A VLM ANSWER IS NOT AN ANOMALY SCORE: RANK COMPRESSION ACROSS IMAGE AND VIDEO ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Anomaly detection aims to identify observations that deviate from normal patterns. Recent work has increasingly used pretrained vision–language models (VLMs) for training-free image and video anomaly detection without task-specific retraining. Anomaly detection is commonly evaluated by how well anomaly scores rank anomalous images or video frames above normal ones. Generative VLMs, however, assign probabilities to possible answers and then decode a single answer. This decoding step may discard ordering information. We call this loss **decoded-answer rank compression**. To isolate this effect, we compare two ways of scoring the same VLM output: one uses only the decoded answer, while the other computes a probability-weighted score over all possible answers. Probability-weighted scoring consistently outperforms decoded-answer scoring across image and video anomaly detection benchmarks, using different VLMs and answer scales. The mean gains range from 7.66 to 19.95 points on the primary metrics. Most of this gap comes from decoded-answer ties. Breaking decoded-answer ties with answer probabilities recovers at least 95% of the average performance gap on every benchmark. We further investigate how these ties affect evaluation. We find that evaluating tied scores one input at a time makes the reported AUROC depend on input order, and that linear PR interpolation can inflate reported performance. These results reveal that both how VLM outputs are scored and how discrete anomaly scores are evaluated affect reported anomaly-detection performance.

1 INTRODUCTION

In computer vision, anomaly detection (AD) aims to identify deviations from normal patterns, such as defects in images or unusual events in videos. Most image and video anomaly detectors learn a task-specific scoring function from normal examples or labeled anomalies (Luo et al., 2025; Song & Lee, 2026a). Adapting these detectors to new environments or anomaly definitions typically requires additional training. To avoid task-specific retraining, recent work uses pretrained vision–language models (VLMs) for training-free AD (Xu et al., 2025; Zhang et al., 2025a; Zanella et al., 2024). These methods are attractive because they can make anomaly judgments and provide natural language explanations without updating model parameters (Yang et al., 2024; Song et al., 2026).

Image and video AD are commonly evaluated by how well anomaly scores rank anomalous images or video frames above normal ones, using metrics such as the area under the receiver operating characteristic curve (AUROC) and average precision (AP). Generative VLMs, however, assign probabilities to possible answers and then decode a single answer, such as a Yes/No judgment or a numerical anomaly rating (Zhang et al., 2024; Höfer et al., 2025; Ye et al., 2025; Ahn et al., 2026). But if only the decoded answer is used as the anomaly score, inputs with different answer probabilities can receive the same score, discarding their relative ordering.

Consider the simplest case with two possible answers, Yes and No. Suppose that "Yes" indicates an anomaly and that the VLM assigns "Yes" probabilities of 0.56 and 0.94 to two inputs. Both inputs decode to "Yes", even though the answer probabilities rank the second input above the first. We call this many-to-one loss of ordering **decoded-answer rank compression**. We ask *how much useful anomaly-ranking information is lost by decoded-answer scoring and where this loss comes from*.

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.

Contributions. To answer these questions, we compare two scoring rules applied to the same VLM output. Decoded-answer scoring assigns a score using only the decoded answer. Probability-weighted scoring instead averages the answer values using their probabilities. We keep the VLM, visual input, prompt, and answer probabilities fixed, changing only how the same output distribution is converted into an anomaly score.

We evaluate decoded-answer and probability-weighted scoring across image and video AD benchmarks, using different VLMs and answer scales. Probability-weighted scoring consistently outperforms decoded-answer scoring, with mean gains across VLMs and answer scales of 19.95 image-level AUROC points on VisA and 10.97 frame-level AP points on the video benchmark XD-Violence. The scoring-rule gap persists with larger VLMs, alternative prompts, reversed answer polarity, and sampled decoding.

We find most of the gap comes from ties created by decoding. To test this, we preserve the order between different decoded answers and use answer probabilities only to break ties. Across all seven answer scales, this tie-breaking recovers at least 95% of the average primary-metric gap on every benchmark. Finer answer scales reduce rank compression, but even on the finest 91-answer scale (0.0–9.0), the VLMs use only a small fraction of the available score values, leaving many inputs tied.

We also find that common evaluation choices can make this ranking loss appear smaller. Evaluating tied scores one at a time can make the reported ROC area depend on input order (Sultani et al., 2018). Trapezoidal PR-AUC can also inflate reported performance by crediting precision values above those attainable by thresholding the anomaly scores (Wu et al., 2020; Zou et al., 2022). We therefore evaluate tied scores together and distinguish trapezoidal PR-AUC from non-interpolated AP, which averages precision attained at the score's thresholds (Fawcett, 2006; Davis & Goadrich, 2006).

These results have two implications for VLM anomaly detection. First, the scoring rule is part of detector design: when answer probabilities are available, probability-weighted scoring retains useful ordering and improves anomaly ranking without changing the VLM or prompt. Second, how these scores are evaluated also matters: tie handling and PR-area calculation can materially affect reported anomaly-detection performance.

2 Related Work

Vision–language models for visual anomaly detection. Recent work uses pretrained vision–language models for visual anomaly detection in both images and videos, reducing the need for task-specific training. CLIP-based methods typically derive anomaly scores from the similarity between visual features and text descriptions of normality and abnormality (Jeong et al., 2023; Zhou et al., 2024; Li et al., 2024; Wu et al., 2024). More recent methods use generative VLMs to make anomaly judgments or provide explanations (Gu et al., 2024; Xu et al., 2025; Zhang et al., 2025a). Several image and video anomaly-detection methods convert decoded binary judgments or numerical ratings into anomaly scores (Zhang et al., 2024; Höfer et al., 2025; Yang et al., 2024; Zanella et al., 2024; Ye et al., 2025; Ahn et al., 2026; Shao et al., 2025; Li et al., 2026; Song & Lee, 2026b). Our work studies this scoring step directly: by holding the VLM answer distribution fixed, we measure how much anomaly-ranking information is lost when that distribution is reduced to a decoded answer.

Probability-based scoring with language models. Language-model probabilities have been used to construct scalar scores for document ranking, text evaluation, and answer verification (Nogueira et al., 2020; Zhong et al., 2022; Liu et al., 2023; Zhuang et al., 2024; Zhang et al., 2025b). Our question is different: how does converting a VLM answer distribution into an anomaly score affect anomaly ranking? We use probability-weighted scoring as a comparison to measure this effect, analyze where the ranking differences arise, and study how they affect anomaly-detection evaluation.

3 From a VLM Answer Distribution to an Anomaly Score

A VLM produces a distribution over possible answers, while anomaly detection requires a scalar anomaly score. We compare two ways of obtaining this score from the same answer distribution.

Answer distribution. Let x denote the visual input and let the prompt q specify a finite ordered set of possible answers $\mathcal{A} = a_0, \ldots, a_K$. We assign each answer a an anomaly value $v(a) \in [0, 1]$, with larger values indicating stronger anomaly judgments. For Yes/No, for example, $v(\text{No}) = 0$ and $v(\text{Yes}) = 1$.

For each answer, we compute the log-probability of the complete answer string:

$$\ell_x(a) = \log p_\theta(a \mid x, q). \tag{1}$$

For multi-token answers, $\ell_x(a)$ is the sum of the token log-probabilities. We normalize these values over the possible answers:

$$\widetilde{p}_x(a) = \frac{\exp \ell_x(a)}{\sum_{a' \in \mathcal{A}} \exp \ell_x(a')}. \tag{2}$$

We refer to $\widetilde{p}_x$ as the answer distribution.

Two scoring rules. Decoded-answer scoring (DEC) assigns the anomaly value of the most probable answer:

$$a_x^* = \arg\max_{a \in \mathcal{A}} \ell_x(a), \qquad s_x^{\text{DEC}} = v(a_x^*). \tag{3}$$

Probability-weighted scoring (PROB) instead computes the expected anomaly value under the answer distribution:

$$s_x^{\text{PROB}} = \sum_{a \in \mathcal{A}} v(a) \widetilde{p}_x(a). \tag{4}$$

Thus the two scoring rules start from the same answer distribution and differ only in how to convert to anomaly score.

Decoded-answer rank compression. DEC assigns the same anomaly score to all inputs with the same decoded answer. With $|\mathcal{A}|$ possible answers, DEC can therefore produce at most $|\mathcal{A}|$ distinct scores, even when the answer distributions differ across inputs. We call this many-to-one loss of ordering *decoded-answer rank compression*.

4 EVALUATION SETUP

To isolate the effect of the scoring rule, we compute DEC and PROB from the same answer distribution for every input. The VLM, visual input, prompt, and inference precision are identical within each comparison.

Evaluation metrics. We evaluate anomaly scores using AUROC and non-interpolated average precision (AP). For scores s^+ and s^- of randomly sampled anomalous and normal examples, respectively,

$$\text{AUROC}(s) = \Pr(s^+ > s^-) + \frac{1}{2} \Pr(s^+ = s^-). \tag{5}$$

Thus, a tied anomalous–normal pair receives half credit (Fawcett, 2006).

For AP, let $u_1 > \cdots > u_J$ denote the distinct score values. Let p_j and n_j be the numbers of anomalous and normal examples with score u_j, and define

$$\text{TP}_j = \sum_{i \le j} p_i, \qquad \text{FP}_j = \sum_{i \le j} n_i, \qquad P = \sum_j p_j. \tag{6}$$

We compute

$$\text{AP}(s) = \sum_{j=1}^{J} \frac{p_j}{P} \frac{\text{TP}_j}{\text{TP}_j + \text{FP}_j}. \tag{7}$$

All examples with the same score are therefore evaluated together. We distinguish this non-interpolated AP from trapezoidal PR-AUC, which linearly interpolates between precision–recall points (Davis & Goadrich, 2006; Chen et al., 2024).

Table 2: Primary-metric performance on 4 benchmarks averaged over four VLMs (percentage points). Each DEC and PROB is computed from the same answer probability. Δ is PROB $-$ DEC.

	MVTec AD, I-AUROC			VisA, I-AUROC			UCF-Crime, AUROC			XD-Violence, AP		
Scale ($\lvert\mathcal{A}_k\rvert$)	DEC	PROB	Δ	DEC	PROB	Δ	DEC	PROB	Δ	DEC	PROB	Δ
Yes/No (2)	69.63	**88.94**	+19.31	58.44	**81.06**	+22.62	72.20	**85.32**	+13.12	49.44	**68.98**	+19.54
0–1 integer (2)	61.04	**89.67**	+28.63	54.21	**82.89**	+28.68	73.69	**85.42**	+11.73	50.53	**68.08**	+17.55
0–5 integer (6)	77.07	**89.81**	+12.74	66.54	**83.97**	+17.43	79.05	**84.76**	+5.71	57.58	**66.96**	+9.38
0–9 integer (10)	77.19	**89.48**	+12.28	67.10	**83.79**	+16.70	78.71	**84.38**	+5.67	58.71	**66.94**	+8.23
0–1 decimal (11)	70.72	**89.79**	+19.07	62.82	**84.09**	+21.27	74.91	**83.68**	+8.78	57.32	**65.67**	+8.34
0–5 decimal (51)	78.30	**89.74**	+11.44	66.33	**83.37**	+17.04	80.37	**84.52**	+4.14	59.41	**66.54**	+7.12
0–9 decimal (91)	77.92	**89.70**	+11.77	66.53	**82.47**	+15.94	79.73	**84.22**	+4.49	60.22	**66.85**	+6.63
Mean over scales	73.13	**89.59**	+16.46	63.14	**83.09**	+19.95	76.95	**84.61**	+7.66	56.17	**67.15**	+10.97

Benchmarks and evaluation protocol. We evaluate image anomaly detection on MVTec AD (Bergmann et al., 2019) and VisA (Zou et al., 2022). Each image is scored independently with its category name included in the prompt, and we report category-macro image-level AUROC (I-AUROC) and AP (I-AP).

