MIXATLAS: UNCERTAINTY-AWARE DATA MIXTURE FOR MULTIMODAL LLM MIDTRAINING

Bingbing Wen[*1,2], **Sirajul Salekin**[2], **Feiyang Kang**[3],
Bill Howe[1], **Lucy Lu Wang**[1], **Javier Movellan**[2], **Manjot Bilkhu**[2]
[1]University of Washington [2]Apple [3]Virginia Tech
bingbw@uw.edu, mbilkhu@apple.com

ABSTRACT

Domain reweighting can improve sample efficiency and downstream generalization, but data-mixture optimization for multimodal midtraining remains largely unexplored. Current multimodal training recipes tune mixtures along a single dimension, typically data format or task type. We introduce MixAtlas, a method that produces benchmark-targeted data recipes that can be inspected, adapted, and transferred to new corpora. MixAtlas decomposes the training corpus along two axes: *image concepts* (10 visual-domain clusters discovered via CLIP embeddings) and *task supervision* (5 objective types including captioning, OCR, grounding, detection, and VQA). Using small proxy models (0.5B) paired with a Gaussian-process surrogate and GP-UCB acquisition, MixAtlas searches the resulting mixture space with the same proxy budget as regression-based baselines but finds better-performing mixtures. We evaluate on 10 benchmarks spanning visual understanding, document reasoning, and multimodal reasoning. On Qwen2-7B, optimized mixtures improve average performance by 8.5%–17.6% over the strongest baseline; on Qwen2.5-7B, gains are 1.0%–3.3%. Both settings reach baseline-equivalent training loss in up to 2$\times$ fewer steps. Recipes discovered on 0.5B proxies transfer to 7B-scale training across Qwen model families.

1 INTRODUCTION

Multimodal large language models (MLLMs) (Liu et al., 2023; Chen et al., 2023; Deitke et al., 2025; Beyer et al., 2024) are increasingly used as the backbone for vision–language applications. A central but underexplored question in this setting is how to compose training data across heterogeneous visual concepts and multimodal objectives. The question is particularly relevant during *midtraining*, when models are trained on high-resolution images and curated annotations to acquire a broad set of vision–language capabilities.

However, data mixture optimization for MLLMs is underexplored. Most MLLMs still rely on simple heuristics (Shukor et al., 2025; McKinzie et al., 2024; Bai et al., 2024; Liu et al., 2025; Roth et al., 2024), largely because systematic exploration is expensive: even a modest search over mixture weights can require dozens to hundreds of training runs, each with substantial compute cost, and the resulting mixtures are often difficult to interpret or transfer across model scales.

In this work, we introduce MixAtlas, a framework for interpretable and compute-efficient multimodal mixture optimization. The core idea is to convert an unstructured collection of midtraining datasets into an explicit, controllable data decomposition along two critical and interpretable axes. We optimize each axis independently; this decoupled design isolates the effect of each factor and keeps the search space low-dimensional.

Along the task supervision axis, we synthesize samples by user-defined objective type (e.g., detailed captioning, OCR), reflecting distinct supervision signals. Along the image concept axis, we use large-scale clustering based on embeddings from the vision encoder to discover concepts automatically. A *mixture* in MixAtlas is then a sampling distribution over each axis: at each step, training examples are drawn according to weights over task supervision or image concepts. This

[*]Work done during an Apple internship

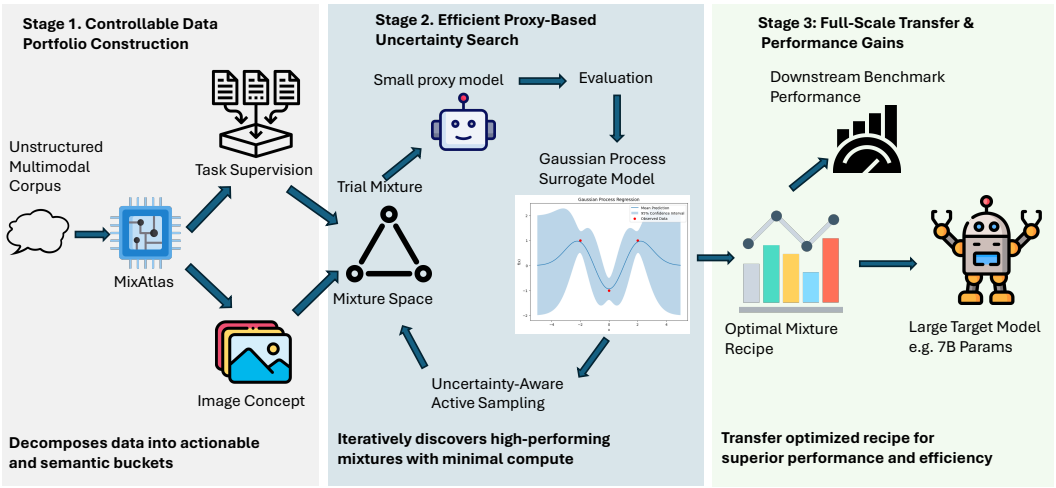


Figure 1: Overview of MixAtlas. Stage 1: MixAtlas converts an unstructured multimodal corpus into a controllable data portfolio by decomposing examples along two interpretable axes: *task supervision* and *image concepts*. Stage 2: We efficiently explore the resulting mixture space using small proxy models and a Gaussian-process surrogate, guided by uncertainty-aware active sampling. Stage 3: The best mixture recipe transfers to full-scale training (e.g., 7B parameters), yielding faster convergence and improved downstream benchmark performance.

decomposition makes mixture design interpretable: rather than treating each dataset as an opaque unit, MixAtlas allows users to diagnose which visual domains or supervision signals are responsible for gains on specific downstream tasks, enabling targeted data collection and informed trade-off decisions when prioritizing capabilities.

To make mixture optimization practical under limited compute, MixAtlas combines small proxy models with an uncertainty-aware policy. We train lightweight proxy models on a small number of selected mixtures, then fit a Gaussian-process surrogate that predicts downstream performance for unseen mixtures while quantifying uncertainty.

