

Dynamic Expert Quantization for Scalable Mixture-of-Experts Inference

Kexin Chu

University of Connecticut
Storrs, CT, USA

Yiwei Yang

University of California, Santa Cruz
Santa Cruz, CA, USA

Dawei Xiang

University of Connecticut
Storrs, CT, USA

Zecheng Lin

Independent Researcher
USA

Zixu Shen

University of Connecticut
Storrs, CT, USA

Wei Zhang*

University of Connecticut
Storrs, CT, USA

Abstract

Mixture-of-Experts (MoE) models scale LLM capacity efficiently, but deployment on consumer GPUs is limited by the large memory footprint of inactive experts. Static post-training quantization reduces storage cost but cannot adapt to shifting activation patterns, causing accuracy loss under aggressive compression. So we present DynaExq, a runtime system that treats expert precision as a first-class, dynamically managed resource. DynaExq combines (1) a hotness-aware precision controller that continuously aligns expert bit-widths with long-term activation statistics, (2) a fully asynchronous precision-switching pipeline that overlaps promotion and demotion with MoE computation, and (3) a fragmentation-free memory pooling mechanism that supports hybrid-precision experts with deterministic allocation. Together, these components enable stable, non-blocking precision transitions under strict HBM budgets.

Across Qwen3-30B and Qwen3-80B MoE models and six representative benchmarks, DynaExq deploys large LLMs on single RTX 5090 and A6000 GPUs and improves accuracy by up to 4.03 points over static low-precision baselines. The results show that adaptive, workload-aware quantization is an effective strategy for memory-constrained MoE serving.

1 Introduction

Recent advances in natural language processing have been driven by the rapid scaling of large language models (LLMs) [20, 27, 28]. Among them, Mixture-of-Experts (MoE) architectures [36] provide an attractive path to efficiency by activating only a small subset of experts per token while preserving a high overall parameter capacity. Modern MoE models, such as Qwen-MoE [34] and DeepSeek-MoE [19], consistently outperform dense counterparts [7], yet their deployment remains constrained by memory footprint. To support dynamic routing, all experts must reside in GPU memory, causing inactive experts to dominate the total parameter load. For example, Qwen3-30B-A3B requires roughly 57GB of weights to be present on device despite using only about 3B (6.1GB) parameters per token. As MoE layers contribute more than 98% of the model size, compressing experts is essential for running such models on memory-limited consumer GPUs like the RTX-5090 or A6000.

Post-training quantization (PTQ) provides a practical means to reduce memory usage via low-bit representations. Methods such as AWQ [18] and GPTQ [9] demonstrate strong compression on dense

LLMs, and recent efforts extend PTQ to MoE experts. Techniques including MoQE [15], MxMoE [6], EaQuant [10], and QMoE [8] push precision down to 2–4 bits or even sub-1-bit regimes. Although it incurs about a 6.7% accuracy drop compared with the full-precision model. In addition, all prior MoE quantization remains *static*: expert bit-widths are fixed offline, based on limited calibration data, and cannot adjust to the highly dynamic activation patterns observed during inference. As a result, hot experts may suffer unnecessary degradation under aggressive quantization, while cold experts retain oversized representations that waste scarce HBM capacity.

This work introduces **DynaExq** (Dynamic Expert Exchange Quantization), a runtime system that treats precision as a *first-class, dynamically managed resource*. DynaExq is motivated by the observation that expert activation patterns shift substantially across inputs and workloads. As shown in Fig. 1, only a small subset of experts remains consistently active within each layer, whereas most experts are rarely selected. By exploiting this temporal sparsity, DynaExq assigns precision adaptively: frequently activated (hot) experts retain high precision to maintain routing accuracy, whereas infrequently used (cold) experts are aggressively quantized or offloaded to reduce memory pressure. This runtime precision realignment enables substantial compression without compromising model fidelity.

DynaExq comprises two core mechanisms and two supporting components. The *Expert Precision Controller* maintains a sliding-window profile of expert activity and issues promotion or demotion decisions based on workload trends and memory feasibility. The *Asynchronous Swapping Pipeline* performs precision transitions in the background by staging weights across SSD, DRAM, and HBM using dedicated CUDA streams, ensuring that expert switching never interrupts the forward path. To guarantee allocator determinism, DynaExq adopts a dual-pool memory layout that isolates high- and low-precision experts, combined with a static threshold initialization procedure that respects per-layer HBM budgets.

Together, these mechanisms allow DynaExq to deliver adaptive quantization, stable routing behavior, and high-throughput inference for large MoE models on a single commodity GPU. The key contributions of this work are:

- We present **DynaExq**, a runtime system that tightly couples expert quantization with activation-aware precision scheduling for single-GPU MoE inference, managing expert precision online based on router statistics and memory budgets.

