预测编码网络在语言建模中的样本效率、机制不可达性与测量纪律：一项单卡架构研究

PRISM 项目 V4-V5 技术报告

2026-09-21 **v6.0** | 覆盖 2026-09-07 至 09-21 的完整研究记录（v4 新增：V7 终端形态；v5 新增：100M 规模验证；v5.1 勘误：超参随规模漂移，「+40.1%」伪影撤回；v5.2 终判：调优 TF 预训练领先 33.7%；v5.3：V8-A 稳定化解锁证伪；v5.4/v5.5：90M 冷启动/流式判决，确立混合架构路径；v5.6：混合最小原型判决成立；v5.7：3-seed 复现全判据命中；v5.8：帕累托前沿 + 主干锚定机制；v5.9：96M 纯 CPU 端侧链路 4/4 判据；**v6.0：混合架构主线定稿——330-354M 外推 3/3 全胜 + 收窄归因解耦（§4.7）、适应主张诚实重述与 PEFT 3-seed 对照（§7.7）、主干锚定消融转正与 step50 机制分解（§7.6/§8.3）、第 7 起测量偏差（§3.6）**。v5.1-v5.9 为未发布的内部里程碑，v6.0 为 v4 之后的首个发布版本——完整变更见附录 E 版本历史）

关键词：非 Transformer 架构、预测编码、混合架构、样本效率、测量纪律

代码与数据已开源：https://github.com/Asaluther/PRISM-V4

单卡 RTX 4080 16GB，PyTorch 2.14，≈209 有效训练 run（本地索引 results/run_index.csv 198 条，另含 75 项作废归档；口径详见附录 B）；100M 规模验证 run 于启智社区 V100S-32GB 完成（24L/90-101M，共 3 次）；330M 档四训练 + 补强七训练全部本地完成（Q5 旧「4080 不可行」判断经实测推翻，附录 D.3）

摘要

我们报告预测编码网络（Predictive Coding Network, PCN）与 Transformer 在语言建模上的一项系统性单卡研究，以及由两者构成的**混合架构**。本版（v6.0）以混合架构为第一叙事，核心发现六项：

1. **规模-数据比规律**：误差流变体（no_gating）相对 Transformer 的优势随 tokens/parameter 比下降呈**倒 U 形**——峰值 ≈1000（难域 WikiText-103 上 +60.8%），数据充裕后被反超（-64.9%），并有一次成功的带外预测（实测 +37.7% ∈ 预测带 [+30%, +55%]）。101M 对称调优终判：纯 PCN 预训练输给调优 TF 33.7%——失败机制经 bf16 对照锁定为**优化动力学内禀失稳**（快速学习区间与自毁区间重合，与数值精度无关），构成纯 PCN 的规模天花板。
2. **混合架构（12TF+12PCN，96M）**：预训练 PPL **63.18 ± 0.22**，以更少参数反超纯 Transformer **36%**（配对 bootstrap 95% CI [-37.7, -33.2]，p≈0）。**该优势经 330-354M 外推保持**：逐 seed 3/3 全胜、调优口径 +11.2%（37.05 vs 41.72）、seed 方差比 34×。**收窄归因解耦实验（v6.0）**：固定数据比下稀缺反而放大优势（96M@470 t/p 达 58.6%，且混合架构用半数据即胜 TF 全数据——1.76× token 效率），**收窄由规模压缩主导**（-47pp）——优势最大区间恰为终端个人模型的「小规模+数据稀缺」象限。机制上，**Transformer 主干锚定了高 lr 误差流区间**（16+8 消融三判读全中，假设转正）：主干阻断误差正反馈的自毁路径，将「快速=自毁」耦合部分解耦。
3. **适应优势的诚实边界（v6.0 重大修正，第 7 起测量偏差触发）**：误差流的真优势是**免配置自限适应**——全参数微调 27 格（3 架构 × 3 seed × 3 lr）符号零翻转，且该结构跨规模保持（330M 复测：冷启动混合 3/3 全正 vs TF 混合，流式 held-out 双正 vs TF 负，多抽签协议确认混合架构均值全正紧凑、TF 方差爆炸）；但补齐 PEFT 3-seed 对照后，TF+bitfit 单发 +74.7 ± 2.4 可恢复甚至反超——**架构优**

势的准确表述是零配置鲁棒性，不是性能上限。PEFT 对照同时钉死 LoRA-TF 单发结构性不可行（-29.8 ± 13.5）并发现 bitfit 流式的架构不对称（方法不可跨架构平移推荐）。

4. **组合性质与对称不可达性**：小样本精确复制能力（5/5 vs 0/30）经 13 变体双向对照 + 2×2×2 全因子，不可分解为任何单子件（限于已测移植策略）；复制优势不可外推到键值绑定（MQAR 否定性）。
5. **局部学习规则的适用边界**：Whittington-Bogacz 式局部规则在 MLP+迭代推断下与 BP 完全等价，在 PCN 结构上存在 2.4-3.7× 结构性差距（K∈{1,2,3}、反馈复用等已测变体内）。
6. **终端链路（端侧验证）**：96M 混合模型纯 CPU 推理 2240 tok/s、生成 43.8 tok/s、**20 秒冷启动达 +85% 峰值。**步数扫描的协议发现：适应峰值在 step 50 后单调回落——机制分解（v6.0）证实为**记忆与通用遗忘并行**（train PPL 背至 5-9，预训练域 PPL 恶化 6-47×），峰值高度是用户属性、峰值步数是 lr 属性——「冻结骨干 + 50 步预算」由此获得机制层根据。

核心结论：PCN 单独是「低数据/参数比区间的优势架构、稳定性天花板以下的最强适应器」；TF 单独赢数据充裕区预训练。**两者的混合架构同时取得两侧能力**：预训练侧在 21-354M 全程为正（96M +36% / 330M +11.2%，且 seed 稳定性高一个量级以上），适应侧零配置自限——终端形态「通用底盘 + 可在线生长的误差流头」的最小原型到端侧链路均已验证。全部七起系统性测量偏差（v6.0 新增第 7 起：适应方法选择偏差）均由预设协议捕获，方法学条款随文给出。

Summary (English)

We report a systematic single-GPU study of Predictive Coding Networks (PCN) versus Transformers for language modeling, culminating in a **hybrid architecture** that takes both sides. (1) The error-stream advantage follows an inverted-U in tokens/parameter (peak ≈1000, +60.8% on WikiText-103, with a successful out-of-band prediction); at 101M, tuned TF beats tuned PCN by 33.7% — the failure is an **intrinsic optimization-dynamics instability** (fast-learning and self-destruction intervals coincide), forming a scale ceiling for pure PCN. (2) The **hybrid (12 TF + 12 PCN, 96M)** reaches PPL 63.18 ± 0.22, beating pure TF by 36% with fewer parameters (paired bootstrap, p≈0); the advantage **survives at 330-354M** (3/3 paired wins, +11.2% tuned-vs-tuned, 34× lower seed variance). A decoupling experiment shows scarcity *amplifies* the advantage at fixed scale (58.6%; the hybrid needs only half of TF's tokens to win — 1.76× token efficiency), while **scale compression dominates the narrowing** (−47pp) — the maximal-advantage quadrant is exactly "small-scale + data-scarce", i.e., personal terminal models. Mechanistically, **the TF trunk anchors the high-lr error-stream regime** (16+8 ablation, three pre-registered readouts all hit). (3) The adaptation advantage is **honestly re-scoped**: full fine-tuning never catastrophically overfits (27/27 sign-stable cells, preserved at 330M in multi-draw protocols), but TF+bitfit (+74.7 ± 2.4, 3-seed) recovers or exceeds single-shot performance — the correct claim is **zero-config robustness, not a performance ceiling**; LoRA-TF single-shot is pinned negative and bitfit streaming is architecture-asymmetric. (4) The copy-task capability is a non-decomposable combinatorial property (13 variants + full factorial; not extrapolable to key-value binding — MQAR negative). (5) Whittington–Bogacz local rules: equivalent on MLP+iterative inference, 2.4–3.7× worse on PCN (within tested variants). (6) Terminal link: 96M hybrid at 2240 tok/s CPU inference, **+85% cold-start adaptation in 20 seconds**; the step-50 peak-and-decline is mechanistically decomposed (memorization parallel to general-domain forgetting, 6–47×). Seven systematic measurement biases (v6.0 adds the seventh: adaptation-method selection) were all caught by pre-registered protocols; methodology clauses

are given throughout. All code and data are open-source.