Empirically, MixAtlas yields significant improvements in both efficiency and accuracy. Across a diverse benchmark suite, optimized mixtures consistently improve average performance by 8.5%–17.6% on Qwen2-7B and 1.0%–3.3% on Qwen2.5-7B across 10 diverse benchmarks compared with the strongest data mixture baselines (e.g., uniform sampling, RegMix (Liu et al., 2024a), Chameleon (Xie et al., 2025)). Moreover, the learned mixtures accelerate training, reaching a target loss in up to 2× fewer optimization steps. Importantly, mixtures discovered on 0.5B proxy models transfer to larger-scale (7B) training, preserving both the convergence and accuracy benefits and enabling practical mixture optimization without expensive large-model search.

In summary, our contributions are:

- **Interpretable benchmark-driven data recipes.** We formulate multimodal midtraining mixture optimization as deriving interpretable, benchmark-targeted recipes and introduce a two-axis data decomposition over *task supervision* and *image concepts*. This turns an unstructured corpus into a controllable mixture space that users can inspect, adapt, and extend to their own corpora and downstream targets.
- **Uncertainty-aware proxy-based search for multimodal mixtures.** We propose a practical mixture optimization procedure that combines small proxy training with a GP surrogate to efficiently explore the two-axis data decomposition under limited compute. Using the same proxy budget, the GP-based search finds better-performing mixtures compared with other baselines.
- **Systematic empirical study of multimodal midtraining mixture optimization.** We adapt the closest LLM-only data mixture baselines to the multimodal setting and benchmark them in this regime, showing that MixAtlas delivers consistent gains in downstream performance, up to 2× faster convergence, and transferable recipes from 0.5B proxy models to 7B-scale models across two model families.

2 Related Work

Data Mixture Optimization in Pretraining. Data composition has long been recognized as a key driver of pretraining efficiency and downstream generalization, and recent work has moved from heuristic mixing toward *principled* mixture optimization—primarily in language-only settings (Kang et al., 2024). DoReMi (Xie et al., 2023) formulates domain reweighting via distributionally robust optimization, learning mixture weights that improve worst-case domain performance. RegMix (Liu et al., 2024a) proposes proxy-based mixture search, showing that mixtures optimized on small models can transfer to larger scales. Chameleon (Xie et al., 2025) estimates domain importance using leverage scores in a learned embedding space, providing another mechanism for data selection and weighting. Complementary analyses study how domain weights shape scaling behavior and performance trade-offs (Shukor et al., 2025).

In multimodal pretraining and midtraining, mixture design is comparatively less explored and often conducted at a coarse granularity. DataComp (Gadre et al., 2023) performs large-scale studies of filtering and mixing for CLIP-style contrastive training, but focuses on image–text matching rather than multi-objective generative training. Other work studies mixture heuristics over *data formats* (e.g., caption vs. interleaved vs. text), finding fixed ratios that work well across several settings (e.g., the 5:5:1 recipe in MM1 (McKinzie et al., 2024)), with synthetic data further improving certain regimes (Bai et al., 2024; Liu et al., 2025; Wen et al., 2023; Yang et al., 2025b; Yao et al., 2025). Separately, Roth et al. (2024) examine the effect of *data ordering* in continual pretraining. Overall, existing multimodal recipes typically optimize mixtures from a *single perspective* (format or task), rely on inherited ratios or expensive sweeps, and provide limited interpretability about which components drive which downstream gains.

Domain Discovery and Characterization. A related line of work aims to define or discover "domains" in large datasets. Prior work in visual domain adaptation studies distribution shift across styles and content (Peng et al., 2019), while foundational vision-language models demonstrate that semantic structure and visual concepts emerge in learned representations (Radford et al., 2021). Nemotron-CLIMB (Diao et al., 2025) embeds and clusters large-scale data in a semantic space, then iteratively searches for mixtures using a proxy model and predictor. Other efforts treat supervision type as the domain (e.g., PixMo (Deitke et al., 2025) and PaliGemma (Beyer et al., 2024)), or organize web-scale corpora into topic/format taxonomies (Wettig et al., 2025). While these approaches provide useful structure, they typically emphasize a *single axis* (semantic clusters *or* task type) and do not directly support fine-grained, jointly controllable mixtures that account for interactions between *what* is seen (visual concepts) and *how* it is supervised.

Efficient Hyperparameter Optimization. Searching over mixture weights is a challenging hyperparameter optimization problem due to the high-dimensional simplex structure and the cost of each training run. Bayesian optimization (Snoek et al., 2012) provides uncertainty-aware search but can struggle as mixture dimensionality grows and evaluations are expensive. Recent work explores gradient-based data-mixture optimization (Albalak et al., 2023) and jointly learning mixture weights with model parameters (Du et al., 2022), trading off scalability, and compute cost.

3 Methodology

MixAtlas consists of three components as shown in Figure 1: (i)Two-axis domain decomposition into task supervision and image concepts (§3.1), yielding an interpretable data collection.(ii)Proxy-based, uncertainty-aware mixture optimization(§3.2) that searches the mixture space using small proxy models and a probabilistic surrogate. (iii) Mixture transfer to full-scale midtraining(§3.3), producing confidence-rated recipes.

3.1 Two-Axis Domain Decomposition

We decompose the midtraining data along two independent axes (task supervision and image concepts) and optimize each axis independently: Given the decomposition, we form two portfolios: (i) a task portfolio $\mathcal{P}_{\text{task}} = \{P_t : t \in \mathcal{T}\}$ and (ii) a image concept portfolio $\mathcal{P}_{\text{concept}} = \{P_c : c \in \mathcal{C}\}$. Each P_t (resp. P_c) is the subset of examples with that task label (resp. concept cluster id). A mixture h then directly corresponds to interpretable sampling weights over tasks or concepts.

3.1.1 TASK AXIS: TRAINING TASK SUPERVISION

We augment each image example by a task type $t \in \mathcal{T}$ and standardize all tasks into a unified instruction-following format: `<image> <instruction> → <response>`. We explain how we synthesize various task types based on the same image corpus as follows:

Detailed Captioning. We run an in-house model to generate a comprehensive natural-language description capturing objects, attributes, relations, and scene context. *Template:* `<image> Describe this image in detail. → <caption>`

OCR. We run an off-the-shelf OCR engine on each image and concatenate the extracted snippets in raster order. *Template:* `<image> What text appears in this image? → <ocr_tokens>`

Grounded Captioning. We pair captions with object regions obtained from an in-house model, encoding each region as normalized coordinate tokens. *Template:* `<image> Describe this image with grounded regions. → <caption> <ymin><xmin><ymax><xmax>`

Detection. We generate object labels using in-house model(Pix2Seq-style serialization of label–box pairs). *Template:* `<image> Detect all objects.→ <label><ymin><xmin><ymax><xmax>`

VQA. We use the labels in LLaVA-Next midtraining corpus as-is for VQA. *Template:* `<image> <question> → <answer>`

We consider the set of tasks above to construct a diverse set of training objectives and capabilities that are typically reflective of multimodal midtraining. However, MixAtlas is not limited to these task types and can be extended to support any task types that are of interest.

