GRAPHPTQ: POST-TRAINING QUANTIZATION FOR GRAPH MODELS, MEMORY SAVINGS AND GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization (PTQ) is essentially unexplored for graph models. Yet graph foundation models (GFMs) are scaling rapidly, straining memory budgets, while high-stakes deployment demands robustness to distribution shift. We address this with **GraphPTQ**, the first unified evaluation of how PTQ affects generalization in message passing models, graph transformers, and pretrained GFMs, featuring hardware-measured GPU peak memory. We evaluate eight models and eleven modes, spanning both weight-dominated and activation-dominated memory regimes across covariate shifts and zero-shot tasks. Our findings reveal clear memory-generalization tradeoffs: weight-only INT8 is the safe default for significant memory reduction, while under INT4 weights or activation quantization, out-of-distribution (OOD) accuracy responses become bidirectional and architecture-dependent. We further show that the in-distribution loss curvature along the rounding direction successfully predicts this OOD risk, with the predictive ordering strengthening as bit-width drops.

1 INTRODUCTION

Graph learning has evolved from message-passing GNNs (Gilmer et al., 2017; Xu et al., 2019) and graph transformers (Rampášek et al., 2022; Ying et al., 2021) to larger pretrained *graph foundation models* (GFMs) (Mao et al., 2024; Wang et al., 2025; Zhao et al., 2025), which increasingly power high-stakes problems such as discovering new drugs and materials (Wu et al., 2018), recommending content to billions, and detecting fraud. Deploying them, however, runs into memory. On large-scale graphs, such as web-scale networks or large molecular libraries, it is often the *activations*, not the weights, that dominate, pushing a single forward pass toward the limits of one GPU (Hu et al., 2021; Bechler-Speicher et al., 2026; Ranjan et al., 2025; Zhang et al., 2024). Graph models remain far smaller than LLMs, a few billion parameters at most, but are steadily growing. The standard solution in LLMs (Frantar et al., 2023; Lin et al., 2024; Dettmers et al., 2022; Xiao et al., 2023) and vision (Gholami et al., 2022) is post-training quantization (PTQ), but for graph models it is unexplored. Distribution shift is common in deployment (Li et al., 2024; Gui et al., 2022), where models meet out-of-distribution (OOD) inputs and GFMs are run on unseen domains, so failures can go unnoticed while in-distribution metrics still look healthy.

This work asks what happens when a graph foundation model is quantized: how much memory it saves, what it costs in in-distribution accuracy, and how it affects OOD covariate-shift generalization and zero-shot domain transfer. Memory, accuracy, and generalization do not move together, and the answers are architecture-dependent and bit-width-dependent in ways that prior graph quantization work did not address. Three gaps in prior work motivate this study.

Gap 1: Absence of PTQ for graph transformers and GFMs. Existing GNN quantization research (Feng et al., 2020; Tailor et al., 2021; Wang et al., 2022; Kwon et al., 2025; Rasool et al., 2025) primarily targets traditional message-passing architectures (e.g., GCN, GIN, GAT) and node-level tasks, typically relying on quantization-aware training (QAT). In contrast, graph transformers and pretrained GFMs are dominated by consecutive linear projections. This linear-heavy compute profile makes their quantization behavior mirror LLMs more closely than traditional GNNs, leaving a critical gap in PTQ exploration.

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.

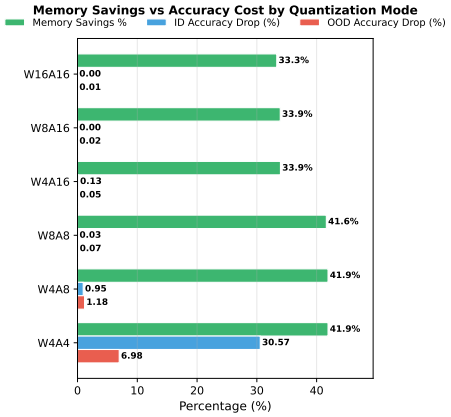

Figure 1: **Weight-only is free, W4A8 degrades, W4A4's ID drops exceed OOD.** Per-mode memory savings, |ID drop|, and |OOD drop| (in %) on the GOOD benchmark (covariate-shift) across five tasks and three backbones. Medians are reported due to heavy-tailed distributions from a few low bit-width failures.

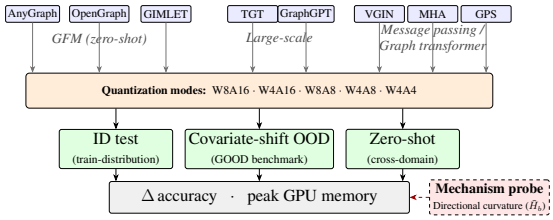

Figure 2: **GraphPTQ.** Eight pretrained models (blue) PTQ'd in WXAY modes (yellow), where X and Y denote weight and activation bit-widths of linear layers. Evaluation spans ID / OOD-covariate (GOOD) / zero-shot (green) regimes. Outputs are ΔAcc% and peak GPU memory (gray), with directional curvature $\tilde{H}_b$ (red, dashed) probing OOD risk.

Gap 2: Theoretical compression vs. measured GPU memory. Prior GNN quantization studies focus solely on compressing weights, reporting theoretical ratios rather than measured GPU peak memory (Feng et al., 2020; Wang et al., 2022; Rasool et al., 2025). Theoretical estimates drastically overstate actual savings because they count bare bits, ignoring framework overheads, workspace allocation, and the activation-dominated bottleneck of large graph data. Without end-to-end hardware measurements, true memory savings remain unknown.

Gap 3: Unexplored quantization-generalization interaction in graph models. Vision and language literature treats this dynamic as a binary question, whether quantization advances or degrades OOD robustness, with conflicting verdicts: some studies find that compression acts as a regularizer (Diffenderfer et al., 2021; Javed et al., 2025), while others show it sharpens the loss landscape and degrades generalization (Lin et al., 2019; Jiang et al., 2025). These discrepancies arise because prior works isolate specific backbones, recipes, and shift types, capturing only single-direction slices of a multidimensional space. For graph models, this interaction remains entirely unmapped: it has not yet been systematically measured under controlled domain shifts, nor under the zero-shot cross-dataset transfer, central to GFMs (Xia & Huang, 2026; Zhao et al., 2023; Xia et al., 2024). Crucially, the mechanistic difference between ID and OOD quantization errors remains unformalized, and it is unknown whether a recipe's OOD impact can be predicted from the ID loss curvature.

We address these gaps with three contributions. First, **GraphPTQ** (Fig. 2) is a unified PTQ framework that analytically decouples quantization noise from distribution shift. By decomposing the forward pass into structured perturbations (weight quantization, activation quantization, and quantization-shift cross-terms), it enables a comprehensive evaluation of graph transformers and GFMs across GOOD (Gui et al., 2022) domain shifts, large-scale regression, and zero-shot transfer.

Second, an **empirical deployment study** (11 modes, 8 architectures, 2 batch-size regimes) reveals the importance of examining both the weight-dominated and activation-dominated memory bottlenecks, charting the memory-vs-accuracy Pareto frontier using hardware-measured GPU peak memory and setting a novel standard for graph PTQ. We establish a universally safe weight-only INT8 default ($\sim$33.9% median memory saving on GOOD, $<$0.1 median GOOD ID drop, $\sim$5% PCQM relative MAE), unlock aggressive recipes for large graph transformers (up to $\sim$71% saving via the fused-kernel $^\circ$ path), and expose severe, configuration-specific risks at INT4 quantization.

Third, we provide the **first quantization-generalization analysis for graph models**, spanning ID, controlled covariate shift, and zero-shot transfer (Xia & Huang, 2026; Zhao et al., 2023; Xia et al., 2024). Its central result is the *Directional-Flatness Ordering* (Hypothesis 1): OOD vulnerability is encoded in the ID loss curvature along the specific PTQ perturbation direction, so a single Hessian-vector product on ID data ranks deployment risk across configurations without requiring OOD evaluation. Two accompanying empirical findings characterize the risk surface: the W4A16

OOD response is bidirectional and architecture-dependent, and W8A16 is the sole recipe where ID preservation and zero-shot generalization reliably coincide across all three GFM families.

The remainder of the paper is organized as follows. §2 reviews related work. §3 formalizes the quantization-shift decomposition, and §4 details the GraphPTQ experimental design. §5 reports the memory and in-distribution accuracy results, and §6 analyzes generalization under quantization.

2 Related Work

Each of the three gaps maps to a distinct prior-work line that falls short of GFMs.

Quantization for graph models. Prior graph quantization is primarily QAT on message-passing GNNs (Feng et al., 2020; Tailor et al., 2021; Wang et al., 2022; Kwon et al., 2025; Rasool et al., 2025), including SGQuant degree-aware bit allocation (Feng et al., 2020), Degree-Quant (Tailor et al., 2021), A^2Q aggregation-aware per-node mixed-precision (Zhu et al., 2023), and Wang et al. (2023) on low-bit / oversmoothing. On the PTQ side, TopGQ (Kwon et al., 2025) uses topology-based node grouping with adjacency-matrix scale absorption, and Rasool et al. (2025) find INT8 the safe threshold on GCN/GIN regression. This line reports theoretical compression rather than measured GPU memory, overlooks graph transformers and GFMs, and does not study the quantization-OOD interaction.

Quantization under distribution shift. In vision and language, compression's robustness effect is bidirectional and implementation-dependent. Lin et al. (2019) show implementation sensitivity (Lipschitz-controlled quantization), and Diffenderfer et al. (2021) show compression *improves* OOD via lottery-ticket subnetworks. In QAT, Jiang et al. (2025) find that OOD degrades via landscape sharpening while Javed et al. (2025) argue quantization noise regularizes. None address graphs under PTQ, where the model cannot adapt to quantization errors, leaving open whether ID loss curvature predicts a configuration's OOD impact.

Flatness and generalization. Two lines tie flat minima to generalization, namely SAM (Foret et al., 2021) (PAC-Bayes *ID* bound motivating flatness-seeking optimization) and SWAD (Cha et al., 2021) (DG-gap extension). Zou et al. (2024) bound the OOD-vs-ID risk gap by a sharpness-dependent constant, but over data-distribution shift on x, not the deterministic weight-space perturbation $\Delta W = Q(W) - W$ PTQ injects, so we invoke it as a heuristic analogy. The Hessian quadratic $\frac{1}{2}\Delta W^\top H \Delta W$ (Nagel et al., 2021) drives GPTQ (Frantar et al., 2023), QuIP (Chee et al., 2023), and the HAWQ family (App. J). On the QAT side, FQAT (Jiang et al., 2025) and QT-DoG (Javed et al., 2025) *find* flat minima via a sharpness-aware penalty in the loss. we instead test an analogous mechanism in a fixed-rounding, calibration-free PTQ regime (§4.3).

3 Preliminaries

We decompose the quantized forward pass into weight, activation, and input-shift perturbations, isolated term-by-term in the experiments.

Setup. Let $f(W, \cdot)$ be a trained graph model with linear-layer weights $W = \{W^{(\ell)}\}$ (normalization and embeddings stay full precision). The activation at layer ℓ is $a^{(\ell)}(x)$, and $\mathcal{L}(W; \mathcal{D})$ is the population loss. We evaluate f and quantized copy f_q on three $(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})$ pairs, namely *ID*, *OOD covariate shift* ($p(y \mid x)$ shared, partitioned on e, e.g. scaffold, graph size), and *zero-shot transfer* (disjoint graph domains).

Structured perturbations. At each linear layer, three noise sources perturb $y^{(\ell)} = W^{(\ell)} a^{(\ell)}$, namely *weight quantization* $\Delta W = Q(W) - W$, *activation quantization* $\Delta a^{(\ell)}(z) = Q(z) - z$, and *distribution shift* $\varepsilon^{(\ell)}(x) = a^{(\ell)}(x) - a^{(\ell)}_{\text{ID}}$, stochastic and unbounded a priori (perturbation taxonomy in App. B).

