MoBiQuant: Mixture-of-Bits Quantization for Token-Adaptive Any-Precision LLM

Dongwei Wang[1*] **Jinhee Kim**[2*] **Seokho Han**[3*]
Denis Gudovskiy[4] **Yohei Nakata**[5] **Tomoyuki Okuno**[5] **KhayTze Peong**[5]
Kang Eun Jeon[6†] **Jong Hwan Ko**[3†] **Yiran Chen**[2†] **Huanrui Yang**[1†]

[1]University of Arizona [2]Duke University [3]Sungkyunkwan University
[4]Panasonic AI Lab [5]Panasonic Holdings Corporation [6]DGIST
{dongweiw,huanruiyang}@arizona.edu
jinhee.kim@duke.edu beppa2396@skku.edu

Abstract

Dynamic runtime latency and memory constraints necessitate flexible large language model (LLM) deployment, where an LLM can be inferred with various quantization precisions based on available computational resources. Recent work on such any-precision quantization either relies on hardware-inefficient vector quantization or induces additional scaling factors when switching between bit-widths. Meanwhile, existing post-training quantization (PTQ) methods calibrated for a fixed low precision show poor generalizability under runtime precision change. In this work, we attribute the source of poor generalization across bit-widths to a precision-dependent *outlier migration* phenomenon where the distribution of PTQ-sensitive tokens changes across precisions. Motivated by this observation, we propose `MoBiQuant`, a novel any-precision Mixture-of-Bits quantization framework that adjusts weight precision for flexible LLM inference based on token sensitivity. Specifically, we propose a many-in-one recursive residual quantization that can iteratively reconstruct higher-precision weights at runtime and mitigates *outlier migration* with a token-aware router to dynamically select the optimal inference precision of each token. Extensive experiments show that `MoBiQuant` matches or surpasses frontier single-precision PTQ while exhibiting strong elasticity, achieving significant memory savings and throughput gains of up to $1.34\times$ over state-of-the-art any-precision methods. Our code is released at this `link`.

1 Introduction

Recent deployments of large language models (LLMs) [1–5] typically use low-precision quantization to meet latency and memory constraints. Although ongoing work on post-training quantization (PTQ) [6, 7] constantly improves LLM performance at predetermined fixed precisions, the tradeoff between precision and performance remains critical.

The impact of such a tradeoff is more significant for edge devices where available computational resources can vary dynamically at runtime. Under sufficient resources, the device can run the model at higher precision for better performance, while under resource contention from other applications, it may need to reduce precision to satisfy latency and memory constraints. Storing LLM weights for each precision is impractical, and, hence, recent research explores any-precision quantization where a single model can support multiple precisions [8–10]. However, existing approaches remain limited

*Equal contribution.
†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

in practice, as they either rely on hardware-unfriendly vector quantization (VQ) [8] or introduce additional calibration parameters that lead to training and inference overhead [9]. Moreover, they often lack flexibility for runtime adaptation, requiring offline repacking and kernel redeployment when switching precisions [10].

In this work, we overcome these limitations and propose a more advanced any-precision framework with three key properties: (1) it supports hardware-friendly scalar quantization, (2) it introduces minimal overhead compared to a single set of PTQ calibration parameters, and (3) it enables efficient runtime precision switching without offline repacking or kernel relaunching. To achieve this, we start by investigating the root cause of why existing single-precision scalar PTQ methods fail to generalize across bit-widths with fixed calibration parameters.

We identify a critical phenomenon of precision-dependent *"outlier migration"* in LLMs. Although it is well-established that outliers in activations govern quantization errors [11], we find that the specific subset of tokens that are responsible for high quantization errors are not static for each precision. Consequently, static PTQ methods that calibrate quantization parameters (e.g., scaling factors or clipping ranges) for a single precision, often fail to generalize. They inevitably overfit to an improper set of tokens when the bit-width changes, leading to suboptimal performance in any-precision inference setting.

Motivated by this, we propose `MoBiQuant`, an any-precision quantization framework that mitigates "*outlier migration*" with a single set of calibration parameters by incorporating token-adaptive precision adjustment during calibration. `MoBiQuant` learns to dynamically assign proper inference precision to tokens on the fly based on their sensitivity while maintaining overall inference cost within an average budget. Our approach consists of two core components. The first one is `MoBiSlice`, a "many-in-one" recursive residual quantization that decomposes weights into a low-precision base and successive residual bit slices. This allows a single model to support multiple precisions (e.g., 2, 4, 6-bit) via activating fixed-precision bit slices (e.g., 2-bit slice) at runtime, thereby supporting any-precision inference with a single type of low-precision bit slice GEMM kernel without the need for repacking and redeployment. The second component is `MoBiRoute`, a lightweight learnable router that implements token-aware inference precision selection. Acting as a gate, the `MoBiRoute` dynamically activates the optimal number of `MoBiSlice` residual components for each token based on learned token sensitivity behavior and current resource conditions, thereby preventing the calibration parameters from overfitting to the outlier distribution of a single precision. Runtime precision switching is achieved by adjusting the routing threshold, without introducing extra scaling factors and training overhead.

In summary, our paper makes the following contributions:

- The `MoBiSlice` recursive quantization explicitly reconstructs higher precisions using low-precision bit slices. It enables a single model to flexibly realize multiple precision levels via hierarchical reconstruction, without requiring offline repacking and redeployment.
- The `MoBiRoute` router mitigates the bit-dependent *outlier migration* phenomenon by dynamically selecting the number of bit slices for each token. It also enables efficient runtime precision switching by a threshold adjustment, eliminating the need to store additional scaling factors and extra training overhead.
- Extensive experiments demonstrate that `MoBiQuant` generalizes effectively across bit-widths, matching or surpassing the performance of state-of-the-art single-precision PTQ methods. Our tailored kernel design achieves $1.34\times$ higher throughput than state-of-the-art any-precision methods [8, 9]. Compared to deploying multiple models, `MoBiQuant` reduces the memory footprint by $3.5\times$.

2 Related Work

Single-precision PTQ. Most existing methods focus on statically-defined quantization schemes, where LLM performance is particularly optimized for the selected bit-width configuration. Representative works include: GPTQ [12] with second-order information to minimize quantization error, OmniQuant [7] with learnable weight clipping to achieve superior results, SmoothQuant [13] and SpinQuant [14] with special linear transformations to mitigate peaking outliers, and AWQ [15] for selectively quantizing salient weights. Recent advancements like the QuIP family of methods [16, 17]

achieve extreme compression ratios using VQ, yet these methods suffer from limited practicality, as their VQ scheme is not well supported by modern GEMM kernels for runtime acceleration. While these methods achieve strong performance at a given precision, they still exhibit an inherent trade-off between accuracy and efficiency across bit-widths. Moreover, they lack flexibility, as switching between precisions typically requires offline recalibration, making them unsuitable for dynamic edge scenarios with changing resource constraints.

Any-precision PTQ. To address the inflexibility of static schemes, recent works propose any-precision quantization, enabling a single model to support multiple bit-widths for dynamic system requirements. AnyPrecisionLLM [8] incrementally upscales a low-bit model to higher precisions, AnyBCQ [9] expands base quantization via per-precision scaling factors, and MatQuant [10] derives low-bit models by truncating higher-bit representations. However, these approaches remain limited in practice: they either rely on hardware-unfriendly vector quantization [8], introduce additional overhead from per-precision calibration parameters [9], or require offline repacking when increasing precision [10]. Moreover, they lack efficiency for runtime adaptation, as switching between precisions requires loading extra scaling factors and launching distinct kernels (e.g., from 3-bit to 4-bit). These limitations motivate our `MoBiQuant` framework, which mitigates precision-dependent outlier migration and enables efficient runtime precision switching without additional overhead.

3 Motivation

Limitations of static PTQ. Typical PTQ calibration process introduces a set of parameters $\mathbf{\Theta}_q$ to minimize the quantization error of the pretrained weights $\mathbf{W}$ at a target bit-width b. For example, OmniQuant [7] learns to clip and rescale $\mathbf{W}$ and SpinQuant [14] learns to rotate $\mathbf{W}$ to reduce outliers, respectively. Given a calibration set $\mathcal{X} = \{\mathbf{X}_i\}_{i=1}^{N}$, where each sequence $\mathbf{X}_i \in \mathbb{R}^{T\times d}$ consists of T tokens, the calibration process for static PTQ methods can be formulated as

$$\arg\min_{\mathbf{\Theta}_q} \mathbb{E}_{\mathbf{X}\sim\mathcal{X}} \mathcal{D}\big(\mathbf{X}, Q(\mathbf{W}|\mathbf{\Theta}_q, b)\big), \tag{1}$$

where $Q(\cdot)$ represents the weight quantizer and $\mathcal{D}(\cdot)$ denotes the quantization error function, respectively.

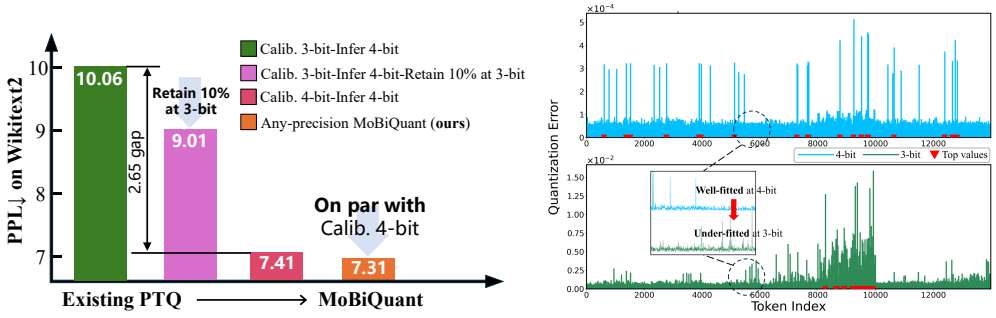

Figure 1: (Left) Existing PTQ methods poorly generalize under calibration–inference precision mismatch, e.g., inference at 4-bit with 3-bit calibrated parameters (green vs. red). Incorporating token-aware bit adjustment partially mitigates this degradation (pink). Our `MoBiQuant` improves generalization across precisions (orange). (Right) Per-token quantization error distribution at layer 5 in LLaMA3-8B: quantization outliers are highly non-uniform at each bit-width. Tokens well-fitted at 4-bit can become outliers at 3-bit, while 4-bit outliers may not be the primary error source at 3-bit.

