COMPSRT: QUANTIZATION AND PRUNING FOR IMAGE SUPER RESOLUTION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Model compression has emerged as a way to reduce the cost of using image super resolution models by decreasing storage size and inference time. However, the gap between the best compressed models and the full precision model still remains large and a deeper understanding of compression theory on more performant models remains unexplored. Prior research on quantization of Large Language Models has shown that Hadamard transforms lead to 'flattened' weight and activation distributions which lower quantization errors. However, we observe that on SwinIR-light, Hadamard transformations on weights and activations do not lead to flatter distributions, but do lead to lower quantization errors. Instead of flattening distributions, we show that lower errors is caused by the Hadamard transforms ability to reduce the ranges, and increase the proportion of values around 0. Based on these findings, we introduce CompSRT, a more performant way to compress the image super resolution transformer network SwinIR-light. We perform Hadamard-based quantization, and we also perform scalar decomposition to introduce two additional trainable parameters. Our quantization performance statistically significantly surpasses the current state-of-the-art in metrics with gains as large as 1.53 dB, and visibly improves visual quality by reducing blurriness at all bitwidths. At 3-4 bits, to show our method is compatible with pruning for increased compression, we also prune 40% of weights and show that we can achieve 6.67-15% reduction in bits per parameter with comparable performance to the state-of-the-art.

1 INTRODUCTION

Image super-resolution (SR), the task of reconstructing high-resolution (HR) images from low-resolution (LR) inputs, plays a critical role in diverse domains such as imagevideo enhancement Hitachi; Takeda et al. (2009); Su et al. (2011), medical imaging Yu et al. (2017); Greenspan et al. (2002); Yu et al. (2018); Robinson et al. (2017), remote sensing Zhu et al. (2018); Murthy et al. (2014), and astronomy Yue et al. (2016). Deep learning has driven significant progress in SR, with convolutional neural networks (CNNs) like EDSR Lim et al. (2017a), RDN Zhang et al. (2018), and SRResNet Ledig et al. (2017) achieving high performance at the cost of substantial parameter counts, resulting in increased storage and inference time. Recently, transformer-based models have emerged as efficient alternatives, offering competitive performance with fewer parameters.

Despite being lighter than traditional CNNs, SwinIR-light still demands significant resources. Model compression addresses this issue through various methods like quantization and pruning. Quantization works by reducing parameter precision (e.g., from 32/16-bit to 2–4-bit), trading minor accuracy loss for substantial gains in memory and compute efficiency. Quantization strategies include quantization-aware training (QAT), which jointly optimizes weights and quantization parameters, and post-training quantization (PTQ), which calibrates quantization on a frozen model. Pruning works through removing nodes that don't offer that much information, eliminating signal noise and enhancing clarity. Pruning methods can be structured, i.e. removing whole blocks/channels or unstructured, removing individual elements. We focus on a Hadamard guided combination quantization with pruning prior to quantization at higher bit-widths.

While most prior work focuses on either quantization or pruning with CNNs, recent PTQ methods like 2DQuant Liu et al. (2024) and CondiQuant Liu et al. (2025) have adapted PTQ to SwinIR. However, the current state-of-the-art (SOTA) in PTQ, CondiQuant, does not provide deeper understanding of

Figure 1: Qualitative visual comparison for 2-bit $(\times 4)$ SR on a challenging example. LR denotes low resolution image. FP denotes the output of the FP model. The comparative example is taken from 2DQuant Liu et al. (2024)[1]. SOTA (2DQuant) suffers from excessive blurriness, while our method is significantly more clear.

theory nor directly modify distributions of weights and activations, which are known to be critical to quantization performance. As a result, it exhibits a notable gap from the full-precision (FP) SwinIR-light model.

Previous literature on quantization of LLMs Sun et al. (2024) introduced how "flatness" of weights and activations is important for quantization and Tseng et al. (2024) and Sun et al. (2024) stated that Hadamard transformations can increase flatness of distributions. However, we find that after applying the Hadamard transform to the weights and activations of SwinIR-light, the weights tend to be more normally distributed and bell curve shaped. Therefore, the mechanism for the Hadamard's lowering of the errors remains unexplored. We show that the Hadamard transformation lowers the ranges of values and increases the proportion of values being concentrated around 0 which lowers quantization errors. Using statistical tests for both activations and weights, we find that Hadamard transformations produce more normally distributed distributions, statistically significantly reduce the ranges, and move more values closer to 0 to give improved quantization performance. Given these, our main contributions are as follows:

- We provide a deeper understanding of how the Hadamard functions, and apply the transforms to directly modify weights and activations pre-quantization to statistically significantly reduce ranges and compact the signal around 0 which lowers quantization errors.
- We reparameterize the quantization scalars and zero offsets by decomposing both terms into two learnable variables, introducing two additional degrees of freedom that enable finer optimization of quantization parameters.
- We statistically significantly outperform SOTA in PSNR and SSIM across $\times 2$, $\times 3$, and $\times 4$ scale factors for all bitwidths. Specifically, we have gained **+1.53 dB** PSNR, and **+0.03** SSIM over CondiQuant Liu et al. (2025) on Manga109 at 2-bit $\times 4$. Qualitative results reveal sharper image reconstructions.
- We implement weight pruning with our quantization strategy at 3-4 bits. Using the Hadamard transform's ability to concentrate more values around 0, we prune **40**% of weights per quantized layer and have comparable performance with CondiQuant, but with **6.67**% and **15**% less bits per parameter for 3 and 4 bits respectively.

2 Related Work

2.1 Image Super Resolution

EDSR (Enhanced Deep Super-Resolution) Lim et al. (2017b), is a CNN-based architecture that improves upon traditional residual networks by removing unnecessary modules and stabilizing the training procedure. SRResNet Ledig et al. (2017) employs residual learning and partial convolution based padding to generate high-quality images with fine details. SwinIR Liang et al. (2021) is a transformer-based model that has demonstrated superior performance with a reduced number of

[1] Since CondiQuant does not have open-sourced code, we compare our visual results with 2DQuant.

parameters compared to CNN-based approaches by leveraging shallow and deep feature extraction along with the self-attention mechanism. SwinIR is made up of Residual Swin Transformer Blocks (RSTB), which are in turn made up of Swin Transformer Layers (STL). SwinIR-light is the SwinIR model designed for lightweight SR made up of 4 RSTBs that each contain 6 STLs. In our work we focus on this model for it's lightweight quality, which makes it more ideal for further compression and edge device deployment.

