When Does Data Leakage Translate into Evaluation Gains? A Study of VLM Post-Training

Anonymous authors
Paper under double-blind review

Abstract

Why does evaluation-data exposure yield substantial score gains in some post-training settings but only limited average changes in others? We study train–evaluation overlap in vision–language models (VLMs) across training objectives and task settings. On natural tasks, increasing overlap from 0% to 100% raises fixed-set accuracy by 21.2 percentage points for supervised fine-tuning and by 14.0 and 14.2 points for DPO and PPO, but by only about 3.7 points for both GRPO and RLOO. In the adaptive visual-token setting, the five post-training objectives instead show fixed-set changes from -0.7 to $+1.3$ percentage points, including a near-zero response for supervised fine-tuning. Limited average responses therefore occur in distinct post-training settings and cannot be assigned to an optimizer category alone. To analyze this variation, we combine local propagation analysis with an exact finite-horizon decomposition connecting exposure-induced training differences, their propagation through later updates, and whether the resulting behavioral changes affect the evaluation score. This framework distinguishes a change in training from an observable evaluation benefit without assuming that similar terminal responses share one mechanism. Together, the results show that exposure responses must be interpreted in relation to the training setting instead of being inferred from an optimization label when considered in isolation.

1 Introduction

Vision–language models (VLMs) are commonly post-trained through supervised fine-tuning, preference optimization, and reinforcement learning (RL) (Liu et al., 2023; Rafailov et al., 2023; Ouyang et al., 2022), and evaluated on public multimodal benchmarks (Srivastava et al., 2022; Liang et al., 2022). Evaluation-data leakage can affect these results when evaluation samples enter the post-training data, as benchmark instances or derivatives may occur in large-scale corpora even after filtering (Brown et al., 2020; Dodge et al., 2021; Achiam et al., 2023). Yet knowing that a model encountered an evaluation item during training is not enough to determine how much that encounter will matter: different objectives use sample information differently, so the same train–evaluation overlap need not produce the same score change. The central question is not merely whether an evaluation item appeared, but whether its training signal can alter the model and create an additional advantage at evaluation. Prior work also shows that memorization depends on repetition and measurement, while extraction and membership tests capture different behavioral signatures (Tirumala et al., 2022; Carlini et al., 2022b; 2021; 2022a). To avoid conflating overlap with its consequences, we call contact with evaluation samples during training *exposure*, and the evaluation change observed as exposure conditions vary an *exposure response*. We ask why comparable exposure produces markedly different responses across VLM post-training settings.

Our results show that limited average responses arise in more than one training setting. Across the natural multimodal tasks, increasing overlap from 0% to 100% raises fixed-set accuracy by only about 3.7 percentage points for GRPO and RLOO, compared with 21.2, 14.0, and 14.2 points for SFT, DPO, and PPO. In the adaptive visual-token setting, all five objectives have fixed-set changes between -0.7 and $+1.3$ points, although some disaggregated combinations shift more. SFT can therefore show either a strong or limited average response across the two settings, while policy-optimization methods also do not respond uniformly. The shared observation is not that one

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.

algorithm class is insensitive to leakage, but that exposure does not always become a large terminal score change; we therefore treat these outcomes as instances of one question without presuming that the same mechanism produced them.

To understand this variation, we develop a task-conditioned framework for how training exposure becomes an evaluation change. Exposure must first alter the training process. Subsequent updates may then amplify, attenuate, or cancel that alteration, and only the influence retained in the final model can affect evaluation. Moreover, the remaining behavioral change must be distinguishable by the scorer. A visual-selection policy, for example, may shift attention to a different region while both regions still lead to the same answer; behavior has changed, but answer accuracy need not. We formalize this process through a local propagation analysis for differentiable updates, together with an exact finite-horizon decomposition for bounded scores in the appendix, and derive corresponding response and behavioral-observability bounds. The analysis preserves differences among post-training methods in supervision, sampling, and update rules: it neither forces them into a common reward-gradient form nor presupposes which optimizer must be more susceptible to leakage. It therefore separates the magnitude of an exposure-induced update from the magnitude of the exposure benefit ultimately observed in the resulting final evaluation.

Guided by this framework, we distinguish empirical regularities in exposure response from possible mechanisms that could produce them. Our contributions are:

1. a systematic characterization of strong and limited average exposure responses across post-training objectives and task settings, including limited responses that arise in distinct settings but are not universally zero; and
2. a task-conditioned theoretical framework connecting training changes, their downstream influence, and final scoring, which explains why "seen during training" need not imply "benefits at evaluation."

2 Related Work

Training-data contamination and sample memorization. Research on data contamination commonly asks whether a model encountered evaluation data and whether that contact affected evaluation. Prior work detects training-data exposure through prompted completion, sequence likelihood, and membership signals (Golchin & Surdeanu, 2024; Oren et al., 2024; Shi et al., 2024). Related memorization studies show that repetition, model scale, and the training trajectory affect retention of individual samples, while deduplication can reduce some memorization risk (Tirumala et al., 2022; Carlini et al., 2022b; Kandpal et al., 2022; Lee et al., 2022; Biderman et al., 2023). Exposure, memorization, and downstream use are nevertheless distinct: prior work separates retaining contaminated information from exploiting it and surveys the broader leakage landscape (Magar & Schwartz, 2022; Xu et al., 2024; Balloccu et al., 2024). Memorized information may also remain latent under ordinary prompts, and observed regurgitation depends on the elicitation and evaluation protocol (Sakarvadia et al., 2025; Duan et al., 2025; Aerni et al., 2025). We connect these stages in post-training by asking how strongly different objectives and task settings convert repeated contact with evaluation samples into a terminal evaluation response.

Post-training objectives and reward feedback. Visual instruction tuning, preference optimization, and RL provide distinct routes for adapting VLMs (Liu et al., 2023; Rafailov et al., 2023; Ouyang et al., 2022). Policy-optimization methods such as PPO, GRPO, and RLOO differ in sampling, advantage estimation, baselines, and regularization, so the same training sample need not induce the same update (Schulman et al., 2017; Shao et al., 2024; Ahmadian et al., 2024). Optimizing an imperfect reward can produce reward overoptimization (Gao et al., 2023; Dai et al., 2025), while changing the feedback design can qualitatively alter learned behavior (Kumar et al., 2025). VLM-R1 likewise reports gains in visual reasoning alongside reward-exploitation behavior (Shen et al., 2025). These findings establish the importance of training feedback, but reward overoptimization and evaluation-data leakage are different questions: the former concerns how a model exploits a reward objective, whereas the latter asks whether exposure to evaluation samples produces an additional evaluation change. We do not propose a new optimizer; we compare leakage responses across post-training methods and analyze why exposure-induced updates need not themselves become persistent evaluation advantages in the final model.

Multimodal tasks and behavioral evaluation. Modern VLMs connect language models to visual representations and extend this interface across resolutions, sequence lengths, images, and video (Alayrac et al., 2022; Li et al., 2023; 2024; Bai et al., 2023). Our backbone panel spans InternVL3.5, LLaVA-OneVision, Qwen2.5-VL, and Qwen3-VL (Wang et al., 2025; Li et al., 2024; Bai et al., 2025b;a); variation in visual pretraining further cautions against treating backbones as interchangeable replications (Keh et al., 2025). The evaluated tasks cover diagrams and charts (Kembhavi et al., 2016; Masry et al., 2022), document and scene text (Mathew et al., 2021; Liu et al., 2024b; Singh et al., 2019), multimodal knowledge (Liu et al., 2024a; Yue et al., 2024), and integrated visual reasoning (Chen et al., 2024; Yu et al., 2023; Lu et al., 2024); the behaviors altered during training therefore need not be the same across tasks. Beyond direct answer generation, visual tool use and evidence selection make "what to inspect" or "what action to take" part of the learned policy (Wu et al., 2026; Sarch et al., 2026), while OSWorld and WebRL extend this setting to multi-step interaction trajectories (Xie et al., 2024; Qi et al., 2025). For leakage analysis, answer-level scores may consequently conceal distinct behavioral changes. Our theory separates behavioral change from score change: with downstream processing fixed, a model may alter its visual selection or answer distribution while moving only among outcomes that receive the same score. We therefore analyze how each task maps behavior to evaluation, rather than treating a small terminal score difference as proof that training exposure had no effect on model behavior.

Evaluation aggregation and interpretation. Cross-task evaluation must retain task structure rather than treating every model and sample as a homogeneous replicate. Statistical evaluation work shows that ignoring dataset structure can conceal task-specific changes and make conclusions depend on the aggregation rule (Demšar, 2006; Dror et al., 2018). Paloma similarly emphasizes domain-aware analysis (Magnusson et al., 2024), while dynamically refreshed benchmarks such as LiveBench address a complementary source of contamination (White et al., 2025). We connect this issue to exposure-response analysis: a small average response may reflect uniformly small changes or cancellation between positive and negative responses across conditions. We therefore report both aggregate trends and disaggregated task results, then use the local response analysis and exact finite-horizon decomposition to characterize how such responses can arise. Our aim is not to label any optimizer or task setting as inherently leakage-resistant, but to characterize variation in exposure response and relate it to training updates, their subsequent influence, and task scoring. These conditions provide an interpretive framework; terminal scores alone do not prove a particular internal memorization mechanism within the model under evaluation.

3 A Task-Conditioned Analysis of Exposure Response

Limited average exposure responses can arise under different objectives and task settings. We do not assume that they share a mechanism. Instead, we analyze two links common to each case: how exposure-induced training differences reach the final model, and when the resulting behavioral changes become score changes. Later training can attenuate or cancel a changed update, and behavioral changes may escape the evaluation metric. Without assuming which post-training method is more susceptible to leakage, we provide formal protocol definitions and an exact stochastic decomposition for bounded terminal scores in Appendix A.1.

3.1 From Exposure to a Final Score Change

Consider two training trajectories with the same initial state and number of steps. Superscripts 0 and r denote a clean condition and a condition that includes the relevant evaluation samples. Write their updates as

$$w_{t+1}^0 = U_t^0(w_t^0), \qquad w_{t+1}^r = U_t^r(w_t^r). \tag{1}$$

Here w_t includes model parameters and continuous optimizer state. This local analysis assumes differentiable updates; Appendix A.1 covers general stochastic training.

Because the two current states may already differ, their executed updates cannot be attributed entirely to exposure at step t. We instead compare the two update rules at the same exposed-trajectory state:

$$d_t = U_t^r(w_t^r) - U_t^0(w_t^r). \tag{2}$$

Thus, d_t is the extra change induced by using the exposed rather than clean update rule at w_t^r. If the conditions also change training budget, data volume, or other settings, d_t includes those changes and cannot be interpreted as the isolated influence of one evaluation example.

Let $\Delta w_t = w_t^r - w_t^0$. A first-order expansion of the clean update near w_t^0 gives

$$\Delta w_{t+1} \approx J_t \Delta w_t + d_t, \qquad J_t = \left.\frac{\partial U_t^0(w)}{\partial w}\right|_{w=w_t^0}. \tag{3}$$

The direct term d_t is introduced at the current step, whereas $J_t \Delta w_t$ describes how that step transforms differences accumulated earlier. Define the subsequent propagation operator

$$P_{t\to T} = J_{T-1} J_{T-2} \cdots J_{t+1}, \qquad P_{T-1\to T} = I. \tag{4}$$

For a differentiable task score or score proxy M_τ that reads only the model-parameter coordinates of the training state, the final response has the local approximation

$$\Delta M_\tau \approx \sum_{t=0}^{T-1} \nabla M_\tau(w_T^0)^\top P_{t\to T} d_t, \qquad \Delta M_\tau = M_\tau(w_T^r) - M_\tau(w_T^0). \tag{5}$$

This expression separates three roles. The direct difference d_t determines whether exposure changes the current update; $P_{t\to T}$ describes whether later training preserves, amplifies, attenuates, or redirects that difference; and the score gradient identifies locally visible directions. A visible update can therefore yield a small terminal response if later training suppresses it, if it points mostly in a score-insensitive direction, or if stepwise contributions cancel.

Exposure count, update magnitude, and final score change therefore have no one-to-one relation. Post-training methods may differ in both exposure-induced direct updates and the subsequent evolution of earlier differences, while the task shapes training feedback and final scoring. Equation 5 imposes no algorithm ranking; any ranking remains task- and protocol-dependent.

The approximation is local, not a global convergence theorem. Fixed-decoding answer accuracy is generally non-differentiable and should not be substituted directly for M_τ; a differentiable expected score or proxy must be distinguished from the reported accuracy. Appendix A.2 states the smoothness assumptions and bounds the approximation error, while Appendix A.1 gives an exact finite-horizon decomposition for bounded, potentially non-smooth terminal scores.

3.2 When Behavioral Changes Affect Scores

After a training influence reaches the final model, behavioral change must still be distinguished from score change. For a fixed evaluation sample i, let a denote the relevant behavior, such as an answer or a selected visual region, and let π_i^r and π_i^0 be its distributions under the two conditions. If downstream processing is shared, define the bounded scoring map

$$q_i(a) = \mathbb{E}[\text{score} \mid i, a]. \tag{6}$$

The expected score difference satisfies

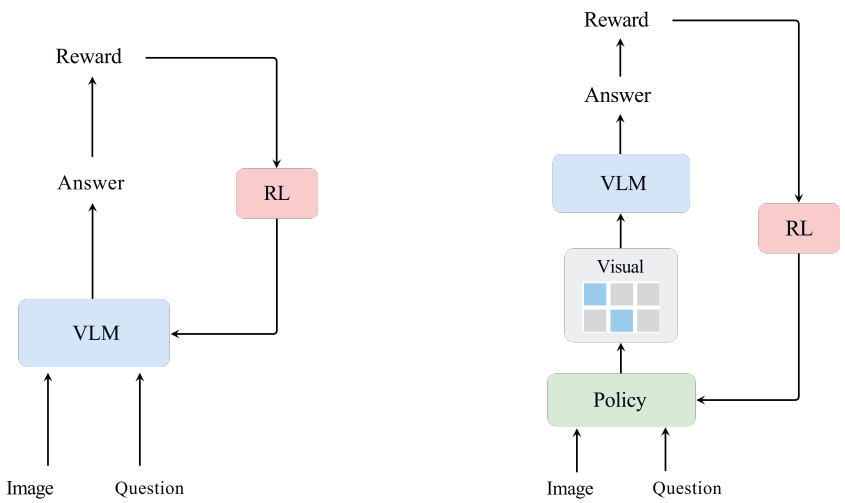
$$\delta_i = \int q_i(a)\, \mathrm{d}(\pi_i^r - \pi_i^0)(a), \qquad |\delta_i| \le \operatorname{osc}(q_i)\, \mathrm{TV}(\pi_i^r, \pi_i^0), \tag{7}$$

where $\operatorname{osc}(q_i) = \sup_a q_i(a) - \inf_a q_i(a)$ and $\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)|$.

