Meeting Follow-Up Report (Unified Edition)

非技术版详细说明 + 技术版 Summary（统一、详尽）

Kecen Li

February 3, 2026

Contents

1 一页式总览（先给“听得懂的人”）

一句话目标：用 **Neuron Activation Coverage (NAC)** 评估模型在不同扰动下的内部神经元行为，判断“模型是否在走原来的正确路径”，并验证它与鲁棒性/泛化能力的关系。

本次工作完成了什么?

- 已**完整跑通官方 NAC-UE 评测流程**（OpenOOD / ood_coverage），并与非 APS（不使用 OOD 校准）对照。
- 已**完成 OODRobustBench 的 CIFAR-10-C 全量评估**（15 corruption × 5 severity）。
- 已**完成自然分布偏移（CIFAR-10.1 / 10.2）评估**。
- 已**完成 Phase3 组合扰动与顺序敏感性实验**并生成可视化。
- 已对核心结论与局限性进行归档,论文素材整理到 `archive/2026-02-02_paper_materials/`。

一句话结论：NAC 在**明显分布偏移/噪声类扰动**上表现较好，但对**近域自然偏移**和**部分对抗扰动**的区分度有限；APS 提供上界，但非 APS / ID-only 更接近真实应用场景。

2 先把 NAC 讲清楚（结合论文原始定义）

这部分严格依据 ICLR 2024 论文 *Neuron Activation Coverage: Rethinking Out-of-Distribution Detection and Generalization*。以下解释既保留数学定义，也尽量用直白语言讲清楚。

2.1 通俗解释：我们到底在看什么?

传统 OOD 检测看的是**模型输出的置信度**，比如最大 softmax。NAC 不仅看输出，而是看**模型内部某一层的神经元行为**，特别是“这些神经元在做出当前判断时到底有多重要”。

简化理解：

- 如果一个输入让模型内部的神经元激活方式**和训练数据相似**，它就更像“正常输入”。
- 如果激活方式**偏离训练时的分布**，就更像“异常输入”。

2.2 论文公式（核心定义）

论文首先定义一个神经元的“激活状态”（不仅是输出值，还包含其对决策的影响）：

$$\hat{z} = \sigma\left(z \odot \frac{\partial D_{KL}(u||p)}{\partial z}\right) \tag{1}$$

其中：

- z 是某层神经元的原始输出；
- $D_{KL}(u||p)$ 是模型输出 p 与均匀分布 u 的 KL divergence；
- $\partial D_{KL}/\partial z$ 衡量该神经元对模型决策的影响；
- σ 是 sigmoid，α 控制其陡峭程度。

接下来定义 NAC（覆盖度函数）：

$$\Phi_i^X(\hat{z}_i; r) = \frac{1}{r}\min\left(\kappa_i^X(\hat{z}_i), r\right) \tag{2}$$

- $\kappa_i^X(\cdot)$ 表示在训练集 X 上的概率密度（PDF）；
- r 是覆盖度“饱和阈值”：超过 r 就认为“充分覆盖”；
- 直觉：**训练数据经常出现的神经元状态 → 高覆盖；罕见状态 → 低覆盖**。

2.3 NAC-UE（用于 OOD 检测）

NAC-UE 就是把所有神经元的覆盖度平均：

$$S(x^*; \hat{f}, X) = \frac{1}{N} \sum_{i=1}^{N} \Phi_i^X(\hat{f}(x^*)_i; r) \tag{3}$$

若 S 较低，则说明该输入触发了训练中很少见的神经元状态 → 更可能是 OOD。

2.4 NAC-ME（用于模型评估/泛化）

论文还提出 NAC-ME：通过对 NAC 分布积分来衡量模型的“覆盖面积”，从而预测其泛化能力（尤其是 OOD 泛化）。

论文重要细节：

- 参数：α（sigmoid 陡峭度），r（覆盖饱和阈值），M（PDF 分桶数）。
- 参数敏感性：论文指出 r 很敏感，α 越陡通常越好。
- 模型选择：ResNet 通常用 layer4，ViT 用 block-11 attention 作为 NAC 层。
- 训练数据：论文强调使用**训练数据（正确分类样本）**建立 NAC 统计分布。