2.2 Model Quantization

Previous research has focused mainly on PTQ for CNN based architectures or vision transformers Hong & Lee (2024a); Makhov et al. (2024); Tu et al. (2023); Ding et al. (2022); Yuan et al. (2024) Li et al. (2023); Liu et al. (2023), QAT Tian et al. (2023); Hong & Lee (2024b); Hong et al. (2022a;b); Li et al. (2020); Zhong et al. (2022); Wang et al. (2021) or unique quantized architectures Qin et al. (2023). 2DQuant Liu et al. (2024) and CondiQuant Liu et al. (2025) represent the SOTA PTQ methods for SwinIR-light. 2DQuant performs Distribution-Oriented Bound Initialization (DOBI) to search for optimal clipping ranges from input distributions, then finetunes these bounds on a calibration dataset to minimize discrepancy with the full precision model's output. CondiQuant identifies that quantization errors primarily stem from activation quantization and uses the condition number of weight matrices to measure how sensitive outputs are to small input changes, employing proximal gradient descent to minimize these condition numbers while preserving model outputs. While both methods are effective, they do not directly address large ranges in weight and activation distributions. 2DQuant initializes parameters based on distributions but doesn't modify them, while CondiQuant focuses on condition numbers rather than distribution properties. Our method goes further by reducing the ranges of the weights and activations and compacting the signal through Hadamard transforms, making distributions more quantization-friendly.

2.3 Model Pruning

Prior work has experimented with pruning, although like quantization, the focus has been on convolutional models like EDSR. However, there has been some work on transformers and SwinIR. Chen et al. (2023) leverage activation sparsity in window-based vision transformers to prune activations while preserving the regular batching structure, enabling speedups on monocular 3D detection, 2D instance segmentation, and semantic segmentation with negligible accuracy loss. Prasetyo et al. (2023) apply Sparse Regularization and Pruning methods to the Vision Transformer architecture for image classification tasks and find that sparse regularization increases performance. Kim et al. and Jiang et al. (2023) focused on SwinIR specifically and experiment with knowledge distillation and pruning of the network. Their results show that the model compression could reduce computational costs and number of parameters without losing the performance, but their performance does not exceed ours. Lastly, Wang et al. (2025) also focus on knowledge distillation and pruning with SwinIR, letting the teacher guide channel selection during pruning for better accuracy and efficiency than sequential pipelines. It uses a learnable, differentiable auto-pruning module and a Multiscale Wavelet Refine Module to transfer high-frequency details. While these methods are effective, none implement two methods of compression at the same time. In our work we implement both quantization and pruning for more compression at higher bitwidths.

3 Methodology

3.1 Hadamard Transformation

Hadamard transformations have been used with effectiveness in quantization of LLMs to lower quantization errors. They have been said to increase "flatness" but this concept has not been rigorously evaluated. Hadamard matrices are recursively defined with entries in $\{\pm 1\}$ and implement linear, orthogonal, involutive transforms for dimensions that are powers of two, typically scaled by $1/\sqrt{n}$ (where n is the last dimension) for reversibility. To implement this, we first pad each weight and activation tensor with zeros so that their dimensions become powers of two. Once padded, we perform matrix multiplication between the tensors and a Hadamard matrix of the appropriate size in full precision. This is given in $X' = (H \cdot X)/\sqrt{\dim(X)}$ where $\dim(X)$ returns the last dimension of X. This operation has been said to flatten the matrices by distributing the magnitude of outliers

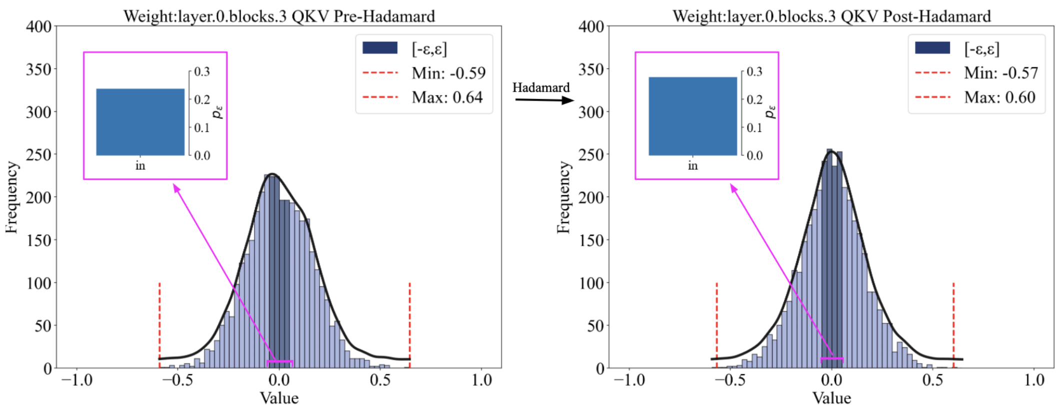

Figure 2: The left histogram shows the weight distribution prior Hadamard with a larger range and a sharper peak. The right histogram shows that the Post-Hadamard weight distribution is not flatter. The dark blue region indicates the values within $[-\epsilon, \epsilon]$ which now have more concentration after the transform as shown by the increase in the probability of a value being in $[-\varepsilon, \varepsilon]$, p_ε.

and that is how the errors get reduced, but the exact mechanism has not been explored nor has the flatness or normality of distributions been tested.

We test whether the weight and activation distributions are more normally distributed after the transformation. To perform all of our statistical tests, since tensor elements are not individually matchable across the Hadamard because each post-coordinate is a linear combination of many pre-coordinates, we treat each whole tensor as the experimental unit and form paired pre/post Hadamard summaries on the same tensor. For all experiments, we have 144 pairs of weight tensors and 240 pairs of activation tensors. For every paired pre/post tensor (p_j, q_j), we flatten the matrices, randomly sample 1 million elements and calculate the Shapiro-Wilk W-statistic prior to and after the transformation. We then calculate $\Delta_{wj} = W_{qj} - W_{pj}$ (positive means more normal). Aggregating over matrices $j = 1, \ldots, N$, we use a one-sided Wilcoxon signed-rank test with $H_0 : \text{median}(\Delta_w) = 0$ vs. $H_1 : \text{median}(\Delta_w) > 0$, and our findings in Table 1 show that W statistically significantly increases, so distributions become statistically significantly more normal after the transformation. Plotting a distribution also confirms this as shown in Figure 2.

Thus, there must be another reason for the Hadamards efficacy in quantization, since they do not make distributions flatter. We show that the Hadamard transformations reduce quantization errors in matrices by reducing the ranges of the values, and concentrating values around 0. This reduces the error because if values are closer together and closer to 0, quantizing them to a fixed value incurs less errors. We perform statistical tests to measure whether the Hadamard does reduce ranges and concentrate more values around 0.

To test whether the Hadamard transform statistically reduces the range of values in activations and weights, for every paired pre/post tensor (p_j, q_j), we flatten the arrays and align dimensions by right-padding the pre tensor with zeros to the post length, ensuring summaries live in the same ambient space as the transform. We then compute per-tensor ranges $R_j^{\text{pre}} = \max(p_j) - \min(p_j)$ and $R_j^{\text{post}} = \max(q_j) - \min(q_j)$, and form paired differences $\Delta_{rj} = R_j^{\text{pre}} - R_j^{\text{post}}$ (positive means a reduction). Aggregating over matrices $j = 1, \ldots, N$, we assess normality of $\{\Delta_{rj}\}$ using a Shapiro–Wilk test. Finding non-normality, we use a one-sided Wilcoxon signed-rank test with ties dropped to test $H_0 : \text{median}(\Delta_r) = 0$ vs. $H_1 : \text{median}(\Delta_r) > 0$. For interpretability, we compute a paired effect size, Cohen's $d_z = \bar{\Delta}_r / s_{\Delta_r}$. Table 1, shows both activations and weights have statistically significant range reductions at $\alpha = 0.05$.

