OCP-R: STABILIZING OPTION-LEVEL READOUT FOR VLM ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

VLM anomaly-detection benchmarks such as MMAD and AnomalyCoT test industrial inspection with multiple-choice questions, matching the closed defect vocabularies of industrial quality assurance, so a frozen VLM must turn its hidden state into answer-option scores, a step we call the option-level readout. In this paper, we show this step is unreliable: frozen VLMs often hold anomaly evidence that their option logits do not expose. Two measurements motivate the repair. Weak frozen readouts spend most of their MCQ supervision on answer-format rather than answer tokens, which points at the readout, not at capacity; and a replacement readout discards decisions the frozen head already scores correctly, so the correction should add to the frozen logits rather than replace them. We propose Option-Conditional Residual Readout (OCP-R), a small trained head that scores each candidate option against the frozen state and adds the result to the frozen logits, zero at initialization so training departs from the frozen readout only where supervision demands. Across nine general-purpose and four anomaly-specialized VLMs, OCP-R raises mean MMAD accuracy on seven general-purpose backbones from 62.17% to 78.27%, preserving 93% of Qwen2.5-VL's correct predictions, and on AnomalyCoT it is the only method with no catastrophic held-out drop on any of the nine backbones, against one for internal-weight adaptation and three for a replacement readout.

1 INTRODUCTION

Vision-language models (VLMs) (Hurst et al., 2024; Liu et al., 2023; Awadalla et al., 2023) are increasingly asked to perform industrial anomaly detection (AD): not only deciding whether a product image is normal or defective, but identifying the defect type, location, and likely cause (Roth et al., 2022; Defard et al., 2021; Zavrtanik et al., 2021; Gudovskiy et al., 2022; Yu et al., 2021). Benchmarks such as MMAD (Jiang et al., 2025) and AnomalyCoT (Cheng et al., 2025) evaluate these abilities with multiple-choice questions (MCQs) (Liu et al., 2024c; Lu et al., 2024; Li et al., 2023a): the model sees an image, an inspection question, and several candidate answers, and must score answer symbols such as A, B, C, and D. This format matches how inspection is already organized in practice, since industrial quality assurance maintains closed defect vocabularies, including defect codes, work-instruction checklists, and disposition categories, so the candidate set an MCQ requires is typically the taxonomy an inspection line already uses. The format does presume such a set exists; open-ended description and pixel-level localization fall outside its scope, and outside ours. Our subject is the final step this interface forces on every VLM: turning an internal visual-language representation into answer-option scores, a step we call the option-level readout.

Our starting observation is that this final step fails independently of the representation behind it. A frozen VLM can encode the visual evidence of a defect and still assign unstable or biased scores to the answer options (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Zhao et al., 2021; Atabuzzaman et al., 2025). The standard response, internal-weight adaptation (Hu et al., 2022; Dettmers et al., 2023; Houlsby et al., 2019; Lester et al., 2021; Li & Liang, 2021), raises accuracy but modifies the model everywhere, including components that were not broken. If the failure is confined to the final option scores, the proportionate fix is to keep the VLM frozen and repair only the readout. Can option logits be stabilized without replacing the LM head or touching any internal weight?

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.

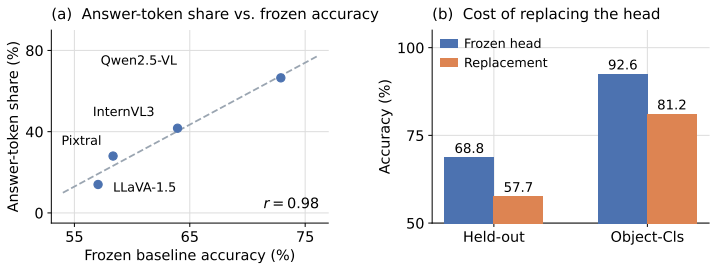


Figure 1: Two measurements motivate the form of the correction. (a) Answer-token supervision share tracks frozen baseline accuracy ($r = 0.98$, $N = 4$ backbones; we read this as a consistent trend rather than an estimated effect size): a weak readout spends its supervision on answer format rather than on the answer itself. (b) On Qwen2.5-VL a replacement readout lowers held-out and Object-Classification accuracy, discarding decisions the frozen head already scores correctly.

We begin by measuring how much the frozen readout leaves behind. Layer-wise probing (Alain & Bengio, 2016; Kim et al., 2018; Belrose et al., 2023; Meng et al., 2022) shows that a readout added to the frozen option logits recovers substantially more anomaly-discriminative accuracy from the final-layer hidden state than those logits report on their own. We call this gap *option-level readout instability*: the answer can be wrong even when the evidence for the right answer is present in the state, because that evidence is not read out into the option scores.

Two further measurements, summarized in Figure 1, point to what the repair must look like and where it must act. First, decomposing MCQ supervision into answer-format and answer tokens (Robinson et al., 2023) shows that the weaker a VLM's frozen readout, the more of its supervision falls on format tokens, from 33.5% on Qwen2.5-VL up to 86.0% on LLaVA-1.5 (details in the Appendix): what is missing is at the readout, not additional internal capacity. Second, replacing the frozen head with a learned option-conditioned readout overshoots. On Qwen2.5-VL it lowers held-out accuracy from 68.85% to 57.67% and Object Classification from 92.62% to 81.21%, discarding decisions the frozen head already scores correctly (Table 4 and Table 9): the repair should add to the frozen logits rather than replace them. The same measurements locate the intervention. Under the letter-verbalizer interface, the frozen readout separates any two options along a direction set by their answer symbols alone, shared by every question that reuses the same letters; making the decision boundary depend on the candidate set requires a term that reads the option texts at the readout. Backbone tuning reaches this only indirectly, by moving the representation rather than re-orienting a readout normal it leaves fixed, displacing decisions the frozen readout already places correctly. Freezing the backbone is therefore not only an efficiency choice but what the diagnosis points to.

We propose **Option-Conditional Residual Readout** (OCP-R), a lightweight trained head that keeps the VLM backbone and LM head frozen and adds an option-conditioned correction to the frozen option logits (He et al., 2016; Wortsman et al., 2022), zero at initialization (Zhang et al., 2023; Goyal et al., 2017; Bachlechner et al., 2021), so an untrained OCP-R is exactly the frozen readout and training departs from it only where supervision demands, unlike a replacement readout, which offers no such starting point. Throughout, *Vanilla OCP* denotes the same probe used as a replacement, without the residual anchor. Every component here is individually known; the contribution is the finding that this combination, not the simpler corrections practitioners reach for first (calibration, unconditional probes, replacement heads), is what the diagnosis points to. Our contributions are summarized as follows.

- We **diagnose and measure an option-level readout instability** in MCQ-style VLM anomaly benchmarks: layer-wise probing shows frozen states carry anomaly evidence their option logits under-expose, and two measurements, on supervision concentration and replacement forgetting, narrow the admissible correction to a single shape.
- We propose **OCP-R**, an identity-initialized option-conditioned residual readout that corrects frozen option logits without touching the backbone or LM head, re-orienting each option-pair decision by an increment the option texts set, anchored at the frozen readout.

- We give a **confound-isolating evaluation** across MMAD and AnomalyCoT with nine general-purpose and four anomaly-specialized VLMs, against matched readout and internal-weight adaptation baselines with controls for parsing, answer-letter priors, and visual grounding.

2 Related Work

VLMs for Industrial Anomaly Detection. Traditional AD models normal appearance and flags deviations (Bergmann et al., 2019; Defard et al., 2021; Zavrtanik et al., 2021; Gudovskiy et al., 2022; Yu et al., 2021; Roth et al., 2022; Zheng et al., 2022; Li et al., 2024b). Recent work moves to Generalist AD (Jeong et al., 2023; Zhu & Pang, 2024; Yao et al., 2024) and VLM-based anomaly understanding (Bai et al., 2025; Li et al., 2024a; Chen et al., 2024b), evaluated by AnomalyCoT (Cheng et al., 2025) and MMAD (Jiang et al., 2025) through MCQ tasks (Liu et al., 2024c; Lu et al., 2024). Existing VLM-AD methods add visual experts, vision-side adapters, or cross-modal alignment (Li et al., 2023b; Gu et al., 2024; Zhang et al., 2024b), or inject external anomaly knowledge and explicit reasoning (Zhang et al., 2026). All reshape how the representation is computed. None examines the final readout step, how frozen hidden states become answer-option logits, where we place the failure.

Probing Frozen Representations. Linear probing asks what a hidden state already carries (Alain & Bengio, 2016; Kim et al., 2018; Tenney et al., 2019), and shows language models store truthfulness, factual, and reasoning-related properties before the output layer (Zou et al., 2023; Marks & Tegmark, 2024; Belrose et al., 2023; Burns et al., 2023). We ask the same question of VLM anomaly detection, then push further: whether the frozen LM head turns that content into reliable option scores (Zheng et al., 2024; Robinson et al., 2023; Zheng et al., 2023). Closest in setting, Atabuzzaman et al. (2025) debias LVLM selection at the logit level and Yang et al. (2025) trace option-symbol bias inside the model; these correct a prior over answer symbols, while our correction is conditioned on each question's option texts.

Adaptation and Residual Readout. PEFT methods including LoRA (Houlsby et al., 2019; Lester et al., 2021; Hu et al., 2022; Dettmers et al., 2023; Chen et al., 2024a; Sheng et al., 2023; Sung et al., 2022; Lialin et al., 2023) add trainable components inside the frozen backbone, changing inference-time computation. On short-answer MCQ this signal can mix visual evidence with answer-format effects (Yang et al., 2025; Wortsman et al., 2022; Kumar et al., 2022; Biderman et al., 2024), so we use internal-weight adaptation as a stress test rather than the only comparison. A readout on frozen features (Radford et al., 2021; Zhang et al., 2022; Gao et al., 2024) is the alternative, and adapters of that family add a term to frozen logits with no per-question candidate set, but a replacement probe discards predictions the frozen head already makes, forgetting under adaptation (Kirkpatrick et al., 2017; Luo et al., 2024; Li & Hoiem, 2017). OCP-R takes a third position: an identity-initialized, option-conditioned residual correction added to the frozen option logits (He et al., 2016; Zhang et al., 2023; Goyal et al., 2017; Mou et al., 2024), keeping frozen predictions.

3 Proposed Method

The two measurements above shape OCP-R's design: the correction reads the option texts and adds to the frozen logits rather than replacing them, starting from an identity initialization that coincides with the frozen readout and never touching the backbone or the LM head.

3.1 Problem Setting: Existing MCQ Anomaly Evaluation

We formalize the option-level readout on the existing MCQ anomaly-evaluation interface of Figure 2: a visual input, an inspection question, and candidate option texts. We consider a dataset $\mathcal{D} = \{(X_j, Q_j, \mathcal{O}_j, y_j)\}_{j=1}^{N}$, where X_j is the visual input, a single query image or a query-reference pair depending on the task, Q_j is the inspection question, $\mathcal{O}_j = \{o_{j,1}, \ldots, o_{j,C_j}\}$ is the set of C_j candidate option texts, and y_j is the correct option index. Let $\mathcal{M}_\theta$ be a pretrained VLM with frozen parameters θ, vocabulary $\mathcal{V}$, input embedding matrix $E_{\text{in}} \in \mathbb{R}^{|\mathcal{V}| \times d}$, final normalization layer $\text{Norm}_{\text{final}}$, and LM head $W_{\text{lm}} \in \mathbb{R}^{|\mathcal{V}| \times d}$. For the i-th option, $v_i \in \mathcal{V}$ is the token index of its answer symbol under a fixed prompt template, typically a letter such as A, B, C, or D. The same verbalizer, tokenizer, and template are used across all option-logit methods.

