ANSWER-LEVEL TRUST SELECTION FOR PHYSICAL VISION-LANGUAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can estimate physical quantities such as duration, speed, and acceleration from visual observations, but existing benchmarks primarily assess overall model performance against annotated ground truth. In deployment, a key question is whether an individual prediction can be trusted when its ground truth is unavailable. Self-consistency alone may fail to capture important failure modes: a VLM may produce stable-but-wrong estimates or rely on textual priors rather than visual evidence. We formulate answer-level selective prediction for quantitative physical reasoning and propose Answer-Level Trust Selection (ATS), a post-hoc, model-agnostic framework for accepting or rejecting individual VLM predictions. ATS requires no fine-tuning, auxiliary verifier, or access to the model's internal logits. Instead, it aggregates eight interpretable behavioral diagnostic scores derived from repeated queries and controlled interventions into a unified trust score. We evaluate ATS in depth on Qwen2.5-VL-7B and across 20 VLM backbones, examining selective performance, diagnostic behavior, and targeted failure modes. Our results show that intervention-based diagnostics help identify stable-but-wrong and prior-tracking predictions that repeated agreement alone may miss. However, improved failure-case rejection can come at the cost of lower retention of correct predictions. ATS therefore complements model-level capability evaluation with answer-level reliability assessment for quantitative VLM predictions. Code will be released upon publication.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly used as perception and reasoning modules in systems that interact with the physical world Chow et al. (2025); Puyin et al. (2026). Beyond object recognition and scene description, these models are now expected to reason about measurable physical quantities such as size, duration, speed, and acceleration Chow et al. (2025); Puyin et al. (2026). Their numerical predictions may serve as interfaces between visual perception and downstream planning, monitoring, measurement, and decision-support systems. This shift makes quantitative physical reasoning an important test of whether VLM outputs are not only semantically plausible but also sufficiently reliable for downstream use. Recent benchmarks such as QuantiPhy Puyin et al. (2026) evaluate whether VLMs can infer physical quantities of moving objects from visual observations, extending VLM evaluation from scene description to numerically grounded visual reasoning.

However, existing benchmarks for quantitative VLM reasoning primarily evaluate model-level performance by comparing predictions with annotated ground truth and averaging accuracy over a dataset. Although such evaluation is essential for measuring overall capability, deployment raises a more actionable question: once a VLM produces a particular numerical estimate, is that specific answer sufficiently trustworthy to retain? Numerical VLMs may exhibit hidden answer-level failures. A model can be stable but wrong, repeatedly producing similar estimates that remain far from the ground truth. It may also track textual numerical priors: with the visual input held fixed, its prediction may shift toward an altered prior. We examine these behaviors in Table 1. Figure 1 summarizes the diagnostic-score profiles that motivate this analysis. Average benchmark performance does not identify which individual answers exhibit these behaviors, and repeated-output agreement alone may not expose them.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

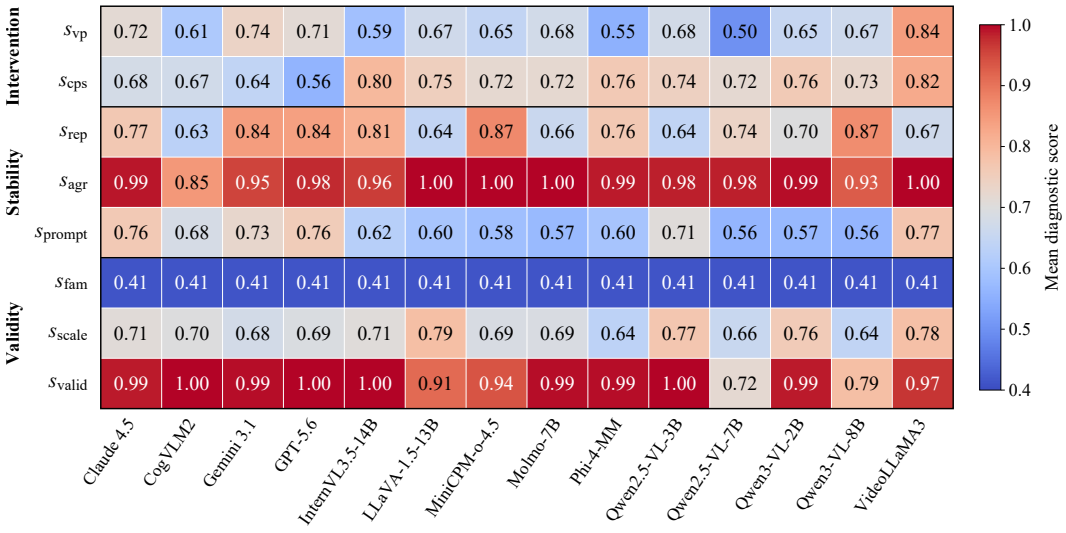

Figure 1: Mean diagnostic-score profiles for 14 of the 20 evaluated VLMs. The complete 20-backbone results are reported in Table 9.

These limitations motivate a complementary perspective on evaluating the reliability of numerical VLM predictions. Building on the principle of selective prediction, an answer-level selection mechanism should retain a prediction only when its estimated reliability is sufficient and otherwise abstain from using it. Unlike conventional model-level evaluation, which characterizes overall capability using aggregate benchmark metrics, this answer-level perspective focuses on whether a particular prediction should be accepted or rejected when its ground truth is unavailable at deployment time. The two perspectives are complementary: model-level evaluation measures how well a VLM performs on average, whereas answer-level selection supports operational decisions about individual outputs. This distinction directly addresses the practical need to separate numerical answer generation from post-hoc accept-or-reject decisions.

We propose Answer-Level Trust Selection (ATS), a post-hoc, black-box selective trust framework for quantitative physical reasoning with VLMs. ATS offers two practical advantages. First, it is model-agnostic and minimally invasive: it requires no fine-tuning of the VLM, no additional neural verifier, no access to internal logits, and no modification or replacement of the original answer. Second, ATS provides an interpretable basis for answer-level selection. It evaluates each prediction using eight behavior-based diagnostic scores, each capturing a distinct aspect of the model's predictive behavior, and aggregates them into a unified trust score for accepting or rejecting the original answer. We conduct an in-depth analysis of ATS using Qwen2.5-VL-7B as the primary model, examining its selective performance and the contribution of each diagnostic component. To further characterize the current state of quantitative physical reasoning, we additionally benchmark 20 VLMs, providing a broad comparison of their numerical reasoning capabilities.

Our contributions are threefold: we formulate answer-level selective prediction for quantitative physical reasoning with VLMs; we propose Answer-Level Trust Selection (ATS), a model-agnostic, post-hoc black-box framework that aggregates eight interpretable behavioral diagnostics to accept or reject predictions without modifying the underlying VLM or its answers; and we evaluate ATS in depth on Qwen2.5-VL-7B, alongside a benchmark of 20 representative VLMs for quantitative physical reasoning.

2 Related Work

2.1 Quantitative Physical Reasoning in VLMs

Visual physical reasoning benchmarks evaluate models' understanding of object interactions, dynamics, causal relations, and latent physical properties. CLEVRER Yi et al. (2020) and Physion Bear et al. (2021) study collision reasoning and physical scene evolution, while ComPhy Chen et al. (2022) and CRIPP-VQA Patel et al. (2022) require inference of properties such as mass, charge, friction,

and initial velocity from observed motion and interactions. More recent benchmarks extend this evaluation to general-purpose VLMs. PhysBench Chow et al. (2025) covers physical properties, relations, and dynamics, whereas QuantiPhy Puyin et al. (2026) focuses on video-based numerical estimation with explicit textual physical priors. This setting aligns directly with ATS, which studies answer-level selective trust for numerical predictions and uses controlled prior interventions to probe prediction reliability. Unlike benchmark-level accuracy evaluation, our focus is whether an estimate produced by a fixed VLM should be retained or rejected without revising its value.

2.2 Behavioral Reliability Estimation

Black-box reliability methods estimate uncertainty from model behavior rather than internal confidence. Repeated-sampling approaches use answer agreement Wang et al. (2023), response inconsistency Manakul et al. (2023), or uncertainty over semantic response clusters Kuhn et al. (2023). Although broadly applicable, these signals may assign high confidence to consistently repeated but numerically incorrect answers. Multimodal methods additionally measure reliability under visual or textual perturbations Zhang et al. (2024), compare reference-free hallucination signals Li et al. (2024), or incorporate cross-modal interventions into uncertainty calibration Padhi et al. (2025). ATS builds on this behavioral perspective but introduces diagnostics tailored to quantitative physical reasoning, including a counterfactual probe of numerical-prior sensitivity. Selective prediction converts reliability estimates into accept-or-abstain decisions characterized by the risk–coverage trade-off El-Yaniv & Wiener (2010). Prior work has studied post-hoc and jointly learned rejection in classification Geifman & El-Yaniv (2017; 2019), as well as learned or calibrated abstention in VQA Whitehead et al. (2022); Dancette et al. (2023); Eisenschlos et al. (2024). In contrast, ATS performs post-hoc selective prediction for numerical VLM outputs using only black-box behavioral diagnostics.

3 Problem Formulation

We study answer-level selective prediction for quantitative physical reasoning with vision-language models (VLMs). Each example is represented as $x_i = (v_i, q_i, p_i)$, where v_i denotes the visual observation, q_i the question, and p_i the textual context containing a numerical prior. Given x_i, a VLM f produces an original numerical prediction $\hat{y}_i = f(v_i, q_i, p_i)$, whose ground-truth physical quantity is denoted by y_i.

Our objective is to perform post-hoc selection without modifying the underlying VLM or correcting its original prediction. A selection method assigns each prediction a trust score $s_i = S(x_i, \hat{y}_i; f)$, where S may obtain additional behavioral evidence through black-box queries to f but does not use y_i at inference time. Given a selection threshold τ, the selector is

$$g_\tau(x_i, \hat{y}_i) = \mathbb{I}[s_i \geq \tau], \tag{1}$$

where $g_\tau = 1$ indicates acceptance and $g_\tau = 0$ rejection. Varying τ trades off the fraction of retained answers against their predictive quality. For ATS, $s_i = T_{\mathrm{ATS},i}$; we reserve τ for a selection threshold. The practical event-conditional ATS rule additionally applies diagnostic conditions, as specified in Section 4.5.

