

DLLMQuant: Quantizing Diffusion-based Large Language Models

Chen Xu¹, Dawei Yang¹

Houmo AI

Abstract

Diffusion-based large language models (DLLMs) have shown promise for non-autoregressive text generation, but their deployment is constrained by large model sizes and heavy computational costs. Post-training quantization (PTQ), a widely used method for compressing and accelerating Large Language Models (LLMs), suffers from severe accuracy degradation and reduced generalization performance when directly applied to DLLMs (e.g., AWQ suffers a 16% accuracy drop on LLADA under W4A4). This paper explores how DLLMs’ key mechanisms—dynamic masking, iterative generation, bidirectional attention—clash with quantization. We identify three core issues: 1) Iterative generation and dynamic masking ratios lead to distinct token distributions across decoding steps, which are not adequately captured by existing PTQ calibration methods; 2) Quantization errors are accumulated and amplified progressively during iteration in DLLMs, causing quantized models to perform worse as decoding steps progress; 3) Unmasked tokens stabilize while masked remain probabilistic, making overall feature distribution incompatible with existing PTQ methods. To address these issues, we propose DLLMQuant, a PTQ framework tailored for DLLMs, which incorporates three novel techniques: 1) Temporal-Mask Adaptive Sampling (TMAS), a calibration method that accounts for both time and mask factors, with the capacity to capture distributions across timesteps. 2) Interaction-Aware Activation Quantization (IA-AQ), which utilizes bidirectional attention’s interaction signals to dynamically allocate quantization resources. 3) Certainty-Guided Quantization (CGQ), which integrates mask status and token scores as key weighting criteria into error compensation, making weight quantization more suitable for DLLMs. Experiments show that DLLMQuant achieves significant performance gains (e.g., over 10-point accuracy improvement on GSM8K for LLADA under 4-bit quantization) while enhancing efficiency.

Introduction

Diffusion-based large language models (DLLMs) have recently attracted growing attention due to their unique advantages and potential applications. Drawing inspiration from diffusion (Rombach et al. 2022) processes, they leverage forward masking and reverse recovery to predict masked tokens. By reframing text generation as a denoising task, DLLMs enable parallel decoding while enhancing control over output structure. Notably, they demonstrate strong scal-

ability and even outperform autoregressive-based large language models (LLMs) (Kasneci et al. 2023; Bai et al. 2023; Touvron et al. 2023) in specific scenarios—such as addressing the reversal curse (Berglund et al. 2023)—highlighting the potential of diffusion models in handling complex language tasks.

However, DLLMs (Nie et al. 2025; Ye et al. 2025) still face issues in practical deployment. Simultaneous decoding of multiple tokens tends to degrade generation quality, yet decoding fewer tokens at once leads to a multiplicative increase in the average computational cost per token—tens or even hundreds of times that of autoregressive-based LLMs of comparable scale (Bai et al. 2023; Touvron et al. 2023; Dubey et al. 2024). This dilemma arises from the diffusion mechanism inherent in DLLMs, which introduces significant computational burdens. DLLMs initialize an entire response sequence upfront and perform iterative generation with bidirectional attention, resulting in enormous computational overhead. Additionally, these models are characterized by large parameter sizes: to enable sufficient interaction between tokens, they are designed with feed-forward network (FFN) layers that contain a substantial number of parameters. Thus, compressing DLLMs and reducing their computational footprint become critical for lowering inference costs and deployment on resource-constrained devices with limited memory and bandwidth.

Post-Training Quantization (PTQ), which quantizes weights and activations into low-precision formats, effectively reduces memory usage and computational overhead, achieving notable success in LLMs (Hu et al. 2025; Xu et al. 2025a,b; Frantar et al. 2022; Xiao et al. 2023). However, directly applying these PTQ approaches to DLLMs leads to substantial performance degradation, particularly in generalization capabilities. For instance, applying AWQ (Lin et al. 2024) to LLADA-8B (Nie et al. 2025) model leads to more than 16% accuracy decline.

We perform a comprehensive analysis of what undermines the quantization performance of DLLMs, and identify three critical issues. Firstly, DLLMs decode a fixed-length sequence that is initialized entirely with mask tokens through multiple iterations. This iterative process leads to divergent input distributions across time steps. For example, as shown in Fig. 1, feature distributions at early steps differ markedly from those at later ones. This temporal dis-

tribution shift poses a significant challenge for PTQ methods due to the difficulty of capturing distributions across all time steps and varying mask ratios. Secondly, the iterative generation mechanism introduces another barrier: the output at each time step serves as input for the next prediction, causing quantization errors to propagate and potentially amplify over iterations—a phenomenon illustrated in Fig. 2. As a result, the performance of quantized models undergoes a progressive decline as iterations proceed. Thirdly, DLLMs employ unique masking and remasking strategies: tokens already decoded remain fixed across iterations, while masked are selectively decoded based on model confidence scores. The evolving process introduces significant disparities in feature distributions across both the token and channel dimensions within certain layers, which undermines the effectiveness of GPTQ-based methods. This is because GPTQ uniformly treats all tokens when computing the Hessian matrix—used as a weighting factor for error compensation during weight quantization. However, this uniform assumption fails to account for the intrinsic variability in token importance, leading to substantial performance degradation when GPTQ is applied to DLLMs.

To this end, we propose DLLMQuant, an effective and efficient PTQ framework tailored for DLLMs. DLLMQuant incorporates three novel techniques: 1) Temporal-Mask Adaptive Sampling (TMAS), which is a calibration sampling scheme tailored for the iterative generation process of DLLMs. It captures temporal variations and masking ratio changes during decoding. By strategically selecting calibration data, it restores most of the performance of INT4 quantized models after calibration, emerging as an effective sampling strategy for correcting quantization errors. 2) Interaction-Aware Activation Quantization (IA-AQ), which mitigates the accumulation of errors in iterative steps. Our analysis identifies that quantization of the matrix multiplication following softmax operation in the attention mechanism is a primary source of error propagation. IA-AQ resolves this by computing quantization parameters for the attention module’s value matrix via interaction-aware metrics, sharply reducing errors at this critical point. 3) Certainty-Guided Quantization (CGQ), which is a weight quantization strategy that leverages DLLMs’ unique masking and re-masking mechanisms to alleviate the adverse effects of weight quantization. By integrating these three methods, DLLMQuant bridges existing quantization techniques with DLLM architectures, reconciling the performance of quantized systems with the unique requirements of DLLMs. Our contributions are summarized as follows:

- We identify three critical factors that affect the quantization performance of DLLMs: issues in calibration selection, temporal accumulation of quantization errors, and distinct feature distributions induced by unique decoding and re-masking mechanisms.
- We propose TMAS, a calibration scheme adapted to iterative generation in DLLMs; CGQ and IA-AQ, which leverage interaction-aware metrics and certainty guidance to facilitate activation and weight quantization tailored to DLLMs.

