TAG: TASK-CONDITIONED ADAPTIVE DIAGNOSIS FOR MULTIMODAL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Diagnosing incorrect answer annotations is important for reliable training and evaluation of multimodal models. Existing approaches use direct judgments from vision–language models (VLMs) or representation-based detectors, but differences in datasets, representations, and noise conditions complicate understanding when each strategy is effective. We compare these approaches under a common protocol across three multimodal benchmarks with controlled answer errors. The comparison reveals potential gains from selecting strategies for individual tasks over the best shared strategy. On AVA-Bench, encoder preferences vary across abilities but remain relatively stable across the tested noise levels, whereas detector preferences vary with both ability and noise. To demonstrate the value of these findings, we introduce Task-conditioned Adaptive Diagnosis (TAG), which identifies tasks and estimates noise conditions to select and combine diagnostic strategies, with confidence-based fallback to a shared strategy. Cross-validation shows that TAG improves AUROC over the best shared strategy by 1.24 percentage points on AVA-Bench, with smaller gains on ScienceQA and MMMU, demonstrating that task-conditioned selection can translate diagnostic preference differences into improved diagnosis.

1 INTRODUCTION

Multimodal models support image–text understanding, visual question answering, and instruction following (Radford et al., 2021; Bai et al., 2025). Their training data and evaluation benchmarks can contain degraded images (Chiu et al., 2020), missing image–text associations (Chun et al., 2022), and incorrect labels (Northcutt et al., 2021b; Wei et al., 2022). Label errors corrupt training supervision and can distort benchmark comparisons (Patrini et al., 2017; Northcutt et al., 2021a). We focus on identifying questions whose annotated answers are inconsistent with the image and question, using multiple-choice benchmarks to construct controlled answer errors.

Two complementary routes support the diagnosis of incorrect answer annotations. The first uses VLMs to judge question–answer examples directly (Chen et al., 2025), alongside broader curation methods based on image–text similarity (Gadre et al., 2023), instruction difficulty (Chen et al., 2024a), and batch learnability (Evans et al., 2024). The second applies data-centric detectors to learned representations, flagging inconsistencies through feature neighborhoods (Zhu et al., 2022), model predictions (Northcutt et al., 2021b), or training dynamics (Pleiss et al., 2020; Swayamdipta et al., 2020). Reported performance reflects the detector, representation, dataset, task composition, and noise conditions. Variation in these factors across evaluations limits our understanding of when particular components are effective and how to combine their strengths. A common evaluation protocol is therefore needed to compare diagnostic components across tasks and noise levels.

We conduct this comparison on three multimodal benchmarks under a common data construction and evaluation protocol. AVA-Bench (Mai et al., 2025) decomposes visual understanding into atomic abilities, providing the main setting for examining task-related diagnostic preferences; ScienceQA (Lu et al., 2022) and MMMU (Yue et al., 2024) provide contrasts through broader task groupings and discipline categories, respectively. We compare eight encoders combined with eleven detectors, alongside direct VLM judgment. Reversing the correctness labels of selected answer pairs marks an incorrect option as correct and the reference answer as incorrect, enabling comparison against the same known answer errors within each task and noise condition.

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.

The comparison yields three observations. First, selecting the best strategy post hoc for each task and noise state outperforms the best shared strategy on all three benchmarks, revealing an *oracle gap* that quantifies the potential benefit of task-specific selection within the candidate pool. The gaps are 1.25, 0.28, and 0.44 AUROC percentage points on AVA-Bench, ScienceQA, and MMMU, respectively (Table 1). Second, on AVA-Bench, encoder preferences vary across abilities but remain relatively stable across the tested noise levels. Third, detector preferences vary with both ability and noise: some tasks favor neighborhood-consistency signals at low noise and training-based signals at higher noise, while others consistently favor training-based signals (Figure 2). These observations motivate selecting representations by task and detectors by both task and noise condition.

Building on these observations, we propose **T**ask-conditioned **A**daptive dia**G**nosis (TAG; Figure 1), which identifies the task of each question, estimates the noise state of the batch, and selects the encoder and detectors accordingly, falling back to a shared strategy when the task assignment is uncertain. Cross-validated experiments show clear gains on AVA-Bench and small gains on ScienceQA and MMMU.

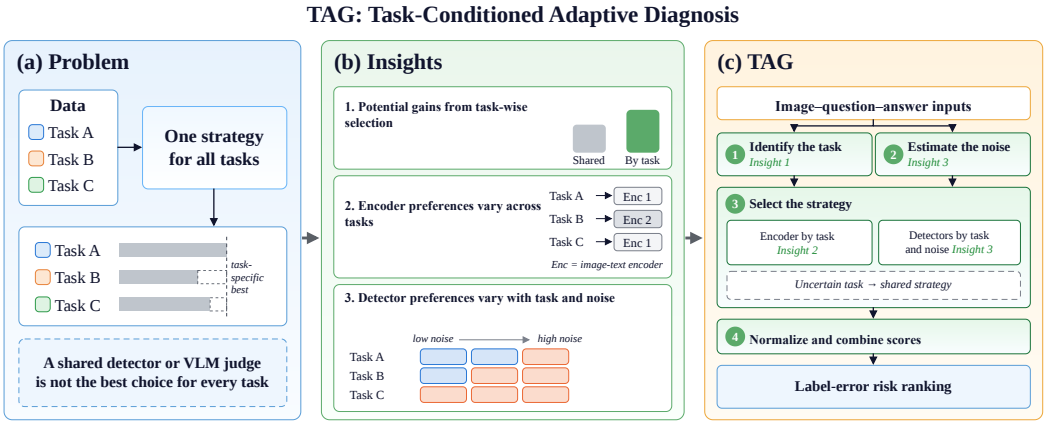

Figure 1: TAG overview for diagnosing incorrect answer annotations in multimodal data. (a) Diagnostic performance varies across tasks. (b) Task-level comparisons reveal an oracle gap between task-specific best strategies and the best shared strategy, along with distinct encoder and detector preferences. (c) TAG selects encoders by task and combines detectors by task and estimated noise state, with confidence-based fallback to Global. Bars and preference changes are schematic.

Our contributions are:

- We establish a common protocol for diagnosing incorrect answer annotations in multimodal data, comparing direct VLM judgment and representation-based detectors across tasks and noise conditions (Figure 1a).
- We identify an oracle gap over the best shared strategy on three benchmarks. On AVA-Bench, encoder preferences vary across abilities but remain relatively stable across noise levels, while detector preferences vary with both ability and noise (Figure 1b).
- We propose TAG to exploit these preferences through task-conditioned strategy selection and confidence-based fallback (Figure 1c). TAG improves AUROC over Global on AVA-Bench, with smaller gains on ScienceQA and MMMU, demonstrating the diagnostic value of task-conditioned selection.

2 Related Work

Label-error detection. Learning with noisy labels studies class-conditional and instance-dependent noise (Natarajan et al., 2013; Patrini et al., 2017; Wei et al., 2022). Diagnostic signals include forgetting events (Toneva et al., 2019), confidence trajectories in Dataset Cartography (Swayamdipta et al., 2020), margins in AUM (Pleiss et al., 2020), and early prediction errors in EL2N (Paul et al.,

2021). Confident Learning uses model predictions to estimate label errors (Northcutt et al., 2021b), while SimiFeat exploits agreement among neighboring representations (Zhu et al., 2022). HOC estimates noise transitions from higher-order label consensuses (Zhu et al., 2021), and Docta connects credibility estimation with label-error identification in language datasets (Zhu et al., 2024). These signals capture different aspects of model–label agreement. We compare their ability to identify incorrect answer annotations across tasks and controlled noise conditions.

VLM-based data curation and judgment. Multimodal models support data assessment through both learned representations and generated judgments. DataComp compares image–text filtering procedures under a standardized protocol (Gadre et al., 2023); Self-Filter learns instruction difficulty jointly with a VLM (Chen et al., 2024a); and JEST selects jointly learnable batches (Evans et al., 2024). PreSel allocates task-wise budgets and selects representative images before generating instructions (Safaei et al., 2025), while TaskGalaxy uses multimodal models to screen generated question–answer examples (Chen et al., 2025). These studies assess training-data utility or generated-example quality. We focus on detecting annotated answers inconsistent with their images and questions, and compare direct VLM judgment with representation-based detection, including comparisons using the same checkpoint.

Task-aware selection. Task information guides data assessment and selection: task-amenable quality assessment learns from downstream performance (Saeed et al., 2021; 2022), DataSifter optimizes task-driven quality objectives (Wang et al., 2021), LESS selects instruction data for target capabilities (Xia et al., 2024), and TADS combines quality, task relevance, and diversity (Cheng et al., 2026). Algorithm selection chooses a method from problem characteristics (Rice, 1976). MetaOD applies this idea to outlier detection using dataset meta-features (Zhao et al., 2021), while evaluations of label-free selection highlight the difficulty of estimating detection utility (Ma et al., 2021). TAG connects task-aware assessment with algorithm selection: it selects encoder–detector combinations by task and noise state using preferences learned from training-fold batches with controlled answer errors.

3 Problem Setup

Given an image, a question, and its annotated answer, diagnosis assesses whether that answer is consistent with the image and question. Multiple-choice benchmarks provide reference answers and incorrect options for constructing controlled answer errors. A *task* describes a question's ability or metadata-derived category; a *data condition* describes the answer-error rate in the batch being diagnosed.

3.1 Task Formulation

For an image–question input (I_i, Q_i), let a_{i0} be the reference answer and a_{i1} an incorrect original option. The answer key assigns this pair the correctness labels $(\ell_{i0}, \ell_{i1}) = (1, 0)$. We consider $\rho \in \{10\%, 20\%, 30\%, 40\%\}$; at each level, a fraction ρ of pairs receives reversed labels $(\widetilde{\ell}_{i0}, \widetilde{\ell}_{i1}) = (0, 1)$, while the remaining pairs retain $(1, 0)$. Each reversal marks the incorrect option as correct and the reference answer as incorrect. Images, questions, and candidate-answer contents stay fixed; detectors receive the resulting correctness labels. Flip records identify the injected answer errors and provide the evaluation ground truth, following synthetic-corruption protocols (Natarajan et al., 2013; Patrini et al., 2017).