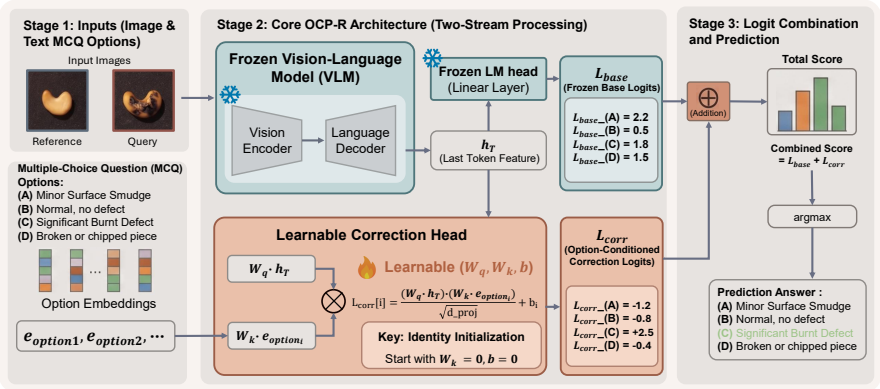


Figure 2: OCP-R overview. Under a forced-prefix option-logit protocol the frozen LM head maps h_T to option logits L_{base}; OCP-R adds an option-conditioned residual L_{corr} from h_T and the candidate option texts, identity-initialized so $L_{\text{corr}} = 0$ at step zero and $s_i = L^{(i)}_{\text{base}} + L^{(i)}_{\text{corr}}$.

3.2 Parser-Free Frozen Option Readout

Free-form generation followed by heuristic parsing can mix visual decision errors with formatting or parsing errors, and we later verify that this protocol change is itself nearly neutral, so the readout is measured on its own. We therefore define a parser-free frozen baseline, the frozen branch of Figure 2, using the same forced-prefix protocol as OCP-R. Given $(X, Q, \mathcal{O})$, we prompt the frozen VLM with the visual input, question, candidate options, and a fixed assistant prefix, e.g., `The answer is (`, without providing the answer token. We extract the next-token hidden state $h_T^{\text{raw}} \in \mathbb{R}^d$ and normalize it as $h_T = \text{Norm}_{\text{final}}(h_T^{\text{raw}})$. For candidate i the frozen option logit is

$$L^{(i)}_{\text{base}} = W_{\text{lm}}[v_i]^\top h_T + b_{\text{lm}}[v_i], \tag{1}$$

where b_{lm} is included only if the LM head has a bias term. All VLM parameters remain frozen, and the same extraction is used during OCP-R training, so the gold answer token is never visible to h_T.

3.3 Option-Conditional Residual Readout

OCP-R keeps the frozen logits of Eq. (1) and adds a small correction for each candidate option, computed from the option text so that it cannot rely only on a global answer-letter bias. For candidate i with token sequence $(t_{i,1}, \ldots, t_{i,S_i})$ we keep at most $T_{\max}$ tokens, $M_i = \min(S_i, T_{\max})$, and mean-pool the VLM's frozen input embeddings, $e_i = M_i^{-1} \sum_{m=1}^{M_i} E_{\text{in}}[t_{i,m}]$, with $T_{\max} = 8$ by default. This introduces no external text encoder and keeps the residual branch tied to the same vocabulary interface as the frozen LM head.

We map the prediction state h_T and the option descriptor e_i into a shared interaction space using $q = W_q \text{LN}_q(h_T)$ and $k_i = W_k \text{LN}_k(e_i)$, where $W_q, W_k \in \mathbb{R}^{d_{\text{proj}} \times d}$ and $d_{\text{proj}} = 64$. Here, LN_q and LN_k are affine layer-normalization modules applied only inside the readout head.

The residual branch scores the compatibility between the current prediction state and each candidate option, then adds this correction to the frozen readout:

$$L^{(i)}_{\text{corr}} = \frac{q^\top k_i}{\sqrt{d_{\text{proj}}}} + b_i, \tag{2}$$

$$s_i = L^{(i)}_{\text{base}} + L^{(i)}_{\text{corr}}. \tag{3}$$

Here, b_i is a small option-slot bias, one scalar per slot over the $C_{\max} = 5$ slots the benchmarks use; because it can absorb answer-position priors, we evaluate bias-only and no-bias variants, and invalid option slots are masked before the softmax. Because the correction enters additively, the softmax over s_i multiplicatively reweights the frozen posterior by $\exp(L^{(i)}_{\text{corr}})$; at initialization the correction vanishes, so the frozen predictions are not behavior to be relearned but a starting point.

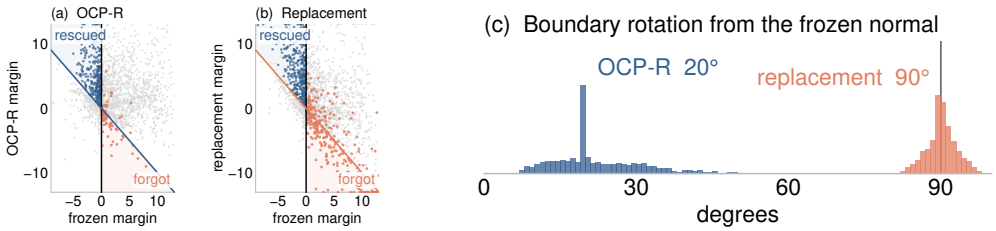


Figure 3: The measured option-pair boundary (Qwen2.5-VL, MMAD, N=1844, final layer): frozen pair margin (x) against the margin the readout adds (y), so the frozen boundary is vertical and the readout's diagonal. Coral marks samples the frozen boundary places correctly and the readout loses; blue the reverse. (a) OCP-R acts almost entirely in the rescue wedge. (b) A replacement readout rescues comparably but forgets nearly four times as many. (c) OCP-R turns the local boundary normal by a median 20° from u_{ij}; a replacement normal is near-orthogonal.

Eqs. (2) and (3) have a direct reading in the space of h_T, measured in Figure 3: options i and j are separated by the sign of $s_i - s_j$, so every option pair defines a decision boundary. Writing $M = W_q^\top W_k / \sqrt{d_{\text{proj}}}$ and $\Delta_{ij} = \text{LN}_k(e_i) - \text{LN}_k(e_j)$,

$$s_i - s_j = u_{ij}^\top h_T + \delta_{ij}^\top \text{LN}_q(h_T) + c_{ij}, \tag{4}$$

with $u_{ij} = W_{\text{lm}}[v_i] - W_{\text{lm}}[v_j]$, $\delta_{ij} = M\Delta_{ij}$, and $c_{ij} = (b_{\text{lm}}[v_i] - b_{\text{lm}}[v_j]) + (b_i - b_j)$ collecting the head biases. The frozen normal u_{ij} carries no information about the option texts, while the added δ_{ij} is a function of those texts alone, so the boundary is rotated by an amount the candidate answers determine, within the column space of M, whose rank is at most d_{proj}. This makes precise the introduction's intuition: since u_{ij} is shared by every question reusing the same letters, only δ_{ij} can make the boundary depend on the candidate set, while backbone adaptation leaves u_{ij} fixed and can act only by moving h_T, at the cost Section 4 measures. Differentiating LN_q gives the added normal as $\sigma_T^{-1} P_T(\gamma_q \odot \delta_{ij})$, with P_T projecting out $\mathbf{1}$ and the normalized state; both are shared by all options, so Eq. (4) is a local linearization whose option-dependent part is δ_{ij}, as Figure 3 measures it (Appendix D). The measured rotations follow the option texts rather than the image: the 386 Anomaly Detection samples share one option set and receive essentially one rotation (median 19.8°, 99% within ±0.5°), with per-task medians from 17.2° to 30.4°. Eq. (4) also names what each readout control we compare against removes: letter-bias calibration changes only c_{ij}, an unconditional probe learns one normal shared by every option pair, and a replacement readout keeps δ_{ij} while discarding u_{ij}, which is why it loses the decisions the frozen arrangement already places correctly, at nearly four times the rate of OCP-R for a comparable number of rescues (Figure 3b).

3.4 Identity Initialization and Training Objective

The residual branch is initialized to leave the frozen readout unchanged: W_k and the option-slot bias vector $\mathbf{b}$ are set to zero, W_q follows the default initialization, and the affine parameters of LN_q and LN_k are $\gamma = \mathbf{1}$, $\beta = \mathbf{0}$. At training step zero this gives $k_i = \mathbf{0}$, $b_i = 0$, and $L^{(i)}_{\text{corr}} = 0$, so $s_i = L^{(i)}_{\text{base}}$ for every candidate and the initial prediction is exactly the frozen option-logit baseline. In the form of Eq. (4) this sets $\delta_{ij} = \mathbf{0}$ for every option pair, so training starts on the frozen arrangement and re-orients a boundary only where supervision demands it. Given the correct option index y, we train only the readout parameters $\phi = \{W_q, W_k, \mathbf{b}, \gamma_q, \beta_q, \gamma_k, \beta_k\}$ with cross-entropy over valid options:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp(s_y)}{\sum_{c=1}^{C} \exp(s_c)}. \tag{5}$$

Identity initialization therefore makes OCP-R a conservative residual readout rather than a replacement classifier: where the frozen logits are confident and correct the correction gradients are small, and where they are uncertain or wrong the residual branch learns a targeted fix. At inference time, OCP-R uses the same forced-prefix protocol as the frozen baseline: after one frozen forward pass the head computes $L^{(i)}_{\text{corr}}$, forms s_i using Eq. (3), and predicts $\hat{y} = \arg\max_i s_i$. Because no backbone weights change, heads can be swapped without reloading or merging the VLM when several MCQ-style tasks share one backbone: each task costs one small head on a shared resident model, rather than one adapted backbone per task.

Table 1: MMAD source-held-out evaluation, GoodsAD held out (aggregate comparison in text); the held-out gain over LoRA holds on 6 of 7 VLMs, InternVL-3 excepted.

VLM	Baseline		LoRA		OCP-R (Ours)	
	Overall	Held	Overall	Held	Overall	Held
Qwen2.5-VL	72.89	67.25	77.77	74.12	**81.62**	**76.04**
Qwen2-VL	71.64	67.89	77.66	74.44	**81.40**	**75.88**
LLaVA-1.5	57.05	61.34	61.28	62.94	**70.99**	**65.34**
Pixtral	58.35	56.87	66.81	62.14	**78.64**	**73.72**
InternVL-3	63.94	60.70	77.55	**74.12**	**80.31**	73.16
InternVL-2.5	69.69	66.29	76.79	71.09	**78.52**	**71.73**
LLaVA-NeXT	41.65	43.13	65.67	68.05	**76.42**	**71.57**
Mean (N=7)	62.17	60.50	71.93	69.56	**78.27**	**72.49**
Mean Δ	-	-	$+9.76$	$+9.06$	$\mathbf{+16.10}$	$\mathbf{+11.99}$

4 EXPERIMENTS

Our experiments ask two questions: whether OCP-R improves option scores with the backbone unchanged, with the gain coming from the residual readout rather than parser changes, answer-letter bias, generic probing, or internal adaptation; and whether the improvement is stable across backbones, measured by the spread of held-out changes rather than mean accuracy alone.

4.1 EXPERIMENTAL SETTING

Benchmarks. We evaluate on MMAD (Jiang et al., 2025) and AnomalyCoT (Cheng et al., 2025). MMAD covers defect presence, category, localization, description, object recognition, and analysis over five industrial source domains: DS-MVTec and MVTec-AD (Bergmann et al., 2019), MVTec-LOCO (Bergmann et al., 2022), VisA (Zou et al., 2022), and GoodsAD (Zhang et al., 2024a). Its evaluation split has $1{,}844$ questions, of which the 626 GoodsAD questions form the held-out source, and the seven task categories, Anomaly Detection, Defect Classification, Defect Localization, Defect Description, Defect Analysis, Object Classification, and Object Analysis, carry 386, 221, 227, 218, 227, 149, and 416 questions respectively. AnomalyCoT provides a complementary MCQ-style anomaly-evaluation benchmark; we use the same forced-prefix readout, held-out-source construction, and head recipe as on MMAD. Its held-out split carries 247 questions, so one held-out question is worth 0.40 pp there.

Models. We evaluate thirteen VLMs: nine general-purpose models spanning multiple families and scales, Qwen2.5-VL-7B (Bai et al., 2025), Qwen2-VL-7B (Wang et al., 2024), InternVL3-8B (Zhu et al., 2025), InternVL2.5-8B (Chen et al., 2024b), LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), LLaVA-OneVision-7B (Li et al., 2024a), Pixtral-12B (Agrawal et al., 2024), and Idefics3-8B (Laurençon et al., 2024), and four anomaly-specialized models, AD-Copilot (Jiang et al., 2026), EMIT-8B (Guan et al., 2025), AnomalyGPT (Gu et al., 2024), and AnomalyR1 (Chao et al., 2025), whose evaluation protocols differ where their visual interfaces require and are detailed in the Appendix. All nine general-purpose models appear on AnomalyCoT; the MMAD comparison (Table 1) covers seven, omitting LLaVA-OneVision and Idefics3 whose LoRA runs did not evaluate stably. Pixtral is kept on AnomalyCoT despite a LoRA run that fails badly (33.42%), reported as measured.