*Corresponding author. Email: wei.13.zhang@uconn.edu

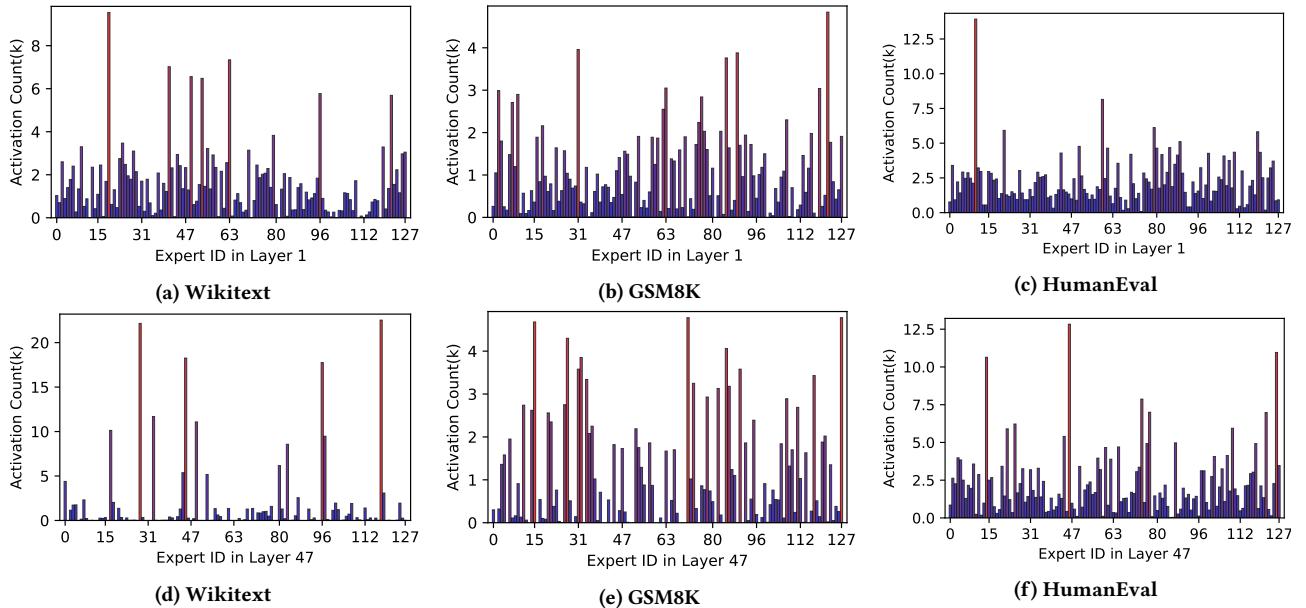


Figure 1: Expert activation distributions across layers and workloads. Each subplot shows routing frequency for 128 experts in layer 1 and layer 47.

- We introduce a multi-stage **asynchronous swapping pipeline** that overlaps data movement, quantization, and computation to ensure non-blocking precision transitions.
- We develop a **structured memory pool** that eliminates fragmentation and enables efficient concurrent expert promotions and demotions under tight memory constraints.

2 Background

2.1 Mixture-of-Experts (MoE) Architecture

Mixture-of-Experts (MoE) architectures scale model capacity by activating only a small subset of expert subnetworks for each input token. This sparse-computation paradigm enables models to reach hundreds of billions or even trillions of parameters without incurring proportional compute costs. Formally, the output of an MoE layer is:

$$y = \sum_{i=1}^m E_i^s(x)g_i(x) + \sum_{j \in \mathcal{K}} E_j^r(x)g_j(x), \quad (1)$$

where $\mathcal{K} = \text{topk}(\{g_i(x)\}_{i=1}^m)$ denotes the set of top- k routed experts selected by the gating network. $E_i^s(x)$ and $E_j^r(x)$ represent shared and routed expert functions, respectively, and $g_i(x)$ is the routing probability.

Although this design offers substantial representational flexibility, it introduces inherent system challenges. Expert activations are highly skewed and shift with input distribution, creating heavy-tailed utilization patterns across layers.

2.2 Motivations

While only a few experts are activated per token, all experts must remain resident in GPU memory to allow unrestricted routing.

This requirement dominates the HBM footprint, as inactive experts consume memory without contributing computation. Low-bit post-training quantization (PTQ) appears attractive for reducing this overhead, but current MoE PTQ methods apply a *static* bit-width to all experts. Such rigidity prevents the precision allocation from reflecting runtime activation behavior: heavily used experts may experience accuracy degradation under aggressive compression, whereas rarely used experts still occupy high-precision storage.

Recent MoE-serving systems mitigate memory pressure through expert offloading and caching, but they similarly assume that expert precision is fixed. As a result, expert placement and scheduling decisions remain decoupled from precision choices, limiting their ability to exploit activation sparsity and to adapt when workload distributions shift. This mismatch reveals a fundamental gap: MoE inference requires a mechanism that manages expert precision as a *runtime resource*, ensuring accuracy for hot-path experts while respecting strict memory budgets.

A practical solution must therefore:

- (1) *adapt precision according to sustained activation trends.*
- (2) *maintain strict memory safety during precision changes.*
- (3) *carry out all transitions without interfering with the forward pass.*

These constraints motivate the design of DynaExq.