The Eq. (1) objective causes the parameters $\mathbf{\Theta}_q$ to overfit for the selected target bit-width b. We illustrate this limitation in Fig. 1 (left): if we use the calibration parameters of 3-bit PTQ to quantize the LLaMA-3-8B model for 4-bit inference, the perplexity increases by 2.65 points compared to a 4-bit calibrated model. This clearly highlights the lack of generalization across quantization precisions.

Analysis of the poor generalization. Previous work [18] has shown that the token sensitivity is highly non-uniform. Consequently, the calibration process optimizes $\mathbf{\Theta}_q$ to specifically fit the distribution of sensitive tokens using the overall objective in (1). We pinpoint the poor generalization

of calibration parameters to the distribution shift of outlier tokens. We plot the per-token quantization error distributions when the target weight precisions are 3/4-bit for the same query in Fig. 1 (right) and observe very different distributions. Specifically, tokens that are accurately represented by 4-bit weights may emerge as dominant outliers for 3-bit weights, whereas outliers with 4-bit precision may no longer be the primary source of quantization error for 3-bit PTQ. We define this phenomenon as *"outlier migration"* and it appears universally across different PTQ calibration methods either gradient-based OmniQuant [7] or Hessian-based AWQ [15] (see Appendices E.1 through E.3).

To further motivate our approach, we revisit the 3-bit calibrated 4-bit inference experiment, but now retain the top 10% of token outliers in the 3-bit calibration to be inferred with 3-bit weights. Counterintuitively, it shows that inferring tokens at a lower precision can yield higher overall performance (pink bar). This observation suggests a promising direction: ***token-adaptive precision adjustment*** *can improve the generalization of PTQ calibration parameters across bit-widths.*

4 The MoBiQuant Framework

Inspired by our previous observations, we present `MoBiQuant`, an any-precision quantization framework that mitigates *outlier migration* through token-aware bit-width adaptation while enabling efficient runtime precision switching. An overview of the framework is shown in Fig. 2.

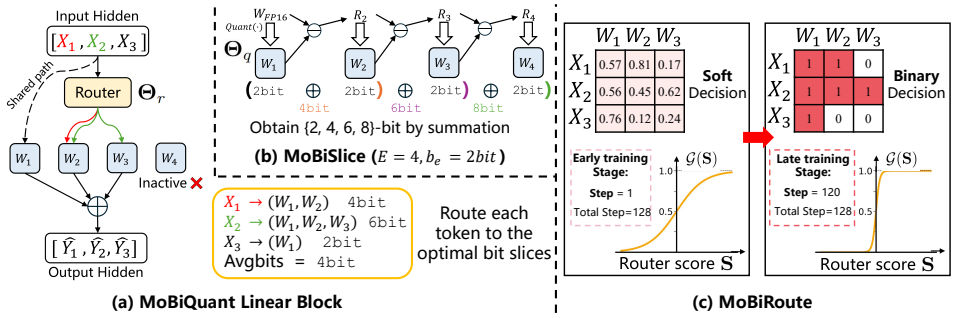

Figure 2: `MoBiQuant` enables token-adaptive any-precision inference for linear blocks in LLMs (a). It consists of two main components: `MoBiSlice` that supports multiple precisions by decomposing FP16 weights into bit slices via recursive quantization (b); and `MoBiRoute` with token-adaptive precision adjustment that is trained to route each token to the optimal bit slice under a budget (c).

4.1 Bit Slicing via Recursive Quantization

Achieving token-level granularity requires a model to efficiently support multiple quantization levels (e.g., 2–8-bit range). A naïve solution with separate models for each precision incurs high memory overhead. Existing any-precision approaches [8, 19, 10] explicitly derive multiple precision levels from a single seed model by weight slicing or scaling. However, each precision remains a monolithic unit: switching a token to a different precision requires an offline reloading of the seed model and re-slicing, followed by redeployment. This is ill-suited for latency-sensitive settings where all adaptation is expected to occur at runtime.

To address this limitation, we propose a novel many-in-one quantization scheme `MoBiSlice`, where multiple precisions can be efficiently obtained at runtime. Our `MoBiSlice` decomposes the weight matrix $\mathbf{W}$ of each linear layer in the LLM into E slices $\{\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_E\}$, each containing a slice of b_e bits from the quantized weight. This decomposition is implemented by recursively quantizing the residuals as

$$\mathbf{R}_1 = \mathbf{W},\ \mathbf{W}_e = Q(\mathbf{R}_e \mid \mathbf{\Theta}_q, b_e),\ \text{and}\ \mathbf{R}_{e+1} = \mathbf{R}_e - \mathbf{W}_e,\ \forall e \geq 1, \tag{2}$$

where the first shared most significant bit (MSB) slice $\mathbf{W}_1$ is obtained by quantizing the original $\mathbf{W}$ and the remaining slices $\mathbf{W}_{e>1}$ quantize the residuals $\mathbf{R}_e$ between the weight and existing slices in a recursive fashion. All slices are quantized using scalings derived from the shared $\mathbf{\Theta}_q$, where the scaling factor of the $(e+1)$-th slice is obtained by dividing that of the e-th slice by 2^{b_e-1}. This avoids the training overhead of multiple sets of scaling factors in existing methods [8, 9]. Moreover,

dequantization can be performed via efficient shift-and-add operations (Sec. 4.3). Consequently, a weight $\mathbf{W}^{(b)}$ at a target precision b can be constructed by summing a selection of k bit slices as

$$\mathbf{W}^{(b)} = \sum_{e=1}^{k} \mathbf{W}_e, \text{ where } b = \sum_{e=1}^{k} b_e. \tag{3}$$

In this paper, we adopt a default configuration with $E = 4$ slices, each quantized to 2-bit, as shown in Fig. 2 (b). In this way, weights with 4-bit and 6-bit precision can be constructed using 2 and 3 slices, respectively. This design enables any-precision inference using a fixed 2-bit kernel, thereby avoiding kernel relaunch overhead. In practice, `MoBiQuant` naturally learns to select the prefix $2k$-bit slices for inference without explicit constraints; further details are provided in Appendix B. `MoBiSlice` can be paired with existing PTQ pipelines, and we use OmniQuant [7] as our PTQ backbone.

4.2 Runtime Precision Switching via Learned Token Routing

Given bit slices provided by `MoBiSlice`, we propose a lightweight `MoBiRoute` router that dynamically assigns a different number of slices to each token with binary decisions. Specifically, given a sequence of T tokens $\mathbf{X} \sim \mathcal{X}$, the routing score matrix $\mathbf{S} \in \mathbb{R}^{T \times E}$ is obtained as

$$\mathbf{S} = \mathcal{R}(\mathbf{X}, \mathbf{\Theta}_r), \tag{4}$$

where $\mathcal{R}(\cdot)$ is a learnable 2-layer MLP with the router parameters $\mathbf{\Theta}_r$, and each element of the score matrix $\mathbf{S}_{i,j}$ represents the confidence that the i-th token will require the j-th slice.

Two key challenges exist when training the router in Eq. (4):

Challenge 1: Learning binary routing decisions. To determine whether a token i should be routed to the j-th slice, one simple approach is to adopt a threshold-based routing strategy. However, fixed thresholds cannot adapt to a varying score distribution during the training stage. Therefore, our `MoBiRoute` learns a decision boundary implicitly using a differentiable gating function $\mathcal{G}(\cdot)$. When training is complete, $\mathcal{G}(\mathbf{S})$ acts as a learned binary mask applied to tokens. We formulate the proposed gating function as

$$\mathcal{G}(\mathbf{S}) = \sigma\left(\tau(t) \cdot \mathbf{S}\right), \ \tau(t) = \ln(L) / \left(\ln(L) - \ln(t)\right), \tag{5}$$

where σ is the conventional sigmoid, and τ is the temperature defined by the current training step t and the total number of steps L.

At the final step, the temperature $\tau(L) = \infty$, effectively converting the continuous gating function into a binary mask $\mathcal{G}(\mathbf{S}) = \mathbb{I}\left(\mathbf{S} > 0\right)$. The change in scores and the gating function during training are illustrated in Fig. 2 (c). Using the learned mask, the input token $\mathbf{X}_i$ is routed through the selected slices, and the corresponding outputs are aggregated to produce the output token $\hat{\mathbf{Y}}_i$ as

$$\hat{\mathbf{Y}}_i = \sum_{e=1}^{E} \mathbf{W}_e^{\top} \left(\mathcal{G}(\mathbf{S})_{i,e} \odot \mathbf{X}_i\right), \tag{6}$$

where input tokens are gated by the router's binary scores, and $\mathbf{W}_e$ is the bit slice with b_e bits.

Challenge 2: Generalizing across precisions. Any-precision inference requires the router to effectively accommodate a dynamic target precision budget (i.e., average bit-width). To achieve this, we apply a *target precision scheduling* mechanism in the training process to encourage the router to explore different precision combinations. Specifically, we control router convergence through a budget-aware regularization objective defined by

$$\mathcal{L}_{\text{reg}}(t) = \left(\text{AvgBits} - b(t)\right) \cdot \left\|\mathcal{G}(\mathbf{S})\right\|_1, \ b(t) = b_{\text{init}} - \left(b_{\text{init}} - b\right) \cdot \ln(t) / \ln(L), \tag{7}$$

where the term $\left(\text{AvgBits} - b(t)\right)$ adaptively encourages bit slice pruning when the current average bit-width exceeds the target, and promotes slice activation otherwise. The AvgBits in (7) is estimated during training by counting activated bit slices per token. We define entries with scores exceeding 0.5 in $\mathcal{G}(\mathbf{S})$ as active, which can be formulated as

$$\text{AvgBits} = \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{E} \mathbb{I}(\mathcal{G}(\mathbf{S})_{i,j} > 0.5)\, b_j, \tag{8}$$

where b_j denotes the bit slice size (e.g., 2 bits). The scheduling term $b(t)$ starts with a higher precision b_{init} (e.g., 8-bit) and gradually decays to the target b. This forces the router to explore a wide spectrum of token-slice assignments before converging to binary routing decisions. During training, the target b is a hyperparameter and is set to 3-bit by default in our experiments. While `MoBiQuant` remains robust across different target precisions, varying the target bit-width may introduce minor performance variations, which we analyze in Appendix D.3.

