assessment of this 2D brain MRI image indicates absence of any visible brain tumor or intracranial mass.	Image-level assessment of this 2D brain MRI image indicates an intracranial mass lesion.
The observable pattern on this 2D brain MRI image is consistent with absence of any visible brain tumor or intracranial mass.	The observable pattern on this 2D brain MRI image is consistent with one or more visible tumor regions.
The scan appearance demonstrates absence of any visible brain tumor or intracranial mass.	The scan appearance demonstrates MRI evidence of a brain tumor.
An example of 2D brain MRI image showing absence of any visible brain tumor or intracranial mass.	An example of 2D brain MRI image showing a focal tumor within the brain.
This image is interpreted as showing absence of any visible brain tumor or intracranial mass on 2D brain MRI image.	This image is interpreted as showing a brain MRI appearance containing a tumor on 2D brain MRI image.

The donor condition uses the donor's complete normal and abnormal banks unchanged; no sentence is rewritten for the target.

A.6 Industrial prompt banks

MVTec-AD and VisA use the fixed two-level WinCLIP-style prompt ensemble (Jeong et al., 2023). For category noun o_t, each of seven normal and seven abnormal state phrases is inserted into each

of 22 photo templates, giving $7 \times 22 = 154$ sentences per state and category. The frozen OpenAI CLIP ViT-L/14 text encoder (Radford et al., 2021) encodes the sentences individually. We average the 154 normal embeddings and the 154 abnormal embeddings separately, then unit-normalize the two means to obtain $(\boldsymbol{a}_t^N, \boldsymbol{a}_t^A)$. No prompt token or anchor is trained. Tables 12–14 give the complete construction, including literal punctuation and object nouns; together they determine every sentence in every industrial bank exactly. Expansion is state-major and template-minor in the displayed orders.

Table 12: Exact state phrases used to construct every industrial prompt bank. The object placeholder {} is replaced by the noun in Table 14. The leading space in the sixth abnormal phrase is retained from the reference implementation.

Normal state phrase	Abnormal state phrase
{}	damaged {}
flawless {}	broken {}
perfect {}	{} with flaw
unblemished {}	{} with defect
{} without flaw	{} with damage
{} without defect	␣disease {}
{} without damage	abnormal {}

Table 13: Exact photo templates used for the industrial prompt ensemble. The state placeholder {} is replaced by one of the state phrases in Table 12.

	Template
1	a cropped photo of the {}.
2	a cropped photo of a {}.
3	a close-up photo of a {}.
4	a close-up photo of the {}.
5	a bright photo of a {}.
6	a bright photo of the {}.
7	a dark photo of the {}.
8	a dark photo of a {}.
9	a jpeg corrupted photo of a {}.
10	a jpeg corrupted photo of the {}.
11	a blurry photo of the {}.
12	a blurry photo of a {}.
13	a photo of the {}
14	a photo of a {}
15	a photo of a small {}
16	a photo of the small {}
17	a photo of a large {}
18	a photo of the large {}
19	a photo of the {} for visual inspection.
20	a photo of a {} for visual inspection.
21	a photo of the {} for anomaly detection.
22	a photo of a {} for anomaly detection.

A.7 Anchor cosines and effective scales

For every correct anchor pair, Tables 15 and 16 report

$$\rho = \langle \boldsymbol{a}_N, \boldsymbol{a}_A \rangle, \qquad c = \sqrt{1-\rho}, \qquad \lambda_{\text{eff}}^{\text{train}} = 100\sqrt{2-2\rho}, \qquad \lambda_{\text{eff}}^{\text{eval}} = 20\sqrt{2-2\rho}. \tag{24}$$

Matched orthogonal anchors multiply the training and evaluation similarity scales by c; because their anchor difference has norm $\sqrt{2}$, this gives exactly the two reported λ_{eff} values. Unmatched

Table 14: Exact object nouns substituted into the industrial prompt templates. Numbered VisA categories deliberately retain their index.

Dataset	Object nouns
MVTec-AD	bottle; cable; capsule; carpet; grid; hazelnut; leather; metal nut; pill; screw; tile; toothbrush; transistor; wood; zipper
VisA	candle; capsules; cashew; chewing gum; fryum; macaroni 1; macaroni 2; printed circuit board 1; printed circuit board 2; printed circuit board 3; printed circuit board 4; pipe fryum

orthogonal anchors instead have $\lambda_{\text{eff}}^{\text{train}} = 100\sqrt{2} = 141.42$ and $\lambda_{\text{eff}}^{\text{eval}} = 20\sqrt{2} = 28.28$ for every dataset or category. They therefore inflate the correct-prompt effective scale by the directly readable factor $1/c$. All entries are computed from the unrounded frozen OpenAI CLIP ViT-L/14 anchors; only their display is rounded.

Table 15: Correct-prompt geometry for the seven medical datasets. The reported values determine both the Gram-matched random construction and the rescaling of matched orthogonal anchors.

Quantity	BraTS 2021	Liver CT	RESC	OCT 2017	BUSI	CheXpert	Kaggle BrainMRI
ρ_{correct}	0.9255	0.9711	0.9524	0.9408	0.9771	0.9241	0.9742
c	0.2730	0.1700	0.2183	0.2432	0.1513	0.2755	0.1608
$\lambda_{\text{eff}}^{\text{train}}$	38.61	24.04	30.87	34.40	21.40	38.97	22.74
$\lambda_{\text{eff}}^{\text{eval}}$	7.72	4.81	6.17	6.88	4.28	7.79	4.55

Table 16: Per-category correct-prompt geometry for the unified industrial detector. Each category's shared- and independent-rotation controls preserve its displayed ρ_t, c_t, and $\lambda_{\text{eff},t}$; independent rotations change only cross-category off-diagonal geometry. The participation-ratio effective rank of the correct-prompt difference Gram matrix $\boldsymbol{M}$ (Section 6.4) is $r_{\text{eff}} = 1.79$ for MVTec-AD and 1.78 for VisA. Shared rotations preserve $\boldsymbol{M}$ and hence r_{eff}; independent rotations raise it to 14.28–14.29 of 15 (MVTec-AD) and about 11.10 of 12 (VisA) across the three geometry seeds.

MVTec-AD					**VisA**				
Category	ρ_t	c_t	$\lambda_{\text{eff},t}^{\text{train}}$	$\lambda_{\text{eff},t}^{\text{eval}}$	Category	ρ_t	c_t	$\lambda_{\text{eff},t}^{\text{train}}$	$\lambda_{\text{eff},t}^{\text{eval}}$
bottle	0.9754	0.1567	22.17	4.43	candle	0.9743	0.1604	22.68	4.54
cable	0.9771	0.1514	21.41	4.28	capsules	0.9742	0.1605	22.70	4.54
capsule	0.9715	0.1688	23.87	4.77	cashew	0.9771	0.1514	21.41	4.28
carpet	0.9767	0.1526	21.58	4.32	chewing gum	0.9801	0.1410	19.95	3.99
grid	0.9735	0.1629	23.03	4.61	fryum	0.9701	0.1730	24.47	4.89
hazelnut	0.9696	0.1744	24.67	4.93	macaroni 1	0.9738	0.1617	22.87	4.57
leather	0.9711	0.1700	24.05	4.81	macaroni 2	0.9732	0.1638	23.16	4.63
metal nut	0.9830	0.1305	18.45	3.69	printed circuit board 1	0.9854	0.1209	17.09	3.42
pill	0.9779	0.1485	21.00	4.20	printed circuit board 2	0.9857	0.1194	16.89	3.38
screw	0.9834	0.1289	18.22	3.64	printed circuit board 3	0.9850	0.1224	17.31	3.46
tile	0.9699	0.1735	24.54	4.91	printed circuit board 4	0.9868	0.1150	16.27	3.25
toothbrush	0.9817	0.1354	19.15	3.83	pipe fryum	0.9744	0.1600	22.63	4.53
transistor	0.9759	0.1551	21.94	4.39					
wood	0.9730	0.1645	23.26	4.65					
zipper	0.9802	0.1407	19.90	3.98					

A.8 ADAPTER COVERAGE AND ROTATION CLOSURE

Table 17 checks the exact scope of Theorem 1 against representative systems. A "no" does not say that the method uses no related symmetry; it says that its published parameterization does not satisfy all assumptions of our exact result. Coverage can fail on either of two sides, and the table names which one applies in each case. On the *anchor* side the pair must be prescribed and held fixed within the run: if the anchors are produced from learned tokens by a frozen text encoder, the reachable set of pairs is the image of the token space under that encoder, so an arbitrarily rotated pair need not be reachable, and ρ—hence λ_{eff}—is fitted rather than prescribed. On the *visual* side the admissible set $\mathcal{S}$ of maps into text-embedding space must be closed under left multiplication by $O(d)$ (Lemma 2); failure is caused by a specific constraint that prevents $\boldsymbol{QL}$ from remaining admissible, not by a family label alone.

