WHAT DO TEXT PROMPTS ACTUALLY CARRY IN ZERO-SHOT ANOMALY DETECTION? GEOMETRY, NOT SEMANTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot anomaly detection with CLIP is standardly credited to the defect semantics of text prompts. This assumption is central for two reasons. The text branch is the only part that ever mentions defects, and prompt design is the field's main axis of progress. We argue instead that the text branch contributes only geometry, not semantics: a single direction from normal to anomalous in feature space. Vision alone supplies a functionally equivalent one, to the same effect. (1) **Reduction.** The text branch of both representative pipelines provably reduces to a monotone function of a single projection onto one direction (rank correlation > 0.99), which turns the question of what text contributes into the question of what makes this direction work: geometry or semantics. (2) **Substitution.** A direction computed from a few normal images, swapped in for the text pair with no other change, matches the complete pipeline on all combinations of pipeline and benchmark. The direction proves not only sufficient but exhaustive: across 5k diverse prompts, directional alignment alone determines performance ($R^2 \geq 0.8736$) and semantics add nothing ($\Delta R^2 \leq 0.0012$). (3) **Prediction.** For 10k entirely unseen prompts, we first forecast each prompt's performance, one forecast per prompt, from a single geometric quantity, without running the pipeline. We then run the pipeline on all 10k prompts and measure the actual performance of each. The per-prompt forecasts match the per-prompt measurements closely, ranking unseen prompts at correlation 0.8440 to 0.8643 ($p \approx 0$) with no systematic bias (calibration slope within 0.04 of unity). Across these analyses, text prompts in zero-shot anomaly detection carry geometry, not semantics. The geometry is a direction from normal to anomalous, and vision alone supplies an equivalent one, to the same effect. We call this the **modal load-bearing** of text.

1 INTRODUCTION

Zero-shot anomaly detection (ZSAD) aims to identify anomalies in categories unseen during training, with applications in various domains including industrial visual inspection (Ma et al., 2026), medical diagnosis (Zhang & Zhang, 2026), and remote sensing (Lin et al., 2026). The dominant pipelines are built on frozen CLIP (Radford et al., 2021), and they fall into two families according to how the text input is produced. The first family uses explicit hand-written text input, represented by WinCLIP (Jeong et al., 2023). It encodes normal and anomalous prompt templates, averages each side into one text embedding, and scores an image by its similarity to the two embeddings. The second family, represented by AnomalyCLIP (Zhou et al., 2024), replaces the hand-written templates with textual prompts learned on auxiliary data. The success of both families is standardly credited to the defect semantics of text prompts: words like damaged, flawed, and defective are assumed to tell the frozen vision encoder what a defect is. This assumption is central for two reasons. First, the text branch is the only part of these pipelines that ever mentions defects. The remaining components, a frozen vision encoder and a scoring head, are designed to be defect-agnostic. Second, prompt design is a main axis of progress in this literature: from hand-written templates (Jeong et al., 2023) to learned textual prompts (Zhou et al., 2024; Hu et al., 2026), careful text design has repeatedly improved detection performance. In short, both the design of these pipelines and the progress of the field rest on one assumption: text prompts carry the defect semantics.

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.

Yet the assumption has never been tested directly, and an observation in the literature sits awkwardly with it: removing the text branch entirely and learning two visual prototypes in its place matches or slightly exceeds the text-based pipeline, with 99% fewer parameters (Hou et al., 2026). If text were supplying defect semantics, removing it should hurt. It does not. A preview experiment sharpens the doubt. We compute a single direction from sixteen normal training images, with no labels and no text, and construct a language-free normal/anomalous pair around it, shaped to slot into the text pair's place. Substituted into WinCLIP, this language-free pair matches the full ensemble of hand-written templates. Substituted into AnomalyCLIP, it matches the learned prompts. Both results are verified end-to-end on MVTec AD (Bergmann et al., 2019) and VisA (Zou et al., 2022), and further extend to five more benchmarks spanning medical (Salehi et al., 2021; Hamada, 2020), industrial (Jezek et al., 2021), and logical domains (Bergmann et al., 2022) (Secs. 3.2 and 5). If a language-free direction from a few normal images can do the text prompt's job, what exactly does the text prompt contribute?

Two hypotheses compete. Under the **geometry-load hypothesis**, words serve only to define a geometric direction, and an equivalent one can be extracted from image features. Under the **semantic-load hypothesis**, words contribute defect semantics that the image features lack. The quantity at stake, meaning the part of the detection decision that the text branch, and only the text branch, can carry, is what we call the modal load-bearing of text. Our results support the geometry-load hypothesis. The text branch contributes only geometry, not semantics: a single direction from normal to anomalous in feature space. Vision alone supplies a functionally equivalent one, to the same effect.

We establish this verdict in three steps. (1) **Reduction** (Sec. 2). The text branch of both representative pipelines provably reduces to a monotone function of a single projection onto one direction, which turns the question of what text contributes into the question of what makes this direction work: geometry or semantics. (2) **Substitution** (Sec. 3). A direction computed from a few normal images, swapped in for the text pair with no other change, matches the full pipeline on every pipeline-benchmark combination. The direction proves not only sufficient but exhaustive: across 5k diverse prompts, directional alignment alone determines performance and semantics add nothing. (3) **Prediction** (Sec. 4). For 10k entirely unseen prompts, we first forecast each prompt's performance, one forecast per prompt, from a single geometric quantity, without running the pipeline. We then run the pipeline on all 10k prompts and measure the actual performance of each. Across multiple statistical metrics, the per-prompt measurements confirm that the per-prompt predictions closely match. Section 5 then maps where the geometric account stops, along two axes. In breadth, the replacement holds on five further benchmarks spanning medical, industrial, and logical domains. In depth, across six tasks of rising semantic demand, the visual direction keeps pace with text far beyond standard detection, and the only genuine advantage of text appears where both sides are near zero.

Our contributions are summarized as follows.

- **Verdict.** We establish that text prompts in zero-shot anomaly detection contribute only geometry, not semantics, and that vision alone supplies a functionally equivalent one, to the same effect. The verdict holds across seven benchmarks and a six-task spectrum of rising semantic demand.
- **Protocol.** We support the verdict with a reusable testing protocol that combines algebraic reduction, component substitution, causal interventions, and large-scale prompt studies.
- **Prediction.** We turn the verdict into a predictive rule: a single geometric quantity forecasts the performance of an entirely unseen prompt, without running the pipeline. We compute one forecast per prompt for 10k such new prompts, then measure the actual performance of each by running the pipeline. The per-prompt forecasts prove highly accurate against the measurements.

2 Reduction: the Text Branch Is Exactly One Direction

Before asking what the text branch contributes, we first make the question mathematical. We prove that in both representative pipelines, one with hand-written prompts and one with learned prompts, the text branch influences the score through a single direction in feature space, and nothing else. The broad question "what does text contribute?" therefore reduces to a specific one, "what makes this direction work?" We establish the reduction twice over, first by algebra and then by experiments.

We first prove the reduction by algebra. Both pipelines score an image by the anomalous channel of a two-class softmax over cosine similarities to a normal and an anomalous text embedding. A short derivation shows that this score is a strictly monotone function of a single projection $\hat{f} \cdot d$, where

Table 1: The reduction holds. Replacing the text apparatus by $\hat{f} \cdot d_t$, we report the per-class rank correlation ρ between the two scores and the worst per-class I-AUROC change.