Option-logit evaluation protocol. All trained option-logit methods, including internal-weight adaptation, use the same forced-prefix protocol, verbalizer, and prompt template of Section 3, so Tables 1 and 2 isolate where the adaptation acts rather than how the answer is read out; the zero-shot frozen column is the benchmark's generate-and-parse readout, and Appendix C measures that switch on Qwen2.5-VL (-0.49 pp overall, $+1.60$ pp held-out). Accuracy is over valid candidate options using the benchmark-provided answer index. Stability is reported as the cross-backbone spread of the held-out change Δ_{he} relative to the frozen baseline, our primary stability metric; we mark catastrophic held-out drops with the descriptive threshold $\Delta_{\mathrm{he}} < -5$ pp as a secondary marker. Absolute numbers on the same split differ slightly across the runs reported below; each table states the setting it uses.

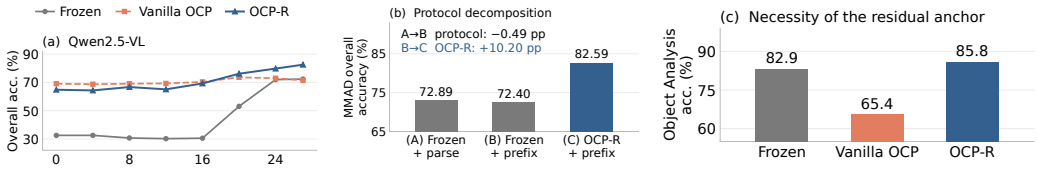


Figure 4: Supporting diagnostics. (a) Layer-wise readout on Qwen2.5-VL: the frozen curve stays low until the final quarter of the network, while option-conditioned readouts recover evidence from layer 0 (InternVL-3 and full values in Appendix A). (b) Protocol decomposition: the protocol switch is nearly neutral, so OCP-R's gain is the residual correction (numbers in text). (c) Necessity of the residual anchor: on Object Analysis, a replacement readout degrades the strong frozen baseline while OCP-R improves on it.

Table 2: AnomalyCoT accuracy across nine VLMs (discussion and stability claim in text; see Table 15 for catastrophic-drop counts).

VLM	Baseline		LoRA		Vanilla OCP		OCP-R (Ours)	
	Overall	Held-out	Overall	Held-out	Overall	Held-out	Overall	Held-out
Qwen2.5-VL	44.92	64.37	**77.61**	**69.64**	76.04	68.42	75.90	67.21
Qwen2-VL	46.54	70.45	**80.56**	**76.52**	75.63	61.13	77.26	67.21
LLaVA-1.5	30.46	53.85	59.72	57.49	73.66	57.89	**74.27**	**61.13**
LLaVA-NeXT	32.95	50.20	**74.38**	57.89	69.37	58.70	71.41	**59.92**
InternVL3	40.08	58.70	**78.41**	**73.68**	76.79	65.18	77.60	69.64
InternVL2.5	42.03	69.23	76.46	66.80	75.63	62.35	**77.20**	**68.83**
Pixtral	34.10	60.73	33.42	52.63	**73.52**	62.75	69.37	**65.18**
Idefics3	35.31	62.75	59.05	59.92	**74.47**	61.13	72.36	**62.35**
LLaVA-OV	47.14	68.42	**79.96**	**72.47**	75.43	60.32	77.74	68.02
Mean (N=9)	39.28	62.08	68.84	65.23	74.50	61.99	**74.79**	**65.50**
Mean Δ vs. Baseline	-	-	+29.56	+3.15	+35.22	−0.09	**+35.51**	**+3.42**

Compared methods and controls. Starting from frozen option logits, we compare OCP-R with readout-family baselines (unconditional probes, non-residual and linear/residual readouts, letter-bias calibration, no-bias, shuffled/random option, and visual-input controls) and with internal-weight adaptation as a stress test. Every table labelled *LoRA* uses one fixed configuration: low-rank adaptation (Hu et al., 2022) with r=4, α=16 on the query, key, and value projections of the last four decoder layers, $\sim$0.246M trainable parameters, trained with the same option-level cross-entropy as OCP-R on each benchmark's own training split; only the updated parameters differ. The Appendix rank sweep and one seven-task breakdown depart from it and say so in place. The frozen baseline is zero-shot, so it measures the interface we correct rather than a matched-supervision competitor. The Appendix gives the full recipe.

4.2 Evaluation Results

Option-level readout diagnostic. Across layers of Qwen2.5-VL and InternVL-3 we compare frozen LM-head option logits, an option-conditioned readout trained on frozen states, and OCP-R (Figure 4a). The early-layer level partly reflects the candidate texts rather than recoverable anomaly content, so we read this as an option-level readout mismatch, not as proof that frozen representations are universally sufficient.

Main results on MMAD. On MMAD with GoodsAD held out (Table 1), OCP-R raises mean overall accuracy over seven VLMs from 62.17% for the frozen baseline and 71.93% for LoRA to 78.27%, and mean held-out accuracy from 60.50% and 69.56% to 72.49%. The held-out gain over LoRA holds on six of seven VLMs; this is not a universal OOD claim, since InternVL-3 is slightly better with LoRA on the held-out split.

Cross-VLM stability on AnomalyCoT. On AnomalyCoT (Table 2) OCP-R attains the highest mean overall/held-out accuracy, 74.79%/65.50%, though not where the claim rests: LoRA is higher

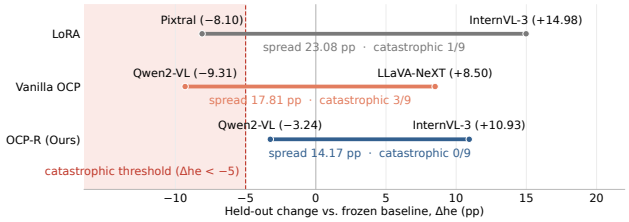


Figure 5: Cross-VLM held-out stability on AnomalyCoT, the primary stability result: each bar spans the held-out change Δ_{he} relative to the frozen baseline across nine VLMs (spreads and catastrophic-drop counts in the text below).

Table 3: OCP-R on four anomaly-specialized VLMs (MMAD, GoodsAD excluded from head training). The two pairs use different visual interfaces (Appendix C) and are not directly comparable. Held-out applies to head training; these backbones saw MMAD upstream.

Model	**Overall**		**Held-out**		**Δ**	
	Baseline	OCP-R	Baseline	OCP-R	Overall	Held
AD-Copilot	80.31	**83.79**	78.59	**79.23**	+3.48	+0.64
EMIT-8B	77.93	**82.10**	76.20	**78.12**	+4.17	+1.92
AnomalyGPT	41.7	**63.3**	41.2	**58.4**	**+21.6**	**+17.2**
AnomalyR1	72.7	**77.6**	72.0	**72.2**	+4.9	+0.2

overall on five of nine backbones and its lower mean is driven by one failed Pixtral run, while the $+0.27$ pp held-out margin is below the 0.40 pp one held-out question is worth. The spread of the held-out column (Figure 5) separates the methods: across the nine backbones, OCP-R has the narrowest held-out spread, 14.17 pp with no catastrophic drop under $\Delta_{\text{he}} < -5$ pp, against 23.08 pp with one drop for LoRA and 17.81 pp with three for Vanilla OCP. The frozen baseline also scores higher held-out than in-domain here, so its held-out split is the easier one.

AD-specialized VLMs and VLM-based AD methods. The nine VLMs above are general-purpose; Table 3 asks whether the mismatch survives anomaly-specific adaptation on four specialized models. On AD-Copilot (Jiang et al., 2026) and EMIT-8B (Guan et al., 2025), OCP-R improves overall accuracy by $+3.83$ pp on average with smaller held-out gains; on the two VLM-based AD methods that bracket the baseline-strength range, AnomalyGPT (Gu et al., 2024), whose single pooled visual token leaves a frozen readout at 41.7%, gains $+21.6$ pp overall, while the GRPO-tuned AnomalyR1 (Chao et al., 2025) starts at 72.7% and gains $+4.9$ pp. The recoverable gain tracks readout weakness, as with the general-purpose backbones, and is not an artifact of using models never trained for the domain (Appendix C).

Readout-family dissection. Table 4 tests each constraint against the parser-free variant that drops it on Qwen2.5-VL, from a frozen logit-lens baseline of 72.40% / 68.85%. Option-content-blind variants map onto established techniques but fail to improve: letter-bias calibration matches the baseline at 72.29%, while tuned-lens-style unconditional probes fall below it (67.03% linear, 70.34% MLP); Vanilla OCP (replacement) degrades Object Analysis where OCP-R improves on it (Figure 4c). The two probe rows carry a caveat: they are fit on AnomalyCoT and evaluated on MMAD, so part of their deficit is the cross-benchmark gap measured separately (-15.44 pp, Appendix). The matched evidence that option conditioning is load-bearing is a same-benchmark AnomalyCoT comparison: an unconditional probe reaches 47.27% against more than 75% for both option-conditioned readouts (Appendix B). Identity initialization is a refinement inside that family rather than a third requirement: a Kaiming-initialized W_k still reaches 82.27% / 75.72%, a gap of $+0.32$ pp on both metrics that is of the same order as the seed-to-seed variation we measure across three seeds (Appendix C).

Prediction preservation. On Qwen2.5-VL MMAD, OCP-R preserves 1246 of 1335 baseline-correct examples and rescues 283 baseline-wrong ones, a rescue rate of 15.35% against a 4.83%

Table 4: Readout-family dissection on Qwen2.5-VL (MMAD source-held-out; layer-wise cached pipeline, cf. the 81.62% paired-image result in Table 1). Each middle-block variant drops one of Section 3's two constraints—option-text conditioning, or adding to the frozen logit rather than replacing it—while the bottom block keeps both, varying only initialization and the option-slot bias. UCP rows are fit on AnomalyCoT (†); part of their deficit is the transfer gap of Table 21.

Readout Variant	**Drops**	**Overall**	**Held-out**
Forced-prefix LM-head (frozen)	–	72.40	68.85
Letter-bias-only (calibration)	option text	72.29	69.01
Linear UCP† (tuned-lens)	option text	67.03	65.34
MLP UCP† (nonlinear remap)	option text	70.34	67.89
Vanilla OCP (replacement)	residual anchor	71.53	57.67
Bilinear + residual (Kaiming W_k)	identity init	82.27	75.72
OCP-R swap-init (W_q=0)	–	80.86	74.28
OCP-R without bias	option-slot bias	82.48	75.72
OCP-R (Full)	–	**82.59**	**76.04**

readout-induced forgetting rate, with the rescues concentrated in the defect-centric tasks where the frozen readout is weakest. Further breakdowns are in Appendix B.

What the correction reads. Switching from generate-and-parse to forced-prefix logits is nearly neutral while the residual correction contributes $+10.20$ pp (Figure 4b), so the gain is not a parser artifact. Removing each input in turn bounds what the correction uses: denied the images, OCP-R loses 20.28 pp on Qwen2.5-VL and 18.71 pp on Qwen2-VL; denied the option texts, an otherwise identical readout falls to 47.27%, as above. The correction is also not tied to the option arrangement: permuting text-to-letter assignment moves accuracy by $+0.54$ pp on Qwen2.5-VL and -1.30 pp on Qwen2-VL, and random distractors move it by -0.11 pp (Appendix B). Meaning-preserving rewrites of the correct option are the one axis we do not vary, the main confound this evidence leaves open.