For video anomaly detection, we evaluate UCF-Crime (Sultani et al., 2018) and XD-Violence (Wu et al., 2020). Following VERA (Ye et al., 2025), we score non-overlapping 16-frame segments using frames sampled from their temporal context, assign each segment score to its frames, and compute frame-level AUROC and AP over the test set. Additional details are given in Appendix A.1.

Models and answer scales. We evaluate four frozen VLMs: Qwen3-VL-8B (Bai et al., 2025a), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3.5-8B (Zhu et al., 2025), and MiniCPM-V-4.5-8B (Yu et al., 2026). For each model, we use seven answer scales ranging from binary Yes/No to a 91-answer decimal scale as shown in Table 1. These scales let us compare DEC and PROB from coarse binary outputs to fine-grained numerical ratings. Within each domain, the prompts describe anomalies in the same terms while changing the allowed answers and required response format. The exact prompts are given in Appendix A.2.

Table 1: Answer scales used in our experiments.

Scale	Format	$\lvert\mathcal{A}\rvert$
Yes/No	binary	2
0–1	integer	2
0–5	integer	6
0–9	integer	10
0–1	one decimal	11
0–5	one decimal	51
0–9	one decimal	91

Uncertainty estimates. We report paired 95% nonparametric bootstrap confidence intervals using 10,000 replicates. For image benchmarks, we resample test images within each category before macro-averaging, while for video benchmarks we resample complete videos. Each bootstrap sample is shared between DEC and PROB, preserving the paired comparison between the two scoring rules.

5 EVALUATION

5.1 MAIN RESULTS

Probability-weighted scoring improves performance across all tested settings. PROB outperforms DEC in all 112 model-scale-benchmark comparisons, covering every VLM and answer scale on all four benchmarks (Appendix B.1). Table 2 summarizes the primary-metric results averaged over the four VLMs. Across models and answer scales, the mean improvement ranges from 7.66 percentage points on UCF-Crime to 19.95 points on VisA. The paired 95% bootstrap interval for the mean improvement is strictly above zero on every benchmark . Because each comparison is computed from the same answer distribution, the observed difference comes from the scoring rule.

The improvement holds across image categories and video event classes. We next ask whether the benchmark-level gains are concentrated in only a few categories or event classes. We evaluate this in the Yes/No setting, averaging each category- and class-level metric over the four VLMs. Table 8 shows that PROB outperforms DEC on both I-AUROC and I-AP in all 27 image categories across

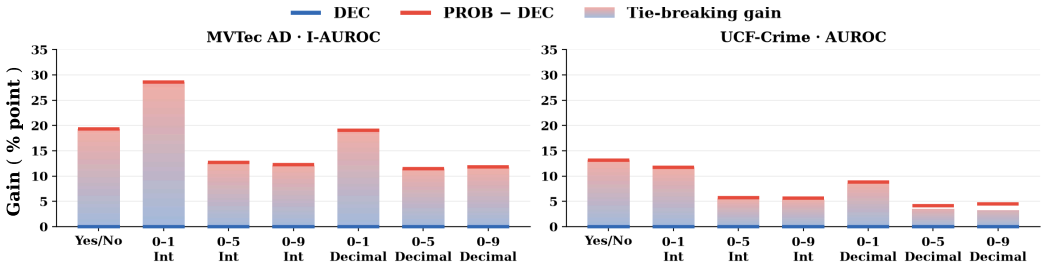

Figure 1: Breaking DEC ties recovers nearly all of the PROB–DEC performance gap. For each answer scale, the blue line marks the DEC baseline at zero gain, the red line marks the full PROB–DEC gain, and the shaded bar marks the tie-breaking gain. Results are averaged over four VLMs.

MVTec AD and VisA. Similarly, Table 9 shows higher AUROC and AP for PROB in all 19 evaluated video event classes across UCF-Crime and XD-Violence. The benchmark-level improvements are therefore not driven by a small subset of categories or event classes.

The improvement persists across model sizes. To test whether this result depends on model size, we repeat the Yes/No comparison with Qwen3-VL and InternVL3.5 models ranging from 2B to 38B parameters. Table 10 shows that PROB outperforms DEC on both metrics in every evaluated model and benchmark comparison. On UCF-Crime, PROB improves AUROC by 28.53 points for Qwen3-VL-2B and by 10.87 points for InternVL3.5-38B. Thus, the improvement is not limited to the 7B and 8B models used in the main evaluation.

The improvement extends to low false-positive rates. AUROC and AP summarize performance across score thresholds, but do not characterize performance at any particular operating point. We therefore compare attainable true-positive rate (TPR) at false-positive-rate (FPR) budgets of 1%, 5%, and 10%. PROB achieves higher TPR on all four benchmarks at every tested FPR budget. At the most restrictive 1% budget, for example, the mean TPR increases from 34.91 to 61.36 on MVTec AD and from 1.00 to 16.46 on UCF-Crime. Complete results are reported in Appendix B.4.

5.2 RANK COMPRESSION

Where does the improvement come from? Compared with DEC, PROB can change the ordering of examples with the same decoded answer or of examples with different decoded answers.

Breaking decoded-answer ties recovers nearly all of the gap. To isolate the first effect, we preserve DEC's ordering between different decoded answers. For examples with the same decoded answer, we break ties by sorting them in descending order of their PROB scores. Under Yes/No, all examples decoded as "Yes" remain above those decoded as "No", while examples within each group are ordered by PROB.

Figure 1 shows how much of the full PROB-DEC performance gap is recovered by breaking decoded-answer ties. For the two binary answer scales, tie-breaking reproduces the PROB ranking by construction. On the five finer scales, PROB can also change the order between examples with different decoded answers. Even so, tie-breaking recovers at least 90.7% of the average primary-metric gap across these scales on every benchmark. Across all seven scales, it recovers at least 95.2% of the average gap on every benchmark. Thus, decoded-answer ties explain nearly all of the average performance gap. Full results are reported in Appendix C.1.

Answer probabilities preserve useful order within decoded-answer ties. To test whether PROB preserves useful information within a decoded-answer tie, we focus on examples decoded as "No" under Yes/No. DEC assigns all of these examples the same score. We sort them by PROB and divide them into ten equally sized groups (deciles). Figure 2 shows that higher-PROB groups contain a larger fraction of anomalies on all four benchmarks.

To quantify this ordering, we compute AUROC within the "No" group for each VLM. On every benchmark, all VLMs achieve above-chance, with values ranging from 71.3 to 89.4 (Table 14). Thus, answer probabilities preserve useful ordering information even among examples decoded as "No".

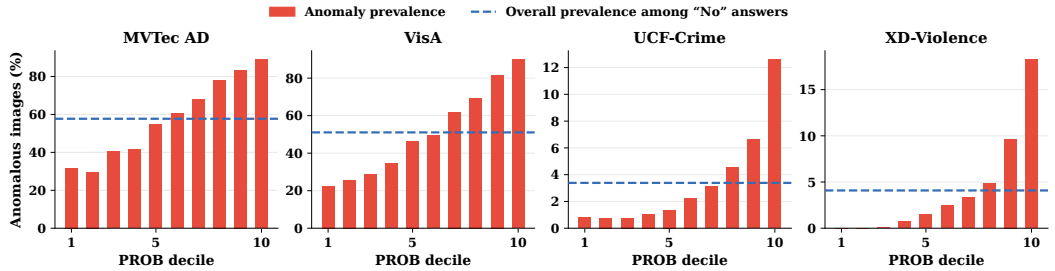

Figure 2: Answer probabilities preserve useful order among inputs decoded as "No". For each VLM, we sort the "No" inputs by PROB and divide them into deciles. Bars show anomaly prevalence within each decile, averaged over four VLMs, and the dashed line shows the corresponding prevalence over all "No" inputs. Higher-PROB deciles contain more anomalies on all four benchmarks.

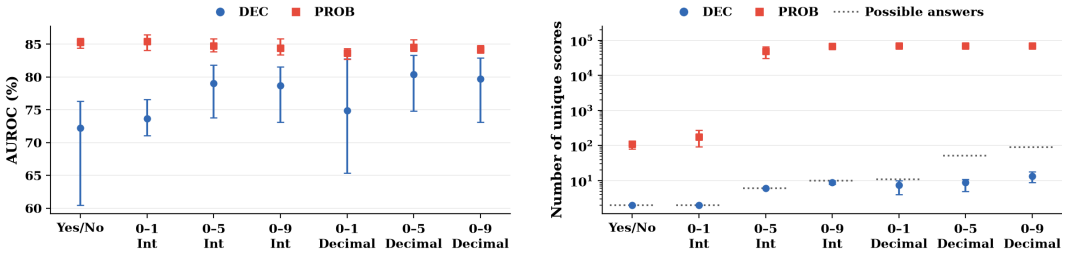

Figure 3: Answer-scale performance and number of distinct scores on UCF-Crime. Left: frame-level AUROC across the seven answer scales. Right: the number of distinct scores across the 69,634 UCF-Crime segments, shown on a logarithmic scale. Dotted gray lines indicate the number of possible answers $|\mathcal{A}_k|$. Markers report the mean over four VLMs, and vertical bars span the minimum and maximum across models.

Even an oracle binary threshold does not close the gap. One possible explanation for the PROB–DEC gap is that decoded scoring uses poorly placed decision boundaries rather than discarding useful ordering information. We test this explanation in the binary setting, where DEC is exactly obtained by thresholding PROB at 0.5. For each VLM and benchmark, we use the test labels to choose the threshold that maximizes the primary metric of the resulting binary scores. This oracle threshold improves binary scoring by 0.89–27.58 points, but a residual gap of 0.41–17.31 points remains across all 16 model–benchmark settings (Table 15). Thus, in the binary setting, poor threshold placement explains part of the performance loss but not the full PROB–DEC gap.

5.3 ANSWER SCALES

If decoded-answer ties account for most of the performance gap, a natural way to reduce rank compression is to use a finer answer scale. We therefore compare seven answer scales ranging from 2 to 91 possible answers. We ask whether finer scales improve DEC and narrow the PROB–DEC gap, and whether the VLMs actually use the additional answer choices.

Finer answer scales help, but most additional answers are not used. Averaged over the four VLMs, all five nonbinary scales improve DEC over Yes/No and narrow the PROB–DEC gap on every benchmark's primary metric (Table 2). Figure 3 illustrates this behavior on UCF-Crime and shows how the number of distinct scores changes across answer scales. The improvement in DEC is not monotonic in the number of possible answers. To understand why, we examine how many distinct DEC scores the models actually produce. Under the 6- and 10-answer integer scales, DEC uses all or nearly all possible answers. The decimal scales allow 11, 51, and 91 possible answers, but DEC produces only 7.5, 8.8, and 13.5 distinct scores on average, respectively. In contrast, PROB produces 69,538–69,542 distinct scores for the 69,634 UCF-Crime segments under the same decimal scales. The other three benchmarks show the same pattern (Appendix C.4). Thus, increasing the number of possible answers does not translate into a comparable increase in the number of distinct DEC scores.

