

Phase 3 可视化审计与图链说明

项目：Yucheng_Project

阶段：Phase 3 (MTUS cross-national)

对齐基准：PHASE1_COMPLETE_REPORT.md 的“主图 + 稳健性图 + 对表图”结构

更新时间：2026-04-03

1) 总体结论

Phase 3 的可视化已从“主表完整但图证不足”扩展为一套 **12 张** 的成体系图组，而且这套图组是基于 **补齐 KR 两个 A1 fine 补跑后重新生成的完整 summary**，不是旧版 38 条 A1 记录的残缺版本。

当前这套图链已经覆盖：

- A1 各国主结果
- seed 稳定性
- quick -> full 修正
- B1 grouped scatter
- country x dimension heatmap
- specific subgroup heatmaps
- B1 quick -> full 修正
- sample-size sensitivity
- country-level CI forest
- grouped full-run distribution boxplots
- fine/coarse sensitivity
- Phase1/2/3 统一对照

也就是说，Phase 3 不仅主表完整，图链也已经能够独立支撑完整叙事。

2) 当前 Phase 3 专属图组 (12 张)

目录：`results/phase3_figures/`

图号	文件	在主报告中的位置	作用
P3-F1	phase3_fig1_a1_delta_by_country.png	A1 主线	各国 A1 delta (SGD vs Transformer)
P3-F2	phase3_fig2_fine_vs_coarse_transformer.png	敏感性	fine vs coarse 的国家级对照
P3-F3	phase3_fig3_b1_group_delta_scatter.png	B1 grouped	age/sex 分组点云分布
P3-F4	phase3_fig4_cross_phase_transformer_vs_persistence.png	统一对照表	Phase1/2/3 主线对照
P3-F5	phase3_fig5_a1_seed_stability.png	A1 主线	Transformer seed 稳定性
P3-F6	phase3_fig6_a1_quick_vs_full_transformer.png	quick/full	A1 quick -> full uplift
P3-F7	phase3_fig7_b1_country_groupby_heatmap.png	B1 grouped	country x group_by 热力图
P3-F8	phase3_fig8_b1_specific_group_heatmaps.png	B1 grouped	specific subgroup 热力图
P3-F9	phase3_fig9_b1_quick_vs_full_delta.png	quick/full	B1 quick -> full uplift
P3-F10	phase3_fig10_sample_size_vs_delta.png	稳健性	sample size 与 delta 的关系
P3-F11	phase3_fig11_a1_country_ci_forest.png	A1 主线	country mean + seed + CI 的 forest-style 视图
P3-F12	phase3_fig12_b1_distribution_boxplots.png	B1 grouped	age_bin/sex 的 grouped full-run 分布箱线图

生成脚本： `generate_phase2_phase3_figures.py`

2.5) 这 12 张图背后压缩了什么实验足迹

Phase 3 的图链看上去比 Phase 1 更“克制”，但这不是因为工作量更小，而是因为最终汇报层把大量 country / seed / group 结果压缩进了更少但更聚焦的图里。

就当前 results/ 目录而言，Phase 3 的核心足迹至少包括：

- 105 个 phase3_mtus_a1 JSON；
- 522 个 phase3_mtus_b1 JSON；
- 11 个 phase3_cross_country CSV；
- 12 张 phase3_figures PNG。

也就是说，12 张最终图对应的不是 12 个实验，而是 627 个 MTUS 结果 JSON 加上一层 cross-country summary / master table 汇总。

这也是为什么 Phase 3 的可视化策略必须更“精选”：

1. 如果把国家、seed、group 的中间结果逐张展开，图量会迅速失控；
 2. 导师真正需要的是能够判断机制是否成立的最终证据链，而不是阅读数百张近似重复的小图；
 3. 因此最终保留下来的 12 张图，承担的是“把 627 个结果 JSON 压缩成可审阅叙事”的任务。
-

3) 图组相对于早期版本的增强

旧版 Phase 3 的最大问题，并不是“没有结果”，而是：

1. 图不够，导致跨国外部效度叙事看起来更像一堆表格摘要。
2. 没有 seed stability / quick-full / sample-size / CI 这类能够直接回应质疑的证据图。
3. 没有把 B1 grouped 的 country-level 结构与 full-run 分布一起可视化出来，因此“跨组稳定性”更多停留在文字层面。

这次补齐后的提升很明确：

1. **A1 不只是一张柱状图**：现在有 country delta、seed stability、quick-full 修正、country-level CI forest 四层图。
 2. **B1 不只是一句‘都为正’**：现在有 scatter、country x dimension heatmap、specific subgroup heatmap、quick-full 修正图，以及 grouped boxplots。
 3. **方法论解释不再只靠文字**：sample-size、fine/coarse 与 grouped full-run 分布现在都有明确图像支撑。
 4. **跨阶段闭环可视化完成**：Phase1 / Phase2 / Phase3 可以一图连起来讲。
-

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

现在这 12 张图已经足以回答导师最可能追问的核心问题：

1. Transformer 在 7 国是不是都保持正增益？
2. 这些增益是不是 seed 偶然？
3. quick 到 full 到底修正了什么？
4. grouped analysis 是不是也成立？
5. fine 为什么应该保留为主线？
6. 整个项目从 Phase 1 到 Phase 3 到底是不是一条连续主线？

如果进一步把图链与 full summary 合并来看，这套“完整性”现在还能被更精确地量化：A1

Transformer 的 country mean 已达到 **7/7 国家为正**，而 `age_bin/sex` 的 grouped full-run 切片已达到 **35/35 个 country×group cells 为正**。因此，这 12 张图现在承担的已经不是“试着讲故事”，而是把正式结论压缩成最可审阅的证据链。

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：

- 更长的 caption 版本
- 更贴近论文风格的 multi-panel summary figure
- 若投稿版式要求，可把 grouped boxplot 拆成 appendix 级 country panels

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) 与 Phase 1 的功能对照

如果拿 Phase 1 的 17+ 张图作标准，Phase 3 现在仍然更精简；但两者的任务也不同：

- **Phase 1**：要承担框架建立、理论解释、应用段与错误分析，因此图更多；
- **Phase 3**：主要承担外部效度验证，图的功能更集中于 cross-country robustness。

因此更准确的判断是：

- Phase 3 已达到能够独立支撑导师审阅的成熟度；
 - 论文终稿阶段主要需要进一步精修 caption 与排版，而不是重新搭建图链。
-

6) 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 3 已形成完整的 12 图主链，并且全部基于最新的完整 summary 重生成，KR 的两个 A1 fine 补跑也已正式进入汇总。

这些图共同支持一个清晰结论：在 7 国 full runs 下，Transformer 相对 persistence 的正增益不仅在总体样本中成立，在 `age_bin` 和 `sex` 分组中也保持稳定，而且 fine 比 coarse 更能承载项目主叙事。