4.3 Discussion

Transfer limitations. Cross-benchmark transfer is partial: AnomalyCoT-trained heads lose -15.44 pp overall on MMAD without adaptation. Direct cross-model transfer is where the residual anchor separates OCP-R from replacement readouts: a Qwen2.5-VL head on Qwen2-VL without retraining reaches $75.76\%/71.73\%$, while a transferred Vanilla OCP head collapses to $51.41\%/45.69\%$ (Appendix C); we claim OCP-R as a stabilizer for one benchmark interface on one backbone, not a portable readout. This works because the target's frozen term is preserved under transfer, so the transferred head does not fall below it here.

Limitations. Two further limits follow (Appendix F): the option descriptor is a mean-pooled frozen embedding truncated to $T_{\max} = 8$ tokens, weakest on negations (e.g. "normal, no defect"), with richer descriptors unablated. Finally, every benchmark here is an anomaly benchmark, so whether the readout gap is specific to anomaly content or general to letter-verbalizer MCQ evaluation remains open; since Eq. (4)'s argument nowhere uses anomaly content, it may apply beyond the MCQ-style anomaly evaluation we confine claims to here.

5 Conclusion

We introduced OCP-R, an option-conditional residual readout that stabilizes the answer-option interface of VLM anomaly-detection benchmarks. Frozen logits under-expose what the state holds because the option-pair boundary's normal is set by the answer symbols, shared by every question using the same letters, so only a readout-side term can re-orient it per candidate set. Across MMAD and AnomalyCoT with many VLMs, it improves option-level accuracy, preserves correct decisions, and alone avoids catastrophic held-out drops, though adaptation remains stronger per backbone.

ETHICS STATEMENT

VLM-based industrial anomaly detection can improve quality-control efficiency and reduce human exposure to hazardous inspection environments, but it also carries societal risk: workforce displacement in manual quality assurance, and accountability concerns where the model produces silent failures (false negatives) that reach end-users. Our transition analysis (Appendix B) shows the predictions OCP-R changes for the worse concentrate in attribute-discrimination tasks rather than the binary defect-presence question, and that it changes far fewer of them than a replacement readout does. This bounds how often the correction can introduce a new false negative but does not eliminate it, and we make no safety guarantee on that basis. These systems should therefore be deployed as human-in-the-loop assistive tools rather than autonomous decision-makers.

REPRODUCIBILITY STATEMENT

All experiments use frozen, publicly released VLM checkpoints and the publicly available MMAD and AnomalyCoT benchmarks (Section 4); no new data is introduced. Appendix E lists the full OCP-R training and inference recipe: optimizer, learning rate, schedule, readout dimension d_{proj}, option truncation length $T_{\max}$, the identity-initialization scheme (Section 3), and the fixed random seed, together with the matched LoRA configuration used as the internal-weight adaptation baseline. Appendix C reports a three-seed variance study (Table 16) bounding the run-to-run spread of the main result. The readout head is a 132-line addition on top of a frozen backbone with no external dependencies beyond the base VLM; the readout module, forced-prefix evaluation harness, and scripts reproducing every table and figure will be released under a permissive license upon publication.

AI USE STATEMENT

We used generative AI tools for manuscript-level assistance only: light proofreading and copyediting, preparing and laying out figures from numbers already produced by our experiments, and minor coding assistance (boilerplate scripting and debugging) when running experiments. We did not use generative AI to propose or refine hypotheses, design our methodology or experiments, implement the method, generate or clean data, formulate or prove mathematical claims, or interpret results; the paper's claims and all reported numbers come from our own experiments. All AI-assisted text, figures, and code were checked against the underlying results by the authors, who take full responsibility for the final content of this work.

REFERENCES

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Md. Atabuzzaman, Ali Asgarov, and Chris Thomas. Benchmarking and mitigating MCQA selection bias of large vision-language models. In *EMNLP*, pp. 33548–33562, 2025.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. ReZero is all you need: Fast convergence at large depth. *UAI*, 2021.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 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.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. LoRA learns less and forgets less. *TMLR*, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *ICLR*, 2023.

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.

Lequn Chen, Zihao Ye, Yongji Wu, Danyang Zhuo, Luis Ceze, and Arvind Krishnamurthy. Punica: Multi-tenant LoRA serving. *MLSys*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Jiaxi Cheng, Yuliang Xu, Shoupeng Wang, Tao Ma, Yuchen He, Jinghe Zhang, Sihang Cai, Jiawei Zhen, Jingyi Jia, Yao Wan, et al. AnomalyCoT: A multi-scenario chain-of-thought dataset for multimodal large language models. In *NeurIPS*, 2025.

Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *ICPR*. Springer, 2021.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*, 2023.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *IJCV*, 2024.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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

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 *WACV*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

Xi Jiang, Jian Li, Hanqiu Deng, et al. MMAD: A comprehensive benchmark for multimodal large language models in industrial anomaly detection. In *ICLR*, 2025.

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.

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 *ICML*, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL*, 2021.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. PromptAD: Learning prompts with only normal samples for few-shot anomaly detection. In *CVPR*, 2024b.

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*, 2023b.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *PAMI*, 2017.

Vladislav Lialin, Vijeta Deshpande, Xiaowei Yao, and Anna Rumshisky. Scaling down to scale up: A guide to parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.15647*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *ECCV*, 2024c.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *TASLP*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *COLM*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *NeurIPS*, 2022.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of NAACL*, pp. 2006–2017, 2024.

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

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. In *ICLR*, 2023.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter V. Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-LoRA: Serving thousands of concurrent LoRA adapters. *arXiv preprint arXiv:2311.03285*, 2023.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. VL-Adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *CVPR*, 2022.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *ACL*, 2019.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

Zhen Yang, Ping Jian, and Chengzhi Li. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In *NAACL*, pp. 1902–1917, 2025.

Xincheng Yao, Zixin Chen, Chao Gao, Guangtao Zhai, and Chongyang Zhang. ResAD: A simple framework for class generalizable anomaly detection. In *NeurIPS*, 2024.

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. DRAEM-a discriminatively trained reconstruction embedding for surface anomaly detection. In *ICCV*, 2021.

Jian Zhang, Runwei Ding, Miaoju Ban, and Linhui Dai. Pku-goodsad: A supermarket goods dataset for unsupervised anomaly detection and segmentation. *IEEE Robotics and Automation Letters*, 9 (3):2008–2015, 2024a.

Kai Zhang, Zekai Zhang, Xihe Sun, Anpeng Wang, Jingmeng Nie, Qinghui Chen, Han Hao, Jianyuan Guo, and Jinglin Zhang. ADSeeker: A knowledge-grounded reasoning framework for industry anomaly detection and reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21379–21388, 2026.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In *ECCV*, 2022.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2024b.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *ICML*, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *NeurIPS*, 2023.

Ye Zheng, Xiang Wang, Rui Deng, Tianpeng Bao, Rui Zhao, and Liwei Wu. Focus your distribution: Coarse-to-fine non-contrastive learning for anomaly detection and localization. In *ICME*, 2022.

Jiawen Zhu and Guansong Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

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*, pp. 392–408. Springer, 2022.

A Extended Diagnostics on Option-Level Readout

This section provides detailed diagnostic results supporting the option-level readout analysis in the main paper. The goal is not to prove that frozen VLM representations are universally sufficient, but to measure how much anomaly-discriminative evidence is recoverable from frozen states before it is exposed by the frozen answer-option logits.

Table 5: Three-variant layer-wise probing on the MMAD benchmark. Evaluation across the transformer blocks (Layer ℓ) of two distinct VLM architectures. The same pattern holds across both models: a flat Probe-only curve, an S-shaped Baseline, and a rising OCP-R. This is consistent with anomaly-discriminative evidence being recoverable from frozen states well before the pre-trained LM head exposes it through option logits, rather than proving that frozen representations are sufficient on their own. The Probe-only level at layer 0, where the state has passed through no decoder block, is set by the option texts and answer priors rather than by the image, so the quantity to read from this table is the gap between the two readouts at each layer and the shape of the Baseline curve. (Note: Baseline here uses the forced-prefix pipeline on cached states; see Table 20 for the delta vs. Table 1.)

Layer ℓ	**Baseline-only (Frozen)**		**Probe-only (Option-Cond.)**		**OCP-R (Full Method)**	
	Overall	**Held**	**Overall**	**Held**	**Overall**	**Held**
Qwen2.5-VL (28 Decoder Blocks)						
0	32.54	31.79	69.14	56.39	64.86	55.91
4	32.54	31.79	68.76	54.79	64.37	54.79
8	30.69	30.83	69.14	55.91	66.76	55.27
12	30.21	31.15	69.31	56.55	65.13	53.67
16	30.53	31.47	70.39	57.03	69.31	56.07
20	53.09	54.63	73.64	57.19	76.19	62.30
24	72.02	68.69	72.99	58.79	79.83	73.48
27 (Last)	72.40	68.85	71.53	57.67	**82.59**	**76.04**
InternVL-3 (28 Decoder Blocks)						
0	30.10	30.83	69.36	56.87	65.73	56.71
4	32.00	31.79	70.01	56.55	67.25	55.75
8	28.96	29.55	70.99	59.90	68.71	57.67
12	29.12	29.23	70.93	58.63	70.82	59.11
16	30.31	31.47	72.72	59.11	71.37	58.31
20	53.04	47.92	73.97	59.11	76.30	63.90
24	67.68	64.86	72.99	59.11	78.74	72.04
27 (Last)	68.49	65.81	72.07	58.31	**80.69**	**74.44**

Layer-wise readout diagnostic. Table 5 reports the full layer-wise results for Qwen2.5-VL and InternVL-3 on MMAD. Across both VLMs, the frozen LM-head readout remains low in early and middle layers. For Qwen2.5-VL, the frozen readout stays around 30–33% until layer 16, then rises to 53.09% at layer 20, 72.02% at layer 24, and 72.40% at the final layer. In contrast, the option-conditioned readout already reaches about 69% from the earliest layers, with 69.14% at layer 0 and 71.53% at the final layer. OCP-R combines the frozen readout with a residual correction and reaches 82.59% overall and 76.04% held-out at the final layer. InternVL-3 shows the same qualitative pattern: the frozen readout stays between 28.96% and 32.00% through layer 16, while the option-conditioned readout stays between 69.36% and 73.97% across every layer, and final-layer OCP-R reaches 80.69% overall and 74.44% held-out. Two features of these curves need care. First, the option-conditioned readout is nearly flat from layer 0, where the prediction state has passed through no decoder block and can carry almost no image-specific evidence. Its early-layer level therefore reflects what the candidate option texts and answer priors supply on their own, not anomaly evidence read out of the state; on the held-out split it is close to the text-only control of Table 11 (56.39% against 57.51%), while the overall column sits about 8 pp above it. The quantity we read from this sweep is the gap between the two readouts on the same state and the shape of the frozen curve, which stays flat until layer 20 and only then rises. Second, OCP-R sits below the option-conditioned readout in the early layers, because a residual anchored on frozen logits that are themselves near chance inherits that starting point. The residual anchor is useful where the frozen readout carries signal, which on both VLMs is the last quarter of the network. With that reading, these results support the readout-mismatch interpretation: the frozen LM-head option logits expose usable evidence later and less reliably than a readout with access to the option texts. Figure 6 plots these two readouts, together with OCP-R, for both backbones and, in panel (c), contrasts the residual and replacement readouts on the strong frozen predictions of the final layer.

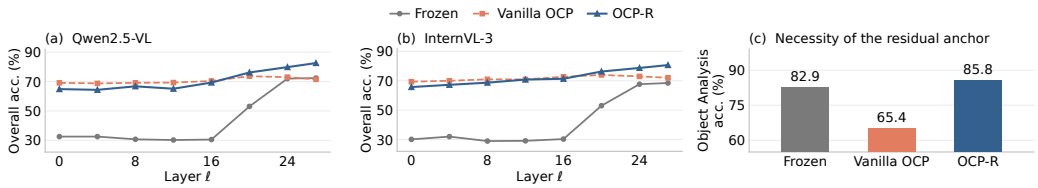


Figure 6: Layer-wise readout diagnostic and residual preservation. (a,b) Option-conditioned readouts recover anomaly evidence earlier than frozen LM-head logits expose it. (c) Non-residual readout can degrade strong frozen predictions, while OCP-R improves option-level prediction through a residual correction; this panel is the one reproduced as Figure 4c in the main text.