We measure numerical prediction quality using Mean Relative Accuracy (MRA) Puyin et al. (2026). Let $\mathcal{C} = \{0.1, 0.2, \ldots, 0.9, 0.95\}$ and let $\epsilon_y > 0$ be a small constant for denominator stabilization. The per-example MRA is

$$\mathrm{MRA}(\hat{y}_i, y_i) = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} \mathbb{I}\left[\frac{|\hat{y}_i - y_i|}{\max\{|y_i|, \epsilon_y\}} < 1 - \theta \right]. \tag{2}$$

This metric averages prediction correctness across multiple relative-error tolerances, assigning higher values to estimates closer to the ground truth.

For a test set of N examples, coverage is the fraction of accepted predictions, and retained-answer MRA is their average MRA:

$$\mathrm{Cov}(\tau) = \frac{1}{N} \sum_{i=1}^{N} g_\tau(x_i, \hat{y}_i), \qquad \mathrm{MRA}_{\mathrm{ret}}(\tau) = \frac{\sum_{i=1}^{N} \mathrm{MRA}(\hat{y}_i, y_i)\, g_\tau(x_i, \hat{y}_i)}{N\, \mathrm{Cov}(\tau)}. \tag{3}$$

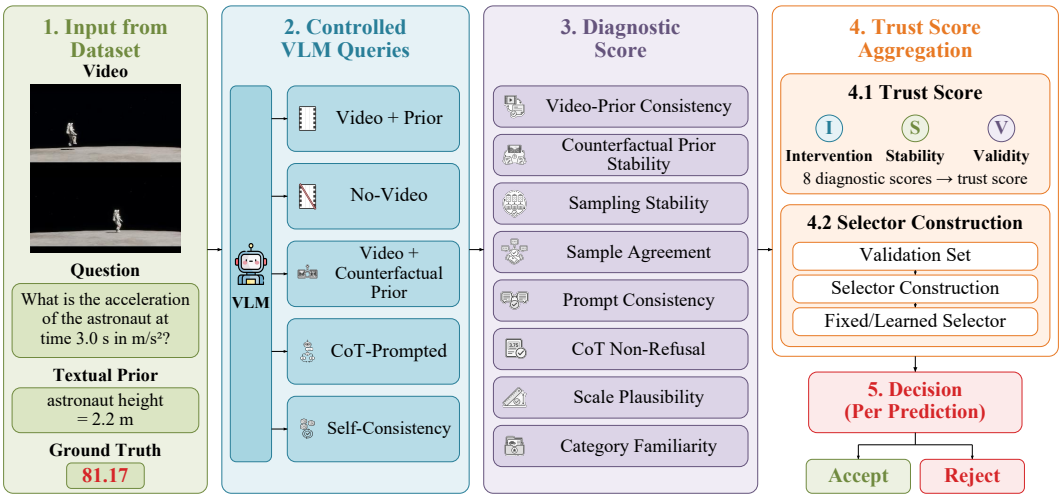

Figure 2: Overview of ATS. Controlled VLM queries produce eight diagnostic scores for answer-level selection. Ground truth is used for validation and evaluation only; it is not an input to the VLM or the test-time selector.

The retained-answer MRA is defined for $\mathrm{Cov}(\tau) > 0$, and the corresponding selective risk is $R(\tau) = 1 - \mathrm{MRA}_{\mathrm{ret}}(\tau)$. A reliable selector should achieve lower selective risk at a given coverage, or equivalently higher coverage at a given risk level.

4 Methods

4.1 Overview

As illustrated in Figure 2, ATS is a post-hoc framework for answer-level selective trust in numerical physical reasoning. For each example, the VLM is evaluated under five controlled conditions: Video + Prior, Prior Only, Video + Counterfactual Prior, CoT-Prompted, and Self-Consistency Sampling. The resulting responses are converted into eight behavioral diagnostic scores and aggregated into an overall trust score. The validation set is then used to construct an answer-level selector that determines whether the original numerical prediction should be accepted for downstream use.

4.2 Controlled VLM Probing

For each example $x = (v, q, p)$, ATS obtains the original Video + Prior prediction, a prior-only prediction without visual input, and a counterfactual-prior prediction:

$$\hat{y}_{\mathrm{full}} = f(v, q, p), \qquad \hat{y}_{\mathrm{prior}} = f(\varnothing, q, p), \qquad \hat{y}_{\mathrm{cf}} = f(v, q, p^{\mathrm{cf}}). \tag{4}$$

Here, p^{cf} changes only the numerical value in the textual prior while preserving the video, question, and target physical quantity. This intervention is a diagnostic perturbation rather than a new task: $\hat{y}_{\mathrm{full}}$ remains the answer under assessment, while $\hat{y}_{\mathrm{cf}}$ reveals sensitivity to the altered textual cue.

To assess prompt sensitivity, ATS obtains $\hat{y}_{\mathrm{cot}} = f(v, q, p; \pi_{\mathrm{cot}})$, where π_{cot} denotes the chain-of-thought prompting instruction. To assess sampling stability, it draws M stochastic generations and summarizes them by their median and dispersion:

$$\hat{y}_{\mathrm{sc}}^{(m)} \sim f(v, q, p; \xi_m), \quad m = 1, \ldots, M, \qquad \tilde{y}_{\mathrm{sc}} = \mathrm{median}_{m=1,\ldots,M}\, \hat{y}_{\mathrm{sc}}^{(m)}. \tag{5}$$

Here, ξ_m denotes sampling randomness. The dispersion is used to compute the sampling stability score.

4.3 Diagnostic Score Construction

ATS compares outputs across controlled black-box queries to construct eight complementary diagnostic scores. These scores describe behavioral evidence about individual answers rather than calibrated probabilities of correctness. Throughout, we use the stabilized log magnitude $\ell(z) = \log(\max\{|z|, \epsilon_{\log}\})$, with $\epsilon_{\log} > 0$.

The **video-prior consistency score** is $s_{\mathrm{vp}} = (1 + |\ell(\hat{y}_{\mathrm{full}}) - \ell(\hat{y}_{\mathrm{prior}})|)^{-1}$. A larger value indicates closer agreement in log magnitude between the Video + Prior and Prior-Only predictions.

The **counterfactual prior stability score** measures whether the counterfactual prediction stays closer to the original answer than to an idealized prior-tracking reference. Let $\rho = \mathrm{num}(p^{\mathrm{cf}})/\mathrm{num}(p)$ be the numerical scale ratio between the counterfactual and original priors, and define

$$d_{\mathrm{ref}} = |\ell(\hat{y}_{\mathrm{cf}}) - \ell(\hat{y}_{\mathrm{full}})|, \qquad d_{\mathrm{track}} = |\ell(\hat{y}_{\mathrm{cf}}) - \ell(\rho\hat{y}_{\mathrm{full}})|. \tag{6}$$

The score is $s_{\mathrm{cps}} = d_{\mathrm{track}}/\max\{\zeta, d_{\mathrm{track}} + d_{\mathrm{ref}}\}$, where $\zeta > 0$ is a denominator floor. Larger values indicate greater stability relative to the prior-tracking reference.

For self-consistency sampling, write $\hat{y}^{(m)} = \hat{y}_{\mathrm{sc}}^{(m)}$ and let $\mathcal{V} = \{m \in \{1, \ldots, M\} : \hat{y}^{(m)} \in \mathbb{R}_{>0}\}$ index valid positive predictions. The **sampling stability score** and **sample agreement score** are

$$\sigma_{\log} = \mathrm{Std}\Big(\{\ell(\hat{y}^{(m)})\}_{m\in\mathcal{V}}\Big), \qquad s_{\mathrm{rep}} = \frac{1}{1+\sigma_{\log}}, \qquad s_{\mathrm{agr}} = a_{\mathrm{valid}} = \frac{|\mathcal{V}|}{M}. \tag{7}$$

A larger s_{rep} indicates tighter concentration of valid predictions in log space. The score s_{agr} measures the valid-output rate, not agreement around a dominant numerical value.

The **prompt consistency score** is $s_{\mathrm{prompt}} = (1 + \sigma_{\mathrm{prompt}})^{-1}$, where

$$\sigma_{\mathrm{prompt}} = \mathrm{Std}(\ell(\hat{y}_{\mathrm{prior}}), \ell(\hat{y}_{\mathrm{full}}), \ell(\hat{y}_{\mathrm{cot}}))\,. \tag{8}$$

Larger values indicate greater numerical consistency across the Prior-Only, Video + Prior, and CoT-Prompted conditions.

The **scale-plausibility score** is $s_{\mathrm{scale}} = \exp(-|\ell(\hat{y}_{\mathrm{full}})|/\kappa_{\mathrm{scale}})$, where $\kappa_{\mathrm{scale}} = 5$ in all experiments. A larger value indicates a less extreme log magnitude under this scale heuristic.

The **category familiarity score** measures the frequency of an example's event family in a fixed reference pool. For example i, let h_i denote its event family, c_{h_i} its empirical count, and $N_{\mathrm{ref}} > 0$ the reference-pool size. The novelty risk and corresponding familiarity score are

$$r_{\mathrm{unk},i} = 1 - \frac{\log(1 + c_{h_i})}{\log(1 + N_{\mathrm{ref}})}, \qquad s_{\mathrm{fam},i} = 1 - r_{\mathrm{unk},i}. \tag{9}$$

A larger $s_{\mathrm{fam},i}$ indicates a more frequent event family.

The **CoT non-refusal score** is $s_{\mathrm{valid}} = \mathbb{I}[\mathrm{match}(R_{\mathrm{cot}}, \mathcal{R}_{\mathrm{refusal}}) = 0]$, where R_{cot} is the CoT-Prompted response and $\mathrm{match}(R, \mathcal{R}_{\mathrm{refusal}})$ equals 1 if R matches any predefined refusal pattern and 0 otherwise. This score measures non-refusal; it does not establish numerical parseability or correctness.

4.4 Trust-Score Aggregation

ATS organizes the eight scores into intervention-related, stability-related, and validity-related groups:

$$\begin{aligned} &\mathbf{s}_{\mathrm{int},i} = [s_{\mathrm{vp},i}, s_{\mathrm{cps},i}]^\top, && \mathbf{s}_{\mathrm{stab},i} = [s_{\mathrm{rep},i}, s_{\mathrm{agr},i}, s_{\mathrm{prompt},i}]^\top, \\ &\mathbf{s}_{\mathrm{val},i} = [s_{\mathrm{fam},i}, s_{\mathrm{scale},i}, s_{\mathrm{valid},i}]^\top. \end{aligned} \tag{10}$$

These groups capture responses to controlled input interventions, stability and valid-output behavior across repeated generations and prompting conditions, and evidence concerning event familiarity, numerical scale, and non-refusal, respectively. After clipping each score to $[0, 1]$, the fixed-weight ATS variants compute

$$T_{\mathrm{ATS},i} = \mathbf{w}_{\mathrm{int}}^\top \mathbf{s}_{\mathrm{int},i} + \mathbf{w}_{\mathrm{stab}}^\top \mathbf{s}_{\mathrm{stab},i} + \mathbf{w}_{\mathrm{val}}^\top \mathbf{s}_{\mathrm{val},i}. \tag{11}$$

Here, $T_{\mathrm{ATS},i} = T_{\mathrm{ATS}}(\mathbf{s}_i)$ is the aggregate score for example i, and $\mathbf{s}_i$ contains all eight diagnostics. The score is distinct from an acceptance threshold.

