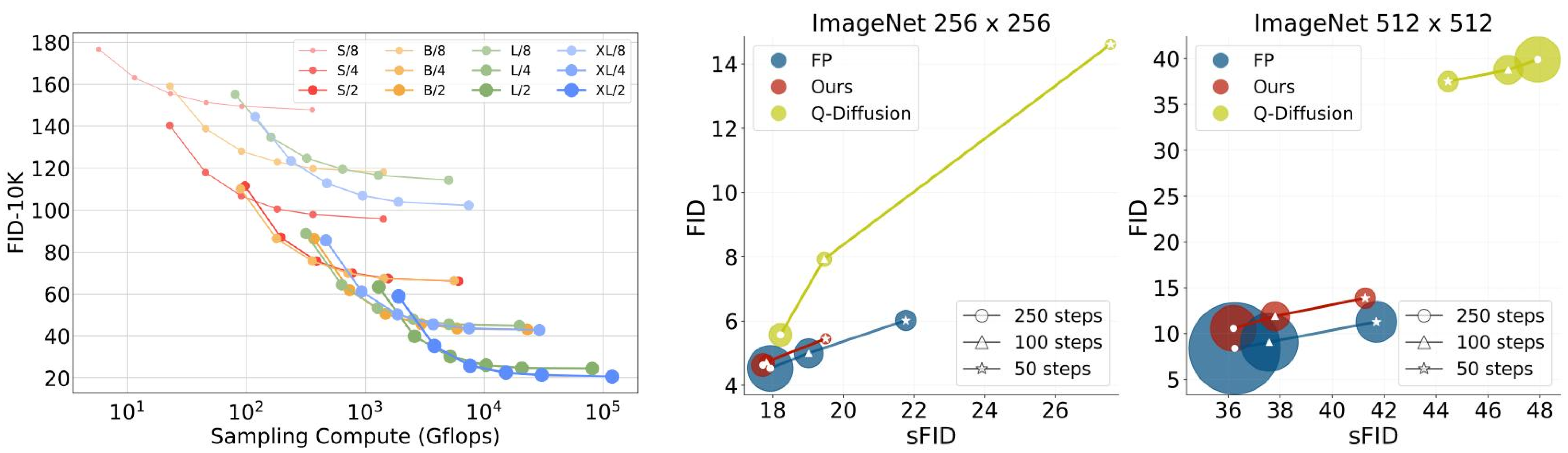
PTQ4DiT: Post-training Quantization for Diffusion Transformers

Junyi Wu*, Haoxuan Wang*, Yuzhang Shang, Mubarak Shah, Yan Yan^

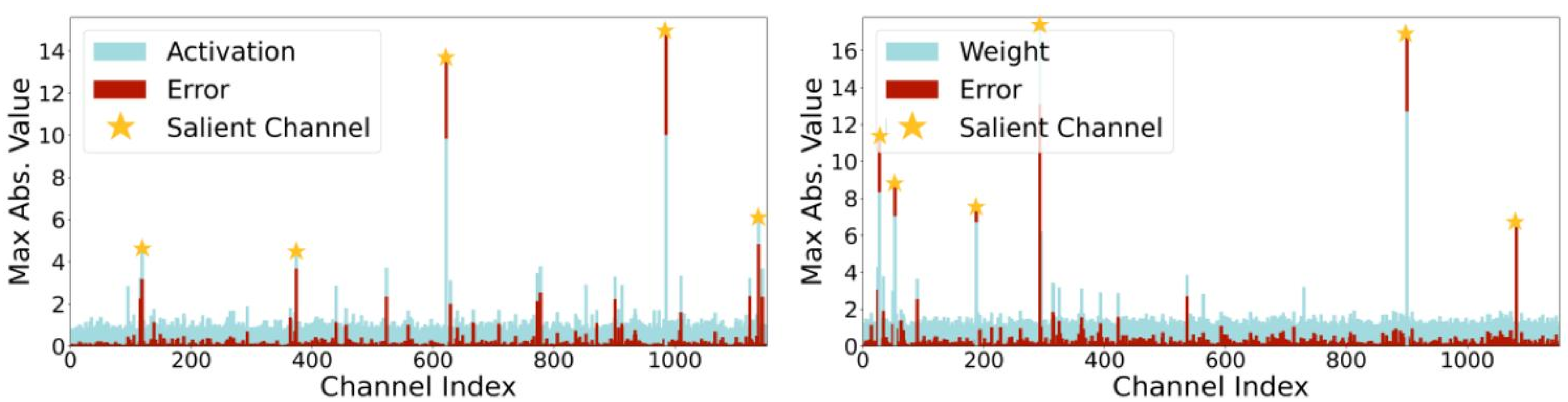
Diffusion Transformers (DiTs) inherit the scaling property but incur increasing computational cost