2.3 Challenges

Enabling online precision adjustment for MoE experts on resource-constrained GPUs raises several system-level challenges:

- (1) **Workload-aware precision scheduling:** The system must track expert usage continuously and adjust precision decisions without relying on offline calibration or fixed activation assumptions.

- (2) **Non-blocking precision transitions:** Upgrades and downgrades must proceed asynchronously and in parallel with inference to prevent stalls or latency spikes.
- (3) **Fragmentation and bandwidth management:** Precision changes alter expert memory size, requiring structured memory management and efficient data movement to prevent fragmentation and bandwidth contention.
- (4) **Resource-adaptive thresholding:** Promotion and demotion policies must align with available HBM capacity while remaining robust to short-term activation fluctuations.

These challenges shape the design of DynaExq, whose architecture and runtime mechanisms are presented in the next section.

3 Design

3.1 Overview

DynaExq provides adaptive expert precision control for MoE inference by treating precision as a runtime-managed resource rather than a fixed configuration. Instead of relying on offline PTQ decisions that assume static activation distributions, DynaExq continuously monitors expert usage and adjusts precision according to long-term activation salience. This enables hot experts to retain accurate high-precision weights, while cold experts are aggressively compressed to reduce the overall HBM footprint without affecting the forward path.

As shown in Fig. 2, DynaExq consists of four coordinated components:

- (1) **Expert Precision Controller**, which aggregates router statistics, identifies persistently activated experts, and issues promotion or demotion commands.
- (2) **Asynchronous Precision Switching Pipeline**, which transfers expert weights across SSD, DRAM, and HBM with full compute–communication overlap.
- (3) **Memory Pool Manager**, which maintains deterministic, fragmentation-free memory allocation for mixed-precision experts.
- (4) **Resource-Aware Initialization Module**, which establishes per-layer precision budgets that guarantee memory stability under variable workloads.

Together, these components support accuracy-aware precision allocation, non-blocking transitions, and stable throughput. By decoupling precision switching from the critical compute path, DynaExq sustains aggressive memory savings even when expert usage shifts over time.

3.2 Runtime Expert Precision Controller

The controller maintains a lightweight profile of expert behavior by sampling router outputs during inference. After each forward pass, DynaExq records the selected experts and accumulates gating probabilities, providing a workload-driven view of expert importance rather than relying on synthetic calibration data.

For each expert i , the exponential-moving-average (EMA) hotness score within a window of length T is:

$$S_i^{(t)} = \alpha S_i^{(t-1)} + (1 - \alpha)g_i(x_t), \quad (2)$$

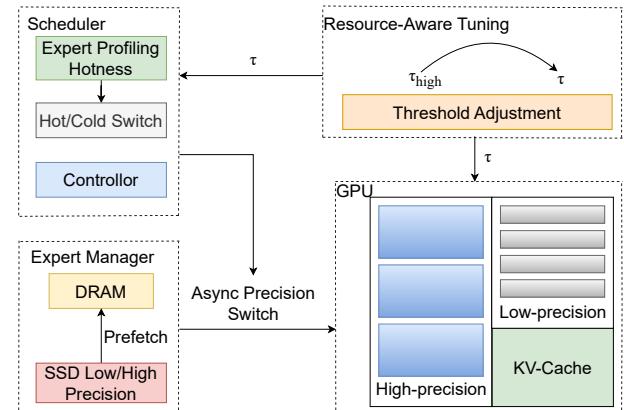


Figure 2: System architecture of DynaExq.

Algorithm 1 Runtime Precision Scheduling for MoE Experts

```

1: Input: Router probabilities  $g_i(x_t)$  for experts  $i = 1..N$ 
2: State: EMA hotness scores  $S_i$ , precision state  $\text{state}[i] \in \{\text{HIGH}, \text{LOW}\}$ 
3: Parameter: Update period  $T$ , EMA factor  $\alpha$ , threshold  $\tau$  (derived from memory budget)
4: procedure UPDATEHOTNESS( $t$ )
5:   for each activated expert  $i$  at step  $t$  do
6:      $S_i \leftarrow \alpha S_i + (1 - \alpha)g_i(x_t)$ 
7:   end for
8:   for each inactive expert  $j$  do
9:      $S_j \leftarrow \alpha S_j$                                  $\triangleright$  passive decay
10:  end for
11: end procedure
12: procedure PRECISIONSCHEDULE( $t$ )
13:   if  $t \bmod T = 0$  then
14:     Sort experts by  $S_i$  in descending order
15:      $\mathcal{H} \leftarrow$  top- $K$  experts with  $S_i \geq \tau$ 
16:     for each expert  $i$  do
17:       if  $i \in \mathcal{H}$  and  $\text{state}[i] = \text{LOW}$  then
18:         ENQUEUEUPGRADE( $i$ )
19:       else if  $i \notin \mathcal{H}$  and  $\text{state}[i] = \text{HIGH}$  then
20:         ENQUEUEDOWNGRADE( $i$ )
21:       end if
22:     end for
23:   end if
24: end procedure

```

where $g_i(x_t)$ is the router probability at step t . This formulation captures both activation frequency and magnitude, while smoothing short-term spikes. Experts with S_i above the high-precision threshold τ_h are promoted, while the remaining experts form the low-precision candidate pool (as shown in Algo-1).