- We present DLLMQuant, which seamlessly integrates TMAS, IA-AQ, and CGQ with existing PTQ methods, significantly boosting DLLM quantization performance. As one of the first studies in this domain, we will release the code to facilitate further exploration and advance research in this field.

Related Work

Large Language Diffusion Models

To address issues such as slow generation speed and reversal curse in autoregressive LLMs, LLaDA (Nie et al. 2025) first proposed DLLM. Inspired by diffusion models (Croitoru et al. 2023), LLaDA characterizes distributions via two processes: a forward data masking process and a reverse process parameterized by a vanilla Transformer (Fedus, Zoph, and Shazeer 2022) to predict masked tokens. The core of LLaDA is a *mask predictor*, a parametric model $p_\theta(\cdot | x_t)$ that takes x_t as input and predicts all masked tokens (denoted \mathbf{M}) simultaneously. Cross-entropy loss is applied to the masked tokens:

$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{t, x_0, x_t} \left[\frac{1}{t} \sum_{i=1}^L \mathbf{1}[x_t^i = \mathbf{M}] \log p_\theta(x_0^i | x_t) \right] \quad (1)$$

where x_0 is sampled from the training data, t is sampled uniformly from $[0, 1]$, and x_t is sampled from the forward process. The indicator function $\mathbf{1}[\cdot]$ ensures that the loss is computed only for masked tokens. This enables DLLMs to decode multiple tokens simultaneously while maintaining excellent context-aware capabilities. DiffuLLaMA (Gong et al. 2024) introduces an ingenious “transformation” approach: it converts pretrained autoregressive models (e.g., LLaMA (Touvron et al. 2023)) into DLLMs via adaptive training, significantly reducing the cost compared to training from scratch. LLaDA-1.5 (Zhu et al. 2025) successfully applies RLHF-like preference alignment techniques to DLLMs, solving the core problem of large variance in diffusion models’ ELBO estimation and significantly improving the model’s alignment ability. Multimodal models based on DLLM—such as LaViDa (Li et al. 2025) and LLaDA-V (You et al. 2025)—have achieved state-of-the-art performance in multimodal understanding tasks, demonstrating the great potential of the end-to-end diffusion paradigm in the multimodal domain.

Quantization

Quantization involves mapping floating-point numbers to discrete intervals using integer values. When it comes to weight quantization, our focus lies on per-channel symmetric uniform quantization, which is a scheme that has been widely adopted. The quantization process is defined in the following manner:

$$\mathcal{Q}(\mathbf{W}) = \text{clamp} \left(\left\lfloor \frac{\mathbf{W}}{s} \right\rfloor, q_{\min}, q_{\max} \right) \quad (2)$$

Here, $\mathbf{W} \in \mathbb{R}^{oc \times ic}$ denotes the weight matrix, $s \in \mathbb{R}^{oc}$ represents the channel-wise quantization step size, and q_{\min}, q_{\max} specify quantization bounds.

For the quantization of activations, we adopt the widely-used per-tensor asymmetric uniform quantization. The quantization process is expressed as follows:

$$\mathcal{Q}(\mathbf{X}) = \text{clamp} \left(\left\lfloor \frac{\mathbf{X} - \mathbf{z}}{s} \right\rfloor, q_{\min}, q_{\max} \right) \quad (3)$$

Here, $\mathbf{X} \in \mathbb{R}^{b \times ic}$ denotes the activation matrix, \mathbf{z} represents the asymmetric quantization zero point, which is computed as X_{\min}/s . For a linear layer, the loss introduced by quantizing both \mathbf{W} and \mathbf{X} can be formulated as:

$$\mathcal{L}(\mathbf{W}_q, \mathbf{X}_q) = \|\mathbf{W}\mathbf{X} - \text{Deq}(\mathbf{W}_q)\text{Deq}(\mathbf{X}_q)\|_F^2 \quad (4)$$

Here, Deq is the de-quantization process, \mathbf{X}_q and \mathbf{W}_q represent the quantized versions of \mathbf{W} and \mathbf{X} . Notable methods like AWQ (Lin et al. 2024) leverage such loss functions to guide selection of smoothing coefficients and weight pruning. GPTQ (Frantar et al. 2022) builds on OBQ (LeCun, Denker, and Solla 1989), which uses the Hessian matrix to compensate for quantization error. Combined with Eq. 4, the Hessian can be computed as:

$$\mathbf{H} = \mathbf{X}\mathbf{X}^\top \quad (5)$$

Post-Training Quantization for LLMs

Most large language models (LLMs) are constructed on the Transformer (Fedus, Zoph, and Shazeer 2022) framework, which is inherently characterized by high memory usage and substantial computational demands. Post-training quantization (PTQ) has established itself as a widely employed strategy for compressing LLMs, as it can effectively cut down memory and computational consumption while maintaining the model’s accuracy. Among the various PTQ techniques, GPTQ (Frantar et al. 2022) and AWQ (Lin et al. 2024) stand out and have undergone extensive research. GPTQ makes use of Hessian-based error compensation to reduce quantization errors, allowing for high compression ratios. AWQ, on the other hand, takes into account how activation distributions influence weight quantization, thereby improving the performance of the quantization process. Beyond these foundational approaches, several advanced techniques have been developed to enhance PTQ further. QuaRot (Ashkboos et al. 2024) utilizes Hadamard transformations to get rid of outliers without changing the output, which in turn improves the effectiveness of GPTQ. GPTVQ (Van Baalen et al. 2024) delves into non-uniform quantization schemes from a vector viewpoint, providing better adaptability to weight distributions.

However, these methods fail to account for the unique challenges inherent in DLLM architectures, resulting in significant accuracy degradation. Our proposed DLLMQuant, grounded in the interplay between quantization and the core mechanisms of DLLMs, is orthogonal to existing PTQ approaches. This characteristic enables its seamless integration with prior methods, thereby facilitating the effective quantization of DLLMs.

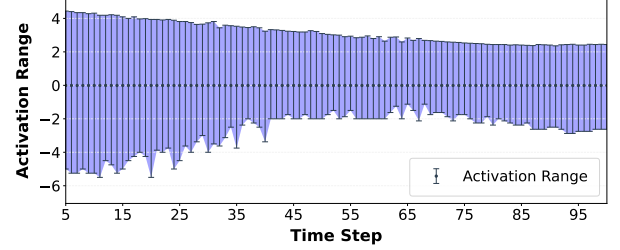


Figure 1: Activation range of outputs from the the first block in LLADA-8B across different time steps, showing significant variations.