Definition 1 (Rotation-closed adapter family)**.** A family $\mathcal{G}$ of maps into $\mathbb{R}^d$ is rotation-closed if $g \in \mathcal{G}$ implies $\boldsymbol{Q} \circ g \in \mathcal{G}$ for every $\boldsymbol{Q} \in O(d)$.

Lemma 2 (A free final map implies rotation closure)**.** *Suppose every $g \in \mathcal{G}$ factors as $g = \boldsymbol{L} \circ h$, where $\boldsymbol{L}$ ranges over a set $\mathcal{S} \subseteq \mathbb{R}^{d\times m}$ closed under left multiplication by $O(d)$, independently of the admissible h. Then $\mathcal{G}$ is rotation-closed.*

Proof. For any $\boldsymbol{Q} \in O(d)$, $\boldsymbol{Q} \circ g = (\boldsymbol{QL}) \circ h$, and $\boldsymbol{QL} \in \mathcal{S}$ by assumption. □

Linear. Let $\tilde{\boldsymbol{z}} = [\boldsymbol{z}; 1]$ and $g(\boldsymbol{z}) = \widetilde{\boldsymbol{W}}^\top \tilde{\boldsymbol{z}}$, so $\widetilde{\boldsymbol{W}}$ contains both weights and bias. Then $\boldsymbol{Q}g(\boldsymbol{z}) = (\widetilde{\boldsymbol{W}}\boldsymbol{Q}^\top)^\top \tilde{\boldsymbol{z}}$, and $\widetilde{\boldsymbol{W}}\boldsymbol{Q}^\top$ is another unconstrained affine map.

MLP. For $g(\boldsymbol{z}) = \boldsymbol{W}_2\mathrm{GELU}(\boldsymbol{W}_1\boldsymbol{z})$, replace the free final matrix $\boldsymbol{W}_2$ by $\boldsymbol{QW}_2$.

Transformer. For $g(Z) = \boldsymbol{U}_{\text{out}}T(\boldsymbol{U}_{\text{in}}Z + E_{\text{pos}})$, replace $\boldsymbol{U}_{\text{out}}$ by $\boldsymbol{QU}_{\text{out}}$. Patch interaction inside T does not affect closure.

A.9 FINITE-TRAINING RESIDUAL

Theorem 1 and Corollary 1 equate attainable probability maps and objective-value sets, not arbitrary optimization paths. Under the reparameterization $\boldsymbol{W} \mapsto \boldsymbol{WQ}^\top$, an empirical difference between exactly matched conditions can arise from components of the finite training or evaluation procedure that are not equivariant to this rotation, or from stochastic elements that are not fully coupled across the compared runs.

Weight decay. Frobenius weight decay is invariant:

$$\left\|\boldsymbol{WQ}^\top\right\|_F^2 = \mathrm{tr}(\boldsymbol{QW}^\top\boldsymbol{WQ}^\top) = \|\boldsymbol{W}\|_F^2 . \tag{25}$$

Initialization. Equivariance requires the distribution of $\boldsymbol{W}^{(0)}$ to be invariant under right multiplication by $O(d)$. A coordinate-aligned uniform initialization is not exactly invariant.

Optimizer. Gradient descent is equivariant because its gradient rotates with $\boldsymbol{Q}$. Adam is not: its diagonal second-moment preconditioner is tied to the coordinate frame. Initialization and Adam are therefore the principal algorithmic sources of a finite-training discrepancy under exact arithmetic and fully coupled randomness. The reported experiments also use bfloat16 arithmetic, and seed identifiers are not common to every reported condition (Appendix B); finite-precision effects and uncoupled stochastic variation can therefore contribute as well. Observed residuals are not attributed uniquely to initialization or Adam.

A.10 PROOF OF PROPOSITION 1

Theorem 1 constructs its correspondence by rotating every layer-wise adapter. If the family is not rotation-closed, the transformed adapters need not be admissible. If the adapter is frozen rather than trained against the replacement anchors, the compensating parameterization is unavailable. In the category-conditioned case, Corollary 1 requires one rotation for all T difference vectors;

Table 17: Scope check for representative VLM anomaly detectors. Exact coverage requires prescribed text embeddings held fixed within a run and a visual scoring path whose admissible set $\mathcal{S}$ is closed under left multiplication by $O(d)$. The last column names the specific constraint that blocks—or admits—$\boldsymbol{QL} \in \mathcal{S}$. A "no" is not a judgment about the method, and where a published parameterization admits more than one reading we state the condition under which the verdict holds.

Method	Text embeddings	Visual path used for scoring	Cov.	Constraint that blocks or admits $\boldsymbol{QL} \in \mathcal{S}$
APRIL-GAN (Chen et al., 2023)	prescribed prompts	free linear projections	yes†	None on either side: the per-layer map into text-embedding space is a free linear map, which absorbs $\boldsymbol{Q}$. †The verdict covers the trained linear-projection branch; the few-shot mode adds a text-free reference-memory branch that our statement does not describe.
MVFA (Huang et al., 2024)	prescribed prompt ensemble	residual multi-level adapters	no‡	Visual side: the scored feature retains a fixed, untrained component, so absorbing $\boldsymbol{Q}$ requires the trainable branch to cancel that component's rotation in addition to supplying its own output. ‡This holds when the trainable branch is capacity-constrained; a freely parameterized full-rank residual branch would instead be closed, since $\{\boldsymbol{I} + \boldsymbol{A} : \boldsymbol{A}$ free$\}$ is all of $\mathbb{R}^{d\times d}$.
MediCLIP (Zhang et al., 2024)	learned prompts	visual adapters	no	Anchor side: prompt-token bottleneck. The pair is not prescribed, so ρ and λ_{eff} are fitted, and the rotated pair need not lie in the image of the token space under the frozen text encoder.
MadCLIP (Shiri et al., 2025)	learned prompts	dual residual adapters	no	Both sides: the same prompt-token bottleneck as above, together with a fixed untrained component in the residual visual path.
Ours (few-shot medical)	prescribed	free output map	yes	None: free final linear map, so Lemma 2 applies directly. Table 18 records the verdict for each parameterization we train.
Ours (full-shot unified)	prescribed	one free shared adapter	yes	None on either side; the multi-category statement additionally requires the replacement bank to match $\boldsymbol{M} = \boldsymbol{D}\boldsymbol{D}^{\top}$ (Corollary 1).
WinCLIP (Jeong et al., 2023)	prescribed prompts	no trained projection	no	Neither side is at issue: there is no $\boldsymbol{L}$ to rotate, so the premise of an adapter trained against the anchors fails before closure arises.

Table 18: Rotation closure of the three adapter families.

Family	Parameterization	Closed	Reason
linear	$\boldsymbol{y} = \widetilde{\boldsymbol{W}}^{\top}\tilde{\boldsymbol{z}}$, $\widetilde{\boldsymbol{W}}$ free	yes	absorb $\boldsymbol{Q}$ into $\widetilde{\boldsymbol{W}}$
MLP	$\boldsymbol{y} = \boldsymbol{W}_2\,\mathrm{GELU}(\boldsymbol{W}_1\boldsymbol{z})$	yes	absorb $\boldsymbol{Q}$ into $\boldsymbol{W}_2$
Transformer	$\boldsymbol{y} = \boldsymbol{U}_{\text{out}}T(\boldsymbol{U}_{\text{in}}Z + E_{\text{pos}})$	yes	absorb $\boldsymbol{Q}$ into $\boldsymbol{U}_{\text{out}}$

changing $\boldsymbol{D}\boldsymbol{D}^\top$ removes that construction. These are failures of a sufficient equivalence argument, not sufficient conditions for a performance difference. In particular, none establishes that linguistic meaning is beneficial.

B Internal Training, Checkpoint, and Evaluation Protocol

This section consolidates the settings shared by our internal experiments; Table 19 lists them. The DINOv3-L/16, DINOv2-L/14, and CLIP ViT-L/14 vision backbones (Siméoni et al., 2025; Oquab et al., 2024; Radford et al., 2021), as well as the OpenAI CLIP ViT-L/14 text encoder, are frozen. Inputs are resized to 224×224, and visual features are retained from transformer blocks $5, 11, 17, 23$. Each retained layer has its own adapter into the 768-dimensional anchor space. The prompt context length is zero: prompt ensembles are encoded once into prescribed anchors, and neither prompt tokens nor anchors are optimized. Unless a table explicitly names another family, the adapter is affine linear. The MLP and Transformer parameterizations are specified in Appendix A.8.

