WINNING THE PRUNING GAMBLE: A UNIFIED APPROACH TO JOINT SAMPLE AND TOKEN PRUNING FOR EFFICIENT SUPERVISED FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

As supervised fine-tuning (SFT) evolves from a lightweight post-training step into a compute-intensive phase rivaling mid-training in scale, data efficiency has become critical for aligning large language models (LLMs) under tight budgets. Existing data pruning methods suffer from a fragmented design: they operate either at the sample level or the token level in isolation, failing to jointly optimize both dimensions. This disconnect leads to significant inefficiencies—high-value samples may still contain redundant tokens, while token-level pruning often discards crucial instructional or corrective signals embedded in individual examples. To address this bottleneck, we introduce the *Error–Uncertainty (EU) Plane*, a diagnostic framework that jointly characterizes the heterogeneous utility of training data across samples and tokens. Guided by this insight, we propose *Quadrant-based Tuning (Q-Tuning)*, a unified framework that strategically coordinates sample pruning and token pruning. Q-Tuning employs a two-stage strategy: first, it performs sample-level triage to retain examples rich in informative misconceptions or calibration signals; second, it applies an asymmetric token-pruning policy, using a context-aware scoring mechanism to trim less salient tokens exclusively from misconception samples while preserving calibration samples in their entirety. Our method sets a new state of the art across five diverse benchmarks. Remarkably, on SmolLM2-1.7B, Q-Tuning achieves a +38% average improvement over the full-data SFT baseline using only 12.5% of the original training data. As the first dynamic pruning approach to consistently outperform full-data training, Q-Tuning provides a practical and scalable blueprint for maximizing data utilization in budget-constrained LLM SFT.

1 INTRODUCTION

The explosive growth of alignment datasets—now routinely spanning billions of tokens—has fundamentally transformed Supervised Fine-Tuning (SFT) from a lightweight post-training step into a compute-intensive phase rivaling mid-training in scale (Ouyang et al., 2022; Dong et al., 2024; Yang et al., 2025; Achiam et al., 2023; Team et al., 2023). In this new regime, the primary challenge is no longer simply reducing data volume, but maximizing the utility of every retained token and sample—a task that demands accurate, on-the-fly estimation of data value. Yet despite the emphasis on data efficiency, recent work reveals a troubling paradox: even sophisticated dynamic pruning heuristics often underperform *simple random sampling* (Xia et al., 2024b). This starkly exposes a fundamental disconnect between current strategies and the true utility of alignment data.

At the heart of this challenge lies the fragmented design of existing pruning strategies. *Sample-level pruning* methods (Qin et al., 2024; Zhou et al., 2023a; Wang et al., 2025a; Yang et al., 2024) identify high-potential examples but treat all tokens within them as equally valuable—retaining redundant or even harmful content that dilutes alignment signals. Conversely, *token-level pruning* approaches (Lin et al., 2024; Xia et al., 2025; Chen et al., 2024; Zhang et al., 2024b) apply context-agnostic heuristics uniformly across the dataset, blind to the semantic role of each sample. Such a one-size-fits-all policy fails to differentiate, for instance, between a sample containing a correctable misconception—where only specific tokens need refinement—and one serving as a calibration anchor,

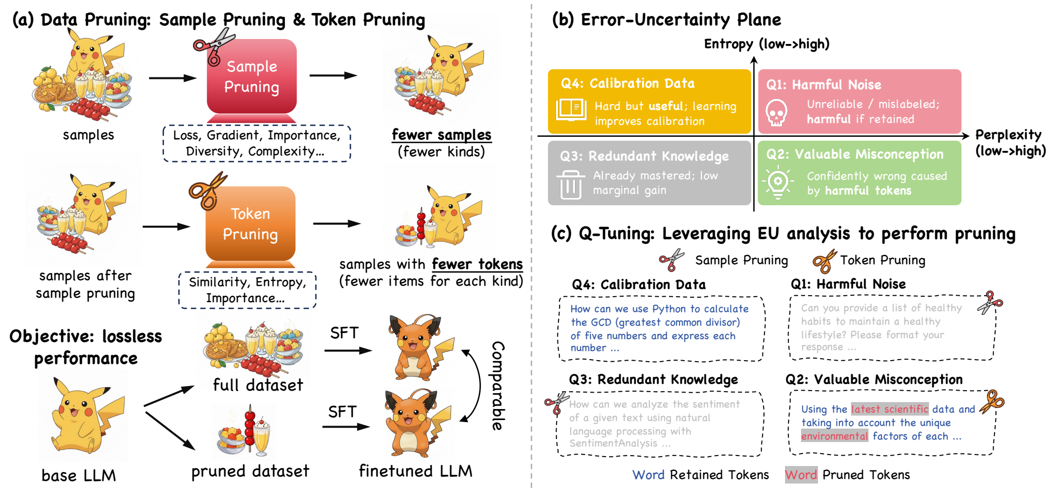

Figure 1: **(a) Data pruning.** *Sample pruning* selects training instances using signals such as loss, gradients, importance, diversity, or complexity; *token pruning* removes low-value tokens within retained samples (*e.g.*, via similarity, entropy, or token importance). **(b) Error–Uncertainty (EU) plane.** We partition samples by perplexity and entropy into four regions: Q1 (harmful noise), Q2 (valuable misconceptions), Q3 (redundant knowledge), and Q4 (calibration data). **(c) Q-Tuning.** Q-Tuning performs joint pruning guided by the EU plane: it drops Q1 and Q3, selectively prunes tokens in Q2, and retains Q4 in full.

which must be preserved holistically to maintain model stability. By operating in isolation, neither paradigm captures the interdependent nature of sample and token utility. This raises a central question: *How can we dynamically coordinate sample selection and token pruning within a unified framework to maximize the true learning utility of limited data?*

To address this, we first formalize the problem as **Generalized Dynamic Data Pruning**, a bilevel optimization framework for jointly optimizing sample and token pruning. We then introduce a novel diagnostic lens, the **Error-Uncertainty (EU) Plane**, which categorizes training instances by mapping model error (perplexity) against model uncertainty (entropy). Specifically, as shown in Figure 1(b), training samples are categorized into four quadrants based on perplexity (higher indicating more wrong) and entropy (higher indicating more uncertain): (i) *Q1 (Harmful Noise)* — unreliable or mislabeled data that actively harms learning and should therefore be removed via sample-level pruning; (ii) *Q2 (Valuable Misconception)* — confidently wrong responses that, when pruned surgically at the token level, can be transformed into powerful teaching signals, making them ideal candidates for token-level pruning; (iii) *Q3 (Redundant Knowledge)* — mastered content offering diminishing returns, best eliminated through sample-level pruning to improve efficiency without sacrificing performance; and (iv) *Q4 (Calibration Data)* — hard but reliable samples essential for improving model confidence and robustness, which should be preserved in full — neither sample nor token pruning should be applied.

This insight motivates our solution: **Quadrant-based Tuning (Q-Tuning)**, the first principled, integrated method for dynamic data pruning. Guided by EU Plane analysis, Q-Tuning implements a two-stage, context-aware strategy, as shown in Figure 1(c). First, at the sample level, it acts as an intelligent triage system: retaining samples that offer clear signals for error correction or calibration (Q2 and Q4), while discarding those classified as harmful noise or redundant knowledge (Q1 and Q3). Crucially, Q-Tuning then applies an asymmetric token policy: for confidently wrong samples (valuable misconceptions), it performs token pruning to isolate the core misconception and amplify the learning signal; for uncertain but correct samples (calibration data), it preserves full token sequences to ensure robust uncertainty modeling. Our contributions are as follows:

1. We formulate the joint sample–token pruning problem as **Generalized Dynamic Data Pruning**, a bilevel optimization framework for hybrid pruning.
2. We introduce the **Error-Uncertainty (EU) Plane**, a tool that reveals and quantifies the heterogeneous value of data across error and uncertainty dimensions. Based on the EU analysis, we propose **Q-Tuning**, the first integrated, diagnosis-driven algorithm for dynamic pruning that coordinates sample and token decisions.
3. Experiments demonstrate that Q-Tuning exceeds full-data training and all existing pruning baselines across 4 models, 5 benchmarks, and 6 different kinds of data budgets. Particularly, with

LLaMA3-8B on GSM8K, Q-Tuning reaches 48.07 using only 35% of the data, outperforming full-data training by 6.0 points and the strongest baseline by 9.9 points.

2 Generalized Dynamic Data Pruning: A Unified Framework

We first propose *Generalized Dynamic Data Pruning*, a framework for accelerating model training where samples and their constituent tokens are selectively and adaptively omitted at each step.

Specifically, the framework first considers the coarse-grained *sample level pruning*, which involves identifying and discarding examples from a mini-batch deemed less informative for the model's state. Subsequently, *token-level pruning* operates on the retained samples, performing a finer-grained selection to keep a critical subset of tokens within each. This two-stage process is inherently *dynamic*: pruning criteria can be re-evaluated for each new mini-batch, allowing the data distribution used for gradient updates to evolve alongside the model's parameters θ. The overarching objective is to focus computation on a "doubly-pruned" data subset to maximize training efficiency while preserving or enhancing the model's generalization.

To formalize this framework, consider a model f_θ with parameters θ. At each training step t, a mini-batch $\mathcal{B}_t$ is drawn from the training distribution $\mathcal{D}$. The hierarchical pruning process can be modeled by two abstract operators: a sample-level pruner Ψ and a token-level pruner Φ.

Stage 1: Sample-Level Pruning. The operator Ψ splits mini-batch $\mathcal{B}_t$ into kept and discarded samples, governed by keep ratio $r_{\text{sample}} \in [0, 1]$. The retained set, $\mathcal{B}'_t$, is defined as:

$$\mathcal{B}'_t = \mathcal{B}_t \setminus \Psi(\mathcal{B}_t) \subseteq \mathcal{B}_t, \tag{1}$$

where $\Psi(\mathcal{B}_t)$ represents the set of discarded samples, and the size of the retained set is constrained such that $|\mathcal{B}'_t| = \lfloor r_{\text{sample}} \cdot |\mathcal{B}_t| \rfloor$.

Stage 2: Token-Level Pruning. For each sample in the retained set $\mathcal{B}'_t$, the operator Φ determines which tokens to keep, guided by a token keep ratio $r_{\text{token}} \in [0, 1]$. This is modeled by generating a binary mask $m(x) \in \{0, 1\}^{L(x)}$ for each sample $x \in \mathcal{B}'_t$, where $L(x)$ is its sequence length. The final, doubly-pruned mini-batch $\tilde{\mathcal{B}}_t$ is constructed by applying these masks:

$$\tilde{\mathcal{B}}_t = \Phi(\mathcal{B}'_t) = \{m(x) \odot x \mid x \in \mathcal{B}'_t\}, \tag{2}$$

where $\odot$ denotes element-wise product, and each mask satisfies $\|m(x)\|_1 = \lfloor r_{\text{token}} L(x) \rfloor$.

Generalized Dynamic Data Pruning. We now put all things together into a unified framework. The central problem of this framework is to identify the optimal dynamic pruning operators, (Φ, Ψ), that guide the training process to a final model θ^* with the best possible generalization performance. These operators are applied at each step t to transform a mini-batch $\mathcal{B}_t$ into a computationally cheaper, pruned version $\tilde{\mathcal{B}}_t$, while adhering to predefined keep ratios.

This problem is naturally captured as a *bi-level optimization problem*. The outer loop seeks optimal pruners, while the inner loop represents the iterative training procedure that produces the final model parameters under the guidance of these pruners. Formally, the objective is as follows:

$$\begin{aligned} &\min_{\Phi, \Psi} \quad \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}} \left[\mathcal{L}_{\text{test}}(f_{\theta^*}(x), y)\right] \\ \text{s.t.} \quad &\theta^* = \arg\min_{\theta} \sum_{t=1}^{T} \mathbb{E}_{\mathcal{B}_t\sim\mathcal{D}} \mathbb{E}_{(x,y)\sim\tilde{\mathcal{B}}_t} \left[\mathcal{L}_{\text{train}}(f_{\theta_t}(x), y)\right] \end{aligned} \tag{3}$$

where at each step t: $\tilde{\mathcal{B}}_t = \Phi(\mathcal{B}_t \setminus \Psi(\mathcal{B}_t))$, with keep ratios $(r_{\text{sample}}, r_{\text{token}})$.

Specifically, the *outer objective* defines the quality metric for any pair of pruners (Φ, Ψ): the final test performance of the model they produce. The *inner objective* defines the training process itself, where the final parameters θ^* result from cumulatively minimizing the loss over a sequence of dynamically pruned mini-batches. All existing methods that instantiates these operators Φ and Ψ can be seen as a specific solution to this alignment problem.

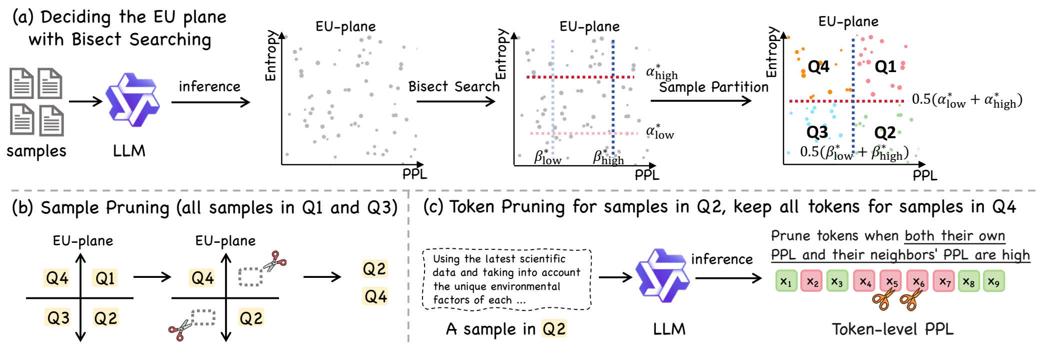

Figure 2: **(a) Constructing the Error-Uncertainty (EU) Plane via Bisect Search.** We run the base LLM to compute sample-level perplexity (PPL) and entropy, and use bisect search to set thresholds $(\alpha^*_{\text{low}}, \alpha^*_{\text{high}}, \beta^*_{\text{low}}, \beta^*_{\text{high}})$ that partition the EU plane into Q1–Q4. **(b) Sample Pruning for Q1 and Q3.** Samples in Q1 and Q3 are pruned at the sample level, while those in Q2 and Q4 are retained. **(c) Token-Level Pruning for Q2 Samples.** For retained Q2 samples, token-level pruning is applied based on both the token's own perplexity and the average perplexity of its neighboring tokens. Tokens with high local PPL are removed, preserving only the most informative ones.

3 Winning the Pruning Gamble

Building on the previous analysis, we now introduce **Q-Tuning**, a dynamic pruning method guided by the Error–Uncertainty (EU) Plane. As shown in Figure 1, Q-Tuning proceeds in two coordinated stages: it first prunes harmful noise and redundant knowledge at the sample level, and then applies targeted token-level pruning to valuable misconceptions while preserving calibration data in full. This integrated strategy avoids the failure modes of one-dimensional heuristics and transforms pruning from a gamble into a principled, performance-enhancing process.

3.1 The Error-Uncertainty Plane: A Diagnostic Lens

We now formalize the **Error–Uncertainty (EU) Plane** introduced in Figure 1(b). Each training sample is positioned by two orthogonal axes: *error*, quantified by perplexity (PPL), and *uncertainty*, quantified by predictive entropy (Ent).

Measure Error via Perplexity. For a training sample (x, y) with sequence length $L(x)$, let $T(x) \subseteq \{1, \ldots, L(x)\}$ denote the set of trainable token positions. We define sample-level error as the perplexity (PPL), derived from the average token-level negative log-likelihood:

$$\text{PPL}(x, y; f_\theta) = \exp\Big(\sum_{i \in T(x)} \frac{-\log p(y_i \mid x, y_{<i}; f_\theta)}{|T(x)|} \Big) \tag{4}$$

High perplexity indicates the model finds the ground-truth continuation highly surprising, which is a sign of either genuine difficulty or misconception.

Measure Uncertainty via Predictive Entropy. Independent of correctness, we quantify the model's uncertainty using entropy (Ent), computed as the average token-level entropy over trainable positions, where $v \in \mathcal{V}$ denotes a vocabulary token:

$$\text{Ent}(x, y; f_\theta) = \frac{\sum_{i \in T(x)} H(p_\theta(\cdot \mid x, y_{<i}))}{|T(x)|}. \tag{5}$$

High entropy indicates that the model spreads probability mass broadly, reflecting persistent uncertainty even when the top prediction is correct. Taken together, PPL and Ent map each sample onto the EU Plane, providing a principled and interpretable basis for partitioning data into the four quadrants illustrated in Figure 1(b).

3.2 Q-TUNING: GENERALIZED DYNAMIC DATA PRUNING

Q-Tuning operationalizes the insights of the EU Plane into a two-stage, per-batch dynamic pruning strategy, as illustrated in Figure 2. In the first stage, bisect search determines quantile-based thresholds that partition samples into four quadrants, enabling sample-level pruning of uninformative data. In the second stage, token-level pruning is selectively applied within retained misconception samples to isolate informative signals, while calibration samples are preserved in full. The pseudocode of our method is illustrated in Algorithm 1 in Appendix D.

Stage 1: Constructing the EU Plane via Binary Search and Pruning Samples Accordingly. At each training step, we compute the perplexity (PPL) and entropy (Ent) for every sample $x \in \mathcal{B}_t$ using a gradient-free forward pass of the current model f_{θ_t}. These statistics map each sample to a point on the EU Plane (Figure 2(a)).

Our objective is to determine quantile thresholds (α^*, β^*) on the PPL and Entropy axes such that the retained proportion of samples in $Q_2 \cup Q_4$ exactly matches the target sample retention ratio r_{sample}. Here $\text{Quantile}_\gamma(X)$ denotes the γ-quantile of a variable X, *i.e.*, the smallest value q such that at least a fraction γ of samples satisfy $X \leq q$.

To locate (α^*, β^*), we perform a **bisect search** on both axes. For each axis, the search interval is initialized as $[0, 0.5]$. At each iteration we set $\alpha = \frac{1}{2}(\alpha_{\text{low}} + \alpha_{\text{high}})$, $\beta = \frac{1}{2}(\beta_{\text{low}} + \beta_{\text{high}})$, and compute the proportion of samples that would fall into the tentative $Q_2 \cup Q_4$ region defined by the current thresholds (α, β). The intervals are then updated as

$$(\alpha_{\text{low}}, \alpha_{\text{high}}, \beta_{\text{low}}, \beta_{\text{high}}) = \begin{cases} (\alpha_{\text{low}}, \alpha, \beta_{\text{low}}, \beta), & \text{if } \gamma > r_{\text{sample}}, \\ (\alpha, \alpha_{\text{high}}, \beta, \beta_{\text{high}}), & \text{otherwise.} \end{cases} \tag{6}$$

The search converges in $O(\log(1/\varepsilon))$ iterations (typically < 10), incurring negligible overhead. The resulting thresholds (α, β) partition the EU Plane (Figure 2(b)), from which Q2 and Q4 samples are retained and Q1 and Q3 discarded.

Stage 2: Token-Level Pruning for Confident Errors. While Stage 1 removes entire samples deemed unhelpful, not all retained examples are equally homogeneous inside. In particular, samples in Q2 (Valuable Misconceptions) often contain a mix of informative context and locally harmful tokens that mislead the model. To extract the useful signal, we apply token-level pruning that discards only the most detrimental tokens while preserving the surrounding context. In contrast, samples in Q4 (Calibration Data) are challenging yet reliable, and every token contributes to improving model calibration. Therefore, Q4 sequences are kept intact without any token pruning.

For a retained sample $x \in Q_2$ with target sequence y, let $T(x)$ denote the set of trainable token positions. For each token $i \in T(x)$(Figure 2(c)), we compute its token-level perplexity PPL_i, which measures how surprising the ground-truth token y_i is to the model.

To avoid pruning rare but meaningful tokens based on isolated spikes, we compute a **smoothed importance score** that incorporates local context:

$$s_i(x, y; f_\theta) = (1 - \lambda)\,\text{PPL}_i(x, y; f_\theta) + \lambda\big[\text{PPL}_{i-1}(x, y; f_\theta) + \text{PPL}_{i+1}(x, y; f_\theta)\big], \tag{7}$$

where $\lambda \in [0, 1]$ balances the contribution of neighboring tokens (default $\lambda = 0.5$). This smoothing reduces the risk of mistakenly removing single high-PPL tokens that may still be semantically critical, and we analyze the sensitivity of this choice in the ablation study (see Figure 5 in Section 4.2).

A token is deemed detrimental if both its own PPL_i and the average PPL of its immediate neighbors exceed a percentile-based threshold (*e.g.*, the median). All tokens in $T(x)$ are subsequently ranked by their smoothed scores s_i, and only the top-r_{token} fraction are retained to construct a binary mask $m(x)$. This mask selectively removes locally noisy tokens while preserving the informative context essential for learning. By contrast, no token-level pruning is applied to Q4 samples, as each token therein contributes to reliable calibration and must be preserved in full.

