
Proxy Metrics Can Hide Downstream Collapse: A Case Study on Compressed Cellular Segmentation

Sebastián A. Cruz Romero*
Capicú Technologies
scruzromero@capicu.ai

Abstract

Post-training compression of vision models is commonly screened using output-level proxies such as normalized RMSE (NRMSE), cosine similarity, and flow angular error. We evaluated post-training quantization of Cellpose–SAM, a promptless Segment–Anything model for cellular instance segmentation, on a stratified 245-field panel spanning BBBC038, BBBC039, and NIST iPSC images across density regimes. These proxies failed to reliably predict downstream mask retention under activation quantization. Angular error saturated at 90° from A8 through A2, while cosine and NRMSE preserved rank-order information without providing an interpretable acceptance threshold. At W8A8-cal-QDQ, cosine similarity was 0.944 despite 89/245 empty masks; moreover, performance varied sharply by modality, with BBBC039 yielding 0/47 empty masks versus 63/99 for NIST iPSC fields at the same proxy footprint. Across five modality–density strata, cosine similarity and empty-mask rate showed weak rank agreement ($\rho \approx 0.3$), with the highest-cosine stratum producing the second-worst downstream outcome. Below A8, all modalities collapsed to 245/245 empty fields. By contrast, ternary weight quantization produced a $12.08\times$ storage reduction and was correctly identified as catastrophic by the proxies. These results show that FP32-referenced output proxies are insufficient as standalone acceptance criteria for promptless dense-prediction models; compression should instead be validated against downstream retention stratified by deployment-relevant imaging conditions.

1 Introduction

Large vision models are increasingly being deployed under strict compute, memory, and latency constraints, motivating post-training quantization (PTQ) as a practical route to efficient inference. Across convolutional and transformer architectures, PTQ workflows typically rely on numerical proxies—including weight reconstruction error, activation discrepancy, NRMSE, and cosine similarity relative to FP32 outputs—to screen candidate compression schemes before downstream evaluation [Jacob et al., 2018, Krishnamoorthi, 2018, Nagel et al., 2020, Li et al., 2021, Frantar et al., 2023, Dettmers et al., 2022, Yao et al., 2022, Lin et al., 2024, Ma et al., 2024b]. These proxies are attractive because they are efficient, architecture-agnostic, and independent of task-specific evaluation pipelines. Their implicit assumption is that numerical agreement with FP32 provides a reliable surrogate for downstream model behavior.

For dense prediction systems, however, this assumption remains largely untested. Segmentation foundation models used in scientific imaging do not only require accurate intermediate representations; they require those representations to remain compatible with downstream decoding procedures. Small numerical deviations may therefore be benign, while changes that alter decoder behavior

*Corresponding author.

can produce categorical failures despite apparently acceptable proxy values. This distinction is particularly important for cellular imaging workflows, where compressed models must operate across heterogeneous modalities and biological regimes.

Here, we evaluate on Cellpose–SAM [Stringer and Pachitariu, 2025], a promptless segmentation foundation model, under a deployment-oriented quantization protocol spanning 245 microscopy fields from BBBC038, BBBC039, and NIST iPSC datasets. We find that standard output-level proxies do not provide reliable deployment criteria for this setting. Under whole-graph activation quantization, flow angular error saturates at the same failure regime across activation widths, while cosine similarity and NRMSE retain numerical variation without yielding an interpretable acceptance threshold. More importantly, identical proxy footprints can produce sharply different downstream outcomes across imaging modalities: configurations that preserve segmentation quality in one biological context can collapse in another. These results indicate that compression safety for dense-prediction foundation models must be evaluated through downstream retention criteria stratified by deployment conditions, rather than by FP32-referenced similarity metrics alone.

2 Related Work

Post-training quantization and its proxy metrics. Modern PTQ recipes screen candidate schemes with cheap output-tensor proxies. Jacob et al. [2018] formalized integer-arithmetic quantized inference and introduced the reconstruction-error framing that has organized subsequent PTQ; Krishnamoorthi [2018] and Nagel et al. [2021] extended this into per-layer and per-tensor evaluation practices. Weight-side refinements including AdaRound [Nagel et al., 2020] and BRECQ [Li et al., 2021] minimize block-level reconstruction; on the transformer side, GPTQ [Frantar et al., 2023], LLM.int8() [Dettmers et al., 2022], ZeroQuant [Yao et al., 2022], AWQ [Lin et al., 2024], and the more recent 1.58-bit regime [Ma et al., 2024b] all rank candidates by activation- or weight-side reconstruction proxies. Sub-bit weight quantization studies [Rastegari et al., 2016, Hubara et al., 2016] follow the same convention. Gholami et al. [2022] surveys the field, and the shared assumption throughout is that small numerical distortion at the tensor level implies small downstream distortion at the task level. Our case study addresses the direction in which that assumption fails.

