STRUCTURALLY AWARE POST-TRAINING QUANTIZATION OF FOURIER NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fourier neural operators (FNOs) provide an efficient framework for learning mappings between function spaces, but deploying trained operators on resource-constrained hardware requires effective model compression. Post-training quantization (PTQ) is particularly relevant here because, unlike quantization-aware training (QAT), it requires no retraining and can be applied directly to a trained model. Despite this, the effectiveness of PTQ for neural operators remains largely unexplored. This paper presents a systematic study of PTQ for FNOs on two three-dimensional problems: forced isotropic turbulence and Darcy flow. We evaluate quantization strategies across model capacities and bit-widths at equal storage and find that quantizer design has a substantial impact on accuracy. In particular, different Fourier modes have very different sensitivity to quantization, and the measured input energy closely tracks quantization sensitivity. Allocating bits according to this input energy reduces error across all configurations tested, with a median improvement over uniform allocation of $5.2\times$ overall and of $14.9\times$, $5.9\times$ and $3.1\times$ at two, three and four bits. We also evaluate three other aspects of quantization: weight representation, scaling granularity, and quantization range. Once the bit budget is allocated according to input energy, these choices have substantially smaller effects. Nevertheless, at a fixed bit-width, different quantizer designs can produce errors that differ by up to two orders of magnitude. These results show that exploiting the structure of Fourier Neural Operators can substantially improve post-training quantization.

1 INTRODUCTION

Neural operators learn mappings between function spaces and serve as surrogates for parameterized partial differential equations, so that one trained model replaces repeated numerical simulation (Li et al., 2021; Kovachki et al., 2023). Such a surrogate is trained once and evaluated many times, and its size often determines where it can be deployed. In a Fourier neural operator (FNO), nearly all parameters are complex weights that act on individual Fourier modes of the input. In the three-dimensional models studied here they account for over 99.9% of the parameters.

Quantization reduces this size by storing the weights with fewer bits. Quantization-aware training has been developed for neural PDE solvers (van den Dool et al., 2023); it fine-tunes the model with the quantizer in the loop. Post-training quantization instead quantizes a trained model directly, using at most a small unlabelled calibration set. It is therefore particularly attractive when a trained model must be compressed without retraining.

Standard post-training quantization gives every weight in a tensor the same number of bits and sets quantization ranges from the weights themselves (Nagel et al., 2021). This implicitly treats all weights as equally sensitive to rounding. In an FNO they are not. Each weight acts on a single Fourier mode, so an error in that weight matters roughly in proportion to how much of the input signal lies at that mode.

This makes bit allocation across frequency bands a central part of PTQ for FNOs. We group the Fourier modes into frequency bands indexed by $\mathrm{round}(|k|)$, and use *mode* for an individual Fourier coefficient and *band* for the unit that receives a bit-width. Because each weight acts on one mode, the quantization cost separates across modes, and each band is weighted by the input energy it

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

receives. Distributing bits under this cost is a reverse water-filling problem (Huang & Schultheiss, 1963; Gersho & Gray, 1992), and bands that receive almost no input energy are pruned. The input spectrum needed for the allocation is measured with a few forward passes on unlabelled data.

We test this view on two three-dimensional problems, forced isotropic turbulence and Darcy flow, at two model sizes and at two to four bits per weight component. If input energy predicts sensitivity, allocating bits by it should help in every setting, and the remaining quantizer choices should matter much less once it is used. Both predictions hold. The allocation reduces the error in every configuration tested and brings the larger models close to full-precision accuracy. Once the bits are allocated, the weight representation, including a polar representation of the complex weights, the quantization granularity and the clipping rule have much smaller effects.

Our contributions are the following.

(a) The observation that Fourier modes differ widely in their sensitivity to quantization, together with a one-layer cost model in which the input energy of a mode predicts the cost of perturbing its weights, and experiments that support it.
(b) A formulation of post-training quantization for FNOs as bit allocation across frequency bands, solved by reverse water-filling on the measured input spectrum, which reduces the error in every configuration tested.
(c) A study, on two PDEs, two model sizes and three bit-widths at equal storage, of how much weight representation, quantization granularity and clipping matter once the bits are allocated.

Figure 1 summarizes the approach and its outcome.

2 Related work

Neural operators and their cost. FNOs (Li et al., 2021; Kovachki et al., 2023) parameterize a kernel integral operator in the Fourier domain. Their cost has been reduced through factorized and tensorized spectral weights (Kossaifi et al., 2024) and through reduced numerical precision during training (Tu et al., 2024). Quantization-aware training (van den Dool et al., 2023) and fully quantized training (Lu et al., 2026) have been applied to neural PDE solvers, and the post-training setting studied here builds on that work.

Second-order, activation-aware and mixed-precision quantization. Post-training quantization combines bit allocation across weights with discrete rounding. Second-order rounding methods such as Optimal Brain Surgeon (Hassibi & Stork, 1993) and GPTQ (Frantar et al., 2023) operate on the rounding axis for a fixed bit-width, which composes orthogonally with our allocation. For bit allocation, HAWQ (Dong et al., 2019) uses Hessian eigenvalues across whole layers, and AWQ (Lin et al., 2024) protects salient weights from activation magnitudes. For the single-layer squared-output loss, the Hessian with respect to the spectral weights is block-diagonal across modes, with blocks proportional to the mode input energy $e(k)$; reverse water-filling is thus the natural Fourier-domain analogue of Hessian-aware mixed precision, operating at intra-layer frequency-band granularity without backpropagation. Allocating bits by the logarithm of component variance is the classical solution of transform coding (Huang & Schultheiss, 1963; Gersho & Gray, 1992) and extends to Gaussian-mixture sources (Park et al., 2026). Reverse water-filling has recently been used to allocate bits across the key-value caches of language models (Zhang et al., 2026; Zuo et al., 2026), and a closed-form allocation from Hessian sensitivities across their weights (Zhang et al., 2025).

Polar representations. Quantizing a two-dimensional source by modulus and phase dates to Pearlman (1979) and Swaszek & Ku (1986). PolarQuant (Han et al., 2025) quantizes the key-value cache of a language model in polar coordinates after a random preconditioning that makes the angle distribution known, so that no per-block scale is stored. This work motivated the polar representations studied here. Trained spectral weights are not isotropic and their variance changes with $|k|$, so we store scales. Complex-valued networks have also been quantized jointly in their real and imaginary parts, with bit widths learned during training (Geng et al., 2026).

Clipping. Choosing the clipping range by minimizing the mean squared quantization error, rather than from the absolute maximum, is standard practice for real-valued networks (Banner et al., 2019; Nagel et al., 2021). We apply it to complex weights through their modulus.