Forward-pass decomposition. The per-layer output perturbation $\Delta y^{(\ell)} := y_q^{(\ell)} - y^{(\ell)}$ is given by

$$y_q^{(\ell)} = (W + \Delta W)\big(a_{\text{ID}}^{(\ell)} + \varepsilon + \Delta a^{(\ell)}\big), \tag{1}$$

$$\begin{aligned} \Delta y^{(\ell)} = {} & \underbrace{W\,\varepsilon}_{\text{OOD shift}} + \underbrace{W\,\Delta a}_{\text{act. quant}} + \underbrace{\Delta W\, a_{\text{ID}}}_{\text{wt. quant}} \\ & + \underbrace{\Delta W\,\varepsilon}_{\text{cross-term}} + \underbrace{\Delta W\,\Delta a}_{\text{2nd order}}. \end{aligned} \tag{2}$$

The decomposition isolates five terms. The *OOD shift* term $W\varepsilon$ governs FP32 OOD, the *act. quant* term $W\Delta a$ and the *wt. quant* term $\Delta W\, a_{\text{ID}}$ govern ID quantization (ε=0), and two *cross-terms* $\Delta W\,\varepsilon$ and $\Delta W\,\Delta a$ exist only under joint quantization and shift. The *cross-term* $\Delta W\,\varepsilon$ is the OOD-specific weight-quant penalty, vanishing at FP32 or ID and composing multiplicatively across layers. It also carries the sign in Δ_{OOD}. The OOD shift term cancels (same with or without quantization), the wt.-quant term tends one-signed (since a_{ID} is the direction the trained minimum was built around), and $\Delta W\,\varepsilon$ remains sign-unconstrained because ΔW and ε are not co-trained. Bit-width sets its magnitude, alignment sets its sign. Because quantization noise and domain shift are independently generated, identical rounding can push predictions in opposite directions on the same task across different backbones. This is the mechanistic source of the bidirectional OOD responses documented in §6.3.

Second-order loss analysis. The forward-pass decomposition identifies $\Delta W\,\varepsilon$ as the output-space coupling between quantization noise and domain shift. Lifting to the loss, we Taylor-expand at the trained W. W is a local minimum on $\mathcal{D}_{\text{ID}}$ (ID gradient vanishes, Hessian quadratic dominates) but not on $\mathcal{D}_{\text{OOD}}$ (a first-order term $g_{\mathcal{D}_{\text{OOD}}}^{\top}\Delta W$ survives). We abbreviate $\mathcal{L}(\cdot;\mathcal{D}_{\text{ID}})$ as $\mathcal{L}$ and drop the $\mathcal{D}_{\text{ID}}$ subscript on $\Delta\mathcal{L}$, g, $\mathbf{H}$ when the ID split is implied. For any split $\mathcal{D}$, set $\Delta\mathcal{L}(\mathcal{D}) := \mathcal{L}(W_q;\mathcal{D}) - \mathcal{L}(W;\mathcal{D})$ and define

$$g_{\mathcal{D}} := \nabla_W \mathcal{L}|_{\mathcal{D}}, \quad \mathbf{H}_{\mathcal{D}} := \nabla_W^2 \mathcal{L}|_{\mathcal{D}}, \tag{3}$$

all at the trained W and over the linear-layer weights, with $\mathbf{H}$ materialized via Hessian-vector products. These quantities depend on data only and are bit-free. Taylor-expanding the ID loss,

$$\Delta\mathcal{L} = g^{\top}\Delta W + \tfrac{1}{2}\,\Delta W^{\top}\mathbf{H}\,\Delta W + O(\|\Delta W\|^3), \tag{4}$$

gives the canonical PTQ second-order form (Nagel et al., 2021), with $g \approx 0$ at the ID minimum so the quadratic dominates (cubic remainder empirically negligible (Hutchinson, 1990)). On $\mathcal{D}_{\text{OOD}}$ the same expansion holds at W but $g_{\mathcal{D}_{\text{OOD}}}^{\top}\Delta W$ does not vanish, so the OOD loss inherits an additional first-order contribution the ID minimum cannot suppress. This surviving term is the loss-level signature of the cross-term $\Delta W\,\varepsilon$. It couples the sign of the OOD loss change to the alignment between ΔW and ε, and is the mechanism behind bidirectional OOD responses. In §4.3 we project this quadratic onto the quantization direction ΔW_b and test it as a deployment-risk probe across bit-widths.

4 GraphPTQ: Experimental Design

GraphPTQ (Fig. 2) evaluates eight pretrained models across 11 quantization modes under ID, OOD covariate shift, and zero-shot transfer. To resolve the gaps of §1, the framework spans diverse architectures, measures hardware peak memory directly, and applies a unified, recipe to cleanly isolate the quantization-generalization surface. We detail the model suite (§4.1), evaluation protocol (§4.2), and Directional-Flatness hypothesis (§4.3) below.

4.1 Models

We evaluate eight models in three groups (Fig. 2). The *GOOD backbones* (trained from scratch on the GOOD (Gui et al., 2022) datasets) are Virtual GIN (VGIN) (Xu et al., 2019; Gilmer et al., 2017), multi-head attention (MHA), and GPS (Rampášek et al., 2022). The *zero-shot GFMs* are AnyGraph (Xia & Huang, 2026) (∼10M, mixture-of-experts on 38 datasets, cross-group Link1↔Link2), GIMLET (Zhao et al., 2023) (∼64M, T5-small (Raffel et al., 2020), five MoleculeNet (Wu et al., 2018) tasks with scaffold splits), and OpenGraph (Xia et al., 2024) (∼25M, three independently pretrained checkpoints `gen0`/`gen1`/`gen2`, each trained on a separate LLM-generated synthetic graph whose seed prompts target a different downstream domain, namely e-commerce, collaboration / ratings, and drug-interaction respectively). The *large PCQM4Mv2 transformers* (Hu et al., 2021) are GraphGPT (Zhao et al., 2025) (∼227M, Eulerian-path serialization) and the Triplet-Attention Graph Transformer (TGT (Hussain et al., 2024), ∼203M). PCQM4Mv2 lacks public covariate-shift splits, so the latter two carry ID only while OOD and zero-shot rest on the GOOD and GFM families. This selection follows the deployable cross-section of public checkpoints (the largest publicly released GraphGPT checkpoint is ≤227M, the ∼2B variant reported in the paper is not released, AnyGraph and GIMLET each ship one checkpoint, OpenGraph ships three) (see App. C).

4.2 Experimental Setup

Datasets. The evaluation suite combines five GOOD (Gui et al., 2022) tasks of the form *dataset-shift* (HIV-size/scaffold, SST2-length, Motif-size/basis) × three backbones (15 configurations),

PCQM4Mv2 ID-only (GraphGPT, TGT), zero-shot link prediction across 33 AnyGraph datasets (Link1↔Link2), five MoleculeNet zero-shot prompted classification tasks (GIMLET), and synthetic-to-real zero-shot link prediction over three OpenGraph checkpoints. GOOD's sixth, colour-biased task is excluded throughout (its colour-shortcut confound makes loss curvature reflect a spurious feature rather than the OOD shift).

Quantization modes. WxAy denotes x-bit weights and y-bit activations. We evaluate eleven modes: FP32, BF16 (W16A16), weight-only (W8A16, W4A16), and weight+activation quantization (W8A8, W4A8, W4A4). Each explicit INT mode has a twin marked $^\circ$ (W4A16°, W8A8°, W4A8°; W4A4° fails on PCQM and is excluded, App. A). Both explicit and torchao paths use `torch.compile` (Ansel et al., 2024) during evaluation, differing in kernel fusion strategy (explicit inductor-fused dequant-GEMM vs. torchao's precompiled kernels) and dequantization precision (FP32 vs. native BF16). Quantization exclusively targets linear layers (Q/K/V, projections, feedforward); normalizations and embeddings remain FP32. Activation modes apply dynamic per-tensor abs-max (Dettmers et al., 2022; Xiao et al., 2023). We reserve the $^\circ$ path for PCQM4Mv2 transformers: its fused kernel keeps dequantized weight tiles in registers, unlocking maximum memory savings and sidestepping standard-kernel instabilities at g=128 (App. E). All variants are calibration-free (detailed in App. A, Table 3).

Memory and reporting. We evaluate small and large batch regimes to reflect distinct deployment constraints. As OOD benchmarks often feature small graphs, large batches in `torch_geometric` (Fey & Lenssen, 2019) serve as a scale proxy. Since activation memory scales with total node count, this simulates large-scale deployment footprints while maintaining rigorous OOD evaluation. ΔAcc$\%$ is reported in absolute (GOOD) or relative (zero-shot) percentages; peak memory is measured via `torch.cuda.max_memory_reserved()`.

4.3 The Directional-Flatness Ordering Hypothesis

The Hessian quadratic in Eq. (4) estimates the second-order loss change under a generic weight perturbation. However, standard sharpness metrics average this curvature across all directions, ignoring that PTQ is a deterministic, lattice-aligned shift. We propose that projecting the ID Hessian specifically along the actual quantization direction ΔW_b reveals the true driver of performance degradation. If OOD vulnerability is inherited from the ID loss curvature in this specific direction, it becomes estimable strictly from ID data (evaluated in §6.3). We formalize this as a falsifiable prediction:

Hypothesis 1. ***(Directional-Flatness Ordering.)*** *Let ΔW_b be the b-bit quantization perturbation (§3) and $\mathbf{H}_\mathcal{D}$ the loss Hessian on dataset $\mathcal{D}$. We define the* directional curvature*:*

$$\tilde{H}_b(\mathcal{D}) \;:=\; \left| \Delta W_b^\top \mathbf{H}_\mathcal{D}\, \Delta W_b \right|, \tag{5}$$

abbreviated as $\tilde{H}_b$. Across configurations:

(i) $\tilde{H}_b(\mathcal{D})$ orders $|\Delta_{Acc}|$ on both ID and OOD splits, and

(ii) this ordering strengthens as the weight bit-width b drops.

Within $\tilde{H}_b$, only ΔW_b depends on bit-width. Since ΔW_b is invariant to activation precision, two values (b=8 for W8A*, b=4 for W4A*) cover all five retained INT modes. We pair these mode-matched perturbations with the FP32 Hessian $\mathbf{H}$, where OOD effects enter via the $\Delta W\,\varepsilon$ cross-term (§3).

From Optimization to Diagnosis. Prior work either consumes curvature to drive a rounding/bit-width choice (GPTQ (Frantar et al., 2023), the HAWQ family (Dong et al., 2019; 2020; Yao et al., 2021), QuIP (Chee et al., 2023)) or modifies training to reach flatter minima (FQAT (Jiang et al., 2025), QT-DoG (Javed et al., 2025)). We instead use $\tilde{H}_b$ as a diagnostic at the unmodified FP32 minimum, requiring only ID data to compute. Two distinctions from prior work: ΔW_b is the realized perturbation from a fixed scheme (not a variable optimized as in AdaRound (Nagel et al., 2020)/GPTQ), and $\tilde{H}_b$ is a scalar projection onto that specific direction (not a direction-averaged score as in SAM). We find that it orders $|\Delta_{\text{Acc}}|$ on both ID and OOD axes, with the ordering strengthening as the weight bit-width drops (§6.3), providing mechanistic grounding for why quantization degrades OOD generalization on some configurations and not others.

Baseline sharpness metrics. To benchmark $\tilde{H}_b$ against direction-agnostic alternatives, we also evaluate SAM (Foret et al., 2021) loss-landscape sharpness $\mathcal{S}(W)$ on $\mathcal{D}_{\text{ID}}$, Hessian trace $|\text{Tr}(\mathbf{H})|$, and absolute top eigenvalue $|\lambda_{\max}|$ (definitions in App. I). Regression metrics scale continuously with $\Delta\mathcal{L}$, while classification and ranking metrics add a task-specific amplifier near the decision boundary.

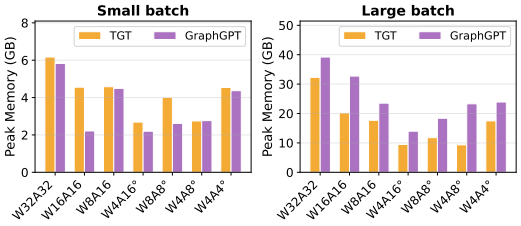

Figure 3: Peak GPU memory on PCQM4Mv2 (TGT, GraphGPT) per mode; small-batch (left), large-batch (right).