Behavioral change need not change the score. A model may switch visual regions while both selections yield the same answer. Probability mass may also move among equally scored answers without producing any corresponding change in expected answer accuracy.

A small answer-level response thus cannot distinguish an almost unchanged policy from a changed policy whose consequences are invisible to the current scorer. This does not establish unelicited memory; it only shows that terminal scores do not uniquely determine what changed during training. If the downstream answer component also changes, the two conditions need not share the same q_i; Appendix A.3 gives the general two-term decomposition.

These relations permit several explanations for a weak exposure response but do not identify the mechanism in any experiment. We therefore compare tasks and training methods while separating theory-permitted explanations from facts empirically observed in these experiments. We use terminal scores to characterize response patterns and local variation, not to determine which theoretical link dominates within any particular experimental setting.



(a) Answer policy

(b) Visual selection policy

Figure 1: Reinforcement-learning processes used in this study. (a) In the answer-policy setting, the VLM maps an image–question pair directly to an answer, and answer-derived reward is returned through the RL objective to update the answer policy. (b) In the adaptive visual-selection setting, a policy first selects the visual information supplied to the VLM; the VLM produces the answer, and the resulting reward is returned through the RL objective to update the visual-selection policy.

4 Experiments

4.1 Experimental Setup

We study the effect of evaluation-data leakage in natural multimodal tasks and an adaptive visual-token task, comparing the responses of supervised fine-tuning, preference optimization, and reinforcement learning. The natural tasks cover chart understanding, text recognition, knowledge question answering, and visual reasoning. The adaptive visual-token task provides a second, controlled task setting. Together, the studies support two complementary comparisons: objectives within a setting and the response of a given objective across settings. They use the same combinations of models, training objectives, and overlap levels. Differences in task and protocol mean that the cross-setting contrast is not a single-factor causal test in this design.

The backbone panel comprises InternVL3.5, LLaVA-OneVision-7B, Qwen2.5-VL, and Qwen3-VL-8B. We compare SFT, DPO, PPO, GRPO, and RLOO across AI2D, ChartQA, DocVQA, MMBench-EN, MMMU, MMStar, MMVet, MathVista, OCRBench, and TextVQA. Complete model configurations, training settings, data processing, and scoring implementations are provided in the appendix and the accompanying reproducibility materials for complete specification.

For the policy-optimization objectives, the two reinforcement-learning feedback pathways are summarized in Figure 1. SFT and DPO instead retain their corresponding supervised and preference-learning objectives rather than using either RL feedback pathway.

We use three train–evaluation overlap levels: 0%, 50%, and 100%, indicating that none, half, or all of the fixed evaluation samples enter post-training. Different training conditions use the same fixed evaluation samples. We also use a holdout set that never enters post-training. For the natural tasks, we report accuracy on both evaluation sets and trace how it changes with overlap. For the adaptive visual-token task, we additionally use the zero-overlap condition as a baseline and subtract the holdout-set change from the change on the exposed samples:

$$\widehat{\Delta}_{\mathrm{DiD}}(r) = [\mathrm{Acc}_r(\mathcal{S}_r) - \mathrm{Acc}_0(\mathcal{S}_r)] - [\mathrm{Acc}_r(\mathcal{H}) - \mathrm{Acc}_0(\mathcal{H})]\,. \tag{8}$$

Here, $\mathcal{S}_r$ is the evaluation subset included in post-training at overlap level r, $\mathcal{H}$ is the holdout set, and subscripts r and 0 denote training with and without overlap. This quantity measures the

Table 1: Overlap responses across two task settings. Each objective occupies two rows, reporting the natural-task and adaptive visual-token results under aligned overlap conditions. Fixed, Holdout, Seen, and Unseen are macro-average accuracy percentages. Seen and Unseen refer to the exposed and unexposed subsets of the fixed evaluation set at 50% overlap. ΔFixed and ΔHoldout denote changes from 0% to 100% overlap, in percentage points. Paired response reports the difference-in-differences defined in the experimental setup and is shown only for the adaptive setting. Entries are three-seed means; standard deviations and complete disaggregated results are provided in the appendix. Differences are calculated before rounding. A dash means that the statistic is not reported for that setting, not that its value is zero.

Objective	Setting	0% overlap		50% overlap				100% overlap		100%−0%		Paired response	
		Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout	ΔFixed	ΔHoldout	50%	100%
SFT	Natural	61.5	57.3	73.7	57.6	84.3	63.0	82.7	58.0	**+21.2**	+0.7	—	—
	Adaptive	40.1	42.4	40.3	42.8	37.6	43.1	40.2	42.2	**+0.0**	−0.3	−0.2	+0.3
DPO	Natural	58.7	55.1	66.3	55.1	73.0	59.5	72.7	55.3	**+14.0**	+0.2	—	—
	Adaptive	42.7	44.4	42.5	43.7	40.4	44.6	44.0	44.7	**+1.3**	+0.3	+0.5	+0.9
PPO	Natural	58.2	54.3	66.2	54.3	72.7	59.8	72.4	54.9	**+14.2**	+0.7	—	—
	Adaptive	42.5	44.2	41.8	43.8	39.4	44.1	42.5	44.8	**+0.0**	+0.5	−0.2	−0.5
GRPO	Natural	57.6	53.8	59.8	53.7	60.5	59.1	61.2	53.1	**+3.7**	−0.7	—	—
	Adaptive	42.9	44.7	43.8	44.7	42.2	45.4	42.2	43.9	−0.7	−0.9	+0.9	+0.2
RLOO	Natural	57.0	53.3	59.1	53.1	59.8	58.4	60.7	53.0	**+3.7**	−0.3	—	—
	Adaptive	45.1	46.2	45.0	46.2	42.9	47.1	45.0	45.7	−0.1	−0.5	+0.0	+0.5

additional score change on exposed samples relative to the holdout set; it is not treated as a direct measure of memorization. We report both aggregate trends and model- and task-level results so that consistently weak responses can be distinguished from cancellation between positive and negative changes. Its statistical decomposition and interpretation conditions are given in Appendix A.4; protocol and execution details are provided with the reproducibility materials.

4.2 Exposure Responses Across Objectives and Task Settings

Table 1 summarizes overlap responses across both task settings. Using the common ΔFixed measure, limited average responses appear in both settings. Natural-task GRPO and RLOO each increase by $3.7\,\mathrm{pp}$, while all five adaptive-setting objectives change by only $-0.7\,\mathrm{pp}$ to $+1.3\,\mathrm{pp}$. Their holdout changes are also small. These are descriptive averages rather than a pooled "insensitive" category: the settings remain distinct, and individual model–benchmark responses vary.

The natural-task results provide a strong-response contrast: SFT, DPO, and PPO increase by $21.2\,\mathrm{pp}$, $14.0\,\mathrm{pp}$, and $14.2\,\mathrm{pp}$, respectively, from 0% to 100% overlap, while no objective improves holdout accuracy by more than $0.7\,\mathrm{pp}$. SFT's different response across settings shows that limited response is not specific to RL, and the spread among PPO, GRPO, and RLOO shows that policy optimization does not imply one response magnitude. Although the Fixed–Holdout separation is inconsistent with a uniform improvement equally visible on the holdout set, it does not identify the causal mechanism.

The adaptive paired response provides a secondary within-setting analysis. It subtracts the corresponding holdout change from the change on exposed samples, and all five objectives remain within $0.9\,\mathrm{pp}$ of zero at both overlap levels. Across objectives and overlap levels, these values range from $-0.5\,\mathrm{pp}$ to $+0.9\,\mathrm{pp}$; uncertainty estimates are reported in the appendix. Because this paired statistic differs from the raw accuracy changes used for the common cross-setting comparison, it is not directly equated with the natural-task $3.7\,\mathrm{pp}$ responses. Together, the results show that overlap alone does not determine the magnitude of the observed response here.

4.3 Variation Across Models and Benchmarks

Figure 2 expands the objective-level summary into the complete model–benchmark response map for both task settings. Each cell shows the change in reported mean accuracy at 50% or 100% overlap relative to the corresponding 0% condition. A common color scale retains every negative result and permits direct comparison across Fixed and Holdout sets. The natural-task Fixed block contains broad positive regions for SFT, DPO, and PPO, whereas the GRPO and RLOO blocks are more con-

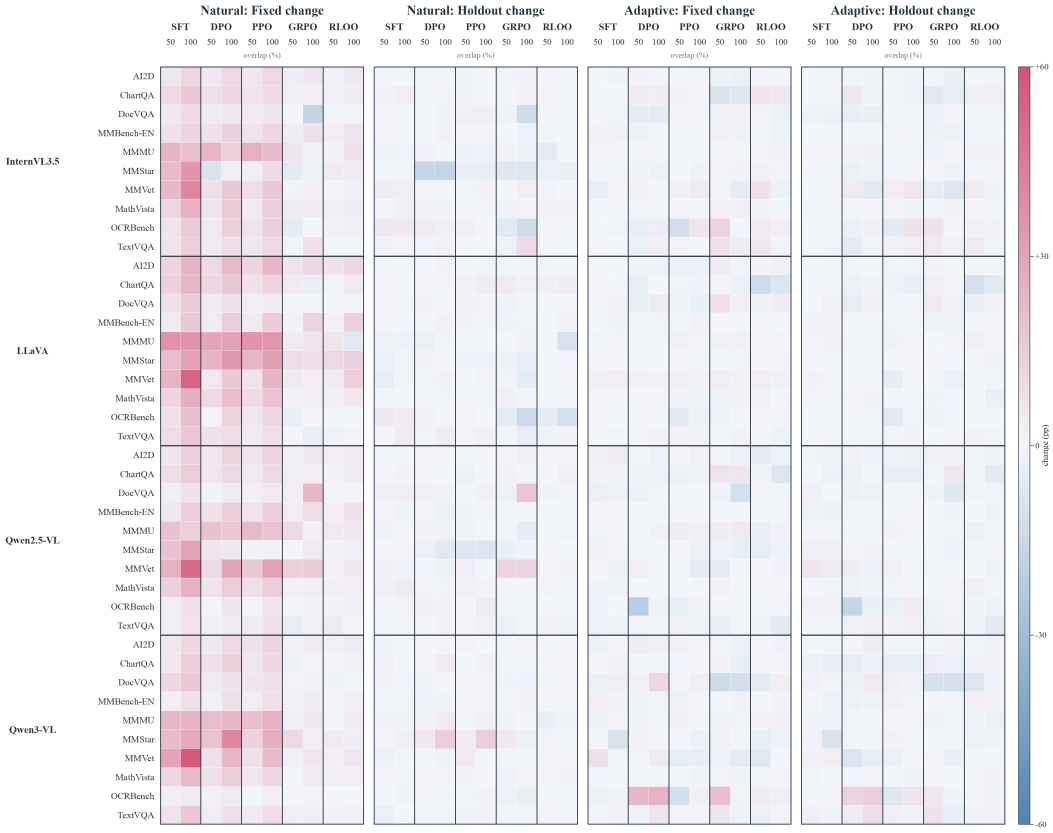


Figure 2: Complete overlap-response map. Rows enumerate all four backbones and ten benchmarks; columns enumerate all five objectives at 50% and 100% overlap. The four panels report Fixed and Holdout accuracy changes from the corresponding 0% condition for the natural and adaptive settings. All panels use the same color scale in percentage points. Blue indicates a decline and pink an improvement; values near zero are pale, and color intensity increases with the absolute response. Cells are differences between displayed means, with uncertainty reported in the appendix.

centrated around zero and include localized declines. Holdout and adaptive responses are generally smaller, but the disaggregated map also shows exceptions that are hidden by the macro averages.

In the Natural-task method-level macro averages, Seen accuracy exceeds Unseen accuracy at 50% overlap. However, because these subsets are defined by source order and may differ in difficulty, this pattern remains descriptive. Complete per-model and per-benchmark results with uncertainty estimates appear in Appendix A.5 for direct inspection and comparison.

The heatmap shows response magnitude cell by cell, while Figure 3 shows whether Fixed changes occur with corresponding Holdout movement across both task settings and overlap levels. Rows separate the Natural and Adaptive settings, while columns hold the objective fixed; each panel overlays the 50% and 100% responses using the two palettes shown in the legend. Each overlap level contains the same four backbones and ten benchmarks, plotted as changes from the corresponding 0% condition on common axes within each setting. Points near the diagonal indicate common movement; points above it indicate a larger Fixed response, and points near the origin indicate small changes on both sets. The Natural SFT, DPO, and PPO panels concentrate above the diagonal, whereas Natural GRPO and RLOO lie closer to the origin. Most Adaptive responses are smaller, although the disaggregated results retain several directional exceptions.

The pooled values do not describe every model equally well. Figure 4 shows model-level results before macro aggregation. Most estimates remain near zero relative to their across-seed dispersion, but some model–objective combinations are directional. For example, Qwen3-VL with DPO

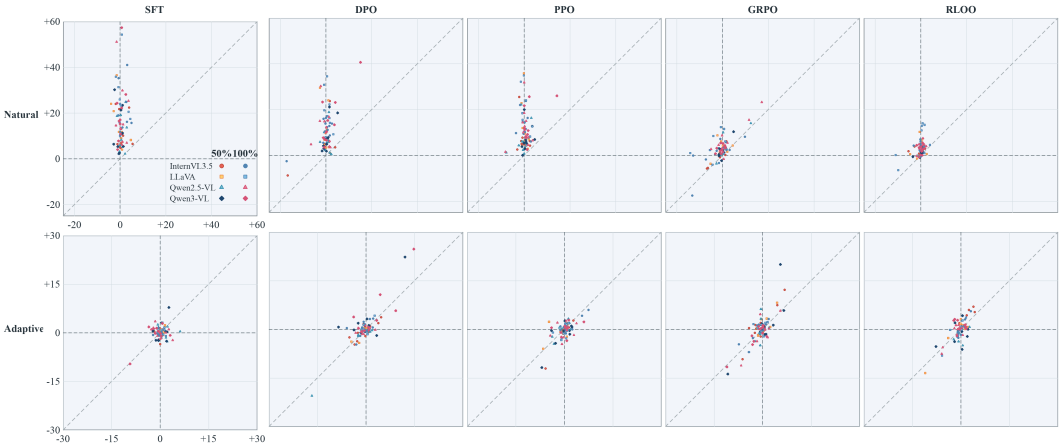


Figure 3: Fixed–Holdout responses across tasks and objectives. Points represent backbone–benchmark combinations at 50% and 100% overlap, showing Holdout changes horizontally and Fixed changes vertically relative to 0% overlap. Changes are measured in percentage points; axis ranges differ between tasks.

reaches $+2.7 \pm 0.7$ pp at 50% overlap and $+2.2 \pm 1.8$ pp at 100% overlap. Thus, the "within 0.9 pp" statement applies to the objective-level macro means in Table 1, not to every disaggregated result.