Let $\mathcal{T}$ denote the set of task categories in a dataset, and let $t_i \in \mathcal{T}$ denote the task of question i. A basic strategy $s = (e, d)$ uses encoder e to obtain vector representations and detector d to score suspected errors in candidate-answer annotations. Averaging the two direction-aligned candidate scores gives question risk $r_s(i)$. Higher scores indicate greater suspected inconsistency between the answer marked as correct and the image–question input. *Shared-strategy diagnosis* uses $s_i = s_G(c)$ across tasks for condition c; *task-conditioned diagnosis* allows $s_i = \pi(t_i, c)$. TAG applies the learned rule using identified tasks and estimated noise states. We examine diagnostic preferences (Section 4) and whether they support gains on held-out questions (Section 5).

3.2 Datasets and Diagnostic Strategies

AVA-Bench organizes visual understanding by atomic ability (Mai et al., 2025), making it our main setting for examining task dependence. We retain ten finite-choice categories covering 116,453 questions, including depth, orientation, texture, and emotion recognition. ScienceQA and MMMU provide contrasting category structures. ScienceQA contributes 10,332 image-grounded questions in ten metadata-derived task patterns plus Others; MMMU contributes 9,864 questions in six disciplines (Lu et al., 2022; Yue et al., 2024). ScienceQA's groups combine skill, subject, topic, and category metadata (Appendix A.1).

We compare two diagnostic paradigms. **Representation-based detection** combines eight frozen encoders with eleven detectors, yielding 88 strategies. The encoders range from text-only MPNet to multimodal representations. Detectors use neighborhood consistency (KNN, SimiFeat), prediction–label agreement (Confident Learning, OOF Self-Confidence), or training dynamics and loss (AUM, DataMap, EL2N, Forgetting Events, Small-loss, Loss-GMM); Docta-HOC supplies a consensus-derived signal. **Direct VLM judgment** gives answer-correctness probability p_{ij}, candidate risk $|p_{ij} - \widetilde{\ell}_{ij}|$, and its pair mean. We abbreviate Qwen2.5-VL-7B-Instruct and Qwen3-VL-Embedding as Qwen2.5-VL and Qwen3-VL, respectively. Qwen2.5-VL supplies both representations and judgments, enabling matched-checkpoint comparisons. The adaptive candidate pool consists of the 88 encoder–detector combinations. Appendix A specifies their implementations.

3.3 Evaluation

We compute AUROC (Fawcett, 2006) and average precision (AP) (Davis & Goadrich, 2006) within each task using the recorded answer errors as ground truth. AUROC measures how often a question with an incorrect annotated answer receives a higher risk score than one with a correct annotated answer, assigning half credit to ties; AP summarizes precision over increases in recall. Both lie in $[0, 1]$, with higher values indicating better ranking; random ranking has expected AUROC of 0.5. Task scores are weighted by question counts and averaged over four noise rates and three corruption seeds (0, 1, and 2), with datasets reported separately. Section 4 describes full-data preferences; Section 5 evaluates training-derived selection through five-fold, association-group-isolated batch cross-validation. Base detectors use representations and observed labels within the batch being diagnosed. For TAG and the learned baselines, evaluation-batch task labels and flip records are withheld from fitting and routing. These annotations support task-wise evaluation and the construction of post-hoc references. Appendix A.3 details the splits and seeds, and Appendix A.5 specifies aggregation and bootstrap intervals.

4 Task and Noise Dependence in Diagnosis

Does a strategy selected across a dataset also suit its individual tasks? We examine the encoder–detector candidate pool at multiple noise levels, first comparing shared and task-conditioned choices, then tracing the preferences underlying their differences. These comparisons use all five batches to characterize diagnostic preferences. We focus on AVA-Bench's ability structure and use ScienceQA and MMMU to assess how the pattern extends to other category structures.

4.1 Potential Gains from Task Conditioning

Shared diagnosis applies one configuration across tasks within each dataset and noise state. Global is the configuration maximizing question-weighted within-task AUROC; the task–condition oracle selects the best configuration per task from the same pool, based on diagnostic performance evaluated against the injected-error ground truth. Their performance difference, the *oracle gap*, describes the benefit of task-conditioned selection when both strategies are selected and evaluated on the full dataset. The gap is measured in AUROC. Selection uses three states: Low (10/20%), Medium (30%), and High (40%); reported performance averages the four rates equally.

Oracle gaps are positive on all three datasets (Table 1). Measured in AUROC, the largest gap is 1.25 percentage points on AVA-Bench; ScienceQA and MMMU show smaller gaps of 0.28 and 0.44 points. This variation is visible in the preferred signals: at low noise, depth tasks favor neighborhood consistency, whereas fine-grained recognition favors training-based signals (Figure 2a).

Table 1: Post-hoc task-conditioned selection versus the best shared strategy. Oracle gap is the AUROC difference in percentage points (pp). Null mean/95th denote the mean/95th percentile of gaps from 500 random task groupings. This comparison uses the full data; Table 2 reports held-out performance.

Dataset	Global	Oracle	Oracle gap (pp)	Null mean	Null 95th	p
AVA-Bench	0.9338	0.9463	1.25	0.00	0.01	0.002
ScienceQA	0.9752	0.9780	0.28	0.01	0.01	0.002
MMMU	0.6775	0.6819	0.44	0.16	0.32	0.006

Selecting within more groups also creates more opportunities to fit observed differences. We assess this effect by repeating the comparison under 500 association-group task-label permutations. AVA-Bench's observed gap exceeds the null's 0.01-point 95th percentile; ScienceQA and MMMU also exceed their respective nulls.[1] MMMU's larger null mean indicates that random partitions reproduce more of its observed oracle gap. Across the three datasets, actual task categories capture diagnostic differences beyond those reproduced by the tested random groupings.

(a) Category-specific detector shortlists

Ability	10% noise			20% noise			30% noise			40% noise		
Absolute Depth	**Simi**	KNN	HOC	**Simi**	KNN	HOC	**Simi**	KNN	GMM	**GMM**	DM	AUM
Relative Depth	**Simi**	KNN	AUM	**GMM**	AUM	DM	**GMM**	DM	AUM	**DM**	AUM	EL2N
Orientation	**KNN**	Simi	HOC	**KNN**	Simi	GMM	**GMM**	DM	AUM	**DM**	AUM	EL2N
Spatial	**KNN**	Simi	AUM	**KNN**	AUM	DM	**AUM**	DM	EL2N	**DM**	AUM	EL2N
Texture	**KNN**	Simi	CL	**KNN**	CL	OOF	**DM**	AUM	GMM	**EL2N**	DM	AUM
Fine-grained	**DM**	AUM	EL2N	**DM**	AUM	EL2N	**DM**	EL2N	AUM	**SL**	EL2N	DM
Action	**AUM**	DM	EL2N	**AUM**	DM	EL2N	**EL2N**	AUM	DM	**EL2N**	SL	AUM
Emotion	**AUM**	DM	KNN	**AUM**	DM	EL2N	**AUM**	DM	EL2N	**DM**	AUM	EL2N
Recognition	**DM**	AUM	CL	**DM**	AUM	EL2N	**DM**	AUM	EL2N	**DM**	AUM	EL2N
Scene	**AUM**	DM	EL2N	**AUM**	DM	EL2N	**AUM**	DM	EL2N	**SL**	EL2N	DM

Method Category: Neighborhood Training / loss Prediction HOC

(b) Examples with Qwen3-VL fixed

Ability	10%	20%	30%	40%
Absolute Depth	Simi	Simi	Simi	Simi
Relative Depth	Simi	AUM	AUM	AUM
Texture	KNN	AUM	DM	DM

Figure 2: Detector preferences on AVA-Bench. Rows denote abilities; columns denote label-noise rates. (a) Top-3 detectors in decreasing AUROC order, each using its best measured encoder. Colors indicate method categories by diagnostic signal: neighborhood consistency, training dynamics/loss, prediction–label consistency, and HOC-derived consensus. (b) Best-detector examples with Qwen3-VL fixed; full Top-3 lists are in Appendix B, Table 8. Both panels use all five batches and seeds 0/1/2. Abbreviations: KNN: nearest-neighbor consistency; Simi: SimiFeat-Raw (Zhu et al., 2022); AUM: Area Under the Margin (Pleiss et al., 2020); DM: DataMap Hardness (Swayamdipta et al., 2020); EL2N: early-training error norm (Paul et al., 2021); SL: Small-loss (Han et al., 2018); GMM: Loss-GMM (Li et al., 2020); CL/OOF: Confident Learning/out-of-fold self-confidence (Northcutt et al., 2021b); HOC: Docta-HOC (Zhu et al., 2021; 2024). Definitions and adaptations appear in Appendix A.

[1]The permutation p-values are 0.002 for AVA-Bench and ScienceQA, rounded from the minimum attainable $1/501$, and 0.006 for MMMU. Appendix A.4 gives the permutation constraints.

4.2 Detector Preferences across Noise Levels

To understand the diagnostic differences across tasks, we compare detector preferences at different noise levels. Figure 2(a) ranks detectors by AUROC, using each detector's best measured encoder. We first examine the highest-ranked detector: depth, orientation, spatial relations, and texture favor neighborhood-consistency methods at low noise and training-based methods at higher noise, whereas recognition, scene, action, and emotion favor training-based methods across all tested levels. The relationship between noise level and detector preference thus varies across tasks.

Extending the comparison to the Top-3 reveals the same pattern in the composition of high-ranking methods. Across the ten abilities, neighborhood-consistency methods and HOC account for 43% of the Top-3 positions at the lowest noise level and none at the highest, where all Top-3 positions are occupied by training-based methods.

Some of these patterns persist with the encoder fixed. Under Qwen3-VL, the best detectors for relative depth and texture still use neighborhood-consistency signals at low noise and training-based signals at higher noise, while SimiFeat ranks first for absolute depth throughout (Figure 2b). Across all ten abilities, neighborhood-consistency methods and HOC account for 30% of the Top-3 positions at the lowest noise level and 10% at the highest. Together, these comparisons reveal consistent task- and noise-related patterns in detector preferences, some of which persist under a fixed representation, with their specific form also affected by encoder choice. Figure rows group related abilities for presentation; the original categories remain the analysis units.

4.3 Task-Specific Encoder Preferences

The leading configurations use VLM-derived representations, but their preferred encoders differ across abilities. In AVA-Bench's noise-averaged task profiles, InternVL3.5 leads for depth and spatial relations, Qwen2.5-VL for fine-grained recognition and texture, and InternVL2.5 for action and orientation. The complete profiles appear in Appendices B and D.

Within a task, the preferred encoder changes in 4 of 30 adjacent-rate comparisons. Detector changes include both signal-family changes and reordering among related methods.[2] On AVA-Bench, encoder preferences thus differ across abilities but change little across noise levels, while detector preferences vary with both ability and noise.