---

1. 引言

1.1 动机

Transformer 之外的有效计算范式仍待发现：现有替代（Mamba-2、RWKV、RetNet 等）要么与 attention 数学等价，要么在精确回忆上存在信息论限制（V4_PLAN 文献调研，基于 arXiv 2312.00752, 2405.21060, 2402.01032, 2504.13173 等）。本项目的长期目标是**终端私有智能**——个人设备上低功耗、可在线学习、数据不出端的架构——该场景天然处于「数据少、参数相对多」的区间，因此我们关心的核心量是**样本效率的架构依赖性**，而非绝对性能。

1.2 贡献概述

- 一条 7 数据点、经外推验证的规模-数据比规律（§4）
- 一个 13 变体双向对照的机制分解，导出「组合性质」结论；及其在权重约束上的再现（第二涌现性质，§5/§7.1）
- 局部学习规则从等价到失效的完整边界刻画（§6）
- 四次测量危机的完整记录与提炼出的方法学（§3；第四次为 v2.2 的合成基准捷径排查）
- 终端适应的真实边界：跨域冷启动成立、域内个性化否定、遗忘权衡与参数经济性重定性（§7，v2.2）
- 终端效率三问的补完：能耗、CUDA Graph 终端口径、int4 混合精度配方与 W_res 机制发现（§7.4，v2.2）
- 稀疏激活与冷启动兼容性验证 + CPU 端侧部署 + 安全框架（§8，v4）
- **混合架构主线（v6.0）**：12TF+12PCN 最小原型 96M 反超纯 TF 36%（3-seed 配对统计）→ 330-354M 外推 3/3 全胜 + 收窄归因解耦（§4.7）→ 主干锚定机制消融转正 + h2e4 终端默认（§7.6）→ 适应主张诚实重述与 PEFT 3-seed 对照（§7.7）→ step50 机制分解与 96M 纯 CPU 端侧链路（§8.3）

---

2. 架构与实现修正

2.1 PCN 架构

每层包含：自底向上提取 u = W_bu(LN(x))；自顶向下预测 p = W_td(h_above)；预测误差 e = clamp(u − p, ±10)；**误差驱动的跨位置门控聚合**——压缩特征交互 [e_i⊙e_j, e_i+e_j] 经 sigmoid 独立门（非 softmax 竞争）、因果掩码、Top-K 稀疏化后加权 W_lat(e)；状态更新 h = GELU(W_up(LN(e + lat))) + W_res(x)（W_res 恒等初始化）。

消融变体（贯穿全文）：

- **no_gating**：门控聚合换成因果 attention（误差流保留）——本文最强的「误差流」载体
- **two_pass**：真 top-down（pass 2 用更深一层状态做预测来源）
- 反馈消融、门控 softmax 化、门来源/值对象开关（§5）

2.2 三轮实现修正（首次报告，附教训）

轮次	缺陷	症状	修复
R1	W_res 随机初始化（谱范数 ≈0.6）→ 12 层梯度衰减 316×；门控双重小尺度压缩 → g≡0.5 死锁；evaluate 未透传消融参数	PPL 卡 406 五千步不动	恒等初始化 + 门控 LayerNorm/Xavier + 参数透传
R2	**非因果泄露**（attention 未传 attn_mask、门控无掩码）——两个架构都能「看未来」	22M 模型 held-out WikiText top-1 准确率 99.6%	双侧因果化（§3.1）
R3	**基线超参次优**（lr=3e-4 对 Transformer 严重次优，其最优为 1e-3）	假性 +56.8% 优势	对称 lr 扫描（§3.2）

全部双向时代数据（75 项）作废归档；本文所有数字均为因果修正后。

---

3. 测量方法学（四次危机的提炼）

3.1 因果性是公理，不是配置项

任何「预测下一 token」的跨位置交互必须显式因果化。**作弊探测器**：训练后模型在 held-out 真实语料上 top-1 准确率 >50% 或 PPL 异常低（<5），几乎必然非因果——本研究的泄露正是由「PPL 2.24 好到不可能」暴露，而非代码审查发现。

3.2 对照超参必须对称调优

统计稳定性（多 seed、小方差）不能补偿系统性偏差：我们的假性优势有 5-seed、std 0.012 的完美统计形态，仍然源于基线 lr 次优。任何架构对比必须双方各自完成同规模超参扫描后取各自最优。**架构最优超参本身是结果**：本研究的 Transformer 最优 lr（1e-3）与误差流变体（5e-4）相差 2-3 倍，且互换代价巨大（-83% 至发散）。

复发与勘误（v2.2 补注）：该陷阱在 V6 收官系列复发——demo 与骨架系列全用了 lr 次优的 TF checkpoint（WT PPL 557 vs 公平版 215），一度得出「TF 微调退化 -19.7%」的结论；公平复测实为 +5.4% 改善。教训：**基线勘误不一定推翻结论**（ΔW 约束幅度在公平对下依然成立，1.91×），但必须先勘误再引用。

3.3 自建对照结构必须先用已知结论校准对照格

为分解实验在外部重写的「简化结构」= 未声明的多变量改动。本研究三次自建结构全部对照失效（含已知□格复现失败）。正确做法：在原实现上加开关，并先跑对照格校准。

3.4 同预算协议

一切对比固定 batch × seq × steps 的 token 总量与等效 batch；单变量三臂消融（full / 去 X / 换 X）。

3.5 合成基准的「高于随机」必须先过捷径天花板（v2.2 新增）

自建合成任务上某臂准确率「高于随机」时，必须先在**固定测试集上计算平凡启发式的得分天花板**（众数值/多数类/最近重复/unigram 等）——臂的得分 ≤ 天花板则信号被捷径解释。实例：MQAR 值有放回采样下「众数值」启发式天花板 0.1263 > 门控臂实测 0.115，且后者得分随 n 的衰减模式与天花板一致（§7.5）；改值唯一采样后臂精确回到随机。采样方式决定捷径的存在与否，**优先无放回/去捷径构造。**

3.6 适应方法选择偏差（v6.0 新增，第 7 起）

以全参数微调为唯一对照会系统性高估架构间适应差异。本研究的混合架构草稿期适应主张即基于 full-FT 对照；预注册判据（若 PEFT-TF 单发 ≥ +30% 则主张需重述）在补齐 LoRA/bitfit 基线后被触发——TF+bitfit 单发 +74.7 ± 2.4（3-seed）。**检测协议对本研究自身的新结论同样有效**（§7.7 重述全文）。配套教训：LoRA 适配器初始化必须播种——未播种的 init 抽样曾使 TF-LoRA 单发在 -26.5% 与 -0.4% 之间摆动（9 个固定抽样后钉死为 -29.8 ± 13.5）；冷启动协议的用户抽签亦未播种（对 12+12 系影响 ±4 内，薄头与 330M 档可达 33pp，见 §7.7 多抽签协议）。

---

4. 主结果：规模-数据比规律

4.1 协议

TinyStories / WikiText-103（GPT-2 tokenizer，117.9M tokens 磁盘缓存），seq 256，41M tokens 固定预算（5000 步 × 等效 batch 32），双方各自最优 lr 与 clip，多 seed。模型 12 层，d ∈ {256, 384, 512}（21-63M 参数）。

口径澄清（响应 V3 审稿 Q2）：所有规模点（d256/384/512）保持完全相同的 token 总量、步数与等效 batch；tokens/param = 训练 token 总量 ÷ n_params（含 embedding/LayerNorm 在内的全部参数，逐 run 记录于 results/<exp>/results.json 的 n_params 与 budget 字段）；每个 run 的最优超参映射同样逐 run 记录于该 JSON 的 config 字段（lr/warmup/clip/wd）。WT d512 预测验证点：results/s512_*/（各 seed 目录含完整 config 与 val 曲线）。

4.2 规律曲线（v2：二维不对称，11 数据点，bootstrap 95% CI）

减数据轴（固定 d256，数据量递减）——优势单调上升，无左端回落：

tokens	TF [95% CI]	NG [95% CI]	NG 优势	稳健性
2.5M	402.67	63.95	**+84.1%**	全胜
5M	268.63	42.93	+84.0%	全胜
10M	128.82	28.27	+78.1%	全胜
20M	32.88	19.92	+39.4%	全胜
41M	15.18 [14.66,15.86]	16.50 [16.45,16.55]	−8.8%	负

加参数轴（固定 41M tokens）——倒 U：

tokens/param	TS	WT	稳健性
~7500	—	−64.9%	负
~1900	−8.8%	+20.7%（TF [210,238] / NG [175.6,176.4]）	全胜
~1025	+43.3%	+59.2%（n=5；TF [290,447] / NG [128,168]）	全胜
~650	+35.5%	+48.5%（n=5；全胜）	全胜

逐点 CI 与 run 清单：analysis_v2/scale_curve_with_ci.json。**「全胜」= NG 最差 seed 优于 TF 最好 seed**——所有正优势点均达此标准（强于 CI 不重叠）。

