

DecoupleMix: Decoupled Ratio Search and Convex Allocation for Scalable VLM Data Recipes

Jiahao Xie* Zhongbin Guo*[†] Qianle Wang Ruiqi Lu Dongling Xiao
Wanxuan Sun[✉] Cheng Yang[✉]

ByteDance

Abstract

While data curation for Vision Language Models (VLMs) is increasingly active, public practice for constructing pretraining mixtures remains largely heuristic: practitioners stack datasets that pass quality filters, set cross-domain ratios by intuition, and lack a principled, attributable criterion for admitting new data, while frontier recipes remain undisclosed. We formulate data construction as a **systematic mixture-optimization** problem and turn it into a reproducible engineering discipline by decoupling the mixture into two orthogonal sub-problems: **inter-class** ratios across capabilities and **intra-class** ratios within a category. For inter-class allocation, we use a single-variable iterative search; for intra-class composition, we apply a multidimensional, dataset-level assessment scoring Quality and Difficulty, and formulate selection as a constrained convex optimization with a diversity objective. The **DecoupleMix** framework delivers two critical capabilities: guiding *what data to collect next* and rendering dataset validation a controlled, attributable experiment. Experiments show our approach consistently surpasses heuristic baselines. Moreover, optimal ratios discovered on small-scale proxies transfer seamlessly to larger scales without retuning. Using 80B additional multimodal continue-pretraining tokens, our VLM is competitive with strong open-source models trained with substantially larger multimodal budgets.

✉ Email: guozhongbin66@gmail.com

1. Introduction

Vision Language Models (VLMs) have become a central paradigm for building general-purpose multimodal systems, combining visual perception, language understanding, and cross-modal reasoning within a unified framework. Proprietary frontier models, such as GPT-5.6 [1], Seed-2.0 [2], Claude Fable 5 [3], and Gemini-3.5-Flash [4], demonstrate remarkable capabilities across a wide range of vision-language tasks. Meanwhile, the open-source community has made substantial progress with models including LLaVA series [5, 6], Qwen3-VL [7] and Molmo2 [8]. Nevertheless, a considerable performance gap persists between frontier proprietary models and their open-source counterparts. Beyond differences in architecture and scale, growing evidence highlights the Data Mixture Recipe—the selection, composition, and allocation of training data—as a critical determinant of model capability and data efficiency [9–12]. However, proprietary recipes remain largely undisclosed, while open-source pipelines still depend heavily on quality filtering, dataset stacking, and empirically chosen sampling ratios. This practice leads to two critical limitations:

*Equal contribution.

[†]Work done during an internship at ByteDance Commercial AI Team.

✉Corresponding author.

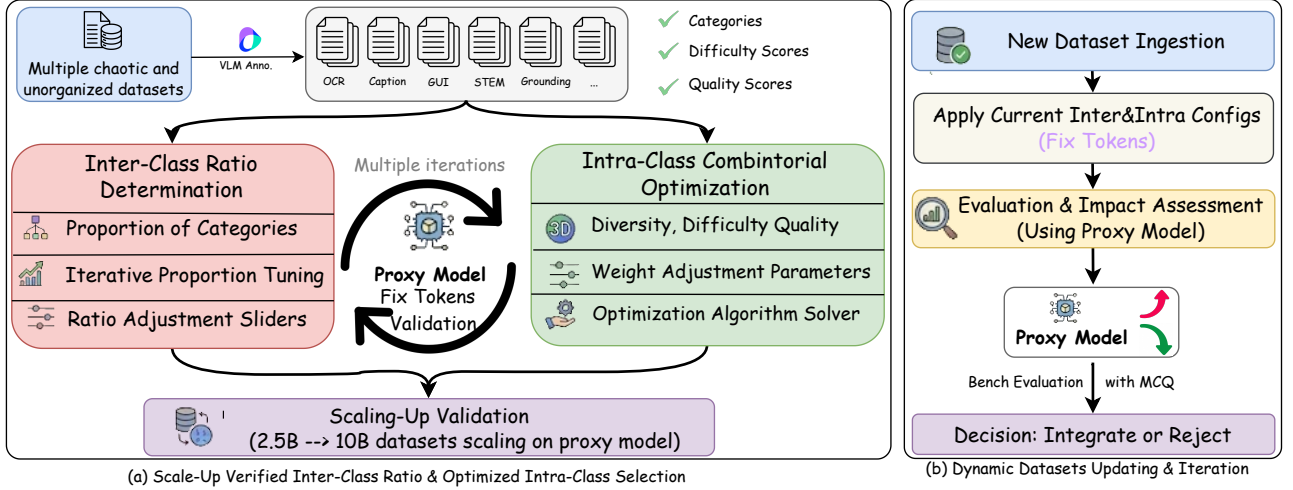


Figure 1: Overview of DecoupleMix. (a) *Recipe search*: annotated datasets are optimized along two decoupled axes—inter-class budget allocation and intra-class convex selection—then validated and scaled up on a cheap proxy model (2.5B→10B tokens). (b) *Dynamic admission*: a new dataset is evaluated under the fixed configuration, turning its acceptance into a controlled, attributable integrate-or-reject decision.

- **Over-reliance on Quality Filtering:** Prevailing methods focus on cleaning “noisy” data, but lack systematic assessment of *Data Difficulty* and *Diversity*.
- **Unsystematic Mixture Strategies:** Ratios for **inter-class** settings (e.g., Grounding vs. Caption) are typically set by intuition. For **intra-class** settings (e.g., multiple OCR datasets), the standard practice is brute-force stacking of all filtered data, ignoring redundancy.

To address these challenges, we recast the construction of continue-pretraining data as a reproducible engineering discipline. Our central insight is that the monolithic question of “what to mix” can be **decoupled** into two hierarchical sub-problems: **inter-class** allocation—how to split the token budget *across* capability categories, and **intra-class** composition—how to mix candidate sources *within* a single category. This mirrors industrial data organization and sidesteps the intractability of high-dimensional search. Rather than sample-level filtering, we score each candidate dataset as a whole unit, keeping optimization tractable across hundreds of sources.

Beyond producing an optimized recipe, this decoupling yields the **attributable** evaluation of dataset admission. By holding the inter-class ratio, total budget, and intra-class strategy fixed, adding a candidate dataset becomes a single intervention whose total effect includes the resulting within-class reallocation. This avoids the broader confounding introduced by expensive “re-stack and retrain” ablations.

Our contributions are summarized as follows:

- **Decoupled Mixture Methodology (DecoupleMix):** We decompose data construction into inter- and intra-class ratios using single-variable search and constrained convex optimization, built on an automated, multi-dimensional dataset-level assessment.
- **Attributable Data Acceptance:** Fixing inter-class ratios and intra-class strategies turns dataset validation into a controlled intervention whose effect can be measured with substantially less global confounding.