To ensure robust scheduling under fluctuating workloads, the controller operates at coarse granularity and applies smoothing to avoid oscillations. Precision updates occur periodically rather than per batch, preventing thrashing when gating behavior is unstable.

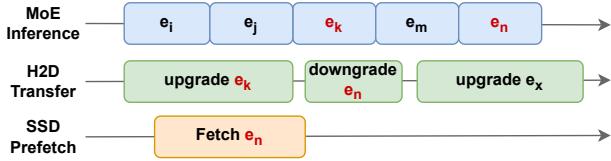


Figure 3: An example of Asynchronous Expert Swapping Pipeline.

Because all operations execute on the CPU and do not block GPU kernels, the controller imposes no latency overhead on inference.

These mechanisms allow DynaExq to detect persistent hot experts reliably, enforce stable precision decisions, and preserve routing fidelity under strict memory constraints.

3.3 Asynchronous Expert Swapping Pipeline

To support continuous inference while precision changes occur in the background, DynaExq employs a multi-stage asynchronous pipeline that overlaps promotion, demotion, and prefetching with MoE computation. High-precision and low-precision weights are stored on SSD and cached in DRAM, and all transfers to HBM use dedicated CUDA streams.

When a cold expert becomes hot, the runtime triggers a non-blocking upgrade sequence: the high-precision weight is fetched from SSD (if not cached), buffered in DRAM, and streamed into HBM. After the transfer completes, a lightweight registration step binds the new weight to the expert’s compute context, and the low-precision version is reclaimed immediately. Demotion mirrors this process in reverse.

During inference, DynaExq always uses the last stable version of each expert. If a promotion is in flight, the previous representation remains active, ensuring that pipeline latency never affects the forward pass. This provenance-consistent strategy prevents correctness issues and avoids compute stalls even under highly dynamic workloads.

To further hide data movement latency, DynaExq proactively prefetches experts predicted to become hot by leveraging cross-layer activation correlations. As illustrated in Fig. 3, SSD prefetch, DRAM buffering, H2D transfers, and demotions proceed concurrently beneath ongoing computation, enabling smooth precision transitions.

3.4 Memory Pool Management

Because precision switching changes expert storage sizes, naively using the GPU allocator can cause fragmentation and unpredictable memory growth. To prevent this, DynaExq partitions HBM into two dedicated pools for high- and low-precision experts. Each pool is subdivided into fixed-size blocks aligned with the corresponding data format (e.g., FP16 or INT4), enabling constant-time and fragmentation-free allocation.

During precision transitions, DynaExq immediately reclaims the old representation once the new weight is successfully registered, ensuring that no expert holds dual copies longer than necessary. A small transient buffer absorbs transfer bursts, allowing multiple in-flight promotions without exceeding the memory budget.

A compact CPU-side index records expert locations, precision tiers, and pool occupancy. Because this index is managed outside the GPU allocator, DynaExq avoids runtime `cudaMalloc` calls and eliminates allocator-induced jitter. This deterministic memory structure maintains stable throughput even when experts frequently change precision.

By imposing structure on the memory hierarchy and ensuring deterministic allocation rules, DynaExq provides predictable memory behavior and safe operation under aggressive dynamic quantization.

3.5 Resource-Aware Threshold Initialization

Before inference begins, DynaExq computes safe precision budgets for each MoE layer to ensure that dynamic switching never violates HBM constraints. Let M_{HBM} denote available memory after reserving space for KV-cache, activations, and shared weights. Given the average storage sizes of high-precision and low-precision experts, S^{hot} and S^{cold} , and the number of experts N , DynaExq computes feasible values of n_{hot} and n_{cold} that satisfy:

$$n_{\text{hot}} \cdot S^{\text{hot}} + n_{\text{cold}} \cdot S^{\text{cold}} \leq M_{\text{HBM}},$$

Where $n_{\text{hot}} + n_{\text{cold}} = N$ and $\tau_h = \text{top}(n_{\text{hot}}, S_i)$. A brief warmup phase collects initial hotness scores, and the top- n_{hot} experts initialize the high-precision set. The threshold τ_h remains fixed while expert identities may change at runtime based on observed activity.

By separating memory feasibility analysis from runtime scheduling, DynaExq guarantees that all precision transitions remain memory-safe even under shifting workloads. This initialization step forms the foundation for predictable and stable mixed-precision execution.

4 Evaluation

4.1 Setup

We implement DynaExq using PyTorch [22] and HuggingFace Transformers [31], with AutoRound [16] providing low-bit expert quantization. All experiments are conducted on **RTX-5090** and **A6000** GPUs (32-48 GB HBM), representing realistic memory-constrained single-device settings targeted by this work.