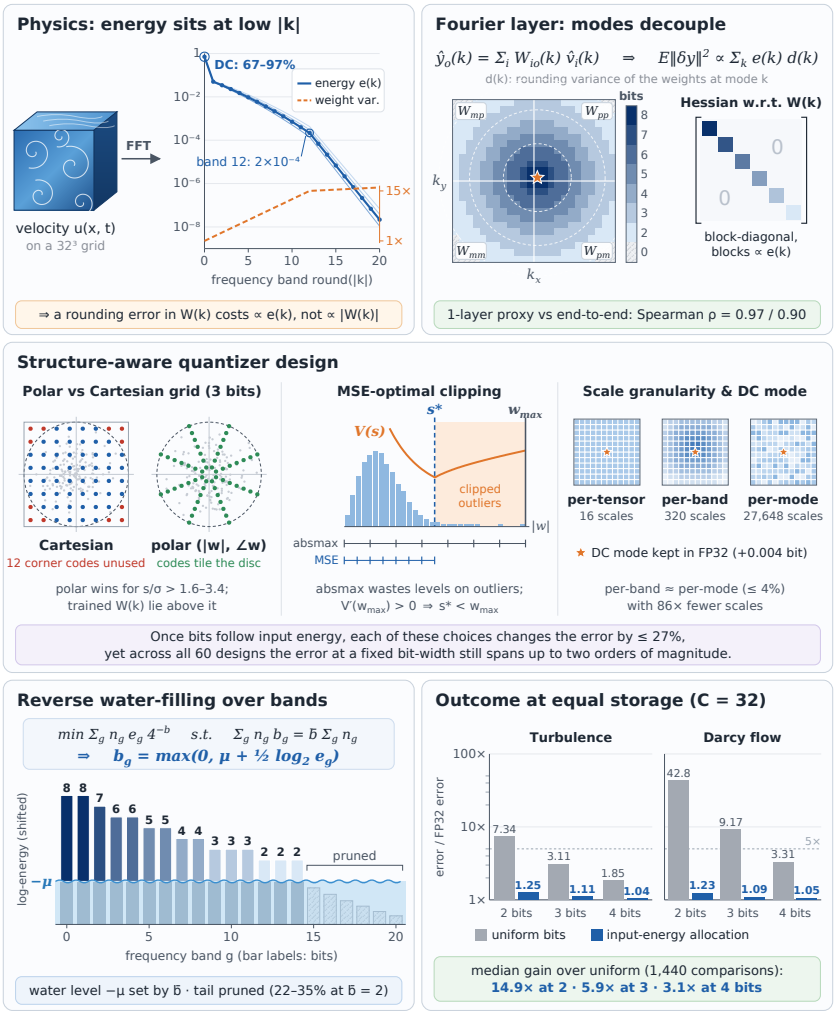


Figure 1: Overview of the approach. Top: input energy $e(k)$ concentrates at low $|k|$ while weight variance grows; since each weight acts on one mode, the layer error is weighted by $e(k)$. Middle: once bits are allocated, the other quantizer choices matter less. Bottom: reverse water-filling allocation and its error ratio at equal storage, $C = 32$ (Table 1). Only the bottom-right panel shows measured results; the rest are schematic.

3 Setup

Turbulence. We learn the solution map of the three-dimensional incompressible Navier–Stokes equations for forced isotropic turbulence on a periodic $[0, 2\pi]^3$ domain at 32^3, mapping a velocity field to the field 0.5 time units (0.168 eddy turnovers) later. Data are generated in float64 with a Fourier–Galerkin solver at $\nu = 0.025$ under constant-power forcing at $k_f = 2$, giving $\mathrm{Re}_\lambda \approx 28.3$. Splits are disjoint by trajectory, and results use the validation split of 384 pairs. Predicting no change gives a relative L^2 error of 0.3591. We train five seeds at channel width $C = 8$ and three at $C = 32$ for 300 epochs with AdamW, batch size 4, learning rate 10^{-3} with a cosine schedule and weight decay 10^{-4}. At seed 0 they reach 0.0557 and 0.0289.

Darcy flow. We learn the map $a \mapsto u$ for $-\nabla \cdot (a \nabla u) = f$ on the periodic $[0, 2\pi]^3$ domain at 32^3, with a fixed mean-zero forcing f and a log-normal coefficient $a = \exp(0.5\,\xi)$, where ξ is a Gaussian random field with spectrum $(|k|^2 + 9)^{-3/2}$. Solutions are computed by preconditioned conjugate gradients in float64, and splits of 8000, 1000 and 1000 samples are drawn from independent seeds. The coefficient is strictly positive, so the input has a large mean, whereas turbulent velocity has none. Two input normalizations are trained on the same data: *raw* divides by the training-split

standard deviation, and *centred* also subtracts the training-split mean. We train three seeds of each at $C \in \{8, 32\}$ (Appendix G), which reach 0.017 to 0.025 and 0.0071 to 0.0084, against 0.1712 for the training-split mean solution.

Model and storage. We use FNO3d with $m = 12$ modes per axis and four spectral layers, each followed by a pointwise bypass and GELU. A real-to-complex FFT stores only $k_z \geq 0$, so each layer needs one complex weight block per sign combination of (k_x, k_y), denoted W_{pp}, W_{mp}, W_{pm} and W_{mm}, and holds $4C^2m^3$ complex entries. Only W_{pp} contains the DC component $k = (0, 0, 0)$. The sixteen spectral blocks hold 99.947% of the parameters at $C = 8$ and 99.984% at $C = 32$, where they occupy $226.5\,\mathrm{MB}$ in FP32. Only the spectral blocks are quantized. The Darcy models also take coordinate channels as input, because the fixed forcing makes the Darcy map not translation-equivariant.

Default quantizer. The mean bit-width $\bar{b}$ counts bits per real component, so a complex weight costs $2\bar{b}$ bits on average. Quantization is uniform with round-to-nearest. Bands are formed within each weight block, 20 per block. In the default configuration, fixed in advance, all weights of a band share one FP32 scale (per-band granularity), the scale is chosen by MSE-optimal clipping, and the DC component is kept in FP32. Results are the median over three weight representations, defined in Section 7. Keeping the DC component in FP32 costs about 0.004 bits per parameter and favours the uniform baseline (Section 4).

Metrics. We report the relative L^2 error and the error ratio ρ, the quantized model's error divided by the same model's FP32 error, so that $\rho = 1$ means no loss. For visualization, we mark $\rho > 5$ as a high-error regime; this threshold is not intended as a usability criterion. Storage is counted in effective bits per parameter, charging the quantized payload, the FP32 remainder of the network, one FP32 scale per group and every entry kept in FP32 (Appendix A). Every allocation comparison is at equal storage. Evaluation corresponds to one step of the operator; because physical simulations demand structural fidelity, we also monitor the divergence error on turbulence (Appendix C, Table 5); reverse water-filling suppresses divergence blow-up ($1.42\times$ to $1.70\times$ vs. $2.05\times$ to $7.05\times$ for uniform at $C = 32$), correlating with relative L^2 error at $r = 0.892$. All results in the main text use the validation split. On Darcy flow a held-out test split was read once, under a pre-registration, for an allocation with a two-bit floor at three and four bits (Appendix G).

4 Sensitivity of Fourier modes to quantization

The primary objective of post-training bit allocation is to minimize the expected error of the operator at the final output field, $\mathbb{E}\|\delta u\|^2$, under a total bit budget. A first-order linear approximation treats this error as the sum of output perturbations from each layer.

Within a single spectral layer with input v and output y, both linear maps—the spectral product $\hat{y}_o(k) = \sum_i W_{io}(k)\,\hat{v}_i(k)$ and the $1 \times 1 \times 1$ pointwise bypass—act on each wavenumber independently. By Parseval's theorem, the spatial L^2 error equals the Fourier ℓ^2 error, so the layer's output error separates additively across modes: $\|\delta y\|^2 = \sum_k \|\delta\hat{y}(k)\|^2$. Modeling rounding errors as independent and zero-mean with variance $d(k)$ per weight entry at mode k, the expected layer error is then

$$\mathbb{E}\|\delta y\|^2 \;\propto\; \sum_k e(k)\,d(k), \qquad e(k) = \sum_i \mathbb{E}\big|\hat{v}_i(k)\big|^2, \tag{1}$$

where i indexes input channels and $e(k)$ is the input energy of mode k.[1] A weight error therefore costs in proportion to the input energy of its mode, not to the size of the weight. MSE-optimal clipping minimizes the weight error $\sum_k d(k)$ alone, which is Equation 1 with every input energy set equal.

Using Equation 1 for allocation assumes that the operator's output is equally sensitive to layer-output errors at every wavenumber. Across layers, the pointwise GELU nonlinearity couples modes, so

[1] The constant is $(2\pi)^3$ times the number of output channels. A stored mode with $k_z > 0$ also represents its conjugate and is counted twice by Parseval's theorem. The allocation omits this factor, which is at most 2 and shifts the bit-width of a band in Equation 2 by at most half a bit.