Joint optimization. Our framework consists of two sets of trainable parameters: $\mathbf{\Theta}_q$ and $\mathbf{\Theta}_r$ for `MoBiSlice` and `MoBiRoute`, respectively. They are jointly optimized during the PTQ process by

$$\arg\min_{\mathbf{\Theta}_{q,r}} \mathbb{E}_{\mathbf{X}\sim\mathcal{X}} \left[\mathcal{D}\left(\mathbf{W}^\top \mathbf{X}, \hat{\mathbf{Y}} | \mathbf{\Theta}_{q,r} \right) + \lambda \mathcal{L}_{\text{reg}} \right], \tag{9}$$

where $\mathcal{D}$ is the L_2 quantization error i.e. $\|\mathbf{Y} - \hat{\mathbf{Y}}\|_2$. We freeze all weights from the pretrained LLM and calibrate only $\mathbf{\Theta}_q$ and $\mathbf{\Theta}_r$. We also treat $\mathbf{W}_1$ as a *shared-expert slice* such that tokens always pass through for stable training. We replace all linear layers in LLM transformer blocks with the proposed `MoBiQuant` block. Furthermore, we adopt a layer-wise calibration strategy from [7]. A detailed calibration strategy is presented in the Appendix A.

Efficient runtime precision switching. At the end of training, the score is away from zero if the router is confident about selection or exclusion of a bit slice, whereas the score remains at 0 with the lack of confident decision. Importantly, this separation allows us to achieve arbitrary precision adjustment at runtime by shifting the threshold of the score conversion process. During inference, we convert the router score into a binary selection mask using an adjustable threshold δ as

$$\mathcal{G}_\delta(\mathbf{S}) = \mathbb{I}\left((\mathbf{S} - \delta) > 0\right), \tag{10}$$

where δ can be globally adjusted for all layers at runtime. Increasing δ reduces the number of activated slices per token, thereby lowering the effective precision, and vice versa. This design enables `MoBiQuant` to adapt precision at runtime without repeated weight repacking or dequantization.

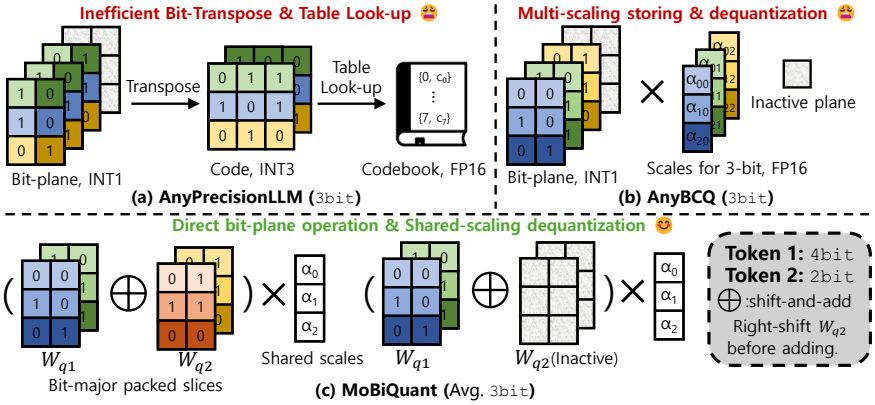

Figure 3: (a) AnyPrecisionLLM [8] with the inefficient bit-plane transposes and table lookups, (b) AnyBCQ [9] with the multi-scale storage and dequantization overhead, and (c) our `MoBiQuant` that performs direct bit-plane operations with the shared scales for efficient any-precision dequantization.

4.3 MoBiQuant Kernel Design

The `MoBiQuant` kernel in Fig. 3 enables efficient inference by addressing the following challenges:

❶ **Redundant Memory Access.** Conventional static low-bit kernels typically load all weight slices regardless of the runtime precision, resulting in unnecessary memory traffic and limiting inference speedups. To enable on-demand fetching, we adopt the bit-major packing representation from [8, 9], which decomposes quantized weights into independent bit-planes and stores each bit-plane in a separate contiguous memory region. The kernel can therefore independently fetch only the bit-planes required by the selected runtime precision. Within each active bit-plane, warp threads perform aligned and fully coalesced accesses, ensuring that skipping inactive bit-planes reduces the actual amount of data transferred from global memory. Tab. 1 reports the measured DRAM read traffic of a linear projection (4096×11008) operation using the `MoBiQuant` kernel at different runtime precisions. The traffic decreases approximately proportionally as the bit-width is reduced, confirming that inactive bit-planes are not fetched from global memory.

❷ **Non-contiguous Memory Access and Load Imbalance.** During prefilling and batched decoding, tokens are dynamically assigned to different bit slices, requiring gathers from non-contiguous memory locations and reducing memory bandwidth utilization. To address this, we permute tokens after

routing so that tokens assigned to the same bit slice are stored contiguously. Tab. 2 quantifies the permutation overhead and the resulting prefilling latency reduction for a 4-bit MLP layer in LLaMA-2-7B, showing that improved memory locality and load balancing outweigh the permutation overhead and yield a net speedup.

Table 1: DRAM read traffic across bit-wdiths.

Bit	Planes	DRAM Read	Δ/Plane
2	1	12.73 MB	–
4	2	24.05 MB	+11.32 MB
6	3	35.40 MB	+11.35 MB
8	4	46.68 MB	+11.28 MB

Table 2: Contribution of token permutation.

Length	Method	Perm. (ms)	Total (ms)
2048	w/o perm.	–	5.857
	w/ perm.	0.134 (2.9%)	**4.609**
4096	w/o perm.	–	9.916
	w/ perm.	0.179 (2.3%)	**7.954**

❸ **Routing Overhead.** `MoBiQuant` improves kernel efficiency in two aspects: First, unlike Any-PrecisionLLM [8], `MoBiQuant` avoids bit transposition and centroid lookup by performing Binary Matrix Multiplication (BMMA) directly on packed bit-planes. Second, compared with AnyBCQ [9], `MoBiQuant` shares scaling factors across slices and combines slices at the bit-plane level before scaling, reducing dequantization overhead (Fig. 3). However, `MoBiQuant` introduces token routing. To mitigate the additional routing overhead, we fuse all router projections within each layer into a single GPU kernel with shared-memory input reuse. Fig. 7 reports the prefill latency breakdown and end-to-end batched-decoding performance of the three methods.

5 Experiments

5.1 Experimental Setup

Models and benchmarks. To cover a broad range of model scales, we conduct experiments on LLaMA2-7B/13B [1], LLaMA3-8B, LLaMA3.2-1B/3B, Gemma2-9B [20], and Qwen3-32B [21]. We evaluate perplexity on WikiText2 [22] and accuracy on MMLU [23]. We further report zero-shot performance on BoolQ [24], PIQA [25], HellaSwag [26], WinoGrande [27], ARC-Easy, and ARC-Challenge [28], as well as reasoning performance on GSM8K [29].

Baselines and implementation. We primarily compare `MoBiQuant` with the any-precision baselines AnyPrecisionLLM [8] and AnyBCQ [9]. We further compare against single-precision PTQ methods, including GPTQ [12], OmniQuant [7], AWQ [15], SmoothQuant [13], SpinQuant [14], and QuaRot [30], as well as the VQ-based methods QUIP# [17] and QTIP [31]. We use 128 sequences from WikiText2 for calibration, and measure all kernel results on NVIDIA A100 GPUs with CUDA 12.9. Additional experimental details are provided in Appendix C.

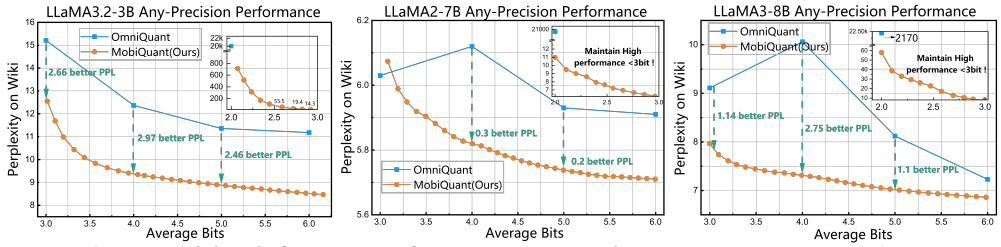

Figure 4: **Any-precision inference performance comparison.** `MoBiQuant` outperforms the PTQ baseline across multiple unseen target bit-widths on various LLaMA models, maintaining smooth and fine-grained precision scaling even at extremely low 2–3-bit precision range.

5.2 Main Results

Improved Cross-Bit Generalization. We first investigate whether `MoBiQuant` improves the cross-bit generalization of our integrated PTQ baseline, OmniQuant [7]. Specifically, we set the calibration bit-width for both OmniQuant and `MoBiQuant` to 3-bit and evaluate their performance across unseen higher and lower bit-widths. As shown in Fig. 4, `MoBiQuant` consistently outperforms OmniQuant at all unseen bit-widths across various models. Importantly, even at extremely low 2–3-bit range, `MoBiQuant` maintains smooth and stable precision scaling, whereas OmniQuant exhibits significant

performance degradation. We show that MoBiQuant can also improve rotation-based PTQ methods (Appendix E.3). Similar trend is also observed in activation quantization (Appendix E.4).

Comparison with any-precision baselines. We compare MoBiQuant with AnyPrecisionLLM [8] and AnyBCQ [9] in Tab. 3 in terms of perplexity, accuracy, and decoding throughput. On LLaMA2-7B at 2-bit, MoBiQuant reduces WikiText2 perplexity from 2e3 and 15.38 to 10.91, while improving MMLU accuracy from 25.11% and 29.87% to 33.20%. The improvements generalize to larger models Gemma2-9B and Qwen3-32B. Similar advantages are observed at 3- and 4-bit, where MoBiQuant generally provides a favorable accuracy–precision trade-off. Moreover, by avoiding centroid lookup and multi-scalar dequantization, the MoBiQuant kernel achieves an average decoding-throughput improvement of 33.8% over AnyPrecisionLLM and 22.8% over AnyBCQ on the LLaMA models. We also compare the calibration cost in Appendix E.5.

Table 3: End-to-end comparison with the any-precision baselines AnyPrecisionLLM (AP) and AnyBCQ (ABC). MoBiQuant achieves higher decoding throughput while maintaining competitive perplexity and accuracy.

