

Phase 2 完整技术报告：US ATUS 复现、社会分层与方法稳健性

项目：Yucheng_Project

阶段：Phase 2 (US-first replication & expansion)

数据：ATUS 2024 (quick) + ATUS pooled 2003–2024 (full)

核心通道：activity + location

核心口径：accuracy / macro-F1 / persistence baseline / delta vs persistence

可视化：10 张

更新时间：2026-04-03

状态：Phase 2 完整长报告 (Advisor Review Version) 实验足迹：49 个 Phase 2 原始实验 JSON
(其中 41 个完成结果 + 8 个 skipped_small_group 占位) + 2 个 tracker JSON +
results/phase2_summary/ 统一 summary 层 + 10 张最终主图

目录

1. 执行摘要
2. Phase 2 在全项目中的角色：它究竟回答什么问题
3. 数据、通道、分组与可比性边界
4. 实验矩阵与完成度：哪些已经完整，哪些必须诚实交代
5. A1 基准实验：US 的惯性上限与模型可学习空间
6. B1 分组实验：社会分层在 US 中如何出现
7. pooled 复核：为什么 quick 阶段的负 delta 不能被直接当成结论
8. 稳健性实验：粒度、时间尺度与样本选择的机制含义
9. 与 Phase 1 的统一对表：相同机制，不同幅度
10. 可视化完备度评估：现在的 Phase 2 是否已经“像一份完整报告”
11. 方法论贡献、局限与风险控制
12. 结论：Phase 2 给 Phase 3 留下了什么
13. 附录：图表索引与关键文件

1) 执行摘要

Phase 2 的目标不是把 US 分数“卷到最高”，而是回答一个更关键的问题：**Phase 1 在 UK 上识别出的那套机制，到 US 这样一个不同制度与编码环境中，是否仍然成立。**

这一阶段最终给出了四个最重要的结论：

1. **短期惯性具有跨国普遍性。** US `activity` 的 persistence baseline 为 **88.76%**，与 UK Phase 1 的 **88.76%** 几乎完全一致。这说明“下一时隙大概率延续上一时隙”的基本事实，并不是 UK 独有现象。
2. **US `activity` 上 Transformer 在 quick 阶段几乎追平 persistence，但不稳定超越它。** ATUS 2024 quick 数据中，Transformer 为 **88.56%**，仅低于 persistence **0.20pp**；SGD 则低 **2.28pp**。这表明在更异质的 US 环境里，线性模型明显更脆弱，而深度序列模型已经逼近惯性上限。
3. **一旦进入 pooled 分组复核，很多 quick 阶段的“负 delta”会显著收敛甚至转正。** 以 `income_mid` 为例，delta 从 **-3.68pp** 修正到 **+0.41pp**；`weekday` 从 **-1.53pp** 修正到 **+0.49pp**。这意味着 quick 阶段不能被当成“机制盖棺定论”，它首先是一个样本条件下的近似观测。
4. **US 不推翻 Phase 1，而是把 Phase 1 的叙事从“高可预测性”推进到“样本充分性决定可学习增益能否显形”。** 也就是说，惯性依然强，但模型额外能学到的那部分规律，在 US 中更依赖样本量、分组规模、变量可用性与编码异质性。

从全项目角度看，Phase 2 的价值非常明确：

- 它把 Phase 1 从“单国成立”推进到“跨制度环境具有可迁移性”；
 - 它逼出了一个关键的方法论修正：**不能把 quick 阶段的负 delta 直接解释成 persistence 不可超越**；
 - 它为 Phase 3 提前筛掉了一批不稳的主线设定，最终把 `activity + persistence + delta + sex/age_bin` 这套口径推成了多国验证的默认主线。
-

2) Phase 2 在全项目中的角色：它究竟回答什么问题

2.1 三阶段逻辑中的中间桥梁

如果把整个项目看成一个递进式论证，那么三阶段分工非常清楚：

- **Phase 1 (UK)**：证明这个框架本身成立，并在一个高质量、多通道、分层充分的数据环境里提炼出理论主叙事。
- **Phase 2 (US)**：检验 Phase 1 识别出来的核心机制，在另一个国家、另一种调查结构、另一套变量约束下是否还能成立。
- **Phase 3 (MTUS)**：把这一套机制推向多国并行验证，建立更强的外部效度。

因此，Phase 2 不是“再做一个小号的 Phase 1”，而是整个项目中最关键的**迁移检验环节**。

2.2 Phase 2 需要回答的三个核心问题

Phase 2 实际上集中回答三个问题：

1. **惯性是否普适**：如果 UK 的高可预测性主要来自短期惯性，那么 US 是否也一样？
2. **模型增益是否可迁移**：在新的制度环境与编码体系下，Transformer 相对 persistence 的额外增益是否依然存在？
3. **社会分层是否方向一致**：收入、性别、经济状态等因素在 US 中会不会改变“谁更可预测”的排序与幅度？

