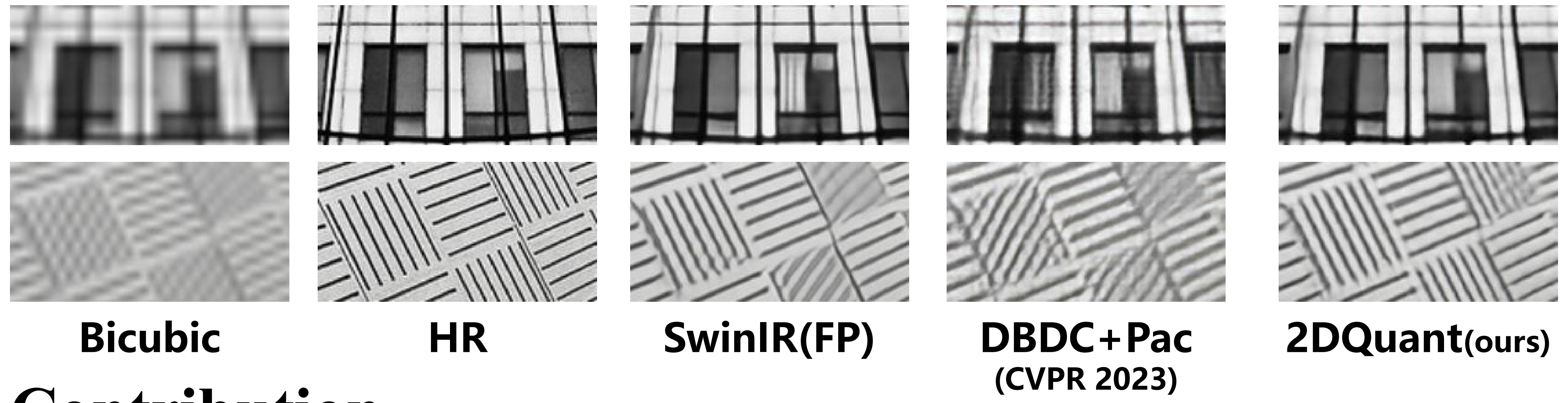
# 2DQuant: Low-bit Post-Training Quantization for Image Super-Resolution

Kai Liu[1], Haotong Qin[2], Yong Guo[3], Xin Yuan[4], Linghe Kong[1], Guihai Chen[1]*, Yulun Zhang[1]*

[1]Shanghai Jiao Tong University, [2]ETH Zürich, [3]Max Planck Institute for Informatics, [4]Westlake University

## Introduction

**Vision Transformers (ViTs)** excel in SR tasks but face high costs. Low bit post-training quantization (**PTQ**) reduces memory and computation. However, the deterioration of self-attention in quantized transformers limit its application. To tackle this, we propose **2DQuant** a novel PTQ for ViT in SR.