We evaluate three aggregation strategies. *Prior-weighted ATS*, our default, uses heuristic weights fixed before test evaluation. *Equal-weight ATS* assigns the same weight to all eight scores. *Learned ATS* learns their combination using lightweight logistic regression on the validation set. We also perturb the prior-weighted weights and recompute retained-answer MRA at fixed coverage levels and AURC to assess sensitivity. Complete analyses are provided in the *supplementary material*.

4.5 Selector Construction

ATS uses the validation set to construct event-conditional accept/reject rules. Let $\mathcal{H}$ denote the event families and $h_i = H(x_i) \in \mathcal{H}$ the family assigned to example i according to its physical inference type and video type. For each family h and candidate trust threshold $\lambda \in \Lambda_h$, the complete candidate parameters are $\gamma_{h,\lambda} = (\alpha_{h,\lambda}, \beta_{h,\lambda}, \eta_{h,\lambda}, \lambda)$. Here, λ is the event-family counterpart of the generic threshold τ and applies to $T_{\mathrm{ATS},i}$; the other parameters specify three diagnostic conditions.

Validation selects $\widehat{\lambda}_h \in \Lambda_h$ and the corresponding rule $\widehat{\gamma}_h = (\widehat{\alpha}_h, \widehat{\beta}_h, \widehat{\eta}_h, \widehat{\lambda}_h)$. The deployed selector is

$$\begin{aligned} g_{\widehat{\gamma}_{h_i}}(x_i, \hat{y}_i) = \mathbb{I}\big[&s_{\mathrm{vp},i} \geq \widehat{\alpha}_{h_i} \ \wedge \ s_{\mathrm{cps},i} \geq \widehat{\beta}_{h_i} \\ &\wedge \ r_{\mathrm{unk},i} \leq \widehat{\eta}_{h_i} \ \wedge \ T_{\mathrm{ATS},i} \geq \widehat{\lambda}_{h_i}\big]. \end{aligned} \tag{12}$$

Candidate rules are evaluated using validation coverage and empirical selective risk, defined as one minus the mean per-example MRA among accepted examples. Let $K = \sum_{h \in \mathcal{H}} |\Lambda_h|$ be the total number of candidates. We additionally compute the Hoeffding-style upper-confidence screening statistic

$$\overline{R}_h(\gamma_{h,\lambda}) = \widehat{R}_h(\gamma_{h,\lambda}) + \sqrt{\frac{\log(K/\delta)}{2n_{h,\lambda}}}. \tag{13}$$

Here, $n_{h,\lambda} > 0$ is the number of accepted validation examples, and $\delta \in (0,1)$ is the nominal family-wise failure-probability parameter. This statistic penalizes candidates supported by few accepted examples. The selected rules are fixed before test evaluation, and their test coverage and selective risk are reported as empirical performance measures. Theoretical details, proofs, and additional analyses are provided in the *supplementary material*.

4.6 Test-Time Selection

During validation, ATS uses the validation set to construct one selector for each event family. Each selector contains family-specific acceptance thresholds and is fixed before test evaluation. For each test example, ATS applies the selector corresponding to its event family to the original numerical prediction. The prediction is accepted only if all prespecified selection conditions are satisfied. Otherwise, it is rejected. No test labels are used, and ATS neither revises nor replaces the original prediction.

4.7 Theoretical Motivations

ATS is primarily an empirical answer-selection framework. We provide two analytical motivations: a fixed-rule concentration result for the validation sample-size penalty and a monotonicity property of the counterfactual prior stability score. Because practical ATS rules are constructed and evaluated on the same validation split, these results do not constitute finite-sample guarantees for the experimental procedure.

Let $L(\hat{y}, y) = 1 - \mathrm{MRA}(\hat{y}, y) \in [0,1]$ be the numerical loss. For each event family h, let $\Gamma_h = \{\gamma_{h,1}, \ldots, \gamma_{h,K_h}\}$ be a finite collection of complete accept/reject rules. Each rule specifies all diagnostic definitions, aggregation weights, thresholds, event-family structure, and acceptance conditions. For $\gamma \in \Gamma_h$, define

$$R_h(\gamma) = \mathbb{E}\Big[L(\hat{Y}, Y) \mid g_\gamma(X, \hat{Y}) = 1, \ H(X) = h\Big]. \tag{14}$$

Let $\widehat{R}_h(\gamma)$ be the empirical risk over $n_{h,\gamma} > 0$ accepted risk-calibration examples, and let $K = \sum_{h\in\mathcal{H}} K_h$.

Proposition 1 (Conditional finite-class selective-risk bound)**.** *Assume calibration units are conditionally i.i.d. within each event family and the finite collection of complete candidate rules $\Gamma = \bigcup_{h\in\mathcal{H}} \Gamma_h$ is fixed independently of the risk-calibration sample. Then, with probability at least $1-\delta$, simultaneously for every candidate with $n_{h,\gamma} > 0$,*

$$R_h(\gamma) \le \widehat{R}_h(\gamma) + \sqrt{\frac{\log(K/\delta)}{2n_{h,\gamma}}}. \tag{15}$$

Any data-dependent selection from this prespecified family therefore inherits its corresponding simultaneous upper bound.

When each $\lambda \in \Lambda_h$ indexes a prespecified complete rule, $K = \sum_{h\in\mathcal{H}} |\Lambda_h|$, matching the correction in the practical screening statistic. The proposition bounds average loss among accepted examples, not answer-specific correctness or a prediction interval. In the practical two-way procedure, some thresholds are derived from the same validation data used to estimate risk, violating the independence condition. We therefore use the Hoeffding-style quantity Hoeffding (1963) as a screening statistic and interpret test results as empirical evidence rather than a formal finite-sample guarantee. Further analysis appears in the *supplementary material*.

Proposition 2 (Monotonicity of counterfactual prior stability)**.** *Assume $\rho > 0$, $\rho \neq 1$, and all relevant magnitudes exceed $\epsilon_{\log}$. Let $\hat{y}_{\mathrm{cf}} = \rho^{\alpha}\hat{y}_{\mathrm{full}}$, where $\alpha \in [0,1]$ interpolates between reference stability ($\alpha = 0$) and full prior tracking ($\alpha = 1$). Writing $A = \log\rho \neq 0$, we have*

$$s_{\mathrm{cps}}(\alpha) = \frac{(1-\alpha)|A|}{\max\{\zeta, |A|\}}. \tag{16}$$

Thus, $s_{\mathrm{cps}}(\alpha)$ decreases strictly with prior tracking, from $s_{\mathrm{cps}}(0) = |A|/\max\{\zeta,|A|\}$ to $s_{\mathrm{cps}}(1) = 0$.

5 Experiments

5.1 Dataset and task

We evaluate ATS on the official validation split of QuantiPhy Puyin et al. (2026). Following the original benchmark, we measure numerical answer quality using Mean Relative Accuracy (MRA) Puyin et al. (2026). Beyond evaluating numerical estimation itself, we consider an answer-level selective-prediction setting in which ATS determines whether each VLM prediction should be accepted or rejected. To the best of our knowledge, no other publicly available video benchmark directly combines real-world physical-quantity estimation with the annotations required for this evaluation. We therefore focus our experiments on QuantiPhy.

5.2 Baselines

We compare Prior-weighted ATS with SC-only, three Generic UQ baselines, Equal-weight ATS, Learned ATS, and Oracle ranking. SC-only ranks predictions by the log-space stability of repeated numerical outputs, motivated by self-consistency (Wang et al., 2023). Generic UQ uses only s_{rep}, s_{agr}, and s_{prompt}: the mean and rank-fusion variants aggregate these scores directly, whereas Learned Generic UQ fits a standardized logistic-regression aggregator on the validation set using a binary pseudo-correctness target indicating whether the per-example MRA is at least 0.4. Equal-weight ATS assigns uniform weights to the full diagnostic set. Learned ATS fits the same aggregator using all diagnostic scores. Oracle ranking orders predictions by ground-truth answer quality and serves as an unattainable upper bound.

5.3 Experimental Results

We report selective-ranking performance and targeted failure-mode analyses. Appendix B provides the evaluation protocol, cross-backbone comparisons, diagnostic studies, ablations, weight-sensitivity analyses, paired effect estimates, and query-budget details.

5.4 Cross-backbone Diagnostic Profiles

Table 1: Failure-case rejection and correct-case retention at 50% coverage, averaged across 20 backbones. Reject and Retain refer to failure cases and correct cases, respectively.

	Stable-but-wrong		Prior tracking	
Selector	Reject ↑	Retain ↑	Reject ↑	Retain ↑
SC-only	0.013	**0.986**	0.654	0.491
Learned Generic UQ	0.271	0.830	0.638	**0.523**
Equal-weight ATS	**0.424**	0.723	0.906	0.248
Prior-weighted ATS	<u>0.409</u>	0.727	**0.958**	0.111
Learned ATS	0.379	0.777	0.574	0.512

Repeated-output stability and validity-related scores are generally high, whereas intervention- and prompt-related scores, particularly s_{vp}, s_{cps}, and s_{prompt}, show greater cross-backbone variation. We use these profiles as descriptive behavioral evidence to motivate the targeted failure-mode analysis, rather than as a direct performance comparison. More detailed analyses are provided in the *supplementary material*.

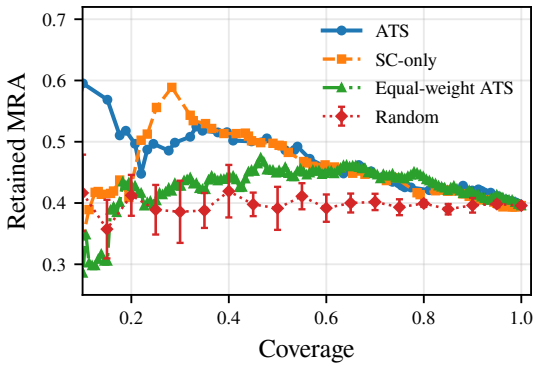

Figure 3: Coverage-retained-MRA curves on Qwen2.5-VL-7B.