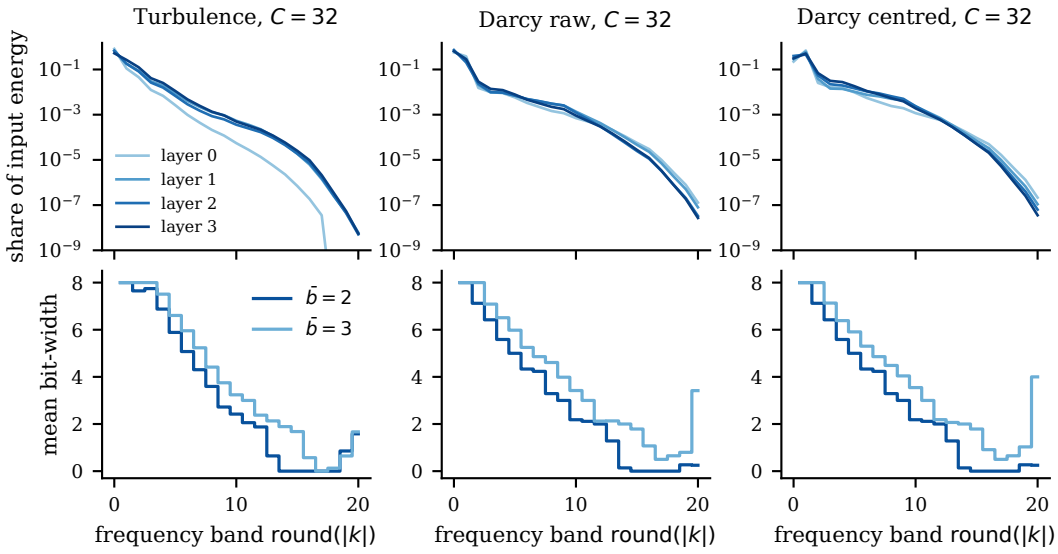


Figure 2: Top: share of the input spectral energy per frequency band entering each of the four spectral layers, measured on 128 training inputs, at $C = 32$ and seed 0. Band 0 is the DC component $k = (0, 0, 0)$. Bottom: the bit-width the allocation of Section 5 assigns to each band at $\bar{b} = 2$ and 3, averaged over the 16 blocks with weights equal to their entry counts. A bit-width of zero means the band is pruned. The rise in the outermost bands is the residual budget, placed in the bands with the fewest entries.

errors at one wavenumber reach other frequencies and can compound across layers. To test whether this approximation is sound, we perform a marginal sensitivity analysis: holding all other weights in FP32, we quantize one band of one layer at a time (Appendix D). The single-layer proxy predicts the end-to-end output damage with high rank correlation (Spearman $\rho = 0.97$ on turbulence and 0.90 on Darcy flow at $C = 32$, reaching 1.00 within individual layers). Downstream layers amplify errors most at the lowest wavenumbers, which the allocation already codes at maximum width (7–8 bits), while the final layer is essentially flat across modes. On turbulence, errors add nearly linearly across the network (within 7.4% of simultaneous quantization).

Measured input spectrum. We measure the input spectral energy with forward hooks that reproduce each layer's transform and block slices (Figure 2). On turbulence the DC component carries 0.667 to 0.973 of the input energy entering a spectral layer across eight checkpoints, and by band 12 the share is 2×10^{-4}, while the weight variance rises by roughly $15\times$ over the same bands. Turbulent velocity has zero mean, so this concentration is produced inside the network: with the lifting bias set to zero, the share at the first spectral layer is 10^{-17} at $C = 32$. On Darcy flow the data supply it. With the bias set to zero the share is 0.41 to 0.57 in the raw arm and below 5×10^{-6} in the centred arm.

Modes kept in full precision. We quantize without protection and then keep K modes per block in FP32, selected by input energy, by weight magnitude or at random, with the number of FP32 entries matched across criteria. Scales are calibrated before the selection, so it changes no scale. Keeping modes selected by input energy helps more than keeping the same number selected by weight magnitude in all 80 turbulence comparisons and in 466 of 480 on Darcy flow. On turbulence at $C = 32$, three bits and the polar representation, the energy-selected improvement rises from $1.033\times$ at $K = 1$ to $1.192\times$ at $K = 16$. Magnitude-selected modes give $1.0000\times$ at every K, and a random mode changes the error by at most 2×10^{-5}. The same ordering holds for bit allocation. At equal storage, allocating bits by the weight variance of each mode is worse than a uniform bit-width (median gain $0.81\times$ on turbulence at $C = 32$ and three bits), because weight variance rises with $|k|$ while input energy falls. Keeping the DC component in FP32, the mode with the largest input energy, lowers the turbulence error by $1.06\times$ to $1.27\times$ under uniform bit-widths; the allocation below assigns that mode 7 or 8 bits itself, and keeping it in FP32 then changes the error by under 0.5%.

5 INPUT-ENERGY BIT ALLOCATION

Let g index the frequency bands across blocks and layers, e_g be the mean of $e(k)$ over modes with $\mathrm{round}(|k|) = g$, and n_g be the number of weight entries in band g. We model the error variance of a weight entry at bit-width b as $D(b) = 4^{-b}$, treating pruned entries ($b = 0$) as contributing their full weight energy ($D(0) = 1$). Minimizing the total weighted distortion $\sum_g n_g e_g D(b_g)$ subject to a mean bit-width budget $\sum_g n_g b_g = \bar{b} \sum_g n_g$ yields the continuous reverse water-filling solution (Huang & Schultheiss, 1963):

$$b_g = \max\Big(0,\ \mu + \tfrac{1}{2}\log_2 e_g\Big), \tag{2}$$

where the water level μ is set by the budget. Any band with $e_g \le 4^{-\mu}$ is pruned ($b_g = 0$). Because the Fourier layer decouples modes, a single measured input spectrum determines the optimal bit allocation across all blocks simultaneously.

Integer solver and pruning. We solve the discrete allocation on $\{0\} \cup \{2, \dots, 8\}$ using a greedy solver: each increment is assigned to the band yielding the largest reduction in weighted distortion per bit until the budget is met (Appendix B). The input spectrum is measured with forward hooks on 32 unlabelled training inputs (taking under 5 seconds), and the greedy solver runs in under two seconds. At $\bar{b} = 2$, the allocation prunes 22% to 35% of the modes, while at $\bar{b} \in \{3, 4\}$ it prunes fewer than 11% and recovers at least 97% of the gain of an allocation with a two-bit floor. Thus, pruning operates primarily at the lowest bit-widths where budget constraints are tightest.

Alternative allocation rules. Frequency truncation is the simplest allocation that also prunes bands: it assigns zero bits to every band with $\mathrm{round}(|k|) \ge k_c$ and spreads the budget uniformly below the cut. We compared it with the allocation at $\bar{b} = 2$. Reverse water-filling allocates bits directly from the measured input spectrum, adapting to whatever distribution of energy the operator encounters. Because both benchmark problems have smooth solutions whose energy decays steeply with $|k|$, a static rule $(1 + |k|)^{-4}$ with no measurement recovers at least 90% of the measured allocation's gain in 317 of 320 comparisons. This only works for the specific PDEs considered, as they are smooth. Fixed rules depend critically on assuming the correct decay rate: setting $\alpha = 2$ or 1 forfeits 13% to 75% of the gain, collapsing to flat uniform precision at two bits (Appendix C). In physical systems with shallower energy cascades (such as high-Reynolds turbulence or active scalar transport) or multiscale permeability, a steep static rule prematurely starves active higher modes. Reverse water-filling avoids guessing an exponent and adapts automatically to the spectrum; isolating its advantage on non-smooth or non-monotone flows is a natural direction for future benchmarks.

6 RESULTS

6.1 INPUT-ENERGY ALLOCATION AGAINST UNIFORM BIT-WIDTHS

The primary comparison uses a fixed default quantizer configuration; no quantizer design is selected using the reported error. In this setting, input-energy allocation reduces the error in every configuration tested.

We then test whether this effect persists across the remaining quantizer choices. Across 1,440 comparisons, covering 20 checkpoints, three bit-widths and 24 configurations of the other choices, the allocation lowers the error by a median of 5.2× and by at least 1.6×. The gain is largest when bits are scarcest, with medians of 14.9× at two bits, 5.9× at three and 3.1× at four.

In the default configuration (Table 1, Figure 3), the larger models stay within 1.15 times the FP32 error at three and four bits on both problems. At two bits, the models are about 16× smaller than FP32, while their error is 1.18 to 1.34 times that of the FP32 model. A uniform bit-width at the same storage gives 5.69 to 8.21 times the FP32 error on turbulence and 33.4 to 84.5 times on Darcy flow.

Comparison with the best uniform design. Table 2 compares the best design of each kind, selected on the validation split, so that the other choices can adapt to a uniform bit-width. At three and four bits the best allocated design uses per-mode scales, MSE-optimal clipping, the DC component

Table 1: Error ratio to FP32 with uniform bit-widths and input-energy allocation at equal storage, in the default configuration: per-band scales, MSE-optimal clipping, DC component in FP32. Medians, with ranges over the three weight representations and all seeds (five for turbulence at $C = 8$, three for turbulence at $C = 32$, three in each of the two Darcy normalizations). Effective storage at $\bar{b} = 2$ is 2.008 to 2.021 bits per parameter. Validation split.