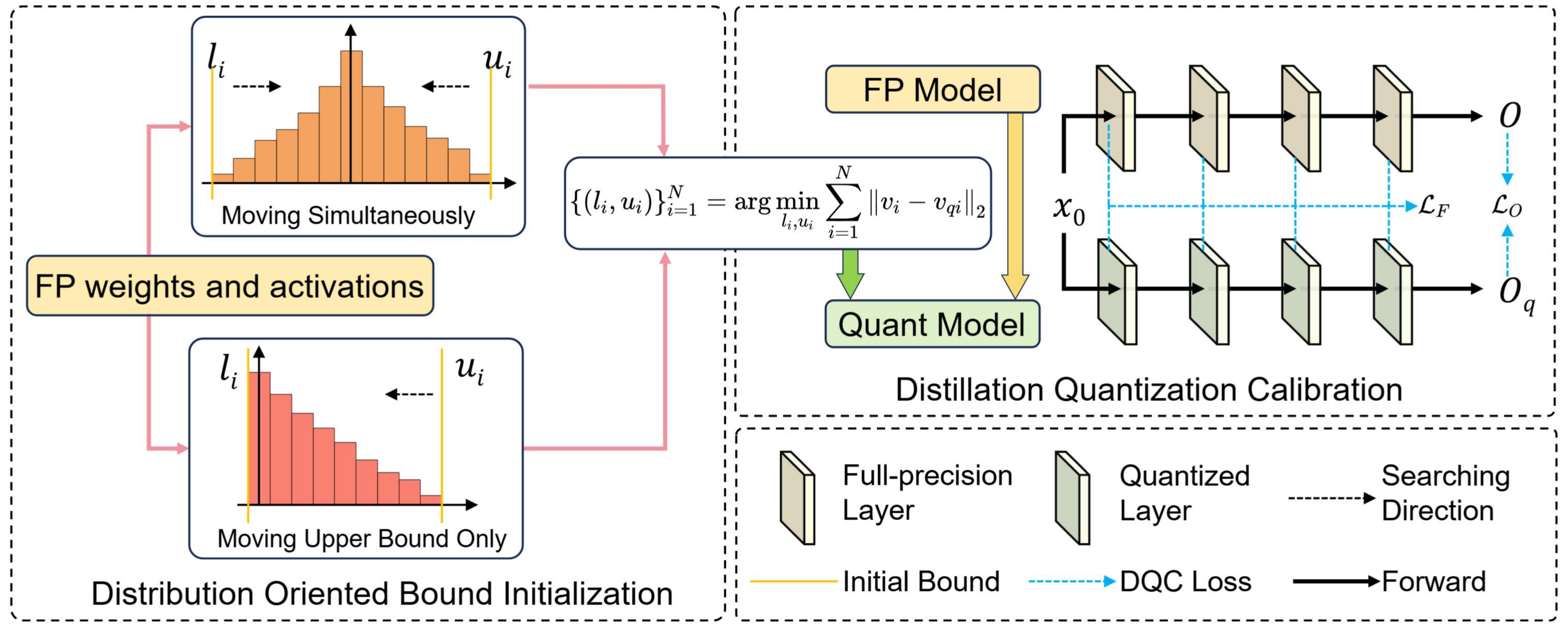
## Contribution

- **Exploration:** We are the first to explore PTQ with ViT models in SR thoroughly.
- **Pipeline:** Design two-stage PTQ method for SR, **DOBI** for fast rough bound search while **DQC** for fine-grained sophisticated bound search.
- **Performance:** Compress the model to 4, 3, and 2 bits with speedup ratio being 3.99x, 4.47x, and 5.08x respectively. Surpass existing SOTA on all benchmark and visual effects.