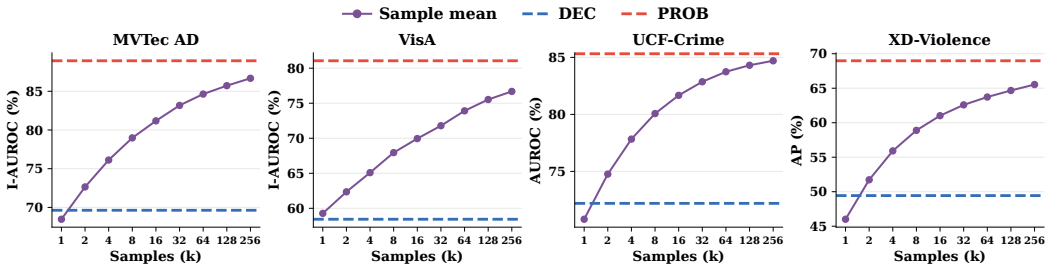

Figure 4: Averaging sampled answers narrows the gap to PROB on all four benchmarks. For each input, the sample mean is the fraction of Yes answers among k draws from the saved Yes/No answer distribution. The sample count k doubles from 1 to 256. Each point averages the benchmark metric over three sampling seeds for each VLM and then over four VLMs. Dashed lines show the means over four VLMs for DEC and PROB.

More distinct scores do not imply better ranking. The previous result does not imply that producing more distinct scores is itself sufficient for better ranking performance. We test this using the remaining ties in binary PROB scores, comparing bfloat16 and float32 runs of Qwen3-VL-8B with Yes/No. On UCF-Crime, the number of distinct segment scores increases from 137 to 65,885, while AUROC changes only from 85.26 to 85.27. On MVTec AD, the mean number of distinct scores per category increases from 53.1 to 115.0, while I-AUROC changes from 91.51 to 91.78. Thus, the number of distinct scores alone does not determine how well those scores rank anomalous examples above normal ones (Appendix C.4).

5.4 ALTERNATIVE SCORING STRATEGIES

Can the remaining PROB–DEC gap be removed by changing other choices in how the VLM output is converted into an anomaly score? We test alternative answer selection and aggregation rules, temporal smoothing, and prompt and output variations.

Alternative answer selection does not close the gap. We first ask whether the gap is specific to selecting the most probable answer. Using the saved Yes/No answer distributions, we simulate temperature sampling, nucleus sampling, and majority voting on both video benchmarks. Each strategy still reduces the model output to a single binary score. Averaged over the four VLMs, none outperforms DEC on either benchmark, and all remain below PROB. For example, the best tested strategy on UCF-Crime reaches 72.13 AUROC, compared with 72.20 for DEC and 85.32 for PROB.

We then average the sampled answer values instead of selecting a single answer. For Yes/No, the fraction of "Yes" answers among k samples estimates the probability of "Yes", which PROB uses directly. Figure 4 shows that performance improves with the number of samples on all four benchmarks. On MVTec AD, mean I-AUROC increases from 81.18 with 16 samples to 86.68 with 256 samples, compared with 88.94 for PROB. Even with 256 samples, mean performance remains below PROB on every benchmark (Appendix D.1). Thus, changing which single answer is selected does not recover the lost ordering, while averaging multiple answers progressively approaches probability-weighted scoring.

Temporal smoothing narrows but does not close the mean gap. Temporal smoothing can reduce decoded-score ties by averaging scores from neighboring video segments. We apply the same Gaussian smoothing to DEC and PROB within each video, using $\sigma \in \{1, \ldots, 9\}$, and average results over the four VLMs and seven answer scales. As smoothing increases, DEC produces more distinct scores and the mean PROB–DEC gap narrows. At $\sigma = 9$, the mean number of distinct DEC scores on UCF-Crime increases from 7.0 to about 31,745, while the mean AUROC gap falls from 7.66 to 2.28 points. On XD-Violence, the mean AP gap falls from 10.97 to 3.01 points. PROB nevertheless remains higher on average at every tested value of σ on both benchmarks (Appendix D.2).

Prompt and output variations do not reverse the comparison. We finally test whether the PROB–DEC comparison depends on the base prompt or output format. Across three Yes/No paraphrases

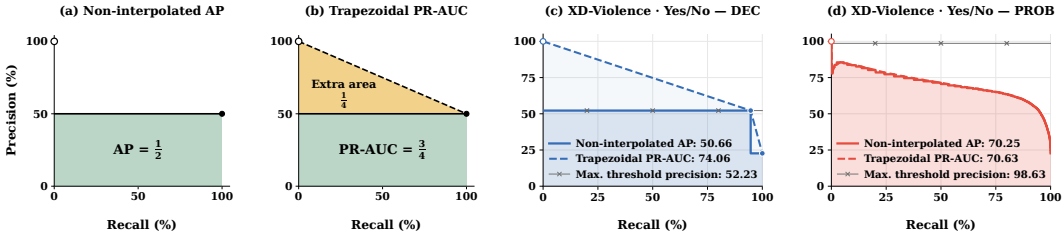

Figure 5: (a,b) Schematic comparison of non-interpolated AP and trapezoidal PR-AUC calculations. (c,d) Qwen3-VL-8B with two scoring methods on XD-Violence. Both calculations evaluate equal scores together. Open circles mark the conventional endpoint at zero recall. Gray lines mark maximum precision across thresholds that predict at least one anomaly.

with Qwen3-VL-8B and InternVL3.5-8B on MVTec AD and UCF-Crime, PROB remains higher on both metrics in every comparison (Appendix D.3). The same holds when we reverse the question and ask whether the example is normal, so that "No" indicates an anomaly (Appendix D.3). We also add a brief factual description before or after the answer, using Qwen3-VL-8B and MiniCPM-V-4.5 on both binary scales across all four benchmarks(Appendix D.4). When the description comes first, both scoring rules use the answer distribution conditioned on the same description. These changes affect absolute performance, but PROB remains higher in all comparisons across both output orders.

6 How Ties Affect Evaluation

Decoded-answer scoring often assigns the same anomaly score to many examples. This raises two evaluation choices: how equal scores are handled and how area is computed from precision–recall points. We examine these two choices.

Evaluate equal scores together. When multiple frames receive the same anomaly score, the detector provides no ordering among them. As a score threshold passes that value, all tied frames should be included together. The UCF-Crime evaluation instead processes tied frames one at a time and updates the ROC curve after each frame (Appendix E.1). This treats tied frames as if they had an order, even though their scores do not distinguish them. It also creates intermediate ROC points that do not correspond to any score threshold. Because these points depend on the order within each tie, the resulting ROC area can also depend on that order (Fawcett, 2006).

Let τ denote the fraction of anomalous–normal pairs that receive the same score. Let AUROC^- and AUROC^+ denote the smallest and largest item-wise ROC areas obtainable by ordering frames within ties. Relative to the half-credit AUROC in Equation (5), these extrema are

$$\mathrm{AUROC}^- = \mathrm{AUROC} - \frac{\tau}{2}, \qquad \mathrm{AUROC}^+ = \mathrm{AUROC} + \frac{\tau}{2}. \tag{8}$$

The difference between these extrema is therefore exactly τ. Averaged over the four VLMs and seven answer scales on UCF-Crime, this difference is 31.32 percentage points for DEC and 0.23 points for PROB (Table 22). We therefore compute AUROC by evaluating equal scores together, using the half-credit definition in Equation (5).

Linear PR interpolation can inflate reported performance. Even when equal scores are evaluated together, AP and trapezoidal PR-AUC can differ on the same precision–recall points. Let $q_j = \mathrm{TP}_j/(\mathrm{TP}_j + \mathrm{FP}_j)$ denote precision at threshold u_j. In Equation (7), the recall increase at this threshold is p_j/P, and AP uses q_j for this increase. Trapezoidal PR-AUC instead averages q_{j-1} and q_j, with the conventional endpoint $q_0 = 1$ at zero recall (Davis & Goadrich, 2006; Chen et al., 2024). The difference between the two calculations is

$$\text{PR-AUC}_{\text{trapezoids}}(s) - \mathrm{AP}(s) = \frac{1}{2}\sum_{j=1}^{J} \frac{p_j}{P}(q_{j-1} - q_j). \tag{9}$$

Figure 5(a,b) illustrates the corresponding rectangles and trapezoids. When precision drops as recall increases, the trapezoid adds area above the AP rectangle, with more added area for larger recall increases and larger precision drops.

Figure 5(c,d) shows a concrete example for Qwen3-VL-8B with Yes/No on XD-Violence. For DEC, no threshold that predicts at least one frame as anomalous attains precision above 52.23%. Yet its trapezoidal PR-AUC is 74.06%, compared with an AP of 50.66%. The line from the conventional $(0, 1)$ endpoint to the first measured PR point spans a large recall increase and includes precision values above this attainable maximum. For PROB, AP and trapezoidal PR-AUC are 70.25% and 70.63%, respectively. Thus, AP favors PROB, whereas trapezoidal PR-AUC favors DEC, without changing either set of anomaly scores.

This reversal is not specific to this example. Holding the anomaly scores fixed, we compare AP and trapezoidal PR-AUC for each answer scale, averaging over the four VLMs. On both image benchmarks, trapezoidal PR-AUC narrows the mean PROB–DEC difference, but PROB remains higher under every scale. On the video benchmarks, it reverses the mean comparison under four of seven scales on UCF-Crime and six of seven on XD-Violence (Table 23).

7 DISCUSSION AND LIMITATIONS

Decoded-answer performance can understate useful anomaly information. An anomaly score derived from a decoded answer does not necessarily reflect all of the useful ordering information in the VLM output. Our results show that examples receiving the same decoded answer can still differ systematically in their answer probabilities. Poor performance with decoded-answer scoring should therefore not, by itself, be interpreted as evidence that the VLM provides little useful information for anomaly detection.

Tied scores make evaluation choices consequential. When examples receive the same anomaly score, different evaluation conventions can assign different credit without any change in the scores. Processing tied scores one at a time introduces an ordering that the scores do not specify and can change AUROC. Linear interpolation in precision–recall space can likewise add area between threshold points and, in our experiments, even reverse method comparisons. These differences arise from the evaluation procedure rather than from changes in the anomaly scores themselves. Evaluation results should therefore state how ties are handled and distinguish non-interpolated AP from trapezoidal PR-AUC.

Access to answer probabilities. Probability-weighted scoring requires the probabilities assigned to the possible answers. Our experiments use open VLMs for which we can compute these probabilities. We do not evaluate proprietary model APIs, where the available model outputs may differ. Our sampled-answer experiments draw from the answer distributions computed for the open VLMs and therefore do not evaluate repeated sampling through such APIs.

Comparisons between answer scales. Within an answer scale, DEC and PROB operate on the same answer distribution and differ only in how it is converted into an anomaly score. Across answer scales, however, the prompt, answer format, and meaning of numerical answers can also change. Our comparisons across scales therefore measure the combined effect of these changes, not the effect of the number of possible answers alone.

8 CONCLUSION

We show that decoded-answer rank compression can substantially degrade anomaly ranking. Across the image and video benchmarks we evaluate, decoded-answer scoring consistently yields lower AUROC and AP than probability-weighted scoring. Using answer probabilities only to break decoded-answer ties recovers at least 95% of the average primary-metric gap on every benchmark. We also show that trapezoidal PR-AUC can mask this ranking loss and even reverse the comparison between DEC and PROB without changing the anomaly scores. Evaluations should group equal scores and distinguish non-interpolated AP from trapezoidal PR-AUC.

