CoRAL Synthesis: Learner-Relative Calibration of Data Complexity Composition for Improving VLM Generalization

Anonymous authors
Paper under double-blind review

Abstract

The data complexity of training examples can substantially affect optimization and generalization. Yet, data complexity remains an underspecified axis of synthetic data design, with complexity typically controlled using simple heuristics rather than principled criteria. This raises two questions: how data complexity affects downstream performance and how it should be leveraged to promote generalization. To study this in vision-language tasks, we construct hundreds of controlled VQA post-training settings spanning multiple complexity levels, using matched distinct subsets that isolate complexity from data scale, quality, and diversity. We find that complexity substantially affects out-of-distribution (OOD) generalization under a controlled setting, with some smaller subsets outperforming larger training sets. Moreover, our empirical analysis shows that appropriately mixing Easy, Medium, and Hard examples improves generalization over relying on any single complexity level, but realizing this benefit requires defining these levels relative to the target learner. We also find that existing complexity measures provide limited guidance for synthetic data curation. Building on these findings, we introduce **CoRAL Synthesis**, a scalable framework that operationalizes learner-relative complexity to govern synthetic data composition. CoRAL Synthesis consistently outperforms naive few-shot synthesis across data scales, with its 1K setting surpassing the vanilla 16K setting. Using CoRAL Synthesis, we release CoRAL-GeneralVQA and CoRAL-ReasoningVQA, two large-scale synthetic datasets for vision-language reasoning. Models trained on these datasets achieve strong OOD generalization; in particular, CoRAL-ReasoningVQA-4B outperforms the strongest evaluated 4B baseline by 5.30%p on average across 9 unseen vision-language reasoning benchmarks.[1]

1 Introduction

Identifying which training examples most effectively improve model generalization is a central problem in data curation (Paul et al., 2021; Mindermann et al., 2022; Jung et al., 2025; Sachdeva et al., 2026). This question becomes particularly important for synthetic data, where scalable generation shifts the challenge from acquiring more data to determining what data should be generated and retained (Jung et al., 2025; Guha et al., 2026; Davidson et al., 2026; Acuna et al., 2026). Beyond data scale alone, prior work has shown that the difficulty and learning dynamics of individual examples can substantially affect optimization and generalization (Swayamdipta et al., 2020; Mindermann et al., 2022; Ethayarajh et al., 2022; Liu et al., 2025; Kessler et al., 2026). Yet, while synthetic data generation has increasingly emphasized data quality and diversity, it remains unclear how data complexity relates to downstream performance and how it can be leveraged to improve generalization.[2]

In this work, we study the role of data complexity in the broader multimodal setting of VLM post-training, focusing on the following two questions. First, we ask how strongly downstream performance—averaged across unseen benchmarks—varies with data complexity when data scale,

[1] Dataset and code will be published.

[2] To avoid terminological ambiguity, we follow Havrilla et al. (2024); Davidson et al. (2026) and view *data complexity* as a characteristic of data that captures notions of a sample's "difficulty" or "compositionality."

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

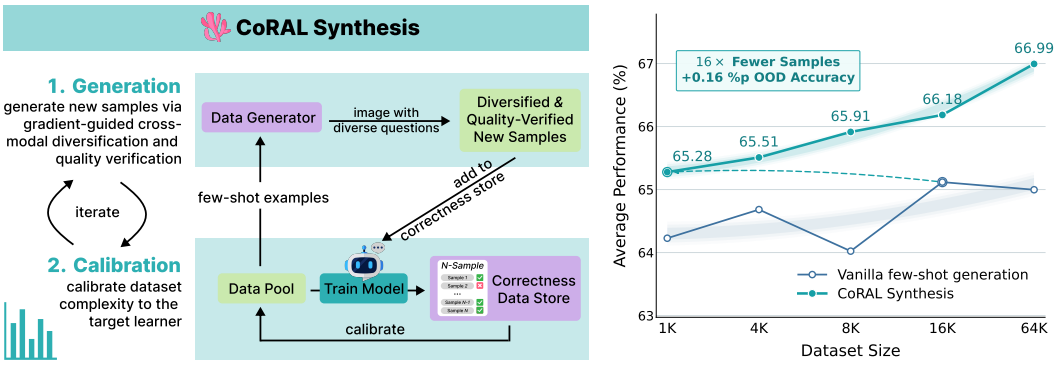


Figure 1: (Left) **Overview of CoRAL Synthesis.** The full process is described in § 4.1. (Right) OOD accuracy on General VQA at scale. By jointly controlling quality, diversity, and complexity, CoRAL Synthesis exhibits strong scaling behavior that continuously improves OOD generalization, surpassing vanilla few-shot generation while requiring up to 16× fewer synthetic samples.

quality, and diversity are carefully controlled. Second, we examine how different complexity levels themselves affect OOD generalization, asking whether heuristically favoring a particular level is sufficient or whether combining multiple levels yields better performance across training stages. To answer these questions, we construct more than 2 million synthetic samples spanning two complementary settings: *General VQA* and *Reasoning VQA* covering mathematical, scientific, and chart-based reasoning. Using these large-scale data pools, we post-train over 300 models on distinct subsets spanning complexity levels from Easy to Hard while controlling for data scale, quality, and diversity, allowing us to isolate the effect of complexity on generalization. From this large-scale empirical study, we derive three key insights:

- Even under controlled data scale, quality, and diversity, different complexity levels lead to systematic differences in OOD generalization. In several cases, these differences are large enough that models trained on smaller subsets consistently outperform those trained on substantially larger ones, with some 16K subsets even surpassing the model trained on the full 1M pool, highlighting complexity as an important factor for improving data efficiency in OOD generalization.
- Rather than treating a single complexity level as optimal, appropriately mixing Easy, Medium, and Hard examples can yield better performance on unseen benchmarks than relying on any one level alone. Crucially, complexity is relative to the learner: examples that are easy for one model may be hard for another, and even when selecting examples based on the target model's own Easy, Medium, and Hard assignments, retaining the same mixture ratio restores the benefit.
- Existing complexity measures often show only weak-to-moderate correlations with downstream performance, limiting their reliability as standalone signals for complexity-aware data curation.

Leveraging these findings, we propose **CoRAL Synthesis** (**CoRAL**; **Co**mplexity calibration **R**elative to **A** **L**earner), a scalable iterative framework that improves data quality and diversity while dynamically calibrating complexity to the target learner, as illustrated in Fig. 1. To calibrate complexity, we operate in two phases. (1) For each problem in the data pool, we sample N responses from the target learner and count the number of correct responses. We partition problems into Easy, Medium, or Hard according to their number of correct rollouts, using three equally sized correctness ranges. (2) We downsample overrepresented complexity tiers based on the size of the most underrepresented tier, keeping the intended Easy–Medium–Hard composition. Repeating this generation–calibration process maintains the target complexity composition throughout synthesis, with the complexity of each problem defined relative to the target learner.

By applying our pipeline to large-scale synthesis, we release two synthetic datasets comprising 624K samples for General VQA and 569K samples for Reasoning VQA: *CoRAL-GeneralVQA* for general visual question answering and *CoRAL-ReasoningVQA* for vision-language reasoning, with the latter comprising approximately 193K math, 110K science, and 264K chart reasoning examples. Models trained on these datasets achieve strong OOD generalization across both settings; for example, CoRAL-ReasoningVQA-4B improves the average score across 9 unseen reasoning benchmarks by 5.30%p over the best baseline. Experimental results show that CoRAL Synthesis consistently

improves OOD performance at scale and outperforms naive few-shot generation at every scale; even with 1K samples, it surpasses the vanilla 16K setting. Taken together, our results suggest that effective synthetic data scaling depends not on quantity alone, but on controlling not only quality and diversity, but also learner-relative complexity.

2 Related Work

Effective data curation and generation require consideration of three key factors: quality, diversity, and complexity (Havrilla et al., 2024; Liu et al., 2024a; Davidson et al., 2026). Prior literature on instruction tuning has emphasized both data quality and diversity as key determinants of training effectiveness (Zhou et al., 2023; Liu et al., 2024a; Bukharin et al., 2024; Sachdeva et al., 2026). Quality typically refers to sample correctness or noise (Havrilla et al., 2024; Gou et al., 2025; Davidson et al., 2026), whereas diversity captures variation across a dataset, ranging from lexical or embedding-based similarity (Wu et al., 2023; Ni et al., 2024) to recent gradient-based measures of learning-signal diversity (Pan et al., 2024; Xia et al., 2024; Jung et al., 2025). Complexity has been operationalized through diverse signals: instruction-side structure such as token length, LM-derived semantic-tree size, and semantic tags (Liu et al., 2024a; Zhao et al., 2024; Lu et al., 2024a); visual statistics such as edge density, subband entropy, and C-score (Rosenholtz et al., 2007); model-based signals including PPL, IFD, GraNd, and EL2N (Li et al., 2024a; Paul et al., 2021); and psychometric formulations such as Item Response Theory (Rasch, 1993; DeMars, 2010). Beyond such static measures, complexity has also been characterized relative to model learning behavior, through training dynamics (Swayamdipta et al., 2020) or model-specific information (Ethayarajh et al., 2022). Despite extensive work on data selection, how difficulty should inform selection remains unresolved: prior work has found benefits from prioritizing increasingly complex or high-loss examples (Liu et al., 2024a; Zhao et al., 2024; Tong et al., 2024; Kessler et al., 2026), while other studies show that extreme difficulty can reflect noise, unlearnability, or a mismatch with model capacity (Swayamdipta et al., 2020; Paul et al., 2021; Mindermann et al., 2022; Gao et al., 2025). We therefore examine how to measure and use complexity for synthetic vision-language data curation.

Recent advances in vision-language understanding have been driven by increasingly sophisticated post-training pipelines, combining large-scale reasoning supervision with supervised fine-tuning and reinforcement learning (Yang et al., 2025; Huang et al., 2026; Zhang et al., 2026; Lin et al., 2026; Sarch et al., 2026). In language reasoning, OpenThoughts (Guha et al., 2026) systematically derives synthetic reasoning data recipes through over 1,000 controlled studies of the generation pipeline, while Prismatic Synthesis (Jung et al., 2025) introduces gradient-based diversity and synthesizes examples targeting underrepresented regions of gradient space to improve OOD generalization. In the multimodal setting, LongPerceptualThoughts (Liao et al., 2025) synthesizes long reasoning traces for perceptual tasks through a staged vision-to-reasoning pipeline, while Long Grounded Thoughts (Acuna et al., 2026) extends this direction to million-scale vision-centric synthesis with an explicit focus on increasing problem complexity. More recently, DeltaPrompts (Jung et al., 2026) targets teacher–student capability gaps through answer divergence, synthesizing prompts that provide informative learning signals for multimodal distillation. In contrast, CoRAL Synthesis adapts diversification to cross-modal learning signals, improves quality through strict filtering, and dynamically calibrates complexity composition to the target learner.

3 Understanding Data Complexity

Our main goal is to understand how data complexity affects downstream generalization and how it can be leveraged in synthetic data generation. We first establish a controlled large-scale evaluation setting across General VQA and Reasoning VQA (§ 3.1), and analyze complexity–generalization trends (§ 3.2). We then examine how complexity can be leveraged through optimization and composition (§ 3.3), and finally investigate whether existing complexity measures can serve as proxies for data curation (§ 3.4).

3.1 Controlled Evaluation of Data Complexity

Motivated by the large-scale controlled evaluation paradigm of Jung et al. (2025), we examine whether differences in complexity correspond to systematic differences in downstream generaliza-

Table 1: Learner-relative complexity estimation and diversity control across constructed subsets. The colored bars show the distribution of correct counts over $N=8$ rollouts across the full 1M+ synthetic data pool, with numbers indicating sample percentages. Text, image, and multimodal embedding diversity remain consistent within each data scale, while correctness rollouts span a broad range of complexity.

VQA Task	Correct Rollout Count Distribution (Easy: 8 → Hard: 0)	Scale	Text Embedding Vendi	Image Embedding Vendi	Multimodal Embedding Vendi
General		16K	47.356 ± 1.699	37.734 ± 0.569	434.31 ± 26.82
		32K	47.755 ± 1.764	37.898 ± 0.558	442.91 ± 28.55
		64K	48.223 ± 1.634	38.045 ± 0.409	452.37 ± 25.40
Reasoning		16K	65.966 ± 1.118	18.210 ± 1.228	280.56 ± 7.89
		32K	66.767 ± 1.113	18.245 ± 1.140	285.41 ± 8.59
		64K	67.499 ± 0.905	18.199 ± 0.753	287.10 ± 10.14

tion. To evaluate this relationship reliably, we construct distinct subsets that span a broad range of complexity within a controlled experimental setting. We next describe how these subsets are constructed and how complexity is controlled at scale.

Synthesizing million-scale vision-language datasets. Motivated by off-the-shelf Visual Question Answering (VQA) datasets that pair each image with multiple questions $\{(x_i, I)\}_{i=1}^{N_I}$ (Antol et al., 2015; Kahou et al., 2017; Hudson & Manning, 2019; Qiao et al., 2026), we adopt few-shot generation to synthesize problems for each image. Starting from a unique seed image–problem pair, we iteratively generate new problem–solution pairs for the same image using five textual demonstrations, always retaining the seed QA and randomly selecting up to four from previously generated QAs when available, to avoid redundant generation.[3] We consider two task families: *General VQA*, seeded from VQAv2 and GQA (Antol et al., 2015; Hudson & Manning, 2019), and *Reasoning VQA*, covering math (GeoQA3-English, UniGeo, MathVista, VLMsAreBlind, and We-Math2.0-Standard (Chen et al., 2021; 2022; Lu et al., 2024b; Rahmanzadehgervi et al., 2024; Qiao et al., 2026)), science (AI2D, TQA, and ScienceQA (Kembhavi et al., 2016; 2017; Lu et al., 2022)), and chart reasoning (FigureQA (Kahou et al., 2017)).[4] For each task family, we use Qwen3.5-27B (Qwen Team, 2026) as the generator and synthesize a data pool at approximately the 1M scale. The generation details and synthesized examples are provided in § A.1 and § A.2.