5.5 Cross-backbone Selective Comparison

Across the 20 backbones, Prior-weighted ATS has mean paired effects of $+0.0192$, $+0.0023$, and -0.0037 in MRA at 30%, 50%, and 70% coverage, respectively, relative to SC-only. For AURC, where a positive difference favours ATS, the mean effect is $+0.0133$, with positive effects on 14 of 20 backbones. We therefore treat the aggregate comparison as a descriptive cross-backbone consistency check rather than evidence of universal superiority. AURC shows the most consistent positive direction, whereas fixed-coverage effects remain small and backbone dependent. More analyses of the cross-backbone experiments are provided in the *supplementary material.*

5.6 Targeted Failure-mode Analysis

Overall selective-ranking metrics do not reveal whether a selector identifies predictions that remain consistent but numerically wrong. Table 1 evaluates rejection of stable-but-wrong and prior-tracking cases at 50% coverage, together with retention of the corresponding correct-case groups. Equal-weight and Prior-weighted ATS reject 42.4% and 40.9% of stable-but-wrong cases, compared with 27.1% for Learned Generic UQ and 1.3% for SC-only. Prior-weighted ATS also rejects 95.8% of wrong prior-tracking cases. In the table, bold indicates the best non-oracle result in each column, while underlining marks the second-highest stable-but-wrong rejection rate. These rejection gains come with lower correct-case retention and must therefore be interpreted as rejection-retention trade-offs rather than unqualified improvements. Nevertheless, they show that controlled interventions reveal answer-level behavioral evidence that self-consistency alone may not capture.

5.7 Representative Qwen2.5-VL-7B Results

Figure 3 and Table 2 summarize selective-ranking performance on Qwen2.5-VL-7B. The curves show that selector differences are more pronounced at lower coverage and narrow as coverage approaches one, where all methods retain the same predictions. No method dominates across metrics: SC-only achieves the highest retained MRA at 30% and 50% coverage, while Generic UQ rank fusion obtains the lowest AURC (0.5125). Prior-weighted ATS has a slightly lower AURC than SC-only (0.5300

Table 2: Selective-ranking performance on Qwen2.5-VL-7B. Retained MRA is reported at three coverage levels. Bold indicates the best value in each column.

	Retained MRA ↑			AURC ↓
Method	30%	50%	70%	
Selection baselines				
SC-only	**0.5429**	**0.4989**	0.4426	0.5338
Generic UQ mean	0.4677	0.4526	0.4533	0.5151
Generic UQ rank fusion	0.4909	0.4709	0.4504	**0.5125**
Learned Generic UQ	0.5063	0.4605	0.4387	0.5593
ATS variants				
Equal-weight ATS	0.4374	0.4441	0.4444	0.5655
Prior-weighted ATS	0.5187	0.4889	0.4476	0.5300
Learned ATS	0.4069	0.4120	0.4028	0.5977
Prior-weighted ATS: ablations				
w/o s_{vp}	0.4879	0.4734	0.4352	0.5406
w/o s_{cps}	0.4780	0.4886	**0.4671**	0.5308
Prior-weighted ATS: weight sensitivity				
Random perturbations: mean	0.5192	0.4881	0.4492	0.5306
Random perturbations: 10th percentile	0.5109	0.4804	0.4437	0.5363

versus 0.5338), despite lower MRA at these two operating points. This distinction highlights the importance of evaluating both fixed-coverage performance and the overall risk–coverage trade-off.

Within the ATS variants, prior-weighted aggregation outperforms equal weighting at all three reported coverage levels and reduces AURC from 0.5655 to 0.5300. Learned ATS performs worse than both fixed-weight variants, indicating that validation-based weight learning does not improve ranking on this backbone under the evaluated setup. Random weight perturbations yield mean MRA@50 of 0.4881 and mean AURC of 0.5306, close to the default values of 0.4889 and 0.5300. These averages suggest limited sensitivity to the tested perturbations, although they do not establish robustness to every weight configuration.

The ablations reveal different contributions from the two intervention-related diagnostics. Removing s_{vp} reduces MRA at all three coverage levels and increases AURC from 0.5300 to 0.5406, supporting its usefulness in this configuration. Removing s_{cps}, however, increases MRA@70 from 0.4476 to 0.4671 while reducing MRA@30 and slightly worsening AURC to 0.5308. Its contribution therefore depends on the operating point. Overall, these results support the utility of the chosen diagnostic combination while showing that its benefits vary with coverage and the evaluation criterion.

6 Conclusion

We presented ATS, a post-hoc framework for answer-level selective prediction in quantitative physical reasoning with VLMs. ATS integrates complementary behavioral evidence from controlled interventions, repeated-generation stability, and response validity, without requiring access to model internals or additional fine-tuning. Across 20 VLM backbones, SC-only and Generic UQ remain strong baselines on aggregate selective-ranking metrics. Targeted analyses show that ATS rejects substantially more stable-but-wrong and prior-tracking cases, two failure modes that self-consistency sampling or generic uncertainty estimation alone may not fully capture. These gains come with reduced retention of the corresponding correct-case groups and therefore represent a rejection–retention trade-off rather than a uniform improvement. Overall, our results position intervention-based diagnostics as a complementary source of answer-level uncertainty evidence and motivate future selectors that better leverage such diagnostic scores to improve aggregate selective performance.

AI USE DISCLOSURE

Generative AI tools assisted with correcting typographical errors and improving the clarity of the writing. All AI-assisted content was reviewed by the authors, who take responsibility for the final content, research methods, and conclusions of this paper.

REFERENCES

Daniel M Bear, Elias Wang, Damian Mrowca, Felix J Binder, Hsiao-Yu Fish Tung, RT Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, et al. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*, 2021.

Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos. *arXiv preprint arXiv:2205.01089*, 2022.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. Physbench: Benchmarking and enhancing vision-language models for physical world understanding. In *International Conference on Learning Representations*, volume 2025, pp. 97959–98108, 2025.

Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, and Marcus Rohrbach. Improving selective visual question answering by learning from your peers. *arXiv preprint arXiv:2306.08751*, 2023.

Julian Martin Eisenschlos, Hernán Maina, Guido Ivetta, and Luciana Benotti. Selectively answering visual questions. *arXiv preprint arXiv:2406.00980*, 2024.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, and Fakhri Karray. Reference-free hallucination detection for large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4542–4551, 2024.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Trilok Padhi, Ramneet Kaur, Adam D. Cobb, Manoj Acharya, Anirban Roy, Colin Samplawski, Brian Matejek, Alexander M. Berenbeim, Nathaniel D. Bastian, and Susmit Jha. Calibrating uncertainty quantification of multi-modal llms using grounding. *arXiv preprint arXiv:2505.03788*, 2025.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. Cripp-vqa: Counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9856–9870, 2022.

Li Puyin, Tiange Xiang, Ella Mao, Shirley Wei, Xinye Chen, Adnan Masood, Li Fei-Fei, and Ehsan Adeli. Quantiphy: A quantitative benchmark evaluating physical reasoning abilities of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33174–33184, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. *arXiv preprint arXiv:2204.13631*, 2022.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

A Additional Formulation

This section expands the formulation of ATS using the same organization and terminology as the main paper. We first formalize trust-score aggregation, then describe validation-based selector construction and test-time selection, and finally state the assumptions and proofs underlying the theoretical motivations.

A.1 Trust-Score Aggregation

Let

$$\mathcal{D}_{\mathrm{val}} = \{(x_i, \hat{y}_i, y_i, h_i)\}_{i=1}^{n_{\mathrm{val}}}$$

denote the validation split used by the practical procedure, where n_{val} is the number of validation examples. Each input

$$x_i = (v_i, q_i, p_i)$$

contains the video v_i, quantitative question q_i, and textual prior p_i. The fixed numerical prediction produced by the VLM is denoted by $\hat{y}_i$, and y_i denotes the corresponding ground-truth physical quantity. The event-family label is defined as

$$h_i = H(x_i) \in \mathcal{H},$$

where H is the event-family mapping and $\mathcal{H}$ is the finite set of event families.

Let

$$\mathbf{s}_i = \mathbf{s}(x_i, \hat{y}_i)$$

denote the diagnostic-score vector associated with validation example i. Following the main paper, ATS organizes the eight diagnostic scores into intervention-related, stability-related, and validity-related groups:

$$\mathbf{s}_{\mathrm{int},i} = \begin{bmatrix} s_{\mathrm{vp},i} & s_{\mathrm{cps},i} \end{bmatrix}^\top,$$

$$\mathbf{s}_{\mathrm{stab},i} = \begin{bmatrix} s_{\mathrm{rep},i} & s_{\mathrm{agr},i} & s_{\mathrm{prompt},i} \end{bmatrix}^\top,$$

and

$$\mathbf{s}_{\mathrm{val},i} = \begin{bmatrix} s_{\mathrm{fam},i} & s_{\mathrm{scale},i} & s_{\mathrm{valid},i} \end{bmatrix}^\top.$$

Here, $s_{\mathrm{vp},i}$ and $s_{\mathrm{cps},i}$ are the video–prior consistency and counterfactual prior stability scores, respectively. The unknown-risk score used by the selector is

$$r_{\mathrm{unk},i} = 1 - s_{\mathrm{fam},i}.$$

After clipping each diagnostic score to $[0, 1]$, the fixed-weight aggregation variants produce the aggregate trust score

$$T_{\mathrm{ATS},i} = T_{\mathrm{ATS}}(\mathbf{s}_i) = \mathbf{w}_{\mathrm{int}}^\top \mathbf{s}_{\mathrm{int},i} + \mathbf{w}_{\mathrm{stab}}^\top \mathbf{s}_{\mathrm{stab},i} + \mathbf{w}_{\mathrm{val}}^\top \mathbf{s}_{\mathrm{val},i}.$$

The aggregation weights are fixed before test evaluation for the prior-weighted and equal-weight variants. The learned variant fits the aggregation function on the validation split and then freezes the fitted parameters before test evaluation, as described in the main paper.

The problem formulation in the main paper uses $s_i = S(x_i, \hat{y}_i; f)$ as a generic scalar trust score for an arbitrary selection method. In ATS, this generic scalar score is instantiated by the aggregate trust score $T_{\mathrm{ATS},i} = T_{\mathrm{ATS}}(\mathbf{s}_i)$. The symbol τ remains a generic selection threshold, and λ denotes a candidate event-family threshold. $\mathbf{s}_i$ denotes the underlying vector of behavioral diagnostic scores. The VLM f is fixed throughout, and the diagnostic scores may use additional black-box queries to f without using y_i at inference time.

