Zero-Shot Quantization via Weight-Space Arithmetic

Daniele Solombrino[1*] **Antonio Andrea Gargiulo**[1] **Alessandro Zirilli**[1]
Luca Zhou[1] **Adrian Robert Minut**[1] **Emanuele Rodolà**[1,2]

[1] GLADIA Research Lab, Sapienza University of Rome
[2] Paradigma

Abstract

We show that robustness to post-training quantization (PTQ) is a transferable direction in weight space. We call this direction the *quantization vector*: extracted from a donor task by simple weight-space arithmetic, it can be used to patch a receiver model and improve PTQ Top-1 accuracy by up to 60% in a 3-bit setting, without receiver-side quantization-aware training (QAT). When using no receiver validation data, the method provides a zero-shot, low-cost alternative to QAT for extremely low-bit deployment. Across multiple vision and language models and more than 30 tasks, donor quantization vectors often yield substantial gains even when donor and receiver tasks differ markedly. We further prove rigorously that quantization vectors are well-defined and do not suffer from reparameterization symmetries, and provide a local geometric account of their effect. Together, these results suggest that quantization robustness can be partially isolated, reused, and transferred through simple weight-space algebra.

Code: https://github.com/gladia-research-group/qat-transfer
Checkpoints: https://huggingface.co/gladia/qat-transfer-{timm, open_clip, text, vae}
Research graph: https://flywheel.paradigma.inc/node/de486ca0-8e83-4041-bbf6-0b0be565435b

1 Introduction

Deep neural networks are typically trained and stored in high-precision floating-point formats, which impose substantial memory footprint and bandwidth demands during deployment. Integer quantization addresses this bottleneck by representing model parameters with extremely low-bit integers, reducing storage and inference costs. Among the available quantization strategies, post-training quantization (PTQ) [61, 27] is particularly attractive because it can be applied to an already trained model without requiring additional optimization or access to the original training data. However, at extremely low bit-widths, PTQ can severely distort the learned parameters, causing a marked degradation in downstream task performance.

Quantization-aware training (QAT) [37] mitigates this issue by exposing the model to quantization effects during optimization, allowing the parameters to adapt to low-bit perturbations. As a result, QAT often recovers a large fraction of the accuracy lost by naive PTQ. The drawback is that it must be run separately for each receiver task, increasing the data, compute, and engineering cost of deployment. This creates a practical trade-off between efficiency and robustness: PTQ is cheap but fragile at very low bit-widths, whereas QAT is robust but costly. This raises a natural question: can the robustness learned through QAT on one task be isolated and *reused* on another? If so, one could improve low-bit performance on a receiver model without receiver-side quantization training.

We investigate this question through the lens of weight-space arithmetic [36]. Given a donor task with both standard and QAT checkpoints, we define the *quantization vector* (QV) as the displacement between them. We then patch a receiver checkpoint by adding the scaled donor QV. The central

*Corresponding author: solombrino@di.uniroma1.it

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

hypothesis is that QAT leaves behind a reusable displacement toward a more quantization-favorable region of weight space; as illustrated in Figure 1, donor QV patching is intended to occupy the middle ground between vanilla PTQ and full receiver-side QAT.

Two theoretical questions arise immediately. First, is QV itself a legitimate vector in a common parameter space, or can fine-tuning inject hidden symmetries that make coordinate-wise subtraction ambiguous? Second, once a donor QV is well-defined, what determines whether it transfers successfully to a receiver?

We address both aspects in this paper. We show that quantization vectors are generically well-defined: continuous linear reparameterization symmetries are broken by the training dynamics, while permutation symmetries occur only on a measure-zero set of common initializations.

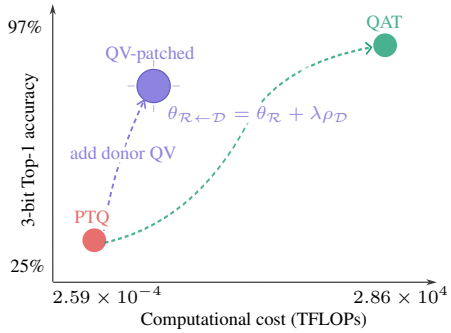

Figure 1: **Zero-shot QV patching.** A donor quantization vector $\rho_{\mathcal{D}} := \theta_{\mathcal{D},\mathrm{QAT}} - \theta_{\mathcal{D}}$, extracted as the weight-space displacement between a standard fine-tuned donor checkpoint and its QAT counterpart, is added to a receiver checkpoint to obtain the patched model $\theta_{\mathcal{R}\leftarrow\mathcal{D}} = \theta_{\mathcal{R}} + \lambda\rho_{\mathcal{D}}$. The plot is not to scale but uses true numbers; it illustrates the intended operating regime of our method, namely improving low-bit accuracy over PTQ without paying the full receiver-side cost of QAT.

We then show that donor transfer admits a simple local theory: under a quadratic model of the receiver's post-quantization objective, optimal donor patching becomes a projection problem in the receiver's geometry: the optimal scale is the projection coefficient onto the donor direction, and the recoverable fraction of the receiver's own QAT gain is exactly a cosine-squared alignment term between donor and receiver QVs. This gives a concrete explanation for two empirical patterns that recur throughout the paper: some donors transfer much better than others, and the scaling factor λ can determine whether the same donor direction helps or hurts. This perspective complements recent geometry-based views of QAT and weight-space arithmetic [36, 101, 77]. However, rather than modeling QAT as a task-specific optimization procedure that must be repeated from scratch, we ask whether its net effect can be isolated as a reusable displacement and transferred across tasks.

We evaluate this framework across multiple ViT scales [20], text encoders, vision, and text tasks. Empirically, donor QVs often improve receiver robustness to 3-bit PTQ, and the gains can be large even when donor and receiver tasks differ substantially and no receiver validation data is used to select the best scaling factor for the QV. Moreover, the role of the scaling factor predicted by the theory is clearly visible in practice: unit-scale transfer is already often effective, while tuning λ, when receiver validation data is available, removes much of the destructive interference that appears when a fixed step size is imposed across all donor–receiver pairs.

The full codebase is available on GitHub, including machine-checked Lean 4 [15] proof artifacts for our theoretical results. We also release checkpoints on the Hugging Face Hub for all model families in our experiments, organized into collections for learned-head vision models, CLIP-style vision models, text models, and VAE decoders [30, 28, 29, 31], including the staged-schedule QAT and QAD donors of Appendix G.5. Finally, we make the research graph available on Flywheel.

Our contributions are fourfold:

- We introduce *quantization vectors*, weight-space displacements between matched standard and QAT checkpoints that operationalize QAT-induced robustness as a transferable object.
- We rigorously prove that quantization vectors are *well-defined*: coordinate-wise subtraction does not generically suffer from hidden symmetry injection during fine-tuning; we also extend the result to ordinary task vectors outside the quantized regime.
- We propose a *zero-shot* cross-task patching framework and a local geometric theory of donor transfer: under a quadratic model of the receiver objective, the best donor patch recovers a cosine-squared fraction of the receiver's own QAT gain.
- Across multiple vision/text tasks and model scales, we show that QV patching often substantially improves 3-bit PTQ over vanilla PTQ, that scale calibration makes transfer more reliable, and how benefits vary across bit-widths and stronger quantizers.

2 Related Work

2.1 Loss Landscape Geometry & Low-Bit Quantization Robustness

Quantization has long been studied to reduce the memory and computational costs of neural network inference, with post-training quantization and quantization-aware training as the two main branches. PTQ is attractive because it can be applied after standard training, does not require access to the original training data, and often uses a small set of calibration data. However, its accuracy can deteriorate sharply at very low bit-widths (3 bits and below). QAT addresses this limitation by incorporating quantization effects during optimization, typically yielding better low-bit performance at the expense of additional data and compute requirements [51].

This trade-off becomes particularly challenging in ViTs [20], where PTQ suffers due to non-standard activation distributions. As a result, many methods have focused on specialized quantizer design, calibration rules [98, 47, 89], or data-free approximation strategies such as MimiQ [9] and DFQ-ViT [79]. Despite all these advancements, such approaches often still require optimization. Recent literature suggests that the effect of quantization is tightly linked to the local geometry of the loss landscape [6]. Hessian-based methods show that curvature can predict sensitivity to low-precision perturbations [18, 19], while other studies [62] have demonstrated that aggressive low-bit quantization can induce highly non-smooth optimization landscapes. In ViTs, recent studies have further emphasized the irregularity of quantized loss surfaces and the role of geometry in determining quantization difficulty [25].

On the training side, Liu et al. [50] connects quantization to sharpness-aware optimization. Instead, recent works argue that QAT, or quantization-induced noise, can guide optimization toward flatter minima or a lower Hessian norm [85, 38, 6], and other weight perturbations applied during training can induce other structural properties in the weights [78]. Another work by Tabesh et al. [77] further formalizes QAT as a multi-objective optimization problem that seeks a Pareto-optimal point between task loss minimization and quantization constraints, achieved through a curvature-aware correction term derived from the local Hessian. Our method is fully consistent with these views; it differs only in its purpose: rather than optimizing a task-specific training recipe, we ask whether the *displacement* induced by such geometry-aware adaptation can be isolated as a reusable direction in parameter space and transferred across tasks.

2.2 Arithmetic in Weight Space

Recent studies indicate that neural networks trained from a common initialization often remain in a compatible basin [88]. Task Arithmetic [36] defines task vectors (TVs) as the difference between a fine-tuned and a pretrained model, showing that they can be used to modify model behavior. Generally, this line of work treats coordinate-wise differences as well-defined as long as the initialization is shared. As an additional contribution, we state this assumption and provide a proof in our setting. Similarly, Cai et al. [4] uses weight-space directions to inject robustness to input corruptions. Other methods improve weight-space arithmetic by finding optimal combinations of TVs [95], mitigating sign disagreement [94], randomly dropping updates [97], using evolutionary strategies [1, 59], or working at the layer level to better respect network structure [76, 26, 57].

We extend this perspective to quantization: instead of encoding a *task*, our QV encodes the *structural robustness* needed to survive low-precision quantization. Our method also differs from Kim et al. [41] where task vectors are quantized to achieve cheaper model merging; instead, in this manuscript, we define *non-quantized displacements to improve quantization robustness*. To our knowledge, this is the first work to isolate the displacement induced by QAT as a reusable, zero-shot patch for cross-task transfer of quantization robustness.

3 Background

In this section, we review the foundational concepts necessary to formalize our framework: a fixed *fake-quantization* operator, working at low-bit precision, and vector arithmetic, operating in weight-space, between checkpoints that share a common pretrained initialization.

3.1 Symmetric Per-Channel Weight Quantization

In this section we focus on weights-only quantization, as the quantity we transfer later represents a displacement in parameter space. Consider a linear layer with a weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. In the setting of this paper, each output channel of W is quantized independently using a signed symmetric integer grid. For a given bit-width b, the range of representable integers is defined as:

$$q_{\min} = -2^{b-1}, \qquad q_{\max} = 2^{b-1} - 1. \tag{1}$$

For channel i, we set the scale based on the largest absolute weight in that channel and quantize by:

$$s_i = \frac{\max_j |W_{ij}|}{q_{\max}}, \qquad \widehat{W}_{ij} = \text{clip}\left(\left\lfloor \frac{W_{ij}}{s_i} \right\rceil, q_{\min}, q_{\max} \right). \tag{2}$$

This quantization process ensures that the weights are scaled appropriately, with values clipped to a specified minimum and maximum. To convert the integer tensor back to floating point, we use the dequantization process defined as follows:

$$\widetilde{W}_{ij} = s_i \widehat{W}_{ij}, \qquad \text{FQ}(W) = \widetilde{W}. \tag{3}$$

Here $\lfloor \cdot \rceil$ denotes rounding to the nearest integer, while the clip function truncates values to fit the representable range. By extension, $\text{FQ}(\theta)$ refers to the checkpoint obtained by applying fake quantization (FQ) to every quantized linear weight tensor in θ, while leaving the remaining parameters unchanged. Note that per-channel quantization commutes with permutations g, i.e. $\text{FQ}(g \cdot W) = g \cdot \text{FQ}(W)$. This desirable property will be instrumental in the theorems that follow.

The specific quantizer matters for our methodology. A different bit-width, granularity, or quantization rule could lead to varying perturbations during QAT, and consequently results in different weight-space displacements. We focus on the symmetric per-channel case because is the operator employed both to create donor QAT checkpoints and to evaluate patched receivers. Throughout the paper, FQ specifically refers to 3-bit symmetric per-channel weight quantization. During QAT, the same fake-quantization operator is incorporated into the forward pass while training, meaning the model is optimized under the perturbation induced by Eq. 2. Since rounding and clipping are not differentiable operations, gradients are approximated with the straight-through estimator [2, 37]. In the next section, we will isolate the parameter displacement between a standard fine-tuned checkpoint and a checkpoint trained to withstand this same fake-quantization noise.

3.2 Weight-Space Arithmetic

The second component of our framework is weight-space arithmetic, which studies linear operations between models that share a common initialization. These checkpoints often remain within a compatible basin, allowing their parameter differences to be composed. Let θ_{pre} denote a pretrained checkpoint and let θ_{task} be the result of fine-tuning it for a specific downstream task. Task arithmetic [36] represents this adaptation by the displacement

$$\tau_{\text{task}} = \theta_{\text{task}} - \theta_{\text{pre}}\,, \tag{4}$$

which is referred to as a *task vector* (TV). When checkpoints share the same architecture and initialization, TVs are treated as vectors in a common parameter space. They can be added to other compatible checkpoints, producing meaningful and controllable changes in model behavior [88]. Our method adopts this viewpoint but focuses on a different source of variation. Instead of measuring the change from pretraining to task adaptation, we assess the change *from standard fine-tuning to QAT on the same task* (as we will show in Eq. (5) below).

Well-definedness of the TV. It is worth noting that, while the difference $\theta_{\text{task}} - \theta_{\text{pre}}$ is algebraically valid, fine-tuning could in principle inject hidden *symmetries* (e.g., neuron permutations) that would place the two checkpoints in different gauges. To our surprise, no theoretical result in the broad task arithmetic literature rigorously rules this out, so we set out to fill this gap. In Theorem 2 (Appendix B.3), we prove that, in the ViT regime studied here, this potential issue is generically absent as AdamW dynamics break continuous linear reparameterization symmetries. We prove this also holds true under QAT fine-tuning, and furthermore, mismatched permutation gauges between matched full-precision and QAT endpoints happen with probability zero, as they occur only on a measure-zero set of common initializations. This ensures that our construction below (in particular Eq. (5)) is well-defined in the ambient backbone parameter space.