结论（范围限定表述，响应审稿 W1）：在本研究参数范围（21-63M）、token 预算（≤164M）与两个域内：**误差流优势随数据稀缺单调增大（至 +84%）；随参数加大先增后减（峰值 ≈1025 tokens/param）**。「减数据比加参数更能放大优势」直接指导终端个性化部署（冻结中等模型 + 少量用

户数据）。>1B 参数与 >164M tokens 未测。

附带发现（稳定性）：NG 的 seed 标准差在两个低数据点上比 TF 小 1-2 个数量级（WT d256: 0.53 vs 17.90；WT d512: 2.39 vs 65.66）——**误差流不仅更准，训练还显著更稳定**，这对端侧部署（不可挑选 seed）是独立于 PPL 的实际优势。

域难度主轴：同比 1025 tokens/param，难域（WT）优势高于简单域（TS）约 17 个点；v1 的「倒 U」表述在数据量轴上不成立（见上表：单调），仅在参数轴成立——v2 已按二维不对称规律改写。

外推验证：WT d512 数据点在曲线其余部分确定后预测（预测带 [+30%, +55%]），实测 +37.7% 落在带内——规律的第一次成功预测。

充分训练翻转：WT d256 @164M tokens（20K 步）下 Transformer 53.1 vs no_gating 87.5——数据充裕后 TF 渐近优势接管。误差流的优势严格属于低数据/参数比区间。

4.3R 超参敏感性对照（v2，响应审稿 W2/W4）

warmup	TF (lr 1e-3)	NG (lr 5e-4)
100	14.98	14.85
500（基线）	15.14	14.13
1500	**13.55**	**14.01**

发现：warmup 对 TF 敏感（500→1500 改善 10.5%）、对 NG 不敏感（±1%）。两边各自最优时收敛区（TS d256）差距为 TF 领先 3.4%（此前单一 warmup 下的 8.7% 含超参不对称成分）。**关键定量论证**：warmup 敏感性的量级（~10%）远小于低数据区优势（+59%~+84%），规律的主结论不受此混杂威胁。

4.4 第三架构对照（09-10 补充；v2.2 升级 n=5/域，响应 V3 审稿 W2/W3）

以**可学习因果衰减注意力**（softmax(QK□+λ(t−i))，RetNet/GLA 线性注意力家族的最小代理；Mamba/RWKV 在 Windows 不可编译，纯 torch 复刻的数值风险大于代理价值）作为第三族（lr 1e-3，另有 3e-4 两点对照；n=5 seed/域，decay_{ts,wt}_lr1e3_s0-4）：

域	decay 最优	Transformer	no_gating
TinyStories	15.83±0.67	15.18±0.80	16.50±0.07
WikiText-103（低数据区）	**219.4±24.6**	**221.97±17.90**	**175.96±0.53**

decay 均值略优于 TF（WT 上 +1.2%，CI 重叠）但**远不具备误差流的低数据区优势**（NG 对 decay 领先 19.8%）——误差流的 +20.7% **不是「替换 softmax」或「添加衰减先验」类单子件改动可以获得的**，与 §5 的组合性质结论形成跨实验呼应。

4.5 定位含义

终端场景（用户数据有限 × 相对大参数）恰好落在峰值区——该规律为「什么区间用什么架构」提供了可操作的决策依据。诚实边界：43-61% 是改进幅度而非数量级跃迁；规律在 <650 与 >1B 参数两个方向均未测。**v2.2 修正**：该区间的增益可兑现场景是**跨域/冷启动适应**（§7.3）——域内已收敛模型的个性化微调无适应空间（§7.2），部署建议按 §9.1 的双准则执行。

4.6 规模验证终判：优势未延续，稳定性天花板（v5-v5.2，启智 V100）

24L/d=512 双架构、除架构与 lr 外全对齐：warmup 2000、eff batch 32、seq 256、10K steps = 82M tokens（~900 tokens/param，落在小规模倒 U 峰区同位段）、同 tokenizer 与数据切分、

n_heads=4。完整网格与轨迹见 results/SCALE_090M_VALIDATION.md。

架构	lr	Best PPL	状态
TF 101.52M	3e-4	158.65	稳定
TF 101.52M	**5e-4**	**99.66**	**最优（两侧夹逼）**
TF 101.52M	1e-3	236.53 / 250.85	方差 6.0%
TF 101.52M	2e-3	452.46	过热
PCN 90.48M	**1e-4**	**150.13 / 150.37**	**最优（四点闭合，方差 0.16%）**
PCN 90.48M	5e-5	207.17	稳定
PCN 90.48M	2e-4	185.44	~6000 步后软发散（零 NaN）
PCN 90.48M	3e-4	290.97*	2000 步后劣化 + fp16 溢出

- **终判：调优 TF 领先调优 PCN 33.7%**——小规模低数据区误差流优势（§4.2，21-63M）未随规模延续，在 63M→100M 间反转。v5 版「PCN -40.1%」系双方沿用 63M 旧扫描最优（TF 1e-3 劣于本规模调优值 2.4 倍）的伪影，v5.1 已撤回——本系列第 6 起系统性偏差（LESSONS #2 的规模升级复发），由「先扫描后主张」协议捕获，并升格为条款：**跨规模对比前双方超参必须各自在新规模重扫**
- **稳定性天花板（而非容量天花板）**：PCN@3e-4 早期轨迹全场最佳（step 1000/2000：469/291；同期 TF 最好 1244/766，PCN@1e-4 为 688/383），step ~2500 起劣化，step 5292 起进入 fp16 激活溢出死循环——eval 在 fp32 下仍正常（step 6000 输出 352.06），权重非 NaN，NaN 恢复机制对「激活溢出、权重干净」模式存在盲区。可用 lr 上限随深度收紧（小规模最优 5e-4 → 24L 稳定最优被压至 1e-4），压缩了 PCN 的调优空间
- **V8-A 稳定化解锁——假说证伪（v5.3）**：bf16 重跑 PCN@3e-4（动态范围≈fp32，前向不可能激活溢出；恢复盲区已补：连续 50 次 NaN-loss → 回滚 + 紧急 lr×0.5，该分支经本地压力测试验证）——早期轨迹与 fp16 逐位一致（287.75 峰@2000），劣化**完整复现**（319→478→493@5000），且平安跨过 fp16 死亡点 5292（零 NaN），尾部 loss 平台化。同时 PCN@2e-4 fp16 以零 NaN 完成软发散（峰值 185.44@中段，尾部 494）。**结论：失稳是优化动力学内禀性质（误差正反馈：e↑→更新↑→e↑），与数值精度无关；「快速学习区间与自毁区间是同一个区间」**。剩余解锁方向仅剩动架构（qk-norm/误差阻尼），属新实验周期决策
- 复现观察：TF@1e-3 方差 6.0% vs PCN@1e-4 的 0.16%（「误差流稳定性优势」限于 PCN 稳定区间内）；TF 最优 lr 随规模下移（1e-3 → 5e-4）
- 诚实边界：限于当前训练配方（fp16 autocast + GradScaler，无 qk-norm）；PCN 2e-4 中间点未跑（翻盘需 34% 改善，不影响判定）；各点单 seed

4.7 330-354M 外推与收窄归因（v6.0 新增，判决 W1）

d1024/24L、20K 步（164M tokens ≈ 470 tokens/param，数据稀缺峰区左侧）、fp16、全部本地 4080 完成（附录 D.3 记录了 Q5 旧「4080 不可行」判断被实测推翻的过程）。完整数据与 v1 配置失误的透明修复记录见 results/SCALE_330M_VALIDATION.md（v2）。

模型	lr（骨干/头）	seed	best PPL	备注
HYB 12+12（330.4M）	5e-4 / 1e-4	0/1/2	**39.50 / 39.60 / 40.14**	σ=0.34，零 NaN
HYB 12+12	**1e-3** / 1e-4	0	**37.05**	调优最优（骨干 lr 上漂）

模型	lr（骨干/头）	seed	best PPL	备注
HYB 12+12	2.5e-4 / 1e-4	0	45.26	夹逼左臂
TF 24L（354.0M）	5e-4	0/1/2	41.72 / 42.37 / **62.24**	σ=11.67，s2 慢收敛，零 NaN
TF 24L	2.5e-4 / 1e-3	0	63.52 / **发散**	右侧上探即死（PPL 736@14K）
PCN 24L（306.8M）	1e-4	0	峰值 147.77@~6K → 终点 222	中途软发散（零 NaN）