Method

In this paper, we propose DLLMQuant, a framework designed for efficient quantization of DLLMs. It specifically addresses three core issues: Quantization errors accumulate across iterations, distinct token distributions across decoding steps and significant disparities in feature distributions across both token and channel dimensions. DLLMQuant tackles these issues from three aspects: optimizing calibration via Temporal-Mask Adaptive Sampling (TMAS), improving weight quantization with Certainty-Guided Quantization (CGQ), and enhancing activation quantization through Interaction-Aware Activation Quantization (IA-AQ). TMAS generates calibrations with proportionally selected data across time steps and masking ratios, ensuring the quantized model performs well throughout iterative generation. CGQ refines weight quantization compensation by incorporating token mask positions along with their final confidence scores. IA-AQ mitigates quantization error accumulation by leveraging bidirectional attention patterns during activation quantization.

All of the aforementioned solutions are plug-and-play, allowing seamless integration with other quantization techniques to enhance the quantization performance of DLLMs. These solutions are detailed in subsequent sections.

Temporal-Mask Adaptive Sampling

Current PTQ methods typically rely on calibration constructed by collecting activation information through random or uniform sampling. While these sampling methods can preserve reasonable generalization capabilities for standard LLMs, their direct application to DLLMs often leads to significant performance degradation. This performance degradation stems from the failure of existing methods to account for two key traits inherent to DLLMs: iterative decoding processes and dynamic masking ratios. These two factors collectively lead to variations in output distributions across different timesteps.

Given that the DLLMs use the same mask prediction network to process inputs at all time steps, determining an effective calibration sampling policy becomes a significant challenge. We begin by analyzing the output distributions of the model’s first block across different time steps. Specifically, we conduct an experiment on the LLADA-8B model

with 100 denoising steps and 4 blocks, plotting the activation ranges of 1,000 random samples across all time steps on the PIQA (Bisk et al. 2020a) dataset. It should be explained that DLLMs divide the total time steps into a number of blocks (such as 4 in this case) and then decode each block sequentially. As shown in Fig. 1, feature distributions gradually change, with neighboring time steps being similar and distant ones being distinctive.

Considering both the high similarity of output distributions across consecutive time steps and the block-based inference decoding mechanism of DLLMs, we propose a time- and mask-aware calibration method. Specifically, as detailed in Alg. 1, we sample inputs at specific intervals and proportions, ensuring they cover diverse masking ratios and span different time steps. This approach has the capability to represent distributions across all time steps. Empirically, we observe that the sampled calibration can restore most of the performance of INT4 quantized models after calibration, rendering it an effective sampling scheme for collecting calibration data in quantization error correction.

Algorithm 1: Temporal-Mask Adaptive Sampling (TMAS)

Require: Inputs \mathcal{X} , Block count B , Time steps T

Ensure: Calibration dataset \mathcal{D}_c

```

1:  $s \leftarrow \lfloor T/B \rfloor$  {Steps per block}
2:  $n \leftarrow \lfloor 512/B \rfloor$  {Samples per block}
3:  $\mathbf{p} \leftarrow n \cdot [0.3, 0.2, 0.2, 0.3]$  {Target proportion of per mask ratio interval}
4:  $\mathbf{C} \leftarrow \text{zeros}(B, 4)$  {Sampling counter matrix}
5:  $\mathcal{D}_c \leftarrow \emptyset$  {Initialize calibration data}
6: Function ClassifyMaskRatio( $r$ ):
7:   return  $[0, 1, 2, 3][\lfloor (r \geq 0.2) + (r \geq 0.5) + (r \geq 0.8) \rfloor]$ 
8: for  $x \in \mathcal{X}$  do
9:   for  $t = T - 1$  to 0 do
10:     $y_t \leftarrow \text{Model}(x)$ 
11:     $r_t \leftarrow |y_t|_{\text{unmasked}} / |y_t|_{\text{total}}$ 
12:     $m \leftarrow \text{ClassifyMaskRatio}(r_t)$ 
13:     $\text{block} \leftarrow \lfloor t/s \rfloor$ 
14:    if  $\mathbf{C}[\text{block}, m] < \mathbf{p}[m]$  then
15:       $\mathcal{D}_c \leftarrow \mathcal{D}_c \cup \{x\}$ 
16:       $\mathbf{C}[\text{block}, m] \leftarrow \mathbf{C}[\text{block}, m] + 1$ 
17:    end if
18:  end for
19: end for
```

Interaction-Aware Activation Quantization

Previous research has identified that quantization errors tend to accumulate across layers (Dao et al. 2022; Hu et al. 2025), making deeper neural networks more difficult to quantize. In DLLMs, at any time step t , the input to the mask prediction model (denoted as x_t) is derived from x_{t+1} , which is the model’s output at the previous time step $t + 1$. Quantization errors, which inherently accumulate across layers, are further compounded by the number of denoising steps in this iterative process. This leads to a geometric growth of total error as the model progresses through later denoising steps. As shown in Eq. 6, the quantization error $L(x_{t+1})$ at time step $t + 1$ propagates through the model’s iterations to

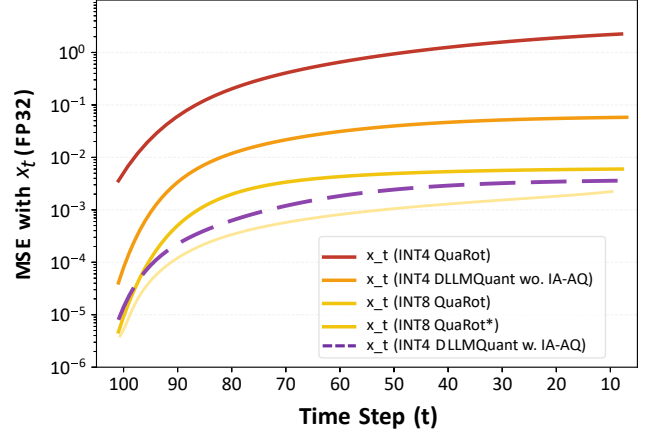


Figure 2: Cumulative quantization error of LLADA-8B over time steps under different methods (*QuaRot** denotes that *matmul* after softmax operation in attention remains unquantized.)

time step t , causing a further increase in the quantization error $L(x_t)$ at time step t . Here, $\mathcal{Q}_{\text{model}}$ denotes the quantized model, and Deq represents the dequantization operation.

$$\begin{aligned}
 L(x_t) &= x_t - \text{Deq}(\mathcal{Q}(x_t + L(x_{t+1}))) \\
 &= \mathcal{Q}_{\text{Model}}(x_{t+1}) - \mathcal{Q}_{\text{Model}}(\text{Deq}(\mathcal{Q}(x_{t+1})))
 \end{aligned} \tag{6}$$

We conduct experiments using the LLADA-8B model on the PIQA dataset, comparing the mean squared error (MSE) differences between the full-precision model and models quantized to INT8 and INT4 using different methods at each time step. As illustrated in Fig. 2, when the model is quantized to 4-bit using QuaRot, quantization errors increase significantly during iteration. This makes it difficult to preserve the performance of the model.

