

Phase 2 可视化审计与图链说明

项目：Yucheng_Project

阶段：Phase 2 (US / ATUS)

对齐基准：PHASE1_COMPLETE_REPORT.md 的“图表嵌入 + 章节叙事 + 方法解释”风格

更新时间：2026-04-03

1) 总体结论

Phase 2 的可视化已经不再是“有几张零散图”，而是形成了一套可以直接嵌入正式主报告的 10 图链：

- baseline：有
- Macro-F1 / delta 误差分解：有
- quick vs pooled stratification：有
- pooled grouped delta：有
- pooled heatmap：有
- methodology sensitivity：有
- sample-size correction：有
- pooled support-width intervals：有
- pooled support-size map：有
- UK vs US 主线对照：有

换句话说，Phase 2 已具备与 Phase 1 同口径的“图驱动叙事”能力。虽然总图量仍少于 Phase 1，但证据结构已经足以支撑导师审阅、阶段性讨论，以及对 negative delta / sample-size effect 的更成熟解释。

2) 当前 Phase 2 专属图组 (10 张)

目录：`results/phase2_figures/`

图号	文件	在主报告中的位置	作用
P2-F1	<code>phase2_fig1_us_baseline_activity_location.png</code>	A1 baseline	US <code>activity/location</code> 基准总览
P2-F2	<code>phase2_fig2_us_stratified_activity.png</code>	B1 grouped	quick vs pooled 的分组 对照
P2-F3	<code>phase2_fig3_us_pooled_group_delta.png</code>	pooled 复核	各分组 delta（按通道 拆开）
P2-F4	<code>phase2_fig4_uk_vs_us_activity_baseline.png</code>	跨阶段 对表	UK vs US <code>activity</code> 主线对照
P2-F5	<code>phase2_fig5_us_baseline_macroF1_delta.png</code>	A1 baseline	baseline 的 Macro-F1 与 delta 分解
P2-F6	<code>phase2_fig6_us_pooled_delta_heatmap.png</code>	pooled 复核	pooled heatmap，突出 位置与活动通道差异
P2-F7	<code>phase2_fig7_us_methodology_sensitivity.png</code>	稳健性	<code>fine/coarse</code> 、 <code>10/30min</code> 、 <code>weekday/full</code>
P2-F8	<code>phase2_fig8_us_sample_size_delta_shift.png</code>	pooled 复核	sample-size correction 证据图
P2-F9	<code>phase2_fig9_us_pooled_delta_uncertainty.png</code>	pooled 复核	用 grouped <code>n_test</code> 形成的保守 support- width 区间
P2-F10	<code>phase2_fig10_us_pooled_support_vs_delta.png</code>	pooled 复核	全部 pooled 切片的 support-size map

生成脚本：`generate_phase2_phase3_figures.py`

2.5) 这 10 张图背后压缩了什么实验足迹

如果只看可视化报告，很容易把“10 张图”误读成“工作量大概也就 10 组实验”。这并不准确。Phase 2 的图链本质上是最终汇报层，它压缩的是一串更长的 US 结果矩阵。

当前 `results/` 中与 Phase 2 直接相关的结果层至少包括：

- 49 个原始实验 JSON（其中 41 个完成结果 + 8 个 `skipped_small_group` 占位）；
- 2 个 tracker JSON；
- 6 个 `phase2_summary` CSV + 1 个 summary 说明 Markdown；
- 10 张正式 PNG 图；
- 以及 quick / pooled / full-support / methodology / cross-phase comparison 这些不同层级的解释链。

其中，证据层现在可以明确拆成四层：

1. quick headline layer：A1 quick、B1 quick、C1/D1/E1 quick；
2. pooled correction layer：`income/sex/age_bin/is_weekend/survey_period` 的 activity/location grouped 矩阵；
3. full-support layer：A1 full、income/sex/econstat 的大样本辅助验证；
4. summary layer：`results/phase2_summary/`，把 JSON 结果统一汇总成可被绘图脚本直接复用的表层。

因此，这 10 张图不是 10 次运行，而是对 US baseline、grouped quick、pooled grouped、full-support、methodology sensitivity 与 UK-US 对表的**最终压缩表达**。

3) 图组相对于早期版本的增强

旧版 Phase 2 的问题不是“完全没图”，而是图不够成体系：

1. 大部分图仍借用 Phase 1 语境，缺少专门服务 Phase 2 章节的问题意识。
2. 没有把 quick 与 pooled 的关系做成可视化，因此“negative delta 是否只是小样本伪象”只能靠文字硬说。
3. 缺少 uncertainty / support-size 这类会直接提升说服力的证据图。

这次补齐后的优势在于：

1. **A1 讲清楚了**：不仅有 baseline accuracy，还有 Macro-F1 与 delta 解释。
 2. **B1 讲清楚了**：quick 与 pooled 可以直接对照，避免过度解释 quick。
 3. **方法论讲清楚了**：不仅展示 sample-size correction，还展示 pooled support-width 与 support-size pattern。
 4. **对表讲清楚了**：UK vs US 主线可以一图看懂，不用来回翻文件。
-

4) 当前版本的适用性判断

导师审阅与阶段汇报：证据已经完整

目前这 10 张图已经能回答导师最可能追问的几个问题：

1. US baseline 到底是什么水平？
2. 分组差异是 quick 偶然现象，还是 pooled 后仍成立？
3. Transformer 相对 persistence 的 gain 到底稳不稳？
4. US 与 UK 的关系是“同机制但不同幅度”还是完全不同故事？

如果再把图链和 summary 层一起量化，当前 Phase 2 的“证据已经完整”并不是泛泛而谈：pooled activity 的 grouped layer 现在已经是 **11 个切片里 10 个为正**，均值约 **+0.33pp**；真正需要谨慎书写的，只剩像 income_low 这样贴近零附近的小残差，而不再是 quick 阶段那种视觉上非常夸张的大负值。

若进入论文投稿版式阶段：只剩表达层精修

如果继续往前推进，重点也只会落在表达层，而不会改变“当前版本已经完整”的判断：

- transition-specific diagnostics；
- confusion / top-error 类型图；
- figure caption 更长、可直接贴进论文 Results。

这些内容属于投稿版式层的表达优化，不影响当前版本作为正式阶段性汇报材料的完整性。

5) 可直接引用的摘要表述

下述表述可直接用于邮件摘要、封面说明或口头汇报：

Phase 2 已形成完整的 10 图主链：baseline、误差分解、quick-vs-pooled 分组对照、pooled delta、heatmap、方法稳健性、样本量修正、pooled support-width、support-size map，以及 UK-US 主线对表。

这些图共同支持一个稳定结论：US 中 persistence 依旧强，Transformer 的额外增益在 quick 阶段未必立即显形，但在 pooled 与更充分样本下会呈现出更稳定的正向证据。