3.1.2 IMAGE CONCEPT AXIS: CLIP-SPACE CONCEPT CLUSTERING

To obtain an interpretable notion of "visual domain" beyond dataset names, we cluster images by semantic concept shown in Appendix Figure 7.

CLIP embeddings. For each image x^v, we compute a vision embedding $z^v = \phi(x^v) \in \mathbb{R}^d$ using a pretrained CLIP (Radford et al., 2021) vision encoder (e.g., ViT-L/14 at 336 resolution). We L2-normalize embeddings prior to clustering, which improves stability for cosine similarity.

Scalable clustering. We fit k-means with $k = 10$ on the embedding set (or on a large random subset if the corpus is very large), then assign every image to its nearest centroid. We choose $k = 10$ as a balance between granularity and interpretability.

Cluster naming and verification. To map clusters to human-readable domains, we inspect random samples from each cluster and assign semantic names based on dominant content. In our data, we identify: (1) informational graphics (charts/diagrams/infographics), (2) natural landscapes, (3) artistic/stylized imagery, (4) documentary photography, (5) close-up/product photography, (6) architecture/interiors, (7) human-centric content, (8) text-heavy documents/UIs, (9) scientific/technical imagery, and (10) entertainment/media.

3.2 PROXY-BASED, UNCERTAINTY-AWARE MIXTURE OPTIMIZATION

Directly searching for $h^\star$ by repeatedly midtraining a large target model is prohibitively expensive. MixAtlas therefore performs search using small proxy models and an uncertainty-aware surrogate.

3.2.1 CANDIDATE POOL GENERATION

We first generate a large pool of candidate mixtures $\mathcal{H} = \{h^{(1)}, \ldots, h^{(M)}\} \subset \Delta^{d-1}$ to ensure coverage of the simplex. We combine:

- Latin hypercube sampling on Δ^{d-1} for space-filling coverage, and
- Dirichlet sampling $h \sim \mathrm{Dir}(\alpha)$ with multiple concentration parameters α to include both specialized mixtures (small α) and near-uniform mixtures (large α).

This pool acts as a discrete approximation to the continuous mixture space that we can score and search efficiently.

3.2.2 PROXY TRAINING AND BENCHMARK EVALUATION

For any candidate mixture h, we train a proxy model M_h (e.g., 0.5B parameters) using the same training recipe as the target model (architecture family, tokenizer, loss, and data formatting), but at reduced scale. All proxy runs use a fixed midtraining budget (e.g., a fixed number of samples/steps), ensuring fair comparison across mixtures.

We then evaluate the proxy on the evaluation suite $\mathcal{V}$ aligned with the downstream target(s). Let

$$s(h) = \text{Eval}(M_h, \mathcal{V}) \tag{1}$$

denote the scalar objective used for search (e.g., a weighted aggregate of normalized benchmark scores, or a single benchmark score for a targeted recipe).

3.2.3 SURROGATE MODEL WITH UNCERTAINTY

We fit a probabilistic surrogate $g(h)$ to predict proxy performance from mixture weights, using the set of evaluated points $\mathcal{S} = \{(h^{(i)}, s^{(i)})\}$ collected so far. We use a Gaussian process (GP) regressor, which yields a posterior predictive mean $\mu(h)$ and standard deviation $\sigma(h)$:

$$g(h) \mid \mathcal{S} \;\Rightarrow\; \mu(h) \;\;, \sigma(h) \;\; . \tag{2}$$

Large $\sigma(h)$ indicates regions poorly supported by existing observations, which is exactly where additional proxy evaluations are most informative.

Interpretability via regression. In addition to the GP used for search, we optionally fit a simple second-order regression model on the same evaluated set,

$$\hat{f}(h) = \beta_0 + \sum_j \beta_j h_j + \sum_{j<k} \beta_{jk} h_j h_k, \tag{3}$$

where β_j captures the marginal contribution of domain j and β_{jk} captures pairwise synergies/antagonisms. This model supports the domain-sensitivity analyses and heatmaps reported in experiments.

3.2.4 UNCERTAINTY-AWARE ACTIVE SAMPLING POLICY

Given the candidate pool $\mathcal{H}$ and current GP posterior, MixAtlas selects which mixtures to evaluate next under a limited proxy budget.

Exploration–exploitation via GP-UCB. To prioritize mixtures that are both promising and uncertain, we use an optimistic acquisition:

$$a_{\text{UCB}}(h) = \mu(h) + \kappa\, \sigma(h), \tag{4}$$

where κ controls exploration.

3.3 MIXTURE TRANSFER TO FULL-SCALE MIDTRAINING

Finally, we transfer the best proxy-discovered recipe $\hat{h}$ to the full-scale target model. The key assumption is that the *relative ranking* of mixtures is preserved across model scales—that is, a mixture that outperforms alternatives at 0.5B parameters will also outperform them at 7B. This assumption is grounded in a growing body of evidence (Liu et al., 2024a; Kang et al., 2024; Diao et al., 2025) that data-level choices transfer more reliably across scales than optimizer hyperparameters. The intuition is that mixture weights control the *distribution of supervision signals* seen during training, which shapes what the model learns to attend to. We empirically validate this transfer assumption in Section 5, showing that recipes discovered on Qwen2-0.5B proxies yield consistent gains when applied to both Qwen2-7B and Qwen2.5-7B, including across model families (Qwen2 $\rightarrow$ Qwen2.5), with the largest improvements on benchmarks compared with proxy-based mixture methods.