These results show that converting a VLM output into an anomaly score is not a neutral post-processing step. This conversion can discard useful ordering information, and evaluation choices can further change how that loss appears in reported performance. Both score construction and evaluation are therefore important when interpreting the performance of VLM-based anomaly detectors.

AI USE STATEMENT

Generative AI tools were used to assist with sentence editing, grammar checking, typographical checks, and isolated utility functions in the inference code. The authors reviewed every AI-assisted change, verified the mathematical definitions and all reported values against the evaluation outputs, and take responsibility for the final text, claims, code, and artifacts. No reported experimental result was generated or selected by a language model.

ETHICS STATEMENT

This work evaluates scoring rules on public image and video anomaly-detection benchmarks and redistributes no benchmark data. MVTec AD and VisA depict industrial objects and textures. UCF-Crime and XD-Violence include real events involving people. We use them only to evaluate aggregate frame-level rankings and do not identify, profile, or characterize individuals. We do not release generated descriptions of identifiable people. As with any anomaly detector, deployment in industrial inspection or surveillance requires domain validation, human oversight, and safeguards against false alarms and misuse.

REPRODUCIBILITY STATEMENT

Every main result uses frozen public checkpoints and deterministic scoring, with no training and, outside the explicitly seeded sampling controls, no stochastic decoding. The checkpoints are `Qwen/Qwen3-VL-8B-Instruct`, `Qwen/Qwen2.5-VL-7B-Instruct`, `OpenGVLab/InternVL3_5-8B`, and `openbmb/MiniCPM-V-4_5`. Each image is scored independently and image metrics are computed within category before macro-averaging. For video, non-overlapping 16-frame units are judged from eight frames sampled uniformly over a centered 300-frame window and expanded back to the annotation grid (Appendix A.1). The released artifact will include the exact domain-specific prompts, answer strings, launch commands, and evaluation scripts; Appendix A.2 documents the prompt structure and all currently recorded variants.

Both scores are computed from one set of answer sequence likelihoods per visual unit. Candidate likelihoods are evaluated without special tokens; multi-token candidates use the sum of their token log-likelihoods. Evaluation uses the grouped-threshold AP of Equation (7) and tie-aware AUROC of Equation (5). Video bootstrap replicates resample whole videos; image replicates resample normal and anomalous images separately within category; every replicate stream is shared across the paired scores, models, and scales. Code and exact commands for every table will be released.

REFERENCES

Sunghyun Ahn, Youngwan Jo, Kijung Lee, Sein Kwon, Inpyo Hong, and Sanghyun Park. Anyanomaly: Zero-shot customizable video anomaly detection with lvlm. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3026–3035, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL `https://arxiv.org/abs/2511.21631`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL `https://arxiv.org/abs/2502.13923`.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592. IEEE, 2019.

Wenyu Chen, Chen Miao, Zhenghao Zhang, Cathy Sin-Hang Fung, Ran Wang, Yizhen Chen, Yan Qian, Lixin Cheng, Kevin Y Yip, Stephen Kwok-Wing Tsui, et al. Commonly used software tools produce conflicting and overly-optimistic auprc values. *Genome Biology*, 25(1):118, 2024.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 1932–1940, 2024.

Sebastian Höfer, Dorian F Henning, Artemij Amiranashvili, Douglas Morrison, Mariliza Tzes, Ingmar Posner, Marc Matvienko, Alessandro Rennola, and Anton Milan. Kaputt: A large-scale dataset for visual defect detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24224–24233. IEEE, 2025.

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

Wenlong Li, Yifei Xu, Yuan Rao, Zhenhua Wang, and Shuiguang Deng. Vadtree: Explainable training-free video anomaly detection via hierarchical granularity-aware tree. *Advances in Neural Information Processing Systems*, 38:148372–148404, 2026.

Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16848–16858. IEEE, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Wei Luo, Yunkang Cao, Haiming Yao, Xiaotian Zhang, Jianan Lou, Yuqi Cheng, Weiming Shen, and Wenyong Yu. Exploring intrinsic normal prototypes within a single image for universal anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 9974–9983, June 2025.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 708–718, 2020.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 2586–2595, 2025.

Inpyo Song and Jangwon Lee. Bounding-box trajectories matter for video anomaly detection. *arXiv preprint arXiv:2605.21957*, 2026a.

Inpyo Song and Jangwon Lee. Rethinking open-world video anomaly detection: Diagnosing definition blindness. *arXiv preprint arXiv:2607.20780*, 2026b.

Inpyo Song, Minjun Joo, Joonhyung Kwon, Eunji Jeon, and Jangwon Lee. Instance-aligned captions for explainable video anomaly detection. *arXiv preprint arXiv:2601.08155*, 2026.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pp. 322–339. Springer, 2020.

Peng Wu, Xuerong Zhou, Guansong Pang, Lingru Zhou, Qingsen Yan, Peng Wang, and Yanning Zhang. Vadclip: Adapting vision-language models for weakly supervised video anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 6074–6082, 2024.

Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20370–20382. IEEE, 2025.

Yuchen Yang, Kwonjoon Lee, Behzad Dariush, Yinzhi Cao, and Shao-Yuan Lo. Follow the rules: reasoning for video anomaly detection with large language models. In *European Conference on Computer Vision*, pp. 304–322. Springer, 2024.

Muchao Ye, Weiyang Liu, and Pan He. Vera: Explainable video anomaly detection via verbalized learning of vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8679–8688. IEEE, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Ranchi Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11715, 2026.

Luca Zanella, Willi Menapace, Massimiliano Mancini, Yiming Wang, and Elisa Ricci. Harnessing large language models for training-free video anomaly detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527–18536. IEEE, 2024.

Jiangning Zhang, Haoyang He, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, Lei Xie, and Yong Liu. Gpt-4v-ad: Exploring grounding potential of vqa-oriented gpt-4v for zero-shot anomaly detection. In *International Joint Conference on Artificial Intelligence*, pp. 3–16. Springer, 2024.

Jinjin Zhang, Guodong Wang, Yizhou Jin, and Di Huang. Towards training-free anomaly detection with vision and language foundation models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15204–15213. IEEE, 2025a.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Seyed Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, volume 2025, pp. 12476–12505, 2025b.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2023–2038, 2022.

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, volume 2024, pp. 49705–49737, 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.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond yes and no: Improving zero-shot llm rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies (volume 2: short papers)*, pp. 358–370, 2024.

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European conference on computer vision*, pp. 392–408. Springer, 2022.

A Experimental Details

A.1 Inputs and scoring computation

We use the frozen checkpoints `Qwen/Qwen3-VL-8B-Instruct`, `Qwen/Qwen2.5-VL-7B-Instruct`, `OpenGVLab/InternVL3_5-8B`, and `openbmb/MiniCPM-V-4_5`. All main runs use bfloat16 inference without temporal smoothing.

For images, we use 1,725 MVTec AD test images across 15 categories (467 normal and 1,258 anomalous) and 2,162 VisA test images across 12 categories (962 normal and 1,200 anomalous). MVTec AD uses the category test directories with `good` labeled normal. VisA uses test rows in `split_csv/1cls.csv`. Each input is EXIF-oriented, converted to RGB, and passed to the model's image processor as a single image. The prompt includes the category name but no labels, defect subtypes, paths, normal-reference images, or pixel masks. Image AUROC and AP are computed within category and macro-averaged.

For videos, the UCF-Crime and XD-Violence test splits contain 290 and 800 videos. We score 1,111,808 and 2,335,801 frames, including 88,094 and 529,797 anomalous frames, respectively. Non-overlapping 16-frame segments give 69,634 and 146,449 predictions. Each segment uses eight frames sampled uniformly over a centered 300-frame window clipped at video boundaries. Qwen3-VL, Qwen2.5-VL, and InternVL3.5 receive an ordered image sequence. MiniCPM-V-4.5 receives the same frames in a 4×2 contact sheet. We repeat each segment score over its frames and concatenate videos before evaluation.

Single-token answers use logits at the answer position. Multi-token answers use teacher-forced continuation scoring, summing token log-probabilities without adding special tokens. We normalize these full-sequence probabilities over the answer set at $T = 1$ and compute both scores from the same distribution. The main comparison does not sample answers.

Agreement with unconstrained decoding. For Qwen3-VL-8B and InternVL3.5-8B with Yes/No, the unconstrained top-1 token agrees with DEC on every UCF-Crime segment and MVTec AD image. Mean probability mass on Yes/No is 0.9999999 and 0.9999750 on UCF-Crime, respectively. Minimum mass exceeds 0.99978 on MVTec AD. This check validates the constrained binary answer selection for these two models and benchmarks.

A.2 Prompts and answer scales

The following templates instantiate the seven answer scales in Table 1. Integer scales use $M \in \{1, 5, 9\}$ and decimal scales use $M \in \{1.0, 5.0, 9.0\}$. In image prompts, `<category>` replaces underscores with spaces and inserts a space before a trailing numeric suffix, as in `metal nut` and `macaroni 1`. The article "a" is retained for every category.

Image Yes/No.

```
You are inspecting an image of a <category>.
Assess whether the object or material is anomalous or
defective, for example damaged, contaminated, malformed,
misplaced, or missing a part.
Answer with exactly one word:  Yes or No.
```

Image integer scales.

```
You are inspecting an image of a <category>.
Assess whether the object or material is anomalous or
defective, for example damaged, contaminated, malformed,
misplaced, or missing a part.
Rate the anomaly severity from 0 to M.
0 means completely normal.
M means clearly anomalous or defective.
Answer with exactly one integer from 0 to M and no other
text.
```

Image one-decimal scales.

```
You are inspecting an image of a <category>.
Assess whether the object or material is anomalous or
defective, for example damaged, contaminated, malformed,
misplaced, or missing a part.
Rate the anomaly severity from 0.0 to M.
0.0 means completely normal.
M means clearly anomalous or defective.
Answer with exactly one float rounded to one decimal place
from 0.0 to M and no other text.
```

The image base prompt version is `industrial-query-v1`. The decoder-agreement and float32 controls reuse its Yes/No wording.

Video Yes/No.

```
You are reviewing ordered frames from a surveillance video
clip.
Is the clip anomalous, suspicious, violent, criminal, or
dangerous?
Answer with exactly one word:  Yes or No.
```

Video integer scales.

```
You are reviewing ordered frames from a surveillance video
clip.
Rate how anomalous, suspicious, violent, criminal, or
dangerous the clip is.
Use a single integer from 0 to M:
0 means completely normal.
M means clearly anomalous or dangerous.
Answer with exactly one integer from 0 to M and no other
text.
```

Video one-decimal scales.

```
You are reviewing ordered frames from a surveillance video
clip.
Rate how anomalous, suspicious, violent, criminal, or
dangerous the clip is on a scale from 0.0 to M.
0.0 means completely normal.
M means clearly anomalous or dangerous.
Answer with exactly one float rounded to one decimal place
from 0.0 to M and no other text.
```

A.3 Paired uncertainty estimates

We use 10,000 nonparametric bootstrap replicates shared across scoring rules, models, and answer scales. Video replicates resample complete videos with replacement and recompute frame-level metrics. Image replicates keep categories fixed, resample normal and anomalous images separately within each category, and macro-average the recomputed metrics. The image seed is 20260903 with deterministic dataset/category offsets. The intervals measure test-set sampling uncertainty conditional on the evaluated models and categories. Table 7 reports the primary-metric intervals.