5 The TAG Framework

5.1 Framework Design

Building on these preference patterns, TAG diagnoses incorrect answer annotations by selecting an encoder for each task and combining detectors selected for each task and noise state (Figure 1). Given a test batch without task labels, TAG identifies each question's task from its image–question–answer representation and estimates the batch's noise state. A confidence gate routes uncertain task assignments to the shared strategy.

A frozen Qwen3-VL encoder represents the image, question, and answer marked as correct in the observed labels. We average the normalized training-reference embeddings within each task category and normalize the resulting prototype. An incoming question is assigned to the category whose prototype has the highest cosine similarity. This adopts the class-mean prototype principle (Snell et al., 2017) with a frozen multimodal encoder and cosine matching. Appendix A.3 specifies the reference condition used to build these prototypes.

The noise estimate supplies a coarse condition for detector selection. We use three fixed states, $\mathcal{C} = \{\mathrm{Low}, \mathrm{Medium}, \mathrm{High}\}$: Low pools the 10% and 20% conditions, while Medium and High correspond to 30% and 40%. This grouping shares a selection rule across the two lower rates and retains separate rules at higher noise levels. A fixed AUM probe (Pleiss et al., 2020) provides a direction-aligned mean risk for the incoming batch. We match it to the closest state reference,

[2]The leading detector changes 13 times: 8 changes reorder training-based methods, while 5 switch from neighborhood to training-based signals.

computed by averaging training-batch probe means over that state's rates and seeds. The resulting state is a discrete routing input; evaluation continues to report all four corruption rates.

We select one encoder for each task by its average diagnostic utility across states. Let $U(e, d \mid t, c)$ denote the within-task AUROC of encoder e paired with detector d on task t, measured on training batches and averaged over the noise rates and seeds in noise state c. The selected encoder is

$$e^*(t) = \arg\max_e \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \max_d U(e, d \mid t, c). \tag{1}$$

The three states receive equal weight. Within the selected encoder, we retain the three detectors with the highest utility for each task–state pair. This allows the diagnostic signal to adapt to noise while keeping the representation tied to the task. To combine scores with different numerical scales, we transform each direction-aligned risk into its rank within the corresponding training-reference distribution and average the three normalized ranks. The transform uses the empirical CDF with the finite-reference normalization specified in Appendix A.3. A higher normalized score places the question further toward the high-risk end of the training-reference distribution.

Task-dependent selection also requires a reliable task assignment. Following the use of the gap between the two leading class scores as a confidence measure (Scheffer et al., 2001), we take the difference between the two largest prototype similarities. Inner leave-one-batch-out fitting chooses the smallest margin threshold τ at which retained assignments achieve at least 95% empirical task accuracy. Questions below this threshold use the shared strategy for the estimated noise state, with the same rank normalization. This gate links fallback to uncertainty in task identification; its effect on diagnostic performance is evaluated in Section 5.4. Fitting conditions and threshold edge cases appear in Appendix A.3.

5.2 Baselines

Our primary comparator is the best shared strategy (Global), selected within each dataset for each noise state and applied across tasks according to the estimated state. Selection maximizes question-weighted within-task AUROC on the training folds, matching the evaluation objective used for TAG. This comparison tests the benefit of allowing diagnosis to depend on the task.

Two intermediate variants separate task conditioning from score averaging. Global-top3 uses one dataset-level encoder and averages three globally selected detectors per state. The dataset-encoder TAG variant retains that encoder but selects its detectors separately for each task and state. The main method then allows the encoder to vary by task, reselecting detectors within the chosen representation. These comparisons trace the contributions of task-conditioned detector selection and task-specific representations.

We also compare with direct VLM judgment using Qwen2.5-VL, including matched-checkpoint comparisons with representation-based detection. Unlike the full-data, task–state Oracle in Section 4.1, TaskBest selects the best single configuration for each test task, rate, and seed within each fold using evaluation annotations. Its difference from the training-selected Global is the *TaskBest reference gap*, used as the denominator of Recovery. Both Oracle and TaskBest are post-hoc references within the single-configuration pool; TAG combines detectors selected from training data.

5.3 Diagnostic Performance

Table 2 shows that TAG raises within-ability AUROC from 0.9339 to 0.9464 and AP from 0.8595 to 0.8845, recovering 93.3% of the reference gap between Global and TaskBest. The AUROC gain over Global is 1.24 percentage points, with a 95% paired-bootstrap confidence interval of 1.15–1.33 points. Global-top3 improves on Global by just 0.01 points, while TAG improves on Global-top3 by 1.23 points. The gain therefore comes primarily from task-conditioned selection. Reviewing the highest-risk 5% of questions within each task gives 94.77% precision for TAG versus 93.39% for Global on AVA-Bench (Appendix C.2).

Table 2: Diagnostic performance on AVA-Bench. AUROC: mean ± standard deviation across five folds; higher AUROC and AP are better. Gain is the AUROC difference from Global in percentage points. TaskBest is a post-hoc single-configuration reference. Paired-bootstrap results: Appendix C, Table 14.

Method	AUROC	AP	Gain (pp)
Shared-strategy diagnosis			
VLM judge	0.9003 ± 0.0023	0.7964	-3.36
Global	0.9339 ± 0.0006	0.8595	—
Global-top3	0.9340 ± 0.0006	0.8597	$+0.01$
Task-conditioned diagnosis			
TAG, dataset encoder	0.9395 ± 0.0011	0.8700	$+0.56$
TAG	$\mathbf{0.9464 \pm 0.0010}$	**0.8845**	$\mathbf{+1.24}$
TaskBest (reference)	0.9473 ± 0.0013	0.8827	$+1.33$

With the encoder and averaging rule fixed, task–state detector selection improves on Global-top3 by a bootstrap mean of 0.54 percentage points. Task-specific encoders, with detector lists reselected accordingly, add a further 0.69 points. Retaining task conditioning but pooling noise states for strategy selection and score calibration reduces AUROC from 0.9464 to 0.9442 on AVA-Bench, supporting noise conditioning in this setting (Appendix C.1). We adopted task-specific encoders after inspecting full-data profiles. This retrospective evaluation uses one batch partition; bootstrap intervals condition on fitted rules. Selecting designs by cross-validation within the training data preserves the 1.24-point AUROC gain over Global on the held-out AVA-Bench test folds (Appendix C.1).

Gains are positive at all four noise rates: 1.61, 1.19, 0.85, and 1.33 points (Figure 3). The largest gains align with the preference profiles. Absolute depth and orientation gain 3.89 and 4.31 points on average, with large low-noise gains where neighborhood signals lead. Fine-grained recognition changes from -0.45 points with the dataset encoder to $+6.97$ with the task encoder; texture changes from $+0.35$ to $+1.81$. Recognition, emotion, and scene classification have near-zero mean gains, including scene classification despite its encoder change.

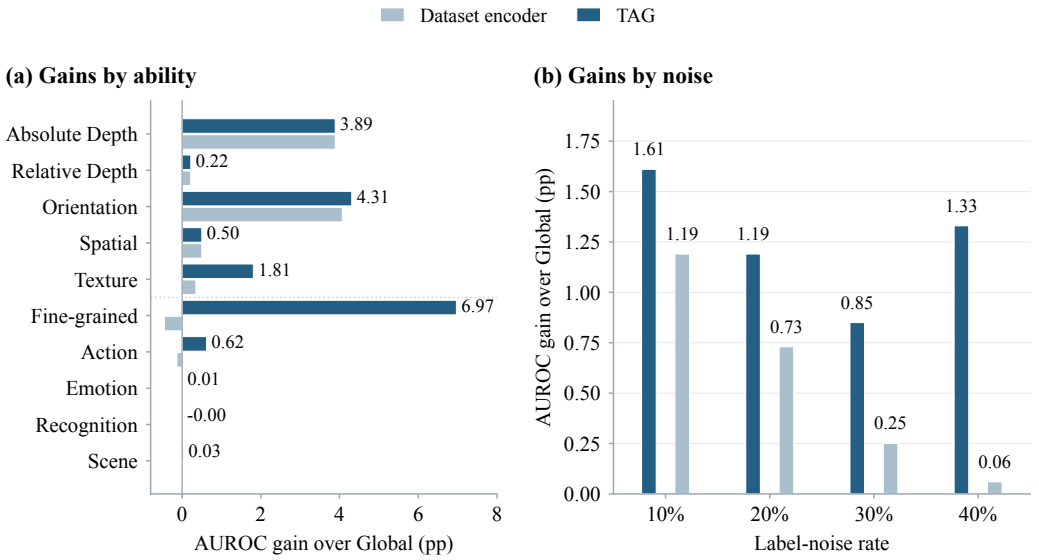

Figure 3: AUROC gains over Global on AVA-Bench (percentage points). Dark/light blue: task-specific/dataset-level encoders. (a) Horizontal axis: gain; vertical axis: ability. (b) Horizontal axis: noise rate; vertical axis: gain. Bars average three seeds and five held-out folds; panel (a) also averages four noise rates.

The same-checkpoint comparisons also favor representation-based diagnosis on AVA-Bench: Qwen2.5-VL representations with a training-selected detector achieve 0.9245 AUROC versus 0.9003 for direct judgment; InternVL2.5 gives 0.9253 versus 0.9025. Prompts and diagnostic procedures differ even when weights are shared. An equal-weight ensemble of the two judges reaches 0.9141 AUROC, below TAG's 0.9464 (Appendix C.3).

AVA-Bench selects $\tau = 0$ in all five folds, so the gate leaves these results unchanged. We next examine datasets with different category structures.

5.4 When Does Task-Conditioned Diagnosis Help?

The same training-based criterion yields fallback rates of 0% on AVA-Bench, approximately 16% on ScienceQA, and 69% on MMMU (Table 3); thresholds are fitted separately within each known dataset.

Table 3: TAG performance and fallback across datasets. Gains are AUROC differences from Global (pp). Fallback rate counts questions routed to Global; "without gate" disables fallback. Oracle gaps use full data (Table 1); other results use cross-validation. Statistical results: Appendix C, Tables 14 and 18.

Measure / setting	**AVA-Bench**	**ScienceQA**	**MMMU**
Category basis	Visual ability	Task patterns	Discipline
Oracle gap (pp)	1.25	0.28	0.44
Fallback rate	0%	16%	69%
Global AUROC	0.9339	0.9749	0.6771
TAG AUROC	0.9464	0.9765	0.6788
TAG − Global (pp)	+1.24	+0.16	+0.17
TAG without gate − Global (pp)	+1.24	+0.11	+0.41
VLM judge − Global (pp)	−3.36	−5.36	+3.22

On ScienceQA, the gate increases the observed gain from 0.11 to 0.16 percentage points; the paired-bootstrap interval excludes zero only with the gate. On MMMU, it reduces the observed gain from 0.41 to 0.17 points, and gains remain uncertain with or without it. Forced task-conditioned scoring on evaluable fallback questions underperforms Global on ScienceQA but outperforms it on MMMU. The gate therefore offers a conservative routing choice whose diagnostic effect differs between the two datasets. Full intervals and fallback-subset results appear in Appendix C.