Model	Bits	WikiText2, PPL ↓				MMLU, Acc. ↑				Throughput, tokens/sec. ↑		
		FP16	AP	ABC	MoBiQ	FP16	AP	ABC	MoBiQ	AP	ABC	MoBiQ
LLaMA2-7B	2	5.47	2e3	15.38	**10.91**	45.30	25.11	29.87	**33.20**	279	312	**404**
	3	5.47	6.13	6.25	**6.07**	45.30	41.85	41.56	**43.35**	241	268	**311**
	4	5.47	6.05	5.91	**5.82**	45.30	41.74	**43.84**	43.28	211	225	**256**
LLaMA3-8B	2	6.12	1e3	**19.01**	58.12	65.02	24.66	35.32	**40.13**	228	245	**352**
	3	6.12	8.60	8.08	**7.97**	65.02	55.53	58.28	**59.71**	194	212	**257**
	4	6.12	**6.70**	6.84	7.31	65.02	**64.04**	63.15	63.29	171	186	**209**
Gemma2-9B	2	6.84	19.62	12.72	**12.21**	70.74	33.09	43.36	**44.28**	163	185	**266**
	3	6.84	7.94	7.95	**7.73**	70.74	65.30	65.38	**66.41**	144	164	**199**
	4	6.84	**7.06**	7.23	7.07	70.74	69.23	68.81	**69.08**	125	146	**164**
Qwen3-32B	2	7.61	12.58	12.07	**11.50**	81.85	71.23	70.70	**72.68**	43	52	**75**
	3	7.61	8.93	8.42	**8.14**	81.85	78.95	79.72	**80.43**	40	46	**56**
	4	7.61	7.87	7.81	**7.75**	81.85	80.82	80.70	**80.97**	35	42	**45**

Comparison with single-precision PTQ baselines. We evaluate prior static scalar PTQ baselines that have been separately trained for fixed 3- and 4-bit precisions. For fair comparison, we keep our MoBiQuant elastic during inference but restrict its average precision to be either 3- or 4-bit. As shown in Tab. 4, MoBiQuant consistently matches the performance of state-of-the-art PTQ methods on LLaMA2 models and even surpasses them on LLaMA-3 models. Quantitative results demonstrate that MoBiQuant attains the performance of static quantization methods while preserving flexible precision adjustment. Additional zero-shot and GSM8K results, along with comparisons against VQ-based methods, are provided in Appendix E.6.

Table 4: The proposed any-precision MoBiQuant **matches** or **surpasses** the static scalar quantization (SQ) PTQ baselines with the same number of average bits on WikiText2, PPL.

Method	Type	LLaMA2-7B		LLaMA2-13B		LLaMA3.2-1B		LLaMA3.2-3B		LLaMA3-8B	
		3-bit	4-bit	3-bit	4-bit	3-bit	4-bit	3-bit	4-bit	3-bit	4-bit
FP16	—	5.5		5.0		13.4		9.2		6.1	
SmoothQuant	Static SQ	8.62	7.50	5.43	6.10	2e2	2e2	3e2	1e2	12.50	10.70
AWQ		6.24	5.83	5.32	5.07	20.69	15.02	15.51	12.76	8.22	7.36
GPTQ		6.29	5.90	5.42	5.60	20.81	15.19	15.77	13.14	8.28	7.43
SpinQuant		6.13	**5.60**	5.24	**5.00**	19.62	14.45	**14.73**	11.17	8.14	**6.40**
QuaRot		6.13	5.61	**5.22**	5.03	20.15	14.77	14.82	11.55	9.01	7.20
OmniQuant		**6.03**	5.74	5.28	5.02	23.92	16.79	15.21	12.71	9.11	7.41
MoBiQuant	Any-prec.	6.07	5.82	5.31	5.08	**19.40**	**14.02**	14.76	**9.41**	**7.97**	7.31

5.3 Analytical Results

Causal link between Token Routing and Outlier Migration. As discussed in Section 3, *outlier migration* exposes a calibration–inference mismatch: tokens well preserved at one precision can become disproportionately sensitive at another. To examine whether MoBiQuant mitigates this mismatch, we compare token-wise error increments when switching OmniQuant from 4-bit calibration to 3-bit inference with MoBiRoute's router scores. Fig. 5 (left) shows that tokens with larger error increments tend to receive higher router scores and higher precision, while less sensitive tokens are routed to 2-bit computation, indicating that the router captures sensitivity to precision reduction. Fig. 5

(right) shows more consistent token-wise error distributions across bit-widths than static PTQ in Fig. 1, indicating reduced *outlier migration*. These observations support the proposed token-adaptive routing allocates higher precision to vulnerable tokens and mitigates the calibration–inference mismatch.

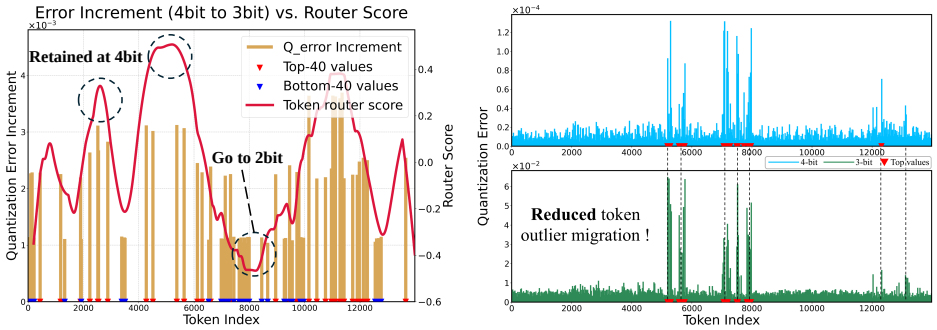

Figure 5: **Visualization of outlier migration mitigation.** `MoBiQuant` router scores correlate with the actual token-wise error increments under precision switching (left), and it produces more consistent token-wise error distributions across bit-widths with reduced outlier migration (right).

Token- and block-wise assignments. We collect statistics on token bit-width assignments for any-precision inference. As shown in Fig. 6 (right), tokens have assignment distributions that shift with target precision e.g., most tokens are assigned to 2- or 4-bit precisions under 3-bit inference, whereas the majority of tokens shift toward the 4–6-bit range for the 5-bit inference. Overall, the average precision remains within the prescribed target token-aware bit adjustment. We also observe that the average precision varies across different linear blocks, even though each layer is trained for the same target bit-width. As shown in Fig. 6 (left), MLP blocks exhibit divergent assignments across layers, and the first layers tend to have deviations in bit-width assignments. This pattern is consistent with the [32] study that allocates different precisions to blocks based on their sensitivity, indicating that `MoBiQuant` also benefits from block-level precision adjustments.

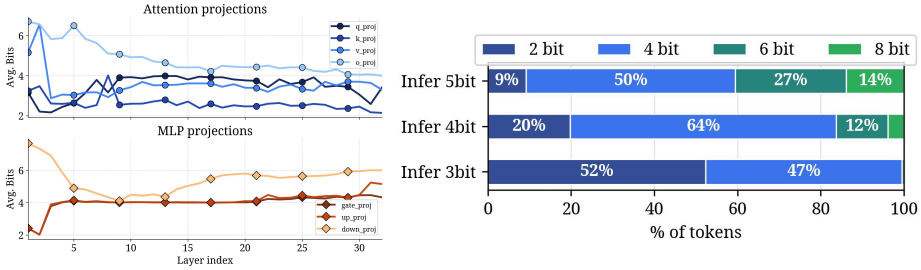

Figure 6: (Left) Average precision assignments of different linear blocks by `MoBiQuant`, showing block-level variation and higher deviations in initial layers. (Right) Precision distributions for any-precision inference, showing heterogeneous token assignments across target bit-widths.

5.4 Kernel Evaluation

We evaluate the `MoBiQuant` kernel under multi-batch decoding and prefilling workloads using a GEMM shape of (14336×4096). Figure 7 (left) compares the latency of the 4-bit GEMM against cuBLAS FP16, AnyPrecisionLLM (APLLM) [8], and AnyBCQ (ABCQ) [9] across different decoding batch sizes. `MoBiQuant` consistently achieves the lowest latency, providing 1.13–$1.63\times$ speedup over APLLM and 1.12–$1.49\times$ over ABCQ for batch sizes ranging from 16 to 64. Figure 7 (right) further reports the prefill latency breakdown with a context length of 2048. Despite introducing token routing and permutation, `MoBiQuant` limits their overhead to only 7.90% of the total latency. Its shared-scalar dequantization also incurs less overhead than the bit-transposition and lookup-table operations in APLLM and the multi-scalar dequantization in ABCQ. Consequently, `MoBiQuant` achieves the lowest overall prefill latency. Additional results on memory savings are provided in Appendices E.7.

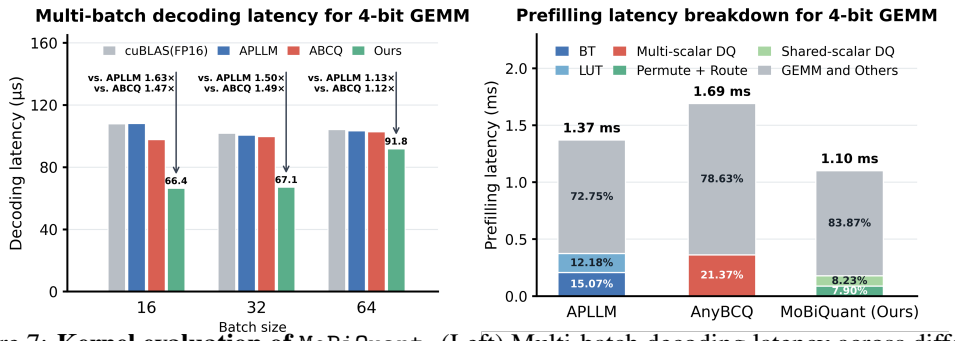

Figure 7: **Kernel evaluation of `MoBiQuant`.** (Left) Multi-batch decoding latency across different batch sizes. (Right) Prefill latency breakdown of any-precision methods.

5.5 Ablation Study

We evaluate the sensitivity to the calibration data by replacing WikiText2 with two alternative datasets: C4 [33] and PTB [34]. `MoBiQuant` consistently achieves better results in most calibration–evaluation pairs, demonstrating robustness to calibration set selection. The full results are shown in the Appendix D.1. We also report ablation studies on router regularization scheduling in Appendix D.2, and examine router generalization for different target precision training in Appendix D.3.

6 Conclusion

In this paper, we identify the "*outlier migration*" in conventional PTQ methods, which limits generalization of calibration across precisions. To address it, we proposed `MoBiQuant`, a quantization framework that integrates `MoBiSlice` for fine-grained precision selection and `MoBiRoute` for token-adaptive routing. Extensive experiments demonstrate its effectiveness and practicality.