To test whether the Hadamard transform increases mass near zero, we evaluate the proportion of entries within $[-\varepsilon, \varepsilon]$. In our work, we set $\varepsilon = 0.05$ to be close to 0. For each pre/post pair (p_j, q_j), we compute per-tensor in-band proportions $\hat{p}_j^{\text{pre}} = \frac{1}{n_{pj}} \sum_{i=1}^{n_{pj}} \mathbf{1}\{|p_{j,i}| \le \varepsilon\}$ and $\hat{p}_j^{\text{post}} = \frac{1}{n_{qj}} \sum_{i=1}^{n_{qj}} \mathbf{1}\{|q_{j,i}| \le \varepsilon\}$, then form the paired difference $\Delta_{pj} = \hat{p}_j^{\text{post}} - \hat{p}_j^{\text{pre}}$. Aggregating over tensors $j = 1, \ldots, N$ (weights and activations analyzed separately), we test $H_0 : \text{median}(\Delta_p) = 0$ vs

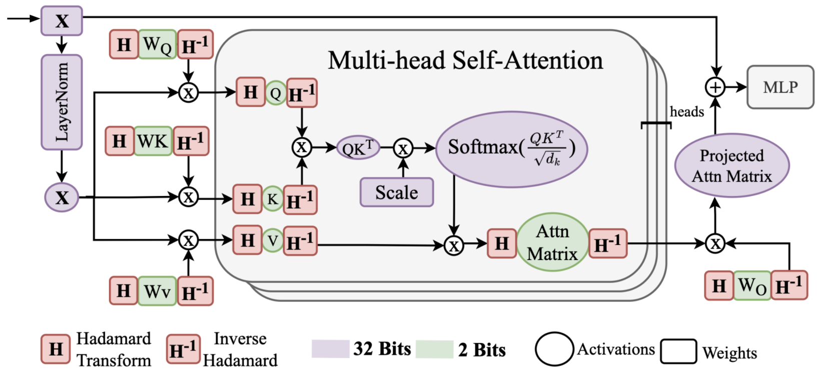

Figure 3: Architecture and Quantization scheme for Swin Transformer Layer (STL). X denotes the input. The quantized weights & activations per component are in green and the FP operations are in purple. The Hadamard and inverse Hadamard transforms are shown with red boxes.

H_1 : $\text{median}(\Delta_p) > 0$ using a one-sided Wilcoxon signed-rank test with ties dropped. We also compute a paired effect size (Cohen's d_z) on $\{\Delta_{pj}\}$.

As shown in Table 1, the ε-band proportion increases significantly after the Hadamard, consistent with shrinkage toward zero at $\alpha = 0.05$ with a significant effect size. All of these findings are further validated in Figure 2. The moving in of the red bars and lowering of min and max indicates range reduction and the bar graph to the left which calculates the proportion of values within the $[-\varepsilon, \varepsilon]$ band, p_ε, shows that p_ε increased after the transformation. Overall, we find that Hadamards statistically significantly make distributions more normal, reduce the range, shrink values, move more values closer within $[-\varepsilon, \varepsilon]$.

3.2 Scalar Decomposition

We quantize attention layers, linear layers and batch matrix multiplication weights and activations. The full pipeline of our method and what is quantized is shown in Figure 3 and Figure 4. To quantize the parameters to 2-4 bits, we use fake-quantization Jacob et al. (2017) i.e. quantization dequantization that simulates the loss of information through the quantization process by restricting the values to be representable by 2-4 bits but then immediately reverting them back to floats. The standard approach to fake quantization involves determining the quantization scalar, $S = \frac{\text{u}-\text{l}}{2^b-1}$, based on the upper and lower bounds of the clipping range. The equations are as follows: $v_c = \text{Clip}\,(x,\, l,\, u)$, $v_q = \text{Round}\,(\frac{2^b-1}{u-l}\,(v_c - l))$, $v_{deq} = \frac{u-l}{2^b-1}\,v_r + l$, where x is the tensor to be quantized, b is the bit-width, l is the lower bound, and u is the upper bound. The lower bound represents the zero offset, which is used to offset the quantization range, ensuring that the smallest value is 0 and so zero is exactly representable by an integer in the quantized range. In our work, we modify this value and the quantization scalar S which is $\frac{u-l}{2^b-1}$. Following 2DQuant, we have a step for searching for the upper and lower bounds of the clipping range to minimize quantization error, and a step for finetuning the quantization parameters given those starting points. We additively decompose the zero offset l and the quantization scalar $S = \frac{\text{u}-\text{l}}{2^b-1}$ into $S' = S + \alpha$ and $l' = l + \beta$ to be able to fine-tune each by one additional parameter. The idea behind introducing an additional predictor for the quantization scalar and zero offset serves two complementary purposes.

First, both variables are crucial for quantization, necessitating a method with high representation capabilities. By expanding the representational capacity of the model, the added predictor helps reduce the bias inherent in simpler models that rely on fewer predictors. Second, the additional parameters provide an alternative pathway through which gradients can flow more freely during backpropagation which allows for better optimization. To this end, we introduce α, which is initialized at 0 and adjusts the quantization scalar S, and β which is also initialized at 0 and we add the additional parameters to the scalar and zero point as shown above. When we search for the upper and lower bounds of the clipping range, we set these values to 0. However, during the parameter finetuning phase, we finetune these parameters starting from 0. With these adjustments, the quantization process then proceeds

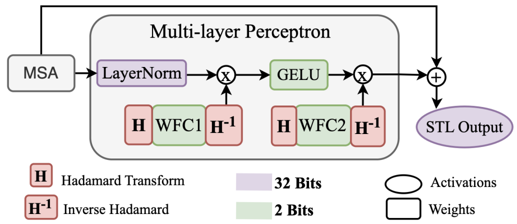

Figure 4: Architecture and quantization scheme for Swin Transformer Layer MLP. The Hadamard transformation is in full precision.