Cellpose and Segment–Anything foundation models. Cellpose introduced dynamics-based flow integration as the decoder for cellular instance segmentation [Stringer et al., 2021, Pachitariu and Stringer, 2022], and Cellpose–SAM [Stringer and Pachitariu, 2025] combines that decoder with the Segment–Anything visual encoder [Kirillov et al., 2023, Ravi et al., 2024]. The promptless flow-integration decoding path is what places Cellpose–SAM in the same design regime as other decoder-heavy biomedical segmentation foundation models entering laboratory workflows [Ma et al., 2024a], and it is also what makes it sensitive to a class of numerical failure that mask-token or end-to-end differentiable decoders do not necessarily exhibit [Weigert et al., 2020]. Broader cellular deep learning has documented the growing role of segmentation foundation models in imaging pipelines [Moen et al., 2019, Meijering, 2020].

Evaluation methodology and metric-validation guidance. The medical-image-analysis community has, over the past several years, produced explicit guidance on the pitfalls of scalar per-pixel proxies and pooled reporting. Maier-Hein et al. [2018] showed that biomedical image-analysis competition rankings are unstable when hidden metric choices are varied; Maier-Hein et al. [2024] formalize per-modality and per-task metric recommendations under the Metrics Reloaded framework; and Reinke et al. [2024] enumerates specific proxy-vs-downstream failure modes. Segmentation-metric sample-size and confidence-interval heuristics [Islam, 2018, Varoquaux and Colliot, 2022] and cluster-bootstrap inference [Efron and Tibshirani, 1994, DiCiccio and Efron, 1996] inform the retention criterion we use here. Pre-registration norms [Nosek et al., 2018] motivate fixing the retention margin before any hold-out data are examined (in our study the margin is pre-specified internally rather than externally registered), and reproducibility reporting norms [Pineau et al., 2021] inform the release path for audit artifacts that make the retention verdicts re-derivable on independent data.

3 Methods

Model. The compressed model in this study is Cellpose–SAM (checkpoint `cpsam_v2`, ~ 305 M parameters, FP32 storage 1,162.07 MiB), a promptless variant of the Segment–Anything family that emits per-pixel dY , dX , and cell-probability channels. Instance masks are produced by the dynamics-based seeding cascade introduced with Cellpose: local integration of the flow field yields attractor basins that are thresholded and reconciled into instance labels.

Panel. Evaluation is performed on a stratified public-dataset panel: BBBC038 nuclei ($n = 99$ fields spanning fluorescence, brightfield, and phase contrast [Caicedo et al., 2019, Ljosa et al., 2012]), BBBC039 U2OS fluorescence ($n = 47$ [Ljosa et al., 2012]), and NIST induced-pluripotent-stem-cell tiles stratified over low, medium, and high culture-density regimes ($n = 33$ per regime, 99 total, aggregated over 3 source-image experimental units [Bajcsy et al., 2018]). Whole-graph evaluations use all 245 hold-out fields per scheme; mask outcomes are reported against both the FP32 counterpart (proxy-vs-outcome discordance) and supplied ground truth. Cluster-bootstrap catastrophic-failure intervals for the weight-only and mixed schemes are computed on a 176-field subset (107 experimental units; BBBC038 $n = 70$, BBBC039 $n = 34$, NIST iPSC $n = 72$) whose per-modality sizes were pre-specified to target the segmentation-metric sample-size design of Islam [2018], Varoquaux and Colliot [2022]. The subset is smaller than the whole-graph evaluation panel because proxy diagnostics can consume the full panel cheaply while cluster-bootstrap on paired retention requires the more expensive full-mask decode.

Retention criterion. For each (scheme, dataset) pair we compute the mean change Δ in a downstream mask-level metric (paired instance F1 at IoU 0.5; AP at IoU $\{0.5, \dots, 0.95\}$) from FP32 [Lin et al., 2014] and form a 2,000-draw cluster-bootstrap 95% interval for $\mathbb{E}[\Delta]$ over experimental units. A scheme meets the criterion for that dataset if the entire interval lies above the pre-specified -0.02 margin. The margin was fixed before any hold-out evaluation. In addition, we report a per-scheme catastrophic-failure rate, the fraction of evaluation fields on which the compressed model’s mask output falls below an absolute quality floor, with an accompanying bootstrap 95% interval over experimental units.