Acknowledgments and Disclosure of Funding

This work was supported in part by NSF (No. 2625580), ARO (No. W911NF-23-2-0224), and DOE (No. DE-SC0026696). Additional support was provided by the National Research Foundation of Korea (NRF) (No. RS-2025-02216217); the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korean government (MSIT) (No. RS-2019-II190421, AI Graduate School Support Program (Sungkyunkwan University)); and the Korea Institute of Police Technology (KIPoT), funded by the Korean National Police Agency and Korea Customs Service (No. RS-2026-25536784). This work was also supported in part by the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) (No. RS-2026-25499095).

References

[1] Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[2] OpenAI et al. Gpt-4 technical report, 2023.

[3] Gemini-Team et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[4] DeepSeek-AI et al. Deepseek-v3 technical report, 2025.

[5] Soumye Singhal et al. Llama-Nemotron: Efficient reasoning models. In *The Exploration in AI Today Workshop at ICML*, 2025.

[6] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *ICLR*, 2023.

[7] Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *ICLR*, 2024.

[8] Yeonhong Park, Jake Hyun, SangLyul Cho, Bonggeun Sim, and Jae W. Lee. Any-precision LLM: Low-cost deployment of multiple, different-sized LLMs. In *ICML*, 2024.

[9] Gunho Park, Jeongin Bae, Beomseok Kwon, Byeongwook Kim, Se Jung Kwon, and Dongsoo Lee. AnyBCQ: Hardware efficient flexible binary-coded quantization for multi-precision LLMs. In *ICLR*, 2026.

[10] Pranav Ajit Nair, Puranjay Datta, Jeff Dean, Prateek Jain, and Aditya Kusupati. Matryoshka quantization. In *ICML*, 2025.

[11] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *NeurIPS*, 2022.

[12] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv:2210.17323*, 2022.

[13] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *ICML*, 2023.

[14] Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations. In *ICLR*, 2025.

[15] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, 2024.

[16] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *NeurIPS*, 2023.

[17] Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. In *ICML*, 2024.

[18] Mengzhao Chen, Yi Liu, Jiahao Wang, Yi Bin, Wenqi Shao, and Ping Luo. PrefixQuant: Static quantization beats dynamic through prefixed outliers in LLMs. *arXiv:2410.05265*, 2024.

[19] Haodong Wang, Qihua Zhou, Zicong Hong, and Song Guo. D2MoE: Dual routing and dynamic scheduling for efficient on-device MoE-based LLM serving. In *Proceedings of the Annual International Conference on Mobile Computing and Networking*, 2025.

[20] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[21] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[22] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *ICLR*, 2017.

[23] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv:2009.03300*, 2020.

[24] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[25] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*, 2020.

[26] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *AAAI*, 2019.

[27] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 2021.

[28] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv:1803.05457*, 2018.

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

[30] Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs. In *NeurIPS*, 2024.

[31] Albert Tseng, Qingyao Sun, David Hou, and Christopher De Sa. QTIP: Quantization with trellises and incoherence processing. In *NIPS*, 2024.

[32] Coleman Hooper, Charbel Sakr, Ben Keller, Rangharajan Venkatesan, Kurt Keutzer, Sophia Shao, and Brucek Khailany. FGMP: Fine-grained mixed-precision weight and activation quantization for hardware-accelerated LLM inference. *arXiv:2504.14152*, 2025.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[34] Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[35] Jinhee Kim, Seoyeon Yoon, Taeho Lee, Joo Chan Lee, Kang Eun Jeon, and Jong Hwan Ko. TruncQuant: Truncation-ready quantization for dnns with flexible weight bit precision. In *IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2025.

[36] Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *NeurIPS*, 2024.

Appendix

Contents

A. Overall Algorithm of MoBiQuant 13
B. MoBiSlice Construction: Formulation and Further Analysis 13
C. Experiment Details 16
C.1. MoBiQuant Training Setup 16
C.2. Flexible Inference and Effective Bit Control 16
C.3. Baseline Results 17
D. Ablation Studies 17
D.1. Ablation of Calibration Dataset 17
D.2. Ablation of Scheduling Methods for Router Regularization 17
D.3. Ablation of Different Training Target Bit 18
E. Additional Experiments 18
E.1. Outlier migration in Non-Gradient-Based PTQ 18
E.2. Outlier migration on Mistral-7B 19
E.3. Compatibility with rotation-based methods 19
E.4. Any-precision performance with Activation Quantization 19
E.5. Additional Comparison with Static PTQ Methods 20
E.6. Memory Savings under Multi-Precision Deployment 21
F. Limitations 22

A Overall Algorithm of MoBiQuant

As shown in Alg. 1, `MoBiQuant` performs layer-wise quantization by matching each quantized layer to a full precision reference output using a two-stage iterative process. For each layer, it first runs the original full-precision layer to obtain full precision outputs, then initializes the shared quantization parameters $\mathbf{\Theta_q}$ and the router $\mathbf{\Theta_r}$. In *Stage 1*, it stabilizes the first slice path by optimizing the quantization parameters to minimize the reconstruction error to the full-precision target. In *Stage 2*, it initializes residual slices' quantization parameters from *Stage 1*, computes the router score for each token, and produces the overall quantized outputs. This stage jointly optimizes the objective that combines reconstruction loss with a router score regularization loss. After the optimization converges for the layer, `MoBiQuant` commits the learned quantization parameters, propagates both full-precision and quantized outputs to the next layer, and repeats the same procedure until all layers are quantized.

B MoBiSlice Construction: Formulation and Further Analysis

In this section, we formalize the `MoBiSlice` quantization scheme and describe how we decompose a full-precision weight into a set of composable low-bit slices, enabling a many-in-one representation for multiple precisions.

Quantizer design for bit slice accumulation. `MoBiSlice` requires that precision be adjustable by either accumulating residual bits or truncating them. To satisfy this property, we adopt a floor-aligned mapping following the truncation-ready quantization principle [35], where a lower precision code is obtained by simply dropping least significant bits (LSB) rather than re-rounding. Concretely, for

bit-width b, scale s, and continuous zero point z, we quantize as

$$x_{\text{int}} = \text{clamp}\Big(\Big\lfloor \frac{x}{s} + z \Big\rfloor, 0, 2^b - 1\Big), \tag{11}$$

$$x_{\text{deq}} = s\,(x_{\text{int}} - z + 0.5)\,. \tag{12}$$

The floor operator enforces hierarchical nesting of integer codes, so switching bit width corresponds to adding or dropping bit slices without changing previously formed higher order bits. The $+0.5$ dequantization shift centers each bin, which reduces bias when residual slices are accumulated.

Algorithm 1 MoBiQuant Calibration

Input: Input tokens $\mathbf{X}$, model $\mathcal{M}$ with L layers $\{f_{\text{fp}}^{(\ell)}\}_{\ell=1}^{L}$, epochs `ep`, number of bit slices E, regularization weight λ

Output: Quantized model $\widehat{\mathcal{M}}$

Notation: We omit the layer superscript (ℓ) for readability.
$\mathbf{H}_{\text{fp}}, \mathbf{H}_{\text{q}}$: Full precision and quantized input for layer ℓ
$\mathbf{Y}_{\text{fp}}, \mathbf{Y}_{\text{q}}$: Full precision and quantized output for layer ℓ
$\mathbf{G}$: Router scores, $\odot$: element wise product

Initialize activations $\mathbf{H}_{\text{fp}}, \mathbf{H}_{\text{q}} \leftarrow \mathbf{X}$
for $\ell = 1$ **to** L **do**
  FP output $\mathbf{Y}_{\text{fp}} \leftarrow f_{\text{fp}}(\mathbf{H}_{\text{fp}}; \mathbf{W})$
  Initialize quantization parameters $\mathbf{\Theta_q}$ and router parameters $\mathbf{\Theta_r}$
  for $t = 1$ **to** `ep` **do**
    Stage 1: First slice (FS) stabilization
    $\mathbf{Y}_{\text{FS}} \leftarrow f_{\text{FS}}(\mathbf{H}_{\text{q}}; \mathbf{\Theta_q})$ ▷ First slice-only forward pass
    $\mathcal{L}_{\text{FS}} \leftarrow \|\mathbf{Y}_{\text{FS}} - \mathbf{Y}_{\text{fp}}\|_2^2$ ▷ match FP reference output
    Update $\mathbf{\Theta_q}$
    Stage 2: Joint training
    Slices $\{\Theta_{q,e}\}_{e=1}^{E} \leftarrow \text{Derive}(\mathbf{\Theta_q})$ ▷ construct MoBi slices
    Router score $\mathbf{S} \leftarrow \text{Router}(\mathbf{H}_{\text{q}}; \mathbf{\Theta_r})$ ▷ token wise gating weights
    $\mathbf{Y}_{\text{q}} \leftarrow \mathbf{Y}_{\text{FS}} + \sum_{e=1}^{E} \mathbf{G}_e \odot f_e(\mathbf{H}_{\text{q}}; \Theta_{q,e})$ ▷ sum all slices' outputs
    $\mathcal{L}_{\text{MoBi}} \leftarrow \|\mathbf{Y}_{\text{q}} - \mathbf{Y}_{\text{fp}}\|_2^2 + \lambda\,\text{Reg}(\mathbf{S}, \mathbf{\Theta_r})$ ▷ reconstruction loss and router regularization
    Update $\mathbf{\Theta_q}, \mathbf{\Theta_r}$ ▷ joint update of quantization and router params
  end for
  $\widehat{\mathbf{W}} \leftarrow \text{Quantize}(\mathbf{W}; \mathbf{\Theta_q}, \mathbf{\Theta_r})$
  $\mathbf{H}_{\text{fp}} \leftarrow \mathbf{Y}_{\text{fp}}$
  $\mathbf{H}_{\text{q}} \leftarrow f_{\text{q}}(\mathbf{H}_{\text{q}}; \widehat{\mathbf{W}})$
end for
return $\widehat{\mathcal{M}}$

To ensure consistent refinement across slices, `MoBiSlice` iteratively refines the scaling factor so that each slice represents progressively finer increments. Let (s_0, z_0) be the calibrated parameters of the first slice. After assigning b_e bits to slice e, the next slice refines the resolution as $s_{e+1} = s_e / 2^{b_e}$. Finally, while the first slice uses the calibrated zero point z_0, slices fix $z_e = 2^{b_e - 1}$ for $e \geq 1$, placing the midpoint code at the center of the integer range so that positive and negative residual corrections are represented symmetrically, which avoids systematic drift during slice accumulation. Together, the floor aligned mapping, scale refinement, and centered residual quantization make the slice decomposition compatible with truncation based precision control, enabling stable elastic reconstruction by activating a subset of slices.