Test	Weights		Acts	
	p	d_z	p	d_z
$\text{median}(\Delta_w) > 0$	$3.9e^{-10}$	0.51	$1.09e^{-6}$	0.31
$\text{median}(\Delta_r) > 0$	$7.4e^{-13}$	0.44	$1.6e^{-36}$	0.97
$\text{median}(\Delta_p) > 0$	$3.0e^{-25}$	2.35	$1.2e^{-4}$	0.61

Table 1: Combined statistical tests for weights and activations. For all tests, $N_{\text{weights}} = 144$, $N_{\text{activations}} = 240$.To measure the effect size, we report Cohen's d_z. All $p < 0.05$.

as follows with the clipping step omitted for brevity: $v_q = \text{Round}\left(\left(\frac{2^b-1}{u-l} + \alpha\right)(v_c - (l+\beta))\right)$, $v_{deq} = (\frac{u-l}{2^b-1} + \alpha)v_q + (l+\beta)$. For the Clip and Round operations, we use the Straight Through Estimator Courbariaux et al. (2016) technique for calculating the gradients during backpropagation.

3.3 Pruning

To achieve more compression at 3 and 4 bits, prior to quantization but after the Hadamard transformation on each weight, we take advantage of the Hadamard's ability to move more values closer to 0 to perform pruning on the weights. As shown in Figure 2, the transform allows for more values to be closer to 0, allowing us to remove them and compress the weights further. We prune 40% of weights as this is the cutoff to actually give storage gains at 3 and 4 bits and results in minimal performance degradation. To perform this pruning, we take a sample of the matrix, take the absolute value of it, and calculate the desired pruning percentile, $T_{n\%}$. To prune 40% of the matrix, then we calculate the 40th percentile. Then, we set to 0 all values whose absolute values are below the threshold, corresponding to the n percent of the matrix with the smallest absolute magnitude. This follows the equation: $X_P = X \odot \mathbf{1}_{\{|X| \geq T_{n\%}\}}$, where X_P is the pruned tensor, and X is the input tensor. To actually lower bits per parameter by this pruning, we propose storing the smaller, pruned weight matrices along with a 1-bit per element mask matrix of 1s and 0s the size of the original tensor to store the indices of the pruned values in the full-sized matrix. In Section 4.6, we show that we can gain bits per parameter by performing our method.

4 Experiments

4.1 Data and Evaluation

We use DF2K Timofte et al. (2017); Lim et al. (2017a) as the training data, which combines DIV2K Timofte et al. (2017) and Flickr2K Lim et al. (2017a). We use the Set5 Bevilacqua et al. (2012) as the validation set. We test our method on five commonly used benchmarks in the SR field: Set5 Bevilacqua et al. (2012), Set14 Zeyde et al. (2012), B100 Martin et al. (2001), Urban100 Huang et al. (2015), and Manga109 Matsui et al. (2017). The evaluation metrics we used are Peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) Wang et al. (2004), which are calculated on the Y channel (i.e., luminance) of the YCbCr space. For both metrics, higher indicates better performance. The implementation details of our method are given in Section 7.1.

4.2 Results

CompSRT demonstrates superior performance to SOTA across all experimental configurations and benchmarks. Table 2 presents comprehensive comparisons of our quantization method and previous SOTA across scale factor $(\times 4)$ and 2-4 bit widths. Additional results for other scale factors are included in Section 7.2. To assess whether CompSRT statistically significantly outperforms CondiQuant, for a fixed scale and bit width and for each metric $m \in \{\text{PSNR}, \text{SSIM}\}$, we form per-dataset differences $\Delta_i = \text{score}^{(m)}_{\text{CompSRT},i} - \text{score}^{(m)}_{\text{CondiQuant},i}$ for $i = 1, \ldots, N$ (with $N = 5$ datasets). We then test $H_0 : \text{median}(\Delta) = 0$ versus $H_1 : \text{median}(\Delta) > 0$ using the paired Wilcoxon signed-rank test; exact ties ($\Delta_i = 0$) are excluded. We present the one-sided Wilcoxon p-value and the

Method (×4)	Bit	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR-light	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
PTQ4ViT	4	31.49	0.8831	28.04	0.7680	27.20	0.7240	25.53	0.7660	29.52	0.8940
NoisyQuant	4	31.09	0.8751	27.75	0.7591	26.91	0.7151	25.07	0.7500	28.96	0.8820
2DQuant	4	31.77	0.8867	28.30	0.7733	27.37	0.7278	25.71	0.7712	29.71	0.8972
CondiQuant	4	32.09	0.8923	28.50	0.7792	27.52	0.7345	25.97	0.7831	30.16	0.9054
CompSRT (ours)	4	**32.41**	**0.8969**	**28.74**	**0.7849**	**27.68**	**0.7399**	**26.39**	**0.7953**	**30.81**	**0.9131**
PTQ4ViT	3	29.77	0.8337	27.00	0.7248	26.21	0.6735	24.22	0.6983	27.94	0.8479
NoisyQuant	3	28.90	0.7972	26.50	0.6970	26.16	0.6628	23.86	0.6667	27.17	0.8116
2DQuant	3	30.90	0.8704	27.75	0.7571	26.99	0.7126	24.85	0.7355	28.21	0.8683
CondiQuant	3	31.62	0.8855	28.20	0.7715	27.31	0.7269	25.39	0.7624	29.29	0.8915
CompSRT (ours)	3	**32.31**	**0.8956**	**28.69**	**0.7839**	**27.64**	**0.7387**	**26.27**	**0.7918**	**30.60**	**0.9108**
PTQ4ViT	2	27.23	0.6702	25.38	0.5914	25.15	0.5621	22.94	0.5587	24.66	0.6132
NoisyQuant	2	25.94	0.5862	24.33	0.5067	24.16	0.4718	22.32	0.4841	23.82	0.5403
2DQuant	2	29.53	0.8372	26.86	0.7322	26.46	0.6927	23.84	0.6912	26.07	0.8163
CondiQuant	2	30.64	0.8671	27.59	0.7567	26.93	0.7136	24.54	0.7282	27.67	0.8613
CompSRT (ours)	2	**31.44**	**0.8820**	**28.15**	**0.7696**	**27.28**	**0.7253**	**25.38**	**0.7585**	**29.20**	**0.8881**

Table 2: Performance comparison with state-of-the-art methods for scale factor $(\times 4)$ across different bit widths. All comparative results are taken from SwinIR-light Liang et al. (2021), PTQ4VIT Yuan et al. (2024), NoisyQuant Liu et al. (2023), 2DQuant Liu et al. (2024), and CondiQuant Liu et al. (2025)as reported in their papers. Our method achieves superior performance across all configurations.

effect size given by Cohen's $d_z = \bar{\Delta}/s_{\Delta}$, where $\bar{\Delta}$ is the mean of the paired differences and s_{Δ} is their sample standard deviation. This procedure is run independently for each (scale, bit, metric) configuration using only dataset-level PSNR and SSIM averages. As presented in Table 5, the resulting p-values for all pairwise comparisons fall below the significance threshold of $\alpha = 0.05$. This allows us to reject the null hypothesis and confirms that CompSRT statistically significantly outperforms CondiQuant across all evaluated conditions. Notably, our 4-bit quantized model delivers performance remarkably close to the full-precision baseline. For $(\times 4)$, the difference ranges from 0.01–0.11 dB across datasets (0.04 on Set5). These results highlight that CompSRT achieves almost full-precision quality for 4-bits while reducing the model size by a factor of 8 times. Furthermore, we have gained +1.53 dB PSNR, and +0.03 SSIM over CondiQuant Liu et al. (2025) on Manga109 at 2-bit $\times 4$, higlighting our large gains.