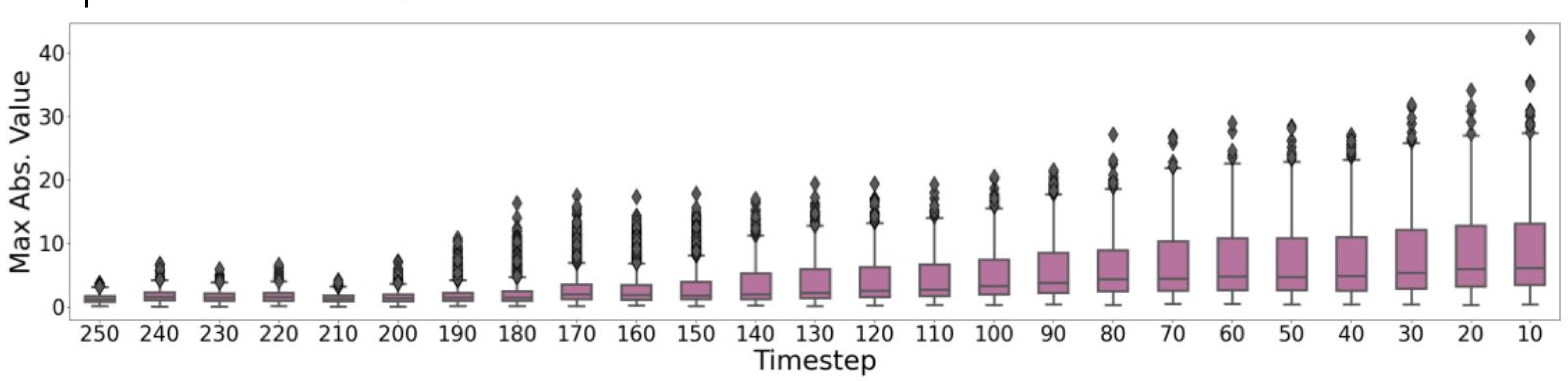
PTQ4DiT is the first effective DiT quantization method, offering a practical deployment solution.