4 EXPERIMENT SETUP

Midtraining corpus. We construct a multi-task midtraining corpus by aggregating open-source multimodal datasets. In our current setup, caption-style image–text supervision is primarily sourced from Conceptual Captions (CC3M (Sharma et al., 2018) and CC12M (Changpinyo et al., 2021)).

We build on top of the mid-training dataset proposed by LLaVA-NeXT (Liu et al., 2023) and use the same image corpus but generate labels for other tasks (captioning, OCR, grounded captions, detection) using a combination of open-source and in-house models. The original labels in the dataset are used for the VQA task as-is. For all other tasks, we generate labels for the entire dataset and use them to create different types of task supervision. All datasets are converted into a unified instruction-following format (image + textual prompt $\rightarrow$ text tokens), so that the same autoregressive loss can be used across tasks (see §3). We sample a fixed training budget per run (1 epoch, $\sim$4M samples; see below) so that comparisons differ *only* in the sampling distribution (mixture).

Separate mixture optimization along task and concept axes. We study mixture optimization along two interpretable axes: (i) a *task* portfolio with $|\mathcal{T}| = 5$ supervision types (Dense captioning, OCR, Grounded captioning, Detection, VQA), and (ii) an *image-concept* portfolio with $|\mathcal{C}| = 10$ semantic clusters discovered by running k-means on CLIP image embeddings (§3.1.2). We empirically set k=5, 10, 20 and found that k=10 balanced granularity and search tractability. For concept clustering, we embed images using `openai/clip-vit-large-patch14-336` and L2-normalize embeddings before clustering. For either axis, a mixture is a probability vector over the corresponding portfolio, and we sample training examples according to these mixture weights.

Although the full data distribution can be viewed as a joint mixture over *both* task types and image concepts, directly optimizing this joint space would require substantial compute. We therefore decouple the axes and optimize them separately for controlled and interpretable comparisons: (i) a *task recipe* that optimizes task weights while sampling concepts uniformly, and (ii) a *concept recipe* that optimizes concept weights while sampling tasks uniformly. This design isolates the effect of each factor, enables axis-specific sensitivity analyses, and keeps proxy search tractable; we leave joint optimization over both axes to future work.

Single-task vs. multi-task setup. For the single-task midtraining experiments, we train using only the original VQA task type in the LLaVA-Next datasets, while keeping the total training budget fixed to 1 epoch ($\sim$4M samples). To ensure a fair comparison, we match the total number of training samples across single-task and multi-task runs. In the single-task case, all samples are drawn from that task domain; in the multi-task case, samples are drawn uniformly across the five task types (unless otherwise specified). Therefore, differences between single-task and multi-task results reflect supervision diversity rather than total training volume.

Models. We evaluate MixAtlas on the LLaVA-Next (Liu et al., 2023) training pipeline with two 7B-scale language backbones: Qwen2-7B and Qwen2.5-7B (Yang et al., 2025a). For the vision encoder, we use the same CLIP model (`openai/clip-vit-large-patch14-336`) as in LLaVA-Next, and we keep the overall architecture and training recipe fixed across mixtures. To make mixture optimization compute-efficient, we use a smaller proxy model (Qwen2-0.5B) to predict optimal mixtures and apply them to train bigger models (Qwen2-7B).

Training setup. We follow the standard LLaVA-Next midtraining recipe. Across all compared methods, we keep *everything* fixed except the mixture weights: model architecture, tokenizer, image resolution, maximum sequence length, optimizer/schedule configuration, and total training budget. All full-scale midtraining runs are trained for 1 epoch ($\sim$4M samples) using $64\times$ H100 GPUs.

Evaluation suite and metrics. We evaluate models on a broad suite spanning four categories: (i)Visual understanding: AI2D (Kembhavi et al., 2016), GQA (Ainslie et al., 2023), ScienceQA (Lu et al., 2022). (ii) Document text: DocVQA (Mathew et al., 2021), TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022). (iii) Multimodal reasoning: MMMU (Yue et al., 2024), MM-Bench (Liu et al., 2024b), MM-Vet (Yu et al., 2024). (IV) Mathematical reasoning: MathVista (Lu et al., 2024).

We use the standard metrics for each benchmark (e.g., accuracy for most VQA-style tasks, ANLS for DocVQA, and GPT-based evaluation where applicable), and report all metrics on a 0–100 scale for readability. In this paper, we focus on the generalist recipe experiments. We use a uniform-weighted average of normalized benchmark scores as the scalar objective $s(h)$.

Baselines. To the best of our knowledge, no prior work addresses proxy-based mixture optimization for multimodal midtraining; existing methods (Liu et al., 2024a; Xie et al., 2025) target language-only pretraining. We adapt the most relevant of these to our multimodal pipeline, matching candidate pools, proxy budgets, and evaluation suites for fair comparison:

Table 1: Comparison of Mid-training Strategies on Qwen2-7B on both Task Axis and Image Concept Axis. Gain is computed relative to the strongest baseline for each axis. Improvements (positive gains) and best performing methods are in bold.

	Task Axis				Image Concept Axis				
Benchmark	**Unif**	**Reg**	**MixAtlas**	**Gain**	**Unif**	**Reg**	**Cham**	**MixAtlas**	**Gain**
Visual Understanding									
AI2D	58.16	58.10	**59.84**	**+2.9%**	54.30	53.00	56.50	**57.20**	**+1.2%**
GQA	18.95	20.90	**30.65**	**+46.6%**	25.14	23.35	23.90	**36.48**	**+45.1%**
ScienceQA	72.53	72.30	**74.81**	**+3.1%**	71.60	**72.90**	72.30	71.50	-1.9%
Document and Text									
DocVQA	60.17	58.30	**62.98**	**+4.7%**	3.50	15.80	3.00	**46.90**	**+196.8%**
TextVQA	50.03	48.10	**54.51**	**+9.0%**	29.90	30.00	25.60	**50.90**	**+69.7%**
ChartQA	43.44	46.00	**50.16**	**+9.0%**	33.30	38.20	34.60	**41.80**	**+9.4%**
Multimodal Reasoning									
MMBench-EN	63.40	63.50	**67.35**	**+6.1%**	53.78	**68.39**	59.54	61.89	-9.5%
MMMU Val	39.78	39.80	**41.67**	**+4.7%**	37.40	40.00	38.10	**40.30**	**+0.8%**
MM-Vet	**28.12**	26.10	27.75	-1.3%	27.39	26.15	23.21	**30.37**	**+10.9%**
Mathematical									
MathVista	32.60	31.80	**37.10**	**+13.8%**	35.59	33.40	**36.20**	34.50	-4.7%
AVERAGE	46.72	46.50	**50.68**	**+8.5%**	37.19	40.12	37.30	**47.18**	**+17.6%**

Table 2: Comparison of Mid-training Strategies on Qwen2.5-7B on Task Axis vs. Image Concept Axis. Gain is computed relative to the strongest baseline for each axis. Improvements (positive gains) and best performing methods are in bold.