Proxies. We study three complementary output-tensor proxies used across the PTQ literature. Per-field NRMSE is the L2 error against the FP32 reference on the concatenated dY , dX , cell-probability output normalized by FP32 L2 norm [Jacob et al., 2018, Nagel et al., 2021, Gholami et al., 2022]. Per-field cosine similarity is the cosine between the flattened compressed and FP32 outputs; and per-field median flow angular error is a decoder-aware proxy on the direction of the decoded flow field, computed as the median per-pixel angle between the compressed and FP32 flow vectors.

Compression schemes. We evaluate three weight-only schemes (W8A16, W4A16-G64 with group size 64, and ternary W2A16-G64); a calibrated whole-graph W8A8-cal-QDQ together with a whole-graph integer W8A8; and four low-bit-width whole-graph configurations (W8A4, W8A4 with softmax-A8 stabilization, W8A2, and W8A2 with softmax-A8). A mixed W4/W8 scheme retains at INT8 the four most sensitive operators identified by a one-operator-at-a-time W4 perturbation over 100 eligible Linear/Conv2d operators on the development split. W8A8-cal-QDQ was calibrated using five separate microscopy fields: BBBC038 nuclei, BBBC039 U2OS nuclei, and NIST iPSCs at low, medium, and high cell densities.

4 Results

Table 1 and Figure 2 report the weight-only bit-width sweep. On this path the three proxies and the downstream retention outcome are broadly monotone with bit width: median NRMSE rises from 0.008 at W8A16 to 0.095 at W4A16-G64 to 1.038 at W2A16-G64, cosine similarity falls from 1.000 to 0.997 to 0.423, and the bootstrap 95% catastrophic-failure interval widens from $[0.00, 0.00]$ to $[0.00, 0.001]$ to $[0.95, 1.00]$. On this path a standard PTQ workflow is defensible: a proxy screen at NRMSE ~ 1 would reject W2A16-G64 for the same reason the retention criterion does, and the $12.08\times$ storage reduction that ternary weights promise does not override the catastrophic-failure rate on this task, bootstrap point estimate 97.9% over experimental units with raw fraction $169/176 = 96.0\%$.

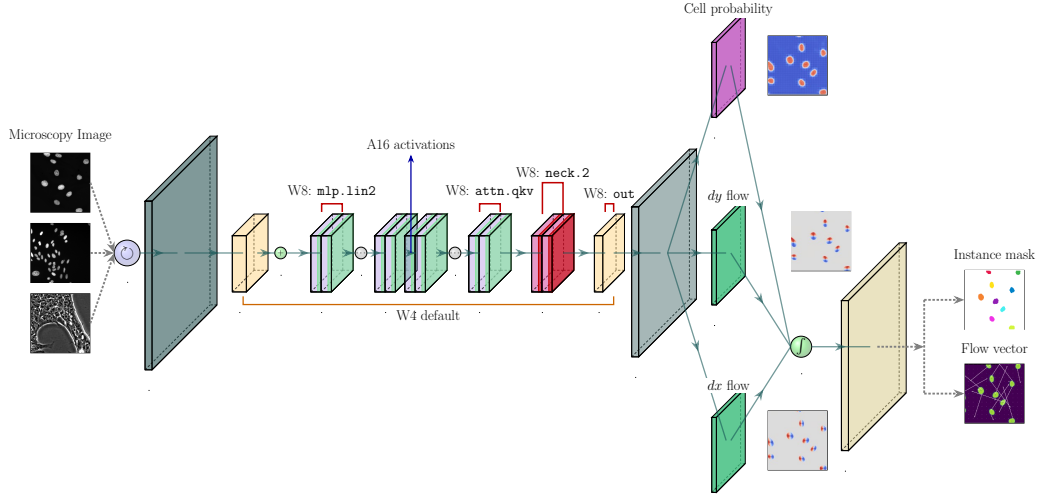


Figure 1: **Where the proxies live inside Cellpose-SAM.** A microscopy field (fluorescence, bright-field, or phase contrast) passes through the SAM visual encoder into the flow-integration decoder. The default weight schedule is W4 (orange bracket); illustrating as an example, the sensitivity-guided mixed scheme retains four operator classes at W8 (mlp.lin2, attn.qkv, neck.2, and out) and activations are held at A16. The decoder emits a cell-probability map and a (dY, dX) flow field; flow integration $\int$ produces the decoded instance mask. Output-tensor proxies (NRMSE, cosine similarity, median flow angular error) are computed on the raw pre-integration tensors marked here, i.e., before the seed-detection step of the flow-integration cascade where whole-graph W8A8 collapse localizes.