Text input	Pipeline	Dataset	ρ (min / median / max)	max $\lvert\Delta$I-AUROC$\rvert$
hand-written	WinCLIP	MVTec AD	0.9922 / 0.9986 / 1.0000	0.0116
	WinCLIP	VisA	0.9952 / 0.9992 / 0.9998	0.0105
learned	AnomalyCLIP	MVTec AD	1.0000 / 1.0000 / 1.0000	0.0000
	AnomalyCLIP	VisA	1.0000 / 1.0000 / 1.0000	0.0000

d is the normalized difference of the two text embeddings, so every rank-based metric (including I-AUROC) is invariant to replacing the whole text apparatus by d. The full derivation is given in Appendix A.

We then corroborate the algebra experimentally. Table 1 confirms the reduction on both running pipelines. Per-image scores of the original head and of $\hat{f} \cdot d_t$ agree at rank correlation $\rho \geq 0.9922$ on WinCLIP and exactly $\rho = 1$ on AnomalyCLIP, and the worst-case I-AUROC change per class is at most 0.0116. Per-class values are given in Table 8 of Appendix D. The reduction is exact for AnomalyCLIP but near-exact for WinCLIP, and the difference has a structural cause. AnomalyCLIP's head is a pure cosine contrast, so the algebra of equation 8 applies without residue. WinCLIP's ensemble head contains a minuscule non-monotone term, which the single-direction form cannot capture, so the agreement is near-exact rather than perfect.

Through both algebra and experiments, everything the text branch contributes to the score reduces to one direction. No matter how many templates are ensembled or how the prompt learner is trained, the entire effect on the ranking passes through one unit vector d. The question of what text contributes therefore becomes the question of what makes this direction work, and two hypotheses compete. On the geometry-load hypothesis, a good direction is an intrinsic geometric property of the feature space itself, independent of semantics. On the semantic-load hypothesis, the direction works because defect words pointed it the right way. The next section decides between the two.

3 Substitution: Geometry, Not Semantics, Makes the Direction Work, and Vision Supplies an Equivalent One

Section 2 reduced the text branch to a single direction, but left one question open: what makes this direction work, geometry or semantics? This section shows that it is geometry, not semantics: a single direction from normal to anomalous in feature space. Vision alone supplies a functionally equivalent one, to the same effect. We establish this in four steps. First, frozen image features contain a usable normal/anomaly direction of their own, which already achieves strong detection performance (Sec. 3.1). Second, swapping this direction in place of the text pair, with no other change to the pipeline, matches the performance of the complete text-based pipeline (Sec. 3.2). Third, erasure interventions show that the visual direction is no shadow of the text direction: the two directions are nearly orthogonal, and the text score does not depend on the visual axis (Sec. 3.3). Fourth, an exhaustive study of 5k diverse prompts shows that directional alignment alone determines prompt performance, and semantics add nothing on top of the direction (Sec. 3.4).

3.1 Existence: a few normal images suffice to read a usable direction

On the geometric account, a good direction is a property of the feature space itself, so it should be readable from images as much as from words. We therefore ask the most basic question: do frozen image features, without any text, already contain a usable normal/anomaly direction? We test this in its strongest form, by asking how little supervision the extraction needs. A supervised probe sets the ceiling, a fully label-free procedure sets the floor, and few-shot extraction from k normal images traces the curve between them.

We extract visual directions under three levels of supervision. In the fully supervised setting, a logistic probe trained on pixel-level labels (downsampled to the patch grid) provides a fully supervised direction, marking what the features can offer at best. In the few-shot setting, we are given k normal

training images together with the unlabeled test set of the class, a batch setting shared by recent zero-shot detectors (Li et al., 2026). We estimate the normal mean μ and covariance Σ from the k normal images, Cholesky-whiten all test features with them, average the whitened vectors whose Mahalanobis norm falls in the top 1%, and map the result back to the original space:

$$\hat{d}_v \;=\; \text{normalize}\Big(L^{-\top}\,\text{normalize}\Big(\tfrac{1}{|\mathcal{T}|}\textstyle\sum_{w\in\mathcal{T}} w\Big)\Big)\,, \qquad \mathcal{T} = \text{top-1\% of } \|w\|\,. \tag{1}$$

In the zero-shot setting, no training data of the class are used as reference at all.

Figure 1 presents the results as supervision decreases from full labels to none. Full numbers with error bars are given in Table 9 of Appendix D. Three observations. (1) The direction exists: the fully supervised direction reaches 0.9824 on MVTec AD and 0.9323 on VisA, so frozen features do contain a strong normal/anomaly axis. (2) A few normal images suffice: directions extracted from k=16 and k=4 normal images already beat both text directions on both datasets. (3) Even zero-shot works: with no training images of the class at all, the extracted direction still reaches 0.8498 on MVTec AD and 0.7376 on VisA, far above the chance level of 0.5.

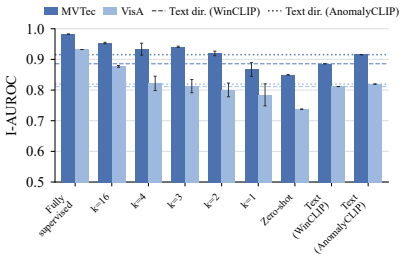

Figure 1: I-AUROC of visual directions under decreasing supervision, against the two text directions.

Experiments show that the direction behind the score does not have to come from text: it can be read directly from images. The visual route is also by far the simpler one. Four normal images are already enough, whereas the text direction is either distilled from hundreds of hand-written, domain-specific prompts or learned from extra training data. In short, whatever text contributes, vision appears able to supply as well.

3.2 Replacement: the visual direction ties the full pipeline

The previous section showed that frozen visual features contain a usable normal/anomaly direction. This opens a direct test of our central claim. If all that text contributes to the pipeline is a direction rather than semantics, then a visually extracted direction, installed in the text pair's place, should cost the pipeline nothing. If text instead carried semantics beyond the direction, the swap should degrade performance. We therefore remove the original text pair, substitute a pair constructed from visual features, keep everything else unchanged, and measure the performance difference.

Given a direction d, an anchor m, and the original differential norm $c = \|t_a - t_n\|$, we orthogonalize and rebuild a unit pair,

$$d \leftarrow \text{normalize}\big(d - (d\cdot\hat{m})\,\hat{m}\big), \qquad \tilde{t}_n = \hat{m}, \qquad \tilde{t}_a = \text{normalize}\big(m + c\,d\big). \tag{2}$$

The anchor m is the mean of the L2-normalized features of the given normal training images. Table 10 summarizes the six variants we compare, which differ in the provenance of the direction, its sign, and the anchor. R1 uses random directions, averaged over 10 seeds. R2 uses the zero-label visual direction. R3 augments R2 with one borrowed bit of text information, the sign of $\cos(d, d_t)$. R4 uses few-shot visual directions with $k \in \{16, 4, 3, 2, 1\}$, averaged over 10 sampling seeds. R5 pairs the zero-shot direction with the text anchor. Each variant is compared against R0 using 1000 image-level bootstrap resamples with percentile 95% confidence intervals.