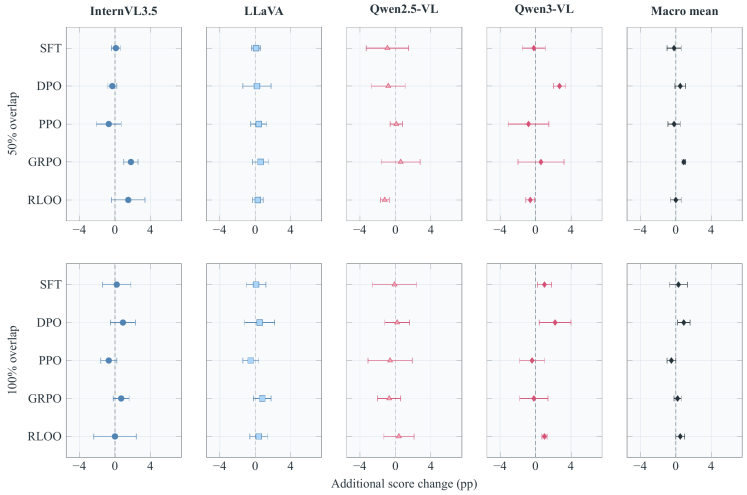


Figure 4: Adaptive visual-token additional score change by backbone and objective. Columns show four backbones and the macro mean; rows compare 50% and 100% overlap. Error bars show sample standard deviations across three training seeds.

Together, these results show why a small average must be read with its distribution: it may reflect uniformly small changes or cancellation across conditions. We retain the full model–benchmark map and cross-seed dispersion rather than rely on exceptions or a universal macro average.

5 Discussion

Limited average exposure responses occur both for GRPO and RLOO on the natural tasks and across objectives in the adaptive visual-token setting. We connect these results because their terminal response patterns are similar, not because they have been shown to share a learning mechanism. The stronger natural-task responses of SFT, DPO, and PPO, together with SFT's

different response across settings, show that leakage sensitivity must be interpreted for a specific task and implementation rather than inferred from an "SFT" or "RL" label.

The analysis in Section 3 connects this variation to three distinct links: exposure must change a current update, that influence must survive subsequent training, and the resulting behavior must affect the final score. A weak link or cancellation across steps can produce a small terminal response; the behavior–score relation further shows that answer accuracy may remain unchanged even when behavior changes. These are theory-permitted explanations, not mechanism estimates from our experiments. Terminal scores do not reveal whether a weak response arose from a small update difference, attenuation during later training, or limited score sensitivity, and they do not establish latent memory.

The conclusions are bounded by the tasks and implementations studied. In the natural-task landscape, training-set composition, sample order, and some decoding and answer-extraction settings differ across conditions, so the full between-method gap cannot be attributed to the optimization objective alone. Backbone–benchmark combinations are not independent replications, and the reported standard deviations describe seed variation rather than population uncertainty. Finally, answer-level primary outcomes do not directly characterize memory for layouts, actions, or interaction trajectories; further work should distinguish whether exposure-induced updates form and persist before asking whether final answers express them during evaluation.

6 Conclusion

We studied how evaluation-data exposure affects VLM post-training and found limited average responses both for natural-task GRPO and RLOO and across objectives in the adaptive visual-token setting, while natural-task SFT, DPO, and PPO responded more strongly. Local propagation analysis and an exact finite-horizon decomposition connect exposure-induced training differences, their subsequent influence, and final-score visibility, showing why training contact and evaluation benefit must be analyzed separately. These results shift the emphasis from assigning optimizers a fixed leakage sensitivity to understanding exposure response in each training setting, without treating similar terminal patterns as evidence of a common mechanism.

AI Use Statement

This research was conducted under the direction of the authors. Generative AI tools were used to assist with literature discovery; discussion and refinement of research hypotheses, methodology, and experimental design; development of the theoretical framing; mathematical derivations and proof drafts; method implementation and software-code editing and inspection; organization, analysis, and interpretation of results; preparation of tables and figures; and manuscript drafting, translation, and language editing. The authors defined the research question and scope, selected the experimental protocols and analyses, made all final methodological and interpretive decisions, and determined which AI-generated suggestions to accept, revise, or reject. All AI-assisted material included in the submission was subject to substantive human review and revision. This review covered theoretical assumptions and proof logic, experimental comparisons and statistical interpretation, and consistency among the text, equations, tables, and figures. AI outputs were treated as provisional assistance rather than as evidence. The authors take responsibility for the final manuscript, mathematical statements, experimental claims, citations, code, and research artifacts.

Ethics Statement

This study uses publicly available multimodal benchmarks and pretrained models without collecting primary human-subject data. The principal ethical consideration concerns evaluation validity: we explicitly document protocol limitations and do not present descriptive patterns from confounded conditions as causal evidence of model capabilities.

Reproducibility Statement

The appendix provides complete per-model and per-benchmark results, formal definitions of all estimands, aggregation procedures, and protocol boundaries. The supplementary materials will include experimental configurations, sample identifier manifests, scoring-function implementations,

cell-level metrics, and scripts sufficient to reproduce all reported tables. Model checkpoints and prediction outputs will be released subject to applicable model and benchmark licenses.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Michael Aerni, Javier Rando, Edoardo Debenedetti, Nicholas Carlini, Daphne Ippolito, and Florian Tramèr. Measuring non-adversarial reproduction of training data in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 54143–54185, 2025.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 67–93, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022a.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Juntao Dai, Taiye Chen, Yaodong Yang, Qian Zheng, and Gang Pan. Mitigating reward over-optimization in rlhf via behavior-supported regularization. In *International Conference on Learning Representations*, volume 2025, pp. 7527–7561, 2025.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 1286–1305, 2021.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1383–1392, 2018.

Sunny Duan, Mikail Khona, Abhiram Iyer, Rylan Schaeffer, and Ila Fiete. Uncovering latent memories in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 27339–27361, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Shahriar Golchin and Mihai Surdeanu. Time travel in llms: Tracing data contamination in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 43008–43029, 2024.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Sedrick Keh, Jean Mercat, Samir Yitzhak Gadre, Kushal Arora, Igor Vasiljevic, Benjamin Burchfiel, Shuran Song, Russ Tedrake, Thomas Kollar, Ludwig Schmidt, et al. Should vlms be pre-trained with image data? In *International Conference on Learning Representations*, volume 2025, pp. 20548–20572, 2025.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, 2022.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya H Jha, Oyvind Tafjord, Dustin Schwenk, Evan Walsh, Yanai Elazar, Kyle Lo, et al. Paloma: A benchmark for evaluating language model fit. *Advances in Neural Information Processing Systems*, 37:64338–64376, 2024.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208. IEEE, 2021.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations*, volume 2024, pp. 16354–16372, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Zehan Qi, Xiao Liu, Iat Long Iong, Hanyu Lai, Xueqiao Sun, Jiadai Sun, Xinyue Yang, Yu Yang, Shuntian Yao, Wei Xu, et al. Webrl: Training llm web agents via self-evolving online curriculum reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 79791–79821, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Mansi Sakarvadia, Aswathy Ajith, Arham Khan, Nathaniel Hudson, Caleb Geniesse, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W Mahoney. Mitigating memorization in language models. In *International Conference on Learning Representations*, volume 2025, pp. 91698–91741, 2025.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *Advances in Neural Information Processing Systems*, 38:150977–151013, 2026.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8309–8318. IEEE, 2019.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290, 2022.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. Livebench: A challenging, contamination-limited llm benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 91595–91631, 2025.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, Chengxiang Zhai, and Klara Nahrstedt. Vtool-r1: Vlms learn to think with images via reinforcement learning on multimodal tool use. In *International Conference on Learning Representations*, volume 2026, pp. 78298–78319, 2026.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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*, pp. 9556–9567, 2024.

A Supplementary Material

A.1 Formal Exposure Response and Exact Finite-Horizon Decomposition

Let Z_t^r and Z_t^0 be complete training states on a common measurable state space. A state contains every variable required to continue training, including model and optimizer parameters, reference models, data position, samplers, and random-number-generator state. The two conditions use the same finite horizon T, and their transitions are Markov kernels K_t^r and K_t^0 on that common space. For any bounded measurable terminal score F, define

$$\mathcal{R}_F = \mathbb{E}[F(Z_T^r)] - \mathbb{E}[F(Z_T^0)]. \tag{9}$$

This definition does not require a particular leakage measure. To connect it to the evaluation used in this paper, one possible terminal score is

$$F(z) = \mathrm{Acc}(z; \mathcal{S}_r) - \mathrm{Acc}(z; \mathcal{H}), \tag{10}$$

where $\mathcal{S}_r$ contains the exposed identities and $\mathcal{H}$ is a disjoint holdout set. The corresponding additional-score measure is defined in Section 4, and its interpretation conditions are given in Appendix A.4.

For a kernel K acting on a bounded function f, write

$$Kf(z) = \int f(z')\, K(z, \mathrm{d}z'). \tag{11}$$

Define the expected clean-continuation score by

$$V_T^0(z) = F(z), \qquad V_t^0(z) = K_t^0 V_{t+1}^0(z). \tag{12}$$

Let the initial distributions be μ_0^r and μ_0^0, and use $\mu f = \int f \,\mathrm{d}\mu$. The exact finite-horizon identity is

$$\mathcal{R}_F = (\mu_0^r - \mu_0^0)V_0^0 + \sum_{t=0}^{T-1} \mathbb{E}_{Z_t^r}\left[\big((K_t^r - K_t^0)V_{t+1}^0\big)(Z_t^r)\right]. \tag{13}$$

Under a common initial-state distribution, the first term is zero. If the initial distributions differ, that term must remain separate and cannot be attributed to exposure during later updates.

To verify the identity, telescope along the exposed trajectory:

$$\mathbb{E}[V_T^0(Z_T^r)] - \mathbb{E}[V_0^0(Z_0^r)] \tag{14}$$

$$= \sum_{t=0}^{T-1} \mathbb{E}\left[V_{t+1}^0(Z_{t+1}^r) - V_t^0(Z_t^r)\right] \tag{15}$$

$$= \sum_{t=0}^{T-1} \mathbb{E}_{Z_t^r}\left[\big((K_t^r - K_t^0)V_{t+1}^0\big)(Z_t^r)\right], \tag{16}$$

where the last equality uses $V_t^0 = K_t^0 V_{t+1}^0$. Adding $\mu_0^r V_0^0 - \mu_0^0 V_0^0$ gives Equation 13.

For

$$\beta_t(z) = \mathrm{TV}\big(K_t^r(z,\cdot), K_t^0(z,\cdot)\big), \tag{17}$$

with $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$ and $\mathrm{osc}(f) = \sup f - \inf f$, the standard inequality $|Pf - Qf| \le \mathrm{osc}(f)\,\mathrm{TV}(P,Q)$ gives, under common initialization,

$$|\mathcal{R}_F| \le \sum_{t=0}^{T-1} \mathbb{E}_{Z_t^r}\left[\beta_t(Z_t^r)\,\mathrm{osc}(V_{t+1}^0)\right]. \tag{18}$$

With unequal initial distributions, the absolute initial-state term must also be added to the bound. This is an upper bound, not a reverse identification result: a small terminal response does not imply that every transition difference is small.

The identity is exact for bounded, non-smooth terminal scores and requires no differentiability or convergence assumption. It shares the explanatory goal of the local analysis in Section 3, but it does not guarantee that the local approximation remains accurate for a large change such as 0%–100% exposure.

A.2 Local Linearization and Approximation Error

Assume that the continuous state w_t lies in a finite-dimensional normed space and includes both model parameters and all continuous optimizer state that affects later updates. Using the definitions in Section 3, the choice $d_t = U_t^r(w_t^r) - U_t^0(w_t^r)$ yields the exact identity

$$\Delta w_{t+1} = U_t^0(w_t^r) - U_t^0(w_t^0) + d_t. \tag{19}$$

Let $J_t = DU_t^0(w_t^0)$ and define

$$\rho_t = U_t^0(w_t^r) - U_t^0(w_t^0) - J_t \Delta w_t. \tag{20}$$

Then

$$\Delta w_{t+1} = J_t \Delta w_t + d_t + \rho_t. \tag{21}$$

If DU_t^0 is L_t-Lipschitz on a neighborhood containing the line segment between w_t^0 and w_t^r, Taylor's theorem gives

$$\|\rho_t\| \le \frac{L_t}{2} \|\Delta w_t\|^2. \tag{22}$$

Under common initialization, unrolling Equation 21 with the propagation operators from Equation 4 yields

$$\Delta w_T = \sum_{t=0}^{T-1} P_{t\to T} d_t + \sum_{t=0}^{T-1} P_{t\to T} \rho_t. \tag{23}$$

Suppose additionally that ∇M_τ is L_M-Lipschitz on a neighborhood containing the segment between w_T^0 and w_T^r. With a vector norm, its dual norm for gradients, and the corresponding induced operator norm for $P_{t\to T}$, the error in the main-text approximation obeys

$$\begin{aligned} &\left| \Delta M_\tau - \sum_{t=0}^{T-1} \nabla M_\tau(w_T^0)^\top P_{t\to T} d_t \right| \\ &\quad \le \frac{\|\nabla M_\tau(w_T^0)\|}{2} \sum_{t=0}^{T-1} \|P_{t\to T}\| L_t \|\Delta w_t\|^2 + \frac{L_M}{2} \|\Delta w_T\|^2. \end{aligned} \tag{24}$$

The approximation is therefore controlled only when trajectory differences remain local and later propagation does not excessively amplify the remainders. A local expansion is not an automatic guarantee for the full 0%–100% exposure contrast.

For optimizers with history, omitting the relevant optimizer state from w_t invalidates the one-step map. Discrete sampling, clipping boundaries, and non-smooth decoding can also prevent pathwise differentiation. Such cases are covered by the stochastic-kernel formulation in Appendix A.1; no differentiability of the realized training path is required there.