	Task Axis				Image Concept Axis				
Benchmark	**Unif**	**Reg**	**MixAtlas**	**Gain**	**Unif**	**Reg**	**Cham**	**MixAtlas**	**Gain**
Visual Understanding									
AI2D	58.65	**60.80**	59.97	-1.4%	55.31	**62.21**	58.16	57.19	-8.1%
GQA	32.61	39.00	**45.05**	**+15.5%**	25.14	**44.27**	23.90	36.48	-17.6%
ScienceQA	**74.62**	73.80	73.18	-1.9%	**74.07**	73.53	73.33	71.49	-3.5%
Document and Text									
DocVQA	56.03	66.40	**74.31**	**+11.9%**	22.68	24.05	25.86	**46.83**	**+81.1%**
TextVQA	53.24	56.70	**64.24**	**+13.3%**	33.10	48.76	40.60	**50.92**	**+4.4%**
ChartQA	52.84	53.50	**53.96**	**+0.9%**	38.40	41.16	39.08	**41.88**	**+1.7%**
Multimodal Reasoning									
MMBench-EN	60.40	**72.80**	61.68	-15.3%	63.23	**65.89**	62.97	60.82	-7.7%
MMMU Val	41.11	42.00	**42.67**	**+1.6%**	38.67	**42.56**	39.00	40.33	-5.2%
MM-Vet	31.38	22.60	**31.51**	**+0.4%**	25.60	27.16	27.89	**31.42**	**+12.7%**
Mathematical									
MathVista	37.50	**37.80**	36.10	-4.5%	33.90	**38.00**	34.90	34.70	-8.7%
AVERAGE	49.84	52.54	**54.27**	**+3.3%**	41.01	46.76	42.57	**47.21**	**+1.0%**

- **Uniform**: equal sampling probability over all portfolio elements (uniform over 5 task types for task-axis experiments; uniform over 10 image concepts for concept-axis experiments). This serves as the standard no-optimization reference.
- **Chameleon** (Xie et al., 2025): estimates domain importance via inverse scores computed in a learned image embedding space, then reweights sampling accordingly. Since leverage scores are

defined over image representations, this baseline applies only to the concept axis; it has no natural analogue for the task axis where domains differ in supervision type rather than visual content.
- **RegMix** (Liu et al., 2024a): fits a LightGBM regression model that maps mixture weights to the average benchmark score, splits all proxy evaluations into training data and test data, then selects the mixture with the highest predicted score from the candidate pool. We adapt RegMix to our multimodal setting by replacing its original text domain with our task and concept axis.
- **MixAtlas (ours)**: fits a Gaussian-process surrogate over the same mixture-to-benchmark-score mapping and selects candidates via GP-UCB, which balances exploitation of high-predicted regions with exploration of high-uncertainty regions. To ensure a fair comparison, MixAtlas and RegMix share the same candidate pool and the same proxy evaluation budget (50 runs for the task axis, 200 for the concept axis); the only difference is how each method selects which candidates to evaluate and how it models the performance landscape.

5 RESULTS

MIXATLAS consistently improves both effectiveness and efficiency across our evaluation suite. Across 7B-scale midtraining runs, MixAtlas-learned mixtures (i) improve average benchmark performance over uniform and widely-used data mixture optimization baselines and (ii) improve sample efficiency by reaching the same training loss in much fewer optimization steps. Gains are largest on benchmarks whose requirements align strongly with specific visual concepts, while broad-coverage benchmarks tend to benefit from more diversified concept mixtures.

MIXATLAS recipes outperform strong baselines at 7B scale We evaluate MixAtlas along each axis separately to isolate the effect of each axis, producing two recipes: a *task recipe* (task-supervision weights) and a *concept recipe* (image-concept weights). The task-recipe experiments optimize supervision-type weights while fixing concept weights to uniform; the concept-recipe experiments optimize concept weights while fixing task weights to uniform. The evaluation suite is identical in both cases (same benchmarks and metrics). The difference between the result tables lies solely in which mixture axis was optimized.

- **Task recipe** Table 1 shows that the MIXATLAS task-optimal mixture improves the overall average from 46.72/46.50 (Uniform/RegMix) to 50.68 (+8.5% relative gain over the stronger baseline) and outperforms the stronger of Uniform and RegMix on 9 of 10 benchmarks. The largest improvements occur on GQA (30.65 vs. 20.90, +46.6%), MathVista (37.10 vs. 32.60, +13.8%), and ChartQA/TextVQA (both +9.0% relative). The only degradation is on MM-Vet (-1.3%), highlighting that task reweighting can introduce trade-offs depending on the target. Crucially, both MIXATLAS and RegMix use the same budget of **50 proxy runs** on the 5-dimensional task simplex, yet RegMix's regression-based selection underperforms MIXATLAS on average by a clear margin (46.50 vs. 50.68). This gap highlights the value of uncertainty-aware search: rather than fitting a fixed regression model to all 50 observations and selecting the predicted optimum, MIXATLAS's GP-UCB policy actively steers proxy evaluations toward the most promising and uncertain regions of the simplex, extracting more information from the same compute budget.
- **Concept recipe** Table 1 shows that the MIXATLAS concept-optimal mixture achieves the best overall average (47.18) compared to Uniform (37.19), RegMix (40.12), and Chameleon (37.30), a +17.6% relative gain over the strongest baseline. The largest gains are on document/text benchmarks: DocVQA (46.90 vs. 15.80, +196.8%), TextVQA (50.90 vs. 30.00, +69.7%), and ChartQA (41.80 vs. 38.20, +9.4%), while also improving MM-Vet (30.37 vs. 27.39, +10.9%). The advantage of GP-UCB over regression-based search is even more pronounced on this harder problem: both methods draw from the same pool of **200 proxy runs** on the 10-dimensional concept simplex, yet RegMix's second-order regression model cannot adequately capture the performance landscape in this higher-dimensional space, leaving a 7-point gap to MIXATLAS (40.12 vs. 47.18). The GP surrogate's ability to model non-linear interactions and focus exploration on high-uncertainty, high-reward regions becomes increasingly important as the dimensionality of the mixture space grows. Together, these results show that controlling both supervision composition and concept composition is important; mixtures that rely on fixed heuristics or lower-capacity surrogate models cannot match the task- and domain-aware optima identified by MIXATLAS.