Table 19: Internal training and evaluation settings. Settings not separated by column are shared.

Setting	**Medical**	**MVTec-AD**	**VisA**
Training images	regime-specific K support	all normal train images	all normal train images
Epochs	200	50	30
Training seeds	three per condition		
Input size	224×224		
Retained blocks	$5, 11, 17, 23$		
Optimizer	Adam, learning rate 10^{-3}, weight decay 10^{-3}		
Adam parameters	$\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$		
Schedule	linear decay after 10% linear warm-up		
Batch / accumulation	$4/2$		
Gradient clipping	global norm 0.5		
Numerical precision	bfloat16 automatic mixed precision		
Support augmentation	16 deterministic views	none	none
Training similarity scale	$100c$		
Evaluation similarity scale	$20c$		
Reported checkpoint	completed final epoch; no checkpoint selection		
Image score	mean of the highest-scoring 1% of pixels		
AUROC aggregation	pooled within dataset	category-macro	category-macro

Here $c = 1$ except for matched orthogonal anchors, where $c = \sqrt{1 - \rho_{\text{correct}}}$. Training minimizes the sum, over the four layer maps, of focal loss with $\gamma = 2$ and Dice loss. Logits are bilinearly upsampled before the binary softmax. Evaluation averages the four abnormal-probability maps before computing pixel AUROC and the top-1% image score. Industrial AUROCs are first computed separately for each object category and then macro-averaged; medical AUROC is computed over the applicable dataset test split.

Medical supervision and evaluation masks. In the real-anomaly-support regime, each support set contains K normal and K real abnormal images. BraTS2021, LiverCT, RESC, and OCT2017 use annotated dense ground-truth training masks, which enter the same pixel-level focal and Dice losses. OCT2017 is evaluated at image level only, so the paper reports only its image AUROC. In the normal-only regime, anomalies and their masks are generated online from normal support images. BUSI does have evaluation masks, so its pixel AUROC could be reported. We nevertheless omit BUSI pixel AUROC from the normal-only anchor tables to retain a common image-only endpoint across BUSI, CheXpert, and Kaggle BrainMRI; the latter two do not have evaluation masks. Thus this table-level omission must not be read as absence of BUSI masks.

B.1 SYNTHETIC-ANOMALY GENERATION

The medical normal-only and industrial full-shot experiments use different synthesis procedures. Both operate online: every access pairs the unchanged normal image with a newly synthesized image and its dense target, so the focal and Dice losses see clean and synthetic examples together.

Medical normal-only synthesis. BUSI, CheXpert, and Kaggle BrainMRI share one task mixture adapted from MediCLIP (Zhang et al., 2024). Each selected normal support image is first expanded into a fixed set of 16 deterministic views spanning small rotations, translations, and zooms, brightness and contrast jitter, and the two flips; synthesis is then resampled on the drawn view at every access. One of three tasks is then drawn: CutPaste (Li et al., 2021), which copies a rotated and rescaled crop of a normal support image into the view; a Gaussian intensity change, which adds a smooth signed perturbation to it; or the identity, which returns the view unchanged. Both non-identity tasks confine their edit to one or two Perlin-noise regions (Perlin, 1985) of randomly drawn size, orientation, and position, and neither consults a foreground mask. CutPaste derives a soft target from the local source-to-destination intensity difference, the Gaussian task uses the binary region mask, and the identity uses an all-zero target.

Industrial full-shot synthesis. MVTec-AD and VisA use foreground-guided natural synthetic anomalies (Schlüter et al., 2022). For each normal destination image, a source image is drawn from the same object category and a rectangular crop of it is Poisson-blended into the destination. Crop height and width are drawn independently from a heavy-tailed distribution with category-specific bounds, capped against the destination's foreground bounding box, and then rescaled; one to a few patches are pasted per image. Placement is restricted to the object: a fixed per-image foreground mask, thresholded offline from the normal training images, supplies the paste center, a placement is accepted only when the patch lies mostly on foreground, and the final patch mask is intersected with that foreground. The dense target is a per-category calibrated sigmoid of the resulting mean absolute RGB intensity change, median-filtered to suppress isolated pixels, and a synthesis that leaves the destination effectively unchanged is dropped from the synthetic half of the batch. The crop bounds, patch counts, and target-curve parameters are fixed per object category and held identical across every anchor condition.

Internal checkpoint provenance. Every result for our medical, industrial, adapter-family, and transfer experiments comes from the checkpoint saved after the completed final training epoch. None of these internal values selects an epoch using validation or test performance. The external MVFA and MadCLIP replications follow their released selection rules, documented separately in Appendix C.

Randomness and pairing. Medical anchor-substitution tables average three training runs per condition. The random-anchor construction couples each anchor draw to its run seed. Because seed identifiers are not common to every reported condition, the main anchor comparison is made between condition means and is not presented as a seed-paired test. The architecture ablation pairs its three seeds across conditions. In the industrial randomized conditions, anchor seed equals training seed, so the three seeds represent three geometry draws as well as three training runs. Correct-prompt anchors are deterministic.

C EXTERNAL VALIDATION PROTOCOLS

These experiments retain the released visual, optimization, scoring, and evaluation pipelines of MVFA (Huang et al., 2024) and MadCLIP (Shiri et al., 2025); only the anchor path is changed. Both methods use one CLIP image encoder with multiple adapter branches. They are empirical extensions: MVFA's residual parameterization and MadCLIP's jointly learned prompts are outside the exact fixed-anchor, rotation-closed premise of Theorem 1.

MVFA architecture and controls. MVFA inserts residual adapters at multiple levels of one CLIP image encoder, then uses its published, separate detection and segmentation readouts. Its released prompt ensemble is encoded once into unit anchors $(\boldsymbol{a}_N, \boldsymbol{a}_A)$ with cosine ρ. For every random arm, Gram–Schmidt applied to two seeded Gaussian draws produces an orthonormal frame $(\boldsymbol{e}_1, \boldsymbol{e}_2)$; the frame seed equals the training seed. Correct prompts use the released anchors. Gram-

Table 20: Complete MVFA anchor-control results across support sizes (AUROC %).

		OCT2017	BraTS2021		LiverCT		RESC		Avg	
Anchor condition	K	AC	AC	AS	AC	AS	AC	AS	AC	AS
Correct prompts	2	92.0±8.2	92.3±1.0	96.8±0.6	80.5±7.9	98.9±0.9	90.9±1.3	97.5±0.8	88.9	97.7
Gram-matched random	2	93.0±6.2	90.8±1.2	96.6±0.8	79.0±7.8	98.6±1.0	90.4±2.0	97.5±0.7	88.3	97.6
Unmatched orthogonal	2	90.8±6.5	60.4±4.8	93.3±1.3	58.6±4.0	95.3±0.7	68.9±16.6	91.4±2.3	69.7	93.3
Matched orthogonal	2	92.2±6.7	90.6±1.2	96.7±0.6	77.7±7.9	98.7±1.0	91.5±0.2	97.8±0.6	88.0	97.7
Correct prompts	4	98.7±0.4	90.6±1.4	96.6±0.4	80.1±1.3	99.1±0.6	93.2±4.0	98.2±0.9	90.7	97.9
Gram-matched random	4	98.4±0.5	91.4±1.6	95.2±1.6	77.8±1.1	98.9±0.7	93.4±3.7	98.2±1.0	90.2	97.5
Unmatched orthogonal	4	94.2±2.6	61.2±3.4	93.3±0.9	59.8±8.0	96.0±0.1	73.8±16.3	92.4±4.8	72.2	93.9
Matched orthogonal	4	98.6±0.2	91.9±1.2	96.2±1.1	80.6±2.2	99.1±0.5	92.6±4.5	98.1±0.8	90.9	97.8
Correct prompts	8	99.2±0.8	92.3±2.7	97.2±0.3	80.8±3.8	98.7±1.1	97.6±0.3	99.1±0.2	92.5	98.3
Gram-matched random	8	99.5±0.3	91.7±3.1	97.2±0.3	77.9±9.9	98.3±2.1	97.7±0.1	99.0±0.0	91.7	98.2
Unmatched orthogonal	8	89.8±5.4	62.9±9.9	93.7±0.8	60.4±3.3	96.1±0.4	65.8±1.6	90.4±0.3	69.7	93.4
Matched orthogonal	8	99.5±0.2	92.0±2.7	97.2±0.2	83.0±4.7	98.7±1.0	97.6±0.2	99.2±0.1	93.0	98.4
Correct prompts	16	99.9±0.0	95.0±0.6	97.5±0.9	86.1±1.5	98.8±1.4	98.2±0.4	99.4±0.0	94.8	98.6
Gram-matched random	16	99.9±0.0	94.1±1.5	97.6±0.1	83.7±2.5	99.1±0.9	98.0±0.2	99.4±0.1	93.9	98.7
Unmatched orthogonal	16	92.2±2.1	66.8±4.0	93.1±1.1	63.0±5.2	96.5±0.3	72.9±9.1	89.9±2.4	73.7	93.2
Matched orthogonal	16	99.9±0.1	94.5±1.1	97.5±0.6	85.5±1.8	99.0±1.0	98.0±0.3	99.3±0.1	94.5	98.6

matched random anchors use $\boldsymbol{a}_N = \boldsymbol{e}_1$ and $\boldsymbol{a}_A = \rho\boldsymbol{e}_1 + \sqrt{1-\rho^2}\boldsymbol{e}_2$. Unmatched orthogonal anchors use $(\boldsymbol{e}_1, \boldsymbol{e}_2)$. Matched orthogonal anchors reuse this pair but multiply the similarity scale by $\sqrt{1-\rho}$, restoring the correct effective scale. Cosines and matched difference norms are checked to 10^{-6} before training, and the published visual, detection, and segmentation paths are otherwise unchanged.

