Catnip Tensorwise FP8 Outlier 影响的统计研究

基于 baseline27FPS 的 576 个 Linear 层、13.690B 权重与真实 Streaming 激活

Catnip 工程统计审计

2026 年 7 月 21 日

基线工程	`/workspace/week2-improved-fp8_scaled`（团队口径 baseline27FPS；后续 week 修改不作为实现来源）
模型权重	`/workspace/LTX/models/ltx_rl16/merged_sst1500_rl16.safetensors`
目标集合	48 个 Transformer blocks × 12 类 Linear = 576 个模块，共 13.690 B 个权重
量化路径	BF16 source → tensorwise E4M3FNweight；动态 tensorwise E4M3FNactivation；`aten._scaled_mm` 输出 BF16
实验标识	`catnip_fp8_outlier_statistics_v1`

摘要

本报告回答一个具体问题：Catnip 从 BF16 转为 tensorwise-scaled FP8 时，权重或激活中的 outlier 到底造成了多大影响，以及这种影响表现为下溢、舍入误差、局部激活误差还是可见的压缩收益。我们对基线 manifest 中全部 576 个 Linear 的 13.690 B 个权重做逐元素精确 E4M3 反事实；对两个真实 20 秒 Streaming 请求的 63,360 个 module-call 记录做激活长尾、rowwise 对照与跨 prompt 稳定性分析；并用 family-stratified bootstrap、Kruskal–Wallis、Wilcoxon 与统一 INT8 对照拆解机制。

最主要的结论是：**权重 outlier 对 E4M3 的“计数型下溢”影响很强，但不是当前能量加权权重重构误差的主导因素**。输出通道 amax 偏斜与非零值量化为零的 Spearman $\rho = 0.812$，而 tensor amax/p99.99 与 FP8 相对 L2 的 $\rho = -0.046$，置信区间覆盖零。全局生产权重 RelL2 为 2.6497%；换成 output-channel scale 只降低 0.476% 的 MSE，却减少 93.89% 的下溢。相反，真实激活 tail 与 tensorwise 激活误差的 $\rho = 0.380$，与 rowwise MSE 收益的 $\rho = 0.442$，说明下一步应优先做**有 heldout 的、按层定向的激活侧因果实验**，而不是全局剪掉权重 outlier。

一句话结论。Catnip 当前 tensorwise E4M3 的约 2.65% 权重相对 L2，主要呈现为 E4M3 三位尾数带来的广泛舍入噪声；权重 outlier 主要把更多小权重推到 subnormal/zero 区，而不是显著增加总误差能量。激活 outlier 的局部影响更强，但现有数据只有两个 prompt，尚不足以批准任何 production 修改。

目录

1 研究问题、边界与决策标准

1.1 研究问题

本轮将“outlier 的影响程度”拆成五个可检验问题：

1. 权重 tail severity 是否增加生产 tensorwise FP8 的重构误差？
2. 它是否主要增加非零小值变为零的 underflow，而 underflow 本身只携带很小能量？
3. 更细粒度的 weight scale、显式 clipping 或 sparse residual 能否形成有意义的误差–存储 Pareto 改善？
4. 真实 Streaming 激活的 outlier 是否比权重 outlier 更能解释局部量化误差，以及这些 outlier channel 是否跨 prompt 稳定？
5. 哪些 block/family 同时暴露较高权重与激活风险，适合进入下一轮定向因果实验？

1.2 本报告可以与不可以证明什么

本报告的权重侧结论覆盖全部目标权重，因而能精确描述 weight-space E4M3 重构误差与下溢。激活侧使用真实运行统计，但 quantization probe 是有界采样，且仅有两个 prompt、同一 seed。因此：

- 可以证明各 outlier 指标与局部权重/激活重构误差的关联强弱；
- 可以计算固定量化反事实的精确权重 MSE 与静态存储上界；
- 不可以仅凭这些关联断言某层导致最终视频或音频质量下降；
- 不可以把 sparse-residual 的离线理论上界当成已有 kernel 的吞吐表现；
- 不可以用两个 calibration prompt 批准新的 production 量化策略。

本文刻意区分三层对象：(1) 权重重构误差，(2) Linear 输入/输出局部误差，(3) 端到端流式生成质量。只有完成配对的 heldout 端到端实验，才能把前两层的统计信号升级为 production 因果结论。

2 基线实现与数据血缘

2.1 为什么选择 week2 工程

本研究将 `/workspace/week2-improved-fp8_scaled` 作为唯一代码基线。锁定的 derived manifest 为：

Manifest	`experiments/split_fp8_dit_vae_profile/manifests/fp8_scaled_mm_attn1_kv_o5_cuda1.json`
SHA256	`c085f53dbdac584a6f53b000ba5efb59892d512dcd9b42a4a8f8f53eb87acdf9`
核心实现	`experiments/split_fp8_dit_vae_profile/scripts/fp8_scaled_mm_fused.py`
实现 SHA256	`05f09ec23860ab448bdce88463cdfbcfbc7ff93031288c05180c081ba8d17226`

后续 week 文件只作为只读的统计资产或反证来源，不把其候选 kernel、scale 粒度或质量决策合并回本基线。特别地，本报告复用 week4 observer 已生成的 activation bundles，但所有 FQN 都与 week2 的 576-layer manifest 做一一交集校验。

2.2 生产量化公式

基线使用 `torch.float8_e4m3fn`，其最大有限值记为 $F = 448$。对某个二维权重 $W \in \mathbb{R}^{N\times K}$：

$$a = \max_{j,k} |W_{jk}|, \tag{1}$$

$$s = \begin{cases} a/F, & a > 0, \\ 1, & a = 0, \end{cases} \tag{2}$$

$$Q_s(W) = s \cdot \mathrm{cast}_{\mathrm{E4M3FN}}\left(\mathrm{clip}(W/s, -F, F)\right). \tag{3}$$

运行时存储的是一字节 FP8 weight 与一个标量 FP32 inverse scale。激活按每次完整二维输入动态计算同样的 tensorwise scale，随后调用 `aten._scaled_mm`，输出 BF16，并使用 fused BF16 bias。图 1 对应基线代码真实路径，而不是抽象的通用量化器。

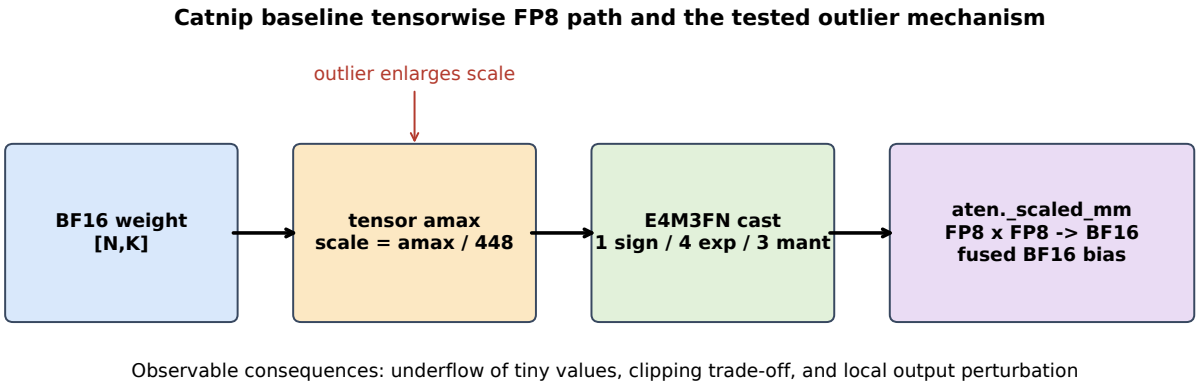

图 1: Catnip baseline tensorwise FP8 数据路径与本轮检查的 outlier 作用链。

2.3 目标层与数据完整性

目标集合为 48 个 block 中固定的 12 类 Linear：self-attention-1 的 Q/K/V/O，self-attention-2 的 Q/K/V/O，audio-to-video attention 的 Q/O，以及 feed-forward up/down。总权重数为

$$\sum_{\ell=1}^{576} N_\ell K_\ell = 13{,}690{,}208{,}256.$$

checkpoint 大小为 37,978,195,728 bytes，其实测 SHA256 为 `d09adb8c85c1ae9681958e4c45b4f6bdcfc3e596d3782ab2f76bd9cd9824734c`。全量 FP8 production 误差重新计算后，与既有独立 audit 的逐层 error-squared 最大相对差为 2.28×10^{-7}，作为实现一致性检查。