A.3 Behavioral Observability and Changing Downstream Components

Let π_i^r and π_i^0 be probability measures on the same measurable behavior space and let q_i be bounded and measurable. The expected score difference is the signed-measure integral in Equation 7. Applying $|Pf - Qf| \le \mathrm{osc}(f)\,\mathrm{TV}(P,Q)$ with $P = \pi_i^r$, $Q = \pi_i^0$, and $f = q_i$ proves the bound. The convention $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$ is the same as in the main text.

If downstream processing also differs between conditions, write the two scoring maps as q_i^r and q_i^0. Adding and subtracting $\int q_i^0 \,\mathrm{d}\pi_i^r$ gives the exact decomposition

$$\delta_i = \int q_i^0(a)\,\mathrm{d}(\pi_i^r - \pi_i^0)(a) + \int \left[q_i^r(a) - q_i^0(a)\right]\,\mathrm{d}\pi_i^r(a). \tag{25}$$

The first term compares behavior distributions on a shared clean scoring map; the second captures a change in downstream response itself. A common scoring map is an assumption of the main-text relation, not an empirically established property of the adaptive task. Small answer-level effects therefore cannot be used to infer that behavioral changes were hidden by the scorer.

A.4 Interpretation of the Additional Score Change

Equation 8 compares the change on exposed samples with the change on the holdout set. To state its interpretation boundary, let $d_r(i) = a_r(i) - a_0(i)$ denote the expected score change for evaluation sample i, and consider the conceptual decomposition

$$d_r(i) = g_r(i) + \mathbf{1}\{i \in \mathcal{S}_r\}\mu_r(i), \tag{26}$$

where $g_r(i)$ is a response not specific to exposure of sample i and $\mu_r(i)$ is the sample-specific component under this decomposition. Then

$$\Delta_{\mathrm{DiD}}(r) = \overline{\mu}_r(\mathcal{S}_r) + \underbrace{\overline{g}_r(\mathcal{S}_r) - \overline{g}_r(\mathcal{H})}_{\zeta_r}. \tag{27}$$

Thus, the additional score change equals the average sample-specific component only under the extra restriction $\zeta_r = 0$. Evaluating the same samples under the different training conditions removes direct sample-composition changes from each within-set contrast, but it does not establish that the remaining non-sample-specific response is equal on $\mathcal{S}_r$ and $\mathcal{H}$. The measure therefore describes an additional score change and is not, by itself, a pure measure of memorization.

A.5 Supplementary Tables

The following tables use one continuous appendix sequence. The natural-task landscape and adaptive visual-token task remain separate experiments, but the tables are not split into separate appendix parts or table namespaces. Complete per-model and per-benchmark results are included here so that the main text can retain only compact summaries.

A.5.1 NATURAL-TASK OVERLAP LANDSCAPE

Table 2: Complete per-model results for InternVL3.5. Fixed and Holdout accuracy are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics.

SFT

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	89.0±0.5	86.2±0.8	94.3±0.8	86.0±0.5	99.7±0.6	89.0±1.0	99.7±0.3	86.3±1.2
ChartQA	64.2±1.3	67.7±0.3	74.3±1.9	69.0±1.3	78.7±2.3	70.0±3.0	80.3±1.9	70.7±1.3
DocVQA	47.0±0.5	36.8±0.6	51.7±1.0	37.0±0.0	61.7±0.6	41.7±1.5	57.0±0.0	36.8±1.3
MMBench-EN	88.0±0.5	85.7±0.8	94.7±0.6	86.0±0.9	100.0±0.0	89.3±1.2	99.7±0.3	85.8±0.3
MMMU	43.0±15.7	36.7±18.8	67.0±18.4	34.8±19.5	90.4±7.9	43.7±29.7	62.0±30.7	35.9±18.0
MMStar	64.5±2.2	66.2±1.3	86.2±0.3	65.8±0.3	99.7±0.6	72.7±0.6	99.7±0.3	65.5±1.5
MMVet	45.0±3.6	36.4±3.6	67.4±0.0	40.3±2.7	81.8±0.0	52.4±0.0	86.0±0.0	39.5±2.3
MathVista	57.0±0.5	62.0±1.8	68.3±1.3	62.3±1.2	81.3±2.3	55.3±1.5	81.8±1.0	64.7±0.8
OCRBench	77.0±0.5	60.2±0.3	83.5±0.9	65.7±1.6	94.7±0.6	72.3±1.5	92.7±0.8	65.2±1.3
TextVQA	70.3±1.3	75.7±1.0	76.7±0.8	74.5±1.3	84.7±0.6	68.7±1.2	84.0±0.9	76.7±1.6

DPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	89.3±1.9	87.3±0.8	94.2±0.3	86.7±0.3	100.0±0.0	88.3±0.6	99.8±0.3	87.0±0.9
ChartQA	62.7±0.6	66.0±1.0	70.0±1.3	66.8±1.6	69.3±1.2	70.7±2.1	73.8±1.0	66.7±0.8
DocVQA	44.8±0.8	37.0±1.0	48.5±0.5	36.8±1.3	54.7±0.6	42.3±1.2	50.3±0.3	35.8±0.3
MMBench-EN	87.2±0.3	85.7±0.3	94.2±0.8	85.7±0.6	100.0±0.0	88.3±1.5	100.0±0.0	87.2±0.3
MMMU	39.4±18.3	33.3±20.2	63.1±20.7	35.0±20.8	83.3±17.5	43.0±27.5	53.3±39.0	33.1±20.4
MMStar	41.8±20.1	46.0±17.0	33.0±46.8	29.3±32.7	39.3±52.5	26.7±41.0	39.2±52.8	28.8±33.5
MMVet	42.6±1.3	43.4±3.6	50.4±1.3	43.4±3.6	45.5±0.0	55.6±2.7	58.9±1.3	43.4±1.3
MathVista	55.3±1.3	58.3±0.8	61.3±0.3	59.8±2.1	70.7±1.5	52.0±1.0	71.2±1.9	57.2±0.8
OCRBench	76.3±2.0	61.7±0.6	79.8±0.3	65.7±1.3	89.7±1.2	70.0±1.0	90.2±0.6	64.3±0.3
TextVQA	68.5±2.3	72.8±1.0	71.8±1.0	72.0±0.9	77.0±1.0	66.7±1.2	77.2±0.3	74.5±1.0

PPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	89.5±1.0	87.2±0.3	94.5±0.5	86.0±0.9	100.0±0.0	89.0±1.0	99.7±0.3	87.5±1.3
ChartQA	63.8±0.6	67.3±0.8	70.0±1.0	66.0±0.5	69.0±1.0	71.0±1.7	73.2±1.3	67.0±0.9
DocVQA	44.3±1.3	33.8±0.3	48.5±0.5	36.3±0.8	53.7±0.6	43.3±0.6	50.3±0.3	36.7±1.3
MMBench-EN	88.2±0.3	85.7±0.3	94.5±0.0	86.0±0.5	100.0±0.0	89.0±0.0	99.7±0.6	86.3±0.3
MMMU	40.4±18.4	36.1±18.4	65.9±19.0	33.9±21.2	86.7±10.6	45.2±28.2	62.4±30.2	34.3±19.4
MMStar	37.0±26.5	37.8±23.1	39.5±41.6	34.8±27.9	45.7±46.2	33.3±37.0	46.7±46.0	34.2±28.5
MMVet	41.1±2.7	41.9±2.3	48.8±0.0	41.1±1.3	45.5±0.0	52.4±0.0	57.4±1.3	43.4±2.7
MathVista	56.0±0.5	58.0±1.0	60.2±2.1	58.7±1.8	68.0±1.7	52.3±2.5	70.5±0.9	57.3±0.3
OCRBench	74.5±0.9	61.8±1.3	80.7±0.8	64.2±0.3	89.7±0.6	71.7±1.2	88.8±0.8	62.0±0.9
TextVQA	67.0±0.9	73.5±0.5	71.8±1.3	73.3±1.0	76.3±2.1	67.3±0.6	77.2±0.8	73.3±0.6

GRPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	88.5±0.0	85.5±0.9	91.7±0.3	86.3±1.5	95.3±0.6	88.0±0.0	94.2±1.0	87.0±0.5
ChartQA	63.2±2.8	66.3±1.3	66.0±0.5	66.8±0.6	61.0±0.0	71.0±1.0	65.5±1.3	68.0±1.0
DocVQA	49.8±1.2	38.5±1.0	50.0±0.0	39.7±2.0	53.3±1.5	46.7±1.5	32.2±27.9	25.3±22.0
MMBench-EN	87.3±0.8	86.3±0.3	90.5±1.0	86.3±0.6	93.7±1.2	87.3±1.2	94.5±0.5	85.3±0.6
MMMU	40.6±17.9	33.9±21.2	46.7±26.4	35.6±18.8	49.3±24.7	44.1±28.6	42.6±32.4	31.3±20.6
MMStar	47.3±16.3	46.3±15.0	41.5±30.0	39.5±22.7	41.0±33.9	42.0±26.1	45.5±29.4	39.0±20.0
MMVet	46.5±2.3	42.6±4.8	47.3±1.3	42.6±5.9	39.4±2.6	55.6±5.5	48.8±6.2	45.7±1.3
MathVista	51.7±2.9	54.8±0.3	55.2±2.8	56.2±3.2	58.3±2.1	52.0±3.5	55.3±4.2	55.5±2.2
OCRBench	81.3±0.8	64.7±0.3	75.8±10.9	58.3±5.4	79.0±13.5	72.7±8.4	81.2±4.3	51.8±2.3
TextVQA	68.2±1.0	71.2±0.6	67.3±2.3	70.0±1.3	70.7±3.5	64.0±1.0	76.3±1.3	81.0±0.9

RLOO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	89.2±1.5	86.8±0.3	91.0±0.5	87.0±0.5	94.7±0.6	87.3±1.5	94.0±0.5	86.7±0.3
ChartQA	64.5±0.0	66.0±1.3	66.5±0.0	66.7±0.8	62.7±1.2	70.3±1.2	68.0±0.5	66.3±0.6
DocVQA	47.0±1.0	36.3±0.6	46.7±0.3	36.5±0.0	49.7±0.6	43.7±0.6	47.2±0.8	36.0±0.5
MMBench-EN	87.7±0.3	86.3±0.3	90.7±0.8	86.0±0.0	93.0±1.7	88.3±0.6	93.8±1.0	86.2±0.3
MMMU	36.5±24.3	34.3±20.7	38.3±35.5	28.9±24.3	35.6±34.2	41.1±36.7	43.5±32.6	34.1±20.0
MMStar	33.8±27.0	38.8±22.2	37.7±33.6	36.0±23.8	39.0±38.1	36.3±29.2	38.0±36.6	35.7±25.2
MMVet	48.1±3.6	43.4±1.3	48.8±2.3	42.6±1.3	36.4±0.0	61.9±4.8	50.4±1.3	43.4±1.3
MathVista	52.8±0.8	53.3±0.8	55.7±1.4	55.2±0.8	59.7±1.5	51.7±1.5	56.5±1.0	54.7±1.3
OCRBench	81.3±0.3	64.2±0.8	83.3±0.3	64.7±0.3	89.7±0.6	77.0±0.0	83.5±1.8	63.7±1.5
TextVQA	67.0±1.3	71.8±1.6	67.2±1.2	71.5±0.0	67.7±0.6	66.7±2.1	66.5±0.9	71.8±0.8

Table 3: Complete per-model results for LLaVA-OneVision-7B. Fixed and Holdout accuracy are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics.