Table 6: Format/answer-token supervision decomposition on MMAD across VLMs with different frozen baseline strengths. We report the share of the MCQ supervision that falls on answer-format tokens versus on the answer token itself. Base Overall is the frozen generate-and-parse MMAD accuracy from Table 1. Models with stronger frozen baselines tend to show a lower format share and a higher answer-token share, suggesting that MCQ supervision can emphasize different token groups depending on baseline strength. This table is used as supporting diagnostic evidence for option-level readout instability, not as a causal explanation of all adaptation behavior.

VLM Backbone	Base Overall (%)	Format Share (%)	Answer Share (%)
LLaVA-1.5-7B	57.05	86.0	14.0
Pixtral-12B	58.35	72.0	28.0
InternVL3-8B	63.94	58.3	41.7
Qwen2.5-VL-7B	72.89	33.5	66.5

Format and content signal decomposition. Table 6 decomposes MCQ supervision into format-token and answer-token shares across VLMs with different frozen baseline strengths. The weakest baseline, LLaVA-1.5-7B, has 57.05% base overall accuracy and an 86.0% format share. Pixtral-12B has 58.35% base accuracy and a 72.0% format share, while InternVL3-8B has 63.94% base accuracy and a 58.3% format share. The strongest baseline in this diagnostic, Qwen2.5-VL-7B, has 72.89% base accuracy and a much lower format share of 33.5%. This trend suggests that MCQ supervision can emphasize different token groups depending on frozen baseline strength, and it is the measurement that motivates the first design constraint stated in the introduction. The four backbones give a rank correlation of 1 and a Pearson r of 0.98 between base accuracy and answer-token share, but with $N = 4$ this is a consistent trend rather than an estimated effect size, and we use it as such. We use this table as supporting evidence for readout-level analysis, not as a complete causal explanation of adaptation.

Definition of format and answer-token share. For each MMAD training sample using the response template "`The answer is (X).`", we partition the tokens into an answer set $\mathcal{A} = \{X\}$ (the answer letter) and a format set $\mathcal{F}$ (the remaining template tokens). We compute the per-token cross-entropy loss $\ell_t = -\log p_\theta(t \mid t_{<t}, X_{\text{visual}}, Q)$ under the frozen VLM. The format share is the average fraction of supervision mass concentrated on format tokens:

$$\text{format-share} = \frac{1}{N}\sum_{j=1}^{N}\frac{\sum_{t\in\mathcal{F}_j}\ell_t}{\sum_{t\in\mathcal{F}_j\cup\mathcal{A}_j}\ell_t},$$

$$\text{answer-share} = 1 - \text{format-share},$$

averaged over $N = 300$ samples from seen sources. This measures how much CE supervision a frozen VLM allocates to format compliance versus content discrimination. This metric is purely observational and does not modify the training loss.

Baseline strength versus residual gain. Table 7 reports how frozen baseline strength relates to OCP-R gain on MMAD. The largest gains occur on weaker frozen baselines. For example, LLaVA-NeXT improves from 41.65% to 76.42% overall, a $+34.77$ pp gain, and from 43.13% to 71.57% held-out, a $+28.44$ pp gain. Pixtral also receives a large correction, improving by $+20.29$ pp overall

Table 7: Baseline strength versus OCP-R residual readout gain on MMAD. The gain (Δ) decreases as the frozen baseline strengthens: the weakest baseline (LLaVA-NeXT) receives the largest residual correction and the strongest (Qwen2.5-VL) the smallest. The trend is not monotone—Pixtral has a slightly stronger baseline than LLaVA-1.5 and still gains more—so we read it as a decreasing trend rather than a fitted relationship, which is what a residual corrector should show and a uniform classifier boost should not.

VLM	Frozen Baseline		OCP-R (Ours)		Residual Gain (Δ)	
	Overall	Held-out	Overall	Held-out	Overall	Held-out
LLaVA-NeXT	41.65	43.13	76.42	71.57	+34.77	+28.44
LLaVA-1.5	57.05	61.34	70.99	65.34	+13.94	+4.00
Pixtral	58.35	56.87	78.64	73.72	+20.29	+16.85
InternVL-3	63.94	60.70	80.31	73.16	+16.37	+12.46
InternVL-2.5	69.69	66.29	78.52	71.73	+8.83	+5.44
Qwen2-VL	71.64	67.89	81.40	75.88	+9.76	+7.99
Qwen2.5-VL	72.89	67.25	81.62	76.04	+8.73	+8.79

and $+16.85$ pp held-out. Stronger baselines receive smaller but still positive corrections: Qwen2.5-VL improves by $+8.73$ pp overall and $+8.79$ pp held-out, while Qwen2-VL improves by $+9.76$ pp overall and $+7.99$ pp held-out. This pattern is consistent with a residual correction mechanism: less correction is needed when the frozen option logits are already relatively reliable. The same trend extends to the four anomaly-specialized models of the main text, where AnomalyGPT, the weakest of those four frozen readouts, receives the largest correction.

B Extended Ablation Studies and Controls

This section provides detailed ablations and control experiments that separate option semantics, residual anchoring, answer-prior effects, and visual grounding. These results complement the main-text ablations by showing where OCP-R improves, where non-residual readouts fail, and which confounds are not sufficient to explain the observed gains.

Task-level behavior across VLMs. Table 8 reports per-task MMAD accuracy for seven VLMs using OCP-R. The mean overall accuracy across the seven VLMs is 78.35%. Object-centric tasks remain strong: mean Object Classification is 91.6% and mean Object Analysis is 83.0%. Defect-centric tasks also show stable performance, with mean Defect Classification 73.7%, Defect Localization 75.1%, Defect Description 82.7%, and Defect Analysis 87.1%; Anomaly Detection, the binary presence question, is the weakest category at 65.3% and also the most variable across backbones. Its weakness aligns with the negation stratum of Table 14: the correct option on the presence question is often a lexicalized negation such as "normal, no defect", which is where the mean-pooled option descriptor is least informative. Among the task categories, Defect Localization has the narrowest cross-model spread, 7.5 pp, suggesting stable spatial option prediction across architectures. This table supports the claim that residual readout correction can improve defect-centric option decisions while preserving strong object-centric predictions.

Residual anchoring on Qwen2.5-VL. Table 9 gives a detailed seven-task breakdown for Qwen2.5-VL. The frozen baseline is already strong on object-centric tasks, with 92.62% Object Classification and 82.93% Object Analysis. Vanilla OCP improves some defect-centric tasks, such as Defect Classification (71.95%), but substantially degrades object-centric tasks, dropping Object Classification to 81.21% and Object Analysis to 65.38%. OCP-R avoids this replacement-readout behavior, reaching 95.30% on Object Classification and 85.82% on Object Analysis, while also improving Defect Classification to 77.83% and Defect Localization to 75.77%. Overall, OCP-R reaches 82.27%, compared with 72.61% for the frozen baseline, 78.25% for LoRA at $r{=}8$, and 71.64% for Vanilla OCP. The LoRA column here is the $r{=}8$ configuration rather than the $r{=}4$ default of Table 1, so this breakdown holds OCP-R against the stronger adapter (Table 17). This supports residual anchoring as a preservation mechanism.

Table 8: Per-VLM Task-Level Accuracy (%) on MMAD using OCP-R. OCP-R maintains high accuracy on object-centric tasks (ObjCls, ObjAnly), and spatial option prediction (DefLoc) has the narrowest cross-model spread of the seven task categories. (Note: Unlike Table 1 which holds out GoodsAD, models here are trained on all five MMAD domains to measure full in-distribution accuracy.)

VLM	AnomDet	DefCls	DefLoc	DefDesc	DefAnly	ObjCls	ObjAnly	Overall
Qwen2.5-VL	73.3	77.8	75.8	86.2	89.4	95.3	85.8	82.27
Qwen2-VL	71.8	78.3	77.1	84.9	86.8	97.3	85.1	81.67
LLaVA-1.5	60.9	69.2	70.5	78.0	79.7	88.6	74.5	72.72
Pixtral	64.4	71.0	75.8	78.1	89.6	88.2	83.7	77.58
InternVL-3	64.0	74.7	78.0	85.3	90.7	89.3	85.3	79.66
InternVL-2.5	63.2	73.3	77.1	83.5	86.3	87.9	84.1	78.09
LLaVA-NeXT	59.2	71.4	71.7	82.8	87.0	94.6	82.4	76.48
Mean (N=7)	65.3	73.7	75.1	82.7	87.1	91.6	83.0	78.35
Spread	14.1	9.1	7.5	8.2	11.0	9.4	11.3	9.55

Table 9: Seven-task MMAD breakdown for Qwen2.5-VL. Vanilla OCP improves several defect-centric tasks but substantially degrades object-centric tasks where the frozen VLM is already strong, especially Object Classification and Object Analysis. OCP-R preserves these object-centric predictions while improving defect-centric tasks, supporting the role of residual anchoring. (Note: the adapted columns here, LoRA included, are retrained under the full five-source training setting rather than the source-held-out setting of Table 1, so overall scores are not directly comparable across the two tables. The LoRA column is the r=8 configuration of Table 17, not the r=4 default used in Table 1, and its overall score coincides with the source-held-out r=8 run in Table 17, which is a separate run; this breakdown compares OCP-R against the stronger of the two adapters.)

Task Category	Baseline	LoRA (r=8)	Vanilla OCP	OCP-R (Ours)
Anomaly Detection	68.13	69.95	69.43	**73.32**
Defect Classification	57.47	68.33	71.95	**77.83**
Defect Localization	54.63	69.16	68.72	**75.77**
Defect Description	70.18	83.49	72.94	**86.24**
Defect Analysis	83.26	86.34	81.94	**89.43**
Object Classification	92.62	93.29	81.21	**95.30**
Object Analysis	82.93	83.65	65.38	**85.82**
Overall	72.61	78.25	71.64	**82.27**

Table 10: Option-conditioning ablation on Qwen2.5-VL with AnomalyCoT overall accuracy. Option-conditioned readouts substantially outperform the unconditional probe, while Vanilla OCP is slightly higher than OCP-R on this overall metric. Residual anchoring is evaluated separately through held-out and preservation analyses.

Architecture	Option-Cond.	Residual	AnomalyCoT (%)
Unconditional Probe (UCP)	×	×	47.27
Vanilla OCP	✓	×	**76.04**
OCP-R (Ours)	✓	✓	75.90

Identity-initialization asymmetry. Identity initialization is worth less than the two constraints it sits inside. Dropping it entirely, by Kaiming-initializing W_k so that the correction is nonzero at step zero, costs 0.32 pp overall and 0.32 pp held-out (Table 4), which is comparable to the seed-to-seed spread in Table 16. We therefore present it as the safe default rather than as a source of accuracy: it is what makes the untrained readout provably equal to the frozen VLM, which is the property the cross-model transfer result of Appendix C relies on. Given that choice, which projection to zero is not symmetric: swapping the setup to $W_q = 0$ with Kaiming-initialized W_k preserves the step-zero identity but degrades held-out accuracy to 74.28%, compared with 76.04% for the default $W_k = 0$. Because zeroing both matrices creates a gradient deadlock that degenerates into the bias-only baseline, at least one must remain nonzero, with $W_k = 0$ proving to be the empirically stronger configuration. A plausible reading of the asymmetry is that h_T is contextually rich while e_i is a static embedding average, so silencing the option side at step zero lets it grow only where supervision indicates useful text directions; we note this as a hypothesis rather than an established mechanism.

Option conditioning on AnomalyCoT. Table 10 isolates option conditioning on the AnomalyCoT overall metric for Qwen2.5-VL. The unconditional probe reaches only 47.27%, while both option-conditioned readouts exceed 75%, confirming that access to the candidate option text is what carries the correction. On this single overall number, Vanilla OCP is marginally higher than OCP-R (76.04% vs. 75.90%). Residual anchoring is therefore not what maximizes overall accuracy on one backbone. Its effect appears in held-out behavior, where Vanilla OCP incurs three catastrophic held-out drops across the nine AnomalyCoT backbones and OCP-R incurs none (Table 15).

