NORMAL-QUOTIENT RESIDUAL READOUT FOR STRUCTURED INSPECTION QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured industrial inspection QA verifies which candidate defect hypothesis is supported by an image, rather than generating an unrestricted report. These candidates can come from inspection sheets, taxonomies, retrieval modules, or human operators. Large multimodal language models provide a flexible interface, but their cost, latency, and deployment constraints motivate lighter evidence-grounded alternatives. We ask whether a frozen CLIP backbone can support this structured setting with a better visual readout. Our key observation is that comparing a query image with a matched normal reference exposes residual defect evidence. This evidence is not just scalar abnormality: defect families form coherent residual directions. However, raw residuals also vary with the chosen normal reference. We therefore introduce a normal-quotient residual readout that removes empirical normal-variation directions and explains the remaining evidence with shared-private defect subspaces. Candidate hypotheses are scored by how well routed subspaces explain the residual evidence, and the same evidence is reused for patch-wise grounding. On AnomalyCoT and MMAD, our readout improves over vanilla CLIP and reaches a competitive regime on structured localization and description, with much lower latency and memory than 4B–8B MLLMs.

1 INTRODUCTION

Industrial visual inspection usually starts with anomaly detectors that output scores, heatmaps, or masks. These outputs are useful, but inspectors often need a more direct answer: what defect is present, where is the evidence, and which candidate explanation is supported by the image? This motivates *structured inspection question answering* (inspection QA), where the model verifies candidate defect hypotheses for reporting and decision-making. In real inspection workflows, these candidates are usually not arbitrary free-form answers. They come from inspection sheets, defect taxonomies, retrieval modules, or human operators. This setting narrows the output space, but it is not a fixed-label classification. The option sheet changes for each query; many descriptions may be unseen during training, and hard distractors differ only by small local visual cues. If no candidate is supported, the interface should allow a null or escalation option, rather than a defect label.

Large multimodal language models are flexible, but they are costly for many on-premises inspection settings. A lightweight CLIP-based readout is attractive, but standard CLIP scoring is not enough. Global image-text similarity, scalar anomaly scores, and point prototypes can detect coarse normal/abnormal differences, but they often miss weak, local, fine-grained defect evidence. Our starting point is simple: compare the query image with a matched normal image. The difference between their CLIP features suppresses shared object appearance and exposes residual defect evidence. As shown in Fig. 1, this residual is not just an anomaly magnitude. Different defect families form different residual directions. However, the residual also changes when the normal reference changes, so raw residuals are not stable enough for answer scoring.

We therefore introduce a normal-quotient residual readout. Standard CLIP scoring lacks this evidence-separating step. We first estimate normal-reference variation from normal-normal pairs and remove those directions from the query residual. The remaining evidence is then read with shared-private defect subspaces: shared directions capture common abnormality, and private directions capture defect-specific patterns. Candidate hypotheses are scored by how well these subspaces explain the residual evidence and match the option text. The same evidence is reused patch-wise

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.

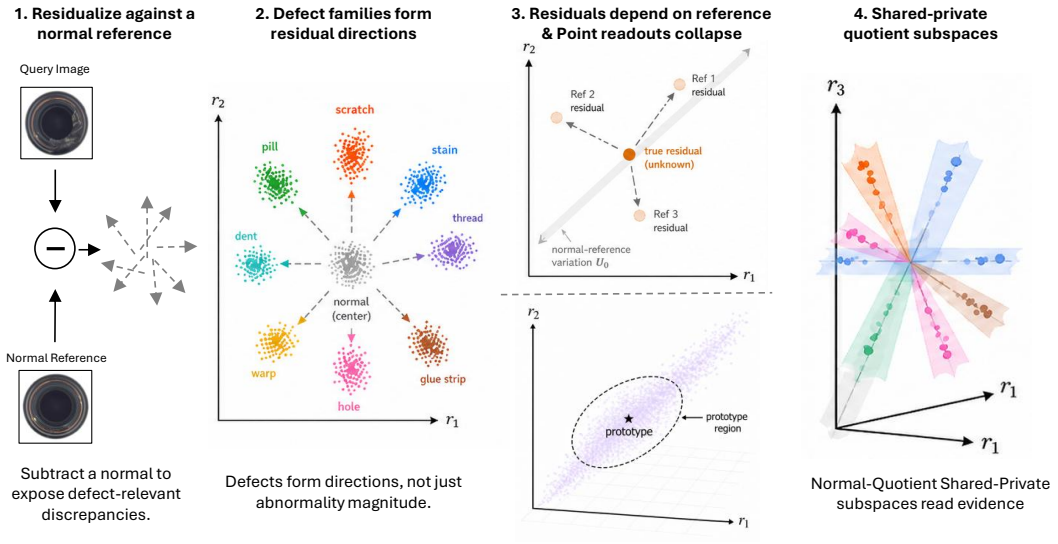


Figure 1: Normal-quotient residual readout. Comparing a query image with a normal reference exposes defect evidence, but the residual also changes with the chosen reference. We remove empirical normal variation and read candidate hypotheses from the remaining residual evidence, which also gives patch-wise grounding.

for answer-grounded localization, without an extra localization head. This turns CLIP from a global image-text matcher into a residual evidence reader for structured inspection QA.

On AnomalyCoT Cheng et al. (2025) and MMAD Jiang et al. (2024), our readout substantially improves over vanilla CLIP, raising AnomalyCoT accuracy from 28.74 to 41.19 and MMAD classification/description from $23.04/31.24$ to $69.79/74.59$. It places a frozen 0.15B CLIP backbone in a competitive regime for structured localization and description, while running at 56.30 ms P50 latency with 1.62 GB memory. These results suggest that practical inspection QA can improve not only by scaling generative models, but also by reading localized residual evidence better.

Our contributions are threefold.

- We identify a residual readout bottleneck in structured inspection QA: CLIP residuals contain defect-family directions, but global, scalar, and point-prototype readouts collapse the structure needed for fine-grained verification.
- We propose a normal-quotient shared-private residual readout that removes normal-reference variation, scores candidate hypotheses with residual subspaces, and reuses the same evidence for answer-grounded localization.
- We validate the readout on AnomalyCoT and MMAD, showing large gains over vanilla CLIP, competitive structured localization and description performance, and substantially lower latency and memory than larger MLLM-based systems.

2 Related Works

Industrial Anomaly Detection. Industrial anomaly detection (IAD) learns normal appearance under the assumption that defects are rare. Benchmarks have grown from MVTec-AD Bergmann et al. (2019) to logical anomalies Bergmann et al. (2022), larger scale Zou et al. (2022); Wang et al. (2024), and robustness Heckler-Kram et al. (2026). Methods model normal patches or memories Defard et al. (2020); Roth et al. (2022), synthesize anomalies Li et al. (2021); Zavrtanik et al. (2021), use flows or distillation Gudovskiy et al. (2022); Yu et al. (2021); Tien et al. (2023), or target efficiency Batzner et al. (2024). These outputs identify anomalous regions but do not directly verify query-dependent candidate defect explanations.

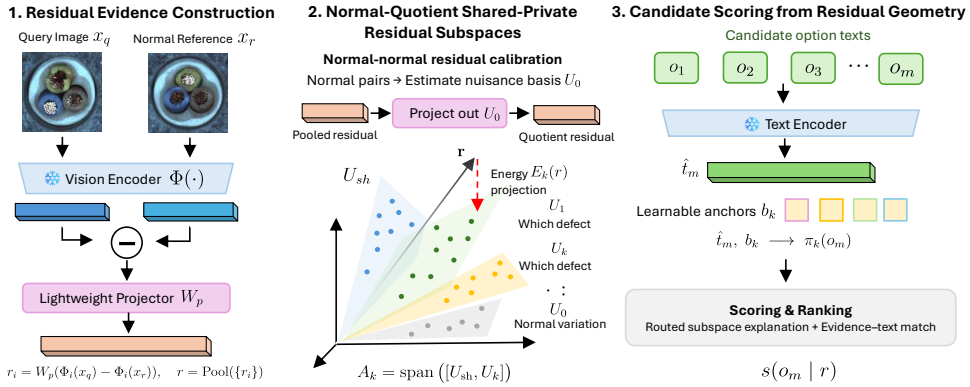


Figure 2: Overview of the normal-quotient residual readout. Normal-normal residuals are used to estimate an empirical normal-variation subspace U_0, and a query residual r is represented by the orthogonal canonical representative $\bar{r} = P_0^{\perp} r$. Shared-private residual subspaces are constrained to the complement of U_0, so candidate hypotheses are scored after removing the empirically estimated normal-reference variation. The same routed energies are applied patch-wise to produce decision-aligned localization without an additional localization head.

CLIP-based Anomaly Inspection. CLIP Radford et al. (2021) enables zero-shot image-text matching, and many works adapt it to anomaly inspection via prompts, patch scoring, visual context, or alignment refinement Jeong et al. (2023); Deng et al. (2023); Zhou et al. (2024); Li et al. (2024); Zhu & Pang (2024); Cao et al. (2024); Qu et al. (2024); Ham et al. (2024); Yang et al. (2025); Ma et al. (2025); Gong et al. (2025); Cai et al. (2024). Grounding and localization methods Lüddecke & Ecker (2022); Liu et al. (2023); Li et al. (2022); Gu et al. (2022); Yao et al. (2023); Lee et al. (2024); Gu et al. (2024a) further link language to image regions. However, these methods output anomaly scores, segmentation cues, or phrase-conditioned maps rather than rankings over query-dependent defect hypotheses, and their readouts typically rely on global similarity, scalar scores, prompt contrasts, or point-like prototypes. In inspection QA, where the correct option may hinge on a weak local residual, such readouts easily dilute or collapse the evidence. We instead study how CLIP residual evidence should be read for structured inspection QA.

Multimodal Large Language Models for Industrial Inspection. Recent benchmarks cast industrial inspection as MLLM evaluation: MMAD Jiang et al. (2024) uses multiple-choice questions, and AnomalyCoT Cheng et al. (2025) adds chain-of-thought samples with region-grounded annotations. Many datasets and adapted models have followed Xu et al. (2025); Wang et al. (2026); Jiang et al. (2026); Gu et al. (2024b); Li et al. (2023); Deng et al. (2024); Zhang et al. (2025); Zhao et al. (2025); Chao et al. (2025); Guan et al. (2025); Liao et al. (2025); Li et al. (2025); Kang et al. (2026); Miao et al. (2025); Peng et al. (2026); Chen et al. (2026). MLLMs suit tasks requiring open-ended language, external knowledge, or process-level reasoning, but they are costly to deploy on-premise and may answer from language priors without verifiable region-level evidence Zhao et al. (2025); Jiang et al. (2026). When candidate hypotheses are already given and the question is whether the image supports them, a lighter alternative suffices. Complementary to MLLMs, we target structured hypothesis verification with a lightweight CLIP readout that selects candidate defects and grounds them in localized visual evidence.

