Supplementary Material: Understanding Pruning Regimes in Vision-Language Models through Domain-Aware Layer Selection

Saeed Khaki
Microsoft AI
saeedkhaki@microsoft.com

Nima Safaei
The Ohio State University
safaei.3@osu.edu

Kamal Ginotra
Microsoft AI
kamalginotra@microsoft.com

This document provides supplementary material for the main paper, including baseline reference scores, probing prompts, and additional analysis figures.

1. Baseline and Baseline+SFT Reference

Table 1. Reference accuracies for unpruned baselines. "Baseline" is without fine-tuning. "Baseline+SFT" applies the same fixed SFT used for post-pruning stabilization.

Dataset	2B Base	2B +SFT	4B Base	4B +SFT
Snapask	40.47	32.83	50.75	44.96
NuminaMath	63.45	52.61	73.49	64.06
LLaVA-OV	59.12	62.00	67.26	68.00
ChartQA	79.72	32.60	84.00	35.12
RWQA	65.23	66.93	71.50	72.29
ScienceQA	86.47	87.80	92.71	93.46
VStar	78.01	78.01	81.68	82.20

The baseline+SFT results do not represent an attempt to improve task performance. The brief fine-tuning stage is designed solely to restore generation stability after structural pruning. Because the SFT mixture is dominated by general multimodal instruction-following data rather than task-specific supervision, it can shift behavior away from the original distribution. This explains why baseline+SFT sometimes underperforms the purely unpruned baseline on certain datasets.

2. Activation Capture and Processing Details

All activation collection was performed on frozen models without gradient tracking.

Forward-pass instrumentation. For each decoder layer ℓ, we register a pre-forward hook to capture its input hidden states H_ℓ^{in} and a post-forward hook to capture its output hidden states H_ℓ^{out}. Both tensors are collected in float precision and stored on CPU memory after each forward step. No modifications are made to the model architecture or weights.

Tokenization and modality formatting. Each example consists of a single user message containing both an image and a text prompt. Images are represented using the model's built-in visual token format (vision start tokens, image patch tokens, padding tokens, and vision end tokens). The text prompt is appended using the model's standard chat template, including role tags and system tokens if required by the underlying model.

Pooling. To reduce storage and enable baselines to reuse activations, we compute mean-pooled representations $\bar{h}_\ell^{\text{in}},\ \bar{h}_\ell^{\text{out}} \in \mathbb{R}^d$ by averaging across both batch and token dimensions. These vectors preserve coarse information about layer transformations and enable methods such as CKA and Interlace to operate without access to full activations.

Domains and sample sizes. We collect activations separately for math and non-math domains. The math domain uses 5,000 image-prompt examples, while the non-math domain uses 4,000 examples, spanning captioning, entity listing, count-based VQA, and grounding tasks.

3. Probing Prompts

We use a fixed set of task prompts to ensure consistent activation probing across math and non-math domains. Each prompt is paired with an image and inserted into the model's multimodal chat template.

Math domain. **Math-CoT:** "Solve the following mathematical problem step by step. Explain your reasoning clearly and provide the final answer. Problem: Look at the math problem in the image and solve it." **Math-Direct:** "Solve

the following mathematical problem. Provide only the final answer, without explanation. Problem: Look at the math problem in the image." **Math-Rephrase:** "Rephrase the following math problem in simpler words, without solving it. Problem: Look at the math problem in the image." **Math-Formalize:** "Convert the following math problem into a set of mathematical equations. Do not solve the equations. Problem: Look at the math problem in the image." **Math-Verify:** "A person claims they solved the math problem in the image. Look at the problem and determine if it appears solvable. Respond with your assessment and explain briefly."

Non-math domain. **Captioning:** "Describe the image in one complete, factual sentence. Avoid opinions or speculation." **Entity Listing:** "List the main objects or entities visible in the image. Return a comma-separated list using short noun phrases." **Counting VQA:** "Question: How many main objects are visible in the image? Answer with a single number." **Grounding:** "Identify the object referred to by the phrase 'the most prominent object' and describe it briefly using one short phrase."

4. Additional Figures

We provide additional visualizations: relative accuracy drops (Figure 1), per-benchmark curves for 2B (Figure 2) and 4B (Figure 3), and the best-method summary (Figure 4).