Constructing subsets spanning different complexity levels under controlled data scale, quality, and diversity. Because our analysis aims to isolate differences in complexity, we control three major confounding factors: data scale, quality, and diversity. For quality, we apply a single LM-judge (Qwen3.5-27B) pass uniformly to all samples to assess the validity of the generated problem, solution, and final answer, ensuring a consistent quality standard across all constructed subsets.

A key challenge is constructing controlled subsets that span a broad range of complexity. Following prior work that characterizes sample difficulty from model outcomes (Swayamdipta et al., 2020; Ethayarajh et al., 2022), we adopt multiple correctness rollouts for large-scale complexity estimation (Wu et al., 2026; Shi et al., 2026; Acuna et al., 2026). We use the train model (Qwen3.5-0.8B) as a complexity estimator, and sample $N=8$ responses for each synthetic problem, recording the number of correct rollouts from 8 to 0 as a learner-relative complexity signal. Using this *correctness datastore* as shown in Tab. 1, we construct distinct subsets spanning different complexity levels at each fixed data scale (n = 16K, 32K, 64K) by sampling separately according to correctness count (8, 7, ..., 0), while preserving the original task/domain proportions. For example, a 16K Reasoning-VQA subset targeting correctness count c is constructed by sampling 16K examples with c correct rollouts while matching the Math/Science/Chart proportions of the full 1M pool. For each correctness count, we sample at least 5 subsets with different random seeds while minimizing problem overlap across subsets, totaling 50 subsets per scale.

During subset construction, we also control diversity, which is particularly important in vision-language data, where variation arises jointly from visual inputs and their associated text. We therefore maximize the number of unique images when constructing each subset and, when image reuse

[3] To preserve potentially important image-specific information, we retain the seed QA.

[4] Multiple seed datasets are required to ensure sufficient unique-image coverage at million scale.

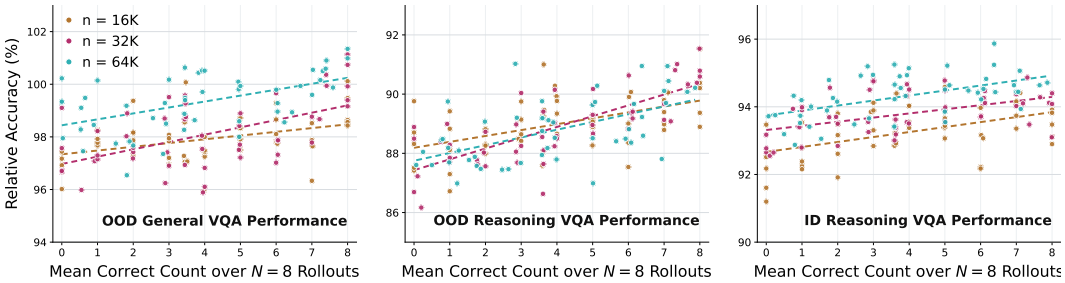


Figure 2: Complexity and OOD/ID performance relative to $P_{\text{ref}}(\mathcal{M})$. **Under controlled quality and diversity, subsets with different complexity levels show consistent differences in OOD/ID performance, suggesting that complexity provides an additional signal for synthetic data curation.** Additional $P_{\text{base}}(\mathcal{M})$ results are provided in § A.4.

is unavoidable, retain only semantically distinct problems for the same image. We then verify that diversity remains comparable across subsets using the Vendi score (Friedman & Dieng, 2022), computed from Qwen3-Embedding-8B (Zhang et al., 2025) text embeddings, InceptionV3 (Szegedy et al., 2016) image embeddings, and Qwen3-VL-Embedding-8B multimodal embeddings. As shown in Tab. 1, the subsets span different correctness ranges while all three Vendi scores remain comparable at each data scale, supporting our controlled setting. The full subset sampling procedure is provided in § A.3.

Baselines. For OOD evaluation, we use 4 unseen benchmarks for General VQA: RealWorldQA, MMStar, MMBench EN, and HallusionBench (xAI, 2024; Chen et al., 2024; Liu et al., 2024b; Guan et al., 2024); and 7 for Reasoning VQA: ChartQA, InfoVQA, M3Exam, MMMU, MathVision, MMSci, and DynaMath (Masry et al., 2022; Mathew et al., 2022; Zhang et al., 2023; Yue et al., 2024; Wang et al., 2024a; Li et al., 2024b; Zou et al., 2025), covering mathematical, scientific, and chart-based reasoning. For in-distribution (ID) evaluation, we evaluate each model on held-out problems from the corresponding seed datasets used to construct the synthetic data. All evaluations are performed after 1 epoch of fine-tuning and conducted using VLMEvalKit (Duan et al., 2024). To capture both gains over the initial model and performance normalized by benchmark-specific difficulty, we evaluate each trained subset against two complementary references: the base model $\mathcal{M}_{\text{base}}$ and the reference model $\mathcal{M}_{\text{ref}}$ trained on the entire data pool (*i.e.*, 1M scale).

$$P_{\text{base}}(\mathcal{M}) = \frac{1}{K}\sum_{k=1}^{K}\frac{A_k(\mathcal{M})}{A_k(\mathcal{M}_{\text{base}})}, \qquad P_{\text{ref}}(\mathcal{M}) = \frac{1}{K}\sum_{k=1}^{K}\frac{A_k(\mathcal{M})}{A_k(\mathcal{M}_{\text{ref}})}, \tag{1}$$

where K denotes the number of evaluation benchmarks and $A_k(\mathcal{M})$ denotes the accuracy of model $\mathcal{M}$ on the k-th benchmark.

3.2 Analysis

Complexity provides an additional axis for improving data efficiency and generalization. As shown in Fig. 2, OOD generalization varies consistently with complexity within each fixed data scale under controlled data quality and diversity. This relationship persists across the 16K, 32K, and 64K scales, suggesting that complexity remains relevant even as the amount of training data increases. In some cases, smaller-scale subsets with appropriate complexity profiles outperform larger-scale subsets in OOD/ID performance. Notably, several curated subsets even exceed the $\mathcal{M}_{\text{ref}}$, suggesting that data scale alone does not determine generalization and that complexity provides an additional axis for efficient synthetic data curation.

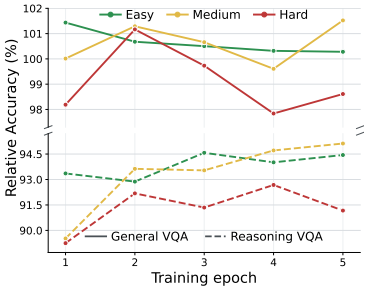


Figure 3: $P_{\text{ref}}(\mathcal{M})$ across E/M/H levels and epochs.

Fig. 2 appears to suggest that easier subsets often yield better performance after 1 epoch. However, since sample utility can change over training, hard subsets may simply require more optimization. We therefore examine

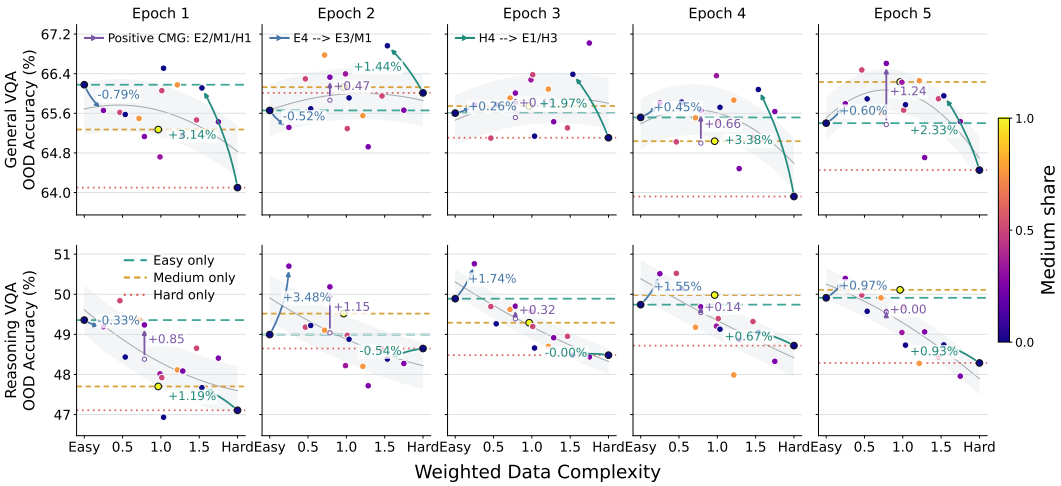


Figure 4: Complexity calibration through task-dependent composition. Point color denotes the share of Medium examples in each mixture. We compare Easy–Medium–Hard complexity mixtures across five epochs for General and Reasoning VQA. Positive CMG indicates that a mixture outperforms the weighted expectation of its corresponding pure-complexity components. **Appropriate mixtures of complexity levels can outperform training on any single level.** Additional results are in § A.5.

the OOD generalization of representative Easy (E), Medium (M), and Hard (H) subsets (correctness counts 8, 4, and 0) across training stages. As shown in Fig. 3, the best-performing complexity level varies across training epochs and task families, making it difficult to favor a single level a priori. In light of these observations and prior work (Swayamdipta et al., 2020; Kordi et al., 2026), we examine whether controlling complexity composition can provide a more effective alternative to single-level selection.

3.3 Impact of Complexity Composition

We examine how different complexity levels affect generalization in synthetic vision-language data generation. Using the same correctness datastore and subset-sampling procedure as in § 3.1, we fix the data scale at 64K and construct new composition-controlled subsets from four 16K blocks sampled from the Easy, Medium, and Hard regions of the datastore (correctness counts 8, 4, and 0, respectively); unlike the fixed-complexity subsets in § 3.1, we vary how these four blocks are combined (*e.g.*, E:E:M:H; hereafter E2/M1/H1). This gives 15 possible four-block compositions in 25% increments; we additionally evaluate an equal 1:1:1 mixture. Overall, we evaluate 160 model variants across 2 tasks, 16 compositions, and 5 training epochs. To distinguish genuine composition effects from the performance expected by independently combining each regime, we define the expected mixture performance and Complexity Mixture Gain (CMG) at epoch e as

$$G_{\text{exp}}(w, e) = \sum_{k \in \{E,M,H\}} w_k G_k(e), \qquad \text{CMG}(w, e) = G_{\text{mix}}(w, e) - G_{\text{exp}}(w, e), \tag{2}$$

where $w = (w_E, w_M, w_H)$ denotes the composition ratio with $\sum_k w_k = 1$. Here, $G_k(e)$ denotes the OOD performance at epoch e of a model trained on a single complexity level k, and $G_{\text{mix}}(w, e)$ denotes that of mixed composition w at epoch e. For example, for E2/M2/H0, where $w = (0.5, 0.5, 0)$, $G_{\text{exp}}(w, e) = 0.5G_E(e) + 0.5G_M(e)$. We additionally report the average mixture gain across five epochs, $\overline{\text{CMG}}(w) = \frac{1}{5}\sum_{e=1}^{5} \text{CMG}(w, e)$. Together, these comparisons assess whether $G_{\text{mix}}(w, e)$ exceeds the single-level baselines $G_E(e)$, $G_M(e)$, and $G_H(e)$, and whether the mixture provides additional gain beyond the composition-weighted linear expectation as measured by $\text{CMG}(w, e)$.

Appropriate complexity composition improves generalization beyond its linear expectation. As shown in Fig. 4, no single complexity level is universally optimal; instead, mixture gains depend strongly on both task and optimization stage. The E2/M1/H1 composition is the most consistent across tasks, yielding $\overline{\text{CMG}} = +0.511\%$p for General VQA and $+0.493\%$p for Reasoning VQA, with positive gains in 9 of 10 task–epoch combinations. More specialized mixtures can achieve

Table 2: Correlation between model-based complexity measures and downstream performance relative to $P_{\text{ref}}(\mathcal{M})$ at the 64K scale. Additional measures and visualizations are provided in § A.6.

Complexity Measure	General VQA OOD Perf.		Reasoning VQA OOD Perf.		Reasoning VQA ID Perf.	
	$\mathbf{R^2}$	Spearman's ρ	$\mathbf{R^2}$	Spearman's ρ	$\mathbf{R^2}$	Spearman's ρ
PPL	0.328	-0.539	0.321	-0.580	0.242	-0.494
IFD	0.333	-0.565	0.322	-0.579	0.263	-0.504
Gradient Norm	0.319	0.471	0.335	0.560	0.270	0.501
Conflict Severity	0.308	0.513	0.074	0.212	0.085	0.242

larger task-specific gains—E1/M0/H3 reaches $+1.343\%$p for General VQA, whereas E3/M1/H0 reaches $+0.798\%$p for Reasoning VQA—but do not transfer equally across tasks. Notably, when the existing correctness store constructed using Qwen3.5-0.8B is directly transferred to LFM2.5-VL-450M (AI, 2025), these composition patterns largely disappear. In contrast, recomputing learner-specific correctness for Qwen3-VL-4B-Instruct (Team, 2025) and InternVL3.5-1B (Wang et al., 2025) recovers positive composition gains; full results are provided in § A.5. In particular, training on mixed compositions outperforms the best single-level baseline at multiple epochs, E2/M1/H1 repeatedly yields positive CMG, and E1/M0/H3 consistently improves over H4 in General VQA. Together, these results suggest that the utility of complexity is learner-dependent and should be calibrated through composition and optimization rather than reduced to a single preferred level.