Table 1: Weight-only compression evaluation on the 176-field retention subset (107 experimental units). Compression statistics, reconstruction proxies, and catastrophic-failure rates are reported for each quantization scheme. Catastrophic rates use 2,000-draw cluster-bootstrap percentile intervals over experimental units.

Scheme	Compression		Reconstruction quality		
	Size (MiB)	Ratio	NRMSE	Range	Catastrophic rate [95% CI]
FP32 (reference)	1162.07	1.00×	—	—	—
W8A16	295.78	3.93×	0.008	0.002–0.059	0.000 [0.000, 0.028] [†]
W4A16-G64	168.47	6.90×	0.095	0.027–0.722	0.001 [0.000, 0.001]
W2A16-G64	96.21	12.08×	1.038	0.680–4.696	0.979 [0.949, 1.000]

[†] Zero observed failures; percentile bootstrap is replaced by the rule-of-three upper bound 3/107 over experimental units.

Table 2 together with Figures 3 and 4 reports the whole-graph path, and this is where the three proxies fail in complementary ways. Angular error is a resolution failure: median flow angular error is $\sim 90^\circ$ (uniform-random against FP32) at W8A8-cal-QDQ, W8A8-full-int, W8A4, and W8A2 alike, so at the pooled panel level the proxy cannot separate partial from categorical downstream collapse. Cosine and NRMSE are a threshold failure: cosine cleanly separates A8 (0.944) from A4 (~ 0.2) and A2 (~ 0.0), and NRMSE moves $\sim 2\times$ from 0.496 at W8A8-cal-QDQ to 1.100 at W8A2, so both carry rank-order information, but a cosine of 0.944 reads as excellent to a practitioner while coexisting with 89/245 empty masks, and neither has a defensible accept threshold near the operating regime. Downstream, at the W8A8-cal-QDQ proxy footprint the model emits zero instances on 89/245 cases with median AP_{50} (against FP32) = 0.100, a partial collapse. At W8A4 and W8A2 emission drops to 0 instances on 245/245 cases with median AP_{50} = 0.000, so at essentially the same angular-error proxy value the outcome shifts from partial to categorical. Also visible in Table 2 but unexplained by the calibration construction: W8A8-full-int has a median AP_{50} of 0.521 against 0.100 for the calibrated W8A8-cal-QDQ control on the same 245 fields, a $\sim 5\times$ reversal opposite to what the