In cross-validation, Global reaches 0.9749 AUROC on ScienceQA, with a TaskBest reference gap of 0.59 pp. On MMMU, the gap reaches 2.52 pp, but TAG realizes just 0.17 pp (Appendix C). In the dataset-encoder stability analysis, detector Top-1 agreement is 81.1% on MMMU versus 95.3% on AVA-Bench.

On MMMU, the Qwen2.5-VL judge reaches 0.7093 AUROC, above the 0.7023 TaskBest single-strategy reference. With Qwen2.5-VL held fixed, direct judgment also exceeds representation-based diagnosis on ScienceQA (0.9213 versus 0.8817) and MMMU (0.7093 versus 0.6204). ScienceQA's best shared configuration across encoders remains stronger than that judge (0.9749). The preferred diagnosis mode thus varies with the dataset and representation.

6 Conclusion

Diagnosing incorrect answer annotations in multimodal data reveals task- and noise-dependent preferences that shared strategies leave unused. TAG turns these patterns into task-specific representations and noise-conditioned detector selection, improving held-out diagnosis on AVA-Bench. Confidence-based fallback keeps performance close to the shared strategy on ScienceQA and MMMU, where gains remain small with or without it. Extending this evaluation to natural annotation errors and downstream training is future work.

AI USE STATEMENT

LLM-based assistants were used to aid writing and improve the grammar, clarity, organization, and LaTeX formatting of the manuscript. An LLM-based coding assistant was also used to assist with writing and editing experiment and analysis code based on specifications provided by the authors. All AI-assisted text and code were reviewed, tested, and verified by the authors, who remain responsible for the accuracy, originality, and final content of the paper.

REPRODUCIBILITY STATEMENT

Sections 3 and 5 describe the data construction, evaluation protocol, and TAG framework. Appendix A details the models, data grouping, fold construction, and statistical procedures. Appendices B–D provide diagnostic preference analyses, cross-validation results, and selection rules. The supplementary ZIP contains experiment and analysis scripts, configuration files, and instructions for obtaining the datasets and models, preparing the inputs, and running the pipeline. Result tables are reported in the paper and appendices.

REFERENCES

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. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Jiankang Chen, Tianke Zhang, Changyi Liu, Haojie Ding, Yaya Shi, Feng Cheng, Huihui Xiao, Bin Wen, Fan Yang, Tingting Gao, and Di Zhang. Taskgalaxy: Scaling multi-modal instruction fine-tuning with tens of thousands vision task types. In *International Conference on Learning Representations*, 2025.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4156–4172, 2024a. doi: 10.18653/v1/2024.findings-acl.246.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*, 2024b. URL https://arxiv.org/abs/2412.05271.

Guanjie Cheng, Boyi Li, Lingyu Sun, Mengying Zhu, Yangyang Wu, Xinkui Zhao, and Shuiguang Deng. Tads: Task-aware data selection for multi-task multimodal pre-training. *arXiv preprint arXiv:2602.05251*, 2026.

Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. doi: 10.1109/CVPR42600.2020.00370.

Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, and Seong Joon Oh. Eccv caption: Correcting false negatives by collecting machine-and-human-verified image-caption associations for ms-coco. In *Computer Vision – ECCV 2022*, pp. 1–19. Springer, 2022.

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. doi: 10.1145/1143844.1143874.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzić, and Olivier J. Hénaff. Data curation via joint example selection further accelerates multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4485.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, et al. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2Vec: Training Vision-Language Models for Massive Multimodal Embedding Tasks. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05160.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://github.com/lupantech/ScienceQA.

Martin Q. Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. A large-scale study on unsupervised outlier model selection: Do internal strategies suffice? *arXiv preprint arXiv:2104.01422*, 2021. URL https://arxiv.org/abs/2104.01422.

Zheda Mai, Arpita Chowdhury, Zihe Wang, Sooyoung Jeon, Lemeng Wang, Jiacheng Hou, Jihyung Kil, and Wei-Lun Chao. Ava-bench: Atomic visual ability benchmark for vision foundation models. *arXiv preprint arXiv:2506.09082*, 2025.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems*, volume 26, pp. 1196–1204, 2013.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021a. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/f2217062e9a397a1dca429e7d70bc6ca-Abstract-round1.html.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021b. doi: 10.1613/jair.1.12125.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Geoff Pleiss, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

John R. Rice. The algorithm selection problem. In *Advances in Computers*, volume 15, pp. 65–118. Academic Press, 1976. doi: 10.1016/S0065-2458(08)60520-3. URL https://www.sciencedirect.com/science/article/pii/S0065245808605203.

Shaheer U. Saeed, Yunguan Fu, Zachary M. C. Baum, Qianye Yang, Mirabela Rusu, Richard E. Fan, Geoffrey A. Sonn, Dean C. Barratt, and Yipeng Hu. Learning image quality assessment by reinforcing task amenable data selection. In *Information Processing in Medical Imaging*, pp. 755–766. Springer, 2021.

Shaheer U. Saeed, Yunguan Fu, Vasilis Stavrinides, Zachary M. C. Baum, Qianye Yang, Mirabela Rusu, Richard E. Fan, Geoffrey A. Sonn, J. Alison Noble, Dean C. Barratt, and Yipeng Hu. Image quality assessment for machine learning tasks using meta-reinforcement learning. *Medical Image Analysis*, 78:102427, 2022. doi: 10.1016/j.media.2022.102427.

Bardia Safaei, Faizan Siddiqui, Jiacong Xu, Vishal M. Patel, and Shao-Yuan Lo. Filter images first, generate instructions later: Pre-instruction data selection for visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14247–14256, 2025.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Mining the web with active hidden markov models. In *Proceedings of the IEEE International Conference on Data Mining*, 2001. doi: 10.1109/ICDM.2001.989591. URL https://fusion.cs.uni-magdeburg.de/pubs/icdm01.pdf.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://arxiv.org/abs/2004.09297.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 9275–9293, 2020. doi: 10.18653/v1/2020.emnlp-main.746.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Tianhao Wang, Yi Zeng, Ming Jin, and Ruoxi Jia. A unified framework for task-driven data quality management. *arXiv preprint arXiv:2106.05484*, 2021.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Bowen Zhou, Weijie Su, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025. URL https://arxiv.org/abs/2508.18265.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 54104–54132. PMLR, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://mmmu-benchmark.github.io/.

Yue Zhao, Ryan Rossi, and Leman Akoglu. Automatic unsupervised outlier model selection. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4489–4502. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/23c894276a2c5a16470e6a31f4618d73-Abstract.html.

Zhaowei Zhu, Yiwen Song, and Yang Liu. Clusterability as an alternative to anchor points when learning with noisy labels. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 12912–12923. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhu21e.html.

Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 27412–27427. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhu22a.html.

Zhaowei Zhu, Jialu Wang, Hao Cheng, and Yang Liu. Unmasking and improving data credibility: A study with datasets for training harmless language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.11202.

A Experimental Details

A.1 Data, categories, and corruption

Table 4: Dataset scale and retained data. Each retained question forms one pair with two candidate answers.

Dataset	Official total	Retained pairs	Candidate answers
ScienceQA	21,208 questions	10,332 questions	20,664
MMMU	11,550 questions	9,864 questions	19,728
AVA-Bench	218K image–question pairs	116,453	232,906

AVA-Bench retains the ten finite-choice ability subsets listed in Figure 2; Localization, OCR, Color, and Counting are excluded because their native outputs lack a fixed option set. ScienceQA retains all 10,332 image-grounded questions. MMMU retains 9,864 of the 9,873 multiple-choice questions in the official test split after checking option validity, referenced-image availability, and the modality consistency of selected answers. Each retained question supplies two candidate rows, with a shared pair identifier.

ScienceQA's ten task patterns are constructed from skill, subject, topic, and category metadata; its residual state is displayed as Other/Unresolved and stored as Others. MMMU uses six discipline labels, and AVA-Bench uses ten native ability labels. The router predicts these categories directly. Noise sampling uses ten AVA abilities, six coarse ScienceQA strata, and six MMMU disciplines. These sampling strata need not equal the finer ScienceQA analysis categories. At each rate, reversing both correctness labels within a selected pair marks the incorrect option as correct and the reference answer as incorrect. The image, question, and candidate-answer contents remain unchanged; flip records identify the injected answer errors for evaluation.

The ScienceQA mapping assigns each of 185 native skills a primary operation using its full description and subject/topic/category context. Explicitly listed ambiguous skills were reviewed with representative questions. The primary operation determines the category; secondary operations and ambiguity flags are retained in the mapping record. Detector scores, winners, and performance metrics were excluded from this semantic mapping. Table 5 summarizes the frozen definitions. The factual/conceptual group supplies the stored Others category, including content-only descriptions with insufficient evidence for a more specific operation.

Table 5: ScienceQA primary-operation mapping. Short names are those used in the analysis. Rules summarize the frozen skill-level taxonomy; the assignment is made from the full skill and its documented context.

Category	Assignment criterion
Localization	Locate or orient using maps, coordinates, directions, regions, or grids.
Attribute Recognition	Recognize depicted objects, materials, attributes, structures, or visible functional adaptations.
Categorization	Match entities or descriptions to classes using defining criteria.
Comparison	Infer qualitative relations, compatibility, ordering, or correspondence.
Quantitative	Use measurement, magnitude, ratio, probability, or calculation.
Causality	Explain or predict effects of mechanisms, interventions, or processes.
Data Interpretation	Extract and use patterns in tables, graphs, observations, or data.
Experiment	Identify or assess experimental questions, variables, controls, evidence, claims, or solutions.
Diagram	Interpret structured models, food webs, cycles, timelines, or symbolic diagrams.
Language	Interpret prose, vocabulary, rhetorical strategies, or language details.
Others	Retrieve or describe facts or concepts when a more specific operation is unsupported.

A.2 ENCODERS AND DIAGNOSTIC SIGNALS

A strategy consists of an encoder and a detector, $s = (e, d)$. Given an image–question–answer example x_i, the encoder supplies a representation $h_i = E_e(x_i)$; the detector combines representations with observed labels, predictions, or training statistics to produce risk $q_i = F_s(x_i)$. Here F_s includes any dataset-level computation required by the detector. This separation lets us examine both how examples are represented and which signal is used to judge their labels.