4 Method

The core idea behind our method is to treat robustness to low-bit quantization not as a per-task training-time constraint that must be relearned for each downstream model, but rather as a *transferable* geometric alignment in weight space of a fixed architecture. We approach this in two stages: first, we define the quantization vector as the difference between the matched full-precision and QAT checkpoints, and then we describe how to transfer this vector to a receiver model.

4.1 Quantization Vector

Let $\mathcal{D}$ denote a *donor* dataset, and let $\theta_{\mathcal{D},\mathrm{QAT}}$ and $\theta_{\mathcal{D}}$ be the checkpoints obtained by fine-tuning the same pretrained backbone θ_{pre} on $\mathcal{D}$ with and without QAT, respectively. Both fine-tunings additionally share the same random head initialization through a common training seed.[1]

We define the *quantization vector* (QV) as the displacement given by the equation:

$$\rho_{\mathcal{D}} = \theta_{\mathcal{D},\mathrm{QAT}} - \theta_{\mathcal{D}} \,. \tag{5}$$

By comparing Eqs. 4 and 5, we can see that this definition of the QV parallels the notion of TVs in weight-space arithmetic; however, the two displacements represent different effects. The TV in Eq. 4 captures the change in parameters from a pretrained model to a task-specific solution. In contrast, the QV in Eq. 5 represents the change from a standard fine-tuned solution to one that is more robust to low-bit quantization. Our hypothesis is that this displacement isolates *structural robustness* to quantization noise without substantially altering the learned task behavior.

This motivates a geometric view of QAT: it moves parameters toward a solution that better balances task performance and quantization robustness. This is consistent with recent work framing QAT as a multi-objective problem [77]; however, unlike that work, we do not seek to model or improve the optimization process; instead, we focus on isolating its net effect in weight space and study the *transferability* of quantization awareness. The QV is our operational representation of this effect.

Well-definedness of the QV. As anticipated, a potential concern with any displacement in weight space is gauge mismatch: two functionally comparable checkpoints may differ by hidden parameter symmetries, which can make coordinate-wise subtraction ambiguous. In our setting, this issue is generically absent, as demonstrated by the following theorem.

Theorem 1 (Quantization vectors are generically well-defined)**.** *Fix the donor task $\mathcal{D}$ and any realization of the stochastic components of training. Let $\Phi_{\mathcal{D}}^{\mathrm{FT}}$ and $\Phi_{\mathcal{D}}^{\mathrm{QAT}}$ denote the resulting finetuning and its corresponding QAT endpoint maps on the trainable backbone parameters. Assume that the only exact symmetries preserved by the regime form a finite permutation group $\mathcal{G}$, and that $\Phi_{\mathcal{D}}^{\mathrm{FT}}$ and $\Phi_{\mathcal{D}}^{\mathrm{QAT}}$ are piecewise real-analytic. If no nontrivial $g \in \mathcal{G}$ identifies the two maps on an open set, then for almost every common initialization θ_{pre}, the matched endpoints*

$$\theta_{\mathcal{D}} = \Phi_{\mathcal{D}}^{\mathrm{FT}}(\theta_{\mathrm{pre}}), \qquad \theta_{\mathcal{D},\mathrm{QAT}} = \Phi_{\mathcal{D}}^{\mathrm{QAT}}(\theta_{\mathrm{pre}})$$

lie in the same gauge. Consequently,

$$\rho_{\mathcal{D}} = \theta_{\mathcal{D},\mathrm{QAT}} - \theta_{\mathcal{D}}$$

is a well-defined parameter-space displacement.

Proof sketch. Appendix B.3 shows that AdamW admits no nontrivial continuous linear equivariances, so the only surviving exact gauge is discrete; because we use per-channel quantization, the fake quantizer (FQ) of our QAT commutes with that same finite permutation gauge; and the event that the finetuning (FT) and its matched QAT land in different gauges is the zero set of a nontrivial piecewise analytic function, hence measure zero.

Theorem 1 justifies treating Eq. (5) as an ordinary vector difference in the ambient backbone parameter space. Proposition 1 below studies the complementary question of when a well-defined donor QV transfers successfully to a receiver.

[1] We write θ for the shared backbone coordinates, since all quantization and vector operations act only on the backbone.

4.2 Zero-Shot Patching

Let $\mathcal{R}$ denote a *receiver* task for which we only possess a standard checkpoint $\theta_{\mathcal{R}}$ trained without QAT. Given a donor QV $\rho_{\mathcal{D}}$, we construct a patched receiver, with improved resilience to PTQ-induced noise, by adding the donor displacement to the receiver checkpoint:

$$\theta_{\mathcal{R}\leftarrow\mathcal{D}} = \theta_{\mathcal{R}} + \lambda\rho_{\mathcal{D}}\,. \tag{6}$$

Here, $\lambda \in \mathbb{R}$ is a scaling coefficient modulating the intensity of the robustness patch. Intuitively, this operation moves the receiver parameters in a direction that was previously learned to improve robustness to PTQ noise on the donor task.

This protocol assumes that the receiver and donor share the same pretrained backbone initialization. Task-specific heads may differ across tasks and are not part of the transferred vector. Under this assumption, the corresponding backbone checkpoints lie within a compatible loss basin of the weight space [88], making linear transfer meaningful. The resulting procedure is zero-shot and data-free on the receiver side, since it requires *no access to $\mathcal{R}$'s training data*, relying entirely on the pre-computed weight-space arithmetic.

Geometric interpretation. The patching rule in Eq. 6 admits a simple local interpretation that helps explain two empirical patterns that will recur in Section 6: first, some donor vectors transfer much better than others; second, the scaling factor λ can affect if a transfer is helpful or harmful.

The next proposition makes this precise under a local quadratic model of the receiver's post-quantization objective.

Proposition 1 (Alignment controls donor transfer)**.** *Let $g_{\mathcal{R}}(\delta)$ denote the receiver's post-quantization objective after a displacement δ, and assume that near the receiver's own quantization vector $\rho_{\mathcal{R}}$ the objective is smooth enough (locally quadratic):*

$$g_{\mathcal{R}}(\delta) = g_{\mathcal{R}}(\rho_{\mathcal{R}}) + \frac{1}{2}\|\delta - \rho_{\mathcal{R}}\|^2_{H_{\mathcal{R}}},$$
$$\|u\|^2_{H_{\mathcal{R}}} := u^\top H_{\mathcal{R}} u, \qquad H_{\mathcal{R}} \succ 0.$$

For a donor quantization vector $\rho_{\mathcal{D}}$, the best scaled donor patch is

$$\theta_{\mathcal{R}} + \lambda^\star\rho_{\mathcal{D}}, \qquad \lambda^\star = \frac{\rho_{\mathcal{D}}^\top H_{\mathcal{R}}\rho_{\mathcal{R}}}{\rho_{\mathcal{D}}^\top H_{\mathcal{R}}\rho_{\mathcal{D}}}.$$

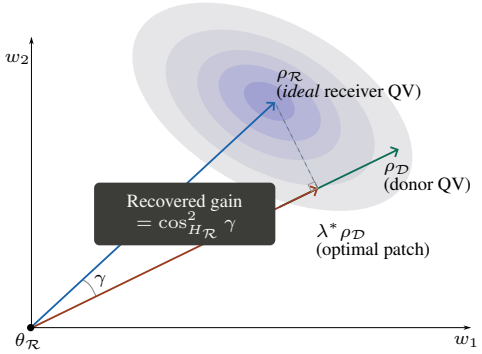

Figure 2: **Geometric view of donor patching.** The blue vector $\rho_{\mathcal{R}}$ is the receiver's own backbone QV (*unknown* in our setting), the green ray is the donor direction $\rho_{\mathcal{D}}$, and the red vector $\lambda^\star\rho_{\mathcal{D}}$ is the orthogonal projection of $\rho_{\mathcal{R}}$ onto that donor line. Proposition 1 states that the fraction of receiver-side QAT gain recovered by this best donor patch is exactly $\cos^2\gamma$, where γ is the angle between $\rho_{\mathcal{D}}$ and $\rho_{\mathcal{R}}$. Illustration in whitened coordinates where the Hessian $H_{\mathcal{R}}$ defines the local geometry.

Moreover, the fraction of the receiver-side QAT gain recovered by this best donor patch is exactly

$$\cos^2_{H_{\mathcal{R}}}(\rho_{\mathcal{D}}, \rho_{\mathcal{R}}), \qquad \cos_H(u,v) := \frac{u^\top H v}{\sqrt{u^\top H u}\sqrt{v^\top H v}}.$$

Therefore:

1. *donor transfer helps if and only if $\rho_{\mathcal{D}}$ is not $H_{\mathcal{R}}$-orthogonal to $\rho_{\mathcal{R}}$;*
2. *donor transfer matches receiver-side QAT if and only if $\rho_{\mathcal{D}}$ is a nonzero scaling of $\rho_{\mathcal{R}}$;*
3. *under this local quadratic model, the best scaled donor patch cannot outperform the ideal receiver QV.*

Proof. See Appendix B.1.

Proposition 1 turns donor patching into a projection problem in the receiver's local geometry; see Figure 2. The optimal scale $\lambda^\star$ is the projection coefficient of the receiver QV onto the donor

direction, and the recoverable fraction of receiver-side QAT gain is exactly the squared cosine of the angle between the two in the $H_{\mathcal{R}}$-geometry (i.e. all inner products are weighted by $H_{\mathcal{R}}$).

This theoretical result captures the second-order term of the local transfer geometry, thus it is exact for a purely quadratic local model. One can further show that the same picture persists beyond that idealization: if the receiver objective is *locally smooth* with Lipschitz Hessian near $\rho_{\mathcal{R}}$, then the cosine-squared law remains accurate up to a cubic remainder. We refer to Appendix B.2 for the complete result with proof.

4.3 Evaluation

To assess the transferability of the QV across different tasks, we compare the downstream performance of the patched receiver checkpoint against the PTQ performance of the unpatched counterpart. Specifically, we apply fake quantization to both models using Eq. 3 and measure Top-1 accuracy on the receiver task. For a donor-receiver pair $(\mathcal{D}, \mathcal{R})$, we define the transfer gain as:

$$\Delta(\mathcal{D}, \mathcal{R}) = \mathrm{Acc}(\mathrm{FQ}(\theta_{\mathcal{R}\leftarrow\mathcal{D}})) - \mathrm{Acc}(\mathrm{FQ}(\theta_{\mathcal{R}}))\,. \tag{7}$$

Here, $\mathrm{Acc}(\cdot)$ denotes the Top-1 accuracy of a model on the evaluation set of the receiver task, and $\mathrm{FQ}(\cdot)$ denotes the fake-quantization operator from Eq. 3, which applies 3-bit symmetric per-channel weight quantization to the model parameters.

For checkpoints trained with QAT, evaluation is always performed after removing the training-time QAT wrappers and applying the same PTQ operator used throughout the paper. This approach ensures that comparisons between FT, PTQ, and QAT are made under common inference-time conditions. The transfer gain, denoted as Δ, provides a clear metric for assessing transfer success: a positive Δ indicates that the transfer successfully improves robustness. In this context, the donor QV effectively mitigates PTQ-induced performance degradation in the receiver task. Conversely, a negative Δ suggests that patching is harmful: the donor QV causes destructive interference in the receiver's parameter space, worsening performance relative to vanilla PTQ. A value of Δ close to zero implies that patching has a negligible impact: the donor QV neither improves nor degrades the receiver's resilience to extremely low-bit PTQ. See Figures 5, 6 and 7 for detailed baseline performance.

5 Experimental Setup

We evaluate QV patching on a broad benchmark spanning both vision and language modalities. In total, our experimental setting covers **14 pre-trained backbones**, finetuned on **22 vision** and

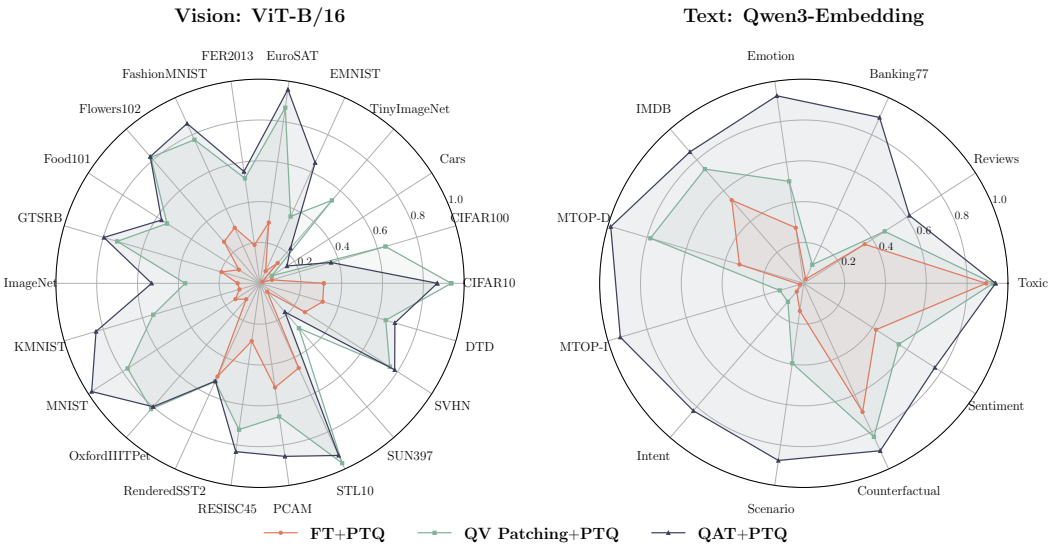

Figure 3: **QV patching across vision and text classifiers.** Receiver Top-1 accuracy under 3-bit PTQ for vanilla fine-tuned checkpoints (FT+PTQ), QV-patched checkpoints (QV Patching+PTQ), and receiver-side QAT checkpoints (QAT+PTQ) on `ViT-B/16` and `Qwen3-Embedding`.