2.4 真实激活资产

激活数据来自两个成功的 20 秒请求：`person_car_seed12345` 与 `person_ceramics_seed12345`。每个 Linear 在 11 个 streaming chunks 中经历 4 个 NFE 与 1 个 commit，共 55 次 forward；因此每个 cell 有 $576 \times 55 = 31{,}680$ records，两格合计 63,360。

每个 call 的 amax、元素数和有限性检查覆盖完整输入；absolute p50/p99.9 来自最多 8,192 个确定性均匀位置的 128-bin log2 histogram；tensorwise/rowwise E4M3 probe 最多均匀抽取 64 行；每模块每输入通道的跨 call amax 完整保存到 safetensors。原始 activation tensor 未落盘。

3 实验设计与统计方法

3.1 Outlier 指标

对权重层 ℓ 定义两种互补指标：

$$T_\ell^{(w)} = \frac{\max |W_\ell|}{\widehat{Q}_{0.9999}(|W_\ell|)}, \tag{4}$$

$$S_\ell^{(w)} = \frac{\max_j a_{\ell j}}{\operatorname{median}_j a_{\ell j}}, \qquad a_{\ell j} = \max_k |W_{\ell,jk}|. \tag{5}$$

$T^{(w)}$ 衡量 tensor 最大值相对主体尾部的突出程度；$S^{(w)}$ 衡量输出通道间 amax 不均衡。amax 与所有误差和 underflow 都是全元素精确值；分位数、kurtosis 与 INT8 control 使用每层最多 262,144 个 flatten 后等距索引样本。

对 activation call c 定义

$$T_c^{(x)} = \frac{\operatorname{amax} |X_c|}{\widehat{Q}_{0.999}(|X_c|)}.$$

模块级 $T^{(x)}$ 是先在每个 cell 对 55 个 call 取 p95，再在两格间取均值。该定义降低单次偶发峰值对模块排序的支配，但仍保留运行尾部行为。

3.2 误差、下溢与反事实

精确相对 L2 为

$$\operatorname{RelL2}(W, \widehat{W}) = \sqrt{\frac{\sum_i (W_i - \widehat{W}_i)^2}{\sum_i W_i^2}}.$$

非零到零下溢率只在 source nonzero 元素中计数。所有全局 MSE 都先汇总每层 error-squared 与 source-squared，再取比值，不做“层均值冒充全局误差”。

共评估十个 scale multiplier

$$\alpha \in \{1, .875, .75, .625, .5, .375, .25, .1875, .125, .0625\},$$

将 scale threshold 改为 α amax $|W|$。两种反事实分别为：

Dense clipping 直接保留被 clip 的 FP8 重构，测量减少步长与剪裁能量损失的净效果；

Sparse residual upper bound 仅对 $|W_i| > \alpha a$ 的位置保存精确 FP32 residual，假设 uint32 index + FP32 value，即每个 outlier 额外 8 bytes。它是可达到误差的乐观上界，不包含 row pointer、alignment 与 kernel overhead。

另计算 output-channel weight scale 的全元素精确反事实，并计算每层最多 262,144 样本的 symmetric uniform INT8（$q_{\max} = 127$）对照，用来区分浮点指数范围与均匀定点步长的作用。

3.3 统计检验

主关联使用 Spearman rank correlation。95% 区间来自 4,000 次按 12 个 Linear families 分层的有放回 bootstrap；这保留每类 48 个 block 的样本量结构。family 差异使用 Kruskal–Wallis，并报告

$$\epsilon^2 = \frac{H - k + 1}{n - k}$$

作为描述性效应量。两个 activation cells 的模块 median error 使用配对 Wilcoxon；跨 cell 排序使用 Spearman。七个相关性检验同时展示原始 p 值，未把 $p < .05$ 自动解释为工程重要性。特别地，weight tail 与 activation tail 的原始 $p = .0415$ 在 Holm 校正后不再显著，应视为弱线索。

4 权重 Outlier：强烈影响下溢，却几乎不影响 E4M3 总误差

4.1 生产误差的全量基准

生产 tensorwise E4M3 在全部 13.690B 权重上的全局 RelL2 为 2.6497%。逐层值高度集中在约 2.64%–2.66%。这与一个只有 3 个显式 fraction bits 的浮点格式在 normal range 内提供近似恒定相对精度的机制相符 [1]。

图 2(a) 显示 $S^{(w)}$ 与下溢有很强单调关系：$\rho = 0.812$，family-stratified 95% CI $[0.773, 0.846]$。但图 2(b) 中 $T^{(w)}$ 与相对 L2 的 $\rho = -0.046$，95% CI $[-0.129, 0.037]$，$p = .271$。也就是说，outlier 确实把更多小值挤入 E4M3 subnormal/zero 区，却没有让这些小值携带足够能量来主导全层 L2。

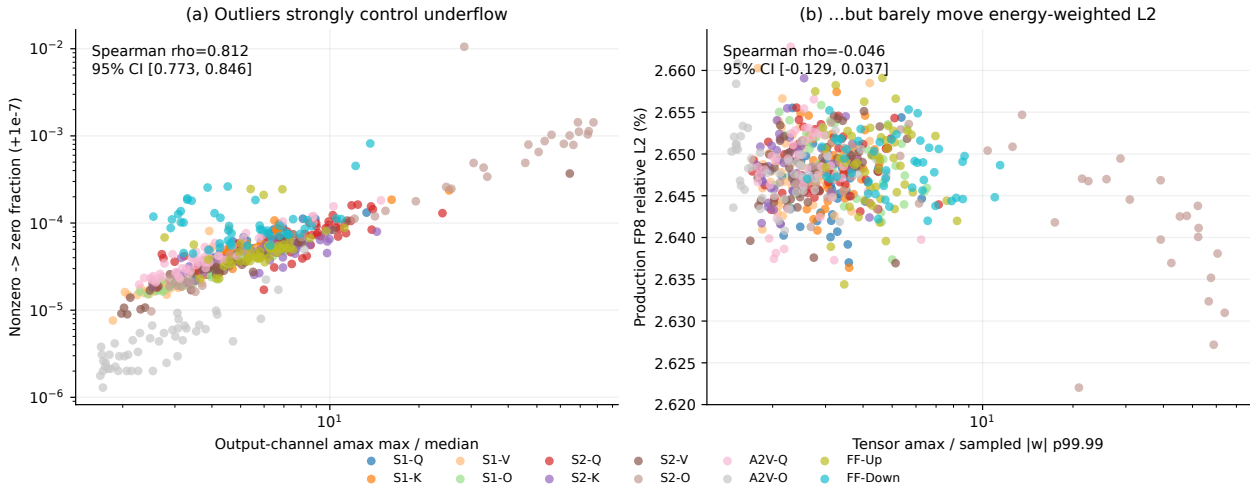

图 2: 权重 outlier 的两种后果。左：输出通道 amax 偏斜强烈控制 nonzero-to-zero；右：tensor amax/p99.99 几乎不能解释生产 FP8 相对 L2。散点颜色为 12 个 Linear families。