4.3 Qualitative Results

We show the visual results comparing the performance of our 2-bit $(\times 4)$ model with the 2-bit $(\times 4)$ 2DQuant model, using the original low-quality image as input, as illustrated in Figure 5. These qualitative comparisons highlight the effectiveness of our approach in enhancement. Across both natural images and manga illustrations, our method consistently outperforms 2DQuant by generating images that exhibit significantly less blurriness, more clarity and sharper lines. For example, in examples from Manga109, our method produces less grainy images, making our method applicable to animation. In images from Urban100, our model's output exhibits the lines sharply, while 2DQuant's rendering has blur. On the Set14 image with writing, our method more clearly enhances words. Lastly, As shown by the images of the man from B100 and the woman from Set5, our method can more clearly enhance close ups of human eyes, making our method more applicable to security and face identification domains.

4.4 Pruning Results

We show the performance of adding weight pruning to our CompSRT method on the 3 and 4-bit quantized $\times 4$ SwinIR models on Set5. We add this step after the Hadamard transformation but prior to quantization. We experiment with varying percentages of pruning, with the results for 0% to 99.5% being show in Figure 7. The figure shows that performance in terms of PSNR and SSIM degrades after pruning any percentage for 4 bits, but at 3 bits, the performance slightly increases from 10 to 30%. This is because the small values close to 0 might have been closer to noise, and for a model with less representational capacity, setting them to 0 allows for learning a smoother more easily

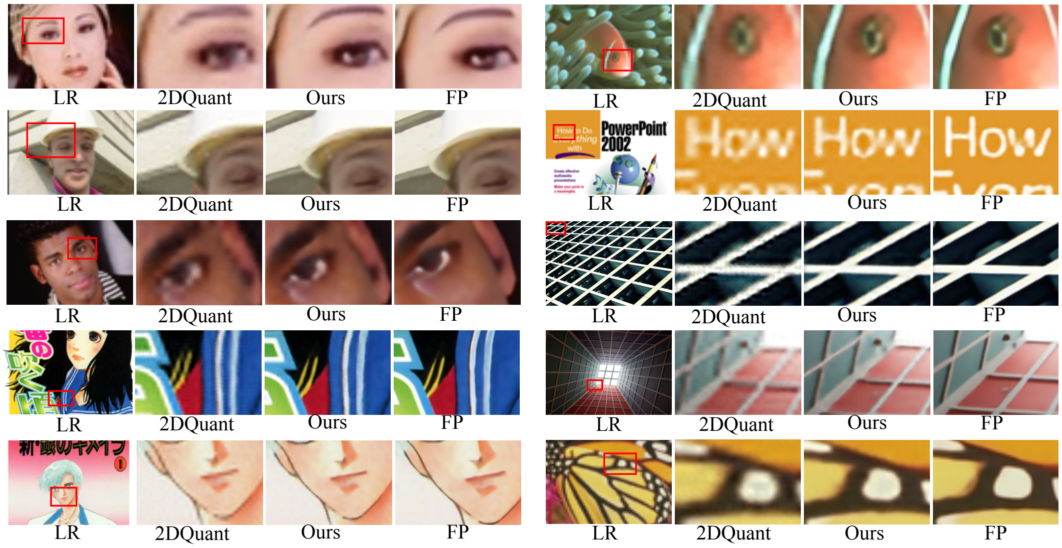

Figure 5: Qualitative visual comparison for 2-bit $(\times 4)$ SR across all five benchmark datasets. LR denotes the low resolution image. Comparative examples are taken from the FP model SwinIR-light Liang et al. (2021), and 2-bit $(\times 4)$ 2DQuant Liu et al. (2024).

representable signal during finetuning. However, as the representational capacity of the 4-bit model is exponentially larger than the 3-bit model, the 4-bit model suffers after the pruning. After 40% for both bitwidths, there is a slight drop, but our performance is comparable with SOTA CondiQuant. That is why we chose 40% pruning in our method; to achieve maximum space reduction without any cost in performance. To further explore the performance of the 40% weight pruned model, we examine the specifics of its performance across varying quantization bitwidths given in Table 3. We find that performance degraded from our quantization method by 0.39 dB for 4 bits and by 0.47 dB for 3 bits, but stayed on par with Condiquant in terms of PSNR and SSIM. This shows that 40% of the weight signal can be extraneous and can be removed without a great loss in performance. The loss at higher percentages of pruning is more pronounced, with any percentage after 40 performing worse, as show in Figure 7.

4.5 Ablation Studies

We conduct ablation studies with the two parts of our quantization method to find the element with the most impact. The results presented in Figure 6 indicate that the most significant performance improvements stem from incorporating additional trainable parameters via scalar decomposition. The trainable parameters alone contributed to 0.64 dB increase in Set5 PSNR over the CondiQuantLiu et al. (2025) baseline. Applying Hadamard transformations to both weights and activations also yields a 0.62 dB gain over the 2DQuant baseline, proving its standalone efficacy. Furthermore, applying the Hadamard transformation to only the weights or adding trainable parameters with Hadamard transforms only on the weights did not lead to large improvements, indicating that weights benefit less from range reduction and compaction. Given these results, our analysis supports the conclusion that the primary advantage of the Hadamard transformation lies in its ability to reduce extreme values in activation distributions.

4.6 Model Complexity

We evaluate the time and space complexity of our quantization and pruning method. To evaluate the time complexity, we measured the time required to complete a single forward pass on our 2-bit quantized $\times 4$ model and our 4-bit quantized and pruned model on an image from the Set5 benchmark dataset in seconds. We also measured the size of the model's trainable parameters in megabytes (MB) and in bits per parameter b. To calculate the size for pruned models, we take into account the unpruned portion of the model, along with the storage cost of storing a 1-bit mask per matrix to store the location of 0s. The calculation for bits per parameter in this case at 4 bits is $(4 * 0.6 + 1) = 3.4$ vs.

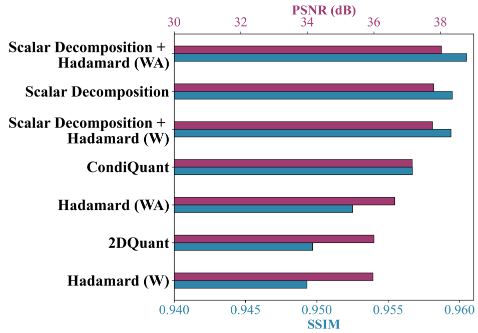

Figure 6: Ablation studies. Blue bars show SSIM and purple bars show PSNR. W = weights; WA = weights+activations. All models are $(\times 2)$, 2-bit PTQ, evaluated on Set5.