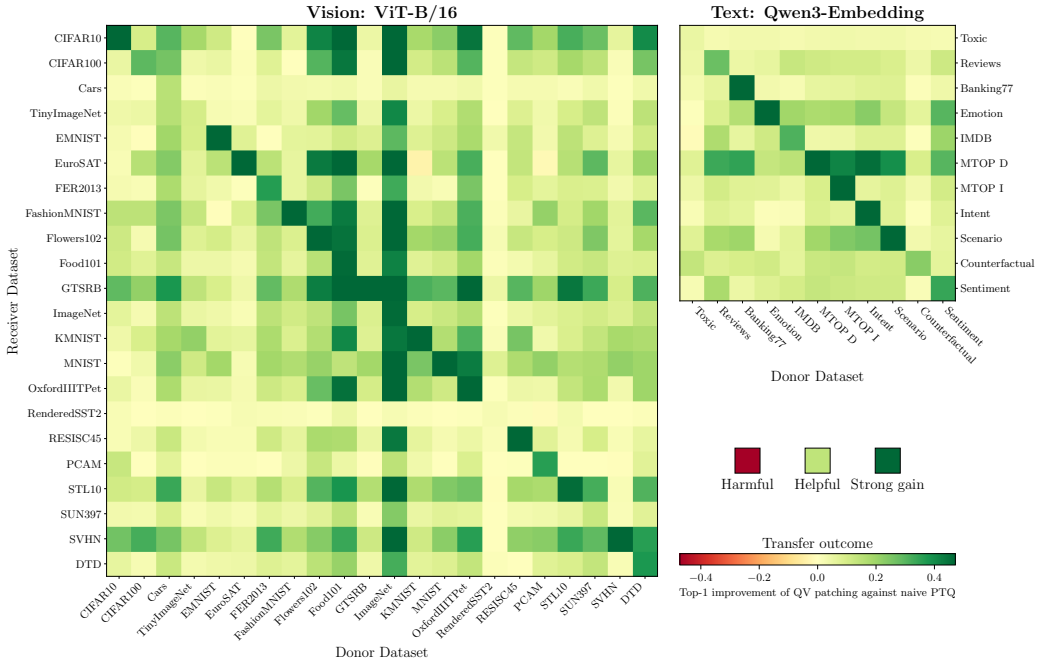

Figure 4: **Cross-task QV transfer for vision (left) and text (right) classifiers.** Top-1 accuracy change from patching receiver r with donor d quantization vector, relative to vanilla 3-bit PTQ.

11 language tasks, allowing us to assess the robustness of our method across model scales and architectures. The complete list of models and datasets is presented in Appendix C.

Quantization setup. We primarily study 3-bit weight-only quantization, which represents the regime in which vanilla PTQ is severely degraded while QAT can still recover substantial performance. Both PTQ and QAT target the weights of every `nn.Linear` PyTorch [69] module in the backbone, excluding the classification head ([67]). The biases, LayerNorm parameters, patch embeddings, and all non-linear modules are left intact. We employ symmetric channel-wise quantization but not activation quantization, activation smoothing, or rotation-based pre-processing [90, 53]. This design choice isolates the effect of QV transfer by preventing advanced pre-processing techniques from introducing confounding signals.

Fine-tuning setup. All weight-space arithmetic is performed between checkpoints sharing the same pretrained initialization. Each task is finetuned with and without QAT with identical optimizer, schedule, batch size, number of epochs, seed, and head initialization; the only difference is the insertion of fake-quantization modules during QAT. QVs are computed and transferred only on the shared backbone or encoder parameters, excluding task-specific heads. We evaluate this protocol on vision and text classifiers across both **learned-head** and **CLIP-style** [70] regimes. Complete dataset lists, model identifiers, and implementation details are provided in Appendix C. Please refer to Appendix G.5 for experiments on additional experimental settings.

6 Results

Quantization vector direction. To isolate the effect of the QV direction from its magnitude, we first set the scaling factor $\lambda = 1$ from Eq. 6. Figures 8, 9 provide distinct insights into transferability when respectively viewed row-wise or column-wise. The rows show how a receiver dataset behaves when patched with available donors, revealing its receptiveness to robustness patching. Conversely, the columns depict the effectiveness of a donor QV across all possible receivers, highlighting its generalizability and strength as a robustness injector. Consequently, the intersection of row r and column d measures the net accuracy change when receiver r is patched with donor d's QV, compared to the standard PTQ described in Eq. 7.

Quantization vector magnitude. Using a fixed-length vector across different tasks might cause the receiver model to either overshoot or undershoot the most useful region of weight space. To investigate whether a suitable magnitude exists for each donor–receiver pair, we treat λ in Eq. 6 as a scalar hyperparameter, similar to the use of task-vector coefficients in Ilharco et al. [36]. This approach is consistent with Proposition 1, which predicts that a donor direction can still be useful even when unit-scale transfer fails, as long as its optimal projection coefficient differs from 1.

In the updated protocol, λ is selected using only the receiver validation split. For each donor–receiver pair, we sweep a fixed grid of candidate coefficients in $[0.15,\ 1.5]$ on validation data with a step size of 0.15, choose the coefficient that maximizes validation Top-1 accuracy after 3-bit PTQ, and then report the corresponding performance on the receiver test split. Formally, we select $\lambda^{\mathrm{val}}_{\mathcal{D}\to\mathcal{R}} = \arg\max_{\lambda\in\Lambda} \mathrm{Acc}_{\mathrm{val}}\left(\mathrm{PTQ}(\theta_{\mathcal{R}} + \lambda\rho_{\mathcal{D}})\right)$ and report test accuracy using this frozen value. Seen through Proposition 1, the validation sweep estimates a practical projection coefficient for applying an already useful donor direction to a given receiver.

The contrast with unscaled transfer is significant, as can be seen in Figures 8, 9. First, validation-tuned scaling mitigates destructive interference. The pronounced negative transfer areas observed under the $\lambda = 1$ protocol are substantially reduced when the magnitude is tuned. In this setting, donor QVs rarely degrade the receiver's baseline PTQ test-set performance. Second, validation-tuned scaling amplifies positive transfer. For pairs that already exhibited gains at unit scale, calibrating λ yields additional improvements, suggesting that the QV often points toward a quantization-robust region, but the appropriate step length depends on the donor–receiver pair. Overall, these results support a directional interpretation of QV transfer. Negative transfer at $\lambda = 1$ is often due to an incorrect step size rather than a fundamentally incompatible donor direction. At the same time, the calibrated setting is not fully data-free: unlike unit-scale patching, it uses a held-out validation split to select one scalar coefficient. It remains receiver-training-free, since no receiver-side QAT or gradient-based optimization is performed.

Highlights. For each receiver task, we apply QV patching using the best donor task (excluding itself) and the best validation-tuned λ. For both vision and text domains, we find that QV patching achieves significant overall gains (see *FT+PTQ* vs *QV Patching + PTQ* in Figure 3), closely approaching the ideal, fully trained QAT+PTQ setting in vision and **even surpassing it on a few occasions**. Because it selects both donor and λ on receiver validation data, this is an upper bound on the data-free setting. A cost analysis in Appendix F shows that QV reuse can amortize the one-time donor QAT cost across multiple receivers, becoming cheaper than independent receiver-side QAT after sufficiently many deployments.

Bit widths and stronger quantizers. Our main experiments focus on 3-bit weight quantization, where vanilla PTQ incurs substantial but still recoverable degradation. At 4 bits, PTQ already remains close to full precision and therefore leaves little room for improvement, whereas at 2 bits even receiver-side QAT struggles to recover performance (Appendix G.2). QV-patched weights remain compatible with stronger PTQ methods such as AWQ [48] and GPTQ [24]; QV patching is not a competitor to them, and its benefit on top of them is small (Appendix G.1).

Remark. Because Theorem 1 rules out generic gauge mismatch between matched FT and QAT endpoints in our regime, the results shown in Figure 3 should be interpreted as genuine geometry transfer rather than as an artifact of subtracting incompatible checkpoints.

In Figure 4, we find that the performance gain across all donor-receiver pairs is non-negative if the proper λ is applied, highlighting the generalizability of the method. Interestingly, ImageNet emerges as an empirically near-universal donor on our benchmark, already at unit scale. We hypothesize that this stems from the breadth and diversity of tasks (over 20k classes), which induces activation and weight distributions with broader dynamic ranges and more representative quantization patterns. As a result, the corresponding QV captures outlier structures that are not overly specialized to a single downstream task, making it effective as a transferable correction signal.

Please refer to Appendix F for an extended analysis of results.

7 Conclusions

In this work, we investigated whether the robustness acquired during QAT can be isolated as a transferable property. We introduced the quantization vector (QV), defined as the weight-space displacement between a standard fine-tuned checkpoint and its QAT-enabled counterpart. Through extensive cross-task evaluations on vision and language models, we demonstrated that patching a standard model with a donor QV can substantially improve its robustness to 3-bit post-training quantization.

More broadly, our results suggest that QAT can induce a stable, partially transferable direction in weight space. This expands the scope of weight-space arithmetic. While previous work has largely focused on manipulating semantic capabilities, such as tasks, languages, or styles, our results show that weight-space directions can also encode structural and computational properties, like resilience to low-precision quantization.

The two theoretical results clarify why this picture is coherent. Theorem 1 shows that, in our regime, the QV is generically a genuine displacement in parameter space rather than an artifact of hidden symmetry mismatch, while Proposition 1 explains why successful transfer depends on donor–receiver alignment and on the patch scale. Consistent with this view, rescaling the donor QV removes most destructive interference and reveals that much of the useful transfer signal lies in the direction itself.

Taken together, these results support a simple perspective: part of quantization robustness can be isolated, stored as a weight-space object, and reused across different tasks without the need to repeat full receiver-side QAT. This suggests a broader understanding of low-bit adaptation, in which robustness is not only trained, but can also be transferred.

Acknowledgments and Disclosure of Funding

This work is supported through the MUR FIS2 grant n. FIS-2023-00942 "NEXUS" (cup B53C25001030001). This work was carried out while Daniele Solombrino was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome. We thank and acknowledge ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy). We thank Fabrizio Fabbri for generously hosting and supporting part of the computing infrastructure used throughout this work. The authors declare no competing interests.

References

[1] Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2):195–204, 2025.

[2] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – Mining Discriminative Components with Random Forests. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 446–461, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10599-4. doi: 10.1007/978-3-319-10599-4_29.

[4] Ruisi Cai, Zhenyu Zhang, and Zhangyang Wang. Robust weight signatures: Gaining robustness as easy as patching weights? In *International Conference on Machine Learning*, pages 3495–3506. PMLR, 2023.

[5] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In Tsung-Hsien Wen, Asli Celikyilmaz, Zhou Yu, Alexandros Papangelis, Mihail Eric, Anuj Kumar, Iñigo Casanueva, and Rushin Shah, editors, *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlp4convai-1.5. URL https://aclanthology.org/2020.nlp4convai-1.5.

[6] Albert Catalan-Tatjer, Niccolò Ajroldi, and Jonas Geiping. Training dynamics impact post-training quantization robustness. *arXiv preprint arXiv:2510.06213*, 2025.

[7] Chan-Y. PixelArt-308K, 2026. URL https://huggingface.co/datasets/Chan-Y/pixelart-308k.

[8] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, October 2017. ISSN 1558-2256. doi: 10.1109/JPROC.2017.2675998. URL https://ieeexplore.ieee.org/document/7891544/?arnumber=7891544. Conference Name: Proceedings of the IEEE.

[9] Kanghyun Choi, Hyeyoon Lee, Dain Kwon, SunJong Park, Kyuyeun Kim, Noseong Park, Jonghyun Choi, and Jinho Lee. Mimiq: Low-bit data-free quantization of vision transformers with encouraging inter-head attention similarity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16037–16045, 2025.

[10] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing Textures in the Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, Columbus, OH, USA, June 2014. IEEE. ISBN 978-1-4799-5118-5. doi: 10.1109/CVPR.2014.461. URL https://ieeexplore.ieee.org/document/6909856.

[11] cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. Jigsaw unintended bias in toxicity classification, 2019. URL https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification.

[12] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep Learning for Classical Japanese Literature, November 2018. URL http://arxiv.org/abs/1812.01718. arXiv:1812.01718 [cs, stat].

[13] Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, June 2011. URL https://proceedings.mlr.press/v15/coates11a.html. ISSN: 1938-7228.

[14] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, May 2017. doi: 10.1109/IJCNN.2017.7966217. URL https://ieeexplore.ieee.org/document/7966217/?arnumber=7966217. ISSN: 2161-4407.

[15] Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, pages 625–635. Springer, 2021. doi: 10.1007/978-3-030-79876-5_37.

[16] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854, 2017.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[18] Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proc. CVPR*, pages 293–302, 2019.

[19] Zhen Dong, Zhewei Yao, Daiyaan Arfeen, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. *Advances in neural information processing systems*, 33:18518–18529, 2020.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[21] Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. Mmteb: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

[22] Zhiwen Fan, Lingjie Zhu, Honghua Li, Xiaohao Chen, Siyu Zhu, and Ping Tan. Floorplancad: A large-scale cad drawing dataset for panoptic symbol spotting. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10108–10117. IEEE, 2021.

[23] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

[24] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

[25] Natalia Frumkin, Dibakar Gope, and Diana Marculescu. Jumping through local minima: Quantization in the loss landscape of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16978–16988, 2023.

[26] Antonio Andrea Gargiulo, Donato Crisostomi, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, and Emanuele Rodolà. Task singular vectors: Reducing task interference in model merging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18695–18705, 2025.

[27] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022.

[28] Gladia Research Group, Daniele Solombrino, Antonio Andrea Gargiulo, Alessandro Zirilli, Luca Zhou, Adrian Robert Minut, and Emanuele Rodolà. qat-transfer-open_clip (revision edff648), 2026. URL `https://huggingface.co/gladia/qat-transfer-open_clip`.

[29] Gladia Research Group, Daniele Solombrino, Antonio Andrea Gargiulo, Alessandro Zirilli, Luca Zhou, Adrian Robert Minut, and Emanuele Rodolà. qat-transfer-text (revision 9367c60), 2026. URL `https://huggingface.co/gladia/qat-transfer-text`.

[30] Gladia Research Group, Daniele Solombrino, Antonio Andrea Gargiulo, Alessandro Zirilli, Luca Zhou, Adrian Robert Minut, and Emanuele Rodolà. qat-transfer-timm (revision 2e5cb7f), 2026. URL `https://huggingface.co/gladia/qat-transfer-timm`.

[31] Gladia Research Group, Daniele Solombrino, Antonio Andrea Gargiulo, Alessandro Zirilli, Luca Zhou, Adrian Robert Minut, and Emanuele Rodolà. qat-transfer-vae (revision 722fd17), 2026. URL `https://huggingface.co/gladia/qat-transfer-vae`.

[32] Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In Minho Lee, Akira Hirose, Zeng-Guang Hou, and Rhee Man Kil, editors, *Neural Information Processing*, pages 117–124, Berlin, Heidelberg, 2013. Springer. ISBN 978-3-642-42051-1. doi: 10.1007/978-3-642-42051-1_16.