3 本项目如何对齐论文（从概念到实现）

3.1 项目 pipeline（通俗解释 + 技术映射）

通俗解释：我们把流程拆成 3 步：

1. **扰动器**：把图片变“坏”，包括噪声、压缩、几何变形、对抗攻击。
2. **分析器**：让模型前向 + 反向，算出神经元覆盖度。
3. **对比分析**：对比 clean vs. 各类扰动，看 NAC 是否能区分它们。

技术映射：

- NAC 计算：`src/nac_efficient.py` + `src/official_nac.py`。
- 攻击/扰动：AutoAttack、Advex-UAR、DeepG、OODRobustBench。
- 模型：RobustBench 的标准/鲁棒模型。
- 实验脚本:`scripts/run_phase3_correct.py`, `run_total_benchmark.py`, `scripts/run_oodr`

3.2 与论文一致的关键点（已落实）

- 使用论文定义的 $\hat{z}$（含 KL 梯度）和 NAC 函数。
- 使用训练集构造覆盖度分布（profiling），测试集做评估。
- 多层 NAC 评估（ResNet block3 / layer4），并可扩展到 ViT。
- APS（Automatic Parameter Search）与非 APS 两条流程都跑通。

3.3 数据集、模型与评估设置（必须明确的“材料清单”）

为避免误读，以下列出本报告中**实际使用**的数据集与设置：**ID 数据：**

- CIFAR-10（训练集用于 NAC profiling；测试集用于评估与比较）。

OOD 数据（OpenOOD / ood_coverage）：

- Near-OOD：CIFAR-100、TinyImageNet (TIN)。
- Far-OOD：MNIST、SVHN、Texture、Places365。

Corruption 数据（OODRobustBench / CIFAR-10-C）：

- 15 种腐蚀 × 5 个严重度，共 75 条结果（完整表格见 CSV）。

自然分布偏移（OODRobustBench Natural Shift）：

- CIFAR-10.1、CIFAR-10.2。

模型与评估设置：

- 模型：RobustBench 标准 ResNet18 与鲁棒 WideResNet (Gowal2021)。
- Phase3 设置：profiling=1000，test=2000，batch=64，目标层 =block3，AA 版本 =standard。
- OODRB 设置：batch=64，profiling=1000，ID 测试 =10000。
- 指标：AUROC、FPR@95、AUPR_IN、AUPR_OUT、Accuracy、NAC-Acc 相关性。

4 实验结果（结合图表，逐块讲清楚）

4.1 A. 官方 OOD Coverage（APS vs Non-APS）

通俗解释：APS 像“拿到了部分 OOD 的答案再调参数”，因此更强；非 APS 更接近现实。我们必须报告两者差异。

关键结果（Non-APS）：

- Near-OOD（CIFAR-100/TIN）：AUROC **89.83–92.02**，FPR@95 **26.56–35.10**。
- Far-OOD（MNIST/SVHN/Texture/Places365）：AUROC **91.85–96.05**，FPR@95 **14.34–26.73**。
- 汇总指标：Near-OOD 平均 AUROC ≈ **90.93**；Far-OOD 平均 AUROC ≈ **94.60**。

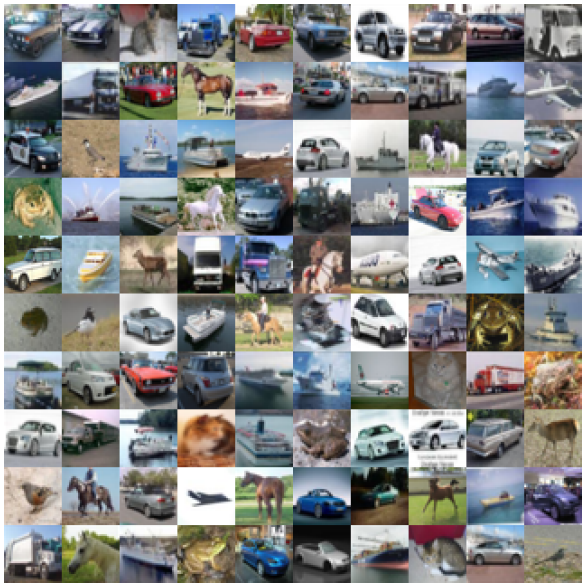
NAC APS vs Non-APS (CIFAR-10)