Fortunately, our experiments reveal a key factor in error accumulation under low-bit quantization. Specifically, this factor refers to the quantization error introduced by the matrix multiplication (*matmul*) between the output of the softmax operation and value matrix in attention. In the case of INT8 QuaRot quantization, we conduct comparative experiments where this specific component was either quantized or left unquantized. As shown in Fig. 2, the accumulated error when this component remains unquantized is significantly lower than when it is quantized.

To explore this further, we visualize the output features of this component, as shown in Fig. 3. We observe that the distribution of the value matrix V exhibits substantial variability across both token and channel dimensions, which poses issues for quantization. Additionally, the softmax output exhibits pronounced sparsity: larger values are concentrated near the diagonal and only within a small subset of tokens, while the rest are negligible. This phenomenon is closely tied to the unique interaction mechanisms of DLLMs, including their carefully designed bidirectional attention and large key-value (KV) heads. These features enable sufficient token interaction, endowing DLLMs with reverse reasoning

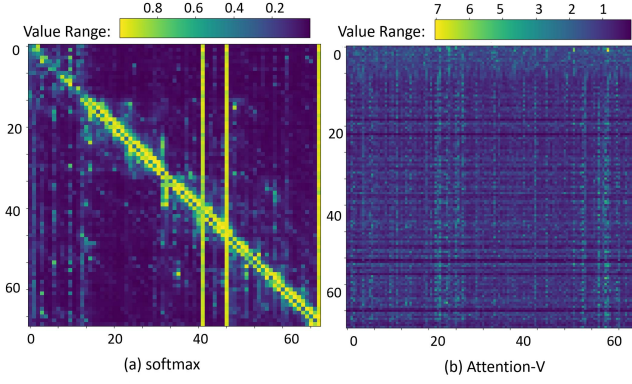


Figure 3: Output distributions of softmax (a) and value matrix (b) in LLADA - 8B attention. Softmax exhibits notable sparsity, while the value matrix shows significant distribution discrepancies across channels and tokens.

and contextual awareness.

$$\mathcal{L}(s) = \left\| \left(\left\lfloor \frac{V - z}{s} \right\rfloor - V \right) \cdot \text{Deq}(O_{\text{softmax}}) \right\|_F^2 \quad (7)$$

To mitigate the cumulative errors arising from quantization, we propose Interaction-Aware Activation Quantization (IA-AQ). Specifically, as described in Eq. 7, when calculating the quantization parameters for value matrix V prior to matrix multiplication, we redesign the quantization error metric by treating the softmax output as a weighting term. In Eq. 7, z denotes the zero-point, s represents the scaling factor and O_{softmax} represents the output of the quantized softmax function. To determine the optimal scaling factor s , we begin with the standard quantization scaling \hat{s} and test α values (stepping by 0.2 from 1.0 to 0.8) to minimize $L(\alpha \odot \hat{s})$:

$$\hat{s} = (V_{\max} - V_{\min}) / (Q_{\max} - Q_{\min}) \quad (8)$$

$$s = \alpha \odot \hat{s} = \arg \min_{\alpha \in \{1.0, 0.8\}} L(\alpha \odot \hat{s}) \quad (9)$$

This approach dynamically allocates quantization resources to suppress interference from features tied to tokens with weak interactions, while ensuring accurate quantization of features associated with tokens exhibiting both high interaction frequency and critical importance.

Certainty-Guided Quantization

As previously described, DLLMs perform iterative decoding with a fixed input-output length. Tokens that have been unmasked remain unchanged in subsequent iterations, while the masked portion is decoded based on the model’s final output scores. Therefore, treating masked and unmasked tokens equally during quantization is inappropriate. Specifically, as illustrated in Fig. 4, errors in the unmasked or low-score regions do not propagate through iterations and thus do not affect subsequent decoding steps.

We analyze the statistical distribution of output scores and find that only a small subset of tokens have relatively high scores, while the majority exhibit low scores. Notably, tokens with high scores are precisely those decoded in the current iteration, and their variations directly influence the input for the next iteration. In contrast, both low-score tokens and already decoded tokens do not affect subsequent iterations—the positions corresponding to low-score tokens remain masked in the following step.

$$H = (X \odot (\mathbf{1}[X_t = M] + \sqrt{sc_t})) \times (X \odot (\mathbf{1}[X_t = M] + \sqrt{sc_t}))^\top \quad (10)$$

Hessian-based PTQ methods typically quantize weights column-wise and adjust subsequent unquantized columns using statistically computed Hessian matrices to compensate for already quantized ones. However, conventional Hessian statistical computation fails to account for the aforementioned characteristics of DLLMs, leading to suboptimal performance. Based on this insight, we propose the Certainty-Guided Quantization (CGQ) method for optimizing weight quantization. Specifically, during quantization, we place greater emphasis on the unmasked regions with higher scores. In implementation, CGQ leverages weighted Hessian matrices to guide compensation during weight quantization. As shown in Eq. 10, when computing Hessian matrix, CGQ integrates coefficients derived from mask regions and token final scores, thereby guiding weight updates to minimize quantization errors.

