ULTRASPARSE: EXTREME COMPRESSION OF UPCYCLED MOE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture-of-Experts (MoE) has become a promising architecture for scaling large models through sparse activation. In particular, upcycling dense models into MoE models has emerged as an effective approach for constructing powerful sparse models, motivating the need for effective compression techniques for upcycled MoE models. Recent studies reveal that upcycled MoE models exhibit a structure of expert-shared base weights and expert-specific delta weights, where the delta weights contain substantial redundancy and compression potential. However, how to achieve extreme compression of expert-specific delta weights while preserving model capability remains largely unexplored. In this work, we conduct the first systematic study on the extreme sparsification of expert-specific delta weights and reveal that both weight distribution and knowledge distribution play critical roles in governing model performance under ultra-high sparsity regimes. Building upon these insights, we present UltraSparse, the first framework for extreme compression of upcycled MoE models. A Weight Distribution Preservation module is proposed to maintain consistency between the weight distributions before and after sparsification. A Knowledge Distribution-Guided Hybrid Sparsification module is proposed to preserve critical knowledge associated with high-magnitude weights while retaining complementary knowledge contributed by low-magnitude weights. To comprehensively evaluate compression methods for upcycled MoE models, we establish a large-scale benchmark covering diverse upcycling approaches, model architectures of different scales, and datasets spanning different tasks and modalities. Extensive experiments demonstrate that UltraSparse consistently outperforms existing compression methods across various settings, and achieves unprecedented compression ratios, reaching up to 1000$\times$ compression while maintaining strong model performance.

1 INTRODUCTION

Upcycling dense models into Mixture-of-Experts (MoE) models has recently emerged as an effective paradigm for constructing powerful sparse models (Komatsuzaki et al., 2023; Nakamura et al., 2025; Li et al., 2025; Liao et al., 2025; Jiang et al., 2024; Lin et al., 2026). By leveraging pretrained dense models as initialization, upcycling avoids the expensive training process of MoE models from scratch while maintaining strong model capability (Cao et al., 2025; Komatsuzaki et al., 2023; Zhu et al., 2024; Hui et al., 2025). Recent studies show that upcycled MoE models can be decomposed into expert-shared base weights and expert-specific delta weights (Ilharco et al., 2023; Huang et al., 2024), where the delta weights capture expert-specific adaptations on top of the shared base weights (Huang et al., 2025b). Unlike directly compressing full expert parameters (Frantar & Alistarh, 2024; Lu et al., 2024; Yang et al., 2024), this decomposition separates expert-specific information into delta weights (Ilharco et al., 2023; Ye et al., 2025), making them a natural target for reducing the parameter overhead of MoEs.

Despite this unique structure, the compression limit of expert-specific delta weights remains largely unexplored. Existing approaches primarily improve delta compression through parameter selection or parameter representation techniques, including sparsification and quantization (Yu et al., 2024a; Liu et al., 2024b; Huang et al., 2025b;a; Wang et al., 2025b). Although these methods substantially reduce the parameter overhead, model capability can degrade rapidly under ultra-high compression ratios (Deng et al., 2025; Wang et al., 2025b), as also shown in Figure 1. This phenomenon indicates that extreme compression is not simply a matter of selecting fewer parameters or representing them

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

more efficiently. Instead, it raises a fundamental question: what information encoded in expert-specific delta weights determines the compression limit, and how can such information be preserved under extreme compression?

To answer this question, we conduct the first systematic study on the extreme sparsification of expert-specific delta weights in upcycled MoE models. Our analysis uncovers two complementary factors governing the compression-performance trade-off. First, preserving the weight distribution is crucial for maintaining the characteristics of expert-specific delta weights, as distribution shifts could weaken expert performance. Second, preserving the knowledge distribution is equally important, as expert-specific delta weights learned from different tasks exhibit distinct knowledge patterns, where high- and low-magnitude weights jointly contribute complementary expert knowledge. Together, these findings indicate that the compression limit of expert-specific delta weights is governed by preserving both weight distribution and knowledge distribution, rather than solely by parameter selection or representation.

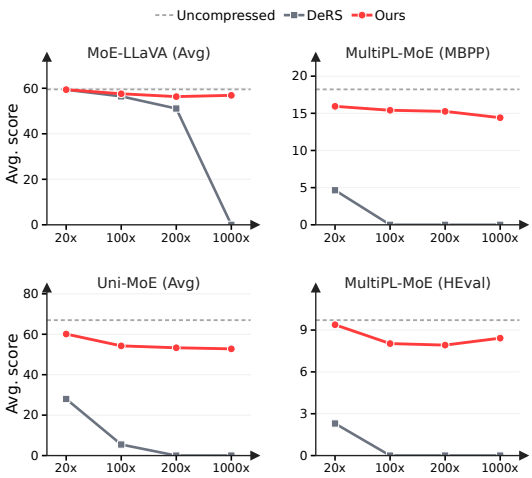


Figure 1: Comparison of the average score of our UltraSparse (red) and DeRS-SP (Huang et al., 2025b) (gray) on four experimental settings at sparsity ratios 0.95, 0.99, 0.995, and 0.999 (corresponding to $20\times$, $100\times$, $200\times$, and $1000\times$ expert-delta compression).

Motivated by these findings, we propose UltraSparse, the first framework for extreme compression of upcycled MoE models. UltraSparse builds upon these findings through two complementary components. A Weight Distribution Preservation module explicitly maintains consistency between the original and sparse weight distributions, mitigating performance degradation caused by distribution shifts. A Knowledge Distribution-Guided Hybrid Sparsification module jointly preserves critical knowledge associated with high-magnitude weights while retaining complementary knowledge contributed by low-magnitude weights. By explicitly preserving both weight distribution and knowledge distribution during compression, UltraSparse enables substantially higher compression ratios with minimal degradation in expert capability, as shown in Figure 1.

To facilitate systematic research on upcycled MoE compression, we establish a comprehensive benchmark covering diverse upcycling approaches, model architectures of different scales, and datasets spanning multiple tasks and modalities. Extensive experiments demonstrate that UltraSparse consistently outperforms existing compression methods across a wide range of settings, and achieves unprecedented compression ratios of up to $1000\times$ while maintaining strong model performance. Beyond the empirical improvements, our study provides new insights into the compression limit of delta weights and establishes distribution preservation as a fundamental principle for achieving extreme compression of upcycled MoE models.

Our contributions are summarized as follows:

- Building upon recent observations that upcycled MoE models can be decomposed into expert-shared base weights and expert-specific delta weights and that delta weights exhibit substantial redundancy and compression potential, we conduct the first systematic study on the extreme sparsification of expert-specific delta weights within upcycled MoE models. Our analysis reveals that both weight distribution and knowledge distribution play critical roles in governing model performance under extreme sparsity regimes.
- We propose UltraSparse, the first framework for achieving extreme compression of upcycled MoE models with extremely high sparsity ratios. Specifically, UltraSparse introduces a Weight Distribution Preservation module that explicitly maintains consistency between the weight distributions before and after sparsification. Furthermore, we develop a Knowledge Distribution-Guided Hybrid Sparsification module that preserves critical knowledge associated with high-magnitude weights while retaining complementary knowledge contributed by low-magnitude weights, enabling aggressive compression with minimal performance degradation.

- We establish a comprehensive benchmark for evaluating upcycled MoE compression methods, covering 5 upcycling approaches, 7 models with diverse architectures and scales, and 30 datasets spanning different tasks and modalities. Extensive experiments demonstrate that UltraSparse consistently outperforms existing compression methods across various settings and achieves unprecedented compression ratios, reaching up to $1000\times$ compression while maintaining strong model performance.

2 Related Work

Upcycled MoE models convert a pretrained dense model into a sparse MoE by replicating each MLP layer as multiple experts and learning a router together with the experts (Komatsuzaki et al., 2023). By reusing the knowledge in a dense checkpoint, they reach competitive performance with substantially less training than constructing an MoE from scratch. Recent work improves expert specialization through partial re-initialization (Nakamura et al., 2025) and extends upcycling to different fields including code generation (Ding et al., 2024; Wang et al., 2025a), general multimodal understanding (Lin et al., 2026; Li et al., 2025), and medical vision-language modeling (Jiang et al., 2024). These approaches cover both experts cloned from a common MLP layer from a pre-trained model, and experts specialized on different domains or modalities, leading to markedly different expert-delta statistics and significant potential for compression. In this paper, we leverage this characteristic of upcycled MoE models and realize an extremely high compression ratio.

Compression of MoE Models has been studied to reduce the storage and deployment costs of MoEs due to their significantly higher parameter counts compared to dense models. General methods target generic MoEs and typically quantize entire expert matrices, remove experts, or alter their effective routing capacity. QMoE co-designs sub-bit quantization and execution kernels for trillion-parameter MoEs (Frantar & Alistarh, 2024); ExpertPrune prunes and dynamically skips low-importance experts (Lu et al., 2024), while MoE-I^2 combines inter-expert pruning with intra-expert low-rank decomposition (Yang et al., 2024). However, these methods focus on common MoE models without utilizing the extreme delta redundancy characteristics of upcycled MoE models, which were found in dense models following the pretrain-finetune paradigm (Deng et al., 2025; Liu et al., 2024b; Wang et al., 2025b). Recently, DeRS (Huang et al., 2025b) exploits the lineage of an upcycled MoE by decomposing each expert into an expert-shared base weight and an expert-specific delta weight, then compresses the delta weights through independent random sparsification or quantization. However, independent compression neither controls the realized proportions retained from different magnitude ranges nor guarantees the survival of the largest delta entries. To tackle these issues, UltraSparse explicitly targets this extreme-sparsity regime through a weight distribution preservation module and a knowledge distribution-guided hybrid sparsification module.

3 Methodology

3.1 Preliminaries

Consider an upcycled MoE model (Komatsuzaki et al., 2023) with L MoE layers, E experts per layer, and P linear modules per expert. For layer $l \in \{1, \dots, L\}$, expert $e \in \{1, \dots, E\}$, and module type $p \in \{1, \dots, P\}$, such as the gate, up, and down projections of an FFN expert, we denote the trained expert weight by $\mathbf{W}^{\text{moe}}_{l,e,p} \in \mathbb{R}^{R_p \times C_p}$, where R_p and C_p are the output and input dimensions of the p-th module. Since all experts of a layer are initialized from the corresponding FFN weight $\mathbf{W}^{\text{base}}_{l,p} \in \mathbb{R}^{R_p \times C_p}$ of the pretrained dense model, the trained weight can be decomposed as (Huang et al., 2025b):

$$\mathbf{W}^{\text{moe}}_{l,e,p} = \mathbf{W}^{\text{base}}_{l,p} + \mathbf{\Delta}_{l,e,p}, \tag{1}$$

where $\mathbf{W}^{\text{base}}_{l,p}$ serves as the expert-shared base weight and $\mathbf{\Delta}_{l,e,p} \in \mathbb{R}^{R_p \times C_p}$ is the expert-specific delta. Throughout this section, bold symbols denote matrices or tensors, and scalar functions applied to them act element-wise unless otherwise stated. We use $\mathbb{1}[\cdot]$ for the indicator function, which equals one when the condition holds and zero otherwise. Delta compression keeps a single copy of the base weights and compresses only the deltas, constructing an expert's weight when needed. In particular, random sparsification (Yu et al., 2024a; Huang et al., 2025b) independently zeroes each element of $\mathbf{\Delta}_{l,e,p}$ with probability θ^* and rescales the kept elements by $1/(1-\theta^*)$ to keep the expectation unbiased. Since the identical procedure is applied to every delta independently, we drop the subscript (l, e, p) below and write $\mathbf{\Delta} \in \mathbb{R}^{R \times C}$, where R and C are the row and column dimensions R_p and C_p of this module and $N = R \cdot C$ is its total number of elements.

3.2 WEIGHT DISTRIBUTION PRESERVATION MODULE

Random sparsification gives every delta element the same keep probability, but it does not control how many elements are kept from each magnitude range in one sparse delta. At extreme sparsity, random changes in these counts can make the magnitude histogram differ greatly from the original one, especially for bins with few elements. Given the target sparsity ratio θ^*, we denote the total keep ratio by $k^* = 1 - \theta^*$ and the target keep count by $K^* = \mathrm{round}(k^* \cdot N)$. We represent the weight magnitudes using a histogram of $|\mathbf{\Delta}|$ over B equal-width intervals. Based on this description, we introduce **W**eight **D**istribution **P**reservation (WDP) as an operator $\mathrm{WDP}(\mathbf{\Delta}, \mathcal{A}, K)$ that controls how many positions are kept in each magnitude interval. Here, $\mathcal{A} \subseteq \Omega = \{1, \ldots, R\} \times \{1, \ldots, C\}$ is a set of candidate positions and $K \leq |\mathcal{A}|$ is the target number of positions to keep. WDP preserves how the kept positions are distributed across the original magnitude bins. It does not preserve the value distribution after the sparse delta is rescaled. For a non-empty candidate set, let $\delta^{\mathcal{A}}_{\min}$ and $\delta^{\mathcal{A}}_{\max}$ be the smallest and largest values of $|\Delta_{uv}|$ over $(u, v) \in \mathcal{A}$. We divide this value range into B equal-width bins:

$$\begin{aligned} \mathcal{B}_j(\mathcal{A}) &= \{(u, v) \in \mathcal{A} : b_{j-1}(\mathcal{A}) \leq |\Delta_{uv}| < b_j(\mathcal{A})\}, \\ b_j(\mathcal{A}) &= \delta^{\mathcal{A}}_{\min} + \frac{j}{B}\left(\delta^{\mathcal{A}}_{\max} - \delta^{\mathcal{A}}_{\min}\right), \end{aligned} \tag{2}$$

for $j \in \{1, \ldots, B\}$. The first bin includes the minimum value $\delta^{\mathcal{A}}_{\min}$. For the final bin, replace the right inequality by $|\Delta_{uv}| \leq b_B(\mathcal{A})$ so that it includes the maximum value. These bins have equal width but may contain different numbers of positions. We also compare other binning methods in the experiments.