dataset	FPR@95_APS	AUROC_APS	AUPR_IN_APS	AUPR_OUT_APS	ACC_APS	FPR@95_NonAPS	AUROC_NonAPS	AUPR_IN_NonAPS	AUPR_OUT_NonAPS	ACC_NonAPS
cifar100	99.00	99.00	99.00	99.00	95.32	35.10 ± 0.32	89.83 ± 0.29	87.06 ± 0.35	90.46 ± 0.21	95.06 ± 0.30
farood	19.74	94.53	96.79	88.93	95.32	18.32 ± 0.90	94.60 ± 0.49	96.48 ± 0.52	89.89 ± 0.71	95.06 ± 0.30
mnist	17.74	94.36	98.74	85.66	95.32	15.14 ± 2.64	94.86 ± 1.36	98.76 ± 0.43	87.23 ± 2.33	95.06 ± 0.30
nearood	99.00	99.00	99.00	99.00	95.32	30.83 ± 0.24	90.93 ± 0.22	87.61 ± 0.34	92.13 ± 0.16	95.06 ± 0.30
places365	27.10	92.54	97.40	81.60	95.32	26.73 ± 0.74	91.85 ± 0.29	96.97 ± 0.18	82.14 ± 0.23	95.06 ± 0.30
svhn	17.19	95.35	97.96	90.93	95.32	14.34 ± 1.26	96.05 ± 0.46	98.10 ± 0.36	92.69 ± 0.68	95.06 ± 0.30
texture	16.93	95.87	93.05	97.54	95.32	17.07 ± 0.65	95.64 ± 0.44	92.10 ± 1.32	97.51 ± 0.20	95.06 ± 0.30
tin	99.00	99.00	99.00	99.00	95.32	26.56 ± 0.38	92.02 ± 0.20	88.16 ± 0.39	93.80 ± 0.13	95.06 ± 0.30

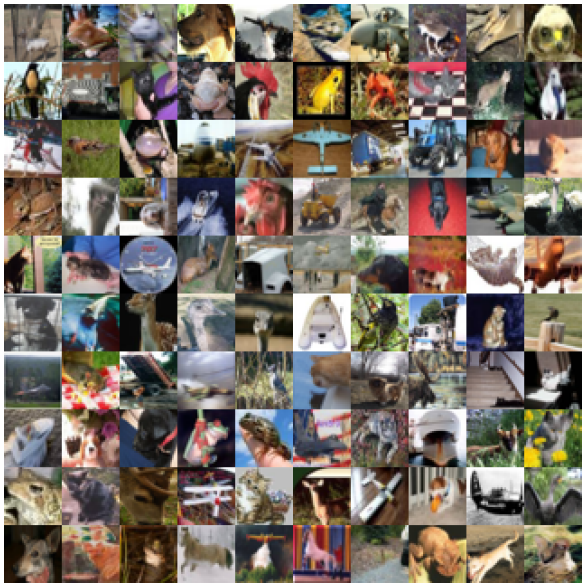
Figure 1: APS vs Non-APS 结果对比（CIFAR-10, NAC-UE）