3.4 Evaluating Complexity Measures

As data-centric measures could be used to guide data curation, we next investigate whether existing complexity measures can serve as proxies for complexity-aware data curation. Following the measure evaluation paradigm of Jung et al. (2025), we assess how well complexity measures predict downstream generalization on the same controlled subsets (§ 3.1). Given the learner-relative effects observed in § 3.3, we compare 4 model-based complexity measures: Perplexity, Instruction Following Difficulty (IFD) (Li et al., 2024a), Gradient Norm (Paul et al., 2021), and Gradient Conflict Severity.[5] Finally, we average sample-wise complexity values within each dataset to obtain a single complexity score, and quantify its association with performance (Eq. 1) using Spearman's ρ and R^2, reporting the better fit between linear and log-linear trends.

As shown in Tab. 2, existing complexity measures show weak-to-moderate correlations with generalization. However, no single measure consistently emerges as a strong predictor. In light of this result, we retain the learner-relative rollout estimate for complexity-aware data curation in CoRAL Synthesis (§ 4). Additional results and details for a broader set of complexity measures, including text- and image-based measures, are provided in § A.6.

4 CoRAL Synthesis

4.1 Method

Building on the empirical findings in § 3, we propose **CoRAL Synthesis**, a scalable synthetic data generation framework for vision-language reasoning that jointly controls data complexity, diversity, and quality. As illustrated in Fig. 1 (left), our framework iteratively follows a *generation* $\rightarrow$ *calibration* process, generating diversified and quality-verified data pools at scale and sequentially calibrating their complexity to the target learner.[6] Below, we describe each stage of the framework.

Generation, Diversification, and Quality Filtering. Following § 3.1, we use the same image-conditioned few-shot synthesis setup with Qwen3.5-27B as both the problem and solution generator

[5] Prior work (Yu et al., 2020) identifies gradient conflict through negative cosine similarity between gradients. We adapt this notion to a dataset-level statistic, $F_{\text{severity}} = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \max(0, -\cos(g_i, g_j))$, where $\mathcal{P}$ denotes the set of gradient pairs. To mitigate quadratic all-pair computation, we estimate the mean using 100K uniformly sampled distinct pairs per subset.

[6] Motivated by Davidson et al. (2026), which shows that quality filtering matters more as complexity increases, we apply quality filtering before complexity calibration.

in a two-stage process (Guha et al., 2026), additionally including ChartBench (Xu et al., 2023) among the Reasoning VQA source datasets.

For diversity selection, we adapt the gradient-space diversification of Prismatic Synthesis (Jung et al., 2025) to vision-language data. Using Qwen3.5-0.8B as a proxy model, we compute normalized VLM loss gradients, reduce them to d=1024 with Rademacher random projection (Rakhshan & Rabusseau, 2020; Park et al., 2023), and cluster the current data pool in gradient space, globally for General VQA and separately for each Reasoning domain; newly generated candidates are retained only when they fall into sparsely occupied clusters. We empirically determine N_{domain}, the number of problems synthesized per image, based on the point at which G-Vendi (Jung et al., 2025) saturates, and accordingly generate at most $(x_i, I)_{i=1}^{N_{\text{domain}}}$ for each image I, with $N_{\text{General,Math,Science,Chart}} = (8, 12, 8, 4)$; details are provided in § B.2.

After diversification, all selected samples undergo a full quality verification using Double-Critic filtering (Davidson et al., 2026) with Qwen3.5-27B, which independently examines evidence for correctness and incorrectness to mitigate critic sycophancy (Sharma et al., 2024). Only diversified samples passing this verification are admitted to the final training pool. Further details of the verification and benchmark decontamination procedures are provided in § B.3.

Learner-Calibrated Complexity. Our results in § 3.3 show that generalization depends on the composition of complexity levels rather than on a single preferred level. We therefore calibrate the generated data toward an appropriate Easy–Medium–Hard (E/M/H) composition for the target learner. Based on our analysis of existing measures in § 3.4, we estimate complexity directly from the target learner's response correctness.

We calibrate dataset complexity in two stages: estimating the learner-relative complexity of each candidate and selecting samples to match a target complexity composition.

- **Phase 1: Constructing the correctness datastore.** For each training example x, we sample N responses from the target learner $\mathcal{L}$ and compute the number of correct responses, $c_{\mathcal{L}}(x) \in \{0, \ldots, N\}$, as its learner-relative correctness score (we set $N = 8$). This follows the same procedure as in § 3.1. We divide the range of the $N+1$ possible correctness counts into three equal-width intervals. Examples in the high-count range are labeled Easy, those in the middle range Medium, and those in the low-count range Hard. Each example is then assigned a complexity label $y \in \{\text{E}, \text{M}, \text{H}\}$, yielding the correctness datastore $\mathcal{D} = \{(x, y)\}$.
- **Phase 2: Calibrating the complexity composition.** Given a target complexity composition ratio $\mathbf{r} = (r_E, r_M, r_H)$, we select verified candidates from $\mathcal{D}$ to match $\mathbf{r}$, using the most underrepresented tier as the limiting tier. Guided by our empirical findings in § 3.3, we adopt $\mathbf{r} = (0.5, 0.25, 0.25)$ (*i.e.*, 2 : 1 : 1 for E, M, and H), while $\mathbf{r}$ remains configurable for different settings.

By iteratively executing this *generation* → *calibration* pipeline, we construct our final synthetic datasets, CoRAL-GeneralVQA and CoRAL-ReasoningVQA, comprising 624K General VQA and 569K Reasoning VQA samples, respectively, with the latter consisting of 193K math, 110K science, and 264K chart reasoning examples.

Experimental Setup. To evaluate the effectiveness of CoRAL-GeneralVQA and CoRAL-ReasoningVQA, we fine-tune Qwen3-VL-4B-Instruct (Team, 2025) and compare it against recent strong off-the-shelf VLMs of comparable scale, including InternVL3.5-4B, Molmo2-4B, Mage-VL-4B, and Gemma-4-E4B-IT (Wang et al., 2025; Clark et al., 2026; Yang et al., 2026; Team et al., 2026). For General VQA, we use the benchmarks from § 3.1 and additionally include SimpleVQA (Cheng et al., 2025). For Reasoning VQA, we retain DynaMath, $\text{MathVision}_{\text{testmini}}$, MMMU_{val}, $\text{MMSci}_{\text{dev}}$, and $\text{InfoVQA}_{\text{val}}$, and further include $\text{MathVerse}_{\text{testmini}}$ (Zhang et al., 2024), MMMU-Pro-Vision (Yue et al., 2025), and $\text{CharXiv}_{\text{val}}$ (DQ/RQ, both included in the aggregate score) (Wang et al., 2024b), extending the evaluation to more challenging reasoning tasks.

4.2 Results

Main Results. The results in Tab. 3 show that our 4B models achieve strong OOD performance across both General and Reasoning VQA benchmarks. The consistent gains across both evalua-

Table 3: OOD benchmark performance of representative open VLMs and synthetic-data-trained models. All scores are mean pass@1 over 5 runs. (Top) General VQA results across 5 unseen benchmarks, showing competitive performance against other 4B open VLMs. (Bottom) Reasoning VQA results across 9 unseen benchmarks, with our CoRAL-ReasoningVQA-4B improving the average score by 5.30%p over the best fully evaluated 4B baseline. Results with standard errors are provided in § B.4.

Model	RealWorldQA	MMStar	MMBenchEN	HallusionBench	SimpleVQA	Avg.
Qwen3-VL-4B-Instruct	71.29	65.96	88.80	62.02	47.37	67.09
Molmo2-4B	**75.95**	62.04	87.92	59.47	34.83	64.04
Mage-VL-4B	73.15	66.35	87.54	63.29	31.53	64.37
Gemma-4-E4B-IT	54.88	59.09	74.88	61.10	30.81	56.15
InternVL3.5-4B	66.20	65.72	86.00	60.14	38.93	63.40
CoRAL-GeneralVQA-4B	74.20	**68.79**	**89.25**	**65.72**	**47.52**	**69.10**

Model	DynaMath	MathVision	MathVerse	MMMU	MMMU Pro-V	MMSci	CharXiv		InfoVQA	Avg.
							DQ	RQ		
Qwen3-VL-4B-Instruct	59.10	36.97	42.65	58.13	26.31	73.05	77.53	34.06	75.23	53.67
Molmo2-4B	47.93	24.28	30.58	51.60	19.90	64.69	77.23	33.90	59.77	45.54
Mage-VL-4B	56.24	23.88	40.95	51.87	26.00	66.12	75.74	35.28	77.92	50.44
Gemma-4-E4B-IT	**66.33**	41.45	**54.64**	57.04	28.44	63.80	**83.44**	36.18	45.23	52.95
InternVL3.5-4B	59.81	32.89	37.78	57.62	31.79	70.29	76.11	**42.60**	73.22	53.57
CoRAL-ReasoningVQA-4B	66.05	**42.17**	52.23	**58.31**	**34.84**	**73.72**	81.60	42.08	**79.69**	**58.97**

tion suites suggest that the benefit of our synthetic data extends beyond task-specific improvements to broader OOD generalization. Importantly, the improvements are not limited to a single reasoning domain, suggesting that CoRAL-ReasoningVQA provides broadly useful supervision on math, science, and chart reasoning benchmarks. Overall, the results suggest that synthetic data design benefits from considering not only quality and diversity, but also complexity, and that such a synthesis framework boosts generalization across unseen benchmarks spanning general visual question answering and vision-language reasoning.

Consistent Scaling and Data Efficiency. Beyond unseen benchmark performance, an effective synthetic data generation pipeline should be scalable. To assess whether CoRAL Synthesis provides consistent scaling gains over vanilla few-shot generation, we compare the two approaches across data scales on General VQA, using the same OOD benchmarks used in § 3.1. At each scale, we keep the target dataset size (n = 1K, 4K, 8K, 16K, 64K) and seed-source composition fixed; vanilla data are generated with standard few-shot prompting (as in § 3.1), whereas ours are generated with CoRAL Synthesis, with all models post-trained from Qwen3.5-0.8B. Additional experimental details are provided in § B.4. As shown in Fig. 1 (right), under this setting, naively increasing the amount of synthetic data does not yield consistent OOD improvements. In contrast, our framework improves steadily with scale, from 65.28 at 1K to 66.99 at 64K. Notably, CoRAL Synthesis already surpasses the best-performing vanilla setting, despite using $16\times$ fewer synthetic examples. Overall, these results highlight both the consistent scaling behavior and the data-efficiency benefit of composition-aware curation of CoRAL Synthesis.

5 Conclusion

We establish data complexity as an important axis of synthetic vision-language data design, showing that it affects downstream generalization even when data scale, quality, and diversity are controlled. Our analysis shows that appropriately mixing Easy, Medium, and Hard examples improves generalization in a learner-relative manner, while existing complexity measures provide limited guidance as standalone objectives for composition-aware curation. We introduce CoRAL Synthesis, which jointly promotes data quality and diversity while calibrating complexity to the target learner, and release CoRAL-GeneralVQA and CoRAL-ReasoningVQA. Models trained on these datasets achieve strong OOD generalization, while CoRAL Synthesis exhibits consistent scaling gains and improved data efficiency over vanilla few-shot generation.

AI USE STATEMENT

In this work, we used generative AI tools to "generate synthetic data sets, clean and reformat dataset". We have not used generative AI tools to "assist with translation, implement methods, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, interpret results", and "formulate mathematical claims, assist in the writing of proofs, provide critical ingredients for proving mathematical claims" are not applicable to this work. Additionally, we used generative AI tools to "create or edit software code, draft parts of a research paper, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature". We have reviewed all AI-assisted work. VLM was used to generate and perform quality checks on synthetic data as part of our data synthesis pipeline. All AI-assisted code, manuscript drafts, and edits were reviewed multiple times by all authors to verify their correctness and appropriateness. For literature search, generative AI was used only to assist in retrieving relevant literature; the authors independently reviewed the original sources and made all judgments regarding their relevance and content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our synthetic VQA datasets reuse images from existing sources, whose licenses and underlying image rights differ; we retain source identifiers and image provenance rather than treating the combined data as having a single license (Tab. 9). In particular, MathVista-derived samples are subject to the source's restriction on commercial model-training use, and any distribution or reuse of derived data should respect the applicable source terms. We also used paid human judgments to assess the semantic correctness verifier; the worker qualifications, compensation, and quality-control procedure are reported in § A.1. Synthetic questions and answers may contain errors or reflect biases in the source images and generator, so automated verification does not remove the need for care when using or releasing these data.

REPRODUCIBILITY STATEMENT

The main text describes the controlled complexity experiments and the generation–selection procedure in § 3 and § 4.1. The appendix documents the seed sources, generation prompts and settings, correctness verification, and training setup in § A.1 and § B.1. Subset construction and diversity controls are specified in § A.3, while the quality filters and benchmark evaluation protocol are detailed in § B.3 and § B.4.

REFERENCES

David Acuna, Chao-Han Huck Yang, Yuntian Deng, Jaehun Jung, Ximing Lu, Prithviraj Ammanabrolu, Hyunwoo Kim, Yuan-Hong Liao, and Yejin Choi. Long grounded thoughts: Synthesizing visual problems and reasoning chains at scale. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=ZipucsUgc6.