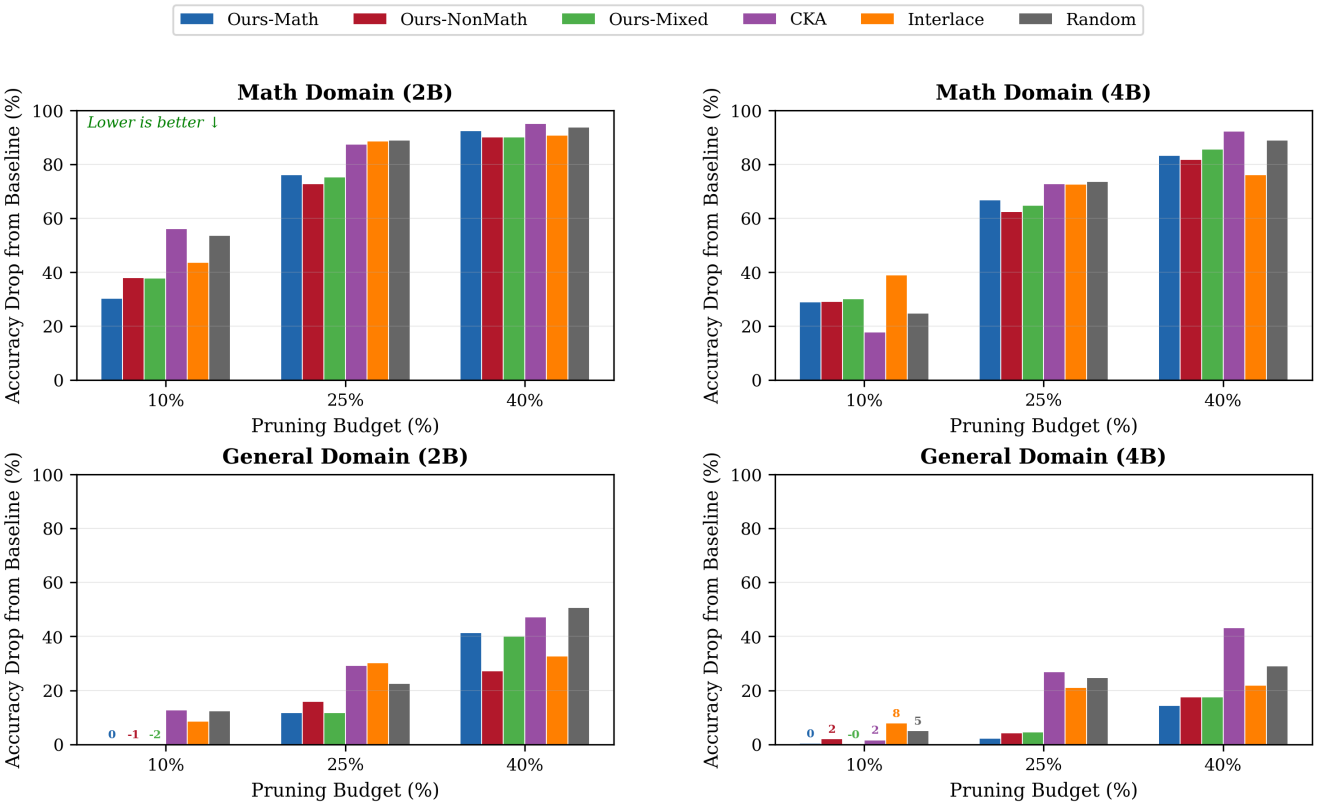

Figure 1. Relative accuracy drop from baseline by domain, model size, and pruning budget.