3 METHODOLOGY

We formulate structured inspection QA as *evidence-grounded hypothesis verification* over normal-referenced residuals. Given a query image x_q, a matched normal reference x_r, and a query-dependent hypothesis set $\mathcal{O} = \{o_1, \ldots, o_M\}$, the goal is to select the hypothesis best supported by localized visual evidence. The readout follows three steps: compare the query with a normal reference, remove normal-reference variation, and score candidate hypotheses.

3.1 Residual Evidence and Normal-Quotient Projection

Let $\Phi(\cdot)$ be a frozen CLIP vision encoder that returns N patch-level features, and let d be the readout dimension. For each patch i, we subtract the normal-reference feature from the query feature and project the difference:

$$r_i = W_p\big(\Phi_i(x_q) - \Phi_i(x_r)\big), \qquad r = \mathrm{Pool}\big(\{r_i\}_{i=1}^{N}\big), \tag{1}$$

where W_p is a lightweight projector shared across patches, $r_i \in \mathbb{R}^d$ is the patch residual, and $r \in \mathbb{R}^d$ is the pooled residual evidence vector. In all main experiments, Pool denotes full-image average pooling for option scoring, while the unpooled patch residuals are retained for grounding. Residualization removes much of the shared normal appearance and exposes possible defect evidence. The pooled residual r is not used as a scalar anomaly score. As illustrated in Fig. 3, normal-referenced residuals form defect-family directions with low-rank within-family variation. This motivates a readout that keeps residual structure instead of collapsing it into global similarity, scalar abnormality, or a prototype.

Raw residuals still depend on which normal reference is chosen. This is a problem because pose, texture, illumination, or normal instance variation can look like residual signal even when they are not defect evidence. To remove this reference-dependent variation, we estimate empirical normal-reference directions from normal-normal residuals. Let $\mathcal{P}_0$ be the set of normal image pairs from the support or deployment calibration pool, and let r_{ij}^0 denote the pooled residual between two normal images x_i^0 and x_j^0, computed by Eq. 1. We estimate a normal-variation basis $U_0 \in \mathbb{R}^{d\times r_0}$ and project the query residual away from it:

$$C_0 = \frac{1}{|\mathcal{P}_0|}\sum_{(i,j)\in\mathcal{P}_0} r_{ij}^0 (r_{ij}^0)^\top, \qquad U_0 = \mathrm{TopEig}_{r_0}(C_0), \qquad \bar{r} = (I - U_0U_0^\top)r. \tag{2}$$

Here, C_0 is the covariance of normal-normal residuals, r_0 is the normal-variation rank, and $\bar{r}$ is the nuisance-controlled residual used by all downstream scores. Because U_0 is estimated only from normal-normal pairs, it captures reference-induced variation rather than defect labels. The normal bank does not use test defect labels, test masks, test answer options, or target answer information; it is used only to estimate nuisance directions.

3.2 Nuisance-Orthogonal Shared-Private Subspaces

We organize residual evidence by defect families, with K defined by the support-set defect-type taxonomy. This family index only organizes residual subspaces; the test-time option sheet remains query-dependent and may contain descriptions not seen verbatim during training.

A single prototype is too limited because one defect family can appear with different severity, shape, location, and material. We therefore use a shared abnormality basis $U_{\mathrm{sh}} \in \mathbb{R}^{d\times r_s}$ and a private basis $U_k \in \mathbb{R}^{d\times r_p}$ for each family k, where r_s and r_p are the shared and private ranks. The shared basis captures abnormality directions useful across families, while the private basis captures family-specific residual patterns. The quotient projection in Eq. 2 removes nuisance from the input residual, but the learned bases themselves must also avoid nuisance directions. Otherwise, the model could explain normal-reference variation as if it were defect evidence. At every forward pass, we project the shared basis away from the normal-variation subspace and each private basis away from both the normal and shared subspaces:

$$\tilde{U}_{\mathrm{sh}} = \mathrm{orth}\big((I - U_0U_0^\top)U_{\mathrm{sh}}\big), \qquad \tilde{U}_k = \mathrm{orth}\big((I - Q_{0,\mathrm{sh}}Q_{0,\mathrm{sh}}^\top)U_k\big), \tag{3}$$

where $Q_{0,\mathrm{sh}} = \mathrm{orth}([U_0, \tilde{U}_{\mathrm{sh}}])$. Below, U_{sh} and U_k refer to these effective projected bases.

This deterministic projection removes the empirically estimated normal-reference directions from the effective bases. For each family, let $Q_k = \mathrm{orth}([U_{\mathrm{sh}}, U_k])$ and $P_{A_k} = Q_kQ_k^\top$ denote the projection onto its shared-private residual subspace. We define the explanation energy of family k as

$$E_k^Q(r) = \|\bar{r} - P_{A_k}\bar{r}\|_2^2. \tag{4}$$

A lower energy means that family k explains more of the nuisance-controlled residual evidence. Thus, Eq. 4 replaces prototype matching: it scores a family by subspace explanation, not centroid distance.

3.3 Hypothesis Scoring and Null-Option Scoring

Structured inspection QA scores candidate texts, not fixed labels. Let o_m be the m-th candidate and let $\hat{t}_m = g(o_m)/\|g(o_m)\|_2$ be its frozen CLIP text embedding. Because one description can match multiple defect families, we softly route each option to family subspaces using learnable anchors b_k and temperature τ:

$$\pi_k(o_m) = \frac{\exp(b_k^\top \hat{t}_m/\tau)}{\sum_{j=1}^{K} \exp(b_j^\top \hat{t}_m/\tau)}. \tag{5}$$

Here, $\pi_k(o_m)$ measures how relevant family k is to option o_m; it does not decide the answer alone. Using the family energy $E_k^Q(r)$ and projection P_{A_k} from Eq. 4, the final option score is

$$s(o_m \mid r) = \sum_{k=1}^{K} \pi_k(o_m) \left[-E_k^Q(r) + \lambda_{\text{sem}} \,\text{sim}\big(W_k^\top P_{A_k} \bar{r}, \hat{t}_m\big) \right]. \tag{6}$$

The energy term checks whether the routed subspaces explain the residual evidence, while the semantic term checks whether the explained evidence matches the option text. Here, W_k is a family-specific projection head, sim is cosine similarity, and λ_{sem} is a learnable scalar that balances the two terms. The `no-defect` and `Unknown/Other` options use a calibrated null score $s_\varnothing$ rather than an extra defect family. At inference, the null option is selected when $s_\varnothing$ exceeds all candidate scores.

3.4 Training and Grounding

The CLIP image and text encoders are frozen throughout. The normal-reference variation basis U_0 is estimated once from normal-normal residuals using Eq. 2 and then held fixed. Only the lightweight readout is trained: the residual projector W_p, the pre-projection shared and private bases $\{U_{\text{sh}}, U_k\}_{k=1}^K$, the semantic projection heads $\{W_k\}_{k=1}^K$, the family anchors $\{b_k\}_{k=1}^K$, and the scalar scale λ_{sem}. The projections in Eq. 3 are deterministic and add no learned parameters. The ranks r_0, r_s, and r_p are fixed throughout the main experiments; sensitivity is reported in Appendix E.5.

Training fits the residual readout, not the CLIP backbone. We use option-level cross-entropy for the candidate answer, an auxiliary family-energy loss when support-set family labels are available, and the null calibration loss described in Appendix A. These losses organize the residual subspaces offline without turning the model into a closed-set classifier. At inference, the model only needs the query image, a matched normal reference, and the candidate hypotheses.

The same learned subspaces used for scoring also provide patch-wise evidence for localizing the selected answer. For each patch residual r_i from Eq. 1, we remove the same normal-reference variation as in Eq. 2, giving $\bar{r}_i = (I - U_0 U_0^\top) r_i$. For the selected hypothesis $\hat{o}$, the answer-supporting evidence map is defined as:

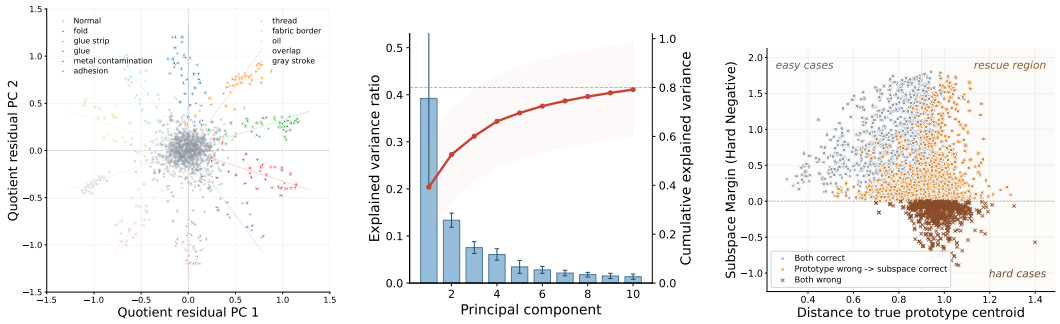
$$A_i(\hat{o}) = \sum_{k=1}^{K} \pi_k(\hat{o}) \, \|P_{A_k} \bar{r}_i\|_2^2. \tag{7}$$

A patch receives a high score when its residual is well absorbed by the same routed subspaces that support the selected answer in Eq. 6. Thus, grounding is the patch-wise expression of the decision evidence, not a separate detector, post-hoc attribution module, or mask-supervised output head. The full training objective and loss weights are given in Appendix A.

4 Experiments

4.1 Experimental Setting

Benchmarks and tasks. We evaluate structured inspection QA on AnomalyCoT Cheng et al. (2025) and MMAD Jiang et al. (2024). For AnomalyCoT, we report option accuracy and answer-grounded localization metrics, IoU, GIoU, and DIoU. For MMAD, we evaluate five defect-related subtasks: detection, classification, localization, description, and analysis. Analysis is reported for completeness, but our primary target is visually grounded hypothesis verification rather than functional or causal reasoning beyond visible evidence.



(a) Quotient residual directions (b) Low-rank family variation (c) Subspace rescue

Figure 3: Residual evidence is directional and low-rank after normal-reference variation removal. (a) Quotient residuals show defect-family directions; labels only color the projection, which uses no learned shared–private subspaces. (b) Within-family variation concentrates in a few principal components. (c) Samples far from the true prototype can still have positive hard-negative subspace margins, revealing a subspace rescue regime.