Table 1: **Per-mode accuracy on GOOD and PCQM4Mv2.** GOOD: 15 configurations (3 backbones × 5 tasks). PCQM: TGT and GraphGPT under ° variants at INT4 weights and activation quantization. Medians across recipe variants. GOOD absolutes: Table 5. Absolute PCQM MAE: Table 4.

	GOOD ΔAcc%				PCQM ΔMAE%	
Mode	med Δ_{ID}	worst Δ_{ID}	med Δ_{OOD}	worst Δ_{OOD}	TGT	GraphGPT
W32A32	0.00	0.00	0.00	0.00	0.0	0.0
W16A16	0.00	−0.77	−0.01	−0.73	−5.0	−3.2
W8A16	0.00	−0.09	−0.02	−0.90	−4.9	−2.7
W4A16	−0.13	−2.77	−0.05	−8.60	−7.0	−4.1
W8A8	−0.03	−3.54	−0.04	−42.63	−4.7	−1.0
W4A8	−0.33	−3.14	−1.11	−41.50	−7.0	−1.5
W4A4	−30.17	−66.73	−5.07	−52.73	n/a	n/a

5 Memory and ID Accuracy

We analyze the tradeoff between peak GPU memory and in-distribution fidelity (Tables 1, 2, Fig. 3). OOD is in §6.

5.1 Memory Reduction

GPU inference memory splits between bit-width-compressed weights and batch-scaled or graph-scaled activations, so the dominant component shifts with architecture and data, and we report both regimes.

Small batch (weight-dominated). On GOOD, W8A16 and W4A16 yield ∼39% median peak-memory savings. On the PCQM4Mv2 large-scale models, W4A16° is the deployable INT4 recipe at this scale (App. E), reaching 56 to 62% memory saving with near-baseline MAE (GraphGPT −4% and TGT −7% MAE degradation). At W8A16, MAE drift is within ∼3% on both backbones.

Large batch (activation-dominated). Although BF16 autocast (W16A16) theoretically halves activation precision, peak memory drops by only up to ∼ 30% on activation-dominated GOOD configurations because FP32 reductions and other graph overheads remain. Weight-only INT8/INT4 yields essentially no additional savings, as peak memory is dominated by activations rather than stored weights at large batch sizes. Activation-quantization in MHA and GPS backbones enables ∼ 60% memory saving, while VGIN which is a message passing model isn't benefit from activation-quantization. Graph transformers thus offer a *double advantage*, stronger OOD generalization on most GOOD configurations (Niv & Rabin, 2025) and deeper quantization reach, since their linear-heavy compute chains into fusion-friendly flows that stay quantized end-to-end where message-passing GNNs break the chain with un-quantized scatter-gather, a structural bypass absent from attention-only architectures. Further details are given in App. D. The deepest savings come from the ° path on large transformers, with TGT W4A8° at 71%, and GraphGPT W4A16° at 64%, all with near-zero quality loss (§5.2). At ∼200M-parameter scale this is the difference between datacenter-class and single commodity GPU.

Across both regimes, the savings trend is clear. Weight bit-width sets the small-batch ceiling, the activation budget opens up only at large batch, and the 64 to 71% deepest savings appear only at the intersection of the ° kernel path, large transformers, and large batch. The deployment Pareto therefore collapses to three modes (Figs. 1 and 3): W16A16 as a free baseline, W8A16 as the universal safe default, and W4A16° on large transformers (64 to 71%), which matches or exceeds every activation-quantized configuration without their accuracy risk (§5.2).

5.2 Accuracy Under Quantization

Weight-only is safe. W8A16 holds <0.1 median ID drop across all 15 GOOD configurations and is lossless on W16A16 (Table 1). PCQM4Mv2 relative MAE drift is 2 to 5%, and AnyGraph in-domain link prediction is preserved (Table 2). On PCQM4Mv2 transformers W4A16° maintains near-baseline MAE at up to 71% savings, at a precision where standard explicit W4A16 is unstable on GraphGPT (further details in App. E). On GOOD, W4A16 keeps Δ_{ID} near zero.

Activation quantization scales with model size. W8A8 and W4A8 hold every GOOD ID configuration within −3.5 (Table 1). On GFMs, W8A8 keeps every cell within ∼3 % (Table 2), and W4A8

stays within the same envelope with one exception, GIMLET hiv at $-12.8\,\%$. The bulk of the failure budget concentrates at W4A4: on small GOOD models (<1M params), $\mathrm{med}\,\Delta_{\mathrm{ID}} = -30.2$, with ten of 15 configurations losing more than 10 ID (per-configuration breakdown in App. Table 5). On the GFM models, W4A4 failures are task-localized: GIMLET hiv $-28.7\,\%$, bace $-23.1\,\%$, AnyGraph L2→L2 $-11.5\,\%$, while the rest stays within $\sim 8\,\%$. W4A4° fails on GraphGPT ($\sim 36\times$ baseline MAE). Combined with §5.1, the per-mode tradeoffs land on the deployment Pareto of Fig. 1, with W8A16 at the universal-safe knee and W4A16° at the large-transformer knee.

6 Generalization Under Quantization

The deployment tables in §5 are silent on input drift. Because identical INT4 rounding improves some configurations while degrading others on the same OOD task, ID fidelity alone cannot reveal which side of the ledger a model lands on. Prior vision and language studies reached conflicting verdicts here, each studying a single backbone; the multi-backbone graph setting makes the architecture-level driver directly visible. We map this bidirectional landscape on GOOD, extend it to zero-shot transfer, and validate the directional curvature hypothesis (§6.3) as a predictive mechanism.

6.1 OOD Covariate-Shift Landscape on GOOD

W8A16 is stable; W4A16 is bidirectional. Weight-only INT8 produces negligible Δ_{OOD} degradation. At W4A16, the axes decouple (per-configuration breakdown in App. Table 5). ID degrades slightly (median $\Delta_{\mathrm{ID}} = -0.13$), but OOD accuracy disperses into a bidirectional distribution (5 of 15 configurations improve, 2 substantially by >1pp and 3 mildly, while 3 degrade), dropping by up to -8.60 (Motif-size/MHA). The mechanism is the cross-term $\Delta W\,\varepsilon$ in Eq. (2). At INT8, $|\Delta W|$ keeps this term below the noise floor. At INT4, it grows ~ 3 to $4\times$, and its sign (unconstrained because ΔW and ε are not co-trained) dictates the per-configuration OOD outcome. The practical consequence is that the same INT4 rounding can move OOD accuracy in either direction on the same task, so the near-zero Δ_{ID} gives no signal about which way, or how far, OOD will move. This unpredictability is what motivates the directional-curvature probe of §6.3, which recovers the missing signal from ID data alone.

Activation quantization splits by architecture. Under dynamic input quantization, only 2 of the 15 GOOD configurations collapse at W8A8, Motif-size/VGIN and Motif-size/GPS (-14.5 and -42.6); the rest, including Motif-size/MHA, stay within ~ 3.5.

ID vs. OOD asymmetry at W4A4. At extreme low-bit quantization (W4A4), the axis ordering reverses unexpectedly. ID degrades more than OOD across all 15 GOOD configurations (median $|\Delta_{\mathrm{OOD}}|/|\Delta_{\mathrm{ID}}| = 0.47$, mapping every cell below the $y{=}x$ diagonal in Fig. 4), with AnyGraph echoing the same dynamic. This inverts the intuitive expectation that distribution shift always suffers more. The driver is mechanistic: the Hessian quadratic in Eq. (4) loads the ID loss directly, while OOD degradation enters via the sign-indeterminate cross-term $\Delta W\,\varepsilon$, which can cancel rather than amplify. At W4A4, the quadratic dominates and drives ID to the floor, while cross-term cancellation keeps the OOD drop comparatively smaller. This completes a three-regime structure: at W8A16 both terms are negligible, at W4A16 the cross-term sign drives the bidirectional OOD distribution (Fig. 5), and at W4A4 the quadratic overwhelms it. The zero-shot panel mirrors this: 8 of 10 GFM configurations degrade at W4A4 (Fig. 5, right).

6.2 Zero-Shot Cross-Dataset Generalization

W8A16 is uniformly safe; W4A16 splits by architecture. On the zero-shot GFMs, weight-only INT8 preserves FP32 performance uniformly across all three model families (Table 2). At W4A16, however, a three-outcome structure emerges that is governed by the architecture family rather than the task.

Activation quantization exposes extreme architecture-level variance. As bit-widths drop toward W4A4, the worst-task failure envelope spans an entire order of magnitude across architectures. OpenGraph is remarkably resilient, absorbing the full activation-quantization grid across all three checkpoints without exceeding a $\pm 0.6\,\%$ deviation from FP32. Conversely, AnyGraph degrades progressively at each step, drifting -7.8 to $-11.5\,\%$ on cross-domain Recall at W4A4. We do not claim a specific mechanism for this architectural-level divergence, as the models differ simultaneously in pretraining corpus, encoder architecture, and prompt conditioning.

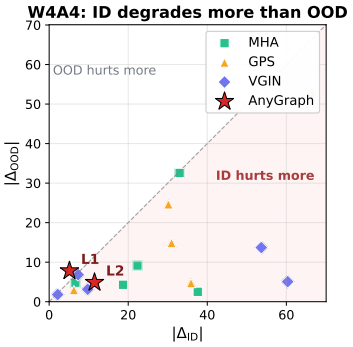

Figure 4: **ID degrades more than OOD at W4A4.** Per-configuration $|\Delta_{\text{ID}}|$ (x-axis) vs. $|\Delta_{\text{OOD}}|$ (y-axis) at W4A4. The diagonal marks equal degradation ($y{=}x$); points below it indicate that ID accuracy drops more than OOD. GOOD configurations (GPS △, MHA □, VGIN ◇, 5 tasks each) and AnyGraph ⋆ (ID vs zero-shot).

Table 2: **Zero-shot performance change (ΔAcc%) across three GFMs.** Positive values = improvement (↑). Evaluated on AnyGraph (33 tasks), OpenGraph (5 checkpoints), and GIMLET (mean of 5 tasks). Absolute scores, task breakdowns, and NDCG@20 metrics are in App. G.

	GIMLET	AnyGraph R@20				AnyGraph N@20		OpenGraph R@20		
Mode	avg.	L1→L1	L1→L2	L2→L1	L2→L2	L1→L2	L2→L1	Books	ML/Co.	DDI
W32A32	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
W16A16	0.0	0.0	−0.4	0.0	0.0	−0.1	−0.1	0.0	−0.6	+0.4
W8A16	0.0	−0.1	−0.1	−0.1	+0.1	−0.1	−0.1	+0.2	+0.1	0.0
W4A16	−4.6	−0.6	−0.1	−0.1	0.0	−0.1	−0.1	+0.2	0.0	+0.1
W8A8	0.0	−0.8	−2.9	−0.8	−2.1	−2.0	−0.7	0.0	−0.2	0.0
W4A8	−4.4	−0.9	−3.3	−0.8	−2.4	−2.3	−0.7	+0.1	−0.1	+0.1
W4A4	−14.5	−5.1	−7.8	−4.9	−11.5	−6.1	−6.8	−0.1	−0.4	0.0

Within-checkpoint task asymmetry. GIMLET exhibits a unique fragility. A single set of pretrained weights fails on specific downstream tasks while preserving others. At W4A16, it drops $-13.2\,\%$ on the HIV task but holds bace, bbbp, muv, and tox21 within $\sim 5\,\%$. This task-localized vulnerability persists deep into activation quantization (W4A8 and W4A4). (AnyGraph's in-domain vs. cross-domain split and the W4A16 NDCG residuals are tabulated in App. K and Table 7.)

6.3 Tests of Hypothesis 1

Testing Hypothesis 1 provides a mechanistic understanding of how quantization interacts with generalization. Projecting the ID loss curvature strictly along the quantization vector isolates the precise vulnerability that direction-blind metrics miss, translating our theoretical decomposition into a deployment tool. We validate this on the GOOD pool ($n{=}15$ configurations) using the mode-matched directional curvature $\tilde{H}_b$ defined in Eq. (5) (per-cell values in App. I.1).