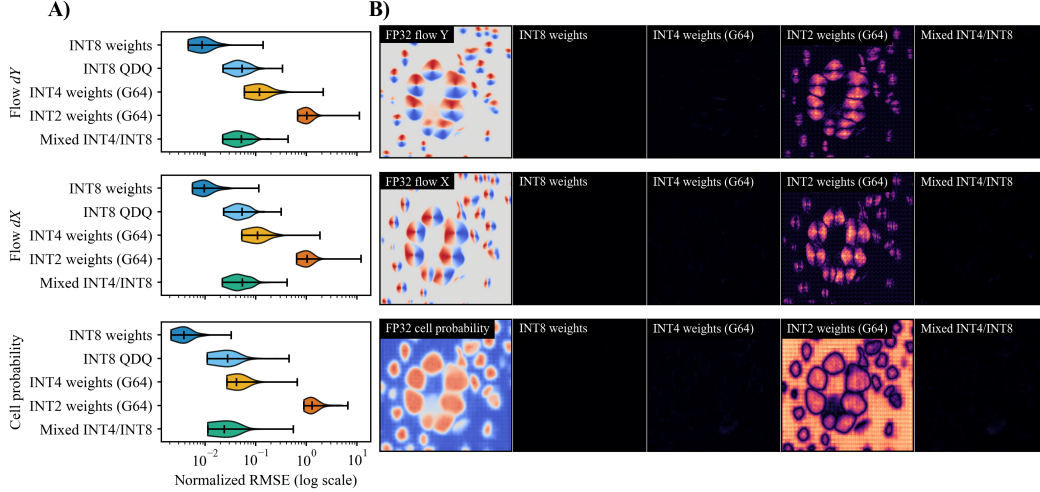


Figure 2: **Raw-output proxies on the weight-only path.** A) per-field raw-output normalized RMSE against FP32 (log scale) split by output channel (dY flow, dX flow, cell probability) for INT8 weights (W8A16), INT8-QDQ, INT4 weights with group size 64 (W4A16-G64), INT2 weights with group size 64 (W2A16-G64), and the sensitivity-guided mixed INT4/INT8 scheme. B) matched qualitative composite of FP32 flow-Y, flow-X, and cell-probability maps against the four quantized configurations on a representative field. NRMSE separates ternary weights from the other schemes but does not resolve the mixed INT4/INT8 candidate from INT4, motivating the downstream retention check reported in Table 1.

Table 2: Whole-graph activation quantization evaluation on 245 hold-out fields. Activation agreement is measured against the FP32 reference using NRMSE, cosine similarity, and flow angular error; downstream performance is reported by empty-mask frequency and instance detection quality.

Scheme	Activation agreement				Detection outcome	
	NRMSE	Range	Cosine	Flow err.	Empty masks	AP ₅₀
W8A8-cal-QDQ	0.496	0.11–2.69	0.944	90°	89/245	0.100
W8A8-full-int	0.372	0.09–2.36	0.963	90°	83/245	0.521
W8A4-full-int-S8	1.088	1.00–1.20	0.219	90°	245/245	0.000
W8A4-full-int	1.088	1.00–1.20	0.204	90°	245/245	0.000
W8A2-full-int-S8	1.100	1.00–1.22	0.000	90°	245/245	0.000
W8A2-full-int	1.100	1.00–1.22	0.000	90°	245/245	0.000

harsher rounding path of full-integer arithmetic on top of QDQ would predict. We report the ordering but do not resolve it in this study.

The discordance is biology-conditioned, and stratification recovers signal the pooled row of Table 2 hides. Table 3 stratifies the W8A8-cal-QDQ result by (dataset, regime); at the stratum level median angular error is 58° on BBBC039 and 87° on NIST high density against 90° on the other three strata, so the panel-level 90° figure is a pooling artifact and the proxy is not saturated everywhere. Even so, at the stratum level the proxy is non-monotone in the outcome. Sorted by cosine (best proxy first), the five strata rank as NIST low (0.972, $24/33 = 72.7\%$ empty), BBBC039 (0.963, $0/47 = 0\%$), NIST medium (0.956, $11/33 = 33.3\%$), BBBC038 (0.942, $26/99 = 26.3\%$), NIST high (0.421, $28/33 = 84.8\%$); Spearman rank correlation between cosine and empty-mask rate across the five strata is $\rho \approx 0.3$, near-random agreement, and the best-cosine stratum in the panel produces the second-worst downstream outcome. BBBC039 U2OS fluorescence retains full instance quality (0/47 empty, median AP₅₀ = 0.927) and is also the one stratum where the angular-error proxy has not saturated. NIST iPSC high density (cosine 0.421) is the only stratum where cosine registers damage, and it collapses regardless. The axis on which the proxies mislead is imaging modality $\times$ density regime, precisely the axis a compressed encoder is deployed against.

Table 3: Biology-conditioned proxy discordance for W8A8-cal-QDQ on the 245-field hold-out panel relative to FP32. Rows are grouped by imaging modality and density regime. Metrics include activation agreement proxies (NRMSE, cosine similarity, and flow angular error) and downstream mask retention.

Modality	Regime	n	NRMSE	Cosine	Flow err.	Empty (AP ₅₀ med)
BBBC039	U2OS fluorescence	47	0.482	0.963	58°	0/47 (0.927)
BBBC038	Heterogeneous nuclei	99	0.563	0.942	90°	26/99 (0.536)
NIST iPSC	Medium density	33	0.381	0.956	90°	11/33 (0.000)
NIST iPSC	Low density	33	0.420	0.972	90°	24/33 (0.000)
NIST iPSC	High density	33	0.927	0.421	87°	28/33 (0.000)