- **Scalability and Efficiency:** Ratios found at the proxy scale transfer across the data and model scales without additional tuning. With 80B additional multimodal continued-pretraining tokens, our VLM is competitive with strong open-model baselines.

2. Methodology

The DecoupleMix framework forms a closed-loop system with three core modules: (1) **Automated Dataset-Level Assessment**, which classifies capabilities and quantifies quality and difficulty at the dataset level; (2) **Decoupled Mixture Optimization**, which uses a heuristic search for the inter-class ratio and constrained convex optimization for intra-class mixture planning; and (3) **Standardized Data Acceptance Criteria**, which measures newly collected datasets under a fixed recipe and supports continuous iteration of the data pool. We use Multiple-Choice-Question (MCQ) validation to evaluate proxy checkpoints while reducing dependence on free-form instruction following.

2.1. Automated Dataset-Level Assessment

Our framework quantifies three properties: capability, quality, and difficulty. We decompose VLM capabilities into an extensible hierarchical taxonomy to identify gaps and guide data production. Its taxonomy is unbounded, and its assignment is automated: practitioners specify capability categories with natural-language criteria, and the model classifies each candidate dataset. This avoids manual labeling and allows the taxonomy to be extended when new capabilities are introduced.

To implement this, we use datasets (rather than individual samples) as the minimal unit, which suits industrial pipelines where datasets often exhibit strong cohesion, reducing sample-level annotation noise. We adopt an LLM-as-a-Judge paradigm: we sample 150 instances per dataset and use Seed-1.6 [13] as the evaluator. The **Quality** score q_i is a weighted average of four dimensions (penalizing hallucinations), and the **Difficulty** score d_i aggregates six dimensions (prioritizing cross-modal synthesis and prior knowledge). **Diversity** is not scored per dataset; instead, it is enforced structurally during mixture optimization via an entropy term (§2.2). The evaluation prompts are provided in Appendix B.

To confirm reliability, we also conducted a human-alignment study. Results show that automated scores are strongly correlated with human consensus (Spearman $r = 0.82$ for Quality, $r = 0.75$ for Difficulty), indicating that the score at dataset level is stable enough to drive downstream optimization (Appendix B.1).

2.2. Decoupled Mixture Optimization

We solve the two decoupled dimensions with different tools: the inter-class taxonomy allocates the token budget across capability categories, while the intra-class objective captures finer-grained semantic variation *within* a category through an explicit diversity term.

Inter-class Strategy. Global ratios between broad categories (e.g., Image-Text vs. Video-Text) shape the model’s capability profile [14]. We treat these macro-distributions as hyperparameters and search them with a **coordinate-style single-variable procedure**. Starting from a warm-start configuration $\mathbf{r}^{(0)}$ (the average of reference recipes), each round sweeps one category’s ratio r_c over a candidate set while holding others fixed (renormalizing the remainder), and keeps the value maximizing validation performance. Fixing other variables isolates the effect of c and avoids multi-variable confounding. We iterate until stabilization, optimizing ratios in linear rather than exponential time.

Intra-class Convex Optimization. For the more granular challenge of selecting data within a category, we reject naive approaches like size-proportional sampling [15] or greedy quality selection [16]. We formulate intra-class sampling as a **Constrained Convex Optimization** problem, solved independently per category c under its allocated budget $T_c = r_c T$. Within a category of N datasets, let $\mathbf{w} = [w_1, \dots, w_N]^\top$ denote the normalized allocation weights. We solve

$$\begin{aligned} \max_{\mathbf{w} \in \Delta^{N-1}} \quad & \sum_{i=1}^N w_i (\alpha q_i + \beta d_i) + \gamma \mathcal{H}(\mathbf{w}) \\ \text{s.t.} \quad & \ell_i \leq w_i \leq u_i, \quad i = 1, \dots, N, \end{aligned} \quad (1)$$

where

$$\Delta^{N-1} = \left\{ \mathbf{w} \in \mathbb{R}_+^N \mid \sum_{i=1}^N w_i = 1 \right\}, \quad \mathcal{H}(\mathbf{w}) = - \sum_{i=1}^N w_i \log w_i. \quad (2)$$

Here, q_i and d_i denote the quality and difficulty scores of dataset i ; $\alpha, \beta, \gamma \geq 0$ control their relative contributions; and $\ell_i = L_i/T_c$ and $u_i = M_i/T_c$ are the normalized lower and upper allocation bounds, where L_i denote the corresponding token-level bounds and M_i denotes the total number of available tokens in dataset i . The resulting token allocation is

$$t_i = T_c w_i, \quad \sum_{i=1}^N t_i = T_c. \quad (3)$$

The entropy regularizer discourages excessive concentration on a small number of high-scoring datasets and promotes broader coverage of the candidate pool. Since $\mathcal{H}(\mathbf{w})$ is concave and the feasible region is convex, Eq. (1) defines a convex optimization problem, which we solve using ECOS [17].

2.3. Data Acceptance via Attributable Validation

Validating a candidate dataset still requires an empirical training run. The fundamental challenge is to define a comparison in which admission does not simultaneously alter the global budget and inter-class composition. Our protocol creates such a controlled admission comparison, making dataset admission mathematically **identifiable**.

Let performance be $P = f(\mathbf{r}, \mathbf{x}, T)$, where $\mathbf{r}$ is the inter-class ratio, T the budget, and $\mathbf{x}$ the intra-class allocation. To first order,

$$dP = \nabla_{\mathbf{r}} f \cdot d\mathbf{r} + \sum_c \nabla_{\mathbf{x}_c} f \cdot d\mathbf{x}_c + \frac{\partial f}{\partial T} dT. \quad (4)$$

In “re-stack and retrain”, all terms move simultaneously, making dP confounded. Our protocol fixes $\mathbf{r}$, T , and the optimizer g , while admitting the candidate into exactly one class c . The induced shift $d\mathbf{x}_c$ is therefore part of the response to a single admission intervention:

$$\Delta P_{d^*} = f(\mathbf{r}, g(\mathcal{D} \cup \{d^*\}), T) - f(\mathbf{r}, g(\mathcal{D}), T) \quad (5)$$

This captures the total effect of admitting the dataset—its direct contribution plus the budget-preserving re-balancing within its class. Because other factors are constant, the effect is attributable to the