	Allocation	$\bar{b} = 2$	$\bar{b} = 3$	$\bar{b} = 4$
Turbulence, $C = 8$	uniform	12.3 (9.64–15.6)	7.50 (6.47–9.23)	4.21 (3.36–5.20)
	input energy	**3.49** (3.24–4.05)	**2.11** (1.74–2.55)	**1.45** (1.24–1.66)
Turbulence, $C = 32$	uniform	7.34 (5.69–8.21)	3.11 (2.76–3.90)	1.85 (1.64–2.20)
	input energy	**1.25** (1.18–1.32)	**1.11** (1.07–1.14)	**1.04** (1.02–1.05)
Darcy, $C = 8$	uniform	15.8 (11.1–25.4)	5.80 (4.97–8.27)	2.89 (2.55–3.83)
	input energy	**1.12** (1.07–1.17)	**1.07** (1.03–1.10)	**1.03** (1.02–1.06)
Darcy, $C = 32$	uniform	42.8 (33.4–84.5)	9.17 (6.84–20.4)	3.31 (2.48–10.1)
	input energy	**1.23** (1.19–1.34)	**1.09** (1.06–1.15)	**1.05** (1.02–1.09)

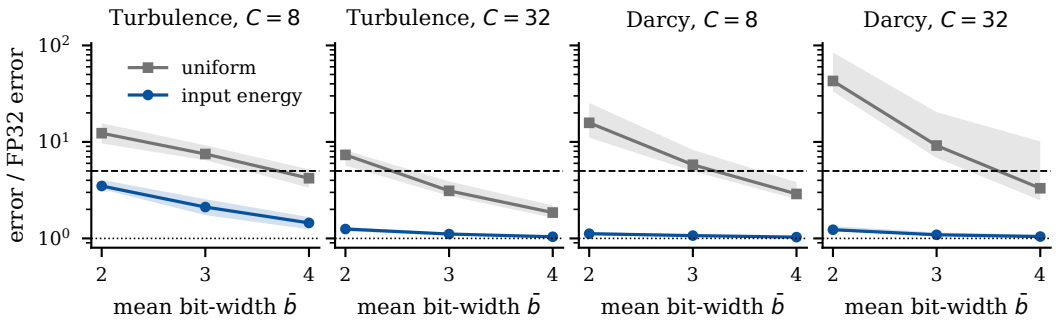


Figure 3: Error ratio to FP32 against mean bit-width with uniform bit-widths and input-energy allocation at equal storage, in the default configuration. Lines are medians and bands the range over weight representations and seeds. The dashed line marks an error five times that of the FP32 model.

in FP32 and the polar representation in 7 of 8 groups; at two bits the best design differs between groups. At $\bar{b} = 2$ the best allocated design improves on the best uniform one by $3.3\times$ and $3.8\times$ on turbulence and by $10\times$ and $13\times$ on Darcy flow. At $\bar{b} = 4$ the margin is $1.6\times$ to $2.6\times$.

Spread across designs. Design choices still matter in absolute terms. At a fixed bit-width and model, the worst of the 36 uniform designs screened has 20 to 205 times the error of the best allocated design, and the conventional design (uniform bit-width, per-tensor scales, absmax scaling, Cartesian representation, DC component quantized) has 26 to 124 times the FP32 error at two and three bits.

Table 2: Best uniform design (of 36) against the best allocated design (of 24), by median error ratio over seeds. The two designs of a pair can differ in the storage of their scales and DC component: per-mode scales add 0.25 bits per component at $C = 8$ and 0.016 at $C = 32$, and an FP32 DC component about 0.004. The allocated design never stores more than 0.02 bits per component above the uniform design it is compared with. The selection uses the validation split that also scores it.

	$\bar{b} = 2$			$\bar{b} = 3$			$\bar{b} = 4$		
	uniform	energy	gain	uniform	energy	gain	uniform	energy	gain
Turbulence, $C = 8$	10.9	3.34	3.3×	6.62	1.92	3.4×	3.39	1.28	2.6×
Turbulence, $C = 32$	4.58	1.19	3.8×	2.91	1.07	2.7×	1.64	1.02	1.6×
Darcy, $C = 8$	10.9	1.10	10.0×	4.49	1.04	4.3×	2.38	1.02	2.3×
Darcy, $C = 32$	15.9	1.20	13.2×	6.19	1.06	5.8×	2.70	1.02	2.6×

7 OTHER QUANTIZER DESIGN CHOICES

If input energy is the main source of sensitivity, the remaining choices should matter less once the bits follow it. Table 3 tests this by changing one choice at a time. None of the three changes the error by more than 27% (Figure 4), whereas returning to a uniform bit-width multiplies it by 3.6 to 40 at two bits and 2.8 to 7.5 at three.

Clipping. Absmax scaling sets each scale to the largest modulus in the group. MSE-optimal clipping searches multiples of that value and keeps the one with the smallest mean squared quantization error at the operative bit-width. Absmax spends most quantization levels on a few outlying weights, and under a clipping model the continuous MSE optimum of every group lies below its largest modulus (Proposition 2). On turbulence MSE-optimal clipping lowers the error by 5% to 26% at two and three bits, most at two bits and at $C = 8$, and by about 2% at four bits. On Darcy flow it changes the error by under 2%.

Weight representation. The Cartesian representation quantizes $\mathrm{Re}(w)$ and $\mathrm{Im}(w)$ on independent uniform grids, which tile a square. The polar representation quantizes modulus and phase separately, with $2^b - 1$ magnitude levels and 2^b phase bins, which tile a disc, and is charged for all 4^b codes. A magnitude-dependent polar variant varies the number of phase levels with the modulus so that $r^2\Delta\theta^2$ is constant. A square grid spends levels on the corners of the complex plane, which a weight of bounded modulus cannot occupy. For circular Gaussian weights with scale s and per-component standard deviation σ, polar has lower distortion when s/σ exceeds a threshold of 1.6 to 3.4, depending on the bit-width and the level convention (Proposition 1), and trained spectral weights at $C = 32$ lie above it. With the allocation in place polar lowers the error by 2% to 9% on turbulence and by under 3% on Darcy flow. Under uniform bit-widths the margin is larger, $1.36\times$ to $1.82\times$ on turbulence, and most of it comes from designs with absmax scaling, so the polar representation and MSE-optimal clipping remove the same loss.

Quantization granularity. A scale can be shared across a whole weight block (per-tensor), across one frequency band (per-band) or across one mode (per-mode), which gives 16, 320 and $27{,}648$ stored scales; a scale always covers all C^2 entries at a mode. Per-band scales store 86 times fewer scales than per-mode scales. With the allocation in place the two differ by at most 4.0% on turbulence and 2% on Darcy flow. Per-tensor scales leave no groups to allocate across; under uniform bit-widths they can erase the bands at small $|k|$, because at $C = 8$ the block-wide maximum modulus is 11.71 against 0.66 for the largest weight in band 0.

Table 3: Leave-one-out over the allocation and the three other design choices. Columns A, C, P and G mark input-energy allocation, MSE-optimal clipping, polar representation and per-mode granularity as on (•) or off (○); each row replaces one choice by the alternative named. The DC component is in FP32 throughout. The last row is the conventional design: uniform bit-width, per-tensor scales, absmax scaling, Cartesian representation, DC component quantized. Median error ratio over seeds for turbulence (T) and Darcy flow (D) at $C = 8$ and 32. Validation split. The full factorial is in Appendix E.