Project

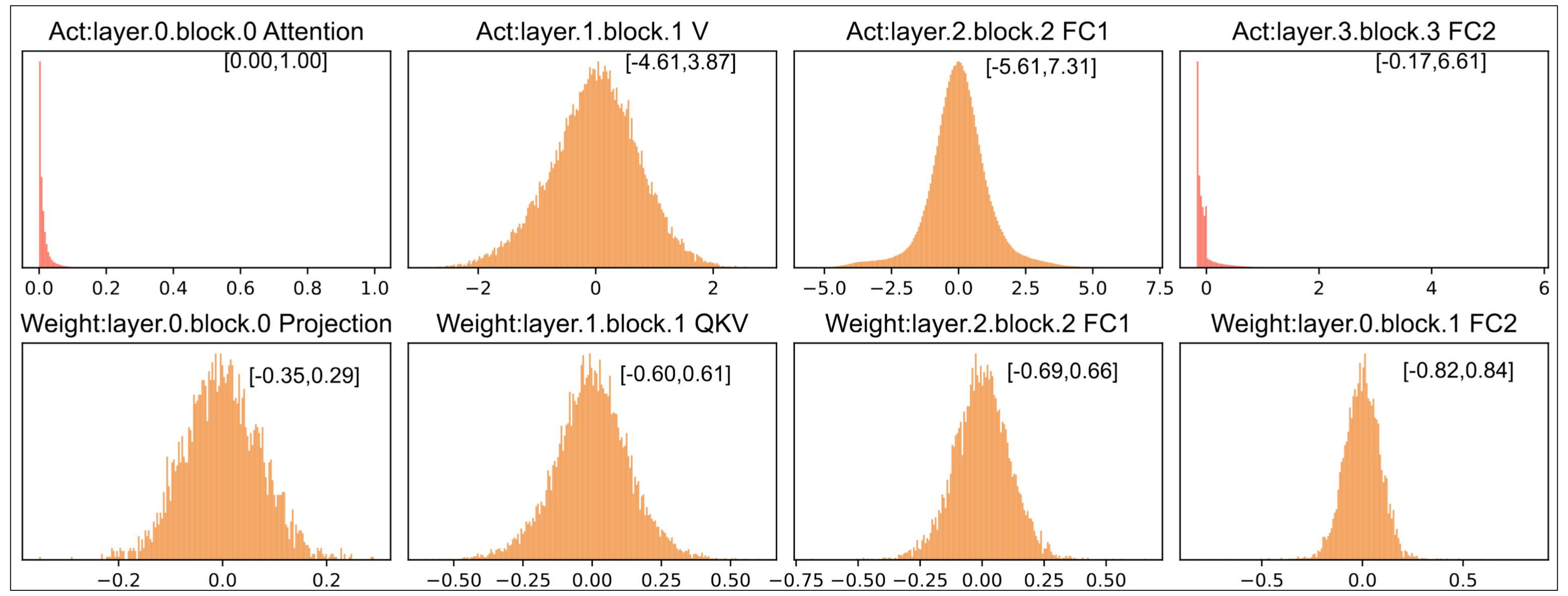
## Method

### ❖ Overall



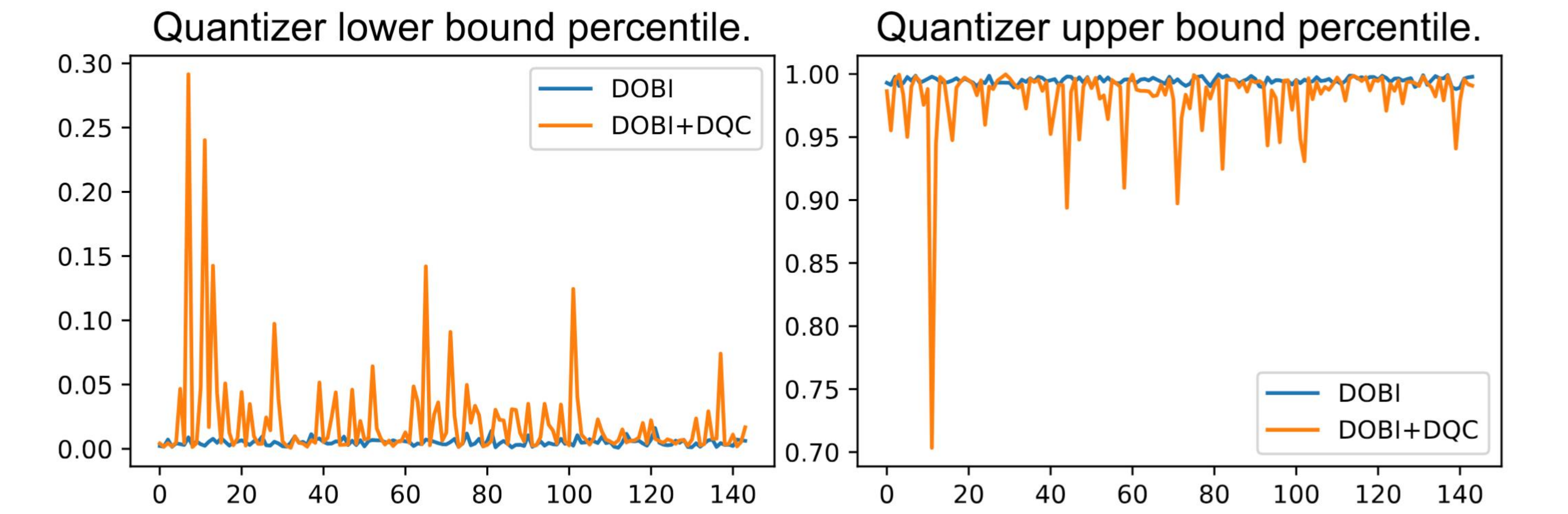
### ❖ Algorithm

- **Challenge I: Long Tail distribution.** The distribution of weight and activation of ViT present long tail distribution. The hugh values are crutial but hinders quantized model's performance.
- **Challenge II: Mismatch between Quant loss and Task loss.** Optimizing quant loss from the local perspective is not always align with the task loss, making



- **DOBI:** We note that the data distribution falls into two categories: one resembling a bell-shaped distribution and the other resembling an exponential distribution. Different searching directions are used for different shapes to guarantee fast and accurate search.
- **DQC:** Distillation quantization calibration between the FP model and the quantized model further adjust the boud and improve the model's performance.

### ❖ Clipping Bound Distribution



- The local search result of DOBI is still around the min value and the max value.
- After DQA, the bound presents more extreme distribution. The most extreme one leaves only 46% data in clipping range and the values beyond are all clipped. This shows that mearly local search can not guarantee low task loss, namely high performance.