Recipes transfer across model families and scales A key advantage of MixAtlas is that recipes can be discovered using smaller proxy models and then transferred to much larger target models. Table 2 shows that transferring the learned recipe from Qwen2-0.5B proxies to Qwen2.5-7B yields clear gains on multiple benchmarks, especially in document/text and fine-grained understanding:

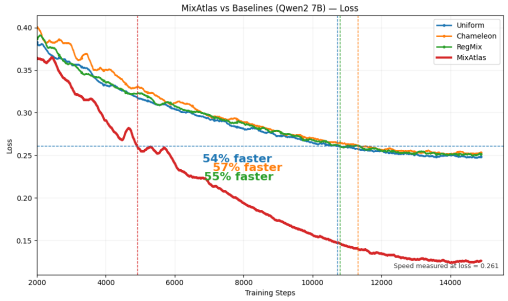


Figure 2: MixAtlas convergence efficiency in Qwen2 7B. It matches baseline final losses at least 54% fewer steps.

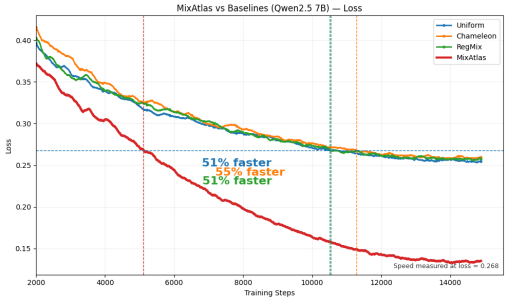


Figure 3: MixAtlas convergence efficiency in Qwen2.5 7B. It matches baseline final losses at least 51% fewer steps.

DocVQA improves from 56.03 (Uniform) to 74.31 (Optimal), TextVQA from 53.24 to 64.24, and GQA from 32.61 to 45.05. We also observe trade-offs: MMBench-EN decreases relative to the strongest baseline (72.80 to 61.68), and MathVista variants see modest drops. These patterns are consistent with the sensitivity maps above: mixtures that prioritize text/document concepts can trade off broad-coverage benchmarks. We note that on the concept axis with the stronger base model (Qwen2.5-7B), gains are concentrated in document/text tasks (DocVQA +81.1%, TextVQA +4.4%, ChartQA +1.7%) while visual understanding and reasoning benchmarks decline, narrowing the overall advantage over RegMix to +1.0%. This suggests that as the base model improves, the marginal benefit of concept reweighting becomes more task-specific, reinforcing the value of MixAtlas as a target-aware recipe tool rather than a uniform-gain method.

MIXATLAS significantly improves training efficiency by reaching the same loss with at least 51% fewer steps. MixAtlas does not only improve final accuracy; it also accelerates optimization. On the target model (LLaVA-Next Qwen2-7B) in Figure 2 , the MixAtlas-optimal mixture reaches the same loss in 54% fewer steps than Uniform, 57% fewer steps than Chameleon and 55% fewer steps than RegMix. Figure 3 shows that at a matched loss level (dashed reference), the optimized mixture reaches the target in roughly 4k steps, while Uniform and Chameleon require close to 11k steps. The loss gap widens as training progresses, indicating that mixture optimization improves both early learning speed and overall optimization efficiency.

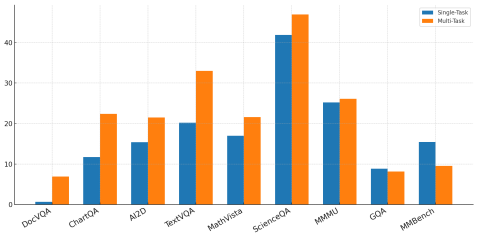


Figure 4: Single-task and multi-task midtraining strategies across benchmark tasks, revealing the benefits of incorporating diverse supervision signals during midtraining.

6 ANALYSIS

Multi-task mixtures surpass single-task midtraining. Beyond optimizing weights, we examine whether supervision diversity itself provides gains. Specifically, we compare single-task midtraining (same total data budget, drawn from only one task type) against uniform multi-task midtraining (same total budget, distributed evenly across tasks).

As shown in Figure 4, multi-task midtraining (captioning + OCR + grounding + VQA + Detection) outperforms single-task training on 8 of 9 benchmarks, with especially large gains on text-heavy benchmarks: ChartQA increases from 12.0 to 22.4 (+10.4), TextVQA from 20.3 to 33.3 (+13.0), and DocVQA from 1.0 to 7.0 (+6.0, from a low baseline). This supports the hypothesis that complementary supervision signals (especially OCR and grounding) add capabilities that single-task VQA supervision does not reliably induce. This highlights the value of studying multi-task supervision effects in multimodal midtraining, emphasizing the importance of quantifying impact of diverse tasks on downstream performance.

Benchmark performance is highly domain- and task-sensitive. To understand *why* different mixtures help different targets, we analyze benchmark sensitivity along two independent axes: task type

Figure 5: Image domain benchmark sensitivity. Heatmaps show per-domain contribution scores for various benchmarks. Cell values indicate the relative contribution of each visual domain to benchmark performance (higher is better).

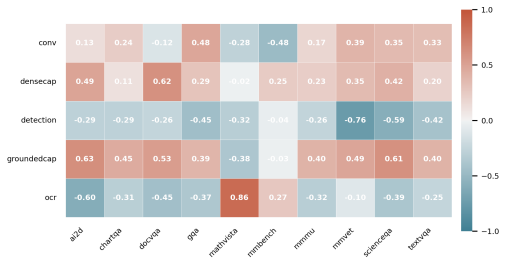


Figure 6: Task Domain–benchmark sensitivity. Heatmaps show per-task domain contribution scores for 10 benchmarks. Cell values indicate the relative contribution of each task domain to benchmark performance (higher is better).