这三个问题决定了 Phase 2 的评价标准不应仅仅是一个单点准确率，而应同时看：

- baseline 是否对齐；
- delta 是否稳定；
- 分组差异是否具有方向上的一致性；
- 结论是否经得起 pooled 大样本复核。

2.3 为什么 Phase 2 比看上去更难

表面上看，ATUS 只是“把 UK 框架搬到 US”；但实际难点不少：

- US ATUS 缺失 `enjoyment` 与 `with_whom` 两个通道，天然削弱了多通道叙事的完整性；
- 原始数据是事件流，不是天然的 10 分钟面板，需要额外对齐；
- `income`、`econstat` 这类变量的可用性与稳定性并不如 UK 一致；
- `pooled` 与 `quick` 的结论差异很容易把人带到过度解释的陷阱里。

也正因为如此，Phase 2 不能写成一份“跑完脚本的流水账”，而必须写成一份**机制辨析报告**。这也是本次增强版 Markdown 的核心改动方向。

3) 数据、通道、分组与可比性边界

3.1 数据来源与时间对齐

Phase 2 使用两层 US 数据：

- quick 层**：ATUS 2024，用于快速验证主线方向；
- pooled 层**：ATUS 2003-2024，用于检验分组结果是否受样本量强影响。

数据处理口径与 Phase 1 保持一致：

- 事件流展开为 10min 时间槽；
- 每天固定为 144 槽；
- 采用按人或按人-天分层的切分思路，尽量避免泄漏；
- 结果解释统一使用 accuracy、macro-F1、persistence baseline 与 delta vs persistence。

3.2 通道可用性边界

Phase 2 最大的结构性边界，是 US 数据并不具备 UK 那种完整的四通道条件。

通道	UK Phase 1	US Phase 2	在 Phase 2 中的角色
activity	□	□	主通道，承担跨阶段主叙事
location	□	□	机制通道，检验可学习的低噪转移
enjoyment	□	□	US 不可用，不应强行平行叙述
with_whom	□	□	US 不可用，不应强行平行叙述

这意味着一个很重要的方法论态度：**Phase 2 的任务不是伪造与 UK 完全对称的故事，而是在真实数据边界内把可比较的部分做扎实。**

3.3 分组变量与优先级

Phase 2 主要使用以下分组维度：

- income_bin
- sex
- econstat
- age_bin（主要在 pooled）

- `is_weekend` (主要在 `pooled`)
- `survey_period` (主要在 `pooled`)

就后续项目价值而言，它们的优先级并不相同：

- **高优先级**：`income_bin`，`sex`，`age_bin`
- **中优先级**：`is_weekend`，`survey_period`
- **谨慎使用**：`econstat`，因为 `pooled` 场景中一些组样本过小，不能轻率上升为主结论

3.4 质量控制与 bug 修复

Phase 2 的一个重要价值，在于它逼出了几处如果不修正就会污染结论的关键问题。

问题	风险	修复结果
US 中 <code>enjoyment</code> / <code>with_whom</code> 不存在，但早期实现可能给出假性 <code>acc=1.0</code>	伪造“完美结果”	已改为 <code>NaN</code> 处理，并在 runner 侧加通道有效性检查
30min <code>downsample</code> 结果曾被固定 <code>144 slots/day</code> 逻辑误伤	D1 结果为空或偏差	已支持 <code>48 slots/day</code>
A1 full SGD 全量 flat feature 容易 OOM	无法对大样本给出稳定结论	通过 <code>sample cap</code> 与 <code>pooled grouped</code> 复核来替代僵硬全量 flat baseline

这部分内容必须写进正式报告，因为它直接关系到导师对数据质量和实验可信度的判断。Phase 2 的“数据完整性”不仅是“有多少表”，也是“有没有把不该信的结果及时剔除掉”。

4) 实验矩阵与完成度：哪些已经完整，哪些必须诚实交代

4.1 当前 Phase 2 的完整交付板块

板块	内容	状态	是否进入主报告
A1 quick	US activity/location baseline	☐ 完成	☐
B1 quick	income_bin / sex / econstat	☐ 完成	☐
pooled grouped	income_bin / sex / age_bin / is_weekend / survey_period	☐ 完成	☐
C1/D1/E1	coarse/fine、10/30min、weekday/full	☐ 完成	☐
pooled econstat full	仅部分组稳定	⚠ 部分完成	仅作边界说明
可视化图组	Phase 2 专属 10 图	☐ 完成	☐