Liquid AI. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*, 2025.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3411–3425, 2024.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, 2021.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3313–3323, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.218. URL https://aclanthology.org/2022.emnlp-main.218/.

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.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4637–4646. IEEE, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Tim R. Davidson, Benoit Seguin, Enrico Bacis, Cesar Ilharco, and Hamza Harkous. Reasoning-driven synthetic data generation and evaluation. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=NALsdGEPhB.

C. DeMars. *Item Response Theory*. Understanding Statistics. Oxford University Press, 2010. ISBN 9780199703845. URL https://books.google.co.kr/books?id=KOADeYBt7sIC.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, pp. 5988–6008. PMLR, 2022.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and Zhiqiang Xu. Principled data selection for alignment: The hidden risks of difficult examples. *arXiv preprint arXiv:2502.09650*, 2025.

Yunhao Gou, Hansi Yang, Zhili Liu, Kai Chen, Yihan Zeng, Lanqing Hong, Zhenguo Li, Qun Liu, Bo Han, James Kwok, and Yu Zhang. Corrupted but not broken: Understanding and mitigating the negative impacts of corrupted data in visual instruction tuning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25934–25960, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1317. URL https://aclanthology.org/2025.emnlp-main.1317/.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. Openthoughts: Data recipes for reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 108059–108130, 2026.

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, et al. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*, 2024.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=R0dC7Xzwbk.

Jaehun Jung, Hyunwoo Kim, Brandon Cui, Ximing Lu, David Acuna, Prithviraj Ammanabrolu, and Yejin Choi. Deltaprompts: Escaping the zero-delta trap in multimodal distillation, 2026. URL https://arxiv.org/abs/2605.15532.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

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.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5376–5384. IEEE, 2017.

Samuel Kessler, Menglin Xia, Daniel Madrigal, Dongge Han, Helia Hashemi, Saravan Rajmohan, Victor Rühle, and Jordan T. Ash. Towards active synthetic data generation for finetuning language models. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=c2pJvSu6Qo.

Yeganeh Kordi, Nihal V. Nayak, Max Zuo, Ilana Nguyen, and Stephen Bach. Revisiting generalization across difficulty levels: It's not so easy. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7014–7042, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.330. URL https://aclanthology.org/2026.eacl-long.330/.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, 2024a.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*, 2024b.

Yuan-Hong Liao, Sven Elflein, Liu He, Laura Leal-Taixé, Yejin Choi, Sanja Fidler, and David Acuna. Longperceptualthoughts: Distilling system-2 reasoning for system-1 perception. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=SrKdi4MsUW.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods. *arXiv preprint arXiv:2601.21821*, 2026.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 22353–22373, 2024a.

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, 2024b.

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, et al. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *arXiv preprint arXiv:2508.08221*, 2025.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 36456–36474, 2024a.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

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, 2024b.

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, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591. IEEE, 2022.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pp. 15630–15649. PMLR, 2022.

Surabhi S Nath, Kevin Shen, Aenne Annelie Brielmann, and Peter Dayan. Simplicity in complexity. In *ICLR 2024 Workshop on Representational Alignment*, 2024. URL `https://openreview.net/forum?id=DHvVdakpqO`.

Xinzhe Ni, Yeyun Gong, Zhibin Gou, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Exploring the mystery of influential data for mathematical reasoning. *arXiv preprint arXiv:2404.01067*, 2024.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-dig: Towards gradient-based diverse and high-quality instruction data selection for machine translation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15395–15406, 2024.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Runqi Qiao, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Guanting Dong, Shiqiang Lang, Sitong Zhou, Yida Xu, Yuchen Zeng, et al. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 150171–150218, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind. In *Asian Conference on Computer Vision*, pp. 293–309. Springer, 2024.

Beheshteh Rakhshan and Guillaume Rabusseau. Tensorized random projections. In *International Conference on Artificial Intelligence and Statistics*, pp. 3306–3316. PMLR, 2020.

Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

Ruth Rosenholtz, Yuanzhen Li, and Lisa Nakano. Measuring visual clutter. *Journal of vision*, 7(2): 17–17, 2007.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Cheng. How to train data-efficient llms. In *International Conference on Learning Representations*, volume 2026, pp. 45491–45545, 2026.

Gabriel Sarch, Linrong Cai, Qunzhong Wang, Haoyang Wu, Danqi Chen, and Zhuang Liu. Vero: An open rl recipe for general visual reasoning. *arXiv preprint arXiv:2604.04917*, 2026.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Taiwei Shi, Yiyang Wu, Linxin Song, Tianyi Zhou, and Jieyu Zhao. Efficient reinforcement finetuning via adaptive curriculum learning. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=UEhpyq41b9.

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Advances in Neural Information Processing Systems*, 37:7821–7846, 2024.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-vision dataset. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=QWTCcxMpPA.

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.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024b.

Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*, 2023.

Zijian Wu, Jinjie Ni, Xiangyan Liu, Zichen Liu, Hang Yan, and Michael Qizhe Shieh. SynthRL: Scaling visual reasoning with verifiable data synthesis. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42456–42476, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2107. URL https://aclanthology.org/2026.findings-acl.2107/.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.

Senqiao Yang, Kaichen Zhang, Zhaoyang Jia, Jinghao Guo, Yifei Shen, Xinjie Zhang, Xiaoyi Zhang, Haoqing Wang, Xiao Li, Peng Zhang, et al. Mage-vl: An efficient codec-native streaming multimodal foundation model. *arXiv preprint arXiv:2607.24904*, 2026.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2376–2385. IEEE, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

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.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in neural information processing systems*, 36:5484–5505, 2023.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yi Zhang, Bolin Ni, Xin-Sheng Chen, Hengrui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Winston Hu, Meng-Hao Guo, and Shi-Min Hu. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. In *International Conference on Learning Representations*, volume 2026, pp. 13769–13836, 2026.

Yingxiu Zhao, Bowen Yu, Binyuan Hui, Haiyang Yu, Minghao Li, Fei Huang, Nevin L Zhang, and Yongbin Li. Tree-instruct: A preliminary study of the intrinsic relationship between complexity and alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16776–16789, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, volume 2025, pp. 48337–48383, 2025.

DISCUSSION

Limitations and Future Work. Our learner-relative calibration is computationally expensive because each candidate requires eight rollouts from the target learner. This cost limits how readily the method can be applied to larger data pools or recalibrated during training. The E/M/H ratio we adopt is empirical and may change with the task, learner, or a finer search over mixtures, so it should not be viewed as a global optimum. Our released datasets are curated for Qwen3-VL-4B-Instruct; using this approach for another learner requires either synthesizing data with our framework for that model or re-estimating difficulty and curating the existing pool accordingly. Moreover, correctness counts capture the outcome of an attempt but do not distinguish visual perception failures from ambiguous questions or noisy supervision. Future work could develop cheaper, more diagnostic difficulty signals and use them to adapt the mixture as the learner changes.

Our empirical evaluation centers on image-based General VQA and Reasoning VQA. These tasks cover a range of visual questions, but they do not represent every way a vision-language model is used. For example, web navigation requires decisions across changing interfaces, while embodied control links visual understanding to actions. Long-form video understanding adds temporal dependencies that are largely absent from our current setting. Future work should test whether learner-relative calibration remains useful in these settings and how their complexity tiers and mixtures should be defined.

Our synthesis pipeline generates new questions and solutions while reusing images from existing datasets. This lets us vary textual supervision for a fixed visual input, but it does not create new visual scenes. Consequently, the range of visual concepts and compositions remains tied to the seed image pool. A more complete multimodal pipeline could generate images with diffusion models or modify them through structured image editing, then verify that the resulting image, question, and answer remain consistent. Future work could integrate these tools into the generation–selection loop to control visual and textual complexity jointly.

Alternative Synthetic Data Recipe. An inexpensive extension would estimate learner-relative Easy/Medium/Hard tiers on a modest seed set and allocate generation across these tiers, rather than obtain eight rollouts for every synthetic candidate. Seed selection can steer the aggregate difficulty of synthetic data, although individual generated examples need not retain their seeds' difficulty (Kessler et al., 2026). Visual-reasoning synthesis can also shift the resulting distribution toward intermediate difficulty (Wu et al., 2026). We could therefore score a small pilot batch from each seed tier, estimate how its outputs are distributed across the three tiers, and adjust the tier-specific generation budgets to approximate the desired composition. Periodic audits of newly generated examples could detect further drift as the generator or target learner changes, without rescoring the full data pool. This seed-guided recipe offers a way to test whether learner-relative composition can be maintained at substantially lower inference cost.

A Complexity Experiment Details

A.1 Synthetic Data Generation Details

Generation Statistics. We synthesize data from a diverse set of seed datasets spanning General and Reasoning VQA in § 3. We use Qwen3.5-27B as data generator and take a 2-step process, generating problems and then solving each problems Guha et al. (2026). As summarized in Tab. 4, General VQA uses 194,833 unique seed images from GQA and VQAv2, yielding 1,077,591 samples after verification. Reasoning VQA uses 187,495 unique seed images across nine source datasets, yielding 998,075 verified samples.

Table 4: Statistics of the synthetic data pools used for the complexity experiments. We report the number of unique seed images and samples retained after verification.

Seed Dataset	Unique Seed Images	QA Samples
General VQA		
GQA	72,140	423,379
VQAv2	122,693	654,212
Total	**194,833**	**1,077,591**
Reasoning VQA		
MathVista	6,107	78,657
We-Math2.0-Standard	5,843	66,048
UniGeo	9,543	72,893
GeoQA3-English	4,985	37,751
VLMsAreBlind	3,876	49,019
ScienceQA	10,332	127,251
AI2D	4,060	51,947
TQA	2,749	37,936
FigureQA	140,000	476,573
Total	**187,495**	**998,075**

Prompts.

Problem generation

Given an image and an example visual problem, create exactly {num_outputs} more similar or harder problems for the same image.
The new problem must require using the image to answer.
Each new problem should be formatted exactly as:

```
---
[[Problem]]
your new problem - come up with a non-multiple choice problem, even if the provided example is multiple choice.
---
```

Output exactly {num_outputs} [[Problem]] blocks and nothing else. Do not write analysis, answers, rationales, labels, bullets, or placeholders. Examples:

Example 1 *(seed question)*

```
---
[[Problem]]
{seed_question}
---
```

Example 2

```
---
[[Problem]]
{example_question}
---
```

⋮

Solution generation

Solve the following problem. Make sure to put your final answer in \boxed{}.

```
{problem}
```

Visual QA verification

Task

Decide whether to keep this visual QA example.

Keep it only if the visual problem, proposed solution, and boxed answer are factually correct and logically sound. Reject it if any part is incorrect, unsupported, or unclear.

Problem:
{problem}

Proposed solution:
{solution}

Proposed boxed answer:
{boxed_answer}

Output format

Output exactly one of these two labels:

\boxed{YES} = keep
\boxed{NO} = reject

Final output only:

Correctness Verification. Although responses are constrained to a \boxed{} answer format, exact matching can be unreliable for non-deterministic answers; we therefore use Qwen3.5-27B as a semantic verifier to determine whether each of the $N = 8$ sampled answers is semantically consistent with the synthesized gold answer. We additionally validate the semantic verifier on 1,000 manually annotated examples collected via Amazon Mechanical Turk (MTurk)[7]. We recruit 50 qualified workers with at least 10,000 approved HITs and an approval rate of at least 98%. Each worker evaluates 20 examples together with two trap questions for quality control, receiving $1.50 for an estimated 10–12 minutes of work. After trap-based quality filtering, 920 annotations remain. The semantic verifier achieves 89.78% agreement with human judgments, with Cohen's $\kappa = 0.7925$.

[7] https://www.mturk.com

A.2 Qualitative Examples

Example 1 — General VQA

Easy

Target Learner: Qwen3-VL-4B-Instruct; correct rollouts: 8/8.



Problem

Identify the specific green and white vegetable toppings visible on the pizza to the left of the pepperoni pizza.

Solution

Looking at the pizza on the left (the one without pepperoni), we can clearly see:
• Green toppings: These are sliced green bell peppers — identifiable by their curved, ring-like shape and vibrant green color. • White toppings: These are thinly sliced onions — visible as translucent white rings or arcs scattered across the pizza.
These are common vegetable toppings for pizzas and match the visual evidence in the image.

Final answer

green bell peppers and onions

Example 2 — General VQA

Medium

Target Learner: Qwen3-VL-4B-Instruct; correct rollouts: 4/8.



Problem

Describe the relative positions of the two giraffes in the image.

Solution

In the image, the two giraffes are standing close together on the left side of the frame, silhouetted against a warm sunset sky.
• The giraffe on the left is slightly behind and to the left of the other giraffe. • The giraffe on the right is positioned slightly in front and to the right of the first one. • Their bodies overlap slightly, with the right giraffe's neck and head appearing more forward in the composition. • Both are facing toward the right side of the image, suggesting they may be looking in the same direction.
Thus, their relative positions can be described as: The right giraffe is slightly in front of and to the right of the left giraffe.

Final answer

The right giraffe is slightly in front of and to the right of the left giraffe.

Example 3 — General VQA

Hard

Target Learner: Qwen3-VL-4B-Instruct; correct rollouts: 0/8.



Problem

Describe the relative position of the orange fish plush toy compared to the brown teddy bear.