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
SFT								
AI2D	75.3±0.6	84.7±0.3	87.0±0.9	85.5±0.5	99.0±0.0	75.0±1.7	99.2±0.3	84.7±0.6
ChartQA	53.0±0.9	49.5±0.9	66.2±1.8	50.0±1.3	70.7±0.6	61.7±3.5	76.3±1.5	49.8±0.3
DocVQA	33.7±0.3	26.5±1.3	41.0±0.0	27.5±0.5	50.7±1.5	31.3±1.5	49.5±1.3	26.3±0.3
MMBench-EN	83.8±0.3	86.8±0.6	87.8±0.6	87.7±0.6	100.0±0.0	75.7±1.2	100.0±0.0	86.5±1.3
MMMU	32.2±15.5	30.6±15.0	68.7±11.8	29.1±12.4	97.0±1.7	40.4±21.9	68.0±23.6	28.7±14.6
MMStar	65.5±1.3	61.0±0.5	86.3±0.8	58.2±1.2	94.7±0.6	78.0±1.0	96.8±1.0	60.3±0.8
MMVet	31.8±1.3	18.6±8.4	55.8±2.3	14.7±1.3	81.8±0.0	28.6±4.8	86.0±0.0	19.4±2.7
MathVista	49.3±1.4	47.5±0.9	61.0±1.7	47.0±2.6	74.0±2.0	48.0±2.0	75.2±2.3	47.3±1.2
OCRBench	74.8±1.6	42.5±3.6	83.0±0.9	47.2±0.8	96.0±1.7	70.0±0.0	95.3±0.3	46.2±1.2
TextVQA	62.8±0.3	68.8±1.3	71.8±1.0	69.3±0.8	78.0±0.0	65.7±2.1	80.2±0.3	73.3±0.3
DPO								
AI2D	77.3±0.6	85.2±0.3	87.7±1.0	84.8±0.8	99.0±0.0	76.3±2.1	99.5±0.5	86.7±0.3
ChartQA	50.2±0.6	50.2±1.3	61.5±0.0	50.7±0.6	63.7±0.6	59.3±0.6	67.3±1.0	50.8±0.8
DocVQA	32.7±0.8	26.0±0.5	36.2±0.8	27.2±1.3	41.3±1.5	31.0±1.0	40.7±0.8	25.8±0.3
MMBench-EN	85.8±0.6	86.8±0.8	89.0±0.5	86.7±0.6	100.0±0.0	78.0±1.0	100.0±0.0	88.5±0.5
MMMU	36.1±11.6	30.6±14.1	65.6±14.5	28.0±12.9	92.6±5.7	38.5±23.5	66.9±25.4	29.8±12.7
MMStar	63.8±1.4	59.7±2.0	87.8±0.3	60.8±1.2	96.7±0.6	79.0±0.0	98.3±0.6	60.2±1.8
MMVet	31.8±2.7	28.7±2.7	36.4±1.3	27.9±2.3	39.4±2.6	33.3±0.0	48.1±2.7	30.2±4.7
MathVista	48.0±1.3	43.7±2.0	57.8±1.6	43.5±0.9	69.3±1.5	46.3±2.1	67.7±0.6	43.0±1.8
OCRBench	78.2±1.0	45.5±0.5	79.2±1.5	47.3±0.3	92.7±0.6	65.7±2.5	90.5±0.9	46.0±1.3
TextVQA	62.5±0.5	67.3±0.3	69.2±0.8	66.7±1.6	72.0±1.0	66.3±2.3	71.7±1.3	70.5±0.5
PPO								
AI2D	75.8±0.3	85.0±0.9	87.7±0.8	84.3±1.5	99.3±0.6	76.0±1.7	99.5±0.0	85.7±1.5
ChartQA	53.2±1.2	45.7±1.0	60.2±0.8	47.7±1.0	60.3±1.2	60.0±1.0	65.8±0.8	49.2±0.8
DocVQA	33.3±1.9	25.0±1.3	36.2±2.0	26.5±0.5	41.7±2.1	30.7±2.1	37.7±1.0	26.0±0.9
MMBench-EN	84.3±0.3	87.5±0.0	88.3±0.8	86.5±1.3	100.0±0.0	76.7±1.5	99.8±0.3	86.8±0.3
MMMU	34.4±15.4	29.6±14.8	70.2±10.0	29.4±14.2	97.4±1.7	43.0±18.6	69.3±21.8	29.4±13.5
MMStar	64.2±0.8	60.2±1.0	87.0±0.9	59.0±1.8	95.0±1.0	79.0±1.0	96.3±0.3	58.7±0.8
MMVet	31.0±1.3	25.6±2.3	37.2±2.3	25.6±2.3	42.4±2.6	31.7±5.5	55.0±1.3	26.4±1.3
MathVista	47.3±1.5	42.0±1.3	56.8±1.3	42.3±1.8	67.7±1.5	46.0±3.5	65.7±0.8	41.2±1.3
OCRBench	76.3±1.0	46.2±1.0	81.0±0.5	47.0±1.5	91.7±0.6	70.3±1.2	87.5±0.0	46.3±2.5
TextVQA	63.3±0.8	68.2±0.6	67.5±0.9	69.0±0.9	69.7±0.6	65.3±1.2	71.5±1.5	70.8±2.1
GRPO								
AI2D	78.2±1.5	85.3±1.8	84.3±0.8	84.7±1.4	91.0±0.0	77.7±1.5	89.7±1.4	86.0±0.9
ChartQA	47.7±0.6	40.8±1.3	52.0±1.3	45.3±1.5	48.3±2.1	55.7±1.5	44.3±12.1	39.2±5.8
DocVQA	34.2±0.3	23.3±2.6	34.8±3.3	21.8±5.5	36.3±1.5	33.3±5.5	34.8±2.9	23.8±1.8
MMBench-EN	83.5±0.5	86.2±0.8	85.0±1.3	85.0±1.3	96.7±1.5	73.3±1.2	95.8±1.6	85.5±1.3
MMMU	35.0±15.4	31.9±11.9	39.3±22.8	32.0±12.2	40.0±23.1	38.5±22.5	41.1±26.0	33.1±14.1
MMStar	62.8±2.4	59.2±1.2	71.5±5.1	57.0±2.8	68.7±8.3	74.3±2.1	71.2±8.3	55.2±2.4
MMVet	34.9±2.3	26.4±1.3	38.8±2.7	27.1±4.8	39.4±2.6	38.1±4.8	37.2±2.3	20.9±2.3
MathVista	40.8±1.9	41.2±1.5	42.3±1.3	43.3±1.5	43.3±4.0	41.3±1.5	43.8±3.1	41.7±1.9
OCRBench	82.5±0.5	47.3±1.2	78.8±1.9	42.0±3.3	87.3±3.2	70.3±1.2	83.5±0.9	33.2±1.5
TextVQA	60.3±0.6	62.2±1.3	60.8±1.8	61.7±3.1	59.0±2.6	62.7±1.2	56.2±3.3	57.8±5.1
RLOO								
AI2D	77.3±0.8	85.2±0.3	84.2±0.8	85.8±0.8	91.7±2.1	76.7±0.6	88.2±1.5	85.5±0.5
ChartQA	50.2±1.3	45.2±4.5	50.8±1.8	47.8±1.8	48.0±1.0	53.7±3.1	54.8±1.8	49.2±0.3
DocVQA	36.8±0.8	25.8±1.5	36.7±1.5	25.3±1.3	38.0±1.7	35.3±1.5	36.0±0.5	25.0±1.3
MMBench-EN	83.0±1.0	84.7±0.6	84.7±0.3	84.7±0.8	96.0±1.0	73.3±0.6	96.3±0.8	86.3±1.2
MMMU	34.8±16.1	31.9±13.8	40.7±23.9	32.2±13.0	40.4±25.7	41.1±22.3	28.3±41.4	22.0±23.5
MMStar	61.8±1.0	57.3±2.3	72.0±4.5	56.8±2.1	69.7±8.6	74.3±1.5	74.7±5.5	57.7±0.8
MMVet	31.0±1.3	31.0±1.3	34.9±2.3	31.0±1.3	39.4±2.6	30.2±2.7	45.0±1.3	31.8±2.7
MathVista	44.5±2.8	44.3±0.3	47.2±1.2	44.0±0.5	49.0±3.6	45.3±2.5	49.5±2.6	43.8±0.3
OCRBench	82.2±0.6	52.0±1.0	82.7±0.8	47.3±2.1	89.3±0.6	76.0±1.0	82.3±0.8	41.0±6.1
TextVQA	58.7±1.8	61.0±3.0	57.3±4.6	59.3±6.5	57.7±4.2	57.0±5.2	59.7±1.8	62.5±3.1

Table 4: Complete per-model results for Qwen2.5-VL. Fixed and Holdout accuracy are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics. The GRPO–OCRBench entry uses the accepted three-seed rerun.

SFT

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	86.5±0.9	86.3±0.3	92.3±0.8	86.2±1.8	99.0±0.0	85.7±1.5	99.5±0.5	86.7±0.6
ChartQA	64.7±1.3	66.2±1.4	73.0±1.0	66.7±0.3	75.3±1.5	70.7±0.6	79.7±0.3	67.5±0.5
DocVQA	49.5±0.9	38.8±1.0	51.7±0.3	41.0±0.5	57.7±0.6	45.7±0.6	56.5±1.3	41.7±1.6
MMBench-EN	85.5±0.9	80.5±1.5	90.7±0.3	81.5±0.5	100.0±0.0	81.3±0.6	100.0±0.0	78.7±1.0
MMMU	41.3±14.8	30.4±15.6	60.0±17.4	29.1±17.7	76.7±13.9	43.3±21.2	54.6±32.6	30.6±17.3
MMStar	63.2±1.3	61.0±1.3	82.2±0.6	61.3±2.8	89.3±1.2	75.0±0.0	93.0±1.8	62.0±0.5
MMVet	34.9±6.2	39.5±2.3	59.7±3.6	41.9±4.0	81.8±0.0	36.5±7.3	86.0±0.0	38.0±1.3
MathVista	52.5±0.5	56.3±1.6	66.2±1.0	59.2±2.1	76.7±0.6	55.7±2.3	77.8±1.0	60.0±2.6
OCRBench	86.3±0.6	66.7±1.6	89.2±0.8	66.3±0.8	94.0±0.0	84.3±1.5	93.7±1.5	67.7±0.3
TextVQA	79.7±1.0	78.5±0.9	82.8±1.5	78.3±0.8	86.0±1.7	79.7±2.1	85.8±1.0	79.3±1.0

DPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	87.8±1.0	87.3±1.2	93.2±0.6	87.7±0.3	99.7±0.6	86.7±0.6	100.0±0.0	87.5±0.0
ChartQA	64.3±1.6	65.0±1.7	69.7±4.3	65.5±1.7	69.3±6.7	70.0±2.0	74.3±4.5	65.0±0.9
DocVQA	48.7±2.0	38.5±1.7	49.8±4.0	40.2±0.3	54.7±5.5	45.0±2.6	52.2±2.9	40.5±2.0
MMBench-EN	85.0±1.3	81.0±1.0	91.8±0.6	81.8±0.6	100.0±0.0	83.7±1.2	100.0±0.0	78.3±1.3
MMMU	39.1±12.9	28.9±15.9	58.0±20.1	30.2±15.4	70.0±23.6	45.9±18.8	55.9±32.9	30.2±16.7
MMStar	29.7±31.0	30.3±28.3	35.5±43.7	27.7±29.0	41.0±46.8	30.0±40.7	34.5±52.0	23.8±30.0
MMVet	28.7±2.7	38.0±1.3	38.8±4.8	37.2±2.3	51.5±6.9	25.4±2.7	58.9±5.4	35.7±3.6
MathVista	53.5±0.5	52.5±1.5	59.3±0.6	53.5±1.0	64.7±2.5	54.0±1.7	68.3±1.0	51.8±1.0
OCRBench	83.8±1.5	59.7±5.9	84.2±2.9	62.0±4.8	91.7±2.9	76.7±3.1	90.0±4.8	61.3±2.8
TextVQA	72.0±5.2	75.0±1.8	74.2±6.4	74.7±3.3	75.0±7.8	73.3±4.9	75.8±7.5	76.5±1.8

PPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	86.7±1.5	86.0±0.5	92.5±0.5	85.7±1.3	99.0±0.0	86.0±1.0	99.3±0.3	86.2±1.0
ChartQA	63.0±0.9	64.7±2.8	67.3±5.3	62.3±2.4	68.0±7.2	66.7±5.1	73.0±5.2	64.3±3.3
DocVQA	47.5±0.9	38.5±0.0	49.8±2.5	38.0±2.2	55.0±3.0	44.7±2.1	52.0±2.6	40.2±1.4
MMBench-EN	85.7±0.8	81.2±0.3	91.8±1.0	80.2±0.3	100.0±0.0	83.7±2.1	100.0±0.0	80.5±0.9
MMMU	37.6±16.1	30.6±15.4	59.8±15.7	29.3±18.8	77.0±12.4	42.6±19.0	56.1±31.3	30.6±18.0
MMStar	38.8±22.7	36.2±22.4	40.0±38.7	28.3±29.3	48.7±38.2	31.3±39.6	40.3±43.6	28.0±31.2
MMVet	24.8±1.3	36.4±3.6	41.1±5.9	39.5±4.0	50.0±9.1	31.7±2.7	56.6±1.3	36.4±1.3
MathVista	51.5±1.3	51.7±2.8	59.2±2.1	53.5±0.0	64.7±3.8	53.7±0.6	66.5±1.3	52.5±1.3
OCRBench	81.0±0.5	59.8±2.4	81.7±1.0	59.8±1.9	87.7±0.6	75.7±1.5	87.5±4.8	63.5±3.9
TextVQA	70.3±10.3	73.2±4.4	73.8±9.0	73.5±4.8	74.3±10.1	73.3±8.1	78.2±7.7	75.0±3.1

GRPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	87.3±1.0	84.3±0.3	90.2±1.0	85.3±1.0	94.7±1.5	85.7±1.2	92.7±0.8	86.5±0.5
ChartQA	56.5±4.8	59.7±1.0	59.3±1.8	61.5±0.9	54.0±0.0	64.7±3.5	59.5±3.5	59.5±0.9
DocVQA	47.0±0.5	42.0±1.0	50.0±0.9	40.3±2.0	52.7±0.6	47.3±1.2	70.2±12.3	59.2±11.4
MMBench-EN	86.5±0.5	80.5±0.0	89.5±1.8	80.8±0.3	95.7±1.5	83.3±4.0	95.2±0.3	80.3±0.6
MMMU	31.9±25.4	30.6±17.2	41.9±22.2	30.9±13.3	42.2±26.0	41.5±18.8	34.6±34.0	25.6±23.4
MMStar	23.0±35.9	25.5±33.5	23.5±40.7	22.3±32.2	22.3±38.7	24.7±42.7	27.5±42.1	23.8±33.5
MMVet	21.7±18.9	25.6±20.3	35.7±3.6	38.0±3.6	33.3±2.6	38.1±4.8	37.2±4.0	37.2±2.3
MathVista	50.0±0.9	53.2±1.5	52.0±2.0	53.0±1.0	49.0±3.6	55.0±1.0	53.0±1.8	53.3±0.3
OCRBench	83.0±0.0	73.0±0.5	83.3±0.3	72.3±0.8	86.7±0.6	80.0±0.0	84.5±0.5	73.2±0.3
TextVQA	67.8±5.8	68.0±4.3	64.2±5.0	65.3±5.7	64.0±5.6	64.3±4.7	66.2±1.6	68.5±2.0

RLOO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	87.2±0.6	84.5±1.3	89.3±0.6	85.0±0.9	94.3±1.5	84.3±0.6	91.7±1.3	85.3±0.8
ChartQA	58.3±0.6	61.2±0.3	59.8±1.3	61.5±1.5	55.7±2.1	64.0±2.6	62.2±1.0	63.5±0.5
DocVQA	49.2±1.4	42.0±0.5	49.0±0.9	40.7±1.0	51.0±1.0	47.0±1.0	48.3±1.5	41.2±1.0
MMBench-EN	86.8±0.8	80.8±0.3	89.8±0.6	80.8±0.3	97.0±2.0	82.7±1.2	94.2±1.9	81.7±0.6
MMMU	20.6±32.2	18.0±27.7	25.0±39.5	18.0±27.7	25.6±42.3	24.4±36.6	25.4±40.1	17.2±26.9
MMStar	23.2±36.7	23.5±33.3	24.8±41.3	23.7±33.7	25.0±40.7	24.7±41.9	26.0±43.7	23.5±33.8
MMVet	33.3±1.3	34.9±0.0	34.9±0.0	35.7±1.3	36.4±0.0	33.3±0.0	38.0±1.3	36.4±1.3
MathVista	52.2±0.3	53.7±0.3	54.3±1.0	54.3±1.5	55.0±2.6	53.7±0.6	55.2±0.6	53.0±2.0
OCRBench	81.3±0.3	63.3±0.8	82.7±0.3	63.3±2.1	84.0±1.0	81.3±1.2	83.7±1.5	64.0±1.5
TextVQA	59.3±6.6	68.0±4.4	63.8±4.2	69.3±6.0	65.3±3.8	62.3±4.6	59.0±6.5	66.2±3.5

Table 5: Complete per-model results for Qwen3-VL-8B. Fixed and Holdout accuracy are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics.