4.1.1 为什么最终主文只放 10 张图，但工作量并不薄

Phase 2 的最终交付故意采用“少而强”的图链策略，而不是把所有中间运行结果逐张堆入正文。换句话说，**10 张图是经过筛选后的最终证据层，不是全部实验层。**

如果只看导师包中的主文与图链，容易低估实际工作量；但从 results/ 目录回看，Phase 2 直接相关的 US 结果产物至少包括：

层级	结果足迹	说明
raw experiment layer	49 个 JSON	41 个完成结果 + 8 个 skipped_small_group 占位，覆盖 quick、pooled、full-support 与辅助探测
tracker layer	2 个 JSON	us_phase2_expansion_index.json 、 us_phase2_expansion_summary.json
unified summary layer	6 个 CSV + 1 个 MD	results/phase2_summary/ ，把 quick / pooled / methodology / support 统一收束
final figure chain	10 张 PNG	results/phase2_figures/

因此，Phase 2 呈现为 10 张图，并不意味着 Phase 2 只做了 10 组实验；更准确的说法是：**Phase 2 把 49 个原始实验 JSON、2 个 tracker JSON 与一层新的 phase2_summary 汇总表，压缩成了一条可供导师快速判断的 10 图证据链。**

4.2 为什么这份完成度已经具备正式汇报条件

如果要判断 Phase 2 是否已经从“实验执行阶段”进入“正式汇报阶段”，主要看三件事：

1. 有没有一个清晰的 baseline ；
2. 有没有一组可信的 grouped evidence ；
3. 有没有对 negative result、样本量效应与可比性边界做充分解释。

当前版本已经同时具备这三点：

- baseline 有；
- grouped evidence 有，而且 quick 与 pooled 两层都有；
- 负 delta 的解释与样本校正证据也有。

也就是说，Phase 2 已不再是“结果碎片堆积”，而是一个**有机制、有反证、有边界意识**的阶段性正式报告。

4.3 哪些地方仍然不应过度承诺


为了让这份报告更像 Phase 1，而不是更像“宣传稿”，这里必须把边界说清楚：

- **US 不能做完整四通道叙事**，所以不要把缺失的 `enjoyment` / `with_whom` 强行写成“只是还没展开”；
- **econstat 在 pooled 下不够稳**，所以它不应成为 Phase 2 的中心发现；
- **A1 full 全量 US 并非所有模型都能无代价跑完**，所以 pooled 证据在 Phase 2 中主要体现为 grouped 复核与 sample-size correction，而不是完全对称的“大一统全样本总表”。

这种诚实并不会削弱 Phase 2，反而会让它在导师那里显得更成熟。

5) A1 基准实验：US 的惯性上限与模型可学习空间

5.1 US baseline 总览

 Phase 2 Figure 1: US baseline

Phase 2 的第一张主图直接回答最基础的问题：US 的两个可用通道上，模型相对 persistence 的位置分别在哪里。

通道	模型	Accuracy	Macro-F1	Δ vs Persistence
activity	Persistence	88.76%	68.75%	0.00pp
activity	SGD	86.48%	40.12%	-2.28pp
activity	Transformer	88.56%	61.06%	-0.20pp
location	Persistence	92.84%	72.08%	0.00pp
location	SGD	94.30%	57.11%	+1.46pp
location	Transformer	94.40%	62.36%	+1.56pp

5.2 Activity：US 中“惯性仍然强，但额外增益更难显形”

发现 5.2.1

activity 的 persistence 仍然高达 **88.76%**，这与 UK Phase 1 的水平几乎一致。这一结果非常重要，因为它说明：

- 日常行为在 10 分钟尺度上的延续性，并不是 UK 样本特有；
- 高可预测性的第一来源依旧是惯性，而不是某个特定国家的制度结构；
- Phase 1 中“高准确率首先是高 stay-rate 的反映”这一逻辑，在 US 并没有被推翻。

发现 5.2.2

Transformer 在 activity 上只比 persistence 低 **0.20pp**，但 SGD 低 **2.28pp**。这说明：

- 线性模型在 US 环境下更容易被异质性拖垮；
- 深度序列模型并没有轻松“碾压惯性”，但已经非常接近它；
- Phase 2 里最值得关注的，不是“Transformer 有没有立刻大幅超越”，而是它相对 SGD 明显更接近上限。