Structured option-ranking protocol. All methods receive the same query image and candidate hypothesis set $\mathcal{O}$, and each model selects one option. For CLIP-based readouts, the prediction is the highest-scoring candidate hypothesis. For MLLM baselines, we rank candidate option letters by conditional likelihood under a fixed multiple-choice prompt. This aligns candidate selection while retaining different adaptation protocols; generation and abnormal-support controls appear in Appendices D.2 and D.3.

k-shot semantics and data usage. For MLLMs, k-shot denotes adding k normal in-context exemplars at inference time, following benchmark convention. For our method, k-shot denotes using k defect images per family for offline residual-readout fitting. At inference, each query uses one object-matched normal reference selected by a fixed index rule, and we take the first entry from the benchmark-provided object-wise normal-reference list. The assignment is fixed before evaluation and does not use defect labels, masks, answer options, or target answers.

Metrics and implementation. We report option accuracy for answer selection and IoU, GIoU, and DIoU for answer-grounded localization. Grounding boxes are derived from the top 15% patches of the evidence map in Eq. 7, with at least 5 patches. Because our native output is a patch-level evidence map, box metrics are used only as coarse compatibility proxies; box-free faithfulness metrics are discussed in Sec. 4.3. All experiments use a frozen OpenCLIP ViT-B/16-plus-240 backbone on a single NVIDIA RTX A6000 Ada Generation GPU. Implementation details are in Appendix A.

4.2 Main Evaluation

We first verify that normal-referenced CLIP residuals provide a useful evidence space. Fig. 3 shows three patterns: defect families form coherent residual directions after normal-variation removal, within-family variation is low-rank, and many prototype-failure cases retain positive subspace margins. These diagnostics motivate reading candidate hypotheses from nuisance-controlled residual subspaces rather than scalar anomaly scores or point prototypes.

Table 1 reports interface-aligned ranking results. Our readout improves vanilla CLIP substantially on AnomalyCoT, raising accuracy from 28.74 to 41.19 and IoU from 0.1562 to 0.1742, with consistent gains in GIoU and DIoU. On MMAD, it improves classification, localization, description, and analysis by large margins. Under the same ranking interface, the 0.15B CLIP readout is competitive with 4B–8B MLLMs on MMAD detection, classification, localization, and description, while trailing the strongest MLLM on AnomalyCoT accuracy and MMAD analysis. As benchmark-reported reference points, GPT-4o attains: AnomalyCoT Acc. 49.19; MMAD Det./Cls./Loc./Desc./Anal. $68.63/65.80/55.62/73.21/83.41$. Further, Appendix D reports diagnostic native-condition reference points for CLIP-based IAD methods (WinCLIP Jeong et al. (2023), AnomalyCLIP Zhou et al. (2024), AdaCLIP Cao et al. (2024)), preserving their intended anomaly scoring/localization

Table 1: Ranking under a shared option sheet. 4N-ICL uses four normal in-context images; 4D/F uses four defect supports per family for offline fitting. Grounding applies the same box conversion to MLLM attention maps and our evidence maps. Additional MLLM controls use object-matched all-family ICL with one normal and 4.1 abnormal images per object on average (Appendix D.3).

Model	**Scale**	**AnomalyCoT**				**MMAD**				
		Acc	IoU	GIoU	DIoU	Det.	Cls.	Loc.	Desc.	Anal.
InternVL2	8B	39.53	0.1452	0.1422	0.1140	44.44	42.54	47.63	55.24	74.91
Qwen2.5-VL	7B	46.31	0.1423	0.1419	0.1141	67.59	58.65	59.31	69.46	71.35
LLaVA-1.5	7B	23.95	0.1471	0.1426	0.1172	67.05	39.45	41.85	51.86	69.85
Gemma3	4B	32.85	0.1421	0.1420	0.1141	51.47	37.32	29.01	54.55	71.38
InternVL2: 4N-ICL	8B	39.10	0.1450	0.1426	0.1146	47.69	43.36	48.52	56.64	76.06
Qwen2.5-VL: 4N-ICL	7B	47.27	0.1426	0.1423	0.1159	67.36	58.77	61.37	66.55	77.72
LLaVA-1.5: 4N-ICL	7B	23.10	0.1468	0.1427	0.1170	67.52	40.05	39.54	52.10	71.35
Gemma3: 4N-ICL	4B	32.17	0.1420	0.1419	0.1140	39.58	38.39	28.50	53.03	72.61
Vanilla-CLIP	0.15B	28.74	0.1562	0.1482	0.1233	42.82	23.04	20.15	31.24	24.64
Ours: 4D/F	0.15B	41.19	0.1742	0.1625	0.1363	68.98	69.79	61.88	74.59	73.76

Table 2: Comparison with recent anomaly-aware VLMs in adaptation burden, additional supervision, inference compute, and reported MMAD performance. FLOPs are measured per end-to-end QA inference sample under the same profiling protocol.

Model	Scale	Training Burden / Extra Supervision	MMAD Avg. ↑	FLOPs (G) ↓
EIAD	8B	mask supervision + 280K QA pairs	62.35	5976.66
OmniAD	7B	full-parameter SFT + GRPO	75.35	2452.03
AnomalyR1	3B	RL post-training (GRPO + ROAM)	71.91	1303.15
Ours: 4-shot	0.15B	frozen CLIP + residual readout	69.80	248.06
Ours: 16-shot	0.15B	frozen CLIP + residual readout	72.84	248.06

pipelines when lifting their outputs to structured inspection QA. Table 2 shows a favorable accuracy–compute trade-off: the 4-shot model reaches 69.80 MMAD average with 248.06G FLOPs, and the 16-shot variant reaches 72.84 with unchanged inference cost. Even when MLLMs receive object-matched defect examples from every available family through ICL, our MMAD average remains higher (69.80 vs. at most 60.46), while Qwen2.5-VL reaches 54.98 AnomalyCoT accuracy (Appendix D.3).

Table 3a and Table 3b summarize the two practical assumptions behind our setting. First, the readout is lightweight: it runs at 56.30 ms P50 latency with 1.62 GB peak VRAM on a single RTX A6000 Ada Generation GPU, making it 4.4× to 17.3× faster and 8.1× to 17.0× more memory-efficient than open MLLM baselines. Second, the option sheets are not closed-set labels: defect-description questions show very low train-to-test overlap, with SMR 0.97%, OOR 4.18%, and AOR 1.85%.

Figure 4a confirms this distinction more directly. Known/novel subsets are defined within the held-out test set by train-answer overlap, and are shared across all methods. Answer-frequency, text-only, fixed-label, and global CLIP controls all show positive known-to-novel gaps, with the answer-frequency prior dropping from 66.48% to 15.68%. In contrast, our readout is the only method with a negative gap, reaching 71.98% on known answers and 75.30% on novel answers. These controls support image-conditioned ranking beyond answer-string memorization. Four-family holdout yields family-macro AnomalyCoT/MMAD accuracies of 46.07/68.24, versus 57.14/79.03 with all-family support on the same subsets, indicating partial transfer (Appendix C.3). Figure 4b isolates readout effects under matched support: all variants share the same backbone, support data, normal-reference, optimizer, and option-ranking loss. Prototypes perform poorly, linear readouts improve, and the full shared-private subspace readout performs best, reaching 41.19 AnomalyCoT accuracy and 69.80

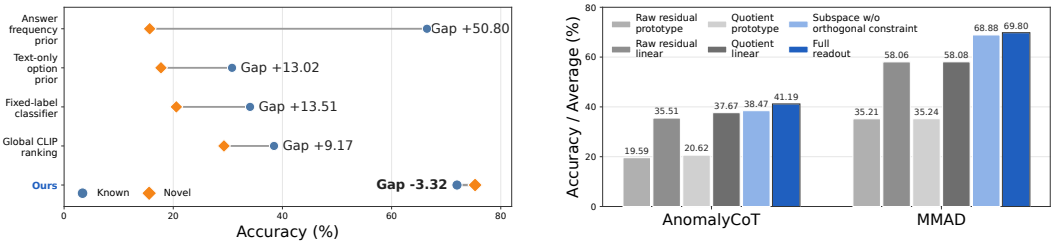
Table 3: Efficiency and option-sheet overlap analysis. (a) reports per-sample latency and peak GPU memory on a single RTX A6000 Ada Generation GPU with batch size 1 in BF16. (b) reports MMAD train-to-test option overlap, showing especially low overlap for defect descriptions and supporting dynamic hypothesis ranking over closed-set classification.

(a) Efficiency.

Model	**Scale**	**Latency (ms/sample)**		**Peak VRAM (GB)**
		P50	P95	
InternVL2	8B	249.09	320.32	15.30
LLaVA1.5-VL	7B	357.05	810.52	27.57
Qwen2.5-VL	7B	575.85	622.69	16.03
Gemma3	4B	971.91	1706.69	13.14
Ours	0.15B	56.30	121.43	1.62

(b) Option overlap.

Metric	Classification	Description
SMR	14.02%	0.97%
OOR	9.28%	4.18%
AOR	5.53%	1.85%



(a) Known–novel answer split.

(b) Matched-support readout isolation.

Figure 4: (a) Only our readout achieves Novel $>$ Known (gap -3.32), while closed-set and text-prior baselines all collapse on novel answers. (b) With matched backbone, support, and training, shared-private subspace readouts dominate prototype/linear variants on both benchmarks.

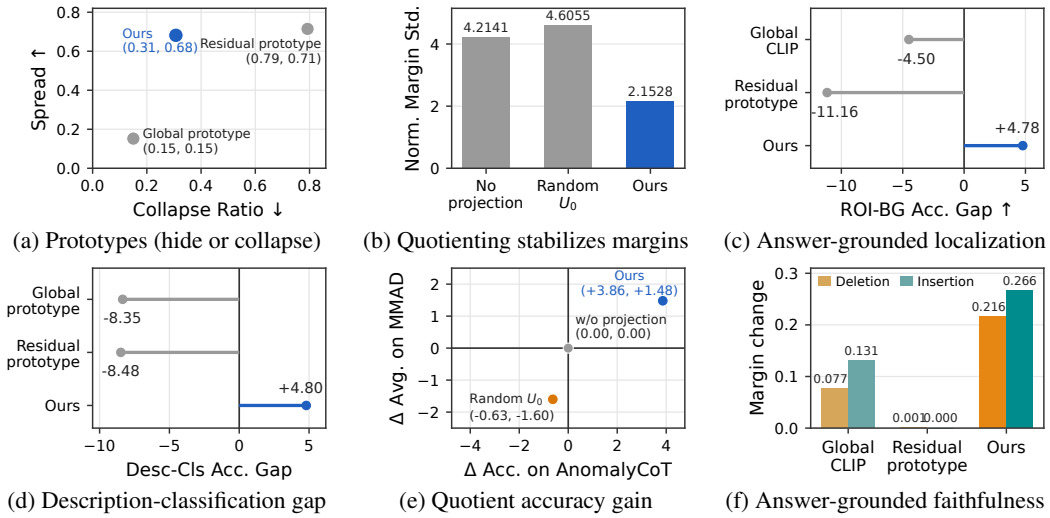
MMAD average. This supports that the gain comes from how residual evidence is read, not merely from offline support data.