Models and Datasets. We evaluate DynaExq on **Qwen3-30B-A3B** [1] and **Qwen3-80B-A3B** [2]. The 30B model contains 48 layers and 128 experts per layer (8 active), while the 80B variant increases to 512 experts (10 active) with a comparable active parameter count of ~3B. Both models exhibit substantial expert-level redundancy, making them strong candidates for runtime precision adaptation.

We benchmark across diverse tasks: WikiText-2 [21], MMLU-Pro [30], GPQA [24], AIME25 [3], GSM8K [5], HumanEval [4], and MMLU-ProX [32]. These datasets jointly stress different aspects of model behavior, including linguistic and mathematical reasoning, multilingual understanding, and code generation.

To isolate the effect of expert-level quantization, all non-expert and routing components remain in full precision. We evaluate DynaExq against uniform FP16, INT4, and INT2 baselines, ensuring consistent routing dynamics and enabling a controlled study of expert precision strategies.

4.2 Perplexity under Expert Quantization

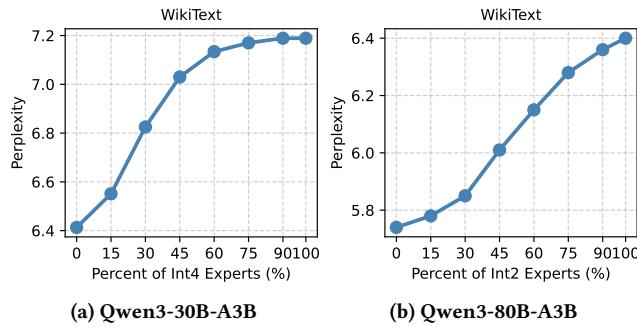


Figure 4: Perplexity Impact of Varying Low-Precision Expert Ratios Across Layers in DynaExq (Qwen3-30B-A3B: FP16 vs. Int4, Qwen3-80B-A3B: Int4 vs. Int2)

We first examine how varying the proportion of low-precision experts affects language modeling quality. Using WikiText-2, we test 128 prompts (2048 input tokens, 256 output tokens) while progressively increasing the number of demoted experts per layer. As Figure 4 shows, perplexity increases smoothly rather than abruptly, indicating that the model degrades predictably under more aggressive compression.

Crucially, DynaExq’s demotion policy prioritizes rarely activated experts, yielding significantly flatter perplexity curves relative to uniform quantization. This demonstrates that activation-aware precision assignment captures expert salience effectively and avoids unnecessary degradation on frequently used experts.

4.3 Accuracy Comparison

Table 1 summarizes end-task accuracy across six benchmarks. For Qwen3-MoE-30B, FP16 achieves 65.95% accuracy. Static INT4 reduces memory from 57 GB to 17 GB with modest accuracy loss. DynaExq closely matches FP16 accuracy on both GPUs (65.15% on A6000, 64.37% on RTX 5090), and retains robustness on challenging tasks such as HumanEval and MMLU-ProX.

For Qwen3-MoE-80B, full-precision inference is infeasible due to its 152 GB footprint, so we compare against INT4 (41 GB) and INT2 (21 GB) baselines. While static INT2 compresses aggressively, it suffers substantial degradation on reasoning-heavy evaluations. In contrast, DynaExq attains 76.68% accuracy on A6000 and 73.88% on RTX-5090, consistently outperforming static INT2 and approaching INT4 performance, demonstrating that runtime precision alignment is particularly beneficial at larger scales.

Across both models, DynaExq reduces accuracy loss induced by static low-bit quantization while remaining deployable under tight HBM constraints, validating its ability to allocate precision where it most impacts task quality.

4.4 Performance Metrics

We further assess latency and throughput: time-to-first-token (TTFT), time-per-output-token (TPOP), prefill throughput, and decode throughput (Figures 5). DynaExq achieves performance between high-precision (FP16/INT4) and low-precision (INT2/INT4) baselines, typically retaining more than 85% of INT2 throughput.

Despite selectively assigning high precision only to hot-path experts, DynaExq maintains stable performance across both GPUs and both model sizes. This indicates that its asynchronous precision switching pipeline effectively decouples background transitions from the forward path, avoiding runtime stalls.

Taken together, these results show that DynaExq achieves a favorable balance: it recovers much of the accuracy benefit of high-precision experts while incurring only moderate overhead relative to uniformly low-precision inference. The resulting hybrid-precision execution enables large MoE models to run on single-GPU hardware otherwise unable to host full-precision variants. As model scales continue to grow, this precision-aware approach provides a practical path toward sustaining deployability under increasingly stringent memory limits.

5 Related Work

5.1 Quantization for Mixture-of-Experts

Recent MoE quantization techniques demonstrate that low-bit PTQ can substantially reduce expert memory cost. MoQE [35], QMoE [8], and MxMoE [6] show that 2–4 bit or sub-1-bit formats can preserve expert functionality, while MoEQuant [12] and EAQuant [10] refine sampling and calibration to reduce accuracy loss. However, these methods set expert precision *offline* and keep it fixed during inference, making them unable to react to dynamic or highly skewed expert activation patterns.