四条判决（预注册）：

1. **优势保持**：逐 seed 配对 3/3 全胜（+5.3/+6.5/+35.5%）；调优对调优 **+11.2%**（37.05 vs 41.72，双方夹逼完成，HYB 参数少 6.7%）
2. **稳定性不对称放大**：seed σ 比 12×（96M）→ **34×**（330M）；TF 坏 seed 在预训练与适应侧全线脆弱（适应侧在线 -100.5%、held-out -497%，§7.7）
3. **lr 容忍度分化**：HYB 骨干最优点上漂（1e-3 最优）而 TF 上探即死——主干锚定机制的预训练侧再现（§7.6）
4. **PCN 规模天花板实测**：306.8M 下 3.4× 参数 + 2× tokens 峰值仅改善 1.5% 且中途软发散——§4.6 的预测转为实测事实

收窄归因解耦（96M @ 470 t/p，3+3 seed）：HYB 83.97 ± 0.35 vs TF 202.62 ± 7.92——差距 **58.6%**（3/3 全胜）。分解：**数据比效应（固定 96M，853→470）放大优势 +22pp**（36.4%→58.6%，与纯 PCN 的倒 U 左坡方向相反——混合结构在深稀缺区持续拉开差距）；**规模效应（固定 470，96M→330M）压缩优势 -47pp，是收窄的主导因素。**附带 token 效率强数字：**HYB@46.7M tokens（84.0）优于 TF@82M tokens（99.3）——1.76× token 效率**；330M 档无此跨预算优势。对 500M+ 的预测边界如实标注：若压缩持续存在归零风险，500M 点为判决性实验（需云算力）；而**终端/个人模型区间（≤100M + 数据稀缺）恰为优势最大区间**——与本研究的目标场景重合。

---

5. 机制分析：组合性质与对称不可达性

5.1 现象

小样本精确复制任务（H 个随机 token + 分隔符 + 复制，1000 步预算，12L-64d）： 因果 PCN **5/5 seed 学会**（copy loss 0.17-0.32），同规模因果 Transformer **0/30**（恒 unigram 3.43；因果版三长度 ×5 seed 0/15 + lr 调优复检 0/15）。

5.2 十三变体双向对照（3-seed each）

已排除的必要条件假说（加法线失败 + 反向线存活 = 双向排除）：

假说	标准块+子件	PCN−子件
门来源（feat 交互 vs q·k 相似度）	—	qk 门在 PCN 内 0.047-0.092 □
值对象（误差 e vs 输入 h）	—	wlat_h 0.022 □（比 e 更好）
误差残差性（e=u−p）	—	e=u（no_feedback）0.277 □
FFN 干扰	去 FFN 3.437 □	+FFN 0.074 □
更新通路形态	S2/S3 3.435 □	原生 □
门对比度（头平均）	单头门 3.435 □	原生 □

假说	标准块+子件	PCN−子件
自信息直注增量	S5 3.435 □	—
独立 sigmoid（vs softmax 竞争）	—	softmax 化 1.109（3/5 不稳定）□□

softmax 竞争化是唯一使 PCN 退化的单改动（不稳定而非完全失败）——独立门是该列表中**唯一存留的单子件必要条件**，但仅有它不够（tf_sigmoid 有独立门仍失败）。

实现澄清（响应 V3 审稿 Q3）：移植 = 在标准块结构中**直接添加对应模块并随机初始化**，随后与对照完全相同的从零训练协议（复制任务无预训练阶段，不涉及预训练权重移植或微调）；拆除 = 原实现上以开关参数旁路该子件（教训 9：不在外部重写结构）。渐进插值、初始化统计匹配、蒸馏式迁移等更强的移植策略**未测试**，列入中期实验清单（附录 D）。

5.2R 全因子补格（v2，响应审稿 W3：交互效应检验）

2×2×2 全因子矩阵（门来源 × 值对象 × FFN，各 3 seed，factorial_fill.json）：

	e 值	h 值
feat 门 −FFN	0.21 □	0.022 □
feat 门 +FFN	0.31 □	0.27 □
qk 门 −FFN	0.09 □	0.047 □
qk 门 +FFN	0.14 □	0.074 □

八格全部成功——PCN 骨架内不存在能破坏复制能力的子件组合；主效应与交互效应均为「无害」。结合标准块侧五连败（§5.2），组合性质结论从「检验过的变体集」升级为**完整因子空间内确认。**

5.3 结论：组合性质

复制能力的必要条件是 **PCN 骨架多个子件的组合**——提取-误差-融合-恒等残差的耦合结构，不可分解为单一组件。

5.4 对称不可达性

- 性能层（V1-V3，20+ 实验）：从 Transformer 修补出发 → 退回 Transformer
- 机制层（本研究）：从标准块移植子件（五种）→ 救不回复制能力；从 PCN 拆除已测子件 → 能力保持

□ **在我们检验的变体集、训练预算（1000 步）与移植策略下，该能力未能通过子件移植复现，也未能通过拆除子件消除。** 我们无法断言此结论在更大规模、更多干预方式（如渐进式移植、初始化匹配、辅助损失训练等）下仍然成立——但这组双向证据支持「换基线而非打补丁」的架构研究策略。

5.5 行为签名（补充）

跨模型 CKA：纯 PCN（two_pass）0.112 / 混合版 0.469——表示结构差异显著。误差范数与 token 惊讶度的相关四次测量 ≈ 0（0.10/0.05/0.015/−0.06）：「误差编码惊讶」的经典预测编码假说在本设置下**未获支持。**

---

6. 负结果：局部学习规则的适用边界

级别	设置	结果
0	两层 MLP + K 步迭代推断	**完全等价**（0.0002 vs 0.0002；K=1 输出层梯度 cos = 0.9995）

级别	设置	结果
1	PCN 的 W_td（玩具任务）	零敏感度（冻结/局部/全 BP 的 loss 全同——任务不依赖反馈）
2	PCN 主路径（LM）	**2.4-3.7× 差距**（m1 = 39.8-49.6 vs 基线 13.6-16.6）

级别 2 的两个混杂因素已对称排除：局部 lr 四点扫描（全平台 2.4-2.6×）；K ∈ {2,3} 迭代推断矩阵（数值稳定但性能更差，且 K-pass 纯 BP 对照 13.60 证明推断本身无代价）。

结论（限定范围表述）：在本研究的 W&B 实现变体、更新方案（K ∈ {1,2,3}）与优化设置下，等价性在「MLP+推断收敛」与「PCN 结构」之间存在 2.4-3.7× 的性能差距。我们**不能**由此推广到所有生物合理的局部学习规则——其他变体（如松弛权重对称性、不同阻尼、更大 K 值、反馈对齐等）可能在 PCN 上表现不同。「以局部规则实现终端在线学习」的路径需要探索非 W&B 形式或等待新理论。

实现澄清（响应 V3 审稿 Q4）：本实现的权重对称性内建于架构——局部规则复用前向的 W_td 本体（p = W_td(h_above)），无独立反馈矩阵；K-pass 的逐层更新顺序、lr_loc 与阻尼等全部超参逐 run 记录于 local_rule_level* 系列 JSON 的 config 字段。反馈对齐（随机反馈矩阵）、K>3、阻尼调度等变体矩阵**未测试**；V6 已按预案关闭该线（V6_CLOSURE.md），变体筛选列入中期清单（附录 D）。能耗口径：BP 训练/推理的 J/token 已有测量（§7.4 T4），逐局部规则变体的 wall-clock/能耗对比随该线关闭不再补充。

附带发现：grad_clip 0.5 对 no_gating 更优（16.56→14.13）；所谓 K-pass 增益（13.84）实为 clip 混杂（差 2% 在噪声内）。V6 按失败出口关闭，替代路径（样本效率微调的终端适应）见 §7；终局文档 results/V6_CLOSURE.md。

---

7. V6 收官：权重约束、遗忘与终端适应（v2.2 新增）

> 本章全部实验使用**公平 checkpoint 对**（PCN ts/wt_causal_no_gating_s{1-4}，最优 lr 3e-4；TF ts_tf_lr1e3_s{1-4} / wt_tf_lr1e3_s{1-4}，最优 lr 1e-3——与 §4 fairness 同源）。
>
> 微调协议（除非另注）：冻结 embedding + 前 6 层，AdamW，clip 1.0，6 个 256-token 训练块。

7.1 权重约束 = 第二涌现性质（幅度稳健，含义重定性）

现象：微调时 PCN 的可训练层权重变化幅度始终约为 TF 的 1/2（E1：后半层 ΔW 0.85-1.17 vs TF 1.65-2.00）。

双向排除（协议同 §5：PCN 侧去子件应增大 ΔW、TF 侧加子件应减小 ΔW）：误差减法（E3，泛化 72.9% vs 74.7% 仍持平）、GELU 门控（0.86×/1.06×）、norm_update LN（0.97×）、W_res 恒等残差（0.93×/0.94×）——四个单子件假说全部排除。