Validating the curvature mechanism. Figure 6 reveals the hypothesized bit-width amplification cell by cell. At W8A16, configurations sit within a tight no-effect band. At W4A16, cells fan out into the bidirectional spread predicted by Eq. (2). By W4A4, the highest-$\tilde{H}_b$ cells clearly cluster at the largest $|\Delta_{\text{OOD}}|$ extremes, while low-curvature cells remain near zero. Consequently, the Spearman correlation between $\tilde{H}_b$ and $|\Delta_{\text{Acc}}|$ strengthens as the bit budget drops (Tab. 8, Fig. 13), supporting the link between directional flatness and robustness.

Predicting OOD risk via directional ID curvature. Crucially, the ID axis tracks consistently above the OOD axis across modes. This follows directly from the Taylor decomposition (Eq. (4)): $\tilde{H}_b$ captures the Hessian quadratic governing the ID loss change, while the OOD axis introduces the $\Delta W\,\varepsilon$ cross-term. Since this cross-term's sign varies unpredictably, a residual fraction of OOD variance remains unresolved. **The practical added value is direct: a model's OOD vulnerability under quantization is estimable from ID data alone, before any OOD evaluation.** Practitioners can rank-order deployment risk across configurations using only ID calibration data at zero additional computational cost.

Unmasking the cross-term: architecture drives OOD direction. While $\tilde{H}_b$ predicts degradation magnitude, the cross-term $\Delta W\,\varepsilon$ explains INT4's bidirectional chaos. Its sign remains unconstrained because ΔW and ε are not co-trained. At W4A16, this yields a mixed distribution (Fig. 5), with backbone architecture emerging as the fundamental driver. Across the five GOOD datasets, VGIN averages $+1.5$ (mostly improving/neutral), GPS averages -0.6, and MHA averages -1.7 (mostly degrading). For example, on Motif-size at W4A16 under identical INT4 rounding, VGIN gains $+8.77$, GPS gains $+1.53$, and MHA loses -8.60 (App. H, Fig. 12). Because only 12% of 75 OOD cells improve >0.1, these positive alignments are architectural artifacts rather than reliably exploitable benefits.

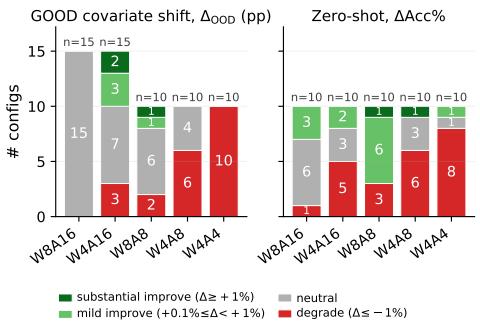

Figure 5: **Sign distribution of Δ_{OOD} per mode.** Configs colored by improve / no change / degrade using a shared threshold. *Left:* GOOD covariate shift (absolute, n=15). *Right:* Zero-shot transfer (relative %, n=10 overarching task/checkpoint groups).

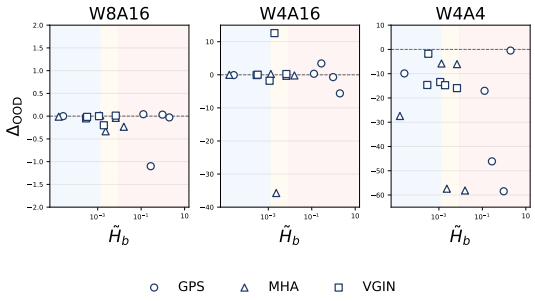

Figure 6: **Mode-matched directional curvature on GOOD (n=15).** Signed Δ_{OOD} vs. $\tilde{H}_b$ (log) matching b=8 (W8A16) and b=4 (W4A*). Background shading splits configurations into low (blue), mid (yellow), and high (red) curvature terciles. Markers: GPS ○, MHA △, VGIN □.

Directional projection is essential. Finally, we demonstrate why prior literature struggles to map this surface by benchmarking $\tilde{H}_b$ against standard direction-blind probes (SAM (Foret et al., 2021), $|\mathrm{Tr}(\mathbf{H})|$ (Hutchinson, 1990), and $|\lambda_{\max}|$). Because none of these direction-blind metrics replicate the bit-width-amplified ordering, the results confirm that projecting the Hessian along the exact quantization step is strictly necessary to unlock this predictive signal (full ablations in Fig. 14, App. I.2).

7 Conclusion

This work investigates the effects of post-training quantization on graph foundation models across three primary dimensions, namely memory efficiency, in-distribution (ID) fidelity, and OOD/zero-shot transfer performance. Our results reveal that these axes are largely decoupled, governed by three distinct mechanisms, namely kernel execution paths, cross-term alignment, and directional loss curvature.

Memory Savings: Beyond Nominal Bit-Width. Hardware-measured peak memory, the first such baseline for graph PTQ, collapses the practical recipe space to two safe operating points. Weight-only W8A16 is the universal safe default. For large graph transformers, the deployable frontier is W4A16°, where kernel-level fused dequantization unlocks up to ∼71% memory savings, matching or exceeding activation-quantized modes (W8A8/W4A8) without their severe accuracy risks (§5.1).

INT4 Decouples ID from OOD. INT8 weight quantization is invisible on every axis, but at INT4 weights the cross-term $\Delta W\,\varepsilon$ in Eq. (2) lifts into a regime where its sign, unconstrained because ΔW and ε are not co-trained, drives OOD outcomes per configuration. Identical INT4 rounding can move same-task OOD accuracy in opposite directions across backbones with ID barely moving for either, on both covariate shift and zero-shot transfer (§6.1, §6.2). At W4A4 the asymmetry inverts: the Hessian quadratic dominates and ID degrades faster than OOD (Fig. 1).

Directional Curvature as an OOD Diagnostic. The directional curvature $\tilde{H}_b$, a single Hessian-vector product in the actual quantization direction at the FP32 minimum, orders $|\Delta_{\mathrm{Acc}}|$ on both axes, with strength increasing as the weight bit-width drops, while isotropic probes (SAM, $|\mathrm{Tr}(\mathbf{H})|$, $|\lambda_{\max}|$) do not (§6.3). Projecting along the precise quantization direction is what makes the screen work, since direction-blind probes average over harmless and harmful axes alike. The practical recipe is therefore W8A16 by default, W4A16° on large graph transformers, and a $\tilde{H}_b$ screen on ID data before committing to any INT4 weight recipe.

Limitations. Billion-parameter GFM checkpoints remain unreleased (Bechler-Speicher et al., 2026; Galkin & Doguparty, 2025), so ours evaluations are only up to 200M parameter scale, and reported memory savings are a lower bound on what larger transformers would achieve (especially at the weight-dominated regime). Node-level tasks are untested. Calibration-aware PTQ (GPTQ, AWQ, SmoothQuant) is out of scope by design, since a calibration set fit on ID data would adapt the quantized model toward ID and confound the interaction we study. Extending $\tilde{H}_b$ to capture Δa, and isolating the cross-term sign mechanism with controlled ablations, remain open. Reproducibility details are in App. C.

Ethics Statement

PTQ for graph foundation models lowers the memory barrier for deploying drug-discovery and fraud-detection systems on commodity hardware. Our finding that quantization effects are bidirectional and architecture-dependent carries a safety implication: in-distribution success is not a proxy for OOD reliability, and OOD or zero-shot validation should be a routine check for quantized graph models in high-stakes applications. All experiments use public benchmarks (GOOD, PCQM4Mv2, and the zero-shot datasets of §6.2) and publicly released pretrained checkpoints; no human subjects, private, or personally identifiable data are involved. Other societal implications are consistent with those of graph machine learning generally.

Reproducibility Statement

All models, datasets, quantization modes, and evaluation protocols are specified in §4; the full quantization pipeline and per-mode backend scope are given in App. A; and hardware, software versions, seeds, and measurement methodology for the reported GPU peak-memory numbers are given in App. C. All benchmarks and pretrained checkpoints used are publicly available. Anonymized code to reproduce the quantization pipeline, the memory measurements, and the directional-curvature probe is submitted as supplementary material.

LLM Usage Statement

Large language models were used solely as a writing aid: polishing prose for clarity and concision, formatting LaTeX tables and figures, and formatting references. LLMs were not used to generate research ideas, formulate or refine the hypothesis, design the experiments, implement the quantization pipeline or evaluation code, or interpret results. The authors reviewed and verified all LLM-assisted text and take full responsibility for the content of this paper.

References

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, et al. PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2024.

Maya Bechler-Speicher, Yoel Gottlieb, Andrey Isakov, David Abensur, Ami Tavory, Daniel Haimovich, Ido Guy, and Udi Weinsberg. Billion-scale graph foundation models. *arXiv preprint arXiv:2602.04768*, 2026.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, 2021.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022.

James Diffenderfer, Brian R. Bartoldson, Shreya Chaganti, Jize Zhang, and Bhavya Kailkhura. A winning hand: Compressing deep networks can improve out-of-distribution robustness. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ: Hessian aware quantization of neural networks with mixed-precision. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. arXiv:1905.03696.

Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. In *Advances in Neural Information Processing Systems*, 2020. arXiv:1911.03852.

Boyuan Feng, Yuke Wang, Xu Li, Shu Yang, Xueqiao Peng, and Yufei Ding. SGQuant: Squeezing the last bit on graph neural networks with specialized quantization. In *IEEE 33rd International System-on-Chip Conference (SOCC)*, 2020.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

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*, 2023.

Mikhail Galkin and Pramod Doguparty. Graph foundation models for relational data. Google Research Blog, 2025. `https://research.google/blog/graph-foundation-models-for-relational-data/`.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *International Journal of Computer Vision*, 130:535–560, 2022.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 2017.

Shurui Gui, Xiner Li, Limei Wang, and Shuiwang Ji. GOOD: A graph out-of-distribution benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Md Shamim Hussain, Mohammed J. Zaki, and Dharmashankar Subramanian. Triplet interaction improves graph transformers: Accurate molecular graph learning with triplet graph transformers. In *International Conference on Machine Learning*, 2024.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics – Simulation and Computation*, 19(2):433–450, 1990.

Saqib Javed, Hieu Le, and Mathieu Salzmann. QT-DoG: Quantization-aware training for domain generalization. In *International Conference on Machine Learning*, 2025.

Jiacheng Jiang, Yuan Meng, Chen Tang, Han Yu, Qun Li, Zhi Wang, and Wenwu Zhu. Quantization meets OOD: Generalizable quantization-aware training from a flatness perspective. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

Dain Kwon, Kanghyun Choi, Hyeyoon Lee, SunJong Park, Sukjin Kim, and Jinho Lee. TopGQ: Post-training quantization for GNNs via topology based node grouping. In *International Conference on Learning Representations*, 2025.

Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. arXiv:2202.07987.

Ji Lin, Chuang Gan, and Song Han. Defensive quantization: When efficiency meets robustness. In *International Conference on Learning Representations*, 2019.

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 *MLSys*, 2024. Best Paper Award.

Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Mikhail Galkin, and Jiliang Tang. Position: Graph foundation models are already here. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34670–34692, 2024.

Shmulik Markovich-Golan, Daniel Ohayon, Itay Niv, and Yair Hanani. Automatic mixed precision for optimizing gained time with constrained loss mean-squared-error based on model partition to sequential sub-graphs. In *OPT-ML 2025: 17th Annual Workshop on Optimization for Machine Learning at NeurIPS*, 2025. URL https://opt-ml.org/papers/2025/paper83.pdf.

Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, 2020. arXiv:2004.10568.

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.

Itay Niv and Neta Rabin. Exploring graph-transformer out-of-distribution generalization abilities. *arXiv preprint arXiv:2506.20575*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Ladislav Rampášek, Mikhail Galkin, Vijay Prakash Dwivedi, Anh Tuan Luu, Guy Wolf, and Dominique Beaini. Recipe for a general, powerful, scalable graph transformer. In *Advances in Neural Information Processing Systems*, 2022.

Aditya K. Ranjan, Siddharth Singh, Cunyang Wei, and Abhinav Bhatele. Plexus: Taming billion-edge graphs with 3d parallel full-graph gnn training. *arXiv preprint arXiv:2505.04083*, 2025.

