

---

# **FreeAct: Freeing Activations for LLM Quantization**

---

**Xiaohao Liu<sup>1</sup> Xiaobo Xia<sup>1</sup> Manyi Zhang<sup>2</sup> Ji-Fu Li<sup>2</sup> Xianzhi Yu<sup>2</sup>**  
**Fei Shen<sup>1</sup> Xiu Su<sup>3</sup> See-Kiong Ng<sup>1</sup> Tat-Seng Chua<sup>1</sup>**

# Background

- 大语言模型（LLM）参数量巨大（如 LLaMA-2 70B 需数百 GB 显存），推理成本极高；限制了在资源受限设备上的部署；
- 量化技术将模型权重与激活从高精度（FP16/FP32）压缩到低比特（如 INT8、INT4）以节省资源；
  - **显著降低显存占用**（FP16  $\rightarrow$  INT4 节省 4 $\times$ ）
  - **加速推理速度**（可在整数矩阵乘法上执行）
  - **降低功耗、支持端侧部署**

# Question

- 核心问题：在LLM的低比特量化中，如何有效处理动态激活分布？
- 传统假设的局限性：传统变换式量化（如QuaRot）假设激活分布是静态的，采用“一对一”的正交变换约束，缺乏灵活性。
- 真实场景的复杂性：在多模态LLM、扩散LLM等场景中，不同类型Token的激活分布差异巨大，呈现高度动态性。

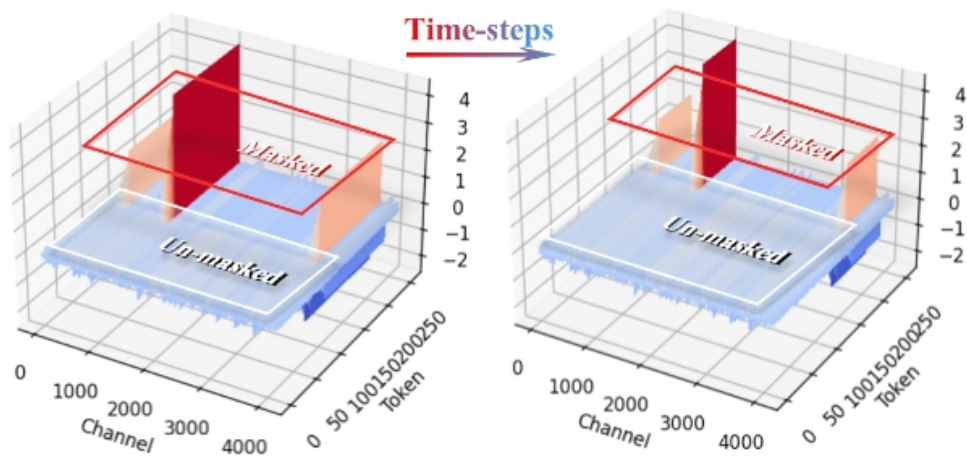


Figure 2. Illustrative example of activations between masked and unmasked tokens along the diffusion time-steps (dLLMs).