公平基线复核：原始测量用的 TF checkpoint 为 lr 次优（WT PPL 557 vs 公平版 215）；换公平对复测 **TF/PCN ΔW = 1.91×（公平）vs 1.88×（次优）——幅度结论对基线质量稳健。**

含义重定性（由 7.2 的遗忘数据强制）：ΔW 小不等于功能稳定。权重约束应表述为**「参数经济性/高函数敏感度」**——单位参数变化带来更大的函数移动（适应快、遗忘也快）——而非「抗遗忘保护」。

7.2 灾难性遗忘基准（GOAL V6#2 首次执行）

双臂 A→B→A 交替任务流（域内主题切换：TS 奇幻池□校园池；跨域远征：TS□WT），双口径，4 seed。

域内校准网格（lr×steps，24 格 × 双架构）无一为正——即使最温和设置（lr 5e-5 × 60 步）也让 held-out 恶化（PCN -7.4% / TF -4.0%），随强度单调恶化。**结论：22M/40M-token 收敛模型在**

1.5K token 同域数据上没有可挖的适应空间；任何「个性化微调」优势必须来自域差闭合（见 7.3）。

匹配适应水平的跨域比较（calibrated 口径，各自最优超参）：适应增益匹配（held_WT PCN +89.0% vs TF +85.1%）时，PCN 的家域遗忘**更多**（TS 通用 PPL +31.9% vs TF +16.7%，0/4 seed 占优）——与「权重约束→低遗忘」的预测相反，构成 7.1 重定性的直接证据。

激进域内重复微调（aggressive 口径）：PCN 抗损能力显著更强（forget +43.3% vs TF +90.8%，3/4 seed；通用漂移 5.1× vs 9.0×）——高可塑性架构的另一面。

规模边界：以上均为 22M、300 步内、1.5K-token 任务的快速范式（4 seed），不外推到长程在线学习。

7.3 跨域冷启动适应：单发与流式

单发（1.5K token 用户数据，300 步冻结微调，WT base → TS 用户，5 用户）： PCN 平均 PPL 改善 **+58.5%** vs 公平 TF **+5.4%**（PCN 5/5 全胜，且起点 PPL 更差 2392 vs 2257；适应耗时 4.6s）。 （内部报告曾记录 TF 为 -19.7% 恶化——该数字基于次优 TF checkpoint，公平复测后修正为小幅改善；教训 8 的复发实例，见 §3.2。）

流式（prequential：8 批 × 256 token 逐批到达，每批先评在线分再适应 60 步，一次性通过）：

	PCN	TF
在线增益（vs 冻结基线）	**+24.0%**	**-80.9%**
流末用户 held-out	+27.8%	-183.8%
累积 ΔW	0.599	1.120（1.87×）

TF 对每个 256-token 小批灾难性过拟合（末批在线 PPL 12470）；域内对照流两架构双输（-189%/-209%）。**部署准则：跨域/冷启动场景用误差流架构边用边学；域内已收敛模型应冻结。**

90M 规模外推（v5.4 新增，判决性；v5.5 补流式）：在 §4.6 的 90-101M 对称调优检查点上重跑本协议（本地复训，跨 GPU 复现 <1.4%）：**单发：PCN +62.5%（3 用户全正，+41~+83）vs TF -127.4%（3 用户全负，微调后 PPL 暴涨）**——优势不仅保持且放大（小规模 +58.5% vs +5.4%）。**流式（v5.5）：PCN 在线增益 +34.3% vs TF -139.5%**（小规模 +24.0% vs -80.9%），用户 held-out +31.5% vs -149.9%，累积 ΔW 0.78 vs 1.66（2.14×）。两协议各配对称 lr 扫描（单发 5e-4/1e-4/5e-5；流式 1e-4/5e-5/2.5e-4）排除超参伪影：**PCN 六档全正，TF 六档全负。**机制：误差驱动更新在少样本域适应中天然自限，标准 AdamW 微调已收敛大模型则灾难性过拟合。附带观察：TF 冻结基线在线 PPL（1379）远优于 PCN（2843）——TF 是更好的语言模型，但每次流式适应都在摧毁它而 PCN 越用越好；PCN 适应后对源域漂移更大（与 §7.2 遗忘结论一致）。**该性质经受住规模与双协议检验，是终端在线学习的核心依据；与 §4.6 预训练终判不矛盾——预训练属数据充裕区（TF 区），个性化属数据稀缺区（PCN 区），支持混合架构路径。**数据：results/v7/coldstart_090m{,_ftlr}.json、results/v7/streaming_090m.json

7.4 终端效率三问补完（T4-T6）

能耗（T4，首次测量）：训练/服务端口径 PCN 门控 bmm 每 token 能耗最低（训练 1.148 vs TF 1.253 mJ；b32 推理 0.618 vs 0.709）；裸 eager 的 b1 终端口径曾为 TF 的 2 倍。

CUDA Graph（T5）：整前向捕获重放（静态形状）后，b1 终端口径吞吐 36.7K→197K tok/s（×5.4），达 TF-graph 的 **102.2%**（no_gating 91.5%），能耗反转为 **0.51 vs 0.71 mJ/tok（-28%）**；数值位级等价（|Δlogit|=0 全配置）。**「低功耗终端」论点在 graph 部署口径下成立。**（边界：自回归逐 token 生成需按长度预捕图；无 graph 的 eager 口径 PCN 仍慢于 TF。）

int4 量化（T6）：TF 全配置近无损（g64 +0.1~0.6%）；PCN 的 RTN 敏感性**全部来自 W_res**——仅量化 W_res（占参数 0.4%）单独造成 +4.2~4.5% 损失，将其除外后 int4-g64 近无损（+0.2~0.4%）。机制：PCN 残差主干穿过 W_res matmul（**非加性残差**），量化误差直接进入主干并跨层复合；TF 的加性残差天然隔离误差。这与 V4 的 W_res 恒等初始化（梯度直通）是**同一设计选择的两面**。端侧配方：int4-g64 非对称 + W_res/embedding 保 fp16（32.0 MB @22M；22M 档 embedding 占 61% 参数，int4 收益在 0.5B+ 档才充分兑现）。GPTQ 无需立项。

7.5 MQAR 否定性结果（防过度主张红线）

标准 Multi-Query Associative Recall（键值绑定）四阶段测试（4 臂 × n∈{16,32,48}，预算至 8000 步 × b64 = 8× 基准，D64-128）：**TF / 衰减注意力代理 / no_gating / PCN 门控全部随机水平。**阶段一至三中门控臂 n=16 的 0.115（7× 随机）经混淆排查确认为**采样捷径伪影**——值有放回时「众数值」启发式在同一测试集天花板 0.1263 > 0.115；值唯一（无捷径）复测中门控臂精确回到随机。

结论：§5 的复制任务优势不可外推到键值绑定；本文的召回主张锚定复制任务。键值绑定的解出需要远超玩具尺度的容量（与 Zoology/Based 的 d256+ 设置一致），留待规模扩展期。此混淆已提炼为方法学教训（先算捷径天花板，再解读「高于随机」）。

7.6 混合架构最小原型：一个模型同时拿到两种能力（v5.6；v6.0 定稿）

§7.3 双判决与 §4.6 预训练终判的直接推论：**TF 赢预训练（数据充裕区）、误差流赢适应（数据稀缺区）→ 混合。**最小原型 HYB-96M（12 TransformerBlock 骨干 + 12 PCNBlock 误差流头，d=512，接口依据：PCN 层对输入无 embedding 假设），训练用组件级分组 lr（骨干 5e-4 / PCN 头 1e-4，各为本规模扫描最优），判据预注册。完整记录见 results/HYBRID_MVP.md。

	TF 24L	PCN 24L	HYB 12+12	判据
预训练 PPL（3-seed）	99.34 ± 2.62	149.95 ± 0.48	**63.18 ± 0.22**	✅ ≤130 超额（**比纯 TF 好 ~36%**）
单发冷启动（3-seed）	-57.5 ± 10.5%	+74.1 ± 4.0%	**+60.3 ± 4.3%**	✅ ≥+30%
流式在线增益 @1e-4	全 seed 负	+31.6~+34.3%	-6.7~-74.3%	✅（阈值移动）
流式在线增益 @5e-5	全 seed 负	+44.3~+45.1%	**+26.1~+43.0%**	✅ 保守步长命中

- **主结果（意外但机制连贯）**：混合架构以更少参数（96M vs 101.5M）在预训练上大幅超过纯 TF——82M tokens ≈ 853 tokens/param 正处数据稀缺峰区，纯 TF 无法利用误差流加速、纯 PCN 被稳定性压住，12+12 各取所长。与 2D 定律自洽而非矛盾
- **流式阈值移动**：hyb 的 PCN 头流式稳定阈值比纯 PCN 低 2-5×（2.5e-4 档：+41.9 vs -47.9）；5e-5 档恢复纯 PCN 最佳水平。3-seed 复现（v5.7）全判据命中（results/SEED_REPLICATION.md）