Models	Params.	General					K&H			OCR			M&L			Video		Avg
		MMBench	MME	MMStar	R.WorldQA	MMMU	BLINK	HalBench	AI2D	OCRBen.	InfoVQA	CharXiv	LogicVis	VisuLog	PuzzleVQA	V.MME	MVBench	
Qwen3-VL-2B-Instruct	2B	77.4	72.5	56.1	64.1	49.6	55.3	45.1	75.8	82.2	58.2	51.6	35.6	24.4	13.4	47.8	50.2	53.7
Our-1B	1B	72.2	72.0	54.2	62.1	41.0	52.9	44.6	71.2	80.4	52.9	48.6	31.5	24.8	30.2	45.1	53.6	52.3
Qwen3-VL-4B-Instruct	4B	83.7	82.0	66.9	70.1	60.0	64.4	54.1	82.4	84.8	67.2	59.9	39.4	26.0	43.0	55.4	60.7	62.5
Our-4B	4B	83.1	80.5	66.1	70.5	59.1	61.0	55.8	80.7	82.7	67.3	63.7	40.3	25.8	49.0	52.9	61.1	62.5

Table 1: Comprehensive Performance Comparison. Our models after multimodal pretraining and before any additional multimodal instruction tuning, compared with Qwen3-VL-Instruct. Our 1B model is smaller than the 2B baseline, while the 4B models are parameter-matched. Best score within each size group and benchmark is in **bold**; ours are shown in light red. Avg is the mean over all 16 benchmarks.

candidate dataset d^* alone. Furthermore, the convex program also provides a low-cost screening step: a source that receives zero allocation under the fixed rule need not proceed to an empirical training run.

3. Experiments

We evaluate DecoupleMix at multiple data and model scales, separating end-to-end comparison from controlled tests of inter-class search, intra-class allocation, and dataset admission. We then scale the selected recipe to 80B additional multimodal continued-pretraining tokens for comparison with open-source model baselines.

3.1. Experimental Setup

Model & Training. To obtain a controlled multimodal initialization without inheriting a pretrained cross-modal projector, we assemble our models from pretrained unimodal-stage components rather than fine-tuning an existing VLM. Following the Qwen3-VL architecture [7], we initialize the visual encoder with Qwen3-VL-ViT and the language backbone with the text-only Qwen3-4B-Instruct [18] (an instruction LLM never exposed to images), and randomly initialize the ViT-LLM projector. All subsequent vision-language alignment and multimodal pretraining are then performed by us. Unless otherwise noted, all mixture-search, ablation, scaling, and attribution experiments are conducted on this 4B model, and the final full-scale run reuses the recipe discovered at the 4B proxy scale. Detailed hyperparameters are provided in Appendix C.

Training Stages: 1) *Alignment (8B tokens):* High-quality image-caption pairs align visual features with the LLM. Only the projector is trainable. 2) *Mixture Validation:* Full-parameter fine-tuning on mixed datasets to assess comprehensive capabilities. 3) *Annealing:* A short, high-quality refinement stage with learning rate decaying to zero.

3.2. Evaluation Metrics

To holistically assess model capabilities within space constraints, we use 16 benchmarks categorized into five domains matching our capability taxonomy. These domains are: (1) General, (2) Knowledge &

¹MMBench/CharXiv average their sub-scores.

²Normalization: MME scaled to 0–100 (SUM / 2800).

Hallucination, (3) OCR & Document, (4) Math & Logic, (5) Video. Detailed benchmark descriptions, evaluation setups and citations are in Appendix A. For fair comparison, all reported numbers including those of external baselines are obtained by us under a single unified evaluation protocol rather than copied from their original papers; in particular, the video benchmarks are evaluated with a fixed sampling of 8 frames per video for every model.

To validate metrics without instruction-tuning alignment, for benchmarks that can be cast as multiple-choice questions we adopt Closed-Task Validation with MCQs [19], which probes internal knowledge and reflects pretraining data efficacy by decoupling foundational understanding from instruction-following.

4. Experiment Results and Analysis

4.1. Main Results

To evaluate end-to-end efficacy, we train VLMs with our pipeline by assembling a pretrained vision encoder and language model, randomly initializing their projector, and performing the subsequent multimodal alignment ourselves. We compare the resulting models with Qwen3-VL in Table 1. This experiment tests whether the pipeline’s end product is competitive with models trained using substantially larger multimodal budgets; controlled comparisons with heuristic stacking are reported separately in Section 4.3.

For comparability, all numbers are obtained under the same evaluation protocol (§3.2). Our 1B model is compared with a larger 2B baseline, whereas the 4B comparison is parameter-matched. Both of our models use 80B additional multimodal continued-pretraining tokens and are evaluated before any additional multimodal instruction tuning (§3.1). Under this setting, the 4B model matches Qwen3-VL-4B in overall average (62.5 vs. 62.5), and the 1B model approaches Qwen3-VL-2B (52.3 vs. 53.7). These results demonstrate the end-to-end competitiveness of the resulting recipe, without isolating data mixture from the other differences between systems.

Where the gains concentrate. The per-benchmark pattern aligns with our design objective: our 4B model is strongest on several reasoning- and difficulty-sensitive benchmarks (PuzzleVQA 49.0 vs. 43.0; CharXiv 63.7 vs. 59.9; as well as RealWorldQA, HallusionBench, LogicVista, MVBench), while trailing slightly on several perception-heavy benchmarks (OCR-Bench, MME and AI2D). A similar asymmetry appears at 1B, where our model surpasses the larger Qwen3-VL-2B on PuzzleVQA (+16.8) and MVBench (+3.4). Because the external baselines differ in data, scale, and training, this comparison is descriptive rather than causal. The matched-budget scaling study and intra-class ablation (§4.3, §4.4) provide the controlled support for the recipe design.

Method	Admit	Per-domain Δ					Δ_{Avg}	Max $ \Delta _{\text{off}}$
		OCR	Gen.	K&H	M&L	Video		
Ours	OCR set	+0.3	-0.1	+0.0	+0.6	+0.2	+0.2	0.6
Stack	OCR set	+1.5	-1.1	+0.8	-0.4	+0.7	+0.1	1.1
Displace	OCR set	+0.5	-2.3	+0.2	+0.9	+0.6	-0.3	2.3

Table 2: Attributable validation of single-dataset admission. Per-domain performance changes after admitting one OCR dataset. Our protocol produces substantially less off-target variation than *Stack* and *Displace*. Full per-benchmark results are in Table 8.