4.3 Discussion

Why point prototypes fail. Fig. 5a and Fig. 5d show that point prototypes are a poor fit for fine-grained inspection QA. The global prototype appears compact because it suppresses within-family variation, while the residual prototype exposes larger spread but also high collapse, indicating that one centroid cannot organize the residual variation it reveals. Our shared-private subspace readout retains residual spread while keeping collapse moderate, and it changes the description-classification gap from negative under prototypes to positive. The gain is not merely better coarse defect recognition, but preserving residual variation needed for description-sensitive hypothesis verification.

Why the normal quotient matters. Fig. 5b and Fig. 5e test whether quotienting stabilizes decisions across matched normal references. Without quotienting, normalized margin instability remains high. Replacing the nuisance subspace with a random U_0 further increases instability and yields negative accuracy changes relative to no projection. In contrast, empirical U_0 reduces Norm. Margin Std to 2.1528 and improves accuracy over the same no-projection baseline by $+3.86$ on AnomalyCoT and $+1.48$ on MMAD. This indicates that the quotient does not merely suppress all variation; it stabilizes discriminative margins while preserving defect evidence. Further reference-selection and mismatch tests appear in Appendix E.4.

Component ablation. Table 4 isolates each component's contribution. Removing shared-private decomposition, orthogonal constraint, or soft routing consistently degrades performance on AnomalyCoT and MMAD. Along with quotient diagnostics in Figs. 5b and 5e, these results support that the method works as a coupled nuisance-controlled residual readout rather than a heuristic component.



(a) Prototypes (hide or collapse) (b) Quotienting stabilizes margins (c) Answer-grounded localization

(d) Description-classification gap (e) Quotient accuracy gain (f) Answer-grounded faithfulness

Figure 5: Diagnostics of residual-evidence readout. (a,d) Point prototypes hide or collapse within-family residual variation, while subspaces preserve description-sensitive evidence. (b,e) Empirical normal quotienting stabilizes margins and improves accuracy over no projection and random U_0. (c,f) Selected-answer evidence maps carry more decision-relevant signal in ROI–BG and deletion/insertion tests. Definitions and protocols are in Appendix B.

Table 4: Targeted ablation across AnomalyCoT and MMAD.

Intervention	**AnomalyCoT**		**MMAD**				
	Acc.	IoU	Det.	Cls.	Loc.	Desc.	Anal.
Full model	41.19	0.1742	68.98	69.79	61.88	74.59	73.76
w/o Shared-Private Decomp.	35.56	0.1662	68.75	64.92	61.87	73.24	71.35
w/o Orthogonal constraint	38.47	0.1626	68.67	68.14	61.87	72.54	73.18
w/o Soft family routing	38.68	0.1626	68.67	69.31	61.45	72.07	72.95

Answer-grounded localization and faithfulness. Fig. 5c and Fig. 5f evaluate whether selected answers are supported by spatial evidence. Our evidence maps yield a positive ROI-background accuracy gap, while global and residual-prototype readouts rely more heavily on background-only cues. Deletion and insertion tests further show that highlighted regions affect the answer margin more strongly for our method than for global or prototype readouts. Thus, the grounding map is the spatialization of decision evidence, not a separate detector or post-hoc visualization.

Scope. Our readout verifies supplied hypotheses; it cannot name omitted defects or reason procedurally. Under the fixed decision rule, benchmark null precision/recall/F1 are $78.85/73.69/76.18\%$ (Appendix E.3); rejection when the correct hypothesis is deliberately removed remains untested.

5 Conclusion

We presented a lightweight normal-quotient residual readout for structured industrial inspection QA. Normal-referenced CLIP residuals contain defect-family directions beyond scalar abnormality, but must be separated from normal-reference variation for reliable hypothesis verification. Our readout removes empirical normal-variation directions, explains the remaining evidence with shared-private defect subspaces, and reuses the same routed subspaces for answer-grounded localization. On AnomalyCoT and MMAD, it improves substantially over vanilla CLIP and places a frozen 0.15B backbone in a competitive regime for structured localization and description, with much lower latency and memory than larger MLLM systems. These results show that practical inspection QA can improve by better reading localized residual evidence, not only by scaling generative models.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the English text, grammar correction, and improving readability. We have not used generative AI tools for generating core research ideas, deriving mathematical proofs, synthesizing primary literature, or formulating novel scientific claims, and generating experimental data with AI is not applicable to this work. Additionally, we used generative AI tools for formatting LaTeX tables and generating boilerplate code for data visualization. We have reviewed all AI-assisted work. Specifically, all AI-assisted text revisions were meticulously cross-checked by the authors to ensure scientific accuracy and preserve the original meaning. Furthermore, any AI-generated code snippets were thoroughly verified, executed, and tested for correctness 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.

ETHICS STATEMENT

This work studies structured industrial inspection QA using the publicly available AnomalyCoT and MMAD benchmarks. The study does not involve human participants, collection of personal data, or biometric information.

REPRODUCIBILITY STATEMENT

Section 3 describes the proposed framework, including normal-referenced residual evidence construction, empirical normal-quotient projection, nuisance-orthogonal shared-private defect subspaces, option scoring, and answer-grounded patch-wise localization. Section 4 specifies the benchmarks, evaluation tasks, metrics, option-ranking protocols, and MLLM comparison settings. Appendix A provides complete implementation details, including the OpenCLIP vision backbone, lightweight projector architecture, optimization settings, loss function parameters, subspace rank configurations, and normal-reference selection rules. Appendix defines the exact mathematical formulations for metric calculations, geometry diagnostics, normal-reference stability tests, and answer-grounded localization faithfulness.

REFERENCES

Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 128–138, 2024. doi: 10.1109/WACV57701.2024.00020.

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, 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. doi: 10.1007/s11263-022-01578-9.

Yuxuan Cai, Xinwei He, Dingkang Liang, Ao Tong, and Xiang Bai. Anomaly detection by adapting a pre-trained vision language model. *arXiv preprint arXiv:2403.09493*, 2024.

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, 2024. doi: 10.1007/978-3-031-72761-0_4.

Yuhao Chao, Jie Liu, Jie Tang, and Gangshan Wu. Anomalyr1: A grpo-based end-to-end mllm for industrial anomaly detection. *arXiv preprint arXiv:2504.11914*, 2025.

Peng Chen, Chao Huang, Yunkang Cao, Chengliang Liu, Wenqiang Wang, Mingbo Yang, Li Shen, Wenqi Ren, and Xiaochun Cao. Reason-iad: Knowledge-guided dynamic latent reasoning for explainable industrial anomaly detection. *arXiv preprint arXiv:2602.09850*, 2026.

Jiaxi Cheng, Yuliang Xu, Shoupeng Wang, Ma Tao, Yuchen He, Jinghe Zhang, Sihang Cai, Jiawei Zhen, Jingyi Jia, Yao Wan, Yan Xia, and Zhou Zhao. Anomalycot: A multi-scenario chain-of-thought dataset for multimodal large language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: A patch distribution modeling framework for anomaly detection and localization. *arXiv preprint arXiv:2011.08785*, 2020.

Hanqiu Deng, Zhaoxiang Zhang, Jinan Bao, and Xingyu Li. Anovl: Adapting vision-language models for unified zero-shot anomaly localization. *arXiv preprint arXiv:2308.15939*, 2023.

Huilin Deng, Hongchen Luo, Wei Zhai, Yang Cao, and Yu Kang. Vmad: Visual-enhanced multimodal large language model for zero-shot anomaly detection. *arXiv preprint arXiv:2409.20146*, 2024.

Tao Gong, Qi Chu, Bin Liu, Zhou Wei, and Nenghai Yu. Fe-clip: Frequency enhanced clip model for zero-shot anomaly detection and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations (ICLR)*, 2022.

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)*, 2024a. doi: 10.1145/3664647.3680685.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024b.

Wei Guan, Jun Lan, Jian Cao, Hao Tan, Huijia Zhu, and Weiqiang Wang. Emit: Enhancing mllms for industrial anomaly detection via difficulty-aware grpo. *arXiv preprint arXiv:2507.21619*, 2025.

Denis 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 (WACV)*, pp. 98–107, 2022.

Jiyul Ham, Yonggon Jung, and Jun-Geol Baek. Glocalclip: Object-agnostic global-local prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2411.06071*, 2024.

Lars Heckler-Kram, Jan-Hendrik Neudeck, Ulla Scheler, Rebecca König, and Carsten Steger. The mvtec ad 2 dataset: Advanced scenarios for unsupervised anomaly detection. *International Journal of Computer Vision*, 134:175, 2026. doi: 10.1007/s11263-026-02743-0.

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, 2023. doi: 10.1109/CVPR52729.2023.01878.

Xi Jiang, Jian Li, Hanqiu Deng, Yong Liu, Bin-Bin Gao, Yifeng Zhou, Jialin Li, Chengjie Wang, and Feng Zheng. Mmad: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. *arXiv preprint arXiv:2410.09453*, 2024.

Xi Jiang, Yue Guo, Jian Li, Yong Liu, Bin-Bin Gao, Hanqiu Deng, Jun Liu, Heng Zhao, Chengjie Wang, and Feng Zheng. Ad-copilot: A vision-language assistant for industrial anomaly detection via visual in-context comparison. *arXiv preprint arXiv:2603.13779*, 2026.

Hyunju Kang, Woohyun Lee, Jaewon Kim, and Hogun Park. Judo: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa. In *International Conference on Learning Representations (ICLR)*, 2026.

Jungbeom Lee, Sanghyuk Chun, and Sangdoo Yun. Toward interactive regional understanding in vision-large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6416–6429, 2024. doi: 10.18653/v1/2024.naacl-long.356.

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 (CVPR)*, pp. 9664–9674, 2021.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, 2022.

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.