Areen Rasool, Jamshaid Ul Rahman, and Rongin Uwitije. Enhancing molecular property prediction with quantized GNN models. *Journal of Cheminformatics*, 17:81, 2025.

Shyam A. Tailor, Javier Fernandez-Marques, and Nicholas D. Lane. Degree-quant: Quantization-aware training for graph neural networks. In *International Conference on Learning Representations*, 2021.

Shuang Wang, Bahaeddin Eravci, Rustam Guliyev, and Hakan Ferhatosmanoglu. Low-bit quantization for deep graph neural networks with smoothness-aware message propagation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023.

Yuke Wang, Boyuan Feng, and Yufei Ding. QGTC: Accelerating quantized graph neural networks via GPU tensor core. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2022.

Zehong Wang, Zheyuan Liu, Tianyi Ma, Jiazheng Li, Zheyuan Zhang, Xingbo Fu, Yiyang Li, Zhengqing Yuan, Wei Song, Yijun Ma, Qingkai Zeng, Xiusi Chen, Jianan Zhao, Jundong Li, Meng Jiang, Pietro Lio, Nitesh Chawla, Chuxu Zhang, and Yanfang Ye. Graph foundation models: A comprehensive survey. *arXiv preprint arXiv:2505.15116*, 2025.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Lianghao Xia and Chao Huang. AnyGraph: Graph foundation model in the wild. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 882–896, 2026.

Lianghao Xia, Ben Kao, and Chao Huang. OpenGraph: Towards open graph foundation models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Guangxuan Xiao, Ji Lin, Michaël Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 2023.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Zhewei Yao, Zhen Dong, Zhangcheng Zheng, Amir Gholami, Jiali Yu, Eric Tan, Leyuan Wang, Qijing Huang, Yida Wang, Michael W. Mahoney, and Kurt Keutzer. HAWQ-V3: Dyadic neural network quantization. In *International Conference on Machine Learning*, 2021. arXiv:2011.10680.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? In *Advances in Neural Information Processing Systems*, 2021.

Meng Zhang, Jie Sun, Qinghao Hu, Peng Sun, Zeke Wang, Yonggang Wen, and Tianwei Zhang. TorchGT: A holistic system for large-scale graph transformer training. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2024.

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. GIMLET: A unified graph-text model for instruction-based molecule zero-shot learning. In *Advances in Neural Information Processing Systems*, 2023.

Qifang Zhao, Weidong Ren, Tianyu Li, Hong Liu, Xingsheng He, and Xiaoxiao Xu. GraphGPT: Generative pre-trained graph eulerian transformer. In *International Conference on Machine Learning*, 2025.

Zeyu Zhu, Fanrong Li, Zitao Mo, Qinghao Hu, Gang Li, Zejian Liu, Xiaoyao Liang, and Jian Cheng. A^2Q: Aggregation-aware quantization for graph neural networks. In *International Conference on Learning Representations*, 2023.

Yingtian Zou, Kenji Kawaguchi, Yingnan Liu, Jiashuo Liu, Mong-Li Lee, and Wynne Hsu. Towards robust out-of-distribution generalization bounds via sharpness. In *International Conference on Learning Representations*, 2024.

APPENDIX CONTENTS

A Quantization Modes and Pipeline **14**

B Perturbation Type Classification **14**

C Reproducibility Details **15**

D GOOD Memory Breakdown **15**

E PCQM4Mv2 Memory vs. MAE Detail **16**

F GOOD Per-Configuration Absolute Accuracy **17**

G GFM Zero-Shot Heatmap **17**

H Ablation: Backbone Sensitivity **19**

Table 3: Quantization modes and backend scope. WXAY: X-bit weights, Y-bit activations. All INT modes apply group-wise symmetric weight quantization at g=128; the ° superscript marks the torchao kernel path (fused dequant-GEMM, BF16 compute). **GOOD/GFM** and **PCQM** columns mark retained modes per analysis policy (§4); PCQM4Mv2 transformers use ° variants in place of standard INT4/INT8-activation modes.

Mode	Mechanism	GOOD/GFM	PCQM4Mv2
W32A32	FP32 baseline	✓	✓
W16A16	BF16 autocast (weights & activations)	✓	✓
W8A16	INT8 group-wise Linear weights, BF16 Linear inputs	✓	✓
W4A16	INT4 group-wise Linear weights, BF16 Linear inputs	✓	-
W8A8	INT8 group-wise weights + dynamic INT8 Linear input quant	✓	-
W4A8	INT4 group-wise weights + dynamic INT8 Linear input quant	✓	-
W4A4	INT4 group-wise weights + dynamic INT4 Linear input quant	✓	-
W4A16°	torchao group-wise INT4 weight-only (BF16 tinygemm)	-	✓
W8A8°	torchao group-wise INT8 weights + INT8 Linear input quant	-	✓
W4A8°	torchao group-wise INT4 weights + INT8 Linear input quant	-	✓

I Validating the Directional-Flatness Hypothesis **20**

I.1 Validating Directional Curvature ($\tilde{H}_b$) 20

I.2 Failure of Direction-Agnostic Probes 20

J Extended Related Work **21**

K Granularity Ablation: Per-Channel vs. Per-Group **22**

A Quantization Modes and Pipeline

Table 3 enumerates the 11 quantization modes and backend scopes, while Figure 7 illustrates the quantized linear-layer mechanism. The FP32→BF16 transition (W32A32 → W16A16) yields the largest single memory saving. Under BF16 autocast, every linear-layer GEMM and raw `torch.matmul` (including $QK^\top$ and softmax$(\cdot)V$) runs in half precision. While this roughly halves activation memory without quality loss, overall peak GPU memory is not strictly halved, as it depends on the full computation graph and parameter overheads. Numerically sensitive reductions (layer-norm statistics, softmax accumulators, and loss) remain in FP32 by default. Finally, the INT modes further compress the linear layers' weight parameters and, depending on the mode, their activations.

The ° variants in Figure 7 differ from our explicit implementation in two primary ways: (i) kernel fusion strategy (torchao's fused kernels vs. explicit dequant-then-GEMM with `torch.compile` inductor fusion), and (ii) numeric precision (native BF16 compute for W4A16° vs. FP32 dequant-then-matmul in explicit W4A16). Both pipelines apply `torch.compile` during evaluation and use group-wise symmetric weight quantization at g=128 to all linear layers (the explicit pipeline additionally wraps the fused QKV projection of `MultiheadAttention`), and per-channel is reported only as an ablation in App. K. Per-group adds one FP16 scale per 128 weights, ∼3% scale-storage overhead on top of the bit-width-driven memory savings, well within the reporting precision of the memory tables.

B Perturbation Type Classification

The three noise sources of Section 3 differ in determinism, data-dependence, and how their magnitudes are bounded.

- *Weight quantization noise* ΔW is a function of W and the quantizer choice only, bounded per group by the group-wise quantizer bin width $\|\Delta W_g^{(\ell)}\|_\infty \le s_g^{(\ell)}/(2Q_{\max})$ with $s_g^{(\ell)} = \max_{i\in g}|W_i^{(\ell)}|$ over group g of 128 weights. Direction is lattice-aligned, not random.
- *Activation quantization noise* is a deterministic operator on its argument whose magnitude varies with whichever z the layer receives, bounded per-input by $\|\Delta a^{(\ell)}(z)\|_\infty \le \|z\|_\infty/(2Q_{\max})$ for

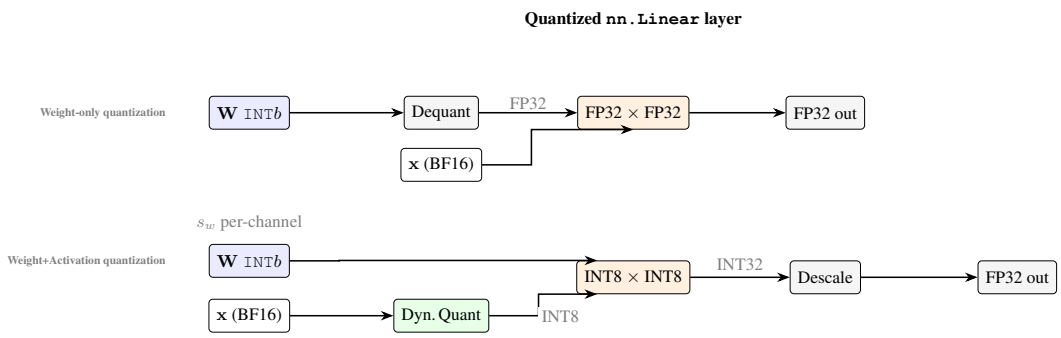

Figure 7: Quantized linear-layer pipeline. All INT modes apply group-wise symmetric weight quantization (g=128). Weight-only: dequantize INTb weights to FP32 before matmul. Activation modes: also quantize input to INT8, run INT8×INT8 GEMM, rescale INT32 output by $s_w s_x$. INT4 weights are nibble-packed. Explicit modes upcast BF16 inputs to FP32 around the dequant-matmul; ° variants stay in BF16 end-to-end.

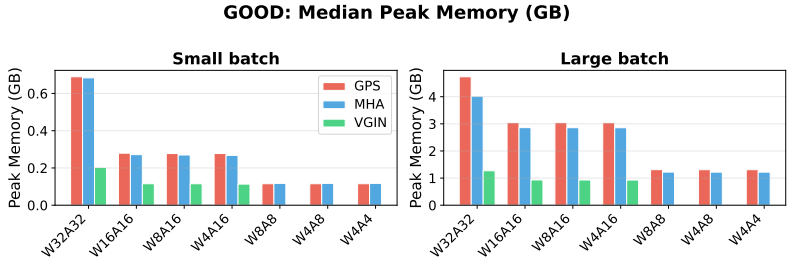

Figure 8: Median peak GPU memory on GOOD by mode and backbone; small/large batch.

the dynamic per-tensor scheme. Unlike common assumptions in robustness or signal processing literature, this quantization error is a deterministic structural artifact rather than stochastic, Gaussian, or adversarial noise.

- *Distribution-shift noise* is a random vector when $x \sim \mathcal{D}_{\text{test}}$, with distribution determined by the pushforward of $\mathcal{D}_{\text{test}}$ through f up to layer ℓ. It is not bounded a priori, and its magnitude scales with shift severity and is characterized empirically rather than by a closed-form bin-width bound. The same form covers OOD covariate shift on GOOD ($p(y \mid x)$ shared) and zero-shot cross-dataset evaluation, where the pushforward gap is typically larger and label-conditionals may also shift.

The two quantization terms are deterministic structured perturbations, while the shift term is distribution-dependent.

C Reproducibility Details

Packages used are: PyTorch ≥2.3 / CUDA 11.3. Datasets are the public GOOD splits and OGB-PCQM4Mv2 (Hu et al., 2021), and for each backbone we keep the originally published seeds, and dataset sub-sampling unchanged so that FP32 baselines match the published numbers, and the quantized modes differ only in the wrapped linear-layer precision. The two memory regimes batch sizes: GOOD backbones use 256 (small) and 4096 (large); TGT uses 32 (small) and 128 (large); GraphGPT uses 16 (small) and 64 (large). We run each configuration on a single NVIDIA A6000 GPU and report peak memory via `torch.cuda.max_memory_reserved()`.

D GOOD Memory Breakdown

Figure 8 details median peak GPU memory for the GOOD backbones by quantization mode and batch size. We focus on PCQM4Mv2 (Figure 3) in the main text because its larger scale yields more pronounced activation quantization savings, though GOOD models exhibit similar underlying trends.