SFT

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	88.5±0.9	90.5±1.8	93.8±0.6	89.5±1.0	100.0±0.0	87.7±1.2	99.3±0.3	90.7±1.0
ChartQA	60.8±0.3	48.8±1.9	66.2±1.6	50.2±1.8	68.3±3.8	64.0±3.6	76.0±1.3	48.8±1.9
DocVQA	21.8±1.3	17.7±1.6	33.0±1.0	18.8±0.3	37.3±1.2	28.7±2.3	38.3±1.9	17.3±0.8
MMBench-EN	92.2±1.0	89.0±0.9	94.7±0.3	88.7±0.6	100.0±0.0	89.3±0.6	100.0±0.0	88.3±0.8
MMMU	38.5±15.8	33.9±17.9	61.9±15.0	35.0±16.0	84.4±10.9	39.3±20.9	62.8±26.9	32.6±16.1
MMStar	70.2±1.8	64.8±1.0	91.5±0.9	65.2±2.8	98.7±1.5	84.3±0.6	98.3±0.8	67.3±1.2
MMVet	28.7±2.7	26.4±1.3	58.9±3.6	24.0±1.3	81.8±0.0	34.9±7.3	86.0±0.0	27.1±1.3
MathVista	54.5±3.0	57.5±1.0	65.2±1.8	57.8±0.8	74.7±1.2	55.7±2.5	76.7±1.8	58.2±1.6
OCRBench	92.8±0.6	64.8±0.8	95.0±0.0	64.3±0.6	98.0±0.0	92.0±0.0	96.8±0.3	63.7±1.3
TextVQA	61.3±1.8	62.5±1.3	67.7±1.3	59.7±1.2	78.3±2.1	57.0±1.0	78.3±0.3	62.5±3.3

DPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	89.0±0.5	90.0±0.5	93.8±0.6	90.3±0.6	99.7±0.6	88.0±1.0	99.3±0.3	90.3±1.3
ChartQA	59.5±4.1	52.0±9.1	67.3±4.5	53.0±8.7	66.3±9.2	68.3±1.5	72.8±5.3	55.7±8.5
DocVQA	30.7±10.8	20.8±8.0	34.8±12.3	19.7±7.2	36.7±16.8	33.0±7.8	37.3±11.1	21.0±8.7
MMBench-EN	92.3±0.8	89.2±0.6	94.2±0.8	87.5±1.3	100.0±0.0	88.3±1.5	100.0±0.0	88.0±1.3
MMMU	34.1±19.6	29.8±20.4	55.2±25.9	31.3±20.3	75.6±20.3	34.8±31.5	57.2±32.3	33.9±17.8
MMStar	40.7±24.1	40.7±21.6	59.2±28.8	45.8±19.2	56.7±38.0	61.7±19.7	81.3±17.3	55.8±12.8
MMVet	30.2±2.3	30.2±0.0	38.0±1.3	30.2±2.3	36.4±0.0	39.7±2.7	53.5±2.3	28.7±1.3
MathVista	52.0±0.5	55.0±0.5	59.5±1.3	55.3±0.3	64.3±0.6	54.7±2.3	67.2±0.8	55.8±0.6
OCRBench	91.5±0.5	65.3±0.6	93.3±1.0	65.5±2.3	97.3±1.5	89.3±0.6	95.0±0.5	64.5±0.9
TextVQA	60.8±2.8	59.7±2.0	63.8±0.3	58.8±1.3	73.3±1.2	54.3±1.5	68.0±2.3	59.0±0.9

PPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	87.5±1.8	90.5±0.5	94.2±0.6	91.2±0.8	99.7±0.6	88.7±0.6	99.3±0.3	90.5±0.5
ChartQA	57.0±1.3	48.3±1.6	62.3±1.8	48.2±3.3	58.7±1.5	66.0±2.0	67.8±1.3	50.0±0.9
DocVQA	23.7±0.8	15.7±1.5	29.5±1.8	16.5±2.0	30.0±2.6	29.0±2.0	31.3±0.6	16.5±1.0
MMBench-EN	91.8±0.8	88.5±0.5	94.7±0.3	88.8±0.3	100.0±0.0	89.3±0.6	100.0±0.0	89.2±0.8
MMMU	37.4±20.0	30.9±21.4	57.4±21.3	30.9±21.4	80.0±16.7	34.8±27.6	63.0±27.0	33.1±18.9
MMStar	38.5±25.1	34.7±26.7	51.2±33.8	38.2±24.9	47.0±44.4	55.3±23.2	64.5±30.5	49.0±20.9
MMVet	32.6±4.0	28.7±1.3	39.5±4.0	33.3±1.3	37.9±5.2	41.3±2.7	54.3±1.3	28.7±3.6
MathVista	56.0±2.0	54.8±1.8	61.0±0.9	56.0±2.8	65.3±1.5	56.7±0.6	67.2±2.0	55.0±0.5
OCRBench	92.2±0.8	63.5±0.5	92.2±1.3	62.8±1.0	94.3±1.5	90.0±1.0	94.7±0.8	64.8±0.3
TextVQA	60.2±2.3	59.8±1.8	64.0±1.3	58.8±0.3	71.3±4.2	56.7±1.5	69.3±1.8	61.2±1.2

GRPO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	88.3±1.2	90.0±2.3	90.2±1.4	90.2±2.3	94.3±1.2	86.0±1.7	92.3±1.0	91.0±1.0
ChartQA	59.0±1.0	48.7±0.8	56.7±0.6	48.3±1.0	50.7±3.1	62.7±3.1	59.7±1.4	48.0±0.9
DocVQA	23.8±2.0	18.2±1.2	24.5±1.8	17.0±1.0	21.7±1.2	27.3±2.5	25.8±1.9	17.2±1.5
MMBench-EN	91.0±0.9	87.7±0.6	92.2±0.6	88.2±1.0	96.3±0.6	88.0±1.0	94.2±1.0	89.0±1.0
MMMU	36.7±18.9	32.0±18.0	39.6±27.7	30.7±20.5	40.0±28.9	39.3±26.6	41.9±28.1	32.2±18.4
MMStar	39.7±24.2	37.5±24.7	50.0±23.8	42.3±14.6	39.0±30.3	61.0±17.4	42.8±27.9	35.0±24.2
MMVet	27.9±4.0	24.8±2.7	30.2±4.7	24.0±1.3	25.8±6.9	34.9±2.7	33.3±1.3	24.0±1.3
MathVista	44.5±3.1	48.3±2.8	44.2±5.7	46.8±6.0	41.3±7.8	47.0±3.6	48.8±4.9	49.3±2.5
OCRBench	92.5±0.0	67.3±1.3	92.8±0.3	65.7±1.2	97.0±0.0	88.7±0.6	93.2±0.3	64.2±0.8
TextVQA	59.8±4.8	62.8±5.5	62.2±8.5	62.2±7.3	67.7±4.7	56.7±12.7	62.0±5.0	62.5±5.8

RLOO

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
AI2D	88.5±0.5	90.8±0.6	90.2±0.8	90.8±0.8	93.7±0.6	86.7±1.5	92.2±0.3	91.0±1.3
ChartQA	59.5±0.5	49.3±1.3	60.3±0.8	49.7±0.8	57.3±0.6	63.3±1.2	61.2±1.5	49.7±1.8
DocVQA	26.3±1.0	17.2±0.3	25.3±0.8	16.8±1.3	22.7±0.6	28.0±1.0	26.3±0.6	17.8±0.6
MMBench-EN	91.8±0.3	87.8±0.8	92.7±0.8	88.0±0.5	97.3±0.6	88.0±1.0	94.3±0.8	88.5±1.0
MMMU	39.3±16.5	32.2±19.3	38.7±26.1	28.7±19.9	33.0±29.2	44.4±23.1	43.7±26.8	30.4±17.5
MMStar	36.8±27.0	33.5±26.4	40.3±30.5	34.2±25.4	30.3±33.5	50.3±27.4	40.2±31.9	32.8±25.3
MMVet	27.9±0.0	25.6±0.0	30.2±4.7	26.4±1.3	19.7±6.9	41.3±2.7	33.3±1.3	26.4±1.3
MathVista	41.3±2.3	43.8±1.4	43.7±1.3	44.5±1.5	39.0±2.0	48.3±0.6	44.3±2.8	45.3±2.8
OCRBench	92.7±0.3	64.7±0.6	92.8±0.3	64.8±0.3	96.3±0.6	89.3±0.6	92.8±0.3	64.5±0.0
TextVQA	57.5±0.9	57.0±0.5	58.7±1.8	58.0±0.9	67.3±3.1	50.0±1.0	59.5±1.7	57.3±1.6

A.5.2 ADAPTIVE VISUAL-TOKEN RL

Table 6: Complete three-seed results for InternVL3.5. Fixed and Holdout are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics.

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
SFT								
AI2D	71.2±4.8	78.5±0.5	71.2±4.8	78.5±0.9	70.0±5.3	72.3±4.9	70.2±5.5	79.0±1.0
ChartQA	37.3±2.8	33.8±3.1	37.2±2.1	33.7±3.3	32.0±2.0	42.3±5.0	37.0±3.0	34.5±3.3
DocVQA	28.8±3.3	25.6±1.5	29.2±1.8	23.8±1.4	26.6±1.2	31.8±4.8	28.1±0.7	23.4±2.3
MMBench-EN	70.3±2.4	80.3±1.6	71.7±3.0	80.2±1.3	79.0±3.0	64.3±3.1	71.7±2.0	79.8±1.3
MMMU	53.3±1.5	49.4±0.0	54.3±2.2	50.0±0.0	41.5±1.7	67.0±2.8	53.5±1.8	49.8±1.4
MMStar	45.2±10.6	42.0±12.6	45.3±9.8	42.2±12.9	44.3±6.4	46.3±13.4	44.7±12.4	41.3±12.1
MMVet	18.6±6.2	24.8±2.4	15.1±7.1	24.8±3.7	9.1±9.1	21.4±7.1	19.0±4.7	25.2±4.4
MathVista	22.2±1.2	27.0±2.0	22.2±0.8	26.3±1.2	17.3±1.5	27.0±0.0	22.5±1.8	26.2±0.6
OCRBench	4.3±0.3	17.5±0.5	4.2±0.6	17.5±0.5	5.0±1.0	3.3±0.6	3.7±0.8	17.3±1.3
TextVQA	28.7±1.4	37.5±2.3	28.9±1.3	37.5±2.3	26.5±2.7	31.3±0.8	28.7±1.1	36.8±2.2
DPO								
AI2D	64.2±4.5	75.5±3.9	65.8±1.0	74.7±1.3	64.3±0.6	67.3±2.3	64.8±2.8	74.3±1.4
ChartQA	32.3±5.0	31.3±1.3	36.0±6.0	36.0±4.8	31.0±7.2	41.0±5.3	35.5±6.9	32.5±3.5
DocVQA	29.2±5.4	24.7±4.2	24.5±2.7	21.2±1.6	23.5±1.3	25.5±4.2	24.5±4.0	21.7±1.0
MMBench-EN	76.3±5.3	79.7±3.1	74.2±1.8	79.5±3.0	80.7±2.1	67.7±1.5	75.7±5.5	79.8±2.6
MMMU	53.7±3.9	49.6±3.6	54.6±4.1	50.9±3.8	41.5±5.3	67.8±4.0	54.3±0.8	51.7±2.9
MMStar	49.3±13.0	42.0±12.1	48.2±13.7	39.5±13.3	46.0±8.9	50.3±18.5	47.3±12.4	40.2±10.4
MMVet	19.0±3.7	26.4±7.7	20.9±8.1	30.2±3.5	12.1±6.9	30.2±9.6	19.4±4.1	20.5±2.9
MathVista	21.5±0.5	25.8±1.2	20.7±1.6	26.7±0.3	18.0±1.7	23.3±4.2	21.0±0.5	26.3±2.4
OCRBench	9.5±11.0	21.0±1.5	5.8±3.8	18.3±4.9	7.3±3.1	4.3±4.9	12.8±17.5	23.2±3.1
TextVQA	35.3±9.5	42.3±6.5	32.8±5.9	37.2±1.0	31.9±7.7	33.6±4.3	37.6±8.9	41.6±9.8
PPO								
AI2D	68.8±4.8	78.5±4.4	68.7±6.9	78.3±3.8	65.7±5.7	71.7±9.0	68.5±5.6	77.3±3.8
ChartQA	37.8±11.9	37.0±9.5	39.3±13.1	37.7±11.3	35.3±12.1	43.3±14.5	38.7±11.9	35.8±9.1
DocVQA	28.1±6.8	21.4±4.7	28.6±9.0	21.6±4.3	25.9±8.7	31.4±9.3	28.8±5.2	21.3±3.9
MMBench-EN	74.5±6.1	81.0±0.5	74.5±7.5	81.8±0.6	79.7±7.0	69.3±8.4	74.2±8.2	81.5±0.5
MMMU	55.7±1.7	49.6±1.7	55.2±3.1	48.9±2.9	41.9±2.6	68.5±3.6	55.2±1.3	50.2±1.2
MMStar	48.8±15.0	43.2±14.5	48.7±14.7	42.3±15.3	49.0±10.4	48.3±19.1	47.5±15.3	43.7±14.4
MMVet	19.4±3.7	26.7±10.1	22.1±3.1	29.5±13.0	10.6±2.6	34.1±5.0	23.6±4.8	32.2±10.6
MathVista	21.0±2.6	27.5±2.6	20.7±1.3	28.5±2.6	17.7±1.5	23.7±4.0	20.7±1.9	28.7±2.9
OCRBench	18.8±26.6	27.8±10.6	6.7±5.9	22.0±0.9	8.3±5.9	5.0±6.1	24.8±38.3	35.3±23.1
TextVQA	34.2±7.3	41.4±6.8	34.8±8.2	42.8±9.4	33.0±10.1	36.6±6.4	34.2±7.4	43.9±7.6
GRPO								
AI2D	67.2±1.8	75.8±1.0	65.0±0.0	75.2±2.4	63.0±1.0	67.0±1.0	64.5±1.5	72.8±0.3
ChartQA	42.0±7.1	36.5±7.5	32.8±3.3	30.3±2.9	26.7±4.0	39.0±2.6	35.0±3.5	32.3±1.9
DocVQA	27.3±5.6	22.3±4.2	26.2±1.3	19.9±2.4	24.6±2.2	27.8±0.9	28.8±8.3	23.3±4.8
MMBench-EN	74.0±6.2	81.7±2.0	75.3±2.8	79.5±2.0	83.3±2.3	67.3±3.5	75.2±5.8	80.0±0.5
MMMU	54.4±3.8	50.7±3.9	54.6±2.1	48.7±3.3	41.9±1.7	67.4±2.6	54.8±2.9	50.9±3.7
MMStar	48.3±15.1	42.0±14.9	49.3±14.2	42.5±14.0	46.3±9.3	52.3±19.3	49.7±13.2	40.8±13.0
MMVet	17.8±4.4	24.0±3.7	18.2±5.2	19.8±5.1	12.1±6.9	24.6±3.6	13.2±4.7	16.7±3.7
MathVista	19.3±1.4	27.3±0.6	21.3±3.3	26.7±1.9	18.7±6.4	24.0±2.0	21.5±2.0	28.8±2.0
OCRBench	3.0±0.5	17.0±2.0	15.2±21.1	23.8±4.5	22.0±30.3	8.3±11.8	2.7±1.4	17.3±6.5
TextVQA	29.8±5.7	37.2±4.7	37.2±10.2	41.8±7.8	38.9±10.1	35.5±10.7	33.0±6.4	39.1±3.6
RLOO								
AI2D	71.2±6.9	77.7±4.5	71.5±7.9	77.3±3.6	70.7±10.7	72.3±5.1	70.3±6.3	76.8±4.6
ChartQA	33.2±2.6	33.7±3.9	39.2±11.1	36.2±8.2	32.0±12.3	46.3±10.1	38.8±11.0	36.8±6.3
DocVQA	31.8±9.7	25.2±5.6	32.5±10.0	25.0±6.1	30.4±9.4	34.5±10.5	29.2±7.9	24.4±4.3
MMBench-EN	76.5±5.9	80.2±0.8	76.3±7.1	79.3±1.9	83.0±2.6	69.7±11.9	74.8±6.6	80.2±1.2
MMMU	55.2±2.1	50.2±2.6	56.1±2.8	51.5±3.3	43.7±2.3	68.5±3.4	56.5±3.4	51.5±3.4
MMStar	45.3±13.9	42.2±14.4	47.2±12.9	41.8±13.3	45.7±8.4	48.7±17.4	46.2±12.9	42.5±16.2
MMVet	18.6±2.0	24.4±4.2	25.6±5.3	28.3±12.8	18.2±0.0	33.3±10.9	16.3±3.1	23.3±5.1
MathVista	21.0±2.0	27.0±4.0	22.2±1.0	27.3±2.8	19.0±1.0	25.3±3.1	20.8±2.0	26.5±2.3
OCRBench	24.5±37.2	33.5±27.3	28.8±43.9	35.3±27.0	32.7±46.2	25.0±41.6	27.5±42.0	34.8±26.2
TextVQA	36.6±13.6	42.9±10.5	41.9±8.3	47.2±10.6	42.9±8.6	40.8±8.3	35.8±10.6	42.1±10.7