4.2 B. Top/Bottom 质检（定性解释）

通俗解释：我们不是只看数字，还看“分数最高/最低的样本长什么样”。如果低分样本确实“奇怪”，说明 NAC 是可信的。



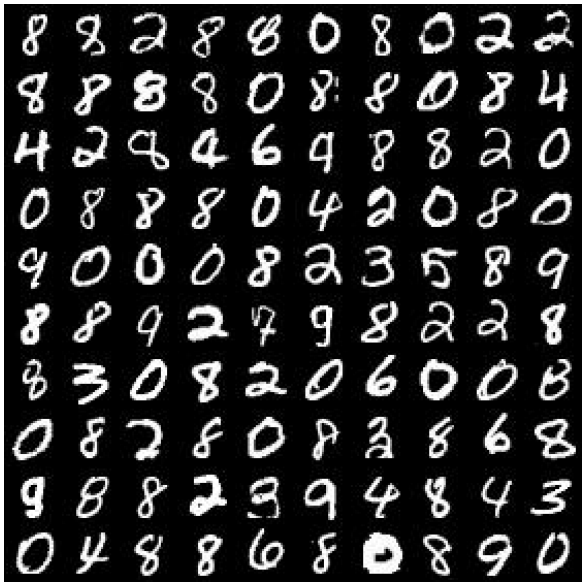
(a) ID Top-100（高 NAC）



(b) ID Bottom-100（低 NAC）

Figure 2: ID 样本的 Top/Bottom 质检

(a) OOD Top-100（MNIST, Non-APS）



(b) OOD Bottom-100（MNIST, Non-APS）

Figure 3: OOD 样本质检（MNIST）

4.3 C. ID-only 阈值（真实场景的最低保障）

通俗解释：真实应用中我们无法用 OOD 样本调参，所以我们只用 ID 数据设阈值，看 OOD 能否被检测出来。**关键数值（来自 threshold_table.csv）：**

- 1% ID-FPR 阈值：Near-OOD (CIFAR-100/TIN) TPR **0.118–0.152**。
- 5% ID-FPR 阈值：Near-OOD TPR **0.452–0.514**。
- Far-OOD（SVHN/Texture）在 5% 阈值时 TPR 通常 **0.66–0.78**。

NAC ID-only Thresholds (val) vs OOD Detection

dataset	thr_p1	id_fpr@thr_p1	ood_tpr@thr_p1	thr_p5	id_fpr@thr_p5	ood_tpr@thr_p5	thr_fpr95	id_fpr@thr_fpr95	ood_tpr@thr_fpr95
cifar100	0.6243	0.0119	0.1181	0.6494	0.0536	0.4524	0.7175	0.3557	0.9500
tin	0.6243	0.0119	0.1515	0.6494	0.0536	0.5135	0.7046	0.2653	0.9500
mnist	0.6243	0.0119	0.1014	0.6494	0.0536	0.6600	0.6796	0.1516	0.9500
svhn	0.6243	0.0119	0.2952	0.6494	0.0536	0.7759	0.6737	0.1282	0.9501
texture	0.6243	0.0119	0.3112	0.6494	0.0536	0.7064	0.6863	0.1799	0.9500
places365	0.6243	0.0119	0.1469	0.6494	0.0536	0.4824	0.7034	0.2572	0.9500

Figure 4: ID-only 阈值（1% / 5%）下的 OOD 检出率

4.4 D. OODRobustBench：CIFAR-10-C 全量结果

结果总结：

- 完整评估 75 条（15 corruption × 5 severity）。
- 平均 AUROC = **0.698**，平均 FPR@95 = **0.779**（见 oodrb_nac_20260202_054121.csv）。
- 最容易被检出的：强噪声 / 对比度 / 结构破坏类（如 impulse_noise, contrast）。
- 最难被检出的：亮度 / 雾化（低强度接近 0.5，接近随机）。