Table 1: Evaluation on WizardLM with a sample ratio of 12.5% and a token ratio of 50%, where ↑ and ↓ respectively denote improvements or degradations over the *Random-Random* baseline. Additional results under more sample and token pruning ratios are provided in the Table 8 and Table 9 in Appendix C.1 and results on Qwen3-8B trained on the OpenHermes are reported in Appendix C.2.

Sample Pruner	Token Pruner	**LLaMA2-7B**						**Mistral-7B**					
		ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.	ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.
Zero-Shot		53.44	38.98	5.31	12.18	43.00	30.58	66.67	46.10	18.35	10.01	43.77	36.98
12.5% *Samples*, 50% *Tokens*													
Random	Random	59.25	41.02	8.11	12.75	48.75	33.98	70.55	48.14	22.74	19.57	52.63	42.73
	PPL	$60.49^{\uparrow 1.24}$	$43.39^{\uparrow 2.37}$	$7.20^{\downarrow 0.91}$	$12.20^{\downarrow 0.55}$	$48.04^{\downarrow 0.71}$	$34.26^{\uparrow 0.28}$	$70.72^{\uparrow 0.17}$	$48.47^{\uparrow 0.33}$	$25.78^{\uparrow 3.04}$	$21.36^{\uparrow 1.79}$	$53.92^{\uparrow 1.29}$	$44.05^{\uparrow 1.32}$
	FastV	$59.96^{\uparrow 0.71}$	$42.37^{\uparrow 1.35}$	$5.76^{\downarrow 2.35}$	$11.31^{\downarrow 1.44}$	$46.42^{\downarrow 2.33}$	$33.17^{\downarrow 0.81}$	$70.72^{\uparrow 0.17}$	$46.44^{\downarrow 1.70}$	$18.80^{\downarrow 3.94}$	$19.14^{\downarrow 0.43}$	$51.56^{\downarrow 1.07}$	$41.33^{\downarrow 1.40}$
	SparseVLM	$54.32^{\downarrow 4.93}$	$37.97^{\downarrow 3.05}$	$7.35^{\downarrow 0.76}$	$12.76^{\uparrow 0.01}$	$44.65^{\downarrow 4.10}$	$31.41^{\downarrow 2.57}$	$67.02^{\downarrow 3.53}$	$44.75^{\downarrow 3.39}$	$20.24^{\downarrow 2.50}$	$10.97^{\downarrow 8.60}$	$44.61^{\downarrow 8.02}$	$37.52^{\downarrow 5.21}$
Longest	Random	$59.96^{\uparrow 0.71}$	$\mathbf{44.41}^{\uparrow 3.39}$	$7.51^{\downarrow 0.60}$	$15.34^{\uparrow 2.59}$	$48.91^{\uparrow 0.16}$	$35.22^{\uparrow 1.24}$	$74.25^{\uparrow 3.70}$	$48.81^{\uparrow 0.67}$	$28.73^{\uparrow 5.99}$	$17.66^{\downarrow 1.91}$	$55.73^{\uparrow 3.10}$	$45.04^{\uparrow 2.31}$
	PPL	$61.19^{\uparrow 1.94}$	$43.73^{\uparrow 2.71}$	$6.82^{\downarrow 1.29}$	$16.33^{\uparrow 3.58}$	$48.16^{\downarrow 0.59}$	$35.24^{\uparrow 1.26}$	$\mathbf{75.49}^{\uparrow 4.94}$	$50.17^{\uparrow 2.03}$	$27.98^{\uparrow 5.24}$	$21.49^{\uparrow 1.92}$	$56.55^{\uparrow 3.92}$	$46.33^{\uparrow 3.60}$
	FastV	$59.25^{\uparrow 0.00}$	$43.05^{\uparrow 2.03}$	$5.69^{\downarrow 2.42}$	$13.64^{\uparrow 0.89}$	$46.98^{\downarrow 1.77}$	$33.72^{\downarrow 0.26}$	$74.43^{\uparrow 3.88}$	$49.15^{\uparrow 1.01}$	$25.70^{\uparrow 2.96}$	$22.89^{\uparrow 3.32}$	$54.15^{\uparrow 1.52}$	$45.26^{\uparrow 2.53}$
	SparseVLM	$54.32^{\downarrow 4.93}$	$38.31^{\downarrow 2.71}$	$7.13^{\downarrow 0.98}$	$10.92^{\downarrow 1.83}$	$43.77^{\downarrow 4.98}$	$30.89^{\downarrow 3.09}$	$69.49^{\downarrow 1.06}$	$46.10^{\downarrow 2.04}$	$28.89^{\uparrow 6.15}$	$8.62^{\downarrow 10.95}$	$50.30^{\downarrow 2.33}$	$40.68^{\downarrow 2.05}$
InfoBatch	Random	$60.31^{\uparrow 1.06}$	$41.36^{\uparrow 0.34}$	$5.38^{\downarrow 2.73}$	$15.71^{\uparrow 2.96}$	$47.74^{\downarrow 1.01}$	$34.10^{\uparrow 0.12}$	$69.31^{\downarrow 1.24}$	$45.76^{\downarrow 2.38}$	$18.95^{\downarrow 3.79}$	$21.23^{\uparrow 1.66}$	$50.39^{\downarrow 2.24}$	$41.13^{\downarrow 1.60}$
	PPL	$59.43^{\uparrow 0.18}$	$40.34^{\downarrow 0.68}$	$5.91^{\downarrow 2.20}$	$13.18^{\uparrow 0.43}$	$48.31^{\downarrow 0.44}$	$33.44^{\downarrow 0.54}$	$70.72^{\uparrow 0.17}$	$47.12^{\downarrow 1.02}$	$18.12^{\downarrow 4.62}$	$24.10^{\uparrow 4.53}$	$51.26^{\downarrow 1.37}$	$42.26^{\downarrow 0.47}$
	FastV	$58.90^{\downarrow 0.35}$	$43.39^{\uparrow 2.37}$	$3.34^{\downarrow 4.77}$	$12.37^{\downarrow 0.38}$	$46.88^{\downarrow 1.87}$	$32.98^{\downarrow 1.00}$	$69.14^{\downarrow 1.41}$	$45.42^{\downarrow 2.72}$	$14.86^{\downarrow 7.88}$	$23.19^{\uparrow 3.62}$	$50.58^{\downarrow 2.05}$	$40.64^{\downarrow 2.09}$
	SparseVLM	$54.67^{\downarrow 4.58}$	$40.00^{\downarrow 1.02}$	$7.73^{\downarrow 0.38}$	$12.41^{\downarrow 0.34}$	$45.07^{\downarrow 3.68}$	$31.98^{\downarrow 2.00}$	$68.25^{\downarrow 2.30}$	$45.08^{\downarrow 3.06}$	$23.43^{\uparrow 0.69}$	$10.17^{\downarrow 9.40}$	$45.34^{\downarrow 7.29}$	$38.46^{\downarrow 4.27}$
Entropy	Random	$60.31^{\uparrow 1.06}$	$42.37^{\uparrow 1.35}$	$6.44^{\downarrow 1.67}$	$14.10^{\uparrow 1.35}$	$48.09^{\downarrow 0.66}$	$34.27^{\uparrow 0.29}$	$72.13^{\uparrow 1.58}$	$48.81^{\uparrow 0.67}$	$20.09^{\downarrow 2.65}$	$17.55^{\downarrow 2.02}$	$54.69^{\uparrow 2.06}$	$42.66^{\downarrow 0.07}$
	PPL	$60.49^{\uparrow 1.24}$	$43.73^{\uparrow 2.71}$	$6.90^{\downarrow 1.21}$	$14.53^{\uparrow 1.78}$	$48.76^{\uparrow 0.01}$	$34.88^{\uparrow 0.90}$	$72.84^{\uparrow 2.29}$	$47.80^{\downarrow 0.34}$	$24.18^{\uparrow 1.44}$	$22.80^{\uparrow 3.23}$	$54.69^{\uparrow 2.06}$	$44.46^{\uparrow 1.73}$
	FastV	$58.91^{\downarrow 0.34}$	$43.05^{\uparrow 2.03}$	$6.37^{\downarrow 1.74}$	$13.03^{\uparrow 0.28}$	$47.05^{\downarrow 1.70}$	$33.68^{\downarrow 0.30}$	$73.90^{\uparrow 3.35}$	$47.12^{\downarrow 1.02}$	$24.56^{\uparrow 1.82}$	$23.96^{\uparrow 4.39}$	$54.67^{\uparrow 2.04}$	$44.84^{\uparrow 2.11}$
	SparseVLM	$55.20^{\downarrow 4.05}$	$38.98^{\downarrow 2.04}$	$7.51^{\downarrow 0.60}$	$12.65^{\downarrow 0.10}$	$46.14^{\downarrow 2.61}$	$32.10^{\downarrow 1.88}$	$68.08^{\downarrow 2.47}$	$44.07^{\downarrow 4.07}$	$24.87^{\uparrow 2.13}$	$10.72^{\downarrow 8.85}$	$47.00^{\downarrow 5.63}$	$38.95^{\downarrow 3.78}$
Q-Tuning (Ours)		$\mathbf{64.20}^{\uparrow 4.95}$	$42.03^{\uparrow 1.01}$	$\mathbf{10.54}^{\uparrow 2.43}$	$\mathbf{18.79}^{\uparrow 6.04}$	$\mathbf{53.12}^{\uparrow 4.37}$	$\mathbf{37.74}^{\uparrow 3.76}$	$71.60^{\uparrow 1.05}$	$48.14^{\uparrow 0.00}$	$\mathbf{29.34}^{\uparrow 6.60}$	$\mathbf{27.75}^{\uparrow 8.18}$	$\mathbf{57.78}^{\uparrow 5.15}$	$\mathbf{46.92}^{\uparrow 4.19}$
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84

4 EXPERIMENTS

We conducted experiments on language models from multiple families and scales, including LLaMA-series (LLaMA2-7B (Touvron et al., 2023), LLaMA3-8B (Grattafiori et al., 2024)), Qwen3-series (Qwen3-8B (Yang et al., 2025), Qwen3-14B, Qwen3-32B), Mistral-7B (Jiang et al., 2023a), and SmolLM2-1.7B (Allal et al., 2025). We fine-tuned models for two settings: (i) alignment on OpenHermes (Teknium, 2023) and WizardLM (Xu et al., 2024), and (ii) reasoning on MathInstruct (Yue et al., 2023). For evaluation, we used ARC-E, ARC-C (Clark et al., 2018), GSM8K (Cobbe et al., 2021), SQuAD (Rajpurkar et al., 2016), and TriviaQA (Joshi et al., 2017) for alignment task, and GSM8K and MATH (Hendrycks et al., 2021) for reasoning task. We compared against baselines that pair sample-level pruning methods (Random, Longest, Entropy, InfoBatch (Qin et al., 2024), Alpagasus (Chen et al., 2023), Deita (Liu et al., 2023), DS2 (Pang et al., 2024), and LESS (Xia et al., 2024a)) with token-level pruning methods (Random, PPL, FastV (Chen et al., 2024), SparseVLM (Zhang et al., 2024b), Rho-1 (Lin et al., 2024), and TokenCleaning (Pang et al., 2025)). More detailed experimental settings are in Appendix B.

4.1 MAIN RESULTS

Results on alignment datasets. Table 1 summarizes results in our alignment setting under a fixed low-budget regime (12.5% samples, 50% tokens) on WizardLM. Across both LLaMA2-7B and Mistral-7B, Q-Tuning achieves the best overall averages (37.74 and 46.92), outperforming all evaluated combinations of existing sample-pruning and token-pruning methods, and it also exceeds full-data fine-tuning (35.36 and 45.84). Many baseline pairings improve one or two benchmarks but exhibit noticeable trade-offs on others (*e.g.*, degradation on GSM8K or SQuAD for several token-pruning choices), indicating that naively combining independently designed sample- and token-level pruners is insufficient under a tight compute budget. In contrast, Q-Tuning delivers consistent gains across all benchmarks, suggesting that jointly coordinating sample triage with targeted token reduction better preserves high-utility supervision in this alignment evaluation. We report additional results under more pruning ratios in Appendix C.1 (Table 8 and Table 9), showing similar trends, while Appendix C.2 presents detailed results and analysis on Qwen3-8B with larger performance gains of up to 8.23 points and improvements of up to 5.55 over the strongest baseline. Overall, these results highlight that Q-Tuning is robust to heterogeneous task mixes and remains effective even when both sample and token budgets are simultaneously constrained across diverse model families.

Results on reasoning dataset. Table 2 compares Q-Tuning against diverse baselines on math reasoning datasets under constrained sample and token budgets. Several methods, especially FastV-based token pruning, degrade sharply (e.g., driving LLaMA3-8B average accuracy below 6% in

Table 2: Evaluation of pruning strategies on GSM8K and MATH under 25% samples with 50% tokens. ↑ and ↓ respectively indicate improvements or degradations over the *Random-Random* baseline under the same sample and token keep ratio. Additional results under more sample and token pruning ratios are provided in the Table 12 in Appendix C.3.

Sample Pruner	Token Pruner	LLaMA3-8B GSM8K	LLaMA3-8B MATH	LLaMA3-8B Avg.	Mistral-7B GSM8K	Mistral-7B MATH	Mistral-7B Avg.	SmolLM2-1.7B GSM8K	SmolLM2-1.7B MATH	SmolLM2-1.7B Avg.
Zero-Shot		27.82	2.26	15.04	19.86	3.30	11.58	15.47	2.20	8.83
25% *Samples*, 50% *Tokens*										
Random	Random	23.96	2.56	13.26	23.35	1.54	12.45	14.33	2.56	8.44
	PPL	24.18$^{\uparrow 0.22}$	2.58$^{\uparrow 0.02}$	13.38$^{\uparrow 0.12}$	24.94$^{\uparrow 1.59}$	2.02$^{\uparrow 0.48}$	13.48$^{\uparrow 1.03}$	14.18$^{\downarrow 0.15}$	2.08$^{\downarrow 0.48}$	8.13$^{\downarrow 0.31}$
	FastV	12.13$^{\downarrow 11.83}$	2.32$^{\downarrow 0.24}$	7.23$^{\downarrow 6.03}$	12.36$^{\downarrow 10.99}$	1.24$^{\downarrow 0.30}$	6.80$^{\downarrow 5.65}$	9.86$^{\downarrow 4.47}$	1.92$^{\downarrow 0.64}$	5.89$^{\downarrow 2.55}$
	SparseVLM	22.97$^{\downarrow 0.99}$	4.72$^{\uparrow 2.16}$	13.85$^{\uparrow 0.59}$	19.26$^{\downarrow 4.09}$	4.58$^{\uparrow 3.04}$	11.92$^{\downarrow 0.53}$	13.19$^{\downarrow 1.14}$	3.48$^{\uparrow 0.92}$	8.34$^{\downarrow 0.10}$
Longest	Random	22.14$^{\downarrow 1.82}$	3.18$^{\uparrow 0.62}$	12.66$^{\downarrow 0.60}$	21.91$^{\downarrow 1.44}$	2.18$^{\uparrow 0.64}$	12.05$^{\downarrow 0.40}$	12.89$^{\downarrow 1.44}$	2.06$^{\downarrow 0.50}$	7.47$^{\downarrow 0.97}$
	PPL	24.94$^{\uparrow 0.98}$	2.78$^{\uparrow 0.22}$	13.86$^{\uparrow 0.60}$	22.90$^{\downarrow 0.45}$	1.86$^{\uparrow 0.32}$	12.38$^{\downarrow 0.07}$	13.19$^{\downarrow 1.14}$	1.78$^{\downarrow 0.78}$	7.49$^{\downarrow 0.95}$
	FastV	9.48$^{\downarrow 14.48}$	2.26$^{\downarrow 0.30}$	5.87$^{\downarrow 7.39}$	7.13$^{\downarrow 16.22}$	1.46$^{\downarrow 0.08}$	4.29$^{\downarrow 8.16}$	12.36$^{\downarrow 1.97}$	1.82$^{\downarrow 0.74}$	7.09$^{\downarrow 1.35}$
	SparseVLM	26.91$^{\uparrow 2.95}$	4.68$^{\uparrow 2.12}$	15.80$^{\uparrow 2.54}$	24.34$^{\uparrow 0.99}$	**4.84**$^{\uparrow 3.30}$	14.59$^{\uparrow 2.14}$	12.43$^{\downarrow 1.90}$	3.60$^{\uparrow 1.04}$	8.02$^{\downarrow 0.42}$
InfoBatch	Random	26.23$^{\uparrow 2.27}$	2.42$^{\downarrow 0.14}$	14.33$^{\uparrow 1.07}$	27.14$^{\uparrow 3.79}$	2.24$^{\uparrow 0.70}$	14.69$^{\uparrow 2.24}$	14.33$^{\uparrow 0.00}$	1.66$^{\downarrow 0.90}$	7.99$^{\downarrow 0.45}$
	PPL	26.91$^{\uparrow 2.95}$	2.66$^{\uparrow 0.10}$	14.79$^{\uparrow 1.53}$	27.90$^{\uparrow 4.55}$	2.52$^{\uparrow 0.98}$	15.21$^{\uparrow 2.76}$	14.71$^{\uparrow 0.38}$	1.90$^{\downarrow 0.66}$	8.30$^{\downarrow 0.14}$
	FastV	7.58$^{\downarrow 16.38}$	1.88$^{\downarrow 0.68}$	4.73$^{\downarrow 8.53}$	6.44$^{\downarrow 16.91}$	1.34$^{\downarrow 0.20}$	3.89$^{\downarrow 8.56}$	7.51$^{\downarrow 6.82}$	1.62$^{\downarrow 0.94}$	4.56$^{\downarrow 3.88}$
	SparseVLM	14.63$^{\downarrow 9.33}$	3.26$^{\uparrow 0.70}$	8.95$^{\downarrow 4.31}$	11.90$^{\downarrow 11.45}$	1.94$^{\uparrow 0.40}$	6.92$^{\downarrow 5.53}$	11.90$^{\downarrow 2.43}$	**4.36**$^{\uparrow 1.80}$	8.13$^{\downarrow 0.31}$
Entropy	Random	30.02$^{\uparrow 6.06}$	3.66$^{\uparrow 1.10}$	16.84$^{\uparrow 3.58}$	26.61$^{\uparrow 3.26}$	2.08$^{\uparrow 0.54}$	14.35$^{\uparrow 1.90}$	14.18$^{\downarrow 0.15}$	2.24$^{\downarrow 0.32}$	8.21$^{\downarrow 0.23}$
	PPL	32.98$^{\uparrow 9.02}$	2.92$^{\uparrow 0.36}$	17.95$^{\uparrow 4.69}$	30.17$^{\uparrow 6.82}$	1.76$^{\uparrow 0.22}$	15.97$^{\uparrow 3.52}$	16.38$^{\uparrow 2.05}$	2.40$^{\downarrow 0.16}$	9.39$^{\uparrow 0.95}$
	FastV	17.29$^{\downarrow 6.67}$	2.66$^{\uparrow 0.10}$	9.97$^{\downarrow 3.29}$	14.56$^{\downarrow 8.79}$	1.12$^{\downarrow 0.42}$	7.84$^{\downarrow 4.61}$	12.59$^{\downarrow 1.74}$	2.36$^{\downarrow 0.20}$	7.47$^{\downarrow 0.97}$
	SparseVLM	20.85$^{\downarrow 3.11}$	5.12$^{\uparrow 2.56}$	12.98$^{\downarrow 0.28}$	19.56$^{\downarrow 3.79}$	4.20$^{\uparrow 2.66}$	11.88$^{\downarrow 0.57}$	14.18$^{\downarrow 0.15}$	2.90$^{\uparrow 0.34}$	8.54$^{\uparrow 0.10}$
Q-Tuning (Ours)		**36.32**$^{\uparrow 12.36}$	**5.54**$^{\uparrow 2.98}$	**20.93**$^{\uparrow 7.67}$	**41.47**$^{\uparrow 18.12}$	4.0$^{\uparrow 2.46}$	**22.74**$^{\uparrow 10.29}$	**21.83**$^{\uparrow 7.50}$	3.90$^{\uparrow 1.34}$	**12.87**$^{\uparrow 4.43}$
Full Dataset		32.90	3.02	17.96	42.08	3.08	22.58	16.53	2.10	9.31

multiple settings). In contrast, quality-aware sample selection (Entropy, InfoBatch) consistently improves over random sampling: Entropy–Random raises LLaMA3-8B from 13.26 to 16.84 and yields similar gains on Mistral-7B. Across all model scales, Q-Tuning is the strongest and most stable, improving the Random–Random baseline by 7.67 points on LLaMA3-8B and 10.29 on Mistral-7B, reaching 20.93 and 22.74, respectively. Notably, Q-Tuning exceeds full-data GSM8K performance for both models despite using only a quarter of the samples and half of the tokens, indicating strong data efficiency and better training-resource utilization; larger models benefit more from the joint strategy. Additional ratios are reported in Appendix C.3 (Table 12). This suggests that jointly optimized pruning can improve reasoning training signal quality, rather than merely reducing compute. In particular, it preferentially retains error-revealing and correction-rich spans while trimming redundant verbiage. As a result, the model receives cleaner gradient signals and learns more effectively.