Models	Scale	General					K&H			OCR			M&L			Video		Avg
		MMBench	MME	MMStar	R.WorldQA	MMMU	BLINK	HalBench	AI2D	OCRBench	InfoVQA	CharXiv	LogicVista	VisuLogic	PuzzleVQA	V.MME	MVBench	
Baseline	2.5B	79.5	72.8	58.4	66.7	51.0	54.6	66.0	76.0	78.2	58.5	53.1	32.9	25.1	29.9	51.5	52.0	56.6
Ours		82.3	73.2	60.7	66.5	52.2	55.5	68.0	78.0	79.4	60.6	52.6	36.2	23.1	38.9	51.8	51.7	58.2
Baseline	5B	80.7	72.6	59.1	66.3	51.0	55.5	68.8	79.7	77.6	61.7	53.4	33.6	22.6	43.8	52.8	51.9	58.2
Ours		82.5	74.6	62.5	69.3	51.9	57.0	69.3	81.2	81.9	64.8	55.5	38.0	26.4	42.4	52.5	52.6	60.2
Baseline	10B	80.9	78.8	58.8	68.0	53.8	55.3	68.0	78.2	80.1	64.5	57.0	35.3	24.6	45.6	52.4	54.0	59.7
Ours		83.2	77.6	63.1	69.0	54.7	58.5	68.9	80.1	81.6	67.6	57.3	36.0	24.3	48.6	53.7	52.9	61.1

Table 3: Comprehensive Scaling Analysis. We compare our full method (**Ours**: optimized inter-class ratios with convex intra-class allocation) against the **Baseline** (heuristic stacking: all quality-filtered datasets pooled and sampled in proportion to their size) across three training-token scales (2.5B, 5B, 10B).

32B Recipe	General					K&H			OCR			M&L			Video		Avg
	MMBench	MME	MMStar	R.WorldQA	MMMU	BLINK	HalBench	AI2D	OCRBench	InfoVQA	CharXiv	LogicVista	VisuLogic	PuzzleVQA	V.MME	MVBench	
Stacking	83.8	80.8	63.9	69.3	57.3	55.9	50.0	83.4	79.0	62.4	59.2	40.9	24.0	60.4	57.2	56.3	61.5
Ours	85.3	82.2	66.1	70.1	59.4	59.3	51.9	81.8	82.9	63.5	60.6	42.5	25.9	56.8	57.1	57.4	62.7

Table 4: Cross-model-scale transfer to 32B. Full per-benchmark comparison of our proxy-searched recipe, applied without retuning, with heuristic stacking at a matched token budget. Best score per column is in **bold**; ours are shown in light red.

4.2. Attributable Validation: Single-Dataset Admission

A central claim is that admitting a dataset under our protocol yields an effect *attributable* to that dataset, avoiding confounded global shifts. We test whether the admission protocol reduces global confounding by admitting one OCR dataset under three regimes and measuring per-domain performance change Δ : **Ours** (inter-class ratio r and total budget T fixed, with the candidate entering through convex intra-class reallocation); **Stack** (the candidate is concatenated, changing the effective budget and mixture); and **Displace** (r is fixed, but the candidate naively crowds out other OCR sources).

Result. Table 2 reports the outcome. Under our protocol, the response is localized: the target OCR domain moves by +0.3 while every off-target domain stays within 0.6. Naive regimes leak perturbation: stacking drifts the global balance (max off-target 1.1), and displacement regresses General by -2.3 . Stacking shows a larger raw OCR increase (+1.5) because it assigns a larger share of the resulting global mixture to OCR, conflating source admission with global rebalancing. Despite this change, its overall average increases by only +0.1, compared with +0.2 under our budget-preserving protocol; displacement decreases the average by -0.3 . The directions of change are also coherent under our protocol: Math & Logic, which can depend on OCR-style perception, improves by +0.6, whereas stacking changes it by -0.4 as part of the broader mixture shift. This comparison turns

Ablation Settings	General					K&H			OCR			M&L			Video		Avg
	MMBench	MME	MMStar	R.WorldQA	MMMU	BLINK	HalBench	AI2D	OCRBench	InfoVQA	CharXiv	Logic Vista	VisuLogic	PuzzleVQA	V.MME	MVBench	
Baseline (Stacking)	79.5	72.8	58.4	66.7	51.0	54.6	66.0	76.0	78.2	58.5	53.1	32.9	25.1	29.9	51.5	52.0	56.6
<i>Inter-class Ratio Variants (Intra-class fixed as Baseline)</i>																	
• Bee Recipe	77.9	71.8	56.4	64.4	56.2	54.1	67.5	78.1	78.4	60.6	50.5	34.2	25.6	20.4	50.5	51.2	56.1
• LLaVA-OV 1.5 Recipe	79.5	73.6	58.9	66.7	52.7	55.8	65.1	79.0	77.5	59.1	52.0	34.5	24.1	30.9	51.7	50.8	57.0
• Ours (Inter-only)	79.7	72.3	57.7	66.0	50.4	56.5	66.8	77.1	77.0	58.5	51.5	30.6	27.2	37.5	51.7	52.7	57.1
Ours (Inter + Intra)	82.3	73.2	60.7	66.5	52.2	55.5	68.0	78.0	79.4	60.6	52.6	36.2	23.1	38.9	51.7	51.7	58.2

Table 5: Ablation on Decoupled Strategies. We compare our optimization against baseline stacking and alternative inter-class recipes (Bee, LLaVA-OV-1.5), reporting all per-benchmark scores and overall average. Best per column in **bold**; full method in **light red**.

“should this source be ingested?” into a controlled policy-level decision (full per-benchmark scores in Appendix Table 8). The transfer results in §4.3 support reusing the surrounding recipe at larger scales, although the admission effect itself is measured only at the proxy scale.

4.3. Scaling Efficiency and Transferability

We analyze our full method across data scales (2.5B, 5B, 10B tokens) in Table 3.

Consistent Superiority. Our method outperforms the baseline at every scale: the average gap is +1.6 at 2.5B, +2.0 at 5B, and +1.4 at 10B. The advantage is stable rather than monotonically growing, maintaining a consistent edge. Improvements are pervasive across capability domains (e.g., InfoVQA +3.1 at 10B).

Transferability Across Data and Model Scales. Optimal ratios searched only at 2.5B transfer to 5B and 10B scales without retuning and outperform the corresponding baselines. This indicates that the proxy search captures regularities that persist across the data scales. To probe transferability across model sizes, we evaluated the same 2.5B proxy recipe on a 32B model without retuning (Table 4). The transferred recipe improves the overall average by +1.2 (61.5 → 62.7) against a matched-budget stacking baseline and leads on 13 of 16 benchmarks (e.g., OCR +2.1). This reinforces that discovered ratios reflect universal data properties rather than model-specific artifacts.

Search cost is amortized at the proxy scale. This transferability is what makes the method economical in practice. The single-variable inter-class search is conducted at the 2.5B-token proxy scale, where each candidate run has lower cost; the resulting recipe is then reused unchanged across both larger data budgets (5B, 10B) and larger model capacities (32B). Within these experiments, the search cost is paid once and amortized across downstream runs, unlike “re-stack and retune” workflows that repeat the sweep for each configuration.