B Complete Results

B.1 Model and answer-scale comparisons

Tables 3 to 6 report both metrics for every main model–scale–benchmark configuration. All 224 paired differences favor PROB. Table 7 gives the corresponding scale-averaged primary-metric gains and confidence intervals.

Table 3: Category-macro image-level performance on MVTec AD for every model and answer scale (%). Each category is weighted equally. PROB exceeds DEC in every cell.

		Yes/No		0–1 int		0–5 int		0–9 int		0–1 float		0–5 float		0–9 float	
Metric	Model	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB
I-AUROC	Qwen3-VL-8B	75.05	91.51	67.96	91.46	78.59	90.69	78.66	90.76	82.08	91.11	80.03	90.90	78.56	90.81
	Qwen2.5-VL-7B	59.26	90.82	52.05	90.29	81.06	91.39	79.79	91.17	73.81	91.28	81.23	91.37	79.64	91.52
	InternVL3.5-8B	69.47	84.18	51.94	83.74	69.84	84.29	71.26	83.77	70.53	84.74	70.76	84.12	72.11	83.62
	MiniCPM-V-4.5-8B	74.73	89.25	72.20	93.18	78.79	92.88	79.07	92.21	56.46	92.03	81.18	92.57	81.38	92.85
	Mean	69.63	**88.94**	61.04	**89.67**	77.07	**89.81**	77.19	**89.48**	70.72	**89.79**	78.30	**89.74**	77.92	**89.70**
I-AP	Qwen3-VL-8B	85.56	95.67	81.87	95.55	87.62	95.38	87.47	95.52	89.16	95.75	88.43	95.48	87.84	95.51
	Qwen2.5-VL-7B	77.44	95.56	73.44	95.30	88.27	95.88	87.61	95.77	84.58	95.83	88.48	95.92	87.81	95.93
	InternVL3.5-8B	82.16	92.11	73.41	91.90	82.56	92.55	83.22	92.11	83.12	92.62	83.36	92.30	84.13	91.80
	MiniCPM-V-4.5-8B	85.51	95.52	84.30	96.98	87.47	96.88	87.54	96.62	76.04	96.88	88.92	96.89	89.11	96.97
	Mean	82.67	**94.71**	78.25	**94.93**	86.48	**95.17**	86.46	**95.01**	83.22	**95.27**	87.30	**95.15**	87.22	**95.05**

Table 4: Category-macro image-level performance on VisA for every model and answer scale (%). Each category is weighted equally. PROB exceeds DEC in every cell.

		Yes/No		0–1 int		0–5 int		0–9 int		0–1 float		0–5 float		0–9 float	
Metric	Model	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB
I-AUROC	Qwen3-VL-8B	62.00	87.23	58.04	87.79	65.60	87.89	66.24	87.68	71.09	87.61	65.04	87.56	64.76	87.30
	Qwen2.5-VL-7B	53.96	81.27	49.96	78.74	71.64	82.23	71.19	81.98	64.31	82.29	70.19	81.89	68.83	81.85
	InternVL3.5-8B	58.42	79.47	50.75	79.52	63.00	79.99	64.79	79.67	63.97	79.64	63.66	77.90	63.74	74.70
	MiniCPM-V-4.5-8B	59.38	76.27	58.08	85.50	65.91	85.77	66.17	85.84	51.92	86.82	66.45	86.13	68.78	86.02
	Mean	58.44	**81.06**	54.21	**82.89**	66.54	**83.97**	67.10	**83.79**	62.82	**84.09**	66.33	**83.37**	66.53	**82.47**
I-AP	Qwen3-VL-8B	66.40	89.12	62.88	89.28	69.18	89.81	69.59	89.73	73.27	89.71	68.61	89.45	68.32	89.39
	Qwen2.5-VL-7B	59.84	84.08	56.58	81.64	73.15	85.20	72.94	84.95	67.73	85.29	71.95	84.97	71.19	85.09
	InternVL3.5-8B	63.24	83.07	57.08	82.95	66.73	83.58	68.04	83.26	67.36	83.32	67.56	81.91	67.57	80.01
	MiniCPM-V-4.5-8B	63.59	81.14	62.62	87.51	69.07	87.99	69.02	87.81	57.83	88.39	69.62	88.23	71.67	88.28
	Mean	63.27	**84.35**	59.79	**85.34**	69.53	**86.64**	69.90	**86.44**	66.55	**86.68**	69.43	**86.14**	69.69	**85.69**

Table 5: Frame-level performance on UCF-Crime for every model and answer scale (%). PROB exceeds DEC in every cell. Underline marks the winner within each pair, and bold marks the winner in the Mean row.

		Yes/No		0–1 int		0–5 int		0–9 int		0–1 float		0–5 float		0–9 float	
Metric	Model	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB
AUROC	Qwen3-VL-8B	76.28	85.26	75.23	85.49	78.93	84.43	78.97	83.90	82.76	84.31	82.85	84.20	82.91	83.90
	Qwen2.5-VL-7B	60.47	85.64	71.08	85.61	73.78	85.81	73.10	85.87	69.88	82.70	74.81	84.20	73.11	84.42
	InternVL3.5-8B	75.97	85.93	71.83	86.46	81.79	84.96	81.19	84.34	81.65	84.37	83.33	85.73	81.68	84.85
	MiniCPM-V-4.5-8B	76.07	84.45	76.60	84.10	81.70	83.85	81.57	83.40	65.34	83.35	80.50	83.94	81.23	83.70
	Mean	72.20	**85.32**	73.69	**85.42**	79.05	**84.76**	78.71	**84.38**	74.91	**83.68**	80.37	**84.52**	79.73	**84.22**
AP	Qwen3-VL-8B	19.91	40.87	18.98	39.85	26.47	38.90	27.12	37.16	30.52	37.41	31.49	37.34	31.86	36.36
	Qwen2.5-VL-7B	17.91	37.46	23.13	39.44	24.73	39.79	23.99	39.04	24.22	38.46	26.49	39.43	24.32	39.11
	InternVL3.5-8B	22.70	39.39	23.39	39.00	30.15	39.67	29.45	38.38	31.42	39.67	31.25	41.52	31.88	39.72
	MiniCPM-V-4.5-8B	23.34	41.79	20.08	39.72	31.37	40.34	26.54	38.46	22.46	37.09	31.15	40.77	32.24	39.42
	Mean	20.96	**39.88**	21.39	**39.50**	28.18	**39.67**	26.77	**38.26**	27.16	**38.16**	30.10	**39.77**	30.07	**38.65**

Table 6: Frame-level performance on XD-Violence for every model and answer scale (%). PROB exceeds DEC in every cell. Underline marks the winner within each pair, and bold marks the winner in the Mean row.

		Yes/No		0–1 int		0–5 int		0–9 int		0–1 float		0–5 float		0–9 float	
Metric	Model	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB	DEC	PROB
AUROC	Qwen3-VL-8B	84.64	91.40	80.35	91.20	86.78	90.17	87.16	89.85	88.03	89.30	88.74	89.62	88.82	89.66
	Qwen2.5-VL-7B	64.33	88.62	80.25	88.87	83.45	88.50	83.04	88.70	80.59	88.43	84.84	88.36	84.83	88.85
	InternVL3.5-8B	85.31	91.67	85.91	91.59	89.15	91.76	90.22	91.73	89.07	91.35	88.84	91.45	89.38	91.44
	MiniCPM-V-4.5-8B	84.92	91.51	84.42	91.20	88.33	90.66	88.35	90.73	79.85	90.16	87.93	90.85	88.41	90.72
	Mean	79.80	**90.80**	82.73	**90.71**	86.93	**90.27**	87.20	**90.25**	84.39	**89.81**	87.59	**90.07**	87.86	**90.17**
AP	Qwen3-VL-8B	50.66	70.25	43.35	69.42	54.60	66.88	55.66	66.01	57.64	63.17	59.42	63.58	59.84	63.45
	Qwen2.5-VL-7B	36.81	59.19	51.13	61.75	54.28	62.11	53.91	62.22	53.77	63.22	56.28	62.79	55.89	64.07
	InternVL3.5-8B	55.32	73.31	56.61	69.31	60.81	69.13	64.37	70.12	63.97	69.49	60.88	69.46	63.86	70.87
	MiniCPM-V-4.5-8B	54.97	73.17	51.01	71.84	60.64	69.73	60.91	69.40	53.92	66.80	61.06	70.31	61.30	69.03
	Mean	49.44	**68.98**	50.53	**68.08**	57.58	**66.96**	58.71	**66.94**	57.32	**65.67**	59.41	**66.54**	60.22	**66.85**

Table 7: Primary-metric gains and paired 95% bootstrap intervals in percentage points, using 10,000 replicates. Each scale averages four models. The final row averages 28 configurations.

Scale	MVTec AD	VisA	UCF-Crime	XD-Violence
Yes/No	$+19.31$ $[18.21, 20.39]$	$+22.62$ $[21.53, 23.71]$	$+13.12$ $[10.08, 15.88]$	$+19.54$ $[17.61, 21.40]$
0–1 int	$+28.63$ $[27.58, 29.65]$	$+28.68$ $[27.59, 29.75]$	$+11.73$ $[9.25, 14.36]$	$+17.55$ $[15.23, 19.84]$
0–5 int	$+12.74$ $[11.87, 13.62]$	$+17.43$ $[16.45, 18.42]$	$+5.71$ $[3.59, 7.87]$	$+9.38$ $[7.39, 11.47]$
0–9 int	$+12.28$ $[11.40, 13.17]$	$+16.70$ $[15.69, 17.71]$	$+5.67$ $[4.01, 7.42]$	$+8.23$ $[6.61, 9.89]$
0–1 float	$+19.07$ $[18.02, 20.09]$	$+21.27$ $[20.26, 22.25]$	$+8.78$ $[6.05, 11.23]$	$+8.34$ $[7.06, 9.60]$
0–5 float	$+11.44$ $[10.60, 12.29]$	$+17.04$ $[15.99, 18.09]$	$+4.14$ $[2.47, 5.89]$	$+7.12$ $[5.53, 8.72]$
0–9 float	$+11.77$ $[10.91, 12.64]$	$+15.94$ $[14.94, 16.94]$	$+4.49$ $[2.71, 6.45]$	$+6.63$ $[5.00, 8.25]$
All 28 pairs	$+16.46$ $[15.67, 17.26]$	$+19.95$ $[19.02, 20.86]$	$+7.66$ $[5.56, 9.71]$	$+10.97$ $[9.37, 12.58]$

Table 8: Yes/No image results (%), averaged over four models. Each category has equal weight within its benchmark.