Let $n_j = |\mathcal{B}_j(\mathcal{A})|$ be the number of positions in bin j. We obtain the integer counts as

$$q_j = K\frac{n_j}{|\mathcal{A}|}, \qquad K_j = \lfloor q_j \rfloor + \mathbb{1}[j \in \mathcal{H}]. \tag{3}$$

Following standard largest-remainder rounding (Lauwers & Puyenbroeck, 2006; Sathyavageeswaran, 2026), $\mathcal{H}$ contains the bins with the largest fractional remainders $q_j - \lfloor q_j \rfloor$. Exactly $K - \sum_{j=1}^{B} \lfloor q_j \rfloor$ bins are selected for $\mathcal{H}$, with ties broken randomly. For each bin, we randomly select K_j positions without replacement to form a subset $\mathcal{C}_j \subseteq \mathcal{B}_j(\mathcal{A})$. The resulting mask is

$$[\mathrm{WDP}(\mathbf{\Delta}, \mathcal{A}, K)]_{uv} = \mathbb{1}\left[(u, v) \in \bigcup_{j=1}^{B} \mathcal{C}_j\right]. \tag{4}$$

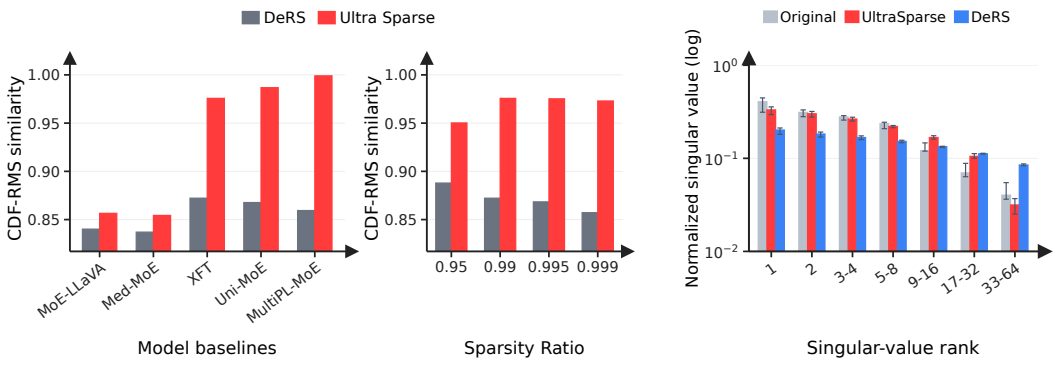
This procedure keeps exactly the target number of positions and makes the bin fractions close to their original fractions. Compared with Bernoulli sampling, fixed bin counts reduce random errors in the fractions of kept positions (Neyman, 1934), but they cannot remove errors from integer rounding. These errors become more noticeable at extreme sparsity when only a few positions are kept across many bins. When all matrix positions are used as candidates, WDP produces the compressed delta:

$$\widehat{\mathbf{\Delta}} = \frac{1}{k^*}\,\mathrm{WDP}(\mathbf{\Delta}, \Omega, K^*) \odot \mathbf{\Delta}. \tag{5}$$

Figure 2a demonstrates that UltraSparse consistently achieves higher CDF-RMS similarity between original and sparse expert deltas than DeRS across diverse MoE baselines and sparsity levels. This indicates that UltraSparse effectively preserves the underlying weight distribution under ultra-high sparsity, validating weight distribution preservation as a key factor for maintaining expert-specific information during extreme compression.

3.3 KNOWLEDGE DISTRIBUTION-GUIDED HYBRID SPARSIFICATION MODULE

The weight distribution preservation module controls how many elements are kept in each magnitude range, but it samples every position within a bin in the same way and cannot guarantee that the largest elements are kept. The information encoded by delta weights is not spread evenly across their magnitudes. A small number of high-magnitude elements make large individual contributions and encode important information (Han et al., 2015), whereas many non-top elements together preserve additional information across the remaining magnitude range. Therefore, WDP may still remove large-magnitude elements, while Top-only sparsification removes all elements outside the Top set.



(a) (b)

Figure 2: (a) Conditional-nonzero CDF-RMS (Anderson, 1962; Bellemare et al., 2017) similarity between original and sparse expert deltas. We compare parameter-weighted discrete CDFs between random sparsification (gray) and UltraSparse (red) within the gate, up, and down projections and average their scores. Note that higher denotes better. (b) Frobenius-normalized singular values of Uni-MoE expert deltas ($100\times$ scaled), binned by rank. Bars represent the mean across non-zero delta blocks, with error bars indicating the interquartile range (IQR). UltraSparse preserves the original spectrum significantly better than DeRS-SP.

Based on this, UltraSparse extends WDP by dividing the total keep ratio between a fixed **Top** branch and a random **Coverage** branch. We refer to this Top–Coverage extension as WDP+:

$$k^{\text{top}} + k^{\text{cov}} = k^* = 1 - \theta^*, \tag{6}$$

where both branch ratios are non-negative and do not exceed the total keep ratio. In WDP+, the Top branch is deterministic, and WDP is applied only to the remaining positions. The Top branch always keeps the $K^{\text{top}} = \text{round}(k^{\text{top}} \cdot N)$ largest-magnitude (Han et al., 2015) elements:

$$\mathcal{T} = \text{TopKIdx}\big(|\mathbf{\Delta}|, K^{\text{top}}\big)\,, \qquad M^{\text{top}}_{uv} = \mathbb{1}[(u,v) \in \mathcal{T}]\,. \tag{7}$$

Let $\mathcal{R} = \Omega \setminus \mathcal{T}$ be the remaining positions and set $K^{\text{cov}} = K^* - K^{\text{top}}$ so that the two branch counts add up to the target count. The Coverage branch applies WDP to this set:

$$\mathbf{M}^{\text{cov}} = \text{WDP}(\mathbf{\Delta}, \mathcal{R}, K^{\text{cov}}), \tag{8}$$

where WDP creates equal-width bins over the magnitude range of the remaining positions and makes the keep rate of each bin close to the average keep rate. The Top and Coverage masks do not overlap, and WDP+ produces the compressed delta as

$$\widetilde{\mathbf{\Delta}} = \mathbf{M}^{\text{top}} \odot \mathbf{\Delta} + \frac{1}{k^*} \cdot \mathbf{M}^{\text{cov}} \odot \mathbf{\Delta}, \tag{9}$$

where $\odot$ denotes element-wise multiplication.

Following prior work that uses a high-precision scale factor to reduce the difference between the original and compressed deltas (Liu et al., 2024b), WDP+ leaves the Top elements unchanged, giving them a scaling factor of one. For the Coverage elements, WDP+ uses the inverse global keep rate $1/k^*$, which includes the budgets assigned to both branches, rather than the inverse of the smaller Coverage keep rate over the remaining positions. Thus, when $k^{\text{top}} > 0$, the Coverage scale $1/k^*$ is already biased, independent of integer rounding. Integer allocation only adds small differences across bins. The smaller global factor avoids much larger scaling at extreme sparsity. We provide a formal analysis of the bias and variance introduced by the global scaling used in WDP+, together with the main properties of its fixed-budget bin allocation, in the Appendix. Each compressed expert is then built as

$$\widetilde{\mathbf{W}}_{l,e,p} = \mathbf{W}^{\text{base}}_{l,p} + \widetilde{\mathbf{\Delta}}_{l,e,p}. \tag{10}$$

When no budget is assigned to Top, WDP+ reduces to WDP. If the Coverage branch further removes binning and uses independent Bernoulli sampling (Huang et al., 2025b), this case becomes standard drop-and-rescale (Yu et al., 2024a). When no budget is assigned to Coverage, the method becomes magnitude pruning (Han et al., 2015). When both branches are used, WDP+ keeps important information encoded by the largest elements and additional information spread across the non-top magnitude range. **An overview of UltraSparse is illustrated in Figure 4 in the Appendix.**

Algorithm 1 The proposed UltraSparse

Input: expert delta $\mathbf{\Delta} \in \mathbb{R}^{R\times C}$; all positions $\Omega = \{1,\dots,R\} \times \{1,\dots,C\}$ with $N = |\Omega| = RC$; total keep ratio k^*; Top keep ratio k^{top}; number of magnitude bins B
Output: compressed delta $\widetilde{\mathbf{\Delta}}$

Knowledge Distribution-Guided Hybrid Sparsification Module
$K^* \leftarrow \text{round}(k^* \cdot N)$, $K^{\text{top}} \leftarrow \text{round}(k^{\text{top}} \cdot N)$
$\mathcal{T} \leftarrow \text{TopKIdx}(|\mathbf{\Delta}|, K^{\text{top}})$, $\mathbf{M}^{\text{top}}_{uv} \leftarrow \mathbb{1}[(u,v) \in \mathcal{T}]$
$\mathcal{R} \leftarrow \Omega \setminus \mathcal{T}$, $K^{\text{cov}} \leftarrow K^* - K^{\text{top}}$, $\mathbf{M}^{\text{cov}} \leftarrow \mathbf{0}$
if $K^{\text{cov}} > 0$ **then**
  Weight Distribution Preservation Module
  Get $\delta^{\mathcal{R}}_{\min}, \delta^{\mathcal{R}}_{\max}$ on $\mathcal{R}$, partition $\mathcal{R}$ into B bins by Eq. (2)
  for $j = 1,\dots,B$ **do**
    $n_j \leftarrow |\mathcal{B}_j(\mathcal{R})|$, $q_j \leftarrow K^{\text{cov}} \cdot n_j / |\mathcal{R}|$
  end for
  Set K_j from $\{q_j\}$ by Eq. (3)
  for $j = 1,\dots,B$ **do**
    $\mathcal{C}_j \leftarrow$ random size-K_j subset of $\mathcal{B}_j(\mathcal{R})$
    for all $(u,v) \in \mathcal{C}_j$ **do**
      $\mathbf{M}^{\text{cov}}_{uv} \leftarrow 1$
    end for
  end for
end if
return $\widetilde{\mathbf{\Delta}} \leftarrow \mathbf{M}^{\text{top}} \odot \mathbf{\Delta} + \frac{1}{k^*}\mathbf{M}^{\text{cov}} \odot \mathbf{\Delta}$

Figure 2b demonstrates that UltraSparse more faithfully preserves the singular value spectrum of original expert deltas than DeRS-SP across different spectral ranks. This indicates that UltraSparse better maintains the information distribution encoded in expert-specific delta weights, preserving complementary contributions from both dominant and less-dominant components.

4 Experiments

The proposed UltraSparse is shown in Alg. 1. We evaluate it on various tasks, covering MoE-LLaVA (Lin et al., 2026), Uni-MoE (Li et al., 2025), Med-MoE (Jiang et al., 2024), $\mathcal{X}$FT (Ding et al., 2024), and MultiPL-MoE (Wang et al., 2025a). Across expert-delta compression ratios from $20\times$ to $1000\times$, UltraSparse consistently outperforms DeRS-SP, with the largest advantages under extreme compression.

Table 1: Performance comparisons on two MoE-LLaVA models upcycled from Phi and Qwen models. Note that Ratio denotes the compression ratio applied to expert parameters.

Method	Ratio	**MoE-LLaVA** *(Phi as Dense)* **4.50B** (Lin et al., 2026)								**MoE-LLaVA** *(Qwen as Dense)* **3.48B** (Lin et al., 2026)							
		SQA^{I}	**VQA^{T}**	**GQA**	**POPE**	**MM-V**	**MMB^{EN}**	**VQA$^{\text{v2}}$**	**Avg.**	**SQA^{I}**	**VQA^{T}**	**GQA**	**POPE**	**MM-V**	**MMB^{EN}**	**VQA$^{\text{v2}}$**	**Avg.**
Vanilla	1x	68.96	51.16	61.42	87.23	32.10	63.74	77.58	63.17	63.16	47.74	61.50	87.81	25.50	54.70	76.18	59.51
DeRS-QT	8x	73.54	51.34	60.92	87.62	32.20	63.27	77.49	63.77	63.78	47.75	61.07	87.80	25.60	54.75	75.91	59.52
DeRS-QT	16x	74.04	50.35	58.94	87.06	32.00	60.99	75.83	62.74	56.64	44.06	56.81	87.24	25.20	51.43	72.75	56.30
DeRS-SP	20x	68.91	51.22	61.38	87.19	34.40	63.38	77.49	63.42	62.17	47.09	61.13	87.74	27.70	53.92	75.73	59.35
DeRS-SP	100x	68.32	50.37	61.18	87.41	34.40	61.19	77.15	62.86	57.96	43.13	58.63	87.46	24.80	49.30	73.79	56.44
DeRS-SP	200x	67.08	49.58	60.24	86.82	32.10	59.90	76.53	61.75	53.00	35.22	54.52	85.99	24.60	35.01	69.36	51.10
DeRS-SP	1000x	69.21	49.79	58.46	86.86	32.10	60.73	75.61	61.82	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UltraSparse	20x	69.31	51.39	61.40	87.34	34.40	63.53	77.58	63.56	62.12	47.67	60.84	88.07	27.50	54.03	75.51	59.39
UltraSparse	100x	70.00	50.54	60.83	87.49	34.90	62.13	77.19	63.30	62.37	45.12	58.96	87.89	25.80	48.73	74.49	57.62
UltraSparse	200x	69.26	51.10	60.95	87.65	32.80	61.92	76.73	62.92	59.99	42.67	57.69	87.64	25.80	48.16	72.33	56.33
UltraSparse	1000x	69.16	49.40	58.46	87.32	33.40	60.36	75.65	61.96	61.97	43.17	55.96	87.07	26.40	51.53	72.16	56.89

4.1 General Multi-modal Task

4.1.1 Settings

We evaluate on two upcycled MoE frameworks, MoE-LLaVA (Lin et al., 2026) and Uni-MoE (Li et al., 2025). Following their original protocols, the compressed MoE-LLaVA models are evaluated on seven vision-language benchmarks, and the compressed Uni-MoE model is evaluated on eight image and speech understanding benchmarks. Each benchmark is reported with its official evaluation metric. The detailed model architecture and dataset settings are included in the Appendix.