(supervision) and image concept. We compute per-domain "contribution" scores using the fitted proxy surrogate/regression model (§3.2), producing interpretable maps from mixture to downstream behavior. We compute the contribution score using the spearman correlation strength between domain weights and downstream performance.

- **Image-concept sensitivity.** As shown in Figure 5, ChartQA and DocVQA are dominated by *Informational Graphics*, with very large positive contributions (0.95 and 0.86), consistent with their reliance on structured charts and document-like layouts; other domains play secondary roles. In contrast, MMMU exhibits a broad, multi-domain dependence with strong positive contributions spread across Natural Landscapes & Scenery (0.55), Entertainment & Fashion (0.52), Artistic & Stylized Content (0.43), and Close-up Photography (0.41), suggesting that it rewards visual diversity and broader world knowledge. MM-Vet shows a different profile: it gains most from Natural Landscapes (0.50), Entertainment & Fashion (0.45), and also Stock Photo (Alamy) (0.44) and Miscellaneous/Uncategorized (0.36), while Informational Graphics contributes almost nothing (0.01). Finally, Stock Photo (Shutterstock) is consistently negative across benchmarks (down to -0.24 on MMMU), indicating that some stock-photo distributions are weakly aligned (or mildly harmful) for these evaluation targets.
- **Task-type sensitivity.** Figure 6 shows that supervision choices also induce sharply different outcomes. OCR supervision is highly specialized: it strongly benefits MathVista (0.86) and mildly helps MMBench (0.27), but is negative for most other benchmarks. Grounded captioning is broadly beneficial: AI2D, ChartQA, MMMU, MM-Vet, ScienceQA, and TextVQA all peak with grounded captions (0.63, 0.45, 0.40, 0.49, 0.61, 0.40). Dense captioning is the strongest signal for DocVQA (0.62, higher than grounded captioning at 0.53), and conversational supervision is most helpful for GQA (0.48). Detection supervision is consistently harmful (negative for every benchmark, with the largest drop on MM-Vet: -0.76). Overall, these maps show that *there is no single best mixture*: chart/document-centric benchmarks reward specialization, while broad-coverage evaluation often rewards diversity.

7 Conclusion

We presented MixAtlas, an uncertainty-aware framework for optimizing multimodal midtraining data mixtures in a way that is both *interpretable* and *compute-efficient*. MixAtlas converts heterogeneous midtraining corpora into a controllable data decomposition with two axis—task supervision and image concept—and uses proxy training with a probabilistic surrogate to explore the mixture space under a limited budget. We demonstrate that recipes discovered on small proxy models can transfer effectively to 7B-scale training, making mixture optimization practical in real-world compute regimes. Across diverse benchmarks, the resulting recipes improve average benchmark performance over existing baselines and increase training efficiency, reaching the same training loss in far fewer optimization steps. Our sensitivity analysis reveals that no single mixture dominates across all benchmarks, suggesting that future MLLM training pipelines may benefit from maintaining a library of specialized recipes selected per-deployment target, rather than searching for a single universal mixture.

REFERENCES

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023.

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. *arXiv preprint arXiv:2312.02406*, 2023. URL https://arxiv.org/abs/2312.02406.

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. *arXiv.org*, 2024.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023. URL https://arxiv.org/abs/2311.12793.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 91–104, 2025.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan SU, Markus Kliegl, ZIJIA CHEN, Peter Belcak, Yoshi Suhara, Hongxu Yin, Mostofa Patwary, Yingyan Celine Lin, Jan Kautz, and Pavlo Molchanov. Nemotron-CLIMB: Clustering-based iterative data mixture bootstrapping for language model pre-training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=aBlqKPkc4a.

Nan Du, Yanping Huang, Andrew M. Dai, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5547–5569. PMLR, 2022. URL https://proceedings.mlr.press/v162/du22c.html.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, and et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. URL https://arxiv.org/abs/2304.14108.

Feiyang Kang, Yifan Sun, Bingbing Wen, Si Chen, Dawn Song, Rafid Mahmood, and Ruoxi Jia. Autoscale: Scale-aware data mixing for pre-training llms. *arXiv preprint arXiv:2407.20177*, 2024.

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016. URL https://api.semanticscholar.org/CorpusID:2682274.

Emmy Liu, Graham Neubig, and Chenyan Xiong. Midtraining bridges pretraining and posttraining distributions. *ArXiv*, abs/2510.14865, 2025. URL https://api.semanticscholar.org/CorpusId:282138804.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. URL https://arxiv.org/abs/2304.08485. NeurIPS 2023 (Oral).

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. *arXiv preprint arXiv:2407.01492*, 2024a. URL https://arxiv.org/abs/2407.01492. ICLR 2025 (to appear).

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.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 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 (ICLR)*, 2024.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 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, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Brandon McKinzie, Zhe Gan, J. Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training. *ArXiv*, abs/2403.09611, 2024. URL https://api.semanticscholar.org/CorpusId:268384865.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Peng_Moment_Matching_for_Multi-Source_Domain_Adaptation_ICCV_2019_paper.html.

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*, pp. 8748–8763. PmLR, 2021.

Karsten Roth, Vishaal Udandarao, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier J Henaff, Samuel Albanie, Matthias Bethge, and Zeynep Akata. A practitioner's guide to continual multimodal pretraining. In *NeurIPS 2024 Workshop on Scalable Continual Learning for Lifelong Foundation Models*, 2024. URL https://openreview.net/forum?id=gkyosluSbR.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*, 2025.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, volume 25, pp. 2951–2959, 2012. URL https://proceedings.neurips.cc/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

Bingbing Wen, Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Bill Howe, and Lijuan Wang. Infovisdial: An informative visual dialogue dataset by bridging large multimodal and language models. *arXiv preprint arXiv:2312.13503*, 2023.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=boSqwdvJVC.

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 *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dcba6be91359358c2355cd920da3fcbd-Paper-Conference.pdf.

Wanyun Xie, Francesco Tonin, and Volkan Cevher. Chameleon: A flexible data-mixing framework for language model pretraining and finetuning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=mDxarRaTY9.

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*, 2025a.