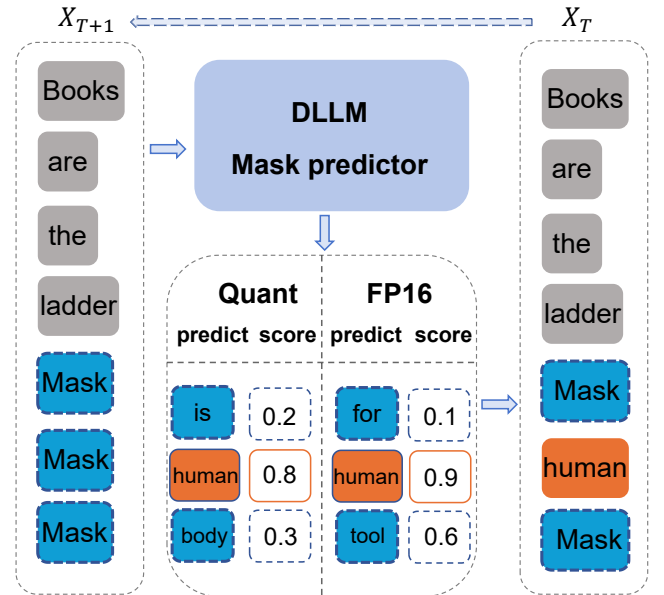


Figure 4: DLLMs iterative inference with masking and re-masking strategies. The quantization errors of unmasked tokens and masked tokens with low confidence scores do not affect the input of the next iteration.

Here, $\mathbf{1}[X_t = M]$ is a custom weighted indicator function. Specifically, for masked regions (i.e., regions where

Model	Method	Truth.	Arc.	Hella.	Wino.	PIQA	MMLU	C-EVAL	Hum.	GSM8K	Avg.
LLADA	FP	47.49	44.03	54.06	74.9	74.65	65.85	69.54	32.92	67.48	59.87
	RTN	40.45	41.83	45.40	64.72	67.95	49.26	57.95	14.02	16.56	44.23
	AWQ	40.87	42.92	46.14	66.88	69.43	51.22	58.43	20.10	36.88	48.09
	textDLLMQuant ⁺	41.53	43.44	46.51	67.87	70.12	51.72	59.38	22.13	40.66	49.26
	QuaRot	42.53	44.20	49.76	69.85	70.75	55.96	56.32	25.33	44.57	51.03
	textDLLMQuant ⁺⁺	43.53	44.18	51.00	71.85	73.94	57.77	61.22	28.92	56.25	54.29
LLADA-1.5	FP	47.2	88.5	74.7	74.8	74.86	66.0	70.05	49.4	83.3	69.86
	RTN	39.51	81.77	56.82	65.27	66.91	48.88	58.96	23.44	36.56	53.12
	AWQ	40.96	82.22	66.84	67.93	69.32	51.12	60.03	30.07	57.95	58.49
	textDLLMQuant ⁺	42.14	83.38	70.09	68.69	70.22	51.63	61.28	32.44	59.55	59.94
	QuaRot	43.21	84.23	65.34	69.55	70.17	56.23	57.66	37.33	65.86	61.06
	DLLMQuant ⁺⁺	43.87	84.18	69.20	71.74	73.58	57.27	60.04	44.58	74.33	64.31
DREAM	FP	49.76	59.80	73.30	74.50	75.66	69.5	64.89	57.9	77.2	66.94
	RTN	41.25	54.34	58.32	64.52	66.26	51.51	49.89	28.90	30.82	49.53
	AWQ	43.66	57.82	65.57	67.13	68.96	55.5	53.27	33.14	48.98	54.89
	textDLLMQuant ⁺	44.14	58.38	66.93	68.66	69.53	57.63	54.21	35.12	51.14	56.19
	QuaRot	47.58	58.18	67.13	70.05	70.36	69.5	53.19	34.48	59.20	58.85
	textDLLMQuant ⁺⁺	47.86	59.43	70.14	71.54	72.09	69.5	55.89	44.50	66.17	61.90

Table 1: Results of RTN, AWQ, QuaRot, and ours DLLMQuant with 4-bit weight and activation quantization among 9 tasks on LLADA-8B, LLADA-1.5-8B, DREAM-7B). DLLMQuant⁺ denotes DLLMQuant based on AWQ, and DLLMQuant⁺⁺ denotes DLLMQuant based on QuaRot.

$X_t = M$), it assigns a weight of 1; for unmasked regions (i.e., regions where $X_t \neq M$), it assigns a weight of 0.7. sc_t denotes the final confidence score assigned to each token in model output. Notably, unmasked regions, though no longer updated, remain non-negligible due to their interaction with masked regions that provide contextual information. Thus, we assign weights of 0.7 to unmasked segments. We acknowledge that more refined parameter tuning may yield better results; however, the current configuration already enables CGQ to effectively account for token masking states and final scores, thereby optimizing weight quantization and improving the performance of quantized models.

Experiments

In this section, we first describe the experimental setup, including the models, datasets, and baselines. We then present the results of comparative experiments across diverse datasets to validate the robustness of DLLMQuant. In addition, we conduct ablation studies and analyze the speed of both float16 and quantized models.

Setup

We adopt symmetric uniform quantization for weights and asymmetric uniform quantization for activations in DLLMs. Specifically, weight quantization is performed with per-channel granularity, while activation quantization uses per-token granularity. All experiments are conducted on NVIDIA A6000 GPUs, unless otherwise specified. As DLLMQuant is an efficient post-training quantization (PTQ) framework, it eliminates the need for any fine-tuning.

Models and Datasets. We conducted experiments on the LLADA-8B (Nie et al. 2025), LLADA-1.5-8B (Zhu et al.

2025), and DREAM-7B (Ye et al. 2025) models. Following the testing methods in the LLADA paper, we evaluate the accuracy metric on TruthfulQA-MC2 (Lin, Hilton, and Evans 2021), Arc-Challenge (Clark et al. 2018), HellaSwag (Zellers et al. 2019), WinoGrande (Sakaguchi et al. 2021), PIQA (Bisk et al. 2020b), MMLU (Hendrycks et al. 2021), and C-EVAL (Huang et al. 2023). Furthermore, we also evaluate DLLMQuant using HumanEval (Chen et al. 2021) and GSM8k (Cobbe et al. 2021). HumanEval evaluates code generation capabilities, while GSM8k assesses multistep mathematical reasoning skills.

Baseline Our primary baselines consist of vanilla RTN and the PTQ methods for LLMs: AWQ (Lin et al. 2024) and QuaRot (Ashkboos et al. 2024). For calibration, 128 segments from the WinoGrande dataset are selected. Floating-point results are provided as references. For QuaRot, following the implementation method in the official repository, we adopted the GPTQ method—a weight compensation approach based on Hessian matrices—to compensate for the quantized weights.

Implementation Details We use the accuracy testing methods provided in the official LLADA repository. To ensure the validity and fairness of the experiments, all experimental configurations are strictly kept consistent with those in the paper. In experiments involving AWQ, we adapt its official repository to support the three DLLMs.