MadCLIP architecture and controls. MadCLIP uses separate normal and abnormal residual adaptation branches within one CLIP image encoder and jointly learns its prompt tokens. The released reference learns two seven-template prompt banks with eight class-specific context tokens at the end, front, and middle token positions; the text encoder is frozen. Because its fitted cosine ρ is an outcome rather than a prescribed input, MVFA's fixed Gram-matched control is unavailable. In the text-free learned-scale arm, the text encoder is absent, a seeded orthonormal frame is frozen, and $\boldsymbol{a}_N = \boldsymbol{e}_1$, $\boldsymbol{a}_A = \rho\boldsymbol{e}_1 + \sqrt{1-\rho^2}\boldsymbol{e}_2$ with $\rho = \tanh\theta$. The scalar θ is the only trainable text-path quantity; it starts at zero and uses Adam with $\beta = (0.5, 0.999)$ and learning rate 0.02. The prompt-scale arm uses the same text-free frame but fixes $\rho = +0.986$, leaving no trainable text-path parameter. All visual branches, losses, data, schedules, and released image and pixel scoring rules remain fixed across the three arms.

Realized geometry. Across all 48 dataset–support–seed cells per MadCLIP arm, the learned prompt cosine averaged -0.984 (range -0.994 to -0.949), and the text-free learned cosine averaged -0.994 (range -1.000 to -0.975). The fixed prompt-scale arm remained at $+0.986$. Thus both fitted paths approached the maximum attainable effective scale, with or without language; the prompt-scale arm deliberately supplied a much smaller scale.

Checkpoint provenance and aggregation. Unlike our internal experiments in Appendix B, which report the completed final epoch, these external results reproduce the released repositories' test-selected checkpoint protocol. Segmentation runs select the checkpoint maximizing test pixel+image AUROC, and detection-only runs maximize test image AUROC. Each method uses three training seeds. Tables 20 and 21 report sample mean±sd. AC is image AUROC and AS is pixel AUROC. OCT2017 pixel AUROC is not reported, so Avg AS covers BraTS2021, LiverCT, and RESC; Avg AC covers all four datasets.

Table 21: Complete MadCLIP learned-prompt and text-free controls by support size (AUROC %).

		OCT2017	BraTS2021		LiverCT		RESC		Avg	
Anchor condition	K	AC	AC	AS	AC	AS	AC	AS	AC	AS
Learned prompts	2	96.5±1.5	92.9±1.0	97.6±0.3	79.0±1.2	97.3±0.3	91.0±2.8	96.8±1.1	89.8	97.2
Learned scale (no text)	2	96.3±1.6	91.8±1.4	97.6±0.2	78.6±2.2	97.5±0.4	90.9±2.7	97.1±0.8	89.4	97.4
Prompt-scale (no text)	2	97.1±1.4	91.1±2.6	97.1±0.8	64.3±16.8	71.1±22.7	87.9±5.4	96.4±0.7	85.1	88.2
Learned prompts	4	98.3±0.8	92.4±1.8	97.4±0.3	84.7±3.2	98.2±1.1	95.0±1.3	98.1±0.5	92.6	97.9
Learned scale (no text)	4	98.5±0.9	91.9±2.4	97.5±0.3	85.1±1.9	98.3±1.1	95.0±0.9	98.2±0.5	92.6	98.0
Prompt-scale (no text)	4	98.2±0.8	90.8±5.1	96.5±0.8	73.7±4.9	87.4±18.7	94.4±0.1	98.2±0.4	89.3	94.0
Learned prompts	8	98.9±0.5	94.8±1.0	97.8±0.3	88.2±2.2	99.2±0.5	96.9±1.1	98.9±0.1	94.7	98.6
Learned scale (no text)	8	98.8±0.5	94.7±1.1	97.8±0.4	87.6±1.5	99.2±0.6	96.9±1.2	98.9±0.1	94.5	98.6
Prompt-scale (no text)	8	98.6±0.3	95.1±1.1	97.7±0.3	83.6±3.8	98.7±1.0	96.3±1.1	98.8±0.2	93.4	98.4
Learned prompts	16	99.6±0.2	96.0±0.5	97.6±0.4	90.5±0.9	99.5±0.5	98.6±0.2	99.4±0.1	96.2	98.8
Learned scale (no text)	16	99.7±0.1	95.8±0.6	97.7±0.6	91.1±0.8	99.1±0.7	98.6±0.2	99.4±0.1	96.3	98.7
Prompt-scale (no text)	16	99.6±0.1	96.1±0.5	98.1±0.1	87.7±3.4	99.1±1.0	97.7±0.2	99.2±0.1	95.3	98.8

D Full Result Tables

The main paper reports the CLIP ViT-L/14 medical anchor results, the combined external-detector summary, aggregate transfer contrasts, and the compact industrial boundary experiment. The tables below give, in order, the industrial replication across all three vision backbones, the individual source–target transfer rows and their DINOv3-L/16 replication, the per-dataset adapter-family results, and the complete three-vision-backbone medical anchor factorials.

Table 22: Full-shot industrial experiment across three vision backbones (%). Each vision backbone uses a linear adapter and full-shot synthetic-anomaly training. Cells are mean±sd over three training seeds; for randomized rows, each training seed also selects the corresponding geometry seed. Pixel columns are AUROC, per-region overlap (PRO) and maximum F1 (F1max); image columns are AUROC, average precision (AUPR) and F1max. The two AUROC columns are category-macro means over the T categories; PRO, F1max and AUPR are pooled over all categories.

VisA

Vision backbone	Anchor condition	Pixel			Image		
		AUROC	PRO	F1max	AUROC	AUPR	F1max
DINOv3-L/16	Correct prompts	88.5±0.4	65.7±0.3	32.2±0.9	83.1±0.3	87.7±0.3	76.3±0.4
	Shared rotation	88.4±0.3	65.5±0.4	32.0±0.9	83.2±0.2	87.8±0.3	76.4±0.3
	Independent rotations	96.8±0.1	76.2±0.5	42.6±0.5	91.9±0.4	92.7±0.5	82.7±0.9
DINOv2-L/14	Correct prompts	92.2±0.1	68.4±1.1	32.8±0.5	89.0±0.3	90.5±0.2	80.0±0.5
	Shared rotation	92.2±0.2	68.5±0.9	32.8±0.3	89.2±0.2	90.7±0.2	80.3±0.4
	Independent rotations	97.1±0.1	76.8±1.5	41.3±0.4	92.5±0.3	93.1±0.5	83.4±0.9
CLIP ViT-L/14	Correct prompts	91.7±0.1	61.0±1.1	32.3±0.3	86.2±0.1	88.7±0.0	78.3±0.3
	Shared rotation	91.6±0.0	60.9±0.5	32.1±0.4	86.3±0.1	88.7±0.1	78.3±0.2
	Independent rotations	96.1±0.1	73.2±0.8	39.9±0.4	91.8±0.1	92.6±0.2	83.0±0.6