Yiwei Yang, Chung Peng Lee, Shangbin Feng, Dora Zhao, Bingbing Wen, Anthony Zhe Liu, Yulia Tsvetkov, and Bill Howe. Escaping the spuriverse: Can large vision-language models generalize beyond seen spurious correlations? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b. URL https://openreview.net/forum?id=es2NkPKFCB.

Jihan Yao, Yushi Hu, Yujie Yi, Bin Han, Shangbin Feng, Guang Yang, Bingbing Wen, Ranjay Krishna, Lucy Lu Wang, Yulia Tsvetkov, et al. Mmmg: a comprehensive and reliable evaluation suite for multitask multimodal generation. *arXiv preprint arXiv:2505.17613*, 2025.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International conference on machine learning*. PMLR, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

A APPENDIX

A.1 LIMITATIONS

While MixAtlas provides an interpretable and compute-efficient approach to multimodal mixture optimization, it has several limitations.

Disjoint optimization across axes. In this work, we treat the two decomposition axes—task type and image concept—as parallel "knobs" and optimize mixtures on each axis separately. We do *not* optimize the full joint cross-product portfolio (task $\times$ concept), which could capture important interaction effects (e.g., OCR on document-like images versus OCR on natural photos). Joint optimization is attractive but increases the dimensionality of the search space; developing scalable joint methods (e.g., hierarchical/low-rank parameterizations or structured priors for the surrogate) is an important next step.

Limited model and pipeline coverage. Our experiments focus on LLaVA-Next style midtraining with Qwen-family backbones (Qwen2-7B and Qwen2.5-7B), and proxy runs based on smaller Qwen variants. Although we demonstrate transfer across scales and across language models (Qwen2 $\rightarrow$ Qwen2.5), it remains to be seen how well the discovered recipes generalize to other language backbones (e.g., LLaMA-family), other vision encoders, and other MLLM training pipelines. This is by design, since we want to discover recipes for a particular model, and believe that MixAtlas allows users to curate benchmark specific optimal recipes, however, discovering mixtures that transfer well across all different model scales and types is something we leave for future work.

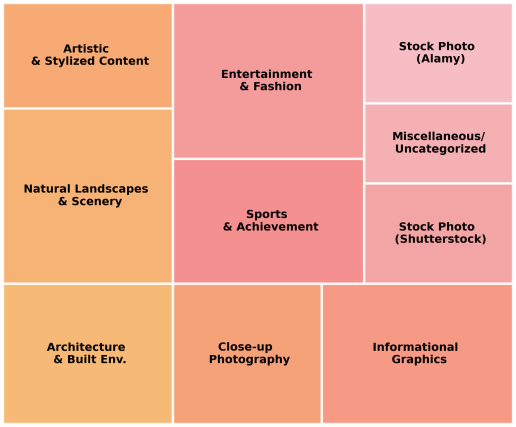
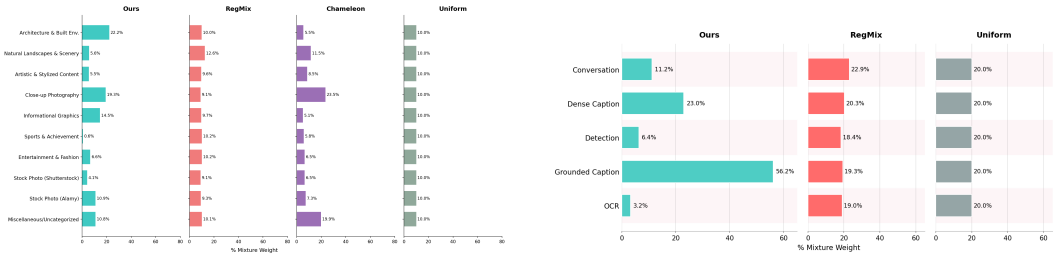


Figure 7: Image concept distribution induced by CLIP-space clustering.

A.2 ADDITIONAL RESULTS AND ANALYSIS

Optimal mixtures analysis MixAtlas learns a sparse, interpretable recipe rather than a near-uniform one shown in Figure 8. On the image-concept axis, it emphasizes Architecture & Built Environment (22.2%), Close-up Photography (19.3%), and Informational Graphics (14.5%), which matches Figure 5: Informational Graphics is the strongest positive domain for ChartQA (0.95) and DocVQA (0.86). This specialization explains the large concept-axis gains in Table 1 on DocVQA (15.8$\rightarrow$46.9), TextVQA (30.0$\rightarrow$50.9), and ChartQA (38.2$\rightarrow$41.8), yielding the best average of 47.18 (+17.6%). On the task axis, MixAtlas puts most weight on Grounded Caption (56.2%) and Dense Caption (23.0%), consistent with Figure 6, where grounded captioning is broadly helpful and dense captioning is strongest for DocVQA; accordingly, the task recipe reaches the best Qwen2-7B average of 50.68 (+8.5%) and also transfers to Qwen2.5-7B with gains on GQA, DocVQA, and TextVQA.

Benchmark trade-off discussion Figure 5 and Figure 6 also explain why MixAtlas does not improve every benchmark simultaneously. Figure 5 shows that chart/document benchmarks reward specialized concepts, whereas broader benchmarks such as MMMU benefit from more diverse visual domains; Figure 6 similarly shows benchmark-specific supervision needs, with grounded cap-



(a) (b)

Figure 8: (a) Image concept domain weight distribution comparison across data mixing strategies. (b) Task type weight distribution comparison across different data mixing strategies. Our method assigns higher weight to Grounded Caption (56.2%) compared to RegMix and Uniform baselines, which maintain near-uniform distributions across tasks.

tioning helping ScienceQA, OCR helping MathVista, and conversational supervision helping GQA. As a result, recipes that favor document/text-oriented concepts and grounded or dense captions deliver large gains on DocVQA, TextVQA, and ChartQA, but can hurt broader benchmarks such as MMBench-EN or MathVista, as seen in Table 1 and Table 2. This becomes even clearer on Qwen2.5-7B, where concept reweighting still helps document tasks but yields only a modest +1.0% average gain because several broader benchmarks decline. These drops are therefore the expected cost of target-specific optimization, and MixAtlas makes that trade-off controllable by allowing users to reweight objectives or impose benchmark constraints during search.