Comparison with stronger baselines. Figure 3 compares Q-Tuning with stronger, task-relevant baselines under low-budget settings. In the top figures with 12.5% samples and 50% tokens, Q-Tuning achieves the best average on both models, scoring 37.74 on LLaMA2-7B and 46.92 on Mistral-7B, which exceeds the strongest sample-selection baseline by +3.44 and +2.00, respectively. In the bottom figures with a fixed 50% token budget, Q-Tuning also outperforms token-level baselines at the same 12.5% sample ratio, improving over the best baseline from 33.82 to 37.74 on LLaMA2-7B (+3.92) and from 43.57 to 46.92 on Mistral-7B (+3.35). These results indicate that Q-Tuning remains competitive against stronger baselines under the same constrained budgets and delivers consistent gains across model families. Additional results under more pruning ratios are reported in Appendix C.4.

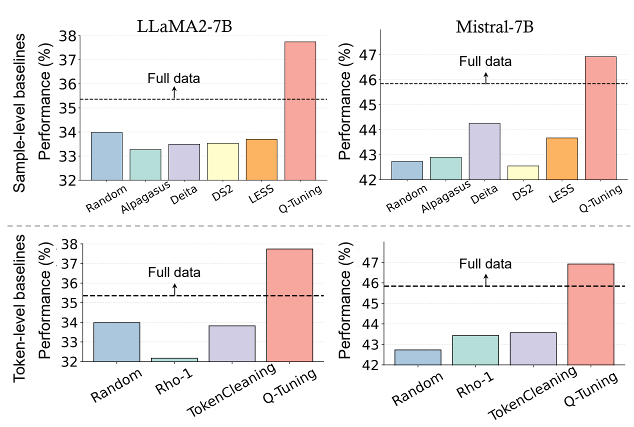

Figure 3: Comparison with stronger task-relevant baselines under matched low-budget settings. Top: Sample-pruner baselines with a sample ratio of 12.5% and a token ratio of 50%. Bottom: Token-pruner baselines with a sample ratio of 12.5% and a token ratio of 50%; the dashed line marks the full-data upper bound. Additional results under more pruning ratios are reported in Appendix C.4.

Robustness to extreme data budgets. Table 3 shows that Q-Tuning remains highly data-efficient even when retaining only 3–9% of the original training tokens on Mistral-7B. Across all four extreme settings, the average score ranges from 46.23 to 46.92, consistently outperforming both the zero-shot baseline (36.98) and the full-dataset SFT baseline (45.84), despite using at most 8.75% of the data and maintaining strong per-task accuracy. Notably, with just 6.25% samples and 50% tokens (*i.e.*, 3.125% data retention), Q-Tuning already achieves an average of 46.23, exceeding full-data fine-tuning. Increasing the token ratio to 70% at the same sample budget further improves the average to 46.62, while still operating in a regime where less than 5% of the original data is used in training epochs.

Table 3: Q-Tuning under extreme data-retention settings on the Wizard dataset using Mistral-7B.

Sample Ratio	Token Ratio	Data Retention	Mistral-7B ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.
Zero-Shot		-	66.67	46.10	18.35	10.01	43.77	36.98
6.25%	50%	3.125%	70.72	46.10	**40.41**	17.25	56.68	46.23
6.25%	70%	4.375%	71.60	**48.47**	39.73	16.13	57.16	46.62
12.50%	50%	6.250%	71.60	48.14	29.34	27.75	**57.78**	**46.92**
12.50%	70%	8.750%	**71.78**	47.12	26.08	**32.79**	56.17	46.79
Full Dataset		100%	71.25	45.76	26.68	31.81	53.67	45.84

Generalization across model scales. Figure 4 studies Q-Tuning on Qwen3-series of increasing scale (8B, 14B, and 32B). Under the same constrained budgets (12.5% samples with 50% or 70% tokens), Q-Tuning consistently matches or exceeds full-dataset SFT across benchmarks, demonstrating strong scalability despite using only a small fraction of the training data. The gains are especially evident on reasoning- and QA-oriented tasks such as GSM8K, ARC-E, and SQuAD. Moreover, for the 32B model, increasing the token budget from 50% to 70% typically yields further improvements and delivers the strongest overall averages.

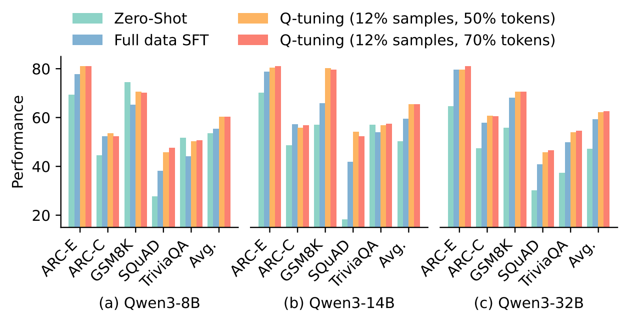

Figure 4: Comparison of Qwen3-series of varying scales (8B, 14B, 32B) across multiple benchmarks and their average. We report performance of Zero-shot, full dataset, and Q-Tuning with 12.5% samples under two token ratios (50% and 70%).

4.2 Further Analysis

Sensitivity to hyperparameters. An important question is how batch size and neighbor awareness influence the performance of Q-Tuning. We first varied the batch size across $8, 16, 32$ and evaluated the method on GSM8K, SQuAD, and TriviaQA. As shown in the upper part of Figure 5, larger batch sizes tend to yield better performance, with more noticeable gains on GSM8K and SQuAD, while TriviaQA remains relatively stable, indicating limited sensitivity to batch size. We further investigated the effect of neighbor awareness controlled by the coefficient λ. As illustrated in the lower part of Figure 5, moderate values of λ lead to improvements on GSM8K and SQuAD, whereas extreme values result in diminishing or unstable gains; in contrast, performance on TriviaQA is largely unaffected across different λ settings. These observations indicate that Q-Tuning benefits from appropriate batch sizes and a moderate degree of neighbor awareness, though the sensitivity varies across tasks. Additional benchmark results are provided in the Figure 9 in Appendix C.5.1.

Effect of token pruning strategy. We compare different token pruning strategies under a fixed $25\% \times 50\%$ budget on Mistral-7B, including Rho-1, PPL, and a reversed PPL variant, as shown in Table 4. The neighbor-aware PPL used in Q-Tuning achieves the best average performance, while alternative pruners are less consistent across tasks. This supports our design choice of locality-aware scoring that preserves salient error-relevant spans.

Table 4: Ablation study on token pruner in Q-tuning

Method	Avg.	ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA
Rho-1	45.03	69.66	46.78	24.03	29.23	55.43
PPL ($\lambda = 0$)	45.92	71.60	46.44	25.32	29.71	**56.54**
Reversed PPL ($\lambda = 0.5$)	44.86	**73.02**	47.12	16.68	32.01	55.47
PPL ($\lambda = 0.5$)	**46.79**	71.78	**47.12**	**26.08**	**32.79**	56.17

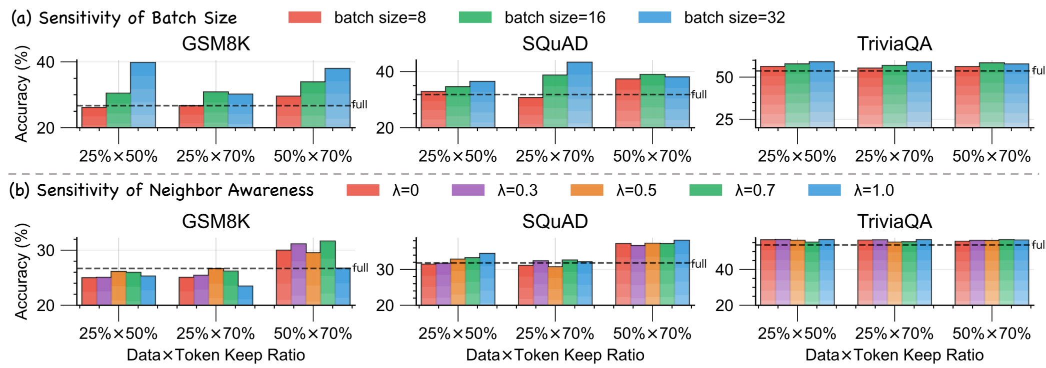

Figure 5: Effect of varying (a) batch size and (b) neighbor awareness for Mistral-7B under three keep ratio configurations. Additional benchmark results (Avg. of five benchmarks, ARC-E, ARC-C) are provided in the Figure 9 in Appendix C.5.1.

Effect of quadrant-wise pruning strategies. Table 5 reports a quadrant-wise ablation on Qwen3-8B under two budget settings. Retaining Q1 consistently yields the lowest averages, with pronounced drops on reasoning and QA tasks (*e.g.*, GSM8K and SQuAD), whereas pruning Q1 leads to more stable performance. Token-level pruning of Q4 is generally harmful, indicating that these low-PPL yet diverse tokens carry useful signal and should be preserved. In contrast, pruning Q2 tokens is crucial, as retaining Q2's high-PPL tokens degrades performance while removing them consistently improves results. Therefore, the best configurations prune Q1 and Q3 at the sample level and prune Q2 at the token level, supporting the design of Q-Tuning. Additional quadrant-wise ablations are provided in Appendix C.5.2.

Table 5: Ablation on four quadrants on OpenHermes and Qwen3-8B under different sample and token ratios, where ✓indicates that the corresponding quadrant is pruned. Additional quadrant-wise ablations are provided in Appendix C.5.2.

Sample pruning Q1	Sample pruning Q3	Token pruning Q2	Token pruning Q4	Qwen3-8B ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.
				12.5% Samples, 50% Tokens					
	✓	✓		79.01	52.20	**74.68**	36.35	48.20	58.09
✓	✓	✓	✓	79.37	52.20	71.80	35.55	47.23	57.23
✓	✓			78.84	**53.90**	68.99	44.32	49.15	59.04
✓	✓	✓		**80.95**	53.56	70.51	**45.76**	**50.16**	**60.19**
				25% Samples, 50% Tokens					
	✓	✓		75.31	50.17	68.08	29.47	45.90	53.79
✓	✓	✓	✓	76.37	50.17	65.28	31.63	45.24	53.74
✓	✓	✓		**79.89**	**52.88**	**69.37**	**43.54**	**48.33**	**58.80**

Case Study: Which Types of Tokens Are Preferentially Pruned in Q2? To better characterize the semantic effects of Q2 pruning, we perform a microscopic analysis within the Q2 quadrant. We randomly sample 300 Q2 instances and use GPT-5 (Singh et al., 2025) to annotate functional roles for 44K+ tokens, then compare category proportions between pruned and kept sets. Figure 6 shows that redundant content is overrepresented among pruned tokens: boilerplate/templates and formatting increase, and non-answer content shifts toward pruning. In contrast, supervision-critical signals are enriched in kept tokens. Gold answer spans account for a higher proportion among retained tokens, and numbers/symbols are substantially more prevalent. We provide quadrant examples in Appendix E.1 and token-level pruning cases in Appendix E.2.

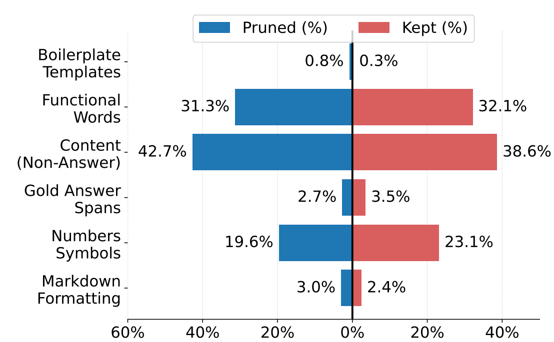

Figure 6: Semantic distribution of pruned vs. retained tokens in the Q2 quadrant, based on 300 instances annotated by GPT-5.

4.3 Efficiency Analysis: Empirical Latency

Table 6 quantifies the end-to-end efficiency of Q-Tuning on LLaMA2-7B under our standard three-epoch setup with 8× A100 GPUs. Compared to full-data SFT (90 minutes, 1.0× FLOPs, 78% peak memory), Q-Tuning with 12.5% samples and 50% tokens reduces training compute to 0.0625× and completes in 65 minutes, yielding a ≈28% wall-clock reduction while lowering peak memory to 27%.

Table 6: Efficiency and overhead breakdown on LLaMA2-7B under different pruning settings.

Pruner		Selection Time (mins)		Training Efficiency	
Sample Ratio (%)	Token Ratio (%)	Sample Selection	Token Selection	Total Time (mins)	Peak Mem (%)
12.5	50	35.88	6.96	65	27
25	50	38.64	7.49	70	35
50	50	44.16	8.56	80	50
100 (Full)	100	0.00	0.00	90	78

The screening overhead introduced by dynamic sample pruning is modest relative to the overall training budget and is effectively amortized by the computational savings gained from skipping backward passes and optimizer updates on pruned data. Even as budget increases (25% and 50% samples at 50% tokens), total time remains below full-data SFT (70–80 minutes) with substantially reduced FLOPs (0.125–0.25×) and peak memory (35–50%), indicating a favorable practical cost–benefit trade-off.

5 Discussion

Can Q-Tuning outperform independent sample or token pruning? To isolate the effect of each pruning dimension, we conduct controlled ablations that apply *only* dynamic sample pruning (retaining all tokens) or *only* dynamic token pruning (retaining all samples). As shown in Figure 7, Q-Tuning consistently outperforms all baseline methods across keep ratios on both LLaMA2-7B and Mistral-7B, indicating that its advantage does not rely on a particular pruning axis but stems from more effective data utility modeling.

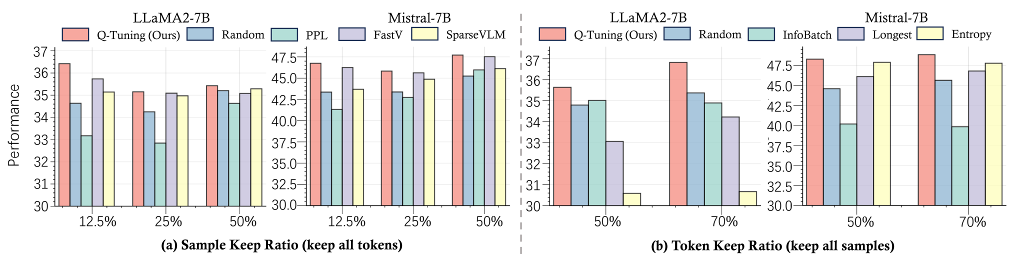

Figure 7: Comparison of independent (a) dynamic sample pruning and (b) dynamic token pruning across different keep ratios for LLaMA2-7B and Mistral-7B.

How does the sample distribution evolve during training? To characterize training dynamics, we track the average perplexity and token entropy on 100 randomly sampled training instances throughout fine-tuning. As shown in Figure 8, Q-Tuning reduces both metrics more rapidly than alternative sample-pruning strategies, suggesting faster stabilization of the training signal and correlating with improved downstream performance.

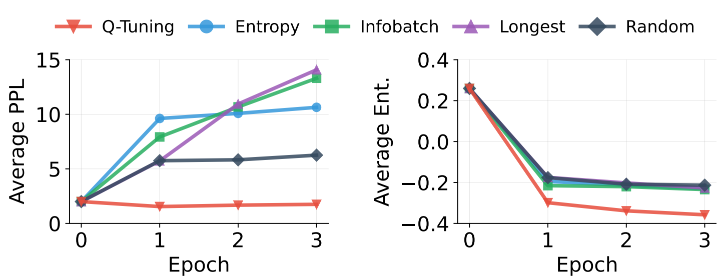

Figure 8: Training dynamics of different sample pruners. Compared to baseline strategies, Q-Tuning consistently reduces both metrics at a faster rate, indicating more efficient learning dynamics.

6 Conclusion

This work turns risky dynamic data pruning for LLM fine-tuning into a reliable, high-utility strategy. By analyzing sample modes via the Error–Uncertainty Plane, we expose data's heterogeneous value and the need for nuance. Our Quadrant-based Tuning (Q-Tuning) uses a two-stage framework to coordinate sample- and token-level pruning, preserving valuable signals while removing noise, boosting efficiency without sacrificing overall stability, and often further improving performance across diverse tasks and models.

References

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

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. Smollm2: When smol goes big–data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, and Agha Ali Raza. Language model-driven data pruning enables efficient active learning. *arXiv preprint arXiv:2410.04275*, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35, 2024.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*, 2023.

Tzu-Chun Chien, Chieh-Kai Lin, Shiang-Feng Tsai, Ruei-Chi Lai, Hung-Jen Chen, and Min Sun. Grounding-aware token pruning: Recovering from drastic performance drops in visual grounding caused by pruning. *arXiv preprint arXiv:2506.21873*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

C Coleman, C Yeh, S Mussmann, B Mirzasoleiman, P Bailis, P Liang, J Leskovec, and M Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations (ICLR)*, 2020.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models, 2023.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving llm alignment via preference data selection. *arXiv preprint arXiv:2502.14560*, 2025.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 177–198, 2024.

Qichen Fu, Minsik Cho, Thomas Merth, Sachin Mehta, Mohammad Rastegari, and Mahyar Najibi. Lazyllm: Dynamic token pruning for efficient long context llm inference. *arXiv preprint arXiv:2407.14057*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yangyang Guo and Mohan Kankanhalli. Scan: Bootstrapping contrastive pre-training for data efficiency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3662–3672, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Xijie Huang, Li Lyna Zhang, Kwang-Ting Cheng, Fan Yang, and Mao Yang. Fewer is more: Boosting llm reasoning with reinforced context pruning. *arXiv preprint arXiv:2312.08901*, 2023.

Hamish Ivison, Noah A Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Data-efficient finetuning using cross-task nearest neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9036–9061, 2023.

Hamish Ivison, Muru Zhang, Faeze Brahman, Pang Wei Koh, and Pradeep Dasigi. Large-scale data selection for instruction tuning. *arXiv preprint arXiv:2503.01807*, 2025.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023a.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023b.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Christopher Keith, Michael Robinson, Francis Duncan, Allan Worthington, Joseph Wilson, and Sofia Harris. Optimizing large language models: A novel approach through dynamic token pruning. 2024.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6342–6353, 2023.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, Weizhu Chen, et al. Not all tokens are what you need for pretraining. *Advances in Neural Information Processing Systems*, 37:29029–29063, 2024.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*, 2023.

Lingkun Long, Rubing Yang, Yushi Huang, Desheng Hui, Ao Zhou, and Jianlei Yang. Sliminfer: Accelerating long-context llm inference via dynamic token pruning. *arXiv preprint arXiv:2508.06447*, 2025.

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

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

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *ACL (Findings)*, 2024.

Jinlong Pang, Jiaheng Wei, Ankit Parag Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. Improving data efficiency via curating llm-driven rating systems. *arXiv preprint arXiv:2410.10877*, 2024.

Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. Token cleaning: Fine-grained data selection for llm supervised fine-tuning. *arXiv preprint arXiv:2502.01968*, 2025.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, Yang You, et al. Infobatch: Lossless training speed up by unbiased dynamic data pruning. In *The Twelfth International Conference on Learning Representations*, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Ravi S Raju, Kyle Daruwalla, and Mikko Lipasti. Accelerating deep learning with dynamic data pruning. *arXiv preprint arXiv:2111.12621*, 2021.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. Data whisperer: Efficient data selection for task-specific llm fine-tuning via few-shot in-context learning. *Annual Meeting of the Association for Computational Linguistics*, 2025a.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*, 2025b.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025.

Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024a.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. *arXiv preprint arXiv:2410.09335*, 2024b.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024.

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.

Yu Yang, Siddhartha Mishra, Jeffrey Chiang, and Baharan Mirzasoleiman. Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *Advances in Neural Information Processing Systems*, 37:83465–83496, 2024.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*, 2025.

Yang Yu, Kai Han, Hang Zhou, Yehui Tang, Kaiqi Huang, Yunhe Wang, and Dacheng Tao. Llm data selection and utilization via dynamic bi-level optimization. *arXiv preprint arXiv:2507.16178*, 2025.

Zichun Yu, Spandan Das, and Chenyan Xiong. Mates: Model-aware data selection for efficient pretraining with data influence models. *Advances in Neural Information Processing Systems*, 37: 108735–108759, 2024.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*, 2023.

Guibin Zhang, Haonan Dong, Zhixun Li, Dingshuo Chen, Kai Wang, Tianlong Chen, Yuxuan Liang, Dawei Cheng, Kun Wang, et al. Gder: Safeguarding efficiency, balancing, and robustness via prototypical graph pruning. *Advances in Neural Information Processing Systems*, 37:50285–50312, 2024a.

Jia Zhang, Chen-Xi Zhang, Yao Liu, Yi-Xuan Jin, Xiao-Wen Yang, Bo Zheng, Yi Liu, and Lan-Zhe Guo. D3: Diversity, difficulty, and dependability-aware data selection for sample-efficient llm instruction tuning. *arXiv preprint arXiv:2503.11441*, 2025.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning*, 2024b.

Yaowei Zheng, Richong Zhang, Junhao Zhang, YeYanhan YeYanhan, and Zheyan Luo. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, 2024.

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

Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. Dataset quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17205–17216, 2023b.