Figure 2 reports, for each variant, the I-AUROC difference against the original text pipeline on all four pipeline-dataset pairs. Full numbers are given in Table 11 of Appendix D. The baselines are healthy: on MVTec AD and VisA, R0 reaches 0.8835 and 0.8111 with WinCLIP and 0.9155 and 0.8194 with AnomalyCLIP. These closely match the original AnomalyCLIP report (0.915 and 0.821) but deviate from the WinCLIP report (0.918 and 0.781), for which no official implementation is publicly available and which we therefore reproduce from the community version. The few-shot direction with k=16 ties the text pipeline on all four pairs, with an average difference of $+0.0164$ I-AUROC and every 95% confidence interval contained in $[-0.0981, +0.1115]$. The tolerance to smaller k varies across pipeline-dataset pairs. Both VisA pairs hold at k=2, while on MVTec AD the deficit is already visible at k=4, turning significant under AnomalyCLIP. Section 5 extends this sweep of k to five further benchmarks spanning medical, industrial, and logical domains. The zero-shot

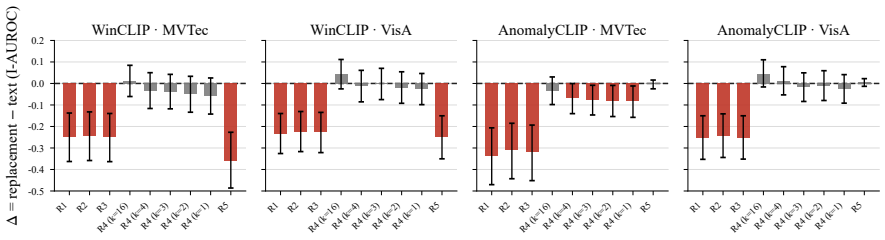

Figure 2: Replacement, end-to-end. I-AUROC difference of each replacement variant against the original text pipeline (Δ=0 dashed), with bootstrap 95% CIs, on all four pipeline-dataset pairs. Bars are red where the interval lies entirely below zero and gray otherwise.

direction fails decisively on all four pairs, and borrowing one bit of text information, the sign of the text direction, does not rescue it (R3). One further variant, R5, tests whether the anchor matters. It keeps the zero-shot direction but borrows the text anchor. Under AnomalyCLIP it now ties the original, the difference essentially zero on both datasets. Under WinCLIP it collapses ($\Delta = -0.3583$ on MVTec AD). A semantic account cannot explain why the identity of the anchor flips the outcome. A geometric account can: in AnomalyCLIP's DPAM feature space the zero-shot direction is already strong, while in WinCLIP's class-token space it is weak. In summary, whether text or vision supplies the better direction depends on the geometry of the feature space, not on the semantics of the words.

Sections 3.1 and 3.2 have shown that vision can replace text: a visual direction, installed in the text pair's place, matches the original pipeline. But the semantic account still has two ways out. First, the visual direction might work only because it happens to coincide with the text direction, so that what really carries the score is still the particular direction the defect words pointed to, and the substitute is merely its shadow. Second, even if a direction is the main carrier, the defect words might contribute semantic information on top of it. The next two subsections close these two exits.

3.3 Erasure: the visual direction is no shadow of the text direction

Section 3.2 left the semantic account with two exits, and this section closes the first: the visual direction might work only because it coincides with the text direction, a shadow of the direction that defect words pointed to. We attack this reading from two angles. Geometrically, we measure how the two directions align. If the visual direction is a shadow, the two should be nearly identical. Causally, we erase the visual axis from the features while scoring with the original text pair. If the text score depends on that axis, removing it should cripple the score.

We intervene on frozen features with dose-controlled projection erasure, $f' = (I - \alpha\, \hat{d}\hat{d}^\top)\, \hat{f}$ for $\alpha \in [0, 1]$, applied both to single directions and to a k-dimensional visual subspace. Besides the visual direction $\hat{d}_v$, we erase two controls: random unit vectors (10 seeds), which exclude the reading that erasing an arbitrary direction hurts, and a wrong-class direction (cyclic class shift), which excludes the reading that erasing any meaningful direction hurts. To show that the erasure itself is effective, we additionally erase the text plane $\mathrm{span}\{t_n, t_a\}$ as a control. If this collapses the score, then a small effect of erasing the visual direction cannot be blamed on a weak intervention.

Geometrically, we first compute the absolute cosine similarity between the two directions. Averaged over classes, it is ≈ 0.1079 (median 0.0847), far below 1, showing that the two directions are nearly orthogonal and that the visual direction is not aligned with the text direction. Figure 3 shows the causal results and full dose-response values are given in Table 12 of Appendix D. We first check that the erasure itself works (Fig. 3a). On all four pipeline–dataset pairs, erasing the text plane $\mathrm{span}\{t_n, t_a\}$ collapses detection to chance (between 0.4798 and 0.5056), while erasing its orthogonal complement leaves the score essentially intact. Both outcomes confirm that the erasure kills the score when it hits the right component and spares it otherwise. The substantive finding concerns the visual direction. Even fully erasing $\hat{d}_v$ barely moves the score, costing between 0.0141 and 0.0547 across the four pairs, so the text score does not depend on the visual axis. This loss is specific: random directions cost nothing, and a wrong-class direction costs at most about one point. Erasing a whole k-dimensional subspace around $\hat{d}_v$, up to k=8, still never brings the score near chance. If the visual direction were a shadow of the text direction, erasing it should cripple the text-based score. It does not.

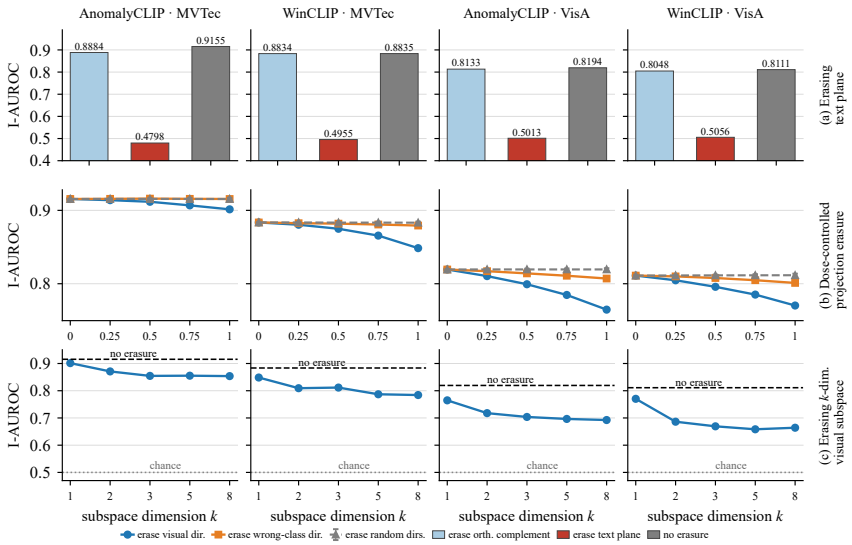

Figure 3: The visual direction is no shadow of the text direction. (a) Erasure of the text plane versus its orthogonal complement. (b) Dose-controlled erasure of the visual direction $\hat{d}_v$, against random and wrong-class controls. (c) Erasing of a k-dimensional visual subspace, up to k=8.

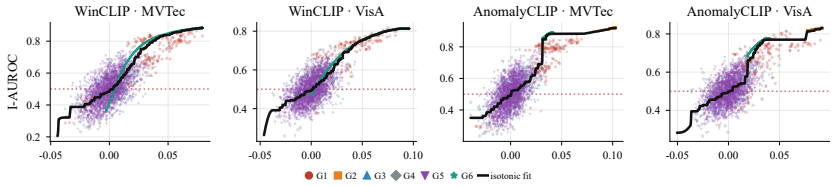

Figure 4: Directional alignment determines prompt performance. I-AUROC of 5k prompts against their alignment with the visual direction, on all four pipeline-dataset pairs. The black curve is the isotonic fit. The dotted line marks chance.