Table 2: Performance comparisons on Uni-MoE (upcycled from Vicuna) for the general multi-modal task. MMBench^{A} and RACE^{A} denote MMBench-Audio and RACE-Audio, respectively. RACE^{A} and EHSL average their Middle/High and Long/Short subsets.

Method	Ratio	OKVQA	AOKVQA	VQA^{v2}	MMBench^{A}	POPE	Seed	RACE^{A}	EHSL	Avg
Uni-MoE *(Vicuna as Dense)* **37.06B** (Li et al., 2025)										
Vanilla	1x	75.88	76.77	86.61	60.25	86.37	66.43	63.43	45.34	70.13
DeRS-QT	8x	74.97	74.41	84.82	60.56	85.51	63.21	58.07	43.67	68.15
DeRS-QT	16x	0.96	0.96	0.03	2.17	67.10	0.14	0.04	0.00	8.92
DeRS-SP	20x	38.92	38.69	67.68	9.32	76.25	32.39	3.70	4.67	33.95
DeRS-SP	100x	0.79	1.48	0.00	1.24	50.49	0.45	0.09	0.00	6.82
DeRS-SP	200x	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeRS-SP	1000x	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UltraSparse	20x	71.76	71.09	82.01	55.59	83.80	56.31	51.27	39.00	63.85
UltraSparse	100x	64.29	58.78	73.67	52.80	79.73	46.09	45.55	38.00	57.36
UltraSparse	200x	63.00	56.77	72.32	52.48	79.24	44.64	44.36	38.00	56.35
UltraSparse	1000x	60.72	55.37	70.95	51.24	78.60	43.43	43.39	40.34	55.50

4.1.2 Results and Analysis

The results on MoE-LLaVA and Uni-MoE are reported in Tables 1 and 2, respectively.

Both baselines originate from DeRS (Huang et al., 2025b), which decomposes each expert and compresses the delta weights either by quantization, denoted DeRS-QT, or by random sparsification, denoted DeRS-SP. DeRS-QT is limited by the BF16 delta bit width and stops at $16\times$ under one-bit quantization; we therefore evaluate DeRS-QT at $8\times$ and $16\times$ and DeRS-SP and UltraSparse at $20\times$–$1000\times$. On MoE-LLaVA, UltraSparse outperforms DeRS-SP at every tested ratio on both backbones while limiting the average-score loss to fewer than 3 points at $1000\times$. On the Qwen backbone, DeRS-SP collapses to zero at this ratio, whereas UltraSparse retains 96% of the vanilla average. Uni-MoE is harder because its modality-specific experts are fine-tuned separately before assembly. DeRS-QT falls to 8.92 at $16\times$ and DeRS-SP to zero from $200\times$ onward, whereas UltraSparse still scores 55.50 at $1000\times$, retaining 79% of the vanilla average.

4.2 Medical Multi-Modal Task

4.2.1 Settings

The medical experiments adopt Med-MoE (Jiang et al., 2024). We evaluate the compressed models on three medical visual question-answering datasets, each providing open- and closed-ended question–answer pairs. Recall is reported for the open-ended questions and accuracy for the closed-ended ones. The detailed model architecture and dataset settings are included in the Appendix.

4.2.2 Results and Analysis

Table 3 shows that UltraSparse outperforms DeRS-SP on both backbones at every matched compression ratio. At $200\times$, its StableLM and Phi averages remain above vanilla at 70.82 and 73.48; even at $1000\times$, they exceed DeRS-SP by 1.03 and 1.65 points, respectively.

4.3 Code Generation Task

4.3.1 Settings

The code generation experiments cover two MoE models whose backbones are sparsified at different granularities, $\mathcal{X}$FT (Ding et al., 2024) and MultiPL-MoE (Wang et al., 2025a). MultiPL-MoE augments Qwen1.5-1.8B with a hybrid MoE architecture. Following the original protocol, the compressed models are evaluated on the Python code generation benchmarks and multilingual code generation tasks. All scores are reported as pass@1. The detailed model architecture and dataset settings are in the Appendix.

Table 3: Performance comparisons on two Med-MoE models, upcycled from (a) StableLM and (b) Phi, respectively, for the medical multi-modal tasks.

Method	Ratio	VQA-RAD		SLAKE		PathVQA		Avg
		Open	Close	Open	Close	Open	Close	
Med-MoE *(StableLM as Dense)* **3.61B** (Jiang et al., 2024)								
Vanilla	1x	47.64	81.62	83.54	85.10	33.38	90.80	70.34
DeRS-QT	8x	52.67	81.25	83.72	84.62	33.08	90.83	71.03
DeRS-QT	16x	47.67	79.78	79.36	83.17	30.91	89.71	68.43
DeRS-SP	20x	48.66	81.62	83.23	83.89	33.23	91.06	70.28
DeRS-SP	100x	49.51	81.25	83.28	84.62	33.11	90.92	70.45
DeRS-SP	200x	49.14	80.88	82.84	84.86	32.77	91.04	70.25
DeRS-SP	1000x	46.05	79.41	78.70	82.69	30.48	89.53	67.81
UltraSparse	20x	52.44	81.99	84.07	84.62	32.98	90.50	71.10
UltraSparse	100x	51.46	82.35	83.54	84.86	33.03	90.56	70.97
UltraSparse	200x	51.04	83.09	82.82	84.62	32.66	90.71	70.82
UltraSparse	1000x	48.25	79.78	81.56	82.69	31.17	89.62	68.84

(a)

Method	Ratio	VQA-RAD		SLAKE		PathVQA		Avg
		Open	Close	Open	Close	Open	Close	
Med-MoE *(Phi as Dense)* **6.45B** (Jiang et al., 2024)								
Vanilla	1x	56.45	83.46	83.69	86.30	34.43	91.33	72.61
DeRS-QT	8x	59.90	84.56	84.40	85.82	34.27	91.36	73.39
DeRS-QT	16x	54.66	84.19	80.22	82.93	32.03	90.47	70.75
DeRS-SP	20x	56.64	82.72	83.53	86.54	34.58	91.30	72.55
DeRS-SP	100x	56.64	81.99	83.24	85.82	34.72	91.21	72.27
DeRS-SP	200x	53.96	84.19	79.69	82.93	31.71	90.33	70.47
DeRS-SP	1000x	54.15	84.19	79.59	82.93	31.61	90.36	70.47
UltraSparse	20x	59.67	85.66	84.52	85.58	34.51	91.36	73.55
UltraSparse	100x	59.95	85.29	84.26	86.06	34.38	91.62	73.59
UltraSparse	200x	59.34	85.66	83.65	86.30	34.52	91.39	73.48
UltraSparse	1000x	55.93	84.56	82.77	86.06	32.78	90.65	72.12

(b)

Table 5: Performance comparisons on MultiPL-MoE upcycled from Qwen for multilingual code generation.

Method	Ratio	Programming Languages						Avg.
		Python	Java	C++	Rust	Go	Ruby	
MultiPL-MoE *(Qwen as Dense)* **10.8B** (Wang et al., 2025a)								**HEval**
Vanilla	1x	16.50	10.76	9.32	8.33	6.49	6.83	9.71
DeRS-QT	8x	15.85	12.03	11.80	4.49	7.79	7.45	9.90
DeRS-QT	16x	13.41	8.86	9.32	3.21	4.55	4.35	7.28
DeRS-SP	20x	3.00	2.53	1.24	2.56	1.95	2.48	2.30
DeRS-SP	100x	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeRS-SP	200x	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeRS-SP	1000x	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UltraSparse	20x	15.90	12.03	11.80	6.41	4.55	5.59	9.38
UltraSparse	100x	12.20	12.03	11.18	4.49	3.90	4.35	8.03
UltraSparse	200x	12.20	10.76	11.18	4.49	3.90	4.97	7.92
UltraSparse	1000x	14.00	9.49	11.18	4.49	4.55	6.83	8.42

(a)

Method	Ratio	Programming Languages						Avg.
		Python	Java	C++	Rust	Go	Ruby	
MultiPL-MoE *(Qwen as Dense)* **10.8B** (Wang et al., 2025a)								**MBPP**
Vanilla	1x	25.40	18.39	18.89	12.99	15.78	17.88	18.22
DeRS-QT	8x	23.56	18.13	16.12	11.86	14.44	15.37	16.58
DeRS-QT	16x	19.30	11.92	17.38	5.37	9.63	9.07	12.11
DeRS-SP	20x	2.90	5.18	5.29	3.67	4.81	6.05	4.65
DeRS-SP	100x	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeRS-SP	200x	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DeRS-SP	1000x	0.00	0.00	0.00	0.00	0.00	0.00	0.00
UltraSparse	20x	25.10	15.03	19.40	10.17	13.64	12.34	15.94
UltraSparse	100x	24.90	14.77	19.65	10.45	13.10	9.57	15.41
UltraSparse	200x	24.60	15.03	20.15	9.60	13.10	9.07	15.26
UltraSparse	1000x	25.10	13.21	18.39	9.89	12.30	7.56	14.41

(b)

4.3.2 Results and Analysis

The results are reported in Tables 4 and 5. On $\mathcal{X}$FT, UltraSparse remains close to the vanilla model across all tested compression ratios and consistently outperforms DeRS-SP, with a 12-point advantage in average score at 200×. This comparison shows that the proposed hybrid selection is more robust when only a small fraction of the expert updates can be retained.

Table 4: Performance comparisons on $\mathcal{X}$FT upcycled from DeepSeek-Coder for Python code generation.

Method	Ratio	HEval	HEval+	MBPP	MBPP+	Avg
$\mathcal{X}$**FT** *(DeepSeek as Dense)* **5.14B** (Ding et al., 2024)						
Vanilla	1x	65.20	61.00	62.20	52.10	60.12
DeRS-QT	8x	57.30	53.70	59.10	49.10	54.80
DeRS-QT	16x	0.00	0.00	0.00	0.00	0.00
DeRS-SP	20x	65.85	61.59	59.90	50.38	59.43
DeRS-SP	100x	57.32	54.27	58.65	50.63	55.21
DeRS-SP	200x	49.39	44.51	49.87	42.86	46.66
DeRS-SP	1000x	62.80	57.93	62.66	52.38	58.94
UltraSparse	20x	64.63	60.37	62.91	52.13	60.01
UltraSparse	100x	61.59	57.32	65.66	54.89	59.86
UltraSparse	200x	62.80	57.32	62.16	52.38	58.66
UltraSparse	1000x	64.02	59.15	62.16	51.88	59.30

MultiPL-MoE further tests the method under token- and segment-level routing. DeRS-SP degrades sharply at 20× and falls to zero on both benchmarks from 100× onward, whereas UltraSparse maintains nonzero performance across all six programming languages even at 1000×. At this extreme ratio, UltraSparse reduces the model to 1.85B parameters while retaining 86.7% and 79.1% of the vanilla HumanEval and MBPP averages, respectively. The consistent results on $\mathcal{X}$FT and MultiPL-MoE demonstrate that UltraSparse is applicable to both conventional top-1 routing and the two-granularity routing for multilingual code generation.

4.4 Ablation Study

4.4.1 Effect of Hybrid Budget Allocation

On the Qwen and Phi backbones of MoE-LLaVA, we fix the total keep ratio k^* and sweep the Top/Coverage allocation from pure Coverage at $0\%{+}100\%$ to pure Top at $100\%{+}0\%$ in steps of 10%. All settings use a sparsity ratio of 0.99, corresponding to $100\times$ expert-delta compression. Figure 3 reports the best 4Avg over interval counts $\{8, 16, 32, 64\}$ for each allocation, where 4Avg averages ScienceQA-IMG, GQA, TextVQA, and POPE; full per-benchmark results are provided in the Appendix. Hybrid allocations outperform both pure branches on both backbones, showing that extreme compression requires both high-magnitude critical elements and information spread across the non-top magnitude range.