Fan Zhou, Zengzhi Wang, Qian Liu, Junlong Li, and Pengfei Liu. Programming every example: Lifting pre-training data quality like experts at scale. *arXiv preprint arXiv:2409.17115*, 2024.

A Related Work

A.1 Sample Pruning

A growing body of work explores sample pruning as a means to reduce training cost by selecting smaller yet higher-quality subsets of data without compromising performance (Ivison et al., 2025; Yang et al., 2024; Xia et al., 2024a; Zhou et al., 2024). For example, LIMA (Zhou et al., 2023a) shows that aligning LLMs requires only a small collection of high-quality instruction–response pairs rather than large-scale corpora. Broadly, existing approaches fall into static and dynamic pruning. In the static setting, data subsets are determined in advance using fixed criteria: SVP (Coleman et al., 2020) leverages proxy models to estimate sample importance, D3 (Zhang et al., 2025) combines diversity, difficulty, and dependability into weighted coresets, and Less is More (Deng et al., 2025) shows that carefully curated preference subsets can outperform full datasets by filtering noisy or redundant examples. Other static approaches target in-context examples: LIMO (Ye et al., 2025) removes redundant demonstrations while distilling essential reasoning patterns, and methods such as DEFT (Ivison et al., 2023), Alpagasus (Chen et al., 2023), and Data Whisperer (Wang et al., 2025a) automatically select or reweight demonstrations based on influence estimation, contribution to performance, or few-shot evaluation.

In contrast, dynamic pruning adapts sample usage throughout training. Early work by (Raju et al., 2021) proposed two methods, ϵ-greedy and UCB, which retain uncertain examples while discarding easier ones. Subsequent approaches explore alternative criteria for adaptive pruning: RHO-LOSS (Mindermann et al., 2022) assigns importance using high-quality reference models, InfoBatch (Qin et al., 2024) removes low-loss examples on the fly, and more recent methods such as MATES (Yu et al., 2024) and DWM (Yu et al., 2025) reweight or select samples according to the evolving model state.Beyond single-example pruning, several methods consider structured or representation-based criteria: ActivePrune (Azeemi et al., 2024) selects demonstrations per input via similarity metrics, SCAN (Guo & Kankanhalli, 2025) iteratively removes ill-matched and redundant data during contrastive pretraining, GDeR (Zhang et al., 2024a) models data as a prototypical graph to prune noisy or redundant samples, and DQ (Zhou et al., 2023b) clusters data in representation space and replaces each cluster with representative subsets.

A.2 Token Pruning

Parallel to sample pruning, recent research (Wang et al., 2025b; Pan et al., 2024; Keith et al., 2024) has increasingly focused on token-level pruning to retain only the most informative portions of input sequences. At pretraining time, methods such as Rho-1 (Lin et al., 2024) estimate token importance using a reference model and remove low-utility tokens. During fine-tuning and inference, pruning is typically dynamic: Token Cleaning (Pang et al., 2025) treats harmful tokens as noisy labels and prunes them based on their influence on parameter updates, while LazyLLM (Fu et al., 2024) and SlimInfer (Long et al., 2025) accelerate long-context inference by dropping tokens or hidden-state blocks using attention signals and cache mechanisms. TokenSkip (Xia et al., 2025) further reduces redundant reasoning by skipping semantically less important tokens. A related line performs input compression before inference, as in LLMLingua (Jiang et al., 2023b) and Selective Context (Li et al., 2023), which apply coarse-to-fine filtering with budget control to prune low-information tokens, phrases, or sentences. Beyond text-only models, multimodal pruning exploits cross-modal attention patterns: FastV (Chen et al., 2024) and SparseVLM (Zhang et al., 2024b) drop redundant visual tokens guided by attention, while LMTL (Huang et al., 2023) prunes unnecessary visual features to adapt computation. Recent methods incorporate task objectives, GAP (Chien et al., 2025) maintains spatial grounding by correcting positional misalignment during pruning, and DART (Wen et al., 2025) reduces computation via confidence-based early stopping.

Despite their progress, prior work typically treats sample pruning and token pruning as independent lines of research. Sample pruning methods focus on reducing the number of training examples but leave token-level redundancy unaddressed, while token pruning techniques emphasize sequence-level efficiency without considering redundancy across training samples. Such separation limits the potential for jointly optimizing efficiency and effectiveness. In this work, we bridge these directions by proposing a unified framework that integrates sample-level and token-level pruning, enabling models to simultaneously filter uninformative data and compress redundant inputs.

B EXPERIMENTAL SETTINGS

B.1 MODELS AND DATASETS.

We conducted experiments on language models of different scales, including *Mistral-7B* (Jiang et al., 2023a), *LLaMA2-7B* (Touvron et al., 2023), *LLaMA3-8B* (Grattafiori et al., 2024), *SmolLM2-1.7B* (Allal et al., 2025), *Qwen3-8B*, *Qwen-14B* and *Qwen-32B* (Yang et al., 2025). For Qwen3-series, we fine-tune on the *OpenHermes* (Teknium, 2023), which contains diverse instruction–response pairs for general-purpose alignment. For other models, we use two complementary datasets: *WizardLM* (Xu et al., 2024) for alignment, which provides high-quality instruction–response pairs for supervised fine-tuning (SFT), and *MathInstruct* (Yue et al., 2023) for reasoning, which contains mathematically focused instructions designed to assess arithmetic and symbolic problem-solving.

B.2 EVALUATION.

For alignment task, we used five standard benchmarks: ARC-E, ARC-C (Clark et al., 2018), GSM8K (Cobbe et al., 2021), SQuAD (Rajpurkar et al., 2016), and TriviaQA (Joshi et al., 2017), covering knowledge-intensive question answering, commonsense reasoning, and reading comprehension. For reasoning task, we evaluated models fine-tuned on MathInstruct using GSM8K and MATH (Hendrycks et al., 2021), two widely used benchmarks for mathematical problem solving.

B.3 BASELINES.

We constructed baselines by pairing sample-level and token-level pruning strategies. For sample-level pruning, we considered eight methods: (i) Random, which dropped samples uniformly at random; (ii) Longest, which removed the longest sequences; (iii) Entropy, which retained high-entropy samples; (iv) InfoBatch (Qin et al., 2024), an information-theoretic data selection method; (v) Alpagasus (Chen et al., 2023), which leveraged a LLM as an automatic quality rater to filter low-quality instruction–response pairs; (vi) Deita (Liu et al., 2023), which selected instruction–response pairs by modeling instruction complexity, quality, and diversity; (vii) DS2 (Pang et al., 2024), a diversity-aware data selection method that corrected LLM-generated quality scores via a transition matrix; and (viii) Less (Xia et al., 2024a), which selected instruction-tuning data by measuring gradient-based influence on few-shot target examples. For token-level pruning, we adopted six methods applied to the retained data: (i) Random, which masked tokens uniformly at random; (ii) PPL, which removed high-perplexity tokens; (iii) FastV (Chen et al., 2024), which pruned tokens receiving the least final-layer attention; (iv) SparseVLM (Zhang et al., 2024b), which combined attention and hidden-state features to score token importance; (v) Rho-1 (Lin et al., 2024), which scored tokens by their excess loss relative to a reference model; and (vi) TokenCleaning (Pang et al., 2025), which removed tokens deemed noisy or uninformative based on token-level statistics.

B.4 IMPLEMENTATION DETAILS

All experiments were conducted using the *LLaMA-Factory* framework (Zheng et al., 2024), which provided utilities for training and evaluation across diverse large language models. For benchmark evaluation, we used the *OpenCompass* framework (Contributors, 2023), which offered a standardized interface to a wide range of tasks. Unless otherwise specified, models were trained for 3 epochs with a learning rate of 1×10^{-4} and a batch size of 8 per device. In all pruning experiments, we applied pruning exclusively to the answers. All experiments were performed on NVIDIA A100 GPUs. For the *Random* baseline, we repeated each experiment five times and reported the averaged results.

B.5 HYPERPARAMETERS

We provide the full hyperparameter settings used in our experiments in Table 7. Unless otherwise noted, all other parameters follow the default settings of the *LLaMA-Factory* framework.

Table 7: Complete hyperparameter configurations employed in our experiments.

Hyperparameter	Value	Hyperparameter	Value
λ	0.5	Batch size	8
Epochs	3	LR	1×10^{-4}
LoRA rank	8	Cutoff length	2048
Grad. accum.	4	Scheduler	cosine

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 ADDITIONAL RESULTS ON ALIGNMENT DATASETS

The detailed results of additional experiments on alignment dataset are presented in Table 8 and Table 9, providing a comprehensive comparison across different pruning strategies.

Table 8: Evaluation on Wizard under different sample ratios (12.5%, 25%) and token ratios (50%, 70%), where ↑ and ↓ respectively denote improvements or degradations over the *Random-Random* baseline.

Sample Pruner	Token Pruner	**LLaMA2-7B**						**Mistral-7B**					
		ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.	ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.
Zero-Shot		53.44	38.98	5.31	12.18	43.00	30.58	66.67	46.10	18.35	10.01	43.77	36.98
12.5% *Samples*, 70% *Tokens*													
Random	Random	59.43	41.02	6.97	13.64	47.97	33.81	71.08	47.46	24.34	21.64	53.15	43.53
	PPL	60.14 ↑0.71	43.39 ↑2.37	6.22 ↓0.75	12.18 ↓1.46	48.18 ↑0.21	34.02 ↑0.21	70.72 ↓0.36	47.80 ↑0.34	25.09 ↑0.75	21.28 ↓0.36	53.83 ↑0.68	43.74 ↑0.21
	FastV	58.20 ↓1.23	41.02 ↑0.00	6.29 ↓0.68	13.42 ↓0.22	45.32 ↓2.65	32.85 ↓0.96	70.72 ↓0.36	46.44 ↓1.02	19.56 ↓4.78	21.38 ↓0.26	53.34 ↑0.19	42.29 ↓1.24
	SparseVLM	54.67 ↓4.76	37.97 ↓3.05	8.04 ↑1.07	13.06 ↓0.58	44.87 ↓3.10	31.72 ↓2.09	67.72 ↓3.36	44.75 ↓2.71	23.65 ↓0.69	11.76 ↓9.88	44.90 ↓8.25	38.58 ↓4.95
Longest	Random	59.44 ↑0.01	43.39 ↑2.37	7.35 ↑0.38	15.59 ↑1.95	50.02 ↑2.05	35.15 ↑1.34	73.37 ↑2.29	48.81 ↑1.35	27.82 ↑3.48	21.31 ↓0.33	55.77 ↑2.62	45.42 ↑1.89
	PPL	60.85 ↑1.42	43.39 ↑2.37	7.73 ↑0.76	16.21 ↑2.57	48.57 ↑0.60	35.35 ↑1.54	**74.96 ↑3.88**	49.83 ↑2.37	28.73 ↑4.39	21.62 ↓0.02	56.59 ↑3.44	46.35 ↑2.82
	FastV	59.44 ↑0.01	42.71 ↑1.69	6.29 ↓0.68	14.53 ↑0.89	47.46 ↓0.51	34.09 ↑0.28	74.07 ↑2.99	49.83 ↑2.37	24.18 ↓0.16	25.74 ↑4.10	55.86 ↑2.71	45.94 ↑2.41
	SparseVLM	54.85 ↓4.58	37.97 ↓3.05	7.05 ↑0.08	11.20 ↓2.44	44.16 ↓3.81	31.04 ↓2.77	69.14 ↓1.94	44.75 ↓2.71	31.01 ↑6.67	6.25 ↓15.39	52.94 ↓0.21	40.82 ↓2.71
InfoBatch	Random	59.26 ↓0.17	42.37 ↑1.35	6.22 ↓0.75	16.10 ↑2.46	47.72 ↓0.25	34.33 ↑0.52	70.19 ↓0.89	47.80 ↑0.34	20.77 ↓3.57	19.03 ↓2.61	52.13 ↓1.02	41.98 ↓1.55
	PPL	60.49 ↑1.06	39.32 ↓1.70	5.76 ↓1.21	14.47 ↑0.83	48.06 ↑0.09	33.62 ↓0.19	70.72 ↓0.36	46.44 ↓1.02	19.03 ↓5.31	23.20 ↑1.56	51.75 ↓1.40	42.23 ↓1.30
	FastV	58.55 ↓0.88	43.39 ↑2.37	5.53 ↓1.44	13.13 ↓0.51	47.64 ↓0.33	33.65 ↓0.16	69.49 ↓1.59	43.39 ↓4.07	16.68 ↓7.66	25.27 ↑3.63	51.47 ↓1.68	41.26 ↓2.27
	SparseVLM	56.61 ↓2.82	38.31 ↓2.71	5.76 ↓1.21	12.47 ↓1.17	44.49 ↓3.48	31.53 ↓2.28	68.25 ↓2.83	44.41 ↓3.05	23.73 ↓0.61	9.07 ↓12.57	45.73 ↓7.42	38.24 ↓5.29
Entropy	Random	61.02 ↑1.59	43.05 ↑2.03	7.66 ↑0.69	14.11 ↑0.47	48.44 ↑0.47	34.86 ↑1.05	73.37 ↑2.29	49.83 ↑2.37	23.05 ↓1.29	16.52 ↓5.12	55.18 ↑2.03	43.59 ↑0.06
	PPL	61.02 ↑1.59	43.39 ↑2.37	6.97 ↑0.00	14.94 ↑1.30	48.94 ↑0.97	35.05 ↑1.24	73.02 ↑1.94	47.46 ↑0.00	24.03 ↑0.68	22.85 ↑1.21	54.89 ↑1.74	44.45 ↑0.92
	FastV	58.73 ↓0.70	**43.39 ↑2.37**	6.14 ↓0.83	14.23 ↑0.59	47.03 ↓0.94	33.90 ↑0.09	74.07 ↑2.99	**50.85 ↑3.39**	24.94 ↑0.60	23.79 ↑2.15	55.94 ↑2.79	45.92 ↑2.39
	SparseVLM	54.85 ↓4.58	37.29 ↓3.73	6.52 ↓0.45	12.73 ↓0.91	46.24 ↓1.73	31.53 ↓2.28	68.08 ↓3.00	44.41 ↓3.05	26.38 ↑2.04	11.06 ↓10.58	46.68 ↓6.47	39.32 ↓4.21
Q-Tuning (Ours)		**64.37 ↑4.94**	42.37 ↑1.35	**10.84 ↑3.87**	**17.63 ↑3.99**	**52.17 ↑4.20**	**37.48 ↑3.67**	71.78 ↑0.70	48.14 ↑0.68	**30.33 ↑6.00**	**28.59 ↑6.95**	**57.93 ↑4.78**	**47.35 ↑3.82**
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84
25% *Samples*, 50% *Tokens*													
Random	Random	60.32	41.69	5.76	13.43	48.41	33.92	70.19	46.10	20.62	24.07	53.74	42.95
	PPL	60.32 ↑0.00	42.03 ↑0.34	7.51 ↑1.75	15.94 ↑2.51	48.58 ↑0.17	34.87 ↑0.95	69.66 ↓0.53	47.46 ↑1.36	19.86 ↓0.76	19.51 ↓4.56	53.74 ↑0.00	42.05 ↓0.90
	FastV	59.08 ↓1.24	41.69 ↑0.00	3.56 ↓2.20	12.78 ↓0.65	45.60 ↓2.81	32.54 ↓1.38	71.78 ↑1.59	47.12 ↑1.02	15.77 ↓4.85	26.97 ↑2.90	50.84 ↓2.90	42.50 ↓0.45
	SparseVLM	54.50 ↓5.82	38.64 ↓3.05	6.44 ↑0.68	12.04 ↓1.39	44.79 ↓3.62	31.28 ↓2.64	67.55 ↓2.64	46.44 ↑0.34	24.41 ↑3.79	11.80 ↓12.27	48.14 ↓5.60	39.67 ↓3.28
Longest	Random	61.20 ↑0.88	42.03 ↑0.34	7.88 ↑2.12	15.40 ↑1.97	48.29 ↓0.12	34.96 ↑1.04	**73.54 ↑3.35**	48.14 ↑2.04	23.73 ↑3.11	26.34 ↑2.27	54.06 ↑0.32	45.16 ↑2.21
	PPL	60.85 ↑0.53	43.39 ↑1.70	7.20 ↑1.44	13.88 ↑0.45	48.48 ↑0.07	34.76 ↑0.84	72.31 ↑2.12	48.14 ↑2.04	24.34 ↑3.72	23.84 ↑-0.23	55.22 ↑1.48	44.77 ↑1.82
	FastV	59.08 ↓1.24	42.71 ↑1.02	5.16 ↓0.60	14.00 ↑0.57	47.47 ↓0.94	33.68 ↓0.24	72.66 ↑2.47	46.10 ↑0.00	18.88 ↓1.74	31.52 ↑7.45	52.13 ↓1.61	44.26 ↑1.31
	SparseVLM	56.61 ↓3.71	37.29 ↓4.40	7.58 ↑1.82	12.09 ↓1.34	44.76 ↓3.65	31.66 ↓2.26	66.84 ↓3.35	44.41 ↓1.69	29.42 ↑8.80	11.22 ↓12.85	48.47 ↓5.27	40.07 ↓2.88
InfoBatch	Random	58.73 ↓1.59	40.68 ↓1.01	6.67 ↑0.91	9.95 ↓3.48	48.98 ↑0.57	33.00 ↓0.92	70.55 ↑0.36	46.44 ↑0.34	21.53 ↑0.91	23.93 ↓0.14	52.14 ↓1.60	42.92 ↓0.03
	PPL	59.96 ↑0.47	42.71 ↑1.02	6.52 ↑0.76	14.58 ↑1.15	48.57 ↑0.16	34.47 ↑0.55	71.08 ↑0.89	47.80 ↑1.70	20.62 ↑0.00	24.88 ↑0.81	51.61 ↓2.13	43.20 ↑0.25
	FastV	59.08 ↓1.24	42.37 ↑0.68	3.03 ↓2.73	11.13 ↓2.30	47.50 ↓0.91	32.63 ↓1.29	69.31 ↓0.88	44.41 ↓1.69	14.48 ↓6.14	23.63 ↓0.44	49.16 ↓4.58	40.20 ↓2.75
	SparseVLM	55.73 ↓4.59	39.66 ↓2.03	5.31 ↓0.45	11.66 ↓1.77	43.25 ↓5.16	31.12 ↓2.80	67.20 ↓3.00	45.76 ↓0.34	23.58 ↑2.96	9.63 ↓14.44	46.09 ↓7.65	38.45 ↓4.50
Entropy	Random	60.49 ↑0.17	41.69 ↑0.00	7.51 ↑1.75	15.94 ↑2.51	48.76 ↑0.35	34.88 ↑0.96	70.19 ↑0.00	47.12 ↑1.02	22.44 ↑1.82	27.35 ↑3.28	54.78 ↑1.04	44.38 ↑1.43
	PPL	60.49 ↑0.17	41.02 ↓0.67	6.60 ↑0.84	14.92 ↑1.49	49.33 ↑0.92	34.47 ↑0.55	71.43 ↑1.24	**48.14 ↑2.04**	21.30 ↑0.68	25.52 ↑1.45	55.60 ↑1.86	44.40 ↑1.45
	FastV	58.91 ↓1.41	41.69 ↑0.00	6.07 ↑0.31	12.79 ↓0.64	46.11 ↓2.30	33.11 ↓0.81	72.31 ↑2.12	47.46 ↑1.36	18.04 ↓2.58	25.96 ↑1.89	52.46 ↓1.28	43.25 ↑0.30
	SparseVLM	54.67 ↓5.65	38.64 ↓3.05	6.90 ↑1.14	11.64 ↓1.79	45.03 ↓3.38	31.38 ↓2.54	68.25 ↓1.94	44.07 ↓2.03	**26.69 ↑6.07**	9.69 ↓14.38	47.24 ↓6.50	39.19 ↓3.76
Q-Tuning (Ours)		**63.14 ↑2.82**	42.03 ↑0.34	**8.87 ↑3.11**	**16.76 ↑3.33**	**51.52 ↑3.11**	**36.47 ↑2.55**	71.78 ↑1.59	47.12 ↑1.02	26.08 ↑5.46	**32.79 ↑8.72**	**56.17 ↑2.43**	**46.79 ↑3.84**
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84

Table 9: Evaluation on Wizard under different sample ratios (25%, 50%) and token ratios (50%, 70%), where ↑ and ↓ respectively denote improvements or degradations over the *Random-Random* baseline.