Neighborhood consistency. KNN and SimiFeat-Raw assess whether an observed label agrees with labels of nearby examples in representation space (Zhu et al., 2022). Local disagreement supplies evidence of an atypical label. Its usefulness depends on whether the representation places relevant examples together and whether their observed labels provide a reliable reference. This makes neighborhood scores a natural candidate for studying task and noise dependence.

Prediction–label consistency. OOF Self-Confidence and Confident Learning use predicted class probabilities to assess observed labels (Northcutt et al., 2021b). Self-Confidence measures the probability assigned to the observed class; Confident Learning uses confident predictions to estimate label inconsistencies. OOF predictions assess a sample using a model fitted outside its fold. These signals capture disagreement between a learned decision rule and the annotation, which can depend on how well the rule represents a particular task.

Training dynamics and loss. These scores describe how a model learns the observed label. Classical AUM averages the margin between the observed-class logit and the strongest alternative during training (Pleiss et al., 2020). DataMap Hardness uses one minus the mean confidence in the observed class (Swayamdipta et al., 2020), and Forgetting Events counts transitions from correct to incorrect prediction relative to that label (Toneva et al., 2019). Small-loss uses early training loss (Han et al., 2018); Loss-GMM uses the posterior of the higher-loss component in a two-component Gaussian mixture (Li et al., 2020). EL2N Early measures early prediction error through its L_2 norm (Paul et al., 2021). These signals can highlight labels that are difficult to fit, but that difficulty alone does not establish that a label is wrong. Their value for diagnosis is evaluated against the known swaps.

Table 6: Diagnostic evidence represented in the candidate pool. The groups organize the signals used by the detectors; they are not mutually exclusive algorithm classes.

Evidence	Detectors
Neighborhood consistency	KNN, SimiFeat-Raw
Prediction–label consistency	OOF Self-Confidence, Confident Learning
Training dynamics and loss	Classical AUM, DataMap Hardness, Forgetting Events, Small-loss, Loss-GMM, EL2N Early
HOC-derived diagnosis	Docta-HOC (multi-round selection frequency)

Table 6 summarizes the eleven detectors, including Docta-HOC scores (Zhu et al., 2021; 2024). These are diagnostic adaptations of the cited methods. We combine them with eight encoders: MPNet (Song et al., 2020), CLIP (Radford et al., 2021), SigLIP2 (Tschannen et al., 2025), VLM2Vec (Jiang et al., 2025), Qwen3-VL-Embedding (Li et al., 2026), Qwen2.5-VL-7B-Instruct (Bai et al., 2025), InternVL3.5 (Wang et al., 2025), and InternVL2.5-8B (Chen et al., 2024b). MPNet is the only text-only encoder in this pool. Our CLIP and SigLIP2 adapters concatenate image and text embeddings, while the other five encoders process image–text inputs jointly. These combinations yield $8 \times 11 = 88$ diagnosis configurations, whose base scores remain fixed during strategy selection.

The batch-isolated experiment uses a single Docta-HOC definition throughout: multi-round HOC selection frequency. Neighborhood methods use $k = 10$. OOF logistic regression uses three group-isolated folds, $C = 1$, at most 300 iterations, and up to 256 PCA components. Training-dynamics probes use 30 epochs, batch size 256, learning rate 10^{-3}, and weight decay 10^{-4}. Loss-GMM uses two full-covariance components, ten initializations, and at most 200 iterations. All encoders are frozen.

A.3 FIVE-FOLD BATCH CROSS-VALIDATION

The fixed association-group partition uses seed 20260924. The retained groups were first split into 80% and 20% partitions, and the former into four batches. The evaluation rotates the held-out role over all five batches, including the original 20% partition. The batches are approximately equal, with task stratification subject to group isolation. Associated groups combine audited image hashes and source-question associations for ScienceQA, source and stored-image associations for MMMU, and source-question/image-reference associations for AVA-Bench. Known group intersections are zero; no additional pixel-content deduplication was performed. Both answer candidates and every rate–seed version of a group stay in one batch.

Neighborhood references, OOF models, and training dynamics were computed separately within each batch. Associated examples are excluded from one another's neighborhood references and kept together across OOF folds. Detectors can fit to the observed labels in the batch being diagnosed: the evaluation is transductive batch diagnosis. Clean correctness labels, flip indicators, and task labels are withheld from evaluation-batch detector fitting. Encoders remain frozen. This analysis reuses the batch-isolated risks; it neither reuses a full-pool fitted detector nor reruns the base detectors for each outer fold.

Each outer fold uses four batches for task prototypes, noise references, risk directions, strategy utilities, and empirical score distributions. Rates are 10/20/30/40%, with noise seeds 0, 1, and 2 throughout. For each task, rate, and seed, training utilities compute AUROC after concatenating the four training batches. Rate–seed utilities are averaged within Low (10/20%), Medium (30%), and High (40%). Task-encoder selection gives the three states equal weight, whereas reported performance gives the four actual rates equal weight.

Risk direction and integration. AVA-Bench and MMMU use a shared training-fitted direction across conditions; ScienceQA first fits rate-specific directions and resolves any Low-state conflict using mean pooled training AUROC. No direction uses evaluation annotations. For each strategy and state, direction-aligned pair risks from the corresponding training conditions form a sorted reference $\mathcal{V}_{s,c}$. A risk z maps to $F_{s,c}(z) = \sum_{v \in \mathcal{V}_{s,c}} \mathbf{1}[v \le z]/(|\mathcal{V}_{s,c}| + 1)$. TAG averages the three selected percentiles. The shared strategy and Global-top3 use the same fold-trained direction and percentile convention. This transformation provides rank normalization without probability calibration.

Routing and noise. Qwen3-VL-Embedding supplies normalized representations of the image, question, and observed-positive answer. Training prototypes are normalized class means at 20% noise, seed 1. Noise probes are Classical AUM with InternVL2.5-8B for ScienceQA and Qwen3-VL-Embedding for AVA/MMMU. One probe direction is shared across rates. Each state reference averages training-batch probe means over its rates and seeds; the evaluation batch is assigned the nearest reference in absolute distance. Dataset identity is known.

Confidence fitting. Within the four outer training batches, each batch is held out in turn while prototypes are built on the other three. Inner validation pools all four rates and all three seeds. For target accuracy a, the gate selects the smallest observed margin whose retained predictions reach a. It uses $\tau = 0$ if all predictions reach the target, and $\tau = \infty$ if no retained set does. Main results use $a = 0.95$; the dataset-encoder sensitivity analysis also uses 0.90 and 0.99. A question with margin below τ receives the fold-trained percentile of the shared strategy. All ties use a fixed lexical order. The retained-set routing accuracy determines this empirical threshold. The task-encoder no-gate comparison sets $\tau = 0$ in every fold and reuses the same encoders, detector shortlists, score transforms, noise estimates, and task predictions.

Baselines and design variants. The shared strategy maximizes training question-weighted within-task AUROC over all 88 configurations separately for each state. The dataset-encoder ablation chooses one encoder by the equally weighted mean over tasks and states of its best-detector utility; its detector lists then depend on task and state. Global-top3 uses that same encoder but selects its three detectors by the task-weighted global objective. Main TAG chooses an encoder separately per task by averaging its best-detector utility equally over the three states. The main method and ablation share task predictions, confidence margins, gate thresholds, noise estimates, and fallback outputs.

Direct judgments. Cached answer-correctness probabilities give candidate risk $|p - \widetilde{\ell}|$, followed by the two-answer mean. The main judge is Qwen2.5-VL-7B-Instruct. Matched-checkpoint com-

parisons fix Qwen2.5-VL for all three datasets and InternVL2.5-8B for AVA, and choose the corresponding representation-based detector using training within-task AUROC. Checkpoint revisions match, while prompts and diagnostic procedures differ.

A.4 Oracle gap and random-grouping control

Table 1 pools the five independently scored batches for descriptive selection. For each state, the shared strategy maximizes sample-weighted within-task AUROC, while the task–condition oracle maximizes within each task before weighting. The Low state has weight one-half and Medium/High one-quarter each, equivalent to averaging four rates equally. Risk directions remain fixed during the permutation analysis.

We generate 500 task-label permutations at association-group level with seed 20260925 and repeat both selections for each. Single-task groups exchange their task labels while preserving the number of such groups per task. Multi-task groups exchange complete label vectors only with multi-task groups of the same size, ordered by canonical pair ID; singleton size strata stay fixed. This preserves group-level constraints; question counts per task may vary. Task weights are recalculated for each permutation. The one-sided value is $(1 + \#\{H_b \geq H_{\text{observed}}\})/501$, where H_b is the permuted oracle gap measured in AUROC. The smallest attainable value is $1/501 \simeq 0.002$. The test concerns the value of the actual grouping under these exchangeability constraints; it does not identify why benchmark categories differ.

A.5 Metrics, paired bootstrap, and references

Within each fold–rate–seed condition, a metric is computed in every true task and weighted by its question count. The twelve conditions are averaged equally within each fold, followed by the mean and sample SD ($\text{ddof} = 1$) across five folds. AUROC and average precision use the same score vector. AP uses the non-interpolated average precision definition. Datasets are reported separately. AUROC differences are computed before rounding, and main-table fallback rates are rounded to whole percentages. Fold SD summarizes variation across the five folds of this partition, whose training sets overlap.

The paired bootstrap uses 1,000 replicates with seed 20260925. Within each outer evaluation batch, association groups are sampled with replacement within majority-task strata, preserving each stratum's group count; ties in a multi-task group's majority label use lexical order. All methods, rates, seeds, and answer candidates share each group's bootstrap multiplicity. Single-class task–rate–seed cells are skipped within that replicate and the remaining task weights renormalized. Their mean skipped fraction is 0.0052 on ScienceQA and zero on AVA/MMMU. Percentile intervals use the 2.5th and 97.5th percentiles. Fold-fitted rules and base scores remain fixed, so intervals describe evaluation-sample uncertainty conditional on those fits; they do not include repeated method design, selection, or detector fitting.

TaskBest selects the best single configuration within each outer test task, actual rate, and seed, using evaluation flip indicators and training-only score transforms. It is neither an upper bound on ensembles nor on judges outside the candidate pool. Recovery is the ratio of aggregated TAG gain over the shared strategy to the aggregated gap between TaskBest and the shared strategy. In particular, the all-data descriptive Oracle and cross-validated TaskBest use different selection units and are reported separately. Statistical intervals including zero do not establish equivalence or non-inferiority.