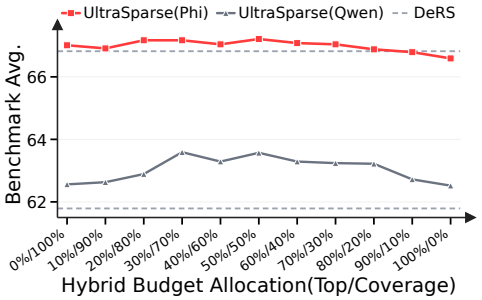


Figure 3: Effect of hybrid budget allocation on MoE-LLaVA at a sparsity ratio of 0.99. Each point denotes the best performance over interval counts $\{8, 16, 32, 64\}$ for a given $k^{\text{top}}{+}k^{\text{cov}}$ split. Note that $0\%{+}100\%$ and $100\%{+}0\%$ recover pure Coverage and pure Top, respectively.

4.4.2 Effect of Interval Granularity in WDP

Under the best hybrid budget allocation of each model, we evaluate four interval counts, $B \in \{8, 16, 32, 64\}$. Table 6a shows MoE-LLaVA, Med-MoE, and $\mathcal{X}$FT with their corresponding DeRS-SP baselines at $100\times$. MoE-LLaVA and Med-MoE vary by at most 0.44 and 0.71 points, whereas $\mathcal{X}$FT decreases from 59.86 at $B{=}8$ to 57.32 at $B{=}64$. Thus, a small-to-moderate number of magnitude intervals is sufficient, and increasing B to 64 does not consistently improve results.

4.4.3 Effect of Interval Construction in WDP

We further compare how WDP constructs its intervals along two axes, equal-width versus equal-count binning, and binning on $|\mathbf{\Delta}|$ versus on signed $\mathbf{\Delta}$. The default method uses equal-width bins on $|\mathbf{\Delta}|$, as in Eq. (2). We also evaluate equal-count bins on $|\mathbf{\Delta}|$, equal-width bins on $\mathbf{\Delta}$, and equal-count bins on $\mathbf{\Delta}$. These four settings are tested on Uni-MoE and on MultiPL-MoE under HumanEval and MBPP. Results are reported in Table 6b.

Table 6: (a) Effect of interval granularity in WDP under the best hybrid budget allocation of each model. Baseline denotes DeRS-SP at $100\times$. All results are average scores across the corresponding benchmarks. (b) Effect of interval construction in WDP at $100\times$. Rows are the four binning constructions; columns report average scores on Uni-MoE and on MultiPL-MoE under HumanEval and MBPP. The default setting is equal-width bins on $|\mathbf{\Delta}|$.

MoE	Dense	Baseline	$B{=}8$	$B{=}16$	$B{=}32$	$B{=}64$
MoE-LLaVA	Phi	62.86	62.90	**63.30**	62.86	63.07
	Qwen	56.44	57.61	57.62	**57.81**	57.75
Med-MoE	StableLM	70.45	70.26	**70.97**	70.72	70.72
	Phi	72.27	**73.59**	73.26	73.06	73.17
$\mathcal{X}$FT	DeepSeek	55.21	**59.86**	58.03	57.41	57.32

(a)

Construction	Uni-MoE	MultiPL-MoE HEval	MultiPL-MoE MBPP
equal-width on $\lvert\mathbf{\Delta}\rvert$	**57.36**	**8.03**	**15.41**
equal-count on $\lvert\mathbf{\Delta}\rvert$	52.40	7.91	15.00
equal-width on $\mathbf{\Delta}$	54.20	8.01	15.22
equal-count on $\mathbf{\Delta}$	54.00	7.60	15.09

(b)

5 Conclusion

In this work, we investigate the extreme compression of upcycled MoE models from the perspective of delta weights. Through systematic analysis, we reveal that the compression limit of expert-specific delta weights is governed not only by parameter selection or representation, but critically by preserving both weight distribution and knowledge distribution. Based on these findings, we propose UltraSparse, a novel framework that enables extreme compression of upcycled MoE models. Extensive experiments demonstrate that UltraSparse consistently outperforms existing compression methods and achieves up to $1000\times$ compression while maintaining strong model performance. Our work highlights distribution preservation as a fundamental principle for unlocking the extreme compression potential of upcycled MoE models. We hope this perspective can inspire future research on efficient scaling and deployment of large-scale sparse models.

AI USE STATEMENT

We used generative AI tools for coding assistance and language polishing. All AI-assisted code and text were carefully reviewed by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript, code, and reported results.

REFERENCES

Theodore W Anderson. On the distribution of the two-sample Cramér-von Mises criterion. *The Annals of Mathematical Statistics*, pp. 1148–1159, 1962.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pp. 4218–4222. European Language Resources Association, 2020. URL https://aclanthology.org/2020.lrec-1.520/.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models, 2021. URL https://arxiv.org/abs/2108.07732.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramér distance as a solution to biased Wasserstein gradients, 2017. URL http://arxiv.org/abs/1705.10743.

Mingyu Cao, Gen Li, Jie Ji, Jiaqi Zhang, Ajay Jaiswal, Li Shen, Xiaolong Ma, Shiwei Liu, and Lu Yin. Condense, don't just prune: Enhancing efficiency and performance in MoE layer pruning. *Transactions on Machine Learning Research*, 2025.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Trans. Software Eng.*, 49(7):3675–3691, 2023. doi: 10.1109/TSE.2023.3267446. URL https://doi.org/10.1109/TSE.2023.3267446.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 1280–1297. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.70. URL https://doi.org/10.18653/v1/2024.acl-long.70.

Wenlong Deng, Yize Zhao, Vala Vakilian, Minghui Chen, Xiaoxiao Li, and Christos Thrampoulidis. DARE the extreme: Revisiting delta-parameter pruning for fine-tuned models. In *International Conference on Learning Representations*, volume 2025, pp. 83658–83692, 2025.

Yifeng Ding, Jiawei Liu, Yuxiang Wei, and Lingming Zhang. $\mathcal{X}$FT: unlocking the power of code instruction tuning by simply merging upcycled mixture-of-experts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16,*

2024, pp. 12941–12955. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.699. URL https://doi.org/10.18653/v1/2024.acl-long.699.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 736–740. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9052990. URL https://doi.org/10.1109/ICASSP40776.2020.9052990.

Elias Frantar and Dan Alistarh. QMoE: Sub-1-bit compression of trillion parameter models. *Proceedings of Machine Learning and Systems*, 6:439–451, 2024.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.*, 127(4):398–414, 2019. doi: 10.1007/S11263-018-1116-0. URL https://doi.org/10.1007/s11263-018-1116-0.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. DeepSeek-Coder: When the large language model meets programming - the rise of code intelligence, 2024. URL https://doi.org/10.48550/arXiv.2401.14196.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both weights and connections for efficient neural network. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 1135–1143, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/ae0eb3eed39d2bcef4622b2499a05fe6-Abstract.html.

Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering, 2020. URL https://arxiv.org/abs/2003.10286.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. EMR-Merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769, 2024.

Chenyu Huang, Peng Ye, Xiaohui Wang, Shenghe Zheng, Biqing Qi, Lei Bai, Wanli Ouyang, and Tao Chen. Seeing delta parameters as JPEG images: Data-free delta compression with discrete cosine transform, 2025a. URL https://arxiv.org/abs/2503.06676.

Yongqi Huang, Peng Ye, Chenyu Huang, Jianjian Cao, Lin Zhang, Baopu Li, Gang Yu, and Tao Chen. DeRS: Towards extremely efficient upcycled mixture-of-experts models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 10056–10066. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.00940. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huang_DeRS_Towards_Extremely_Efficient_Upcycled_Mixture-of-Experts_Models_CVPR_2025_paper.html.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Tingfeng Hui, Zhenyu Zhang, Shuohuan Wang, Yu Sun, Hua Wu, and Sen Su. Upcycling instruction tuning from dense to mixture-of-experts via parameter merging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13012–13031, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-MoE: Mixture of domain-specific experts for lightweight medical vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pp. 3843–3860. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.221. URL https://doi.org/10.18653/v1/2024.findings-emnlp.221.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 119–132. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1011. URL https://doi.org/10.18653/v1/n19-1011.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. The Stack: 3 TB of permissively licensed source code. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=pxpbTdUEpD.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=T5nUQDrM4u.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 785–794. Association for Computational Linguistics, 2017. doi: 10.18653/V1/D17-1082. URL https://doi.org/10.18653/v1/d17-1082.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Luc Lauwers and Tom Van Puyenbroeck. The Hamilton apportionment method is between the Adams method and the Jefferson method. *Math. Oper. Res.*, 31(2):390–397, 2006. doi: 10.1287/MOOR.1060.0192. URL https://doi.org/10.1287/moor.1060.0192.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. MIMIC-IT: multi-modal in-context instruction tuning, 2023a. URL https://doi.org/10.48550/arXiv.2306.05425.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal LLMs with generative comprehension, 2023b. URL https://doi.org/10.48550/arXiv.2307.16125.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36:*

Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023, 2023c. URL http://papers.nips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023d. URL https://proceedings.mlr.press/v202/li23q.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023e. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://doi.org/10.18653/v1/2023.emnlp-main.20.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-MoE: Scaling unified multimodal LLMs with mixture of experts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3424–3439, 2025. doi: 10.1109/TPAMI.2025.3532688. URL https://doi.org/10.1109/TPAMI.2025.3532688.

Ning Liao, Xiaoxing Wang, Zehao Lin, Weiyang Guo, Feng Hong, Shixiang Song, Geng Yu, Zihua Zhao, Sitao Xie, Longxuan Wei, et al. Innovator: Scientific continued pretraining with fine-grained MoE upcycling, 2025. URL https://arxiv.org/abs/2507.18671.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. MoE-LLaVA: Mixture of experts for large vision-language models. *IEEE Trans. Multim.*, 28:4408–4419, 2026. doi: 10.1109/TMM.2026.3654458. URL https://doi.org/10.1109/TMM.2026.3654458.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *18th IEEE International Symposium on Biomedical Imaging, ISBI 2021, Nice, France, April 13-16, 2021*, pp. 1650–1654. IEEE, 2021. doi: 10.1109/ISBI48211.2021.9434010. URL https://doi.org/10.1109/ISBI48211.2021.9434010.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning, 2023a. URL https://doi.org/10.48550/arXiv.2306.14565.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

James Liu, Guangxuan Xiao, Kai Li, Jason D. Lee, Song Han, Tri Dao, and Tianle Cai. BitDelta: Your fine-tune may only be worth one bit. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/187d94b3c93343f0e925b5cf729eadd5-Abstract-Conference.html.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023c. URL http://papers.nips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pp. 216–233. Springer, 2024c. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172, 2024.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. WizardCoder: Empowering code large language models with Evol-Instruct. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=UnUwSIgK5W.

Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.679. URL https://doi.org/10.18653/v1/2024.acl-long.679.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3195–3204. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00331. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446. URL https://doi.org/10.1109/TASLP.2024.3419446.

Taishi Nakamura, Takuya Akiba, Kazuki Fujii, Yusuke Oda, Rio Yokota, and Jun Suzuki. Drop-Upcycling: Training sparse mixture of experts with partial re-initialization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97: 123–150, 1934. URL https://api.semanticscholar.org/CorpusID:86859017.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 527–536. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1050. URL https://doi.org/10.18653/v1/p19-1050.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Ramakrishnan Sathyavageeswaran. FIM-LoRA: Task-informative rank allocation for LoRA via calibration-time gradient-variance estimation, 2026. URL https://doi.org/10.48550/arXiv.2605.16800.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pp. 146–162. Springer, 2022. doi: 10.1007/978-3-031-20074-8_9. URL https://doi.org/10.1007/978-3-031-20074-8_9.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00851. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting GPT-4V for better visual instruction tuning, 2023. URL https://doi.org/10.48550/arXiv.2311.07574.

Qing Wang, Xue Han, Jiahui Wang, Lehao Xing, Qian Hu, Lianlian Zhang, Chao Deng, and Junlan Feng. MultiPL-MoE: Multi-programming-lingual extension of large language models through hybrid mixture-of-experts. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 12817–12828. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.FINDINGS-EMNLP.686. URL https://doi.org/10.18653/v1/2025.findings-emnlp.686.

Xiaohui Wang, Peng Ye, Chenyu Huang, Shenghe Zheng, Bo Zhang, Lei Bai, Wanli Ouyang, and Tao Chen. Breaking the compression ceiling: Data-free pipeline for ultra-efficient delta compression. *Advances in Neural Information Processing Systems*, 38:23167–23195, 2025b.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Yuanlin Duan, Wenqi Jia, Miao Yin, Yu Cheng, and Bo Yuan. MoE-I^2: Compressing mixture of experts models through inter-expert pruning and intra-expert low-rank decomposition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10456–10466, 2024.

Peng Ye, Chenyu Huang, Mingzhu Shen, Tao Chen, Yongqi Huang, and Wanli Ouyang. Dynamic model merging with mixture of weights. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57755–57775. PMLR / OpenReview.net, 2024a. URL https://proceedings.mlr.press/v235/yu24p.html.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR / OpenReview.net, 2024b. URL https://proceedings.mlr.press/v235/yu24o.html.

Bo Zhao, Boya Wu, and Tiejun Huang. SVIT: scaling up visual instruction tuning, 2023. URL https://doi.org/10.48550/arXiv.2307.04087.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew M. Dai, Zhifeng Chen, Quoc V. Le, and James Laudon. Mixture-of-experts with expert choice routing. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/2f00ecd787b432c1d36f3de9800728eb-Abstract-Conference.html.

Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. LLaMA-MoE: Building mixture-of-experts from LLaMA with continual pre-training. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 15913–15923, 2024.

A Method Overview

Figure 4 illustrates the overall pipeline of UltraSparse.

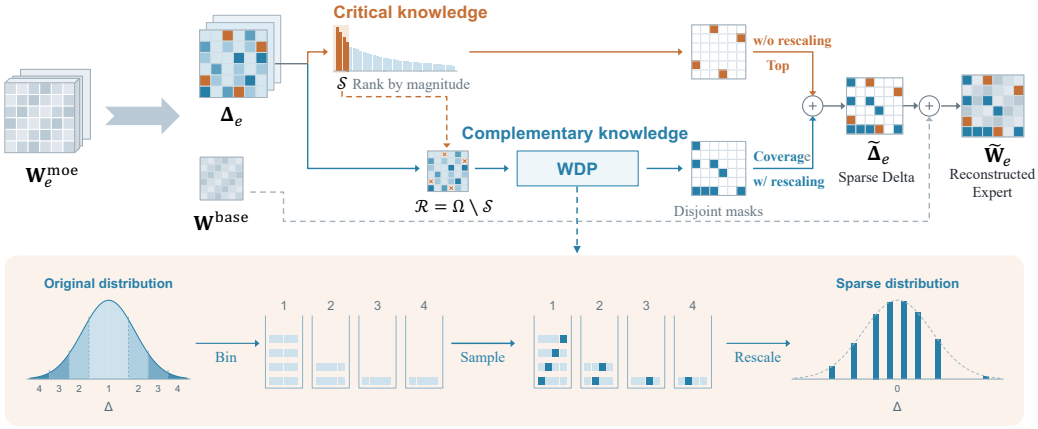


Figure 4: Overview of UltraSparse. The Top branch keeps the largest-magnitude delta elements without rescaling, while the Coverage branch applies WDP to the remaining positions with rescaling. WDP bins the remaining elements by magnitude, samples each bin in proportion to its size, and rescales the kept elements, so that the sparse delta follows the original weight distribution.

B Model and Evaluation Details

This section describes the source architectures, original training recipes, training data, evaluation benchmarks, and metrics for the five upcycled MoE model families considered in the main paper. These details characterize the trained models to which all compression methods are applied.

B.1 General Multi-Modal Models

Model Architectures and Training Recipes

The general multi-modal experiments use MoE-LLaVA (Lin et al., 2026) and Uni-MoE (Li et al., 2025). MoE-LLaVA pairs a CLIP-Large (Radford et al., 2021) visual encoder with Qwen-1.8B or Phi-2.7B as the language backbone. Each backbone is first fine-tuned as a dense model on a mixture of MIMIC-IT-256k (Li et al., 2023a), LRV-331k (Liu et al., 2023a), SViT-157k (Zhao et al., 2023), and LVIS-256k (Wang et al., 2023). The FFN in every other transformer block is then expanded into a 4-expert MoE layer with top-2 routing. The upcycled model is further fine-tuned on LLaVA-mix-665k (Liu et al., 2024a), with only the MoE layers trainable.

Uni-MoE builds a fully sparse backbone from Vicuna-7B through three progressive stages. Cross-modality alignment first trains the modality connectors on WavCaps (Mei et al., 2024), AudioCaps (Kim et al., 2019), MELD (Poria et al., 2019), ClothoV1 (Drossos et al., 2020), and Common Voice (Ardila et al., 2020), while the language model remains frozen. The second stage separately fine-tunes eight modality-specific experts on task-specific mixtures drawn from LLaVA-Instruct-150K (Liu et al., 2023b), RACE (Lai et al., 2017), LibriSpeech (Panayotov et al., 2015), and the audio corpora above. The final stage assembles the experts into 8-way MoE layers that replace the FFNs in all 32 transformer blocks. The assembled model is jointly refined with LoRA (Hu et al., 2022) under top-2 routing on LLaVA-v1.5-665K (Liu et al., 2024a), Video-Instruct (Maaz et al., 2024), and speech-augmented RACE and LibriSpeech data. Images are encoded by CLIP-Large and speech by Whisper-small (Radford et al., 2023). An MLP connector maps image features into the language space, while a Q-Former (Li et al., 2023d) followed by a linear projection maps speech features.

DATASETS AND EVALUATION METRICS

MoE-LLaVA follows its original protocol on seven vision–language benchmarks. VQAv2 (Goyal et al., 2019), GQA (Hudson & Manning, 2019), ScienceQA-IMG (Lu et al., 2022), and TextVQA (Singh et al., 2019) evaluate visual question answering. POPE (Li et al., 2023e) evaluates object hallucination, while MMBench (Liu et al., 2024c) and MM-Vet (Yu et al., 2024b) evaluate integrated multi-modal understanding. Because MM-Vet requires an LLM-as-a-judge evaluation and we observed scoring stochasticity with GPT-5.4, fixed MM-Vet outputs are judged five times and the mean score is reported.

Uni-MoE is evaluated on ten benchmarks spanning image and speech understanding. The image benchmarks include OKVQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), VQAv2, POPE, and SEED-Bench (Li et al., 2023b). The speech benchmarks include MMBench-Audio (Liu et al., 2024c), RACE-Audio (Lai et al., 2017), and the English High School Listening tests (Li et al., 2025). MMBench-Audio and RACE-Audio convert the corresponding textual questions and passages into speech to assess joint image–speech understanding. Each benchmark is reported with its official evaluation metric.

B.2 MEDICAL MULTI-MODAL MODELS

MODEL ARCHITECTURES AND TRAINING RECIPES

The medical experiments follow Med-MoE (Jiang et al., 2024) with StableLM-1.6B and Phi-2.7B as dense backbones. Both backbones retain a CLIP-Large visual encoder and are first fine-tuned on the medical image–language instruction data from LLaVA-Med (Li et al., 2023c). During upcycling, the FFN in every other transformer block is replaced by a parallel universal FFN and a 4-expert MoE layer. The universal FFN processes every token, while the MoE branch selects one expert through top-1 routing. Their outputs are summed to form the final representation.

DATASETS AND EVALUATION METRICS

The upcycled Med-MoE models are fine-tuned and evaluated on VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020). Each medical visual question-answering dataset contains open- and closed-ended question–answer pairs. Recall is reported for open-ended questions and accuracy for closed-ended questions.

B.3 CODE GENERATION MODELS

MODEL ARCHITECTURES AND TRAINING RECIPES

The code generation experiments use $\mathcal{X}$FT (Ding et al., 2024) and MultiPL-MoE (Wang et al., 2025a), whose backbones are sparsified at different granularities. $\mathcal{X}$FT upcycles DeepSeek-Coder-Base-1.3B (Guo et al., 2024) into a Coder-MoE architecture. Every transformer FFN is replaced by a parallel universal FFN and a 4-expert MoE layer with top-1 routing. The model is instruction-tuned on 110K instruction–output pairs from evol-codealpaca-v1 (Luo et al., 2024).

MultiPL-MoE augments Qwen1.5-1.8B with hybrid routing at token and segment levels. The token-level branch contains six experts with one shared expert following the DeepSeekMoE design (Dai et al., 2024), and each token activates the top-2 experts. The segment-level branch partitions the sequence into fixed-length windows and uses expert-choice routing (Zhou et al., 2022) under a fixed capacity. The two branch outputs are combined by a weighted sum. The model is post-pretrained on 3.3B code tokens sampled from The Stack (Kocetkov et al., 2023), with a 9-to-1 ratio of low-resource to high-resource programming languages. Only the newly introduced experts and routers are updated, while the remaining backbone parameters stay frozen.

DATASETS AND EVALUATION METRICS

$\mathcal{X}$FT is evaluated on HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), and the HumanEval+ and MBPP+ variants provided by EvalPlus (Liu et al., 2023c). MultiPL-MoE follows its original protocol on the MultiPL-E (Cassano et al., 2023) translations of HumanEval and MBPP. The

evaluation covers the high-resource languages Python, Java, and C++ together with the low-resource languages Rust, Go, and Ruby. All code generation scores are reported as pass@1.

B.4 Compression Baselines and Common Protocol

The compression baselines are DeRS-QT and DeRS-SP from DeRS (Huang et al., 2025b). DeRS decomposes each trained expert into shared base weights and expert-specific delta weights. DeRS-QT quantizes the delta weights, while DeRS-SP applies independent random sparsification with drop-and-rescale (Yu et al., 2024a). All reported compression ratios apply to expert-specific delta parameters. At $100\times$ expert-delta compression, one percent of the delta elements are retained.

C Full Experimental Results

This section reports the complete benchmark-level results behind the hybrid budget allocation and interval-granularity studies of WDP+ in the main paper. All tables retain the model training and evaluation protocols described above.

C.1 General Multi-Modal Hybrid Budget Allocation Results

Tables S1 and S2 report the complete benchmark results for WDP+ under different hybrid budget allocations at $100\times$ expert-delta compression. All runs use equal-width intervals on $|\mathbf{\Delta}|$, largest-remainder quota allocation, and the global $1/k^*$ rescaling. For each row, *Top* and *Coverage* denote the percentages of the total keep budget assigned to deterministic retention of the largest-magnitude elements and WDP sampling over the remaining magnitude range, respectively. *Intervals* denotes the number of intervals. The reported Avg is the mean of SQA image accuracy, GQA accuracy, TextVQA accuracy, and POPE accuracy. All displayed values are rounded to two decimal places.

C.2 Per-Benchmark Results for Interval Granularity

The main-paper interval-granularity results for Qwen correspond to 30% Top and 70% Coverage in **Table S2**. The Phi results correspond to 50% Top and 50% Coverage in **Table S1**. **Tables S3 and S4** provide the corresponding complete seven-benchmark results. All settings use $100\times$ expert-delta compression and equal-width intervals on $|\mathbf{\Delta}|$. For MM-Vet, the GPT-5.4 judge is run five times on each fixed model output and the mean score is reported to reduce scoring stochasticity. MMBench uses the TEST_EN_V11 split, while VQAv2 reports the official overall accuracy.

C.3 Medical Multi-Modal Interval-Granularity Results

Tables S5 and S6 report the per-benchmark interval-granularity results for WDP+ on Med-MoE at $100\times$ expert-delta compression. Both backbones use the 40% Top and 60% Coverage configuration, which is used for the $100\times$ Med-MoE results in the main paper. All runs use equal-width intervals on $|\mathbf{\Delta}|$, largest-remainder quota allocation, and the global $1/k^*$ rescaling. The reported Avg is the mean of the open- and closed-ended scores on VQA-RAD, SLAKE, and PathVQA.

C.4 Code Generation Interval-Granularity Results

Table S7 reports the complete per-benchmark interval-granularity results for WDP+ on $\mathcal{X}$FT at $100\times$ expert-delta compression. The table uses the 80% Top and 20% Coverage configuration, which is used for the $100\times$ $\mathcal{X}$FT results in the main paper. All runs use equal-width intervals on $|\mathbf{\Delta}|$, largest-remainder quota allocation, and the global $1/k^*$ rescaling. Each benchmark cell reports pass@1 as a percentage. The Avg score is the mean over HumanEval, HumanEval+, MBPP, and MBPP+.

Table S1: Hybrid budget allocation ablation of WDP+ on MoE-LLaVA upcycled from Phi at $100\times$ expert-delta compression. Baseline denotes DeRS-SP at $100\times$. Top and Coverage give the percentages of the total keep budget assigned to deterministic retention of the largest-magnitude elements and WDP sampling over the remaining magnitude range, respectively. The 0%/100% and 100%/0% settings recover the two pure branches. Avg is the arithmetic mean of SQA-IMG, GQA, TextVQA, and POPE. For each interval count, boldface marks the highest Avg across allocations.