Bias and Error Bounds for Bit Slice Activation. We now analyze truncation-based bit slice activation and show that activating residual slices performs a true refinement: it leaves the coarse quantization code unchanged and only adds a bounded zero-mean remainder term that recovers finer

detail. First, we assume a two slice scenario, consisting of a shared base slice with $(b-k)$ bits and a single k bit residual slice that can be activated or dropped.

Let W be the original weight. The base slice produces an MSB code with parameters (s_1, z_1), and the residual slice quantizes the residual R using parameters (s_2, z_2) derived from the base slice:

$$\mathrm{MSB}_{b-k} \;=\; \Big\lfloor \frac{W}{s_1} + z_1 \Big\rfloor, \tag{13}$$

$$\mathrm{LSB}_k \;=\; \Big\lfloor \frac{R}{s_2} + z_2 \Big\rfloor, \qquad s_2 = \frac{s_1}{2^k}, \qquad z_2 = 2^{k-1}. \tag{14}$$

Using the centered dequantization, the two-slice reconstruction is:

$$\widehat{W} = s_1(\mathrm{MSB}_{b-k} - z_1 + 0.5) \;+\; s_2(\mathrm{LSB}_k - z_2 + 0.5)\,. \tag{15}$$

Although the reconstruction in Eq. (15) is written for a base $(b-k)$ bit slice plus a k bit residual slice, our goal is to characterize *any* intermediate precision obtained by activating only a subset of the residual information. We therefore introduce a generic truncation level p that drops p least significant bits from the merged code. This captures all partial activation cases: $p=0$ corresponds to using all k residual bits, and larger p corresponds to deactivating progressively more of the finest bits. We truncate p least significant bits of the merged integer code $\mathrm{INT}_8 = (\mathrm{MSB}_{b-k} \ll k) + \mathrm{LSB}_k$ using a right shift:

$$I \;=\; \mathrm{INT}_8 \gg p \;=\; \Big\lfloor \frac{\mathrm{INT}_8}{2^p} \Big\rfloor, \qquad \mathrm{INT}_8 \;=\; 2^p I + r, \qquad r \in \{0, 1, \ldots, 2^p - 1\} \tag{16}$$

Substituting Eq. (16) into Eq. (15) and noting that $s_2 = s_1/2^k$, we can rearrange terms to express $\widehat{W}$ as a coarse quantizer applied to the truncated code I, plus an additive truncation noise term determined by the dropped remainder r:

$$\begin{aligned}
\widehat{W} &= s_2\big(\mathrm{INT}_8 - 2^k z_1 + 0.5\big) \\
&= s_2\big(2^p I + r - 2^k z_1 + 0.5\big) \\
&= \underbrace{(2^p s_2)}_{s'}\left(I - \underbrace{2^{k-p} z_1}_{z'} + \underbrace{0.5}_{\mathrm{mid}} \right) \;+\; \underbrace{s_2\big(r + 0.5 - 2^{p-1}\big)}_{E_p}.
\end{aligned} \tag{17}$$

Therefore, truncating p least significant bits is equivalent to quantizing with coarser parameters:

$$s' = 2^p s_2, \qquad z' = 2^{k-p} z_1, \qquad \mathrm{mid} = 0.5, \tag{18}$$

Eq. (17) makes the role of p explicit: truncation does not change the coarse code I, and its only effect is an additive perturbation E_p determined by the discarded remainder r. Thus, analyzing E_p directly characterizes the impact of partially activating bit slices. Assuming the residue r is approximately uniform over $\{0, 1, \ldots, 2^p - 1\}$, we obtain:

$$\mathbb{E}[E_p] = s_2\big(\mathbb{E}[r] + 0.5 - 2^{p-1}\big) = s_2\left(\frac{2^p - 1}{2} + 0.5 - 2^{p-1} \right) = 0, \tag{19}$$

This yields $\mathbb{E}[E_p] = 0$, indicating that slice-based reconstruction does not introduce bias in expectation. Moreover, since $r \in \{0, 1, \ldots, 2^p - 1\}$,

$$-\Big(2^{p-1} - 0.5\Big) \;\le\; r + 0.5 - 2^{p-1} \;\le\; \Big(2^{p-1} - 0.5\Big), \tag{20}$$

which implies the error bound:

$$|E_p| \;\le\; s_2\Big(2^{p-1} - 0.5\Big) \;<\; 2^{p-1} s_2 \quad\Longrightarrow\quad \frac{|E_p|}{s_2 2^p} \;<\; \frac{1}{2}. \tag{21}$$

Eq. (21) implies that the truncation error is strictly smaller than one half step of the coarser quantizer. Equivalently, the perturbation E_p cannot move $\widehat{W}$ across a decision boundary of the base MSB quantizer, and therefore cannot flip any bit of the coarse code. As a result, activating residual slices performs a true residual refinement: it only fills in finer bit slices without altering the coarser representation, yielding an accurate and stable activation with controlled error. Therefore, this establishes that bit slice activation yields a guaranteed residual refinement: it adds finer information while preserving the coarse code, and its only effect is an additive truncation noise term that is bounded and zero-mean.

Prefix Selection Is Naturally Learned. The formulation in Eq. 3 requires selecting a prefix of the ordered slices. In practice, we do not enforce prefix selection, allowing token-aware precision routing to use E independent binary gates and enabling run-time precision adjustment through a single global threshold, as in Eq. (10). In contrast, parameterizing the routing decision as a one-hot selection among E prefix lengths complicates such threshold-based adjustment. Nonetheless, we find that an explicit prefix constraint is unnecessary, as prefix selection emerges naturally during optimization. Earlier slices tend to receive stronger optimization signals, encouraging the router to preferentially select them when minimizing the reconstruction loss.

Consider four 2-bit slices $[\mathbf{W}_1, \mathbf{W}_2, \mathbf{W}_3, \mathbf{W}_4]$, ordered from the most significant to the least significant bit-slice. Given an input token $\mathbf{X}$, a shared scale $\mathbf{\Theta}_q$, router parameters $\mathbf{\Theta}_r$, and an upstream gradient g, the routing and reconstruction process can be written as

$$\mathbf{X}\mathbf{\Theta}_r \;\longrightarrow\; \left(2^6\mathbf{W}_1 + 2^4\mathbf{W}_2 + 2^2\mathbf{W}_3 + 2^0\mathbf{W}_4\right)\mathbf{X}\mathbf{\Theta}_q, \tag{22}$$

with branch-wise gradient contributions proportional to

$$\left[2^6\mathbf{W}_1 g\mathbf{X},\; 2^4\mathbf{W}_2 g\mathbf{X},\; 2^2\mathbf{W}_3 g\mathbf{X},\; 2^0\mathbf{W}_4 g\mathbf{X}\right]. \tag{23}$$

Thus, earlier slices have a larger influence on reconstruction-error reduction. At the beginning of training, all slices are active. As the regularization term gradually encourages fewer active slices, the router retains the slices that reduce the L2 loss in Eq. 9 the most. For example, if a token is routed to 4-bit precision using two slices, selecting $\mathbf{W}_1 + \mathbf{W}_2$ preserves the most significant bit-slices, whereas $\mathbf{W}_3$ and $\mathbf{W}_4$ contribute exponentially less to the reconstruction. Therefore, prefix selection naturally yields the smallest reconstruction error, making non-prefix combinations suboptimal. Consequently, `MoBiRoute` learns prefix selection without an explicit constraint. Empirically, after convergence, all routed token–bit assignments satisfy the prefix property, selecting the first k slices for $2k$-bit routing.

C Experiment Details

C.1 MoBiQuant Training Setup

We use weight-only quantization in all reported runs in the main paper, with `abits=16` and `group_size=128`, while the base bit slice uses `wbits=2`. Weight-activation quantization results and their experimental setup are reported in Appendix E.4. Slice decomposition is enabled by `slice_bits_list`, and our default configuration uses four bit slices with `slice_bits_list = 2 2 2 2`. Training proceeds for `epochs=20` and `nsamples=128`, with `batch_size=1` for all models. The total number of router annealing steps is used by the router temperature schedule and calculated as follows:

$$\texttt{global_steps} = \left(\frac{\texttt{nsamples}}{\texttt{batch_size}}\right) \cdot \texttt{epochs},$$

For each layer, we optimize three parameter groups with AdamW: learnable weight clipping parameters (LWC), learnable equivalent transformation parameters (LET), and MoBiQuant parameters. Group specific learning rates are denoted as follows: `lwc_lr`, `let_lr`, and `mobi_lr`. We typically use `lwc_lr` in the range `1e-3` to `1e-2`, and `mobi_lr` in the range `5e-6` to `4e-5`, depending on model size.

C.2 Flexible Inference and Effective Bit Control

A core feature of `MoBiQuant` is flexible inference, where the effective precision is controlled by activating a subset of residual bit slices. At evaluation time, this is exposed by `target_activation_ratio_eval` or `target_bit_eval`. When `target_bit_eval` is used, we convert it into a desired routing ratio based on the base MSB bits and the sum of residual slice bits:

$$\rho \;=\; \frac{\texttt{target_bit_eval} - b_{\text{msb}}}{\sum_{e\geq 1} b_e}, \qquad b_{\text{msb}} = \texttt{slice_bits_list}[0].$$

Hard binary gating is applied in evaluation using router scores and a threshold selected to satisfy the target ratio.

Layer-wise threshold calibration. To make the realized routing ratio stable across layers and modules, we calibrate a layer specific threshold from calibration samples whenever `target_bit_eval` requests residual activation. This calibration collects router scores over the calibration set and sets each layer's threshold to the appropriate quantile so that approximately a fraction ρ of routed experts are active. We set the calibration sample to be the same as `nsamples`.

C.3 Baseline Results

The baseline results reported in our experiments are either taken from the corresponding published papers or obtained by running their official open-source implementations. All reproduced experiments are conducted on NVIDIA A100 GPUs. Specifically, for OmniQuant, SpinQuant, AWQ, GPTQ, SmoothQuant, QUIP#, and QTIP, we run their open-source code to obtain WikiText2 PPL results on the LLaMA-3.2-1B, LLaMA-3.2-3B, and LLaMA-3-8B model series under a context length of 2048. In particular, for QUIP#, we reproduce and report its non-finetuned variant to align its computational cost with the other post-training quantization baselines. These reproduced results are used to ensure a consistent evaluation setting with our method. For kernel measurement, to ensure a consistent evaluation environment with our customized kernel, we uniformly evaluate the kernels of QUIP#, QTIP, AnyPrecisionLLM, and AnyBCQ on an NVIDIA A100 GPU with CUDA 12.9.