B Complete Conditional Preference Profiles

These profiles use all five batches and seeds 0/1/2. At each task and actual noise rate, each detector receives its best measured encoder before distinct detectors are ranked. Entries list Top-3 detectors from left to right. Abbreviations follow Figure 2: Simi (SimiFeat-Raw), DM (DataMap Hardness), AUM (Classical AUM), EL2N (EL2N Early), SL (Small-loss), GMM (Loss-GMM), CL (Confident Learning), OOF (OOF Self-Confidence), FE (Forgetting Events), and HOC (Docta-HOC).

Table 7: AVA-Bench descriptive detector preferences.

Task	10%	20%	30%	40%
Absolute Depth	Simi, KNN, HOC[a]	Simi, KNN, HOC[a]	Simi, KNN, GMM[a]	GMM, DM, AUM[a]
Action	AUM, DM, EL2N[b]	AUM, DM, EL2N[b]	EL2N[b], AUM[c], DM[c]	EL2N[c], SL[b], AUM[c]
Emotion	AUM, DM, KNN[a]	AUM, DM, EL2N[a]	AUM, DM, EL2N[a]	DM, AUM, EL2N[a]
Fine-grained	DM, AUM, EL2N[c]	DM, AUM, EL2N[c]	DM, EL2N, AUM[c]	SL, EL2N, DM[c]
Orientation	KNN, Simi, HOC[c]	KNN[c], Simi[c], GMM[a]	GMM[a], DM[b], AUM[b]	DM, AUM, EL2N[b]
Recognition	DM, AUM, CL[a]	DM, AUM, EL2N[a]	DM, AUM, EL2N[a]	DM, AUM, EL2N[c]
Relative Depth	Simi, KNN, AUM[a]	GMM, AUM, DM[a]	GMM, DM, AUM[a]	DM, AUM, EL2N[a]
Scene Classification	AUM, DM, EL2N[c]	AUM, DM, EL2N[c]	AUM, DM, EL2N[c]	SL, EL2N, DM[c]
Spatial	KNN, Simi, AUM[a]	KNN, AUM, DM[a]	AUM, DM, EL2N[a]	DM, AUM, EL2N[a]
Texture	KNN, Simi, CL[c]	KNN, CL, OOF[c]	DM, AUM, GMM[c]	EL2N, DM, AUM[c]

A single suffix applies to all three detectors. Encoder keys: [a]InternVL3.5; [b]InternVL2.5-8B; [c]Qwen2.5-VL-7B-Instruct.

Table 8: Fixed-encoder examples corresponding to Figure 2(b). Each cell lists the Top-3 detectors in decreasing AUROC order, using all five batches and seeds 0/1/2. Qwen3-VL is fixed across all three tasks and four noise rates.

Task	10%	20%	30%	40%
Absolute Depth	Simi, KNN, HOC	Simi, KNN, HOC	Simi, KNN, HOC	Simi, KNN, HOC
Relative Depth	Simi, AUM, DM	AUM, DM, EL2N	AUM, DM, EL2N	AUM, DM, EL2N
Texture	KNN, AUM, DM	AUM, DM, EL2N	DM, AUM, EL2N	DM, AUM, SL

Table 9: ScienceQA descriptive detector preferences.

Task	10%	20%	30%	40%
Attribute Recognition	EL2N, DM, AUM[b]	EL2N, DM, AUM[b]	EL2N, DM, AUM[b]	SL, EL2N, DM[b]
Categorization	SL, EL2N, AUM[b]	SL, EL2N, AUM[b]	SL, EL2N, AUM[b]	SL, EL2N, AUM[d]
Causality	AUM, DM, EL2N[b]	EL2N, AUM, DM[b]	EL2N, SL, DM[b]	SL, EL2N, DM[b]
Comparison	CL, OOF, AUM[b]	AUM, DM, EL2N[b]	DM, AUM, EL2N[b]	SL, EL2N, DM[b]
Data Interpretation	SL, AUM, DM[b]	SL, EL2N, AUM[b]	SL, EL2N, AUM[b]	SL, EL2N, AUM[b]
Diagram	EL2N, SL, AUM[b]	SL, EL2N, DM[b]	SL, EL2N, DM[b]	SL, EL2N, DM[b]
Experiment	DM, AUM, EL2N[e]	DM, AUM, EL2N[e]	EL2N, DM, AUM[e]	EL2N, SL, DM[e]
Language	SL, EL2N, DM[b]	EL2N, SL, DM[b]	EL2N, SL, DM[b]	SL, EL2N, DM[b]
Localization	AUM, EL2N, DM[b]	EL2N, AUM, DM[b]	EL2N, SL, AUM[b]	SL, EL2N, DM[b]
Others	AUM, DM, EL2N[b]	AUM, DM, EL2N[b]	AUM, DM, EL2N[b]	EL2N, SL, AUM[b]
Quantitative	Simi, CL, OOF[b]	DM, AUM, Simi[b]	DM, AUM, EL2N[b]	EL2N, SL, DM[b]

A single suffix applies to all three detectors. Encoder keys: [b]InternVL2.5-8B; [d]Qwen3-VL-Embedding; [e]VLM2Vec.

Table 10: MMMU descriptive detector preferences.

Task	10%	20%	30%	40%
Art & Design	DM, AUM, EL2N[b]	DM, AUM, EL2N[d]	EL2N, DM, AUM[d]	EL2N, SL, AUM[d]
Business	DM, AUM, EL2N[a]	EL2N, DM, AUM[a]	AUM, DM, EL2N[b]	CL, OOF, AUM[b]
Health & Medicine	SL, EL2N, DM[d]	EL2N, SL, DM[d]	EL2N, SL, DM[d]	EL2N, SL, AUM[d]
Humanities & Social Science	DM, AUM, EL2N[b]	AUM, DM, EL2N[d]	EL2N, AUM, DM[d]	EL2N, SL, AUM[d]
Science	DM, AUM, EL2N[d]	EL2N, SL, DM[d]	EL2N, DM, AUM[d]	EL2N, SL, DM[d]

Table 10 (continued)

Task	10%	20%	30%	40%
Tech & Engineering	AUM, DM, EL2N[d]	EL2N, AUM, DM[d]	SL, EL2N, DM[d]	SL, EL2N, AUM[d]

A single suffix applies to all three detectors. Encoder keys: [a]InternVL3.5; [b]InternVL2.5-8B; [d]Qwen3-VL-Embedding.

Shortlist stability. Top-1 agreement and Top-3 overlap compare each four-batch training shortlist with the all-data reference for the dataset-encoder variant. Top-3 overlap is intersection size divided by three, equally averaged over task–state–fold cells. The shared observations make this a leave-one-batch sensitivity summary.

Table 11: Shortlist stability (dataset encoder) and adjacent-rate winner changes (full candidate pool).

Dataset	Top-1 agreement	Top-3 overlap	Encoder changes	Detector changes
AVA-Bench	95.33%	99.33%	4/30	13/30
ScienceQA	89.70%	94.75%	1/33	15/33
MMMU	81.11%	88.52%	3/18	10/18

B.1 Ability and source grouping

Upstream sources can be traced for 61,056 of the 116,453 AVA-Bench questions (52.43%). On this subset, Global is reselected and ability- and source-conditioned variants use the same outer splits, candidate pool, and evaluation tasks. Both variants receive known group labels: six ability groups versus ten source groups. Table 12 shows comparable gains from both groupings. This comparison leaves the respective contributions of ability and source unresolved; it does not test automatic routing or fully source-isolated detector fitting.

Table 12: Known-group comparison on the source-traceable AVA-Bench subset. Global is reselected on the same subset. AUROC and AP are five-fold means; gains are AUROC percentage points relative to this matched Global. Both conditioned variants use known group labels.

Method	**AUROC**	**AP**	**Gain (pp)**
Global	0.9844	0.9568	—
Global-top3	0.9845	0.9574	+0.01
Ability-conditioned	0.9902	0.9727	+0.59
Source-conditioned	0.9904	0.9732	+0.60

C Complete Cross-Validated Results

TAG uses a task-specific encoder; Dataset encoder denotes the original design retained as an ablation. AUROC/AP entries give five-fold means and sample SDs. Observed gains and bootstrap means are reported separately, preserving their original precision. Grouped tables report each dataset independently.

Table 13: Cross-validated results; gains are observed differences from Global.

Method	AUROC	AP	Gain (pp)
AVA-Bench			
Global	0.9339 ± 0.0006	0.8595 ± 0.0036	0.00
Global-top3	0.9340 ± 0.0006	0.8597 ± 0.0035	0.01
Qwen2.5 judge	0.9003 ± 0.0023	0.7964 ± 0.0054	-3.36
TAG	0.9464 ± 0.0010	0.8845 ± 0.0031	1.24
Dataset encoder	0.9395 ± 0.0011	0.8700 ± 0.0035	0.56
TaskBest	0.9473 ± 0.0013	0.8827 ± 0.0044	1.33

Table 13 (continued)

Method	AUROC	AP	Gain (pp)
ScienceQA			
Global	0.9749 ± 0.0023	0.9333 ± 0.0082	0.00
Global-top3	0.9744 ± 0.0024	0.9324 ± 0.0083	-0.05
Qwen2.5 judge	0.9213 ± 0.0169	0.8419 ± 0.0234	-5.36
TAG	0.9765 ± 0.0020	0.9388 ± 0.0070	0.16
Dataset encoder	0.9753 ± 0.0024	0.9356 ± 0.0080	0.04
TaskBest	0.9808 ± 0.0022	0.9487 ± 0.0081	0.59
MMMU			
Global	0.6771 ± 0.0111	0.4336 ± 0.0074	0.00
Global-top3	0.6771 ± 0.0108	0.4335 ± 0.0073	-0.00
Qwen2.5 judge	0.7093 ± 0.0122	0.4747 ± 0.0190	3.22
TAG	0.6788 ± 0.0108	0.4354 ± 0.0072	0.17
Dataset encoder	0.6771 ± 0.0111	0.4334 ± 0.0072	-0.01
TaskBest	0.7023 ± 0.0048	0.4503 ± 0.0079	2.52

Table 14: Paired group-bootstrap results for the comparisons in Tables 2 and 3, with additional baselines. Gain is the mean AUROC difference across bootstrap resamples in pp and may differ slightly from the observed gain in the main tables. Intervals are 95% percentile intervals conditional on fitted fold rules. Skipped is the mean fraction of single-class task cells.