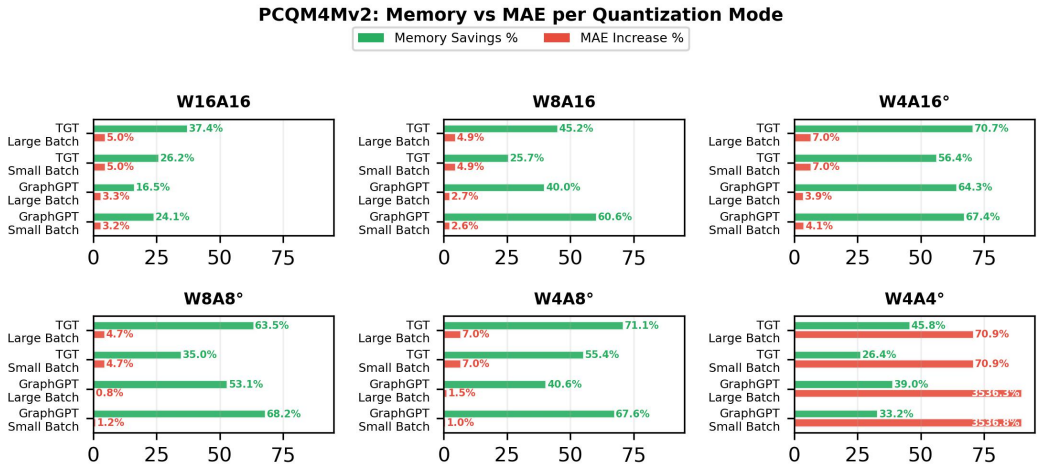

Figure 9: PCQM4Mv2 memory savings (green) vs. MAE degradation (red); TGT, GraphGPT $\times$ both batches.

In the large-batch regime, weight quantization saturates early. Compared to FP32, W8A16 cuts mean peak memory by 42.6% on GPS (5.8 GB to 3.3 GB), 39.3% on MHA (4.3 GB to 2.6 GB), and 39.0% on VGIN (4.3 GB to 2.6 GB), with W4A16 offering negligible additional benefit (within 0.1 MB). Activation quantization effects then diverge by backbone: W8A8 increases savings to 60.2% on GPS and 54.1% on MHA, yet offers no benefit on VGIN, where W8A8 (37.8%) actually trails W8A16 (39.0%). Conversely, at small batch sizes, weight-only quantization dominates peak memory savings (e.g., GPS W8A16 reaches 53.1%).

This supports the weight-only saturation claim in §5.1. GOOD models are small enough that weights form a tiny fraction of an activation-dominated peak. For instance, VGIN's $\sim$21 MB of weights compared to its $\sim$2.6 GB memory peak explains why its activation-quantized modes yield no memory benefit, though their accuracy impacts are still included in the n=15 aggregates (§5.2).

This discrepancy stems from the message-passing architecture. Graph transformers (GPS, MHA) route their large activation tensors through consecutive linear projections that fused INT8 kernels keep in low precision end-to-end, allowing W8A8 savings to compound. In contrast, VGIN interleaves MLPs with gather-scatter aggregations. The gather-expanded edge-message buffer dominates VGIN's memory peak, but standard PTQ stacks (PyG, torchao) do not quantize scatter-gather operations. Consequently, this massive buffer remains in BF16 across all modes. Because the quantizable MLP intermediates are comparatively small, the overhead of per-layer dequantization and requantization slightly outweighs the memory saved, explaining VGIN's W8A8 regression. This systems-level bottleneck mirrors the consecutive-linear versus scatter-aggregation distinction discussed in §1 (Gap 1).

E PCQM4Mv2 Memory vs. MAE Detail

In Figure 9 across both batch regimes and both transformers, the safe region is populated by W4A16° (62.3% peak-memory saving on GraphGPT small-batch and 64.3% on large-batch, MAE 0.065 vs. 0.062 baseline, within 5%) together with the activation-quantized ° modes (W8A8° 53.1% and W4A8° 40.6% on GraphGPT large-batch, W4A8° 71.1% and W8A8° 63.5% on TGT large-batch), all at MAE within 7% of baseline. The unsafe modes tell the kernel-path story directly. Standard W4A16 on GraphGPT explodes to MAE 0.155 (vs. 0.063 baseline, $+147\%$), at *worse* peak memory than W4A16°. W4A4° is omitted from the PCQM analysis because it fails on GraphGPT (MAE 2.28, $\sim$36$\times$ baseline).

The figure resolves Table 1's PCQM columns at per-mode granularity and lands the deployment-tier 2 recipe of §5. On large graph transformers the ° path (per-group weight scales at g=128, `torch.compile`-fused dequant-GEMM) unlocks the deepest savings without the numerical instability the standard kernel exhibits at INT4.

Table 4: Absolute PCQM4Mv2 MAE underlying Table 1 (↓ better). Rows below W8A16 use ° variants (analysis policy).

Mode	TGT MAE	GraphGPT MAE
W32A32 (FP32)	0.068	0.063
W16A16 (BF16)	0.072	0.065
W8A16	0.072	0.064
W4A16	0.073	0.065
W8A8	0.072	0.063
W4A8	0.073	0.063

Table 5: GOOD per-configuration absolute test accuracy underlying Table 1 and Fig. 5. Two rows per (dataset, backbone): top row is ID test, bottom row is OOD test. (**Blue** indicates the best ID accuracy per task and mode; **Red** indicates the best OOD accuracy).

Dataset	Backbone	Split	W32A32	W16A16	W8A16	W4A16	W8A8	W4A8	W4A4
HIV-scaffold	GPS	ID	0.789	0.788	0.788	0.786	0.753	0.770	0.487
		OOD	0.689	0.689	0.689	0.683	0.686	0.685	0.442
HIV-scaffold	MHA	ID	0.737	0.739	0.737	0.733	0.704	0.705	0.513
		OOD	0.574	0.575	0.573	0.577	0.561	0.563	0.482
HIV-scaffold	VGIN	ID	**0.829**	**0.829**	**0.829**	**0.824**	**0.829**	**0.824**	**0.756**
		OOD	**0.715**	**0.715**	**0.715**	**0.717**	**0.714**	**0.717**	**0.647**
HIV-size	GPS	ID	**0.822**	**0.822**	**0.822**	0.795	**0.830**	0.798	0.513
		OOD	0.565	0.565	0.565	0.525	0.577	0.535	0.417
HIV-size	MHA	ID	0.713	0.709	0.713	0.708	0.704	0.698	0.526
		OOD	0.561	0.560	0.560	0.558	0.565	0.548	0.518
HIV-size	VGIN	ID	0.814	0.813	0.814	**0.814**	0.813	**0.814**	**0.792**
		OOD	**0.624**	**0.623**	**0.624**	**0.626**	**0.624**	**0.626**	**0.606**
Motif-basis	GPS	ID	**0.936**	0.929	**0.936**	0.935	0.928	0.932	**0.577**
		OOD	**0.928**	**0.929**	**0.928**	**0.928**	**0.929**	**0.921**	0.881
Motif-basis	MHA	ID	**0.936**	**0.935**	**0.936**	**0.936**	**0.936**	**0.935**	0.560
		OOD	0.925	0.923	0.925	0.924	0.924	0.902	**0.900**
Motif-basis	VGIN	ID	0.926	0.926	0.926	0.926	0.926	0.926	0.323
		OOD	0.675	0.675	0.675	0.656	0.675	0.656	0.624
Motif-size	GPS	ID	0.928	0.928	0.928	**0.929**	0.928	**0.925**	0.261
		OOD	0.806	0.812	0.805	**0.822**	0.380	0.391	0.279
Motif-size	MHA	ID	**0.930**	**0.929**	**0.930**	0.902	**0.929**	0.903	**0.600**
		OOD	**0.904**	**0.897**	**0.902**	0.818	**0.903**	**0.832**	**0.579**
Motif-size	VGIN	ID	0.923	0.923	0.923	0.922	0.922	0.921	0.386
		OOD	0.490	0.491	0.481	0.578	0.345	0.333	0.353
SST2-length	GPS	ID	0.892	0.892	0.892	0.892	0.892	0.893	**0.829**
		OOD	**0.834**	**0.834**	**0.834**	**0.833**	**0.834**	**0.833**	**0.805**
SST2-length	MHA	ID	0.862	0.862	0.862	0.863	0.862	0.862	0.795
		OOD	0.791	0.790	0.790	0.790	0.791	0.791	0.742
SST2-length	VGIN	ID	**0.894**	**0.894**	**0.893**	**0.894**	**0.894**	**0.894**	0.796
		OOD	0.806	0.806	0.806	0.806	0.805	0.804	0.774

F GOOD Per-Configuration Absolute Accuracy

Table 5 reports per-(dataset, backbone) absolute accuracy on both the in-domain (ID) and OOD-covariate (OOD) test splits of GOOD across all seven retained modes. This is the per-configuration view underlying the Δ_{ID} / Δ_{OOD} aggregates in Table 1 and the sign distribution in Fig. 5.

G GFM Zero-Shot Heatmap

Figure 10 visualises the zero-shot accuracy response across every retained quantization mode for GIMLET and AnyGraph, complementing Table 2 in the body. OpenGraph is omitted from the heatmap because its absolute Recall@20 baselines (Books 0.096, DDI 0.076) are small enough that a ΔAcc% view amplifies sub-percent run-to-run noise, and its full numerical record is given

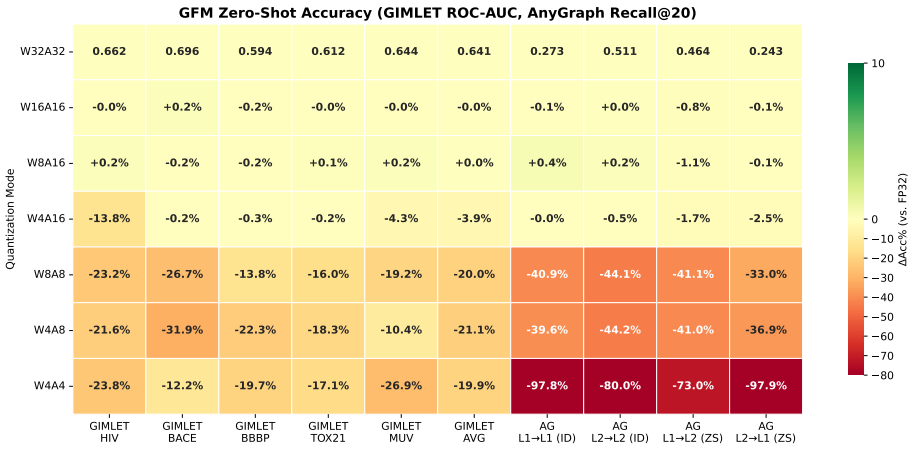

Figure 10: Zero-shot ΔAcc% on GFM backends. *Left:* GIMLET ROC-AUC across 5 MoleculeNet tasks. *Middle:* AnyGraph in-domain Recall@20 (L1→L1, L2→L2). *Right:* AnyGraph cross-domain Recall@20 (L1→L2, L2→L1). W32A32 row reports absolute scores; lower rows report ΔAcc% vs. that baseline. The per-channel pipeline (App. K) shows a much sharper activation-quant collapse: AnyGraph → near-zero at W4A4, GIMLET broad collapse already at W8A8.

Table 6: Absolute zero-shot scores underlying Table 2 (↑ better). All cells aggregated over the full canonical evaluation (AnyGraph Link1 n=15, Link2 n=18; OpenGraph 5 datasets per checkpoint; GIMLET five MoleculeNet ROC-AUC tasks).

	GIMLET ROC-AUC					AnyGraph R@20				OpenGraph R@20		
Mode	bace	bbbp	hiv	muv	tox21	L1→L1	L1→L2	L2→L1	L2→L2	Books	ML/Co.	DDI
W32A32	0.696	0.594	0.662	0.644	0.612	0.327	0.357	0.319	0.402	0.096	0.115	0.076
W8A16	0.697	0.595	0.662	0.643	0.612	0.326	0.356	0.318	0.403	0.098	0.115	0.076
W4A16	0.648	0.572	0.566	0.604	0.610	0.321	0.357	0.318	0.402	0.099	0.115	0.077
W8A8	0.698	0.595	0.666	0.636	0.613	0.318	0.328	0.311	0.381	0.096	0.113	0.077
W4A8	0.646	0.581	0.534	0.616	0.611	0.318	0.324	0.311	0.379	0.097	0.114	0.077
W4A4	0.464	0.524	0.375	0.570	0.548	0.275	0.279	0.270	0.288	0.095	0.111	0.076

in Table 6. The task-level resolution is what makes this view diagnostic. Three signatures stand out. GIMLET W8A8 retains FP32 ROC-AUC on every task (≤ 0.8 % drop on muv, others within 0.4 %), the broad per-channel activation-quant collapse compresses to a W4A8/W4A4 collapse concentrated on HIV/bace, and AnyGraph in-domain ($-0.8/-2.1$ % at W8A8 on L1→L1 / L2→L2) and cross-domain ($-2.9/-0.8$ % on L1→L2 / L2→L1) drift through W4A4 ($-5.1/-11.5$ % in-domain, $-7.8/-4.9$ % cross-domain) without the per-channel-style 50-pp collapse. OpenGraph stays within ± 0.6 % of FP32 across its three independently pretrained checkpoints (gen0/gen1/gen2 on amazon-book / ML+collab / ddi).