Two measurements close the first exit. The two directions are nearly orthogonal, so the visual direction is no copy of the text direction, and erasing the visual axis leaves the text score essentially untouched, so the score does not depend on that axis either. The shadow reading therefore fails: what carries the score is not the one direction the defect words happened to single out, but any geometrically good direction. One exit remains: even if a direction is the main carrier, defect words might still contribute semantic information on top of it. Section 3.4 closes it by exhaustion.

3.4 Exhaustion: 5k prompts collapse onto one curve, semantics add nothing

The previous section ruled out the first exit. But a second possibility remains: even if a direction is the main carrier, defect words might contribute semantic information on top of it. We falsify this by exhaustion: across a large and deliberately diverse pool of prompts, we measure how much of prompt performance directional alignment alone explains, and how much is left for semantics.

Each prompt p contributes a triple (x, y, z): directional alignment $x = \cos(d_p, \hat{d}_v)$, measured I-AUROC y, and a semantic flag $z = \mathbb{1}[\text{anomalous side names a defect}]$. The pool of 5000 prompts spans six provenances. G1 contains manual templates, G2 learned contexts from AnomalyCLIP's prompt learner blended with random initializations, G3 defect-naming prompts in 20 languages, G4 empty and gibberish strings, G5 random token sequences, and G6 language-free vector interpolations $d = \cos\theta\, d_t + \sin\theta\, u$ with $u \perp d_t$ random and $\theta \in [0, \pi/2]$. We then ask whether geometry alone predicts performance, reporting bootstrap 95% CIs on ΔR^2 (1000 resamples):

$$\hat{g} = \arg\min_{g\ \text{isotone}} \sum_i \big(y_i - g(x_i)\big)^2, \qquad \Delta R^2 = R^2_{\text{geo}+z} - R^2_{\text{geo}}. \tag{3}$$

Table 2: Geometry exhausts performance. Directional alignment alone explains at least 87.36% of the variance across 5k prompts. The semantic adds nothing.

Pipeline–dataset	pool	isotonic R^2	ΔR^2 (semantic)	95% CI
WinCLIP·MVTec	4900	0.8945	0.0009	[0.0003, 0.0018]
WinCLIP·VisA	4900	0.9103	0.0001	[0.0000, 0.0004]
AnomalyCLIP·MVTec	5000	0.9162	0.0008	[0.0003, 0.0016]
AnomalyCLIP·VisA	5000	0.8736	0.0012	[0.0004, 0.0022]

Fig. 4 and Table 2 answer the question on all four pipeline–dataset pairs. Isotonic R^2 reaches 0.8736–0.9162: alignment alone explains at least 87.36% of the variance. The semantic increment is $\Delta R^2 \leq 0.0012$, and even the widest confidence interval ends at 0.0022: once alignment is known, whether the prompt carries defect semantics adds nothing. A cross-fitted control, in which the reference direction and the I-AUROC are computed on disjoint halves of each test set, leaves all values essentially unchanged (Appendix B). The group structure corroborates this verdict (group means in Table 13, Appendix D). G3 names defects in 20 languages and scores at chance (≈ 0.51), indistinguishable from G4's empty strings and G5's random tokens; if words carried defect semantics, multilingual prompts could not fall to chance. G6 contains no words at all, yet beats the hand-written templates of G1 on every pipeline–dataset pair, simply by pointing in the right direction. And G2, the learned prompts AnomalyCLIP actually deploys, tops the pool not because learning discovered defect semantics, but because it discovered alignment: its contexts attain the pool's largest $\cos(d_p, \hat{d}_v)$.

Two measurements close the second exit. Alignment alone determines the performance of 5k diverse prompts, and semantics adds nothing on top of it. In sum, the analyses of Sec. 3.1–3.4 converge on a single conclusion. The text branch contributes only geometry, not semantics: a single direction from normal to anomalous in feature space and nothing else. Vision alone supplies a functionally equivalent one, to the same effect.

4 Prediction: One Scalar Forecasts Prompts It Has Never Seen

The analyses so far have shown that text contributes only geometry, not semantics, and that vision supplies a functionally equivalent one, to the same effect. If this account holds, it should not only explain the performance of measured prompts, but also forecast the performance of each individual new prompt it has never seen, from that geometric quantity alone, without running the pipeline.

The protocol has three steps. First, we predict: from a single geometric quantity, we forecast the I-AUROC of each of 10k entirely new prompts, without running the pipeline. Second, we measure: we run the complete unmodified pipeline on each of 10k prompts, obtaining one measured I-AUROC per prompt. Third, we compare the 10k forecasts against their corresponding measurements. The geometric quantity is a directional signal-to-noise ratio, the component of the prompt direction along the visual axis over its orthogonal component,

$$\mathrm{SNR}(d_p) \;=\; \frac{x}{\sqrt{1-x^2}}, \qquad x = \cos(d_p, \hat{d}_v). \tag{4}$$

SNR is strictly increasing in x, so the monotone relationship of Sec. 3.4 carries over without further argument: performance is monotone in SNR. We fit an isotonic calibration $\hat{g}$ on the exhaustion pool and forecast each new prompt as $\hat{y} = \hat{g}(\mathrm{SNR})$. The 10k evaluation prompts are disjoint from the calibration pool by construction, spanning defect prompts in new languages, random-token sequences from new seeds, and interpolation angles interleaved between the pool's angles.

The forecasts land with rank correlation $\rho = 0.8440$ to 0.8643, binned RMSE ≤ 0.0120, and calibration slope within 0.0399 of unity (Table 3 and Fig. 5). The same cross-fitted control leaves the rank correlation at 0.8100 to 0.8670 (Appendix B), so the forecasts do not exploit the overlap between the direction's construction set and the evaluation set. All p-values are so small that they fall below machine precision and are therefore reported as zero.

This completes the prediction step. Even for an entirely new prompt, a single geometric quantity directly predicts its performance without running the CLIP pipeline. The geometric account is therefore not merely a description of measured prompts, but a predictive rule for how the pipeline

Table 3: One scalar predicts unseen prompts. We fit an isotonic calibration on a separate set of prompts, then use directional SNR to predict the I-AUROC of each of the 10k unseen prompts.

Pipeline–dataset	Spearman ρ	p-value	binned RMSE	pointwise RMSE	calib. slope
WinCLIP·MVTec	0.8605	≈ 0	0.0120	0.0674	0.9601
WinCLIP·VisA	0.8440	≈ 0	0.0071	0.0470	0.9773
AnomalyCLIP·MVTec	0.8643	≈ 0	0.0111	0.0674	0.9683
AnomalyCLIP·VisA	0.8519	≈ 0	0.0075	0.0544	0.9645

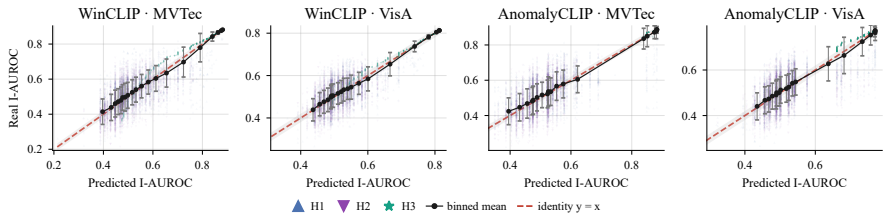

Figure 5: Forecasts track measurements. Each faint point gives the predicted and measured I-AUROC for one of 10k unseen prompts. Black dots show binned means over 20 bins with standard-deviation bars. The dashed diagonal marks perfect agreement. The calibration is fitted on the disjoint prompt pool of Sec. 3.4 and never refitted.

will behave on new ones. As a byproduct, prompt design is no longer limited to trial and error: the quality of a candidate prompt can be read off from the curve before it is ever run.