Option-prior and visual-grounding controls. Table 11 evaluates whether OCP-R relies on visual evidence and candidate-option semantics rather than answer-letter priors. The original reference protocol obtains 81.72% overall accuracy. Shuffling the mapping between option text and answer letter yields 82.27%, a negligible $+0.54$ pp change, suggesting that the method is not relying on fixed letter positions. Replacing only the reference image reduces performance to 76.30%, a -5.42 pp change, while removing visual inputs entirely drops accuracy to 61.44%, a -20.28 pp change. The much larger drop in the text-only setting indicates that visual input remains essential, while the moderate reference-image drop is consistent with the query-centric structure of the benchmark. The same controls repeated on Qwen2-VL-7B (Table 12) show the same pattern from an OCP-R readout of 79.99% overall: shuffling the option-to-letter mapping changes accuracy by only -1.30 pp, shuffling the reference image by -3.52 pp, and removing the visual input entirely by -18.71 pp, so neither the instability nor the controls that verify the correction are confined to one backbone.

Option-set robustness. On the same $1{,}844$ Qwen2.5-VL evaluation samples, replacing each incorrect option with a randomly drawn incorrect option changes overall accuracy by only -0.11 pp, and removing one distractor raises accuracy to 83.30% on the subset where one can be removed, as expected from the lower option count. This bounds what the distractor set contributes, but it is not itself evidence of option conditioning, since the correct option text is unchanged; that evidence is Table 10, where a readout without the option texts reaches 47.27% against more than 75% for both option-conditioned readouts. The correction is therefore driven by the query image and the correct option text, not by a fixed distractor set or option count.

What the option controls bound, and what they do not. The controls above constrain three distinct shortcuts and leave one open, and it is worth separating them. Shuffling the text-to-letter assign-

Table 11: Option-prior and visual-grounding controls for OCP-R on Qwen2.5-VL (MMAD). Shuffling the option-letter mapping (Shuffled Options) yields negligible change ($+0.54$ pp), which is consistent with a readout conditioned on option semantics rather than on memorized letter-position priors. Conversely, removing visual inputs (Text Only) drops accuracy by 20.28 pp, indicating that the correction depends on visual evidence. Replacing only the reference image (Shuffled Ref. Image) drops accuracy moderately, matching the benchmark's query-centric design.

Input Perturbation	**Overall**	**Seen**	**Held-out**	**Δ Overall**
Original (Reference Protocol)	81.72	85.30	74.76	—
Shuffled Options (Text $\leftrightarrow$ Letter)	82.27	85.88	75.24	$+0.54$
Shuffled Ref. Image (Random anchor)	76.30	78.82	71.41	-5.42
Text Only (No query/ref images)	61.44	63.46	57.51	-20.28

Table 12: The same option-prior and visual-grounding controls for OCP-R on a second backbone, Qwen2-VL-7B (MMAD). The pattern matches Qwen2.5-VL in Table 11. Shuffling the option-letter mapping (Shuffled Options) changes accuracy by only -1.30 pp, while removing visual inputs (Text Only) drops it by 18.71 pp. The readout instability and the controls that verify the correction are therefore not specific to one backbone.

Input Perturbation	**Overall**	**Seen**	**Held-out**	**Δ Overall**
Original (OCP-R Readout)	79.99	83.99	72.20	—
Shuffled Options (Text $\leftrightarrow$ Letter)	78.69	82.68	70.93	-1.30
Shuffled Ref. Image (Random anchor)	76.46	80.30	69.01	-3.52
Text Only (No query/ref images)	61.28	64.70	54.63	-18.71

ment ($+0.54$ pp on Qwen2.5-VL and -1.30 pp on Qwen2-VL, Table 11) bounds reliance on a fixed answer-letter prior; replacing every incorrect option with a randomly drawn one (-0.11 pp) bounds reliance on the particular distractor set; and removing the visual input (-20.28 pp and -18.71 pp) shows the correction is not reading the option set alone. Two numbers nevertheless show how much this benchmark can be scored without image-specific evidence: the text-only readout reaches 61.44%, and an option-conditioned readout is already near 69% at layer 0 (Table 5), where the prediction state has passed through no decoder block. What none of these controls varies is the wording of the correct option itself. A meaning-preserving rewrite of the correct answer, holding its meaning and the image fixed while changing its lexical form, is the control that would separate a readout using option semantics from one keyed to the phrasing MMAD happens to use, and we did not run it. We therefore do not claim that the correction is invariant to option wording, and we regard this as the main confound our evidence leaves open.

Sample-level transition analysis. Table 13 reports the sample-level transition matrix for Qwen2.5-VL on the MMAD evaluation split with $N = 1844$. Among the 1335 examples that the frozen baseline predicts correctly, OCP-R preserves 1246, corresponding to a 93.3% preservation rate. It changes 89 baseline-correct examples to wrong predictions, a readout-induced forgetting rate of 4.83% over the full evaluation set. Among the 509 baseline-wrong examples, OCP-R rescues 283, a 55.6% conditional rescue rate and a 15.35% rescue rate over the full evaluation set. Measured on the same full-set denominator, the rescue rate is therefore more than three times the forgetting rate, supporting OCP-R as a conservative residual correction with positive net correction. Qualitatively, the baseline-correct samples forgotten by OCP-R cluster in attribute-discrimination tasks such as Object Analysis rather than in the binary Anomaly Detection question, whose per-task forgetting rate is the lower of the two. These involve class-level confusions among semantically close options (e.g., "brownish" vs. "dark grey"). Forgetting on the binary presence question is not zero, however, so the residual correction does not by construction rule out new false negatives; it reduces how often they arise relative to a replacement readout, which forgets at roughly four times the rate. We do not separate the direction of these binary flips, so this bounds how often a new false negative can be introduced without characterizing which flips they are.

Table 13: Sample-level transition matrix for Qwen2.5-VL on the MMAD evaluation split ($N = 1844$). OCP-R preserves 93.3% of baseline-correct samples ($1246/1335$) and rescues 283 baseline-incorrect samples. Its rescue rate (15.35%) is more than three times its readout-induced forgetting rate (4.83%). The totals of this run correspond to 72.40% overall for the frozen readout and 82.92% for OCP-R.

	OCP-R Correct	**OCP-R Wrong**	**Total**
Baseline Correct	1246 *(Preserved)*	89 *(Forgotten)*	1335
Baseline Wrong	283 *(Rescued)*	226 *(Remained)*	509
Total	1529	315	1844

Table 14: Fine-grained failure mode analysis of OCP-R on Qwen2.5-VL (MMAD source-held-out). An option counts as negation-bearing if it asserts the absence of a defect or feature, which covers the explicit markers `no`, `not`, `without`, and `absent` as well as the lexicalized forms `normal`, `no defect`, and `no anomaly`. The last group is why the negation-bearing stratum is 27.6% of the split: "normal" is a standard answer in defect-presence questions, not a rare construction. Gains are largest when options overlap lexically or are of medium length, which is where option-conditional grounding has the most to separate.

Property	**Category**	**Count**	**Frozen Baseline**	**OCP-R**	**Δ (pp)**
Option Similarity (Pairwise Jaccard)	Low (< 0.10)	1113	71.34	80.77	+9.43
	Mid (0.10–0.30)	470	74.04	87.02	+12.98
	High (≥ 0.30)	261	65.52	79.69	**+14.18**
Option Length (Tokens)	Short (≤ 4)	1168	67.29	78.34	+11.04
	Med (5–8)	470	74.89	87.66	**+12.77**
	Long (> 8)	206	84.95	91.75	+6.80
Negation Presence	Without Negation	1335	72.06	84.64	**+12.58**
	With Negation	509	68.96	75.83	+6.88

Fine-grained failure mode analysis. To better understand where the residual readout succeeds and fails, Table 14 stratifies the 1844 Qwen2.5-VL evaluation samples by candidate-option properties. OCP-R yields its largest gains where the options overlap lexically ($+14.18$ pp) and for medium-length options ($+12.77$ pp), the two strata where frozen letter-level priors have the most trouble separating close candidates. Conversely, gains are smaller on negation-containing samples ($+6.88$ pp vs. $+12.58$ pp without), likely because mean-pooled input embeddings are less informative when the discriminating signal relies on the absence of a feature. Qualitatively, forgotten cases (baseline-correct $\rightarrow$ OCP-R-wrong, 4.83%, see Table 13) concentrate in tasks with subtle visual cues where frozen logits were already correct, while rescued cases (baseline-wrong $\rightarrow$ OCP-R-correct, 15.35%) concentrate in tasks with semantically gradient option texts. The negation strata have the same sizes as the correctness partition of Table 13, 1335 and 509, which is a coincidence of this split and not a shared mask: if the two partitions were the same, the without-negation stratum would be the baseline-correct set and its frozen accuracy would be 100% by construction, whereas it is 72.06%. These strata are computed from a separate evaluation run of the same split, whose aggregate accuracies are 71.20% frozen and 82.21% OCP-R, so the differences in this table should be read within it rather than against Table 1.

Cross-VLM held-out variation. Table 15 provides the numerical held-out stability statistics behind Figure 5 of the main text. LoRA ranges from a -8.10 pp held-out change on Pixtral to a $+14.98$ pp gain on InternVL3, producing a 23.08 pp spread and one catastrophic held-out drop under $\Delta_{\text{he}} < -5$ pp. Vanilla OCP ranges from -9.31 pp on Qwen2-VL to $+8.50$ pp on LLaVA-NeXT, with a 17.81 pp spread and three catastrophic drops. OCP-R has the narrowest spread, 14.17 pp, ranging from -3.24 pp on Qwen2-VL to $+10.93$ pp on InternVL3, and has no catastrophic held-out drops under the threshold. This supports the view that OCP-R stabilizes held-out behavior across VLMs, while still allowing nonzero negative changes. The threshold should be read against the split size: one AnomalyCoT held-out question is worth 0.40 pp, so $\Delta_{\text{he}} < -5$ pp corresponds to roughly twelve questions, and the three smallest OCP-R changes (-0.40 pp) are single-question differences.

Table 15: Cross-VLM held-out variation on AnomalyCoT. We report the minimum and maximum held-out change relative to the frozen baseline, the resulting spread, and the number of catastrophic held-out drops, defined as $\Delta_{\mathrm{he}} < -5$ pp. OCP-R has the smallest spread and no catastrophic held-out drop under this threshold, while still showing a nonzero negative minimum change.

Method	Min Δ_{he} (VLM)	Max Δ_{he} (VLM)	Spread	Catastrophic Drops
LoRA	-8.10 (Pixtral)	$+14.98$ (InternVL3)	23.08 pp	1 / 9
Vanilla OCP	-9.31 (Qwen2-VL)	$+8.50$ (LLaVA-NeXT)	17.81 pp	3 / 9
OCP-R (Ours)	-3.24 (Qwen2-VL)	$+10.93$ (InternVL3)	**14.17 pp**	**0 / 9**

Table 16: Multi-seed variance on Qwen2.5-VL (MMAD source-held-out) across three random seeds ($N = 3$: the default seed 42 plus seeds 1 and 7). Reported as Mean $\pm$ Standard Deviation. The gain of OCP-R over LoRA is large relative to seed-level variation.

Method	Overall Acc (%)	Held-out Acc (%)
LoRA	77.60 ± 0.13	73.64 ± 0.35
OCP-R (Ours)	$\mathbf{81.58 \pm 0.46}$	$\mathbf{75.99 \pm 0.40}$
Δ (OCP-R $-$ LoRA)	+3.98	+2.35

We therefore treat the count of catastrophic drops, not the exact magnitudes, as the quantity this table reports.

C Further Analysis and Stress Tests

Multi-seed variance and statistical stability. To ensure our reported gains are robust to initialization and data-shuffling variance, Table 16 reports a multi-seed evaluation ($N = 3$ random seeds: the default seed 42 plus seeds 1 and 7) comparing OCP-R and LoRA on Qwen2.5-VL. Across the three seeds, OCP-R achieves $81.58 \pm 0.46\%$ overall and $75.99 \pm 0.40\%$ on held-out domains, while LoRA achieves $77.60 \pm 0.13\%$ overall and $73.64 \pm 0.35\%$ held-out. The improvement of OCP-R over LoRA ($+3.98$ pp overall, $+2.35$ pp held-out) is at least $5.8\times$ the larger standard deviation of either method, and their 95% intervals, formed with the t quantile for two degrees of freedom, do not overlap on either metric. With only three seeds we do not claim a formal significance test, but the gap between the residual readout and internal-weight adaptation is far larger than the seed-level variation we observe. Two caveats apply to how this table is used. All other tables report the default seed 42, which for both methods lies above the three-seed mean, so the seed-42 rows should be read as one draw rather than as an expected value. The seed sweep also covers one backbone, so it bounds initialization and shuffling variance for Qwen2.5-VL and not the cross-VLM spread, which Table 15 reports separately.