## Experiments

### ❖ Ablation Study

| Learning rate | PSNR↑ | SSIM↑ |
|---|---|---|
| $10^{-1}$ | 37.82 | 0.9594 |
| $10^{-2}$ | 37.87 | 0.9594 |
| $10^{-3}$ | 37.78 | 0.9592 |
| $10^{-4}$ | 37.74 | 0.9587 |

(a) Learning rate

| Batch size | PSNR↑ | SSIM↑ |
|---|---|---|
| 4 | 37.82 | 0.9594 |
| 8 | 37.83 | 0.9594 |
| 16 | 37.84 | 0.9593 |
| 32 | 37.87 | 0.9594 |

(b) Batch size

| DOBI | DQC | PSNR↑ | SSIM↑ |
|---|---|---|---|
|  |  | 34.39 | 0.9202 |
| ✓ |  | 37.44 | 0.9568 |
|  | ✓ | 37.32 | 0.9563 |
| ✓ | ✓ | 37.87 | 0.9594 |

(c) DOBI and DQC

### ❖ Quantitative Results

| Method | Bit | Set5 (×4) |  | Set14 (×4) |  | B100 (×4) |  | Urban100 (×4) |  | Manga109 (×4) |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
|  |  | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| SwinIR-light [32] | 32 | 32.45 | 0.8976 | 28.77 | 0.7858 | 27.69 | 0.7406 | 26.48 | 0.7980 | 30.92 | 0.9150 |
| Bicubic | 32 | 27.56 | 0.7896 | 25.51 | 0.6820 | 25.54 | 0.6466 | 22.68 | 0.6352 | 24.19 | 0.7670 |
| MinMax [22] | 4 | 28.63 | 0.7891 | 25.73 | 0.6657 | 25.10 | 0.6061 | 23.07 | 0.6216 | 26.97 | 0.8104 |
| Percentile [27] | 4 | 30.64 | 0.8679 | 27.61 | 0.7563 | 26.96 | 0.7151 | 24.96 | 0.7479 | 28.78 | 0.8803 |
| EDSR† [33, 41] | 4 | 31.20 | 0.8670 | 27.98 | 0.7600 | 27.09 | 0.7140 | 25.56 | 0.7640 | N/A | N/A |
| DBDC+Pac [41] | 4 | 30.74 | 0.8609 | 27.66 | 0.7526 | 26.97 | 0.7104 | 24.94 | 0.7369 | 28.52 | 0.8697 |
| DOBI (Ours) | 4 | 31.10 | 0.8770 | 28.03 | 0.7672 | 27.18 | 0.7237 | 25.43 | 0.7631 | 29.31 | 0.8916 |
| 2DQuant (Ours) | 4 | 31.77 | 0.8867 | 28.30 | 0.7733 | 27.37 | 0.7278 | 25.71 | 0.7712 | 29.71 | 0.8972 |
| MinMax [22] | 3 | 19.41 | 0.3385 | 18.35 | 0.2549 | 18.79 | 0.2434 | 17.88 | 0.2825 | 19.13 | 0.3097 |
| Percentile [27] | 3 | 27.55 | 0.7270 | 25.15 | 0.6043 | 24.45 | 0.5333 | 22.80 | 0.5833 | 26.15 | 0.7569 |
| DBDC+Pac [41] | 3 | 27.91 | 0.7250 | 25.86 | 0.6451 | 25.65 | 0.6239 | 23.45 | 0.6249 | 26.03 | 0.7321 |
| DOBI (Ours) | 3 | 29.59 | 0.8237 | 26.87 | 0.7156 | 26.24 | 0.6735 | 24.17 | 0.6880 | 27.62 | 0.8349 |
| 2DQuant (Ours) | 3 | 30.90 | 0.8704 | 27.75 | 0.7571 | 26.99 | 0.7126 | 24.85 | 0.7355 | 28.21 | 0.8683 |
| MinMax [22] | 2 | 23.96 | 0.4950 | 22.92 | 0.4407 | 22.70 | 0.3943 | 21.16 | 0.4053 | 22.94 | 0.5178 |
| Percentile [27] | 2 | 23.03 | 0.4772 | 22.12 | 0.4059 | 21.83 | 0.3816 | 20.45 | 0.3951 | 20.88 | 0.3948 |
| DBDC+Pac [41] | 2 | 25.01 | 0.5554 | 23.82 | 0.4995 | 23.64 | 0.4544 | 21.84 | 0.4631 | 23.63 | 0.5854 |
| DOBI (Ours) | 2 | 28.82 | 0.7699 | 26.46 | 0.6804 | 25.97 | 0.6319 | 23.67 | 0.6407 | 26.32 | 0.7718 |
| 2DQuant (Ours) | 2 | 29.53 | 0.8372 | 26.86 | 0.7322 | 26.46 | 0.6927 | 23.84 | 0.6912 | 26.07 | 0.8163 |

### ❖ Visual Results