全局 element-weighted nonzero-to-zero rate 为 9.601×10^{-5}，即约 0.00960%。其计数虽可观，但能量很低；这是“计数损失”和“能量损失”给出不同结论的根源。

4.2 INT8 对照：为何不能照搬均匀量化直觉

同一组 $T^{(w)}$ 对 symmetric uniform INT8 control 的 sampled relative L2 有 $\rho = 0.817$，95% CI $[0.784, 0.847]$；对 E4M3 则仍是 -0.046。INT8 production-scale sampled relative L2 中位数为 4.849%，且所有 576 层在固定 grid 中都选择 $\alpha < 1$，best clipped INT8 的 sampled relative L2 中位数降为 2.104%，MSE 中位改善 75.87%。

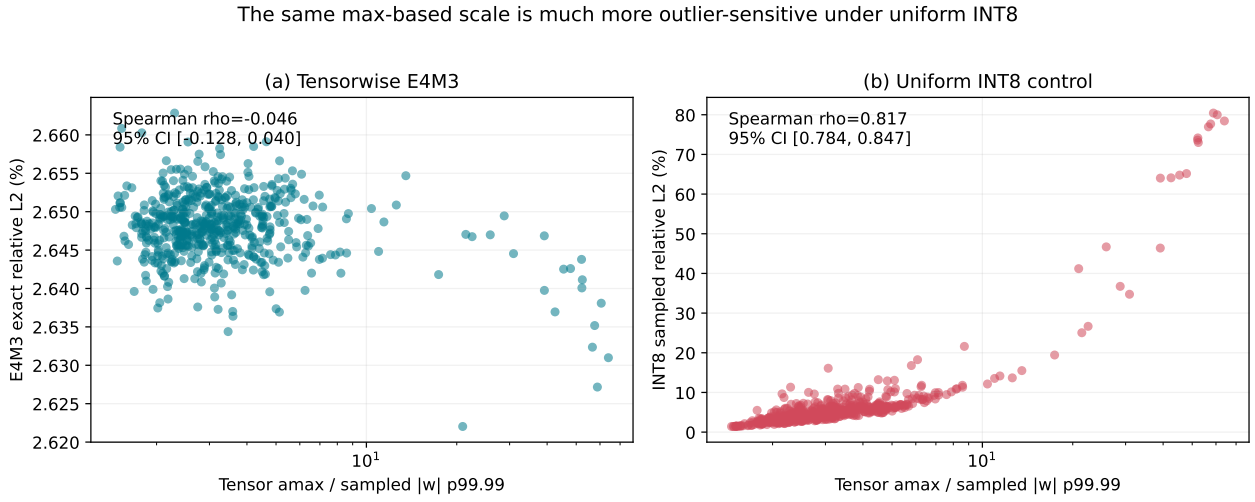

图 3: 相同 max-based tensor scale 在 E4M3 与 uniform INT8 下的 outlier 敏感性完全不同。INT8 仅为确定性抽样机制对照，不是 production 候选。

表 1: E4M3 与 INT8 控制实验。E4M3 error 为全元素精确值；INT8 为每层最多 262,144 权重的确定性样本。

Format	Median rel. L2 (percent)	Tail rho	CI low	CI high
E4M3 (exact all weights)	2.6483	-0.0459	-0.1281	0.0400
Uniform INT8 (sampled control)	4.8489	0.8173	0.7843	0.8468

机制解释是：uniform INT8 的步长 $\Delta = a/127$ 会被单个大值线性放大，主体分布共享同一固定步长；E4M3 在 scale 后仍具有指数编码，不同数量级各自拥有近似相对精度，因此 max scale 对 normal-range 舍入误差的影响弱得多。E4M3 并非对 outlier 免疫：它把代价转移到动态范围底部，表现为更多 subnormal 与 zero。

4.3 family 与 block 结构

权重 channel skew 的 Kruskal–Wallis $H = 201.67$、$\epsilon^2 = 0.338$；underflow 的 $H = 280.28$、$\epsilon^2 = 0.477$，均显示 family 是重要结构变量。S2-O 的 p95 underflow 为 0.132%，远高于大多数 family；其晚层热点在图 4 和图 5 中清晰可见。

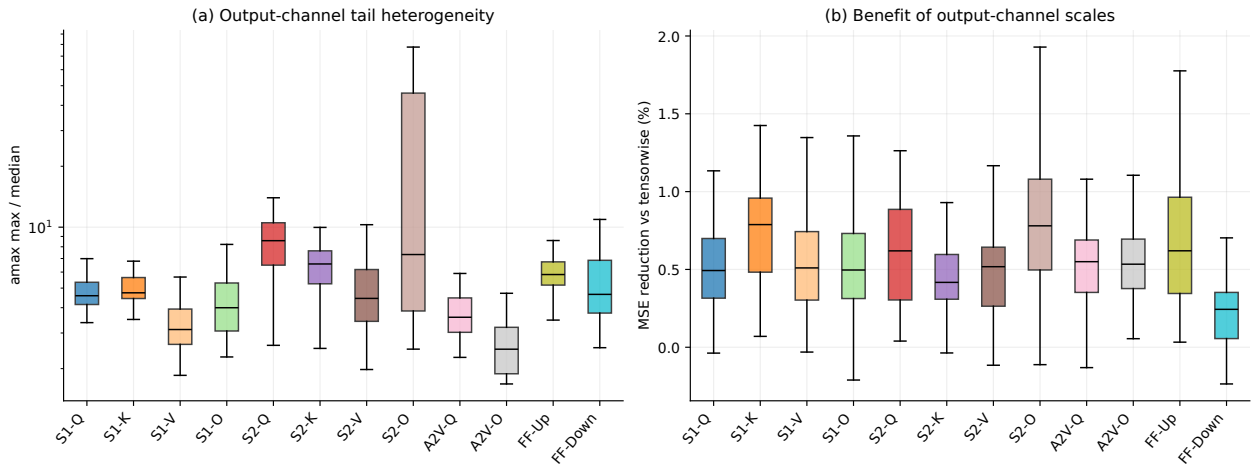

图 4: 按 12 个 family 分层的输出通道 tail heterogeneity 与 output-channel scale MSE 收益。箱线图不绘制离群点，避免遮蔽主体。