5 Boundary: How Far Geometry Goes, and Where It Stops

We now ask how far the geometric account travels, and probe its boundary along two axes. The first axis is breadth. Section 3.2 established replacement on MVTec AD and VisA. Here we extend the WinCLIP replacement test to five additional benchmarks spanning medical imaging (Br35H (Hamada, 2020), BrainMRI (Salehi et al., 2021), and HeadCT (Salehi et al., 2021)), industrial parts (MPDD (Jezek et al., 2021)), and logical constraints (MVTec LOCO (Bergmann et al., 2022)). We sweep $k \in \{1, 2, 4, 16\}$ normal images and report the smallest adequate k on each dataset. The second axis is depth. If text contributed semantics, its advantage should grow as tasks require finer semantic distinctions. Holding the visual side fixed, we vary only the task across six levels of semantic demand, from standard detection, through logical and structural anomalies and coarse and fine defect specification, to open-vocabulary compound intent.

Table 4 maps the first axis. Under the adequacy rule, the visual direction replaces text on all five additional benchmarks. Br35H and MVTec LOCO need only one normal image, BrainMRI and MPDD need two, and HeadCT needs four. Together with Sec. 3.2, this gives replacement on all seven WinCLIP benchmarks tested. The smallest adequate k varies from one to sixteen normal images. Replacement therefore generalizes across domains, but no single k fits all datasets.

Table 5 maps the second axis. The visual direction remains on par with or better than text through fine defect specification. The most diagnostic cell is (c′). CLIP's text embeddings of `broken_large` and `broken_small` nearly coincide, so the text pair has little signal for distinguishing the two sizes, while the visual direction wins by 0.0549. Only cell (d) favors text, by 0.0712, but both sides are near zero on intent compliance. This is a victory in name only. Increasing semantic demand therefore does not reveal a hidden semantic contribution. The boundary is not where semantic demand begins. It is where CLIP's text encoder can no longer express the requested distinction, and neither route can perform the task.

6 Implications

The verdict has three implications. They concern how prompt gains should be interpreted, how detectors should be designed, and what representation convergence does and does not imply.

Table 4: A few normal images replace text across datasets. Each row reports one benchmark, giving the smallest adequate number of normal images k, the I-AUROC difference Δ (visual $-$ text), and its 95% bootstrap CI.

Dataset	Domain	min. k	Δ [95% CI]
Br35H	medical	1	$+0.0097$ [$+0.0049$, $+0.0149$]
BrainMRI	medical	2	-0.0089 [-0.0555, $+0.0322$]
HeadCT	medical	4	$+0.0187$ [-0.0346, $+0.0802$]
MVTec LOCO	logical	1	$+0.0299$ [-0.0712, $+0.1276$]
MPDD	industrial	2	$+0.0714$ [-0.1521, $+0.2972$]

Table 5: Rising semantic demand does not rescue text. Each row reports one task of increasing semantic demand, evaluated with WinCLIP and scored by I-AUROC for tasks (a) to (c$'$) and by intent compliance for task (d).

Task	Semantic demand	Text	Visual dir.	Δ (text $-$ visual)
(a) standard detection	lowest	0.8473	0.8737	-0.0264
(b) logical anomaly	low	0.5368	0.6467	-0.1099
(b$'$) structural anomaly	mid-low	0.5642	0.5644	-0.0002
(c) coarse defect spec.	mid	0.7287	0.7161	$+0.0126$
(c$'$) fine defect spec.	mid-high	0.7532	0.8081	-0.0549
(d) open-vocab. intent	highest	0.1152	0.0440	$+0.0712$

Prompt learning improves alignment, not semantics. In the exhaustion pool, hand-written and learned prompts alike fall on the same alignment curve, and the semantic adds nothing (Sec. 3.4). Gains from learned or ensembled prompts in ZSAD should therefore first be interpreted as gains in alignment. Prompt sensitivity is sensitivity of the induced direction, and prompt quality can be forecast before the pipeline is run (Sec. 4).

Text branches are probe generators. Once the reduction is in place, the practical role of text is to place a probe direction in feature space. Language is not the only possible source of that probe. A few normal images can supply a functionally equivalent direction while keeping the frozen backbone and scoring head unchanged (Secs. 3.2 and 5). The design question therefore shifts from which words describe defects to which feature-space statistics expose a strong normal/anomaly direction. For these pipelines, prompt engineering becomes direction engineering.

Representation convergence can be functional rather than directional. The Platonic Representation Hypothesis predicts that strong models converge toward shared representations (Huh et al., 2024). Our results caution against reading this as convergence of individual decision directions. The text and visual directions are nearly orthogonal, yet either can drive the detector in the same scoring head (Secs. 3.2 and 3.3). Convergence, where it matters for this task, is functional, not directional: two modalities can implement the same decision variable while pointing in different directions.

7 Conclusion

Progress in zero-shot anomaly detection has commonly been attributed to the defect semantics of text prompts. We show instead that the text branch contributes only geometry, not semantics. Text prompts carry a single direction from normal to anomalous, and vision alone supplies a functionally equivalent one, to the same effect. We prove this in three steps. First, the text branch reduces algebraically to a projection onto one direction. Second, a direction read from a few normal images replaces the text pair across pipelines and benchmarks, and directional alignment explains 5k diverse prompts while semantics add nothing. Third, directional SNR forecasts the performance of 10k entirely unseen prompts. Two implications follow. Prompt learning learns geometry, not semantics, and the text branch functions as a probe generator. Representation convergence can be functional rather than directional, because two modalities can implement the same decision without sharing a direction.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of this paper. Specifically, we have not used generative AI tools for any tasks with required disclosure, including generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, or interpreting results. Additionally, we used generative AI tools for editing this research paper to improve readability. We have not used generative AI tools for any other tasks with recommended disclosure. We have reviewed all AI-assisted work. Every AI-polished sentence was checked against the original author-written draft by the authors to confirm that no claim, number, or citation was altered. All research ideas, analyses, experiments, and claims were produced and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release the code, configurations, and evaluation scripts after review. All experiments build on the public AnomalyCLIP and WinCLIP pipelines (CLIP ViT-L-14@336px) and use only public datasets. Both backbones are frozen, so scores are deterministic given a direction; the only stochastic steps are sampling procedures with fixed seeds: three seeds $\{0, 1, 2\}$ for the few-shot ladder, ten seeds $\{0, \ldots, 9\}$ for the score-averaged few-shot replacement rows, and ten random-vector seeds $\{10, \ldots, 19\}$ for random-direction controls. All confidence intervals are paired bootstrap percentile intervals with 1000 resamples. Complete per-experiment protocols are specified in the released configurations.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2017.

Yuhu Bai, Jiangning Zhang, Yunkang Cao, Guangyuan Lu, Qingdong He, Xiangtai Li, and Guanzhong Tian. Bridge feature matching and cross-modal alignment with mutual-filtering for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3584–3593, 2025.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 66044–66063, 2023.

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 (CVPR)*, pp. 9584–9592. IEEE, 2019.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization. *International Journal of Computer Vision*, 130(4):947–969, 2022.

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision (ECCV)*, pp. 55–72. Springer, 2024.

Qingqing Fang, Wenxi Lv, and Qinliang Su. Af-clip: Zero-shot anomaly detection via anomaly-focused clip adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, pp. 4846–4855, 2025.