	I-AUROC			I-AP		
Class/category	DEC	PROB	Δ	DEC	PROB	Δ
MVTec AD						
Bottle	76.73	88.88	+12.15	88.61	96.52	+7.91
Cable	51.63	82.16	+30.53	62.59	88.24	+25.65
Capsule	57.68	81.75	+24.06	85.25	95.33	+10.08
Carpet	83.71	99.37	+15.67	92.20	99.79	+7.59
Grid	85.09	99.24	+14.15	91.97	99.69	+7.72
Hazelnut	73.35	92.28	+18.93	79.96	95.51	+15.55
Leather	83.56	99.65	+16.09	91.51	99.86	+8.35
Metal nut	60.62	91.10	+30.48	84.93	97.64	+12.71
Pill	64.36	76.99	+12.63	88.90	94.99	+6.09
Screw	56.51	86.89	+30.38	77.71	94.67	+16.96
Tile	74.26	98.45	+24.19	85.48	99.33	+13.86
Toothbrush	74.38	77.78	+3.40	84.54	89.91	+5.37
Transistor	60.83	74.92	+14.08	52.38	73.91	+21.53
Wood	77.50	99.68	+22.18	89.18	99.87	+10.70
Zipper	64.24	84.97	+20.73	84.76	95.41	+10.65
VisA						
Candle	62.50	90.11	+27.61	62.38	91.70	+29.32
Capsules	53.25	83.53	+30.28	64.94	89.24	+24.30
Cashew	54.12	74.22	+20.10	69.42	87.49	+18.08
Chewing gum	76.00	96.86	+20.86	84.00	98.51	+14.51
Fryum	65.50	91.02	+25.52	77.00	95.37	+18.37
Macaroni 1	55.75	82.48	+26.73	55.64	84.36	+28.73
Macaroni 2	55.50	67.08	+11.58	53.80	64.83	+11.02
PCB 1	50.75	58.98	+8.23	50.75	62.72	+11.97
PCB 2	53.50	71.90	+18.40	53.50	75.18	+21.68
PCB 3	50.88	69.86	+18.99	50.63	71.46	+20.83
PCB 4	65.38	96.21	+30.84	65.08	96.23	+31.15
Pipe fryum	58.13	90.46	+32.34	72.08	95.16	+23.08

All scale-averaged intervals exclude zero. At the individual-configuration level, 28 of 28 XD-Violence AP intervals and 23 of 28 UCF-Crime AUROC intervals exclude zero. The five UCF-Crime intervals that include zero are MiniCPM-V-4.5-8B at 0–5 integer ($+2.15$ $[-0.28, +4.48]$) and 0–9 integer ($+1.83$ $[-0.00, +3.19]$), and Qwen3-VL-8B at 0–1 decimal ($+1.55$ $[-0.29, +2.98]$), 0–5 decimal ($+1.34$ $[-0.56, +2.75]$), and 0–9 decimal ($+0.99$ $[-0.92, +2.35]$). All image per-model Yes/No intervals exclude zero.

B.2 IMAGE CATEGORIES AND VIDEO EVENT CLASSES

Under Yes/No, PROB improves both metrics in all 27 image categories and all 19 video event classes, averaged over four VLMs. Each video class is evaluated with all normal test videos. XD-Violence is multi-label, so a video contributes to each annotated event type and is compared with all 300 normal videos. The per-class AP values reflect different anomaly prevalences.

B.3 MODEL SIZES

Table 10 extends the Yes/No comparison from 2B to 38B parameters while retaining the domain-specific prompts and visual protocol. Every paired endpoint favors PROB on both metrics. The 2B video runs and the 7–8B and larger image controls use RTX PRO 6000 Blackwell hardware. The 2B/3B image endpoints use RTX 6000 Ada hardware.

Table 9: Yes/No video results (%), averaged over four models. Each event class includes all normal test videos. XD-Violence videos can contribute to multiple classes.

	AUROC			AP		
Class/category	DEC	PROB	Δ	DEC	PROB	Δ
UCF-Crime						
Abuse	65.13	94.73	+29.60	0.14	0.33	+0.18
Arrest	85.02	97.48	+12.46	17.45	47.33	+29.89
Arson	67.87	72.92	+5.05	9.27	26.83	+17.56
Assault	89.58	99.33	+9.75	34.15	82.19	+48.04
Burglary	67.08	89.16	+22.08	10.32	32.46	+22.14
Explosion	85.14	95.84	+10.70	15.27	48.30	+33.04
Fighting	87.95	98.52	+10.57	17.81	57.23	+39.42
RoadAccidents	86.87	96.47	+9.60	9.55	29.06	+19.51
Robbery	78.37	96.32	+17.94	10.89	57.42	+46.53
Shooting	77.93	94.50	+16.57	10.40	23.20	+12.79
Shoplifting	56.15	83.29	+27.14	2.25	4.81	+2.56
Stealing	76.16	96.88	+20.72	11.05	45.03	+33.98
Vandalism	78.68	98.16	+19.48	5.58	21.03	+15.45
XD-Violence						
Abuse	89.90	99.02	+9.12	10.67	49.28	+38.61
Car accident	76.65	91.68	+15.04	19.37	38.01	+18.64
Explosion	89.81	97.51	+7.70	31.10	54.84	+23.74
Fighting	89.25	96.97	+7.73	45.62	66.35	+20.74
Riot	83.91	96.99	+13.08	60.50	87.44	+26.94
Shooting	89.99	97.11	+7.12	27.30	44.35	+17.05

Table 10: Yes/No performance across model sizes (%). MVTec AD uses category-macro image metrics. Video metrics use frame-level scores.

Benchmark	Model	AUROC		AP	
		DEC	PROB	DEC	PROB
UCF-Crime	Qwen3-VL-2B	59.01	87.54	18.40	41.38
UCF-Crime	Qwen3-VL-8B	76.28	85.26	19.91	40.87
UCF-Crime	Qwen3-VL-32B	76.22	86.84	23.48	43.63
UCF-Crime	InternVL3.5-2B	61.23	86.87	19.63	41.91
UCF-Crime	InternVL3.5-8B	75.97	85.93	22.70	39.39
UCF-Crime	InternVL3.5-38B	76.60	87.47	22.43	43.87
XD-Violence	Qwen3-VL-2B	70.09	88.92	43.56	62.60
XD-Violence	Qwen3-VL-8B	84.64	91.40	50.66	70.25
XD-Violence	InternVL3.5-2B	77.94	91.85	53.83	73.22
XD-Violence	InternVL3.5-8B	85.31	91.67	55.32	73.31
MVTec AD	Qwen3-VL-2B	53.92	87.93	74.43	92.91
MVTec AD	Qwen3-VL-8B	75.05	91.51	85.56	95.67
MVTec AD	Qwen3-VL-32B	81.68	92.98	88.77	96.92
MVTec AD	Qwen2.5-VL-3B	68.09	83.81	81.45	91.26
MVTec AD	Qwen2.5-VL-7B	59.26	90.82	77.44	95.56
MVTec AD	InternVL3.5-8B	69.47	84.18	82.16	92.11
MVTec AD	InternVL3.5-38B	73.15	88.65	84.62	94.51

Table 11: Attainable TPR (%) at fixed FPR budgets, averaged over 28 model–scale configurations. Image results are category-macro and video results are frame-level.

Benchmark	Rule	FPR $\leq 1\%$	FPR $\leq 5\%$	FPR $\leq 10\%$
MVTec AD	DEC	34.91	40.55	42.84
	PROB	61.36	70.72	76.10
VisA	DEC	19.68	23.59	24.85
	PROB	42.04	56.63	63.36
UCF-Crime	DEC	1.00	17.37	38.94
	PROB	16.46	43.59	57.86
XD-Violence	DEC	0.33	7.59	24.46
	PROB	10.56	39.61	66.54

Table 12: Primary-metric recovery (%). Rows average four models before computing the ratio of gains. All 7 and Nonbinary 5 average 28 and 20 configurations. Columns use I-AUROC, I-AUROC, AUROC, and AP, respectively.

Scale	MVTec AD	VisA	UCF-Crime	XD-Violence
Yes/No	100.00	100.00	100.00	100.00
0–1 integer	100.00	100.00	100.00	100.00
0–5 integer	98.91	99.95	96.83	97.78
0–9 integer	98.69	100.67	95.35	90.26
0–1 decimal	99.13	100.13	97.84	93.05
0–5 decimal	100.19	100.93	85.16	93.30
0–9 decimal	99.58	102.43	73.26	75.46
All 7 scales	99.57	100.48	95.44	95.19
Nonbinary 5	99.27	100.77	91.50	90.70

B.4 LOW-FALSE-POSITIVE-RATE PERFORMANCE

We report the highest TPR attainable subject to an FPR budget:

$$\mathrm{TPR}^*(\alpha) = \max_{t:\mathrm{FPR}(t)\leq\alpha} \mathrm{TPR}(t). \tag{10}$$

We include the reject-all prediction, process equal scores together, and do not interpolate. Image thresholds are selected within category before macro-averaging. Video thresholds use concatenated frame-level scores. These thresholds are selected using test labels to characterize the attainable operating points. Mean TPR improves on every benchmark at every budget. At the individual-configuration level, PROB is higher in all 28 UCF-Crime comparisons at each budget and in 27, 28, and 28 XD-Violence comparisons at 1%, 5%, and 10%, respectively.

C RANK COMPRESSION

C.1 TIE-BREAKING RECOVERY

We rank by $H = (s_{\mathrm{DEC}}, s_{\mathrm{PROB}})$ lexicographically in descending order. Thus PROB orders inputs only within a DEC tie. For metric M, define

$$\Delta = M(\textsc{Prob}) - M(\textsc{Dec}), \qquad T = M(H) - M(\textsc{Dec}), \qquad R = M(\textsc{Prob}) - M(H).$$

Then $\Delta = T + R$. For a set of configurations S, recovery is $100\,\overline{T}_S/\overline{\Delta}_S$. A value above 100% means that preserving the order between decoded levels outperforms the full PROB ranking. For binary scales, tie-breaking reconstructs the PROB ranking by construction. The nonbinary rows measure the empirical contribution of within-answer ordering.

Recovery varies across configurations. At the finest video scale, the four-model values are 73.26% on UCF-Crime and 75.46% on XD-Violence. The lowest model-level value in Table 13 is 37.90% for Qwen2.5-VL-7B on XD-Violence at this scale.

Table 13: Primary-metric recovery (%) by model. All 7 and Nonbinary 5 are ratios of mean gains across their answer scales. The last column reports the finest scale.

Benchmark	Model	All 7	Nonbinary 5	0–9 decimal
MVTec AD	Qwen3-VL-8B	100.09	100.15	100.06
MVTec AD	Qwen2.5-VL-7B	99.72	99.41	99.62
MVTec AD	InternVL3.5-8B	98.67	97.73	98.95
MVTec AD	MiniCPM-V-4.5-8B	99.84	99.77	99.68
VisA	Qwen3-VL-8B	100.04	100.07	100.12
VisA	Qwen2.5-VL-7B	99.94	99.89	100.22
VisA	InternVL3.5-8B	101.82	103.06	112.76
VisA	MiniCPM-V-4.5-8B	100.31	100.43	100.57
UCF-Crime	Qwen3-VL-8B	99.51	98.85	96.85
UCF-Crime	Qwen2.5-VL-7B	93.56	89.18	68.27
UCF-Crime	InternVL3.5-8B	92.31	79.37	67.92
UCF-Crime	MiniCPM-V-4.5-8B	99.30	98.90	93.57
XD-Violence	Qwen3-VL-8B	99.28	98.36	93.99
XD-Violence	Qwen2.5-VL-7B	85.17	73.01	37.90
XD-Violence	InternVL3.5-8B	99.47	99.00	103.25
XD-Violence	MiniCPM-V-4.5-8B	96.57	93.74	81.32

Table 14: Within-answer AUROC under Yes/No (%). Image results are category-macro, with the number of eligible categories in parentheses. Video results are frame-level.