Yuanze Li, Haolin Wang, Shihao Yuan, Ming Liu, Debin Zhao, Yiwen Guo, Chen Xu, Guangming Shi, and Wangmeng Zuo. Myriad: Large multimodal model by applying vision experts for industrial anomaly detection. *arXiv preprint arXiv:2310.19070*, 2023.

Yuanze Li, Shihao Yuan, Haolin Wang, Qizhang Li, Ming Liu, Chen Xu, Guangming Shi, and Wangmeng Zuo. Triad: Empowering lmm-based anomaly detection with vision expert-guided visual tokenizer and manufacturing process. *arXiv preprint arXiv:2503.13184*, 2025.

Jingyi Liao, Yongyi Su, Rong-Cheng Tu, Zhao Jin, Wenhao Sun, Yiting Li, Dacheng Tao, Xun Xu, and Xulei Yang. Ad-fm: Multimodal llms for anomaly detection via multi-stage reasoning and fine-grained reward optimization. *arXiv preprint arXiv:2508.04175*, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7086–7096, 2022.

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, 2025. doi: 10.1109/CVPR52734.2025.00447.

Junwen Miao, Penghui Du, Yi Liu, Yu Wang, and Yan Wang. Agentiad: Tool-augmented single-agent for industrial anomaly detection. *arXiv preprint arXiv:2512.13671*, 2025.

Xiaomeng Peng, Xilang Huang, and Seon Han Choi. Eagle: Expert-augmented attention guidance for tuning-free industrial anomaly detection in multimodal large language models. *arXiv preprint arXiv:2602.17419*, 2026.

Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 301–317, 2024. doi: 10.1007/978-3-031-72890-7_18.

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 (ICML)*, pp. 8748–8763, 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.

Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T. M. Duong, Chanh D. Tr. Nguyen, and Steven Q. H. Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24511–24520, 2023.

Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22883–22892, 2024.

Zhuonan Wang, Zhenxuan Fan, Siwen Tan, Yu Zhong, Yuqian Yuan, Haoyuan Li, Hao Jiang, Wenqiao Zhang, Feifei Shao, Hongwei Wang, and Jun Xiao. Mau-gpt: Enhancing multi-type industrial anomaly understanding via anomaly-aware and generalist experts adaptation. *arXiv preprint arXiv:2602.07011*, 2026.

Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20370–20382, 2025.

Ziteng Yang, Jingzehua Xu, Yanshu Li, Zepeng Li, Yeqiang Wang, and Xinghui Li. Vip^2-clip: Visual-perception prompting with unified alignment for zero-shot anomaly detection. *arXiv preprint arXiv:2505.17692*, 2025.

Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, and Hang Xu. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23497–23506, 2023. doi: 10.1109/CVPR52729.2023.02250.

Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Dræm – a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8330–8339, 2021. doi: 10.1109/ICCV48922.2021.00822.

Zongyun Zhang, Jiacheng Ruan, Xian Gao, Ting Liu, and Yuzhuo Fu. Eiad: Explainable industrial anomaly detection via multi-modal large language models. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2025.

Shifang Zhao, Yiheng Lin, Lu Han, Yao Zhao, and Yunchao Wei. Omniad: Detect and understand industrial anomaly via multimodal reasoning. *arXiv preprint arXiv:2505.22039*, 2025.

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)*, 2024.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18480–18490, 2024. doi: 10.1109/CVPR52733.2024.01688.

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, 2022. doi: 10.1007/978-3-031-20056-4_23.

A Experimental Protocol and Implementation Details

A.1 Benchmarks and Data Usage

We evaluate on AnomalyCoT and MMAD following Sec. 4. For AnomalyCoT, we report option accuracy and answer-grounded localization; for MMAD, we report detection, classification, localization, description, and analysis. The analysis subtask is included for completeness, while our target setting is visually grounded hypothesis verification.

For MLLMs, k-shot denotes k normal in-context examples at inference time. For our method, it denotes k abnormal images per defect family from the benchmark training split for offline readout fitting; evaluation queries belong to the separate test split and are not used for fitting. Normal images for nuisance calibration are separate from these abnormal supports. The comparison shares test queries, candidate sheets, and metrics, but uses different supervision budgets and adaptation mechanisms. The abnormal-support ICL control in Appendix D.3 additionally gives MLLMs labeled defect examples.

A.2 Normal-Reference Protocol

Normal calibration. We estimate one global normal-variation basis U_0 from support-set normal–normal residual pairs. We sample 8–128 pairs per object, depending on availability, up to $2{,}048$ pairs total, form the empirical second-moment matrix C_0, and compute U_0 once by top-r_0 eigendecomposition. Object-stratified sampling reduces the influence of unequal normal-image availability on the global estimate; it is separate from the abnormal-support supervision budget. U_0 is fixed during training and inference and uses no defect labels, masks, answer options, or target answers.

Inference-time reference selection. Each benchmark provides randomized category-matched normal-reference candidate lists. We use the first entry, with this list position fixed before evaluation and the assignment shared across the corresponding QA rows. Selection uses neither pose-based retrieval nor manual matching and does not use defect labels, masks, candidate text, or target answers. The deployment pool is assumed to be restricted to the relevant product and acquisition view, for example by the inspection recipe or camera station. Alternative-reference and reference-only perturbation tests are reported in Appendix E.4.

A.3 Architecture

We use frozen OpenCLIP ViT-B/16-plus-240 (LAION400M-e32) at 240×240 resolution, yielding a 15×15 patch grid. Trainable parameters are the residual projector W_p, shared/private residual subspaces, semantic projection heads, and family anchors. W_p is a single linear layer without LayerNorm or dropout.

Unless stated otherwise, we use a fixed low-rank setting across all experiments: shared rank $r_s = 4$, private rank $r_p = 8$, and normal-variation rank $r_0 = 4$. These ranks are not tuned on the test set. Option scoring uses full-image average pooling, while unpooled patch residuals are retained for answer-grounded maps.

A.4 Training objective and loss weights

The CLIP image and text encoders are frozen, and only the lightweight residual readout is trained. Let $s_m = s(o_m \mid r)$ denote the score of candidate option o_m in Eq. 6, and let $s_\varnothing$ denote the calibrated null score used for `no-defect` and `Unknown/Other` options.

For samples whose correct answer is one of the candidate defect hypotheses, we use the option-level cross-entropy

$$\mathcal{L}_{\mathrm{QA}} = -\log \frac{\exp(s_y)}{\sum_{m=1}^{M} \exp(s_m)}, \tag{8}$$

where y is the ground-truth option index. When a support-set defect-family label z is available, we also supervise the family-energy logits $a_k = -E_k^Q(r)$ with

$$\mathcal{L}_{\text{type}} = -\log \frac{\exp(a_z)}{\sum_{k=1}^{K} \exp(a_k)}. \tag{9}$$

Finally, $\mathcal{L}_{\text{null}}$ calibrates the separate null score as a binary normal-vs-defect decision. Let $u \in \{0, 1\}$ indicate whether the sample is null-labeled, and let $\ell_{\text{def}} = \log \sum_{k=1}^{K} \exp(a_k)$ denote the aggregate defect-family logit. We use

$$\mathcal{L}_{\text{null}} = -u \log \sigma(s_\varnothing - \ell_{\text{def}}) - (1 - u) \log \left(1 - \sigma(s_\varnothing - \ell_{\text{def}})\right). \tag{10}$$

This loss only calibrates the null score against aggregate defect evidence; candidate-option ranking remains supervised by $\mathcal{L}_{\text{QA}}$. The total objective is

$$\mathcal{L} = \lambda_{\text{QA}} \mathcal{L}_{\text{QA}} + \lambda_{\text{type}} \mathcal{L}_{\text{type}} + \lambda_{\text{null}} \mathcal{L}_{\text{null}}. \tag{11}$$

We use $\lambda_{\text{QA}} = 1.0$, $\lambda_{\text{type}} = 0.5$, and $\lambda_{\text{null}} = 0.1$ in all main experiments; All loss weights were fixed during development and not tuned on the test split.

The semantic scale λ_{sem} is learned jointly with the lightweight readout and initialized to 1.0.

A.5 Optimization

We train the lightweight readout with AdamW for 12 epochs using learning rate $3{\times}10^{-4}$, weight decay 10^{-4}, batch size 32, mixed precision, and seed 42 on a single NVIDIA RTX A6000 Ada Generation GPU. For matched readout ablations, we keep the backbone, support split, normal-reference protocol, optimizer, and option-ranking loss fixed, and change only the ablated readout component.

A.6 Inference and Grounding

At inference, candidate options are scored by the routed shared-private subspace readout and ranked under the benchmark option interface. Grounding uses the same routed subspaces on unpooled patch residuals. Boxes are extracted from the top 15% evidence patches, with at least 5 patches selected.

These boxes are coarse compatibility proxies for coordinate-style evaluation. The native output is a patch-level answer-evidence map, not a detector-level segmentation mask. Masks are used only for evaluation.

Unified grounding protocol. For MLLM baselines, we apply the same map-to-box conversion rule to attention-derived maps: after option ranking, we take last-layer attention from the selected option token to image patches and extract the top 15% patches, again with at least 5 patches. Thus, IoU-family metrics use a shared conversion rule and are less dependent on whether a model natively emits boxes.

B Metric Definitions and Evaluation Protocols

B.1 Option-Sheet Overlap

Let B and T denote the base and test splits. For each question q, let $\mathcal{S}(q)$ be the order-invariant set of canonicalized option strings. We define

$$\mathcal{S}_B = \{\mathcal{S}(q) \mid q \in B\}, \quad \mathcal{S}_T = \{\mathcal{S}(q) \mid q \in T\}.$$

Let $\mathcal{O}_B, \mathcal{O}_T$ be the sets of unique options, and $\mathcal{A}_B, \mathcal{A}_T$ the sets of unique correct answers. We report sheet-match rate (SMR), option-overlap rate (OOR), and answer-overlap rate (AOR):

$$\text{SMR} = \frac{1}{|T|} \sum_{q \in T} \mathbf{1}[\mathcal{S}(q) \in \mathcal{S}_B], \tag{12}$$

$$\text{OOR} = \frac{|\mathcal{O}_T \cap \mathcal{O}_B|}{|\mathcal{O}_T|}, \qquad \text{AOR} = \frac{|\mathcal{A}_T \cap \mathcal{A}_B|}{|\mathcal{A}_T|}. \tag{13}$$

B.2 Residual-Geometry Diagnostics

Spread and Collapse Ratio. For each defect family k, let $\mathcal{D}_k$ be the set of evaluation samples in that family, and let z_i be the evidence representation used by a readout: global image features for global prototypes, residual features for residual prototypes, and the pooled quotient residual $\bar{r}_i$ for our diagnostic.