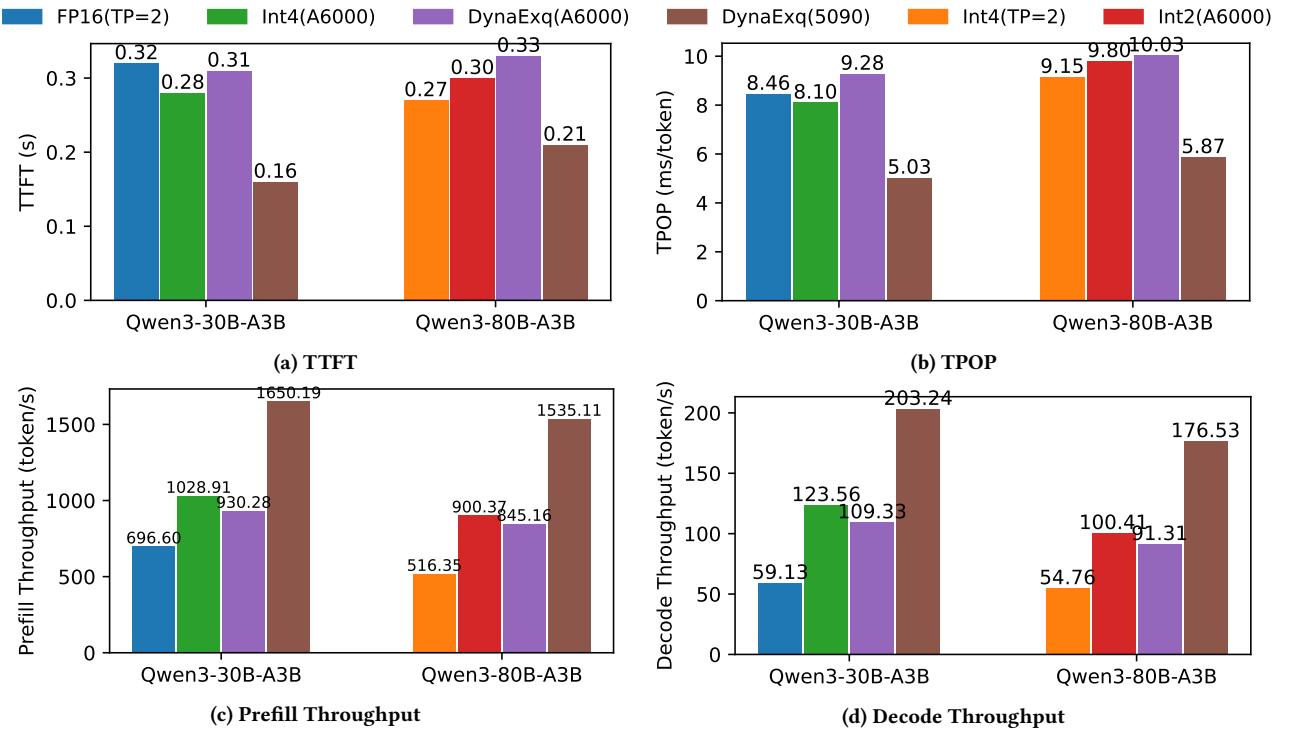
DynaExq differs by treating precision as a *runtime-managed resource*. Expert bit-widths are adjusted online based on EMA-smoothed activation statistics and memory budgets, enabling precision schedules that track long-term workload behavior rather than static calibration decisions.

5.2 MoE Serving Systems and Expert Offloading

MoE serving frameworks improve expert execution and data movement across heterogeneous memory tiers. DeepSpeed-MoE [23], Tutel [14], and Megablocks [11] optimize kernel and routing efficiency, while MoE-Infinity [33], ProMoE [26], and ExpertFlow [25] manage SSD/CPU offloading and expert placement.

These systems optimize *where* experts reside, but assume *fixed-precision* weights. Thus, bandwidth and memory usage scale with the full-precision representation of each expert, even when their activation rates differ dramatically. DynaExq complements these approaches by introducing precision as an additional control dimension: cold experts are aggressively compressed, reducing both HBM footprint and transfer volume, while hot experts retain high-fidelity representations.

MODEL	METHOD	MMLU-Pro	GPQA	AIME25	GSM8K	Human Eval	MMLU-ProX	AVG.
Qwen3-MoE-30B	FP16	73.37	54.55	23.33	90.45	84.76	69.29	65.96
	Int4	72.84	53.54	20.00	89.39	84.15	68.39	64.72
	DynaExq(A6000)	73.09	54.14	20.00	89.91	84.76	68.96	65.15
	DynaExq(5090)	72.58	54.54	20.00	89.54	84.48	68.39	64.37
Qwen3-MoE-80B	Int4	75.92	72.22	70.00	87.04	85.37	75.88	77.74
	Int2	72.75	66.67	63.33	80.97	81.71	70.49	72.65
	DynaExq(A6000)	75.48	71.21	66.67	86.73	84.76	75.22	76.68
	DynaExq(5090)	72.89	68.18	63.33	83.47	82.68	72.73	73.88

Table 1: The comparison of accuracy across model/method.**Figure 5: The comparison of latency/throughput across model/method.**

5.3 Dynamic Quantization and Adaptive Execution

Dynamic quantization methods (e.g., HAQ [29], AdaQuant [13], FlexRound [17]) select bit-widths based on hardware feedback or layer sensitivity, and adaptive inference techniques such as early exiting adjust computation depth per input. However, these approaches target dense models, operate at coarse granularity, and do not address expert-level activation skew in MoE architectures.