	A	C	P	G	$\bar{b} = 2$				$\bar{b} = 3$			
					T8	T32	D8	D32	T8	T32	D8	D32
all on	•	•	•	•	3.45	1.21	1.10	1.21	1.92	1.07	1.04	1.06
uniform bit-width	○	•	•	•	12.5	5.78	20.1	48.9	6.62	3.00	5.82	7.94
absmax scaling	•	○	•	•	4.36	1.41	1.11	1.23	2.17	1.12	1.04	1.07
Cartesian	•	•	○	•	3.72	1.28	1.10	1.24	2.09	1.11	1.04	1.09
per-band scales	•	•	•	○	3.44	1.19	1.10	1.20	2.00	1.07	1.05	1.07
conventional design	○	○	○	○	26.1	28.1	44.4	124	26.5	28.0	39.9	124

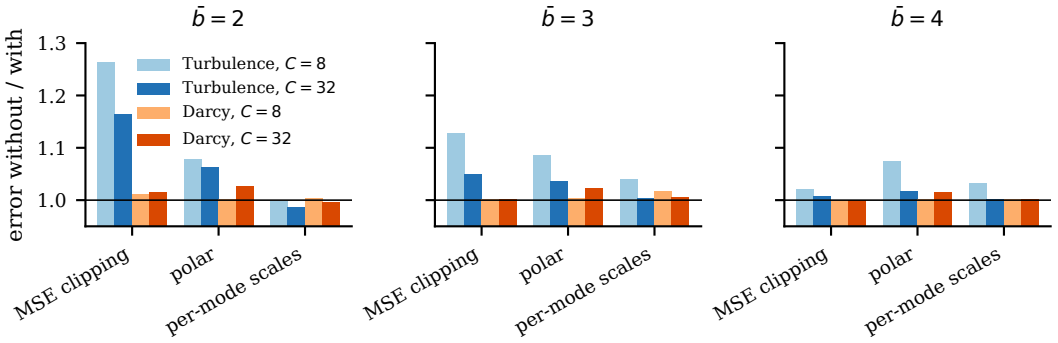


Figure 4: Effect of each other design choice with the allocation in place: error ratio without the choice divided by the error ratio with it, median over seeds. Replacing the allocation by a uniform bit-width gives ratios of 1.6 to 40 over the three bit-widths and is not shown. Validation split.

8 Limitations

The smallest model. At two bits on turbulence at $C = 8$ the default configuration has 3.24 to 4.05 times the FP32 error, and frequency truncation performs better than reverse water-filling allocation. (Appendix C, Table 6). This limitation traces directly to the single-layer decoupling assumption: on this tiny model, cross-coupled quantization errors in the middle bands compounded across downstream layers rather than remaining independent (Section 4, Appendix D).

Mechanism. Equation 1 is a one-layer cost model. The marginal analysis (Section 4, Appendix D) shows it predicts end-to-end damage with high rank correlation ($\rho \geq 0.90$) on larger models, with downstream layers amplifying middle bands by $\approx 4\times$ on turbulence $C = 8$. The allocation treats weight energy per entry as equal across bands, though weight variance rises by about $15\times$ over the first 12 bands. On Darcy flow ($C = 32$), raw FP32 changes error by 0.97–1.02×, which the cost model does not predict.

Evidence scope. Only the allocation with a two-bit floor at three and four bits was tested on the held-out Darcy split (Appendix G); all two-bit and pruned-band results, and all turbulence results, use the validation split, and the best designs of Table 2 are selected on the split that scores them.

Scope and dynamical stability. The study covers two PDEs at 32^3, weight-only quantization of spectral blocks, and single-step maps $u_t \mapsto u_{t+\Delta t}$. In multi-step rollouts, errors compound over time; uniform quantization introduces severe low-frequency distortion ($\rho_g \approx 38\times$ in band 1), while reverse water-filling preserves uniform relative fidelity ($\rho_g \approx 1.1\times$) and suppresses divergence blow-up ($r = 0.892$, Table 5). Long-horizon rollouts remain future work. At higher spatial resolutions ($64^3, 128^3$), additional modes fall in the steep dissipation tail where energy decays rapidly ($e(k) \to 0$), so reverse water-filling will prune or assign minimal bits to an even larger fraction of modes, amplifying compression gains. Storage is counted analytically and for numerical accuracy.

9 Conclusion

Post-training quantization of a Fourier neural operator depends mainly on how the bits are allocated across frequency bands. Fourier modes differ widely in their sensitivity to quantization, and the input energy of a mode predicts the cost of perturbing its weights. Allocating bits by reverse water-filling on the measured spectrum, with pruned bands, reduces the error in every configuration tested, by a median of $5.2\times$ over uniform bit-widths at equal storage. On the larger models it keeps the error within 1.35 times that of the FP32 model at two bits and within 1.15 times at three and four bits, while on the smallest turbulence model, omitting high-frequency modes eliminates low-bit quantization noise. With the bits allocated, weight representation, quantization granularity and clipping change the error by at most 27%, although across all designs the error at a fixed bit-width spans up to two orders of magnitude.

REPRODUCIBILITY STATEMENT

Code, evaluation scripts, model checkpoints, and datasets are provided in the supplementary material to ensure full replicability. Both physical benchmarks (3D Navier–Stokes and Darcy flow) use deterministic dataset generation from fixed seeds. Mathematical proofs (Propositions 2, 1) are in Appendix F; solver details and bit-exact storage accounting are formalized in Appendices A and B.

ETHICS STATEMENT

This work studies numerical quantization for neural operators in physical simulations. It does not involve human participants, personal identifiable data, or sensitive demographic information.

AI USE STATEMENT

We used generative AI tools (LLM coding and writing assistants) to implement experiment and evaluation code, to process logs into tables and figures, and to edit the manuscript for readability. We did not use them to propose the method, formulate hypotheses or design the experiments. The authors reviewed all AI-assisted code and text, and take full responsibility for the content.

REFERENCES

Ron Banner, Yury Nahshan, and Daniel Soudry. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/c0a62e133894cdce435bcb4a5df1db2d-Abstract.html.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian AWare quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 293–302, 2019.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.17323.

Zihan Geng, Zhilin Li, Mi Zhou, Zijie Chen, Ganzhangqin Yuan, Jaebum Noh, Seokho Lee, Cherry Park, Hongya Geng, Xiu Li, Kui Jiang, Mu Ku Chen, and Junsuk Rho. Ultra-efficient physical field computing by complex-valued network quantization. *Nature Communications*, 17:3762, 2026. doi: 10.1038/s41467-026-70319-0.

Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

Insu Han, Praneeth Kacham, Amin Karbasi, Vahab Mirrokni, and Amir Zandieh. PolarQuant: Quantizing KV caches with polar transformation. *arXiv preprint arXiv:2502.02617*, 2025. URL https://arxiv.org/abs/2502.02617.

Babak Hassibi and David G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In *Advances in Neural Information Processing Systems 5*, pp. 164–171, 1993.

J. J. Y. Huang and Peter M. Schultheiss. Block quantization of correlated Gaussian random variables. *IEEE Transactions on Communications Systems*, 11(3):289–296, 1963.

Jean Kossaifi, Nikola Kovachki, Kamyar Azizzadenesheli, and Anima Anandkumar. Multi-grid tensorized Fourier neural operator for high-resolution PDEs. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2310.00120.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research (JMLR)*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.08895.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Jinming Lu, Jiayi Tian, Yequan Zhao, Hai Li, and Zheng Zhang. Tensor-compressed and fully-quantized training of neural pde solvers on edge devices. In *Design, Automation and Test in Europe Conference (DATE)*, 2026. URL https://arxiv.org/abs/2512.09202.

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021. URL https://arxiv.org/abs/2106.08295.

Bumsu Park, Chanho Park, Youngmok Park, and Namyoon Lee. PrismQuant: Rate-distortion-optimal vector quantization for Gaussian-mixture sources. *arXiv preprint arXiv:2605.15507*, 2026. URL https://arxiv.org/abs/2605.15507.

William A. Pearlman. Polar quantization of a complex Gaussian random variable. *IEEE Transactions on Communications*, 27(6):892–899, 1979. doi: 10.1109/TCOM.1979.1094476.

Peter F. Swaszek and Tsu W. Ku. Asymptotic performance of unrestricted polar quantizers. *IEEE Transactions on Information Theory*, 32(2):330–333, 1986. doi: 10.1109/TIT.1986.1057141.

Renbo Tu, Colin White, Jean Kossaifi, Boris Bonev, Nikola Kovachki, Gennady Pekhimenko, Kamyar Azizzadenesheli, and Anima Anandkumar. Guaranteed approximation bounds for mixed-precision neural operators. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2307.15034.

Winfried van den Dool, Tijmen Blankevoort, Max Welling, and Yuki M. Asano. Efficient neural PDE-solvers using quantization aware training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023. URL https://arxiv.org/abs/2308.07350.