A.2 Selector Construction

ATS converts the aggregate trust score into an event-family-specific accept/reject rule using the validation split. For each event family $h \in \mathcal{H}$, define the corresponding validation index set as

$$\mathcal{I}_h = \{i : h_i = h\}.$$

Let Λ_h denote the finite set of candidate aggregate trust-score thresholds considered for event family h. For each $\lambda \in \Lambda_h$, define the preliminary validation subset

$$\mathcal{A}_{h,\lambda}^{T} = \{i \in \mathcal{I}_h : T_{\mathrm{ATS},i} \geq \lambda\}.$$

This subset is used only to estimate the remaining family-specific score thresholds. It is not the final set of accepted examples, because final acceptance additionally depends on the three diagnostic conditions defined below. A candidate value of λ is invalid when $\mathcal{A}_{h,\lambda}^{T} = \varnothing$, because the empirical quantiles required to construct the corresponding rule are then undefined.

For a nonempty finite multiset of real values

$$\mathcal{Z} = \{z_1, \ldots, z_m\},$$

let

$$Q_q(\mathcal{Z}), \qquad q \in (0,1),$$

denote its empirical q-quantile. Using only the label-free diagnostic scores in $\mathcal{A}_{h,\lambda}^{T}$, the practical implementation constructs

$$\alpha_{h,\lambda} = Q_{0.05}\left(\{s_{\mathrm{vp},i} : i \in \mathcal{A}_{h,\lambda}^{T}\}\right),$$
$$\beta_{h,\lambda} = Q_{0.05}\left(\{s_{\mathrm{cps},i} : i \in \mathcal{A}_{h,\lambda}^{T}\}\right),$$

and

$$\eta_{h,\lambda} = Q_{0.95}\left(\{r_{\mathrm{unk},i} : i \in \mathcal{A}_{h,\lambda}^{T}\}\right).$$

Thus, $\alpha_{h,\lambda}$ and $\beta_{h,\lambda}$ are lower acceptance thresholds for the video–prior consistency and counterfactual prior stability scores, respectively, whereas $\eta_{h,\lambda}$ is an upper acceptance threshold for the unknown-risk score. Together with the aggregate trust-score threshold λ, they define the complete candidate rule

$$\gamma_{h,\lambda} = (\alpha_{h,\lambda}, \beta_{h,\lambda}, \eta_{h,\lambda}, \lambda).$$

Here, $\gamma_{h,\lambda}$ denotes the complete rule induced for event family h by the aggregate operating point λ. It does not denote only the scalar threshold λ.

For a generic input–prediction pair $(x, \hat{y})$, with the underlying VLM f held fixed, let

$$\mathbf{s} = \mathbf{s}(x, \hat{y})$$

denote the ATS diagnostic-score vector obtained from the original prediction and the associated black-box behavioral probes. The corresponding event-family-specific selector function is

$$g_{\gamma_{h,\lambda}}(x,\hat{y}) = \mathbb{I}\big[\, s_{\mathrm{vp}}(x,\hat{y}) \geq \alpha_{h,\lambda} \;\wedge\; s_{\mathrm{cps}}(x,\hat{y}) \geq \beta_{h,\lambda} \\ \wedge\; r_{\mathrm{unk}}(x,\hat{y}) \leq \eta_{h,\lambda} \;\wedge\; T_{\mathrm{ATS}}(\mathbf{s}(x,\hat{y})) \geq \lambda\big],$$

where $\mathbb{I}[\cdot]$ is the indicator function. The selector returns 1 only when all four acceptance conditions hold and returns 0 otherwise. Thus, ATS performs post-hoc selection without modifying or replacing the original numerical prediction.

The practical construction can be summarized as

$$(h, \lambda) \longrightarrow \mathcal{A}_{h,\lambda}^{T} \longrightarrow (\alpha_{h,\lambda}, \beta_{h,\lambda}, \eta_{h,\lambda}) \longrightarrow \gamma_{h,\lambda}.$$

For each event family h, the implementation explicitly enumerates only $\lambda \in \Lambda_h$. Each candidate λ determines the preliminary validation subset $\mathcal{A}_{h,\lambda}^{T}$, from which the other three thresholds are deterministically derived. The implementation therefore does not independently enumerate $\alpha_{h,\lambda}$, $\beta_{h,\lambda}$, $\eta_{h,\lambda}$, and λ over a four-dimensional Cartesian grid. Each pair (h, λ) defines one complete candidate selection rule.

For each candidate $\gamma_{h,\lambda}$, the practical validation procedure computes empirical coverage, empirical selective risk, and the Hoeffding-style screening statistic defined below. Because the quantile thresholds are derived from the same validation split, these practical candidate rules are data-adaptive and are not covered directly by the strict finite-class guarantee stated later. The procedure then selects an operating point $\widehat{\lambda}_h \in \Lambda_h$ according to the validation criterion used in the main experiments and fixes the corresponding rule

$$\widehat{\gamma}_h = \gamma_{h,\widehat{\lambda}_h} = \left(\widehat{\alpha}_h, \widehat{\beta}_h, \widehat{\eta}_h, \widehat{\lambda}_h\right).$$

A.3 TEST-TIME SELECTION

During test evaluation, ATS uses the fixed rule associated with the event family of each test example. For a test input–prediction pair $(x_i, \hat{y}_i)$ with $h_i = H(x_i)$, the deployed selector is

$$g_{\widehat{\gamma}_{h_i}}(x_i, \hat{y}_i).$$

The original numerical prediction is accepted only when all four prespecified conditions in $\widehat{\gamma}_{h_i}$ are satisfied. Otherwise, it is rejected. No test labels are used to construct or modify the selector, and ATS neither revises nor replaces the original VLM prediction.

Finite-Class Selective-Risk Formulation. For the strict finite-class result, rules of the same form must instead be constructed on a disjoint design split and frozen before an independent risk-calibration split is examined. This strict theoretical formulation is separate from the practical two-way validation/test procedure described above. Let

$$\mathcal{D}_{\text{risk-cal}} = \{(x_i, \hat{y}_i, y_i, h_i)\}_{i=1}^{n}$$

denote that independent risk-calibration split, where n is its total number of examples. In this strict formulation, Λ_h serves as a finite index set for the complete rules prespecified for event family h. Each pair

$$(h, \lambda), \qquad h \in \mathcal{H}, \quad \lambda \in \Lambda_h,$$

indexes one frozen complete rule $\gamma_{h,\lambda}$.

Define

$$\Gamma_h = \{\gamma_{h,\lambda} : \lambda \in \Lambda_h\},$$

where Γ_h is the finite candidate-rule family associated with event family h. The complete candidate family across all event families is

$$\Gamma = \bigcup_{h \in \mathcal{H}} \Gamma_h, \qquad K = |\Gamma| = \sum_{h \in \mathcal{H}} |\Lambda_h|.$$

Thus, K counts all prespecified event-family and operating-point pairs (h, λ), rather than only the number $|\mathcal{H}|$ of event families.

For candidate $\gamma_{h,\lambda}$, define

$$A_i^{h,\lambda} = g_{\gamma_{h,\lambda}}(x_i, \hat{y}_i) \in \{0, 1\},$$

where $A_i^{h,\lambda} = 1$ means that risk-calibration example i is accepted by $\gamma_{h,\lambda}$, and $A_i^{h,\lambda} = 0$ means that it is rejected. The bounded per-example loss is

$$L_i = L(\hat{y}_i, y_i) = 1 - \text{MRA}(\hat{y}_i, y_i) \in [0, 1].$$

The number of accepted risk-calibration examples belonging to event family h is

$$n_{h,\lambda} = \sum_{i:h_i=h} A_i^{h,\lambda}.$$

When $n_{h,\lambda} > 0$, the empirical selective risk is

$$\widehat{R}_h(\gamma_{h,\lambda}) = \frac{\sum_{i:h_i=h} L_i A_i^{h,\lambda}}{n_{h,\lambda}}.$$

This is the mean loss among the risk-calibration examples that both belong to event family h and are accepted by $\gamma_{h,\lambda}$.

The corresponding selective risk is

$$R_h(\gamma_{h,\lambda}) = \mathbb{E}\Big[L(\hat{Y}, Y) \mid g_{\gamma_{h,\lambda}}(X, \hat{Y}) = 1,\ H(X) = h\Big].$$

It is the expected loss of a future prediction conditional on both belonging to event family h and being accepted by $\gamma_{h,\lambda}$.

For screening over the finite candidate family, define

$$\overline{R}_h(\gamma_{h,\lambda}) = \widehat{R}_h(\gamma_{h,\lambda}) + \sqrt{\frac{\log(K/\delta)}{2n_{h,\lambda}}},$$

where $\delta \in (0,1)$ is the desired total failure-probability level. The factor K accounts for comparison of all prespecified candidate rules, while the term depending on $n_{h,\lambda}$ penalizes rules supported by few accepted examples.

The complete candidate family Γ must be fixed independently of the entire risk-calibration split. In particular, thresholds, aggregation weights, event-family definitions, and acceptance conditions constructed using the same risk-calibration inputs fall outside the scope of the result below unless an additional valid conditional or uniform-convergence argument is supplied.

Conditional finite-class selective-risk bound.

Proposition 3. *Assume that risk-calibration units are conditionally i.i.d. within each event family, that $L(\hat{y}, y) \in [0,1]$, and that the complete finite candidate family Γ is fixed independently of the entire risk-calibration split. Then, with probability at least $1-\delta$, simultaneously for every $\gamma_{h,\lambda} \in \Gamma$ with $n_{h,\lambda} > 0$,*

$$R_h(\gamma_{h,\lambda}) \le \widehat{R}_h(\gamma_{h,\lambda}) + \sqrt{\frac{\log(K/\delta)}{2n_{h,\lambda}}}.$$

Consequently, any candidate selected only from this prespecified family using these simultaneous upper bounds inherits the corresponding bound.

Proof. Fix a candidate $\gamma_{h,\lambda}$. Because the complete selection rule is fixed independently of the risk-calibration split, its acceptance indicator is a fixed measurable function of $(x_i, \hat{y}_i)$.