Method	Reference	Gain (pp)	95% CI	Skipped
AVA-Bench				
TAG	Global	1.24	[1.15, 1.33]	0.0000
TAG	Global-top3	1.23	[1.15, 1.32]	0.0000
TAG	Qwen2.5 judge	4.60	[4.45, 4.76]	0.0000
Dataset encoder	Global	0.55	[0.49, 0.63]	0.0000
Dataset encoder	Global-top3	0.54	[0.48, 0.61]	0.0000
Dataset encoder	Qwen2.5 judge	3.91	[3.75, 4.08]	0.0000
TAG	Dataset encoder	0.69	[0.64, 0.74]	0.0000
ScienceQA				
TAG	Global	0.17	[0.03, 0.35]	0.0052
TAG	Global-top3	0.22	[0.08, 0.40]	0.0052
TAG	Qwen2.5 judge	5.52	[4.27, 6.92]	0.0052
Dataset encoder	Global	0.05	[-0.00, 0.11]	0.0052
Dataset encoder	Global-top3	0.09	[0.05, 0.14]	0.0052
Dataset encoder	Qwen2.5 judge	5.40	[4.20, 6.75]	0.0052
TAG	Dataset encoder	0.12	[-0.00, 0.29]	0.0052
MMMU				
TAG	Global	0.17	[-0.08, 0.45]	0.0000
TAG	Global-top3	0.18	[-0.07, 0.44]	0.0000
TAG	Qwen2.5 judge	-3.06	[-4.15, -2.03]	0.0000
Dataset encoder	Global	-0.00	[-0.03, 0.02]	0.0000
Dataset encoder	Global-top3	-0.00	[-0.03, 0.03]	0.0000
Dataset encoder	Qwen2.5 judge	-3.24	[-4.33, -2.18]	0.0000
TAG	Dataset encoder	0.18	[-0.07, 0.44]	0.0000

Table 15: Matched checkpoint and revision; prompts and diagnosis procedures differ between representation and judgment paths.

Method	AUROC	AP
AVA-Bench		
Qwen2.5-repr-Global	0.9245 ± 0.0017	0.8265 ± 0.0052
Qwen2.5 judge	0.9003 ± 0.0023	0.7964 ± 0.0054
InternVL2.5-repr-Global	0.9253 ± 0.0014	0.8429 ± 0.0043
InternVL2.5 judge	0.9025 ± 0.0020	0.7998 ± 0.0042
ScienceQA		
Qwen2.5-repr-Global	0.8817 ± 0.0122	0.7464 ± 0.0238

Table 15 (continued)

Method	AUROC	AP
Qwen2.5 judge	0.9213 ± 0.0169	0.8419 ± 0.0234
MMMU		
Qwen2.5-repr-Global	0.6204 ± 0.0102	0.3596 ± 0.0115
Qwen2.5 judge	0.7093 ± 0.0122	0.4747 ± 0.0190

Table 16: Confidence thresholds and fallback fractions, overall and by noise rate.

Fold	τ	Overall	10%	20%	30%	40%
AVA-Bench						
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ScienceQA						
1	0.0265	0.1555	0.1536	0.1549	0.1559	0.1576
2	0.0283	0.1764	0.1734	0.1762	0.1770	0.1791
3	0.0250	0.1588	0.1588	0.1591	0.1594	0.1581
4	0.0267	0.1524	0.1535	0.1545	0.1514	0.1500
5	0.0247	0.1565	0.1585	0.1576	0.1561	0.1537
MMMU						
1	0.0880	0.6785	0.6770	0.6782	0.6785	0.6805
2	0.0865	0.6787	0.6795	0.6792	0.6780	0.6783
3	0.0852	0.6837	0.6854	0.6841	0.6829	0.6824
4	0.0892	0.7032	0.7059	0.7043	0.7032	0.6994
5	0.0864	0.6859	0.6854	0.6861	0.6858	0.6863

Table 17: Subset AUROC. Accepted rows use actual scores; fallback rows force task-conditioned scoring. A task cell requires at least 20 flipped and 20 unchanged questions; remaining weights are renormalized. NaN denotes an undefined subset.

Subset	TAG	Dataset encoder	Global	Defined / 60
AVA-Bench				
accepted	0.9464	0.9395	0.9339	60/60
fallback	nan	nan	nan	0/60
ScienceQA				
accepted	0.9823	0.9809	0.9807	60/60
fallback	0.9221	0.9249	0.9269	41/60
MMMU				
accepted	0.6144	0.6136	0.6130	52/60
fallback	0.6863	0.6818	0.6824	60/60

Table 18: Confidence-target sensitivity. D: dataset-encoder ablation (all original rows); T: task-encoder main method without the gate. No gate sets $\tau = 0$ in all folds with other settings unchanged. Gains are bootstrap mean differences from Global in pp; brackets contain 95% percentile intervals conditional on fitted fold rules. The T rows provide the no-gate uncertainty estimates corresponding to Table 3. Dashes indicate intervals not reported in the sensitivity experiment.

Variant/target	AUROC	AP	Fallback	Gain [95% CI]
AVA-Bench				
D: 90%	0.9395 ± 0.0011	0.8700 ± 0.0035	0.0000	—
D: 95%	0.9395 ± 0.0011	0.8700 ± 0.0035	0.0000	—
D: 99%	0.9395 ± 0.0011	0.8700 ± 0.0035	0.0000	—
D: no gate	0.9395 ± 0.0011	0.8700 ± 0.0035	0.0000	—
T: no gate	0.9464 ± 0.0010	0.8845 ± 0.0031	0.0000	1.24 [1.15, 1.33]

Table 18 (continued)

Variant/target	AUROC	AP	Fallback	Gain [95% CI]
ScienceQA				
D: 90%	0.9754 ± 0.0024	0.9368 ± 0.0078	0.0589	—
D: 95%	0.9753 ± 0.0024	0.9356 ± 0.0080	0.1599	—
D: 99%	0.9753 ± 0.0024	0.9349 ± 0.0087	0.3146	—
D: no gate	0.9755 ± 0.0024	0.9379 ± 0.0076	0.0000	—
T: no gate	0.9760 ± 0.0024	0.9395 ± 0.0076	0.0000	0.12 [-0.04, 0.33]
MMMU				
D: 90%	0.6770 ± 0.0109	0.4334 ± 0.0070	0.5526	—
D: 95%	0.6771 ± 0.0111	0.4334 ± 0.0072	0.6860	—
D: 99%	0.6772 ± 0.0111	0.4336 ± 0.0073	0.8930	—
D: no gate	0.6767 ± 0.0113	0.4330 ± 0.0075	0.0000	—
T: no gate	0.6812 ± 0.0060	0.4376 ± 0.0070	0.0000	0.42 [-0.03, 0.86]

Task-level gains. Table 19 retains both design variants at each actual rate. Each cell lists task encoder / dataset encoder (T/D). Values are AUROC differences from the shared strategy in pp, averaged equally over seeds and folds; Mean averages the four rates. Per-ability and per-rate confidence intervals for Figure 3 were not estimated.

Table 19: Complete task-level gains, with both variants in each cell.

Task	10% (T/D)	20% (T/D)	30% (T/D)	40% (T/D)	Mean (T/D)
AVA-Bench					
Absolute Depth	8.15 / 8.15	3.90 / 3.90	2.81 / 2.81	0.71 / 0.71	3.89 / 3.89
Action	0.41 / -0.06	0.50 / -0.02	0.69 / 0.07	0.90 / -0.51	0.62 / -0.13
Emotion	-0.11 / -0.11	-0.08 / -0.08	-0.05 / -0.05	0.29 / 0.29	0.01 / 0.01
Fine-grained	5.07 / -0.17	5.62 / -0.09	7.12 / 0.28	10.08 / -1.84	6.97 / -0.45
Orientation	7.95 / 7.79	5.48 / 5.53	0.14 / 0.56	3.65 / 2.40	4.31 / 4.07
Recognition	-0.00 / -0.00	-0.00 / -0.00	-0.00 / -0.00	-0.01 / -0.01	-0.00 / -0.00
Relative Depth	0.16 / 0.16	0.19 / 0.19	0.15 / 0.15	0.37 / 0.37	0.22 / 0.22
Scene Classification	0.00 / -0.01	0.01 / 0.00	0.03 / 0.00	0.08 / -0.06	0.03 / -0.02
Spatial	1.20 / 1.20	0.70 / 0.70	-0.02 / -0.02	0.11 / 0.11	0.50 / 0.50
Texture	1.36 / 0.94	1.46 / 0.83	1.42 / -0.04	3.00 / -0.32	1.81 / 0.35
ScienceQA					
Attribute Recognition	0.00 / 0.00	0.00 / 0.00	-0.03 / -0.03	-0.23 / -0.23	-0.06 / -0.06
Categorization	0.11 / 0.11	0.07 / 0.07	0.03 / 0.03	-0.50 / -0.50	-0.07 / -0.07
Causality	0.07 / 0.00	0.06 / 0.03	-0.01 / -0.04	-0.08 / -0.09	0.01 / -0.02
Comparison	0.05 / 0.05	-0.09 / -0.09	0.05 / 0.05	-0.04 / -0.04	-0.01 / -0.01
Data Interpretation	-0.03 / -0.03	0.03 / 0.03	-0.01 / -0.01	-0.49 / -0.49	-0.13 / -0.13
Diagram	0.05 / 0.05	0.14 / 0.14	0.03 / 0.03	-0.56 / -0.56	-0.08 / -0.08
Experiment	1.75 / -0.24	1.64 / -0.15	1.83 / 0.49	3.19 / 0.24	2.10 / 0.08
Language	0.03 / 0.03	-0.28 / -0.06	-0.14 / -0.03	-0.52 / -0.50	-0.23 / -0.14
Localization	0.00 / 0.00	0.00 / 0.00	-0.00 / -0.00	-0.03 / -0.03	-0.01 / -0.01
Others	-0.13 / -0.13	-0.10 / -0.10	0.04 / 0.04	0.06 / 0.06	-0.03 / -0.03
Quantitative	1.59 / 1.66	0.64 / 0.69	-0.24 / -0.06	-0.19 / -0.10	0.45 / 0.55
MMMU					
Art & Design	0.01 / 0.01	-0.00 / -0.00	-0.10 / -0.10	-0.05 / -0.05	-0.04 / -0.04
Business	2.76 / -0.01	2.07 / -0.02	1.70 / -0.01	2.12 / 0.01	2.16 / -0.01
Health & Medicine	0.16 / 0.17	-0.04 / -0.02	-0.09 / -0.07	0.00 / 0.05	0.01 / 0.03
Humanities & Social Science	-0.05 / -0.05	-0.05 / -0.05	-0.15 / -0.15	-0.04 / -0.04	-0.07 / -0.07
Science	-0.03 / -0.11	0.01 / -0.06	0.06 / -0.02	-0.02 / -0.03	0.00 / -0.05
Tech & Engineering	-0.50 / 0.06	-0.46 / 0.07	-0.41 / -0.01	-0.13 / 0.09	-0.37 / 0.05

C.1 Design selection and noise conditioning

Design selection. Within each outer training split, leave-one-batch-out validation compares 18 fixed candidate designs: Global, Global-top3, and dataset- or task-specific encoders combined with one or three detectors and routing targets of 0, 0.90, 0.95, or 0.99. The selected design is then evalu-

ated on the outer test batch. Batch-isolated base scores are reused. All five AVA-Bench folds select task-specific encoders, three detectors, and no gate, retaining the main method's 1.24-point gain. The main gate also accepts all AVA-Bench questions. ScienceQA selects different designs across folds; MMMU selects the task-encoder, three-detector, no-gate design throughout. The candidate set comes from the earlier full-data exploration, so this check evaluates selection within that set. Table 20 reports fold means and standard deviations; no additional significance intervals are estimated for these comparisons.