5.3 Location : US 中最清晰的“可学习增益”

location 的情况与 activity 不同：无论 SGD 还是 Transformer，都稳定高于 persistence。

这意味着 location 不是单纯靠“延续上一时隙”就能解释完的；模型确实学到了额外结构。换句话说：

- 在 US 中，位置变化虽然稀少，但其转移模式更规则、更低噪；
- location 是 Phase 2 中最适合作为“模型额外可学习性”展示的机制通道；
- 这也解释了为什么后续 Phase 3 虽然仍以 activity 为主轴，但 location 在 Phase 2 扮演了方法论证明的关键角色。

5.4 仅看 accuracy 远远不够

 Phase 2 Figure 5: Macro-F1 and delta

这一图是本次增强版报告专门加进去的，因为它能弥补“只看 accuracy 会显得过于乐观”的问题。

发现 5.4.1

在 activity 上，Transformer 的 accuracy 几乎追平 persistence，但 macro-F1 仍显著低于 persistence。这说明：

- 它更擅长保持总体预测正确率，而未必同样擅长把小类活动预测好；
- “接近 persistence”不意味着“全面学会了转移结构”；
- Phase 2 的叙事必须同时写 accuracy 与 macro-F1，否则容易把“总体接近”误写成“机制等价”。

发现 5.4.2

在 location 上，SGD 和 Transformer 的 accuracy 已经高于 persistence，但 macro-F1 并没有同步更高。这进一步说明：

- US 中模型可以在主类和常见转移上获得优势；
 - 但少数位置类别的平衡表现仍未足够理想；
 - 因此，Phase 2 的“模型有效”应该被写成**对主流结构有增益、对尾部类别仍需谨慎**，而不是简单的“全面超越”。
-

6) B1 分组实验：社会分层在 US 中如何出现

6.1 quick 分组结果首先告诉了我们什么

先看 quick 阶段（ATUS 2024）的 grouped evidence：

income_bin (Transformer, quick)

组别	Accuracy	Persistence	Δ vs Persistence
low	81.87%	89.93%	-8.06pp
mid	85.69%	89.36%	-3.68pp
high	88.45%	88.80%	-0.35pp
unknown	88.12%	88.37%	-0.25pp

sex (Transformer, quick)

组别	Accuracy
male	88.76%
female	87.45%

econstat (Transformer, quick)

组别	Accuracy	说明
employed	88.2%	主体组，接近 baseline
unemployed	86.1%	明显更难
other	59.9%	极小样本/高异质组，不能直接上升为主结论

6.2 quick 阶段最直观的图像

Phase 2 Figure 2: Stratified activity

这张图的关键价值在于，它把 quick 与 pooled 放在一张图里，让我们不至于把 quick 的点估计误当成最后答案。

发现 6.2.1

如果只看 quick，会很容易得出一个夸张的结论：**低收入群体比高收入群体“难预测得多”**。但这种差距过大，恰恰提示了样本量效应和组内异质性可能在放大结果。

发现 6.2.2

性别差异在 US 中存在，但幅度明显小于收入切片。这与 UK Phase 1 的方向是一致的：性别不是零效应，但也不是最强解释维度。

发现 6.2.3

年龄在 pooled 中才开始变得清楚，这提醒我们：**有些分层不是不存在，而是在 quick 阶段还没有足够样本把它“稳定显影”出来。**

6.3 为什么 Phase 2 不应把 quick 分组直接写成结论

Phase 2 的成熟度恰恰体现在这里：我们没有把 quick 的负 delta 直接写成“US 中模型就是学不到任何额外规律”，而是继续做了 pooled 复核。

这一步非常关键，因为如果跳过它，Phase 2 会变成一份“看上去结论很鲜明、实际上受样本条件严重约束”的文档。导师如果认真看，很容易马上追问：

- 这是不是样本太少导致的？
- 为什么 low income 会比 high income 低这么多？
- 为什么 weekday-only 反而更差？


本报告后面的 pooled 章节，就是专门回答这些追问的。

7) pooled 复核：为什么 quick 阶段的负 delta 不能被直接当成结论


7.1 pooled grouped 是 Phase 2 最重要的“修正层”

如果说 quick 用来回答“方向大致如何”，那么 pooled grouped 用来回答“这个方向经不经得起更多样本与更多切片”。

先看 pooled grouped 的总体分布：

Phase 2 Figure 3: pooled grouped deltas

再看一张更浓缩的 heatmap：

Phase 2 Figure 6: pooled delta heatmap

7.2 pooled 的核心结论可以浓缩成一句话

在 pooled 大样本条件下，US 中许多原本在 quick 中为负的 delta 会显著上移，location 更是对所有分组都呈现稳定正增益。

7.3 pooled 分组的总体统计

按通道汇总的平均 delta

通道	pooled mean Δ	最小值	最大值	解释
activity	+0.33pp	-0.27pp	+0.52pp	小幅但真实的可学习增益开始显形
location	+1.46pp	+1.26pp	+1.70pp	几乎所有分组都稳定超越 persistence