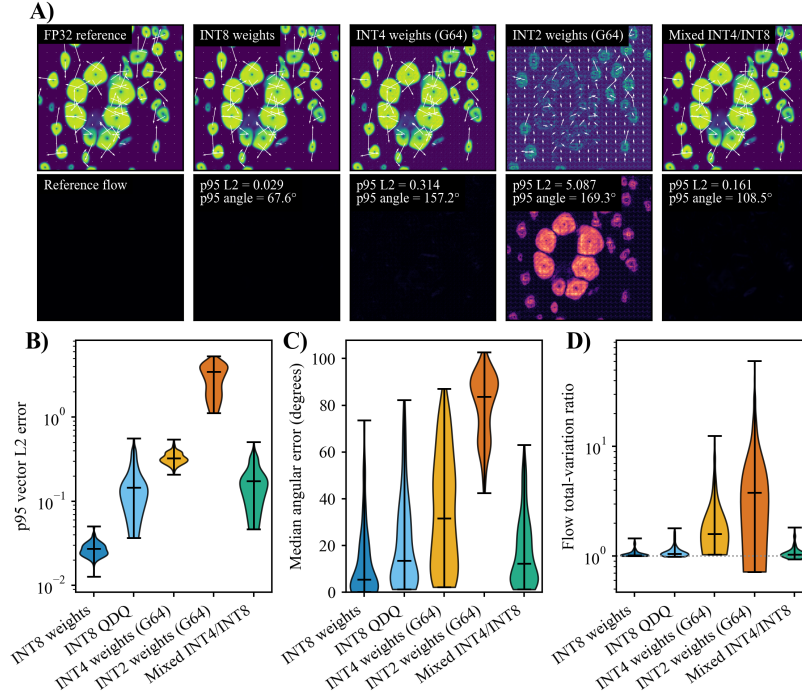


Figure 3: **Decoder-aware flow-field proxies.** **A**, qualitative decoded flow fields for FP32 and four quantized configurations, with per-field 95th-percentile L2 error, angular error, and zero-flow plateau fraction. **B–D**, retention-panel distributions of **B** 95th-percentile vector L2 error, **C** median angular error, and **D** decoded-flow total-variation ratio. W2A16-G64 shifts angular error toward 90°, consistent with a categorical flow-field collapse that is not captured by scalar output-tensor similarity metrics.

A widespread PTQ heuristic identifies the most-sensitive operators by a proxy criterion and preserves their precision, and Table 4 shows where this heuristic still works on our task. A one-operator-at-a-time W4 perturbation over 100 eligible Linear/Conv2d operators isolates four highest-sensitivity operators (`encoder.blocks.0.mlp.lin2`, `encoder.blocks.23.attn.qkv`, `encoder.neck.2.out`); retaining these at INT8 while the remainder use W4 gives a $6.76\times$ weight-storage reduction (171.86 MiB), and delivers 0/176 observed catastrophic fields (rule-of-three upper bound $3/107 \approx 0.028$ over experimental units), the same zero-event outcome as W8A16 at this sample size. The proxy-driven selection therefore transfers within the weight-only regime. Applying the same operator selection under a whole-graph activation-quantization envelope does not recover the collapsed configurations of Table 2: the retained INT8 exceptions do not fix a failure whose mechanism is, we hypothesize, post-decoder seed detection rather than weight reconstruction at any individual operator.

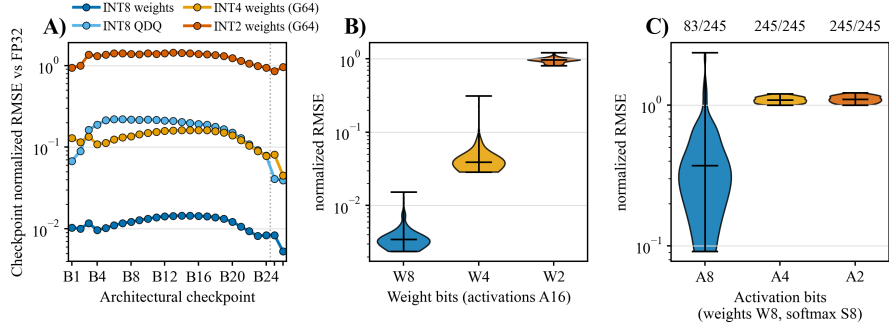


Figure 4: **Where in the graph the quantization budget spends.** **A)** checkpoint-level normalized RMSE against FP32 through the 26 architectural checkpoints (B1–B26, comprising the ViT-L encoder’s 24 transformer blocks together with the patch-embedding and neck taps), stratified by scheme; the calibrated W8A8-cal-QDQ curve tracks the W8 weight-only curve for most of the encoder but diverges near the neck. **B)** per-field raw-output NRMSE aggregated by weight bit width at fixed A16 activations, showing the weight-only sensitivity floor. **C)** activation-bit sweep at fixed W8 (softmax stabilization pinned to A8): 89/245 hold-out fields produce empty masks at A8 (W8A8-cal-QDQ, and 83/245 at W8A8-full-int) and 245/245 produce empty masks at both A4 and A2, so the proxy in panel B, only $\sim 2\times$ larger between A4 and A8, with median flow angular error already at 90° at A8, does not track the downstream shift from partial to categorical collapse reported in Tables 2 and 3.