D Ablation Studies

D.1 Ablation of Calibration Dataset

Table 5: **Ablation study on different calibration datasets for LLaMA3.2-1B-3bit.**

Dataset	Method	Perplexity (PPL ↓)			Downstream Tasks (Acc ↑)							
		Wiki	C4	PTB	ARC E	ARC E(n)	ARC C	ARC C(n)	BoolQ	Hella	Hella(n)	Wino
Wiki	Omni.	21.597	32.860	37.467	0.510	0.472	0.245	0.284	0.593	0.378	0.488	**0.551**
	MoBiQ	**18.947**	**27.924**	**34.915**	**0.521**	**0.481**	**0.261**	**0.294**	**0.613**	**0.382**	**0.497**	0.533
C4	Omni.	**22.611**	**33.198**	**38.096**	0.503	**0.470**	0.245	0.282	0.583	**0.378**	**0.488**	**0.548**
	MoBiQ	26.360	33.382	43.888	**0.522**	0.452	**0.282**	**0.293**	**0.619**	0.369	0.477	0.538
PTB	Omni.	22.689	33.519	35.353	0.503	0.471	0.241	0.281	0.586	0.376	0.488	**0.556**
	MoBiQ	**16.909**	**24.849**	**29.268**	**0.588**	**0.532**	**0.284**	**0.312**	**0.589**	**0.402**	**0.526**	**0.556**
Mix	Omni.	22.335	33.327	36.763	0.502	0.474	0.246	0.277	0.599	0.374	0.488	**0.553**
	MoBiQ	**21.070**	**30.072**	**33.646**	**0.534**	**0.502**	**0.258**	**0.300**	**0.612**	**0.382**	**0.500**	0.530

We study the impact of different calibration datasets in Tab. 5. `MoBiQuant` remains stable across calibration choices and consistently achieves lower perplexity than OmniQuant on most calibration–evaluation pairs. In particular, WikiText2 and PTB calibration provide the strongest improvements across WikiText2, C4, and PTB, while the mixed calibration setting maintains similar average bits with favorable cross-domain generalization.

On zero-shot tasks, `MoBiQuant` generally matches or outperforms OmniQuant across ARC Easy, ARC Challenge, BoolQ, HellaSwag, and WinoGrande [28, 24, 26]. The most consistent gains are again observed with PTB and WikiText2 calibration, indicating that `MoBiQuant` is robust to calibration set selection across both language modeling and downstream reasoning tasks.

D.2 Ablation of Scheduling Methods for Router Regularization

We conduct an ablation study to identify the most effective scheduling strategy for router regularization. We compare four schedules, Linear, Cosine, Exponential, and Logarithmic, and measure their impact on model perplexity (PPL) on WikiText2, C4, and PTB. As shown in Fig. 8, performance differences are most pronounced in the low-precision regime of 2.5 to 3.0 average effective bits. **In this range, both Logarithmic and Exponential scheduling consistently outperform the Linear and Cosine alternatives.** The relative ordering of the two best schedules is dataset dependent: Exponential offers a slight advantage on WikiText2, while Logarithmic performs better on C4 and PTB within the same bit-width range.

We adopt a logarithmic schedule for regularization, as it naturally aligns with the logarithmic temperature annealing used in router gating and converges smoothly to the target precision during

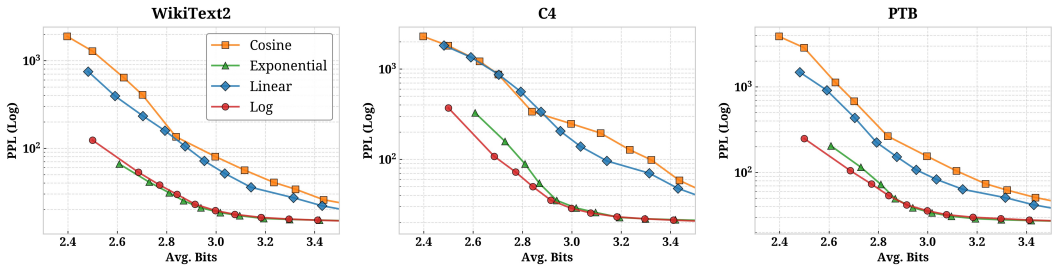

Figure 8: Comparison of PPL (↓) sweep between different scheduling methods across WikiText2, C4, PTB datasets.

inference. This synchronization between routing sharpness and precision allocation stabilizes training and enables smoother adaptation to resource constraints.

D.3 Ablation of Different Training Target Bit

We train LLaMA3.2-1B with five target bit budgets (2.5, 3.0, 3.5, 4.0, and 5.0) to study how the training target affects router generalization across precisions. For each configuration, we sweep inference-time budgets and evaluate PPL on WikiText2, C4, and PTB.

As shown in Fig. 9, consistent trends are observed across datasets, with the largest differences appearing near the transition region around 3.0 average bits. Among all settings, the 3.0-bit target achieves the best overall trade-off. Lower targets (e.g., 2.5-bit) improve low-bit performance but degrade higher-bit PPL, while higher targets (3.5–5.0-bit) significantly worsen sub-3.0-bit performance with only marginal gains at higher precisions. Elastic inference requires the router to accommodate a

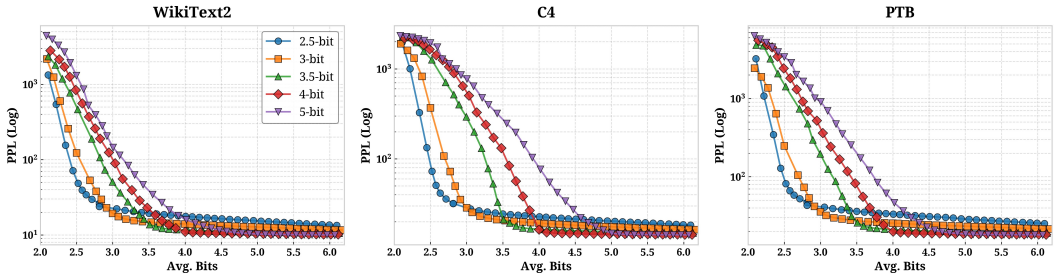

Figure 9: Comparison of PPL (↓) sweep between different target bit-widths for training across WikiText2, C4, PTB datasets.

changing target budget, and the target precision scheduling objective in Eq. 7 is designed to encourage exploration of diverse token slice allocations before converging. However, the training target still sets the operating point around which the router learns stable binary decisions: a low target biases the router toward early and persistent slice pruning, while a high target biases it toward maintaining many active slices. The 3.0 bit target sits near a balanced regime and improves calibration of routing decisions across a broader range of average bits. This motivates us to adopt 3.0 bits as the default training target throughout all experiments.

E Additional Experiments

E.1 Outlier migration in Non-Gradient-Based PTQ

To further validate the generality of our observation, we analyze the behavior of non-gradient-based PTQ methods under calibration–inference mismatch. We take AWQ [15] as a representative baseline and evaluate its generalization performance across different bit-widths. As shown in Table 6, AWQ exhibits noticeable degradation when the inference precision differs from the calibration setting, indicating limited generalization across bit-widths. In addition, we observe that the top outlier

Table 6: Generalization gap (PPL) of AWQ on LLaMA2-7B.

Calibration Bits	Infer @ 3-bit	Infer @ 4-bit
3-bit	8.22	7.85
4-bit	8.45	7.36

tokens overlap by only 41% between 3-bit and 4-bit settings, suggesting that token sensitivity shifts significantly across precisions. This behavior is consistent with the *outlier migration* phenomenon described in the main text.

E.2 Outlier migration on Mistral-7B

We further verify that the outlier migration phenomenon generalizes beyond the LLaMA family. Specifically, we observe similar behavior on Mistral-7B. For example, in a middle layer of Mistral-7B, the overlap of top token outliers between 3-bit and 4-bit is only 16%, indicating significant precision-dependent variation in token sensitivity. This migration leads to poor generalization for existing PTQ methods. As shown in Table 7, when the inference bit-width differs from the calibration setting, performance degrades substantially. In contrast, `MoBiQuant` effectively mitigates this issue.

Table 7: Mistral-7B performance gap (PPL) under calibration–inference mismatch.

Method	Calib 3 $\rightarrow$ Infer 4	Calib 4 $\rightarrow$ Infer 3
OmniQuant	5.61	23.19
MoBiQuant	**5.42**	**19.38**

E.3 Compatibility with rotation-based methods

We further investigate whether our method can be combined with rotation-based PTQ techniques. We consider two representative methods: (1) the non-gradient-based QuaRot [30], and (2) the gradient-based DuQuant [36].

Table 8: Generalization gap (PPL) of QuaRot on LLaMA2-7B.

Method	Calib 4 $\rightarrow$ Infer 4	Calib 4 $\rightarrow$ Infer 3
OmniQ	5.74	32.80
OmniQ + QuaRot	5.65	27.55
MoBiQuant + QuaRot	**5.65**	**22.24**

Table 8 shows that while QuaRot improves static performance, the degradation under bit-width mismatch remains significant. In contrast, combining `MoBiQuant` with QuaRot substantially improves generalization, demonstrating that our approach is complementary to rotation-based techniques. For DuQuant, we calibrate the model at W3A4 and evaluate on unseen weight–activation configurations. As shown in Table 9, `MoBiQuant` consistently improves generalization across all settings, with significant gains on LLaMA3-8B. Overall, these results demonstrate that (1) the outlier migration phenomenon is not limited to a specific PTQ method, and (2) `MoBiQuant` effectively improves generalization when combined with existing PTQ techniques.

E.4 Any-precision performance with Activation Quantization

The elasticity of `MoBiQuant` also extends to activation quantization, as shown in Fig. 10, where we report results on LLaMA2-13B. We keep the same slice decomposition as the weight-only setting and the same calibration hyperparameters. We also keep the same regularization parameter and learning rate. A key practical detail in this setting is the interaction between learnable equivalent transforms (LET) from OmniQuant [7] and router training. LET preserves the main linear path by transforming the input activation and compensating the linear weights so that the layer output remains equivalent. For a linear layer,

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{B}, \qquad \tilde{\mathbf{X}} = (\mathbf{X} - \boldsymbol{\delta}) \oslash \mathbf{s}, \qquad \tilde{\mathbf{W}} = \mathbf{s} \odot \mathbf{W}, \qquad \tilde{\mathbf{B}} = \mathbf{B} + \boldsymbol{\delta}\mathbf{W}, \tag{24}$$

Table 9: Generalization performance of DuQuant and MoBiQuant on unseen W-A configurations.