Sample Pruner	Token Pruner	LLaMA2-7B ARC-E	LLaMA2-7B ARC-C	LLaMA2-7B GSM8K	LLaMA2-7B SQuAD	LLaMA2-7B TriviaQA	LLaMA2-7B Avg.	Mistral-7B ARC-E	Mistral-7B ARC-C	Mistral-7B GSM8K	Mistral-7B SQuAD	Mistral-7B TriviaQA	Mistral-7B Avg.
Zero-Shot		53.44	38.98	5.31	12.18	43.00	30.58	66.67	46.10	18.35	10.01	43.77	36.98
25% *Samples,* 70% *Tokens*													
Random	Random	60.67	41.69	6.22	13.91	48.77	34.25	70.02	46.78	19.71	23.50	52.93	42.59
Random	PPL	60.32 ↓0.35	42.03 ↑0.34	5.91 ↓0.31	15.88 ↑1.97	48.56 ↓0.21	34.54 ↑0.29	69.49 ↓0.53	48.14 ↑1.36	21.08 ↑1.37	20.63 ↓2.87	53.73 ↑0.80	42.61 ↑0.02
Random	FastV	60.32 ↓0.35	42.71 ↑1.02	6.07 ↓0.15	14.92 ↑1.01	47.20 ↓1.57	34.24 ↓0.01	71.25 ↑1.23	47.80 ↑1.02	18.95 ↓0.76	24.29 ↑0.79	52.07 ↓0.86	42.87 ↑0.28
Random	SparseVLM	55.73 ↓4.94	37.97 ↓3.72	7.20 ↑0.98	11.83 ↓2.08	44.96 ↓3.81	31.54 ↓2.71	67.72 ↓2.30	44.75 ↓2.03	25.32 ↑5.61	12.69 ↓10.81	47.46 ↓5.47	39.59 ↓3.00
Longest	Random	61.02 ↑0.35	41.69 ↑0.00	7.20 ↑0.98	15.61 ↑1.70	49.53 ↑0.76	35.01 ↑0.76	**74.43** ↑4.41	47.80 ↑1.02	24.56 ↑4.85	26.79 ↑3.29	55.24 ↑2.31	45.76 ↑3.17
Longest	PPL	61.20 ↑0.53	43.05 ↑1.36	7.28 ↑1.06	14.06 ↑0.15	48.87 ↑0.10	34.89 ↑0.64	72.49 ↑2.47	47.46 ↑0.68	25.55 ↑5.84	23.92 ↑0.42	55.26 ↑2.33	44.93 ↑2.34
Longest	FastV	59.26 ↓1.41	43.05 ↑1.36	5.53 ↓0.69	13.98 ↑0.07	48.03 ↓0.74	33.97 ↓0.28	72.84 ↑2.82	47.80 ↑1.02	19.79 ↑0.08	31.54 ↑8.04	53.24 ↑0.31	45.04 ↑2.45
Longest	SparseVLM	55.73 ↓4.94	36.27 ↓5.42	8.04 ↑1.82	11.39 ↓2.52	44.65 ↓4.12	31.22 ↓3.03	68.08 ↓1.94	43.73 ↓3.05	25.78 ↑6.07	13.47 ↓10.03	48.00 ↓4.93	39.81 ↓2.78
InfoBatch	Random	59.79 ↓0.88	42.71 ↑1.02	5.84 ↓0.38	10.42 ↓3.49	48.99 ↑0.22	33.55 ↓0.70	70.37 ↑0.35	46.44 ↓0.34	21.00 ↑1.29	24.09 ↑0.59	53.75 ↑0.82	43.13 ↑0.54
InfoBatch	PPL	60.32 ↑0.35	**43.39** ↑1.70	6.67 ↑0.45	13.88 ↑0.55	49.03 ↑0.26	34.66 ↑0.41	71.08 ↑1.06	46.10 ↓0.68	22.44 ↑2.73	22.07 ↓1.43	52.17 ↓0.76	42.77 ↑0.18
InfoBatch	FastV	59.61 ↓1.06	42.71 ↑1.02	3.34 ↓2.88	12.18 ↓1.73	47.98 ↓0.79	33.16 ↓1.09	69.66 ↓0.36	45.76 ↓1.02	15.39 ↓5.21	24.75 ↑1.25	49.51 ↓3.42	41.01 ↓1.58
InfoBatch	SparseVLM	55.56 ↓5.11	38.31 ↓3.38	5.38 ↓0.84	10.81 ↓3.10	43.08 ↓5.69	30.63 ↓3.62	65.78 ↓4.24	45.08 ↓1.70	20.70 ↑0.99	10.97 ↓12.53	48.66 ↓4.27	38.24 ↓4.35
Entropy	Random	60.67 ↑0.00	41.02 ↓0.67	5.99 ↓0.23	14.92 ↑1.01	49.34 ↑0.57	34.39 ↑0.14	70.37 ↑0.35	**48.47** ↑1.69	22.21 ↑2.50	27.57 ↑4.07	**56.05** ↑3.12	44.93 ↑2.34
Entropy	PPL	61.55 ↑0.88	41.02 ↓0.67	5.38 ↓0.84	14.90 ↑0.99	49.72 ↑0.95	34.51 ↑0.26	70.90 ↑0.88	48.14 ↑1.36	21.15 ↑1.44	26.22 ↑2.72	55.83 ↑2.90	44.45 ↑1.86
Entropy	FastV	59.79 ↓0.88	42.03 ↑0.34	6.44 ↑0.22	13.21 ↓0.70	47.58 ↓1.19	33.81 ↓0.44	71.43 ↑1.41	47.12 ↑0.34	20.24 ↑0.53	26.06 ↑2.56	54.40 ↑1.47	43.85 ↑1.26
Entropy	SparseVLM	54.50 ↓6.17	37.97 ↓3.72	8.87 ↑2.65	12.25 ↓1.66	45.05 ↓3.72	31.73 ↓2.52	67.20 ↓2.82	45.76 ↓1.02	25.17 ↑5.46	9.64 ↓13.86	48.38 ↓4.55	39.23 ↓3.36
Q-Tuning (Ours)		**62.43** ↑1.76	42.37 ↑0.68	**9.25** ↑3.03	**19.53** ↑5.62	**50.78** ↑2.01	**36.87** ↑2.62	71.60 ↑1.58	47.12 ↑0.34	**26.61** ↑6.90	**30.64** ↑7.14	55.13 ↑2.20	**46.22** ↑3.63
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84
50% *Samples,* 50% *Tokens*													
Random	Random	62.08	41.36	6.75	12.14	48.86	34.24	71.25	46.44	21.53	24.91	54.16	43.66
Random	PPL	61.02 ↓1.06	**43.05** ↑1.69	6.82 ↑0.07	15.08 ↑2.94	49.10 ↑0.24	35.01 ↑0.77	72.49 ↑1.24	46.78 ↑0.34	22.21 ↑0.68	33.28 ↑8.37	53.83 ↓0.33	45.72 ↑2.06
Random	FastV	60.32 ↓1.76	42.03 ↑0.67	4.40 ↓2.35	11.80 ↓0.34	48.35 ↓0.51	33.38 ↓0.86	72.13 ↑0.88	45.42 ↓1.02	16.22 ↓5.31	27.69 ↑2.78	50.58 ↓3.58	42.41 ↓1.25
Random	SparseVLM	55.73 ↓6.35	38.31 ↓3.05	7.20 ↑0.45	11.33 ↓0.81	44.80 ↓4.06	31.47 ↓2.77	67.90 ↓3.35	45.42 ↓1.02	25.78 ↑4.25	11.85 ↓13.06	47.64 ↓6.52	39.72 ↓3.94
Longest	Random	60.85 ↓1.23	42.03 ↑0.67	7.05 ↑0.30	14.90 ↑2.76	49.46 ↑0.60	34.86 ↑0.62	72.49 ↑1.24	47.46 ↑1.02	21.91 ↑0.38	29.41 ↑4.50	55.99 ↑1.83	45.45 ↑1.79
Longest	PPL	62.08 ↑0.00	41.69 ↑0.33	7.51 ↑0.76	14.89 ↑2.75	48.99 ↑0.13	35.03 ↑0.79	71.60 ↑0.35	46.78 ↑0.34	24.11 ↑2.58	29.17 ↑4.26	55.83 ↑1.67	45.50 ↑1.84
Longest	FastV	60.32 ↓1.76	40.68 ↓0.68	4.17 ↓2.58	13.26 ↑1.12	48.24 ↓0.62	33.33 ↓0.91	72.31 ↑1.06	45.76 ↓0.68	16.83 ↓4.70	30.37 ↑5.46	53.99 ↓0.17	43.85 ↑0.19
Longest	SparseVLM	56.97 ↓5.11	39.66 ↓1.70	6.60 ↓0.15	11.93 ↓0.21	44.52 ↓4.34	31.93 ↓2.31	69.31 ↓1.94	42.71 ↓3.73	23.73 ↑2.20	12.72 ↓12.19	49.90 ↓4.26	39.68 ↓3.98
InfoBatch	Random	58.38 ↓3.70	42.71 ↑1.35	5.76 ↓0.99	13.52 ↑1.38	48.95 ↑0.09	33.86 ↓0.38	71.08 ↑0.17	46.44 ↑0.00	21.38 ↓0.15	26.73 ↑1.82	53.67 ↓0.49	43.86 ↑0.20
InfoBatch	PPL	60.67 ↓1.41	42.71 ↑1.35	5.76 ↓0.99	14.29 ↑2.15	49.18 ↑0.32	34.52 ↑0.28	71.43 ↑0.18	47.80 ↑1.36	24.11 ↑2.58	30.22 ↑5.31	53.97 ↓0.19	45.50 ↑1.84
InfoBatch	FastV	59.44 ↓2.64	41.36 ↑0.00	3.87 ↓2.88	12.39 ↑0.25	48.25 ↓0.61	33.06 ↓1.18	70.72 ↓0.53	45.42 ↓1.02	13.87 ↓7.66	27.83 ↑2.92	51.53 ↓2.63	41.88 ↓1.78
InfoBatch	SparseVLM	56.08 ↓6.00	37.97 ↓3.39	6.52 ↓0.23	10.33 ↓1.81	43.84 ↓5.02	30.95 ↓3.29	68.25 ↓3.00	41.69 ↓4.75	**29.72** ↑8.19	9.95 ↓14.96	49.22 ↓4.94	39.77 ↓3.89
Entropy	Random	59.96 ↓2.12	42.71 ↑1.35	7.66 ↑0.91	**15.47** ↑3.33	49.07 ↑0.21	34.97 ↑0.73	72.13 ↑0.88	47.80 ↑1.36	22.67 ↑1.14	25.19 ↑0.28	55.70 ↑1.54	44.70 ↑1.04
Entropy	PPL	60.49 ↓1.59	41.02 ↓0.34	6.67 ↓0.08	13.51 ↑1.37	49.75 ↑0.89	34.29 ↑0.05	72.13 ↑0.88	47.80 ↑1.36	25.17 ↑3.64	30.25 ↑5.34	**56.58** ↑2.42	46.39 ↑2.73
Entropy	FastV	59.26 ↓2.82	42.37 ↑1.01	4.02 ↓2.73	13.81 ↑1.67	48.93 ↑0.07	33.68 ↓0.56	71.60 ↑0.35	47.46 ↑1.02	17.97 ↓3.56	29.82 ↑4.91	54.40 ↓0.24	44.25 ↑0.59
Entropy	SparseVLM	56.97 ↓5.11	38.64 ↓2.72	5.53 ↓1.22	11.28 ↓0.86	44.72 ↓4.14	31.43 ↓2.81	67.55 ↓3.70	43.73 ↓2.71	25.85 ↑4.32	7.09 ↓17.82	48.24 ↓5.92	38.49 ↓5.17
Q-Tuning (Ours)		**62.79** ↑0.71	42.03 ↑0.67	**10.46** ↑3.71	14.53 ↑2.39	**51.05** ↑2.19	**36.17** ↑1.93	**73.37** ↑2.12	**48.14** ↑1.70	28.81 ↑7.28	**36.35** ↑11.44	56.30 ↑2.14	**48.59** ↑4.93
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84
50% *Samples,* 70% *Tokens*													
Random	Random	61.02	41.36	7.43	15.56	48.91	34.85	71.60	47.12	22.59	27.86	53.68	44.57
Random	PPL	60.67 ↓0.35	42.71 ↑1.35	6.75 ↓0.68	14.93 ↓0.63	49.13 ↑0.22	34.84 ↓0.01	72.31 ↑0.71	47.12 ↑0.00	20.70 ↓1.89	32.02 ↑4.16	53.56 ↓0.12	45.14 ↑0.57
Random	FastV	59.61 ↓1.41	42.37 ↑1.01	5.53 ↓1.90	13.36 ↓2.20	48.48 ↓0.43	33.87 ↓0.98	72.31 ↑0.71	45.08 ↓2.04	18.95 ↓3.64	28.60 ↑0.74	51.40 ↓2.28	43.27 ↓1.30
Random	SparseVLM	54.67 ↓6.35	38.98 ↓2.38	7.28 ↓0.15	10.84 ↓4.72	44.68 ↓4.23	31.29 ↓3.56	68.78 ↓2.82	46.10 ↓1.02	23.65 ↑1.06	11.32 ↓16.54	49.04 ↓4.64	39.78 ↓4.79
Longest	Random	61.20 ↑0.18	41.02 ↓0.34	6.37 ↓1.06	14.90 ↓0.66	49.79 ↑0.88	34.65 ↓0.20	72.13 ↑0.53	**48.47** ↑1.35	23.05 ↑0.46	29.60 ↑1.74	55.94 ↑2.26	45.84 ↑1.27
Longest	PPL	62.26 ↑1.24	41.36 ↑0.00	6.82 ↓0.61	14.83 ↓0.73	49.04 ↑0.13	34.86 ↑0.01	70.90 ↓0.70	47.80 ↑0.68	24.87 ↑2.28	27.33 ↓0.53	55.62 ↑1.94	45.30 ↑0.73
Longest	FastV	59.61 ↓1.41	**42.71** ↑1.35	5.53 ↓1.90	14.12 ↓1.44	48.77 ↓0.14	34.15 ↓0.70	72.66 ↑1.06	45.08 ↓2.04	19.94 ↓2.65	31.53 ↑3.67	54.66 ↑0.98	44.77 ↑0.20
Longest	SparseVLM	55.38 ↓5.64	39.66 ↓1.70	6.14 ↓1.29	11.51 ↓4.05	44.56 ↓4.35	31.45 ↓3.40	69.14 ↓2.46	45.42 ↓1.70	23.43 ↑0.84	13.42 ↓14.44	47.86 ↓5.82	39.85 ↓4.72
InfoBatch	Random	**61.55** ↑0.53	40.34 ↓1.02	5.84 ↓1.59	14.99 ↓0.57	49.45 ↑0.54	34.43 ↓0.42	70.90 ↓0.70	45.76 ↓1.36	22.97 ↑0.38	28.85 ↑0.99	54.11 ↑0.43	44.52 ↓0.05
InfoBatch	PPL	60.85 ↓0.17	42.03 ↑0.67	6.75 ↓0.68	14.22 ↓1.34	48.41 ↓0.50	34.45 ↓0.40	71.08 ↓0.52	45.76 ↓1.36	22.82 ↑0.23	32.80 ↑4.94	54.63 ↑0.95	45.42 ↑0.85
InfoBatch	FastV	58.91 ↓2.11	41.36 ↑0.00	4.02 ↓3.41	14.98 ↓0.58	48.76 ↓0.15	33.60 ↓1.25	70.19 ↓1.41	45.76 ↓1.36	19.86 ↓2.73	28.95 ↑1.09	50.92 ↓2.76	43.14 ↓1.43
InfoBatch	SparseVLM	55.91 ↓5.11	37.29 ↓4.07	6.22 ↓1.21	10.40 ↓5.16	43.75 ↓5.16	30.71 ↓4.14	68.96 ↓2.64	44.07 ↓3.05	25.63 ↑3.04	9.82 ↓18.04	49.56 ↓4.12	39.61 ↓4.96
Entropy	Random	60.49 ↓0.53	41.69 ↑0.33	7.13 ↓0.30	15.41 ↓0.15	48.33 ↓0.58	34.61 ↓0.24	72.13 ↑0.53	47.46 ↑0.34	23.43 ↑0.84	28.99 ↑1.13	55.44 ↑1.76	45.49 ↑0.92
Entropy	PPL	60.85 ↓0.17	40.68 ↓0.68	6.29 ↓1.14	13.38 ↓2.18	49.94 ↑1.03	34.23 ↓0.62	72.31 ↑0.71	46.44 ↓0.68	23.20 ↑0.61	31.44 ↑3.58	**56.67** ↑2.99	46.01 ↑1.44
Entropy	FastV	58.91 ↓2.11	42.37 ↑1.01	4.85 ↓2.58	14.27 ↓1.29	48.84 ↓0.07	33.85 ↓1.00	71.96 ↑0.36	48.14 ↑1.02	20.39 ↓2.20	30.30 ↑2.44	54.46 ↑0.78	45.05 ↑0.48
Entropy	SparseVLM	56.79 ↓4.23	37.63 ↓3.73	5.84 ↓1.59	11.85 ↓3.71	43.82 ↓5.09	31.18 ↓3.67	67.55 ↓4.05	46.44 ↓0.68	25.17 ↑2.58	10.36 ↓17.50	48.60 ↓5.08	39.62 ↓4.95
Q-Tuning (Ours)		61.38 ↑0.36	42.03 ↑0.67	**9.55** ↑2.12	**17.69** ↑2.13	**51.04** ↑2.13	**36.34** ↑1.49	**73.90** ↑2.30	47.46 ↑0.34	**29.49** ↑6.90	**37.24** ↑9.38	56.07 ↑2.39	**48.83** ↑4.26
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84

C.2 Results on Qwen3-8B on Alignment Datasets

The detailed results on the OpenHermes dataset using Qwen3-8B (Yang et al., 2025) are summarized in Tables 10 and 11, which systematically evaluate different sample ratios and token ratios.

Strong performance with limited data budgets. Across all evaluated budgets, Q-Tuning consistently outperforms traditional pruning-based fine-tuning methods and surpasses the full-data baseline. For instance, under the extremely low-budget setting of 12.5% samples and 50% tokens, Q-Tuning reaches an average of 60.19, outperforming the full-dataset result (55.45) by more than +4.7 points. Even at 12.5% samples, 70% tokens, Q-Tuning maintains the same strong average (60.29), showing remarkable robustness under aggressive pruning.

Consistent superiority under larger budgets. When the available data increases, Q-Tuning continues to dominate other baselines. At 25% samples, 50% tokens, Q-Tuning attains an average score of 58.80, exceeding the strongest baseline (InfoBatch–Random) by nearly +3.9 points. Similarly, at 25% samples, 70% tokens, it reaches 59.20, still outperforming all competing methods such as InfoBatch, Entropy, and SparseVLM by a clear margin, and exceeding the full-data fine-tuning by +3.75 points.

Performance saturation and stable generalization. At even larger budgets, such as 50% samples with 50% or 70% tokens, Q-Tuning continues to deliver strong and stable improvements. In the 50% / 50% setting, it achieves an average of 58.39, improving over all pruning-based baselines by 3–6 points. Under 50% / 70%, it reaches 57.86, maintaining competitive accuracy and demonstrating consistent generalization despite heavy token-level sparsity.

Table 10: Evaluation on OpenHermes using Qwen3-8B under different sample ratios (12.5%, 25%) and token ratios (50%, 70%), where ↑ and ↓ respectively denote improvements or degradations over the *Random-Random* baseline.