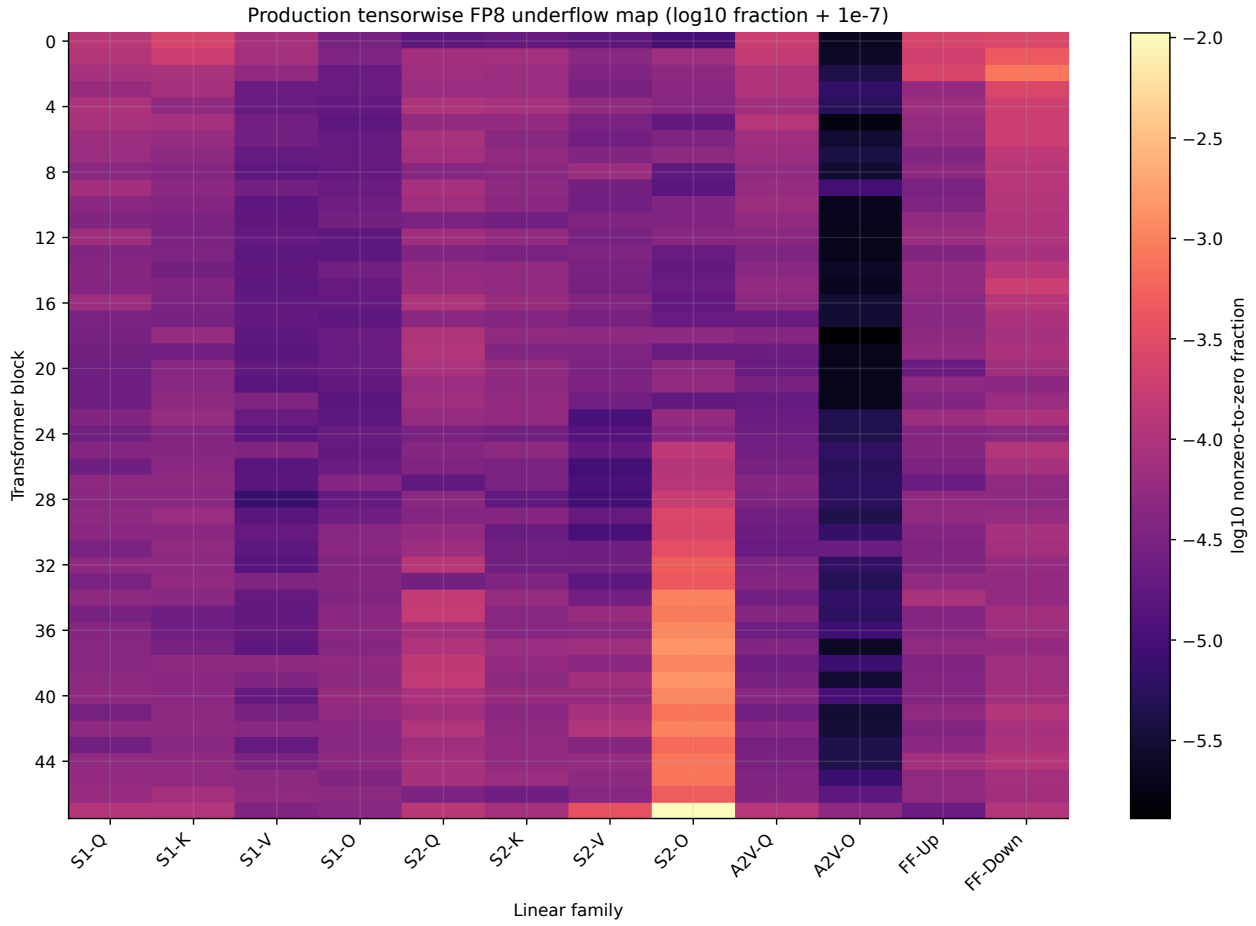

图 5: 48 blocks × 12 families 的生产 tensorwise FP8 underflow heatmap。颜色为 $\log_{10}(\text{fraction} + 10^{-7})$。

表 2: Family 摘要。Weight underflow p95 与两类 MSE gain 的单位为百分比；Activation tail 为 amax/p99.9。

Family	Weight skew median	Weight underflow p95	Channel MSE gain	Activation tail p95	Rowwise MSE gain	Top1 Jaccard
S1-Q	4.585	0.011	0.493	4.151	6.491	0.547
S1-K	4.739	0.010	0.788	4.151	6.491	0.547
S1-V	3.121	0.005	0.510	4.151	6.491	0.547
S1-O	3.999	0.005	0.496	3.350	1.216	0.180
S2-Q	8.575	0.015	0.619	11.947	49.129	0.783
S2-K	6.580	0.007	0.417	4.320	6.965	0.673
S2-V	4.447	0.008	0.518	4.320	6.965	0.673
S2-O	7.325	0.132	0.780	22.821	9.011	0.180
A2V-Q	3.588	0.012	0.550	7.960	31.978	0.708
A2V-O	2.495	0.001	0.534	8.286	5.720	0.183
FF-Up	5.834	0.017	0.620	4.103	2.299	0.426
FF-Down	4.655	0.026	0.244	6.218	1.261	0.176

4.4 Output-channel weight scale 的真实收益

output-channel scale 将全局权重 RelL2 从 2.6497% 降到 2.6434%，MSE 只改善 0.476%；但 underflow 从 9.601×10^{-5} 降到 5.869×10^{-6}，减少 93.89%。额外 scale storage 后约为 1.000833 bytes/weight，静态压缩率从 BF16 的 2.0000× 变为 1.9983×。

这再次说明：channelwise scale 很擅长修复计数型动态范围问题，却只能略微改变能量加权误

差。channel skew 与 channelwise MSE gain 的相关性虽然为 $\rho = 0.200$（95% CI $[0.122, 0.280]$），工程量级仍然很小。

5 Clipping 与 Sparse Residual：误差–压缩率的真实代价

5.1 Dense clipping 不能解决当前 E4M3 问题

在十点 α grid 中，每层取最优 dense clip 是一个乐观 oracle；即便如此，全局 MSE 只下降 0.1479%，且 57.47% 层仍选择 $\alpha = 1$。$T^{(w)}$ 与 dense-clipping gain 的 $\rho = -0.0647$，95% CI $[-0.144, 0.0128]$。当 $\alpha = .25, .125, .0625$ 时，dense MSE 分别变为生产 MSE 的 $3.71\times, 34.88\times, 175.25\times$。减少 scale 并不能抵消被硬剪掉的 outlier energy。

5.2 Sparse residual 能改善误差，但不是免费午餐

如果把被 clip 的元素用 exact FP32 residual 恢复，$\alpha = .25$ 可降低 2.99% 的全局 MSE，同时仍有 1.958× BF16 压缩率；$\alpha = .125$ 可降低 21.49% MSE，但压缩率降至 1.605×；$\alpha = .0625$ 可降低 58.62% MSE，却需要 2.141 bytes/weight，压缩率仅 0.934×，已经比 BF16 更大。