Bin-Bin Gao, Yue Zhou, Jiangtao Yan, Yuezhi Cai, Weixi Zhang, Meng Wang, Jun Liu, Yong Liu, Lei Wang, and Chengjie Wang. Adaptclip: Adapting clip for universal visual anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4095–4103, 2026.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9574–9586, 2021.

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

Ahmed Hamada. Br35h :: Brain tumor detection 2020, 2020. URL `https://dx.doi.org/10.21227/tbkk-q937`.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4129–4138, 2019.

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. Visualad: Language-free zero-shot anomaly detection via vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21346–21356, June 2026.

Ming Hu, Yongsheng Huo, Mingyu Dou, Jianfu Yin, Peng Zhao, Yao Wang, Cong Hu, Bingliang Hu, and Quan Wang. Fb-clip: Fine-grained zero-shot anomaly detection with foreground-background disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35659–35669, June 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 20617–20642. PMLR, 21–27 Jul 2024.

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 (CVPR)*, pp. 19606–19616. IEEE, 2023.

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

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2668–2677. PMLR, 2018.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 2172–2197, 2024.

Xurui Li, Feng Xue, and Yu Zhou. Musc-v2: Zero-shot multimodal industrial anomaly classification and segmentation with mutual scoring of unlabeled samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Shih-Chih Lin, Jia-Xian Jian, YunTung Chu, Wei-Chieh Sun, and Fang-Yi Lin. Bridging semantic and structural manifolds: Zero-shot single-temporal anomaly detection in remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1272–1280, June 2026.

Wenxin Ma, Xu Zhang, Qingsong Yao, Fenghe Tang, Chenxu Wu, Yingtai Li, Rui Yan, Zihang Jiang, and S Kevin Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4744–4754. IEEE, 2025.

Yunfeng Ma, Min Liu, Shuai Jiang, Jingyu Zhou, Yuan Bian, Xueping Wang, and Yaonan Wang. Zuma: Training-free zero-shot unified multimodal anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15645–15655. IEEE, 2023.

Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30398–30408. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256, 2020.

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

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2300–2344, 2022.

Gaowei Zhang and Lihe Zhang. Dlvp-clip: Enhancing fine-grained zero-shot anomaly detection via dynamic local visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35524–35533, June 2026.

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 (CVPR)*, pp. 16795–16804. IEEE, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 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 (ICLR)*, volume 2024, pp. 49705–49737, 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 *European Conference on Computer Vision (ECCV)*, pp. 392–408. Springer, 2022.

A Derivation of the Direction Reduction

Let f be the image feature and $\hat{f} = f/\|f\|$ its L2-normalization, t_n, t_a the normal and anomalous text embeddings, and τ the CLIP temperature. For WinCLIP, t_n, t_a are the means over its normal and anomalous prompt templates. For AnomalyCLIP, they are the two outputs of its prompt learner. Both pipelines score an image by the anomalous channel of a two-class softmax over cosine similarities:

$$S(f) \;=\; \frac{\exp(\hat{f}\cdot\hat{t}_a/\tau)}{\exp(\hat{f}\cdot\hat{t}_n/\tau)+\exp(\hat{f}\cdot\hat{t}_a/\tau)}. \tag{5}$$

Dividing numerator and denominator by $\exp(\hat{f}\cdot\hat{t}_a/\tau)$ shows that S depends on the two embeddings only through the difference of the two similarities:

$$S(f) \;=\; \frac{1}{1+\exp\big(-(\hat{f}\cdot\hat{t}_a-\hat{f}\cdot\hat{t}_n)/\tau\big)}. \tag{6}$$

By linearity of the dot product, this difference is a single projection. Define the text difference direction

$$d \;=\; \frac{\hat{t}_a-\hat{t}_n}{\big\|\hat{t}_a-\hat{t}_n\big\|}, \tag{7}$$

then $\hat{f}\cdot\hat{t}_a-\hat{f}\cdot\hat{t}_n=\hat{f}\cdot(\hat{t}_a-\hat{t}_n)=\big\|\hat{t}_a-\hat{t}_n\big\|\;(\hat{f}\cdot d)$, where $\big\|\hat{t}_a-\hat{t}_n\big\|$ is a fixed positive scalar once the text embeddings are fixed. Equation 6 becomes

$$S(f) \;=\; \sigma\!\left(\frac{\big\|\hat{t}_a-\hat{t}_n\big\|}{\tau}\;\hat{f}\cdot d\right), \qquad \sigma(u)=\frac{1}{1+e^{-u}}. \tag{8}$$

Since $\sigma'(u)=\sigma(u)(1-\sigma(u))>0$ and $\big\|\hat{t}_a-\hat{t}_n\big\|/\tau>0$, $S(f)$ is a strictly increasing function of $\hat{f}\cdot d$. Strictly increasing transformations preserve the ordering of any set of scores, so S and $\hat{f}\cdot d$ induce identical image rankings. I-AUROC depends on scores only through their ordering, since it equals the probability that a random anomalous image outranks a random normal one, and the same holds for other rank-based criteria. Every such metric is therefore exactly invariant to replacing the whole text apparatus by the single direction d. Finally, ensembling more templates changes only the mean embeddings t_n, t_a, hence only the values of d and of the scale, not the form of equation 8, so the reduction still holds.

B Cross-Fitted Control for the Analyses of Secs. 3.4 and 4

The visual reference direction $\hat{d}_v$ is constructed with the help of the unlabeled test batch (Sec. 3.1), a batch setting shared by recent zero-shot detectors (Li et al., 2026), which raises a transduction concern for Secs. 3.4 and 4: the direction is defined by extreme points of the test set, and I-AUROC is measured on the same test set, so the alignment $x=\cos(d_p,\hat{d}_v)$ and the performance y could be spuriously coupled. We cut this coupling by cross-fitting. For each class, the test images are split into two stratified random halves A and B. The direction is constructed using half A only: the whitening statistics still come from the k=16 normal training images, and the top-1% selection is restricted to A. The I-AUROC of every prompt is then measured on half B only. The predictor x and the outcome y therefore share no test image.

Tables 6 and 7 report the cross-fitted results against the original values. The isotonic R^2 moves from 0.8736–0.9162 to 0.8710–0.9246, and the semantic increment remains negligible ($\Delta R^2 \le 0.0005$, widest CI end 0.0011). The prediction rank correlation on the 10k unseen prompts moves from 0.8440–0.8643 to 0.8100–0.8670. Both analyses still hold under cross-fitting, ruling out the concern that the high isotonic R^2 and high prediction rank correlation arise from constructing the visual direction and measuring I-AUROC on the same test images.

C Related Work

Zero-shot anomaly detection. Zero-shot anomaly detection aims to score and localize defects in target categories unseen during training. Existing pipelines predominantly build on CLIP, and they

Table 6: Exhaustion analysis under cross-fitting. The reference direction is built on one stratified half of each class's test set, and I-AUROC is measured on the other half. Original values from Table 2 are repeated for comparison.

Pipeline–dataset	cross-fit R^2	orig. R^2	cross-fit ΔR^2	95% CI	orig. ΔR^2
WinCLIP·MVTec	0.9049	0.8945	0.0001	[0.0000, 0.0004]	0.0009
WinCLIP·VisA	0.8710	0.9103	0.0000	[0.0000, 0.0003]	0.0001
AnomalyCLIP·MVTec	0.9246	0.9162	0.0005	[0.0002, 0.0011]	0.0008
AnomalyCLIP·VisA	0.8967	0.8736	0.0001	[0.0000, 0.0005]	0.0012

Table 7: Prediction analysis under cross-fitting. Using the same split as Table 6, we report Spearman ρ between directional SNR and measured I-AUROC, together with binned RMSE of the forecasts on 10k unseen prompts. The calibration is fitted on the disjoint prompt pool of Sec. 3.4 and is never refitted. Original values from Table 3 are reported for comparison.