AnyGraph Link1 / Link2 group composition. The AnyGraph zero-shot protocol of Xia & Huang (2026) partitions 33 link-prediction datasets into Link1 (15 targets: 7 product/text-feature, 4 citation, 1 PPA, 3 social/p2p) and Link2 (18 targets: 3 e-commerce, 3 recommendation, 5 academic, 4 protein, plus DDI, web-Stanford, roadNet-PA). Each AnyGraph cell in Tables 2 and 6 is the Recall@20 mean over the target group; cross-group L1↔L2 cells use a model pretrained on the source group and evaluated on each member of the target group. The two groups intentionally draw from different sources and do not share feature spaces, which is what makes L1↔L2 a true cross-domain zero-shot protocol; full member roster in Xia & Huang (2026).

NDCG@20 absolutes. Table 7 reports the NDCG@20 absolutes underlying the body claims of §6.2, on the same six configurations and seven modes as Table 6. Recall@20 is set-membership over the top-20, while NDCG@20 weights each held-out edge by $1/\log_2(\text{rank}+1)$, so the two diverge when quantization preserves which edges land in the top-20 but reorders them ($|\Delta_N| > |\Delta_R|$, the AnyGraph cross-transfer regime), or shrinks the top-20 set without disrupting the survivors' rank ($|\Delta_N| < |\Delta_R|$).

Table 7: Absolute zero-shot NDCG@20 (↑ better) for AnyGraph (four cross-group protocols) and OpenGraph (three checkpoints). Full canonical evaluation sets (AnyGraph Link1 n=15, Link2 n=18; OpenGraph 5 datasets per checkpoint). Same modes and aggregation as Table 6; equal-weighted mean across each protocol's target set. GIMLET reports ROC-AUC and is omitted here.

	AnyGraph N@20				**OpenGraph N@20**		
Mode	L1→L1	L1→L2	L2→L1	L2→L2	Books	ML/Co.	DDI
W32A32	0.223	0.269	0.226	0.315	0.140	0.156	0.109
W8A16	0.222	0.268	0.225	0.316	0.144	0.157	0.109
W4A16	0.219	0.269	0.225	0.316	0.143	0.156	0.112
W8A8	0.215	0.250	0.219	0.299	0.140	0.154	0.110
W4A8	0.214	0.246	0.219	0.298	0.140	0.155	0.111
W4A4	0.154	0.208	0.158	0.217	0.138	0.152	0.109

H Ablation: Backbone Sensitivity

Figure 11 visualizes accuracy retention per backbone across quantization modes. Cross-architecture comparison is confounded by different quantization coverage and structural paradigms. MHA relies on pure global attention, VGIN is a pure message-passing network with only $\sim$8 isolated MLP layers, and GPS is a hybrid architecture that interleaves local message-passing with global attention. Because of this, the attention-equipped models (GPS and MHA) share a much larger quantization surface area ($\sim$17 quantized linear layers). Under weight-only modes, all three backbones retain accuracy; VGIN's memory benefit is limited by its small parameter count (median 20 MB model size). This underpins the main-text observation that global attention confers a *double advantage*: stronger OOD generalization on most GOOD configurations and deeper quantization reach. VGIN's activation-quant accuracy cells are reported in the analysis set (Table 5). As highlighted by the blue (ID) and red (OOD) indicators in that table, absolute performance leadership splits cleanly by task type: the pure message-passing VGIN frequently holds the highest accuracy on molecular and text graphs (HIV, SST2), whereas the models equipped with global attention (the hybrid GPS and pure MHA) uniformly dominate the structure-heavy Motif datasets across all quantization modes. However, VGIN's activation-quant memory cells are excluded by the memory-benefit policy (§4), so backbone-intrinsic activation-quant fragility can be inspected on the accuracy axis but cannot be amortized into the memory-savings comparison.

Baseline OOD vs. Quantized OOD Robustness. Comparing the baseline domain shift (FP32) against extreme joint shift (W4A4) exposes a clear architectural divergence. At the FP32 baseline, the models incorporating global attention (both the pure MHA and the hybrid GPS) are natively more robust to domain shift than the strictly message-passing VGIN. For example, on Motif-basis, FP32 VGIN suffers a severe 25.1 OOD penalty ($0.926 \rightarrow 0.675$), while GPS and MHA generalize almost perfectly (> 0.92 OOD). When stressed with W4A4 quantization, the ID accuracies of all three backbones collapse, creating the statistical inversion where W4A4 OOD scores mathematically sit 30 to 34 higher than W4A4 ID. Crucially, the backbones arrive at this inversion via entirely different mechanisms. The attention-equipped models exhibit structural resilience: despite their specifically fitted ID minimum being destroyed (ID dropping to $\sim$0.57), their broader features survive quantization, holding OOD accuracy remarkably high (0.88 to 0.90). VGIN, lacking this structural capacity, achieves the mathematical inversion solely because its ID accuracy completely crashes to the floor (0.323), dragging its already-weak OOD performance down with it (0.624).

Panel (d) reports a non-comparable companion view on the zero-shot GFMs, plotting median score retention vs. FP32 across the three families (AnyGraph, GIMLET, OpenGraph) as used in Fig. 5. Each architecture is evaluated on its own zero-shot set with its own primary metric (AnyGraph Recall@20 across all four cross-group protocols, GIMLET ROC-AUC across the five MoleculeNet classification tasks, OpenGraph link-prediction Recall@20 over the three checkpoints), so absolute retention values across panels are not commensurable, and the panel exists to expose the per-architecture sensitivity profile, not to anchor a cross-architecture comparison. The pattern reproduces the main-text fragility ordering, with GIMLET showing mode-split fragility (W8A8 flat, W4A8 / W4A4 collapse concentrated on HIV / bace), AnyGraph drifting gracefully (-5 to -12 % at W4A4 across protocols), and OpenGraph absorbing activation quantization end-to-end (within ± 0.6 %).

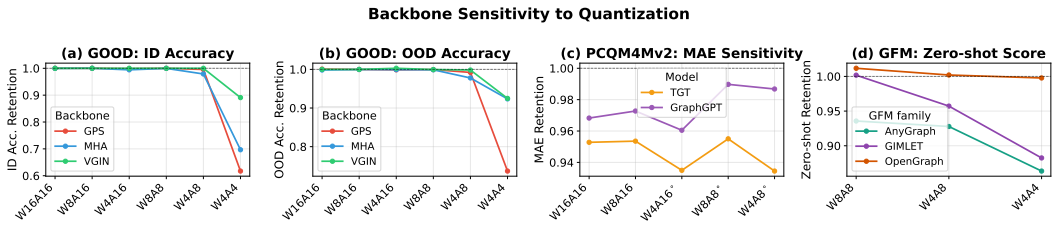

Figure 11: Backbone sensitivity to quantization. VGIN: shown at all modes; activation-quant gives zero memory benefit (S 5.1). Panel (c): retained PCQM4Mv2 modes (° variants below W8A16); standard W4A* and W4A4° excluded per policy (W4A4° fails on PCQM, App. A). Panel (d): non-comparable companion view of the zero-shot GFMs; architecture-median score retention vs. FP32 over the ten zero-shot configurations of Fig. 5, with each architecture scored on its own metric and set, so the y-axis is architecture-internal and should not be read against (a)/(b)/(c) on absolute scale.

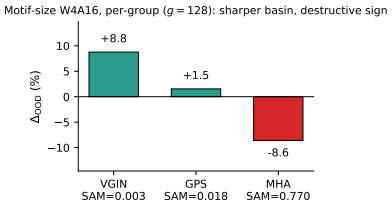

Figure 12: **Cross-backbone Δ_{OOD} example at W4A16 on Motif-size.** Identical quantization yields bidirectional generalization responses across different architectures. The sign of this response is governed by the cross-term $\Delta W\,\varepsilon$, whose alignment is set by backbone architecture (§6.3).

I Validating the Directional-Flatness Hypothesis

This section details the statistical evidence supporting the Directional-Flatness Ordering Hypothesis (Hypothesis 1), validates the failure of direction-agnostic sharpness metrics, and extends the directional curvature analysis to the pretrained zero-shot GFMs.

I.1 Validating Directional Curvature ($\tilde{H}_b$)

Table 8 provides the primary statistical basis for Hypothesis 1 on the GOOD benchmark (n=15). As the quantization perturbation magnitude grows (moving from b=8 to b=4 weight precision), the correlation between the directional curvature $\tilde{H}_b$ and the accuracy degradation strengthens on both axes (W8A16→W4A16: ρ_{ID} $+0.15 \rightarrow +0.28$, ρ_{OOD} $+0.13 \rightarrow +0.15$).

At W8A16, the perturbation is small enough that the curvature ordering sits below the noise floor ($p \geq 0.60$). At W4A4, the directional quadratic form dominates, yielding significant Spearman correlations on the ID axis ($+0.58$, p=0.02). Figure 13 visualizes this bit-width amplification across all retained modes.

Subset robustness. The FP32 checkpoints are not strictly positive-semidefinite minima (6 of 15 configurations sit at saddles where $\lambda_{\max}(\mathbf{H}) < 0$). However, because $\tilde{H}_b$ utilizes the absolute magnitude of the projection, the ordering holds. The correlation separates cleanly by curvature sign at W4A4, with $\rho_{\text{ID}} = +0.89$ (p=0.019) on the negative-curvature subset and $+0.25$ (p=0.52) on the positive subset. Leave-one-cell-out evaluations at W4A4 keep $\rho_{\text{ID}} \in [+0.50, +0.69]$ across all 15 removals, confirming the signal is not driven by a single outlier.

I.2 Failure of Direction-Agnostic Probes

To prove that projecting the Hessian strictly along the quantization vector is necessary, we benchmark $\tilde{H}_b$ against standard direction-agnostic metrics: SAM sharpness $\mathcal{S}(W)$, Hessian trace $\text{tr}(H)$, and absolute top eigenvalue $|\lambda_{\max}|$. Table 10 confirms these metrics form distinct analytical axes.

Crucially, none of the standard isotropic metrics order the OOD accuracy drops effectively. Table 9 splits the GOOD configurations into high and low SAM sharpness groups. Across all modes, the high-vs-low gap confidence intervals cross zero (e.g., the W4A4 OOD gap spans $[-22.12, +23.46]$).

Table 8: **Mode-matched directional curvature ordering of $|\Delta_{\mathbf{Acc}}|$ on GOOD** (n=15)**.** Per-mode Spearman ρ between $\tilde{H}_b$ and the absolute accuracy drops, with b matched to the mode. Bold: $p < 0.05$ (two-sided, critical $|\rho|$=0.514).

Mode	$\rho(\lvert\Delta_{\text{ID}}\rvert)$	p	$\rho(\lvert\Delta_{\text{OOD}}\rvert)$	p
W8A16	+0.15	0.60	+0.13	0.63
W4A16	+0.28	0.31	+0.15	0.59
W8A8	**+0.59**	**0.02**	+0.43	0.11
W4A8	+0.41	0.13	**+0.55**	**0.04**
W4A4	**+0.58**	**0.02**	+0.34	0.21

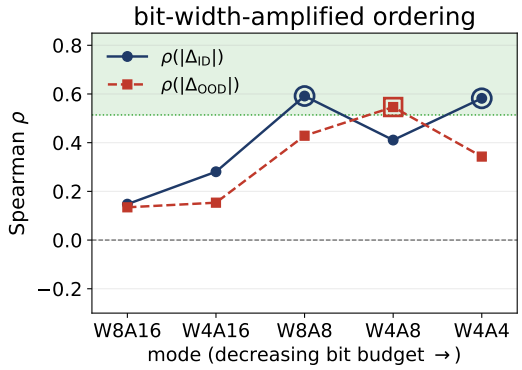

Figure 13: **Mode-matched directional curvature: Spearman ρ across modes** (n=15)**.** ID (solid) and OOD (dashed) ordering of $|\Delta_{\text{Acc}}|$ by $\tilde{H}_b$. The green band marks the n=15, α=0.05 significance threshold. Haloed markers flag $p < 0.05$ cells.