Noise conditioning. The estimated- and true-state conditions use identical fitted strategies and score transforms, replacing only the state assignment. The single-state variant pools all four noise rates with equal weight to select strategies and fit a common empirical score distribution. It therefore tests noise conditioning in both selection and calibration. State identification is correct in 60/60, 59/60, and 55/60 batch–rate–seed conditions on AVA-Bench, ScienceQA, and MMMU, respectively. These are coarse Low/Medium/High assignments. Pooling states reduces TAG performance on AVA-Bench and ScienceQA but slightly improves it on MMMU.

Table 20: Design selection and noise-state ablations. AUROC is mean ± standard deviation across five folds; AP is the fold mean. All conditions use the four noise rates and three seeds. "Design selected by validation" selects among the 18 candidate designs within each outer training split.

Dataset	Method / condition	AUROC	AP
AVA-Bench	Global	0.9339 ± 0.0006	0.8595
	TAG	0.9464 ± 0.0010	0.8845
	Design selected by validation	0.9464 ± 0.0010	0.8845
	TAG, true noise state	0.9464 ± 0.0010	0.8845
	TAG, single noise state	0.9442 ± 0.0014	0.8779
ScienceQA	Global	0.9749 ± 0.0023	0.9333
	TAG	0.9765 ± 0.0020	0.9388
	Design selected by validation	0.9764 ± 0.0023	0.9388
	TAG, true noise state	0.9766 ± 0.0020	0.9388
	TAG, single noise state	0.9755 ± 0.0022	0.9350
MMMU	Global	0.6771 ± 0.0111	0.4336
	TAG	0.6788 ± 0.0108	0.4354
	Design selected by validation	0.6812 ± 0.0060	0.4376
	TAG, true noise state	0.6788 ± 0.0108	0.4355
	TAG, single noise state	0.6794 ± 0.0103	0.4366

C.2 Diagnostic utility at fixed review budgets

Within each true task, we inspect the highest-risk $\lceil qn \rceil$ questions for $q \in \{0.01, 0.05, 0.10\}$. A hit is a reviewed question whose annotated answer is incorrect. Ties at the cutoff receive fractional expected hits. Hits, reviewed questions, and injected errors are summed across tasks before calculating precision and recall, then averaged over conditions and folds. True task labels allocate the evaluation budget; TAG retains its predicted task assignments. Review budgets and risk rankings are defined separately within each true task to identify injected answer errors. Table 21 reports Global and TAG at identical budgets.

Table 21: Precision and recall at fixed within-task review budgets. Budgets are fractions of questions in each true task; precision and recall are percentages averaged across four noise rates, three seeds, and five folds. G and T denote Global and TAG.

Dataset	Budget	Prec. G	Prec. T	Recall G	Recall T
AVA-Bench	1%	97.22	97.61	5.13	5.17
	5%	93.39	94.77	23.84	24.40
	10%	88.57	90.39	43.74	45.18
ScienceQA	1%	99.68	99.68	6.54	6.54

Table 21 (continued)

Dataset	Budget	Prec. G	Prec. T	Recall G	Recall T
	5%	96.14	97.08	25.47	25.85
	10%	92.65	93.21	47.57	47.94
MMMU	1%	72.10	72.14	4.01	4.01
	5%	57.34	57.76	13.45	13.58
	10%	50.15	50.47	22.46	22.64

C.3 Comparison with VLM judge ensembles

We average the cached risks of Qwen2.5-VL and InternVL2.5 on AVA-Bench, and Qwen2.5-VL and InternVL3.5 on ScienceQA, with equal weights fixed before this comparison. No additional prompting or weight tuning is used. Both ensembles remain below TAG; the available single-model judge remains stronger on MMMU. Ensemble compositions differ across datasets, and no two-model ensemble is available for MMMU. Table 22 compares these judgments with TAG.

Table 22: VLM judgment and TAG. Ensembles average Qwen2.5-VL with InternVL2.5 on AVA-Bench and with InternVL3.5 on ScienceQA. MMMU has a single-model judge. AUROC is mean $\pm$ fold standard deviation; AP is the fold mean.

Dataset	Method / condition	AUROC	AP
AVA-Bench	TAG	0.9464 ± 0.0010	0.8845
	VLM judge ensemble	0.9141 ± 0.0020	0.8230
ScienceQA	TAG	0.9765 ± 0.0020	0.9388
	VLM judge (InternVL3.5)	0.9475 ± 0.0130	0.8884
	VLM judge ensemble	0.9513 ± 0.0137	0.8968
MMMU	TAG	0.6788 ± 0.0108	0.4354
	VLM judge (Qwen2.5-VL)	0.7093 ± 0.0122	0.4747

D Full-Data Reference Rules

The following tables give the main TAG rules fitted on all five batches: one encoder per task and a Top-3 detector list per state. These full-data reference rules describe the complete sample; held-out scores use the corresponding four-batch fits. Low pools 10/20%, Medium is 30%, and High is 40%. Detector abbreviations follow Figure 2; IV3.5, IV2.5, Q2.5, Q3E, and V2V denote InternVL3.5, InternVL2.5-8B, Qwen2.5-VL-7B-Instruct, Qwen3-VL-Embedding, and VLM2Vec.

After preparing the inputs following the supplementary README, run `run.py --component insight` and `run.py --component tag` with the specified asset and work directories to generate fold-specific rules for both TAG variants and the shared strategy. In the dataset-encoder sensitivity analysis, Top-3 overlap is 99.33%, 94.75%, and 88.52% on AVA-Bench, ScienceQA, and MMMU, respectively (Table 11).

Table 23: AVA-Bench full-data TAG reference rules.

Task	Encoder	Low Top-3	Medium Top-3	High Top-3
Absolute Depth	IV3.5	Simi, KNN, HOC	Simi, KNN, GMM	GMM, DM, AUM
Action	IV2.5	AUM, DM, EL2N	EL2N, AUM, DM	SL, EL2N, DM
Emotion	IV3.5	AUM, DM, EL2N	AUM, DM, EL2N	DM, AUM, EL2N
Fine-grained	Q2.5	DM, AUM, EL2N	DM, EL2N, AUM	SL, EL2N, DM
Orientation	IV2.5	KNN, Simi, DM	DM, AUM, EL2N	DM, AUM, EL2N
Recognition	IV3.5	DM, AUM, EL2N	DM, AUM, EL2N	SL, EL2N, DM
Relative Depth	IV3.5	GMM, AUM, DM	GMM, DM, AUM	DM, AUM, EL2N
Scene Classification	Q2.5	AUM, DM, EL2N	AUM, DM, EL2N	SL, EL2N, DM
Spatial	IV3.5	KNN, Simi, AUM	AUM, DM, EL2N	DM, AUM, EL2N
Texture	Q2.5	KNN, CL, OOF	DM, AUM, GMM	EL2N, DM, AUM

Table 24: ScienceQA full-data TAG reference rules.

Task	Encoder	Low Top-3	Medium Top-3	High Top-3
Attribute Recognition	IV2.5	EL2N, DM, AUM	EL2N, DM, AUM	SL, EL2N, DM
Categorization	IV2.5	SL, EL2N, AUM	SL, EL2N, AUM	SL, EL2N, DM
Causality	IV2.5	AUM, DM, EL2N	EL2N, SL, DM	SL, EL2N, DM
Comparison	IV2.5	CL, OOF, AUM	DM, AUM, EL2N	SL, EL2N, DM
Data Interpretation	IV2.5	SL, EL2N, AUM	SL, EL2N, AUM	SL, EL2N, AUM
Diagram	IV2.5	SL, EL2N, AUM	SL, EL2N, DM	SL, EL2N, DM
Experiment	V2V	DM, AUM, EL2N	EL2N, DM, AUM	EL2N, SL, DM
Language	IV2.5	EL2N, SL, DM	EL2N, SL, DM	SL, EL2N, DM
Localization	IV2.5	EL2N, AUM, DM	EL2N, SL, AUM	SL, EL2N, DM
Others	IV2.5	AUM, DM, EL2N	AUM, DM, EL2N	EL2N, SL, AUM
Quantitative	IV2.5	Simi, CL, OOF	DM, AUM, EL2N	EL2N, SL, DM

Table 25: MMMU full-data TAG reference rules.

Task	Encoder	Low Top-3	Medium Top-3	High Top-3
Art & Design	Q3E	DM, AUM, EL2N	EL2N, DM, AUM	EL2N, SL, AUM
Business	IV2.5	DM, AUM, EL2N	AUM, DM, EL2N	CL, OOF, AUM
Health & Medicine	Q3E	SL, EL2N, DM	EL2N, SL, DM	EL2N, SL, AUM
Humanities & Social Science	Q3E	AUM, DM, EL2N	EL2N, AUM, DM	EL2N, SL, AUM
Science	Q3E	EL2N, DM, AUM	EL2N, DM, AUM	EL2N, SL, DM
Tech & Engineering	Q3E	EL2N, AUM, DM	SL, EL2N, DM	SL, EL2N, AUM

Table 26: Full-data Global reference choices.

Dataset	State	Encoder	Detector
AVA-Bench	Low	IV3.5	AUM
AVA-Bench	Medium	IV3.5	AUM
AVA-Bench	High	IV3.5	SL
ScienceQA	Low	IV2.5	DM
ScienceQA	Medium	IV2.5	EL2N
ScienceQA	High	IV2.5	SL
MMMU	Low	Q3E	AUM
MMMU	Medium	Q3E	EL2N
MMMU	High	Q3E	EL2N