Top (%)	Coverage (%)	Intervals	SQA^{I}	GQA	VQA^{T}	POPE	Avg
Baseline			68.32	61.18	50.37	87.41	66.82
0	100	8	67.63	60.42	50.73	87.09	66.47
0	100	16	68.42	61.54	50.41	87.65	67.01
0	100	32	67.43	61.20	50.89	87.18	66.68
0	100	64	68.91	61.04	50.66	87.28	66.97
10	90	8	67.58	60.99	51.48	86.95	66.75
10	90	16	67.58	60.85	50.61	86.29	66.33
10	90	32	67.77	60.98	51.22	87.32	66.82
10	90	64	67.92	61.19	51.20	87.34	66.91
20	80	8	68.62	60.96	50.61	87.67	66.96
20	80	16	69.11	60.69	50.87	87.46	67.03
20	80	32	69.31	60.72	50.90	87.76	**67.17**
20	80	64	68.52	60.78	50.78	87.44	66.88
30	70	8	68.62	60.84	50.69	87.71	66.97
30	70	16	69.06	60.69	50.78	87.14	66.92
30	70	32	69.71	60.88	50.82	87.25	**67.17**
30	70	64	69.26	60.77	51.00	87.61	**67.16**
40	60	8	69.31	60.72	50.81	87.32	67.04
40	60	16	69.01	60.63	51.18	87.08	66.98
40	60	32	69.01	60.75	51.07	87.09	66.98
40	60	64	68.91	60.83	50.76	87.15	66.91
50	50	8	68.77	60.51	50.68	87.27	66.81
50	50	16	70.00	60.83	50.54	87.49	**67.21**
50	50	32	68.02	60.84	50.89	87.19	66.74
50	50	64	68.91	60.88	50.84	87.25	66.97
60	40	8	69.16	60.55	51.15	87.46	**67.08**
60	40	16	68.86	60.38	51.05	87.65	66.99
60	40	32	68.62	60.57	51.14	87.31	66.91
60	40	64	68.62	60.38	50.57	87.54	66.78
70	30	8	68.91	60.14	50.72	87.36	66.78
70	30	16	68.62	60.10	50.55	87.55	66.71
70	30	32	69.11	60.27	50.46	87.22	66.76
70	30	64	69.71	60.29	50.74	87.43	67.04
80	20	8	69.51	59.81	50.22	87.47	66.75
80	20	16	69.26	59.84	50.54	87.23	66.72
80	20	32	69.06	60.07	50.73	87.46	66.83
80	20	64	69.61	59.93	50.78	87.22	66.88
90	10	8	68.72	59.39	50.65	87.37	66.53
90	10	16	69.36	59.56	50.21	87.26	66.60
90	10	32	69.46	59.73	50.68	87.30	66.79
90	10	64	69.66	59.39	50.64	87.30	66.75
100	0	8	69.71	59.12	50.43	87.12	66.59
100	0	16	69.71	59.12	50.43	87.12	66.59
100	0	32	69.71	59.12	50.43	87.12	66.59
100	0	64	69.71	59.12	50.43	87.12	66.59

Table S2: Hybrid budget allocation ablation of WDP+ on MoE-LLaVA upcycled from Qwen at $100\times$ expert-delta compression. Baseline denotes DeRS-SP at $100\times$. Top and Coverage give the percentages of the total keep budget assigned to deterministic retention of the largest-magnitude elements and WDP sampling over the remaining magnitude range, respectively. The 0%/100% and 100%/0% settings recover the two pure branches. Avg is the arithmetic mean of SQA-IMG, GQA, TextVQA, and POPE. For each interval count, boldface marks the highest Avg across allocations.

Top (%)	**Coverage (%)**	**Intervals**	**SQA^{I}**	**GQA**	**VQA^{T}**	**POPE**	**Avg**
Baseline			57.96	58.63	43.13	87.46	61.80
0	100	8	57.71	58.25	43.98	86.70	61.66
0	100	16	59.74	59.31	43.45	87.72	62.56
0	100	32	57.56	58.24	43.89	87.32	61.75
0	100	64	57.11	59.02	41.85	87.26	61.31
10	90	8	58.11	59.25	42.70	86.44	61.63
10	90	16	59.84	58.55	44.22	87.90	62.63
10	90	32	57.46	59.48	44.34	86.76	62.01
10	90	64	58.55	58.52	44.67	85.60	61.84
20	80	8	59.64	59.27	45.07	87.36	62.84
20	80	16	57.61	58.92	43.58	87.98	62.02
20	80	32	59.05	58.79	43.73	87.63	62.30
20	80	64	59.84	58.75	45.54	87.43	62.89
30	70	8	59.54	58.62	44.07	87.54	62.44
30	70	16	62.37	58.96	45.12	87.89	**63.59**
30	70	32	60.98	58.89	44.51	88.01	63.10
30	70	64	59.59	59.46	45.40	87.64	63.02
40	60	8	59.79	59.05	44.58	87.77	62.80
40	60	16	60.49	59.42	45.57	87.70	63.29
40	60	32	60.54	58.75	44.38	87.78	62.86
40	60	64	59.99	59.22	45.89	87.93	**63.26**
50	50	8	61.18	58.67	44.56	87.63	63.01
50	50	16	61.13	59.17	45.85	88.11	63.57
50	50	32	61.38	59.10	45.48	88.06	**63.50**
50	50	64	60.44	59.00	44.35	87.68	62.87
60	40	8	60.34	59.10	45.84	87.10	63.10
60	40	16	60.98	58.82	45.69	87.68	63.29
60	40	32	61.28	58.83	44.54	87.67	63.08
60	40	64	59.79	59.10	44.81	87.52	62.80
70	30	8	62.32	57.99	44.98	86.97	63.06
70	30	16	60.68	58.31	45.13	87.70	62.95
70	30	32	61.28	58.53	45.54	87.27	63.16
70	30	64	62.02	58.34	45.19	87.42	63.24
80	20	8	61.73	57.98	45.20	87.62	**63.13**
80	20	16	60.19	57.81	45.09	87.71	62.70
80	20	32	61.77	57.78	45.51	87.82	63.22
80	20	64	60.93	57.84	45.59	87.27	62.91
90	10	8	61.82	57.41	44.29	87.30	62.70
90	10	16	60.78	57.62	44.58	87.08	62.52
90	10	32	61.58	57.18	44.81	87.32	62.72
90	10	64	60.78	56.89	44.58	87.37	62.41
100	0	8	61.92	56.82	44.16	87.16	62.52
100	0	16	61.92	56.82	44.16	87.16	62.52
100	0	32	61.92	56.82	44.16	87.16	62.52
100	0	64	61.92	56.82	44.16	87.16	62.52

Table S3: Interval-granularity ablation on MoE-LLaVA upcycled from Qwen at $100\times$ expert-delta compression with 30% Top and 70% Coverage. Baseline denotes DeRS-SP at $100\times$. The interval count varies over $\{8, 16, 32, 64\}$. Avg is the arithmetic mean of SQA-IMG, GQA, TextVQA, POPE, MM-Vet, MMBench, and VQAv2.

Intervals	SQA^{I}	GQA	VQA^{T}	POPE	MMVet	$\text{MMBench}^{\text{EN}}$	VQA^{v2}	Avg
Baseline	57.96	58.63	43.13	87.46	24.80	49.30	73.79	56.44
8	59.54	58.62	44.07	87.54	28.50	50.86	74.15	57.61
16	62.37	58.96	45.12	87.89	25.80	48.73	74.49	57.62
32	60.98	58.89	44.51	88.01	26.30	51.90	74.10	57.81
64	59.59	59.46	45.40	87.64	27.30	50.65	74.18	57.75

Table S4: Interval-granularity ablation on MoE-LLaVA upcycled from Phi at $100\times$ expert-delta compression with 50% Top and 50% Coverage. Baseline denotes DeRS-SP at $100\times$. The interval count varies over $\{8, 16, 32, 64\}$. Avg is the arithmetic mean of SQA-IMG, GQA, TextVQA, POPE, MM-Vet, MMBench, and VQAv2.

Intervals	SQA^{I}	GQA	VQA^{T}	POPE	MMVet	$\text{MMBench}^{\text{EN}}$	VQA^{v2}	Avg
Baseline	68.32	61.18	50.37	87.41	34.40	61.19	77.15	62.86
8	68.77	60.51	50.68	87.27	33.60	62.34	77.10	62.90
16	70.00	60.83	50.54	87.49	34.90	62.13	77.19	63.30
32	68.02	60.84	50.89	87.19	32.10	61.77	77.22	62.58
64	68.91	60.88	50.84	87.25	31.70	61.82	77.12	62.65

Table S5: Interval-granularity ablation of WDP+ on Med-MoE upcycled from StableLM at $100\times$ expert-delta compression with 40% Top and 60% Coverage. Baseline denotes DeRS-SP at $100\times$. Avg is the arithmetic mean of the open-ended recall and closed-ended accuracy scores on VQA-RAD, SLAKE, and PathVQA. Boldface marks the highest Avg across interval counts $\{8, 16, 32, 64\}$.

Intervals	VQA-RAD		SLAKE		PathVQA		Avg
	Open	Close	Open	Close	Open	Close	
Baseline	49.51	81.25	83.28	84.62	33.11	90.92	70.45
8	49.59	82.35	83.03	83.17	32.89	90.53	70.26
16	51.46	82.35	83.54	84.86	33.03	90.56	**70.97**
32	49.90	83.09	82.77	84.86	33.08	90.62	70.72
64	52.45	80.88	83.70	83.65	33.06	90.56	70.72

Table S6: Interval-granularity ablation of WDP+ on Med-MoE upcycled from Phi at $100\times$ expert-delta compression with 40% Top and 60% Coverage. Baseline denotes DeRS-SP at $100\times$. Avg is the arithmetic mean of the open-ended recall and closed-ended accuracy scores on VQA-RAD, SLAKE, and PathVQA. Boldface marks the highest Avg across interval counts $\{8, 16, 32, 64\}$.

Intervals	VQA-RAD		SLAKE		PathVQA		Avg
	Open	Close	Open	Close	Open	Close	
Baseline	56.64	81.99	83.24	85.82	34.72	91.21	72.27
8	59.95	85.29	84.26	86.06	34.38	91.62	**73.59**
16	58.83	84.93	84.14	85.58	34.49	91.60	73.26
32	58.50	84.56	83.77	85.82	34.30	91.42	73.06
64	58.27	84.93	83.97	86.06	34.26	91.54	73.17

Table S7: Interval-granularity ablation of WDP+ on $\mathcal{X}$FT upcycled from DeepSeek-Coder at $100\times$ expert-delta compression with 80% Top and 20% Coverage. Baseline denotes DeRS-SP at $100\times$. HumanEval, HumanEval+, MBPP, and MBPP+ are evaluated by pass@1, with Avg denoting their arithmetic mean. Boldface marks the highest Avg across interval counts $\{8, 16, 32, 64\}$.

Intervals	**HumanEval**	**HumanEval+**	**MBPP**	**MBPP+**	**Avg**
Baseline	57.32	54.27	58.65	50.63	55.21
8	61.59	57.32	65.66	54.89	**59.86**
16	60.37	56.71	63.41	51.63	58.03
32	59.76	56.10	62.41	51.38	57.41
64	59.15	56.10	61.90	52.13	57.32

D THEORETICAL ANALYSIS OF WDP+

WDP+ has four theoretical properties. Global Coverage scaling shrinks non-Top entries, but its squared bias is lower order than sampling variance in the ultra-sparse regime with a fixed Coverage share. It also reduces variance and MSE relative to Coverage-unbiased scaling and, under ideal allocation, has no higher weight-space MSE than standard drop-and-rescale over the studied range. Fixed-budget allocation preserves the exact Coverage count, with histogram and rounding errors vanishing as the retained budget grows. Finally, the weight-space bounds control linear-module output perturbations but do not directly guarantee task-level performance.

D.1 SETUP

We flatten one expert delta matrix into N scalar entries and denote the total keep ratio by

$$\kappa = k^* \in (0, 1). \tag{S1}$$

We use $\kappa = k^*$, $\tau = k^{\text{top}}$, $\lambda = \tau/\kappa$, and $\eta = K^{\text{cov}}/|\mathcal{R}|$ to match the notation in the main paper. At $100\times$ expert-delta compression, $\kappa = 0.01$. The Top branch consumes an absolute keep ratio $\tau = k^{\text{top}}$. We write its fraction of the total budget as

$$\lambda = \frac{\tau}{\kappa} \in [0, 1]. \tag{S2}$$

We first ignore integer rounding so that the Top branch contains $\tau N = \lambda\kappa N$ entries. The remaining set $\mathcal{R}$ then contains $n = (1 - \lambda\kappa)N$ entries, while the Coverage branch retains $(1 - \lambda)\kappa N$ entries. The average Coverage inclusion probability over $\mathcal{R}$ is therefore

$$\eta = \frac{\kappa - \tau}{1 - \tau} = \frac{\kappa(1 - \lambda)}{1 - \lambda\kappa}. \tag{S3}$$

A nonzero Top budget gives $\eta < \kappa$. Rescaling Coverage entries by $1/\kappa$ instead of $1/\eta$ consequently introduces shrinkage toward zero in expectation.

Let $\mathbf{\Delta}_{\mathcal{T}}$ and $\mathbf{\Delta}_{\mathcal{R}}$ denote the restrictions of $\mathbf{\Delta}$ to the Top and remaining positions. Both restrictions are zero padded to the full matrix shape. Consequently, $\mathbf{\Delta}_{\mathcal{T}} + \mathbf{\Delta}_{\mathcal{R}} = \mathbf{\Delta}$ and their supports do not overlap. For each $i \in \mathcal{R}$, let $M_i \in \{0, 1\}$ denote its Coverage mask. Ideal proportional allocation has $\mathbb{E}[M_i] = \eta$ for every remaining entry. Under this idealization, WDP+ gives

$$\widetilde{\Delta}_i = \begin{cases} \Delta_i & i \in \mathcal{T}, \\ \dfrac{M_i}{\kappa}\Delta_i & i \in \mathcal{R}. \end{cases} \tag{S4}$$

This estimator uses a scale of one on the Top branch and $1/\kappa$ on the Coverage branch, matching the rule defined in the main paper. We treat $\mathbf{\Delta}$ as fixed, and the only remaining randomness comes from uniform Coverage sampling within each bin. All expectations below are conditional on the Top set, any ties in Top selection, the bin boundaries, and the tie-breaking used by largest-remainder rounding. The ideal analysis first sets every bin inclusion probability to η. We then quantify the correction introduced by integer allocation.