Pipeline–dataset	cross-fit ρ	orig. ρ	cross-fit binned RMSE	orig. binned RMSE
WinCLIP·MVTec	0.8498	0.8605	0.0131	0.0120
WinCLIP·VisA	0.8100	0.8440	0.0097	0.0071
AnomalyCLIP·MVTec	0.8670	0.8643	0.0092	0.0111
AnomalyCLIP·VisA	0.8327	0.8519	0.0052	0.0075

take it for granted that text prompts carry the defect semantics, telling the model what is normal and what counts as a defect. WinCLIP (Jeong et al., 2023) introduces prompt-ensemble scoring, AnomalyCLIP (Zhou et al., 2024) and AdaCLIP (Cao et al., 2024) replace hand-written prompts with learned ones, and the same recipe keeps being refined with anomaly-aware text anchors (Ma et al., 2025), visual adapters (Fang et al., 2025; Gao et al., 2026), a learned distribution over the prompt space (Qu et al., 2025), or learnable prompts equipped with positional encodings that explicitly model word-order semantics (Ma et al., 2026). A parallel line removes text entirely: MuSc (Li et al., 2024; 2026) scores unlabeled test images against each other, and VisualAD (Hou et al., 2026) learns normal and anomalous tokens directly inside a frozen ViT. FiSeCLIP (Bai et al., 2025) sits in between, keeping text prompts but adding mutual references among images in the same batch. These language-free and hybrid detectors match or exceed the text-based pipelines on the same benchmarks. We resolve this tension not with another detector but by looking directly at what the text branch carries. We show that it carries geometry, not semantics: the prompts reduce to a single direction from normal to anomalous in feature space, and vision alone supplies a functionally equivalent one, to the same effect.

Prompting and prompt learning. Prompting adapts a frozen pretrained model by editing its input text, and prompt learning optimizes that text directly. Ensembling templates (Radford et al., 2021), generating descriptions (Menon & Vondrick, 2023; Pratt et al., 2023), and learning context tokens (Zhou et al., 2022b;a) all bank on the content of the prompt. The ZSAD literature inherits this assumption, investing in LLM-written defect descriptions (Gu et al., 2024) and learned prompt distributions (Qu et al., 2025). For text classification, Webson & Pavlick (2022) showed that models can ignore prompt meaning. We deliver the analogous verdict for anomaly detection.

Linear structure and its causal status. A mature toolbox asks whether a variable is encoded along a direction in representation space: linear probes (Alain & Bengio, 2017; Hewitt & Manning, 2019), concept vectors (Kim et al., 2018), and Mahalanobis scoring (Lee et al., 2018) read such directions out, concept erasure (Ravfogel et al., 2020; Belrose et al., 2023) deletes them, and causal abstraction (Geiger et al., 2021) characterizes when a direction genuinely plays the role of a variable. We use the same tools to ask a different question: not what a representation contains, but whether the directions supplied by text and by images do the same job. This turns tools built for diagnosing representations into an identity test across modalities.

D ADDITIONAL TABLES

Table 8: The reduction holds class by class. For each class we report the Spearman ρ between the original pipeline score and the reduced linear score $\hat{f} \cdot d_t$, together with the original and reduced I-AUROC. Δ = original − reduced. Raw data underlying Table 1.

	WinCLIP				**AnomalyCLIP**				
Class	ρ	orig.	reduced	Δ	Class	ρ	orig.	reduced	Δ
MVTec AD					*MVTec AD*				
bottle	0.9983	0.8889	0.8817	+0.0072	bottle	1.0000	0.8865	0.8865	+0.0000
cable	0.9990	0.7931	0.8025	−0.0094	cable	1.0000	0.7028	0.7028	+0.0000
capsule	0.9951	0.8604	0.8720	−0.0116	capsule	1.0000	0.8951	0.8951	+0.0000
carpet	1.0000	0.9932	0.9932	+0.0000	carpet	1.0000	0.9996	0.9996	+0.0000
grid	0.9986	0.8680	0.8647	+0.0033	grid	1.0000	0.9783	0.9783	+0.0000
hazelnut	0.9985	0.8007	0.8100	−0.0093	hazelnut	1.0000	0.9718	0.9718	+0.0000
leather	0.9993	0.9983	0.9986	−0.0003	leather	1.0000	0.9983	0.9983	+0.0000
metal_nut	0.9944	0.9384	0.9418	−0.0034	metal_nut	1.0000	0.9238	0.9238	+0.0000
pill	0.9969	0.8500	0.8502	−0.0002	pill	1.0000	0.8110	0.8110	+0.0000
screw	0.9998	0.9031	0.9033	−0.0002	screw	1.0000	0.8211	0.8211	+0.0000
tile	0.9976	0.9794	0.9903	−0.0109	tile	1.0000	1.0000	1.0000	+0.0000
toothbrush	0.9922	0.7639	0.7583	+0.0056	toothbrush	1.0000	0.8528	0.8528	+0.0000
transistor	0.9991	0.8267	0.8304	−0.0037	transistor	1.0000	0.9387	0.9387	+0.0000
wood	1.0000	0.9886	0.9886	+0.0000	wood	1.0000	0.9693	0.9693	+0.0000
zipper	0.9989	0.7999	0.8007	−0.0008	zipper	1.0000	0.9840	0.9840	+0.0000
VisA					*VisA*				
candle	0.9952	0.9474	0.9369	+0.0105	candle	1.0000	0.8458	0.8458	+0.0000
capsules	0.9980	0.8032	0.8083	−0.0051	capsules	1.0000	0.8033	0.8033	+0.0000
cashew	0.9997	0.8806	0.8820	−0.0014	cashew	1.0000	0.7998	0.7998	+0.0000
chewinggum	0.9998	0.9660	0.9658	+0.0002	chewinggum	1.0000	0.9628	0.9628	+0.0000
fryum	0.9965	0.9268	0.9254	+0.0014	fryum	1.0000	0.9436	0.9436	+0.0000
macaroni1	0.9993	0.7847	0.7840	+0.0007	macaroni1	1.0000	0.8604	0.8604	+0.0000
macaroni2	0.9979	0.6658	0.6671	−0.0013	macaroni2	1.0000	0.7165	0.7165	+0.0000
pcb1	0.9993	0.8110	0.8066	+0.0044	pcb1	1.0000	0.7393	0.7393	+0.0000
pcb2	0.9994	0.6748	0.6750	−0.0002	pcb2	1.0000	0.6633	0.6633	+0.0000
pcb3	0.9995	0.6308	0.6329	−0.0021	pcb3	1.0000	0.6753	0.6753	+0.0000
pcb4	0.9984	0.7267	0.7361	−0.0094	pcb4	1.0000	0.9108	0.9108	+0.0000
pipe_fryum	0.9992	0.9158	0.9174	−0.0016	pipe_fryum	1.0000	0.9120	0.9120	+0.0000

Table 9: Existence ladder, full numbers. Macro-averaged I-AUROC per tier, underlying Fig. 1. Error bars give ± 1 standard deviation across the 3 few-shot sampling seeds. The fully supervised, zero-shot, and text tiers are deterministic configurations with no sampling variability. The text tiers are scored by the reduced single-direction form $\hat{f} \cdot d_t$ rather than the original pipeline. For AnomalyCLIP the reduction is exact, so these values coincide with Table 8. For WinCLIP the ensemble head leaves a small reduction residue, so the values here differ slightly from the original-head values in Table 8.