Intra-class Strategy	General					K&H			OCR			M&L			Video		Avg
	MMBench	MME	MMS _{tar}	R. WorldQA	MMMU	BLINK	HalBench	AI2D	OCRBench	InfoVQA	CharXiv	LogicVista	VisuLogic	PuzzleVQA	V.MME	MVBench	
Quantity-Proportional	63.8	72.9	42.7	53.1	39.1	44.2	50.3	60.9	67.0	31.5	35.8	30.6	23.7	32.9	45.6	42.1	46.0
Quality Only	66.3	71.5	44.4	57.4	39.2	43.9	51.9	65.1	70.4	36.9	36.5	31.1	23.4	26.5	45.0	43.7	47.1
Convex Alloc. (Ours)	67.6	74.8	45.0	56.5	38.9	44.9	52.1	64.9	73.0	47.1	38.0	30.0	25.6	30.8	45.5	44.6	48.7

Table 6: Intra-class allocation under a fixed inter-class budget. Full per-benchmark comparison of three sampling strategies under a 1B-token budget. Best score per column is in **bold**; ours are shown in light red.

4.4. Ablation: Decoupling and Intra-class Allocation

To attribute our gains, we decompose the effect of inter-class and intra-class optimization (Table 5), and analyze the intra-class allocator itself (Table 6).

Inter-class ratios provide a strong macro-structure. We compare our inter-class proportions against baseline stacking and two published reference recipes (Bee [20] and LLaVA-OneVision-1.5 [6]). Our Inter-only strategy (Avg 57.1), compared with 56.6 for stacking and 56.1 and 57.0 for the reference recipes, showing our capability-driven taxonomy allocates data better than empirical structures. Adding intra-class allocation further raises the average to 58.2.

Balanced intra-class allocation beats naive sampling. We compare our convex allocation against two baselines under a fixed inter-class budget (Table 6): *Quantity-Proportional* (suboptimal Avg 46.0) and *Quality-Only* (improves to 47.1 overall but underperforms on Math & Logic, 27.0, by over-concentrating on easy high-quality data). Our convex allocation attains the highest average (**48.7**), with its largest domain-level improvement on OCR (52.7 vs. 44.8). This pattern is consistent with the entropy regularizer reducing concentration on a small set of high-scoring datasets and preserves long-tail coverage, as visualized in Fig. 2.

5. Related Works

Multimodal data curation. Large-scale VLM pipelines commonly improve data efficiency through similarity-based filtering [21], model-based quality curation [22], and joint quality–diversity selection [11]. These approaches primarily score or retain individual examples. DecoupleMix is complementary: it treats datasets as allocation units after curation and addresses how the retained sources should share a finite training budget.

Data-mixture optimization. Data-mixture optimization has been studied extensively in language-model pretraining. DoReMi learns domain weights with a proxy model and group distributionally robust optimization [23]; Data Mixing Laws and RegMix use small-scale training runs to predict the performance of unseen mixtures [24, 25]; and RegMix-D extends regression-based mixture optimization to dynamic schedules [26]. Recent work also studies scaling laws for optimal mixtures across language, vision, and native multimodal settings [27]. These studies establish the broader value of mixture optimization and motivate our VLM-specific hierarchical formulation.

Multimodal mixture recipes. MM1 demonstrates that the proportions of caption, interleaved image–text, and text-only data materially affect multimodal pretraining [28], while open pipelines such as LLaVA-OneVision-1.5 and Bee provide reproducible reference recipes [6, 20]. DecoupleMix differs by jointly separating inter-class budget search from intra-class dataset allocation and by evaluating candidate admission under a fixed budget and allocation rule. The contribution is therefore not the first use of mixture optimization, but a dataset-level, hierarchical workflow designed for continuously evolving VLM data pools.

6. Limitations

Several limitations remain and point to future work. **(1) Scaling scope:** our study spans 2.5B–10B tokens with model-scale transfer evaluated at 32B parameters; we have not released the trend at the much larger token budgets of frontier industrial pretraining. **(2) Optimization algorithms:** the single-variable inter-class search and convex intra-class program are deliberately tractable but likely suboptimal—richer joint search or more expressive objectives could improve the recipes. **(3) Modality coverage:** we validate only on vision–language models, leaving the extension to fully omni-modal models (audio, video, etc.) untested. Finally, our judge-based assessment is bounded by the frozen judge’s capability, and treating each dataset as an atomic quality–difficulty unit precludes finer sample-level gains.

7. Conclusion

We present **DecoupleMix**, a systematic data mixture optimization framework that transforms VLM pretraining from trial-and-error into principled engineering. By introducing a multi-dimensional dataset-level assessment and decoupling data composition into inter-class and intra-class optimization, we establish a scalable methodology for constructing high-performance recipes. Validated across 2.5B to 10B tokens and transferring without retuning to larger data and model scales, our approach consistently outperforms heuristic baselines, demonstrating that systematic data curation is the critical lever for efficient VLM development.

References

- [1] OpenAI. Gpt-5.6 preview system card, 2026. URL <https://deploymentsafety.openai.com/gpt-5-6-preview/gpt-5-6-preview.pdf>.
- [2] Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. 2026. URL <https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0%20Model%20Card.pdf>.
- [3] Anthropic. System card: Claude fable 5 & claude mythos 5, 2026. URL <https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf>.
- [4] Google Deepmind. Gemini 3.5 flash model card. 2026. URL <https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-5-Flash-Model-Card.pdf>.
- [5] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.
- [6] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

-
- [7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL <https://arxiv.org/abs/2511.21631>.
- [8] Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.
- [9] Tianyi Bai, Hao Liang, Binwang Wan, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective. *CoRR*, abs/2405.16640, 2024. URL <https://doi.org/10.48550/arXiv.2405.16640>.
- [10] Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876, June 2025.
- [11] Fengze Liu, Weidong Zhou, Binbin Liu, Zhimiao Yu, Yifan Zhang, Haobin Lin, Yifeng Yu, Bingni Zhang, Xiaohuan Zhou, Taifeng Wang, et al. Quadmix: Quality-diversity balanced data selection for efficient llm pretraining. *arXiv preprint arXiv:2504.16511*, 2025.
- [12] Ali Vosoughi, Dimitra Emmanouilidou, and Hannes Gamper. Quality over quantity? llm-based curation for a data-efficient audio-video foundation model. In *2025 33rd European Signal Processing Conference (EUSIPCO)*, pages 286–290. IEEE, 2025.
- [13] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.
- [14] Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying LLM visual priors from language pre-training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=pfw176o1YJ>.
- [15] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [16] Biao Wu and Ling Chen. Curriculum learning with quality-driven data selection. *arXiv preprint arXiv:2407.00102*, 2024.
- [17] Alexander Domahidi, Eric Chu, and Stephen Boyd. Ecos: An socp solver for embedded systems. In *2013 European control conference (ECC)*, pages 3071–3076. IEEE, 2013.
- [18] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [19] Enci Zhang, Z.Q. ZHANG, Jiahao Xie, Ruiqi Lu, Boyan Zhou, and Cheng Yang. Closed-task validation: A more robust and efficient proxy for guiding VLM training. In *1st Workshop on VLM4RWD @ NeurIPS 2025*, 2025. URL <https://openreview.net/forum?id=fNxm8j1EWD>.
- [20] Yi Zhang, Bolin Ni, Xin-Sheng Chen, Heng-Rui Zhang, Yongming Rao, Houwen Peng, Qinglin Lu, Han Hu, Meng-Hao Guo, and Shi-Min Hu. Bee: A high-quality corpus and full-stack suite to unlock advanced fully open mllms. *arXiv preprint arXiv:2510.13795*, 2025.