Table 4: Sensitivity-guided mixed W4/W8 scheme. Four INT8 exceptions are selected from 100 candidate operators by a one-operator-at-a-time W4 perturbation on the development split; the remainder use W4. The scheme achieves a $6.76\times$ weight-storage reduction and 0/176 observed catastrophic fields (rule-of-three upper bound 0.028 over 107 experimental units), the same zero-event outcome as W8A16 at this sample size.

Scheme	Compression		Robustness	
	Size (MiB)	Ratio	Catastrophic rate [95% CI]	INT8 exceptions
mixed-W4-sensitive-W8	171.86	$6.76\times$	0.000 [0.000, 0.028] [†]	4 layers ^a

[†] Rule-of-three upper bound 3/107 over experimental units, in place of the degenerate zero-event bootstrap interval.

^a INT8 exceptions: `blocks.0.mlp.lin2`, `blocks.23.attn.qkv`, `neck.2`, and `out`.

5 Discussion & Conclusion

The three proxies studied (NRMSE, cosine similarity, and median flow angular error) fail in complementary ways under whole-graph activation quantization. Angular error is a resolution failure: median flow angular error is 90° at W8A8-cal-QDQ, W8A8-full-int, W8A4, and W8A2 alike; a decoded flow field that is uniform-random against FP32 is the boundary the angular-error proxy can measure, and every whole-graph configuration we test is already at that boundary at the pooled panel level, so the proxy cannot separate partial (89/245 empty at W8A8) from categorical (245/245 at sub-A8) collapse. Stratification recovers signal the pooled panel hides: median angular error is 58° on BBBC039 and 87° on NIST high density against 90° on the other three strata (Table 3). Cosine and NRMSE are a threshold failure: cosine remains ≥ 0.94 at W8A8-cal-QDQ, compatible on our panel with 89/245 empty-mask cases and a median AP_{50} of only 0.100; NRMSE moves $\sim 2\times$ across activation widths that separate partial from categorical downstream collapse. Both carry rank information (cosine separates A8, A4, and A2 cleanly), but a cosine of 0.944 reads as excellent to a practitioner while coexisting with 36% empty masks, and neither proxy has a defensible accept threshold near the operating regime.

The most consequential piece of the picture is that the downstream discordance is biology-conditioned and non-monotone in the proxy. Sorted by cosine (best proxy first) the five per-modality strata in Table 3 rank as NIST low (0.972, 72.7% empty), BBBC039 (0.963, 0%), NIST medium (0.956, 33.3%), BBBC038 (0.942, 26.3%), NIST high (0.421, 84.8%); Spearman rank correlation between

cosine and empty-mask rate is $\rho \approx 0.3$, near-random agreement, and the highest-cosine stratum in the panel produces the second-worst downstream outcome. On BBBC039 U2OS fluorescence, the one stratum where the angular-error proxy has not saturated (58°), the whole-graph W8A8-cal-QDQ scheme is essentially non-disruptive downstream and the proxies (NRMSE 0.482, cosine 0.963) match. NIST iPSC high-density (cosine 0.421) is the only sub-stratum where cosine registers damage, and it collapses regardless. The axis on which the proxies mislead is the axis on which the compressed model is deployed. This is not a defect in any individual proxy; it is a structural consequence of comparing pixel- or channel-wise residuals against a bounded reference through a non-linear post-hoc decoding cascade whose invariances the proxies do not represent. We hypothesize the specific mechanism is that seed detection in the flow-integration cascade requires a directionally coherent flow field; our data are consistent with this account without isolating it, and a decisive test is a channel-substitution experiment that pairs FP32 flow with quantized cell-probability (and the inverse) to identify which channel’s quantization is dispositive. We report this substitution experiment as an open follow-up.