Paired significance of the main result. Table 13 makes the Qwen2.5-VL MMAD comparison a paired test rather than a comparison of two independent accuracies: across the 1844 evaluation samples OCP-R and the frozen readout disagree on 372, of which 283 resolve in favour of OCP-R and 89 against. McNemar's test on those discordant pairs gives $\chi^2 = 100.1$ with continuity correction ($p < 10^{-22}$; the exact two-sided binomial test gives $p < 10^{-23}$), and the paired accuracy difference on that run is $+10.52$ pp with a 95% interval of $[+8.53, +12.51]$ pp. On AnomalyCoT the pairing is across backbones rather than samples, and there it does not support a mean-accuracy claim: the held-out difference between OCP-R and LoRA is positive on five of the nine backbones with a mean of $+0.27$ pp, and an exact Wilcoxon signed-rank test over the nine pairs is far from significant, which is why Section 4 rests the AnomalyCoT claim on the spread of the held-out column rather than on its mean. The spread comparison is itself descriptive at this sample size, and we prefer to say so. Swapping the two method labels within each backbone, the exact paired permutation test over all 2^9 assignments gives $p = 0.25$ for OCP-R's held-out spread being the narrower one, against LoRA and against Vanilla OCP alike, while the catastrophic-drop counts ($0/9$ against $3/9$ for Vanilla OCP) give Fisher's exact $p = 0.21$. Nine backbones with one run each cannot separate these at conventional levels, so we report the stability result as a consistent pattern rather than as a significance claim. The quantities that do not depend on that test are the worst-case

Table 17: Parameter-efficiency stress test on Qwen2.5-VL (MMAD). LoRA rank sweeps at comparable trainable-parameter scales ($r \in \{4, 8, 16, 32\}$) do not close the gap with OCP-R. At a matched parameter count ($r = 8$, ~ 0.49M parameters), OCP-R outperforms LoRA by $+3.37$ pp overall. Even when quadrupling LoRA's trainable parameters ($r = 32$), it reaches 79.07%, so in this parameter regime the readout-level correction remains ahead while leaving the frozen backbone unmodified. All LoRA configurations target q/k/v projections with depth held constant; rank is the swept axis.

Method / Config	Trainable Params	Overall	Held-out	Δ Overall	Δ Held
LoRA $r = 4$	0.246M	77.77	74.12	-3.85	-1.92
LoRA $r = 8$ *(Matched)*	0.492M	78.25	74.28	-3.37	-1.76
LoRA $r = 16$	0.983M	78.52	74.60	-3.10	-1.44
LoRA $r = 32$	1.970M	79.07	75.72	-2.55	-0.32
OCP-R (Reference)	**0.473M**	**81.62**	**76.04**	**—**	**—**

held-out change, -3.24 pp for OCP-R against -8.10 pp for LoRA and -9.31 pp for Vanilla OCP, and the observation that three of OCP-R's four negative changes are single-question differences of -0.40 pp.

Sensitivity of the stability comparison to one failed run. LoRA's held-out spread on AnomalyCoT is driven by a single run that did not converge, on Pixtral, and the comparison should be read with that in mind. Excluding that backbone, LoRA's spread falls from 23.08 to 17.81 pp and its catastrophic-drop count from $1/9$ to $0/8$, its worst held-out change becomes -2.83 pp against OCP-R's -3.24 pp, and its mean held-out accuracy, 66.80%, exceeds ours, 65.54%. What survives the exclusion in the comparison against internal-weight adaptation is therefore the narrower spread, 14.17 against 17.81 pp, and not the drop count or the mean, and we claim no more than that; we also did not tune a per-model learning rate, so the Pixtral failure may be an optimization artifact rather than a property of adaptation. The comparison that does not depend on this run is the one the residual anchor is meant to address, against the replacement readout: with Pixtral excluded, Vanilla OCP still records three catastrophic drops out of eight and a 17.81 pp spread against OCP-R's 14.17 pp.

Internal-weight adaptation stress test. Table 17 compares OCP-R with LoRA rank sweeps on Qwen2.5-VL MMAD. This experiment tests whether the advantage of OCP-R can be explained simply by using a stronger internal-weight adaptation baseline at comparable parameter scales. All variants are evaluated under the same MMAD protocol. LoRA with rank $r = 4$ uses 0.246M trainable parameters and reaches 77.77% overall and 74.12% held-out. The matched-scale LoRA setting ($r = 8$ with 0.492M parameters) reaches 78.25% overall and 74.28% held-out. Increasing rank improves LoRA gradually: $r = 16$ reaches 78.52% overall and 74.60% held-out, while $r = 32$ reaches 79.07% overall and 75.72% held-out. OCP-R uses 0.473M parameters and reaches 81.62% overall and 76.04% held-out. Thus, increasing LoRA rank at comparable trainable-parameter scales does not close the gap to OCP-R, although LoRA remains a competitive adaptation baseline. We interpret this result as support for readout-level correction in this parameter regime, not as evidence that internal-weight adaptation is generally inferior.

Source-domain decomposition. Table 18 reports leave-one-source-out evaluation on MMAD over DS-MVTec, GoodsAD, MVTec-LOCO, VisA, and MVTec-AD. This analysis is included because aggregate MMAD accuracy can hide source-specific behavior across different industrial visual regimes. Each fold holds out one source domain for testing while using the remaining sources for adaptation, providing a stricter within-benchmark stress test than a single aggregate split. For Qwen2.5-VL, OCP-R achieves the best mean held-out accuracy, 80.57%, compared with 79.52% for LoRA and 77.59% for the frozen baseline. For InternVL-3, LoRA has a slightly higher mean held-out accuracy, 79.35%, compared with 78.90% for OCP-R. At the source level, OCP-R is strongest on DS-MVTec and GoodsAD for both backbones, while LoRA remains competitive or better on MVTec-LOCO and VisA. On MVTec-AD, where the frozen baseline already exceeds 97% on both backbones, both LoRA and OCP-R regress relative to the frozen baseline (Qwen2.5-VL: -8.91 / -4.95 pp; InternVL-3: -4.95 / -6.93 pp). This is consistent with our characterization of OCP-R as a residual readout corrector: when the frozen option logits are already near saturation,

Table 18: Leave-one-source-out evaluation on MMAD. Each source domain is held out for testing while the remaining sources are used for adaptation. OCP-R achieves the best mean held-out accuracy for Qwen2.5-VL, while LoRA is slightly higher for InternVL-3. The source-level results are mixed, with OCP-R strongest on DS-MVTec and GoodsAD, and LoRA remaining competitive on MVTec-LOCO and VisA. We use this table as a source-domain stress test rather than as evidence of uniform domain-invariant generalization.

Held-out Source	**Qwen2.5-VL**			**InternVL-3**		
	Baseline	LoRA	OCP-R	Baseline	LoRA	OCP-R
DS-MVTec	83.33	85.56	**86.71**	74.76	83.40	**84.62**
GoodsAD	65.66	72.31	**73.07**	61.95	72.39	**73.68**
MVTec-LOCO	67.01	**73.35**	72.73	57.02	**70.52**	69.01
VisA	74.90	**78.26**	**78.26**	63.83	**77.38**	76.11
MVTec-AD	**97.03**	88.12	92.08	**98.02**	93.07	91.09
Mean	77.59	79.52	**80.57**	71.12	**79.35**	78.90

Table 19: Additional check on AD-aware VLMs. OCP-R improves overall accuracy on AD-Copilot and EMIT-8B, suggesting that option-level readout correction can remain useful after anomaly-specific adaptation. Held-out changes are smaller, so this result is used as supporting evidence rather than a central claim.

AD-Aware VLM	**Baseline**		**LoRA**		**OCP-R (Ours)**	
	Overall	Held-out	Overall	Held-out	Overall	Held-out
AD-Copilot	80.31	78.59	81.34	**79.71**	**83.79**	79.23
EMIT-8B	77.93	76.20	78.54	77.32	**82.10**	**78.12**
Gain over Baseline	-	-	-	-	**+3.83**	**+1.28**

any additional correction signal mostly displaces correct predictions, so the largest residual gains naturally come from sources where the frozen readout has more headroom. These mixed results are useful because they show where residual readout correction is robust and where internal-weight adaptation or the frozen readout remains competitive. We therefore use this evaluation as a source-domain stress test rather than as evidence of uniform domain-invariant generalization.

Protocols for the AD-specialized VLM evaluation. This paragraph gives the protocol details behind Table 3 of the main text and Table 19 below. Classic score- or mask-based detectors such as WinCLIP (Jeong et al., 2023) and PatchCore (Roth et al., 2022) emit image- or pixel-level anomaly scores rather than answer-option logits, so they cannot be evaluated under the MCQ interface of MMAD, nor equipped with OCP-R. AnomalyGPT (Gu et al., 2024) couples an ImageBind visual encoder with a Vicuna language model and passes the query image to the language model as a single pooled visual token, which is why its frozen option readout is the weakest we measure. It is evaluated on a stratified subset with GoodsAD excluded from head training. That such a weak readout still admits a $+21.6$ pp recoverable gain is a strong instance of option-level readout instability. AnomalyR1 (Chao et al., 2025) is a recent Qwen2.5-VL model fine-tuned with GRPO for industrial AD, evaluated under the standard reference-and-query protocol. Its smaller gain on an already-strong readout matches our baseline-strength analysis, where less correction is needed when the frozen logits are already reliable. Because AnomalyR1 is itself trained on MMAD, we treat its held-out number as in-distribution and report it as supporting evidence, as with AD-Copilot and EMIT-8B. Table 19 adds the LoRA comparison for AD-Copilot and EMIT-8B, which the main-text table omits for compactness.

Protocol decomposition. Table 20 separates OCP-R's gain from parser removal on Qwen2.5-VL MMAD. This control is important because changing from free-form answer generation to option-logit scoring could itself affect measured accuracy. By comparing generate-and-parse, forced-prefix frozen logits, and OCP-R under the same option set, we isolate the contribution of the learned residual correction. The generate-and-parse frozen baseline has 72.89% overall and 67.25% held-out accuracy. Switching to the forced-prefix frozen LM-head protocol gives 72.40% overall and

Table 20: Decomposition of OCP-R performance gains on Qwen2.5-VL (MMAD). To isolate the contribution of the bilinear readout from the overhead of parser failures, we compare the standard generate-and-parse evaluation against a forced-prefix logit evaluation. Switching to the forced-prefix protocol yields a negligible overall change (-0.49 pp). The substantial $+10.20$ pp overall gain originates entirely from the OCP-R bilinear correction. (Note: This table uses the layer-wise cached pipeline (Table 5 here and Table 4 of the main paper), so OCP-R reaches 82.59 overall here; total gain A→C is $+9.71$ pp. See Table 1 for the main 81.62 paired-image result.)

Protocol	Overall	Held-out	Source of Gain
(A) Frozen + Generate-and-parse (Baseline)	72.89	67.25	—
(B) Frozen + Forced-prefix `lm_head` (No probe)	72.40	68.85	$\Delta_{\text{Protocol (A}\to\text{B)}}$: -0.49 (Overall)
(C) OCP-R (Forced-prefix + Bilinear corr.)	**82.59**	**76.04**	$\Delta_{\text{OCP-R (B}\to\text{C)}}$: $\mathbf{+10.20}$ **(Overall)**

Note: Total gain (A → C) is $+9.71$ pp overall and $+8.79$ pp held-out.

Table 21: Cross-benchmark transfer from AnomalyCoT to MMAD without MMAD adaptation. OCP-R readout heads are trained only on AnomalyCoT and evaluated on MMAD. (Note: LLaVA-OneVision-7B is included here to widen the transfer comparison. It is not part of the seven-VLM MMAD comparison in Table 1 because its internal-weight adaptation baseline did not evaluate stably, so only its frozen and OCP-R numbers are available.)