Chao Zhang, Li Wang, Samson Lasaulce, and Merouane Debbah. BAQ: Efficient bit allocation quantization for large language models. *arXiv preprint arXiv:2506.05664*, 2025. URL https://arxiv.org/abs/2506.05664.

Junkai Zhang, Hang Guo, Luca Benini, and Yawei Li. RDKV: Rate-distortion bit allocation for joint eviction and quantization of the KV cache. *arXiv preprint arXiv:2605.08317*, 2026. URL https://arxiv.org/abs/2605.08317.

Fei Zuo, Zikang Zhou, Hao Cong, Xiaoyan Xi, and Ho Fai Leung. RateQuant: Optimal mixed-precision KV cache quantization via rate-distortion theory. *arXiv preprint arXiv:2605.06675*, 2026. URL https://arxiv.org/abs/2605.06675.

A STORAGE ACCOUNTING

Effective bits per parameter charge the quantized payload, the FP32 remainder of the network, one FP32 scale per group and every entry kept in FP32. A per-mode scale costs $16/C^2$ bits per component, 0.25 at $C = 8$ and 0.0156 at $C = 32$ (Appendix F). An FP32 DC component stores $8C^2$ real entries over the four layers, out of $55{,}296\,C^2$ spectral parameters, and costs $(32 - \bar{b})/6912$ bits per parameter net of the code it replaces, about 0.004. The allocated model has the same code-segment size, scales and FP32 entries as the uniform one, and the solver asserts the match before any result is recorded. The bit-width table, one entry per group, is not charged; at per-band granularity it adds 0.007% to the packed artifact. Pruned bands keep their scales, which is conservative. We pack uniform models, and allocated models with a two-bit floor, into a binary artifact and confirm that it matches the counted total component by component, to the byte, and reconstructs the evaluated weights. At $C = 32$ and three bits the artifact occupies 21.30 MB against 226.53 MB in FP32.

B ALLOCATION SOLVER

The greedy solver keeps a priority queue of groups keyed by the cost reduction per bit of their next increment, $n_g e_g[D(b_g) - D(b'_g)]/[n_g(b'_g - b_g)]$, and pops the best group while its increment fits the remaining budget and the 8-bit ceiling. Per bit and per unit of e_g, the reduction is $15/32$ for the first increment from zero and $\frac{3}{4}\,4^{-b}$ from $b \geq 2$, so it decreases along every band's sequence of increments. The residual budget that no increment fits is closed on coded groups by placing it in the groups with the fewest entries, the two outermost corner bands of each block. Frequency truncation at cut k_c uses the same code: every band with $\mathrm{round}(|k|) \geq k_c$ receives zero bits and the budget is spread uniformly below the cut.

C ALLOCATION RULES

Table 4 and Figure 5 compare allocation rules at equal storage in the default configuration, on three turbulence seeds at $C = 8$, three at $C = 32$ and the six Darcy checkpoints at $C = 32$. The first row is the gain of the input-energy allocation with a two-bit floor over a uniform bit-width. Each other row is a rule's gain divided by that gain, as the median over checkpoints of the per-checkpoint ratio, so 1 means equal to the allocation with a two-bit floor. Frequency truncation assigns zero bits to bands with $\mathrm{round}(|k|) \geq k_c$ and at least two bits to every band below the cut, with the best $k_c \in \{12, \ldots, 20\}$ chosen per checkpoint on the validation split; this set was fixed in the pre-registration of these controls, before the two-bit comparison of Table 6 fixed its own. At three bits the best cut beats the allocation with a two-bit floor on turbulence at $C = 8$ by $1.34\times$ and matches it at $C = 32$ (1.01); in the other six comparisons at three and four bits it reaches 0.75 to 0.91 of that allocation's gain. The random-input spectrum is measured with independent standard Gaussian input fields in place of data; it lowers the error of a uniform bit-width by $1.06\times$ on turbulence at $C = 32$ and three bits, against $2.67\times$ for the measured spectrum, so the concentration comes from the data and the trained network rather than the architecture alone. The weight variance of a mode is taken as $\sum_{i,o} |W_{io}(k)|^2$; a constant factor does not change the allocation. Reweighting the clipping search by the measured energy, instead of reallocating bits, changes the error by at most 4.4% at per-band granularity. The fixed rules were fixed before Darcy flow was measured.

Table 4: Allocation rules against the input-energy allocation with a two-bit floor at equal storage. T: turbulence; D: Darcy flow. Validation split.

	$\bar{b} = 3$				$\bar{b} = 4$			
	T8	T32	D32 raw	D32 centred	T8	T32	D32 raw	D32 centred
input energy, 2-bit floor: gain over uniform	$3.54\times$	$2.67\times$	$8.13\times$	$8.27\times$	$2.90\times$	$1.77\times$	$2.94\times$	$3.05\times$
input energy with pruned bands (ours)	0.99	1.03	1.02	1.01	1.00	1.00	1.00	1.00
fixed, $(1 + \lvert k\rvert)^{-4}$	1.07	0.96	0.98	0.99	1.03	0.98	0.99	0.99
fixed, $(1 + \lvert k\rvert)^{-2}$	0.66	0.82	0.78	0.84	0.74	0.89	0.94	0.94
fixed, $(1 + \lvert k\rvert)^{-1}$	0.39	0.57	0.37	0.32	0.47	0.74	0.70	0.75
best frequency truncation	1.34	1.01	0.82	0.87	0.85	0.91	0.75	0.79
random-input spectrum	0.28	0.40	0.47	0.48	0.34	0.58	0.84	0.84
weight variance	0.23	0.27	0.16	0.12	0.25	0.48	0.38	0.33

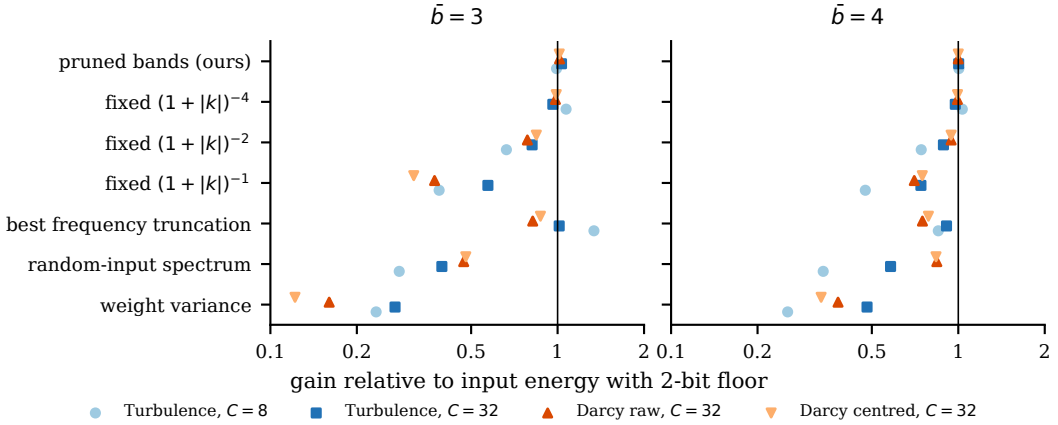


Figure 5: The rules of Table 4. Values to the right of 1 are better than the input-energy allocation with a two-bit floor.

Table 5: Divergence error ratio to FP32 on forced isotropic turbulence for uniform bit-width, reverse water-filling and frequency truncation + INT8. Medians over seeds and representations, with min–max ranges. Pearson correlation with relative L^2 error ratio across all configurations: $r = 0.892$.

Model	$\bar{b}$	Uniform	Water-filling	Truncation
Turbulence, $C = 8$	2	7.05 (6.50–8.50)	4.66 (4.05–5.89)	1.00 (0.97–1.04)
Turbulence, $C = 8$	3	4.79 (4.16–6.13)	3.50 (2.89–4.23)	0.97 (0.95–1.02)
Turbulence, $C = 8$	4	2.94 (2.32–3.99)	2.06 (1.72–2.71)	0.95 (0.93–1.00)
Turbulence, $C = 32$	2	3.02 (2.67–3.84)	1.70 (1.59–2.05)	1.27 (1.23–1.29)
Turbulence, $C = 32$	3	2.05 (1.93–2.27)	1.42 (1.36–1.62)	1.15 (1.13–1.17)
Turbulence, $C = 32$	4	1.49 (1.32–1.59)	1.16 (1.11–1.20)	1.08 (1.07–1.09)

D MARGINAL SENSITIVITY ANALYSIS

To evaluate whether the single-layer cost proxy of Equation 1 predicts end-to-end performance in the multi-layer operator, we measure the marginal sensitivity of each layer-frequency band (l, g) across the validation split on three representative checkpoints: turbulence $C = 8$ (seed 0), turbulence $C = 32$ (seed 0), and Darcy flow $C = 32$ (seed 0, raw). Holding all other weights at FP32, we quantize only band $g \in \{1, \ldots, 20\}$ of layer $l \in \{0, 1, 2, 3\}$ at 3 bits in the default polar configuration (64-bit precision accumulation, TF32 disabled).