Quantization Challenges of DiTs

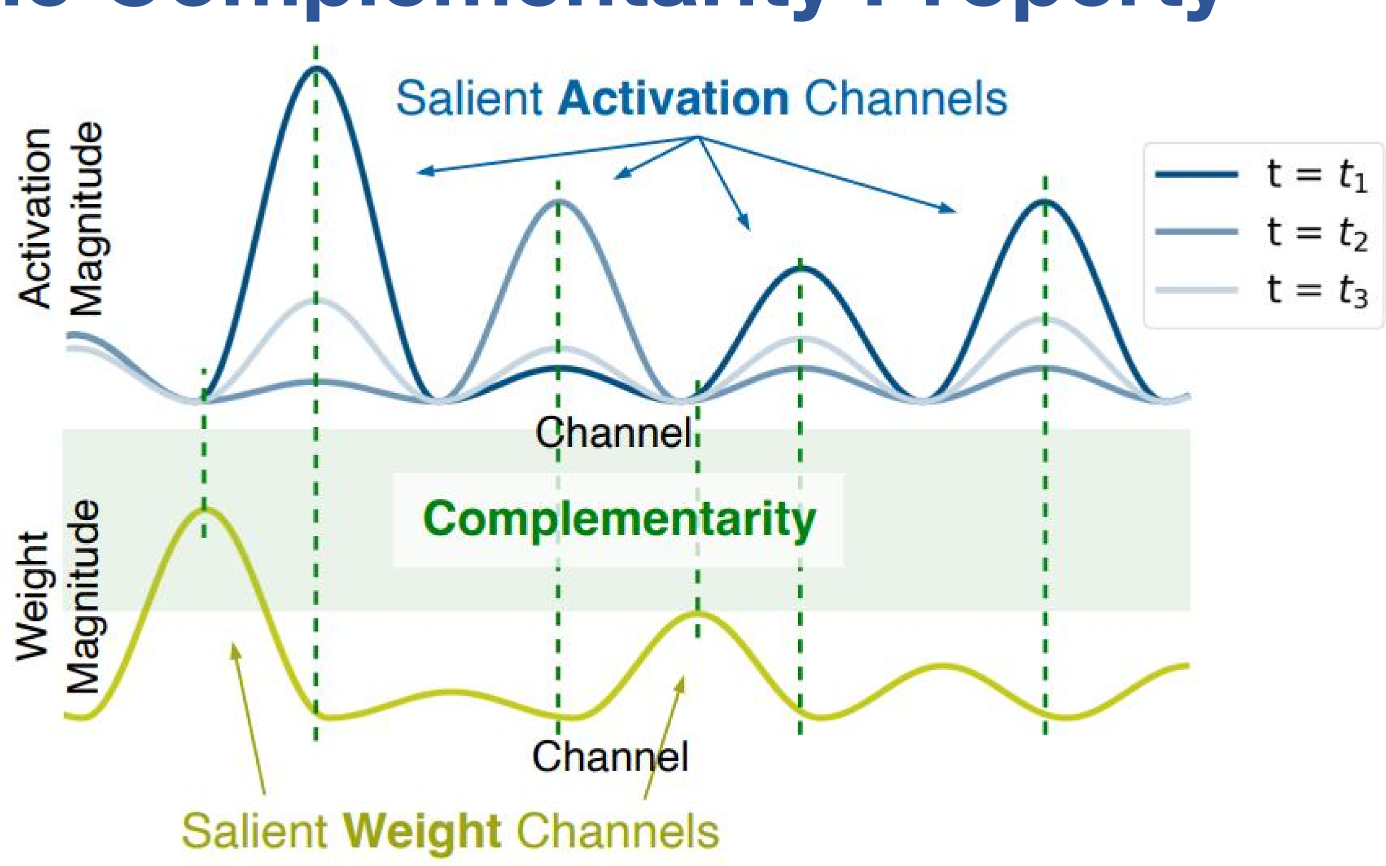
- Pronounced Quantization Error in Salient Channels