W-A	LLaMA2-7B			LLaMA3-8B		
	W3A4	W4A4	W5A4	W3A4	W4A4	W5A4
DuQuant	6.77	6.51	6.49	17.12	15.31	12.12
MoBiQuant + DuQuant	6.81	**6.40**	**6.30**	**13.71**	**12.25**	**11.69**

where $\mathbf{X} \in \mathbb{R}^{T\times d}$ denote the input activations with T tokens and d input channels; $\mathbf{W} \in \mathbb{R}^{d\times m}$ the weight matrix; and $\mathbf{B} \in \mathbb{R}^{m}$ the bias. LET introduces a learnable per-channel shift $\boldsymbol{\delta} \in \mathbb{R}^{d}$ and scale $\mathbf{s} \in \mathbb{R}^{d}$, producing the transformed input $\tilde{\mathbf{X}}$ and the compensated parameters $\tilde{\mathbf{W}}$, $\tilde{\mathbf{B}}$ so that the layer output is preserved. However, the router is a separate linear branch whose weights are not

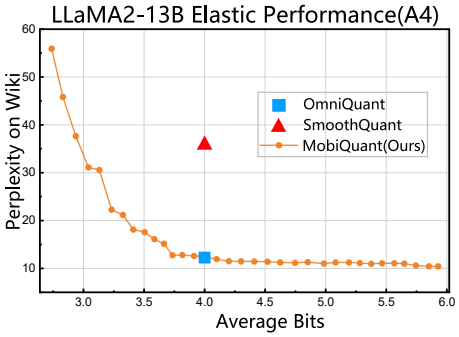

Figure 10: PPL (↓) tradeoff under 4 bit activation quantization on LLaMA2-13B. We compare with SmoothQuant [13] and OmniQuant [7] W4A4 baselines.

re-parameterized by LET, so feeding $\tilde{\mathbf{X}}$ directly changes the routing function and can distort router scores. This is important because LET parameters evolve during calibration, so the router would otherwise see a moving input distribution, which can drive logits into saturation under sharp gating schedules and small calibration budgets, especially when the router input is quantized. For stability, we undo LET for the router input by reconstructing the activation in the original space before routing as follows:

$$\mathbf{X}_r = \tilde{\mathbf{X}} \odot \mathbf{s} + \boldsymbol{\delta}, \tag{25}$$

This keeps the router operating in a fixed activation space while LET continues to re-parameterize the main linear path, improving routing stability.

E.5 Calibration cost comparison

We report in Tab. 10 the calibration time and computational cost required to obtain the $\{2, 3, 4\}$-bit LLaMA2-7B models. MoBiQuant requires only a single calibration run with 128 samples, comparable to calibrating one static PTQ model, to support multiple bit-widths. In contrast, OmniQuant requires three independent calibration runs, substantially increasing both calibration time and total calibration-data usage. MoBiQuant also incurs lower calibration-memory overhead than AnyBCQ. AnyBCQ maintains multiple groups of learnable scaling factors associated with different bit-planes, whereas MoBiQuant shares a single scaling-factor parameterization across slices. This design eliminates precision-specific scaling parameters and their associated optimizer states, reducing the peak calibration memory from 19 GB to 17 GB. Overall, all three methods can be calibrated on a single NVIDIA RTX 4090 GPU with 24 GB of memory, demonstrating that the offline calibration cost remains practical for real-world deployment.

E.6 Additional Comparison with Static PTQ Methods

Vector Quantization Methods. We further compare MoBiQuant with the recent VQ-based methods QuIP# [17] and QTIP [31]. Although these methods achieve strong bit–accuracy trade-offs, their decoding throughput is limited by centroid-table lookup overhead. As shown in Tab. 11, 2-bit QuIP# and QTIP achieve 191 and 218 tokens/sec. on LLaMA2-7B, with perplexities of 12.30 and 5.86,

Table 10: Calibration time and computational cost for obtaining the $\{2, 3, 4\}$-bit LLaMA2-7B models on a single NVIDIA RTX 4090 GPU.

Method	Time	Data Usage	Peak Memory	Calibration Runs
OmniQuant	7.6 h	384	13 GB	3
AnyBCQ	6.3 h	512	19 GB	1
MoBiQuant	**4.5 h**	**128**	17 GB	1

respectively. In comparison, 4-bit `MoBiQuant` achieves a lower perplexity of 5.82 while increasing throughput to 256 tokens/sec. Overall, `MoBiQuant` provides a favorable accuracy–throughput trade-off while retaining run-time precision adjustment.

Table 11: Comparison with the static VQ methods QuIP# and QTIP. `MoBiQuant` provides competitive perplexity with higher decoding throughput while supporting run-time precision adjustment.

Model	Bits	WikiText2, PPL ↓				Throughput, tokens/sec. ↑		
		FP16	QuIP#	QTIP	MoBiQ	QuIP#	QTIP	MoBiQ
LLaMA2-7B	2	5.47	12.30	**5.86**	10.91	191	218	**404**
	3	5.47	6.19	**5.28**	6.07	177	196	**311**
	4	5.47	5.66	**5.17**	5.82	155	189	**256**
LLaMA3-8B	2	6.12	15.20	**7.25**	58.12	157	180	**352**
	3	6.12	8.27	**7.05**	7.97	142	158	**257**
	4	6.12	6.64	**6.06**	7.31	126	154	**209**

Downstream Evaluation. We further compare elastic `MoBiQuant` with static PTQ baselines on downstream zero-shot tasks. Table 12 reports the average accuracy over six commonsense-reasoning tasks. Prior methods use fixed 4-bit precision, whereas the average precision of `MoBiQuant` is restricted to 3.9–4.0 bits for a fair comparison. `MoBiQuant` remains competitive with the strongest static PTQ methods and outperforms several widely used baselines across both LLaMA2 and LLaMA3 models. Table 13 further reports GSM8K performance, where the elastic `MoBiQuant` outperforms static 4-bit OmniQuant. These results demonstrate that `MoBiQuant` preserves run-time precision flexibility while maintaining strong downstream-task performance.

Table 12: Zero-shot comparison with static scalar quantization (SQ) methods. We report average accuracy over six zero-shot commonsense reasoning tasks.MoBiQuant matches or surpasses static PTQ baselines. "-" denotes that no direct results are reported in the published papers.

Method	Avg. Bits	Type	LLaMA-2-7B 0-shot Avg.(↑)	LLaMA-2-13B 0-shot Avg.(↑)	LLaMA-3.2-1B 0-shot Avg.(↑)	LLaMA-3.2-3B 0-shot Avg.(↑)	LLaMA-3-8B 0-shot Avg.(↑)
Floating-point	16	-	69.68	73.19	59.42	66.71	74.19
RTN	4	Static SQ	64.72	64.19	-	-	70.40
SmoothQuant			63.87	67.46	-	-	67.47
AWQ			**69.40**	72.64	-	-	**72.64**
GPTQ			65.54	70.22	-	-	70.21
SpinQuant			68.00	**72.68**	-	-	72.62
OmniQuant			68.41	71.68	**56.99**	**63.8**	71.66
MoBiQuant	3.9-4.0	AnyPrec.	67.22	71.54	55.49	63.2	71.84

E.7 Memory Savings under Multi-Precision Deployment.

We further evaluate the memory efficiency of `MoBiQuant` under multi-precision deployment. Conventional approaches require separately storing a quantized model for each supported precision, causing the total memory footprint to grow with the number of precision options. In contrast, `MoBiQuant` represents multiple precisions within a single model by composing different prefixes of the shared bit-slices. It therefore supports run-time precision adjustment without storing separate weight copies. As shown in Fig. 11, this many-in-one representation reduces the total model memory footprint by up to $3.5\times$ compared with separate multi-precision deployment.

Table 13: GSM8K performance of LLaMA3.2-1B under 4-bit setting.

Method	Flexible-Extract	Strict-Match
FP16	46.47%	42.15%
OmniQuant-4bit	30.86%	30.86%
Ours (Elastic)	**32.83%**	**32.45%**

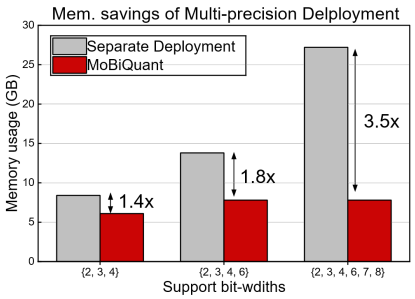

Figure 11: **Memory savings under multi-precision deployment.** Conventional deployment stores a separate quantized model for each supported precision, whereas `MoBiQuant` supports multiple precisions within a single composable representation, reducing the total model memory footprint by up to $3.5\times$.

F Limitations

While `MoBiQuant` demonstrates strong accuracy–efficiency trade-offs across multiple model families and bit budgets, several limitations remain. First, our experiments are primarily conducted on small- to medium-scale LLMs (up to 32B parameters). We have not yet evaluated the method on ultra-large models such as 70B-scale LLMs due to the substantial computational and memory requirements of training and benchmarking elastic multi-precision quantization systems at that scale. We leave large-scale evaluation as future work. Second, although we implement optimized custom kernels for our method, the current system has not yet been fully integrated with mature production inference frameworks such as vLLM or FlashInfer. Finally, the behavior of token-wise adaptive precision under large-scale cloud serving workloads with heterogeneous batching strategies remains an important direction for future investigation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract, section 1

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix F

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 3, 5

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix C

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides a code link in the abstract, along with sufficient instructions to faithfully reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5, Appendix C

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 5, Appendix C, E

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5, Appendix C

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors affirm that the research presented in this paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The proposed method is a systems and efficiency optimization technique for LLM quantization and deployment, and we do not identify any direct potential malicious or unintended uses beyond those already associated with general-purpose language models. Therefore, the paper does not include a dedicated societal impact discussion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets or foundation models with elevated misuse risks, and therefore does not include additional safeguard mechanisms for release. All datasets and pre-trained models used in this work are publicly available open-source resources released by their original providers.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets, codebases, and pre-trained models used in this work are properly credited through citations to their original sources. We also follow the corresponding licenses and terms of use provided by the original authors and organizations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets`

has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This work does not introduce new public datasets, models, or other released assets requiring additional documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects or study participants, and therefore no IRB approval or participant risk disclosure was required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were only used for writing, editing, and formatting assistance, and were not used as part of the core methodology, experiments, or scientific contributions of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.