Table 7 reports the Spearman rank correlation of the single-layer proxy $P(l, g) = \sum_{k \in g} (e(k)/C_{\text{in}}) \|\delta W(k)\|_F^2$ against the measured layer-output perturbation $\Delta_{\text{layer}}(l, g)$ and against the final network output error $\Delta_{\text{final}}(l, g) = \|u_{\text{quant}} - u_{\text{FP32}}\|^2$. Within every individual layer, the proxy ranks frequency bands with near-perfect fidelity ($\rho \geq 0.97$ across all models). Across the entire network, pooled Spearman correlations with end-to-end damage are 0.97 on turbulence $C = 32$, 0.90 on Darcy flow, and 0.86 on turbulence $C = 8$.

Downstream gain and additivity. We compute the downstream gain $G(l, g) = \Delta_{\text{final}}/\Delta_{\text{layer}}$, normalized by each layer's median. In the final layer ($l = 3$), G is flat across all wavenumbers (0.81–1.40). In earlier layers ($l \in \{0, 1, 2\}$), G decreases with $|k|$, meaning subsequent layers amplify low-frequency errors more than high-frequency errors. Because reverse water-filling already assigns maximum bit-widths (7–8 bits) to these lowest modes, this non-uniformity aligns with the energy allocation. On turbulence $C = 8$, inner layers display a gain peak at bands 5–7 ($G \approx 4.5$), indicating that the single-layer proxy under-allocates bits to these middle modes by roughly one bit; this accounts for why frequency truncation, which retains those bands at 8 bits, outperforms water-filling on that specific model. Finally, comparing $\sum_{l,g} \Delta_{\text{final}}(l, g)$ to the error when all bands and layers are simultaneously quantized at 3 bits yields an additivity ratio of 1.074 on turbulence $C = 32$ and 1.057 on $C = 8$, verifying that cross-layer error interaction is minimal.

Table 6: Frequency truncation + INT8 at best cuts against reverse water-filling. Top: summary by model group and bit-width (medians, with min–max in parentheses). Bottom: per-checkpoint best cuts and median damage over representations. Validation split.

(a) Summary by model group

Model group	$\bar{b}$	Best cut k_c	Truncation error ratio	Trunc. eff. bits	Water-filling error ratio	WF eff. bits
Turbulence, $C = 8$	2	$k_c = 9$	1.17 (1.10–1.18)	1.623	3.49 (3.24–4.05)	2.016
	3	$k_c = 10$	1.06 (1.02–1.08)	2.323	2.11 (1.74–2.55)	3.016
	4	$k_c = 11$	1.02 (1.00–1.04)	3.079	1.45 (1.24–1.66)	4.016
Turbulence, $C = 32$	2	$k_c = 10$	1.34 (1.29–1.36)	1.997	1.25 (1.18–1.32)	2.008
	3	$k_c = 11$	1.14 (1.10–1.14)	2.920	1.11 (1.07–1.14)	3.008
	4	$k_c = 12$	1.04 (1.03–1.04)	3.837	1.04 (1.02–1.05)	4.008
Darcy, $C = 8$	2	$k_c = 9$	1.19 (1.17–1.42)	1.628	1.12 (1.07–1.17)	2.021
	3	$k_c = 10 - -12$	1.11 (1.10–1.23)	2.328	1.07 (1.03–1.10)	3.020
	4	$k_c = 11 - -12$	1.08 (1.06–1.09)	3.467	1.03 (1.02–1.06)	4.020
Darcy, $C = 32$	2	$k_c = 10 - -11$	1.80 (1.68–2.08)	1.998	1.23 (1.19–1.34)	2.009
	3	$k_c = 12$	1.38 (1.31–1.44)	3.009	1.09 (1.06–1.15)	3.009
	4	$k_c = 12$	1.23 (1.21–1.28)	3.838	1.05 (1.02–1.09)	4.009

(b) Per-checkpoint best cuts and error ratios

Checkpoint	$\bar{b} = 2$	$\bar{b} = 3$	$\bar{b} = 4$
turb_c8_s0	$k_c = 9$, 1.17	$k_c = 10$, 1.06	$k_c = 11$, 1.02
turb_c8_s1	$k_c = 9$, 1.12	$k_c = 10$, 1.04	$k_c = 11$, 1.01
turb_c8_s2	$k_c = 9$, 1.18	$k_c = 10$, 1.08	$k_c = 11$, 1.04
turb_c8_s3	$k_c = 9$, 1.10	$k_c = 10$, 1.02	$k_c = 11$, 1.00
turb_c8_s4	$k_c = 9$, 1.17	$k_c = 10$, 1.08	$k_c = 11$, 1.04
turb_c32_s0	$k_c = 10$, 1.36	$k_c = 11$, 1.14	$k_c = 12$, 1.04
turb_c32_s1	$k_c = 10$, 1.29	$k_c = 11$, 1.10	$k_c = 12$, 1.03
turb_c32_s2	$k_c = 10$, 1.34	$k_c = 11$, 1.14	$k_c = 12$, 1.04
darcy_raw_w8_s0	$k_c = 9$, 1.19	$k_c = 10$, 1.12	$k_c = 11$, 1.08
darcy_raw_w8_s1	$k_c = 9$, 1.42	$k_c = 12$, 1.23	$k_c = 12$, 1.09
darcy_raw_w8_s2	$k_c = 9$, 1.17	$k_c = 10$, 1.10	$k_c = 12$, 1.06
darcy_centred_w8_s0	$k_c = 9$, 1.19	$k_c = 10$, 1.11	$k_c = 11$, 1.07
darcy_centred_w8_s1	$k_c = 9$, 1.29	$k_c = 10$, 1.16	$k_c = 12$, 1.08
darcy_centred_w8_s2	$k_c = 9$, 1.19	$k_c = 10$, 1.11	$k_c = 11$, 1.07
darcy_raw_w32_s0	$k_c = 10$, 1.68	$k_c = 12$, 1.32	$k_c = 12$, 1.21
darcy_raw_w32_s1	$k_c = 10$, 1.75	$k_c = 12$, 1.34	$k_c = 12$, 1.22
darcy_raw_w32_s2	$k_c = 10$, 2.08	$k_c = 12$, 1.41	$k_c = 12$, 1.28
darcy_centred_w32_s0	$k_c = 10$, 1.81	$k_c = 12$, 1.44	$k_c = 12$, 1.23
darcy_centred_w32_s1	$k_c = 10$, 1.80	$k_c = 12$, 1.31	$k_c = 12$, 1.21
darcy_centred_w32_s2	$k_c = 11$, 2.03	$k_c = 12$, 1.44	$k_c = 12$, 1.26

Table 7: Marginal sensitivity rank correlations (Spearman ρ) of the Equation 1 proxy $P(l, g)$ against local layer perturbation Δ_{layer} and final output damage Δ_{final}, evaluated on validation splits.

Model	P vs. Δ_{layer}	P vs. Δ_{final}				
	Pooled	Pooled	Layer 0	Layer 1	Layer 2	Layer 3
Turbulence, $C = 8$	0.997	0.86	1.00	0.97	0.71	1.00
Turbulence, $C = 32$	0.980	0.97	1.00	1.00	1.00	1.00
Darcy flow, $C = 32$	0.921	0.90	1.00	1.00	1.00	1.00

E Design screen and full factorial

The uniform screen crosses the three weight representations, the three granularities, the two clipping rules and the two DC treatments, 36 designs, at $\bar{b} \in \{2, 3, 4, 8\}$. The allocation is crossed with the 24 per-band and per-mode designs at $\bar{b} \in \{2, 3, 4\}$. Tables 8 and 9 give every combination of the

binary choices at per-band and per-mode granularity, with the DC component in FP32. Columns are as in Table 3.