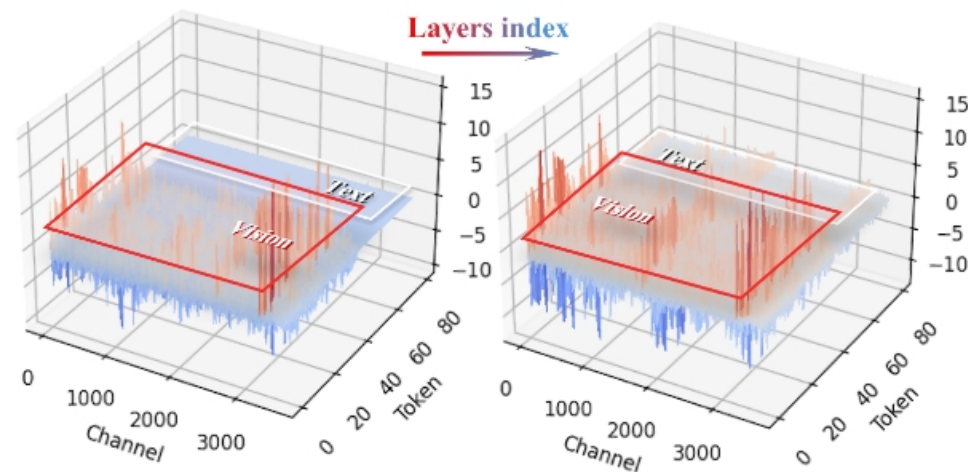


Figure 3. Illustrative example of activation distributions between vision and text tokens at different layers in MLLMs.

# Existing situation

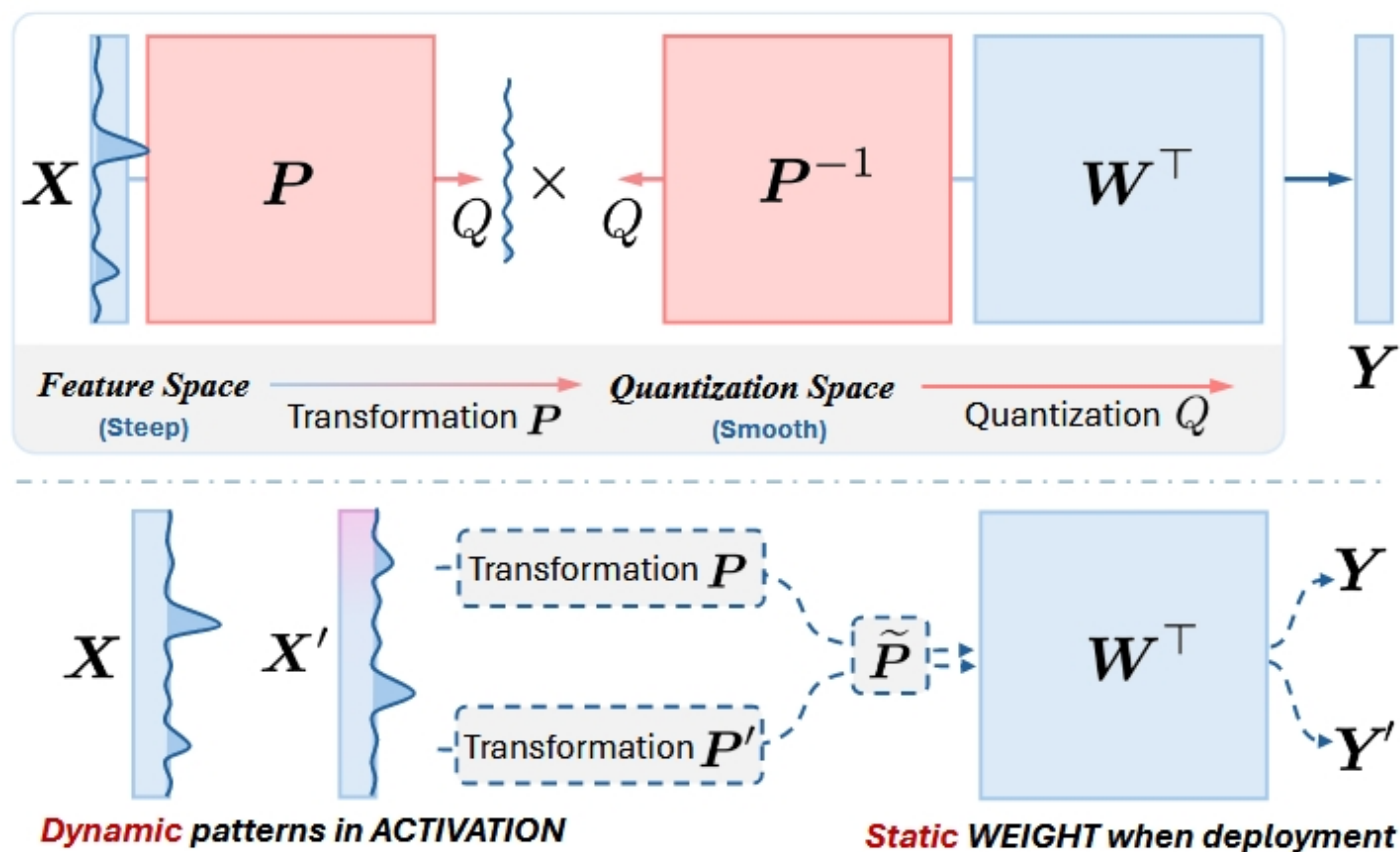
- 主流方法：变换式量化 (Transformation-based Quantization)，如 QuaRot, FlatQuant。
- 核心思想：通过正交矩阵将激活/权重投影到更平滑的空间，降低量化误差。