Conditional on $n_{h,\lambda} = m > 0$, the m accepted losses are drawn from the distribution induced by

$$g_{\gamma_{h,\lambda}}(X, \hat{Y}) = 1, \qquad H(X) = h.$$

They lie in $[0,1]$ and have mean $R_h(\gamma_{h,\lambda})$. Therefore, Hoeffding's inequality gives

$$\Pr\Big(R_h(\gamma_{h,\lambda}) - \widehat{R}_h(\gamma_{h,\lambda}) \ge \epsilon \;\Big|\; n_{h,\lambda} = m\Big) \le \exp(-2m\epsilon^2).$$

Setting

$$\epsilon_m = \sqrt{\frac{\log(K/\delta)}{2m}}$$

makes the conditional failure probability at most δ/K for every $m > 0$. Averaging over the possible values of $n_{h,\lambda}$ yields

$$\Pr\left(R_h(\gamma_{h,\lambda}) > \widehat{R}_h(\gamma_{h,\lambda}) + \sqrt{\frac{\log(K/\delta)}{2n_{h,\lambda}}},\ n_{h,\lambda} > 0\right) \le \frac{\delta}{K}.$$

A union bound over the K prespecified candidates yields the simultaneous statement. Since the selected candidate is one member of the same fixed finite family, it inherits the corresponding bound on the same high-probability event. □

Monotonicity of counterfactual prior stability. We next prove the monotonicity property stated in Proposition 2 of the main paper. Recall that

$$\ell(z) = \log(\max\{|z|, \epsilon_{\log}\}),$$

where $\epsilon_{\log} > 0$ is the log-magnitude floor, and let

$$\rho = \frac{\mathrm{num}(p^{\mathrm{cf}})}{\mathrm{num}(p)}$$

denote the numerical scale ratio induced by the counterfactual prior. The idealized prior-tracking reference is $\rho\hat{y}_{\mathrm{full}}$. Define

$$d_{\mathrm{ref}} = |\ell(\hat{y}_{\mathrm{cf}}) - \ell(\hat{y}_{\mathrm{full}})|,$$

$$d_{\text{track}} = \left|\ell(\hat{y}_{\text{cf}}) - \ell(\rho\hat{y}_{\text{full}})\right|,$$

and

$$s_{\text{cps}} = \frac{d_{\text{track}}}{\max\{\zeta, d_{\text{track}} + d_{\text{ref}}\}},$$

where $\zeta > 0$ is the denominator floor.

Proof of Proposition 2. Assume that all relevant magnitudes exceed $\epsilon_{\log}$, so that $\ell(z) = \log|z|$ for the quantities considered. Let

$$\hat{y}_{\text{cf}} = \rho^{\alpha}\hat{y}_{\text{full}}, \qquad \alpha \in [0, 1],$$

where $\alpha = 0$ corresponds to reference stability and $\alpha = 1$ corresponds to full prior tracking. Since $\rho > 0$ and $\rho \neq 1$, let

$$A = \log\rho \neq 0.$$

Then

$$\begin{aligned} d_{\text{ref}} &= |\ell(\rho^{\alpha}\hat{y}_{\text{full}}) - \ell(\hat{y}_{\text{full}})| \\ &= |\alpha\log\rho| = \alpha|A|. \end{aligned}$$

Similarly,

$$\begin{aligned} d_{\text{track}} &= |\ell(\rho^{\alpha}\hat{y}_{\text{full}}) - \ell(\rho\hat{y}_{\text{full}})| \\ &= |(\alpha - 1)\log\rho| = (1 - \alpha)|A|. \end{aligned}$$

Substituting these expressions into the definition of s_{cps} gives

$$\begin{aligned} s_{\text{cps}}(\alpha) &= \frac{(1-\alpha)|A|}{\max\{\zeta, (1-\alpha)|A| + \alpha|A|\}} \\ &= \frac{(1-\alpha)|A|}{\max\{\zeta, |A|\}}. \end{aligned}$$

The denominator is constant with respect to α. Therefore,

$$\frac{d}{d\alpha}s_{\text{cps}}(\alpha) = -\frac{|A|}{\max\{\zeta, |A|\}} < 0.$$

Hence $s_{\text{cps}}(\alpha)$ is strictly decreasing on $[0, 1]$. At the endpoints,

$$s_{\text{cps}}(0) = \frac{|A|}{\max\{\zeta, |A|\}}, \qquad s_{\text{cps}}(1) = 0.$$

Thus, the score is largest under reference stability and decreases monotonically as the counterfactual prediction moves toward the idealized prior-tracking reference. □

Scope and interpretation. The finite-class proposition is a conditional result for selective risk among accepted predictions within each event family. It is not a conformal prediction guarantee, does not construct a prediction interval, and does not assign an answer-specific probability of correctness. An accepted prediction may therefore still be inaccurate.

The result requires the complete candidate family to be fixed independently of the risk-calibration split. If thresholds, aggregation weights, event-family definitions, or acceptance conditions are adapted using that same split, the proposition does not apply directly without additional data splitting or a suitable uniform-convergence argument.

The practical two-way validation/test procedure used in the main experiments should therefore be interpreted as an empirical selective-trust method. Its Hoeffding-style quantity is used only as a sample-size-penalized screening statistic and not as a formal finite-sample guarantee. A strict design/risk-calibration/test separation would satisfy the required data-separation condition, but such a separation is not used to support the main empirical claims in this work.

	Numerical parsing			Quantitative Reasoning Performance		
Backbone	N	Parsed	Parse rate	Mean MRA	Median MRA	Risk
Claude Sonnet 4.5	159	159	1.000	0.511	0.600	0.489
CogVLM2-Video	159	157	0.987	0.268	0.100	0.732
Gemini 3.1 Pro	159	159	1.000	0.586	0.700	0.414
Gemma 3-12B	159	159	1.000	0.401	0.400	0.599
GPT-5.1	159	159	1.000	0.545	0.600	0.455
GPT-5.6 Terra	159	159	1.000	0.532	0.600	0.468
InternVL3.5-14B	159	159	1.000	0.419	0.400	0.581
InternVL3.5-2B	159	159	1.000	0.213	0.100	0.787
LLaVA-1.5-13B	159	159	1.000	0.200	0.000	0.800
LLaVA-OneVision-7B	159	159	1.000	0.367	0.300	0.633
MiniCPM-o-4.5	159	159	1.000	0.394	0.400	0.606
Molmo-7B	159	159	1.000	0.296	0.200	0.704
Phi-3.5-Vision	159	159	1.000	0.333	0.300	0.667
Phi-4-Multimodal	159	159	1.000	0.256	0.100	0.744
Qwen2.5-VL-3B	159	159	1.000	0.350	0.300	0.650
Qwen2.5-VL-7B	159	159	1.000	0.399	0.400	0.601
Qwen3-VL-2B	159	159	1.000	0.283	0.200	0.717
Qwen3-VL-8B	159	159	1.000	0.395	0.400	0.605
VideoChat-R1.5-7B	159	159	1.000	0.404	0.400	0.596
VideoLLaMA3-7B	159	159	1.000	0.389	0.300	0.611

Table 4: Full-coverage numerical performance and numerical parse rates across the evaluated backbones.

B Additional Experimental Details

This section provides additional experimental details and results that complement the main paper, including the evaluation protocol, cross-backbone comparisons, failure-mode analyses, diagnostic-score studies, ablations, weight-sensitivity analyses, paired cross-backbone effect estimates, and query budgets. Unless otherwise stated, cross-backbone summaries include 20 VLMs, and all split-based results are averaged over the same 15 matched video-level validation/test splits for each backbone. summaries include 20 VLMs, and all split-based results are averaged over the same 15 matched video-level validation/test splits for each backbone. Thresholds, aggregation parameters, and operating points are selected on the validation split and then fixed before evaluation on the corresponding test split.

Probe	Video	Prior	Purpose
Video+Prior	✓	Reference	Standard prediction
Prior-only	–	Reference	Video–prior dependence
Counterfactual prior	✓	Altered	Prior sensitivity
CoT-prompted	✓	Reference	Prompt consistency
Repeated sampling	✓	Reference	Sampling stability

Table 3: Controlled probe variants used by ATS.

B.1 Evaluation Protocol and Execution Pipeline

We generate 15 video-level validation/test splits using different random seeds. All records associated with the same video remain in the same split to prevent data leakage. All validation procedures, including diagnostic-score normalization, learned aggregation, threshold estimation, and operating-point selection, are performed using only the validation set and then applied unchanged to the corresponding test set. Unparsable responses are retained as invalid outputs.

Backbone	MRA@30	MRA@50	MRA@70	AURC ↓	Δ_{50}	Δ_{AURC}
Claude Sonnet 4.5	0.708	0.612	0.587	0.376	-0.042	0.003
CogVLM2-Video	0.341	0.336	0.311	0.657	**0.030**	**0.053**
Gemini 3.1 Pro	**0.746**	**0.720**	**0.674**	**0.309**	**0.031**	**0.043**
Gemma 3-12B	0.601	0.541	0.468	0.475	**0.046**	**0.009**
GPT-5.1	0.661	0.650	0.618	0.366	**0.025**	**0.019**
GPT-5.6 Terra	0.665	0.603	0.576	0.416	-0.011	**0.002**
InternVL3.5-14B	0.637	0.552	0.512	0.450	**0.051**	**0.076**
InternVL3.5-2B	0.262	0.242	0.246	0.750	-0.023	-0.016
LLaVA-1.5-13B	0.298	0.255	0.249	0.741	**0.038**	**0.056**
LLaVA-OneVision-7B	0.431	0.414	0.423	0.574	-0.042	-0.033
MiniCPM-o-4.5	0.476	0.458	0.445	0.541	**0.028**	**0.027**
Molmo-7B	0.414	0.358	0.348	0.591	-0.013	**0.017**
Phi-3.5-Vision	0.429	0.428	0.365	0.591	**0.076**	**0.036**
Phi-4-Multimodal	0.369	0.328	0.290	0.688	**0.028**	**0.059**
Qwen2.5-VL-3B	0.426	0.423	0.383	0.609	-0.060	-0.046
Qwen2.5-VL-7B	0.519	0.489	0.448	0.530	-0.010	**0.004**
Qwen3-VL-2B	0.325	0.303	0.299	0.680	-0.023	-0.010
Qwen3-VL-8B	0.531	0.468	0.437	0.506	-0.027	**0.013**
VideoChat-R1.5-7B	0.476	0.472	0.445	0.531	-0.014	-0.010
VideoLLaMA3-7B	0.445	0.397	0.398	0.595	-0.042	-0.037
Mean	0.488	0.452	0.426	0.549	0.002	0.013
Median	0.461	0.443	0.430	0.558	-0.011	0.011

Table 5: Selective-ranking performance of Prior-weighted ATS across the evaluated backbones.