Solution

The orange fish plush toy is positioned in front of and slightly below the brown teddy bear. The teddy bear is sitting upright on the wooden bench, while the fish plush lies horizontally across the bear's lap or lower body area, with its head pointing toward the left side of the image.
Thus, the fish is in the foreground relative to the bear, and partially overlapping it from the viewer's perspective.

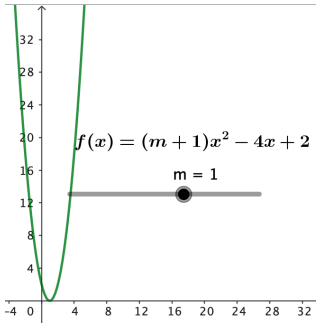
Final answer

in front of and slightly below

Example 4 — Reasoning VQA: Math

Easy

Target Learner: Qwen3.5-0.8B; correct rollouts: 8/8.



Problem

Given the function $f(x) = (m+1)x^2 - 4x + 2$ with the parameter m set to the value indicated by the slider in the figure, calculate the exact value of $f(3)$.

Solution

We are given the function:

$$f(x) = (m+1)x^2 - 4x + 2$$

and from the figure, the slider indicates that $m = 1$.
Substitute $m = 1$ into the function:

$$f(x) = (1+1)x^2 - 4x + 2 = 2x^2 - 4x + 2$$

Now compute $f(3)$:

$$f(3) = 2(3)^2 - 4(3) + 2 = 2(9) - 12 + 2 = 18 - 12 + 2 = 8$$

Final answer

$$\boxed{8}$$

Example 5 — Reasoning VQA: Math

Medium

Target Learner: Qwen3.5-0.8B; correct rollouts: 4/8.

$$\begin{pmatrix} \mathbf{1} & \mathbf{-2} \\ \mathbf{-2} & \mathbf{4} \\ \mathbf{1} & \mathbf{-2} \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \mathbf{3} & \mathbf{2} & \mathbf{1} \end{pmatrix} = \mathbf{?}$$

Problem

Calculate the element in the second row and third column of the resulting matrix from the multiplication shown in the figure.

Solution

...[omitted]
We want row 2, column 3.
So:
• Row 2 of first matrix: $\begin{pmatrix} -2 & 4 \end{pmatrix}$ • Column 3 of second matrix: $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$
Step 2: Compute dot product

$$(-2)(3) + (4)(1) = -6 + 4 = -2$$

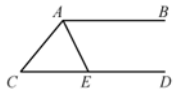
Final answer

$$\boxed{-2}$$

Example 6 — Reasoning VQA: Math

Hard

Target Learner: Qwen3.5-0.8B; correct rollouts: 0/8.



Problem

As shown in the figure, lines AB and CD are parallel. The ray AE bisects the angle $\angle CAB$. If the measure of $\angle C$ is $50°$, calculate the measure of $\angle AED$.

Solution

...[omitted]
Let's analyze the figure step by step.
Since $AB \parallel CD$, and line AC is a transversal, then consecutive interior angles (also called same-side interior angles) are supplementary.
That is:

$$\angle CAB + \angle C = 180°$$

Given $\angle C = 50°$, so:

$$\angle CAB = 180° - 50° = 130°$$

Now, ray AE bisects $\angle CAB$, so:

$$\angle CAE = \angle EAB = \frac{130°}{2} = 65°$$

...[omitted]
So:

$$\angle AED = 180° - \angle AEC = 180° - 65° = 115°$$

...[omitted]

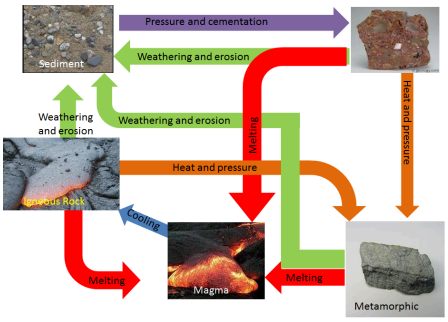
FINAL ANSWER

115

Example 7 — Reasoning VQA: Science

Easy

Target Learner: Qwen3.5-0.8B; correct rollouts: 8/8.



PROBLEM

Identify the specific process labeled on the red arrow that originates from the Igneous Rock image and points towards the Magma image.

SOLUTION

Looking at the diagram:
• The Igneous Rock image is in the bottom left. • A red arrow originates from it and points downward and then rightward toward the Magma image (center bottom). • Along this red arrow, the label clearly reads: "Melting"
This makes sense geologically — igneous rocks can be melted (via heat) to form magma.

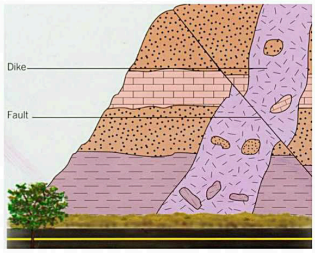
FINAL ANSWER

Melting

Example 8 — Reasoning VQA: Science

Medium

Target Learner: Qwen3.5-0.8B; correct rollouts: 4/8.



PROBLEM

Based on the principle of cross-cutting relationships shown in the diagram, determine which event occurred most recently: the deposition of the horizontal sedimentary layers, the formation of the fault, or the intrusion of the dike.

SOLUTION

To determine the relative timing of geological events, we apply the principle of cross-cutting relationships, which states that any geologic feature that cuts across another is younger than the feature it cuts.
Let's analyze the diagram:
1. Deposition of horizontal sedimentary layers – These are the layered rocks (with different patterns: dotted, brick-like, dashed) that form the background. They must have been deposited first, as they are cut by both the fault and the dike.
2. Formation of the fault – The fault line cuts through the sedimentary layers, meaning the fault occurred *after* those layers were deposited.
3. Intrusion of the dike – The purple dike cuts through *both* the sedimentary layers *and* the fault. This means the dike must be younger than both the layers and the fault.

Therefore, the most recent event is the intrusion of the dike.

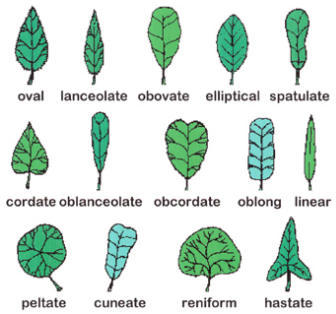
FINAL ANSWER

the intrusion of the dike

Example 9 — Reasoning VQA: Science

Hard

Target Learner: Qwen3.5-0.8B; correct rollouts: 0/8.



PROBLEM

Compare the ‘elliptical’ and ‘obovate’ leaf shapes shown in the top row. Describe the specific difference in the width distribution of the two leaves relative to their midpoints.

SOLUTION

To compare the elliptical and obovate leaf shapes shown in the top row, we examine how their width is distributed relative to their midpoints (i.e., halfway along the length of the leaf from base to tip).
Elliptical Leaf: • The elliptical leaf is widest at or near its midpoint. • It tapers gradually and symmetrically toward both the base and the tip. • So, the maximum width occurs approximately at the center of the leaf’s length.
Obovate Leaf: • The obovate leaf is widest above the midpoint, closer to the tip. • It tapers more sharply toward the base. • So, the maximum width occurs in the upper half of the leaf, not at the center.
...[omitted]

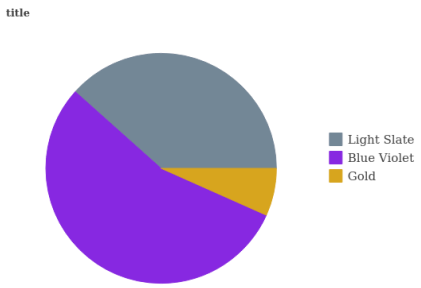
FINAL ANSWER

The elliptical leaf is widest at its midpoint, while the obovate leaf is widest above its midpoint, closer to the tip.

Example 10 — Reasoning VQA: Chart

Easy

Target Learner: Qwen3.5-0.8B; correct rollouts: 8/8.



PROBLEM

List the categories in the pie chart in descending order of their percentage share.

SOLUTION

To determine the descending order of percentage shares in the pie chart, we visually compare the sizes of the slices:

1. Blue Violet — This is the largest slice, clearly occupying more than half of the pie chart. 2. Light Slate — This is the second-largest slice, significantly smaller than Blue Violet but much larger than Gold. 3. Gold — This is the smallest slice, a small wedge compared to the other two.
Thus, in descending order of percentage share:
Blue Violet > Light Slate > Gold

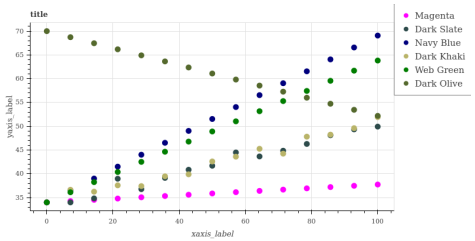
FINAL ANSWER

Blue Violet, Light Slate, Gold

Example 11 — Reasoning VQA: Chart

Medium

Target Learner: Qwen3.5-0.8B; correct rollouts: 4/8.



PROBLEM

Identify the two series that exhibit the steepest positive linear trends across the entire x-axis range.

SOLUTION

To identify the two series with the steepest positive linear trends, we need to look at how much each series increases in y-value as x goes from 0 to 100.
Let's examine the key series:
1. Navy Blue (dark blue dots) • Starts near y=35 at x=0 • Ends near y=69 at x=100 • $\Delta y \approx 69 - 35 = 34$ • Very steep, consistent upward trend
2. Web Green (green dots) • Starts at y=34 at x=0 • Ends at y=64 at x=100 • $\Delta y \approx 64 - 34 = 30$ • Also very steep and consistent
...[omitted]

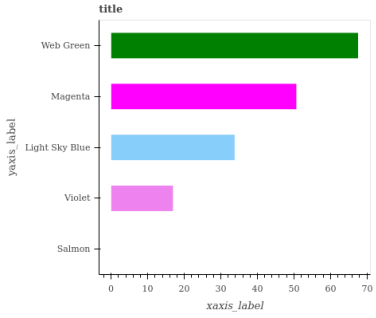
FINAL ANSWER

Navy Blue, Web Green

Example 12 — Reasoning VQA: Chart

Hard

Target Learner: Qwen3.5-0.8B; correct rollouts: 0/8.



PROBLEM

Calculate the total value of the three longest bars.

SOLUTION

To solve this problem, we need to:
1. Identify the lengths (values) of all bars from the bar chart. 2. Select the three longest bars. 3. Sum their values.

From the chart:
• Web Green: extends to approximately 68 • Magenta: extends to approximately 51 • Light Sky Blue: extends to approximately 34 • Violet: extends to approximately 17 • Salmon: has no visible bar → value = 0
So the values are: → Web Green: 68 → Magenta: 51 → Light Sky Blue: 34 → Violet: 17 → Salmon: 0
The three longest bars are: Web Green (68), Magenta (51), and Light Sky Blue (34)
Now sum them:
68 + 51 + 34 = 153

FINAL ANSWER

$$\boxed{153}$$

A.3 SUBSET SAMPLING

Complexity-Guided Subset Sampling. As detailed in Algorithm 1, we vary complexity while controlling for domain composition and image redundancy using the learner-relative correctness scores described in § 3.1. We first convert correctness scores into domain-wise percentile ranks and sample examples from target complexity ranges while preserving the original domain proportions. Within each domain, we preferentially select at most one question per image to maximize image coverage. When additional questions from the same images are required, we use Qwen3-Embedding-8B to prioritize semantically dissimilar questions based on cosine similarity, thereby reducing semantic duplication. This procedure yields subsets that differ primarily in complexity while keeping domain balance and diversity controlled.

Algorithm 1 *Complexity-Guided Subset Sampling*

Input: Dataset $\mathcal{D}$, complexity score (*i.e.*, correctness) $\{c_i\}$, domains $\{d_i\}$, image keys $\{m_i\}$, problem texts $\{p_i\}$, subset size K
Input: Anchor set $\mathcal{A}$, variant set $\mathcal{V}$, similarity threshold τ
Output: A collection of sampled subsets $\mathcal{O}$

$\mathcal{O} \leftarrow \emptyset$ ▷ Initialize the output collection.
$q_i \leftarrow \mathrm{PercentileRank}(c_i \mid d_i)$ ▷ Compute domain-wise complexity percentiles.
$K_d \leftarrow K \cdot |\mathcal{D}_d|/|\mathcal{D}|$ ▷ Preserve the original domain ratio.
for each anchor $a \in \mathcal{A}$ and variant $v \in \mathcal{V}$ **do**
  $\mathcal{S} \leftarrow \emptyset$ ▷ Initialize one sampled subset.
  for each domain d **do**
    $\mathcal{C}_d \leftarrow \mathrm{Candidates}(\mathcal{D}_d, q, a, v)$ ▷ Select samples from the target complexity range.
    $\mathcal{C}_d \leftarrow \mathrm{ExpandUntil}(\mathcal{C}_d, K_d)$ ▷ Expand the quantile range if needed.
    $\mathcal{S}_d \leftarrow \mathrm{UniqueImageSelect}(\mathcal{C}_d, K_d)$ ▷ Prefer at most one sample per image.
    if $|\mathcal{S}_d| < K_d$ **then**
      $\mathcal{S}_d \leftarrow \mathrm{QuestionDiverseFill}(\mathcal{S}_d, \mathcal{C}_d, K_d, \tau)$ ▷ Fill with non-duplicate and semantically diverse questions.
    end if
    $\mathcal{S} \leftarrow \mathcal{S} \cup \mathcal{S}_d$ ▷ Merge domain-level samples.
  end for
  $\mathcal{O} \leftarrow \mathcal{O} \cup \{\mathrm{Shuffle}(\mathcal{S})\}$ ▷ Store the sampled subset.
end for
return $\mathcal{O}$ ▷ Return all sampled subsets.