[33] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, July 2019. ISSN 2151-1535. doi: 10.1109/JSTARS.2019.2918242. URL https://ieeexplore.ieee.org/document/8736785/?arnumber=8736785. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[34] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

[35] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

[36] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

[37] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

[38] Saqib Javed, Hieu Le, and Mathieu Salzmann. Qt-dog: Quantization-aware training for domain generalization. In *International Conference on Machine Learning*, pages 26981–27004. PMLR, 2025.

[39] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[40] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual amazon reviews corpus, 2020.

[41] Youngeun Kim, Seunghwan Lee, Aecheon Jung, Bogon Ryu, and Sungeun Hong. Task vector quantization for memory-efficient model merging. In *Proc. ICCV*, pages 20105–20115, 2025.

[42] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained Categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, Sydney, Australia, December 2013. IEEE. ISBN 978-1-4799-3022-7. doi: 10.1109/ICCVW.2013.77. URL http://ieeexplore.ieee.org/document/6755945/.

[43] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

[44] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. page 3, 2015.

[45] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[46] Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.257. URL https://aclanthology.org/2021.eacl-main.257.

[47] Zhikai Li, Junrui Xiao, Lianwei Yang, and Qingyi Gu. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17227–17236, 2023.

[48] 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. *Proceedings of machine learning and systems*, 6:87–100, 2024.

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[50] Jing Liu, Jianfei Cai, and Bohan Zhuang. Sharpness-aware quantization for deep neural networks. *arXiv preprint arXiv:2111.12273*, 2021.

[51] Kai Liu, Qian Zheng, Kaiwen Tao, Zhiteng Li, Haotong Qin, Wenbo Li, Yong Guo, Xianglong Liu, Linghe Kong, Guihai Chen, Yulun Zhang, and Xiaokang Yang. Low-bit model quantization for deep neural networks: A survey. *arXiv preprint arXiv:2505.05530*, 2025.

[52] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[53] Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. Spinquant: Llm quantization with learned rotations. In *The Thirteenth International Conference on Learning Representations*, 2024.

[54] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`.

[55] Wei Chen Maggie, Phil Culliton. Tweet sentiment extraction, 2020. URL `https://kaggle.com/competitions/tweet-sentiment-extraction`.

[56] Vladimir Malinovskii, Denis Mazur, Ivan Ilin, Denis Kuznedelev, Konstantin Burlachenko, Kai Yi, Dan Alistarh, and Peter Richtarik. Pv-tuning: Beyond straight-through estimation for extreme llm compression. *Advances in Neural Information Processing Systems*, 37:5074–5121, 2024.

[57] Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D Bagdanov, and Joost Van De Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *International Conference on Machine Learning*, pages 43177–43199. PMLR, 2025.

[58] Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[59] Tommaso Mencattini, Robert Adrian Minut, Donato Crisostomi, Andrea Santilli, and Emanuele Rodolà. Merge^3: Efficient evolutionary merging on consumer-grade gpus. In *International Conference on Machine Learning*, pages 43694–43715. PMLR, 2025.

[60] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, 2023.

[61] 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.

[62] Yury Nahshan, Brian Chmiel, Chaim Baskin, Evgenii Zheltonozhskii, Ron Banner, Alex M. Bronstein, and Avi Mendelson. Loss aware post-training quantization. *Machine Learning*, 110 (11):3247–3262, 2021.

[63] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. URL http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf.

[64] Giorgos Nikolaou, Tommaso Mencattini, Donato Crisostomi, Andrea Santilli, Yannis Panagakis, and Emanuele Rodolà. Language models are injective and hence invertible. In *The Fourteenth International Conference on Learning Representations*, 2026.

[65] Maria-Elena Nilsback and Andrew Zisserman. Automated Flower Classification over a Large Number of Classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, December 2008. doi: 10.1109/ICVGIP.2008.47. URL https://ieeexplore.ieee.org/document/4756141.

[66] James O'Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. I wish I would have loved this one, but I didn't – a multilingual dataset for counterfactual detection in product review. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7092–7108, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.568. URL https://aclanthology.org/2021.emnlp-main.568.

[67] Andrew Or, Apurva Jain, Daniel Vega-Myhre, Jesse Cai, Charles David Hernandez, Zhenrui Zhang, Driss Guessous, Vasiliy Kuznetsov, Christian Puhrsch, Mark Saroufim, and Supriya Rao. TorchAO: Pytorch-native training-to-serving model optimization. In *Championing Open-source DEvelopment in ML Workshop @ ICML25*, 2025. URL https://openreview.net/forum?id=HpqH0JakHf.

[68] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, June 2012. doi: 10.1109/CVPR.2012.6248092. URL https://ieeexplore.ieee.org/document/6248092. ISSN: 1063-6919.

[69] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[71] Robin Rombach. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022.

[72] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[73] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404.

[74] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170.

[75] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In *The 2011 International Joint Conference on Neural Networks*, pages 1453–1460, July 2011. doi: 10.1109/IJCNN.2011.6033395. URL https://ieeexplore.ieee.org/document/6033395/?arnumber=6033395. ISSN: 2161-4407.

[76] George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with svd to tie the knots. In *International Conference on Learning Representations*, 2025.

[77] Soroush Tabesh, Mher Safaryan, Andrei Panferov, Alexandra Volkova, and Dan Alistarh. Cage: Curvature-aware gradient estimation for accurate quantization-aware training. *arXiv preprint arXiv:2510.18784*, 2025.

[78] Irene Tallini, Daniele Solombrino, Alberto Cazzaniga, and Emanuele Rodolà. Weight pair encoding: Inducing a smaller grammar in neural network weights. *arXiv preprint arXiv:2609.31564*, 2026.

[79] Yujia Tong, Jingling Yuan, Tian Zhang, Jianquan Liu, and Chuang Hu. Dfq-vit: Data-free quantization for vision transformers without fine-tuning. *arXiv preprint arXiv:2507.14481*, 2025.

[80] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pages 516–533. Springer, 2022.

[81] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation Equivariant CNNs for Digital Pathology. In Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 210–218, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00934-2. doi: 10.1007/978-3-030-00934-2_24.

[82] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

[83] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510, 2021.

[84] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[85] Zheng Wang, Juncheng B. Li, Shuhui Qu, Florian Metze, and Emma Strubell. SQuAT: Sharpness- and quantization-aware training for BERT. *arXiv preprint arXiv:2210.07171*, 2022.

[86] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[87] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[88] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy

without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022.

[89] Zhuguanyu Wu, Jiaxin Chen, Hanwen Zhong, Di Huang, and Yunhong Wang. Adalog: Post-training quantization for vision transformers with adaptive logarithm quantizer. In *European Conference on Computer Vision*, pages 411–427. Springer, 2024.

[90] Guangxuan Xiao, Ji Lin, Mickael 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*, pages 38087–38099. PMLR, 2023.

[91] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms, September 2017. URL `http://arxiv.org/abs/1708.07747`. arXiv:1708.07747 [cs, stat].

[92] Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. SUN Database: Exploring a Large Collection of Scene Categories. *International Journal of Computer Vision*, 119(1):3–22, August 2016. ISSN 1573-1405. doi: 10.1007/s11263-014-0748-y. URL `https://doi.org/10.1007/s11263-014-0748-y`.

[93] Meng Xin, Sweta Priyadarshi, Jingyu Xin, Bilal Kartal, Aditya Vavre, Asma Kuriparambil Thekkumpate, Zijia Chen, Ameya Sunil Mahabaleshwarkar, Ido Shahaf, Akhiad Bercovich, Kinjal Patel, Suguna Varshini Velury, Chenjie Luo, Zhiyu Cheng, Jenny Chen, Chen-Han Yu, Wei Ping, Oleg Rybakov, Nima Tajbakhsh, Oluwatobi Olabiyi, Dusan Stosic, Di Wu, Song Han, Eric Chung, Sharath Turuvekere Sreenivas, Bryan Catanzaro, Yoshi Suhara, Tijmen Blankevoort, and Huizi Mao. Quantization-aware distillation for NVFP4 inference accuracy recovery, 2026. URL `https://arxiv.org/abs/2601.20088`.

[94] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

[95] Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[96] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[97] Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, pages 57755–57775. PMLR, 2024.

[98] Zhihang Yuan, Chenhao Xue, Yiqi Chen, Qiang Wu, and Guangyu Sun. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European conference on computer vision*, pages 191–207. Springer, 2022.

[99] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[100] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pages 1015–1022. IEEE, 2019.

[101] Luca Zhou, Daniele Solombrino, Donato Crisostomi, Maria Sofia Bucarelli, Giuseppe Alessio D'Inverno, Fabrizio Silvestri, and Emanuele Rodolà. On task vectors and gradients. In *UniReps: 3rd Edition of the Workshop on Unifying Representations in Neural Models*, 2025.

A Limitations

Our study represents a foundational step toward transferring quantization robustness, yet certain aspects require further exploration. Theorem 1 addresses the well-defined nature of the displacement itself but does not solve the separate problem of selecting λ without held-out calibration data.

The validation sweep over λ suggests that a significant portion of the useful transfer signal lies in the QV direction, yet it does not establish universality. In practical deployment scenarios, a minimal held-out calibration set should be used rather than the test set. Furthermore, to isolate the effect of the QV without introducing confounding variables, we intentionally employed a basic PTQ setup. Patched weights remain fully compatible with stronger PTQ methods such as AWQ and GPTQ, although in this setting the benefits of QV patching are best obtained by calibrating the scale λ (Appendix G.1). Finally, since a QV is a displacement in parameter space, donor and receiver must share the same architecture and pretrained initialization; transferring QVs across architectures or initializations is left for future work. We acknowledge these limitations and view them as opportunities for potential future work.

B Theorems and proofs

B.1 Proof of Proposition 1

The proof is exact line minimization of a positive-definite quadratic in the receiver's local geometry.

Proof. Because $\theta_{\mathcal{R}}$ is fixed, it is enough to optimize over donor-induced displacements of the form

$$\delta(\lambda) = \lambda \rho_{\mathcal{D}}.$$

Define the $H_{\mathcal{R}}$-inner product and norm by

$$\langle u, v \rangle_{H_{\mathcal{R}}} := u^\top H_{\mathcal{R}} v, \qquad \|u\|^2_{H_{\mathcal{R}}} := u^\top H_{\mathcal{R}} u.$$

Under the local quadratic model,

$$g_{\mathcal{R}}(\delta) = g_{\mathcal{R}}(\rho_{\mathcal{R}}) + \frac{1}{2}\|\delta - \rho_{\mathcal{R}}\|^2_{H_{\mathcal{R}}},$$

so along the donor line we have

$$g_{\mathcal{R}}(\lambda\rho_{\mathcal{D}}) = g_{\mathcal{R}}(\rho_{\mathcal{R}}) + \frac{1}{2}\|\lambda\rho_{\mathcal{D}} - \rho_{\mathcal{R}}\|^2_{H_{\mathcal{R}}}.$$

Expanding the square gives

$$g_{\mathcal{R}}(\lambda\rho_{\mathcal{D}}) = g_{\mathcal{R}}(\rho_{\mathcal{R}}) + \frac{1}{2}\Big(\lambda^2\|\rho_{\mathcal{D}}\|^2_{H_{\mathcal{R}}} - 2\lambda\langle\rho_{\mathcal{D}}, \rho_{\mathcal{R}}\rangle_{H_{\mathcal{R}}} + \|\rho_{\mathcal{R}}\|^2_{H_{\mathcal{R}}}\Big).$$

Since $H_{\mathcal{R}} \succ 0$ and $\rho_{\mathcal{D}} \neq 0$, this is a strictly convex quadratic function of λ, so it has a unique minimizer. Differentiating with respect to λ and setting the derivative to zero yields

$$\lambda^\star = \frac{\langle\rho_{\mathcal{D}}, \rho_{\mathcal{R}}\rangle_{H_{\mathcal{R}}}}{\|\rho_{\mathcal{D}}\|^2_{H_{\mathcal{R}}}} = \frac{\rho_{\mathcal{D}}^\top H_{\mathcal{R}} \rho_{\mathcal{R}}}{\rho_{\mathcal{D}}^\top H_{\mathcal{R}} \rho_{\mathcal{D}}}.$$

Substituting $\lambda^\star$ back into the quadratic gives

$$g_{\mathcal{R}}(\lambda^\star\rho_{\mathcal{D}}) = g_{\mathcal{R}}(\rho_{\mathcal{R}}) + \frac{1}{2}\left(\|\rho_{\mathcal{R}}\|^2_{H_{\mathcal{R}}} - \frac{\langle\rho_{\mathcal{D}}, \rho_{\mathcal{R}}\rangle^2_{H_{\mathcal{R}}}}{\|\rho_{\mathcal{D}}\|^2_{H_{\mathcal{R}}}}\right).$$

Now the receiver-side QAT gain relative to leaving the receiver unpatched is

$$g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\rho_{\mathcal{R}}) = \frac{1}{2}\|\rho_{\mathcal{R}}\|^2_{H_{\mathcal{R}}},$$

while the gain obtained by the best donor patch is

$$g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\lambda^\star\rho_{\mathcal{D}}) = \frac{1}{2}\frac{\langle\rho_{\mathcal{D}}, \rho_{\mathcal{R}}\rangle^2_{H_{\mathcal{R}}}}{\|\rho_{\mathcal{D}}\|^2_{H_{\mathcal{R}}}}.$$

Therefore, the fraction of receiver-side QAT gain recovered by the best donor patch is

$$\frac{g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\lambda^\star \rho_{\mathcal{D}})}{g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\rho_{\mathcal{R}})} = \frac{\langle \rho_{\mathcal{D}}, \rho_{\mathcal{R}} \rangle_{H_{\mathcal{R}}}^2}{\|\rho_{\mathcal{D}}\|_{H_{\mathcal{R}}}^2 \, \|\rho_{\mathcal{R}}\|_{H_{\mathcal{R}}}^2} = \cos_{H_{\mathcal{R}}}^2(\rho_{\mathcal{D}}, \rho_{\mathcal{R}}).$$

The three conclusions now follow immediately.

First, donor transfer helps if and only if the recovered fraction is strictly positive, which holds if and only if

$$\langle \rho_{\mathcal{D}}, \rho_{\mathcal{R}} \rangle_{H_{\mathcal{R}}} \neq 0,$$

that is, if and only if $\rho_{\mathcal{D}}$ is not $H_{\mathcal{R}}$-orthogonal to $\rho_{\mathcal{R}}$.