- Temporal Variation in Salient Activation



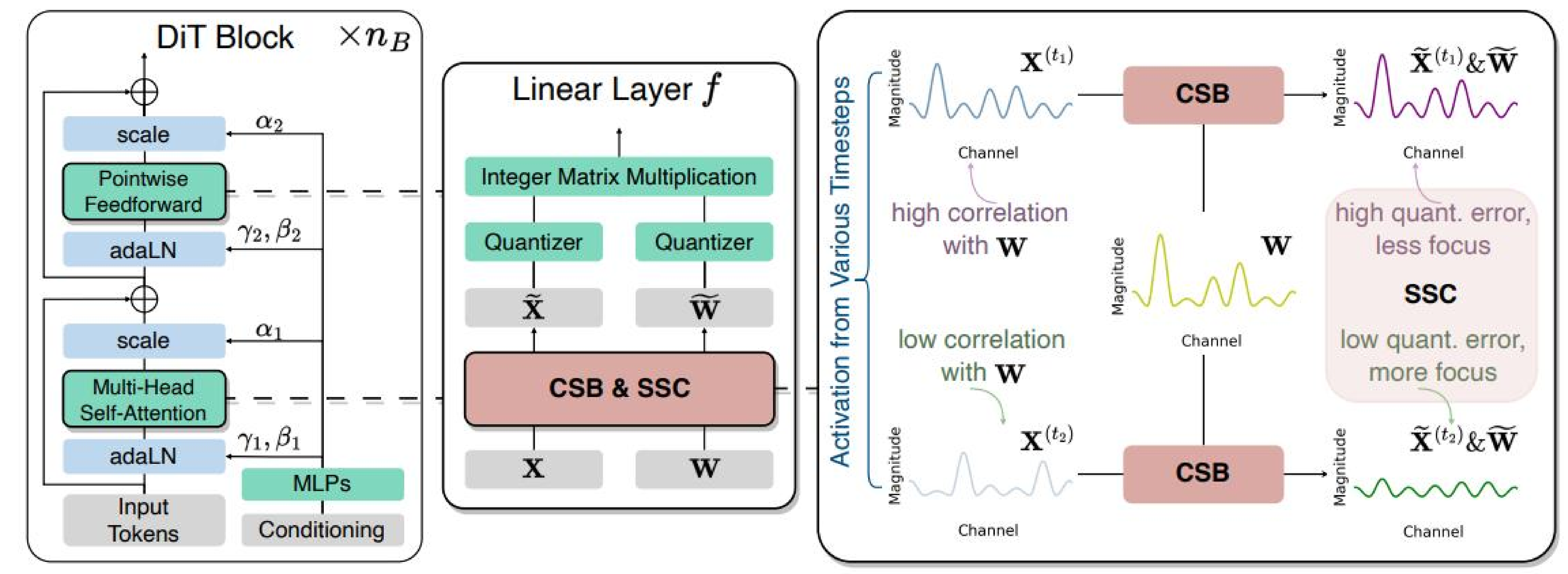
The Complementarity Property



- Observation: Activation and weight channels do not have extreme magnitude simultaneously
- Idea: Redistribute channel salience between weights and activations across various timesteps

The Proposed PTQ4DiT

- Channel-wise Salience Balancing (CSB) and Spearmen's ρ-guided Salience Calibration (SSC)



- Balancing Transformation

$$\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{B}^{\mathbf{X}}, \quad \widetilde{\mathbf{W}} = \mathbf{B}^{\mathbf{W}}\mathbf{W}$$

$$\mathbf{B}^{\mathbf{X}} = \text{diag}\left(\frac{\widetilde{s}(\mathbf{X}_1, \mathbf{W}_1)}{s(\mathbf{X}_1)}, \frac{\widetilde{s}(\mathbf{X}_2, \mathbf{W}_2)}{s(\mathbf{X}_2)}, \ldots, \frac{\widetilde{s}(\mathbf{X}_{d_{in}}, \mathbf{W}_{d_{in}})}{s(\mathbf{X}_{d_{in}})}\right),$$

$$\mathbf{B}^{\mathbf{W}} = \text{diag}\left(\frac{\widetilde{s}(\mathbf{X}_1, \mathbf{W}_1)}{s(\mathbf{W}_1)}, \frac{\widetilde{s}(\mathbf{X}_2, \mathbf{W}_2)}{s(\mathbf{W}_2)}, \ldots, \frac{\widetilde{s}(\mathbf{X}_{d_{in}}, \mathbf{W}_{d_{in}})}{s(\mathbf{W}_{d_{in}})}\right).$$