4.5 E. 自然分布偏移（CIFAR-10.1 / 10.2）

结果：AUROC 平均 **0.563**，FPR@95 平均 **0.930**（见 oodrb_nac_20260202_205623.csv）。

结论（通俗解释）：自然偏移是“很像但不完全一样”的分布变化，NAC 在这种情况下几乎失效，接近随机水平。

4.6 F. Phase3 组合扰动与顺序效应

通俗解释：我们不仅看单一扰动，还看“先转再加噪声”和“先加噪声再转”是否不同。

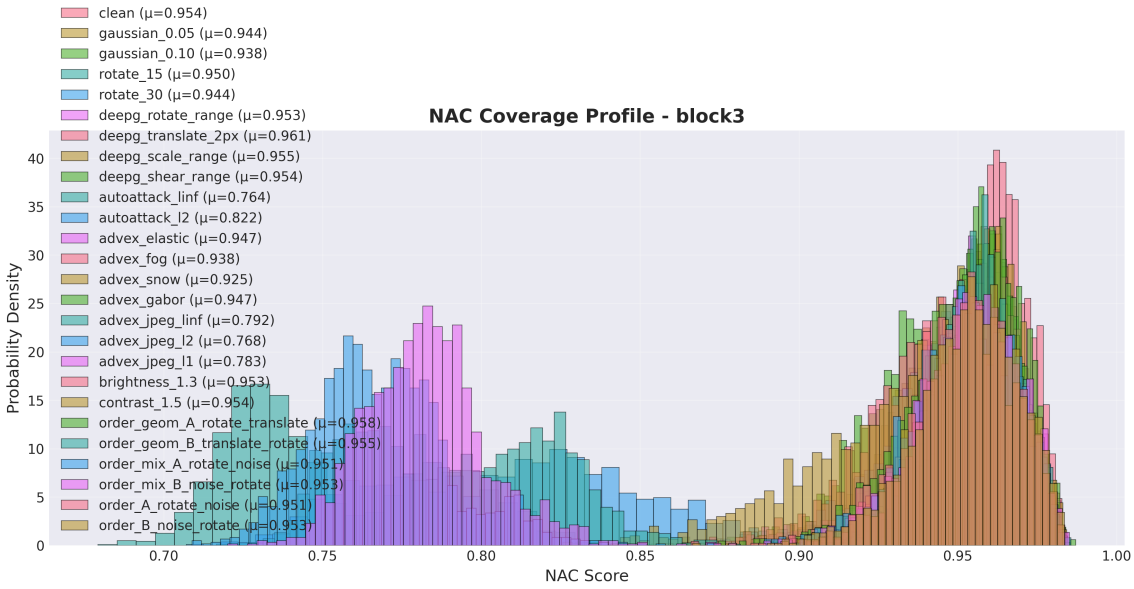


Figure 5: Phase3 NAC 分布直方图（block3）

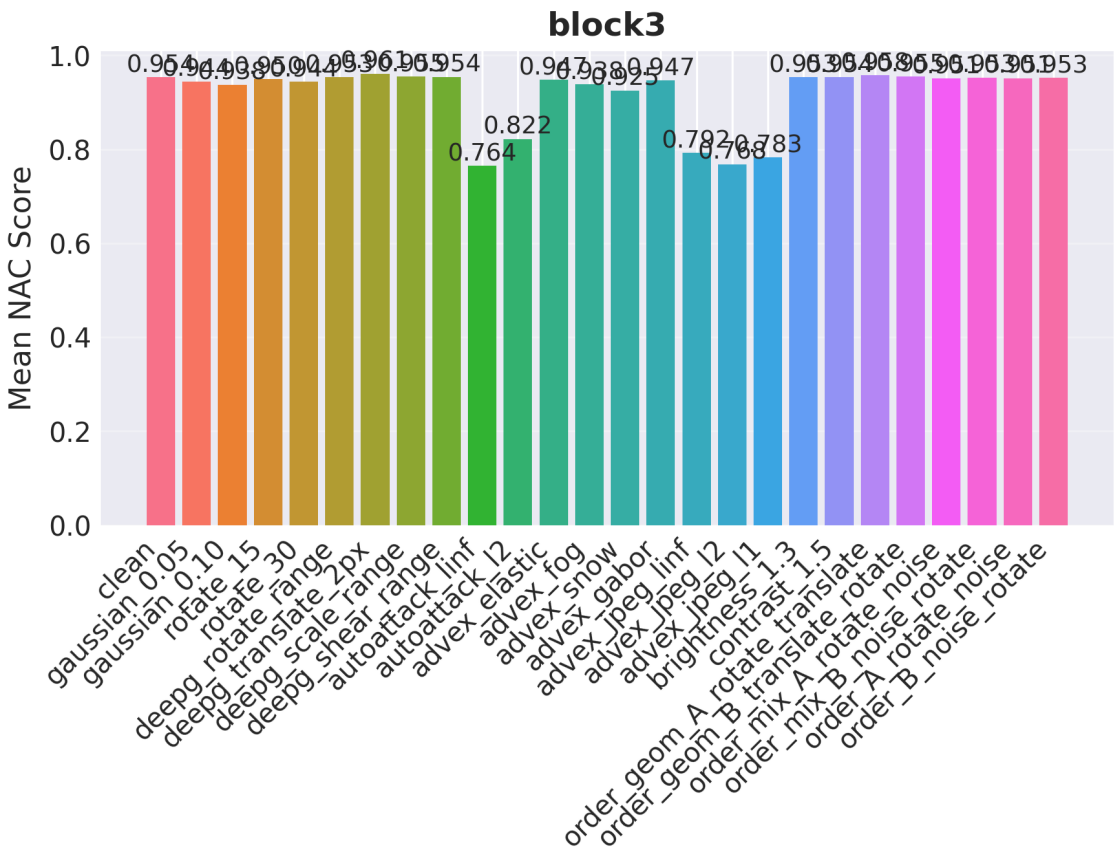
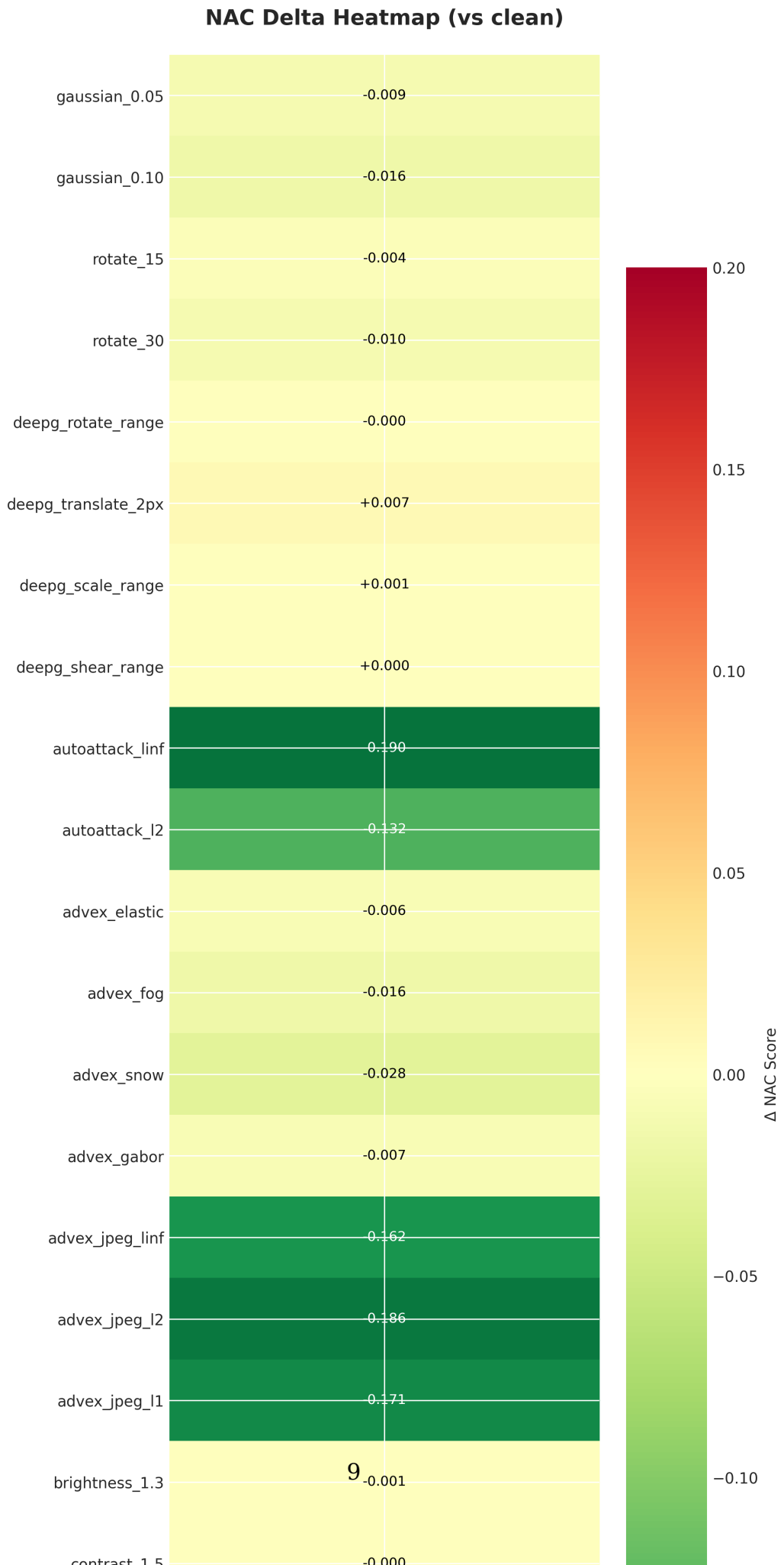