We measure within-family variation by average pairwise cosine distance:

$$\text{Spread} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|\mathcal{D}_k|(|\mathcal{D}_k|-1)} \sum_{\substack{i,j\in\mathcal{D}_k \\ i\neq j}} (1-\cos(z_i, z_j)). \tag{14}$$

To measure whether a point prototype can cover this variation, let μ_k be the prototype of family k, and let $k_i^- = \arg\min_{\ell\neq y_i} d(z_i, \mu_\ell)$ be the nearest negative prototype, where $d(a,b) = 1-\cos(a,b)$. The collapse ratio is

$$\text{Collapse} = \frac{1}{|\mathcal{D}|}\sum_{i\in\mathcal{D}} \min\left(1,\ \frac{d(z_i, \mu_{y_i})}{d(\mu_{y_i}, \mu_{k_i^-}) + \epsilon}\right). \tag{15}$$

Higher values indicate larger within-family deviation relative to prototype separation. We therefore interpret Spread and Collapse jointly: low Spread can reflect hidden variation, while high Collapse indicates that a single centroid poorly organizes the observed evidence.

Description–classification gap. We report the MMAD description–classification gap:

$$\Delta_{\text{Desc}-\text{Cls}} = \text{Acc}_{\text{Desc}} - \text{Acc}_{\text{Cls}}. \tag{16}$$

This diagnostic measures whether a readout remains effective when candidate options require description-level evidence rather than only coarse defect identity.

Pairwise rescue and margin diagnostics. We compare Residual+Prototype and Ours on the same paired valid defect rows, excluding normal/no-defect rows. Let $s_{i,m}^a$ be the score assigned by readout a to option m, y_i the ground-truth option, and $\hat{y}_i^a = \arg\max_m s_{i,m}^a$. The option margin is

$$\gamma_i^a = s_{i,y_i}^a - \max_{m\neq y_i} s_{i,m}^a.$$

Using Residual+Prototype as the reference a and Ours as the target b, we define

$$\mathcal{R}_{a\to b} = \{i : \hat{y}_i^a \neq y_i,\ \hat{y}_i^b = y_i\}, \qquad \mathcal{V}_{a\to b} = \{i : \hat{y}_i^a = y_i,\ \hat{y}_i^b \neq y_i\}.$$

We report the rescue rate

$$\text{Rescue}_{a\to b} = \frac{|\mathcal{R}_{a\to b}|}{|\{i : \hat{y}_i^a \neq y_i\}|},$$

the reverse rate

$$\text{Reverse}_{a\to b} = \frac{|\mathcal{V}_{a\to b}|}{|\{i : \hat{y}_i^a = y_i\}|},$$

and the rescue-only margin change

$$\Delta\gamma_{a\to b} = \frac{1}{|\mathcal{R}_{a\to b}|}\sum_{i\in\mathcal{R}_{a\to b}} (\gamma_i^b - \gamma_i^a).$$

Thus, Rescue measures how often the subspace readout corrects residual-prototype failures, Reverse measures lost residual-prototype successes, and $\Delta\gamma$ is interpreted only within the rescued subset.

B.3 Normal-Reference Stability

For this diagnostic, each query is evaluated with multiple object-matched normal references from the available reference set. The option sheet and model checkpoint are fixed. Let $s_m^{(r)}$ be the score of option m under reference r, and let y be the correct option. We define the normalized margin as

$$\widetilde{\Delta}^{(r)} = \frac{s_y^{(r)} - \max_{m\neq y} s_m^{(r)}}{\text{Std}_m\left(s_m^{(r)}\right) + \epsilon}. \tag{17}$$

The reported margin instability is

$$\text{Norm. Margin Std.} = \frac{1}{|\mathcal{D}|} \sum_{q \in \mathcal{D}} \text{Std}_r \left(\widetilde{\Delta}_q^{(r)} \right). \tag{18}$$

Lower values indicate less sensitivity to the choice of matched normal reference. Accuracy deltas in Fig. 5 are computed relative to the no-projection baseline under the same backbone, support data, scoring rule, and reference set.

B.4 Answer-Grounded Localization and Faithfulness

All metrics in Table 14 are computed from the evidence map associated with the selected answer. For these diagnostics, we average over defect samples with valid ground-truth regions on which the original prediction is correct, since the evidence map is answer-conditioned.

Let a_i denote the nonnegative patch evidence after per-sample shifting and normalization, let $g_i \in [0, 1]$ be the ground-truth region weight at patch i, and let $r_i \in \{0, 1\}$ indicate the top-evidence patch set selected under the fixed area budget. We report

$$\text{Mass@GT} = \sum_i a_i g_i, \qquad \text{GT-Mass@ROI} = \frac{\sum_i a_i g_i r_i}{\sum_i a_i r_i + \epsilon}, \tag{19}$$

and the evidence concentration ratio

$$\text{ECR} = \frac{(\sum_i a_i g_i) \,/\, (\sum_i g_i + \epsilon)}{(\sum_i a_i (1 - g_i)) \,/\, (\sum_i (1 - g_i) + \epsilon) + \epsilon}. \tag{20}$$

These metrics measure how the selected-answer evidence overlaps the annotated region, but the annotation is used only for evaluation.

For intervention-based metrics, ROI-only inputs retain the query evidence inside the selected top-evidence region and replace the complement with the matched normal reference. BG-only inputs apply the complementary intervention. We rerank the original option sheet and report

$$\text{ROI-BG Gap} = \text{Acc}_{\text{ROI-only}} - \text{Acc}_{\text{BG-only}}. \tag{21}$$

Deletion and insertion measure the effect of the same evidence region on the selected-answer margin. Let $\hat{o}$ be the option selected on the original query, fixed during the intervention, and define

$$M(x) = s_{\hat{o}}(x) - \max_{m \neq \hat{o}} s_m(x). \tag{22}$$

Deletion replaces the selected top-evidence patches with the matched normal reference, while insertion starts from the matched normal reference and inserts the same query patches:

$$\text{Deletion} = M(x_q) - M(x_q^{\text{del}}), \qquad \text{Insertion} = M(x_q^{\text{ins}}) - M(x_r). \tag{23}$$

Larger values indicate that the highlighted region has a stronger effect on the selected-answer margin.

C Additional Readout and Shortcut Controls

C.1 Matched-Support Readout Isolation

Discrimination and reference stability. The controlled comparison in Table 5 holds the backbone, abnormal supports, normal-reference assignment, and evaluation protocol fixed; trainable variants use the same optimizer and option-ranking loss. With a linear readout, quotienting improves AnomalyCoT accuracy from 35.51 to 37.67, while MMAD changes little (58.06 to 58.08). The larger MMAD gain comes from shared–private modeling, which reaches 68.88 without the orthogonal constraint and 69.80 with the full readout. These results distinguish the primary discriminative contribution of the subspace readout from the stability role of normal calibration. In the matched-reference diagnostic of Fig. 5b, empirical U_0 reduces normalized margin variability from 4.2141 without projection to 2.1528, whereas a random same-rank basis gives 4.6055. The corresponding accuracy gains over no projection are 3.86 points on AnomalyCoT and 1.48 on MMAD (Fig. 5e). Together, the controls support preserving defect-family residual structure while suppressing variation induced by the normal reference.

Table 5: All variants share the same backbone, support data, normal-reference assignment, and evaluation protocol; trainable variants use the same optimizer and option-ranking loss.

Readout	AnomalyCoT Acc. ↑	MMAD Avg. ↑
Raw residual prototype	19.59	35.21
Raw residual linear	35.51	58.06
Quotient prototype	20.62	35.24
Quotient linear	37.67	58.08
Shared-private subspace w/o constraint	38.47	68.88
Full readout	41.19	69.80

Table 6: Closed-set and text-prior controls on known-answer and novel-answer subsets. Known answers are test ground-truth answer strings observed in the train/support answer vocabulary.

Method	Known Acc. ↑	Novel Acc. ↑	All Acc. ↑	Known–Novel Gap ↓
Answer-frequency prior	66.48	15.68	26.46	50.80
Text-only option prior	30.77	17.75	20.51	13.02
Fixed-label classifier	34.07	20.56	23.43	13.51
Global CLIP ranking	38.46	29.29	31.24	9.17
Ours	71.98	75.30	74.59	**−3.32**

C.2 Closed-Set and Text-Prior Controls

Table 6 evaluates whether the option interface reduces the task to closed-set memorization or text-prior matching. Answer-frequency, text-only, fixed-label, and global CLIP controls provide diagnostic baselines for non-visual or weakly visual shortcuts. The gap between these controls and our readout supports the claim that structured inspection QA requires image-conditioned evidence, rather than only answer-string memorization or text-prior matching.

C.3 Generalization to Held-Out Defect Families

We jointly exclude scratch, crack, logical anomaly, and solder residue from abnormal support fitting and remove their corresponding private subspaces. All-family fitting and both holdout variants are evaluated on the same four family-specific test subsets, with the candidate hypotheses retained at inference. This tests transfer beyond the unseen-answer-string analysis: the held-out family is absent from support fitting, but its hypothesis remains in the test candidate sheet.

Table 7 reports per-family results and equally weighted family-macro averages. The full holdout readout retains 46.07 AnomalyCoT accuracy, 0.1702 IoU, and 68.24 MMAD, compared with 57.14, 0.1860, and 79.03 under all-family fitting. Removing the shared subspace further reduces AnomalyCoT and MMAD to 43.12 and 65.47, respectively. The shared component improves both accuracies for every held-out family, supporting partial transfer of defect evidence without family-specific supports or private bases. The nearly unchanged IoU between the full and w/o-shared variants (0.1702 versus 0.1698) indicates that the main shared-component gain in this test is in answer selection. The gap to all-family fitting also quantifies the remaining dependence on family-specific support.

D Additional Comparisons

D.1 Native CLIP-based Methods for Inspection QA

Table 8 compares anomaly-specialized CLIP variants under the inspection-QA interface. These methods are originally designed for native IAD outputs, such as image-level anomaly scores or patch-level anomaly heatmaps, rather than for ranking dynamic textual hypotheses. We include them because this mismatch is central to our thesis: scalar or heatmap-based anomaly readouts can

Table 7: Joint four-family holdout. The same family-specific test subsets are evaluated with all-family support, four-family holdout, and holdout without the shared subspace. The final rows report equally weighted family-macro averages.