- Mathematically Equivalence

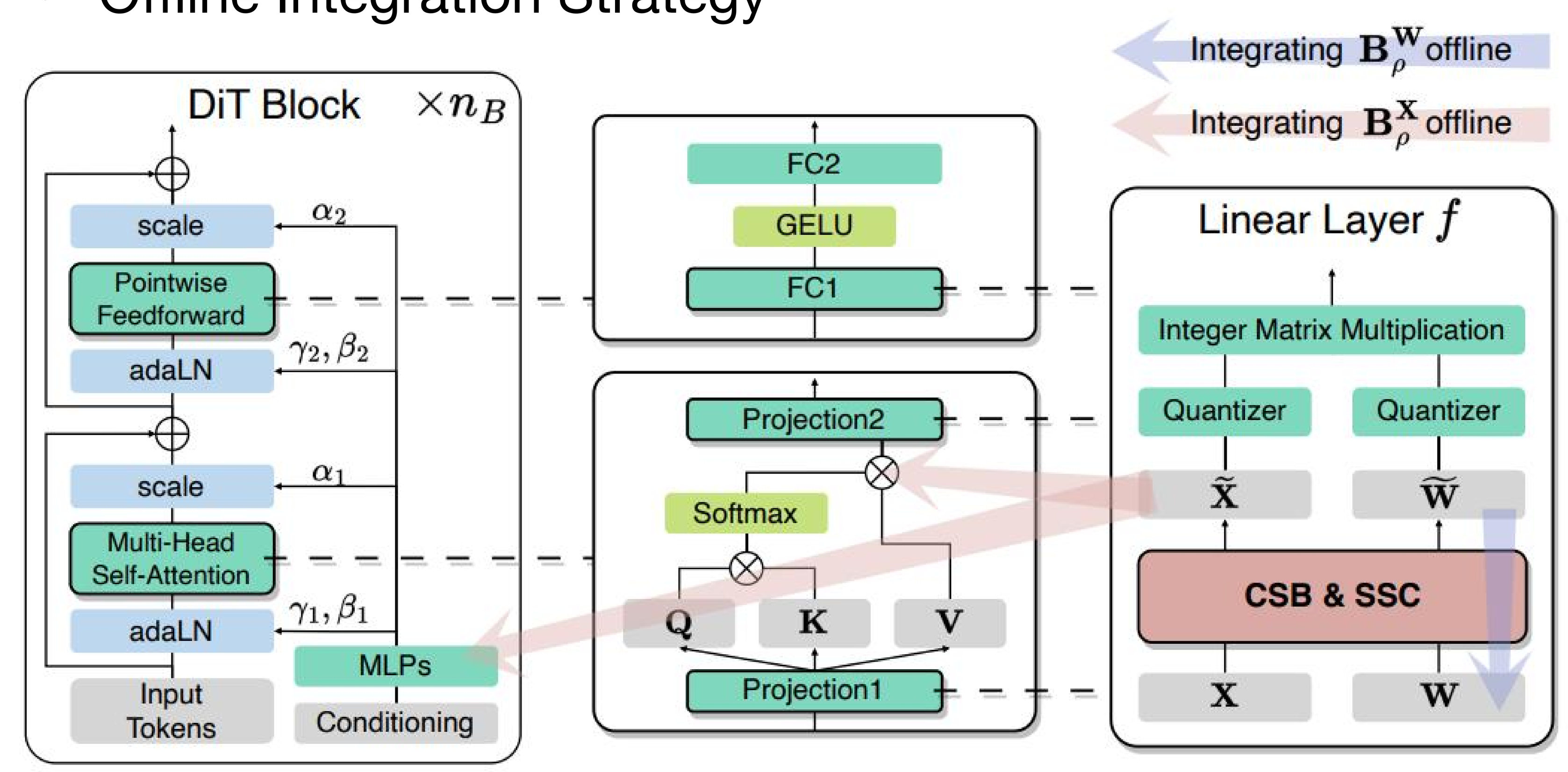
$$\widetilde{\mathbf{X}} \cdot \widetilde{\mathbf{W}} = (\mathbf{X}\mathbf{B}_{\rho}^{\mathbf{X}}) \cdot (\mathbf{B}_{\rho}^{\mathbf{W}}\mathbf{W}) = \mathbf{X} \cdot \mathbf{W}$$