按维度与通道汇总的 mean delta

分组维度	Activity mean Δ	Location mean Δ
income_bin	+0.11pp	+1.37pp
sex	+0.44pp	+1.48pp
age_bin	+0.33pp	+1.52pp
is_weekend	+0.44pp	+1.48pp
survey_period	+0.52pp	+1.51pp

这张表本身就说明了两个重要事实：

1. location 的 learnable regularity 明显强于 activity ；
2. activity 虽然增益不大，但 pooled 下已经不再是“全面负值”。

7.4 income 维度是最值得谨慎解读的例子

income_bin 在 Phase 2 中特别重要，因为它既是最有社会科学吸引力的维度，也是最容易被样本量误导的维度。

quick 到 pooled 的变化

组别	quick Δ	pooled Δ	变化
low	-8.06pp	-0.27pp	+7.79pp
mid	-3.68pp	+0.41pp	+4.09pp
high	-0.35pp	+0.20pp	+0.55pp

发现 7.4.1

如果只看 quick，会以为低收入群体几乎“不可能学到额外结构”；但 pooled 后，low 已经从 **-8.06pp** 修正到 **-0.27pp**。这说明 quick 中的巨大负值，至少有很大一部分来自样本不足和不稳定切片，而不是机制本身。

发现 7.4.2

mid 组从 **-3.68pp** 到 **+0.41pp** 的转变更具方法论意义，因为它告诉我们：

- negative delta 可以被样本量“修正”；
- persistence 并不是一个绝对不可超越的天花板；
- 如果训练样本足够、分组不至于过碎，Transformer 对 activity 的确可以学到少量但稳定的额外规律。

7.5 sex、age、weekday 与 survey_period 提供了“更稳的正值证据”


相比 income，sex、age_bin、is_weekend 与 survey_period 在 pooled 下更稳定。

其中几个有代表性的结果如下：

- sex-female activity : **+0.49pp**
- sex-male activity : **+0.39pp**
- age-middle activity : **+0.49pp**
- survey_period-pre_covid activity : **+0.52pp**

这几组的共同意义是：在 US 中，Transformer 的 activity 增益虽然不大，但在若干较稳定的 pooled 切片里，已经可以持续为正。

7.6 sample-size correction 是 Phase 2 最关键的反证证据

 Phase 2 Figure 8: sample-size correction

这张图比单纯的 grouped bar 更重要，因为它把“small sample bias”画成了可以一眼看懂的轨迹。

发现 7.6.1

income_mid 从 **-3.68pp** 到 **+0.41pp**，weekday 从 **-1.53pp** 到 **+0.49pp**，这已经不是“轻微波动”，而是结论层级的修正。


发现 7.6.2

这意味着 Phase 2 最应该写进方法论部分的一句话是：

在高异质分组任务中，negative delta 首先应被视为一个待复核现象，而不是立即上升为 persistence 的绝对不可超越性。

这句话的重要性甚至超过某个单独的准确率数值，因为它决定了整个项目后续如何读 quick results。

7.7 pooled support-width 告诉我们：哪些正值只是刚冒头，哪些已经有更强支撑

 Phase 2 Figure 9: pooled uncertainty

在这一步里，我们不再只画 pooled mean delta，而是把每个 grouped slice 的 n_test 也转译成一个**保守的支持宽度区间**。它不是严格的 paired significance test，而是一个非常实用的判断标准：当前这个切片的正负方向，到底已经站得多稳。

发现 7.7.1

`income_low activity` 的 pooled mean 只有 **-0.27pp**，其保守支持宽度约为 **±0.60pp**。这意味着它现在更应该被读成“贴近零、仍待谨慎”而不是“US 低收入组存在巨大负增益”。也就是说，quick 阶段那个 **-8.06pp** 的视觉冲击，在 pooled 后已经被压缩回了一个**接近零附近的小残差**。


发现 7.7.2

与此相对，`survey_period-pre_covid activity` 的 pooled mean 为 **+0.52pp**，支持宽度约 **±0.20pp**；`sex-female activity` 为 **+0.49pp**，支持宽度约 **±0.28pp**。这说明 Phase 2 的正值证据并不只是“偶然浮出零线”，而是在若干更大样本切片上已经形成了更可信的正向支持。

发现 7.7.3

`location` 通道的 pooled 结果更稳：各组 mean delta 落在 **+1.26pp ~ +1.70pp** 之间，而支持宽度大多只有 **±0.15pp ~ ±0.47pp**。因此，Phase 2 中最强的“模型确实学到惯性之外结构”的证据，依然首先来自 `location`，而不是 `activity`。