敏感性扫描（v5.8）与 v6.0 定稿：分割比与头 lr 两轴扫描（results/PHASE2_SENSITIVITY.md）：分割比轴预训练严格单调（16+8: 54.4 → 12+12: 63.2 → 8+16: 74.5）但 **PCN 头 <12 层时单发适应塌陷**；头 lr 轴呈**预训练-适应帕累托前沿**——头 2e-4 预训练劣化 20% 但适应全场最强且**流式阈值移动消失。**

- **主干锚定机制（v6.0 消融转正）**：16+8@头2e-4 消融三判读全中（PPL 59.86 零 NaN 不发散 vs 纯 PCN@2e-4 软发散；单发 +18.9%→+52.8%；流式@1e-4 +3.1%→+59.1%）——**「头 2e-4 解锁适应」跨分割比成立，高头 lr 稳定性与 TF 主干存在性绑定。**当初设想的 qk-norm 修复路线不再必需
- **h2e4 终端默认转正（v6.0 3-seed）**：12+12@头2e-4 流式@5e-5 **+72.0 ± 2.0** vs 基线 +32.9 ± 9.0（分离 >4σ）、@1e-4 阈值移动三 seed 全消（+60.1 ± 0.4）、单发 +64.3 ± 2.3 与基线同带——正式接管终端在线学习默认配置（results/v7/h2e4_replication.json）

- **16+8@2e-4 定性修正（v6.0 3-seed 反转）**：预训练 59.53 ± 0.86 与流式 +76.2 ± 1.8 复现，但**单发塌陷**（~27.6 ± 7.1；同检查点两次测量差 33pp，暴露薄头对用户抽签的敏感性）——定性为**流式专精配置**，不作终端默认；「头 ≥12 层是单发必要条件」在 2e-4 下成立
- 附带：**gating 分级对照（v6.0）**——96M 孪生（共享抽签）gating 预训练略优（61.75 vs 63.42）但冷启动腰斩（+34.2 vs +63.1）、@1e-4 崩（-68.2 vs -6.7）；330M 全灭（-47.8/-104.5）——**gating 以 2~3 PPL 预训练换适应可塑性，代价随规模放大**，no_gating 主线双支柱确认

7.7 适应主张的诚实重述与 PEFT 对照（v6.0 新增）

零配置全参数微调符号矩阵（3-seed × 3 lr，27 格零翻转）：PCN/HYB 全格为正（PCN 单发 +74.1 ± 4.0、HYB +60.3 ± 4.3；流式@5e-5 PCN +44.7 ± 0.4、HYB +32.9 ± 9.0），TF 全格为负（单发 -57.5 ± 10.5）；配对统计 HYB−TF 单发 +117.8pp [+101.4, +129.1]。PCN 流式@5e-5 统计显著优于 HYB（+11.8pp，p=0.022，如实报告）。

330M 复测（符号结构跨规模保持）：冷启动 HYB 3/3 全正（3 用户口径 +34~+51）；流式@5e-5 HYB +55~61% 且 held-out 同步为正（+62~68）vs TF +19~26% 且 held-out 零至负（坏 seed 在线 -100.5%、held-out -497%）；ΔW 参数经济性保持（~140 vs ~230）。**多抽签终稿（10 用户 × 6 主体，固定抽签种子）**：HYB 均值 +29.3~+40.2（正用户 27/30）vs TF -45.5~+4.2（σ 至 156）——单点 3 抽签在 330M 档不可靠（用户抽签方差较 96M 放大一个量级），多抽签为最终口径（adapt330_multidraw.json）。

PEFT 对照（3-seed × LoRA 3 个固定 init 抽样，results/v7/peft_3seed.json**）**：

方法 @最优档	单发（TF / HYB）	流式@5e-5（TF / HYB）	流式@1e-4（TF / HYB）
full-FT 5e-4	-57.5±10.5 / **+60.3±4.3**	全负 / +32.9±9	全负 / 负
LoRA r8 5e-4	**-29.8±13.5** / +57.1±6.0	+73.0±1.5 / **+77.7±2.7**	+50.8±7.3 / **+73.8±3.2**
bitfit 1e-3	**+74.7±2.4** / +52.8±3.3	+24.1±0.5 / +6.1±0.2	+37.4±1.0 / +11.1±0.3

四条结论：① **LoRA-TF 单发结构性不可行**（9 抽样钉死为负，rank 16 亦不救；s0 版 -0.4% 系幸运 init 抽样）；② **bitfit-TF 单发 +74.7 ± 2.4 跨 seed 复现**——TF 的问题不在不可适应，在方法×配置选择；③ **bitfit 流式架构不对称**（TF +24~37% vs HYB +6~11%：PCN 头适应靠权重位移，偏置微调推不动）——PEFT 方法不可跨架构平移推荐；④ HYB+LoRA 是流式最稳组合（@1e-4 仅降 4pp）。

重述后的准确主张（预注册判据①触发，第 7 起测量偏差，§3.6）：

1. 误差流架构提供的是**免方法选择、免步长调参的自限适应**——运维鲁棒性性质，终端场景（不能为每用户调参选方法）的直接价值
2. 它**不是性能上限**——TF+正确配置的 PEFT 可恢复并部分反超
3. 部署地图：零配置默认混合架构 full-FT/保守步长；可承担一次性方法选择时 TF+bitfit（单发）或 HYB+LoRA（流式）同级可用

---

8. V7 终端形态：稀疏兼容性 + 端侧部署 + 安全框架（v4 新增）

8.1 稀疏激活与冷启动兼容性（第五次测量偏差修正）

初始结论（后修正）：25% Top-K 稀疏激活「摧毁冷启动能力」（+64% → −257%）。

混杂来源：稠密 PCN 使用 WikiText checkpoint（跨域冷启动 = 强项），稀疏 PCN 使用 TinyStories checkpoint（域内微调 = 已知无优势）——域效应被误判为稀疏效应。

四臂对照修正：

臂	预训练域	微调域	冷启动改善
稠密 × WT	WikiText	TS	+53.2%
稀疏 25% × WT	WikiText	TS	**+57.9%**
稠密 × TS	TinyStories	TS	−317.2%
稀疏 25% × TS	TinyStories	TS	−268.6%

修正后结论：稀疏激活与冷启动**完全兼容**。终端部署推荐更新为「**稀疏 PCN 全能**」。

8.2 CPU 端侧部署

指标	结果	终端判据
batch1 推理吞吐	8,603 tok/s	≥50 □
自回归生成	112.9 tok/s（8.9 ms/token）	实时 □
冷启动适应	+55.0% PPL / 22.7 秒	≤30 秒 □
模型大小	80 MB (21M 参数)	终端可装 □
数据不出端	全流程本地	□

8.3 混合架构 96M 端侧链路（v5.9；v6.0 补机制分解）

12TF+12PCN/96M（366MB）纯 CPU 验证（8 线程 fp32，无网络调用，协议对齐本节 b2/b3）： **推理 2240 tok/s @b1_s256、生成 43.8 tok/s、冷启动 20 秒达峰值改善 +85%**（判据 4/4 命中，results/PHASE3_EDGE.md）。对照 V7 小模型（8.6K/113/22.7s→+55%）：参数 4.6× 换速度 26%，仍远超终端底线。**步数扫描的协议层发现：少样本适应在 step 50（20s）达峰后单调回落**（300 步仅剩 +60%）——既有单发协议只测终点，低估峰值 ~25pp；论文适应数字应报峰值+曲线， 终端默认预算固化为 50 步。300 步终点与 GPU 协议（+60.3±4.3%）精确一致，CPU 适应质量 不打折。

step50 达峰/回落的机制分解（v6.0，results/STEP50_MECHANISM.md**）**：评估点加密到每 10 步 × 3 用户 × lr{5e-4, 2.5e-4, 1e-4} × 四条曲线。① train PPL 在全部 (lr, 用户) 组合下单调降至 5~9（6 条序列被完全记忆，排除动力学失稳）；② test 峰值后回升、**预训练域 PPL 同步恶化 6~47×**——PCN 头参数被重新征用做记忆，held-out 泛化与通用知识同时丧失（记忆假设 + 遗忘假设同时成立）；③ **峰值高度是用户属性**（同一用户三 lr 峰值差 <4pp）、**峰值步数是 lr 属性**（随 lr 降低右移——降低 lr 只拉伸时间轴、不提高峰值）；④ 遗忘从 step10 即与适应并行发生。「冻结 TF 骨干 + 50 步预算」由此获得机制层根据：不是「50 步够用」，而是「20~80 步窗口内到达峰值，其后净收益为负」。