D.2 BIAS FROM GLOBAL SCALING

The Top entries are retained deterministically and remain unchanged. Every Coverage entry instead satisfies

$$\mathbb{E}[\widetilde{\Delta}_i] = \frac{\mathbb{E}[M_i]}{\kappa}\Delta_i = \frac{\eta}{\kappa}\Delta_i. \tag{S5}$$

It follows that

$$\mathbb{E}[\widetilde{\mathbf{\Delta}}] = \mathbf{\Delta}_{\mathcal{T}} + \frac{\eta}{\kappa}\mathbf{\Delta}_{\mathcal{R}}. \tag{S6}$$

The ideal bias is supported only on the remaining positions and can be written as

$$\begin{aligned} \mathbf{b}_0 &= \text{Bias}(\widetilde{\mathbf{\Delta}}) = \mathbb{E}[\widetilde{\mathbf{\Delta}}] - \mathbf{\Delta} = -\gamma\mathbf{\Delta}_{\mathcal{R}}, \\ \gamma &= 1 - \frac{\eta}{\kappa} = \frac{\lambda(1 - \kappa)}{1 - \lambda\kappa}. \end{aligned} \tag{S7}$$

Since $0 \le \gamma \le \lambda$, the global scale introduces a uniform shrinkage over the non-Top entries without changing their relative directions. The normalized bias obeys

$$\frac{\|\mathrm{Bias}(\widetilde{\boldsymbol{\Delta}})\|_F}{\|\boldsymbol{\Delta}\|_F} = \gamma \frac{\|\boldsymbol{\Delta}_{\mathcal{R}}\|_F}{\|\boldsymbol{\Delta}\|_F}. \tag{S8}$$

To verify the bound on γ, the case $\lambda = 0$ is immediate. For $\lambda > 0$, we have

$$\frac{\gamma}{\lambda} = \frac{1-\kappa}{1-\lambda\kappa} \le 1 \tag{S9}$$

because $\lambda \le 1$. The disjoint supports of the Top and remaining sets then give Eq. (S8).

The role of the deterministic branch becomes more explicit through the tail-energy ratio

$$\varepsilon_{\mathcal{R}} = \frac{\|\boldsymbol{\Delta}_{\mathcal{R}}\|_F^2}{\|\boldsymbol{\Delta}\|_F^2}. \tag{S10}$$

Using this notation, the normalized squared bias is

$$\frac{\|\mathrm{Bias}(\widetilde{\boldsymbol{\Delta}})\|_F^2}{\|\boldsymbol{\Delta}\|_F^2} = \gamma^2 \varepsilon_{\mathcal{R}}. \tag{S11}$$

The bias is therefore further reduced when the Top branch captures a large fraction of the delta energy.

D.3 Sampling Variance and the Order of the Bias

The identity $M_i^2 = M_i$ gives the entry-wise variance

$$\mathrm{Var}(\widetilde{\Delta}_i) = \frac{\eta(1-\eta)}{\kappa^2}\Delta_i^2. \tag{S12}$$

Summing these variances over the remaining entries yields the Frobenius variance

$$\begin{aligned} V_0 &= \mathbb{E}\left[\left\|\widetilde{\boldsymbol{\Delta}} - \mathbb{E}[\widetilde{\boldsymbol{\Delta}}]\right\|_F^2\right] \\ &= \frac{\eta(1-\eta)}{\kappa^2}\|\boldsymbol{\Delta}_{\mathcal{R}}\|_F^2. \end{aligned} \tag{S13}$$

No covariance terms appear in Eq. (S13) because the squared Frobenius norm is a sum of entry-wise squared deviations and contains no cross products between distinct entries. Substituting Eq. (S3) and $1-\eta = (1-\kappa)/(1-\lambda\kappa)$ gives

$$\frac{\eta(1-\eta)}{\kappa^2} = \frac{(1-\lambda)(1-\kappa)}{\kappa(1-\lambda\kappa)^2}. \tag{S14}$$

Combining Eqs. (S7), (S13), and (S14) gives the exact ratio

$$\frac{\|\mathrm{Bias}(\widetilde{\boldsymbol{\Delta}})\|_F^2}{V_0} = \frac{\kappa\lambda^2(1-\kappa)}{1-\lambda}. \tag{S15}$$

For every fixed $\lambda < 1$, this ratio is $O(\kappa)$ and approaches zero as $\kappa \to 0$. This is the precise sense in which the global-scaling bias is negligible in the ultra-sparse regime. The claim concerns squared bias relative to stochastic sampling variance and does not assert that the entry-wise bias vanishes in absolute value.

Evaluating Eq. (S15) at $\lambda = 0.8$, which upper-bounds every $\lambda \le 0.8$, gives

$$\frac{\|\mathrm{Bias}(\widetilde{\boldsymbol{\Delta}})\|_F^2}{V_0} \le \begin{cases} 3.17\% & \kappa = 0.01, \\ 0.32\% & \kappa = 0.001. \end{cases} \tag{S16}$$

These keep ratios correspond to $100\times$ and $1000\times$ compression. The condition $\lambda < 1$ excludes the Top-only endpoint, where the Coverage branch and its scale are absent.

D.4 Variance and MSE under Alternative Scaling

Consider a remaining entry Δ_i. Scaling its mask by $1/\eta$ would produce an unbiased estimator with variance

$$V_{\mathrm{unbiased},i} = \frac{1-\eta}{\eta}\Delta_i^2. \tag{S17}$$

The WDP+ scale $1/\kappa$ instead gives

$$V_{\mathrm{global},i} = \frac{\eta(1-\eta)}{\kappa^2}\Delta_i^2. \tag{S18}$$

Their ratio is

$$\frac{V_{\mathrm{global},i}}{V_{\mathrm{unbiased},i}} = \left(\frac{\eta}{\kappa}\right)^2 \le 1. \tag{S19}$$

Thus, a nonzero Top budget makes global scaling strictly lower variance than Coverage-unbiased scaling.

The entry-wise mean-squared errors under the two scales are

$$R_{\mathrm{global},i} = \left(1 - \frac{2\eta}{\kappa} + \frac{\eta}{\kappa^2}\right)\Delta_i^2, \tag{S20}$$

$$R_{\mathrm{unbiased},i} = \frac{1-\eta}{\eta}\Delta_i^2. \tag{S21}$$

Their difference is

$$R_{\mathrm{global},i} - R_{\mathrm{unbiased},i} = -\frac{(\kappa-\eta)(\kappa+\eta-2\kappa\eta)}{\kappa^2\eta}\Delta_i^2 \le 0. \tag{S22}$$

Both factors in the numerator are nonnegative because $0 < \eta \le \kappa < 1$. The WDP+ scale therefore acts as shrinkage. Its bias is more than offset by the reduction in variance, which results in lower total MSE than the unbiased scale $1/\eta$.

We can also compare ideal WDP+ with standard drop-and-rescale at the same total keep ratio. Independent sampling with keep probability κ and scale $1/\kappa$ has expected Frobenius error

$$R_{\mathrm{DR}} = \frac{1-\kappa}{\kappa}\|\mathbf{\Delta}\|_F^2. \tag{S23}$$

For ideal WDP+, the Top entries have zero error. Summing the Coverage bias and variance gives

$$R^{(0)}_{\mathrm{WDP}+} = \frac{(1-\kappa)(1-\lambda+\lambda\kappa)}{\kappa(1-\lambda\kappa)}\|\mathbf{\Delta}_{\mathcal{R}}\|_F^2. \tag{S24}$$

The ratio between the two errors is

$$\frac{R^{(0)}_{\mathrm{WDP}+}}{R_{\mathrm{DR}}} = \frac{\|\mathbf{\Delta}_{\mathcal{R}}\|_F^2}{\|\mathbf{\Delta}\|_F^2}\frac{1-\lambda+\lambda\kappa}{1-\lambda\kappa}. \tag{S25}$$

When $\kappa \le 1/2$, the second factor is at most one because

$$1-\lambda+\lambda\kappa \le 1-\lambda\kappa. \tag{S26}$$

Hence $R^{(0)}_{\mathrm{WDP}+} \le R_{\mathrm{DR}}$ throughout the $20\times$ to $1000\times$ range considered in the main paper, where $\kappa \in [0.001, 0.05]$. Under ideal proportional allocation, the bias introduced by global scaling does not increase expected weight-space MSE relative to standard unbiased drop-and-rescale at the same total keep ratio. The second factor in Eq. (S25) guarantees this nonincrease. The larger reduction typically comes from the tail-energy factor $\varepsilon_{\mathcal{R}}$ when the Top branch captures the high-energy entries.

D.5 Fixed-Budget Bin Allocation

We next restore the exact integer allocation used by WDP. Let WDP construct the bins $\{\mathcal{B}_j\}_{j=1}^{B}$ and let $B_+ = \#\{j \mid \mathcal{B}_j \neq \varnothing\} \le B$ denote the number of nonempty bins. Empty bins receive no

quota and do not participate in largest-remainder allocation. After reindexing the nonempty bins, the remaining set satisfies

$$\mathcal{R} = \bigcup_{j=1}^{B_+} \mathcal{B}_j, \qquad n_j = |\mathcal{B}_j|, \qquad \sum_{j=1}^{B_+} n_j = n. \tag{S27}$$

Writing $K = K^{\mathrm{cov}}$ gives $\eta = K/n$. The real-valued proportional quota of bin j is

$$q_j = K\frac{n_j}{n} = \eta n_j. \tag{S28}$$

Largest-remainder allocation returns an integer K_j for each bin. We denote the allocation residual by

$$e_j = K_j - q_j. \tag{S29}$$

Because $\sum_j q_j = K$, largest-remainder rounding adds one to exactly $K - \sum_j \lfloor q_j \rfloor$ bins. This construction gives

$$\sum_{j=1}^{B_+} K_j = K, \qquad |e_j| < 1. \tag{S30}$$

The global Coverage count is therefore exact rather than random.

Uniform sampling without replacement within bin j gives every entry the marginal inclusion probability

$$p_j = \frac{K_j}{n_j} = \eta + \frac{e_j}{n_j}, \qquad |p_j - \eta| < \frac{1}{n_j}. \tag{S31}$$

Let $r_j = n_j/n$ be the original bin proportion and let $\widehat{r}_j = K_j/K$ be the retained proportion. Their discrepancy satisfies

$$|\widehat{r}_j - r_j| = \frac{|e_j|}{K} < \frac{1}{K}. \tag{S32}$$

Summing across nonempty bins gives

$$\|\widehat{\mathbf{r}} - \mathbf{r}\|_1 < \frac{B_+}{K}, \qquad d_{\mathrm{TV}}(\widehat{\mathbf{r}}, \mathbf{r}) < \frac{B_+}{2K}. \tag{S33}$$

At bin boundaries, the cumulative discrepancy also obeys

$$\max_m \left| \sum_{j \le m} \widehat{r}_j - \sum_{j \le m} r_j \right| < \frac{B_+}{K}. \tag{S34}$$

The allocation-induced histogram and CDF discrepancies are consequently $O(B_+/K^{\mathrm{cov}})$ and vanish when $K^{\mathrm{cov}} \gg B_+$. This bound is crude but sufficient for convergence. It controls the retained-position distribution across the original magnitude bins rather than the magnitude histogram after rescaling.

Fixed-size sampling also removes random fluctuations in every bin count. For distinct $i, h \in \mathcal{B}_j$, uniform sampling without replacement gives

$$\mathbb{P}(M_i = M_h = 1) = \frac{K_j(K_j - 1)}{n_j(n_j - 1)} \tag{S35}$$

and

$$\mathrm{Cov}(M_i, M_h) = -\frac{p_j(1 - p_j)}{n_j - 1}. \tag{S36}$$

These pairwise expressions apply only when $n_j > 1$. The retained bin count therefore has

$$\mathrm{Var}\left(\sum_{i \in \mathcal{B}_j} M_i \right) = 0. \tag{S37}$$

Independent Bernoulli sampling with rate p_j would instead produce count variance $n_j p_j(1 - p_j)$. The fixed-budget construction removes these bin-count fluctuations, which are a direct source of histogram distortion at extreme sparsity.

D.6 WITHIN-BIN SAMPLING ERROR

The remaining randomness arises from which entries are selected within each bin. For a bin with $K_j > 0$ and $n_j > 1$, let z_i be arbitrary fixed values and define the population and sampled means as

$$\overline{z}_j = \frac{1}{n_j} \sum_{i \in \mathcal{B}_j} z_i, \qquad \widehat{z}_j = \frac{1}{K_j} \sum_{i \in \mathcal{C}_j} z_i. \tag{S38}$$

Uniform sampling without replacement gives

$$\mathbb{E}[\widehat{z}_j] = \overline{z}_j, \qquad \mathrm{Var}(\widehat{z}_j) = \frac{1 - p_j}{K_j} S_j^2, \tag{S39}$$

where the finite-population variance is

$$S_j^2 = \frac{1}{n_j - 1} \sum_{i \in \mathcal{B}_j} (z_i - \overline{z}_j)^2. \tag{S40}$$

For the magnitude statistic $z_i = |\Delta_i|$, every value in an equal-width bin lies within an interval of width

$$h = \frac{\delta_{\max}^{\mathcal{R}} - \delta_{\min}^{\mathcal{R}}}{B}. \tag{S41}$$

The finite-population variance and sampled-mean variance are then bounded by

$$\begin{aligned} S_j^2 &\le \frac{n_j}{n_j - 1} \frac{h^2}{4}, \\ \mathrm{Var}(\widehat{z}_j) &\le \frac{1 - p_j}{K_j} \frac{n_j}{n_j - 1} \frac{h^2}{4}. \end{aligned} \tag{S42}$$

When $K_j = 0$, no sampled mean is formed and no within-bin sampling randomness remains. Every entry in that bin is deterministically omitted, and its error is included in the exact bias and MSE below. When $n_j = 1$, the finite-population variance is not formed. If $K_j = 1$, the sampled mean equals the population mean and the within-bin variance is zero. If $K_j = 0$, the deterministic omission case applies. The between-bin composition error is deterministically $O(B_+/K^{\mathrm{cov}})$, while the standard deviation of the remaining within-bin error is $O(h/\sqrt{K_j})$. Increasing B narrows each interval but also reduces the number of samples assigned to it. This trade-off explains why a small or moderate number of intervals is sufficient and why increasing the count to 64 does not consistently improve the reported results.