Figure 6: Phase3 对比条形图


NAC Delta Heatmap (vs clean)
gaussian_0.05
gaussian_0.10
rotate_15
rotate_30
deepg_rotate_range
deepg_translate_2px
deepg_scale_range
deepg_shear_range
autoattack_linf
autoattack_l2
advex_elastic
advex_fog
advex_snow
advex_gabor
advex_jpeg_linf
advex_jpeg_l2
advex_jpeg_l1
brightness_1.3
contrast_1.5
-0.009
-0.016
-0.004
-0.010
-0.000
+0.007
+0.001
+0.000
-0.190
-0.132
-0.006
-0.016
-0.028
-0.007
-0.162
-0.186
-0.171
-0.001
-0.000
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
Δ NAC Score

关键观察（数值化）：

- NAC-Acc 相关系数 = **0.6894**（排除 clean）。
- 最大 NAC 降幅来自 AutoAttack 与 JPEG（如 AutoAttack Linf, JPEG L1/L2）。
- DeepG 几何扰动 NAC 变化接近 0（部分为正）。
- 组合顺序差异非常小（A→B 与 B→A 差异 **约 1e-3**）。

4.7 G. 标准 vs 鲁棒模型对照（归档结果）

现象：

- 鲁棒模型在 AutoAttack 下准确率明显更高。
- 在 Rotation / Fog / Snow 等扰动下，鲁棒模型的 NAC 分离更明显。

5 讨论：这些结果说明了什么？

- NAC 对**明显分布偏移**有效（噪声/结构破坏），对**近域自然偏移**弱。
- APS 可以显著提高表面成绩，但具有过拟合风险（官方 README 明确提醒）。
- 非 APS / ID-only 更接近现实场景，是必须报告的“真实能力”。
- 组合扰动与顺序效应在 NAC 上影响不大，说明 NAC 更多体现“幅度型变化”而非“路径顺序”。

6 局限性与风险（必须提前说明）

- 论文官方 README 有明确“overfitting issue”警告，APS 结果应视作上界。
- NAC 对自然偏移（CIFAR-10.1/10.2）检测能力弱，接近随机。
- 当前 profiling 默认使用训练集，不一定仅保留“正确分类样本”（与论文设定略有差别）。

7 可复现材料路径（给导师或审稿人查看）

- 统一素材目录：`archive/2026-02-02_paper_materials/`
- Phase3 输出：`phase3_output_20260202/`
- OODRB Corruption CSV：`oodrb_results_20260202/oodrb_nac_20260202_054121.csv`
- Natural Shift CSV：`oodrb_results_20260202/oodrb_nac_20260202_205623.csv`
- APS/Non-APS 分析：`ood_coverage/analysis/`

8 后续工作（还可以继续做的）

- JPEG 扫描结果合并，生成完整曲线（已拆 job）。
- ResNet50 / ViT-B/16 层名确认后扩展实验。
- ImageNet 版本 OODRB（需要 ImageNet 数据）。