7.8 把所有 pooled 切片一起摆出来后，support-size pattern 就更清楚

Phase 2 Figure 10: pooled support map

这张图把全部 pooled grouped slices 一起放到 `n_test` 的对数尺度上，目的是回答一个比“有没有正值”更成熟的问题：**正值出现在哪些支持规模上，它们是零散的，还是已经形成结构性的分布模式。**


发现 7.8.1

在这张 support-size map 里，`location` 的所有点都稳稳落在零线上方，而且随着样本规模扩大并没有回落到零附近。这说明 `location` 的 learnable gain 不是某个单独分组的巧合，而是一个横跨 `income / sex / age / weekday / survey period` 的稳定现象。

发现 7.8.2

`activity` 的 pooled 点云则明显更贴近零线，但模式并不混乱：除了 `income_low` 仍略低于零之外，其余切片基本都集中在 **+0.2pp ~ +0.52pp** 之间。换句话说，Phase 2 现在已经不是“activity 普遍为负”，而是“activity 的额外增益很小、但在足够样本条件下会系统性地靠近或越过零线”。这正是 Phase 2 最值得保留下来的方法论结论。

8) 稳健性实验：粒度、时间尺度与样本选择的机制含义

Phase 2 Figure 7: methodology sensitivity

除了 baseline 与 grouped evidence，Phase 2 还必须回答一个问题：**这些结论是不是只是在某个偶然设定下才成立？**

为此，我们保留了三组稳健性实验。

8.1 C1 : fine vs coarse

粒度	Model Acc	Persistence	Δ vs Persistence	Macro-F1
fine	87.72%	88.71%	-0.99pp	48.36
coarse	90.11%	90.12%	-0.00pp	81.70

发现 8.1.1

coarse label 明显更容易，但这并不等于更有科学价值。恰恰相反：

- coarse 把许多真正有意义的细粒度差异压平了；
- coarse 下 persistence 自己就更接近上限；
- 如果论文主线想讨论“行为复杂性”，就必须继续以 fine 作为主轴。

8.2 D1 : 10min vs 30min

时间粒度	Model Acc	Persistence	Δ vs Persistence	Macro-F1
10min	87.72%	88.71%	-0.99pp	48.36
30min	73.47%	73.56%	-0.09pp	31.50

发现 8.2.1

30min 会让总体准确率大幅下降。这是一个很重要的技术与理论双重信息：

- 技术上，说明原始 10min 粒度确实包含更强的预测信号；
- 理论上，说明行为组织的关键结构在更细尺度上已经存在，过度聚合反而把它抹掉了。

8.3 E1 : weekday-only vs full sample

样本范围	Model Acc	Persistence	Δ vs Persistence
full_week	87.72%	88.71%	-0.99pp
weekday_only	86.85%	88.39%	-1.53pp

发现 8.3.1

“工作日更规律，所以更好预测”这个直觉在 quick 阶段并没有自动成立。相反，weekday-only 反而更差。这进一步支持了 Phase 2 的方法论主张：**不能把社会学直觉直接当成统计结果，样本量与样本构成往往先于直觉解释。**

8.4 稳健性实验的总体意义

这三组实验合在一起，给 Phase 2 带来了比单纯 baseline 更深的价值：

- 它说明 fine 粒度保留了真正困难也真正有价值的行为复杂性；
 - 它说明 10min 并不是一个随便选的参数，而是实证上更有信息量的尺度；
 - 它说明样本选择（如 weekday-only）并不会天然简化问题，反而可能带来额外偏差。
-

9) 与 Phase 1 的统一对表：相同机制，不同幅度

Phase 2 Figure 4: UK vs US

9.1 Activity 主线对表

Phase	数据	Persistence	SGD	Transformer	Transformer Δ
Phase 1	UK	88.76%	90.95%	91.00%	+2.24pp
Phase 2	US	88.76%	86.48%	88.56%	-0.20pp

9.2 这张对表应该怎么读

这张表最容易被误读成“UK 成功、US 失败”。这是不对的。

更准确的读法是：

发现 9.2.1

两国的惯性底盘几乎一样。 这说明短期行为延续是跨国普遍机制，而不是某个国家的偶然现象。

发现 9.2.2

UK 中 Transformer 的额外增益更容易显形，US 中则更依赖样本条件与分组规模。 这不意味着 US 没有结构，而是意味着它的结构更难在 quick 条件下一览看清。