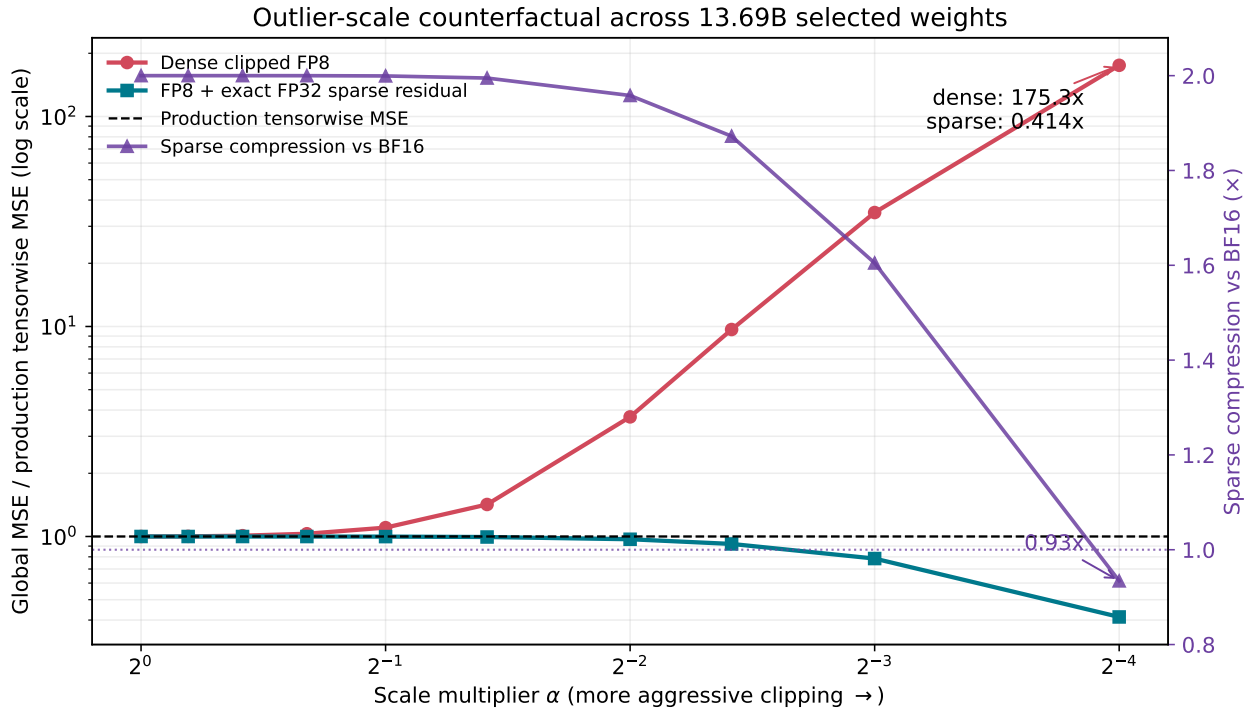

图 6: 全 13.69B 权重的 scale 反事实。左轴为相对生产 MSE 的对数轴；右轴为假设每个 outlier 使用 uint32 index + FP32 residual 时的静态压缩率。

表 3: 全局权重策略对比。MSE change 为相对生产 tensorwise 的降低百分比；sparse residual 是离线上界。

Strategy	Rel. L2 (percent)	MSE change (percent)	Bytes/weight	Compression
Production tensorwise FP8	2.6497	0.0000	1.0000	2.0000
Output-channel FP8	2.6434	0.4764	1.0008	1.9983
Per-layer best dense clip	2.6478	0.1479	1.0000	2.0000
FP8 + exact FP32 sparse residual (alpha=0.5)	2.6479	0.1347	1.0003	1.9994
FP8 + exact FP32 sparse residual (alpha=0.25)	2.6097	2.9948	1.0214	1.9581
FP8 + exact FP32 sparse residual (alpha=0.125)	2.3477	21.4945	1.2463	1.6048
FP8 + exact FP32 sparse residual (alpha=0.0625)	1.7045	58.6185	2.1408	0.9342

图 7 进一步显示 layer-level 收益分布：output-channel scale 是广泛但很小的改善；best dense grid 大量停在零附近；$\alpha = .25$ sparse residual 在多数层有正收益，但其实际价值仍取决于真实稀疏 kernel、索引布局与 end-to-end 质量。

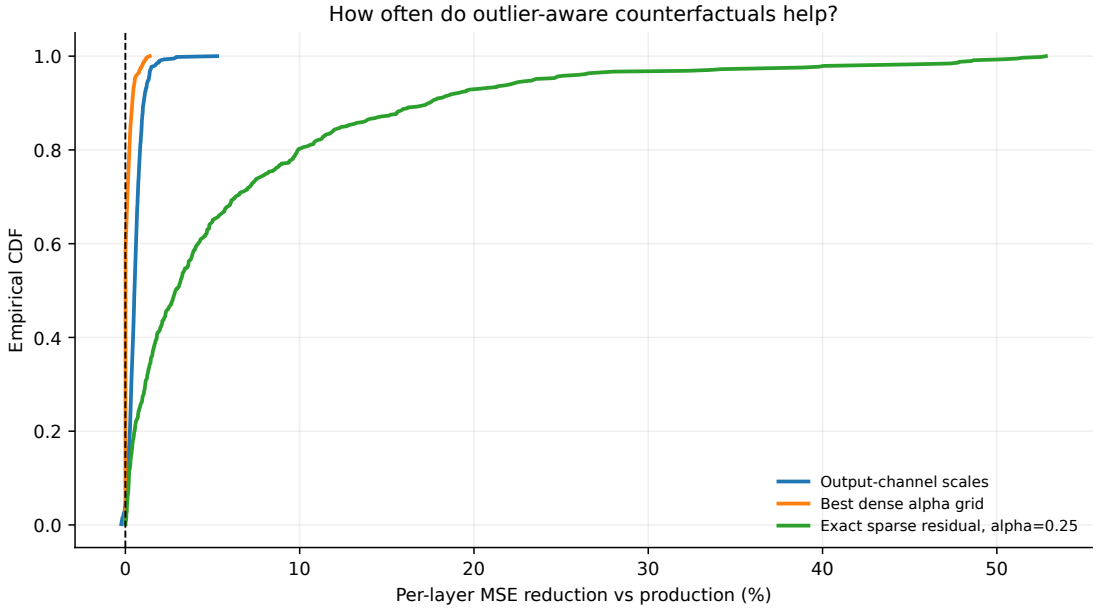

图 7: 三种 outlier-aware 反事实的逐层 MSE 改善经验分布。零线左侧代表该策略在某层反而更差。

6 真实 Streaming 激活：Outlier 与局部误差关系更强

6.1 时间结构

图 8 汇总两个 cells、全部模块在每个 NFE/chunk 的中位数。activation tail 随 NFE 有明显层次：nfe0 最大，commit 最小；而 tensorwise E4M3 相对 L2 仍集中在约 2.64%–2.65% 的窄范围。这说明 tail severity 的巨大视觉变化，不会一比一映射到总体相对 L2，但模块间仍存在可检测的单调关系。

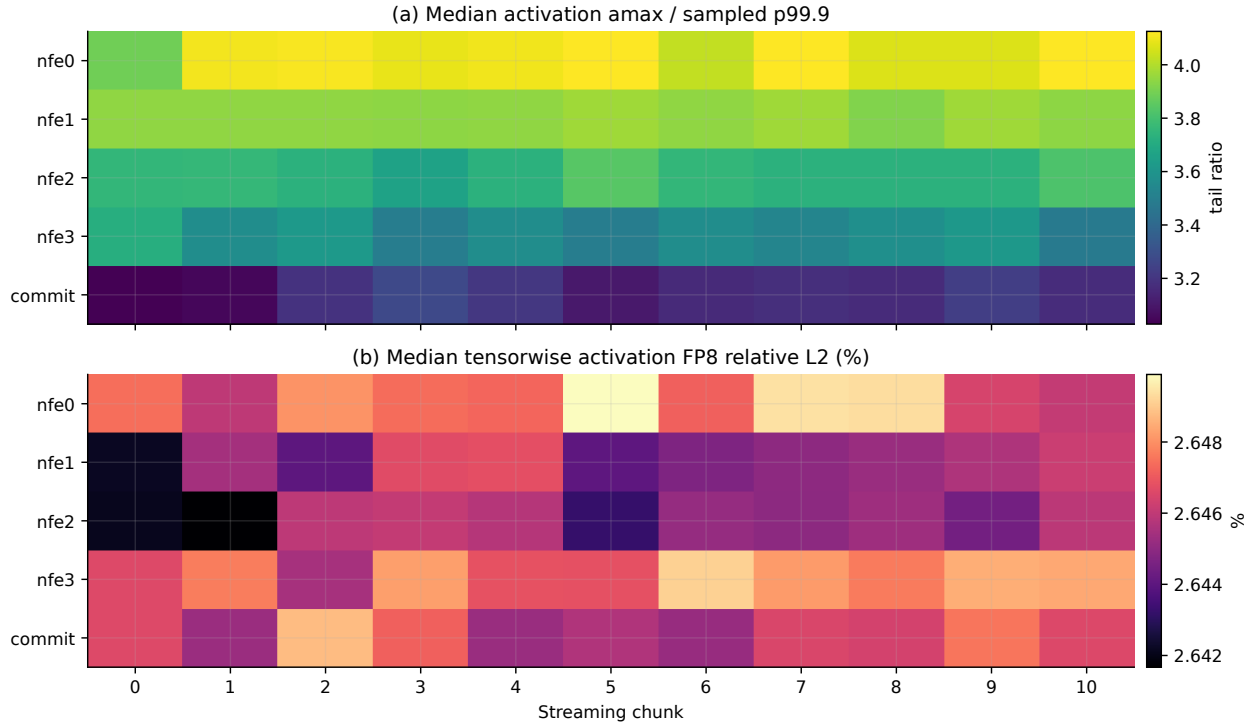

图 8: 真实 Streaming 调用的时间结构。每格为跨两个 prompts 与全部相关模块的 call 中位数；上图是 activation amax/p99.9，下图是 sampled tensorwise E4M3 relative L2。