Second, donor transfer matches receiver-side QAT if and only if the recovered fraction equals 1, i.e. if and only if

$$\cos_{H_{\mathcal{R}}}^2(\rho_{\mathcal{D}}, \rho_{\mathcal{R}}) = 1.$$

By equality in the Cauchy–Schwarz inequality for the $H_{\mathcal{R}}$-inner product, this happens if and only if $\rho_{\mathcal{D}}$ is a nonzero scalar multiple of $\rho_{\mathcal{R}}$.

Third, donor transfer never exceeds receiver-side QAT because

$$\cos_{H_{\mathcal{R}}}^2(\rho_{\mathcal{D}}, \rho_{\mathcal{R}}) \leq 1,$$

again by Cauchy–Schwarz. □

B.2 First deviation in local displacement

Proposition 2 (The cosine-squared law is second-order accurate)**.** *Let $g_{\mathcal{R}}$ be twice continuously differentiable near $\rho_{\mathcal{R}}$, assume*

$$\nabla g_{\mathcal{R}}(\rho_{\mathcal{R}}) = 0, \qquad H_{\mathcal{R}} := \nabla^2 g_{\mathcal{R}}(\rho_{\mathcal{R}}) \succ 0,$$

and let $\lambda^\star$ be the donor scale from Proposition 1. Suppose that $\nabla^2 g_{\mathcal{R}}$ is $L_{\mathcal{R}}$-Lipschitz on a neighborhood containing the line segments from $\rho_{\mathcal{R}}$ to 0 and to $\lambda^\star \rho_{\mathcal{D}}$. Then

$$g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\lambda^\star \rho_{\mathcal{D}}) = \cos_{H_{\mathcal{R}}}^2(\rho_{\mathcal{D}}, \rho_{\mathcal{R}}) \big[g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\rho_{\mathcal{R}}) \big] + \varepsilon_{\mathcal{R}},$$

with

$$|\varepsilon_{\mathcal{R}}| \leq \frac{L_{\mathcal{R}}}{6} \left(\|\rho_{\mathcal{R}}\|^3 + \|\lambda^\star \rho_{\mathcal{D}} - \rho_{\mathcal{R}}\|^3 \right),$$

where $\|\cdot\|$ denotes the Euclidean norm.

In particular, the first deviation from Proposition 1 is cubic in the local displacement scale.

Proof. Write

$$H := H_{\mathcal{R}}, \qquad \rho := \rho_{\mathcal{R}}, \qquad \eta := \rho_{\mathcal{D}}, \qquad \delta^\star := \lambda^\star \eta.$$

Since $\nabla g_{\mathcal{R}}(\rho) = 0$, Taylor's theorem with integral remainder around ρ gives, for every δ on the relevant line segments,

$$g_{\mathcal{R}}(\delta) = g_{\mathcal{R}}(\rho) + \frac{1}{2} (\delta - \rho)^\top H (\delta - \rho) + r(\delta),$$

where

$$r(\delta) = \int_0^1 (1 - t) \, (\delta - \rho)^\top \big(\nabla^2 g_{\mathcal{R}}(\rho + t(\delta - \rho)) - H \big) (\delta - \rho) \, dt.$$

Because $\nabla^2 g_{\mathcal{R}}$ is $L_{\mathcal{R}}$-Lipschitz,

$$\big\| \nabla^2 g_{\mathcal{R}}(\rho + t(\delta - \rho)) - H \big\| \leq L_{\mathcal{R}} \, t \, \|\delta - \rho\|.$$

Therefore

$$|r(\delta)| \leq \int_0^1 (1 - t) \, L_{\mathcal{R}} \, t \, \|\delta - \rho\|^3 \, dt = \frac{L_{\mathcal{R}}}{6} \, \|\delta - \rho\|^3.$$

Applying this expansion at $\delta = 0$ and $\delta = \delta^\star$ yields

$$g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\delta^\star) = \frac{1}{2} \|\rho\|_H^2 - \frac{1}{2} \|\delta^\star - \rho\|_H^2 + r(0) - r(\delta^\star),$$

where $\|u\|_H^2 := u^\top H u$.

Now let

$$c := \cos_H^2(\eta, \rho).$$

By Proposition 1, the quadratic part satisfies

$$\frac{1}{2}\|\rho\|_H^2 - \frac{1}{2}\|\delta^\star - \rho\|_H^2 = c\,\frac{1}{2}\|\rho\|_H^2.$$

Also,

$$g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\rho) = \frac{1}{2}\|\rho\|_H^2 + r(0).$$

Substituting this into the previous identity gives

$$g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\delta^\star) = c\,[g_{\mathcal{R}}(0) - g_{\mathcal{R}}(\rho)] + \varepsilon_{\mathcal{R}},$$

where

$$\varepsilon_{\mathcal{R}} = (1-c)\,r(0) - r(\delta^\star).$$

Since $0 \le c \le 1$,

$$|\varepsilon_{\mathcal{R}}| \le |r(0)| + |r(\delta^\star)|.$$

Using the remainder bound proved above at $\delta = 0$ and $\delta = \delta^\star$, we obtain

$$|\varepsilon_{\mathcal{R}}| \le \frac{L_{\mathcal{R}}}{6}\left(\|\rho\|^3 + \|\delta^\star - \rho\|^3\right),$$

which is exactly the claimed bound. □

B.3 Quantization vectors are well-defined

Throughout this subsection, $g^{\odot 2}$ denotes elementwise squaring, and all divisions are coordinatewise.

Theorem 2 (AdamW kills continuous linear reparameterizations)**.** *Consider an AdamW update of the form*

$$m^+ = \beta_1 m + (1-\beta_1)g, \qquad v^+ = \beta_2 v + (1-\beta_2)g^{\odot 2}, \qquad \theta^+ = a_t\theta - b_t\frac{m^+}{\sqrt{v^+} + \varepsilon\mathbf{1}},$$

where $a_t > 0$, $b_t > 0$, $\beta_1, \beta_2 \in (0,1)$, and $\varepsilon > 0$.

Suppose an invertible linear map $S \in GL_n(\mathbb{R})$ on parameter coordinates is an exact linear equivariance of this update, in the sense that there exists a linear map R on second-moment coordinates such that, for all θ, m, v, g,

$$U_t\big(S\theta,\ S^{-\top}m,\ Rv;\ S^{-\top}g\big) = \big(S\theta^+,\ S^{-\top}m^+,\ Rv^+\big).$$

Then S is a signed permutation matrix. In particular, AdamW cannot admit a nontrivial continuous family of exact linear reparameterization symmetries.

Proof. Let $T := S^{-\top}$. Equivariance of the second-moment update gives

$$R\big(g^{\odot 2}\big) = (Tg)^{\odot 2} \qquad \text{for all } g \in \mathbb{R}^n.$$

We first show that every row of T has at most one nonzero entry. If row i had two nonzero entries $a = T_{ij}$ and $b = T_{ik}$ with $j \ne k$, then for

$$g = te_j + e_k$$

the i-th coordinate of the right-hand side would be

$$\big((Tg)^{\odot 2}\big)_i = (at+b)^2 = a^2t^2 + 2ab\,t + b^2,$$

while the left-hand side would be

$$\big(R(g^{\odot 2})\big)_i = \big(R(t^2e_j + e_k)\big)_i,$$

which is affine in t^2 and therefore contains no linear term in t. Since $ab \ne 0$, this is impossible. Hence, each row of T has at most one nonzero entry.

Because T is invertible, each column must also have exactly one nonzero entry. Therefore

$$T = DP$$

for some permutation matrix P and some diagonal matrix

$$D = \mathrm{diag}(d_1, \ldots, d_n)$$

with each $d_i \neq 0$. Since every vector in $\mathbb{R}^n_{\geq 0}$ can be written as $g^{\odot 2}$ for some g, the identity

$$R\big(g^{\odot 2}\big) = (Tg)^{\odot 2}$$

implies

$$R = D^2 P.$$

Now use equivariance of the parameter update. Since the scalar factor a_t commutes with every linear map, the adaptive term must satisfy

$$S\left(\frac{m}{\sqrt{v} + \varepsilon \mathbf{1}}\right) = \frac{Tm}{\sqrt{Rv} + \varepsilon \mathbf{1}} \qquad \text{for all } m, v.$$

Choose

$$m = P^{-1} e_i, \qquad v = t\, P^{-1} e_i \qquad (t \geq 0).$$

Then

$$Tm = d_i e_i, \qquad Rv = d_i^2 t\, e_i, \qquad S = T^{-\top} = D^{-1} P.$$

Therefore the i-th coordinate of the previous identity becomes

$$\frac{d_i^{-1}}{\sqrt{t} + \varepsilon} = \frac{d_i}{|d_i|\sqrt{t} + \varepsilon} \qquad \text{for all } t \geq 0.$$

Comparing the coefficients of $\sqrt{t}$ and the constant terms forces

$$|d_i| = 1.$$

Since this holds for every i, D has only ± 1 on the diagonal. Hence S is a signed permutation matrix. In particular, no nontrivial continuous linear family can satisfy the equivariance identity. □

Lemma 1 (The fake quantizer preserves the remaining finite gauge)**.** *Let $\mathcal{G}$ be the finite permutation gauge left by the backbone parameterization. For every quantized linear weight tensor W and every $g \in \mathcal{G}$,*

$$FQ(g \cdot W) = g \cdot FQ(W).$$

Consequently, the STE forward rule used in QAT,

$$W \longmapsto W + \big(FQ(W) - W\big)_{\mathrm{stopgrad}},$$

commutes with the same action of g.

Proof. The operator FQ used in this paper is built from row-wise absolute maxima, division by the resulting scales, element-wise rounding, element-wise clipping to a symmetric signed grid, and rescaling. Row and column permutations merely reorder entries and, in the row-wise case, the corresponding scales. Because the same symmetric quantization rule is then applied coordinate-wise, each of these operations commutes with the permutation actions in $\mathcal{G}$. Hence

$$FQ(g \cdot W) = g \cdot FQ(W).$$

The STE rule is affine in W and $FQ(W)$, so it commutes with the same action as well. □

Proof of Theorem 1. The proof follows a standard real-analytic zero-set argument; see also Nikolaou et al. [64] for a recent use of the same measure-zero template in a transformer setting.

Write the piecewise real-analytic assumption using a finite common partition of parameter space into full-dimensional cells on each of which both endpoint maps $\Phi_{\mathcal{D}}^{\mathrm{FT}}$ and $\Phi_{\mathcal{D}}^{\mathrm{QAT}}$ are real-analytic.

By Theorem 2, matched AdamW fine-tuning cannot preserve any nontrivial continuous linear gauge. In the ViT regime considered here, the only exact symmetries that remain are the finite hidden-unit and

attention-head relabelings collected in $\mathcal{G}$. By Lemma 1, the QAT fake quantizer commutes with that same finite permutation gauge, so the FT and QAT endpoint maps live in a common discrete-gauge setting.

Fix $g \in \mathcal{G} \setminus \{e\}$. On each full-dimensional cell C, define

$$h_g(\theta) := \left\| \Phi_{\mathcal{D}}^{\mathrm{QAT}}(\theta) - g\, \Phi_{\mathcal{D}}^{\mathrm{FT}}(\theta) \right\|^2 .$$

By construction, h_g is real-analytic on C. By hypothesis, it is not identically zero on any such cell. Therefore its zero set on C,

$$Z_{g,C} := \left\{ \theta \in C : \Phi_{\mathcal{D}}^{\mathrm{QAT}}(\theta) = g\, \Phi_{\mathcal{D}}^{\mathrm{FT}}(\theta) \right\},$$

has Lebesgue measure zero.

Now consider the bad set

$$\mathcal{B}_D := \bigcup_{g \in \mathcal{G} \setminus \{e\}} \left\{ \theta : \Phi_{\mathcal{D}}^{\mathrm{QAT}}(\theta) = g\, \Phi_{\mathcal{D}}^{\mathrm{FT}}(\theta) \right\}.$$

This set is contained in the union of all cell boundaries and all zero sets $Z_{g,C}$. The cell boundaries have measure zero, and each $Z_{g,C}$ has measure zero. Since there are only finitely many cells and finitely many $g \in \mathcal{G}$, $\mathcal{B}_{\mathcal{D}}$ itself has measure zero.

Therefore, for every common initialization $\theta_{\mathrm{pre}} \notin \mathcal{B}_{\mathcal{D}}$, the matched endpoints

$$\theta_{\mathcal{D}} = \Phi_{\mathcal{D}}^{\mathrm{FT}}(\theta_{\mathrm{pre}}), \qquad \theta_{\mathcal{D},\mathrm{QAT}} = \Phi_{\mathcal{D}}^{\mathrm{QAT}}(\theta_{\mathrm{pre}})$$

lie in the same gauge, and the difference

$$\rho_{\mathcal{D}} = \theta_{\mathcal{D},\mathrm{QAT}} - \theta_{\mathcal{D}}$$

is a well-defined parameter-space displacement. □

C Additional Fine-Tuning Setup Details

Following Ilharco et al. [36], all weight-space arithmetic is performed between checkpoints obtained from a shared pretrained initialization. For each task and architecture, we train matched full-precision and QAT checkpoints with identical settings, including optimizer, learning-rate schedule, batch size, number of epochs, total number of training steps, and random seed. The only difference between the two runs is that QAT inserts fake-quantization modules into the targeted linear layers during training. Additionally, when a task-specific classification head is trained, both the full-precision and QAT models share the same randomly initialized head. We evaluate this protocol in both vision and text classification tasks. A complete list of datasets, along with their canonical citations and Hugging Face dataset identifiers, is provided here.

Vision classifiers. For image classification, we construct classifiers by pairing a pretrained visual backbone with a task-specific classification head. We evaluate two head regimes. In the *learned-head* regime, the classifier consists of a task-specific linear head initialized from a common random seed and trained jointly with the backbone. In the *CLIP-style* regime [70], the classifier uses a frozen bank of label-text embeddings to compute image-text logits, with only the visual backbone being fine-tuned. For Vision Transformers [20], we utilize `ViT-B/16`, `ViT-L/14`, and `ViT-H/14` under both the learned-head and CLIP-style regimes. In addition, for the learned-head regime, we also evaluate `Swin-B` and `Swin-L` [52], as well as `DeiT-III-B` and `DeiT-III-L` [80]. All vision models are fine-tuned at 224×224 resolution.

Text classifiers. We also apply the same fine-tuning protocol to text classification using pretrained text encoders. Specifically, we fine-tune `BERT-base`, `BERT-large` [17], `EmbeddingGemma-300M` [82], and `Qwen3-Embedding-0.6B` [99] as the encoder backbones, each paired with a task-specific classification head.