- Timestep-aware Calibration

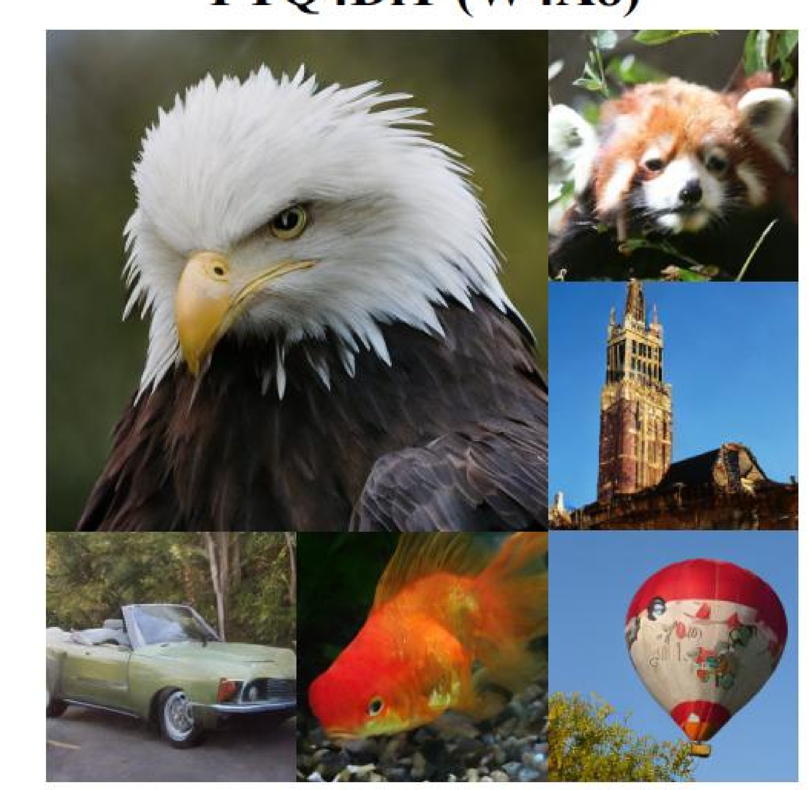
$$\eta_t = \frac{\exp[-\rho(\mathbf{s}(\mathbf{X}^{(t)}), \mathbf{s}(\mathbf{W}))]}{\sum_{\tau=1}^{T} \exp[-\rho(\mathbf{s}(\mathbf{X}^{(\tau)}), \mathbf{s}(\mathbf{W}))]}$$

$$\mathbf{s}_{\rho}(\mathbf{X}^{(1:T)}) = (\eta_1, \eta_2, \ldots, \eta_T) \cdot (\mathbf{s}(\mathbf{X}^{(1)}), \mathbf{s}(\mathbf{X}^{(2)}), \ldots, \mathbf{s}(\mathbf{X}^{(T)}))^{\mathrm{T}} \in \mathbb{R}^{d_{in}}$$

- Offline Integration Strategy

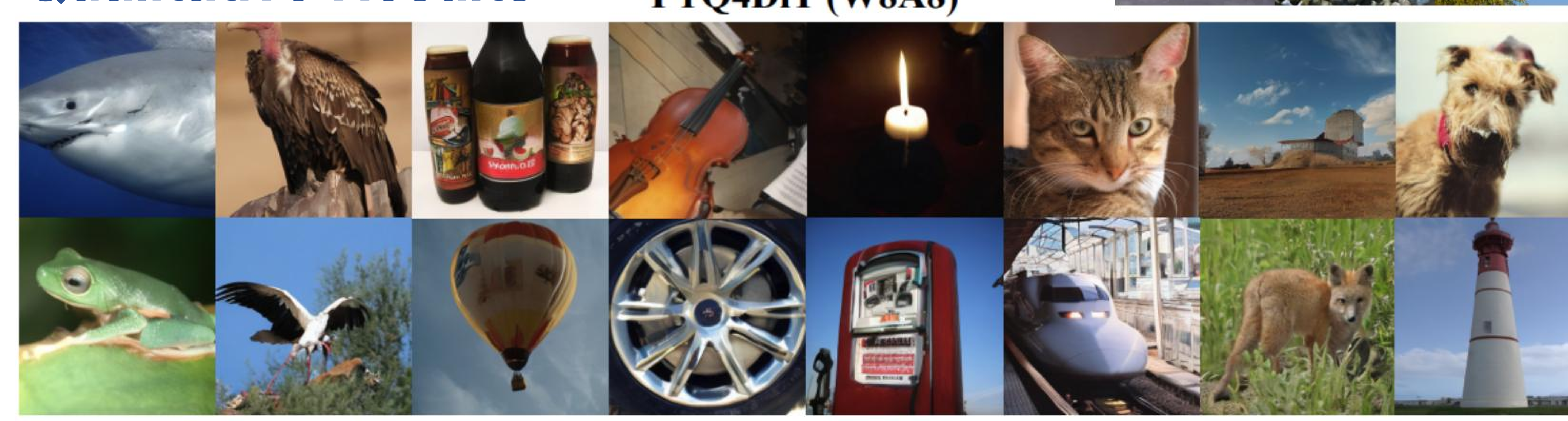


PTQ4DiT (W4A8)



Qualitative Results

PTQ4DiT (W8A8)



Full-Precision