Model	MVTec AD		VisA	
	No	Yes	No	Yes
Qwen3-VL-8B	83.99 (15)	85.24 (3)	85.47 (12)	62.49 (3)
Qwen2.5-VL-7B	88.59 (15)	90.00 (1)	80.61 (12)	50.00 (1)
InternVL3.5-8B	76.80 (15)	72.26 (4)	77.29 (12)	83.42 (2)
MiniCPM-V-4.5-8B	82.81 (15)	98.91 (1)	73.89 (12)	53.87 (1)
Model	UCF-Crime		XD-Violence	
	No	Yes	No	Yes
Qwen3-VL-8B	76.23	71.36	82.43	72.77
Qwen2.5-VL-7B	83.13	60.84	89.38	46.90
InternVL3.5-8B	76.88	69.51	84.22	69.38
MiniCPM-V-4.5-8B	71.33	70.84	83.49	70.00

C.2 Ordering within each decoded answer

Table 14 evaluates PROB among inputs assigned the same Yes/No answer. All 16 model–benchmark "No" results are above chance. Image averages include categories containing both labels in the corresponding answer group, with their counts in parentheses. The "Yes" results include a below-chance value of 46.90 on XD-Violence for Qwen2.5-VL-7B.

C.3 Oracle binary thresholds

We evaluate every binary partition induced by the Yes/No PROB scores, including accept-all and reject-all, and choose the test-set threshold maximizing the primary metric. Equal scores remain grouped. Images use one threshold per category to maximize I-AUROC. Videos use one threshold per model to maximize UCF-Crime AUROC or XD-Violence AP. Table 15 shows that the oracle improves on DEC but remains below PROB in all 16 settings.

C.4 Distinct scores and numerical precision

Under the 0–9 decimal scale, DEC produces only 3.2–5.8 distinct scores per image category and 9–18 over a video benchmark. Table 16 reports the corresponding PROB coverage across the decimal scales. The remaining video duplicates arise from repeated inputs near video boundaries and numerically identical probabilities.

Table 15: Oracle binary-threshold diagnostic under Yes/No (%). Test labels select the threshold maximizing the primary metric. Threshold gain is Oracle binary minus DEC, and residual gap is PROB minus Oracle binary, in percentage points. Differences use unrounded values.

Benchmark	Model	DEC	Oracle binary	PROB	Threshold gain	Residual gap
MVTec AD	Qwen3-VL-8B	75.05	88.07	91.51	13.02	3.44
	Qwen2.5-VL-7B	59.26	86.85	90.82	27.58	3.97
	InternVL3.5-8B	69.47	82.39	84.18	12.91	1.79
	MiniCPM-V-4.5-8B	74.73	87.89	89.25	13.17	1.36
VisA	Qwen3-VL-8B	62.00	82.96	87.23	20.96	4.27
	Qwen2.5-VL-7B	53.96	76.41	81.27	22.46	4.85
	InternVL3.5-8B	58.42	76.57	79.47	18.15	2.91
	MiniCPM-V-4.5-8B	59.38	75.86	76.27	16.48	0.41
UCF-Crime	Qwen3-VL-8B	76.28	77.66	85.26	1.38	7.60
	Qwen2.5-VL-7B	60.47	78.55	85.64	18.08	7.09
	InternVL3.5-8B	75.97	78.89	85.93	2.92	7.03
	MiniCPM-V-4.5-8B	76.07	77.15	84.45	1.08	7.30
XD-Violence	Qwen3-VL-8B	50.66	55.65	70.25	4.99	14.60
	Qwen2.5-VL-7B	36.81	51.20	59.19	14.39	7.99
	InternVL3.5-8B	55.32	56.61	73.31	1.29	16.70
	MiniCPM-V-4.5-8B	54.97	55.86	73.17	0.89	17.31

Table 16: Distinct PROB scores across three decimal scales and four models. Image counts are means per category. Video counts are prediction segments. Ranges span the 12 configurations.

Benchmark	Units	Unique PROB	Unique (%)
MVTec AD	115.0	115.0	100.000
VisA	180.2	180.2	100.000
UCF-Crime	69,634	69,538–69,542	99.862–99.868
XD-Violence	146,449	146,334–146,421	99.921–99.981

Table 17: Sample-mean image performance under Yes/No (%). Metrics are evaluated per model and seed before averaging. DEC and PROB are the four-model means without sampling.

	MVTec AD		VisA	
Scoring rule / samples	I-AUROC	I-AP	I-AUROC	I-AP
Greedy DEC	69.63	82.67	58.44	63.27
$k = 1$	68.47	81.84	59.28	63.61
$k = 2$	72.64	84.25	62.35	66.35
$k = 4$	76.12	86.50	65.09	69.15
$k = 8$	78.98	88.40	67.95	71.87
$k = 16$	81.18	89.90	69.95	74.02
$k = 32$	83.17	91.39	71.79	75.99
$k = 64$	84.64	92.42	73.91	78.10
$k = 128$	85.72	93.12	75.53	79.76
$k = 256$	86.68	93.72	76.69	80.83
Direct PROB	88.94	94.71	81.06	84.35

Precision comparison. For Qwen3-VL-8B with Yes/No on UCF-Crime, switching from bfloat16 to float32 increases distinct PROB scores from 137 to 65,885, while AUROC changes from 85.26 to 85.27. The segment-score Spearman correlation is 0.9985. On MVTec AD, the mean count per category rises from 53.1 to 115.0, I-AUROC changes from 91.51 to 91.78, and I-AP from 95.67 to 95.98. The runs also differ in batch size. The comparison demonstrates that a large increase in distinct-score count need not substantially improve ranking.

D ADDITIONAL SCORING EXPERIMENTS

D.1 SAMPLED ANSWERS

Single-answer video simulations use temperatures 0.6 and 1.0, nucleus top-p values 0.8 and 0.95 at $T = 1$, and majority votes over 5 and 16 samples at $T = 1$. A tied 16-sample vote returns No. All draw from saved answer-set-normalized Yes/No probabilities. Across these strategies, four-model mean performance ranges from 70.85 to 72.13 UCF-Crime AUROC and from 46.03 to 49.09 XD-Violence AP, below the corresponding PROB values of 85.32 and 68.98.

For sample-mean scoring, we use $T = 1$ and $k \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$ and score each input by its fraction of Yes answers. Metrics are computed for each model–seed run before averaging over three seeds and four VLMs. Image seeds are 20260903, 20260904, and 20260905. Models share uniform random streams within each dataset and seed, and smaller budgets reuse prefixes of 256 draws. Video seeds are 0, 1, and 2, with binomial draws for each segment and sample count before frame expansion. Table 17 reports the image results, and Figure 4 shows all four benchmarks.

D.2 TEMPORAL SMOOTHING

We apply matched Gaussian smoothing at $\sigma \in \{1, \ldots, 9\}$ to each video's segment scores before frame expansion. The kernel is truncated at 4σ with reflected boundaries. Mean primary-metric gaps remain positive at every width (Table 18). Across individual model–scale–width comparisons, PROB is higher in 207 of 252 UCF-Crime AUROC comparisons and 230 of 252 XD-Violence AP comparisons. Distinct scores are counted by exact float64 equality without rounding.

D.3 PROMPT WORDING AND ANSWER POLARITY

We evaluate three paraphrases and a reversed-normality question using Qwen3-VL-8B and InternVL3.5-8B on MVTec AD and UCF-Crime. Paraphrases preserve the Yes/No answer set, value map, visual protocol, and batch size. The reversed question assigns $v(\text{No}) = 1$ and $v(\text{Yes}) = 0$. PROB exceeds DEC on both metrics in every row of Table 19.

Video pv1.

Table 18: Matched smoothing: primary-metric gap in percentage points and distinct segment-score counts, averaged over 28 configurations. Rows show representative widths and the unsmoothed baseline.

Benchmark	σ	PROB–DEC gap	Unique DEC	Unique PROB
UCF-Crime	0	7.664	7.0	46,424.9
	1	4.613	4,989.0	69,479.1
	4	2.947	22,979.5	69,555.4
	8	2.350	30,549.3	69,561.2
	9	2.284	31,744.9	69,561.9
XD-Violence	0	10.972	7.7	99,740.4
	1	5.742	10,745.4	145,665.5
	4	3.862	57,905.2	146,256.4
	8	3.131	75,080.5	146,343.2
	9	3.008	77,582.8	146,356.2

```
You are reviewing ordered frames from a surveillance video
clip.
Is anything unusual or dangerous happening in the clip?
Answer with exactly one word:  Yes or No.
```

Video pv2.

```
You are a security analyst monitoring surveillance footage.
Do these ordered frames show an anomalous, suspicious,
violent, criminal, or dangerous event?
Answer with exactly one word:  Yes or No.
```

Video pv3.

```
These are ordered frames from a surveillance video clip.
Does the clip contain an anomaly, such as violence, crime,
or a dangerous situation?
Answer with exactly one word:  Yes or No.
```

Image pv1.

```
You are inspecting an image of a <category>.
Is the <category> unusual or defective in any way?
Answer with exactly one word:  Yes or No.
```

Image pv2.

```
You are a quality-control inspector examining an image of a
<category>.
Does the object or material show an anomaly, damage,
contamination, malformation, misplacement, or a missing
part?
Answer with exactly one word:  Yes or No.
```

Image pv3.

```
This image shows a <category>.
Does it contain a defect, such as damage, contamination, an
incorrect shape, a misplaced component, or a missing part?
Answer with exactly one word:  Yes or No.
```

Image reversed-polarity prompt.

```
You are inspecting an image of a <category>.
```

Table 19: Prompt and polarity controls (%). Base and pv1–pv3 assign anomaly to Yes. Reversed assigns anomaly to No. Image metrics are category-macro and video metrics are unsmoothed frame-level.

Benchmark	Model	Control	AUROC		AP	
			DEC	PROB	DEC	PROB
MVTec AD	Qwen3-VL-8B	base	75.05	91.51	85.56	95.67
MVTec AD	Qwen3-VL-8B	pv1	77.13	91.34	86.70	95.57
MVTec AD	Qwen3-VL-8B	pv2	79.63	92.17	87.61	95.92
MVTec AD	Qwen3-VL-8B	pv3	79.78	92.31	87.67	95.84
MVTec AD	InternVL3.5-8B	base	69.47	84.18	82.16	92.11
MVTec AD	InternVL3.5-8B	pv1	68.48	85.30	81.68	92.87
MVTec AD	InternVL3.5-8B	pv2	71.89	83.92	83.14	91.81
MVTec AD	InternVL3.5-8B	pv3	72.44	83.80	83.45	92.07
UCF-Crime	Qwen3-VL-8B	base	76.28	85.26	19.91	40.87
UCF-Crime	Qwen3-VL-8B	pv1	77.40	85.95	19.82	40.83
UCF-Crime	Qwen3-VL-8B	pv2	74.94	86.03	22.72	43.43
UCF-Crime	Qwen3-VL-8B	pv3	75.96	85.30	20.85	41.40
UCF-Crime	InternVL3.5-8B	base	75.97	85.93	22.70	39.39
UCF-Crime	InternVL3.5-8B	pv1	75.91	84.71	22.12	39.60
UCF-Crime	InternVL3.5-8B	pv2	73.98	84.60	22.41	38.51
UCF-Crime	InternVL3.5-8B	pv3	74.55	85.12	21.89	38.80
MVTec AD	Qwen3-VL-8B	Reversed	72.08	90.63	82.90	95.28
MVTec AD	InternVL3.5-8B	Reversed	71.98	82.15	83.36	90.39
UCF-Crime	Qwen3-VL-8B	Reversed	77.77	85.06	18.31	39.02
UCF-Crime	InternVL3.5-8B	Reversed	76.66	85.84	21.74	35.63

```
Is the object or material normal and defect-free, with no
damage, contamination, malformation, misplacement, or
missing part?
Answer with exactly one word:  Yes or No.
```

For this control, No maps to 1 and Yes maps to 0.