Sample Pruner	Token Pruner	Qwen3-8B ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.
Zero-Shot		69.31	44.41	74.45	27.63	51.76	53.51
12.5% *Samples,* 50% *Tokens*							
Random	Random	76.54	51.53	60.88	35.74	40.13	52.96
	PPL	77.60 $^{\uparrow 1.06}$	52.20 $^{\uparrow 0.67}$	58.00 $^{\downarrow 2.88}$	35.55 $^{\downarrow 0.19}$	40.73 $^{\uparrow 0.60}$	52.82 $^{\downarrow 0.14}$
	FastV	76.19 $^{\downarrow 0.35}$	52.54 $^{\uparrow 1.01}$	56.48 $^{\downarrow 4.40}$	35.04 $^{\downarrow 0.70}$	41.07 $^{\uparrow 0.94}$	52.26 $^{\downarrow 0.70}$
	SparseVLM	70.90 $^{\downarrow 5.64}$	45.42 $^{\downarrow 6.11}$	73.69 $^{\uparrow 12.81}$	24.39 $^{\downarrow 11.35}$	**50.58** $^{\uparrow \mathbf{10.45}}$	53.00 $^{\uparrow 0.04}$
Longest	Random	78.48 $^{\uparrow 1.94}$	53.22 $^{\uparrow 1.69}$	62.70 $^{\uparrow 1.82}$	33.97 $^{\downarrow 1.77}$	42.13 $^{\uparrow 2.00}$	54.10 $^{\uparrow 1.14}$
	PPL	78.48 $^{\uparrow 1.94}$	53.22 $^{\uparrow 1.69}$	56.18 $^{\downarrow 4.70}$	32.05 $^{\downarrow 3.69}$	41.91 $^{\uparrow 1.78}$	52.37 $^{\downarrow 0.59}$
	FastV	77.60 $^{\uparrow 1.06}$	51.86 $^{\uparrow 0.33}$	60.12 $^{\downarrow 0.76}$	32.57 $^{\downarrow 3.17}$	43.01 $^{\uparrow 2.88}$	53.03 $^{\uparrow 0.07}$
	SparseVLM	69.14 $^{\downarrow 7.40}$	44.75 $^{\downarrow 6.78}$	75.59 $^{\uparrow 14.71}$	22.24 $^{\downarrow 13.50}$	50.02 $^{\uparrow 9.89}$	52.35 $^{\downarrow 0.61}$
InfoBatch	Random	80.25 $^{\uparrow 3.71}$	53.56 $^{\uparrow 2.03}$	60.58 $^{\downarrow 0.30}$	38.51 $^{\uparrow 2.77}$	44.36 $^{\uparrow 4.23}$	55.45 $^{\uparrow 2.49}$
	PPL	79.01 $^{\uparrow 2.47}$	53.56 $^{\uparrow 2.03}$	62.47 $^{\uparrow 1.59}$	35.69 $^{\downarrow 0.05}$	42.06 $^{\uparrow 1.93}$	54.56 $^{\uparrow 1.60}$
	FastV	79.01 $^{\uparrow 2.47}$	51.19 $^{\downarrow 0.34}$	45.49 $^{\downarrow 15.39}$	38.71 $^{\uparrow 2.97}$	44.72 $^{\uparrow 4.59}$	51.82 $^{\downarrow 1.14}$
	SparseVLM	68.96 $^{\downarrow 7.58}$	44.41 $^{\downarrow 7.12}$	**77.71** $^{\uparrow \mathbf{16.83}}$	19.88 $^{\downarrow 15.86}$	49.35 $^{\uparrow 9.22}$	52.06 $^{\downarrow 0.90}$
Entropy	Random	76.37 $^{\downarrow 0.17}$	51.53 $^{\uparrow 0.00}$	62.77 $^{\uparrow 1.89}$	33.90 $^{\downarrow 1.84}$	43.07 $^{\uparrow 2.94}$	53.53 $^{\uparrow 0.57}$
	PPL	78.66 $^{\uparrow 2.12}$	52.54 $^{\uparrow 1.01}$	61.94 $^{\uparrow 1.06}$	35.62 $^{\downarrow 0.12}$	42.73 $^{\uparrow 2.60}$	54.30 $^{\uparrow 1.34}$
	FastV	76.90 $^{\uparrow 0.36}$	53.56 $^{\uparrow 2.03}$	65.05 $^{\uparrow 4.17}$	36.14 $^{\uparrow 0.40}$	43.99 $^{\uparrow 3.86}$	55.13 $^{\uparrow 2.17}$
	SparseVLM	71.08 $^{\downarrow 5.46}$	44.07 $^{\downarrow 7.46}$	75.89 $^{\uparrow 15.01}$	22.77 $^{\downarrow 13.00}$	50.47 $^{\uparrow 10.34}$	52.85 $^{\downarrow 0.11}$
Q-Tuning (Ours)		**80.95** $^{\uparrow \mathbf{4.41}}$	**53.56** $^{\uparrow \mathbf{2.03}}$	70.51 $^{\uparrow 9.63}$	**45.76** $^{\uparrow \mathbf{10.02}}$	50.16 $^{\uparrow 10.03}$	**60.19** $^{\uparrow \mathbf{7.23}}$
12.5% *Samples,* 70% *Tokens*							
Random	Random	76.54	52.20	61.64	34.88	41.02	53.26
	PPL	77.07 $^{\uparrow 0.53}$	52.20 $^{\uparrow 0.00}$	59.21 $^{\downarrow 2.43}$	35.20 $^{\uparrow 0.32}$	40.87 $^{\downarrow 0.15}$	52.91 $^{\downarrow 0.35}$
	FastV	76.37 $^{\downarrow 0.17}$	50.51 $^{\downarrow 1.69}$	59.36 $^{\downarrow 2.28}$	35.25 $^{\uparrow 0.37}$	41.01 $^{\downarrow 0.01}$	52.50 $^{\downarrow 0.76}$
	SparseVLM	70.90 $^{\downarrow 5.64}$	45.42 $^{\downarrow 6.78}$	73.69 $^{\uparrow 12.05}$	24.39 $^{\downarrow 10.49}$	50.58 $^{\uparrow 9.56}$	53.00 $^{\downarrow 0.26}$
Longest	Random	79.19 $^{\uparrow 2.65}$	**53.90** $^{\uparrow \mathbf{1.70}}$	61.64 $^{\uparrow 0.00}$	34.14 $^{\downarrow 0.74}$	42.23 $^{\uparrow 1.21}$	54.22 $^{\uparrow 0.96}$
	PPL	78.48 $^{\uparrow 1.94}$	51.53 $^{\downarrow 0.67}$	55.65 $^{\downarrow 5.99}$	32.95 $^{\downarrow 1.93}$	42.45 $^{\uparrow 1.43}$	52.21 $^{\downarrow 1.05}$
	FastV	78.13 $^{\uparrow 1.59}$	51.53 $^{\downarrow 0.67}$	59.21 $^{\downarrow 2.43}$	33.16 $^{\downarrow 1.72}$	42.19 $^{\uparrow 1.17}$	52.84 $^{\downarrow 0.42}$
	SparseVLM	69.14 $^{\downarrow 7.40}$	44.75 $^{\downarrow 7.45}$	75.59 $^{\uparrow 13.95}$	22.24 $^{\downarrow 12.64}$	50.02 $^{\uparrow 9.00}$	52.35 $^{\downarrow 0.91}$
InfoBatch	Random	78.84 $^{\uparrow 2.30}$	51.86 $^{\downarrow 0.34}$	58.38 $^{\downarrow 3.26}$	38.12 $^{\uparrow 3.24}$	39.98 $^{\downarrow 1.04}$	53.44 $^{\uparrow 0.18}$
	PPL	78.48 $^{\uparrow 1.94}$	52.88 $^{\uparrow 0.68}$	62.17 $^{\uparrow 0.53}$	36.28 $^{\uparrow 1.40}$	42.96 $^{\uparrow 1.94}$	54.55 $^{\uparrow 1.29}$
	FastV	79.89 $^{\uparrow 3.35}$	53.22 $^{\uparrow 1.02}$	55.19 $^{\downarrow 6.45}$	38.63 $^{\uparrow 3.75}$	44.23 $^{\uparrow 3.21}$	54.23 $^{\uparrow 0.97}$
	SparseVLM	68.96 $^{\downarrow 7.58}$	44.41 $^{\downarrow 7.79}$	**77.71** $^{\uparrow \mathbf{16.07}}$	19.88 $^{\downarrow 15.00}$	49.35 $^{\uparrow 8.33}$	52.06 $^{\downarrow 1.20}$
Entropy	Random	77.07 $^{\uparrow 0.53}$	51.86 $^{\downarrow 0.34}$	60.73 $^{\downarrow 0.91}$	34.30 $^{\downarrow 0.58}$	42.34 $^{\uparrow 1.32}$	53.26 $^{\uparrow 0.00}$
	PPL	78.84 $^{\uparrow 2.30}$	52.88 $^{\uparrow 0.68}$	61.26 $^{\downarrow 0.38}$	36.44 $^{\uparrow 1.56}$	43.53 $^{\uparrow 2.51}$	54.59 $^{\uparrow 1.33}$
	FastV	77.07 $^{\uparrow 0.53}$	51.86 $^{\downarrow 0.34}$	64.22 $^{\uparrow 2.58}$	36.63 $^{\uparrow 1.75}$	43.94 $^{\uparrow 2.92}$	54.74 $^{\uparrow 1.48}$
	SparseVLM	71.08 $^{\downarrow 5.46}$	44.07 $^{\downarrow 8.13}$	75.89 $^{\uparrow 14.25}$	22.77 $^{\downarrow 12.11}$	50.47 $^{\uparrow 9.45}$	52.85 $^{\downarrow 0.41}$
Q-Tuning (Ours)		**80.95** $^{\uparrow \mathbf{4.41}}$	52.20 $^{\uparrow 0.00}$	70.13 $^{\uparrow 8.49}$	**47.60** $^{\uparrow \mathbf{12.72}}$	**50.58** $^{\uparrow \mathbf{9.56}}$	**60.29** $^{\uparrow \mathbf{7.03}}$
25% *Samples,* 50% *Tokens*							
Random	Random	77.07	51.86	61.03	35.11	40.55	53.13
	PPL	77.78 $^{\uparrow 0.71}$	54.58 $^{\uparrow 2.72}$	58.53 $^{\downarrow 2.50}$	35.43 $^{\uparrow 0.32}$	40.68 $^{\uparrow 0.13}$	53.40 $^{\uparrow 0.27}$
	FastV	79.54 $^{\uparrow 2.47}$	52.54 $^{\uparrow 0.68}$	53.45 $^{\downarrow 7.58}$	36.89 $^{\uparrow 1.78}$	43.13 $^{\uparrow 2.58}$	53.11 $^{\downarrow 0.02}$
	SparseVLM	69.14 $^{\downarrow 7.93}$	45.76 $^{\downarrow 6.10}$	**77.71** $^{\uparrow \mathbf{16.68}}$	20.80 $^{\downarrow 14.31}$	50.20 $^{\uparrow 9.65}$	52.72 $^{\downarrow 0.41}$
Longest	Random	79.01 $^{\uparrow 1.94}$	54.58 $^{\uparrow 2.72}$	51.10 $^{\downarrow 9.93}$	36.30 $^{\uparrow 1.19}$	41.93 $^{\uparrow 1.38}$	52.58 $^{\downarrow 0.55}$
	PPL	79.01 $^{\uparrow 1.94}$	**54.92** $^{\uparrow \mathbf{3.06}}$	55.42 $^{\downarrow 5.61}$	35.10 $^{\downarrow 0.01}$	41.09 $^{\uparrow 0.54}$	53.11 $^{\downarrow 0.02}$
	FastV	**80.95** $^{\uparrow \mathbf{3.88}}$	51.53 $^{\downarrow 0.33}$	53.30 $^{\downarrow 7.73}$	39.73 $^{\uparrow 4.62}$	44.34 $^{\uparrow 3.79}$	53.97 $^{\uparrow 0.84}$
	SparseVLM	72.49 $^{\downarrow 4.58}$	44.41 $^{\downarrow 7.45}$	74.83 $^{\uparrow 13.80}$	22.08 $^{\downarrow 13.03}$	48.93 $^{\uparrow 8.38}$	52.55 $^{\downarrow 0.58}$
InfoBatch	Random	77.25 $^{\uparrow 0.18}$	51.53 $^{\downarrow 0.33}$	63.61 $^{\uparrow 2.58}$	36.99 $^{\uparrow 1.88}$	45.15 $^{\uparrow 4.60}$	54.91 $^{\uparrow 1.78}$
	PPL	78.84 $^{\uparrow 1.77}$	54.58 $^{\uparrow 2.72}$	60.20 $^{\downarrow 0.83}$	35.81 $^{\uparrow 0.70}$	43.53 $^{\uparrow 3.00}$	54.59 $^{\uparrow 1.46}$
	FastV	80.25 $^{\uparrow 3.18}$	52.54 $^{\uparrow 0.68}$	48.52 $^{\downarrow 12.51}$	39.62 $^{\uparrow 4.51}$	44.87 $^{\uparrow 4.32}$	53.16 $^{\uparrow 0.03}$
	SparseVLM	69.49 $^{\downarrow 7.58}$	44.07 $^{\downarrow 7.79}$	76.88 $^{\uparrow 15.85}$	18.29 $^{\downarrow 16.82}$	50.16 $^{\uparrow 9.61}$	51.78 $^{\downarrow 1.35}$
Entropy	Random	77.07 $^{\uparrow 0.00}$	53.22 $^{\uparrow 1.36}$	62.02 $^{\uparrow 1.00}$	35.40 $^{\uparrow 0.29}$	42.91 $^{\uparrow 2.36}$	54.12 $^{\uparrow 0.99}$
	PPL	79.37 $^{\uparrow 2.30}$	54.24 $^{\uparrow 2.38}$	52.24 $^{\downarrow 8.79}$	33.13 $^{\downarrow 1.98}$	40.82 $^{\uparrow 0.27}$	51.96 $^{\downarrow 1.17}$
	FastV	79.54 $^{\uparrow 2.47}$	51.86 $^{\uparrow 0.00}$	52.92 $^{\downarrow 8.11}$	35.68 $^{\uparrow 0.57}$	43.79 $^{\uparrow 3.24}$	52.76 $^{\downarrow 0.37}$
	SparseVLM	70.19 $^{\downarrow 6.88}$	43.39 $^{\downarrow 8.47}$	75.21 $^{\uparrow 14.18}$	19.97 $^{\downarrow 15.14}$	**50.46** $^{\uparrow \mathbf{9.91}}$	51.84 $^{\downarrow 1.29}$
Q-Tuning (Ours)		79.89 $^{\uparrow 2.82}$	52.88 $^{\uparrow 1.02}$	69.37 $^{\uparrow 8.34}$	**43.54** $^{\uparrow \mathbf{8.43}}$	48.33 $^{\uparrow 7.78}$	**58.80** $^{\uparrow \mathbf{5.67}}$
Full Dataset		77.78	52.20	65.13	38.05	44.08	55.45

Table 11: Evaluation on OpenHermes using Qwen3-8B under different sample ratios (25%, 50%) and token ratios (50%, 70%), where ↑ and ↓ respectively denote improvements or degradations over the *Random-Random* baseline.

Sample Pruner	Token Pruner	ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.
		Qwen3-8B					
Zero-Shot		69.31	44.41	74.45	27.63	51.76	53.51
25% *Samples*, 70% *Tokens*							
Random	Random	77.43	52.20	57.85	35.55	41.47	52.90
	PPL	76.90 ↓0.53	53.90 ↑1.70	60.27 ↑2.42	35.70 ↑0.15	40.73 ↓0.74	53.50 ↑0.60
	FastV	79.19 ↑1.76	51.86 ↓0.34	56.03 ↓1.82	36.01 ↑0.46	43.86 ↑2.39	53.39 ↑0.49
	SparseVLM	69.14 ↓8.29	45.76 ↓6.44	**77.71 ↑19.86**	20.80 ↓14.75	50.20 ↑8.73	52.72 ↓0.18
Longest	Random	79.01 ↑1.58	54.58 ↑2.38	53.83 ↓4.02	37.56 ↑2.01	42.40 ↑0.93	53.48 ↑0.58
	PPL	78.48 ↑1.05	**55.59 ↑3.39**	55.72 ↓2.13	35.16 ↓0.39	41.31 ↓0.16	53.25 ↑0.35
	FastV	79.89 ↑2.46	52.54 ↑0.34	53.22 ↓4.63	37.90 ↑2.35	43.79 ↑2.32	53.47 ↑0.57
	SparseVLM	72.49 ↓4.94	44.41 ↓7.79	74.83 ↑16.98	22.08 ↓13.47	48.93 ↑7.46	52.55 ↓0.35
InfoBatch	Random	78.66 ↑1.23	52.54 ↑0.34	63.68 ↑5.83	37.56 ↑2.01	44.51 ↑3.04	55.39 ↑2.49
	PPL	78.84 ↑1.41	53.22 ↑1.02	62.40 ↑4.55	36.39 ↑0.84	44.87 ↑3.40	55.14 ↑2.24
	FastV	**79.89 ↑2.46**	53.90 ↑1.70	55.04 ↓2.82	38.79 ↑3.24	44.28 ↑2.81	54.38 ↑1.48
	SparseVLM	69.49 ↓7.94	44.07 ↓8.13	76.88 ↑19.03	18.29 ↓17.26	50.16 ↑8.69	51.78 ↓1.12
Entropy	Random	76.90 ↓0.53	52.54 ↑0.34	57.47 ↓0.38	34.58 ↓0.97	42.50 ↑1.03	52.80 ↓0.10
	PPL	78.66 ↑1.23	53.90 ↑1.70	52.46 ↓5.39	33.72 ↓1.83	41.16 ↓0.31	51.98 ↓0.92
	FastV	79.01 ↑1.58	52.20 ↑0.00	57.54 ↓0.31	35.72 ↑0.17	42.94 ↑1.47	53.48 ↑0.58
	SparseVLM	70.19 ↓7.24	43.39 ↓8.81	75.21 ↑17.36	19.97 ↓15.58	**50.46 ↑9.00**	51.84 ↓1.06
Q-Tuning (Ours)		79.72 ↑2.29	52.88 ↑0.68	71.11 ↑13.26	**43.70 ↑8.15**	48.58 ↑7.11	**59.20 ↑6.30**
50% *Samples*, 50% *Tokens*							
Random	Random	77.95	51.86	63.61	36.64	42.22	54.46
	PPL	78.48 ↑0.53	53.22 ↑1.36	62.02 ↓1.59	35.96 ↓0.68	42.56 ↑0.34	54.45 ↓0.01
	FastV	79.19 ↑1.24	51.53 ↓0.33	49.73 ↓13.88	40.64 ↑4.00	44.45 ↑2.23	53.11 ↓1.35
	SparseVLM	72.13 ↓5.82	46.10 ↓5.76	73.69 ↑10.08	18.74 ↓17.90	**50.39 ↑8.17**	52.21 ↓2.25
Longest	Random	78.84 ↑0.89	54.24 ↑2.38	59.67 ↓3.94	36.66 ↑0.02	43.05 ↑0.83	54.49 ↑0.03
	PPL	79.72 ↑1.77	**54.24 ↑2.38**	58.15 ↓5.46	39.88 ↑3.24	44.01 ↑1.79	55.20 ↑0.74
	FastV	79.72 ↑1.77	51.86 ↑0.00	47.99 ↓15.62	40.57 ↑3.93	44.89 ↑2.67	53.01 ↓1.45
	SparseVLM	70.19 ↓7.76	43.39 ↓8.47	74.22 ↑10.61	17.90 ↓18.74	49.34 ↑7.12	51.01 ↓3.45
InfoBatch	Random	78.84 ↑0.89	52.88 ↑1.02	64.37 ↑0.76	36.98 ↑0.34	45.49 ↑3.27	55.71 ↑1.25
	PPL	77.95 ↑0.00	53.56 ↑1.70	60.73 ↓2.88	38.10 ↑1.47	44.39 ↑2.17	54.95 ↑0.49
	FastV	**80.42 ↑2.47**	52.88 ↑1.02	50.34 ↓13.27	**40.79 ↑4.15**	45.14 ↑2.92	53.92 ↓0.54
	SparseVLM	66.49 ↓11.46	43.39 ↓8.47	74.30 ↑10.69	15.16 ↓21.48	49.64 ↑7.42	49.80 ↓4.66
Entropy	Random	78.13 ↑0.18	51.53 ↓0.33	61.94 ↓1.67	34.59 ↓2.05	42.23 ↑0.01	53.68 ↓0.78
	PPL	78.66 ↑0.71	52.88 ↑1.02	54.97 ↓8.61	36.11 ↓0.53	42.06 ↓0.16	52.94 ↓1.52
	FastV	79.54 ↑1.59	53.22 ↑1.36	47.08 ↓16.53	40.24 ↑3.58	43.33 ↑1.11	52.68 ↓1.78
	SparseVLM	70.90 ↓7.05	43.39 ↓8.47	**78.70 ↑15.09**	16.25 ↓20.39	50.10 ↑7.87	51.87 ↓2.59
Q-Tuning (Ours)		80.25 ↑2.30	53.90 ↑2.04	70.05 ↑6.44	39.03 ↑2.39	48.70 ↑6.48	**58.39 ↑3.93**
50% *Samples*, 70% *Tokens*							
Random	Random	78.13	52.20	61.49	36.91	42.75	54.30
	PPL	77.95 ↓0.18	52.20 ↑0.00	59.59 ↓1.90	36.14 ↓0.77	42.51 ↓0.24	53.68 ↓0.62
	FastV	79.01 ↑0.88	53.90 ↑1.70	58.30 ↓3.19	40.33 ↑3.42	45.20 ↑2.45	55.35 ↑1.05
	SparseVLM	72.13 ↓6.00	46.10 ↓6.10	73.69 ↑12.20	18.74 ↓18.17	**50.39 ↑7.64**	52.21 ↓2.09
Longest	Random	78.48 ↑0.35	52.54 ↑0.34	59.14 ↓2.35	37.27 ↑0.36	42.38 ↓0.37	53.96 ↓0.34
	PPL	79.19 ↑1.06	53.56 ↑1.36	60.96 ↓0.53	39.78 ↑2.87	44.17 ↑1.42	55.53 ↑1.23
	FastV	**80.25 ↑2.12**	52.20 ↑0.00	56.79 ↓4.70	**40.53 ↑3.62**	45.62 ↑2.87	55.08 ↑0.78
	SparseVLM	70.19 ↓7.94	43.39 ↓8.81	74.22 ↑12.73	17.90 ↓19.01	49.34 ↑6.59	51.01 ↓3.29
InfoBatch	Random	78.31 ↑0.18	52.88 ↑0.68	62.32 ↑0.83	35.55 ↓1.36	45.15 ↑2.40	54.84 ↑0.54
	PPL	78.31 ↑0.18	**54.58 ↑2.38**	63.23 ↑1.74	38.89 ↑1.98	44.16 ↑1.41	55.83 ↑1.53
	FastV	79.01 ↑0.88	52.88 ↑0.68	58.30 ↓3.19	39.05 ↑2.14	46.05 ↑3.30	55.06 ↑0.76
	SparseVLM	66.49 ↓11.64	43.39 ↓8.81	74.30 ↑12.81	15.16 ↓21.75	49.64 ↑6.89	49.80 ↓4.50
Entropy	Random	78.48 ↑0.35	53.22 ↑1.02	60.65 ↓0.84	35.64 ↓1.27	42.31 ↓0.44	54.06 ↓0.24
	PPL	78.13 ↑0.00	53.22 ↑1.02	57.85 ↓3.64	36.85 ↓0.06	41.70 ↓1.05	53.55 ↓0.75
	FastV	79.37 ↑1.24	52.88 ↑0.68	52.62 ↓8.87	39.76 ↑2.85	44.18 ↑1.43	53.76 ↓0.54
	SparseVLM	70.90 ↓7.23	43.39 ↓8.81	**78.70 ↑17.21**	16.25 ↓20.66	50.10 ↑7.79	51.87 ↓2.43
Q-Tuning (Ours)		78.84 ↑0.71	52.88 ↑0.68	69.90 ↑8.41	39.54 ↑2.63	48.15 ↑5.40	**57.86 ↑3.56**
Full Dataset		77.78	52.20	65.13	38.05	44.08	55.45

C.3 Additional Results on Reasoning Datasets

The detailed results of additional experiments on math reasoning are presented in Table 12, providing a comprehensive comparison across different pruning strategies.