Direction	MVTec AD	VisA
Fully supervised	0.9824	0.9323
k=16	0.9534 ± 0.0025	0.8772 ± 0.0030
k=4	0.9333 ± 0.0195	0.8218 ± 0.0235
k=3	0.9414 ± 0.0018	0.8129 ± 0.0214
k=2	0.9195 ± 0.0074	0.8006 ± 0.0224
k=1	0.8671 ± 0.0223	0.7843 ± 0.0359
Zero-shot	0.8498	0.7376
Text (WinCLIP)	0.8858	0.8115
Text (AnomalyCLIP)	0.9155	0.8194

Table 10: Replacement variants. Each row gives one replacement variant: the provenance of the direction, its sign, and the anchor.

Variant	Direction	Sign	Anchor
R0 (original text pair)	Text	Text	Text
R1 (random, 10 seeds)	Random	Random	Visual
R2 (zero-shot)	Visual	Visual	Visual
R3 (zero-shot + text sign)	Visual	Text	Visual
R4 (few-shot, $k \in \{16, 4, 3, 2, 1\}$)	Visual	Visual	Visual
R5 (zero-shot)	Visual	Visual	Text

Table 11: Replacement verdicts, full numbers. For each variant, the difference in I-AUROC against the original text pipeline R0 (Δ = variant $-$ R0), with 95% bootstrap CI. R0 is reported in absolute I-AUROC. R1 averages 10 random seeds. R4 averages 10 few-shot seeds (score-averaged). Raw data underlying Fig. 2.

WinCLIP

Variant	MVTec	VisA
R0 (original text pair)	0.8835	0.8111
R1 (random)	-0.2495 [-0.3629, -0.1375]	-0.2329 [-0.3258, -0.1398]
R2 (zero-shot)	-0.2430 [-0.3583, -0.1322]	-0.2233 [-0.3168, -0.1305]
R3 (zero-shot + text sign)	-0.2502 [-0.3635, -0.1396]	-0.2278 [-0.3215, -0.1345]
R4 (few-shot, k=16)	$+0.0104$ [-0.0607, $+0.0844$]	$+0.0423$ [-0.0252, $+0.1115$]
R4 (few-shot, k=4)	-0.0327 [-0.1164, $+0.0501$]	-0.0124 [-0.0855, $+0.0610$]
R4 (few-shot, k=3)	-0.0378 [-0.1179, $+0.0423$]	-0.0019 [-0.0750, $+0.0703$]
R4 (few-shot, k=2)	-0.0497 [-0.1337, $+0.0333$]	-0.0193 [-0.0922, $+0.0543$]
R4 (few-shot, k=1)	-0.0583 [-0.1423, $+0.0256$]	-0.0262 [-0.0985, $+0.0464$]
R5 (zero-shot + text anchor)	-0.3583 [-0.4862, -0.2269]	-0.2503 [-0.3504, -0.1504]

AnomalyCLIP

Variant	MVTec	VisA
R0 (original text pair)	0.9155	0.8194
R1 (random)	-0.3373 [-0.4705, -0.2064]	-0.2528 [-0.3528, -0.1503]
R2 (zero-shot)	-0.3107 [-0.4430, -0.1851]	-0.2435 [-0.3441, -0.1413]
R3 (zero-shot + text sign)	-0.3198 [-0.4518, -0.1935]	-0.2521 [-0.3517, -0.1503]
R4 (few-shot, k=16)	-0.0328 [-0.0981, $+0.0303$]	$+0.0455$ [-0.0162, $+0.1101$]
R4 (few-shot, k=4)	-0.0692 [-0.1403, -0.0004]	$+0.0113$ [-0.0531, $+0.0780$]
R4 (few-shot, k=3)	-0.0763 [-0.1465, -0.0085]	-0.0183 [-0.0837, $+0.0494$]
R4 (few-shot, k=2)	-0.0801 [-0.1538, -0.0091]	-0.0101 [-0.0792, $+0.0593$]
R4 (few-shot, k=1)	-0.0835 [-0.1576, -0.0117]	-0.0252 [-0.0915, $+0.0411$]
R5 (zero-shot + text anchor)	-0.0043 [-0.0249, $+0.0159$]	$+0.0047$ [-0.0133, $+0.0224$]

Table 12: Erasure, full numbers. I-AUROC on all four pipeline-dataset pairs. (a) Full-dose erasure of the text plane versus its orthogonal complement, against no erasure. (b) Dose-controlled projection erasure $f' = (I - \alpha \hat{d}\hat{d}^\top)\hat{f}$ of the visual direction $\hat{d}_v$, a wrong-class direction, and random directions (mean over 10 seeds). (c) Erasure of a k-dimensional visual subspace. Raw data underlying Fig. 3.

(a) Erasing the text plane

	W·MVTec	W·VisA	A·MVTec	A·VisA
no erasure	0.8835	0.8111	0.9155	0.8194
erase text plane	0.4955	0.5056	0.4798	0.5013
erase complement	0.8834	0.8048	0.8884	0.8133

(b) Dose-controlled projection erasure

	W·MVTec	W·VisA	A·MVTec	A·VisA
erase visual direction $\hat{d}_v$				
α=0.00	0.8835	0.8111	0.9155	0.8194
α=0.25	0.8804	0.8047	0.9139	0.8104
α=0.50	0.8749	0.7958	0.9117	0.7993
α=0.75	0.8656	0.7852	0.9068	0.7847
α=1.00	0.8485	0.7703	0.9014	0.7647
erase wrong-class direction				
α=0.00	0.8835	0.8111	0.9155	0.8194
α=0.25	0.8825	0.8100	0.9157	0.8171
α=0.50	0.8819	0.8078	0.9160	0.8141
α=0.75	0.8807	0.8048	0.9157	0.8109
α=1.00	0.8793	0.8011	0.9155	0.8071
erase random directions (10 seeds)				
α=0.00	0.8835	0.8111	0.9155	0.8194
α=0.25	0.8835	0.8113	0.9156	0.8195
α=0.50	0.8833	0.8114	0.9156	0.8195
α=0.75	0.8833	0.8115	0.9155	0.8195
α=1.00	0.8832	0.8116	0.9155	0.8195

(c) Erasing a k-dimensional visual subspace

	W·MVTec	W·VisA	A·MVTec	A·VisA
k=1	0.8485	0.7703	0.9014	0.7647
k=2	0.8093	0.6862	0.8709	0.7177
k=3	0.8116	0.6693	0.8545	0.7037
k=5	0.7870	0.6584	0.8552	0.6964
k=8	0.7842	0.6641	0.8537	0.6924

Table 13: Semantics add nothing in the 5k prompt pool. Mean I-AUROC per provenance group, underlying Fig. 4. Multilingual defect-naming prompts (G3) sit at chance, indistinguishable from empty strings (G4) and random tokens (G5). Language-free interpolations (G6) reach the top of the pool. G2 is defined only for AnomalyCLIP.

Group mean (I-AUROC)	W·MVTec	W·VisA	A·MVTec	A·VisA
G1 manual templates	0.6976	0.6592	0.6818	0.6571
G2 learned contexts	—	—	0.9153	0.8222
G3 multilingual (defect words)	0.5162	0.5124	0.5170	0.5068
G4 empty / gibberish	0.4851	0.4949	0.4901	0.4987
G5 random tokens	0.4996	0.4986	0.4999	0.4992
G6 vector interpolation (language-free)	0.7884	0.7408	0.8772	0.7574