D.7 FINITE-BUDGET CORRECTIONS

The ideal global-scaling analysis assumed $p_j = \eta$. Largest-remainder allocation instead gives

$$p_j = \eta + \frac{e_j}{n_j}, \qquad |e_j| < 1, \qquad \sum_j e_j = 0. \tag{S43}$$

The exact bias under these probabilities is

$$\begin{aligned} \mathbf{b} &= \sum_j \left(\frac{p_j}{\kappa} - 1 \right) \boldsymbol{\Delta}_{\mathcal{B}_j} \\ &= -\gamma \boldsymbol{\Delta}_{\mathcal{R}} + \frac{1}{\kappa} \sum_j \frac{e_j}{n_j} \boldsymbol{\Delta}_{\mathcal{B}_j}. \end{aligned} \tag{S44}$$

The first term is the global-scaling bias. The second term is the correction caused by integer allocation.

Let $n_{\min} = \min_{1 \le j \le B_+} n_j$. The bins have disjoint support, so the correction obeys

$$\|\mathbf{b} + \gamma \boldsymbol{\Delta}_{\mathcal{R}}\|_F \le \frac{1}{\kappa n_{\min}} \|\boldsymbol{\Delta}_{\mathcal{R}}\|_F. \tag{S45}$$

Define the smallest expected count of a nonempty bin as $q_{\min} = \eta n_{\min}$. Then

$$\frac{1}{\kappa n_{\min}} = \frac{\eta/\kappa}{q_{\min}} \le \frac{1}{q_{\min}}. \tag{S46}$$

The additional bias from integer allocation is therefore $O(q_{\min}^{-1})$ relative to the tail norm.

Although masks within a bin are negatively correlated, their covariances do not enter the Frobenius variance because the squared norm is a sum of squared coordinate deviations without cross products. The exact Frobenius variance is

$$V = \frac{1}{\kappa^2} \sum_j p_j(1-p_j) \|\mathbf{\Delta}_{\mathcal{B}_j}\|_F^2. \tag{S47}$$

The function $p(1-p)$ is one-Lipschitz on $[0,1]$. Equations (S31) and (S13) therefore give

$$|V - V_0| \le \frac{1}{\kappa^2 n_{\min}} \|\mathbf{\Delta}_{\mathcal{R}}\|_F^2. \tag{S48}$$

Relative to the ideal variance, this becomes

$$\frac{|V - V_0|}{V_0} \le \frac{1}{\eta(1-\eta) n_{\min}} = \frac{1}{(1-\eta) q_{\min}}. \tag{S49}$$

The variance correction vanishes when every active bin receives many expected samples.

The exact MSE can be written as

$$R = \sum_j \left(1 - \frac{2p_j}{\kappa} + \frac{p_j}{\kappa^2}\right) \|\mathbf{\Delta}_{\mathcal{B}_j}\|_F^2. \tag{S50}$$

Replacing p_j by η defines the ideal value R_0. Since the coefficient is linear in p_j, the same residual bound gives

$$|R - R_0| \le \frac{|1-2\kappa|}{\kappa^2 n_{\min}} \|\mathbf{\Delta}_{\mathcal{R}}\|_F^2. \tag{S51}$$

After normalization by the ideal MSE, this correction is $O(q_{\min}^{-1})$. More explicitly, Eqs. (S24) and (S51) give

$$\frac{|R - R_0|}{R_0} \le \frac{|1-2\kappa|(1-\lambda)}{(1-\kappa)(1-\lambda+\lambda\kappa)} \frac{1}{q_{\min}}. \tag{S52}$$

The outer count definitions also use rounding. They satisfy

$$K^* = \mathrm{round}(\kappa N), \qquad K^{\mathrm{top}} = \mathrm{round}(\lambda\kappa N), \tag{S53}$$

$$K^{\mathrm{cov}} = K^* - K^{\mathrm{top}}. \tag{S54}$$

Defining K^{cov} by this difference preserves the exact total budget K^*. The continuous Coverage budget $(1-\lambda)\kappa N$ is its unrounded approximation. Each directly rounded rate differs from its real-valued target by at most $1/(2N)$, and the induced Coverage rate differs by at most $1/N$. Relative to the global keep ratio, this changes the scale-level quantities by

$$O\left(\frac{1}{\kappa N}\right) = O\left(\frac{1}{K^*}\right). \tag{S55}$$

The outer rounding correction thus vanishes when the retained matrix budget grows.

D.8 Asymptotic Regime

The preceding bounds combine under a direct asymptotic condition. Suppose N grows while

$$\kappa \to 0, \qquad \kappa N \to \infty. \tag{S56}$$

Assume the budget split obeys $\lambda \le 1 - \delta$ for a fixed $\delta > 0$ and that

$$\frac{B_+}{K^{\mathrm{cov}}} \to 0, \qquad q_{\min} \to \infty. \tag{S57}$$

For every fixed $\lambda \in (0, 1-\delta]$, Eqs. (S7) and (S14) give

$$\|\mathbf{b}_0\|_F^2 = \Theta\big(\|\mathbf{\Delta}_{\mathcal{R}}\|_F^2\big), \tag{S58}$$

$$V_0 = \Theta\big(\kappa^{-1} \|\mathbf{\Delta}_{\mathcal{R}}\|_F^2\big). \tag{S59}$$

When $\lambda = 0$, the global-scaling bias is exactly zero. Equations (S15), (S33), (S49), and (S55) then yield

$$\frac{\|\mathbf{b}_0\|_F^2}{V_0} = O(\kappa) \to 0, \tag{S60}$$

$$d_{\mathrm{TV}}(\widehat{\mathbf{r}}, \mathbf{r}) = O\left(\frac{B_+}{K^{\mathrm{cov}}}\right) \to 0. \tag{S61}$$

$$\frac{|V - V_0|}{V_0} = O(q_{\min}^{-1}) \to 0, \tag{S62}$$

$$\text{outer rounding error} = O((\kappa N)^{-1}) \to 0. \tag{S63}$$

In this regime, the squared global-scaling bias relative to sampling variance is $O(\kappa)$. This relative statement does not claim that the absolute compression error vanishes. When the tail energy remains nonzero, the sampling variance grows as $\Theta(\kappa^{-1})$ relative to that energy, while the squared bias remains a lower-order term. Histogram error, integer allocation error, and outer rounding error vanish with the retained budget under the stated conditions. The dominant residual term is the sampling variance of sparse approximation. The global scale also gives lower variance and lower MSE than the Coverage-unbiased scale $1/\eta$.

D.9 Implication for Linear-Layer Outputs

The weight-space guarantees also control the output of each compressed linear module. Let $\mathbf{A} = \widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}$ denote the weight error. For any fixed input $\mathbf{x}$,

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{A}\mathbf{x}\|_2^2\right] &\le \|\mathbf{x}\|_2^2 \mathbb{E}\left[\|\mathbf{A}\|_2^2\right] \\ &\le \|\mathbf{x}\|_2^2 \mathbb{E}\left[\|\mathbf{A}\|_F^2\right]. \end{aligned} \tag{S64}$$

The expected output bias satisfies

$$\begin{aligned} \|\mathbb{E}[\mathbf{A}\mathbf{x}]\|_2 &\le \|\mathbf{x}\|_2 \, \|\mathbb{E}[\mathbf{A}]\|_2 \\ &\le \|\mathbf{x}\|_2 \, \|\mathbb{E}[\mathbf{A}]\|_F. \end{aligned} \tag{S65}$$

The spectral-norm terms provide the sharper bounds, while the Frobenius terms connect directly to the preceding analysis. These results are weight-space guarantees for a single compressed linear module. They do not directly guarantee task-level performance. Under an additional local Lipschitz condition on the downstream network, the same terms can be propagated to upper bounds on hidden-state or loss perturbation.

E Further Analysis

This section reports additional inference-seed sensitivity results and the per-benchmark interval-construction results summarized in the main paper.

E.1 Per-Benchmark Results of Inference-Seed Sensitivity

Results from running three different seeds on the same configuration in $\mathcal{X}$FT at $100\times$ are reported in **Table S8**. Both DeRS-SP and UltraSparse vary only the inference seed. Each benchmark cell reports pass@1 as a percentage, and Avg is the mean over the four code-generation benchmarks.

E.2 Uni-MoE Interval-Construction Results

Table S9 reports the per-benchmark Uni-MoE results for the four WDP interval constructions at $100\times$ expert-delta compression. The Abs. Width row is the default setting reported in the main-paper Uni-MoE table. Here, *Abs.*/*Signed* denotes binning on $|\boldsymbol{\Delta}|$/signed $\boldsymbol{\Delta}$, and *Width*/*Count* denotes equal-width/equal-count bins.

Table S8: Inference-seed sensitivity of DeRS-SP and UltraSparse on $\mathcal{X}$FT at $100\times$. For both methods, only the inference seed varies, while all other settings remain fixed. HumanEval, HumanEval+, MBPP, and MBPP+ are evaluated by pass@1, with Avg denoting their arithmetic mean.

Method	Infer. Seed	HumanEval	HumanEval+	MBPP	MBPP+	Avg
DeRS-SP with varying inference seed at $100\times$						
DeRS-SP	Seed 1	57.9	53.0	55.1	46.9	53.23
DeRS-SP	Seed 2	57.9	53.0	55.1	47.1	53.31
DeRS-SP	Seed 3	57.3	53.7	55.9	46.4	53.33
UltraSparse with varying inference seed at $100\times$						
UltraSparse	Seed 1	61.59	57.32	65.66	54.89	59.86
UltraSparse	Seed 2	62.80	60.37	62.41	51.38	59.24
UltraSparse	Seed 3	62.80	60.98	63.16	50.63	59.39

Table S9: WDP interval-construction ablation on Uni-MoE at $100\times$ expert-delta compression. Abs. and Signed denote binning by absolute and signed delta values, while Width and Count denote equal-width and equal-count intervals. The shaded Abs. Width row is the default setting. Avg is the arithmetic mean of OKVQA, AOKVQA, VQAv2, MMBench-Audio, POPE, SEED, RACE-Audio, and EHSL, where RACE-Audio and EHSL first average their two subsets, as in the main-paper Uni-MoE table.

Construction	OKVQA	AOKVQA	VQA^{v2}	MMBench^{A}	POPE	Seed	RACE^{A}		EHSL		Avg
							Middle	High	Long	Short	
Abs. Width	64.29	58.78	73.67	52.80	79.73	46.09	47.70	43.40	46.00	30.00	57.36
Abs. Count	59.69	54.15	72.14	40.37	80.62	45.00	38.65	37.16	34.67	24.00	52.40
Signed Width	59.12	57.47	73.00	41.61	82.99	39.86	38.65	36.48	46.00	38.00	54.20
Signed Count	59.55	55.28	72.67	43.17	81.80	46.50	38.44	35.56	40.00	32.00	54.00

E.3 MULTIPL-MOE INTERVAL-CONSTRUCTION RESULTS

Tables S10 and S11 report the full six-language HumanEval and MBPP results for the four WDP interval constructions at $100\times$ expert-delta compression. The Abs. Width row is the default setting, while the remaining rows vary the binning domain and width/count construction.

Table S10: WDP interval-construction ablation on MultiPL-MoE for multilingual HumanEval at $100\times$ expert-delta compression. The four rows combine absolute or signed delta values with equal-width or equal-count intervals. All language scores are pass@1 and Avg is the arithmetic mean over Python, Java, C++, Rust, Go, and Ruby. The shaded Abs. Width row is the default setting.

Construction	Python	Java	C++	Rust	Go	Ruby	Avg
Abs. Width	12.20	12.03	11.18	4.49	3.90	4.35	8.03
Abs. Count	12.80	10.13	11.18	5.13	3.90	4.35	7.91
Signed Width	13.40	10.76	11.18	4.49	3.90	4.35	8.01
Signed Count	12.20	10.13	11.18	3.85	3.90	4.35	7.60

Table S11: WDP interval-construction ablation on MultiPL-MoE for multilingual MBPP at $100\times$ expert-delta compression. The four rows combine absolute or signed delta values with equal-width or equal-count intervals. All language scores are pass@1 and Avg is the arithmetic mean over Python, Java, C++, Rust, Go, and Ruby. The shaded Abs. Width row is the default setting.

Construction	Python	Java	C++	Rust	Go	Ruby	Avg
Abs. Width	24.90	14.77	19.65	10.45	13.10	9.57	15.41
Abs. Count	23.80	14.77	18.89	9.04	13.90	9.57	15.00
Signed Width	24.90	14.77	19.65	9.32	13.10	9.57	15.22
Signed Count	25.40	14.25	19.14	9.04	13.37	9.32	15.09