Table 12: Evaluation of pruning strategies on GSM8K and MATH under 25% samples with 70% tokens, and 50% samples with 70% tokens settings. ↑ and ↓ respectively indicate improvements or degradations over the *Random-Random* baseline under the same sample and token keep ratio.

Sample Pruner	Token Pruner	LLaMA3-8B GSM8K	LLaMA3-8B MATH	LLaMA3-8B Avg.	Mistral-7B GSM8K	Mistral-7B MATH	Mistral-7B Avg.	SmolLM2-1.7B GSM8K	SmolLM2-1.7B MATH	SmolLM2-1.7B Avg.
Zero-Shot		27.82	2.26	15.04	19.86	3.30	11.58	15.47	2.20	8.83
25% Samples, 70% Tokens										
Random	Random	25.09	2.20	13.65	24.11	1.68	12.89	13.80	2.22	8.01
	PPL	23.65 ↓1.44	2.62 ↑0.42	13.14 ↓0.51	25.32 ↑1.21	1.54 ↓0.14	13.43 ↑0.54	13.04 ↓0.76	2.28 ↑0.06	7.66 ↓0.35
	FastV	16.91 ↓8.18	2.16 ↓0.04	9.53 ↓4.12	16.07 ↓8.04	1.60 ↓0.08	8.84 ↓4.05	12.89 ↓0.91	1.94 ↓0.28	7.41 ↓0.60
	SparseVLM	22.97 ↓2.12	4.72 ↑2.52	13.85 ↑0.20	19.26 ↓4.85	4.58 ↑2.90	11.92 ↓0.97	13.19 ↓0.61	3.48 ↑1.26	8.34 ↑0.33
Longest	Random	25.47 ↑0.38	3.34 ↑1.14	14.41 ↑0.76	21.83 ↓2.28	1.76 ↑0.08	11.80 ↓1.09	12.96 ↓0.84	1.96 ↓0.26	7.46 ↓0.55
	PPL	24.87 ↓0.22	3.22 ↑1.02	14.04 ↑0.39	23.65 ↓0.46	2.22 ↑0.54	12.94 ↑0.05	14.33 ↑0.53	1.60 ↓0.62	7.96 ↓0.05
	FastV	18.95 ↓6.14	2.84 ↑0.64	10.90 ↓2.75	12.96 ↓11.15	1.74 ↑0.06	7.35 ↓5.54	12.59 ↓1.21	1.58 ↓0.64	7.08 ↓0.93
	SparseVLM	26.91 ↑1.82	4.68 ↑2.48	15.80 ↑2.15	24.34 ↑0.23	4.84 ↑3.16	14.59 ↑1.70	12.43 ↓1.37	3.60 ↑1.38	8.02 ↑0.01
InfoBatch	Random	26.91 ↑1.82	2.60 ↑0.40	14.76 ↑1.11	28.96 ↑4.85	2.44 ↑0.76	15.70 ↑2.81	13.72 ↓0.08	1.82 ↓0.40	7.77 ↓0.24
	PPL	25.93 ↑0.84	2.48 ↑0.28	14.20 ↑0.55	31.46 ↑7.35	2.18 ↑0.50	16.82 ↑3.93	14.86 ↑1.06	1.90 ↓0.32	8.38 ↑0.37
	FastV	16.83 ↓8.26	2.30 ↑0.10	9.57 ↓4.08	13.87 ↓10.24	2.06 ↑0.38	7.97 ↓4.92	10.84 ↓2.96	1.72 ↓0.50	6.28 ↓1.73
	SparseVLM	14.63 ↓10.46	3.26 ↑1.06	8.95 ↓4.70	11.90 ↓12.21	1.94 ↑0.26	6.92 ↓5.97	11.90 ↓1.90	**4.36 ↑2.14**	8.13 ↑0.12
Entropy	Random	31.92 ↑6.83	2.50 ↑0.30	17.21 ↑3.56	32.37 ↑8.26	1.92 ↑0.24	17.15 ↑4.26	14.94 ↑1.14	1.92 ↓0.30	8.43 ↑0.42
	PPL	33.13 ↑8.04	2.86 ↑0.66	18.00 ↑4.35	30.17 ↑6.06	1.96 ↑0.28	16.07 ↑3.18	14.94 ↑1.14	1.96 ↓0.26	8.45 ↑0.44
	FastV	25.25 ↑0.16	2.40 ↑0.20	13.82 ↑0.17	21.00 ↓3.11	1.32 ↓0.36	11.16 ↓1.73	14.18 ↑0.38	1.98 ↓0.24	8.08 ↑0.07
	SparseVLM	20.85 ↓4.24	**5.12 ↑2.92**	12.98 ↓0.67	19.56 ↓4.55	4.20 ↑2.52	11.88 ↓1.01	14.18 ↑0.38	2.90 ↑0.68	8.54 ↑0.53
Q-Tuning (Ours)		**37.23 ↑12.14**	4.86 ↑2.66	**21.04 ↑7.39**	**42.99 ↑18.88**	**5.08 ↑3.40**	**24.56 ↑11.67**	**22.90 ↑9.10**	3.64 ↑1.42	**13.27 ↑5.26**
50% Samples, 70% Tokens										
Random	Random	27.90	2.50	15.20	32.30	2.46	17.38	14.94	1.76	8.35
	PPL	27.45 ↓0.45	2.50 ↑0.00	14.97 ↓0.23	31.99 ↓0.31	2.04 ↓0.42	17.02 ↓0.36	16.15 ↑1.21	1.80 ↑0.04	8.97 ↑0.62
	FastV	18.20 ↓9.70	1.92 ↓0.58	10.06 ↓5.14	17.44 ↓14.86	1.72 ↓0.74	9.58 ↓7.80	12.13 ↓2.81	1.60 ↓0.16	6.87 ↓1.48
	SparseVLM	10.31 ↓17.59	3.58 ↑1.08	6.95 ↓8.25	12.21 ↓20.09	2.86 ↑0.40	7.53 ↓9.85	11.90 ↓3.04	3.34 ↑1.58	7.62 ↓0.73
Longest	Random	26.23 ↓1.67	2.76 ↑0.26	14.50 ↓0.70	27.82 ↓4.48	2.20 ↓0.26	15.01 ↓2.37	15.69 ↑0.75	1.92 ↑0.16	8.81 ↑0.46
	PPL	30.25 ↑2.35	2.64 ↑0.14	16.45 ↑1.25	31.46 ↓0.84	2.00 ↓0.46	16.73 ↓0.65	16.45 ↑1.51	1.94 ↑0.18	9.20 ↑0.85
	FastV	18.42 ↓9.48	2.28 ↓0.22	10.35 ↓4.85	18.65 ↓13.65	1.82 ↓0.64	10.24 ↓7.14	12.43 ↓2.51	1.94 ↑0.18	7.19 ↓1.16
	SparseVLM	18.95 ↓8.95	**4.84 ↑2.34**	11.90 ↓3.30	19.48 ↓12.82	4.74 ↑2.28	12.11 ↓5.27	11.07 ↓3.87	3.90 ↑2.14	7.48 ↓0.87
InfoBatch	Random	29.42 ↑1.52	2.92 ↑0.42	16.17 ↑0.97	35.03 ↑2.73	2.50 ↑0.04	18.76 ↑1.38	16.00 ↑1.06	1.78 ↑0.02	8.89 ↑0.54
	PPL	28.73 ↑0.83	2.78 ↑0.28	15.76 ↑0.56	38.82 ↑6.52	2.98 ↑0.52	20.90 ↑3.52	15.39 ↑0.45	2.20 ↑0.44	8.80 ↑0.45
	FastV	19.33 ↓8.57	1.76 ↓0.74	10.55 ↓4.65	18.04 ↓14.26	1.94 ↓0.52	9.99 ↓7.39	12.05 ↓2.89	1.86 ↑0.10	6.96 ↓1.39
	SparseVLM	9.10 ↓18.80	3.12 ↑0.62	6.11 ↓9.09	12.05 ↓20.25	2.32 ↓0.14	7.19 ↓10.19	10.84 ↓4.10	**3.96 ↑2.20**	7.40 ↓0.95
Entropy	Random	32.60 ↑4.70	2.14 ↓0.36	17.37 ↑2.17	38.67 ↑6.37	2.10 ↓0.36	20.38 ↑3.00	18.88 ↑3.94	2.18 ↑0.42	10.53 ↑2.18
	PPL	33.97 ↑6.07	3.10 ↑0.60	18.53 ↑3.33	40.18 ↑7.88	2.30 ↓0.16	21.24 ↑3.86	17.21 ↑2.27	2.06 ↑0.30	9.64 ↑1.29
	FastV	24.94 ↓2.96	2.26 ↓0.24	13.60 ↓1.60	21.53 ↓10.77	1.56 ↓0.90	11.55 ↓5.83	16.22 ↑1.28	1.78 ↑0.02	9.00 ↑0.65
	SparseVLM	10.31 ↓17.59	4.56 ↑2.06	7.44 ↓7.76	9.40 ↓22.90	4.40 ↑1.94	6.90 ↓10.48	13.95 ↓0.99	3.68 ↑1.92	8.82 ↑0.47
Q-Tuning (Ours)		**38.21 ↑10.31**	4.30 ↑1.80	**21.26 ↑6.06**	**48.07 ↑15.77**	**6.14 ↑3.68**	**26.57 ↑9.19**	**20.47 ↑5.53**	3.20 ↑1.44	**11.84 ↑3.49**
Full Dataset		32.90	3.02	17.96	42.08	3.08	22.58	16.53	2.10	9.31

C.4 Additional Results on additional baselines

Tables 13 and 14 extend our evaluation to a broader set of pruning strategies under matched sample–token budgets. Table 13 varies sample ratios (12.5%, 25%) and token ratios (50%, 100%) under a fixed random token pruner, while Table 14 fixes the token ratio at 50% and varies the sample ratio (12.5%, 25%, 100%) under a random sample pruner.

Table 13: Evaluation of stronger task-relevant baselines under different sample ratios (12.5%, 25%) and token ratios (50%, 100%) with the same token pruner (*Random*), where ↑ and ↓ respectively denote improvements or degradations over the *Random* baseline.

Sample Pruner	Token Ratio	LLaMA2-7B ARC-E	LLaMA2-7B ARC-C	LLaMA2-7B GSM8K	LLaMA2-7B SQuAD	LLaMA2-7B TriviaQA	LLaMA2-7B Avg.	Mistral-7B ARC-E	Mistral-7B ARC-C	Mistral-7B GSM8K	Mistral-7B SQuAD	Mistral-7B TriviaQA	Mistral-7B Avg.
Zero-Shot		53.44	38.98	5.31	12.18	43.00	30.58	66.67	46.10	18.35	10.01	43.77	36.98
12.5% *Samples*													
Random	50%	59.25	41.02	8.11	12.75	48.75	33.98	70.55	48.14	22.74	19.57	52.63	42.73
Alpagasus	50%	58.91$^{\downarrow 0.34}$	41.69$^{\uparrow 0.67}$	6.67$^{\downarrow 1.44}$	12.81$^{\uparrow 0.06}$	46.28$^{\downarrow 2.47}$	33.27$^{\downarrow 0.71}$	72.84$^{\uparrow 2.29}$	48.47$^{\uparrow 0.33}$	20.55$^{\downarrow 2.19}$	20.41$^{\uparrow 0.84}$	52.21$^{\downarrow 0.42}$	42.90$^{\uparrow 0.17}$
	100%	58.91$^{\downarrow 0.34}$	41.36$^{\uparrow 0.34}$	7.58$^{\downarrow 0.53}$	13.45$^{\uparrow 0.70}$	46.96$^{\downarrow 1.79}$	33.65$^{\downarrow 0.33}$	70.72$^{\uparrow 0.17}$	48.81$^{\uparrow 0.67}$	23.35$^{\uparrow 0.61}$	24.30$^{\uparrow 4.73}$	52.71$^{\uparrow 0.08}$	43.98$^{\uparrow 1.25}$
Deita	50%	58.20$^{\downarrow 1.05}$	42.37$^{\uparrow 1.35}$	7.43$^{\downarrow 0.68}$	13.85$^{\uparrow 1.10}$	45.59$^{\downarrow 3.16}$	33.49$^{\downarrow 0.49}$	72.66$^{\uparrow 2.11}$	**50.51**$^{\uparrow 2.37}$	22.97$^{\downarrow 0.23}$	20.91$^{\uparrow 1.34}$	54.17$^{\uparrow 1.54}$	44.25$^{\uparrow 1.52}$
	100%	58.38$^{\downarrow 0.87}$	42.37$^{\uparrow 1.35}$	5.46$^{\downarrow 2.65}$	13.99$^{\uparrow 1.24}$	45.63$^{\downarrow 3.12}$	33.17$^{\downarrow 0.81}$	72.13$^{\uparrow 1.58}$	50.17$^{\uparrow 2.03}$	22.51$^{\downarrow 0.23}$	24.46$^{\uparrow 4.89}$	55.28$^{\uparrow 2.65}$	44.92$^{\uparrow 2.19}$
DS2	50%	58.73$^{\downarrow 0.52}$	42.37$^{\uparrow 1.35}$	5.99$^{\downarrow 2.12}$	15.05$^{\uparrow 2.30}$	45.51$^{\downarrow 3.24}$	33.53$^{\downarrow 0.45}$	71.25$^{\downarrow 0.70}$	47.46$^{\downarrow 0.68}$	21.23$^{\downarrow 1.51}$	21.57$^{\uparrow 2.00}$	51.23$^{\downarrow 1.40}$	42.55$^{\downarrow 0.18}$
	100%	59.08$^{\downarrow 0.17}$	41.36$^{\uparrow 0.34}$	6.67$^{\downarrow 1.44}$	15.14$^{\uparrow 2.39}$	45.69$^{\downarrow 3.06}$	33.59$^{\downarrow 0.39}$	72.13$^{\uparrow 1.58}$	49.83$^{\uparrow 1.69}$	23.35$^{\uparrow 0.61}$	23.99$^{\uparrow 4.42}$	52.85$^{\uparrow 0.22}$	44.43$^{\uparrow 1.70}$
LESS	50%	58.20$^{\downarrow 1.05}$	42.37$^{\uparrow 1.35}$	6.67$^{\downarrow 1.44}$	14.56$^{\uparrow 1.81}$	46.62$^{\downarrow 2.13}$	33.69$^{\downarrow 0.29}$	73.19$^{\uparrow 2.64}$	49.49$^{\uparrow 1.35}$	22.90$^{\downarrow 0.16}$	19.61$^{\uparrow 0.04}$	53.14$^{\uparrow 0.51}$	43.67$^{\uparrow 0.94}$
	100	58.38$^{\downarrow 0.87}$	**43.39**$^{\uparrow 2.37}$	6.82$^{\downarrow 1.29}$	16.16$^{\uparrow 3.41}$	46.72$^{\downarrow 2.03}$	34.30$^{\uparrow 0.32}$	**73.90**$^{\uparrow 3.35}$	48.47$^{\uparrow 0.33}$	23.12$^{\uparrow 0.38}$	21.67$^{\uparrow 2.10}$	56.00$^{\uparrow 3.37}$	44.63$^{\uparrow 1.90}$
Q-Tuning (Ours)	50%	**64.20**$^{\uparrow 4.95}$	42.03$^{\uparrow 1.01}$	**10.54**$^{\uparrow 2.43}$	**18.79**$^{\uparrow 6.04}$	**53.12**$^{\uparrow 4.37}$	**37.74**$^{\uparrow 3.76}$	71.60$^{\uparrow 1.05}$	48.14$^{\uparrow 0.00}$	**29.34**$^{\uparrow 6.60}$	**27.75**$^{\uparrow 8.18}$	**57.78**$^{\uparrow 5.15}$	**46.92**$^{\uparrow 4.19}$
25% *Samples*													
Random	50%	60.32	41.69	5.76	13.43	48.41	33.92	70.19	46.10	20.62	24.07	53.74	42.95
Alpagasus	50%	59.61$^{\downarrow 0.71}$	42.71$^{\uparrow 1.02}$	5.91$^{\uparrow 0.15}$	14.12$^{\uparrow 0.69}$	47.30$^{\downarrow 1.11}$	33.93$^{\uparrow 0.01}$	70.37$^{\uparrow 0.18}$	47.80$^{\uparrow 1.70}$	20.85$^{\uparrow 0.23}$	22.38$^{\downarrow 1.69}$	52.87$^{\downarrow 0.87}$	42.78$^{\downarrow 0.17}$
	100%	60.49$^{\uparrow 0.17}$	43.39$^{\uparrow 1.70}$	7.51$^{\uparrow 1.75}$	13.16$^{\downarrow 0.27}$	48.41$^{\uparrow 0.00}$	34.59$^{\uparrow 0.67}$	71.60$^{\uparrow 1.41}$	**50.17**$^{\uparrow 4.07}$	23.35$^{\uparrow 2.73}$	22.38$^{\downarrow 1.69}$	55.73$^{\uparrow 1.99}$	44.65$^{\uparrow 1.70}$
Deita	50%	59.08$^{\downarrow 1.24}$	41.69$^{\uparrow 0.00}$	7.66$^{\uparrow 1.90}$	13.23$^{\downarrow 0.20}$	47.55$^{\downarrow 0.86}$	33.84$^{\downarrow 0.08}$	72.13$^{\uparrow 1.94}$	47.80$^{\uparrow 1.70}$	21.00$^{\uparrow 0.38}$	23.16$^{\downarrow 0.91}$	53.97$^{\uparrow 0.23}$	43.61$^{\uparrow 0.66}$
	100%	58.73$^{\downarrow 1.59}$	42.71$^{\uparrow 1.02}$	4.92$^{\downarrow 0.84}$	14.88$^{\uparrow 1.45}$	48.24$^{\downarrow 0.17}$	33.90$^{\downarrow 0.02}$	**72.66**$^{\uparrow 2.47}$	47.80$^{\uparrow 1.70}$	22.29$^{\uparrow 1.67}$	25.87$^{\uparrow 1.80}$	54.40$^{\uparrow 0.66}$	44.60$^{\uparrow 1.65}$
DS2	50%	59.44$^{\downarrow 0.88}$	42.03$^{\uparrow 0.34}$	6.44$^{\uparrow 0.68}$	15.59$^{\uparrow 2.16}$	47.34$^{\downarrow 1.07}$	34.17$^{\uparrow 0.25}$	70.72$^{\uparrow 0.53}$	48.47$^{\uparrow 2.37}$	22.74$^{\uparrow 2.12}$	21.15$^{\downarrow 2.92}$	53.49$^{\downarrow 0.25}$	43.32$^{\uparrow 0.37}$
	100%	59.26$^{\downarrow 1.06}$	41.36$^{\downarrow 0.33}$	7.05$^{\uparrow 1.29}$	15.20$^{\uparrow 1.77}$	48.76$^{\uparrow 0.35}$	34.33$^{\uparrow 0.41}$	71.43$^{\uparrow 1.24}$	48.47$^{\uparrow 2.37}$	22.44$^{\uparrow 1.82}$	21.72$^{\downarrow 2.35}$	53.76$^{\uparrow 0.02}$	43.57$^{\uparrow 0.62}$
LESS	50%	59.44$^{\downarrow 0.88}$	**43.39**$^{\uparrow 1.70}$	7.05$^{\uparrow 1.29}$	15.89$^{\uparrow 2.46}$	46.60$^{\downarrow 1.81}$	34.47$^{\uparrow 0.55}$	71.78$^{\uparrow 1.59}$	47.80$^{\uparrow 1.70}$	22.14$^{\uparrow 1.52}$	24.51$^{\uparrow 0.44}$	53.04$^{\downarrow 0.70}$	43.85$^{\uparrow 0.90}$
	100	59.79$^{\downarrow 0.53}$	41.02$^{\downarrow 0.67}$	6.37$^{\uparrow 0.61}$	14.44$^{\uparrow 1.01}$	48.44$^{\uparrow 0.03}$	34.01$^{\uparrow 0.09}$	71.60$^{\uparrow 1.41}$	49.49$^{\uparrow 3.39}$	21.68$^{\uparrow 1.06}$	30.85$^{\uparrow 6.78}$	53.40$^{\downarrow 0.34}$	45.41$^{\uparrow 2.46}$
Q-Tuning (Ours)	50%	**63.14**$^{\uparrow 2.82}$	42.03$^{\uparrow 0.34}$	**8.87**$^{\uparrow 3.11}$	**16.76**$^{\uparrow 3.33}$	**51.52**$^{\uparrow 3.11}$	**36.47**$^{\uparrow 2.55}$	71.78$^{\uparrow 1.59}$	47.12$^{\uparrow 1.02}$	**26.08**$^{\uparrow 5.46}$	**32.79**$^{\uparrow 8.72}$	**56.17**$^{\uparrow 2.43}$	**46.79**$^{\uparrow 3.84}$
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84

Table 14: Evaluation of stronger task-relevant baselines under different sample ratios (12.5%, 25%, 100%) and fixed token ratio (50%) with the same sample pruner (*Random*), where ↑ and ↓ respectively denote improvements or degradations over the *Random* baseline.