Diversity-Controlled Sampling. To isolate the effects of image and text diversity studied in § B.2, we construct 64K-example subsets in three configurations: low-image/high-text (LI/HT), high-image/low-text (HI/LT), and high-image/high-text (HI/HT).[8] As detailed in Algorithm 2, we embed questions with Qwen3-Embedding-8B and unique images with InceptionV3, then independently cluster the two modalities into 512 groups each within the task pool. Within each source

[8] We omit LI/LT because fixing the data scale at 64K while simultaneously enforcing low diversity in both modalities would require excessive sample reuse, making the condition ill-posed.

domain, high-diversity sampling upweights rare clusters with inverse-frequency power 4, whereas low-diversity sampling upweights dense clusters with frequency power 8 and restricts selection to the densest 15% of occupied clusters. For low image diversity, we initially target image anchors equal to 10% of the source quota and add anchors when needed to supply enough questions; this yields approximately 10.6K unique images in General VQA and 11.5K in Reasoning VQA per 64K-example subset. For high image diversity, we select one question per image first, yielding 64K unique images in both tasks. We multiply the image and question weights and sample without replacement while exactly preserving the original source-dataset proportions within each task.

Algorithm 2 *Modality-Controlled Subset Sampling*

Input: Dataset $\mathcal{D}$, subset size N, condition $c \in \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$, source quotas $\{N_s\}$
Input: diversity parameters $\alpha, \beta, \eta, \rho_I$
Output: Sampled subset $\mathcal{S}_c$

$q_i \leftarrow \mathrm{Normalize}(f_Q(x_i))$, $v_i \leftarrow \mathrm{Normalize}(f_I(I_i))$ ▷ Extract question and image embeddings.
$z_i^Q \leftarrow \mathrm{SphericalKMeans}(\{q_i\})$, $z_i^I \leftarrow \mathrm{SphericalKMeans}(\{v_i\})$ ▷ Cluster each modality independently.

$$(t_I, t_Q) \leftarrow \begin{cases} (\mathrm{low}, \mathrm{high}) & c = \mathrm{LH}, \\ (\mathrm{high}, \mathrm{low}) & c = \mathrm{HL}, \\ (\mathrm{high}, \mathrm{high}) & c = \mathrm{HH}. \end{cases}$$

$\mathcal{S}_c \leftarrow \emptyset$
for each source group s **do**
  $\mathcal{D}_s \leftarrow \{i \in \mathcal{D} : \mathrm{source}(i) = s\}$
  for each modality $m \in \{I, Q\}$ **do**
    $w_i^m \leftarrow \mathrm{ClusterWeight}(z_i^m, t_m; \alpha, \beta, \eta)$ ▷ Favor rare clusters for high diversity and dense clusters for low diversity.
  end for
  $w_i \leftarrow w_i^I w_i^Q$ ▷ Combine modality-specific weights.
  if $t_I = \mathrm{low}$ **then**
    $\mathcal{C}_s \leftarrow \mathrm{SelectImageAnchors}(\mathcal{D}_s, w, \rho_I N_s)$ ▷ Restrict sampling to a small set of high-capacity images.
  else
    $\mathcal{C}_s \leftarrow \mathrm{UniqueImageFirst}(\mathcal{D}_s)$ ▷ Prefer at most one question per image.
  end if
  $\pi_i \leftarrow \log w_i + \epsilon_i$, $\epsilon_i \sim \mathrm{Gumbel}(0, 1)$ for $i \in \mathcal{C}_s$
  $\mathcal{S}_c \leftarrow \mathcal{S}_c \cup \mathrm{TopK}(\mathcal{C}_s, \pi, N_s)$ ▷ Perform weighted sampling without replacement.
end for
return $\mathcal{S}_c$

Synthesis and Training Setup. For data synthesis and correctness verification, we use up to 8 H200 GPUs to parallelize problem generation, answer generation, and semantic verification. For the subset-level experiments, we perform full-parameter fine-tuning for all evaluated models (Qwen3.5-0.8B, Qwen3-VL-4B-Instruct, InternVL3.5-1B) independently on each sampled subset. Owing to the relatively small model sizes used in these experiments, each training and inference run fits on a single RTX 6000 Ada Generation GPU, allowing us to parallelize distinct runs across up to 16 GPUs. The same pool of up to 16 GPUs is used for all auxiliary computations, including $N = 8$ rollout sampling, PPL and IFD scoring, gradient-based measurements, and the extraction of text, image, and multimodal embeddings. Exceptionally, for large-scale training on the full 1M dataset, we use Distributed Data Parallel training across 4 to 8 GPUs.

A.4 Additional Different Complexity Level Performance Results

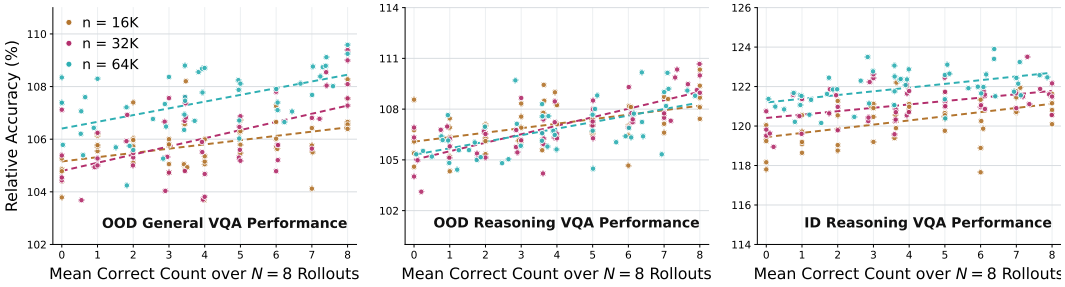
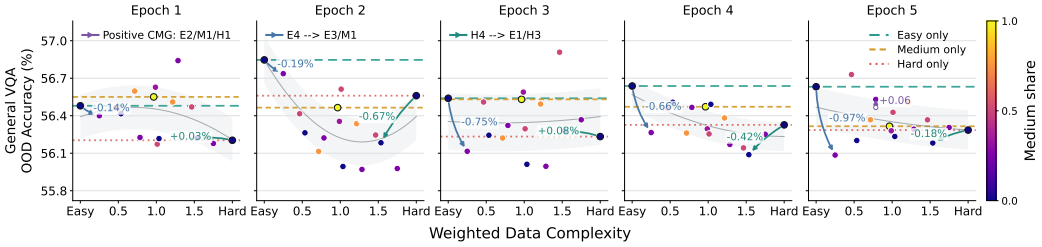
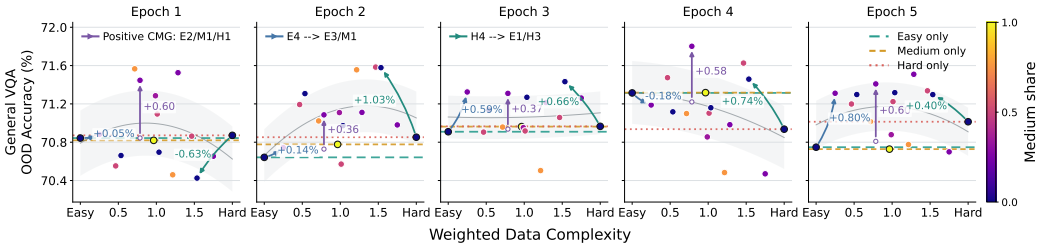


Figure 5: Complexity and OOD/ID performance relative to $P_{\text{base}}(\mathcal{M})$.

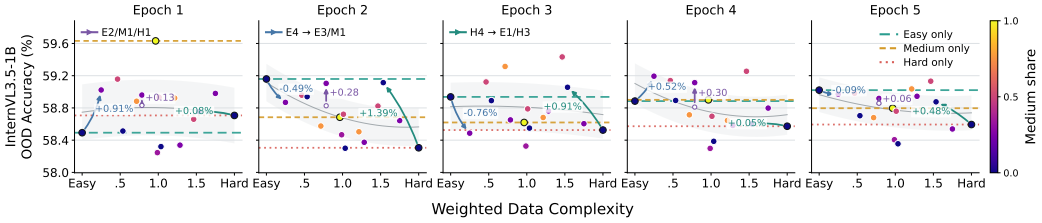
A.5 Additional Complexity Composition Results



(a) LFM2.5-VL-450M with transferred complexity assignments.



(b) Qwen3-VL-4B-Instruct with learner-specific complexity assignments.



(c) Internvl3.5-1B with learner-specific complexity assignments.

Figure 6: Full 5-epoch results for the complexity-composition and learner-transfer analyses.

Table 5: 64K General VQA composition results. Each Epoch reports OOD accuracy (%) over 4 benchmarks, with its CMG (%p) in parentheses. $\overline{\text{CMG}}$ averages CMG over five Epochs (%p). Bold and underline mark the best and second-best OOD and mean-CMG scores.

E/M/H	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5	$\overline{\text{CMG}}$
E4/M0/H0	<u>66.18</u> (0.000)	65.66 (0.000)	65.60 (0.000)	65.52 (0.000)	65.40 (0.000)	+0.000
E3/M1/H0	65.66 (-0.294)	65.32 (-0.461)	65.77 (+0.135)	65.81 (+0.413)	65.79 (+0.184)	-0.004
E2/M2/H0	65.62 (-0.106)	66.30 (+0.403)	65.10 (-0.576)	65.02 (-0.254)	<u>66.48</u> (+0.658)	+0.025
E3/M0/H1	65.58 (-0.081)	65.69 (-0.055)	65.78 (+0.297)	65.83 (+0.710)	65.89 (+0.727)	+0.319
E1/M3/H0	65.50 (-0.002)	<u>66.78</u> (+0.766)	65.91 (+0.197)	65.51 (+0.354)	65.60 (-0.427)	+0.177
E2/M1/H1	65.13 (-0.300)	66.33 (+0.466)	66.01 (+0.494)	65.65 (+0.657)	**66.61** (+1.236)	+0.511
E0/M4/H0	65.27 (0.000)	66.13 (0.000)	65.75 (0.000)	65.04 (0.000)	66.23 (0.000)	+0.000
E1/M1/H1	64.72 (-0.463)	66.40 (+0.464)	66.28 (+0.790)	**66.36** (+1.536)	66.23 (+0.865)	+0.638
E1/M2/H1	66.06 (+0.852)	65.29 (-0.691)	66.38 (+0.829)	65.43 (+0.553)	65.66 (+0.084)	+0.326
E2/M0/H2	**66.51** (+1.373)	65.91 (+0.076)	65.14 (-0.217)	65.72 (+1.003)	65.77 (+0.847)	+0.616
E0/M3/H1	66.17 (+1.196)	65.55 (-0.545)	66.09 (+0.502)	65.87 (+1.109)	66.26 (+0.470)	+0.546
E1/M1/H2	65.16 (+0.250)	64.92 (-1.028)	65.43 (+0.039)	64.48 (-0.115)	64.71 (-0.429)	-0.257
E0/M2/H2	65.47 (+0.780)	65.95 (-0.120)	65.31 (-0.122)	65.12 (+0.646)	65.89 (+0.551)	+0.347
E1/M0/H3	66.11 (+1.494)	**66.96** (+1.041)	<u>66.39</u> (+1.157)	<u>66.08</u> (+1.762)	65.95 (+1.263)	**+1.343**
E0/M1/H3	65.43 (+1.035)	65.66 (-0.378)	**67.02** (+1.752)	65.64 (+1.437)	65.43 (+0.536)	<u>+0.876</u>
E0/M0/H4	64.10 (0.000)	66.01 (0.000)	65.11 (0.000)	63.92 (0.000)	64.45 (0.000)	+0.000

Table 6: 64K Reasoning VQA composition results. Each Epoch reports OOD accuracy (%) over 7 benchmarks, with its CMG (%p) in parentheses. $\overline{\text{CMG}}$ averages CMG over five Epochs (%p). Bold and underline mark the best and second-best OOD and mean-CMG scores.