发现 9.2.3

Phase 2 的真正贡献不是复制出一个和 Phase 1 一样高的 uplift，而是证明：即使 uplift 变小，框架的基本逻辑仍然成立。

也就是说：

- persistence 仍然必须保留；
- delta 仍然是必要的第二口径；
- 只是 Phase 2 把“delta 的稳定性条件”这个问题真正推到了台前。

9.3 从社会科学叙事上，Phase 2 是对 Phase 1 的补强而不是削弱

Phase 1 讲的是：

日常生活高度可预测，但不同群体的可预测性程度不同。

Phase 2 则把这句话扩展成：

日常生活的高惯性具有跨国普适性，但模型想要从惯性之外再多学到一点规律，取决于样本规模、变量口径和制度异质性。

这不是退步，而是更成熟的表述。

10) 可视化完备度评估：现在的 Phase 2 是否已经“像一份完整报告”

10.1 这次补齐后的图链

本报告现在对应的 Phase 2 专属图组共 10 张：

1. phase2_fig1_us_baseline_activity_location.png
2. phase2_fig2_us_stratified_activity.png
3. phase2_fig3_us_pooled_group_delta.png
4. phase2_fig4_uk_vs_us_activity_baseline.png
5. phase2_fig5_us_baseline_macrof1_delta.png
6. phase2_fig6_us_pooled_delta_heatmap.png
7. phase2_fig7_us_methodology_sensitivity.png
8. phase2_fig8_us_sample_size_delta_shift.png
9. phase2_fig9_us_pooled_delta_uncertainty.png
10. phase2_fig10_us_pooled_support_vs_delta.png

10.2 图链增强后的证据结构

相较于早期版本仅以结果摘要为主的组织方式，当前版本已经把 Phase 2 的核心证据整合为一条结构清晰的图链：

- **Fig.1 + Fig.5**：A1 baseline、Macro-F1 与 delta 的误差结构；
- **Fig.2**：quick 与 pooled 的分组现象对照；
- **Fig.3 + Fig.6 + Fig.8 + Fig.9 + Fig.10**：pooled 修正、热力图证据、样本量机制、支持宽度与 support-size map；
- **Fig.7**：粒度、时间尺度与样本边界的稳健性检查；
- **Fig.4**：与 Phase 1 的统一对表。

这使 Phase 2 首次具备了与 Phase 1 相同层级的“章节 - 图像 - 解释”闭环，也让 negative delta、sample-size correction、support-width judgement 与 cross-phase comparison 不再依赖文字孤立支撑，而拥有可直接审阅的视觉证据。

10.3 当前版本的证据完备性

就导师审阅与阶段性交付而言，当前版本已经满足三个关键标准：

- 图链结构完整，能够覆盖 baseline、grouped analysis、sample-size correction 与 cross-phase comparison；
- results/phase2_summary/ 已经把 quick / pooled / methodology / support 汇总成统一 summary 层，形成 报告 -> 图 -> CSV/JSON 闭环；
- 证据类型平衡，既包含支持性结果，也保留了对负 delta 与不稳定分组的谨慎解释；
- 每张图都对应一个明确研究问题，而非仅作为装饰性补图存在。

如果再用这次全项目 closure audit 的口径往回检验，Phase 2 的关键判断也已经能被逐条钉在 summary 层上：ATUS-2024 quick A1 activity 的 Transformer delta 现在锁定在 **+0.02pp** 的近零区间；pooled activity grouped slices 中已有 **10/11** 为正，均值约 **+0.33pp**；10min 相对 30min 仍保留 **+14.25pp** 的清晰优势。也就是说，Phase 2 当前真正需要解决的已经不是“还有没有证据”，而是“如何把现有证据组织得更成熟”。

若进入论文投稿版式阶段，后续只剩表达层精修：

- transition-specific diagnostics / confusion breakdown；
- 更长、更可直接贴入论文 Results 的 caption；
- 若期刊要求，再补 paired prediction trace 层面的严格统计检验。

这些内容属于版式与附录层优化，不构成当前主报告的未完成事项。

11) 方法论贡献、局限与风险控制

11.1 Phase 2 的方法论贡献

贡献 1：把 persistence 正式制度化为必报基线

Phase 2 进一步证明，任何不与 persistence 对照的 US 结果都几乎无法解释。因为如果只看模型 accuracy，很容易误把“高惯性环境中的自然高分”当成“模型真学到了额外结构”。

贡献 2：把 negative delta 重新定义为“待复核现象”

这是本阶段最关键的方法论改进。quick 阶段的负值不是没价值，而是：

- 不能直接当终局结论；
- 必须与 pooled 或 sample-size correction 一起解释；
- 必须区分“真不可学习”与“样本条件不足”。

贡献 3：把跨国比较从“谁更高分”转成“机制是否同向”