In both vision and text experiments, quantization-vector subtraction and donor patching are applied only to the shared backbone or encoder parameters. Task-specific heads are never included in the

Table 1: **Vision datasets.** Dataset names, citations, and Hugging Face identifiers used in the vision experiments.

Dataset	Hugging Face identifier
Stanford Cars [42]	`tanganke/stanford_cars`
CIFAR-10 [43]	`uoft-cs/cifar10`
CIFAR-100 [43]	`uoft-cs/cifar100`
DTD [10]	`tanganke/dtd`
EMNIST [14]	`tanganke/emnist_letters`
EuroSAT [33]	`tanganke/eurosat`
Fashion-MNIST [91]	`zalando-datasets/fashion_mnist`
FER2013 [32]	`clip-benchmark/wds_fer2013`
Flowers102 [65]	`dpdl-benchmark/oxford_flowers102`
Food-101 [3]	`ethz/food101`
GTSRB [75]	`tanganke/gtsrb`
ImageNet [72]	`ILSVRC/imagenet-1k`
KMNIST [12]	`tanganke/kmnist`
MNIST [45]	`ylecun/mnist`
Oxford-IIIT Pet [68]	`timm/oxford-iiit-pet`
PCam [81]	`1aurent/PatchCamelyon`
RESISC45 [8]	`tanganke/resisc45`
Rendered SST-2 [74]	`nateraw/rendered-sst2`
STL-10 [13]	`tanganke/stl10`
SUN397 [92]	`tanganke/sun397`
SVHN [63]	`ufldl-stanford/svhn`
Tiny ImageNet [44]	`zh-plus/tiny-imagenet`

quantization vector. Therefore, for a learned-head receiver, patching modifies the backbone or encoder while keeping the learned head fixed; for a CLIP-style receiver, patching modifies the visual backbone while keeping the frozen text classifier unchanged. Following the work of Ilharco et al. [36] and Gargiulo et al. [26], arithmetic operations are performed only between checkpoints that share compatible parameterizations: the same architecture, the same pretrained initialization, and the same head regime. This compatibility is important because the quantization vector represents a coordinate-wise displacement in the shared backbone or encoder parameter space.

Table 1 presents the vision classification datasets utilized in the experiments, along with their respective Hugging Face identifiers. For the SVHN dataset, the experiments use the `cropped_digits` configuration from `ufldl-stanford/svhn`. Additionally, Table 2 outlines the text classification datasets from MTEB [60, 21] and their Hugging Face identifiers.

Table 3 provides a detailed list of the exact vision backbone identifiers, frameworks, head configurations, and pretrained versions utilized in the experiments. For CLIP-style vision experiments, models are loaded with the `open_clip` framework [35]. In contrast, for learned-head vision experiments, models employ the `timm` framework [86]. Table 4 outlines the text encoders used for learned-head text classification, which are loaded with the `transformers` library [87].

Table 2: **Text datasets.** MTEB classification datasets, citations, and Hugging Face identifiers used in the text experiments.

Dataset	Hugging Face identifier
Emotion [73]	`mteb/emotion`
IMDB [54]	`mteb/imdb`
Banking77 [5]	`mteb/banking77`
AmazonReviewsClassification [40]	`mteb/AmazonReviewsClassification`
AmazonCounterfactualClassification [66]	`mteb/amazon_counterfactual`
MassiveIntentClassification [23]	`mteb/amazon_massive_intent`
MassiveScenarioClassification [23]	`mteb/amazon_massive_scenario`
MTOPDomainClassification [46]	`mteb/mtop_domain`
MTOPIntentClassification [46]	`mteb/mtop_intent`
ToxicConversationsClassification [11]	`mteb/toxic_conversations_50k`
TweetSentimentExtractionClassification [55]	`mteb/tweet_sentiment_extraction`

Table 3: **Vision model identifiers.** Exact vision backbone identifiers, head regimes, frameworks, and pretrained versions used in the experiments.

Head regime	Framework	Model identifier	Pretrained version
CLIP-style	`open_clip`	`ViT-B/16`	`laion2b_s34b_b88k`
CLIP-style	`open_clip`	`ViT-L/14`	`laion2b_s32b_b82k`
CLIP-style	`open_clip`	`ViT-H/14`	`laion2b_s32b_b79k`
Learned head	`timm`	`deit3_base_patch16_224`	`fb_in1k`
Learned head	`timm`	`deit3_large_patch16_224`	`fb_in1k`
Learned head	`timm`	`swin_base_patch4_window7_224`	`ms_in22k_ft_in1k`
Learned head	`timm`	`swin_large_patch4_window7_224`	`ms_in22k_ft_in1k`
Learned head	`timm`	`vit_base_patch16_224`	`orig_in21k`
Learned head	`timm`	`vit_large_patch16_224`	`orig_in21k`
Learned head	`timm`	`vit_huge_patch14_224`	`orig_in21k`

Table 4: **Text encoder identifiers.** Exact Hugging Face identifiers used for learned-head text classification.

Model	Hugging Face identifier
`BERT-base`	`google-bert/BERT-base-uncased`
`BERT-large`	`google-bert/BERT-large-uncased`
`EmbeddingGemma-300M`	`google/EmbeddingGemma-300M`
`Qwen3-Embedding-0.6B`	`Qwen/Qwen3-Embedding-0.6B`

D Compute Setup

Fine-tunings and evaluations were carried out using a mix of local and cloud machines. In all cases, single-GPU setups were used. The local machines were equipped with 64 GB of RAM, 8- or 16-core CPUs, and one of the following GPUs: an NVIDIA RTX 4090, RTX 3090, or RTX 3090 Ti. The cloud machines were equipped with 128 GB of RAM, 8 CPU cores, and an NVIDIA A100 (64 GB). Unless otherwise specified, all experiments were carried out using full precision (float32). The total compute budget was approximately 20,000 GPU-hours.

E Baselines

Figures 5, 6 and 7 report the absolute Top-1 accuracies for all the models, including `ViT-B/16`, ViT-L/16, and `ViT-H/14` under both the learned-head and CLIP-style regimes, as well as for the learned-head regime of `Swin-B`, `Swin-L`, `DeiT-III-B`, and `DeiT-III-L` for vision and `BERT-base`, `BERT-large`, `EmbeddingGemma`, and `Qwen3-Embedding` for text, used under three different conditions: standard fine-tuning (FT), PTQ to 3-bit, and 3-bit QAT. These baseline results serve two main purposes. First, they quantify the impact of the low-bit quantization regime addressed in this study. As shown by the images, across most datasets, PTQ induces a significant drop relative to full-precision (FT), confirming that 3-bit weight-only quantization is a challenging scenario in which robustness to quantization noise is not guaranteed. Second, the QAT results indicate that much of the performance lost due to PTQ can be regained when the model is explicitly trained to mitigate quantization effects. This establishes a notable gap between the results from PTQ and QAT, which our Quantization Vector aims to partially bridge through zero-shot transfer.

Additionally, the images also highlight that the quantization difficulty is highly task-dependent. Some datasets demonstrate relative stability under PTQ, while others suffer severe performance degradation, to the point of collapsing almost entirely, even though they perform strongly at FP. This variability motivates our donor-receiver analysis: if QAT-induced robustness can be transferred across tasks, then tasks that are particularly susceptible to PTQ may benefit from robustness directions extracted from more transferable donor tasks.

F Extended Results

In this appendix, we provide the full set of cross-task QV transfer results that complement the main analysis in Section 6. The main text focuses on representative heatmaps and aggregate trends, while here we report the corresponding results across all remaining architectures, head regimes, and modalities.

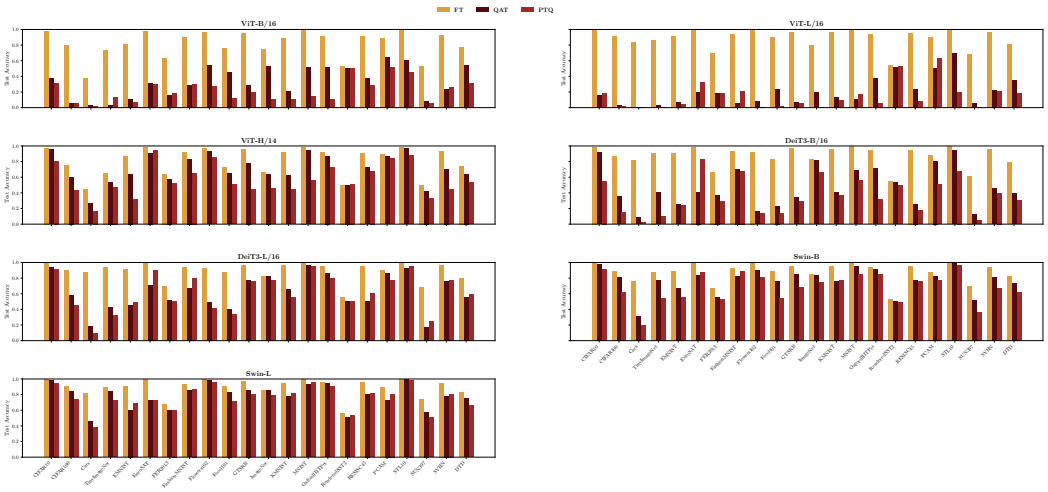

Figure 5: Comparison between accuracy of the full precision finetuned model (FT), the corresponding QAT and FP + PTQ, for the vision backbones used throughout the paper.

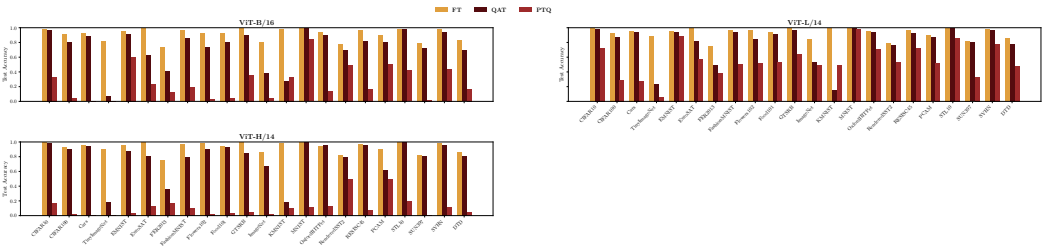

Figure 6: Comparison between accuracy of the full precision finetuned model (FT), the corresponding QAT and FP + PTQ, for the clip-style vision backbones used throughout the paper.

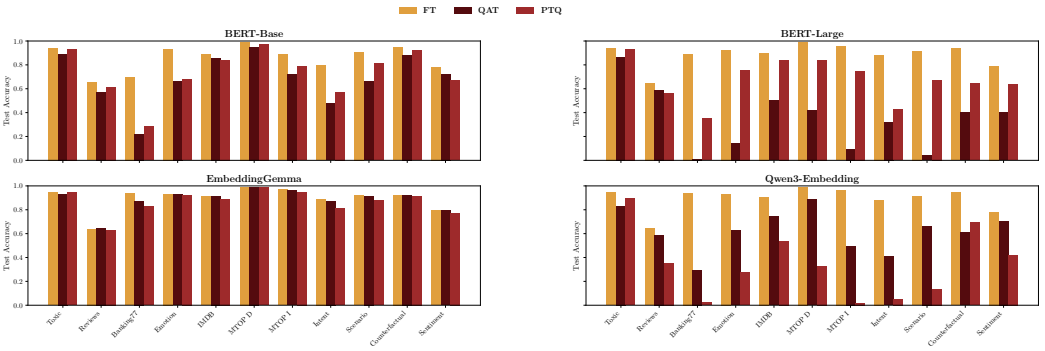

Figure 7: Comparison between accuracy of the full precision finetuned model, the corresponding QAT and FP + PTQ, for the text backbones used throughout the paper.

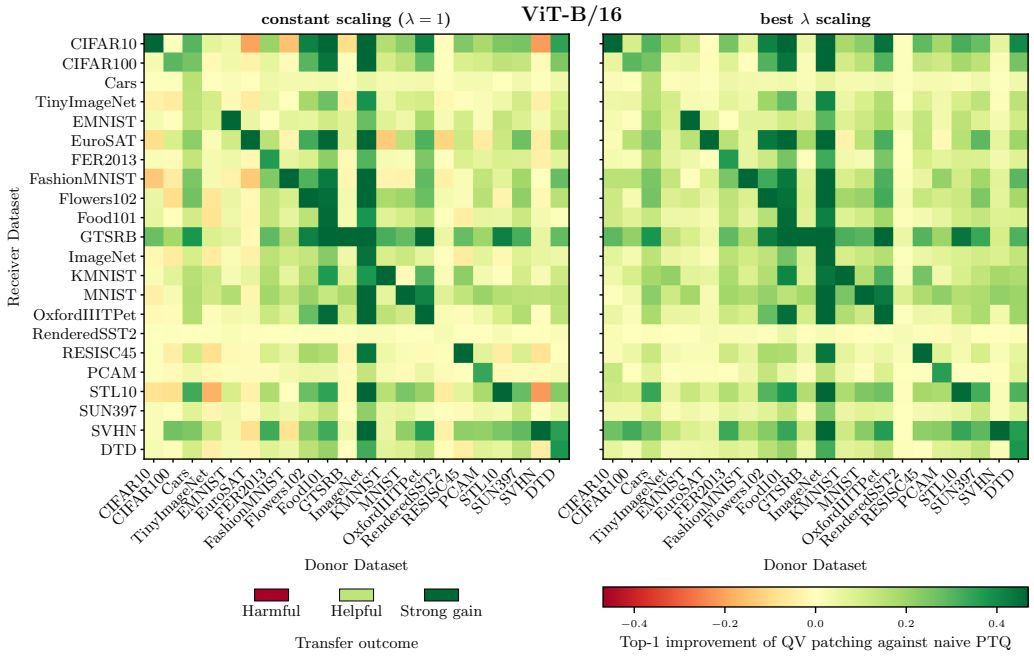

Figure 8: **QV transferability for learned-head `ViT-B/16` (vision).** Top-1 accuracy change (Δ) from patching receiver r with donor d quantization vector, relative to baseline 3-bit PTQ. For each architecture, heatmaps on the left show transfer using a constant scaling factor ($\lambda = 1$), revealing both positive transfer and destructive interference. Heatmaps on the right demonstrate that modulating the magnitude λ eliminates destructive interference while maximizing gains.