MVTec-AD

Vision backbone	Anchor condition	Pixel			Image		
		AUROC	PRO	F1max	AUROC	AUPR	F1max
DINOv3-L/16	Correct prompts	92.8±0.1	82.2±0.6	56.0±0.1	95.0±0.1	97.9±0.1	92.6±0.4
	Shared rotation	92.8±0.0	82.3±0.1	55.9±0.2	95.0±0.0	97.9±0.0	92.6±0.0
	Independent rotations	96.3±0.0	85.6±0.2	54.5±0.6	97.3±0.1	98.6±0.1	94.7±0.0
DINOv2-L/14	Correct prompts	94.0±0.1	85.4±0.5	59.2±0.1	96.1±0.2	98.5±0.1	94.0±0.3
	Shared rotation	94.0±0.1	85.3±0.5	59.2±0.1	96.0±0.2	98.5±0.1	93.9±0.3
	Independent rotations	96.9±0.1	85.8±0.4	55.8±0.9	97.5±0.1	98.9±0.0	95.3±0.3
CLIP ViT-L/14	Correct prompts	93.6±0.1	85.2±0.5	53.6±0.5	96.2±0.1	98.1±0.1	92.9±0.2
	Shared rotation	93.6±0.1	85.5±0.8	53.7±0.2	96.2±0.1	98.1±0.1	92.9±0.2
	Independent rotations	96.6±0.0	88.4±0.4	59.5±0.0	97.4±0.1	98.8±0.0	94.6±0.2

Table 23: Anchor swap cost under a frozen transferred adapter. CLIP ViT-L/14, K=16, image AUROC (%), mean over three seeds. A detector trained on S is evaluated on T twice, differing only in which anchors score the images: **src** uses S's, **tgt** uses T's; both share the same domain shift, so **swap** = src − tgt isolates the substitution. Δ_{swap} is swap(random) − swap(correct). Dataset keys are B=BraTS2021, L=LiverCT, O=OCT2017, R=RESC, M=Kaggle BrainMRI, U=BUSI, and C=CheXpert.

	Real-anomaly support							**Normal-only**					
	Correct prompts			**Gram-matched random**				**Correct prompts**			**Gram-matched random**		
$S\to T$	**src**	**tgt**	**swap**	**src**	**tgt**	**swap**	$S\to T$	**src**	**tgt**	**swap**	**src**	**tgt**	**swap**
B→L	63.4	63.9	−0.5	64.1	50.9	13.1	M→U	73.1	73.7	−0.6	73.9	55.2	18.7
B→O	59.7	61.1	−1.5	58.7	64.7	−6.0	M→C	57.6	56.9	0.7	57.6	54.2	3.4
B→R	45.2	44.5	0.6	44.9	48.8	−3.9	U→M	78.9	79.6	−0.7	79.7	36.8	42.9
L→B	68.9	69.2	−0.3	68.7	54.4	14.3	U→C	55.8	57.2	−1.3	58.7	51.3	7.4
L→O	69.7	62.0	7.8	69.7	55.8	13.9	C→M	75.8	68.0	7.9	74.7	32.1	42.5
L→R	66.1	66.8	−0.8	67.1	58.6	8.5	C→U	68.3	69.6	−1.3	67.0	67.1	−0.1
O→B	55.4	54.2	1.2	54.2	52.5	1.8							
O→L	62.3	60.2	2.1	63.4	56.7	6.7							
O→R	84.3	86.6	−2.4	84.0	61.9	22.1							
R→B	45.6	45.9	−0.3	45.7	54.0	−8.2							
R→L	51.6	51.4	0.2	52.0	49.2	2.8							
R→O	90.3	89.3	1.0	90.5	42.0	48.5							
mean	63.5	62.9	0.6	63.6	54.1	9.5	*mean*	68.3	67.5	0.8	68.6	49.4	19.1
Δ_{swap}				+8.9			Δ_{swap}				+18.4		

Table 24: Anchor swap cost under a frozen transferred adapter. DINOv3-L/16, K=16, image AUROC (%), mean over three seeds. A detector trained on S is evaluated on T twice, differing only in which anchors score the images: **src** uses S's, **tgt** uses T's; both share the same domain shift, so **swap** = src − tgt isolates the substitution. Δ_{swap} is swap(random) − swap(correct). The source-anchor score reuses the pair that trained the adapter and preserves its anchor frame. In the random condition, the target pair uses a different seed and changes the readout direction without co-rotating the adapter; its swap measures coordinate-frame disruption, not a benefit from linguistic meaning. Dataset keys are B=BraTS2021, L=LiverCT, O=OCT2017, R=RESC, M=Kaggle BrainMRI, U=BUSI, and C=CheXpert.

	Real-anomaly support							**Normal-only**					
	Correct prompts			**Gram-matched random**				**Correct prompts**			**Gram-matched random**		
$S\to T$	**src**	**tgt**	**swap**	**src**	**tgt**	**swap**	$S\to T$	**src**	**tgt**	**swap**	**src**	**tgt**	**swap**
B→L	71.2	73.9	−2.6	71.6	47.6	24.0	M→U	71.6	70.0	1.6	70.5	51.3	19.1
B→O	80.2	87.6	−7.3	80.3	64.1	16.2	M→C	49.3	51.1	−1.8	48.6	46.3	2.3
B→R	74.7	79.1	−4.4	75.5	50.1	25.4	U→M	93.0	92.9	0.2	92.1	51.2	40.9
L→B	67.6	67.7	−0.1	63.6	45.9	17.7	U→C	55.7	56.4	−0.8	56.5	48.1	8.4
L→O	72.4	75.6	−3.1	72.1	48.4	23.7	C→M	81.2	80.0	1.2	80.6	51.0	29.6
L→R	75.0	78.0	−3.0	75.3	50.2	25.0	C→U	45.2	47.3	−2.1	44.8	65.7	−20.9
O→B	57.5	58.9	−1.4	61.4	44.5	17.0							
O→L	67.4	64.2	3.2	65.7	57.8	7.9							
O→R	86.5	84.8	1.7	86.6	48.2	38.3							
R→B	52.9	53.0	−0.1	52.2	58.3	−6.1							
R→L	53.8	53.4	0.5	54.0	57.5	−3.5							
R→O	94.5	93.5	1.0	94.3	45.3	49.0							
mean	71.2	72.5	−1.3	71.0	51.5	19.6	*mean*	66.0	66.3	−0.3	65.5	52.3	13.2
Δ_{swap}				+20.9			Δ_{swap}				+13.5		

Table 25: Per-dataset results across adapter architectures, CLIP ViT-L/14, $K=16$, under real-anomaly support (top) and normal-only synthetic-anomaly training (bottom). Cells are mean±sd over three training seeds; AC is image-level and AS pixel-level AUROC (%). Each panel fixes the anchor condition. Avg is the unweighted mean over datasets, AS over BraTS2021, LiverCT, and RESC, for which pixel AUROC is reported. As in Table 28, the normal-only block reports image AUROC only: BUSI has evaluation masks, but its pixel AUROC is omitted to keep a common image-level endpoint, and CheXpert and Kaggle BrainMRI have no evaluation masks. The families are approximately parameter-matched: summed over the four retained layers, the linear ($1024 \to 768$), MLP ($1024 \to 448 \to 768$), and Transformer ($1024 \to 192 \to 768$) adapters have 3,148,800, 3,216,128, and 3,159,552 trainable parameters.

Real-anomaly support

Adapter	OCT2017	BraTS2021		LiverCT (BTCV+LiTS)		RESC		Avg	
	AC	AC	AS	AC	AS	AC	AS	AC	AS
Correct prompts									
Linear	99.6±0.1	96.9±0.2	99.2±0.0	91.6±1.1	99.9±0.0	97.2±0.9	99.0±0.1	96.3	99.4
MLP	99.5±0.2	97.0±0.0	99.1±0.1	92.7±1.4	99.8±0.1	97.7±1.2	99.2±0.1	96.7	99.4
Transformer	99.4±0.1	96.8±0.3	99.0±0.0	90.9±1.2	99.8±0.1	97.3±1.0	99.1±0.0	96.1	99.3
Mismatched-dataset bank									
Linear	99.7±0.1	96.9±0.2	99.3±0.0	90.8±1.3	99.9±0.0	97.2±0.9	99.0±0.1	96.1	99.4
MLP	99.5±0.4	97.1±0.1	99.1±0.1	91.7±1.3	99.8±0.1	97.8±1.1	99.2±0.1	96.5	99.4
Transformer	99.5±0.0	96.8±0.2	99.0±0.0	90.4±1.6	99.8±0.1	97.1±1.2	99.0±0.2	96.0	99.3
Gram-matched random									
Linear	99.6±0.1	96.9±0.2	99.1±0.0	91.5±1.2	99.9±0.0	97.2±0.9	99.0±0.1	96.3	99.3
MLP	99.4±0.3	97.1±0.2	99.0±0.1	93.0±0.7	99.8±0.1	97.7±1.0	99.2±0.1	96.8	99.4
Transformer	99.4±0.2	96.9±0.2	99.0±0.0	91.4±1.0	99.8±0.0	97.2±1.0	99.0±0.1	96.2	99.3
Unmatched orthogonal									
Linear	99.2±0.0	97.0±0.1	99.2±0.0	89.0±0.3	99.8±0.0	97.1±1.0	99.1±0.1	95.6	99.4
MLP	98.6±0.9	97.1±0.2	99.1±0.0	89.4±1.6	99.6±0.4	97.9±1.2	99.1±0.2	95.8	99.3
Transformer	81.7±27.4	96.9±0.2	99.0±0.0	87.3±0.9	98.9±0.8	97.1±1.0	99.1±0.1	90.7	99.0
Matched orthogonal									
Linear	99.6±0.1	96.9±0.2	99.1±0.0	91.6±1.4	99.9±0.0	97.2±0.9	99.0±0.1	96.3	99.3
MLP	99.3±0.2	97.1±0.1	99.1±0.1	92.9±1.4	99.8±0.1	97.8±0.9	99.2±0.2	96.8	99.4
Transformer	99.3±0.1	96.8±0.2	99.0±0.0	91.1±1.3	99.8±0.0	97.4±1.0	99.0±0.1	96.1	99.3