Table 7: Complete three-seed results for LLaVA-OneVision-7B. Fixed and Holdout are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics.

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
SFT								
AI2D	68.8±1.2	77.5±1.5	69.2±2.0	77.7±1.3	70.0±1.7	68.3±2.5	68.8±0.3	76.7±2.1
ChartQA	38.0±3.1	37.3±2.8	38.5±2.8	37.3±3.0	30.7±3.2	46.3±3.2	38.0±3.0	37.2±2.3
DocVQA	41.9±0.6	37.9±2.1	42.5±1.0	37.8±0.2	41.2±3.0	43.8±0.9	43.5±1.1	39.2±2.2
MMBench-EN	76.8±0.8	77.2±1.3	76.8±2.1	78.3±2.1	88.3±3.1	65.3±1.2	75.8±1.8	79.0±1.0
MMMU	47.4±0.3	42.0±0.8	46.7±1.9	43.0±0.3	36.7±1.1	56.7±2.9	48.3±1.1	43.3±1.0
MMStar	49.0±1.8	45.2±1.9	50.0±0.5	45.5±1.8	44.7±1.2	55.3±1.5	50.0±0.9	44.7±2.0
MMVet	13.2±4.8	32.2±6.4	15.5±3.6	33.7±7.1	12.1±2.6	19.0±9.5	16.3±6.2	32.9±10.6
MathVista	23.8±2.3	29.3±1.2	24.3±2.0	30.0±0.5	24.0±3.6	24.7±1.2	24.7±1.5	29.8±0.3
OCRBench	11.7±0.8	28.5±1.3	12.5±0.5	28.5±1.5	10.7±1.5	14.3±1.5	11.7±0.8	28.0±2.5
TextVQA	41.5±1.7	45.9±3.7	41.5±2.2	45.9±3.7	38.9±2.1	44.1±2.2	40.9±2.5	46.9±2.4
DPO								
AI2D	72.2±5.8	82.5±4.3	73.3±2.6	83.3±1.3	74.3±4.6	72.3±1.5	70.7±4.5	83.0±2.2
ChartQA	49.0±15.3	50.3±14.0	44.8±14.6	48.2±15.0	38.3±12.2	51.3±17.8	49.0±14.0	49.2±15.6
DocVQA	67.5±14.0	60.0±10.6	63.7±20.2	55.5±18.6	61.2±18.7	66.1±22.0	71.7±10.2	62.7±7.6
MMBench-EN	81.3±6.0	79.5±1.5	81.7±5.3	79.8±2.0	91.3±2.9	72.0±7.8	80.7±8.0	78.8±2.1
MMMU	47.4±2.0	43.3±1.1	47.4±2.6	42.4±1.2	37.8±1.1	57.0±4.2	47.2±1.1	43.0±2.2
MMStar	54.2±4.5	48.0±3.0	53.7±3.2	47.7±3.2	47.3±2.5	60.0±4.4	54.3±3.9	47.3±3.8
MMVet	20.9±6.2	31.4±9.1	22.5±4.8	31.0±10.7	12.1±5.2	33.3±4.8	22.5±7.1	31.0±12.8
MathVista	30.0±5.7	33.0±4.4	30.2±5.4	34.0±7.1	28.0±4.4	32.3±6.7	30.2±4.6	34.2±5.3
OCRBench	56.2±44.0	49.2±18.3	56.7±44.3	48.8±17.6	61.0±45.9	52.3±42.8	55.8±43.7	49.0±17.8
TextVQA	54.0±16.4	60.7±14.0	54.4±12.9	60.3±14.4	54.9±12.4	53.9±13.5	55.4±12.7	60.0±13.9
PPO								
AI2D	69.3±6.2	78.7±6.5	67.0±6.9	79.5±6.0	65.3±7.8	68.7±6.0	66.5±6.1	79.3±5.1
ChartQA	37.3±9.3	41.5±12.0	36.5±10.8	40.3±10.8	33.0±7.9	40.0±14.0	36.3±7.9	38.0±9.5
DocVQA	45.6±19.4	37.5±15.8	44.2±14.8	37.2±14.7	44.1±12.5	44.3±17.2	42.3±14.9	36.6±15.0
MMBench-EN	77.3±5.9	78.0±2.3	76.7±3.8	77.3±2.4	89.0±2.6	64.3±5.9	79.0±4.8	78.3±2.8
MMMU	47.4±0.8	41.5±2.3	49.3±1.2	42.6±2.2	40.0±1.1	58.5±2.3	48.1±0.8	42.2±1.5
MMStar	50.3±1.5	46.2±1.2	51.2±2.9	46.7±2.5	47.0±2.0	55.3±4.5	49.7±2.5	46.7±2.5
MMVet	17.1±7.5	29.1±12.1	19.4±5.9	24.4±10.3	12.1±2.6	27.0±9.9	18.6±8.1	29.8±10.0
MathVista	26.8±5.8	29.2±4.9	25.7±6.8	29.2±3.5	25.0±6.1	26.3±7.6	25.5±4.4	29.0±5.1
OCRBench	29.2±42.7	37.8±20.1	23.2±31.9	31.3±7.6	26.7±36.7	19.7±27.2	26.5±38.5	37.0±18.7
TextVQA	44.6±10.1	52.6±13.1	45.8±10.9	53.5±13.5	47.3±13.0	44.3±8.9	46.5±12.0	53.9±13.8
GRPO								
AI2D	72.5±4.1	83.0±2.6	75.8±1.3	84.3±0.3	76.7±3.2	75.0±1.7	71.8±5.5	82.5±4.0
ChartQA	49.5±14.4	51.3±14.5	50.2±12.5	53.7±10.5	44.3±10.0	56.0±15.6	47.3±13.0	51.7±10.1
DocVQA	65.4±11.7	58.3±10.5	73.6±4.2	62.8±1.7	73.4±3.4	73.8±5.9	67.6±16.6	59.9±11.5
MMBench-EN	81.3±6.4	81.0±1.3	81.7±5.3	79.2±2.0	91.7±2.3	71.7±8.4	81.7±5.3	79.5±1.7
MMMU	48.0±3.6	43.0±1.2	47.0±1.4	43.1±1.4	37.8±0.0	56.3±2.8	50.0±2.0	43.9±0.0
MMStar	53.8±3.9	47.7±3.2	53.3±3.8	48.0±3.0	48.7±1.2	58.0±7.0	53.2±4.5	47.3±2.1
MMVet	18.6±7.0	32.6±4.0	21.7±7.1	32.2±10.2	16.7±5.2	27.0±11.0	20.2±5.9	29.8±9.8
MathVista	29.3±6.8	34.0±6.5	27.5±6.0	32.7±6.1	25.7±4.5	29.3±7.5	30.7±3.6	33.7±4.5
OCRBench	56.2±44.4	48.7±18.8	54.2±42.5	47.3±18.2	59.0±46.0	49.3±39.3	56.5±44.2	48.3±18.5
TextVQA	54.3±15.2	60.3±13.3	53.9±17.0	60.4±15.2	54.4±17.7	53.4±16.4	55.7±13.2	60.5±15.4
RLOO								
AI2D	69.5±6.1	80.5±5.2	70.8±3.3	81.0±3.1	73.0±2.0	68.7±4.5	72.2±7.6	81.0±3.9
ChartQA	51.8±3.3	52.3±6.5	38.3±13.6	41.3±7.2	34.3±11.2	42.3±16.7	43.7±15.0	46.5±12.3
DocVQA	61.3±27.9	54.7±21.8	58.6±24.4	50.8±20.4	56.1±21.9	61.0±26.9	64.9±22.0	56.3±20.9
MMBench-EN	80.3±5.3	78.7±1.2	80.5±4.5	78.3±1.0	89.0±3.5	72.0±6.2	80.2±4.2	79.0±2.0
MMMU	48.0±1.7	41.9±1.2	48.7±1.4	44.1±0.3	40.4±1.7	57.0±2.3	48.7±2.0	42.2±1.5
MMStar	49.7±3.0	46.0±3.6	50.3±2.8	46.2±3.3	45.0±1.7	55.7±5.0	50.3±2.0	47.2±2.4
MMVet	16.3±7.0	29.1±6.0	19.4±2.7	29.1±9.2	15.2±5.2	23.8±9.5	14.7±1.3	28.3±5.7
MathVista	26.5±7.3	31.3±5.3	28.2±7.4	31.2±4.0	26.3±5.8	30.0±9.2	26.3±4.1	28.0±3.0
OCRBench	31.0±46.8	38.2±18.5	31.5±45.9	38.3±19.6	33.7±47.9	29.3±43.9	31.8±46.5	38.2±18.0
TextVQA	46.0±14.2	49.8±13.8	46.1±13.7	51.1±15.0	43.9±15.4	48.4±11.9	44.1±8.1	48.9±12.5

Table 8: Complete three-seed results for Qwen2.5-VL. Fixed and Holdout are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics.