Token Pruner	Sample Ratio	LLaMA2-7B ARC-E	LLaMA2-7B ARC-C	LLaMA2-7B GSM8K	LLaMA2-7B SQuAD	LLaMA2-7B TriviaQA	LLaMA2-7B Avg.	Mistral-7B ARC-E	Mistral-7B ARC-C	Mistral-7B GSM8K	Mistral-7B SQuAD	Mistral-7B TriviaQA	Mistral-7B Avg.
Zero-Shot		53.44	38.98	5.31	12.18	43.00	30.58	66.67	46.10	18.35	10.01	43.77	36.98
50% *Tokens*													
Random	12.5%	59.25	41.02	8.11	12.75	48.75	33.98	70.55	48.14	22.74	19.57	52.63	42.73
	25%	60.32$^{\uparrow 1.07}$	41.69$^{\uparrow 0.67}$	5.76$^{\downarrow 2.35}$	13.43$^{\uparrow 0.68}$	48.41$^{\downarrow 0.34}$	33.92$^{\downarrow 0.06}$	70.19$^{\downarrow 0.36}$	46.10$^{\downarrow 2.04}$	20.62$^{\downarrow 2.12}$	24.07$^{\uparrow 4.50}$	53.74$^{\uparrow 1.11}$	42.95$^{\uparrow 0.22}$
Rho-1	12.5%	53.97$^{\downarrow 5.28}$	40.34$^{\downarrow 0.68}$	6.67$^{\downarrow 1.44}$	13.27$^{\uparrow 0.52}$	46.62$^{\downarrow 2.13}$	32.17$^{\downarrow 1.81}$	71.60$^{\uparrow 1.05}$	47.46$^{\downarrow 0.68}$	20.62$^{\downarrow 2.12}$	23.74$^{\uparrow 4.17}$	53.71$^{\uparrow 1.08}$	43.43$^{\uparrow 0.70}$
	25%	59.08$^{\downarrow 0.17}$	41.02$^{\uparrow 0.00}$	6.67$^{\downarrow 1.44}$	13.99$^{\uparrow 1.24}$	47.19$^{\downarrow 1.56}$	33.59$^{\downarrow 0.39}$	70.37$^{\downarrow 0.18}$	46.78$^{\downarrow 1.36}$	21.76$^{\downarrow 0.98}$	20.85$^{\uparrow 1.28}$	52.22$^{\downarrow 0.41}$	42.40$^{\downarrow 0.33}$
	100%	60.67$^{\uparrow 1.42}$	41.36$^{\uparrow 0.34}$	7.28$^{\downarrow 0.83}$	14.97$^{\uparrow 2.22}$	49.82$^{\uparrow 1.07}$	34.82$^{\uparrow 0.84}$	63.67$^{\downarrow 6.88}$	43.39$^{\downarrow 4.75}$	11.68$^{\downarrow 11.06}$	15.46$^{\downarrow 4.11}$	31.84$^{\downarrow 20.79}$	33.21$^{\downarrow 9.52}$
TokenCleaning	12.5%	58.73$^{\downarrow 0.52}$	**43.73**$^{\uparrow 2.71}$	7.43$^{\downarrow 0.68}$	12.84$^{\uparrow 0.09}$	46.37$^{\downarrow 2.38}$	33.82$^{\downarrow 0.16}$	71.60$^{\uparrow 1.05}$	**49.49**$^{\uparrow 1.35}$	21.00$^{\downarrow 1.74}$	22.73$^{\uparrow 3.16}$	53.03$^{\uparrow 0.40}$	43.57$^{\uparrow 0.84}$
	25%	59.96$^{\uparrow 0.71}$	42.37$^{\uparrow 1.35}$	5.99$^{\downarrow 2.12}$	13.62$^{\uparrow 0.87}$	46.14$^{\downarrow 2.61}$	33.62$^{\downarrow 0.36}$	70.19$^{\downarrow 0.36}$	47.80$^{\downarrow 0.34}$	19.94$^{\downarrow 2.80}$	24.17$^{\uparrow 4.60}$	52.02$^{\downarrow 0.61}$	42.82$^{\uparrow 0.09}$
	100%	61.20$^{\uparrow 1.95}$	41.02$^{\uparrow 0.00}$	7.73$^{\downarrow 0.38}$	15.41$^{\uparrow 2.66}$	48.29$^{\downarrow 0.46}$	34.73$^{\uparrow 0.75}$	**74.07**$^{\uparrow 3.52}$	46.44$^{\downarrow 1.70}$	24.79$^{\uparrow 2.05}$	33.26$^{\uparrow 13.69}$	54.32$^{\uparrow 1.69}$	46.58$^{\uparrow 3.85}$
Q-Tuning (Ours)	12.5%	**64.20**$^{\uparrow 4.95}$	42.03$^{\uparrow 1.01}$	**10.54**$^{\uparrow 2.43}$	**18.79**$^{\uparrow 6.04}$	**53.12**$^{\uparrow 4.37}$	**37.74**$^{\uparrow 3.76}$	71.60$^{\uparrow 1.05}$	48.14$^{\uparrow 0.00}$	**29.34**$^{\uparrow 6.60}$	27.75$^{\uparrow 8.18}$	**57.78**$^{\uparrow 5.15}$	**46.92**$^{\uparrow 4.19}$
	25%	63.14$^{\uparrow 3.89}$	42.03$^{\uparrow 1.01}$	8.87$^{\uparrow 0.76}$	16.76$^{\uparrow 4.01}$	51.52$^{\uparrow 2.77}$	36.47$^{\uparrow 2.49}$	71.78$^{\uparrow 1.23}$	47.12$^{\downarrow 1.02}$	26.08$^{\uparrow 3.34}$	**32.79**$^{\uparrow 13.22}$	56.17$^{\uparrow 3.54}$	46.79$^{\uparrow 4.06}$
Full Dataset		61.55	42.37	8.64	13.80	50.45	35.36	71.25	45.76	26.68	31.81	53.67	45.84

C.5 Further Ablation study

C.5.1 Ablation with different hyperparameters

Sensitivity of batch size in dynamic sample pruning.

Figure 9 presents whole ablation results on Mistral-7B, evaluating the effects of batch size ($8/16/32$) and neighbor awareness λ (0–1.0) under multiple keep-ratio settings across ARC-E, ARC-C, GSM8K, SQuAD, and TriviaQA, as well as the average over the five benchmarks.

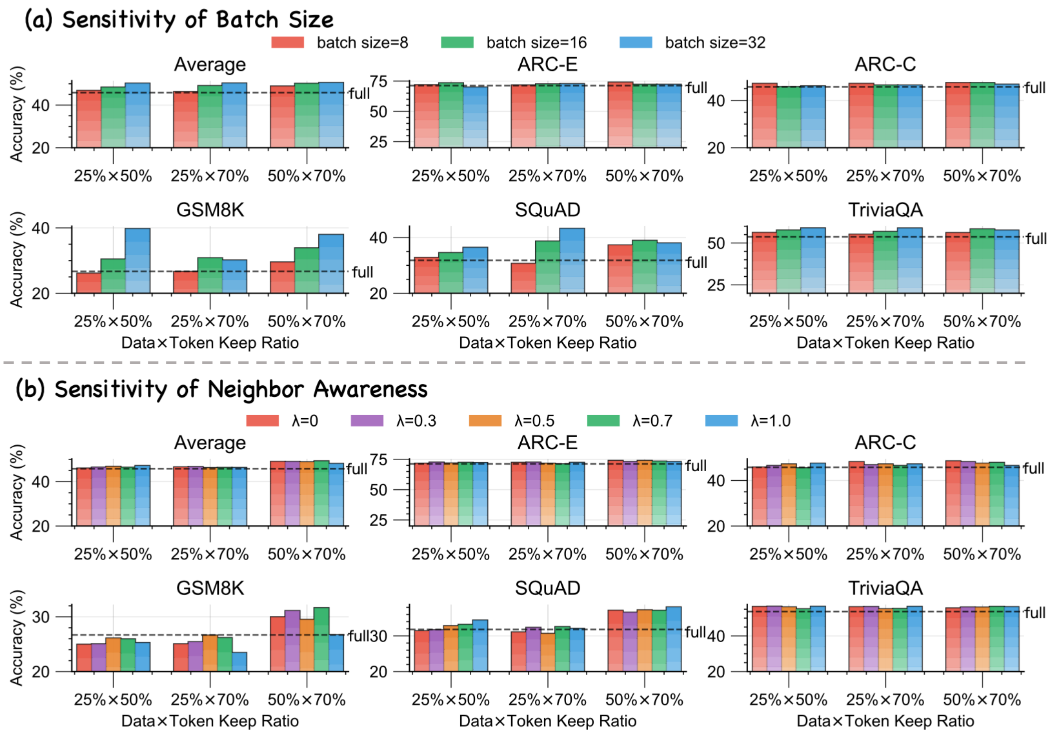

Figure 9: Effect of varying (a) batch size (8, 16, 32) and (b) neighbor awareness λ $(0, 0.3, 0.5, 0.7, 1.0)$ for Mistral-7B under three data–token keep ratio configurations ($25\%\times50\%$, $25\%\times70\%$, $50\%\times70\%$). Dashed lines marked "full" denote models trained on the full data without pruning.

C.5.2 Additional ablation across different quadrants

Table 15 presents additional quadrant-wise ablation results on the Wizard dataset, evaluated on LLaMA2-7B and Mistral-7B, examining how pruning Q1–Q4 at both the sample and token levels impacts downstream performance.

Table 15: Ablation on four quadrants under different sample and token ratios, where ✓indicates that the corresponding quadrant is pruned.

Sample pruning Q1	Sample pruning Q3	Token pruning Q2	Token pruning Q4	LLaMA2-7B ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.	Mistral-7B ARC-E	ARC-C	GSM8K	SQuAD	TriviaQA	Avg.
12.5% *Samples,* 50% *Tokens*															
	✓	✓		60.14	41.02	8.04	13.11	50.23	34.51	71.43	47.80	28.13	**31.09**	54.27	46.54
✓	✓	✓	✓	64.02	41.02	9.86	**19.26**	52.00	37.23	71.60	**48.47**	28.89	26.46	**57.88**	46.66
✓	✓	✓		**64.20**	**42.03**	**10.54**	18.79	**53.12**	**37.74**	**71.60**	48.14	**29.34**	27.75	57.78	**46.92**
25% *Samples,* 50% *Tokens*															
	✓	✓		59.61	41.02	7.51	13.74	50.25	34.43	71.25	46.10	**29.04**	31.55	54.84	46.56
✓	✓	✓	✓	61.02	40.68	**9.10**	**18.66**	**51.68**	36.23	70.72	47.12	28.58	31.85	**56.33**	**46.92**
✓	✓	✓		**63.14**	**42.03**	8.87	16.76	51.52	**36.47**	**71.78**	**47.12**	26.08	**32.79**	56.17	46.79

D PSEUDOCODE OF THE PROPOSED METHOD

To facilitate clarity and reproducibility, we summarize the proposed Q-Tuning algorithm in Algorithm 1. The procedure unfolds within a single training iteration and consists of two tightly coupled stages: (i) *sample pruning*, where instances are dynamically selected based on their position in the error–uncertainty (EU) plane through an efficient bisection-based search of quantile thresholds, and (ii) *token pruning*, where retained samples undergo finer-grained filtering to preserve only the most informative subset of tokens.

Algorithm 1 Q-Tuning: dynamic data pruning in one iteration

Input: Mini-batch $\mathcal{B}_t$, model f_{θ_t}, retention ratios $r_{\text{sample}}, r_{\text{token}}$, smoothing λ.
Output: Pruned mini-batch $\tilde{\mathcal{B}}_t$.
// Stage 1: Sample Pruning via EU Plane
Compute $(\mathrm{PPL}(x, y; f_\theta), \mathrm{Ent}(x, y; f_\theta))$ for each $x \in \mathcal{B}_t$.
Initialize ranges $\alpha_{\text{low}} = 0$, $\alpha_{\text{high}} = 0.49$, $\beta_{\text{low}} = 0$, $\beta_{\text{high}} = 0.49$.
for $k = 1$ to $K_{\max}$ **do** ▷ Bisection iterations on both axes
    $\alpha \leftarrow (\alpha_{\text{low}} + \alpha_{\text{high}})/2$, $\beta \leftarrow (\beta_{\text{low}} + \beta_{\text{high}})/2$.
    Derive thresholds $ppl_{\text{hi}} = Q_{1-\alpha}(\mathrm{PPL})$, $ppl_{\text{lo}} = Q_{\alpha}(\mathrm{PPL})$, $ent_{\text{lo}} = Q_{\beta}(\mathrm{Ent})$, $ent_{\text{hi}} = Q_{1-\beta}(\mathrm{Ent})$.
    Partition samples into quadrants Q_1–Q_4.
    $r \leftarrow \frac{|Q_2|+|Q_4|}{|\mathcal{B}_t|}$.
    if $r < r_{\text{sample}}$ **then**
        $\alpha_{\text{low}} \leftarrow \alpha,\ \beta_{\text{low}} \leftarrow \beta$ ▷ Too few kept, relax thresholds
    else
        $\alpha_{\text{high}} \leftarrow \alpha,\ \beta_{\text{high}} \leftarrow \beta$ ▷ Too many kept, tighten thresholds
    end if
end for
$\mathcal{B}'_t \leftarrow Q_2 \cup Q_4$.
if $|\mathcal{B}'_t| < r_{\text{sample}} \cdot |\mathcal{B}_t|$ **then** ▷ Ensure target sample ratio
    Compute **supp-score** for each $x \in Q_1 \cup Q_3$: $\mathrm{supp}(x) = \max\{\hat{\mathrm{PPL}}(x, y; f_\theta) - \hat{\mathrm{Ent}}(x, y; f_\theta),\ \hat{\mathrm{Ent}}(x, y; f_\theta) - \hat{\mathrm{PPL}}(x, y; f_\theta)\}$, where $\hat{\cdot}$ denotes min–max normalization.
    Select top-scoring samples to augment $\mathcal{B}'_t$ until target size is met.
end if
// Stage 2: Token Pruning
$\tilde{\mathcal{B}}_t \leftarrow \emptyset$.
for each $x \in \mathcal{B}'_t$ **do**
    if $x \in Q_4$ **then**
        Keep all tokens of x. ▷ Calibration samples intact
    else if $x \in Q_2$ **then**
        For each token $i \in T(x)$, compute

$$s_i(x, y; f_\theta) = (1 - \lambda)\,\mathrm{PPL}_i(x, y; f_\theta) + \lambda\,[\mathrm{PPL}_{i-1}(x, y; f_\theta) + \mathrm{PPL}_{i+1}(x, y; f_\theta)].$$

        Keep top-r_{token} fraction of tokens ranked by s_i.
    end if
    $\tilde{\mathcal{B}}_t \leftarrow \tilde{\mathcal{B}}_t \cup \{m(x) \odot x\}$.
end for
return $\tilde{\mathcal{B}}_t$.

E Further Case Study

E.1 Example Samples from Each Quadrant

We report representative samples from each quadrant (Q1–Q4) at the beginning of training, including their perplexity (PPL), entropy, and the corresponding instruction–output pairs.

PPL	Entropy	Example (Instruction / Output)
Q1 Examples		
8.33	2.12	`Instruction: What are the three primary colors? Output: The three primary colors are red, blue, and yellow.`
8.04	1.73	`Instruction: How can we reduce air pollution? Output: There are a number of ways to reduce air pollution, such as shifting to renewable energy sources...`
91.24	1.91	`Instruction: Explain why the fraction 4/16 is equivalent to 1/4. Output: The fraction 4/16 is equivalent to 1/4 because both numerator and denominator are divisible by 4.`
Q2 Examples		
8.65	1.54	`Instruction: Write a short story in third person narration about a protagonist making an important career decision. Output: John was at a crossroads in his life. He had just graduate...`
11.02	1.48	`Instruction: Write a short story in third person narration about a protagonist making an important career decision. Output: John was at a crossroads in his life. He had just graduate...`
15.74	1.50	`Instruction: Identify the odd one out.\nTwitter, Instagram, Telegram Output: Telegram`
Q3 Examples		
4.80	1.45	`Instruction: Describe the structure of an atom. Output: An atom is made up of a nucleus containing protons and neutrons, surrounded by electrons...`
2.41	1.09	`Instruction: How can we reduce air pollution? Output: There are a number of ways to reduce air pollution, such as shifting to renewable energy sources...`
4.73	1.49	`Instruction: Identify the odd one out (Twitter, Instagram, Telegram). Output: Telegram`
Q4 Examples		
6.03	1.83	`Instruction: Give three tips for staying healthy. Output: 1. Eat a balanced diet... 2. Exercise regularly...`
8.04	1.71	`Instruction: What are the three primary colors? Output: The three primary colors are red, blue, and yellow.`
5.37	1.79	`Instruction: Write a short story in third person narration about a protagonist making an important career decision. Output: John was at a crossroads in his life. He had just graduate...`

E.2 PRUNED CASE

We present representative token-pruning cases from MATHINSTRUCT using the MISTRAL-7B model. For each example, we show the model's original `output` and highlight in red the fine-grained segments that are *absent* in the corresponding `pruned_output`. This visualization makes the pruning behavior transparent at the character/substring level and helps diagnose which semantic components are preferentially removed (e.g., boilerplate, redundant explanations, or unstable numeric derivations).

Unless otherwise stated, we keep the original formatting and line breaks of the model outputs to preserve exact alignment. The highlighting is produced by an LCS-based diff between `output` and `pruned_output`, ensuring that only truly deleted spans are marked while retained text remains uncolored. We do *not* introduce any additional edits beyond necessary LaTeX escaping.

These qualitative cases complement the aggregate statistics reported in the main text by revealing typical pruning patterns on mathematical reasoning tasks (definitions and restatements vs. computational steps), and by surfacing edge cases where pruning removes numerically relevant content. Together, they provide an interpretable view of how token-level pruning affects the fidelity and brevity of generated solutions.

SAMPLE 0

Instruction sandy is younger than molly by 20 years . if the ratio of their ages is 7 : 9 , how old is sandy ? Let's write a program.

Output

```
n0 = 20.0
n1 = 7.0
n2 = 9.0
t0 = n2 - n1\n
t1 = n1 / t0
answer =n0 * t1
print(answer)
```

SAMPLE 1

Instruction A manufacturing process produces 5000 units of a particular product. The probability of a certain type of defect occurring in a unit is 0.02. What is the probability that exactly 100 units will have this defect? **Output (deletions highlighted)**

```
To solve this problem, we can use the binomial probability
formula, which is:
P(X = k) = (nCk) * (p^k) * (1-p)^(n-k)
where:
- P(X = k) is the probability of exactly k successes (defective
units in this case) in n trials (total units produced)
- nCk (read as "n choose k") is the number of combinations of
choosing k successes from n trials, calculated as n! / (k! *
(n-k)!)
- p is the probability of success (defect occurring) in a
single trial
- 1-p is the probability of failure (no defect occurring) in a
single trial
- n is the total number of trials (units produced)
- k is the number of successes (defective units) we want to
find the probability for
In this problem, we have:
- n =5000 (total units produced)
- k = 100 (defective units)
```

```
- p = 0.02 (probability of a defect occurring in a unit)
Now, we can plug these values into the binomial probability
formula:
P(X=100)=(5000C100)· (0.02)^100· (0.98)^4900)
First, let's calculate 5000C100:
5000C100 = 5000! / (100! * 4900!)
= 2.031759e+134 / (9.332622e+157 * 1.929713e+126)
= 2.031759e+134 / (1.801713e+284)
= 1.127271e-150
Now, let's calculate 0.02^100 and 0.98^4900:
0.02^100 = 1.267651e-70
0.98^4900 = 2.372472e-44
Finally, let's plug these values back into the formula:
P(X = 100) = 1.127271e-150 * 1.267651e-70 * 2.372472e-44
= 3.394e-264
The probability that exactly 100 units will have this defect is
approximately 3.394e-264, which is an extremely small
probability.
```

F USE OF LLMS

This work did not involve the use of large language models (LLMs) at any stage. The design of experiments, data analysis, and manuscript preparation were conducted entirely by the authors through conventional computational methods and human expertise, without reliance on automated text generation or model-driven reasoning.