Video reversed-polarity prompt.

```
You are reviewing ordered frames from a surveillance video
clip.
Is the clip normal, ordinary, and free of anything
suspicious, violent, criminal, or dangerous?
Answer with exactly one word:  Yes or No.
```

The same reversed value map applies.

D.4 ANSWERS ACCOMPANIED BY DESCRIPTIONS

We evaluate Qwen3-VL-8B and MiniCPM-V-4.5-8B on both binary scales with a factual description before or after the answer. Descriptions use greedy decoding (`do_sample=False`, `num_beams=1`) with at most 64 tokens. Score-first conditions the description on the argmax answer. Explanation-first conditions both scoring rules on the same generated description. Table 20 reports the primary-metric comparisons.

Image joint-output prompt. The following common stem is followed by one score instruction and one output-order instruction, separated by newlines.

```
You are inspecting an image of a <category>.
Judge whether the object or material is anomalous or
defective, for example damaged, contaminated, malformed,
misplaced, or missing a part.
Also describe only what is visibly present in one concise
factual sentence.
```

```
Do not speculate about causes or anything that is not
visible.
```

For Yes/No, the score instruction is:

```
The score must be exactly Yes or No:  No means normal and
Yes means anomalous or defective.
```

For 0–1 integer, it is:

```
The score must be exactly 0 or 1:  0 means normal and 1
means anomalous or defective.
```

Score-first uses:

```
Return exactly two lines in this order:
Score:  <Yes or No>
Description:  <one concise factual sentence>
```

Explanation-first uses:

```
Return exactly two lines in this order:
Description:  <one concise factual sentence>
Score:  <Yes or No>
```

For the integer scale, the literal placeholder `<Yes or No>` is replaced by `<0 or 1>`. These placeholders appear literally in the prompt. Only `<category>` is substituted with a category name. The prompt version is `industrial-image-joint-v1` with the output order recorded in each run.

Video joint-output prompt. For video, replace the image common stem above with:

```
You are reviewing ordered frames from a surveillance video
clip.
Judge whether the clip is anomalous, suspicious, violent,
criminal, or dangerous.
Also describe only what is visibly happening in one concise
factual sentence.
Do not speculate about intent or events that are not
visible.
```

The video Yes/No instruction is `The score must be exactly Yes or No.` The video integer instruction is `The score must be exactly 0 or 1:  0 means normal and 1 means anomalous or dangerous.` Append the same output-order instructions shown above, with newline separators.

PROB remains higher in every paired table comparison. Output order changes absolute performance. Across the eight unsmoothed video configurations, explanation-first reduces PROB by 3.39 primary-metric points on average, including a 10.57-point AP decrease. The paired scoring comparison therefore remains distinct from the effect of generating a description.

E EVALUATION WITH TIES

E.1 BENCHMARK EVALUATION CODE

Table 21 identifies the exact releases and metric entry points inspected for the main-text comparison.

Checks of the MVTec AD and VisA functions and XD-Violence metric statements matched the grouped calculations in 384 tests per release, using 16 score assignments and 24 input permutations. The UCF-Crime check used a literal NumPy translation of the active MATLAB ROC block, not native MATLAB execution. The audit environment was Python 3.11.15, NumPy 2.4.4, scikit-learn 1.9.0, and SciPy 1.17.1.

Table 20: Primary-metric performance with descriptions (%). S-first emits the score first, and E-first emits the description first. MiniCPM-V denotes MiniCPM-V-4.5-8B.

Benchmark	Model	Scale	Base		S-first		E-first	
			DEC	PROB	DEC	PROB	DEC	PROB
UCF-Crime	Qwen3-VL-8B	Yes/No	76.28	85.26	76.08	85.07	75.12	84.46
UCF-Crime	Qwen3-VL-8B	0–1 int	75.23	85.49	76.56	84.96	75.18	84.37
UCF-Crime	MiniCPM-V-4.5-8B	Yes/No	76.07	84.45	74.64	82.34	75.72	83.38
UCF-Crime	MiniCPM-V-4.5-8B	0–1 int	76.60	84.10	76.07	83.39	74.20	81.94
XD-Violence	Qwen3-VL-8B	Yes/No	50.66	70.25	51.11	71.50	48.33	65.03
XD-Violence	Qwen3-VL-8B	0–1 int	43.35	69.42	47.67	64.44	47.56	58.86
XD-Violence	MiniCPM-V-4.5-8B	Yes/No	54.97	73.17	53.96	73.18	53.26	70.92
XD-Violence	MiniCPM-V-4.5-8B	0–1 int	51.01	71.84	48.81	71.90	51.30	67.91
MVTec AD	Qwen3-VL-8B	Yes/No	75.05	91.51	65.62	92.57	67.60	88.49
MVTec AD	Qwen3-VL-8B	0–1 int	67.96	91.46	75.52	92.07	73.49	87.85
MVTec AD	MiniCPM-V	Yes/No	74.73	89.25	75.48	93.12	72.21	91.89
MVTec AD	MiniCPM-V	0–1 int	72.20	93.18	79.90	93.39	79.69	90.28
VisA	Qwen3-VL-8B	Yes/No	62.00	87.23	55.96	87.27	56.92	80.19
VisA	Qwen3-VL-8B	0–1 int	58.04	87.79	61.92	86.75	62.42	81.64
VisA	MiniCPM-V	Yes/No	59.38	76.27	61.29	84.04	58.92	83.79
VisA	MiniCPM-V	0–1 int	58.08	85.50	72.81	86.27	67.59	83.45

Table 21: Metric conventions in the inspected releases. Absent means that the metric is not reported in the specified path.

Benchmark	ROC	PR	Release / metric entry point
MVTec AD	Grouped, half-credit	Absent at image level	Evaluation archive v1.0. `roc_curve_util.py`, `compute_classification_roc`.
VisA	Grouped, half-credit	Grouped, trapezoidal	spot-diff, `e1429125dbaa`. `utils/metrics.py`, `compute_classification_metrics`.
UCF-Crime	Item-wise, order-dependent	Absent	author release, `5e3ee7a991ce`. `Evaluate_Anomaly_Detector.m`, active ROC block.
XD-Violence	Absent	Grouped, trapezoidal	XDVioDet, `f846d8cfc454`. `test.py`, offline/online statements.

Table 22: Mean fraction of tied anomalous–normal pairs (%) across four models and seven scales. Numerically, each value is also the maximum item-wise ROC-area span in percentage points.

Benchmark	DEC	PROB
MVTec AD	49.97	0.47
VisA	71.23	1.10
UCF-Crime	31.32	0.23
XD-Violence	19.29	0.20

E.2 EXACT ROC TIE-ORDER SENSITIVITY

Let p_j and n_j count anomalous and normal examples in tied-score group j, and let $P = \sum_j p_j > 0$ and $N = \sum_j n_j > 0$. The fraction of tied anomalous–normal pairs is

$$\tau = \frac{\sum_j p_j n_j}{PN}. \tag{11}$$

Grouping ties gives the half-credit AUROC in Equation (5). An item-wise ROC traversal instead assigns an order to each tied pair. Putting anomalous examples before all normal examples within each group gives each such pair full credit, while the reverse order gives zero. All untied pairs retain their ordering, so the exact extrema are

$$\mathrm{AUROC}^- = \mathrm{AUROC} - \frac{\tau}{2}, \qquad \mathrm{AUROC}^+ = \mathrm{AUROC} + \frac{\tau}{2}. \tag{12}$$

The maximum item-wise ROC-area span is therefore τ. These label-aware extrema measure worst-case sensitivity to an unspecified tie order, not uncertainty in the grouped AUROC. Table 22 reports the tied-pair fractions on the saved scores. Image fractions are computed within category before macro-averaging. Video fractions use frame-level scores, so segment-to-frame repetition can create mixed-label ties at anomaly boundaries.

E.3 PR-AREA COMPARISON ON FIXED SCORES

Using the grouped precision–recall points and endpoint in Equation (9),

$$\text{PR-AUC}_{\text{trapezoids}} - \mathrm{AP} = \frac{1}{2}\sum_{j=1}^{J}(r_j - r_{j-1})(q_{j-1} - q_j).$$

The difference can have either sign and need not vanish in the absence of ties. We retain all threshold points in score order, including equal-recall points, without applying a precision envelope. Table 23 reports both integration offsets and the resulting PROB–DEC gaps for all benchmarks and scales.

Non-interpolated AP is a weighted mean of threshold precisions and therefore cannot exceed their maximum over nonempty predictions. For image macro-averages, the corresponding bound is the mean of the category-specific maximum precisions. Trapezoidal PR-AUC exceeds this bound in 1 of 28 MVTec AD, 3 of 28 VisA, 11 of 28 UCF-Crime, and 19 of 28 XD-Violence DEC configurations. No PROB configuration exceeds it. An exceedance requires a difference greater than 10^{-12} on the unit-interval scale. These measurements support the fixed-score comparison in Section 6.

Table 23: PR-area calculations on identical scores. Each offset is trapezoidal PR-AUC minus AP for the indicated rule. Each gap is PROB minus DEC. Entries are percentage-point differences averaged over four models before rounding. Images use category-macro metrics, and videos use unsmoothed frame-level metrics.

Benchmark	Scale	Offset DEC	Offset PROB	Grouped AP gap	Trapezoidal PR-AUC gap
MVTec AD	Yes/No	+8.36	+0.25	+12.05	+3.93
	0–1 integer	+10.71	+0.17	+16.68	+6.14
	0–5 integer	+5.84	−0.05	+8.69	+2.80
	0–9 integer	+5.92	−0.06	+8.54	+2.56
	0–1 decimal	+7.56	−0.06	+12.05	+4.43
	0–5 decimal	+5.60	−0.06	+7.85	+2.20
	0–9 decimal	+5.54	−0.06	+7.83	+2.23
VisA	Yes/No	+17.20	+0.79	+21.09	+4.67
	0–1 integer	+18.79	+0.46	+25.55	+7.22
	0–5 integer	+13.24	−0.10	+17.11	+3.78
	0–9 integer	+13.11	−0.11	+16.54	+3.33
	0–1 decimal	+14.50	−0.11	+20.13	+5.52
	0–5 decimal	+14.33	−0.11	+16.71	+2.26
	0–9 decimal	+13.37	−0.11	+16.00	+2.53
UCF-Crime	Yes/No	+25.99	+0.45	+18.91	−6.63
	0–1 integer	+26.08	+0.95	+18.11	−7.02
	0–5 integer	+15.40	−0.01	+11.49	−3.92
	0–9 integer	+11.20	−0.01	+11.48	+0.27
	0–1 decimal	+11.78	−0.02	+11.00	−0.79
	0–5 decimal	+9.46	−0.02	+9.67	+0.20
	0–9 decimal	+4.40	−0.01	+8.58	+4.16
XD-Violence	Yes/No	+20.74	+0.30	+19.54	−0.89
	0–1 integer	+22.27	+1.25	+17.55	−3.47
	0–5 integer	+15.09	−0.00	+9.38	−5.71
	0–9 integer	+9.25	−0.00	+8.23	−1.03
	0–1 decimal	+8.98	−0.00	+8.34	−0.63
	0–5 decimal	+8.84	−0.00	+7.12	−1.72
	0–9 decimal	+5.54	−0.00	+6.63	+1.09