Similarly, MoE offloading systems adapt expert placement but treat weights as static objects with fixed precision. In contrast, DynaExq performs *expert-level precision scheduling* integrated into the serving runtime. By combining online hotness profiling, precision transitions, and a fragmentation-free memory pool, DynaExq enables activation-aware, memory-efficient MoE inference on a single GPU.

6 Conclusion

We present DynaExq, a runtime system that enables dynamic expert-level quantization for Mixture-of-Experts models on memory constrained GPUs. By combining workload-aware precision management, an asynchronous expert swapping pipeline, and a stable mixed-precision execution framework, DynaExq aligns expert precision with real activation behavior rather than static calibration choices. Our evaluation across diverse benchmarks shows that DynaExq preserves model quality while substantially reducing memory usage and improving efficiency relative to static low-bit baselines. These results demonstrate that integrating quantization into the serving runtime, treating precision as a controllable system resource, is an effective direction for scalable and responsive MoE inference on commodity hardware.

Future work includes extending precision scheduling with predictive models, exploring cross-layer correlated activation patterns, and integrating DynaExq into distributed MoE serving frameworks where expert placement and precision co-optimization may yield further gains.

References

- [1] Qwen3-30b-a3b-instruct-2507. <https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507>.
- [2] Qwen3-next-80b-a3b. <https://huggingface.co/Qwen/Qwen3-Next-80B-A3B-Instruct>.
- [3] BALUNOVIĆ, M., DEKONINCK, J., PETROV, I., JOVANOVIĆ, N., AND VECHEV, M. Matharena: Evaluating llms on uncontaminated math competitions, Feb. 2025.
- [4] CHEN, M., TWOREK, J., JUN, H., YUAN, Q., DE OLIVEIRA PINTO, H. P., KAPLAN, J., EDWARDS, H., BURDA, Y., JOSEPH, N., BROCKMAN, G., RAY, A., PURI, R., KRUEGER, G., PETROV, M., KHLAAF, H., SAstry, G., MISHKIN, P., CHAN, B., GRAY, S., RYDER, N., PAVLOV, M., POWER, A., KAISER, L., BAVARIAN, M., WINTER, C., TILLET, P., SUCHI, F. P., CUMMINGS, D., PLAPPERT, M., CHANTZIS, F., BARNES, E., HERBERT-VOSS, A., GUSS, W. H., NICHOL, A., PAINO, A., TEZAK, N., TANG, J., BABUSCHKIN, I., BALAJI, S., JAIN, S., SAUNDERS, W., HESSE, C., CARR, A. N., LEIKE, J., ACHIAM, J., MISRA, V., MORIKAWA, E., RADFORD, A., KNIGHT, M., BRUNDAGE, M., MURATI, M., MAYER, K., WELINDER, P., MCGREW, B., AMODEI, D., MCCANDLISH, S., SUTSKEVER, I., AND ZAREMBA, W. Evaluating large language models trained on code, 2021.
- [5] COBBE, K., KOSARAJU, V., BAVARIAN, M., CHEN, M., JUN, H., KAISER, L., PLAPPERT, M., TWOREK, J., HILTON, J., NAKANO, R., HESSE, C., AND SCHULMAN, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* (2021).
- [6] DUANMU, H., LI, X., YUAN, Z., ZHENG, S., DUAN, J., ZHANG, X., AND LIN, D. Mxmoe: Mixed-precision quantization for moe with accuracy and performance co-design, 2025.
- [7] DUBEY, A., JAURHI, A., PANDEY, A., KADIAN, A., AL-DAHLE, A., LETMAN, A., MATHUR, A., SCHELTEN, A., YANG, A., FAN, A., ET AL. The llama 3 herd of models. *arXiv e-prints* (2024), arXiv-2407.
- [8] FRANTAR, E., AND ALISTARH, D. Qmoe: Practical sub-1-bit compression of trillion-parameter models, 2023.
- [9] FRANTAR, E., ASHKBOOS, S., HOEFLER, T., AND ALISTARH, D. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323* (2022).
- [10] FU, Z., DING, N., HAN, K., YU, X., LI, X., CHEN, X., TANG, Y., AND WANG, Y. Equant: Enhancing post-training quantization for moe models via expert-aware optimization, 2025.
- [11] GALE, T., NARAYANAN, D., YOUNG, C., AND ZAHARIA, M. Megablocks: Efficient sparse training with mixture-of-experts. *Proceedings of Machine Learning and Systems 5* (2023), 288–304.
- [12] HU, X., CHEN, Z., YANG, D., XU, Z., XU, C., YUAN, Z., ZHOU, S., AND YU, J. Moequant: Enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance, 2025.
- [13] HUBARA, I., NAHSAN, Y., HANANI, Y., BANNER, R., AND SOUDRY, D. Accurate post training quantization with small calibration sets. In *International conference on machine learning* (2021), PMLR, pp. 4466–4475.