-
- [21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [22] Hongyuan Dong, Zijian Kang, Weijie Yin, Xiao Liang, Chao Feng, and Jiao Ran. Scalable vision language model training via high quality data curation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33272–33293, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1595. URL <https://aclanthology.org/2025.acl-long.1595/>.
- [23] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=1XuByUeHhd>.
- [24] Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=jjCB27TMK3>.
- [25] Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, volume 2025, pages 38305–38339, 2025.
- [26] Kaiyan Zhao, Zhongtao Miao, Akiko Aizawa, and Yoshimasa Tsuruoka. Regmix-d: Dynamic data mixing via proxy training trajectories. *arXiv preprint arXiv:2606.18663*, 2026.
- [27] Mustafa Shukor, Louis Béthune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablyn. Scaling laws for optimal data mixtures. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=vU1KT0sju>.
- [28] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mml: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024.
- [29] 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*, pages 216–233. Springer, 2024.
- [30] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL <https://openreview.net/forum?id=DgH9YCsqWm>.
- [31] 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.
- [32] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoyi 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*, pages 9556–9567, 2024.
- [33] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

-
- [34] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14375–14385, 2024.
- [35] 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, pages 235–251. Springer, 2016.
- [36] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences, 67(12):220102, 2024.
- [37] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706, 2022.
- [38] 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, 2024.
- [39] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. arXiv preprint arXiv:2407.04973, 2024.
- [40] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. arXiv preprint arXiv:2504.15279, 2025.
- [41] Yew Ken Chia, Vernon Toh, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In Findings of the Association for Computational Linguistics: ACL 2024, pages 16259–16273, 2024.
- [42] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 24108–24118, 2025.
- [43] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22195–22206, 2024.
- [44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
- [45] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In SC22: International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–15. IEEE, 2022.

A. Detailed Evaluation Metrics

To comprehensively evaluate the performance of our models across diverse multimodal capability dimensions, we utilize a robust suite of 16 mainstream benchmarks. These are categorized into five primary domains, directly aligning with the capability taxonomy discussed in our methodology:

General. This domain assesses broad multimodal understanding, real-world knowledge, and complex reasoning capabilities. The evaluation suite includes **MMBench** [29], **MME** [30], **MMStar** [31], **MMMU** [32] (focusing on expert-level multi-discipline knowledge), and **RealWorldQA**.

Knowledge & Hallucination. To evaluate fine-grained visual perception, diagram/knowledge understanding, and the model’s robustness against hallucinating non-existent visual elements or being misled by text prompts, we employ **BLINK** [33], **HallusionBench** [34], and **AI2D** [35].

OCR & Document. This category focuses on text-rich image comprehension, document layout parsing, and complex chart reasoning. It incorporates **OCRBench** [36], **InfoVQA** [37], and **CharXiv** [38].

Math & Logic. To rigorously test the model’s visual logic deduction, spatial intelligence, and geometric problem-solving abilities, we utilize **LogicVista** [39], **VisuLogic** [40] and **PuzzleVQA** [41].

Video. For evaluating dynamic scene understanding, temporal localization, and cross-frame synergies, we extend our evaluation to the temporal domain using **Video-MME** [42] and **MVBench** [43]. For both benchmarks we uniformly sample a fixed 8 frames per video and apply the same protocol to all models, including the external baselines.

B. Dataset-Level Assessment: Prompts and Formulations

In this section, we detail the formulations and the specific system prompts used by the SOTA VLM to evaluate the Quality (Q) and Difficulty (D) of the sampled pretraining data. For each dataset, 150 samples are evaluated to estimate the overall distribution.

B.1. Human Alignment of the Automated Judge

To verify that the LLM-as-a-Judge scores are reliable rather than arbitrary, we conduct a human-alignment study. We randomly sample 10 instances spanning all capability categories and ask three expert annotators to independently rate each instance on the same Quality and Difficulty rubrics used by the automated judge (1–5 scale per dimension). We take the median of the three annotators as the human consensus and measure rank agreement between the automated scores and this consensus using the Spearman correlation coefficient.

Dimension	Spearman r	P -value
Quality (Q)	0.82	0.004
Difficulty (D)	0.75	0.013

Table 7: Agreement between the automated judge and the human consensus over 10 instances scored by three expert annotators. Both correlations are statistically significant ($P < 0.05$).

As shown in Table 7, the automated scores correlate strongly and significantly with the human consensus on both dimensions ($r = 0.82$ for Quality, $r = 0.75$ for Difficulty), indicating that the dataset-level Quality and Difficulty scoring is stable enough to drive the downstream mixture optimization. We will release additional annotated demonstrations in the camera-ready version.

B.2. Quality Assessment Formulation

The final Quality score Q for a single data instance is calculated using a weighted average of four dimensions. Accuracy and Hallucination are given higher weights to ensure the factual correctness of the pretraining signals. The formula is defined as:

$$Q = \frac{2 \cdot S_{acc} + 1.5 \cdot S_{hal} + 1 \cdot S_{cor} + 0.3 \cdot S_{gra}}{4.8} \quad (6)$$

where S_{acc} , S_{hal} , S_{cor} , and S_{gra} represent the scores (1-5 scale) for Accuracy, Hallucination, Correlation, and Grammar, respectively.

Quality Evaluation Prompt

You are a meticulous data quality assessor for VLM training data. Analyze each data sample, which may include both image and text (user question and assistant response) modalities, across multiple quality dimensions and provide numerical scores with clear, evidence-based reasoning. Pay special attention to hallucinations—assertions in the assistant’s response that are not directly supported or verifiable from the image or text context.

1. Evaluation Dimensions

Score each dimension on a 1-5 scale (1=lowest, 5=highest):

- **Accuracy:** Factual correctness and verifiability of the assistant’s response...
- **Grammar:** Linguistic correctness & fluency of text.
- **Correlation:** Relevance of the question and answer to the image content...
- **Hallucination:** Degree of hallucinated content, i.e., claims or inferences not grounded in visible image evidence... Lower score means more hallucination.