Normal-only

Adapter	Kaggle BrainMRI	CheXpert	BUSI	Avg
	AC	AC	AC	AC
Correct prompts				
Linear	92.9±0.8	69.3±1.6	88.2±0.6	83.4
MLP	94.8±0.9	71.6±1.0	89.2±0.4	85.2
Transformer	86.7±2.0	66.7±1.5	86.2±0.6	79.9
Mismatched-dataset bank				
Linear	92.9±0.7	69.0±2.3	87.7±0.9	83.2
MLP	94.9±0.8	72.3±1.5	89.3±0.9	85.5
Transformer	89.8±2.6	65.2±1.5	86.0±0.7	80.3
Gram-matched random				
Linear	92.9±0.2	69.5±1.3	88.4±0.6	83.6
MLP	94.7±0.7	72.3±0.4	89.3±1.0	85.4
Transformer	88.2±2.4	65.1±1.8	84.7±1.8	79.4
Unmatched orthogonal				
Linear	92.5±0.4	69.8±1.3	87.8±0.6	83.4
MLP	94.0±1.3	71.5±1.0	89.0±0.5	84.8
Transformer	92.7±0.2	67.3±2.1	87.4±1.2	82.5
Matched orthogonal				
Linear	92.9±0.7	69.4±1.7	87.8±0.8	83.4
MLP	95.1±0.3	72.0±0.4	89.4±1.0	85.5
Transformer	89.2±3.7	64.2±1.9	85.2±1.4	79.5

Table 26: Per-dataset results across adapter architectures, DINOv3-L/16, $K=16$, under real-anomaly support (top) and normal-only synthetic-anomaly training (bottom). Cells are mean±sd over three training seeds; AC is image-level and AS pixel-level AUROC (%). Each panel fixes the anchor condition. Avg is the unweighted mean over datasets, AS over BraTS2021, LiverCT, and RESC, for which pixel AUROC is reported. As in Table 28, the normal-only block reports image AUROC only: BUSI has evaluation masks, but its pixel AUROC is omitted to keep a common image-level endpoint, and CheXpert and Kaggle BrainMRI have no evaluation masks. This table replicates Table 25 on a second vision backbone.

Real-anomaly support

Adapter	OCT2017	BraTS2021		LiverCT (BTCV+LiTS)		RESC		Avg	
	AC	AC	AS	AC	AS	AC	AS	AC	AS
Correct prompts									
Linear	99.3±0.1	97.2±0.2	99.4±0.1	91.4±0.9	99.9±0.0	97.9±0.7	99.3±0.1	96.4	99.5
MLP	98.9±0.1	97.2±0.3	99.3±0.1	91.1±2.3	99.8±0.1	98.1±0.6	99.4±0.0	96.3	99.5
Transformer	99.1±0.5	97.0±0.2	99.3±0.1	91.1±0.9	99.8±0.1	97.8±0.9	99.2±0.0	96.2	99.4
Mismatched-dataset bank									
Linear	99.3±0.2	97.1±0.2	99.4±0.1	90.9±1.0	99.9±0.0	97.8±0.8	99.2±0.1	96.3	99.5
MLP	99.1±0.2	97.1±0.4	99.3±0.1	91.1±2.4	99.8±0.2	98.0±0.6	99.3±0.1	96.3	99.4
Transformer	99.3±0.3	96.9±0.4	99.3±0.1	90.7±1.7	99.6±0.4	97.9±0.8	99.2±0.0	96.2	99.4
Gram-matched random									
Linear	99.2±0.3	97.2±0.2	99.4±0.1	91.4±0.9	99.9±0.1	97.9±0.7	99.2±0.1	96.4	99.5
MLP	99.2±0.2	97.2±0.1	99.2±0.1	91.1±1.8	99.8±0.1	98.0±0.6	99.3±0.1	96.4	99.5
Transformer	99.1±0.4	97.1±0.2	99.3±0.1	90.9±1.4	99.8±0.1	97.7±1.0	99.1±0.1	96.2	99.4
Unmatched orthogonal									
Linear	98.7±0.1	97.1±0.2	99.4±0.1	88.9±1.6	99.8±0.1	97.7±0.8	99.2±0.1	95.6	99.5
MLP	98.0±0.8	97.2±0.1	99.4±0.1	87.9±1.8	99.7±0.2	98.1±0.9	99.4±0.1	95.3	99.5
Transformer	50.0±0.0	97.1±0.1	99.4±0.1	65.1±19.9	55.2±38.5	97.9±0.6	99.2±0.0	77.5	84.6
Matched orthogonal									
Linear	99.2±0.3	97.1±0.2	99.4±0.1	91.3±0.8	99.9±0.0	97.9±0.7	99.2±0.1	96.4	99.5
MLP	99.0±0.3	97.1±0.3	99.3±0.1	90.8±1.3	99.8±0.1	98.1±0.6	99.3±0.1	96.3	99.5
Transformer	99.2±0.2	97.0±0.1	99.3±0.1	91.3±1.5	99.7±0.2	98.0±0.7	99.2±0.1	96.4	99.4

Normal-only

Adapter	Kaggle BrainMRI	CheXpert	BUSI	Avg
	AC	AC	AC	AC
Correct prompts				
Linear	95.2±0.4	70.8±0.9	89.2±0.8	85.0
MLP	94.5±0.3	69.4±0.7	88.8±0.9	84.2
Transformer	94.5±0.6	65.5±1.3	89.0±0.5	83.0
Mismatched-dataset bank				
Linear	95.0±0.3	70.8±1.1	89.9±1.0	85.2
MLP	94.4±0.6	69.4±0.5	89.0±0.5	84.3
Transformer	94.2±0.5	64.9±0.8	89.1±1.2	82.7
Gram-matched random				
Linear	95.2±0.4	70.8±1.0	89.4±0.6	85.1
MLP	94.5±0.3	69.1±0.5	87.9±1.2	83.8
Transformer	94.3±0.2	64.7±2.4	88.9±0.6	82.6
Unmatched orthogonal				
Linear	93.5±1.0	70.3±1.7	88.1±2.2	84.0
MLP	94.9±0.6	69.5±0.3	89.8±3.2	84.8
Transformer	94.0±0.8	66.1±1.8	89.0±1.0	83.0
Matched orthogonal				
Linear	95.1±0.5	70.6±0.6	89.6±0.7	85.1
MLP	94.6±0.5	69.7±1.3	89.4±0.6	84.5
Transformer	94.4±0.5	64.7±1.1	88.6±0.9	82.6

Table 27: Anchor substitution with real-anomaly support (K normal + K observed abnormal samples). Cells are mean±sd over three training seeds; AC is image-level and AS pixel-level AUROC (%). Avg is the unweighted mean over datasets, AS over BraTS2021, LiverCT, and RESC, for which pixel AUROC is reported. Rows within a block share support samples, training and scoring; only the anchor pair differs.