- [14] HWANG, C., CUI, W., XIONG, Y., YANG, Z., LIU, Z., HU, H., WANG, Z., SALAS, R., JOSE, J., RAM, P., ET AL. Tutel: Adaptive mixture-of-experts at scale. *Proceedings of Machine Learning and Systems 5* (2023), 269–287.
- [15] KIM, Y. J., FAHIM, R., AND AWADALLA, H. H. Mixture of quantized experts (moqe): Complementary effect of low-bit quantization and robustness, 2023.
- [16] LABS, I. Autoround: Advanced quantization algorithm for llms. <https://github.com/intel/auto-round>, 2024.
- [17] LEE, J. H., KIM, J., KWON, S. J., AND LEE, D. Flexround: Learnable rounding based on element-wise division for post-training quantization, 2024.
- [18] LIN, J., TANG, J., TANG, H., YANG, S., CHEN, W.-M., WANG, W.-C., XIAO, G., DANG, X., GAN, C., AND HAN, S. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems 6* (2024), 87–100.
- [19] LIU, A., FENG, B., XUE, B., WANG, B., WU, B., LU, C., ZHAO, C., DENG, C., ZHANG, C., RUAN, C., ET AL. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
- [20] LIU, X., ZHENG, M., CAO, J., WEI, D., LAI, K., LAN, H., SONG, Y., AND LI, X. Efficient partitioning deep learning models for medical image analysis on iot devices. In *Proceedings of the 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2025), IEEE.
- [21] MERITY, S., XIONG, C., BRADBURY, J., AND SOCHER, R. Pointer sentinel mixture models, 2016.
- [22] PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHTEIN, N., ANTIGA, L., ET AL. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems 32* (2019).
- [23] RAJBHANDARI, S., LI, C., YAO, Z., ZHANG, M., AMINABADI, R. Y., AWAN, A. A., RASLEY, J., AND HE, Y. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *International conference on machine learning* (2022), PMLR, pp. 18332–18346.
- [24] REIN, D., HOU, B. L., STICKLAND, A. C., PETTY, J., PANG, R. Y., DIRANI, J., MICHAEL, J., AND BOWMAN, S. R. Gpq: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling* (2024).
- [25] SHEN, Z., CHU, K., ZHANG, Y., XIANG, D., WU, R., AND ZHANG, W. Expertflow: Adaptive expert scheduling and memory coordination for efficient moe inference, 2025.
- [26] SONG, X., ZHONG, Z., CHEN, R., AND CHEN, H. Promoe: Fast moe-based llm serving using proactive caching. *arXiv preprint arXiv:2410.22134* (2024).
- [27] TOUVRON, H., LAVRIL, T., IZACARD, G., MARTINET, X., LACHAUX, M.-A., LACROIX, T., ROZIÈRE, B., GOYAL, N., HAMBRO, E., AZHAR, F., ET AL. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [28] VASWANI, A., SHAZER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., AND POLOSUKHIN, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).
- [29] WANG, K., LIU, Z., LIN, Y., LIN, J., AND HAN, S. Haq: Hardware-aware automated quantization with mixed precision, 2019.
- [30] WANG, Y., MA, X., ZHANG, G., NI, Y., CHANDRA, A., GUO, S., REN, W., ARULRAJ, A., HE, X., JIANG, Z., ET AL. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems 37* (2024), 95266–95290.
- [31] WOLF, T., DEBUT, L., SANH, V., CHAUMOND, J., DELANGUE, C., MOI, A., CISTAC, P., RAULT, T., LOUF, R., FUNTOWICZ, M., ET AL. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (2020), pp. 38–45.
- [32] XUAN, W., YANG, R., QI, H., ZENG, Q., XIAO, Y., FENG, A., LIU, D., XING, Y., WANG, J., GAO, F., ET AL. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497* (2025).
- [33] XUE, L., FU, Y., LU, Z., MAI, L., AND MARINA, M. Moe-infinity: Offloading-efficient moe model serving. *arXiv preprint arXiv:2401.14361* (2024).
- [34] YANG, A., LI, A., YANG, B., ZHANG, B., HUI, B., ZHENG, B., YU, B., GAO, C., HUANG, C., LV, C., ET AL. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
- [35] ZHANG, J., ZHANG, Y., ZHANG, B., LIU, Z., AND CHENG, D. Moqe: Improve quantization model performance via mixture of quantization experts. *arXiv preprint arXiv:2508.09204* (2025).
- [36] ZHOU, Y., LEI, T., LIU, H., DU, N., HUANG, Y., ZHAO, V., DAI, A. M., LE, Q. V., LAUDON, J., ET AL. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems 35* (2022), 7103–7114.