2. Scoring Protocol

- Base scores strictly on concrete, verifiable evidence from both text and image.
- Score based on the percentage of errors: 5 points for completely correct, 4 points for errors within 10%, 3 points for errors within 20%, 2 points for errors within 35%, and 1 point for errors at 40% or above.
- Flag hallucinations, misinformation, or unsupported inferences explicitly.
- Mark samples needing human review if confidence <90%.

3. Output Format

Respond with a JSON dictionary: { "dimension_scores": { "accuracy": <int>, "grammar": <int>, "correlation": <int>, "hallucination": <int> }, "flags": "...", "rationale": "...", "recommendation": "keep/review/discard" }

B.3. Diversity Categorization Prompt

To systematically measure the semantic diversity of candidate datasets, we employ an LLM-based classifier to map each data sample to a precise primary category and subcategory. This process ensures our optimization algorithm (Section 2.3) correctly balances the distribution across various multimodal capabilities.

Diversity Classification Prompt

You are an expert classifier specialized in categorizing multimodal (image-text) or text-only data into defined types, aligned with the data structures of state-of-the-art (SOTA) models. Your goal is to accurately map each sample to a primary category and subcategory based on content and image dependency.

Note on Image Dependency:

- **Multimodal Categories** (Caption, OCR, General VQA, Interleave, STEM, GUI, Grounding): Require an image or <image> placeholder.
- **Text-Only Category**: No image dependency—text contains no <image> placeholder and is fully self-contained.

1. Primary Categories & Subcategories

- **Caption**: textual descriptions of single or tightly arranged multiple images (basic_caption, fine_grained_caption, scene_caption, comparative_caption, multilingual_caption).
- **OCR**: Extracting and using text from images to solve tasks (text_detection, text_recognition, text_understanding, scene_text, document_ocr).
- **Grounding**: Linking text to specific visual elements like bbox or video segments (location_grounding, phrase_grounding, referring_expression, counting_grounding, multimodal_alignment).
- **STEM**: Integrating advanced math/science principles with specialized visuals (scientific_image_qa, complex_math_problem_solving, engineering_schematic).
- **Interleave**: Alternating Q&A interaction between multiple images and text (image_text_alignment, image_guided_text, text_guided_image, multimodal_dialogue).
- **General VQA**: General image/video-based QA not involving Grounding or OCR (factual_vqa, reasoning_vqa, open_domain_vqa, multi_turn_vqa, commonsense_vqa).
- **GUI**: Agent-centric multimodal instructions in interactive interfaces (ui_navigation, ui_element_interaction, ui_task_planning, ui_command_execution).
- **Text-Only**: Self-contained text (plain_text, structured_text, textual_qa, textual_summarization, textual_reasoning).

Other Rules:

- Takes Priority: text_only > OCR = interleave = caption > grounding = gui > stem > general_vqa.
- The core distinction between OCR and STEM lies in reasoning depth: direct correspondence belongs to OCR; complex reasoning belongs to STEM.

2. Output Format

Respond strictly in JSON with: { "tags": { "primary_category": "<caption/.../text_only>", "subcategory": "<corresponding subcategory>" }, "flags": "<comma-separated data features>", "rationale": "<detailed explanation linking data to category>" }

B.4. Difficulty Assessment Formulation

The Difficulty score D identifies samples requiring advanced capabilities. It aggregates six dimensions, with Cross-Modal Synthesis and Prior Knowledge Demand weighted most heavily, as they are key

drivers for VLM scaling. The formula is:

$$D = \frac{0.5 \cdot S_{txt} + 0.5 \cdot S_{img} + 2 \cdot S_{cross} + 1.5 \cdot S_{prior} + 1 \cdot S_{cue} + 0.3 \cdot S_{amb}}{5.8} \quad (7)$$

where the variables correspond to Text Complexity, Image Complexity, Cross-Modal Synthesis, Prior Knowledge Demand, Visual Cue Sensitivity, and Task Ambiguity.

Difficulty Evaluation Prompt

You are an expert evaluator specialized in screening high-difficulty, valuable multimodal (image-answer) training data for large language models. Your goal is to identify samples requiring advanced capabilities (complex reasoning, cross-modal integration, expert knowledge, precise visual perception) while excluding low-value noise.

1. Core Evaluation Dimensions

Rate 6 key difficulty dimensions on a 1-5 scale. Use 2 or 4 for intermediate difficulty:

- a. Image Complexity:** Measures the difficulty of understanding the image on its own.
(1: Simple object; 3: Multiple interacting objects; 5: Abstract/hyper-detailed content)
- b. Text Complexity:** Measures the difficulty of understanding the answer text.
(1: Short phrases; 3: Domain-general sentences; 5: Complex syntax, jargon)
- c. Cross-Modal Synthesis:** Measures the difficulty of connecting image and text.
(1: No meaningful link; 3: Moderate link, 1-2 steps inference; 5: Deep, abstract link, multi-step)
- d. Prior Knowledge Demand:** Measures domain/cultural knowledge required.
(1: No prior knowledge; 3: Basic domain knowledge; 5: Cross-domain expert knowledge)
- e. Visual Cue Sensitivity:** Measures the difficulty of perceiving task-relevant critical visual cues.
(1: Cues are large, high-contrast; 3: Small, low-contrast; 5: Nearly imperceptible, requires magnification)
- f. Task Ambiguity:** Measures the difficulty of resolving uncertainty.
(1: No ambiguity; 3: Requires close inspection; 5: Requires multi-step expert reasoning)

2. Output Format

Respond strictly in JSON with: { "dimension_scores": { "image_complexity": <int>, "text_complexity": <int>, "cross_modal_synthesis": <int>, "prior_knowledge_demand": <int>, "visual_cue_sensitivity": <int>, "task_ambiguity": <int> }, "flags": "...", "rationale": "..." }

C. Detailed Hardware and Hyperparameters

For Stage 2, we use a global batch size of 128 and a peak learning rate of $1e^{-4}$ with cosine decay. We enable sequence packing with a maximum length of 8192 to maximize training efficiency. All models are trained using the AdamW [44] optimizer. To handle the memory overhead of the context length, we utilized DeepSpeed ZeRO-2 optimization [45]. Our efficient data pipeline and training stack allowed for rapid iteration: training the 10B token recipe in Stage 2 required approximately 480 GPU hours, all of our training experiments totally cost about 3500 GPU hours. This efficiency is central to our methodology, proving that systematic data construction enables high-performance VLM training with manageable computational budgets (accessible even within reasonable industrial constraints).