For compression evaluation of promptless dense-prediction foundation models, we recommend pre-registering the downstream retention margin before hold-out evaluation, reporting per-(scheme, dataset, regime) verdicts with cluster-bootstrap intervals rather than pooled retention numbers, and releasing per-field metric tables so that verdicts can be re-derived on independent data. On our task, this protocol accepts weight-only W8A16 and a sensitivity-guided mixed W4/W8 scheme that achieves $6.76\times$ weight-storage reduction with 0/176 observed catastrophic fields (rule-of-three upper bound 0.028 over 107 experimental units); it rejects ternary weight-only quantization at any storage advantage; and it rejects every whole-graph activation-quantization configuration tested despite proxy profiles that would tempt a practitioner to accept the W8A8 configurations, most sharply on the iPSC modality that motivates the deployment.

6 Limitations

The findings scope to promptless flow-integration decoders and to a mask-level AP retention criterion; pipelines with different decoder families or that consume the raw flow field directly should re-derive the mapping from proxy to retention rather than inherit it. The retention margin used here (-0.02 mean AP change from FP32) is calibrated to segmentation-driven monitoring workflows that tolerate a bounded AP drop; deployments with tighter numerical tolerances (regulatory release testing) require a tighter margin, and the protocol scales but the verdicts do not transfer.

Acknowledgments and Disclosure of Funding

Do **not** include this section in the anonymized submission, only in the final paper. You can use the ack environment provided in the style file to automatically hide this section in the anonymized submission.

References

- Peter Bajcsy, Antoine Vandecreme, Julien Amelot, Joe Chalfoun, Michael Majurski, and Mary Brady. Enabling interactive characterization of stem-cell colony morphology using large microscopy images. *Cell Reports Methods*, 2018. Referenced here for the NIST iPSC low/medium/high density regime tiles used in this work.
- Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, 16(12):1247–1253, 2019.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.
- Thomas J. DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3): 189–228, 1996.

- Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*, 2023.
- Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. *Low-Power Computer Vision, Chapman and Hall/CRC*, 2022.
- Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- Mohammed Rashedul Islam. Sample size and its role in the Central Limit Theorem. *Computational and Applied Mathematics Journal*, 4(1):1–7, 2018.
- 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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.
- Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. Technical report, arXiv preprint arXiv:1806.08342, 2018.
- Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations (ICLR)*, 2021.
- 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 LLM compression and acceleration. In *Machine Learning and Systems (MLSys)*, 2024.
- 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. 2014.
- Vebjorn Ljosa, Katherine L. Sokolnicki, and Anne E. Carpenter. Annotated high-throughput microscopy image sets for validation (BBBC). *Nature Methods*, 9(7):637, 2012.
- Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024a.
- Shuming Ma, Hongyu Wang, Lingxiao Ma, Lei Wang, Wenhui Wang, Shaohan Huang, Li Dong, Ruiping Wang, Jilong Xue, and Furu Wei. The era of 1-bit LLMs: All large language models are in 1.58 bits. *arXiv preprint arXiv:2402.17764*, 2024b.
- Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 9:5217, 2018.
- Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettnner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21:195–212, 2024.
- Erik Meijering. A bird’s-eye view of deep learning in bioimage analysis. *Computational and Structural Biotechnology Journal*, 18:2312–2325, 2020.
- Erick Moen, Dylan Bannon, Takamasa Kudo, William Graf, Markus Covert, and David Van Valen. Deep learning for cellular image analysis. *Nature Methods*, 16(12):1233–1246, 2019.

- Markus Nagel, Rana Ali Amjad, Mart van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? Adaptive rounding for post-training quantization. In *International Conference on Machine Learning (ICML)*, 2020.
- 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.
- Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.
- Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature Methods*, 19(12):1634–1641, 2022.
- Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22:1–20, 2021.
- Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. XNOR-Net: ImageNet classification using binary convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, 2016.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- Annika Reinke, Minu D. Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Bänsch, A. Emre Kavur, Carole H. Sudre, Laura Acion, Michela Antonelli, Tal Arbel, et al. Understanding metric-related pitfalls in image analysis validation. *Nature Methods*, 21:182–194, 2024.
- Carsen Stringer and Marius Pachitariu. Cellpose–SAM: a foundation model for generalist cellular segmentation, 2025. Preprint. Cellpose–SAM release notes and model checkpoint cpsam_v2.
- Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, 2021.
- Gaël Varoquaux and Olivier Colliot. How to evaluate performance in medical imaging: sample size, tests, and error bars. *arXiv preprint arXiv:2210.14677*, 2022.
- Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. Star-convex polyhedra for 3D object detection and segmentation in microscopy. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.
- Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. ZeroQuant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.