- 局限一：静态约束

现有方法受限于严格的“一对一”变换约束 ( $P \times P^{-1} = I$ )，这种强数学约束导致模型无法灵活适应动态变化的激活分布，限制了泛化能力。

- 局限二：性能瓶颈

在处理动态激活时，量化精度会出现显著下降，极端情况下甚至会导致模型完全崩溃。



*Figure 1.* The steep activation is smoothed by the designed transformation matrix for quantization (*top*). One weight corresponds to one transformation to ensure equivalence, *i.e.*,  $P \times P^{-1} = I$ . In practice, activations are diverse, reflecting dynamic patterns, necessitating flexible transformation matrices on the activation side, while keeping static on the weight side for quantization (*bottom*).

# Solution: FreeAct

- 核心突破：解绑变换约束

打破激活与权重之间严格的“一对一”变换限制，引入更灵活的解空间。

- 具体方案：动静分离策略

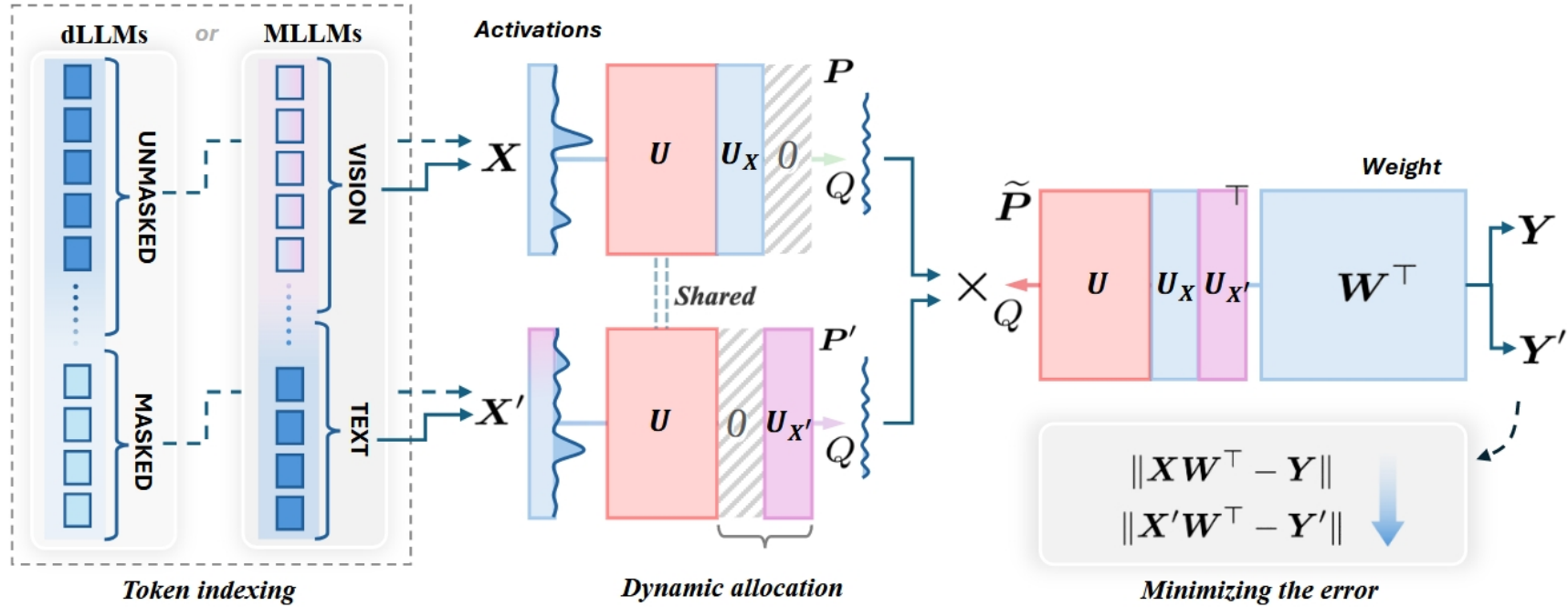
激活侧：动态多变换矩阵，适配不同Token类型。

权重侧：静态统一变换矩阵，确保部署高效无开销。

- 理论基础：激活秩亏缺特性

利用激活的秩亏缺证明了多变换与单变换在计算上的等价性，扩大了解空间。

## FreeAct: Freeing Activations for LLM Quantization



**Figure 4. The overall framework of FreeAct.** Different tokens are indexed to different transformation matrices according to their unique types (vision-text for MLLMs, and unmasked and masked for dLLMs). Different transformation matrices maintain a shared part together while possessing their own uniqueness, where zeros are filled in on another portion. The weight transformation matrix unites both portions to handle all the different activations. And the quantization error is minimized to optimize the quantization parameter.