K	Anchor condition	OCT2017	BraTS2021		LiverCT (BTCV+LiTS)		RESC		Avg	
		AC	AC	AS	AC	AS	AC	AS	AC	AS
2	**DINOv3-L/16**									
	Correct prompts	93.6±6.7	94.9±1.0	98.8±0.3	81.8±7.1	99.4±0.4	93.2±1.9	97.0±1.2	90.9	98.4
	Mismatched-dataset bank	94.9±4.8	95.0±0.8	98.8±0.3	80.5±6.8	99.4±0.4	93.3±1.7	97.2±1.1	90.9	98.5
	Gram-matched random	93.7±5.3	94.8±1.1	98.8±0.3	81.5±7.3	99.4±0.4	93.2±2.1	97.0±1.2	90.8	98.4
	Unmatched orthogonal	81.8±15.9	94.4±1.1	98.5±0.6	78.2±7.7	97.1±4.1	93.2±2.1	97.1±1.1	86.9	97.6
	Matched orthogonal	93.9±6.4	94.7±1.2	98.8±0.3	81.0±7.5	99.4±0.3	93.3±2.0	97.1±1.2	90.7	98.4
	DINOv2-L/14									
	Correct prompts	94.5±6.2	94.8±1.2	98.4±0.6	85.9±2.8	99.3±0.6	90.2±3.2	96.2±2.0	91.4	98.0
	Mismatched-dataset bank	95.5±4.6	95.0±0.9	98.7±0.4	84.6±3.4	99.1±0.8	89.8±3.0	96.2±2.0	91.2	98.0
	Gram-matched random	94.6±5.9	94.7±1.3	98.4±0.5	86.7±2.7	99.4±0.6	89.8±3.0	96.2±1.9	91.5	98.0
	Unmatched orthogonal	93.0±7.1	94.4±1.4	98.2±0.9	80.5±4.9	97.1±4.1	89.9±3.8	96.1±2.4	89.5	97.2
	Matched orthogonal	94.2±6.5	94.7±1.4	98.4±0.5	86.2±3.0	99.3±0.6	89.8±3.3	96.2±1.9	91.2	98.0
	CLIP ViT-L/14									
	Correct prompts	94.0±3.3	94.7±0.6	98.0±0.7	80.7±8.3	99.1±0.7	91.9±2.5	96.8±1.4	90.3	97.9
	Mismatched-dataset bank	94.2±2.2	94.8±0.9	98.4±0.4	79.7±8.5	98.1±2.3	91.6±2.6	96.7±1.5	90.1	97.7
	Gram-matched random	93.8±3.5	94.6±0.5	98.1±0.6	80.3±8.1	99.0±0.7	91.8±2.6	96.8±1.4	90.1	98.0
	Unmatched orthogonal	92.5±4.4	93.6±2.4	97.7±1.4	72.8±5.2	98.8±1.3	91.4±2.7	96.8±1.3	87.6	97.8
	Matched orthogonal	94.5±4.3	94.6±0.6	98.0±0.6	80.4±7.9	99.1±0.7	91.6±2.7	96.7±1.4	90.3	97.9
4	**DINOv3-L/16**									
	Correct prompts	97.2±1.4	96.0±1.1	99.2±0.0	87.0±3.8	99.7±0.2	93.9±2.9	98.0±0.5	93.5	98.9
	Mismatched-dataset bank	98.0±0.8	96.0±0.9	99.2±0.0	85.3±2.6	99.6±0.2	93.8±3.0	97.9±0.6	93.3	98.9
	Gram-matched random	96.9±1.5	95.9±1.0	99.2±0.0	85.8±2.0	99.7±0.1	93.9±2.9	98.0±0.6	93.1	98.9
	Unmatched orthogonal	95.5±1.8	95.8±1.2	99.1±0.1	82.6±2.6	99.4±0.5	94.2±2.2	98.3±0.3	92.0	98.9
	Matched orthogonal	97.6±0.9	95.9±1.0	99.2±0.0	86.7±3.8	99.7±0.1	93.9±2.9	97.9±0.6	93.5	98.9
	DINOv2-L/14									
	Correct prompts	96.5±3.2	95.9±0.6	99.0±0.2	88.8±2.5	99.8±0.1	93.7±2.4	97.8±0.5	93.7	98.9
	Mismatched-dataset bank	97.6±1.7	96.0±0.7	99.1±0.1	88.2±2.5	99.8±0.1	93.3±2.2	97.8±0.4	93.8	98.9
	Gram-matched random	97.4±1.5	95.9±0.6	99.0±0.2	89.0±2.6	99.8±0.1	93.6±2.3	97.8±0.5	94.0	98.9
	Unmatched orthogonal	96.4±1.7	95.9±0.7	99.1±0.1	86.2±2.1	99.5±0.5	93.3±2.4	97.9±0.4	92.9	98.8
	Matched orthogonal	97.4±1.7	95.9±0.6	99.0±0.2	88.9±2.8	99.8±0.1	93.6±2.2	97.8±0.5	93.9	98.9
	CLIP ViT-L/14									
	Correct prompts	97.5±1.7	95.9±0.2	98.9±0.0	84.4±5.0	99.7±0.2	92.8±2.4	97.9±0.4	92.7	98.8
	Mismatched-dataset bank	97.6±1.7	95.8±0.3	99.0±0.1	83.9±4.4	99.7±0.2	92.8±2.3	97.8±0.4	92.5	98.8
	Gram-matched random	97.6±1.5	95.8±0.2	98.9±0.1	84.6±4.9	99.7±0.2	92.8±2.5	97.8±0.5	92.7	98.8
	Unmatched orthogonal	96.2±1.3	96.1±0.3	98.9±0.1	82.8±4.9	99.5±0.3	92.7±2.4	98.0±0.5	91.9	98.8
	Matched orthogonal	97.4±1.8	95.8±0.2	98.9±0.0	84.4±4.7	99.7±0.2	92.9±2.4	97.8±0.4	92.6	98.8
8	**DINOv3-L/16**									
	Correct prompts	97.8±0.8	96.1±0.3	99.2±0.1	89.4±3.0	98.6±1.7	97.3±0.9	98.7±0.4	95.2	98.8
	Mismatched-dataset bank	98.0±0.8	96.0±0.2	99.1±0.1	88.8±3.3	97.5±3.6	97.2±1.0	98.7±0.5	95.0	98.4
	Gram-matched random	97.2±1.6	96.1±0.3	99.1±0.2	89.4±2.8	98.6±1.7	97.2±1.0	98.6±0.4	95.0	98.8
	Unmatched orthogonal	96.1±2.0	96.1±0.2	99.2±0.2	86.1±2.7	95.3±7.4	97.4±0.5	98.7±0.4	93.9	97.8
	Matched orthogonal	96.8±1.8	96.2±0.4	99.1±0.2	89.3±2.8	98.8±1.4	97.2±1.0	98.6±0.4	94.9	98.9
	DINOv2-L/14									
	Correct prompts	98.8±0.8	96.4±0.1	99.0±0.2	89.6±4.2	99.2±0.9	97.3±0.6	98.5±0.4	95.5	98.9
	Mismatched-dataset bank	99.1±0.5	96.5±0.1	99.2±0.1	89.2±3.9	98.8±1.6	97.3±0.6	98.5±0.4	95.5	98.8
	Gram-matched random	99.0±0.5	96.4±0.2	99.0±0.2	89.6±3.9	99.3±0.8	97.3±0.5	98.5±0.4	95.6	98.9
	Unmatched orthogonal	98.4±0.3	96.2±0.1	99.1±0.1	86.5±3.3	96.4±4.7	97.5±0.8	98.8±0.2	94.7	98.1
	Matched orthogonal	98.9±0.4	96.4±0.1	99.0±0.2	89.3±3.9	99.3±0.8	97.3±0.5	98.5±0.4	95.5	98.9
	CLIP ViT-L/14									
	Correct prompts	98.5±0.1	96.0±0.3	98.7±0.5	87.5±3.0	98.8±1.5	95.9±1.2	98.4±0.3	94.5	98.6
	Mismatched-dataset bank	98.6±0.1	96.0±0.4	99.0±0.2	87.0±3.2	98.2±2.3	95.9±1.1	98.2±0.5	94.4	98.5
	Gram-matched random	98.2±0.5	95.9±0.3	98.7±0.5	87.8±2.6	98.7±1.6	96.0±1.1	98.3±0.3	94.5	98.6
	Unmatched orthogonal	97.8±0.3	95.9±0.2	98.9±0.3	84.3±3.2	99.0±0.9	95.8±1.0	98.5±0.3	93.5	98.8
	Matched orthogonal	98.3±0.3	95.9±0.4	98.7±0.5	87.7±2.9	98.7±1.5	95.9±1.2	98.3±0.3	94.5	98.6
16	**DINOv3-L/16**									
	Correct prompts	99.3±0.1	97.2±0.2	99.4±0.1	91.4±0.9	99.9±0.0	97.9±0.7	99.3±0.1	96.4	99.5
	Mismatched-dataset bank	99.3±0.2	97.1±0.2	99.4±0.1	90.9±1.0	99.9±0.0	97.8±0.8	99.2±0.1	96.3	99.5
	Gram-matched random	99.2±0.3	97.2±0.2	99.4±0.1	91.4±0.9	99.9±0.1	97.9±0.7	99.2±0.1	96.4	99.5
	Unmatched orthogonal	98.7±0.1	97.1±0.2	99.4±0.1	88.9±1.6	99.8±0.1	97.7±0.8	99.2±0.1	95.6	99.5
	Matched orthogonal	99.2±0.3	97.1±0.2	99.4±0.1	91.3±0.8	99.9±0.0	97.9±0.7	99.2±0.1	96.4	99.5
	DINOv2-L/14									
	Correct prompts	99.6±0.0	97.3±0.2	99.4±0.1	92.1±1.4	99.9±0.1	98.2±0.6	99.3±0.1	96.8	99.5
	Mismatched-dataset bank	99.7±0.0	97.3±0.2	99.4±0.0	91.4±1.4	99.8±0.1	98.0±0.7	99.3±0.1	96.6	99.5
	Gram-matched random	99.6±0.0	97.3±0.2	99.4±0.1	92.2±1.2	99.9±0.1	98.2±0.6	99.3±0.1	96.8	99.5
	Unmatched orthogonal	99.3±0.1	97.2±0.3	99.4±0.1	89.5±1.1	99.5±0.7	98.0±0.9	99.3±0.1	96.0	99.4
	Matched orthogonal	99.6±0.0	97.2±0.2	99.4±0.1	92.3±1.1	99.9±0.1	98.1±0.6	99.3±0.1	96.8	99.5
	CLIP ViT-L/14									
	Correct prompts	99.6±0.1	96.9±0.2	99.2±0.0	91.6±1.1	99.9±0.0	97.2±0.9	99.0±0.1	96.3	99.4
	Mismatched-dataset bank	99.7±0.1	96.9±0.2	99.3±0.0	90.8±1.3	99.9±0.0	97.2±0.9	99.0±0.1	96.1	99.4
	Gram-matched random	99.6±0.1	96.9±0.2	99.1±0.0	91.5±1.2	99.9±0.0	97.2±0.9	99.0±0.1	96.3	99.3
	Unmatched orthogonal	99.2±0.0	97.0±0.1	99.2±0.0	89.0±0.3	99.8±0.0	97.1±1.0	99.1±0.1	95.6	99.4
	Matched orthogonal	99.6±0.1	96.9±0.2	99.1±0.0	91.6±1.4	99.9±0.0	97.2±0.9	99.0±0.1	96.3	99.3