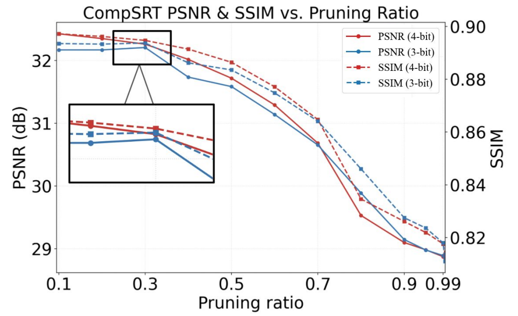

Figure 7: 3 and 4-bit $(\times 4)$ model PSNR and SSIM performance on Set5 vs. pruning percentage. Only quantized layers are pruned. Between 10-30% error lowers, but increases after 40%.

Bits	PSNR	SSIM
4	32.02	0.89
3	31.74	0.89

Table 3: Bitwise performance for CompSRT + 40% weight-pruned $(\times 4)$ models. Performance is comparable with SOTA after pruning.

Method	Bit	s	MB	b
2DQuant	4	0.048	13.99	4
Ours	4	0.092	14.00	4
Ours + Prune	4	0.092	13.95	3.4

Table 4: Time and space complexity for $(\times 4)$ models. All models evaluated on Set5. b denotes bits per parameter and s seconds.

Scale, Bit	PSNR		SSIM	
	p	d_z	p	d_z
×2, 4	0.031	1.57	0.031	1.28
×2, 3	0.031	1.64	0.031	1.15
×2, 2	0.031	1.83	0.031	1.11
×3, 4	0.031	1.63	0.031	1.68
×3, 3	0.031	1.79	0.031	1.60
×3, 2	0.031	1.85	0.031	1.87
×4, 4	0.031	1.89	0.031	2.33
×4, 3	0.031	1.95	0.031	2.08
×4, 2	0.031	1.83	0.031	2.25

Table 5: One-sided Wilcoxon signed-rank tests (CompSRT $>$ CondiQuant) and paired effect sizes (Cohen's d_z). All p-values < 0.05.

4 (15% reduction in bits per parameter). The calculation for 3 bits is $(3 * 0.6 + 1) = 2.8$ vs. 3 (6.67% reduction in bits per parameter). Furthermore, Table 4 shows that our quantization method adds no storage overhead but incorporating Hadamard transformations introduces a modest computational overhead of 0.044 seconds compared to the 2DQuant baseline. This is expected, as this process introduces an additional dense matrix multiplication per quantized matrix. However, there is very minimal additional storage overhead because Hadamard matrices don't need to be stored and can be created when needed and we add only 2 extra parameters. When comparing pruning against 2DQuant, our approach introduces more compression but adds no additional runtime to our method, as pruning is done once per layer. This demonstrates that Hadamards can be applied to both quantization and pruning of Swin-IR light. For MB of storage gains, it can be physically minimal without packing, but the bits per parameter gains can allow us to pack values more tightly.

5 Discussion

5.1 Conclusion

In this work, we propose a novel Hadamard guided approach to improve image SR PTQ. We challenge previous intuitions about the Hadamard transform and find that the Hadamard transform does not make distributions flatter in SwinIR-light. Instead, we hypothesized that there is another mechanism for their function in improving quantization. We find that instead it decreases the ranges and increases the concentration of values around 0 in activations and weights, which is why it lowers quantization errors. These properties improve quantization to low bit widths but also allow us to prune 40% of weights for increased compression without significant loss in performance. We also perform parameter decomposition, which leverages the bias-variance tradeoff. Our quantization method achieves statistically significant gains in quantitative metrics and visible improvements in visual quality over the SOTA quantization method, with minimal storage overhead; our pruned models have comparable performance with SOTA but with 6.67-15% less bits per parameter.

6 REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our results. The model architecture is described in Sections 2 and 3 of the main paper, and the training details are described in Section 4. All datasets used in our experiments are publicly available and evaluation methods are also described in Section 4. Finally, an anonymous link to our source code and scripts for training, evaluation, and statistical tests, together with exact run commands and an environment specification, is provided in the supplementary materials to facilitate faithful replication of our experiments.

REFERENCES

Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *British Machine Vision Conference*, 2012.

Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, and Song Han. Sparsevit: Revisiting activation sparsity for efficient high-resolution vision transformer, 2023. URL `https://arxiv.org/abs/2303.17605`.

Matthieu Courbariaux, Itay Hubara, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Yifu Ding, Haotong Qin, Qinghua Yan, Zhenhua Chai, Junjie Liu, Xiaolin Wei, and Xianglong Liu. Towards accurate post-training quantization for vision transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 5380–5388, 2022.

Hayit Greenspan, G Oz, N Kiryati, and SLBG Peled. Mri inter-slice reconstruction using super-resolution. *Magnetic resonance imaging*, 20(5):437–446, 2002.

L Hitachi. Super-resolution technology to convert video of various resolutions to high-definition.

Cheeun Hong and Kyoung Mu Lee. Adabm: on-the-fly adaptive bit mapping for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2641–2650, 2024a.

Cheeun Hong and Kyoung Mu Lee. Overcoming distribution mismatch in quantizing image super-resolution networks. In *European Conference on Computer Vision*, pp. 380–396. Springer, 2024b.

Cheeun Hong, Sungyong Baik, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Cadyq: Content-aware dynamic quantization for image super-resolution. In *European Conference on Computer Vision*, pp. 367–383. Springer, 2022a.

Cheeun Hong, Heewon Kim, Sungyong Baik, Junghun Oh, and Kyoung Mu Lee. Daq: Channel-wise distribution-aware quantization for deep image super-resolution networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2675–2684, 2022b.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5197–5206, 2015. doi: 10.1109/CVPR.2015.7299156.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017. URL `https://arxiv.org/abs/1712.05877`.

Yuxuan Jiang, Jakub Nawala, Fan Zhang, and David Bull. Compressing deep image super-resolution models. *arXiv preprint arXiv:2401.00523*, 2023.

Dokyun Kim, Gyeongseon Eo, Sooyoun Park, and Soyeon Park. Lightweight transformer super-resolution.

D Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5, pp. 6. San Diego, California;, 2015.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Huixia Li, Chenqian Yan, Shaohui Lin, Xiawu Zheng, Baochang Zhang, Fan Yang, and Rongrong Ji. Pams: Quantized super-resolution via parameterized max scale. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 564–580. Springer, 2020.

Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers, 2023. URL https://arxiv.org/abs/2212.08254.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution, 2017a. URL https://arxiv.org/abs/1707.02921.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017b.

Kai Liu, Haotong Qin, Yong Guo, Xin Yuan, Linghe Kong, Guihai Chen, and Yulun Zhang. 2dquant: Low-bit post-training quantization for image super-resolution. *arXiv preprint arXiv:2406.06649*, 2024.

Kai Liu, Dehui Wang, Zhiteng Li, Zheng Chen, Yong Guo, Wenbo Li, Linghe Kong, and Yulun Zhang. Condiquant: Condition number based low-bit quantization for image super-resolution. *arXiv preprint arXiv:2502.15478*, 2025.