E/M/H	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5	$\overline{\text{CMG}}$
E4/M0/H0	49.36 (0.000)	48.99 (0.000)	<u>49.89</u> (0.000)	49.74 (0.000)	49.91 (0.000)	+0.000
E3/M1/H0	49.19 (+0.249)	**50.70** (+1.576)	**50.76** (+1.020)	<u>50.51</u> (+0.711)	**50.40** (+0.434)	**+0.798**
E2/M2/H0	**49.84** (+1.311)	49.18 (-0.076)	49.69 (+0.104)	**50.52** (+0.661)	49.98 (-0.031)	+0.394
E3/M0/H1	48.43 (-0.360)	49.22 (+0.314)	49.26 (-0.274)	50.15 (+0.661)	49.58 (+0.070)	+0.082
E1/M3/H0	<u>49.37</u> (+1.258)	49.10 (-0.283)	49.62 (+0.182)	49.18 (-0.743)	49.91 (-0.149)	+0.053
E2/M1/H1	49.24 (+0.855)	<u>50.18</u> (+1.146)	49.70 (+0.315)	49.69 (+0.141)	49.56 (+0.005)	<u>+0.493</u>
E0/M4/H0	47.70 (0.000)	49.52 (0.000)	49.29 (0.000)	49.98 (0.000)	<u>50.11</u> (0.000)	+0.000
E1/M1/H1	48.02 (-0.038)	48.22 (-0.833)	49.57 (+0.353)	49.20 (-0.275)	49.05 (-0.390)	-0.237
E1/M2/H1	47.92 (-0.044)	48.98 (-0.193)	49.20 (-0.041)	49.39 (-0.214)	49.68 (+0.075)	-0.083
E2/M0/H2	46.93 (-1.299)	48.88 (+0.057)	48.66 (-0.525)	49.13 (-0.102)	48.73 (-0.368)	-0.447
E0/M3/H1	48.11 (+0.561)	48.20 (-1.098)	48.70 (-0.387)	47.99 (-1.677)	48.28 (-1.379)	-0.796
E1/M1/H2	48.09 (+0.267)	47.72 (-1.233)	48.92 (-0.119)	48.87 (-0.424)	49.06 (-0.088)	-0.319
E0/M2/H2	48.65 (+1.250)	48.65 (-0.430)	48.95 (+0.066)	49.32 (-0.030)	48.44 (-0.759)	+0.019
E1/M0/H3	47.67 (-0.001)	48.38 (-0.350)	48.48 (-0.353)	49.04 (+0.069)	48.74 (+0.044)	-0.118
E0/M1/H3	48.41 (+1.152)	48.27 (-0.592)	48.44 (-0.247)	48.33 (-0.707)	47.96 (-0.785)	-0.236
E0/M0/H4	47.11 (0.000)	48.65 (0.000)	48.48 (0.000)	48.72 (0.000)	48.29 (0.000)	+0.000

A.6 Additional Complexity Measure Results

Table 7: Relationship between complexity measures and model performance. We report R^2 and Spearman's ρ between subset-level complexity measure scores and performance relative to $P_{\text{ref}}(\mathcal{M})$ at the 64k scale.

Category	Complexity Measure	General VQA OOD Performance		Reasoning VQA OOD Performance		Reasoning VQA ID Performance	
		$\mathbf{R^2}$	Spearman's ρ	$\mathbf{R^2}$	Spearman's ρ	$\mathbf{R^2}$	Spearman's ρ
Text	Token Length	0.332	-0.518	0.318	-0.577	0.241	-0.513
	Tree Instruct	0.332	-0.521	0.331	-0.580	0.235	-0.484
	InsTag	0.235	-0.511	0.322	-0.577	0.281	-0.539
Image	Segmentation Mask Count	0.033	0.177	0.116	0.356	0.228	0.435
	Edge Density	0.203	-0.473	0.370	-0.603	0.305	-0.536
	Subband Entropy	0.307	-0.503	0.260	-0.427	0.270	-0.523
Model	PPL	0.328	-0.539	0.321	-0.580	0.242	-0.494
	IFD	0.333	-0.565	0.322	-0.579	0.263	-0.504
	Gradient Norm	0.319	0.471	0.335	0.560	0.270	0.501
	Conflict Severity	0.308	0.513	0.074	0.212	0.085	0.242

Table 8: Relationship between complexity measures and model performance. We report R^2 and Spearman's ρ between subset-level complexity measure scores and performance relative to $P_{\text{base}}(\mathcal{M})$ at the 64k scale.

Category	Complexity Measure	General VQA OOD Performance		Reasoning VQA OOD Performance		Reasoning VQA ID Performance	
		$\mathbf{R^2}$	Spearman's ρ	$\mathbf{R^2}$	Spearman's ρ	$\mathbf{R^2}$	Spearman's ρ
Text	Token Length	0.346	-0.539	0.353	-0.586	0.225	-0.508
	Tree Instruct	0.345	-0.541	0.369	-0.593	0.224	-0.489
	InsTag	0.254	-0.539	0.356	-0.595	0.268	-0.545
Image	Segmentation Mask Count	0.030	0.172	0.132	0.381	0.192	0.429
	Edge Density	0.221	-0.500	0.409	-0.619	0.304	-0.548
	Subband Entropy	0.324	-0.530	0.279	-0.427	0.272	-0.527
Model	PPL	0.343	-0.564	0.352	-0.595	0.225	-0.503
	IFD	0.348	-0.587	0.355	-0.593	0.243	-0.511
	Gradient Norm	0.331	0.490	0.370	0.575	0.255	0.509
	Conflict Severity	0.326	0.536	0.070	0.197	0.073	0.243

Computing complexity measures. We evaluate the same ten measures on the controlled subsets described in § 3.1, independently of the correctness counts used to construct those subsets. The text-based measures are problem Token Length (Liu et al., 2024a), Tree-Instruct (Zhao et al., 2024), and InsTag (Lu et al., 2024a); the image-based measures are Segmentation Mask Count (Nath et al., 2024), Edge Density, and Subband Entropy (Rosenholtz et al., 2007). The model-based measures are response Perplexity (PPL), Instruction Following Difficulty (IFD) (Li et al., 2024a), Gradient Norm (Paul et al., 2021), and Gradient Conflict Severity (Yu et al., 2020). PPL and IFD are computed with Qwen3.5-0.8B, using the image and question as conditioning input; IFD is the ratio of the conditioned solution loss to the solution-only loss. For every measure except conflict severity, we average example-level values over the subset to obtain one subset score. For conflict severity, we instead average $\max(0, -\cos(g_i, g_j))$ over 100K uniformly sampled distinct pairs of example gradients from that subset, avoiding an all-pairs computation.

Relating measure scores to generalization. At the 64K data scale, we compare each subset score with the downstream performance of the same base model fine-tuned on that subset, separately for General VQA OOD, Reasoning VQA OOD, and Reasoning VQA ID evaluation. Performance is normalized benchmark by benchmark against either the full-pool reference model, $P_{\text{ref}}(\mathcal{M})$, or the base model before subset fine-tuning, $P_{\text{base}}(\mathcal{M})$, as defined in Eq. 1. Tab. 7 reports the reference-normalized results, and Tab. 8 reports the corresponding base-normalized results. Fig. 7 and Fig. 8 show these visualizations.

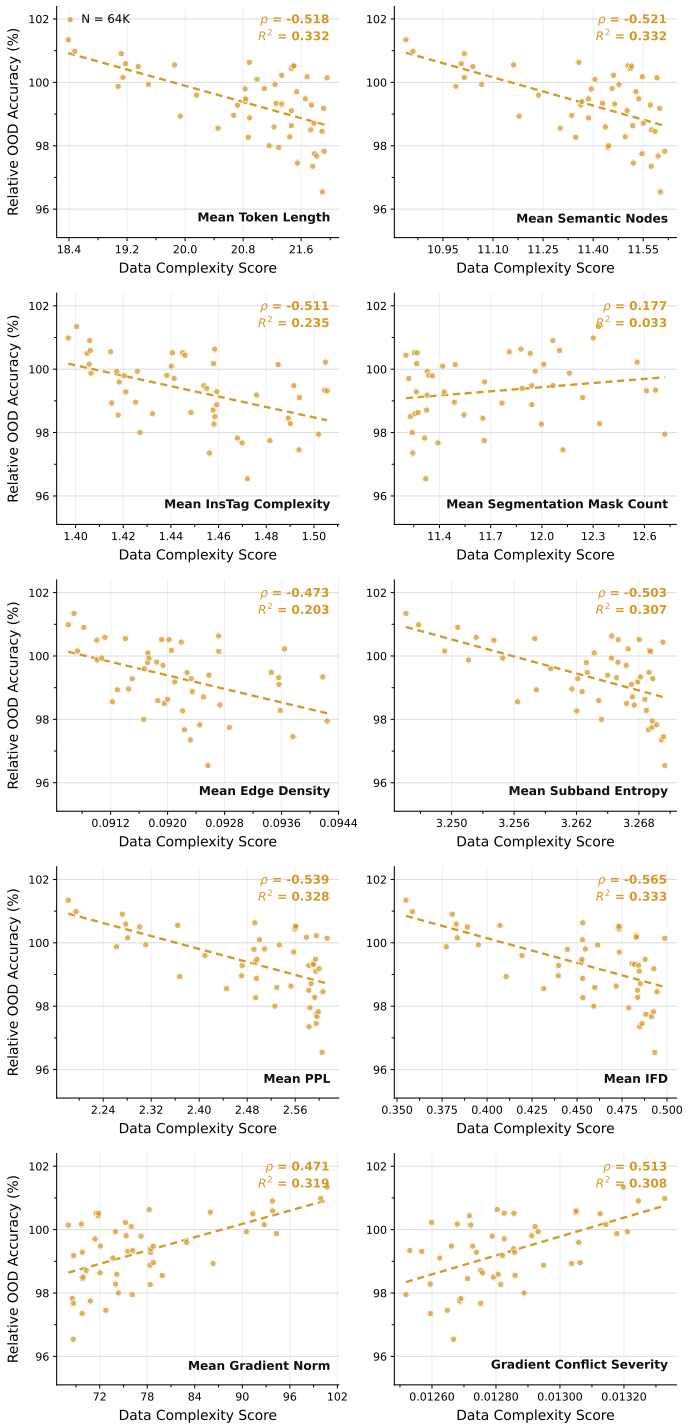


Figure 7: Additional correlations between complexity measures and OOD generalization performance ($P_{\text{ref}}(\mathcal{M})$) on General VQA at the 64K scale.

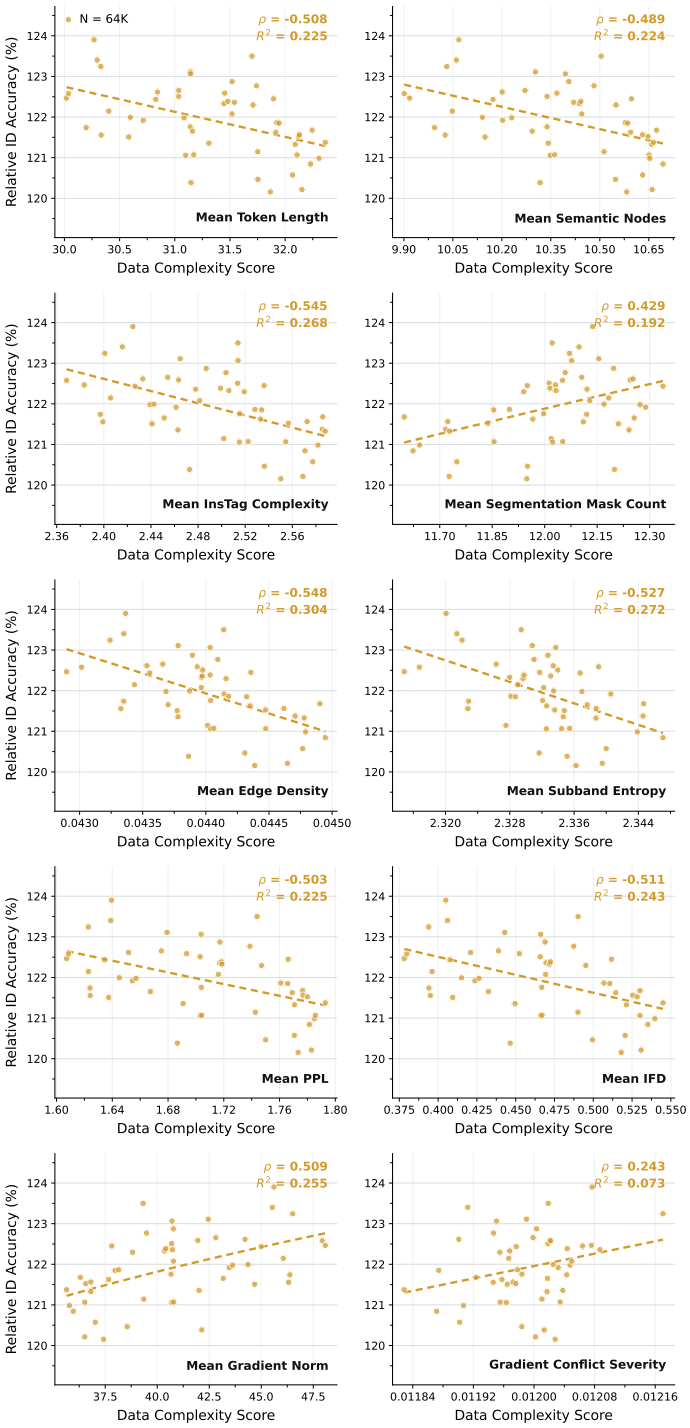


Figure 8: Additional correlations between complexity measures and ID generalization performance ($P_{\text{base}}(\mathcal{M})$) on Reasoning VQA at the 64K scale.

Table 9: Seed image sources and retained synthetic question–answer (QA) samples. The 272,228 source-specific image records correspond to 264,680 distinct files after cross-source deduplication; QA counts are after final composition selection.

Seed source	Images	QA samples	License
General VQA			
VQAv2	75,888	377,716	CC BY 4.0 (annotations)
GQA	49,112	246,412	–
Total	125,000	624,128	
Reasoning VQA: Math			
GeoQA3-English	4,954	30,396	–
UniGeo	9,506	47,219	–
MathVista	6,141	50,832	CC BY-SA 4.0
VLMsAreBlind	3,696	28,927	MIT
We-Math2.0-Standard	5,827	36,565	CC BY-NC 4.0
Reasoning VQA: Science			
AI2D	4,058	27,301	CC BY-SA
TQA	2,749	19,163	CC BY-NC 3.0
ScienceQA	10,297	64,261	CC BY-NC-SA 4.0
Reasoning VQA: Chart			
FigureQA	33,376	131,819	–
ChartBench	66,624	132,778	MIT
Total	147,228	569,261	

License note. "–" indicates no explicit dataset license in the source documentation; AI2D does not specify a CC BY-SA version. VQAv2's CC BY 4.0 covers annotations. MathVista's CC BY-SA 4.0 covers new contributions, while original images and questions retain source rights. Within Reasoning VQA, we prohibit commercial model-training use of MathVista-derived samples, consistent with MathVista's published terms.