$$\min_{P, P', P_e, \text{params}} \|XW^T - Q_A(XP)Q_W(P_eW^T)\|_2^2$$

# Challenges

- 动态变换矩阵设计

如何为不同类型的Token分配合适的变换矩阵，同时在保证效果的前提下维持计算效率？

$$\begin{aligned} (\mathbf{X}, \mathbf{X}') & \begin{cases} \mathbf{P} = [\mathbf{U}, \mathbf{U}_{\mathbf{X}}, \mathbf{0}], & \mathbf{U} \in \mathbb{R}^{d \times r}, \mathbf{U}_{\mathbf{X}} \in \mathbb{R}^{d \times r_1}, \\ \mathbf{P}' = [\mathbf{U}, \mathbf{0}, \mathbf{U}_{\mathbf{X}'}], & \mathbf{U}_{\mathbf{X}'} \in \mathbb{R}^{d \times r_2} \end{cases} \\ (\mathbf{W}) \quad \tilde{\mathbf{P}} &= [\mathbf{U}, \mathbf{U}_{\mathbf{X}}, \mathbf{U}_{\mathbf{X}'}]^\top, \quad r + r_1 + r_2 = d \quad (8) \end{aligned}$$

- 计算等价性保证

在激活使用多变换、权重使用单变换的混合场景下，如何严格确保数学上的计算等价性？

$$\text{秩亏性} \longrightarrow \begin{cases} \mathbf{X}\mathbf{W}^\top = \mathbf{X}\mathbf{P} \times \tilde{\mathbf{P}}\mathbf{W}^\top, \\ \mathbf{X}'\mathbf{W}^\top = \mathbf{X}'\mathbf{P}' \times \tilde{\mathbf{P}}\mathbf{W}^\top, \end{cases} \quad \text{s.t. } \mathbf{P} \neq \mathbf{P}' \quad (6)$$

# Experiment

- 实验设置

模型选择：dLLM (LLaDA, Dream)、MLLM (Qwen2.5-VL, InternVL2.5)