6.2 Tail severity、量化误差与 rowwise 收益

在每个模块先按 cell 聚合、再合并两格后，activation tail p95 与 tensorwise relative-L2 p95 的 $\rho = 0.3798$，95% CI $[0.3165, 0.4463]$；与 rowwise MSE gain median 的 $\rho = 0.4417$，95% CI $[0.3898, 0.4934]$。这两个效应显著强于 weight tail 对 weight L2 的效应。

family 差异更大：activation tail 的 Kruskal–Wallis $H = 382.01$、$\epsilon^2 = 0.658$；rowwise gain 的 $H = 398.48$、$\epsilon^2 = 0.687$。S2-Q 的 rowwise MSE gain 中位数约 49.13%，A2V-Q 为 31.98%；S2-O tail 中位数最高，为 22.82。这些值是局部 sampled reconstruction 指标，不是最终媒体质量提升。

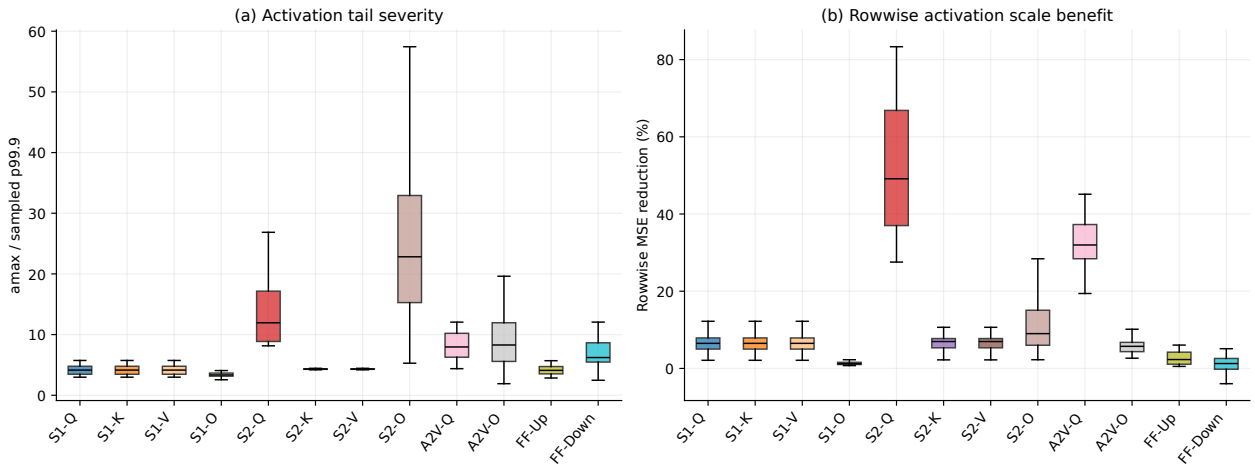

图 9: 按 family 的真实 activation tail 与 rowwise scale 局部 MSE 收益。family 是激活风险的重要结构变量。

6.3 跨 prompt 稳定性

两个 cells 之间，每模块输入通道 log-amax rank Spearman 的中位数为 0.913，说明同一模块内的高风险通道大体可重复；top-1% channel set 的 Jaccard 中位数为 0.547，exact argmax channel 相同的模块占 62.5%。因此固定 channel treatment 有一定统计依据，但并非完全稳定。

另一方面，576 个模块的 median tensorwise error 在两格间没有整体位置偏移（paired Wilcoxon $p = .741$），但模块风险排序的跨 cell Spearman 只有 0.436。这两个结论并不冲突：全体误差分布中心可以相似，同时“哪个模块最危险”会随 prompt 变化。

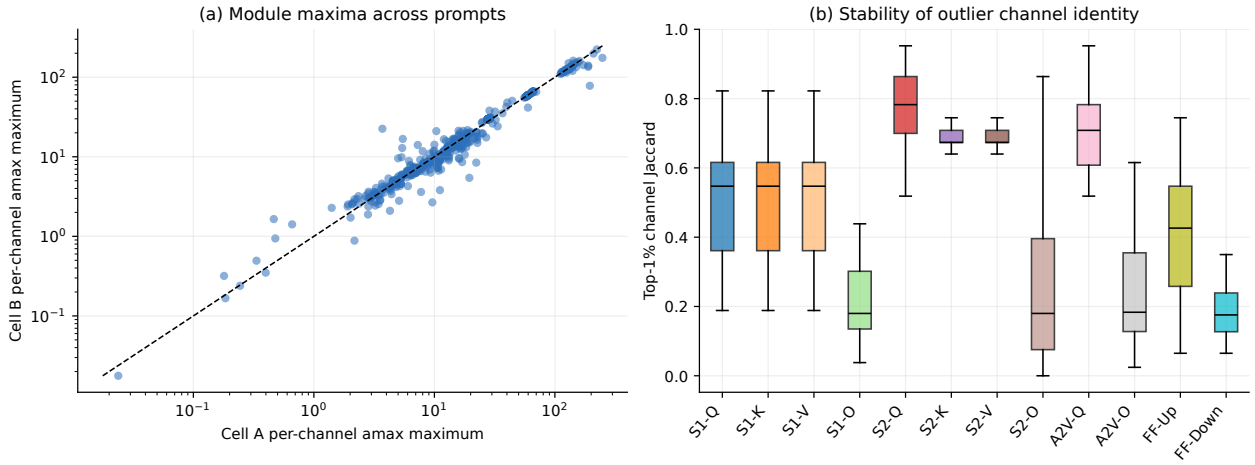

图 10: 跨 prompt 稳定性。左：模块最大 activation amax；右：每模块 top-1% input channels 的 Jaccard，按 family 分层。

6.4 与后续候选历史的交叉解释

只读检查 week4 的既有端到端结果可见：full rowwise activation 虽通过速度门，但跨六格媒体/轨迹质量存在明显尾部回退，最终被拒绝；SmoothQuant 也在 calibration 上改善而 heldout 失败。这与本轮并不矛盾，反而说明“局部 quantization MSE 降低”不足以保证非线性、跨模态、跨 chunk 系统质量。因此，本轮结果支持**缩小定向实验范围**，不支持重新启用全局 rowwise 或 uniform smoothing。

7 权重–激活联合风险定位

为了选择下一轮候选，而不是制造一个未经验证的质量分数，我们构造描述性 rank score：

$$R_\ell = \frac{1}{4}\left[\mathrm{rankpct}(T_\ell^{(w)}) + \mathrm{rankpct}(U_\ell^{(w)}) + \mathrm{rankpct}(T_\ell^{(x)}) + \mathrm{rankpct}(G_\ell^{(x)})\right],$$

其中 $U^{(w)}$ 是生产 underflow，$G^{(x)}$ 是 rowwise sampled MSE gain。每一 rank 都在 576 个模块总体内计算。R 仅用于挑选覆盖 weight/activation 两类症状的实验层，不能解释为失败概率。