B CoRAL Synthesis Details

B.1 Generate Samples

Seed images and source provenance. We construct CoRAL-GeneralVQA and CoRAL-ReasoningVQA from the General VQA and Reasoning VQA seed sources introduced in § 3.1, with ChartBench additionally included for chart reasoning. Tab. 9 reports both the source distribution of the image pool before synthesis (125,000 General VQA and 147,228 Reasoning VQA image records) and the number of retained synthetic question–answer samples from each source (624,128 and 569,261, respectively). Each record has a unique source-specific image identifier; after deduplicating identical image files across sources, the pool contains 264,680 distinct images. Tab. 9 also records the published license or terms for each seed source, distinguishing dataset terms from software licenses and rights in third-party images. We retain source identifiers and image provenance for every generated example rather than assigning a single license to the combined pool.

Generation setup. We use Qwen3.5-27B (Qwen Team, 2026) for both stages of image-conditioned synthesis: generating new questions and then generating their solutions. For each image, the prompt contains its original seed question–answer pair and up to four previously generated pairs as textual demonstrations (same as § A.1). The generation limits are 2,048 tokens for a question and 8,192 tokens for a solution. For both tasks, we use temperature 1.0, top-p 0.95, top-k 20, and presence penalty 1.5. Generation was run with up to 16 H200 GPUs operating in parallel. The subsequent quality verification, learner-relative correctness estimation, and $N{=}8$ response sampling follow the procedures in § A.1.

B.2 Cross-Modal Diversification for Synthetic VLMs

Having established the role of complexity, we next consider diversity as a complementary factor that is particularly important for vision-language synthetic data generation (Acuna et al., 2026). Unlike

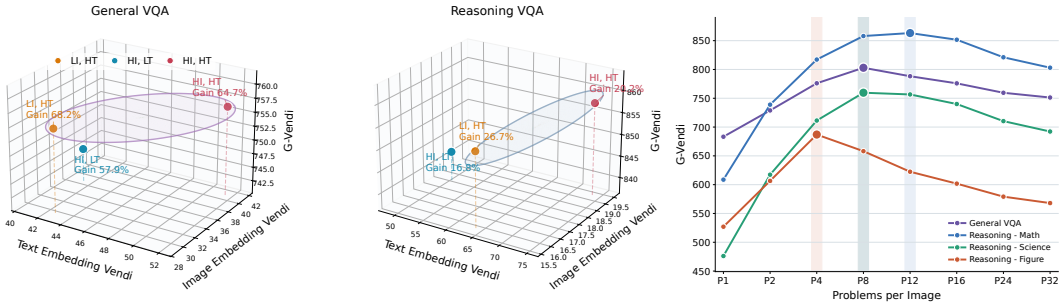


Figure 9: Modality-wise diversity and per-image supervision saturation. (Left, middle) We compare text, image, and gradient-space diversity across LI/HT, HI/LT, and HI/HT configurations for General and Reasoning VQA. (Right) Scaling the number of problems per image increases G-Vendi only up to a task- and domain-specific saturation point. **Overall, text diversification can enrich learning signals, but its useful per-image scaling range varies across tasks and domains.**

text-only settings, where supervision can often be expanded through simple few-shot synthesis, VLM synthesis can diversify either visual inputs or their associated text and therefore requires more careful control over which modality provides useful variation. We therefore examine how image and text diversity affect downstream performance and gradient-space learning signals, and further study how much diverse supervision can be synthesized per image.

Text diversity provides useful cross-modal learning signals but exhibits saturation. We first control the data scale at 64K and construct three configurations with different image/text embedding diversity: low-image/high-text (LI/HT), high-image/low-text (HI/LT), and high-image/high-text (HI/HT), following the sampling procedure in § A.3. We quantify semantic diversity using image/text embedding Vendi scores (Friedman & Dieng, 2022) (following § 3.1) and gradient diversity using G-Vendi (Jung et al., 2025)—a strong measure that predicts generalization (with gradients from a Qwen3.5-0.8B proxy)—while reporting downstream gains relative to $\mathcal{M}_{\text{ref}}$. As shown in Fig. 9 (left, middle), across both General and Reasoning VQA, higher text diversity is consistently associated with improved downstream performance, supporting the value of diverse supervision within the same visual context.

To assess whether gradient diversity depends on the choice of VLM parameters, we compute G-Vendi using gradients from the vision encoder, projector, language model, and their combinations. Tab. 10 reports the dimensionality of each component, and Tab. 11 compares the resulting diversity scores. Although component-specific scores differ, full-VLM gradients show the same key pattern in both tasks: HI/HT yields the highest G-Vendi, while LI/HT exceeds HI/LT. We therefore use gradients from the full VLM for the main diversity analysis, without selecting a particular component.

We next sample 4K unique images and scale the number of synthesized problems per image using the same few-shot generation procedure as in § 3.1, from P1 to P32 (*i.e.*, P1, P2, P4, P8, P12, P16, P24, and P32, corresponding to 4k–128k samples), to examine how long this common per-image expansion strategy continues to produce useful gradient diversity. As shown in Fig. 9 (right), G-Vendi increases only up to a task- and domain-specific point and subsequently saturates or decreases. For example, the early saturation in chart reasoning suggests that the generator exhausts its capacity to produce distinct, learning-relevant problem–solution patterns earlier in this domain than in others. Thus, higher per-image text diversity can benefit synthetic VLM generation, but its useful scaling range depends on the target task and domain.

Table 10: Raw gradient dimensionality of Qwen3.5-0.8B by model component.

Component	Gradient Dimension	Ratio
Vision Encoder	88,004,352	10.32%
Projector	12,588,544	1.48%
Language Model	752,393,024	88.21%
Full VLM	852,985,920	100%

Table 11: Gradient-space diversity across VLM components at the 64K scale. We report G-Vendi computed from gradients over individual components and their combinations under low-image/high-text (LI/HT), high-image/low-text (HI/LT), and high-image/high-text (HI/HT) configurations. Best and second-best results in each row are shown in **bold** and underline, respectively.

Gradient Component	LI/HT	HI/LT	HI/HT
General VQA			
Vision	927.87	928.11	**929.34**
Projector	907.44	904.84	**914.74**
Language Model	505.95	481.48	**511.35**
Vision + Projector	929.95	930.23	**931.46**
Projector + Language Model	507.97	483.60	**513.38**
Vision + Language Model	628.54	607.61	**636.61**
Full VLM	755.28	746.22	**756.32**
Reasoning VQA			
Vision	**745.60**	702.59	707.47
Projector	771.79	760.76	**775.96**
Language Model	620.83	594.80	**645.45**
Vision + Projector	**752.35**	710.22	715.31
Projector + Language Model	622.21	596.35	**646.91**
Vision + Language Model	680.70	653.58	**696.41**
Full VLM	853.43	843.13	**857.53**

B.3 Quality Verification Details

Double Critic. After gradient-based selection, we apply Double-Critic filtering (Davidson et al., 2026) with Qwen3.5-27B: a critic asks whether the visual QA example is fully valid, while an error critic asks whether it contains any error. Both critics receive the same image, problem, proposed solution, and boxed answer, but make their judgments independently without seeing the other's response. We disable thinking and sample one response per critic with temperature zero and a maximum of eight output tokens; the system messages require a single label. A candidate enters the training pool only when the correctness critic returns $\boxed{\text{YES}}$ and the error critic returns $\boxed{\text{NO}}$. Unrecognized labels and generation failures are treated as rejections. The prompts are shown below.

Correctness critic

Is this visual QA example fully correct?

The problem must be clear and answerable from the image. The solution must be factually and logically correct, and the boxed answer must follow from the solution.

Candidate

Problem:
{problem}
Proposed solution:
{solution}
Proposed boxed answer:
{boxed_answer}

Output format

Output exactly:

\boxed{YES} = fully correct
\boxed{NO} = otherwise

Error critic

Does this visual QA example contain any error?

An error includes an unclear or unanswerable problem, unsupported image interpretation, incorrect reasoning, or an incorrect boxed answer.

Candidate

Problem:
{problem}
Proposed solution:
{solution}
Proposed boxed answer:
{boxed_answer}

Output format

Output exactly:

\boxed{YES} = contains an error
\boxed{NO} = no error found

Benchmark Decontamination. We perform VQA benchmark decontamination in two stages. First, we identify training examples whose images exactly match evaluation images using source identifiers or exact image hashes when available. Among these image-matched candidates, we compare the associated questions using normalized 8-gram overlap and remove examples with highly similar questions. Thus, text-based filtering is applied only to examples sharing the same evaluation image, targeting multimodal leakage while avoiding unnecessary removal of generic VQA questions.

Table 12: Full OOD benchmark performance of representative open VLMs and synthetic-data-trained models. Scores are mean pass@1 over 5 runs, with standard errors of the mean shown in parentheses.

Model	RealWorldQA	MMStar	MMBenchEN	HallusionBench	SimpleVQA	Avg.
Qwen3-VL-4B-Instruct	71.29 (0.28)	65.96 (0.12)	88.80 (0.07)	62.02 (0.33)	47.37 (0.29)	67.09 (0.07)
Molmo2-4B	**75.95** (0.23)	62.04 (0.49)	87.92 (0.16)	59.47 (1.35)	34.83 (0.20)	64.04 (0.36)
Mage-VL-4B	73.15 (0.21)	66.35 (0.28)	87.54 (0.23)	63.29 (0.41)	31.53 (0.60)	64.37 (0.35)
Gemma-4-E4B-IT	54.88 (0.33)	59.09 (0.33)	74.88 (0.16)	61.10 (0.38)	30.81 (0.28)	56.15 (0.11)
InternVL3.5-4B	66.20 (0.16)	65.72 (0.16)	86.00 (0.03)	60.14 (0.29)	38.93 (0.05)	63.40 (0.10)
CoRAL-GeneralVQA-4B	74.20 (0.25)	**68.79** (0.49)	**89.25** (0.08)	**65.72** (0.40)	**47.52** (0.39)	**69.10** (0.16)

Model	DynaMath	MathVision	MathVerse	MMMU	MMMU Pro-V	MMSci	CharXiv		InfoVQA	Avg.
							DQ	RQ		
Qwen3-VL-4B-Instruct	59.10 (0.20)	36.97 (3.87)	42.65 (0.12)	58.13 (0.34)	26.31 (0.30)	73.05 (0.24)	77.53 (0.10)	34.06 (0.47)	75.23 (0.11)	53.67 (0.45)
Molmo2-4B	47.93 (0.17)	24.28 (0.85)	30.58 (0.33)	51.60 (0.22)	19.90 (0.16)	64.69 (0.18)	77.23 (0.62)	33.90 (0.92)	59.77 (0.27)	45.54 (0.18)
Mage-VL-4B	56.24 (0.30)	23.88 (0.20)	40.95 (0.37)	51.87 (0.53)	26.00 (0.28)	66.12 (0.31)	75.74 (0.45)	35.28 (0.29)	77.92 (0.55)	50.44 (0.33)
Gemma-4-E4B-IT	**66.33** (0.15)	41.45 (0.63)	**54.64** (0.20)	57.04 (1.05)	28.44 (0.21)	63.80 (1.81)	**83.44** (0.10)	36.18 (0.42)	45.23 (0.20)	52.95 (0.32)
InternVL3.5-4B	59.81 (0.13)	32.89 (0.43)	37.78 (0.10)	57.62 (1.08)	31.79 (0.06)	70.29 (0.40)	76.11 (0.02)	**42.60** (0.15)	73.22 (0.03)	53.57 (0.14)
CoRAL-ReasoningVQA-4B	66.05 (0.20)	**42.17** (0.67)	52.23 (0.20)	**58.31** (0.42)	**34.84** (0.30)	**73.72** (0.12)	81.60 (0.15)	42.08 (0.27)	**79.69** (0.15)	**58.97** (0.08)

B.4 Evaluating CoRAL Synthesis

CoRAL-GeneralVQA and CoRAL-ReasoningVQA. We evaluate the models trained on CoRAL-GeneralVQA and CoRAL-ReasoningVQA alongside the baselines on 5 General VQA and 9 Reasoning VQA benchmarks, respectively, using the selected split and scoring protocol for each benchmark. All models are fine-tuned for 2 epochs. Whereas the controlled subset experiments in § 3.1 compare the relative effects of complexity, this broader OOD evaluation assesses how models trained on the final curated datasets perform against representative open VLMs across unseen tasks. We score answers using VLMEvalKit (Duan et al., 2024) and use GPT-5-mini with low reasoning effort as the judge for benchmarks requiring model-based evaluation. We report mean pass@1 over 5 evaluation runs (different seeds) and the standard error of the mean. Full benchmark-level results are provided in Tab. 12.

Scaling comparison with vanilla few-shot generation. To examine whether the gains of CoRAL Synthesis persist as the training set grows, we compare General VQA subsets drawn from a vanilla few-shot pool (§ 3.1) with subsets drawn from our curated pool (§ 4.1) while matching dataset size and seed-source proportions. For the vanilla baseline, we construct a 64K few-shot subset and randomly sample the smaller 1K–16K subsets from it, without complexity calibration. For CoRAL Synthesis, we first construct the 64K subset from candidates selected for diversity and verified for quality, then form nested smaller subsets while preserving a learner-relative Easy–Medium–Hard composition using the target model's correctness estimates. We post-train the same Qwen3.5-0.8B base model on each subset and evaluate OOD performance using the benchmarks in § 3.1.