量化配置：核心 W4A4，对比 W8A8 与 FP16 基线

对比方法：RTN, SmoothQuant, QuaRot, FlatQuant

- 评估指标

dLLM 生成质量：FID (Fréchet Inception Distance)、CLIP Score

MLLM 理解能力：MM-Vet, MME, SEED-Bench 多模态基准测试

- 核心结果

FreeAct显著超越现有方法，在W4A4极下性能接近 FP16 全精度模型。最高相对提升达到：  
+5.3%

*Table 1. Overall performance comparison of FreeAct against quantization methods across four models. The best result in each case is marked in bold, and the averaged value is under the column “Avg.”.*

|                       | dLLMs        |              |              |              | MLLMs              |              |              |              |
|-----------------------|--------------|--------------|--------------|--------------|--------------------|--------------|--------------|--------------|
|                       | HumanEval    | GSM8K        | Math500      | Avg.         | MMMU               | MMBench      | RealworldQA  | Avg.         |
|                       | <i>LLaDA</i> |              |              |              | <i>Qwen2.5-VL</i>  |              |              |              |
| 16-bit Baseline       | 43.90        | 76.80        | 38.40        | 53.03        | 50.44              | 83.25        | 68.50        | 67.40        |
| RTN (W8A8)            | 39.02        | 77.56        | 37.00        | 51.19        | 49.33              | 82.38        | 68.50        | 66.74        |
| RTN (W4A4)            | 00.00        | 00.00        | 00.14        | 00.05        | 26.33              | 00.09        | 00.65        | 09.02        |
| SmoothQuant           | 00.00        | 00.00        | 00.00        | 00.00        | 24.41              | 00.43        | 01.17        | 08.67        |
| QuaRot                | 28.66        | 65.81        | 24.20        | 39.56        | 43.56              | 72.51        | 56.34        | 57.47        |
| FlatQuant             | 39.63        | 74.37        | 33.60        | 49.20        | 48.89              | 81.61        | <b>66.27</b> | 65.59        |
| <b>FreeAct (ours)</b> | <b>42.68</b> | <b>75.74</b> | <b>34.60</b> | <b>51.01</b> | <b>50.44</b>       | <b>82.22</b> | 66.01        | <b>66.22</b> |
|                       | <i>Dream</i> |              |              |              | <i>InternVL2.5</i> |              |              |              |
|                       | HumanEval    | GSM8K        | Math500      | Avg.         | MMMU               | MMBench      | RealworldQA  | Avg.         |
|                       | <i>LLaDA</i> |              |              |              | <i>Qwen2.5-VL</i>  |              |              |              |
| 16-bit Baseline       | 56.10        | 68.16        | 42.20        | 55.49        | 53.22              | 84.71        | 68.76        | 68.90        |
| RTN (W8A8)            | 48.78        | 68.76        | 40.80        | 52.78        | 44.33              | 74.74        | 63.66        | 60.91        |
| RTN (W4A4)            | 00.00        | 00.53        | 0.018        | 00.18        | 22.78              | 00.20        | 00.78        | 07.92        |
| SmoothQuant           | 00.00        | 00.00        | 0.014        | 00.05        | 23.44              | 00.94        | 01.05        | 08.47        |
| QuaRot                | 39.02        | 56.56        | 25.80        | 40.46        | 42.19              | 63.66        | 54.25        | 53.36        |
| FlatQuant             | 46.34        | 63.53        | <b>34.60</b> | 48.16        | 42.11              | 69.84        | 57.65        | 56.53        |
| <b>FreeAct (ours)</b> | <b>50.00</b> | <b>68.84</b> | 32.60        | <b>50.48</b> | <b>42.44</b>       | <b>70.62</b> | <b>58.69</b> | <b>57.25</b> |

# Contributions

- 方法论创新：FreeAct框架

首次放松LLM量化中激活与权重的“一对一”约束，支持动态激活量化。

- 高效实现：零额外开销设计

通过“共享+专属+零填充”子空间设计，直接从权重变换切片得到激活变换。

$$\tilde{P} = [U, U_X, U_{X'}]^\top$$

- 理论与实践验证

完备的理论证明结合大量实验，在多个主流模型上均取得了SOTA性能。