Results

Comparison results. We comprehensively compare quantization performance across various DLLMs and tasks. As shown in Tab. 1, results from nine tasks demonstrate that our DLLMQuant outperforms other methods on DLLMs,

Method	Avg.
AWQ	48.09
AWQ + TMAS	48.63
AWQ + TMAS + CGQ	49.06
AWQ + TMAS + CGQ + IA-AQ	49.26
GPTQ	51.03
GPTQ + TMAS	52.36
GPTQ + TMAS + CGQ	53.16
GPTQ + TMAS + CGQ + IA-AQ	54.29

Table 2: Ablation study of proposed TMAS, CGQ, IA-AQ on AWQ/GPTQ baselines, evaluating 4-bit weight and activation quantization average performance (Avg.) of the LLADA model across nine tasks.

achieving the highest accuracy on nearly all tasks. Across the three DLLMs, it outperforms the original methods by an average of 2% based on the nine-task mean score. On some tasks such as TruthfulQA-MC2 and Arc-Challenge, the results based on QuaRot are not far behind DLLMQuant. However, on the HumanEval and GSM8k tasks, other methods like QuaRot degrade the model’s reasoning ability after quantization. In contrast, DLLMQuant effectively preserves the reasoning ability in generation tasks, achieving results comparable to full-precision models. This is particularly important as reasoning in complex tasks such as HumanEval is crucial for real-world applications, further highlighting the practical relevance of DLLMQuant’s performance.

Method	GSM8K	Hum.
GPTQ	44.57	25.33
GPTQ+ CGQ w. mask	45.07	25.76
GPTQ+ CGQ w. score	45.22	26.06
GPTQ+ CGQ w. score & mask	45.58	26.85

Table 3: Ablation study of the proposed CGQ method and its components (mask state and score) on the LLADA-8B model, evaluating 4-bit weight and activation quantization performance on GSM8K and HumanEval tasks.

Ablation results. DLLMQuant improves the quantization performance of DLLMs through three primary methods: TMAS, CGQ, and IA-AQ. To evaluate these methods, we conduct decomposition experiments. As can be seen in the Tab. 2, the addition of each individual method yields better metrics than when that method is not included.

In addition, as shown in Tab. 3, we conducted ablation experiments on the two variables of the CGQ method, namely the mask state and score, i.e., whether to include the relevant parts in Eq. 10. It can be observed that considering only one

of these two factors individually does not yield better results than taking both factors into account simultaneously.

Memory and Speedup

The core motivation of DLLMQuant lies in compressing large language diffusion models to a lower bitwidth, which aims to reduce both inference latency and GPU memory consumption while maximizing accuracy retention, thus ensuring practical applicability. As presented in Tab. 4, DLLMQuant achieves an average inference speedup of over 1.6x and memory savings exceeding 3.2x, marking substantial improvements in inference efficiency. These advancements facilitate the deployment of DLLMs on consumer-grade devices such as the Nvidia 4090 GPU.

MODEL	Speed (Tokens/s)			Memory (GB)		
	FP	Quant	Speed Up	FP	Quant	Mem. Sav.
LLADA	34.59	50.14	1.71	15.89	4.91	3.24
LLADA-1.5	35.55	60.43	1.70	15.88	4.90	3.24
DREAM	23.27	35.84	1.54	13.95	4.44	3.14

Table 4: Speedup and memory saving of three DLLMs, compared between our 4-bit implementation and FP16.

Conclusion

In this paper, we address the critical challenge of quantizing DLLMs. DLLMs feature unique mechanisms, including iterative generation, dynamic masking, and bidirectional attention. Conventional PTQ methods, while effective for standard LLMs, perform poorly when directly applied to DLLMs. We identify three core issues behind this failure: existing calibration methods fail to capture token distributions that vary with time steps and masking ratios; quantization errors accumulate and amplify across iterations; and conventional quantization strategies mismatch DLLM feature distributions, where fixed unmasked tokens coexist with probabilistic masked tokens. To address these issues, we propose DLLMQuant, a novel PTQ framework tailored for DLLMs, integrating three key techniques: Temporal-Mask Adaptive Sampling balances the size and representativeness of calibration, ensuring robust quantization throughout iterations. Interaction-Aware Activation Quantization mitigates error accumulation by dynamically allocating quantization resources to critical attention modules, particularly targeting matrix multiplication in the attention mechanism. Certainty-Guided Quantization enhances weight quantization by prioritizing error compensation for high-confidence masked tokens, which are soon to be decoded, and incorporating information from token mask states. Experiments on DLLMs demonstrate DLLMQuant outperforms baselines (RTN, AWQ, QuaRot), with top accuracy on most tasks, preserved reasoning capabilities, and 2% average gains. DLLMQuant bridges PTQ methods and DLLM architectures, enabling efficient compression and acceleration without significant accuracy degradation.

References

- Ashkboos, S.; Mohtashami, A.; Croci, M. L.; Li, B.; Cameron, P.; Jaggi, M.; Alistarh, D.; Hoefler, T.; and Hensman, J. 2024. Quarot: Outlier-free 4-bit inference in rotated llms. *Advances in Neural Information Processing Systems*, 37: 100213–100240.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A. C.; Korbak, T.; and Evans, O. 2023. The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”. *arXiv preprint arXiv:2309.12288*.
- Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020a. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439.
- Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020b. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 7432–7439.
- 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.; Such, 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. 2021. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374*.
- Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafford, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1*.
- 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. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.
- Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10850–10869.
- Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359.
- Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.
- Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.
- Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.
- Gong, S.; Agarwal, S.; Zhang, Y.; Ye, J.; Zheng, L.; Li, M.; An, C.; Zhao, P.; Bi, W.; Han, J.; et al. 2024. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*.
- Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Hu, X.; Cheng, Y.; Yang, D.; Xu, Z.; Yuan, Z.; Yu, J.; Xu, C.; Jiang, Z.; and Zhou, S. 2025. Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. *arXiv preprint arXiv:2501.13987*.
- Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; Fu, Y.; Sun, M.; and He, J. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *arXiv preprint arXiv:2305.08322*.
- Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.
- LeCun, Y.; Denker, J.; and Solla, S. 1989. Optimal brain damage. *Advances in neural information processing systems*, 2.
- Li, S.; Kallidromitis, K.; Bansal, H.; Gokul, A.; Kato, Y.; Kozuka, K.; Kuen, J.; Lin, Z.; Chang, K. W.; and Grover, A. 2025. LaViDa: A Large Diffusion Language Model for Multimodal Understanding. *arXiv preprint arXiv:2505.16839*.
- Lin, J.; Tang, J.; Tang, H.; Yang, S.; Chen, W.-M.; Wang, W.-C.; Xiao, G.; Dang, X.; Gan, C.; and Han, S. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6: 87–100.
- Lin, S.; Hilton, J.; and Evans, O. 2021. TruthfulQA: Measuring How Models Mimic Human Falsehoods. *arXiv:2109.07958*.
- Liu, Z.; Oguz, B.; Zhao, C.; Chang, E.; Stock, P.; Mehdad, Y.; Shi, Y.; Krishnamoorthi, R.; and Chandra, V. 2023. Llm-qat: Data-free quantization aware training for large language models. *arXiv preprint arXiv:2305.17888*.
- Nie, S.; Zhu, F.; You, Z.; Zhang, X.; Ou, J.; Hu, J.; Zhou, J.; Lin, Y.; Wen, J.-R.; and Li, C. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Van Baalen, M.; Kuzmin, A.; Koryakovskiy, I.; Nagel, M.; Couperus, P.; Bastoul, C.; Mahurin, E.; Blankevoort, T.; and Whatmough, P. 2024. Gptvq: The blessing of dimensionality for llm quantization. *arXiv preprint arXiv:2402.15319*.