Method	MRA@30% ↑	MRA@50% ↑	MRA@70% ↑	AURC ↓
Self-consistency and generic UQ baselines				
SC-only	$\mathbf{0.543} \pm 0.082$	$\mathbf{0.499} \pm 0.060$	0.443 ± 0.048	0.534 ± 0.040
Generic UQ mean	0.468 ± 0.051	0.453 ± 0.065	$\mathbf{0.453} \pm 0.050$	0.515 ± 0.037
Generic UQ rank fusion	0.491 ± 0.073	0.471 ± 0.050	0.450 ± 0.049	$\mathbf{0.513} \pm 0.038$
Learned Generic UQ	0.506 ± 0.112	0.460 ± 0.091	0.439 ± 0.072	0.559 ± 0.070
ATS variants				
Equal-weight ATS	0.437 ± 0.081	0.444 ± 0.057	0.444 ± 0.046	0.566 ± 0.027
Prior-weighted ATS	0.519 ± 0.064	0.489 ± 0.051	0.448 ± 0.042	0.530 ± 0.031
Learned ATS	0.407 ± 0.139	0.412 ± 0.073	0.403 ± 0.058	0.598 ± 0.069
Upper bound				
Oracle	0.842 ± 0.031	0.701 ± 0.044	0.557 ± 0.036	0.305 ± 0.027

Table 6: Comparison of selective-prediction methods on Qwen2.5-VL-7B. Results are mean ± standard deviation. Bold indicates the best non-oracle result in each column.

B.2 Cross-Backbone Benchmark Results

Full-coverage base performance. Table 4 reports numerical answer quality and numerical parse rates before selective filtering. Full-coverage MRA varies substantially across backbones, indicating clear differences in base prediction quality. We therefore report both per-backbone selective-prediction performance and cross-backbone summaries to characterize results at the individual-model and aggregate levels.

Cross-backbone selective prediction. Table 5 reports Prior-weighted ATS at 30%, 50%, and 70% coverage, together with AURC and paired differences relative to SC-only. ATS improves MRA at 50% coverage on 9 of 20 backbones and reduces AURC on 14 of 20. Across backbones, the mean differences are $+0.002$ in MRA@50 and $+0.013$ in AURC, where positive differences favour ATS. These results indicate that the fixed-coverage effects vary across backbones, whereas AURC favours ATS on a larger proportion of backbones.

B.3 Baseline and Aggregation Comparisons

Baselines and aggregation strategies. The SC-only baseline estimates uncertainty from self-consistency under the standard Video+Prior input. For each example, it combines the original prediction with three additional stochastic generations from the same input, yielding four outputs in total, and ranks examples using s_{rep}, which measures the log-space stability of valid repeated numerical predictions.

The Generic UQ baselines use s_{rep}, s_{agr}, and s_{prompt}. Here, s_{agr} is the fraction of repeated outputs that are valid positive numerical predictions, whereas s_{prompt} measures consistency across the Prior-Only, Video+Prior, and CoT-Prompted conditions. These baselines exclude s_{vp} and s_{cps} as explicit input features, but still use the Prior-Only response through s_{prompt}.

B.4 Failure-Mode Analysis

Failure-mode definitions. We define all diagnostic groups before evaluating any selector. Ground-truth MRA is used only to construct post-hoc evaluation groups and is never provided to the selector.

For each backbone b and split s, we define a high-stability threshold from the corresponding validation set:

$$q_{\text{stab}}^{(b,s)} = \text{Quantile}_{0.85}\left(\left\{s_{\text{rep},i} : i \in \mathcal{D}_{\text{val}}^{(b,s)}\right\}\right).$$

This threshold is then fixed and applied unchanged to the held-out test set. Predictions with $s_{\text{rep},i} \geq q_{\text{stab}}^{(b,s)}$ are therefore in the highest-stability region for that backbone and split. We use the fixed accuracy thresholds

$$q_{\text{wrong}} = 0.20, \qquad q_{\text{correct}} = 0.40,$$

for all selectors.

A test prediction is classified as stable-but-wrong if it is highly stable under repeated sampling but has low numerical accuracy:

$$\mathcal{F}_{\text{SBW}}^{(b,s)} = \left\{i : s_{\text{rep},i} \geq q_{\text{stab}}^{(b,s)},\ \text{MRA}(\hat{y}_i, y_i) \leq q_{\text{wrong}}\right\}.$$

To identify prior tracking, we compare the counterfactual prediction with two references. The unchanged-answer reference assumes that the prediction remains close to the original Video+Prior output:

$$d_{i,\text{ref}} = \left|\ell(\hat{y}_{i,\text{cf}}) - \ell(\hat{y}_{i,\text{full}})\right|.$$

The idealized prior-tracking reference assumes that the prediction changes in proportion to the prior transformation:

$$d_{i,\text{track}} = \left|\ell(\hat{y}_{i,\text{cf}}) - \ell(\rho_i \hat{y}_{i,\text{full}})\right|.$$

A valid counterfactual pair with $\rho_i > 0$ is classified as prior tracking when the counterfactual prediction is closer to the idealized tracking reference than to the unchanged-answer reference:

$$\mathcal{F}_{\text{PT}}^{(b,s)} = \left\{i : d_{i,\text{track}} < d_{i,\text{ref}}\right\}.$$

Invalid counterfactual pairs are excluded from this analysis because tracking cannot be determined reliably. The wrong prior-tracking subset is

$$\mathcal{F}_{\text{PTW}}^{(b,s)} = \left\{i \in \mathcal{F}_{\text{PT}}^{(b,s)} : \text{MRA}(\hat{y}_i, y_i) \leq q_{\text{wrong}}\right\}.$$

To quantify the cost of rejecting these failures, we also define corresponding correct-case groups. Highly stable correct predictions are

$$\mathcal{F}_{\text{SBC}}^{(b,s)} = \left\{i : s_{\text{rep},i} \geq q_{\text{stab}}^{(b,s)},\ \text{MRA}(\hat{y}_i, y_i) \geq q_{\text{correct}}\right\},$$

and correct prior-tracking predictions are

$$\mathcal{F}_{\text{PTC}}^{(b,s)} = \left\{i \in \mathcal{F}_{\text{PT}}^{(b,s)} : \text{MRA}(\hat{y}_i, y_i) \geq q_{\text{correct}}\right\}.$$

Selector	Stable-but-wrong		Prior tracking		Backbones
	Reject ↑	Correct retain ↑	Wrong reject ↑	Correct retain ↑	
SC-only	0.013	0.986	0.654	0.491	20
Learned Generic UQ	0.271	0.830	0.638	0.523	20
Equal-weight ATS	0.424	0.723	0.906	0.248	20
Prior-weighted ATS	0.409	0.727	0.958	0.111	20
Learned ATS	0.379	0.777	0.574	0.512	20
Learned ATS w/o s_{cps}	0.376	0.792	0.588	0.521	20
Oracle	0.963	0.941	0.916	0.919	20

Table 7: Targeted failure-mode performance at 50% coverage, averaged over all backbones.

Backbone	Stable-but-wrong	Prior tracking	Wrong prior tracking
Claude Sonnet 4.5	4	45	18
CogVLM2-Video	14	21	13
Gemini 3.1 Pro	7	44	12
Gemma 3-12B	8	15	8
GPT-5.1	8	33	10
GPT-5.6 Terra	10	61	18
InternVL3.5-14B	10	18	13
InternVL3.5-2B	13	32	23
LLaVA-1.5-13B	17	19	16
LLaVA-OneVision-7B	7	23	8
MiniCPM-o-4.5	37	21	11
Molmo-7B	5	16	10
Phi-3.5-Vision	9	31	19
Phi-4-Multimodal	19	22	15
Qwen2.5-VL-3B	8	29	16
Qwen2.5-VL-7B	8	28	15
Qwen3-VL-2B	12	27	15
Qwen3-VL-8B	16	13	5
VideoChat-R1.5-7B	8	26	12
VideoLLaMA3-7B	7	20	10

Table 8: Numbers of available prediction-level diagnostic cases across the evaluated backbones.

We evaluate target coverages $c \in \{0.30, 0.50, 0.70\}$, and Table 7 reports the prespecified $c = 0.50$ operating point. Let $g_{i,c} = 1$ indicate that prediction i is retained at coverage c. For a non-empty failure set $\mathcal{F}$, rejection is

$$\mathrm{Reject}_c(\mathcal{F}) = 1 - \frac{\sum_{i \in \mathcal{F}} g_{i,c}}{|\mathcal{F}|}.$$

For a non-empty corresponding correct-case set $\mathcal{F}^+$, retention is

$$\mathrm{Retain}_c(\mathcal{F}^+) = \frac{\sum_{i \in \mathcal{F}^+} g_{i,c}}{|\mathcal{F}^+|}.$$

Failure-case rejection rates and corresponding correct-case retention rates are computed separately for each backbone and split. If a backbone-split pair contains no eligible examples for a diagnostic group, the corresponding rate is treated as undefined and excluded from the average rather than set to zero. We first average valid split-level rates within each backbone and then average the resulting backbone-level rates across backbones. This gives each backbone equal weight regardless of the number of available cases.

Targeted failure rejection. Stable-but-wrong examples directly test a failure mode that consistency-only uncertainty may miss. At 50% coverage, ATS variants achieve the highest non-oracle rejection rates for these cases: Equal-weight ATS rejects 0.424 and Prior-weighted ATS rejects 0.409, compared

Backbone	Intervention		Stability			Validity, scale, familiarity		
	s_{vp}	s_{cps}	s_{rep}	s_{agr}	s_{prompt}	s_{valid}	s_{scale}	s_{fam}
Claude Sonnet 4.5	0.716	0.680	0.765	0.989	0.761	0.987	0.708	0.414
CogVLM2-Video	0.608	0.673	0.631	0.851	0.683	1.000	0.699	0.414
Gemini 3.1 Pro	0.737	0.643	0.844	0.954	0.726	0.994	0.677	0.414
Gemma 3-12B	0.642	0.831	0.850	1.000	0.602	0.981	0.714	0.414
GPT-5.1	0.730	0.684	0.879	0.966	0.793	1.000	0.695	0.414
GPT-5.6 Terra	0.712	0.561	0.840	0.977	0.756	1.000	0.692	0.414
InternVL3.5-14B	0.586	0.797	0.813	0.962	0.624	1.000	0.707	0.414
InternVL3.5-2B	0.536	0.683	0.689	0.975	0.605	0.937	0.774	0.414
LLaVA-1.5-13B	0.670	0.747	0.644	0.998	0.598	0.912	0.786	0.414
LLaVA-OneVision-7B	0.707	0.754	0.627	1.000	0.713	1.000	0.719	0.414
MiniCPM-o-4.5	0.646	0.719	0.873	1.000	0.584	0.937	0.687	0.414
Molmo-7B	0.676	0.719	0.658	0.998	0.572	0.994	0.689	0.414
Phi-3.5-Vision	0.807	0.668	0.667	0.948	0.712	1.000	0.720	0.414
Phi-4-Multimodal	0.554	0.762	0.764	0.987	0.596	0.994	0.635	0.414
Qwen2.5-VL-3B	0.678	0.744	0.642	0.985	0.707	1.000	0.770	0.414
Qwen2.5-VL-7B	0.500	0.717	0.737	0.985	0.562	0.717	0.657	0.414
Qwen3-VL-2B	0.653	0.759	0.699	0.994	0.567	0.987	0.758	0.414
Qwen3-VL-8B	0.666	0.726	0.868	0.929	0.563	0.786	0.643	0.414
VideoChat-R1.5-7B	0.629	0.764	0.737	0.996	0.583	0.918	0.714	0.414
VideoLLaMA3-7B	0.843	0.822	0.668	1.000	0.772	0.969	0.783	0.414

Table 9: Mean diagnostic-score profiles across the evaluated backbones.

with 0.271 for Generic UQ and 0.013 for SC-only. Prior tracking captures cases in which the counterfactual prediction is closer to the idealized prior-scaled reference than to the unchanged original prediction. We report wrong-case rejection together with corresponding correct-case retention to show the resulting rejection–retention trade-off rather than interpreting high rejection on its own.