8.4 安全框架

层	机制	防护目标
输入过滤	长度/重复/字符集/注入检测	投毒适应
速率限制	频率 + 总步数双控	过度适应退化
审计回滚	checkpoint 快照 + 一键回滚	适应不可逆
所有者绑定	PBKDF2 身份 + 权重加密	模型被盗

8.5 稀疏度扫描

act_sparse	PPL	变化
0%	16.50	—
25%	**14.13**	**−14.4%（最优）**
50%	14.14	饱和

---

9. 讨论与局限

9.1 对终端智能愿景的含义

倒 U 规律的峰值区与终端场景重合，且 v2.2 的适应实验把「重合」细化为**可操作准则**：跨域/冷启动（新用户、换域、预训练-部署域差）用误差流架构做边用边学（流式 +24% vs TF -81%）；域内已收敛的模型应冻结（域内流式双输）。在线学习方面，W&B 局部规则已证伪（§6），替代路径是「BP 微调的样本效率 + 权重约束的参数经济性」。效率侧，graph 部署口径下终端吞吐/能耗短板已清偿（§7.4）——但 43-61% 的幅度、+24% 的流式增益不足以单独支撑终端愿景，稀疏化与系统工程的叠加仍是必需。

9.2 局限（诚实清单）

1. ~~规模上限 63M 参数 / 单卡~~ **v6.0 更新**：已测至 354M（96M 3-seed 铁证 + 330M 3-seed + 双侧夹逼 + 解耦，§4.7）；HYB@1e-3 臂、gating 对照、306M PCN 为单 seed；500M-1B 未测（B4 显示规模压缩 -47pp 主导收窄，500M 为判决性实验）
2. 第三架构以衰减注意力为代理（RetNet/GLA 家族精神），完整 Mamba/RWKV 未测；

代理与 TF 重合暗示衰减核家族可能同样不具备该优势，但需正式实现确认

3. 复制任务是合成任务；「1000 步内学不会」≠「永远学不会」——结论限于小样本预算；

且复制优势**不可外推到键值绑定**（MQAR 四臂全随机，§7.5）

4. 倒 U 的机理（为何低数据区占优）只有现象学刻画，无理论模型；B4 显示混合架构在深稀缺区与纯 PCN 的倒 U 左坡方向相反——机理刻画需扩展到混合结构
5. ~~效率限制~~：v2.1 的 bmm 重构之上，v2.2 补齐能耗、CUDA Graph 终端口径（b1 102%/能耗 -28%）、int4 混合精度配方——graph 部署口径下全部解除；残余边界：自回归生成的按长度预捕图、eager 口径仍慢、22M 档 int4 收益受 embedding 占比压制
6. ~~适应结论限于 22M 快速范式~~ **v6.0 更新**：适应符号结构已经 3-seed（96M）+ 330M 复测（多抽签终稿，§7.7）；但流式协议仍为短流（8 批）、单发为 6 序列——真实长时程与真实用户数据未测
7. 遗忘的权衡：匹配适应下 PCN 遗忘更多（高可塑性端）——终端多用户/多域共存的场景需要恢复机制或参数隔离，未在本研究范围内
8. 冷启动协议用户抽签未播种（继承口径）——对 12+12 系影响 ±4 内，薄头与 330M 档可达 33pp（§3.6/§7.7）；跨配置比较单发数字需多抽签支撑
9. 330M 全部 run 于 20K 步未饱和——数字为预算内相对比较，非各自收敛值

9.2R 规模扩展计划（响应审稿 W1；v6.0 更新）

阶段	目标	状态
~~短期~~	~~一个 200-300M 参数点在 WikiText-103 上的验证~~	**已完成（v6.0，本地 4080）**：330-354M × 3-seed + 双侧夹逼 + 适应复测 + 收窄归因解耦（§4.7/§7.7）——Q5 旧「4080 不可行需租用」判断经冒烟实测推翻，全程零云积分
中期（6 个月）	500M 判决性实验（优势存续 vs 规模压缩归零）+ 完整 tokens/param 曲线	B4 后优先级升高；需云算力
长期	>1B 参数验证	需集群级算力

操作建议的条件限定：终端个性化部署建议在 tokens/param ∈ [~470, ~1100] 且参数 ≤3.5e8 的已测区间内直接适用（v6.0 上界自 1e8 上修）；超出此范围需先验证。

9.3 方法学立场

本研究的七次测量危机（非因果泄露、基线超参次优——含 V6 系列复发并勘误、自建对照失效、合成基准频率捷径、超参跨规模漂移、适应方法选择偏差——v6.0 新增）全部被**预设的验证程序**（而非运气）捕获——其中第四次（§7.5）是「高于随机 ≠ 学会任务」的直接示范，第七次（§3.6/§7.7）发生在本研究自己的新主张上。我们主张：架构比较论文应报告其验证程序（因果探测器、超参对称性、对照校准、捷径天花板、init/抽签播种）与结果本身同等重要的地位——尤其在单点「惊人优势」出现时，**先怀疑测量，再相信架构。**

附录 A：关键数字索引

开源仓库：https://github.com/Asaluther/PRISM-V4 — 全部代码、run 脚本、结果 JSON 可复现

V7 数据：results/v7/（a2_scan / a4_fixed / b2_cpu / b3_coldstart / c_security）

数据	文件
规律曲线 7 点	results/analysis_v2/scale_data_ratio_curve.json
因果 5-seed（两域）	results/analysis_v2/{ts,wt}_causal_seed_validation.json
公平判定 + d384/d512	results/analysis_v2/fairness_final.json, d384_confirm.json
机制 13 变体	results/mechanism/{copy_mechanism_s*.json, gate_source_v2.json, block_ablation.json}
局部规则三级别	results/mechanism/{local_rule_test 系, local_rule_level*_*.json}
行为签名	results/analysis_v2/signatures_causal_*.json
骨架约束（第二涌现性质）	results/mechanism/{skeleton_constraint, skeleton_s3s4, skeleton_fair_recheck}.json
遗忘基准（双口径）	results/mechanism/forgetting_benchmark_{calibrated, aggressive}.json + forgetting_calibration.json
跨域适应（单发/流式）	results/demo/{personalization_demo, streaming_demo}.json
终端效率 T4-T6	results/efficiency/{t4_energy, t5_b1_graph, t6_int4}.json
MQAR 四阶段（否定性）	results/mechanism/mqar_{benchmark,phase2,phase3, phase4}.json

数据	文件
双向时代作废归档	results/_invalid_bidirectional/（README 含报警现场）
混合 MVP + 3-seed 复现	results/HYBRID_MVP.md, results/SEED_REPLICATION.md, results/v7/seed_replication.json
帕累托扫描 + 锚定消融 + 16+8 三 seed	results/PHASE2_SENSITIVITY.md, results/v7/{phase2_protocols,abl16_8_h2e4,abl16_8_3seed}.json
h2e4 终端默认 3-seed	results/v7/h2e4_replication.json
PEFT 3-seed + init 方差	results/v7/peft_3seed.json（单 seed 版 peft_baselines.json 保留对照）
step50 机制分解	results/STEP50_MECHANISM.md, results/v7/step50_mechanism.json
330M 外推 + 解耦 + gating 阶梯 + 多抽签	results/SCALE_330M_VALIDATION.md（v2），results/v7/{adapt330,gate96_pair,adapt330_multidraw}.json, results/wt_33*m_*/ 各 results.json
全量 run 索引	results/run_index.csv（198 条：配置/seed/PPL/NaN/耗时/参数量）

附录 B：实验统计

仓库本地 run 索引 results/run_index.csv 共 **198 条**（配置/seed/lr/PPL/NaN 恢复/耗时/参数量逐字段；另 TF@1e-3@354M 发散臂提前终止无 results.json，见 SCALE_330M_VALIDATION v2）+ 启智 11 条 + 75 项双向时代作废归档（results/_invalid_bidirectional/）。GPU 总时长 ≈55 小时（其中 96M 档单 run 25-28 分钟、330M 档约 2 小时）；v6.0 的全部新实验（330M 队列 ×11、PEFT 3-seed、step50、B4 解耦 ×6、gating 孪生、多抽签）在本地 RTX 4080 完成，零云积分。全部实验可由 run_*.sh 与本报告所引脚本幂等复现。Power 分析：预训练 PPL（σ≈2.6）检测 10% 差异需 n=2/组；单发改善（σ≈4.3）检测 30%+ 级效应 n=3 覆盖充分、10% 级需 n=8/组。配对统计明细：analysis_stats.json（bootstrap 20000 次，seed 固定）。