Xiao, G.; Lin, J.; Seznec, M.; Wu, H.; Demouth, J.; and Han, S. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, 38087–38099. PMLR.

Xu, C.; Yue, Y.; Xu, Z.; Hu, X.; Yu, J.; Chen, Z.; Zhou, S.; Yuan, Z.; and Yang, D. 2025a. RWKVQuant: Quantizing the RWKV Family with Proxy Guided Hybrid of Scalar and Vector Quantization. *arXiv preprint arXiv:2505.03803*.

Xu, Z.; Yue, Y.; Hu, X.; Yuan, Z.; Jiang, Z.; Chen, Z.; Yu, J.; Xu, C.; Zhou, S.; and Yang, D. 2025b. Mambaquant: Quantizing the mamba family with variance aligned rotation methods. *arXiv preprint arXiv:2501.13484*.

Ye, J.; Xie, Z.; Zheng, L.; Gao, J.; Wu, Z.; Jiang, X.; Li, Z.; and Kong, L. 2025. Dream 7B.

You, Z.; Nie, S.; Zhang, X.; Hu, J.; Zhou, J.; Lu, Z.; Wen, J.-R.; and Li, C. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhu, F.; Wang, R.; Nie, S.; Zhang, X.; Wu, C.; Hu, J.; Zhou, J.; Chen, J.; Lin, Y.; Wen, J.-R.; et al. 2025. LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models. *arXiv preprint arXiv:2505.19223*.

A. Overall Results of Ablation Results

Tab. 5 presents the complete results of the ablation study on the proposed TMA, CGQ, and IA-AQ based on AWQ/G-PTQ baselines. It evaluates the average performance (Avg.) of 4-bit weight and activation quantization for the LLADA model across nine tasks.

Tab. 6 presents the performance of 4-bit weight and activation quantization for the LLADA model on GSM8K and HumanEval tasks, where the calibration sets are constructed based on different sampling methods and the quantization is based on RTN. Here, LLMQAT (Liu et al. 2023) employs a self-generated calibration approach.

Method	GSM8K	HumanEval
RTN + Random_calib	16.56	14.02
RTN + LLMQAT	15.43	13.34
RTN + Uniform_time	17.44	15.82
RTN + TMA	18.12	16.56

Table 6: The performance of 4-bit weight and activation quantization for the LLADA under different sampling methods.

B. Output distribution of specific layers in DLM

Fig. 5 shows the distribution of softmax output in different blocks of LLADA. It can be observed that, across the entire model, the softmax outputs exhibit a relatively obvious sparsity. Except for the areas near the diagonal and some individual tokens with larger values, the values in other regions are very small.

Fig. 6 shows per-channel distribution of the FFN outputs in the attention mechanism of the first block of LLADA. It can be observed that there is an obvious difference in value distribution between the first iteration (i.e., (a) in the figure) and the last iteration (i.e., (c) in the figure).