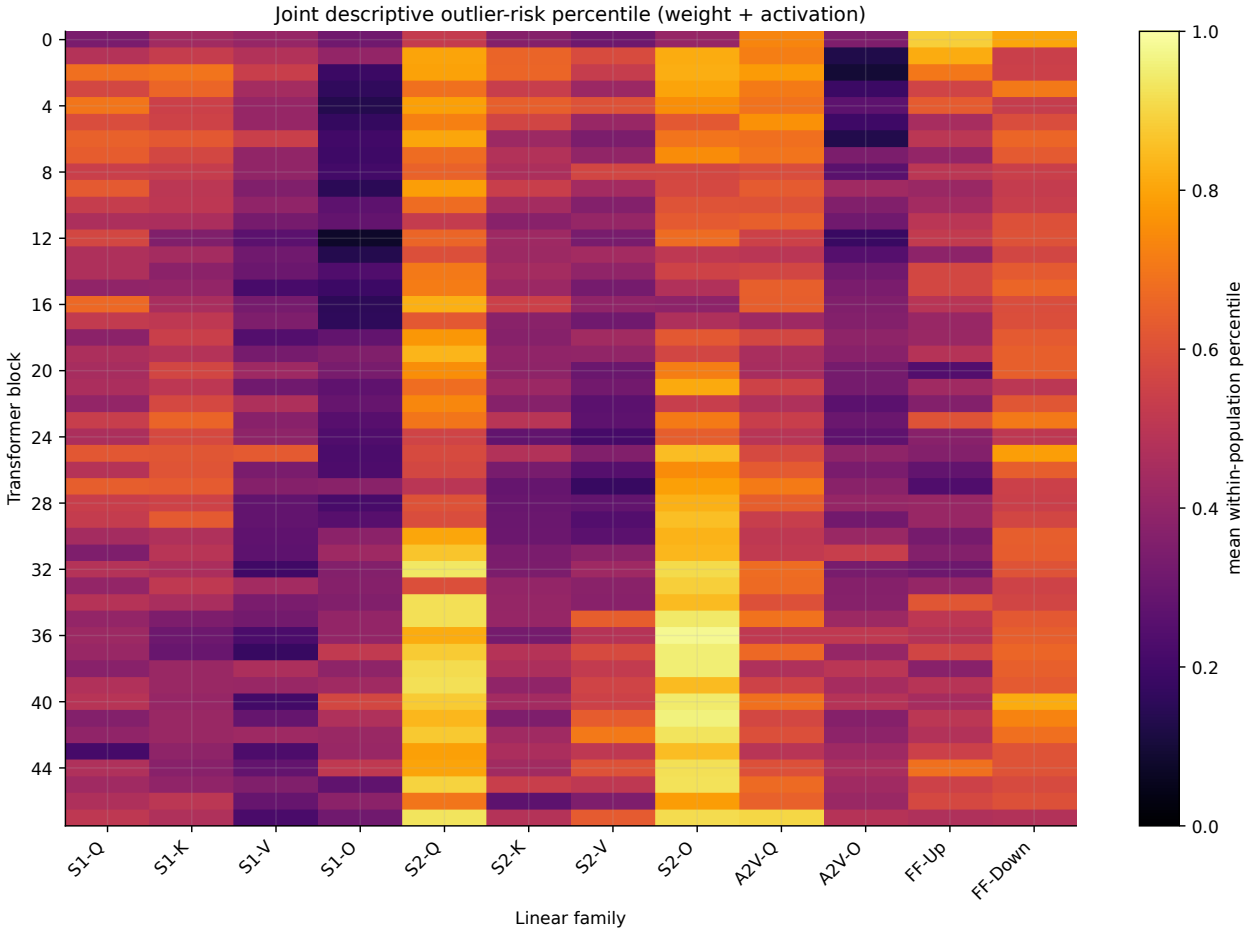

图 11: 联合描述性风险 rank。晚层 S2-O 与 S2-Q 形成最稳定的高风险区域。

最高风险模块以 block 35–45 的 `attn2.to_out.0` 为主，也包括晚层 `attn2.to_q`。例如 block 36 S2-O 的 $R = 97.96$ percentile，weight amax/p99.99 为 56.62、underflow 为 0.112%、activation tail p95 为 110.12、rowwise local MSE gain 为 40.63%。

FQN	Risk pct.	W amax/p99.99	W underflow (%)	Act tail p95	Act row gain (%)
transformer_blocks.36.attn2.to_out.0	97.960	56.616	0.112	110.121	40.628
transformer_blocks.41.attn2.to_out.0	96.007	52.090	0.079	47.686	34.038
transformer_blocks.38.attn2.to_out.0	95.009	60.587	0.104	31.104	27.916
transformer_blocks.37.attn2.to_out.0	94.922	58.820	0.143	45.961	21.517
transformer_blocks.40.attn2.to_out.0	94.314	64.000	0.113	57.443	13.688
transformer_blocks.35.attn2.to_out.0	93.750	45.395	0.087	72.914	13.302
transformer_blocks.32.attn2.to_q	93.576	7.141	0.012	16.674	67.103
transformer_blocks.47.attn2.to_q	92.882	4.500	0.013	49.718	78.517
transformer_blocks.42.attn2.to_out.0	92.795	52.222	0.101	19.665	14.491
transformer_blocks.45.attn2.to_out.0	92.448	39.169	0.079	13.760	28.919
transformer_blocks.34.attn2.to_q	92.057	4.768	0.016	17.321	69.636
transformer_blocks.44.attn2.to_out.0	92.014	52.400	0.081	16.639	18.461
transformer_blocks.35.attn2.to_q	91.927	5.093	0.016	15.946	65.442
transformer_blocks.39.attn2.to_q	91.862	4.842	0.016	17.102	66.189
transformer_blocks.38.attn2.to_q	91.363	4.954	0.014	15.815	62.421

8 统计检验总表与稳健性判断

表 5: 主要 Spearman 相关、family-stratified bootstrap 95% CI 与原始双侧 p 值。

Relationship	n	rho	CI low	CI high	p
Weight channel skew vs underflow	576	0.8123	0.7733	0.8461	1.62e-136
Weight amax/p99.99 vs relative L2	576	-0.04594	-0.1294	0.03746	0.271
Weight channel skew vs channelwise MSE gain	576	0.2	0.1224	0.2797	1.304e-06
Weight amax/p99.99 vs dense clipping gain	576	-0.06469	-0.1444	0.01279	0.1209
Activation tail vs tensorwise relative L2	576	0.3798	0.3165	0.4463	3.289e-21
Activation tail vs rowwise MSE gain	576	0.4417	0.3898	0.4934	6.666e-29
Weight tail vs activation tail	576	0.08497	0.01357	0.1592	0.04151

表 5 可分为三组：

1. **稳定且工程含义明确：** weight channel skew → underflow；activation tail → local error/rowwise gain。区间远离零且效应量中到大。
2. **统计存在但工程收益很小：** weight channel skew → channelwise MSE gain，$\rho = .200$；全局 gain 仍只有 0.476%。
3. **不支持或仅弱线索：** weight tail → FP8 relative L2、weight tail → dense clip gain 均不支持；weight tail ↔ activation tail 的 $\rho = .085$ 很小且多重校正后不成立。

9 工程结论

当前不建议： 全局 dense clipping；仅因 weight outlier 高就改成 output-channel scales；把离线 sparse residual 当成可部署方案；或根据两个 prompt 重新启用全局 rowwise activation。