Phase 2 最终告诉我们，跨国比较不应只看 UK 是不是比 US 高，而应看：

- persistence 是否同样强；
- grouped direction 是否一致；
- delta 是否在更大样本下开始显形。

这套视角直接影响了 Phase 3 的设计。

11.2 局限

1. US 缺少 enjoyment 与 with_whom，因此无法像 Phase 1 那样做真正完整的多通道叙事。
2. pooled 下不是所有维度都同样稳定，尤其 econstat 仍不适合作为主结论支柱。
3. US 的事件流转槽化过程本身会引入额外异质性，这使得 US 结果更容易受编码与边界定义影响。
4. full A1 并非在所有模型上都能无代价对称跑完，因此 Phase 2 的 pooled 主证据更多体现为 grouped 复核与 sample-size correction。

11.3 风险控制

目前已经采取的控制包括：

- 剔除了 US 中不该存在的伪通道高分；

- 用 pooled grouped 对 quick 结果做二次检验；
- 把 coarse/fine、10/30min、weekday/full 纳入统一稳健性检查；
- 在最终叙事中主动避免把 econstat 小组结果写成主结论。

这意味着当前 Phase 2 的强项并不是“结果全都很漂亮”，而是“知道哪些结果漂亮，哪些结果还不能夸”。

12) 结论：Phase 2 给 Phase 3 留下了什么

Phase 2 的最终结论，可以浓缩成下面四句话：

1. **US 没有推翻 Phase 1 的核心机制**。短期惯性依然强，persistence 依然是主基线。
2. **US 把样本量条件的重要性暴露得非常充分**。negative delta 如果不经 pooled 复核，往往容易被过度解释。
3. **location 是 US 中最稳的“模型有额外增益”证据，activity 则是最有理论价值也最需要样本支持的主通道**。
4. **Phase 3 的默认主线应该保留 activity + persistence + delta，同时优先使用更稳的分组维度而不是盲目扩大切片**。

如果从 Phase 1 到 Phase 3 看一条连续主线，那么 Phase 2 的作用就是：

把“在一个国家里成立”的故事，转化为“在不同数据条件下仍然能成立、但解释方式需要更严谨”的故事。

这一步其实非常关键，因为它让最终论文不只是“高分故事”，而是“高惯性、弱增益、样本条件、跨国迁移”四者之间真正有张力的科学叙事。

13) 附录：图表索引与关键文件

13.1 图表索引

编号	文件	对应章节	作用
P2-F1	results/phase2_figures/phase2_fig1_us_baseline_activity_location.png	§5	US baseline 总览
P2-F2	results/phase2_figures/phase2_fig2_us_stratified_activity.png	§6	quick vs pooled 分组对照
P2-F3	results/phase2_figures/phase2_fig3_us_pooled_group_delta.png	§7	pooled grouped delta
P2-F4	results/phase2_figures/phase2_fig4_uk_vs_us_activity_baseline.png	§9	UK vs US 主线对表
P2-F5	results/phase2_figures/phase2_fig5_us_baseline_macroF1_delta.png	§5	baseline 的 Macro-F1 与 delta
P2-F6	results/phase2_figures/phase2_fig6_us_pooled_delta_heatmap.png	§7	pooled 热力图
P2-F7	results/phase2_figures/phase2_fig7_us_methodology_sensitivity.png	§8	粒度/时间尺度/样本稳健性
P2-F8	results/phase2_figures/phase2_fig8_us_sample_size_delta_shift.png	§7	sample-size correction
P2-F9	results/phase2_figures/phase2_fig9_us_pooled_delta_uncertainty.png	§7	pooled support-width 区间
P2-F10	results/phase2_figures/phase2_fig10_us_pooled_support_vs_delta.png	§7	全部 pooled 切片的 support-size map

13.2 核心数据文件

- `results/phase2_summary/phase2_b1_quick_summary.csv`
- `results/phase2_summary/phase2_methodology_summary.csv`
- `results/phase2_summary/phase2_pooled_grouped_summary.csv`
- `results/phase2_summary/phase2_master_table.csv`
- `results/phase2_summary/phase2_json_manifest.csv`
- `results/phase3_cross_country/uk_us_mtus_master_table.csv`

13.3 相关支撑文档

- `PHASE2_VISUALIZATION_REPORT.md`
- `results/phase2_summary/phase2_summary_overview.md`
- `results/PHASE2_EXECUTION_REPORT.md`
- `results/us_phase2_final_report.md`

文档版本：v3.0

生成日期：2026-04-03

状态：Phase 2 Markdown Enhanced — Advisor Ready

图表：10 张已嵌入主文