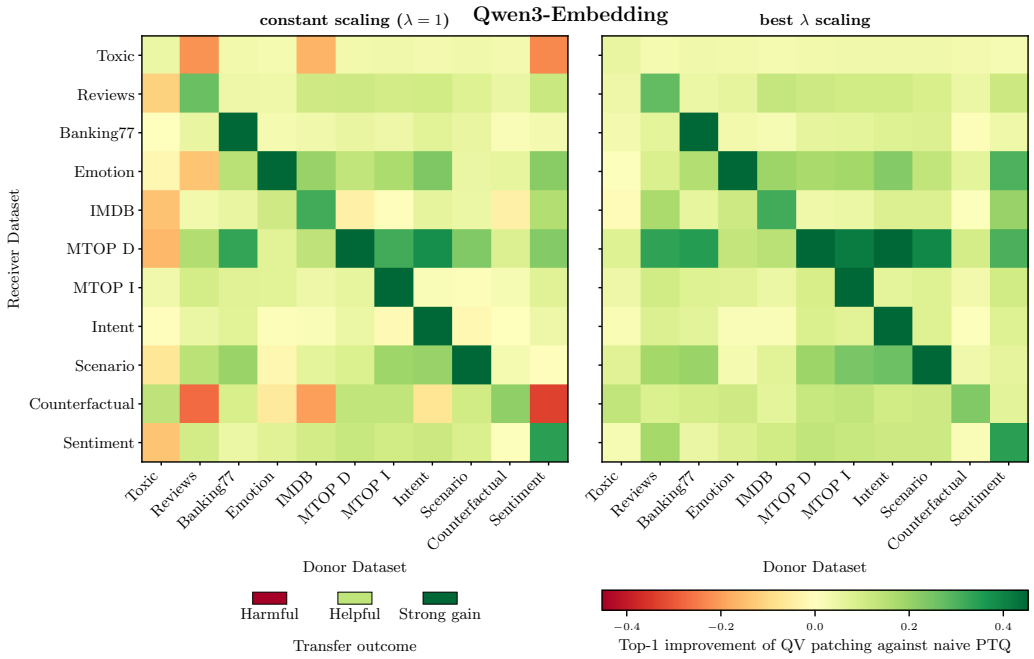

Figure 9: **QV transferability for `Qwen3-Embedding` (text).** Top-1 accuracy change (Δ) from patching receiver r with donor d quantization vector, relative to baseline 3-bit PTQ. For each architecture, heatmaps on the left show transfer using a constant scaling factor ($\lambda = 1$), revealing both positive transfer and destructive interference. Heatmaps on the right demonstrate that modulating the magnitude λ eliminates destructive interference while maximizing gains.

Unit-scale transfer across all models. For vision models, unit-scale QV transfer is broadly beneficial. In the CLIP-style setting, all reported backbones have positive mean transfer, with average gains ranging from $+18.7\%$ to $+43.7\%$ and positive-transfer rates above 92%. The strongest gains appear for the largest CLIP-style model, `ViT-H/14`, whose mean unit-scale transfer reaches $+43.7\%$. Learned-head vision models show the same qualitative trend, although with smaller and less uniform gains. Their mean unit-scale transfer remains positive for all architectures, ranging from $+1.3\%$ to $+9.7\%$, with positive-transfer rates between 60.8% and 76.6%. These results indicate that, in vision, the raw QV direction often already moves the receiver toward a more quantization-robust region of weight space.

The text models exhibit a more heterogeneous pattern at unit scale. `Qwen3-Embedding` shows strong positive transfer, with a mean gain of $+4.8\%$ and 78.2% on positive-transfer pairs. By contrast, `BERT-base`, `BERT-large`, and `EmbeddingGemma` have negative mean transfer at $\lambda = 1$. The strongest failure case is `BERT-large`, whose unit-scale mean transfer is -29.1%. This does not imply that text QVs are not transferable; rather, it shows that applying the donor displacement with a fixed unit magnitude can be unreliable in some text encoders. This observation is consistent with Proposition 1, which predicts that a useful donor direction may still hurt if its scale is not matched to the receiver geometry.

Effect of scale calibration. We next evaluate whether the negative transfer observed at $\lambda = 1$ can be mitigated by tuning only the scalar coefficient in Eq. 6. For each donor–receiver pair, we select the best coefficient from the same fixed grid used in the main text and then evaluate the patched receiver after 3-bit PTQ. We denote this selected coefficient by λ^{best}.

The right-hand heatmaps in Figures 8, 9, 10, 11, and 12 show the resulting transfer matrices after scale calibration on validation data. Compared with the unit-scale setting, the calibrated setting substantially reduces destructive interference and increases the reliability of transfer. This is especially visible in the text experiments, where several models that were negative on average at $\lambda = 1$ become positive after scale selection.

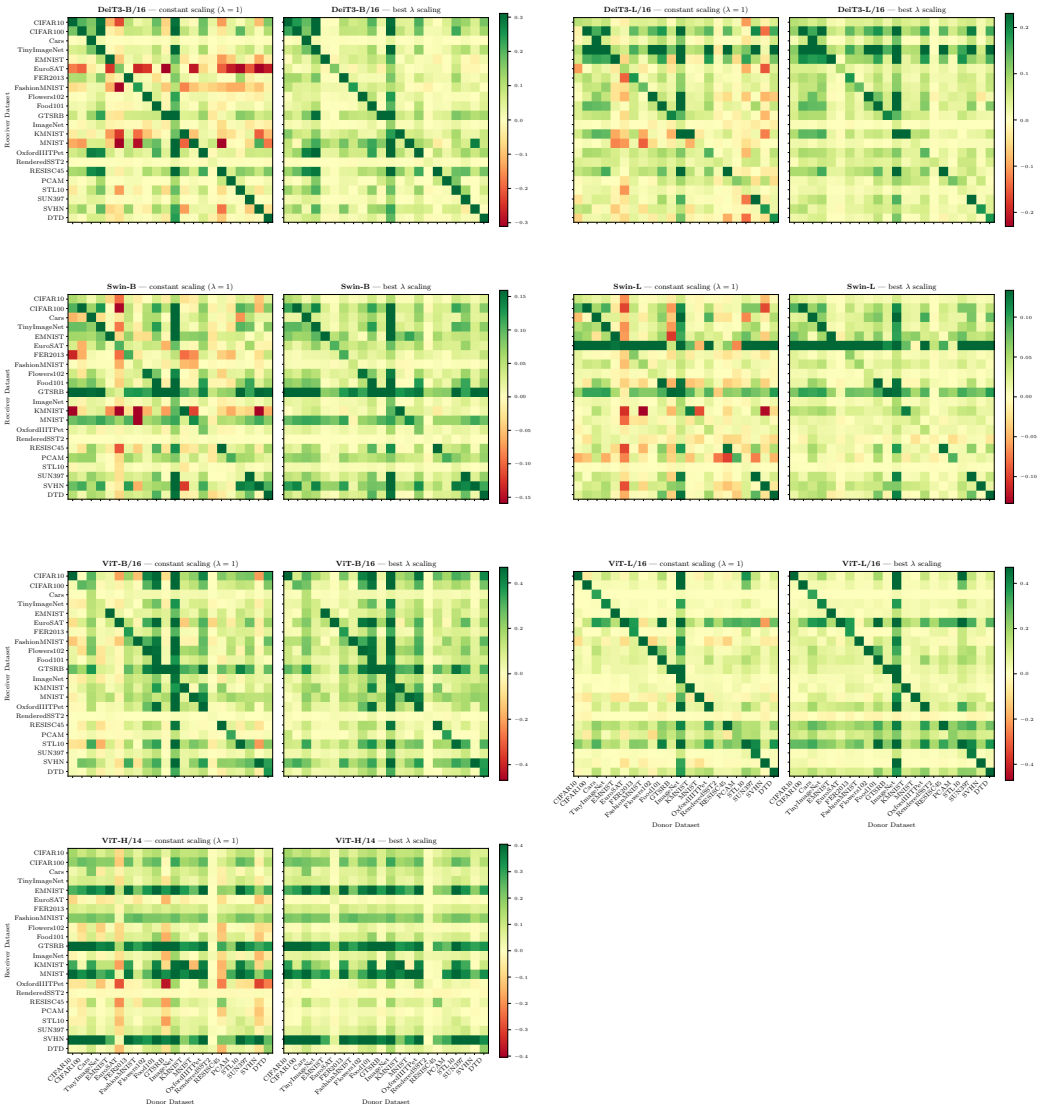

Figure 10: **QV transferability for vision (learned-head)**. Top-1 accuracy change (Δ) from patching receiver r with donor d quantization vector, relative to baseline 3-bit PTQ. For each architecture, heatmaps on the left show transfer using a constant scaling factor ($\lambda = 1$), revealing both positive transfer and destructive interference. Heatmaps on the right demonstrate that modulating the magnitude λ eliminates destructive interference while maximizing gains.

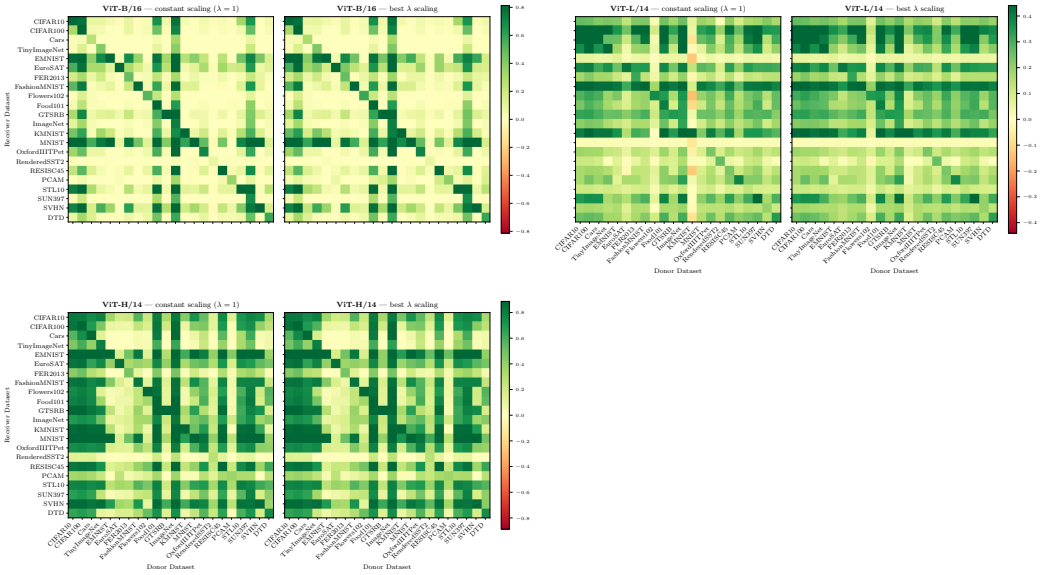

Figure 11: **QV transferability for vision (CLIP-style)**. Top-1 accuracy change (Δ) from patching receiver r with donor d quantization vector, relative to baseline 3-bit PTQ. For each architecture, heatmaps on the left show transfer using a constant scaling factor ($\lambda = 1$), revealing both positive transfer and destructive interference. Heatmaps on the right demonstrate that modulating the magnitude λ eliminates destructive interference while maximizing gains.

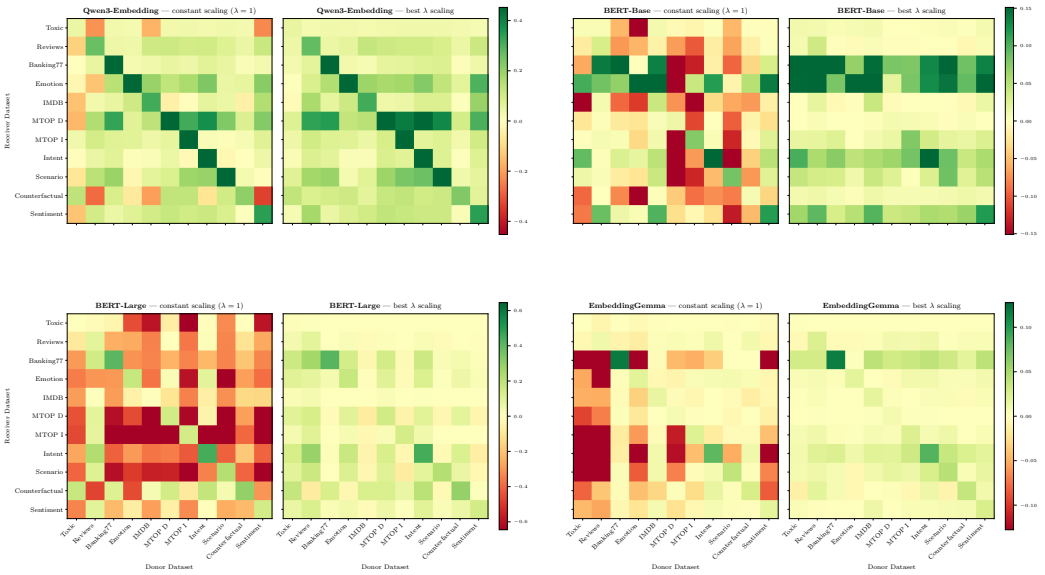

Figure 12: **QV transferability for text**. Top-1 accuracy change (Δ) from patching receiver r with donor d quantization vector, relative to baseline 3-bit PTQ. For each architecture, heatmaps on the left show transfer using a constant scaling factor ($\lambda = 1$), revealing both positive transfer and destructive interference. Heatmaps on the right demonstrate that modulating the magnitude λ eliminates destructive interference while maximizing gains.

In learned-head vision models, the mean gain becomes positive for every architecture and increases relative to unit-scale transfer. For example, DeiT3-B/16 improves from a mean unit-scale gain of $+1.7\%$ to $+6.1\%$, `Swin-B` from $+2.3\%$ to $+4.3\%$, `ViT-B/16` from $+9.7$ to $+13.2$, and `ViT-H/14` from $+8.5$ to $+11.2$. Positive-transfer rates also increase substantially, reaching between 87.7% and 95.5% across learned-head vision architectures.

The CLIP-style models remain the strongest regime. Scale calibration preserves the large gains already observed at unit scale and further improves reliability. The mean calibrated gains range from $+18.1\%$ to $+44.8\%$, with positive-transfer rates between 87.9% and 99.8%. In particular, `ViT-H/14` reaches a mean gain of $+44.8\%$, while `ViT-L/14` achieves positive transfer in 99.8% of donor–receiver pairs. These results suggest that CLIP-style visual backbones contain especially transferable quantization-robust directions.