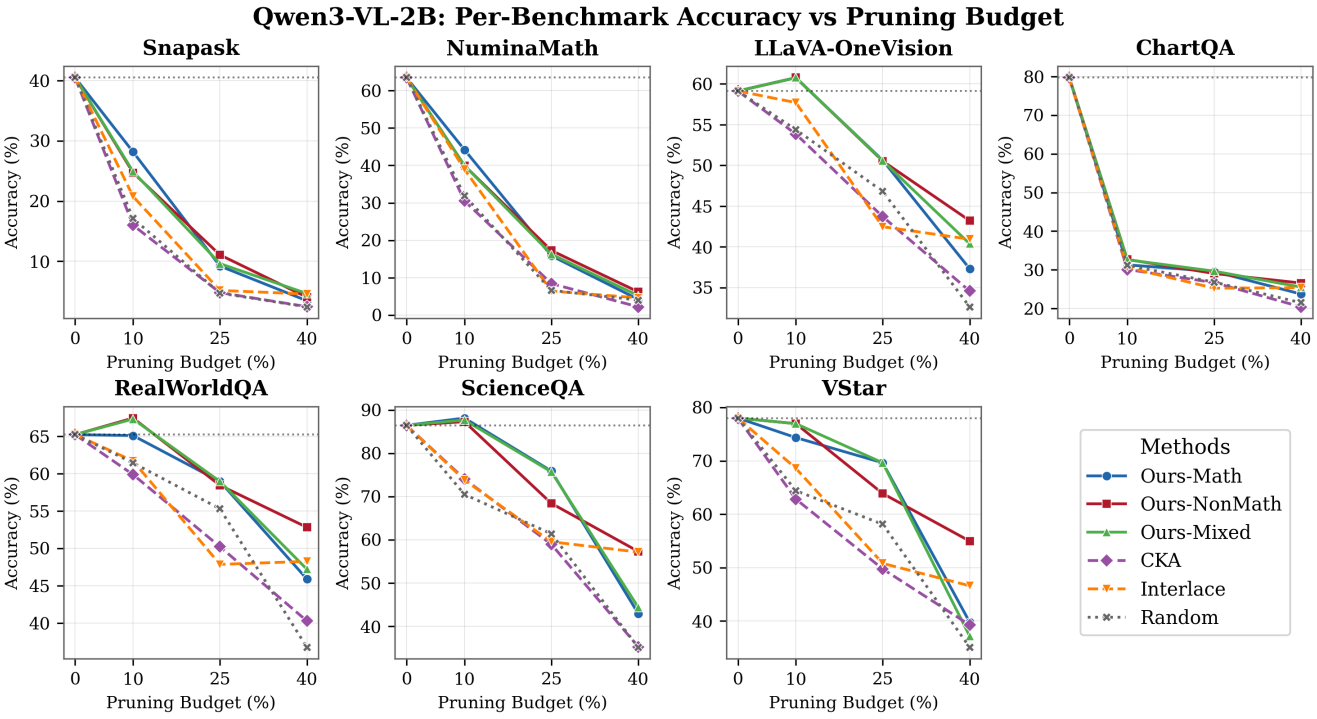

Figure 2. Per-benchmark accuracy curves for Qwen3-VL-2B.