Table 8: Full factorial at per-band granularity. Median error ratio over seeds, validation split.

A	C	P	T8	T32	D8	D32
$\bar{b} = 2$						
●	●	●	3.44	1.19	1.10	1.20
●	●	○	3.58	1.30	1.12	1.23
●	○	●	4.76	1.47	1.14	1.28
●	○	○	5.41	1.65	1.33	1.80
○	●	●	12.7	5.75	19.2	42.6
○	●	○	13.3	7.34	13.6	65.0
○	○	●	19.5	8.23	18.7	103
○	○	○	17.0	26.1	43.6	124
$\bar{b} = 3$						
●	●	●	2.00	1.07	1.05	1.07
●	●	○	2.17	1.12	1.07	1.09
●	○	●	2.61	1.20	1.08	1.10
●	○	○	3.17	1.42	1.19	1.34
○	●	●	6.94	2.91	5.43	8.25
○	●	○	7.81	3.11	5.75	8.39
○	○	●	7.99	3.65	5.39	16.5
○	○	○	16.4	8.06	13.4	113
$\bar{b} = 4$						
●	●	●	1.32	1.02	1.02	1.02
●	●	○	1.45	1.04	1.03	1.05
●	○	●	1.47	1.05	1.02	1.03
●	○	○	2.28	1.19	1.06	1.12
○	●	●	3.50	1.65	2.66	2.97
○	●	○	4.28	1.85	2.90	3.05
○	○	●	3.75	1.84	2.59	3.97
○	○	○	7.00	3.41	4.14	29.6

F THEORY

This appendix states the results used in the main text.

Proposition 1 (Polar and Cartesian distortion)**.** *For a circular complex Gaussian weight with per-component standard deviation σ, quantized with scale s, in the high-resolution regime with negligible clipping, b bits per component and 2^b magnitude levels, $\varepsilon_{\mathrm{cart}} = \frac{2}{3}s^2 4^{-b}$ and $\varepsilon_{\mathrm{polar}} = (\frac{s^2}{12} + \frac{2\pi^2\sigma^2}{3})4^{-b}$. Polar has lower distortion if and only if $s/\sigma > \sqrt{8\pi^2/7} \approx 3.3585$.*

Proof. Each Cartesian component has step $2s/2^b$ and error $s^2 4^{-b}/3$. The modulus has step $s/2^b$ and error $s^2 4^{-b}/12$; the phase has step $2\pi/2^b$ and angular error $\pi^2 4^{-b}/3$, which contributes $\mathbb{E}[r^2]\,\pi^2 4^{-b}/3$ with $\mathbb{E}[r^2] = 2\sigma^2$. Then $\varepsilon_{\mathrm{polar}} < \varepsilon_{\mathrm{cart}}$ is equivalent to $\frac{2\pi^2}{3}\sigma^2 < \frac{7}{12}s^2$. □

The quantizer uses $2^b - 1$ magnitude levels. Measured by bisection on 4×10^5 samples with that convention, the crossover is 1.61, 2.51 and 2.97 at $b = 2$, 3 and 4, against 3.31, 3.36 and 3.37 with 2^b levels. The measured s/σ of trained spectral weights at $C = 32$ is above the crossover at every granularity.

Proposition 2 (Clipping optimum)**.** *Let the moduli of the weights in a group have a bounded density q on $[0, w_{\max}]$, and let a quantizer with scale s have granular error $\gamma(s) \geq 0$ per weight inside the range, with γ differentiable and $\gamma' > 0$, and clipping error $(r - s)^2$ for a modulus $r > s$. The expected squared error*

$$V(s) = \gamma(s)\,H(s) + \int_s^{w_{\max}} (r - s)^2 q(r)\,dr, \qquad H(s) = \int_0^s q(r)\,dr,$$

Table 9: Full factorial at per-mode granularity. Median error ratio over seeds, validation split.

A	C	P	T8	T32	D8	D32
$\bar{b} = 2$						
●	●	●	3.45	1.21	1.10	1.21
●	●	○	3.72	1.28	1.10	1.24
●	○	●	4.36	1.41	1.11	1.23
●	○	○	5.87	1.61	1.22	1.62
○	●	●	12.5	5.78	20.1	48.9
○	●	○	13.3	7.61	10.9	16.4
○	○	●	16.5	7.29	18.9	55.5
○	○	○	19.9	24.1	38.3	124
$\bar{b} = 3$						
●	●	●	1.92	1.07	1.04	1.06
●	●	○	2.09	1.11	1.04	1.09
●	○	●	2.17	1.12	1.04	1.07
●	○	○	3.06	1.32	1.08	1.20
○	●	●	6.62	3.00	5.82	7.94
○	●	○	7.63	3.14	4.49	6.26
○	○	●	7.08	3.24	5.21	8.45
○	○	○	13.0	6.34	8.17	70.6
$\bar{b} = 4$						
●	●	●	1.28	1.02	1.02	1.02
●	●	○	1.38	1.04	1.02	1.04
●	○	●	1.31	1.03	1.02	1.02
●	○	○	1.67	1.11	1.03	1.05
○	●	●	3.39	1.64	2.53	3.05
○	●	○	4.06	1.84	2.44	2.72
○	○	●	3.46	1.71	2.38	2.98
○	○	○	5.32	2.84	2.88	6.78

Table 10: Leave-one-out at $\bar{b} = 4$, as in Table 3.

	A	C	P	G	$\bar{b} = 4$			
					T8	T32	D8	D32
all on	●	●	●	●	1.28	1.02	1.02	1.02
uniform bit-width	○	●	●	●	3.39	1.64	2.53	3.05
absmax scaling	●	○	●	●	1.31	1.03	1.02	1.02
Cartesian	●	●	○	●	1.38	1.04	1.02	1.04
per-band scales	●	●	●	○	1.32	1.02	1.02	1.02
conventional design	○	○	○	○	29.1	23.5	14.4	82.5

satisfies $V'(w_{\max}) > 0$*, so every minimizer of* V *on* $[0, w_{\max}]$ *lies strictly below* $w_{\max}$.

Proof. $V'(s) = \gamma'(s)H(s) + \gamma(s)q(s) - 2\int_s^{w_{\max}}(r-s)q(r)\,dr$. At $s = w_{\max}$ the integral vanishes, $H = 1$ and $\gamma'(w_{\max}) > 0$, so $V'(w_{\max}) \geq \gamma'(w_{\max}) > 0$. □

For a uniform grid with M levels per side, $\gamma(s) \propto s^2/M^2$. The argument uses only that the clipping error and its derivative vanish when nothing is clipped. The implemented search evaluates 50 multiples of the largest modulus and can still return the largest modulus itself.

Scale storage. One FP32 scale per mode, shared over C^2 complex weights, costs $32/(2C^2) = 16/C^2$ bits per real component: 0.25 at $C = 8$, 0.0156 at $C = 32$ and 0.0039 at $C = 64$.

G DARCY FLOW

Data and training. The Gaussian field ξ has spectrum $(|k|^2 + \tau^2)^{-\alpha/2}$ with $\alpha = 3$ and $\tau = 3$. The forcing is $f = \sin x \cos y + \sin y \cos z + \sin z \cos x$ minus its mean, and the solver converges to a relative residual of 10^{-11}. Models are trained for 50 epochs with AdamW, batch size 32, learning rate 10^{-3} with a cosine schedule and weight decay 10^{-4}. Every checkpoint beats the mean-solution baseline by more than the pre-registered factor of 2: by $6.8\times$ to $10.1\times$ at $C = 8$ and $20.3\times$ to $24.1\times$ at $C = 32$.

Pre-registered test. The allocation comparison with a two-bit floor was fixed in a pre-registration before any Darcy data existed, and was measured once on the held-out test split at $b \in \{3, 4\}$, per-band scales, MSE-optimal clipping and the DC component in FP32. The allocation wins in 36 of 36 test cells. At three bits it lowers the median error ratio from 9.29 to 1.12 in the raw arm and from 9.04 to 1.10 in the centred arm. On the validation split the allocation with pruned bands reaches at least 0.97 of this allocation's gain in every comparison at three and four bits. The amendment log records every change to the pre-registration with its time.