Prediction-level failure cases. For each failure case included in the qualitative analysis, we retain the question, model prediction, ground truth, MRA, relevant diagnostic scores, and failure-mode label. Table 8 summarizes the number of such cases available for each backbone. These counts show only how many cases were selected for qualitative analysis. They do not indicate how often each failure mode actually occurs.

B.5 Diagnostic Score Analysis

Cross-backbone diagnostic profiles. Table 9 reports the mean diagnostic scores for each backbone. The intervention scores s_{vp} and s_{cps}, together with the stability scores s_{rep} and s_{prompt}, show substantial variation across backbones, indicating that different models respond differently to prior interventions, repeated generation, and prompt perturbations. By comparison, s_{agr} and s_{valid} are generally high, suggesting that most backbones usually return valid numerical answers across repeated queries. The scale-plausibility score s_{scale} also varies across backbones, whereas s_{fam} remains constant because it is determined by the common benchmark event-family distribution rather than by model responses.

Quantity	Mean	SD	Min	Max
Intervention scores				
s_{vp}	0.665	0.085	0.500	0.843
s_{cps}	0.723	0.063	0.561	0.831
Stability scores				
s_{rep}	0.745	0.091	0.627	0.879
s_{agr}	0.975	0.035	0.851	1.000
s_{prompt}	0.654	0.081	0.562	0.793
Validity scores				
s_{valid}	0.956	0.076	0.717	1.000
s_{scale}	0.711	0.044	0.635	0.786
s_{fam}	0.414	0.000	0.414	0.414
Aggregate trust score				
T_{ATS}	3.759	0.163	3.438	4.108

Table 10: Summary statistics of ATS diagnostic scores and aggregate trust scores across evaluated backbones.

B.6 Ablation Analysis

We assess the contribution of each diagnostic score through leave-one-score-out ablations under both fixed and learned aggregation. In the fixed-weight setting, each score is removed while all remaining weights are kept unchanged. In the learned setting, the aggregation model is retrained after removing the corresponding score, allowing the remaining diagnostics to redistribute their contributions. Tables 11 and 12 summarize the cross-backbone results, while Tables 13 and 14 report the corresponding results for Qwen2.5-VL-7B. Overall, no single diagnostic is uniformly dominant across all settings, although s_{rep} shows a relatively consistent contribution. The remaining effects vary across backbones and aggregation schemes, suggesting partial redundancy and context-dependent utility among the diagnostic scores.

Variant	MRA@ 30% ↑	MRA@ 50% ↑	ΔMRA @50	MRA@ 70% ↑	AURC ↓
Prior-weighted ATS	0.4879	0.4525	0.0000	0.4262	0.5488
w/o s_{vp}	0.4837	0.4411	-0.0114	0.4162	0.5533
w/o s_{cps}	0.4888	0.4646	+0.0121	0.4322	0.5433
w/o s_{fam}	0.4896	0.4530	+0.0005	0.4261	0.5471
w/o s_{rep}	0.4696	0.4403	-0.0122	0.4207	0.5609
w/o s_{agr}	0.4878	0.4523	-0.0002	0.4259	0.5491
w/o s_{scale}	0.4865	0.4496	-0.0029	0.4248	0.5491
w/o s_{prompt}	0.4856	0.4474	-0.0051	0.4243	0.5540
w/o s_{valid}	0.4872	0.4527	+0.0002	0.4259	0.5479

Table 11: Prior-weighted ATS leave-one-score-out ablations averaged across the evaluated backbones.

Variant	MRA@ 30% ↑	MRA@ 50% ↑	ΔMRA @50	MRA@ 70% ↑	AURC ↓
Learned ATS	0.4606	0.4413	0.0000	0.4228	0.5631
w/o s_{vp}	0.4594	0.4435	+0.0022	0.4243	0.5622
w/o s_{cps}	0.4638	0.4445	+0.0032	0.4253	0.5606
w/o s_{fam}	0.4690	0.4435	+0.0022	0.4243	0.5610
w/o s_{rep}	0.4447	0.4319	-0.0094	0.4174	0.5756
w/o s_{agr}	0.4610	0.4413	0.0000	0.4228	0.5610
w/o s_{scale}	0.4666	0.4431	+0.0018	0.4252	0.5604
w/o s_{prompt}	0.4521	0.4344	-0.0069	0.4205	0.5698
w/o s_{valid}	0.4647	0.4435	+0.0022	0.4234	0.5611

Table 12: Learned ATS leave-one-score-out ablations averaged across the evaluated backbones.

Variant	MRA@ 30% ↑	MRA@ 50% ↑	ΔMRA @50	MRA@ 70% ↑	AURC ↓
Prior-weighted ATS	0.5187	0.4889	0.0000	0.4476	0.5300
w/o s_{vp}	0.4879	0.4734	-0.0155	0.4352	0.5406
w/o s_{cps}	0.4780	0.4886	-0.0003	0.4671	0.5308
w/o s_{fam}	0.5344	0.4897	+0.0009	0.4459	0.5296
w/o s_{rep}	0.4984	0.4752	-0.0137	0.4385	0.5477
w/o s_{agr}	0.5187	0.4885	-0.0004	0.4452	0.5307
w/o s_{scale}	0.5142	0.4855	-0.0033	0.4520	0.5321
w/o s_{prompt}	0.5239	0.4901	+0.0012	0.4461	0.5431
w/o s_{valid}	0.5337	0.4865	-0.0024	0.4516	0.5214

Table 13: Prior-weighted ATS leave-one-score-out ablations on Qwen2.5-VL-7B.

Variant	MRA@ 30% ↑	MRA@ 50% ↑	ΔMRA @50	MRA@ 70% ↑	AURC ↓
Learned ATS	0.4069	0.4120	0.0000	0.4028	0.5977
w/o s_{vp}	0.4128	0.3926	-0.0193	0.3994	0.5984
w/o s_{cps}	0.4169	0.4037	-0.0083	0.3973	0.6020
w/o s_{fam}	0.4107	0.4197	+0.0078	0.4186	0.5956
w/o s_{rep}	0.3793	0.3968	-0.0152	0.3991	0.6167
w/o s_{agr}	0.3999	0.4107	-0.0012	0.4064	0.5948
w/o s_{scale}	0.4090	0.4231	+0.0112	0.4073	0.5899
w/o s_{prompt}	0.4052	0.4166	+0.0047	0.4053	0.5945
w/o s_{valid}	0.4486	0.4204	+0.0084	0.4131	0.5821

Table 14: Learned ATS leave-one-score-out ablations on Qwen2.5-VL-7B.

Weight sensitivity. To examine the robustness of Prior-weighted ATS to the manually specified aggregation weights, we randomly perturbed the default weight vector and reevaluated the resulting selectors on Qwen2.5-VL-7B. Performance remained stable under these perturbations: the mean MRA@50 was 0.4881, compared with 0.4889 for the default configuration, while the corresponding AURC values were 0.5306 and 0.5300. The 10th-percentile perturbed configuration achieved an MRA@50 of 0.4804 and an AURC of 0.5363. These results suggest that, on Qwen2.5-VL-7B, Prior-weighted ATS is not highly sensitive to moderate perturbations of the manually specified aggregation weights.

B.7 Cross-Backbone Paired Effect Analysis

This analysis characterizes the direction and variability of the difference between Prior-weighted ATS and SC-only across backbones. We first average the paired split-level differences within each

	Query budget		Probe calls				
Variant	Calls	Total	Base	Prior-only	CF	CoT	SC samples
Base VLM	1	159	1	0	0	0	0
SC-only	4	636	1	0	0	0	3
ATS-lite	3	477	1	1	1	0	0
Full ATS	7	1113	1	1	1	1	3
Learned ATS/UQ	7	1113	1	1	1	1	3

Table 16: Query budget for the 159-example benchmark.

backbone, yielding one effect estimate per backbone and metric. For fixed-coverage MRA, we define $\Delta_{\mathrm{MRA@}c} = \mathrm{MRA@}c_{\mathrm{ATS}} - \mathrm{MRA@}c_{\mathrm{SC}}$. Because lower AURC is preferable, we define $\Delta_{\mathrm{AURC}} = \mathrm{AURC}_{\mathrm{SC}} - \mathrm{AURC}_{\mathrm{ATS}}$. Under both definitions, positive values favour Prior-weighted ATS. We report the mean and median backbone-level differences, the number of positive differences, and 95% percentile bootstrap confidence intervals for the cross-backbone mean difference.

Metric	Mean Δ	Median Δ	95% CI	Wins
MRA@30%	$+0.0192$	$+0.0064$	$[-0.0075, 0.0485]$	$12/20$
MRA@50%	$+0.0023$	-0.0106	$[-0.0137, 0.0186]$	$9/20$
MRA@70%	-0.0037	-0.0043	$[-0.0120, 0.0042]$	$8/20$
AURC	$+0.0133$	$+0.0109$	$[-0.0009, 0.0277]$	$14/20$

Table 15: Backbone-level paired differences between Prior-weighted ATS and SC-only. The 95% confidence intervals are percentile bootstrap intervals for the mean difference.

As shown in Table 15, Prior-weighted ATS exhibits positive mean effects at 30% and 50% coverage across the 20 backbones, while the mean effect is slightly negative at 70% coverage. For AURC, ATS achieves a positive mean difference and yields a positive difference on 14 of the 20 backbones, the strongest directional trend among the reported metrics. However, all bootstrap confidence intervals include zero, so these results should be interpreted as descriptive cross-backbone trends rather than statistically conclusive evidence of superiority.