Ablation of Binning Metric: We verify this failure is not an artifact of choosing SAM. Table 11 replays the analysis using $|\lambda_{\max}|$ and $\text{tr}(H)$. Under both SAM and $|\lambda_{\max}|$, the W4A4 response is non-monotone (the low-sharpness group actually degrades the most). While $\text{tr}(H)$ manages a monotone ordering at W4A4, it is highly sign-ambiguous across the dataset. The failure of SAM and $|\lambda_{\max}|$ confirms that averaging over all perturbation directions washes out the quantization-specific vulnerability signal that $\tilde{H}_b$ cleanly captures.

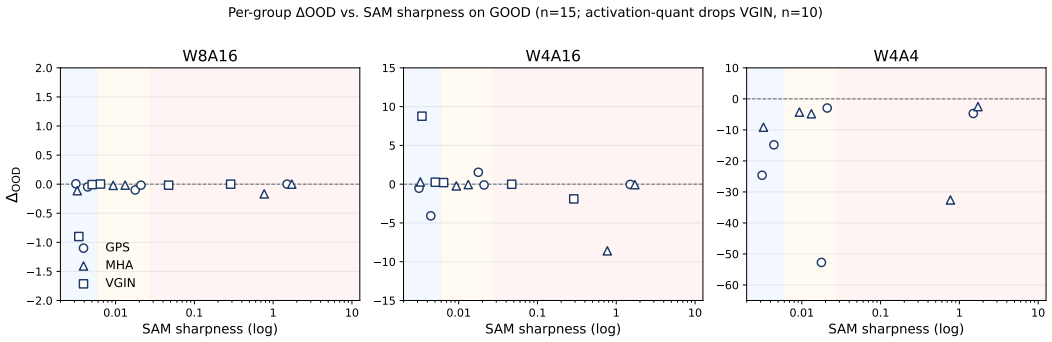

Figure 14: **Signed Δ_{OOD} vs. SAM (log) on GOOD.** Unlike directional curvature (Fig. 6), SAM does not order $|\Delta_{\text{OOD}}|$ at any mode. Markers: GPS ○, MHA △, VGIN □.

J Extended Related Work

PTQ for Sequence and Vision Backbones. Established methods like GPTQ (Frantar et al., 2023), AWQ (Lin et al., 2024), and SmoothQuant (Xiao et al., 2023) leverage dense matmuls in LLMs and vision transformers. However, they do not address the variable-topology operations (scatter-gather aggregation) unique to graph models, which create distinct systems-level bottlenecks.

Table 9: **SAM-binned rank-sum stats (isotropic ablation).** High-vs-low SAM-group medians on $|d_{\text{OOD}}|$ and gap on $|d_{\text{ID}}|$; SAM does not yield a powered ordering at any mode. CI_{95} = 1000-sample percentile bootstrap. p is one-sided.

	$\lvert d_{\text{OOD}}\rvert$ (pp)				$\lvert d_{\text{ID}}\rvert$ (pp)	
Mode	Low	High	Gap [CI_{95}]	p	Gap [CI_{95}]	p
W8A16	0.05	0.00	-0.05 [-0.90, $+0.12$]	0.93	-0.01 [-0.09, $+0.04$]	0.79
W4A16	0.52	0.07	-0.45 [-8.70, $+8.08$]	0.89	-0.26 [-2.70, $+2.63$]	0.87
W8A8	1.21	0.10	-1.11 [-1.27, -0.07]	1.00	-3.26 [-3.51, $+0.12$]	0.95
W4A8	1.11	2.30	$+1.19$ [-2.26, $+6.85$]	0.35	-2.06 [-3.04, $+0.75$]	0.80
W4A4	14.81	4.70	-10.11 [-22.1, $+23.4$]	0.80	$+5.74$ [$+2.03$, $+15.2$]	0.35

Table 10: Spearman rank-correlation matrix between sharpness metrics (n=15).

	SAM	Tr(H)	$\lvert\lambda_{\max}\rvert$	$\lVert g\rVert$	QPS(W4A16)
SAM	1.00	-0.31	0.46	0.91	0.08
Tr(H)	-0.31	1.00	-0.09	-0.27	0.30
$\lvert\lambda_{\max}\rvert$	0.46	-0.09	1.00	0.32	0.14

Curvature: Optimization vs. Sensitivity. Hessian quadratic forms $\frac{1}{2}\Delta W^\top H \Delta W$ (Nagel et al., 2021) are traditionally used as:

- **Optimization targets** for rounding (GPTQ (Frantar et al., 2023), AdaRound (Nagel et al., 2020)).
- **Sensitivity signals** for bit allocation, such as the HAWQ family (Dong et al., 2020) or the automated mixed-precision search by Markovich-Golan et al. (Markovich-Golan et al., 2025), which uses Taylor expansions to optimize within a MSE budget.

Our work is orthogonal: we fix the recipe and use directional curvature as an **analytical diagnostic** to explain why specific rounding directions trigger OOD failure.

Quantization under Distribution Shift. While vision-centric studies suggest compression can regularize OOD performance (Diffenderfer et al., 2021; Javed et al., 2025), others find it sharpens the landscape and degrades robustness (Jiang et al., 2025). Our framework resolves these conflicting views for graph models by identifying that OOD failure is a structural consequence of ID basin-collapse, detectable via directional curvature.

K Granularity Ablation: Per-Channel vs. Per-Group

The body's tables (§5, §6.2) report per-group symmetric weight quantization at g=128. Per-channel scaling (one scale per output row) is the standard baseline in prior graph-PTQ work (Kwon et al., 2025; Rasool et al., 2025) and the LLM literature (Frantar et al., 2023; Nagel et al., 2021), while per-group differs only in scale resolution ($\sim$1:1024 scales per row at g=128 on a typical 1024-input GEMM, $\sim$3% FP16 storage overhead). The choice changes the deployment story on three of the eight backbones and reframes the sharpness binning result.

Where per-group attenuates collapse. (i) On *GOOD W4A16*, switching schemes reduces the worst Δ_{OOD} by $\sim 4\times$ (Motif-size/MHA $-35.90 \rightarrow -8.60\,\%$, Tab. 12). The bidirectional Motif-size sign pattern (§6.3) is preserved (Spearman ρ=0.64 on signed deltas across 15 cells, p=0.010) but destructive-side magnitudes compress, and the W8A16 row is invisible under either scheme (max $|\Delta| < 1.2\,\%$). (ii) On *GOOD WxAy*, the per-channel pipeline leaves `MultiheadAttention.in_proj_weight` unwrapped (a `Parameter`, not a sub-module). With both schemes wrapping it, per-channel still collapses on every Motif-size and HIV-size attention cell at W8A8 / W4A8 / W4A4 (mean $|\Delta_{\text{OOD}}| > 20\,\%$), while per-group holds mean $|\Delta_{\text{OOD}}| < 5\,\%$ at W8A8 / W4A8, with collapse remaining only on Motif-size. (iii) On the *PCQM4Mv2 transformers*, the explicit kernel path with per-channel W4A16 inflates GraphGPT MAE by $+147\%$ (TGT $+38\%$), while the torchao $^\circ$ path (per-group scales at g=128, BF16 fused dequant-GEMM) drops these to $+4\%$ and $+7\%$. The body uses W4A16° as the deployable INT4 path at this scale, and an explicit per-group rerun at INT4 was deferred because the kernel-path difference (FP32 dequant-then-matmul vs. BF16 fused) confounds granularity at INT4. (iv) On *AnyGraph cross-group at W4A4*, per-group drift is $-4.9/-5.1/-7.8/-11.5\,\%$ on L2→L1 / L1→L1 / L1→L2 / L2→L2 against per-channel $-23.8/-26.7/-33.8/-40.9\,\%$, a 3.5-to-7$\times$ attenuation, so the per-channel collapse is rounding-

Table 11: Group medians of signed Δ_{OOD} (pp) under alternative binnings.

Mode	Bin-metric	low-S	mid-S	high-S
W4A16	SAM	$+0.25$	-0.05	-0.07
W4A16	$\lvert\lambda_{\max}\rvert$	$+0.25$	-0.10	-0.07
W4A4	SAM	-13.67	-4.80	-4.70
W4A4	$\lvert\lambda_{\max}\rvert$	-6.86	-4.80	-5.07
W4A4	Tr(H)	-5.57	-9.14	-14.81

rule sensitive, not architectural. (v) On *OpenGraph*, per-checkpoint mean Recall@20 stays within ± 0.6 % of FP32 at every per-group mode.

Where per-group is not strictly better. GIMLET is the only backbone where per-group does not dominate weight-only. At W4A16, per-channel ROC-AUC drifts on bbbp / bace are within 0.2 % absolute while per-group drifts to $-2.17/-4.73$ %, and on HIV both are substantial but per-group incurs a larger penalty (-9.22% per-channel vs. -13.2% per-group). Calibration-aware GPTQ on the same FP32 checkpoints flips bbbp positive ($+4.22$ %) but worsens HIV (-14.88 %), so no rounding rule strictly dominates on this T5-encoder-decoder.

Sharpness binning reframed. The W4A4 SAM-binned high-vs-low $|\Delta_{\text{OOD}}|$ gap of $+52.33$ under per-channel (CI $[+40.34, +58.00]$, p=0.05) collapses to -10.11 under per-group (CI $[-22.12, +23.46]$, p=0.80, Tab. 9). The driver is three HIV / Motif attention cells that floor-saturate to ROC-AUC$\sim$0.50 under per-channel W4A4, while under per-group they stay 9 to 25 above floor. Two residual signals survive the rescaling. First, a within-dataset sign test on the 5 GOOD datasets shows that in 5/5 the flatter (low-SAM) of GPS vs. MHA takes the larger $|\Delta_{\text{OOD}}|$ at W4A4 (one-sided p=0.031), opposite to the textbook "flatter is more robust" intuition. Second, $|\text{Tr}(H)|$-binned medians at W4A4 recover a monotone $-4.3/-4.8/-14.8$ ordering on low/mid/high (Tab. 11), while SAM and $|\lambda_{\max}|$ do not, consistent with $|\text{Tr}(\mathbf{H})|$ being a coarse direction-agnostic proxy for the directional curvature, and the body adopts the directional metric directly (§6.3). On the attention-only subset (n=10), Spearman ρ between $\log_{10}\mathcal{S}$ and $|d_{\text{ID}}|$ is -0.63 at W8A16 (p=0.05) and -0.67 at W8A8 (p=0.03), flipping to $+0.43$ at W4A4 (p=0.21), so a direction-averaged probe cannot track a direction-specific cross-term across bit-widths (full breakdown in App. I.2).

Table 12: Per-channel vs. per-group Δ_{OOD} on the 15 GOOD W4A16 cells (%). Per-group sign distribution preserves the per-channel one (Spearman $\rho = 0.64$ on signed deltas across 15 cells, $p = 0.010$), with destructive-side magnitudes attenuated.

Dataset	Backbone	Per-channel	Per-group (g=128)
HIV-scaffold	GPS	$+0.34$	-0.52
HIV-scaffold	MHA	$+0.11$	$+0.28$
HIV-scaffold	VGIN	-0.06	$+0.20$
HIV-size	GPS	-5.62	-4.08
HIV-size	MHA	$+0.30$	-0.23
HIV-size	VGIN	-0.03	$+0.25$
Motif-basis	GPS	-1.00	-0.03
Motif-basis	MHA	-0.07	-0.07
Motif-basis	VGIN	-1.83	-1.90
Motif-size	GPS	$+3.13$	$+1.53$
Motif-size	MHA	-35.90	-8.60
Motif-size	VGIN	$+12.57$	$+8.77$
SST2	GPS	-0.11	-0.10
SST2	MHA	$+0.01$	-0.05
SST2	VGIN	$+0.25$	-0.01