Held-out family	Training setting	AnomalyCoT Acc. (%)	IoU	MMAD Avg. (%)
Scratch	All-family support	60.75	0.1837	80.29
Scratch	Four-family holdout (full)	57.81	0.1508	74.66
Scratch	Four-family holdout w/o shared	52.30	0.1500	73.15
Crack	All-family support	54.87	0.2252	73.75
Crack	Four-family holdout (full)	37.02	0.2263	68.75
Crack	Four-family holdout w/o shared	36.14	0.2260	62.50
Logical anomaly	All-family support	83.78	0.2378	86.02
Logical anomaly	Four-family holdout (full)	74.19	0.2388	73.09
Logical anomaly	Four-family holdout w/o shared	70.65	0.2382	72.90
Solder residue	All-family support	29.17	0.0974	76.04
Solder residue	Four-family holdout (full)	15.26	0.0649	56.46
Solder residue	Four-family holdout w/o shared	13.39	0.0651	53.33
Family Macro Avg.	All-family support	57.14	0.1860	79.03
Family Macro Avg.	Four-family holdout (full)	46.07	0.1702	68.24
Family Macro Avg.	Four-family holdout w/o shared	43.12	0.1698	65.47

Table 8: Native-condition comparison with CLIP-based IAD methods. Each baseline is evaluated with its intended backbone, input resolution, and dense anomaly-map pipeline, while the final interface is standardized to multiple-choice inspection QA through the same fixed, training-free native-evidence adapter. Results are diagnostic rather than a controlled backbone, resolution, or training-data ablation. A same-readout ViT-L/14 check is reported in Table 18.

Method	**Backbone + Img. Size**	**AnomalyCoT**				**MMAD**				
		Acc	IoU	GIoU	DIoU	Det.	Cls.	Loc.	Desc.	Anal.
Ours	ViT-B/16@240 + 240 input	41.19	0.1742	0.1625	0.1363	68.98	69.79	61.88	74.59	73.76
WinCLIP	ViT-B/16@240 + 240 input	14.82	0.1867	0.1430	0.1362	52.93	20.73	39.28	26.69	32.80
AnomalyCLIP	ViT-L/14@336 + 518 input	28.62	0.1542	0.1539	0.1231	60.80	25.35	20.92	28.44	25.78
AdaCLIP	ViT-L/14@336 + 518 input	24.49	0.1560	0.1529	0.1238	55.86	22.15	32.86	25.41	24.97

identify abnormal evidence, but they do not directly provide the structured option readout needed for fine-grained inspection QA.

We evaluate each method in its native operating condition and standardize only the final multiple-choice interface. This is important because CLIP-based IAD methods often rely on method-specific input resizing, dense feature extraction, and heatmap post-processing to obtain reliable anomaly maps; forcing a single resolution or preprocessing pipeline would not be a clean controlled comparison and could distort their intended behavior. We therefore preserve each method's native anomaly extraction pipeline and apply the same fixed, training-free option-ranking adapter.

For each candidate option o_m, the local evidence term is anomaly-weighted text matching,

$$s_{\text{local}}(o_m \mid x) = \frac{\sum_i A_i(x)\, \cos(\phi_i(x), t_m)}{\sum_i A_i(x) + \epsilon}, \tag{24}$$

where $A_i(x)$ is the native anomaly evidence, $\phi_i(x)$ is the corresponding local visual feature, and t_m is the option embedding from the same CLIP text encoder. For defect options, we combine this local term with the native global image–text score and image-level anomaly presence using fixed coefficients $(1.0, 0.25, 0.5)$ for all datasets and baselines. These coefficients are fixed before evaluation and are not tuned on the test split. The goal is not to optimize these methods for inspection QA, but to provide a diagnostic bridge of their native outputs to the shared option-ranking interface.

For no-defect options, we use the method's native image-level normality score when available, rather than introducing a learned null classifier.

This adapter is meant to provide a direct native-evidence lift to option ranking: it uses the baseline's own anomaly map, local features, text encoder, and normality score, adds no supervision, and uses the same rule across all inspected CLIP-based IAD methods. The comparison therefore asks a favorable but diagnostic question: even when anomaly-specialized CLIP methods are evaluated under their intended conditions, can their native evidence support structured inspection QA? Since public IAD weights are not trained under our inspection-QA train/test protocol, the table is not a controlled training-data leaderboard. Nevertheless, their limited transfer to dynamic option ranking is consistent with our main claim that fine-grained inspection QA benefits from a structured evidence readout beyond scalar anomaly scores or heatmap-style localization.

D.2 MLLM Prompting and Decoding Protocols

Evaluation protocols. We evaluate Qwen2.5-VL, InternVL2, and Gemma 3 under four protocols. *Letter likelihood* ranks option letters using the fixed multiple-choice prompt, specified before evaluation without validation- or test-based prompt selection. *Basic generation* follows the benchmark-specific generation setting and emits an answer and, for AnomalyCoT, box coordinates without an explicit rationale. *CoT generation* adds the benchmark-provided reasoning instruction, while *localization generation* emphasizes grounding the selected hypothesis before answer and coordinate generation. Generated answers are parsed with the benchmark procedure, with letter-likelihood fallback for unparsable outputs. The reported generation conditions therefore include this fallback; generated-box IoU is evaluated only on AnomalyCoT.

Table 9: MLLM evaluation under letter likelihood and three generation protocols. GPT-4o is a benchmark-reported reference.

Configuration		AnomalyCoT		MMAD (%)					
Model	Protocol	Acc. (%)	IoU	Det.	Cls.	Loc.	Desc.	Anal.	Avg.
Qwen2.5-VL	Letter likelihood	47.27	0.1426	67.36	58.77	61.37	66.55	77.72	66.35
Qwen2.5-VL	Basic generation	45.53	0.0265	68.34	58.32	61.11	65.14	81.77	66.94
Qwen2.5-VL	CoT generation	50.07	0.0041	54.21	56.18	57.19	66.08	75.52	61.84
Qwen2.5-VL	Localization generation	48.64	0.0592	58.15	62.74	62.00	72.59	80.01	67.10
InternVL2	Letter likelihood	39.10	0.1450	47.69	43.36	48.52	56.64	76.06	54.45
InternVL2	Basic generation	37.54	0.0081	59.97	43.85	47.91	57.60	78.10	57.49
InternVL2	CoT generation	41.38	0.0278	54.12	44.85	52.47	66.29	76.15	58.78
InternVL2	Localization generation	26.71	0.0272	56.43	46.19	52.88	58.64	76.32	58.09
Gemma 3	Letter likelihood	32.17	0.1420	39.58	38.39	28.50	53.03	72.61	46.42
Gemma 3	Basic generation	22.15	0.0010	46.25	37.50	27.74	51.75	62.92	45.23
Gemma 3	CoT generation	20.45	0.0128	45.31	38.74	32.56	62.18	76.91	51.14
Gemma 3	Localization generation	21.16	0.0201	46.82	36.25	34.69	48.14	74.47	48.07
Ours	Residual readout	41.19	0.1742	68.98	69.79	61.88	74.59	73.76	69.80
GPT-4o	Benchmark-reported	49.19	0.0364	68.63	65.80	55.62	73.21	83.41	69.33

Generation accuracy includes letter-likelihood fallback for unparsable outputs. Letter-likelihood IoU is attention-derived; generation IoU uses generated boxes. GPT-4o is benchmark-reported. Its MMAD average is the mean of the five reported task accuracies.

Task-dependent effects. Table 9 shows that generation materially changes individual MLLM results. Qwen2.5-VL reaches 50.07 AnomalyCoT accuracy with CoT and 67.10 MMAD with localization-focused generation; the best tested generation MMAD averages for InternVL2 and Gemma 3 are 58.78 and 51.14. Our 69.80 MMAD average remains competitive across these settings, while Qwen2.5-VL is stronger on AnomalyCoT accuracy and MLLMs retain advantages on analysis. The efficiency measurements in Table 3a apply to the original option-ranking configuration; they are not timings of these additional generation runs.

Interpretation of sensitivity and grounding. The protocols vary prompting and decoding jointly, so they do not measure variation across several prompt templates within one fixed decoding mode. Letter-likelihood IoU uses an attention-derived map-to-box proxy, whereas generation IoU evaluates emitted coordinates. We therefore report each interface separately and avoid averaging these IoUs into a homogeneous localization-variance estimate.

D.3 Abnormal-Support In-Context Control

Abnormal-support control. To test the effect of providing explicit defect examples to MLLMs, we construct an object-specific bank with one normal image and one abnormal training image from each available family. The same bank is supplied to every test query of that object regardless of its true defect family, providing approximately 4.1 abnormal images per object. Table 10 shows useful gains: InternVL2 improves from 39.10 to 43.72 on AnomalyCoT and from 54.45 to 60.46 on MMAD, while Qwen2.5-VL reaches 54.98 AnomalyCoT accuracy. Our 69.80 MMAD average remains higher under this control, so providing abnormal demonstrations does not eliminate the observed MMAD advantage. The control addresses support access; online ICL and offline fitting still use different sample budgets and adaptation mechanisms.

Table 10: Object-matched abnormal-support ICL control. Each object-specific support bank contains one normal image and one training-split abnormal image from every available defect family.

Model and adaptation			AnomalyCoT		MMAD (%)					
Model	Scale	Adaptation / support protocol	Acc. (%)	IoU	Det.	Cls.	Loc.	Desc.	Anal.	Avg.
InternVL2	8B	4 normal ICL examples	39.10	0.1450	47.69	43.36	48.52	56.64	76.06	54.45
InternVL2	8B	Object-matched all-family 1-shot ICL	43.72	0.0643	62.35	54.98	46.18	60.47	78.34	60.46
Qwen2.5-VL	7B	4 normal ICL examples	47.27	0.1426	67.36	58.77	61.37	66.55	77.72	66.35
Qwen2.5-VL	7B	Object-matched all-family 1-shot ICL	54.98	0.0651	48.58	55.92	55.90	69.30	70.97	60.13
Gemma 3	4B	4 normal ICL examples	32.17	0.1420	39.58	38.39	28.50	53.03	72.61	46.42
Gemma 3	4B	Object-matched all-family 1-shot ICL	34.91	0.0648	47.53	24.64	23.07	42.32	67.27	40.97
Ours	0.15B	4 defect supports/family, offline fitting	41.19	0.1742	68.98	69.79	61.88	74.59	73.76	69.80

The same support bank is supplied to every test query of the object, regardless of its true defect family, and contains approximately 4.1 abnormal images per object. Normal/abnormal online ICL and offline readout fitting are distinct adaptation protocols.

D.4 Repeated Runs and Paired Statistical Comparisons

Random-seed robustness. We repeat the main 4-shot readout over three seeds, keeping the data split, support set, normal-reference selection, architecture, optimizer, and hyperparameters fixed.