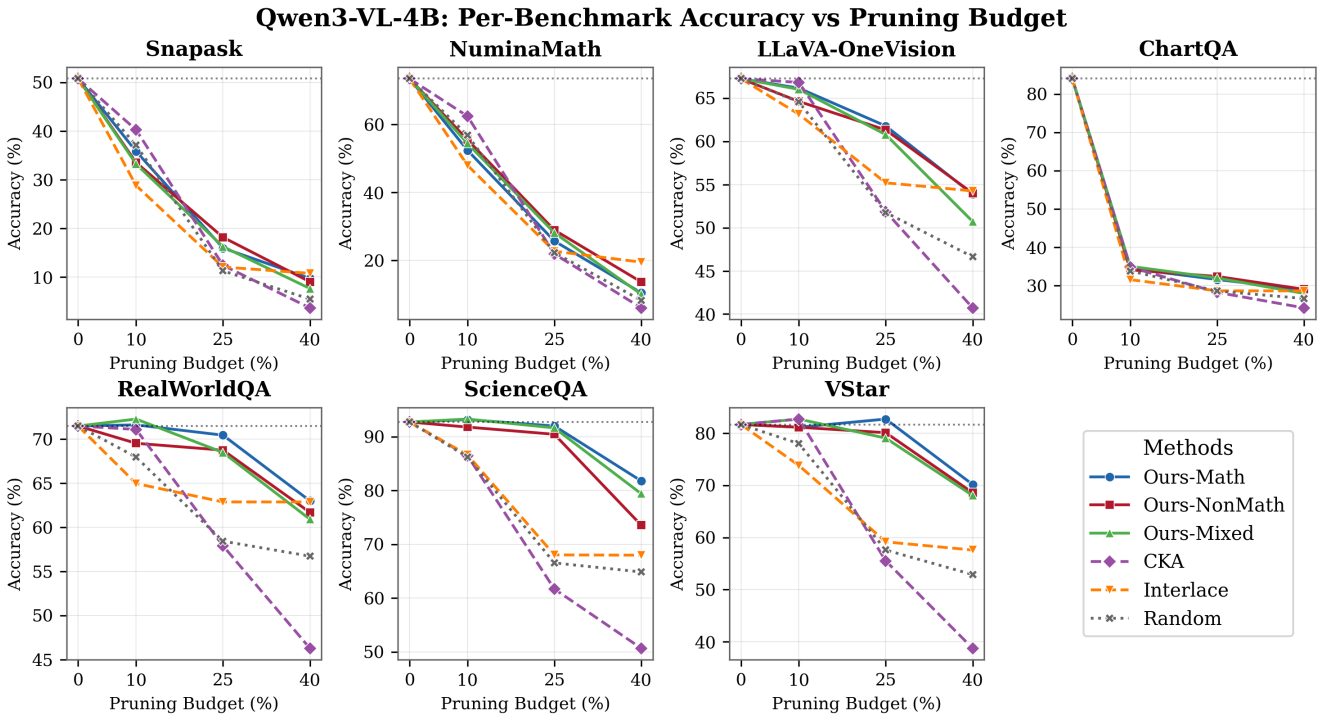

Figure 3. Per-benchmark accuracy curves for Qwen3-VL-4B.

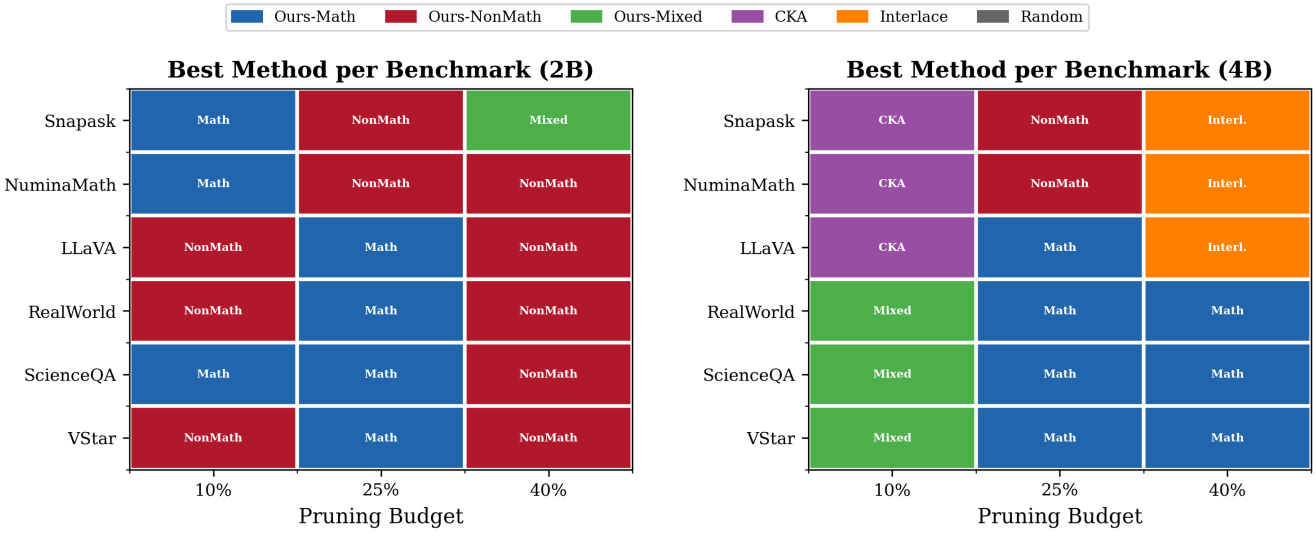

Figure 4. Best method per benchmark and pruning budget for 2B and 4B.