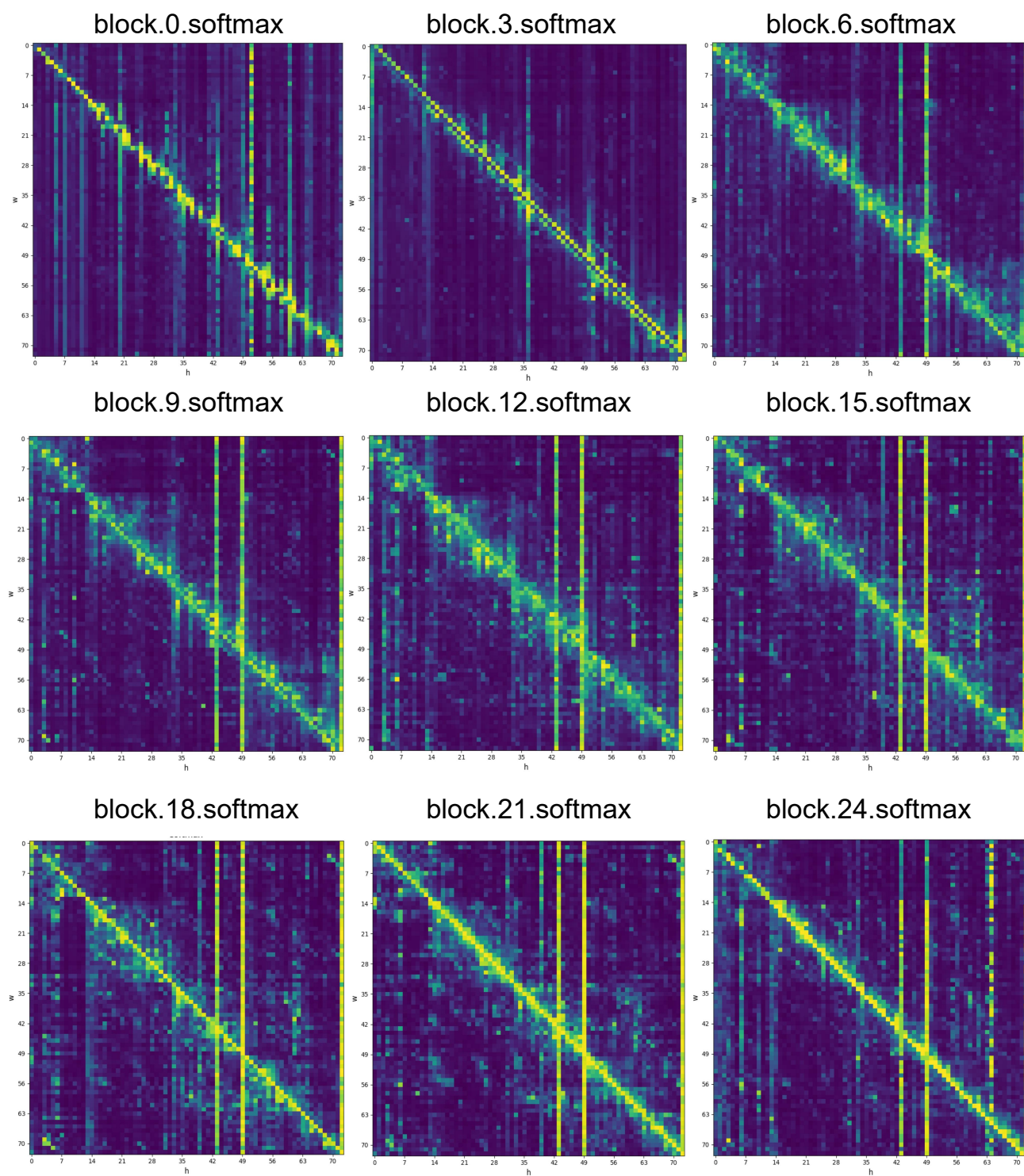
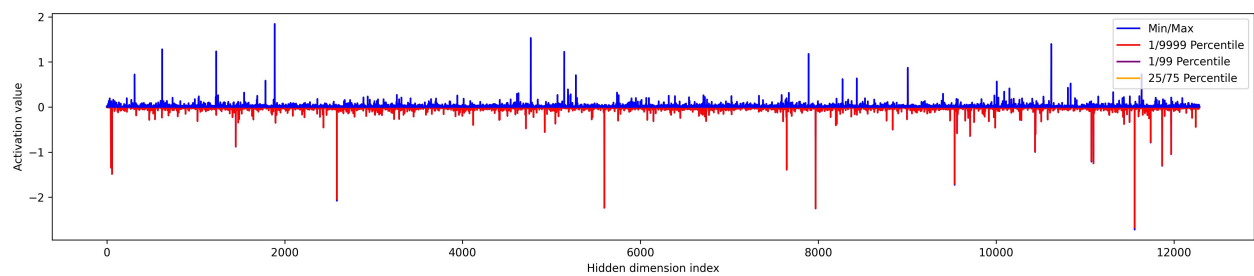
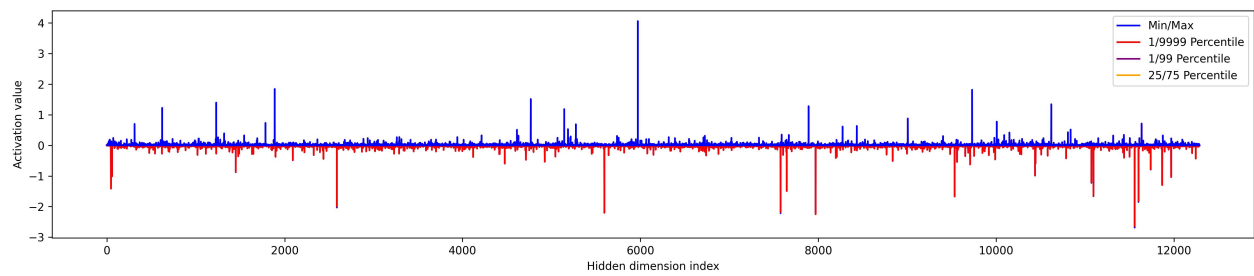


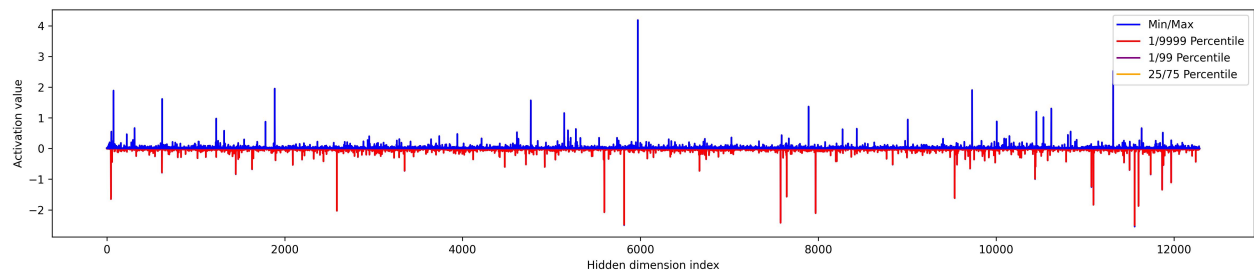
Figure 5: Output distribution of softmax in different blocks of LLADA.



(a) step0.block0.ffn



(b) step30.block0.ffn



(c) step100.block0.ffn

Figure 6: The per-channel output distribution of ffn in the attention mechanism across different iteration steps of LLADA.

Model	Method	Truth.	Arc.	Hel.	Wino.	PIQA	MMLU	C-EVAL	Hum.	GSM8K	Avg.
LLADA	AWQ	40.87	42.92	46.14	66.88	69.43	51.22	58.43	20.10	36.88	48.09
	AWQ + TMA5	41.12	43.22	46.29	67.12	70.03	51.52	59.13	21.18	38.12	48.63
	AWQ + TMA5 + CGQ	41.37	43.12	46.35	67.68	70.23	51.46	59.47	22.10	39.78	49.06
	AWQ + TMA5 + CGQ + IA-AQ	41.53	43.44	46.51	67.87	70.12	51.72	59.38	22.13	40.66	49.26
	GPTQ	42.53	44.20	49.76	69.85	70.75	55.96	56.32	25.33	44.57	51.03
	GPTQ + TMA5	43.18	43.92	50.05	70.85	71.85	56.56	58.33	27.87	48.63	52.36
	GPTQ + TMA5 + CGQ	43.33	44.17	50.76	71.25	72.34	56.96	59.47	28.04	52.18	53.16
	GPTQ + TMA5 + CGQ + IA-AQ	43.53	44.18	51.00	71.85	73.94	57.77	61.22	28.92	56.25	54.29

Table 5: Results of RTN, AWQ, GPTQ, and ours DLLMQuant with 4-bit weight and activation quantization among 9 tasks on LLADA-8B, LLADA-1.5-8B, DREAM-7B) . DLLMQuant⁺ denotes DLLMQuant based on AWQ, and DLLMQuant⁺⁺ denotes DLLMQuant based on GPTQ.