Table 28: Anchor substitution in the normal-only regime (K normal samples; synthetic anomalies generated online). Cells are mean±sd over three training seeds of image-level AUROC (%). BUSI has evaluation masks, but its pixel AUROC is omitted to keep a common image-level endpoint across all three datasets; CheXpert and Kaggle BrainMRI have no evaluation masks. Conditions are as in the real-anomaly-support table; the two supervision regimes are never pooled.

K	Anchor condition	Kaggle BrainMRI	CheXpert	BUSI	Avg
		AC	AC	AC	AC
4	**DINOv3-L/16**				
	Correct prompts	93.6±1.7	71.1±1.3	87.0±2.7	83.9
	Mismatched-dataset bank	94.1±1.6	72.2±1.2	87.3±1.4	84.5
	Gram-matched random	93.5±2.4	71.5±1.3	87.0±2.1	84.0
	Unmatched orthogonal	91.0±2.7	69.9±4.2	84.9±1.8	82.0
	Matched orthogonal	93.7±2.3	71.7±0.9	86.9±1.2	84.1
	DINOv2-L/14				
	Correct prompts	92.6±1.4	66.8±4.7	85.4±2.4	81.6
	Mismatched-dataset bank	91.9±1.6	66.7±3.4	85.7±1.4	81.4
	Gram-matched random	92.3±1.7	67.1±3.6	86.2±1.1	81.9
	Unmatched orthogonal	92.1±2.1	67.2±3.6	85.3±0.7	81.5
	Matched orthogonal	92.5±1.3	66.6±4.1	85.9±1.5	81.7
	CLIP ViT-L/14				
	Correct prompts	91.6±2.0	69.6±3.1	86.2±1.9	82.5
	Mismatched-dataset bank	92.4±1.4	70.0±3.2	86.0±0.4	82.8
	Gram-matched random	92.0±2.6	70.0±3.5	85.9±1.4	82.6
	Unmatched orthogonal	91.9±2.4	70.1±3.6	84.4±0.9	82.1
	Matched orthogonal	91.7±3.0	70.1±2.8	86.2±2.0	82.7
8	**DINOv3-L/16**				
	Correct prompts	95.1±1.0	70.2±1.1	89.5±1.0	84.9
	Mismatched-dataset bank	95.2±0.8	70.0±1.1	89.6±1.1	84.9
	Gram-matched random	95.1±1.0	70.3±0.6	89.5±0.9	85.0
	Unmatched orthogonal	93.1±1.0	67.4±2.4	86.1±1.9	82.2
	Matched orthogonal	95.3±1.0	70.3±0.8	89.2±1.1	84.9
	DINOv2-L/14				
	Correct prompts	93.6±0.2	67.2±3.1	88.2±1.7	83.0
	Mismatched-dataset bank	93.8±0.1	66.5±2.9	88.5±1.5	82.9
	Gram-matched random	93.4±0.4	67.3±3.3	88.2±2.4	82.9
	Unmatched orthogonal	93.7±0.3	68.0±3.4	88.2±1.6	83.3
	Matched orthogonal	93.6±0.3	66.6±2.8	88.6±1.8	82.9
	CLIP ViT-L/14				
	Correct prompts	91.2±0.9	68.7±2.5	87.5±0.9	82.4
	Mismatched-dataset bank	92.5±0.7	69.0±3.0	87.3±0.9	83.0
	Gram-matched random	91.9±0.4	68.8±2.7	87.3±0.1	82.7
	Unmatched orthogonal	91.8±0.8	69.3±2.9	87.2±0.8	82.8
	Matched orthogonal	91.9±0.5	69.1±2.3	87.1±1.3	82.7
16	**DINOv3-L/16**				
	Correct prompts	95.2±0.4	70.8±0.9	89.2±0.8	85.0
	Mismatched-dataset bank	95.0±0.3	70.8±1.1	89.9±1.0	85.2
	Gram-matched random	95.2±0.4	70.8±1.0	89.4±0.6	85.1
	Unmatched orthogonal	93.5±1.0	70.3±1.7	88.1±2.2	84.0
	Matched orthogonal	95.1±0.5	70.6±0.6	89.6±0.7	85.1
	DINOv2-L/14				
	Correct prompts	93.7±0.7	67.0±3.0	90.0±0.7	83.6
	Mismatched-dataset bank	93.6±0.5	66.7±3.0	89.8±0.6	83.4
	Gram-matched random	93.7±0.6	66.8±3.0	89.3±0.6	83.3
	Unmatched orthogonal	94.2±0.5	67.7±2.0	89.7±0.3	83.9
	Matched orthogonal	93.8±0.7	66.5±2.7	89.9±0.8	83.4
	CLIP ViT-L/14				
	Correct prompts	92.9±0.8	69.3±1.6	88.2±0.6	83.4
	Mismatched-dataset bank	92.9±0.7	69.0±2.3	87.7±0.9	83.2
	Gram-matched random	92.9±0.2	69.5±1.3	88.4±0.6	83.6
	Unmatched orthogonal	92.5±0.4	69.8±1.3	87.8±0.6	83.4
	Matched orthogonal	92.9±0.7	69.4±1.7	87.8±0.8	83.4
32	**DINOv3-L/16**				
	Correct prompts	94.7±0.2	69.9±1.7	87.4±0.4	84.0
	Mismatched-dataset bank	94.6±0.2	69.6±1.9	88.0±0.2	84.1
	Gram-matched random	94.2±0.3	69.9±1.9	87.2±0.5	83.8
	Unmatched orthogonal	92.8±1.4	67.6±1.5	87.0±1.5	82.5
	Matched orthogonal	94.2±0.0	69.9±1.3	87.3±0.2	83.8
	DINOv2-L/14				
	Correct prompts	94.2±0.1	65.8±1.9	89.5±0.7	83.2
	Mismatched-dataset bank	94.1±0.1	65.9±1.8	89.6±0.3	83.2
	Gram-matched random	94.1±0.1	65.8±2.0	89.4±0.5	83.1
	Unmatched orthogonal	94.1±0.2	66.7±2.3	88.9±0.1	83.2
	Matched orthogonal	94.0±0.4	65.8±1.9	89.8±0.4	83.2
	CLIP ViT-L/14				
	Correct prompts	93.0±0.3	69.7±1.0	88.1±0.4	83.6
	Mismatched-dataset bank	92.9±0.2	69.3±1.4	87.4±0.2	83.2
	Gram-matched random	92.9±0.3	68.8±1.1	87.8±0.5	83.2
	Unmatched orthogonal	92.8±0.4	70.1±1.1	87.6±0.3	83.5
	Matched orthogonal	92.9±0.1	69.1±1.4	87.8±0.2	83.3