9.1 对 “outlier 影响有多大” 的定量回答

1. **权重 underflow：影响大。** channel skew 与 underflow 的 $\rho = .812$；output-channel scale 可减少 93.89% underflow。
2. **权重能量误差：影响小。** tensor tail 与 E4M3 relative L2 的 $\rho = -.046$；生产全局 RelL2 约 2.6497% 且逐层高度集中。
3. **更细 weight scale：收益小。** output-channel scales 全局 MSE 改善 0.476%，相对 L2 仅从 2.6497% 到 2.6434%。
4. **Dense clipping：负面。** 最优固定 grid oracle 仅改善 0.148% 全局 MSE；激进 clip 会将 MSE 放大到几十或上百倍。
5. **Sparse outlier rescue：可形成 Pareto 点，但代价清晰。** $\alpha = .25$ 可改善 2.995% MSE、保持 1.958× 静态压缩；更大收益迅速消耗压缩率，并尚无 runtime 证据。
6. **激活局部误差：影响中等且更值得研究。** tail 与 tensorwise error 的 $\rho = .380$，与 rowwise gain 的 $\rho = .442$；family 效应量约 .66–.69。

7. **跨 prompt 稳定性：通道级较高，模块级中等**。通道 rank ρ 中位数 .913，top-1% Jaccard .547；但模块 error 排名跨 cell 只有 .436。

9.2 机制判断

当前证据最符合以下图景：E4M3 的指数位缓冲了 max outlier 对主体值相对精度的冲击，而三位尾数形成广泛、近似恒定的相对舍入误差；极端 scale 仍把动态范围底部的小权重推向 zero，但这些元素的平方能量太小，无法主导总 L2。activation 的分布随 NFE、family、prompt 更动态，且不同 row 的 amax 差异能被 rowwise scale 直接利用，所以 activation tail 对局部误差更有解释力。

10 下一轮实验建议：从描述性定位走向因果验证

10.1 最小、可归因的 lane 设计

优先固定 block 35–45 的 S2-O/S2-Q 高风险集合，并构造 family/block 匹配的低风险 control layers。每次只改变一个因素：

Lane	修改	目的
B	576 层 production tensorwise	配对基线
W15	仅 top-15 层 output-channel weight scale	因果测试 weight underflow rescue，避免同时改 activation
A15	仅 top-15 层 dynamic rowwise activation	因果测试 activation tail 机制
BF15	仅 top-15 层 BF16 Linear oracle	给出这些层可恢复的质量上限
R15	15 个 family/block 匹配的低风险层做同样处理	排除“只是修改任意 15 层”的效应

若 A15 通过 heldout，再分别缩到 S2-Q 与 S2-O，避免把两个机制不同的 family 绑定。sparse residual 应在上述因果链成立后再做 kernel prototype；其首个合理点是 $\alpha = .25$，因为静态压缩率仍约 1.958×，且 outlier fraction 只有 0.267%。

10.2 数据矩阵与统计模型

建议至少使用 6 个内容 prompts × 3 个 seeds，共 18 个配对 cells；calibration 只用于选层和固定 policy，最终结论只看未参与选择的 heldout cells。所有 lanes 共享初始 noise、prompt、seed、scheduler 与 chunk geometry。端到端评估同时报告：

- 逐层 output probe 的 relative L2/cosine 与首次显著分叉位置；
- video/audio latent 与 velocity 的配对误差；
- 解码后视频/音频质量指标及最差 cell；

- DiT FPS、峰值显存、fallback/upcast、scale bytes；
- 每个 cell 的 paired delta，而不是只报跨 cell 平均值。

统计上以 cell 为独立重采样单元，module call 为嵌套观测；可使用 mixed-effects 或 cluster bootstrap，模型形式例如

$$\text{local error} \sim \log T^{(x)} + \text{family} + \text{NFE} + \text{chunk} + (1|\text{module}) + (1|\text{cell}).$$

production gate 应预先冻结最差 cell 容忍度、质量主指标与速度预算，避免在看到结果后挑指标。

11 局限性

1. 权重 amax、channel amax、误差、clip fraction 与 underflow 为全元素精确值；weight quantiles/kurtosis 为每层最多 262,144 的确定性等距样本。
2. 激活 amax 是完整 call 精确值，但 p50/p99.9 与 quantization error probe 是有界样本；不能当成所有激活元素的精确分位数或精确误差。
3. 两个 activation prompts 使用相同 seed，无法独立估计 prompt、seed 与其交互；模块级 $n = 576$ 不等同于 576 个独立内容实验。
4. family-stratified bootstrap 控制了 12 类组成，但未完整建模相邻 block 或共享输入带来的相关性；因此区间用于描述层级关联，不用于系统质量批准。
5. weight-space L2 对 task importance 一视同仁。某些低能量权重可能通过高幅 activation 被放大；AWQ、GPTQ 等工作正是通过 activation 或二阶信息重新加权重要性 [2, 3]。
6. sparse residual 参考了“隔离少量 outlier”的一般思路 [4, 5]，但本轮没有实现真实编码、GEMM kernel 或端到端视频评估。
7. INT8 control 是机制对照，采用 sampled uniform symmetric quantizer；不可与 production FP8 的精确全量结果做绝对性能排名。

12 复现方法与产物索引

12.1 运行命令

以下命令从项目根目录运行；Python 环境为 `/workspace/LTX/.venv/bin/python`。第一步会逐层读取 checkpoint，并在 GPU 上执行所有权重反事实：

```
cd /workspace/catnip-fp8-outlier-statistics

/workspace/LTX/.venv/bin/python scripts/analyze_weight_outliers.py \
  --config config/experiment.json \
  --output-jsonl data/raw/weight_outlier_records.jsonl \
  --run-summary data/raw/weight_outlier_run.json \
  --device cuda:0 --resume

/workspace/LTX/.venv/bin/python scripts/analyze_and_visualize.py \
  --config config/experiment.json \
  --weight-jsonl data/raw/weight_outlier_records.jsonl \
  --processed-dir data/processed \
```

```
  --figures-dir figures \
  --tables-dir tables \
  --latex-macros report/generated_metrics.tex

/workspace/LTX/.venv/bin/python scripts/refine_report_figures.py
/workspace/LTX/.venv/bin/python scripts/int8_control_analysis.py
cd report && latexmk -xelatex -interaction=nonstopmode main.tex
```

12.2 主要机器可读产物

全量层记录 `data/raw/weight_outlier_records.jsonl`；576 records；SHA256 `9346a7013b56ba49724e1b7372756057ccdee4ff8f8ae7f225903a2be5a1add5`。

总摘要 `data/processed/analysis_summary.json`。

INT8 对照摘要 `data/processed/int8_control_summary.json`。

逐层权重表 `data/processed/weight_metrics.csv`。

逐 call 激活表 `data/processed/activation_call_metrics.csv`。

模块聚合与稳定性 `data/processed/activation_module_metrics.csv`、`activation_stability.csv`。

联合风险 `data/processed/joint_module_metrics.csv`。

图表 `figures/` 中 11 组 PDF/PNG。

12.3 最终判定

本轮已经把 “outlier 会不会伤害 FP8” 从泛化直觉改写成 Catnip 的量化结论：**weight outlier 显著伤害动态范围底部，但不支配 tensorwise E4M3 的权重 L2；activation outlier 对局部误差更重要，却需要更广 heldout 和按层因果 lane 才能转化为 production 决策**。下一步最合理的工程动作是针对晚层 S2-O/S2-Q 做 W15/A15/BF15/R15 配对实验，而不是全局更换 scale 粒度。

参考文献

[1] Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. Fp8 formats for deep learning. *arXiv preprint arXiv:2209.05433*, 2022.

[2] 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. *arXiv preprint arXiv:2306.00978*, 2023.

[3] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

[4] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

[5] Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023.