The effect of scale calibration is most pronounced in the text setting. `BERT-base` changes from a negative mean transfer of -3.6% at unit scale to a positive mean transfer of $+3.9\%$, while its positive-transfer rate increases from 35.5% to 85.5%. `BERT-large` changes from -29.1% to $+4.0$, and its positive-transfer rate increases from 15.5% to 70.9%. `EmbeddingGemma` also becomes positive on average, although with a smaller mean gain of $+0.7\%$. `Qwen3-Embedding` remains the strongest text model, improving from $+4.8\%$ to $+10.0\%$ mean transfer and reaching 99.1% positive-transfer pairs. These results show that many negative unit-scale text transfers are not caused by an absence of transferable QV structure. To separate a wrong step size from a genuinely anti-aligned direction, we extend the λ sweep to negative values for `BERT-base` and `BERT-large`. The failures turn out to be mixed: for most failing pairs the best scale is positive but smaller than one, i.e. a magnitude mismatch, whereas for a minority, larger in `BERT-large`, the best scale is negative. The latter case is still consistent with Proposition 1, whose recoverable fraction $\cos^2_{H_{\mathcal{R}}}(\rho_{\mathcal{D}}, \rho_{\mathcal{R}})$ does not depend on the sign of the donor QV: a negative optimum only means that the donor direction must be applied with the opposite sign. Transfer is truly impossible only when the two QVs are nearly orthogonal in the receiver's geometry.

Amortized deployment cost. QV reuse incurs a one-time QAT cost on the donor, after which the resulting vector can be applied to multiple compatible receivers. In contrast, the standard approach requires a separate QAT run for every receiver. We quantify this difference by the number of training samples processed during QAT. For a dataset D, let

$$C_{\mathrm{QAT}}(D) = E_D |D_{\mathrm{train}}|$$

denote the QAT cost, where E_D is the number of QAT epochs and $|D_{\mathrm{train}}|$ is the number of training examples. Because donor and receiver runs use the same backbone and training procedure, this measure isolates the compute cost that differs between QV reuse and independent receiver-side QAT.

For N receiver deployments, the total QAT-related cost of QV reuse is

$$C_{\mathrm{reuse}}(N) = C_{\mathrm{QAT}}(D_{\mathrm{donor}}) + N c_{\mathrm{patch}},$$

where c_{patch} is the cost of applying the stored vector to one receiver. By comparison, performing QAT independently for each receiver costs

$$C_{\mathrm{separate}}(N) = \sum_{i=1}^{N} C_{\mathrm{QAT}}(D_i).$$

The break-even deployment count is therefore the smallest N for which QV reuse is no more expensive than independent receiver-side QAT:

$$N^\star = \min\{N \in \mathbb{N} : C_{\mathrm{reuse}}(N) \le C_{\mathrm{separate}}(N)\}\,.$$

If $\bar{C}$ denotes the mean QAT cost across receivers, this can be approximated as

$$N^\star \approx \left\lceil \frac{C_{\mathrm{QAT}}(D_{\mathrm{donor}})}{\bar{C} - c_{\mathrm{patch}}} \right\rceil.$$

In practice, applying a QV only requires a state-dict addition, whose cost is negligible relative to QAT, so $c_{\mathrm{patch}} \approx 0$. On the vision datasets, this yields a break-even point of $N \ge 6$ when ImageNet is used as the donor and $N \ge 3$ for the other evaluated donors. Thus, once the corresponding number of compatible receivers is reached, the initial donor QAT cost has been amortized; each additional receiver avoids an otherwise separate QAT run.

G Additional Experiments

G.1 Stronger PTQ Methods

We study how QV patching interacts with stronger post-training quantizers, namely AWQ [48] and GPTQ [24], on the learned-head `ViT-B/16` backbone with 3-bit symmetric per-channel weight-only quantization. Throughout, RTN denotes the fake-quantization operator of Eq. 3, i.e. the vanilla PTQ used in the rest of the paper.

QV patching is not a competitor to strong PTQ. AWQ and GPTQ are far stronger than RTN, and QV patching followed by RTN does not reach their accuracy on any donor–receiver pair. The contribution of QV patching is not state-of-the-art low-bit accuracy, but the demonstration that quantization robustness can be transferred to new tasks without receiver-side QAT.

Strong PTQ behaves normally on patched weights. A QV-patched checkpoint remains a valid input for strong quantizers: replacing RTN with AWQ or GPTQ on a patched receiver improves accuracy on virtually every cross-task pair, by a margin comparable to what these methods achieve on unpatched weights. The only exception is the same-task case, where the patch at unit scale coincides with the receiver's own QAT checkpoint: its weights already lie on the RTN grid, and methods that deliberately move weights off that grid remove the property QAT installed.

Does the QV help on top of AWQ or GPTQ? Comparing against the same strong quantizer applied to the unpatched receiver, unit-scale patching hurts under both AWQ and GPTQ. This is consistent with Proposition 1: changing the final quantizer changes the receiver's post-quantization objective, hence $H_{\mathcal{R}}$ and the optimal scale $\lambda^\star$. When λ is selected on receiver validation data, the QV yields a small but consistent gain on top of AWQ, and no meaningful gain on top of GPTQ, where improvements on most receivers are offset by a few large losses. The benefit of the QV is therefore bounded by the residual error left by the deployment quantizer: in the strictly data-free setting, QV patching does not help on top of AWQ or GPTQ, and the zero-shot claim of this paper refers only to RTN deployment.

AWQ and GPTQ displacements do not transfer. As an ablation, we replace the QAT-derived QV with the weight displacement that AWQ or GPTQ induces on the donor, and add it to a different task's checkpoint. Both transferred displacements degrade the receiver, both relative to its own AWQ/GPTQ result and relative to the unpatched receiver. We hypothesize that this is because AWQ and GPTQ are activation-space methods: although they output modified weights, their corrections are dictated by the donor's activation statistics on calibration data, through activation-aware channel scaling in AWQ and the input Hessian used for error compensation in GPTQ, and are therefore specific to the donor's input distribution. QAT, in contrast, shapes the weights directly in weight space, which is consistent with its displacement being the one that transfers across tasks.

G.2 Other Bit Widths

QV transfer is largely accuracy-preserving at both extremes of the bit-width range and strongly beneficial in the intermediate regime, where quantization causes substantial degradation that receiver-side QAT can still undo. This is consistent with Proposition 1, which bounds the gain recoverable by a donor patch by the receiver's own QAT gain. We obtain this picture by repeating the cross-task transfer setup of Section 6 at 2 and 4 bits, with the same quantization configuration (Section 5) and the strictly data-free unit-scale protocol ($\lambda = 1$). At 4 bits, where PTQ leaves little room for improvement (Section 6), QV transfer brings no aggregate gain, but nearly all patched receivers stay close to their vanilla PTQ accuracy and none collapses. At 3 bits, the regime studied in the main text, QV transfer yields the gains reported in Section 6. At 2 bits, vanilla PTQ collapses severely, destroying most of the full-precision accuracy, and even receiver-side QAT recovers only a limited share of it. Since a donor QV can at best transfer the robustness that QAT itself achieves, a weak QAT endpoint yields a weak QV: patching leaves the receiver near its poor PTQ baseline, and the transfer inherits the failure of QAT in this regime.

G.3 Empirical Test of Proposition 1

Proposition 1 predicts that transfer is governed by the alignment between donor and receiver QVs. We measure this alignment with the Euclidean cosine between $\rho_{\mathcal{D}}$ and $\rho_{\mathcal{R}}$ of Equation 5, computed on the linear-layer weights that PTQ acts on, and correlate it with the QV patching gain over all cross-task pairs of the setup of Section 6, at each λ of the grid. We find a clear positive correlation between donor–receiver QV alignment and QV patching gain at small λ, followed by a decreasing correlation as λ increases, crossing zero near $\lambda = 1$. We test the validity of this correlation by computing the cosine row-wise instead of globally, by leaving out each task in turn, and by a Quadratic Assignment Procedure test, which accounts for pairs sharing a task: the results are preserved in all cases and are statistically significant. This shows that well-aligned pairs benefit most from small λ, while poorly aligned pairs need larger ones; at $\lambda = 1$ the two effects cancel out, which explains the near-zero correlation there. Finally, we note that although raw cosine similarities are small, they lie far above what unrelated directions would produce in a space of this dimension, indicating that donor and receiver QVs share a non-trivial common component.

G.4 Study on the Quantization Error

A natural explanation for QV transfer is that the patch moves the weights closer to the quantization grid, making them easier to quantize. We test this directly by measuring the relative weight-space quantization error $\|W - \mathrm{FQ}(W)\|/\|W\|$ of patched receivers over all cross-task pairs of the setup of Section 6, both at unit scale and at the validation-selected λ. The error is essentially unchanged by patching and in the receiver's own QAT checkpoint, which is nonetheless far more robust to PTQ; in an ablation comparison, the QV affects this error no differently than a random direction of the same norm. QV patching, therefore, does not appear to work by reducing quantization error. We hypothesize that it instead changes where that error falls relative to the network function, steering the weights toward configurations whose quantization residual lies along directions to which the loss is less sensitive. This would be consistent with the local picture of Proposition 1, in which the post-quantization damage depends on the curvature of the loss along the error, not only on its magnitude; verifying it, e.g. through curvature-aware analyses, is left for future work.

G.5 Further Variants of QV Transfer

QAT optimizer. To check that QV transfer is not an artifact of the straight-through estimator [2] used for all other QAT runs in this paper, we rebuild the donor QVs with PV-Tuning [56], following its original specification and taking the full-precision weights it optimizes, before their final rounding onto the quantization grid, and repeat the cross-task transfer setup of Section 6 at unit scale ($\lambda = 1$). PV-Tuning reproduces the same transfer phenomenon almost exactly, with nearly identical gains and positive-transfer rates; the two QVs are highly aligned and have comparable norms. PV-Tuning thus does not improve transfer, but it shows that the QV is a property of quantization-aware training rather than of the particular implementation of QAT.

Staged schedules. The main experiments train the FT and QAT endpoints side by side from the pretrained backbone. We also tested a staged variant: quantization-aware training starts from an intermediate full-precision checkpoint, at 30, 60 or 90% of the full-precision run, and runs for the remaining steps with its own learning-rate schedule, so the total number of optimization steps stays fixed. The QV is then the displacement between the final quantization-aware and final full-precision checkpoints. Across the 22 vision tasks with `ViT-B/16` [20] at 3 bits and unit scale ($\lambda = 1$), QVs from staged donors transfer consistently, recovering a large share of the receiver's own QA gain. Within this set of runs, staged donors transfer markedly more strongly than donors trained quantization-aware from the pretrained backbone with the same budget, which suggests that QVs extracted closer to the full-precision solution are more reusable.

Quantization-aware distillation. Under the same staged schedules, we also trained donors with quantization-aware distillation (QAD) [34, 93], using the full-precision checkpoint as teacher. QVs from QAD donors transfer about as well as those from QAT donors.

Beyond classification. To check that QV transfer is not specific to discriminative models, we applied it to the decoder of the Stable Diffusion autoencoder (KL-f8) [71] finetuned on 12 image

domains: CelebA-HQ [39], ChartQA [58], chest X-rays [84], COCO [49], floor plans [22], Food-101 [3], LSUN Bedroom and Church [96], pixel art [7], PubLayNet [100], RESISC45 [8] and RICO [16, 83]. Under 3-bit weight PTQ, a QV from a different domain improves the receiver's reconstruction in most donor-receiver pairs and closes roughly half of the gap between vanilla PTQ and the receiver's own QAT. The gains carry over to 4-bit floating-point formats, and QAT and QAD donors are again indistinguishable.

H Societal Impacts

Positive societal impacts. By enabling zero-shot transfer of quantization robustness, this work reduces the compute, data, and engineering costs of deploying models at extremely low bit-widths. This can democratize efficient deployment for resource-constrained practitioners, lower the energy footprint of AI inference, and avoid the need to access sensitive task-specific data on the receiver side, a meaningful privacy advantage in domains like healthcare.

Negative societal impacts. Cheaper deployment of compact models could accelerate adoption in sensitive domains (e.g., surveillance, automated decision-making) without proportionate investment in safety or fairness evaluation. Because patching is zero-shot and data-free, practitioners may deploy models with silently degraded accuracy on critical tasks if transfer quality is not carefully validated. Extensions beyond vision classifiers could amplify these risks.

I Safeguards

The artifacts released with this work consist of quantization vectors, i.e. lightweight weight-space displacements, and the corresponding fine-tuned checkpoints for standard image classification tasks. These are derived from publicly available pretrained, well-established, and openly licensed classification models and datasets. As such, they do not introduce new risks of misuse beyond those already present in the underlying models and datasets. The quantization vectors themselves encode structural robustness to low-bit quantization rather than new task capabilities, and cannot be used in isolation to generate content or perform inference. We do *not* release large-scale generative models, scraped datasets, or personally identifiable data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The four main claims, namely (1) quantization vectors as transferable displacements, (2) well-definedness under symmetries, (3) cosine-squared recovery theory, and (4) empirical gains up to $\sim$60 points across 4 ViT scales and 22 tasks, are each substantiated: Theorem 1 (Section 4.1 and Appendix B.3) supports (2), Proposition 1 (Appendix B.1) supports (3), and the heatmaps and tables in Section 5 and Appendix F support (1) and (4).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper includes a dedicated "Limitations" section (Appendix A) that addresses several concrete limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: All theoretical results (Theorems 1–2, Propositions 1–2, Lemma 1) explicitly state their assumptions and are accompanied by complete proofs in the appendices (Appendices B.1, B.2, B.3), with proof sketches provided in the main text for the key results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Architectures, quantization scheme, optimizer, datasets, and other fundamental reproducibility details are presented in Section 5.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: All datasets used are publicly available, we provide the code in supplementary materials, and we plan to release code and checkpoints on GitHub upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: The full details for the reproduction of the experiments are provided in the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

   Justification: Given a fixed compute budget, we opted to invest it in breadth across multiple datasets and model scales rather than in multiple seeds per dataset. Error bars over different seeds would require substantial additional computation on each dataset; instead, the breadth of the evaluation provides robustness evidence.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix D provides details about compute resources used during the experimental phase.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The authors have reviewed the NeurIPS Code of Ethics and believe the manuscript is aligned with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

Justification: Section H shows potential positive and negative societal impacts stemming from this paper.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Section I describes safeguards that have been put in place to address modern AI concerns.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: Across the manuscript, we have provided all the citations to the original papers for the dataset used. We will point to the specific citation to provide the correct license to each specific case. For our released assets, we provide the license.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: Upon releasing code and a checkpoint on GitHub, we will include instructions for using the released assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs have been used just for writing, editing, or formatting purposes of the manuscript and portions of the experimental source code.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.