Admission Regime	General					K&H			OCR			M&L			Video		Avg
	MMBench	MME	MMStar	R.WorldQA	MMMU	BLINK	HalBench	AI2D	OCRBench	InfoVQA	CharXiv	LogicVista	VisuLogic	PuzzleVQA	V.MME	MVBench	
Baseline (no admission)	82.6	76.9	63.1	71.2	50.2	57.1	50.8	82.5	79.7	55.7	52.2	36.2	25.0	37.0	51.9	52.1	57.8
Stack	82.8	73.5	63.5	69.8	49.1	57.9	51.3	83.5	80.1	57.2	54.8	36.7	22.9	37.5	52.1	53.2	57.9
Displace	81.5	75.5	61.2	65.5	48.9	56.8	51.9	82.3	79.8	56.2	53.2	38.0	21.9	41.0	52.1	53.1	57.4
Ours	82.8	74.6	62.9	72.3	51.0	57.3	50.7	82.5	80.8	56.1	51.5	37.4	25.0	37.5	52.3	52.2	57.9

Table 8: Full per-benchmark attribution results. Absolute scores for admitting one OCR dataset under each regime, from which the per-domain Δ in Table 2 are derived; *Ours* in light red. All runs use the same 4B proxy model and token budget.

D. Intra-class Sampling Visualization Analysis

Figure 2 visualizes token allocations across the Quality and Difficulty dimensions for each strategy. Quantity-proportional sampling assigns substantial mass to lower-quality, lower-difficulty regions, whereas quality-only sampling concentrates toward higher Quality scores without explicitly accounting for Difficulty. Our convex allocation shifts more token mass toward datasets that score highly on both dimensions while retaining broader coverage, illustrating the behavior encouraged by its objective.

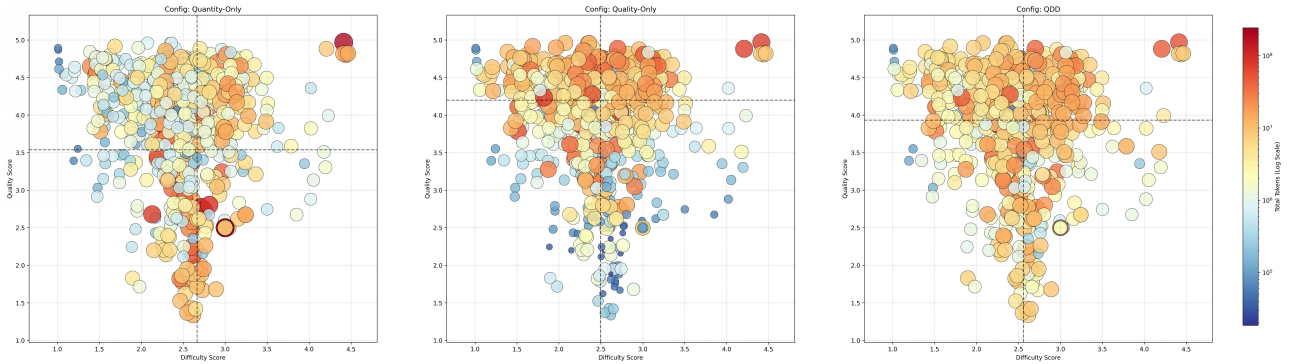


Figure 2: Visualization of Intra-class Sampling Distributions. The X-axis represents the Difficulty score, and the Y-axis represents the Quality score. The size of each point corresponds to the allocated token volume (log-scaled). From left to right: Quantity-Proportional, Quality-Only, and our convex intra-class allocation.

Case 1: Low-Quality OCR Sample



Question: <image> Provide a description of the image content and text.

Answer: person celebrates his penalty during soccer league, match
<ocr> </ocr>

Human Scores

Accuracy	2
Correlation	3
Grammar	2
Hallucination	1

Overall 1.90

Automated Judge Scores

Accuracy	2
Correlation	3
Grammar	3
Hallucination	1

Overall 1.96

Rationale:

1. **Accuracy:** The response incorrectly claims the player is celebrating a “penalty,” which has no basis in the image. While the OCR text is accurate, the penalty assertion is a significant factual error (~30%), justifying a score of 2.
2. **Grammar:** Misplaced comma (“soccer league , match”) and vague “person” instead of “soccer player” lead to awkward phrasing—score 3.
3. **Correlation:** Relevant elements (celebration, OCR) are present but an irrelevant penalty claim detracts—score 3.
4. **Hallucination:** The “penalty” claim is a severe hallucination with no visual evidence—score 1.

Case 2: High-Quality General Sample



Question: <image> Considering the security measures in place at the entrance, what might be the level of security concern for the property, and what does this imply about the importance or value of “Bramble Hall”?

Answer: The presence of electronic access panels on both sides of the gate suggests a moderate to high level of security concern for the property. This implies that “Bramble Hall” is likely to be a property of significance or value, possibly containing assets or individuals that require protection from unauthorized access.

Human Scores

Accuracy	4
Correlation	5
Grammar	5
Hallucination	4
Overall	4.27

Automated Judge Scores

Accuracy	4
Correlation	5
Grammar	5
Hallucination	4
Overall	4.27

Rationale:

The assistant correctly identified an electronic access panel on the right gate pillar (consistent with the image) and reasonably inferred moderate-to-high security concern, property significance, and owner prioritization of privacy from the gate and access features. However, the assistant incorrectly claimed electronic access panels are present "on both sides of the gate"—the left pillar only has a “Bramble Hall” sign and no access panel, which is a hallucinated detail not supported by the image. This error constitutes 4% of the response content.

Case 3: Difficulty Assessment – Grounding Sample



Question: <image> Name and localize each object present in the photo.

Answer: To grow potatoes in small garden <ref-object><bbox>s, as depicted in the image, place three <ref-object><bbox>s filled with young potato seedlings on a pavement or sidewalk. The <ref-object><bbox>s showcase various designs and colors, including a predominantly yellow and white one, a blue <ref-object><bbox> with a pattern, and a black <ref-object><bbox>. The plants inside the <ref-object><bbox>s are in their early growth <ref-object><bbox>...

Dimension	Human	Auto Judge
Image Complexity	3	3
Text Complexity	2	3
Cross-Modal Synthesis	3	3
Prior Knowledge	3	3
Visual Cue Sensitivity	2	1
Task Ambiguity	5	5
Average	3.02	2.84

Rationale:

- **Image Complexity=3:** Multiple interacting objects (three grow bags, potato plants, background elements like a red wall, wood, and concrete)—moderate detail, no abstract content.
- **Text Complexity=3:** Moderate-length sentences using domain-general gardening terms like “young potato seedlings” and “early growth stage.”
- **Cross-Modal Synthesis=3:** Links image elements (colored grow bags, green plants) to gardening context, requiring 1–2 inference steps to recognize bags as planters and plants as potato seedlings.
- **Prior Knowledge=3:** Basic gardening knowledge needed to identify grow bags, potato seedlings, and understand early growth stages.
- **Visual Cue Sensitivity=1:** Task-relevant cues (grow bags and plants) are large, high-contrast, and immediately noticeable.
- **Task Ambiguity=5:** No uncertainty in naming/locating core objects; the task is straightforward.