Yijiang Liu, Huanrui Yang, Zhen Dong, Kurt Keutzer, Li Du, and Shanghang Zhang. Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers, 2023. URL https://arxiv.org/abs/2211.16056.

Denis Makhov, Ruslan Ostapets, Irina Zhelavskaya, Dehua Song, and Kirill Solodskikh. Towards robust full low-bit quantization of super resolution networks. In *European Conference on Computer Vision*, pp. 182–198. Springer, 2024.

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pp. 416–423 vol.2, 2001. doi: 10.1109/ICCV.2001.937655.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using manga109 dataset. *Multimedia tools and applications*, 76:21811–21838, 2017.

Kiran Murthy, Michael Shearn, Byron D Smiley, Alexandra H Chau, Josh Levine, and M Dirk Robinson. Skysat-1: very high-resolution imagery from a small satellite. In *Sensors, systems, and next-generation Satellites XVIII*, volume 9241, pp. 367–378. SPIE, 2014.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Yogi Prasetyo, Novanto Yudistira, and Agus Wahyu Widodo. Sparse then prune: Toward efficient vision transformers. *arXiv preprint arXiv:2307.11988*, 2023.

Haotong Qin, Yulun Zhang, Yifu Ding, Xianglong Liu, Martin Danelljan, Fisher Yu, et al. Quantsr: accurate low-bit quantization for efficient image super-resolution. *Advances in Neural Information Processing Systems*, 36:56838–56848, 2023.

M Dirk Robinson, Stephanie J Chiu, Cynthia A Toth, Joseph A Izatt, Joseph Y Lo, and Sina Farsiu. New applications of super-resolution in medical imaging. In *Super-resolution imaging*, pp. 383–412. CRC Press, 2017.

Heng Su, Liang Tang, Ying Wu, Daniel Tretter, and Jie Zhou. Spatially adaptive block-based super-resolution. *IEEE Transactions on Image Processing*, 21(3):1031–1045, 2011.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, et al. Flatquant: Flatness matters for llm quantization. *arXiv preprint arXiv:2410.09426*, 2024.

Hiroyuki Takeda, Peyman Milanfar, Matan Protter, and Michael Elad. Super-resolution without explicit subpixel motion estimation. *IEEE transactions on image processing*, 18(9):1958–1975, 2009.

Senmao Tian, Ming Lu, Jiaming Liu, Yandong Guo, Yurong Chen, and Shunli Zhang. Cabm: Content-aware bit mapping for single image super-resolution network with large input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1756–1765, 2023.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, Lei Zhang, Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, Kyoung Mu Lee, Xintao Wang, Yapeng Tian, Ke Yu, Yulun Zhang, Shixiang Wu, Chao Dong, Liang Lin, Yu Qiao, Chen Change Loy, Woong Bae, Jaejun Yoo, Yoseob Han, Jong Chul Ye, Jae-Seok Choi, Munchurl Kim, Yuchen Fan, Jiahui Yu, Wei Han, Ding Liu, Haichao Yu, Zhangyang Wang, Honghui Shi, Xinchao Wang, Thomas S. Huang, Yunjin Chen, Kai Zhang, Wangmeng Zuo, Zhimin Tang, Linkai Luo, Shaohui Li, Min Fu, Lei Cao, Wen Heng, Giang Bui, Truc Le, Ye Duan, Dacheng Tao, Ruxin Wang, Xu Lin, Jianxin Pang, Jinchang Xu, Yu Zhao, Xiangyu Xu, Jinshan Pan, Deqing Sun, Yujin Zhang, Xibin Song, Yuchao Dai, Xueying Qin, Xuan-Phung Huynh, Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, Vishal Monga, Cristovao Cruz, Karen Egiazarian, Vladimir Katkovnik, Rakesh Mehta, Arnav Kumar Jain, Abhinav Agarwalla, Ch V. Sai Praveen, Ruofan Zhou, Hongdiao Wen, Che Zhu, Zhiqiang Xia, Zhengtao Wang, and Qi Guo. Ntire 2017 challenge on single image super-resolution: Methods and results. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1110–1121, 2017. doi: 10.1109/CVPRW.2017.149.

Albert Tseng, Jerry Chee, Qingyao Sun, Volodymyr Kuleshov, and Christopher De Sa. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. *arXiv preprint arXiv:2402.04396*, 2024.

Zhijun Tu, Jie Hu, Hanting Chen, and Yunhe Wang. Toward accurate post-training quantization for image super resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5856–5865, 2023.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Hu Wang, Peng Chen, Bohan Zhuang, and Chunhua Shen. Fully quantized image super-resolution networks, 2021. URL `https://arxiv.org/abs/2011.14265`.

Yanzhe Wang, Yizhen Wang, Avinash Rohra, and Baoqun Yin. End-to-end model compression via pruning and knowledge distillation for lightweight image super resolution. *Pattern Analysis and Applications*, 28(2):94, 2025.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Haichao Yu, Ding Liu, Honghui Shi, Hanchao Yu, Zhangyang Wang, Xinchao Wang, Brent Cross, Matthew Bramler, and Thomas S Huang. Computed tomography super-resolution using convolutional neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3944–3948. IEEE, 2017.

Jaesok Yu, Linda Lavery, and Kang Kim. Super-resolution ultrasound imaging method for microvasculature in vivo with a high temporal accuracy. *Scientific reports*, 8(1):13918, 2018.

Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization, 2024. URL https://arxiv.org/abs/2111.12293.

Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal processing*, 128:389–408, 2016.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In Jean-Daniel Boissonnat, Patrick Chenin, Albert Cohen, Christian Gout, Tom Lyche, Marie-Laurence Mazure, and Larry Schumaker (eds.), *Curves and Surfaces*, pp. 711–730, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-27413-8.

Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution, 2018. URL https://arxiv.org/abs/1802.08797.

Yunshan Zhong, Mingbao Lin, Xunchao Li, Ke Li, Yunhang Shen, Fei Chao, Yongjian Wu, and Rongrong Ji. Dynamic dual trainable bounds for ultra-low precision super-resolution networks. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2022.

Hong Zhu, Xinming Tang, Junfeng Xie, Weidong Song, Fan Mo, and Xiaoming Gao. Spatio-temporal super-resolution reconstruction of remote-sensing images based on adaptive multi-scale detail enhancement. *Sensors*, 18(2):498, 2018.

7 Supplementary Materials

7.1 Implementation details

We build our work on top of 2DQuant's open sourced repository Liu et al. (2024), and use SwinIR-light Liang et al. (2021) as the model backbone of our method. For the Hadamard transform and statistical tests, we use SciPy Virtanen et al. (2020). For finetuning, we use the Adam Kinga et al. (2015) optimizer with a learning rate of $1 * 10^{-2}$ and betas set to (0.9, 0.999). We clip all gradient values to between $[-1, 1]$ and we finetune for at most 4000 iterations, or until we reach a nan gradient which we handle by safely exiting. We select the model with the highest PSNR and SSIM. For measuring the time and space complexity of our model, we calculate inference time in seconds and size of model parameters in MB. Our code is written with PyTorch Paszke et al. (2019) and runs for at most 7 hours on one NVIDIA RTX 6000 48G GPU. The anonymous open source code for this paper along with instructions can be found here.