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
SFT								
AI2D	59.5±3.8	72.7±1.3	62.0±2.6	72.0±1.5	59.3±3.2	64.7±2.5	63.0±2.8	71.2±3.2
ChartQA	30.8±1.6	31.3±4.2	32.5±3.0	30.8±5.5	27.3±5.9	37.7±1.5	32.3±0.8	29.2±1.4
DocVQA	34.0±2.3	27.7±1.4	31.3±4.8	29.9±1.9	28.5±5.6	34.2±7.3	32.1±1.1	27.7±3.7
MMBench-EN	77.2±2.1	72.2±0.8	78.2±0.3	70.2±2.3	84.7±1.5	71.7±1.5	78.2±1.0	69.5±0.5
MMMU	42.0±2.6	33.0±1.7	41.3±2.3	32.6±2.1	32.2±2.9	50.4±1.7	42.2±2.2	32.8±1.7
MMStar	39.8±10.3	34.5±12.1	41.7±11.9	36.8±10.9	39.7±5.8	43.7±18.1	40.8±9.0	37.2±8.0
MMVet	24.8±5.7	19.4±4.8	25.2±2.9	25.6±2.3	9.8±4.7	41.3±7.3	22.5±5.9	23.3±6.0
MathVista	18.2±1.0	24.8±1.0	17.2±2.0	24.7±1.3	14.0±2.6	20.3±1.5	17.2±1.8	24.0±1.5
OCRBench	10.5±1.8	20.0±1.3	9.0±2.6	21.8±1.2	10.0±2.0	8.0±3.6	10.5±1.0	21.8±0.3
TextVQA	33.4±3.7	36.7±2.2	31.9±0.7	37.1±1.7	29.3±1.7	34.4±0.9	31.7±2.9	37.0±2.6
DPO								
AI2D	57.5±4.8	68.8±1.3	57.2±3.8	70.7±2.8	55.0±8.0	59.3±0.6	56.0±3.0	67.7±2.0
ChartQA	29.2±2.8	30.7±3.0	27.5±1.7	28.3±3.5	22.7±3.8	32.3±3.2	27.7±3.7	30.0±3.5
DocVQA	26.9±3.8	25.9±2.3	24.8±5.5	23.5±1.9	22.4±4.0	27.2±7.4	26.1±1.3	24.4±2.0
MMBench-EN	79.2±3.2	66.5±4.4	79.2±4.6	67.2±2.8	84.3±4.5	74.0±5.2	79.7±3.8	67.3±2.9
MMMU	41.1±1.1	32.8±1.5	40.7±2.6	31.9±2.0	30.7±3.9	50.7±2.3	42.2±2.9	32.6±1.8
MMStar	40.0±9.7	36.3±9.8	39.2±9.0	36.7±9.8	36.0±3.6	42.3±14.4	40.5±10.6	37.5±10.8
MMVet	20.5±5.2	20.2±5.4	23.3±8.4	23.6±7.7	15.2±6.9	31.7±11.0	20.9±5.1	17.8±3.6
MathVista	17.8±2.0	25.5±1.3	17.7±1.6	24.0±1.7	14.7±3.8	20.7±0.6	17.3±1.2	24.2±1.6
OCRBench	25.8±37.0	35.7±20.6	5.3±0.8	19.0±3.6	7.0±1.0	3.7±1.2	24.7±34.5	32.5±16.0
TextVQA	36.8±3.5	36.2±5.4	36.0±7.0	37.5±7.2	33.9±5.0	38.2±9.7	36.6±7.2	39.1±5.8
PPO								
AI2D	57.8±6.6	71.8±4.3	58.3±8.9	70.3±5.1	56.7±9.6	60.0±8.2	59.8±4.9	69.7±7.3
ChartQA	34.0±13.9	37.5±13.5	31.7±11.1	33.8±10.4	26.0±12.1	37.3±10.2	32.0±11.0	33.2±11.1
DocVQA	32.5±14.4	29.2±8.8	32.7±11.1	28.5±7.8	30.5±12.7	35.0±9.6	33.0±8.0	29.5±7.9
MMBench-EN	81.0±5.6	70.0±1.3	80.2±4.8	69.8±1.3	84.7±4.2	75.7±5.5	80.7±3.3	70.8±1.8
MMMU	40.9±3.4	33.3±1.7	44.1±2.7	34.6±0.3	34.1±0.6	54.1±6.1	43.5±1.3	33.9±2.0
MMStar	42.7±14.1	37.2±12.7	43.0±12.3	37.5±11.1	38.3±8.4	47.7±18.1	40.2±11.4	37.0±9.7
MMVet	24.4±9.5	27.1±8.8	24.4±8.1	29.1±11.5	10.6±2.6	38.9±18.5	19.8±5.3	26.4±9.7
MathVista	18.0±1.8	24.8±4.1	17.5±1.3	25.5±3.6	14.7±2.1	20.3±1.2	18.3±1.5	25.7±3.3
OCRBench	26.0±37.3	32.2±18.5	21.5±30.3	29.0±10.4	26.0±35.5	17.0±25.1	27.8±40.9	36.0±23.0
TextVQA	39.1±8.4	41.0±8.0	41.0±10.6	42.6±13.9	38.9±10.6	43.2±10.8	37.3±11.3	44.0±10.3
GRPO								
AI2D	58.0±6.2	67.7±1.8	58.5±5.8	69.7±1.8	57.3±8.1	59.7±3.5	55.5±2.6	69.3±1.0
ChartQA	27.8±0.3	30.7±1.8	34.0±10.7	30.2±10.8	29.0±11.8	39.0±9.6	33.5±12.2	36.2±7.4
DocVQA	37.7±9.8	32.9±8.3	36.6±12.4	31.9±5.1	34.3±14.6	38.9±10.8	26.5±4.3	26.3±0.8
MMBench-EN	80.2±2.5	68.3±2.8	79.8±2.4	68.5±3.6	85.0±4.6	74.7±0.6	80.2±4.0	67.3±2.8
MMMU	38.7±2.5	32.0±1.6	41.9±1.8	31.5±2.0	31.5±2.3	52.2±2.2	43.3±3.4	31.3±2.1
MMStar	41.7±13.3	37.5±11.1	40.2±11.1	36.8±10.9	37.3±6.4	43.0±15.7	38.0±10.2	35.8±10.3
MMVet	24.4±4.2	24.0±7.5	19.4±4.1	23.3±4.0	12.1±5.2	27.0±9.6	24.4±5.3	23.3±6.2
MathVista	17.2±1.6	25.5±1.3	18.5±1.0	25.0±0.9	16.3±1.5	20.7±0.6	17.3±0.8	26.3±1.0
OCRBench	4.0±1.3	19.5±3.3	4.5±1.3	18.0±3.5	6.0±1.7	3.0±1.0	3.8±0.8	20.3±0.6
TextVQA	39.1±11.8	40.3±11.3	35.4±8.8	38.0±7.6	34.5±8.2	36.4±9.9	39.3±8.5	42.3±8.7
RLOO								
AI2D	66.7±7.8	74.7±5.9	61.7±6.5	75.3±4.6	62.0±7.2	61.3±5.9	65.0±10.9	73.5±7.9
ChartQA	36.3±11.0	38.2±10.9	35.2±13.0	37.7±13.5	31.3±11.8	39.0±14.4	28.7±3.6	32.3±3.7
DocVQA	45.9±16.0	38.6±9.0	46.8±14.0	39.4±8.1	43.1±11.5	50.5±16.5	46.4±15.3	38.2±11.4
MMBench-EN	79.5±7.4	70.3±1.6	79.3±6.8	70.3±4.0	83.0±5.6	75.7±8.1	79.5±6.6	69.3±2.4
MMMU	43.1±4.2	34.6±3.1	40.0±2.0	32.4±2.0	31.1±2.2	48.9±1.9	44.4±4.8	33.9±2.4
MMStar	46.5±7.9	37.7±10.4	42.7±14.6	37.0±12.1	38.7±9.0	46.7±20.5	44.7±13.7	38.3±13.8
MMVet	23.6±8.4	24.0±11.9	24.0±3.6	24.0±9.4	12.1±2.6	36.5±9.9	24.8±5.9	22.5±12.8
MathVista	18.0±1.0	24.8±3.0	18.7±2.8	27.5±2.3	15.3±2.9	22.0±2.6	18.5±2.0	25.2±2.4
OCRBench	30.7±44.5	37.2±25.0	30.0±45.0	36.2±25.9	31.0±44.2	29.0±45.9	31.0±43.7	35.2±25.4
TextVQA	44.1±12.8	47.2±9.9	45.2±10.4	48.4±10.5	43.7±7.5	46.8±13.3	38.5±10.1	41.6±11.4

Table 9: Complete three-seed results for Qwen3-VL-8B. Fixed and Holdout are reported at every overlap ratio; Seen and Unseen are reported at 50% as descriptive subset diagnostics.

Benchmark	0% overlap		50% overlap				100% overlap	
	Fixed	Holdout	Fixed	Holdout	Seen	Unseen	Fixed	Holdout
SFT								
AI2D	72.2±1.3	79.3±1.3	70.0±1.8	78.8±1.9	68.0±1.7	72.0±3.6	72.7±3.8	78.7±1.4
ChartQA	40.5±1.8	42.5±3.8	41.0±2.6	40.3±4.6	34.7±4.0	47.3±3.2	42.3±2.9	39.0±3.3
DocVQA	42.2±3.6	34.6±1.5	39.6±5.7	36.6±5.4	37.1±4.7	42.2±8.7	45.1±2.8	35.3±3.2
MMBench-EN	80.7±0.3	82.7±1.2	83.8±0.6	83.2±2.3	89.3±0.6	78.3±1.5	82.0±1.5	80.7±1.5
MMMU	54.3±2.3	47.8±2.0	52.0±1.7	46.5±2.2	42.2±3.3	61.9±0.6	54.1±1.2	46.3±1.2
MMStar	62.8±1.9	54.5±1.5	63.2±2.5	56.2±0.8	50.3±3.1	76.0±3.0	53.2±17.1	45.2±18.0
MMVet	12.4±1.3	19.4±0.7	20.2±2.7	22.1±5.1	16.7±2.6	23.8±8.2	11.6±2.3	22.5±5.9
MathVista	16.7±1.6	21.7±0.8	16.3±0.8	21.0±0.5	16.0±2.0	16.7±2.1	16.5±1.8	21.8±0.6
OCRBench	22.7±3.3	28.0±2.0	20.3±1.3	27.5±2.3	24.0±1.0	16.7±3.1	21.7±1.9	28.3±1.5
TextVQA	39.0±1.4	44.6±1.7	39.3±2.5	46.7±1.8	40.8±3.7	37.8±1.8	40.4±3.3	43.5±1.7
DPO								
AI2D	66.0±3.1	77.0±0.9	69.2±4.8	77.3±3.5	68.0±7.0	70.3±2.5	68.7±6.4	80.3±0.8
ChartQA	37.0±0.0	40.5±4.0	36.8±2.9	36.8±7.0	34.0±3.6	39.7±2.3	37.3±5.5	38.2±4.5
DocVQA	39.9±7.4	33.7±5.0	41.8±4.3	31.3±4.4	41.9±5.0	41.6±3.8	50.6±4.6	38.1±3.7
MMBench-EN	83.3±4.9	80.8±1.8	82.3±3.3	79.2±3.8	88.3±3.2	76.3±4.0	82.5±4.0	79.5±1.5
MMMU	54.8±2.1	48.1±3.9	55.9±1.6	47.8±2.5	45.6±5.1	66.3±2.3	54.4±2.0	47.6±0.8
MMStar	52.3±15.5	44.3±20.3	53.5±14.3	45.2±19.9	46.7±4.7	60.3±25.4	53.0±12.3	45.0±20.2
MMVet	10.9±3.6	26.7±1.2	11.6±4.7	18.2±1.8	12.1±2.6	11.1±9.9	14.7±3.6	22.9±3.4
MathVista	17.5±1.0	20.3±0.8	16.8±2.0	20.0±0.9	17.3±1.2	16.3±2.9	17.3±1.4	21.2±1.4
OCRBench	8.5±0.0	26.7±1.0	30.8±37.0	38.8±16.2	37.3±42.2	24.3±31.8	33.2±41.9	41.5±22.5
TextVQA	40.9±1.9	43.0±2.8	39.0±9.7	46.3±11.4	40.3±10.2	37.7±9.9	46.6±7.3	52.2±8.1
PPO								
AI2D	69.8±5.9	79.0±2.3	71.5±4.4	79.5±3.8	69.3±3.2	73.7±5.7	70.8±7.7	79.2±2.0
ChartQA	41.2±18.6	44.7±13.8	40.8±10.9	40.8±10.1	36.3±11.2	45.3±10.7	41.7±13.4	40.8±12.6
DocVQA	39.8±10.9	31.2±8.1	40.8±10.5	33.4±7.5	39.3±8.6	42.3±12.4	42.6±7.5	32.6±7.6
MMBench-EN	83.2±4.0	81.0±2.0	81.7±5.5	82.7±2.1	86.0±3.6	77.3±7.5	83.7±3.5	81.8±2.9
MMMU	54.3±0.8	46.1±1.1	56.9±0.8	46.3±2.2	43.3±5.1	70.4±4.5	54.6±0.8	46.7±1.7
MMStar	55.2±17.5	45.7±20.7	54.8±13.5	47.3±22.2	49.3±7.1	60.3±21.1	52.8±14.2	44.8±19.8
MMVet	15.9±3.7	24.4±6.5	11.6±2.3	21.7±3.6	12.1±2.6	11.1±2.7	12.4±2.7	25.2±1.8
MathVista	16.8±1.3	21.2±1.3	16.3±2.0	21.0±1.3	15.3±1.5	17.3±2.5	15.8±1.0	20.2±1.3
OCRBench	30.0±37.2	37.2±14.2	18.2±16.7	30.3±3.3	21.7±15.9	14.7±17.7	32.3±43.0	43.3±21.9
TextVQA	44.6±9.0	49.1±9.7	45.3±10.7	51.5±11.1	46.8±13.0	43.9±9.1	45.3±10.3	50.2±13.4
GRPO								
AI2D	70.0±6.6	77.2±2.1	69.3±7.2	76.5±2.3	69.0±9.2	69.7±5.5	67.8±4.8	78.5±0.9
ChartQA	41.2±10.6	39.5±10.5	42.5±2.8	42.8±7.3	37.3±4.5	47.7±1.5	37.5±0.5	36.8±1.2
DocVQA	51.5±14.3	43.7±13.7	37.7±6.4	33.0±3.3	36.2±6.9	39.3±6.4	39.9±2.6	32.7±2.1
MMBench-EN	82.7±3.3	80.8±2.0	82.8±3.3	80.7±2.8	88.7±4.2	77.0±3.6	81.5±4.0	80.8±4.0
MMMU	55.2±2.3	47.2±0.0	53.7±3.3	48.3±2.9	42.2±6.2	65.2±0.6	55.9±0.8	47.4±0.8
MMStar	52.0±13.0	45.5±19.9	53.7±13.3	45.3±20.0	47.7±6.5	59.7±20.5	52.5±13.4	45.3±18.5
MMVet	14.0±4.0	23.3±2.3	12.4±1.3	24.0±3.6	9.1±0.0	15.9±2.7	9.3±2.3	20.2±8.2
MathVista	16.3±2.4	20.2±1.0	17.0±0.9	20.8±0.8	17.0±1.0	17.0±2.0	17.5±1.0	20.3±0.3
OCRBench	8.8±1.0	27.7±0.3	28.8±32.6	33.2±12.0	37.0±39.8	20.7±25.4	9.7±2.0	26.5±3.0
TextVQA	45.0±7.8	51.2±8.5	50.8±9.4	57.8±12.0	53.4±10.6	48.1±8.6	43.0±4.2	47.9±5.3
RLOO								
AI2D	72.0±10.7	80.5±5.1	72.5±10.5	80.2±3.8	69.3±13.8	75.7±7.2	73.3±11.0	81.7±6.4
ChartQA	44.3±17.1	45.5±16.5	45.0±18.6	47.0±16.0	40.0±21.0	50.0±16.4	44.5±9.8	47.2±9.5
DocVQA	56.2±15.6	46.4±14.5	50.9±14.0	38.7±9.2	49.1±12.6	52.6±15.4	59.3±9.5	44.5±7.1
MMBench-EN	82.8±4.8	80.8±1.5	82.8±3.4	83.0±2.3	88.3±3.2	77.3±5.7	84.2±4.6	82.2±3.3
MMMU	53.0±0.8	48.5±1.4	55.0±1.5	46.7±2.5	44.1±1.7	65.9±1.7	55.7±3.7	46.3±0.8
MMStar	57.5±15.3	45.7±22.3	53.7±14.9	43.8±20.7	50.7±5.9	56.7±24.0	55.5±13.7	44.5±19.1
MMVet	14.0±2.3	23.6±8.6	7.8±3.6	24.0±7.1	9.1±0.0	6.3±7.3	14.0±4.0	23.3±4.7
MathVista	16.2±2.1	20.7±0.8	15.5±1.3	21.0±1.5	15.3±2.1	15.7±1.2	16.8±1.8	22.2±0.3
OCRBench	33.0±41.6	42.8±24.0	37.2±45.8	43.3±27.1	40.0±47.6	34.3±44.0	35.8±46.5	41.8±27.0
TextVQA	45.1±9.7	48.8±14.6	42.9±11.2	50.8±14.7	45.4±13.5	40.4±8.9	45.9±11.5	50.3±16.5