Sources of run-to-run variation. Table 11 additionally reports three runs for the locally evaluated MLLMs and our readout. For MLLMs, the seed changes the normal ICL exemplars while checkpoints, prompt, decoding configuration, and test queries remain fixed. For Ours, it changes readout training while the support split, inference references, architecture, and hyperparameters remain fixed. The standard deviations therefore measure different sources of variation; they do not quantify prompt-template sensitivity or stochastic decoding.

Paired comparisons under the original ranking protocol. Table 12 reports paired bootstrap differences between our readout and MLLMs evaluated with four normal ICL examples under letter likelihood. For MMAD, samples are resampled within each subtask and the five-task macro-average is recomputed in each replicate. The comparison with Qwen2.5-VL gives -6.08 AnomalyCoT points (95% CI: $[-7.38, -4.78]$) and $+3.45$ MMAD points ($[2.10, 4.80]$), supporting the benchmark-dependent interpretation. These intervals are conditional on the evaluated predictions and do not quantify uncertainty in the additional generation or abnormal-ICL comparisons.

Table 11: Mean ± standard deviation over three runs. MLLMs vary normal-ICL exemplar selection; Ours varies readout-training seed. Accuracy and MMAD Avg. are percentages; IoU is on the raw scale.

Model	AnomalyCoT Acc.	IoU	MMAD Avg.
Qwen2.5-VL	48.67 ± 1.25	0.1429 ± 0.0030	65.93 ± 1.12
InternVL2	38.00 ± 2.16	0.1441 ± 0.0050	54.87 ± 0.25
Gemma 3	30.33 ± 2.49	0.1406 ± 0.0010	47.13 ± 0.96
Ours	40.94 ± 2.07	0.1718 ± 0.0070	69.70 ± 0.14

Table 12: Paired bootstrap comparisons on the same evaluated predictions, with differences defined as Ours minus Baseline and 95% confidence intervals.

Comparison	Δ AnomalyCoT Acc. (pp)	95% CI	Δ MMAD Avg. (pp)	95% CI
Ours - Qwen2.5-VL	-6.08	$[-7.38, -4.78]$	$+3.45$	$[+2.10, +4.80]$
Ours - InternVL2	$+2.09$	$[+0.84, +3.34]$	$+15.35$	$[+13.75, +16.95]$
Ours - Gemma 3	$+9.02$	$[+7.82, +10.22]$	$+23.38$	$[+21.78, +24.98]$

MMAD samples are resampled within each of its five subtasks, and the macro-average is recomputed for each replicate. pp denotes percentage points.

E Additional Diagnostics and Robustness

E.1 Prototype Collapse Diagnostics

Table 13: Prototype diagnostics for fine-grained inspection QA. We compare global prototypes, residual prototypes, and our subspace readout in terms of intra-family variation, option-level accuracy, and cases where subspace readout corrects residual-prototype errors. B3 rescue metrics are defined only for Ours relative to the residual point prototype; N/A indicates rows for which this rescue baseline is not defined.

	B1: Intra-family Geometry		**B2: QA Accuracy**		**B3: Rescue Regime**		
Readout Strategy	Spread	Collapse Ratio	Cls Acc ↑	Desc Acc ↑	Rescue % ↑	Reverse % ↓	Δ Margin ↑
Global Point Prototype	0.1521	0.1503	33.18	24.83	N/A	N/A	N/A
Residual Point Prototype	0.7145	0.7927	40.99	32.51	N/A	N/A	N/A
Ours (Subspace)	0.6816	0.3071	69.79	74.59	57.63	19.09	1.9148

Table 13 provides additional diagnostics for the prototype-collapse analysis. Point prototypes summarize each defect family by a single centroid, which can either hide residual variation or expose variation without organizing it. In contrast, the shared-private subspace readout preserves within-family residual spread while constraining the directions used for scoring. These diagnostics support the main claim that fine-grained inspection QA requires reading residual evidence as structured subspace variation rather than isolated prototype distance.

E.2 Answer-grounded Localization and Faithfulness

We evaluate answer-grounded faithfulness operationally: selected-answer evidence should overlap annotated regions, and highlighted patches should affect the option-ranking margin under deletion/insertion. Table 14 reports spatial-overlap and margin-sensitivity diagnostics for the evidence map associated with the selected answer. All metrics are computed from this selected-answer evidence map. ROI-only and BG-only tests replace regions outside the retained subset with the matched normal reference, while deletion and insertion measure score-margin changes after removing or inserting area-matched evidence patches. These evaluations complement box IoU by testing whether the highlighted regions affect the option-ranking decision.

Table 14: Answer-grounded localization and faithfulness. Metrics are computed from the evidence map associated with the selected answer, over correctly answered defect samples with valid ground-truth regions. ROI-only and BG-only intervene using the selected top-evidence region and its complement. Deletion/Insertion measure selected-answer margin changes under area-matched evidence removal/insertion.

Readout	ROI-only Acc. ↑	BG-only Acc.	ROI-BG Acc. Gap ↑	GT-Mass@ROI ↑	Mass@GT ↑	ECR ↑	Deletion ↑	Insertion ↑
Global CLIP	19.48	23.98	-4.50	0.120	0.161	0.729	0.077	0.131
Residual prototype	17.55	28.71	-11.16	0.202	0.211	0.799	0.001	0.000
Ours	53.97	49.19	4.78	0.290	0.238	1.257	0.216	0.266

E.3 Null-Option Evaluation

The relevant hypothesis must still be supplied in the option sheet; the readout cannot autonomously name a missing defect. Under the fixed main decision rule, treating null as positive gives precision/recall/F1 of $78.85/73.69/76.18\%$, without test-time threshold tuning. These metrics characterize the benchmark null branch. They do not establish rejection reliability when the correct hypothesis is deliberately removed from the candidate set.

E.4 Normal-Reference Selection and Robustness to Mismatch

The reference selection and normal-calibration protocols are specified in Appendix A.2.

Alternative references. With the model and queries fixed, five alternative reference assignments yield 38.20 ± 0.99 AnomalyCoT accuracy, 0.1778 ± 0.0016 raw IoU, and 70.48 ± 1.05 MMAD (Table 15). The default gives $41.19/0.1742/69.80$: its AnomalyCoT accuracy is higher, while its IoU and MMAD are lower than the alternative-reference means. Thus, the default is not uniformly favorable, although AnomalyCoT exhibits greater reference sensitivity.

Table 15: Sensitivity to normal-reference selection over five category-matched reference assignments (mean ± standard deviation).

Reference setting	AnomalyCoT Acc. (%)	IoU (%)	Det.	Cls.	Loc.	Desc.	Anal.	MMAD Avg.
Random category-matched normal reference (5 seeds)	38.20 ± 0.99	17.78 ± 0.16	69.20 ± 1.60	73.20 ± 2.23	62.80 ± 2.14	74.40 ± 1.02	72.80 ± 1.47	70.48 ± 1.05

Accuracies and IoU are reported as percentages. The reference assignment varies while the model and evaluation queries remain fixed.

Reference-only perturbations. Table 16 perturbs only the normal reference at inference. Translations up to 28 pixels, rotations up to $10°$, and brightness increases up to 20% cause maximum losses of 0.0064 raw IoU and 3.20 MMAD points. Brightness increases cause at most a 0.0015 IoU loss and do not reduce MMAD in the tested settings. Both reference tables report IoU on the percentage scale ($100 \times$ IoU); the values in the text use raw IoU.

Scope of robustness. These tests cover moderate within-category variation. Severe layout mismatch remains a limitation: in LOCO-style pushpins scenes, valid changes in component positions can make same-index subtraction compare different components or occupied and empty compartments, producing diffuse responses around normal structure. The quotient suppresses variation represented in the calibrated normal subspace; it does not perform geometric registration.

E.5 Sensitivity to the Subspace Ranks

We reduce each rank separately from the default $(r_s, r_p, r_0) = (4, 8, 4)$ and additionally double all three ranks. As shown in Table 17, AnomalyCoT accuracy ranges from 40.49 to 42.00, raw IoU from 0.1737 to 0.1772, and MMAD from 68.98 to 70.44. The largest absolute changes from the default are 0.81 accuracy points, 0.0030 IoU, and 0.82 MMAD points. The shallow, non-monotonic profile supports robustness within the tested capacity range. The default ranks are retained across the main experiments and are not optimal on any individual metric in this analysis.

Table 16: Robustness to perturbations applied only to the normal reference at inference. Changes from the clean reference are shown in parentheses.

Reference condition	IoU (%) (Δ from clean)	MMAD Avg. (%) (Δ from clean)
Clean	$17.42(+0.00)$	$69.80(+0.00)$
Translation +x (14 px)	$16.78(-0.64)$	$68.20(-1.60)$
Translation +x (28 px)	$16.83(-0.59)$	$66.60(-3.20)$
Rotation (+5°)	$17.09(-0.33)$	$68.20(-1.60)$
Rotation (+10°)	$16.92(-0.50)$	$68.40(-1.40)$
Brightness (+10%)	$17.36(-0.06)$	$70.40(+0.60)$
Brightness (+20%)	$17.27(-0.15)$	$70.00(+0.20)$

IoU follows the percentage scale in the original table: 17.42 corresponds to raw IoU 0.1742. Changes are percentage points.

Table 17: Sensitivity to shared, private, and normal-variation subspace ranks. Each rank is reduced separately, and all three are additionally doubled.

Setting	(r_s, r_p, r_0)	AnomalyCoT Acc. (%)	IoU	MMAD Avg. (%)
Default	(4,8,4)	41.19	0.1742	69.80
Lower shared rank	(2,8,4)	41.79	0.1769	69.63
Lower private rank	(4,4,4)	40.64	0.1772	70.44
Lower normal rank	(4,8,2)	40.49	0.1747	68.98
All ranks doubled	(8,16,8)	42.00	0.1737	69.63

F Backbone Scaling Check

Table 18 evaluates the same residual-readout protocol with an additional CLIP backbone. This check tests whether the proposed readout is tied to OpenCLIP ViT-B/16-plus-240 or reflects a more general property of normal-referenced residual evidence.

Table 18: Backbone generalization of the proposed readout. We replace the frozen CLIP backbone while keeping the same option-ranking protocol and training setting.

Method	**Backbone**	**AnomalyCoT**				**MMAD**				
		Acc	IoU	GIoU	DIoU	Det.	Cls.	Loc.	Desc.	Anal.
Ours	ViT-B/16	41.19	0.1742	0.1625	0.1363	68.98	69.79	61.88	74.59	73.76
Ours	ViT-L/14	43.98	0.1757	0.1608	0.1364	69.06	71.80	61.93	76.46	77.68