7.2 Results

Method ($\times 2$)	Bit	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR-light	32	38.15	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.11	0.9781
Bicubic	32	32.25	0.9118	29.25	0.8406	28.68	0.8104	25.96	0.8088	29.17	0.9128
PTQ4ViT	4	37.43	0.9571	33.19	0.9139	31.84	0.8950	31.54	0.9212	37.59	0.9735
NoisyQuant	4	37.50	0.9570	33.06	0.9101	31.73	0.8936	31.31	0.9181	37.47	0.9723
2DQuant	4	37.87	0.9594	33.41	0.9161	32.02	0.8971	31.84	0.9251	38.31	0.9761
CondiQuant	4	38.03	0.9605	33.50	0.9180	32.16	0.8993	32.03	0.9282	38.57	0.9769
CompSRT (ours)	4	**38.13**	**0.9610**	**33.81**	**0.9203**	**32.28**	**0.9009**	**32.57**	**0.9325**	**38.98**	**0.9778**
PTQ4ViT	3	36.49	0.9510	32.49	0.9045	31.27	0.8854	30.16	0.9027	36.41	0.9673
NoisyQuant	3	35.32	0.9334	31.88	0.8911	30.73	0.8710	29.28	0.8835	35.30	0.9537
2DQuant	3	37.32	0.9567	32.35	0.9106	31.60	0.8911	30.45	0.9086	37.24	0.9722
CondiQuant	3	37.77	0.9594	33.21	0.9151	31.94	0.8966	31.18	0.9197	38.01	0.9755
CompSRT (ours)	3	**38.11**	**0.9609**	**33.82**	0.9202	**32.27**	**0.9008**	**32.53**	**0.9321**	**38.90**	**0.9775**
PTQ4ViT	2	33.25	0.8923	30.22	0.8402	29.21	0.8066	27.31	0.8111	32.75	0.9093
NoisyQuant	2	30.13	0.7620	28.80	0.7536	28.26	0.7421	26.68	0.7627	30.40	0.8204
2DQuant	2	36.00	0.9497	31.98	0.9012	30.91	0.8810	28.62	0.8819	34.40	0.9602
CondiQuant	2	37.15	0.9567	32.74	0.9103	31.55	0.8912	29.96	0.9047	36.63	0.9713
CompSRT (ours)	2	**38.03**	**0.9605**	**33.70**	**0.9194**	**32.19**	**0.9294**	**32.22**	**0.9294**	**38.69**	**0.9770**

Table 6: Performance comparison with SOTA methods for scale factor $(\times 2)$ across different bit widths. All comparative results are taken from SwinIR-light Liang et al. (2021), PTQ4VIT Yuan et al. (2024), NoisyQuant Liu et al. (2023), 2DQuant Liu et al. (2024), and CondiQuant Liu et al. (2025). Our method achieves superior performance across all datasets and bitwidths.

Method ($\times 3$)	Bit	Set5		Set14		B100		Urban100		Manga109	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SwinIR-light	32	34.63	0.9290	30.54	0.8464	29.20	0.8082	28.66	0.8624	33.99	0.9478
Bicubic	32	29.54	0.8516	27.04	0.7551	26.78	0.7187	24.00	0.7144	26.16	0.8384
PTQ4ViT	4	33.77	0.9201	29.75	0.8272	28.62	0.7942	27.43	0.8361	32.50	0.9360
NoisyQuant	4	33.13	0.9122	29.06	0.8093	27.93	0.7754	26.66	0.8143	31.94	0.9293
2DQuant	4	34.06	0.9231	30.12	0.8374	28.89	0.7988	27.69	0.8405	32.88	0.9389
CondiQuant	4	34.32	0.9260	30.29	0.8417	29.05	0.8039	28.05	0.8506	33.23	0.9431
CompSRT (ours)	4	**34.56**	**0.9284**	**30.49**	**0.8454**	**29.17**	**0.8075**	**28.50**	**0.8598**	**33.83**	**0.9467**
PTQ4ViT	3	32.75	0.9028	29.14	0.8113	28.06	0.7712	26.43	0.8014	31.20	0.9131
NoisyQuant	3	30.78	0.8511	27.94	0.7624	26.98	0.7153	25.43	0.7481	29.64	0.8792
2DQuant	3	33.24	0.9135	29.56	0.8255	28.50	0.7873	26.65	0.8116	31.46	0.9235
CondiQuant	3	33.92	0.9224	30.02	0.8367	28.84	0.7986	27.37	0.8356	32.48	0.9367
CompSRT (ours)	3	**34.54**	**0.9281**	**30.48**	**0.8451**	**29.16**	**0.8070**	**28.47**	**0.8589**	**33.79**	**0.9465**
PTQ4ViT	2	29.96	0.7901	27.36	0.7030	26.74	0.6590	24.56	0.6460	27.37	0.7390
NoisyQuant	2	27.53	0.6641	25.77	0.5952	25.37	0.5613	23.59	0.5739	26.03	0.6632
2DQuant	2	31.62	0.8887	28.54	0.8038	27.85	0.7679	25.30	0.7685	28.46	0.8814
CondiQuant	2	33.00	0.9130	29.44	0.8253	28.45	0.7882	26.36	0.8080	30.88	0.9203
CompSRT (ours)	2	**34.17**	**0.9248**	**30.21**	**0.8401**	**28.97**	**0.8017**	**27.86**	**0.8456**	**33.11**	**0.9414**

Table 7: Performance comparison with SOTA methods for scale factor $(\times 3)$ across different bit widths. All comparative results are taken from SwinIR-light Liang et al. (2021), PTQ4VIT Yuan et al. (2024), NoisyQuant Liu et al. (2023), 2DQuant Liu et al. (2024), and CondiQuant Liu et al. (2025). Our method achieves superior performance across all datasets and bitwidths.

7.3 LLM Usage

We used LLMs to assist—but not replace—our research workflow. Specifically, LLMs were employed to (i) help draft and refactor code snippets and experimental scripts, (ii) brainstorm and clarify ideas and concepts discussed in the paper, (iii) suggest edits and critiques on early drafts, and (iv) provide limited writing assistance for grammar and phrasing. All model outputs were reviewed, verified, and, where needed, rewritten by the authors; we independently implemented, tested, and validated every algorithmic choice and experimental result reported. No proprietary, confidential, or unreleased data were provided to the models. LLMs were not used to generate or fabricate data, analyses, or citations, and they are not listed as authors. The authors bear full responsibility for the paper's content, including the correctness of code, experiments, and references.