VLM	MMAD-Trained OCP-R Overall / Held-out	AnomalyCoT-Trained OCP-R Overall / Held-out	Transfer Gap (Δ)
Qwen2.5-VL	81.62 / 76.04	71.85 / 68.53	-9.77 / -7.51
Qwen2-VL	81.40 / 75.88	66.70 / 62.30	-14.70 / -13.58
InternVL-3	80.31 / 73.16	68.06 / 65.02	-12.25 / -8.14
InternVL-2.5	78.52 / 71.73	66.76 / 64.38	-11.76 / -7.35
Pixtral	78.64 / 73.72	62.47 / 60.13	-16.17 / -13.59
LLaVA-OV	78.28 / 72.84	59.38 / 57.83	-18.90 / -15.01
LLaVA-NeXT	76.42 / 71.57	58.15 / 63.73	-18.27 / -7.84
LLaVA-1.5	70.99 / 65.34	49.30 / 57.03	-21.69 / -8.31
Mean (N=8)	78.27 / 72.54	62.84 / 62.37	-15.44 / -10.17

68.85% held-out, a negligible -0.49 pp overall change and a $+1.60$ pp held-out change. Adding the OCP-R bilinear residual correction raises performance to 82.59% overall and 76.04% held-out. Thus, the main gain comes from the residual correction, not from simply removing free-form answer parsing. This supports the interpretation that OCP-R improves the option-level readout itself rather than only changing the answer extraction protocol.

Cross-benchmark transfer. Table 21 evaluates readout heads trained on AnomalyCoT and tested on MMAD without MMAD adaptation. This setting tests whether the learned readout correction transfers across benchmark interfaces, source distributions, and option vocabularies. It is a deliberately stringent evaluation because the readout head is not exposed to MMAD training examples. Compared with MMAD-trained OCP-R, the AnomalyCoT-trained heads have a mean transfer gap of -15.44 pp overall and -10.17 pp held-out. The smallest overall transfer gap is on Qwen2.5-VL, -9.77 pp, while the largest is on LLaVA-1.5, -21.69 pp. The transferred heads retain nontrivial performance, but the gaps are substantial, so this should be interpreted as partial transfer rather than a universal anomaly readout. This result clarifies that OCP-R learns useful readout structure, but that the correction remains benchmark-interface dependent.

Cross-model transfer. Applying a Qwen2.5-VL OCP-R head to Qwen2-VL without retraining yields 75.76% overall and 71.73% held-out, above the target's frozen baseline ($+4.12$ pp overall, $+3.84$ pp held-out), while a transferred Vanilla OCP head collapses to $51.41\%/45.69\%$; the $+24.35$ pp gap follows from the residual anchor, which preserves the target VLM's frozen term under transfer (Table 22). A native head remains best (81.40%), and this is a same-family, matched-dimension transfer (both $d = 3584$); broader cross-family transfer is left for future work.

Table 22: Cross-model robustness of the OCP-R readout. We apply an OCP-R head trained on Qwen2.5-VL directly to Qwen2-VL without retraining. Unlike Vanilla OCP which shows a catastrophic drop, the residual structure of OCP-R preserves the target VLM's knowledge and even improves upon the frozen baseline (+4.12 pp). Under identical conditions, this isolates the cross-model stability contributed by the residual anchor relative to a replacement-based readout.

Source Readout	**Target VLM**	**Target Baseline** Overall / Held-out	**Vanilla OCP Xfer** Overall / Held-out	**OCP-R Xfer** Overall / Held-out	**Native OCP-R** Overall / Held-out
Qwen2.5-VL	Qwen2-VL	71.64 / 67.89	51.41 / 45.69	75.76 / 71.73	81.40 / 75.88

D Local Geometry of the Option-Pair Boundary

Section 3 reads Eq. (4) as a local linearization. This section states the exact form and bounds what the linearization leaves out, since LN_q is not affine in h_T and the zero set of $s_i - s_j$ is therefore not a hyperplane in h_T.

Write $\mathrm{LN}_q(h) = \gamma_q \odot z(h) + \beta_q$ with $z(h) = (h - \mu_h \mathbf{1})/\sigma_h$, where $\mu_h = d^{-1}\mathbf{1}^\top h$ and $\sigma_h^2 = d^{-1}\|h - \mu_h \mathbf{1}\|^2$. Differentiating Eq. (4) gives the direction that separates options i and j at a sample as

$$\nabla_h \big(s_i - s_j\big)\Big|_{h=h_T} = u_{ij} + \frac{1}{\sigma_T} P_T\big(\gamma_q \odot \delta_{ij}\big), \qquad P_T = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top - \frac{1}{d} z_T z_T^\top, \tag{6}$$

with $z_T = z(h_T)$. The frozen term u_{ij} is unchanged, and both σ_T and P_T are shared by every option at a given sample, so the candidate texts still enter only through δ_{ij}; what the layer norm adds is one common rescaling and one common projection, not an option-dependent one. The expansion also produces the constant $\delta_{ij}^\top \beta_q$, which joins c_{ij} and offsets the boundary without rotating it. Figure 3 reads this direction on one option pair per sample, the correct option against the runner-up under the frozen readout, so the pair is chosen by the frozen model rather than by the readout under measurement. Figure 2 draws this correction schematically and omits the two layer norms; Eq. (2) is the exact form.

P_T is the orthogonal projection onto the complement of $\mathrm{span}\{\mathbf{1}, z_T\}$, so it removes two directions out of d. For a correction with no particular alignment with those two directions, dropping the projection changes the direction by an angle of order $\sqrt{2/d}$ radians, which is 1.4° at the $d = 3584$ of Qwen2.5-VL; drawing $\gamma_q \odot \delta_{ij}$ at random in that dimension gives a median angular change of 1.1° between the projected and unprojected correction and a median change of 0.7° (maximum 2.7°) in the angle it makes with u_{ij}. The angles we report in Section 3 and Figure 3 are read under this linearization and should be taken at that resolution, which is an order of magnitude finer than the separation they describe: a median near 20° for OCP-R against a near-orthogonal replacement normal. Alignment between $\gamma_q \odot \delta_{ij}$ and $\{\mathbf{1}, z_T\}$ would change the estimate, so we state it as the order of the neglected term rather than as a bound.

E Implementation Details

Training. Across all evaluated VLMs, we use the same readout-head training recipe unless otherwise specified. The OCP-R head is optimized with AdamW, learning rate 5×10^{-4}, weight decay 0.01, and one epoch over $5{,}000$ MMAD training samples. We use batch size 1 with gradient accumulation of 4, a cosine schedule with 5% linear warmup, and gradient-norm clipping at 1.0. The default readout dimension is $d_{\mathrm{proj}} = 64$, the option truncation length is $T_{\max} = 8$, and the residual branch uses identity initialization with $W_k = 0$ and $\mathbf{b} = 0$. The VLM backbone remains frozen.

The LoRA baseline is optimized with the same objective as OCP-R: cross-entropy over the forced-prefix answer-symbol logits of the valid options, on the same training split as OCP-R for each benchmark—the $5{,}000$ MMAD samples for the MMAD tables and the AnomalyCoT training split for Table 2—rather than a language-modelling loss over the whole response. No supervision is therefore spent on the answer-format tokens of the response template, whose share of the cross-entropy mass Table 6 measures separately, and the comparison in Table 1 varies which parameters the gradient updates rather than what is optimized.

The readout head has $132d + 5$ trainable parameters, where d is the VLM hidden dimension. This gives $473{,}093$ trainable parameters for Qwen2.5-VL with $d = 3584$, and up to $675{,}845$ parameters for Pixtral with $d = 5120$. Beyond the parameter count, OCP-R adds negligible runtime and memory overhead. It modifies no backbone weights and reuses the same single forward pass as the frozen baseline. On an RTX A6000 for Qwen2.5-VL, the readout head occupies 1.8 MB, and its per-sample correction takes about 0.35 ms with about 10 MB of peak activation. The comparison with internal-weight adaptation is not a like-for-like latency difference but a structural one. OCP-R's cost is additive: the frozen forward pass is bit-identical to the baseline's, and the 0.35 ms is appended to it, so a served backbone needs no reload and several task heads can share one resident model. LoRA instead changes the forward pass itself, in the query, key, and value projections of the last four decoder layers, so its cost is not a term one can add or remove at inference time, and each adapted task requires either merged weights or a separate adapter path in the backbone. At 0.246M versus 0.473M trainable parameters the two are comparable on paper; the difference that matters at deployment is which computation they touch. We therefore do not report a wall-clock figure for LoRA: with weights merged, its marginal inference latency is zero by construction, and the cost it carries is one adapted backbone per task rather than time per sample. The measured quantity that separates the two is residency, not latency, K OCP-R heads for K tasks cost $K \times 1.8$ MB on top of one shared backbone, while K merged LoRA variants cost K backbones. Option representations are obtained by attention-masked mean pooling over the frozen input embedding lookup.

Throughout the paper, differences and spreads are computed from unrounded accuracies, so a value recomputed from the rounded columns of a table can differ from the reported difference by 0.01 pp.

Inference. We capture h_T with a forward pre-hook on the LM head, after the model's final pre-LM-head normalization. At inference time, we append a deterministic assistant prefix ending immediately before the answer-letter token, such as `The answer is (`, and perform one frozen VLM forward pass. The readout head computes the bilinear correction, adds it to the frozen LM-head option logits, masks absent options to $-\infty$, and selects the highest-scoring option. The MMAD evaluation split excludes GoodsAD from training, and all experiments use random seed 42 unless otherwise specified.

F Limitations and Future Work

While OCP-R effectively stabilizes option-level readout for frozen VLMs, it has several limitations. First, it is designed specifically to correct the multiple-choice question (MCQ) interface and does not replace traditional score- or mask-based anomaly detectors for dense pixel-level localization, nor does it solve open-ended anomaly reasoning. Second, its transferability is constrained. As shown in our cross-benchmark analysis, transferring an AnomalyCoT-trained head to MMAD results in a noticeable performance drop, indicating that the learned residual correction is partially tied to the source distribution and benchmark interface. Direct cross-model transfer is also limited; while it remains stable within the same model family (e.g., Qwen2.5-VL to Qwen2-VL), we have not measured broader cross-family transfer (e.g., Qwen to LLaVA). Third, the residual correction relies on mean-pooled frozen input embeddings of candidate option texts, with truncation to $T_{\max} = 8$ tokens. This encoding is less informative when the discriminating signal lies in the absence of a feature (e.g., negation) or in long descriptive options, as quantified in our fine-grained failure-mode analysis (Table 14).

Scope of the readout gap. Every result in this paper is measured on an industrial anomaly benchmark, so our evaluations do not determine whether the gap is specific to anomaly content or general to letter-verbalizer MCQ evaluation. The boundary reading of Eq. (4) points to the second: the fixed option-pair normal u_{ij} follows from scoring answer symbols, and the derivation nowhere uses anomaly content, so the same constraint should hold for any MCQ task read out this way. Two pieces of our evidence bear on the question without settling it. The mismatch survives anomaly-specific adaptation, appearing on AD-Copilot, EMIT-8B, AnomalyGPT, and the GRPO-tuned AnomalyR1 as well as on nine general-purpose backbones, so it is not an artifact of using models that were never trained for the domain. Conversely, the size of the correction tracks frozen readout strength rather than anything anomaly-specific (Table 7), which is what a general readout effect would predict. If the gap does extend to general MCQ evaluation, the correction would apply more broadly than we

claim, widening its scope rather than weakening the results here; verifying this on a non-anomaly benchmark such as MMBench (Liu et al., 2024c) or SEED-Bench (Li et al., 2023a) is a natural next step, and until then we scope every claim in this paper to MCQ-style anomaly evaluation.

Other open directions. Richer option descriptors are untested: the last hidden state of the option text, or simply a larger $T_{\max}$, would directly target the negation and long-option cases where our gains are smallest, and our current encoder was chosen to avoid introducing a second text model rather than because it is the best available. A confidence-gated residual, scaling the correction by the frozen margin, is the natural response to the MVTec-AD regression, where an already-saturated readout has more to lose than to gain. Beyond these, broader cross-family transferability, adaptation without benchmark-specific training data, multilingual inspection settings, and generative anomaly tasks such as producing defect resolution steps all remain open.