PRISM V4-V6 · 2026-09-07 至 09-21 · 单卡架构研究完整记录

附录 C：术语表（响应审稿 W5）

缩写	全称/定义
PCN	预测编码网络（Predictive Coding Network），本研究的非 Transformer 骨架
NG / no_gating	PCN 的误差流变体：门控聚合替换为因果 attention，误差流保留——最强配置
TF	Transformer 对照基线（Pre-LN + SDPA 因果注意力）
误差流	e = u − p（自底向上提取减自顶向下预测），逐层计算的残差信号
门控通路	PCN 的跨位置聚合机制：压缩特征交互 → sigmoid 独立门 → Top-K → 加权 W_lat 值
对称不可达性	从标准块移植子件无法获得、从 PCN 拆除子件不损失某能力的双向实验性质
tokens/param	训练 token 总量 ÷ 参数量（数据-参数比）

缩写	全称/定义
全胜	NG 的最差 seed 优于 TF 的最好 seed（强于 95% CI 不重叠的稳健性声明）
W&B 局部规则	Whittington–Bogacz 式预测编码局部学习规则（无需全局反向传播）
K-pass	迭代推断：K 遍前向中逐层松弛误差（W&B 等价性的推断条件）
权重约束	PCN 微调时 ΔW ≈ TF/2 的幅度约束（第二涌现性质；含义=参数经济性，§7.1）
公平 checkpoint	各架构在其最优 lr 下训练的对照模型（TF 1e-3 / NG 3e-4，§4 协议）
prequential	先预测后适应的在线评估协议：对未见过批次的预测质量即在线得分（§7.3）
MQAR	Multi-Query Associative Recall：键值绑定标准合成基准（§7.5，否定性）

附录 D：审稿意见回应索引

D.1 V2 审稿（v2.1 时点）

审稿弱点	响应	所在节
W1 规模/外推	措辞限定 + 二维规律改写 + 峰值点 n=5	§4.2
W2 基线/超参对称	warmup 敏感对照 + decay n=3 + 定量论证	§4.3R/§4.4
W3 因子设计	2×2×2 全因子 8 格补齐（全部成功）	§5.2R
W4 统计细节	bootstrap CI + 「全胜」稳健性声明	§4.2
W5 可复现性	术语表（本附录 C）；开源仓库计划：GitHub + smoke test + 相对路径（下版）	附录 C/D
效率局限#5	bmm 重构 6.12×（达 TF 96%）+ 推理 batch32 反超 +14%（v2.1）；v2.2 补能耗/CUDA Graph b1 102%/能耗 -28%/int4 配方——三问解除（§7.4）	§7.4
（v2.2 新增）V6 终端主张	单发/流式跨域适应 + 遗忘边界 + MQAR 否定性（防过度主张）	§7
（v2.2 新增）测量纪律第 4 条	合成基准捷径天花板（MQAR 混淆排查）	§3.5

D.2 V3 审稿（v2.2 时点，针对 v2.1 文本；总体 7.5/10，Soundness 6/10）

V3 弱点/建议	v2.2 响应	状态
W1 规模外推不足；建议 200-300M 或外推不确定度量化	维持资源决策（规模扩展搁置，9.2R 计划在）；§4.5/§9.1 的部署建议已收紧为双准则并显式限定已测区间	部分解决（措辞）+挂账（实验）

V3 弱点/建议	v2.2 响应	状态
W2 基线不全/未对称调优；decay 单 seed	decay 升级为 **n=5/域**（lr1e-3，另有两点 lr 证据 3e-4 vs 1e-3）；Mamba/RWKV 完整实现仍以衰减代理代替（编译限制已说明）；optimizer/wd/init 对称网格未做——挂账	部分解决（decay 补种）+挂账
W3 seed 数不统一、「全胜」判定、run 元数据	「全胜」保留为**补充性**稳健声明（正文标注其定义）；decay 补至 n=5；全局 seed 统一（≥7）与分布检验/smoke test/run_metadata.csv 列入仓库路线图	部分解决+挂账
W4 不可达性外推过强	§5.4 的限定表述维持并加强（§5.2 实现澄清：直接替换+随机初始化+从零训练，渐进/蒸馏未测）；渐进插值与蒸馏列入中期实验清单	措辞解决+挂账（实验）
W5 局部规则变体空间未覆盖	§6 范围限定维持；V6 已按预案关闭该线（决策记录）；K>3/反馈对齐/阻尼变体筛选与逐变体能耗列入中期清单；BP 口径能耗已有（T4）	挂账（明确）
附加：run 数 130 vs ~140 矛盾	附录 B 统一口径（205 文件 = 75 作废 + ≈130 有效 + 机制/效率；另有 V6 ≈40）	**已解决**
长期项：能耗 / batch1 kernel 优化	T4 能耗首次测量 + T5 CUDA Graph（b1 102%、能耗 -28%）+ T6 int4 配方——审稿「长期优先级」两项已在 v2.2 完成	**已解决**
Q1 超参映射	逐 run 记录于 results/<exp>/results.json 的 config 字段（§4.1 澄清）	**已答**
Q2 协议等效性/tokens-param 口径 /d512 run 路径	§4.1 口径澄清；results/s512_*/	**已答**
Q3 移植实现细节	§5.2 实现澄清（直接替换+随机初始化+从零训练；渐进/初始化匹配/蒸馏未测）	**已答**
Q4 W&B 实现细节	§6 实现澄清（对称性内建于 W_td 复用；超参见 local_rule_level* JSON）	**已答**
Q5 最小可行 200-300M 方案	8.2R 短期行 + 最小配置建议：d1024/24L（≈350M）单点 × 3 seed × 20K 步 ≈ 单卡 A100/H100 约 60 GPU·h 或 4080 约不可行（需租用）——资源到位后启动	**已答**（计划）

V3 建议中已列入中期清单的实验（资源/优先级裁决后执行）：渐进插值移植、蒸馏式迁移、复制任务延长预算（5K-20K 步）、局部规则变体筛选矩阵（K≤10、反馈对齐、阻尼调度，d128 玩具规模预筛）、optimizer/wd/init 对称网格、全局 seed≥7 统一与分布检验改造。

D.3 V4 审稿（v6.0 时点，针对已发布 v4 PDF；总体 7.5/10，Clarity 6/10）

V4 弱点/建议	v6.0 响应	状态
W1 规模外推不足；建议 200-300M 中规模点或外推 CI	**已完成**：330-354M × 3-seed + 双侧夹逼 + 适应复测 + **收窄归因解耦**（§4.7/§7.7）——本地 4080 完成，Q5 旧「4080 不可行」判断经实测推翻；条件化措辞全文执行，已测区间上修至 ≤354M	**已解决**
W2 基线不全：RWKV/Mamba 代理 + PEFT 缺失	PEFT 基线**已补齐并超规格**：LoRA r{4,8,16} + bitfit × 双协议 × 3-seed × 3 init 抽样（§7.7）——并触发适应主张重述（第 7 起偏差）；RWKV/Mamba 仍以衰减代理代替（编译限制未变，挂账）	PEFT **已解决**；SSM 家族挂账
W3 不可达性外推过强；建议渐进插值/蒸馏/统计匹配	§5.4 限定表述维持；渐进插值与蒸馏仍在中期清单（未执行）	措辞解决+挂账（实验）
W4 局部规则变体空间未覆盖	§6 范围限定维持；K>3/反馈对齐/阻尼变体筛选仍在清单	挂账（明确）
W5 统计/修辞/章节乱序	配对 bootstrap+t 已补（五组对比全 CI）；run_index.csv 全量索引（198 条）；**章节 8/9 乱序已修复**；术语表更新；修辞按已测范围收敛（「全胜」保留为补充性声明）；新增版本历史（附录 E）	**已解决**
Q1 配对检验	已补：逐 seed 配对 bootstrap 95% CI 与 Welch t（§7.7/附录 B）	**已答**
Q2 移植预热/长预算尝试	未做（§5.2 澄清维持：直接替换+随机初始化+从零训练）；列入清单	已答（范围）
Q3 W&B 实现细节	§6 实现澄清维持（对称性内建于 W_td 复用；超参逐 run 记录）	**已答**
Q4 W_res 量化机制	§7.4 已含（非加性残差主干，int4 豁免配方）；替代残差设计未测	**已答**（配方）+挂账（设计）
Q5 最小可行 200-300M 方案	旧「≈350M 单点 ≈60 A100·h / 4080 不可行」判断**被实测推翻**：d1024/24L 本地 4080 单 run ≈2h（6-14GB VRAM），全套（含 3-seed + 夹逼 + 解耦）≈30 GPU·h 零积分完成	**已答并被超越**

V4 建议中仍挂账的实验（按优先级）：500M 判决性点（B4 解耦后优先级升高，需云算力）、渐进插值/蒸馏迁移、W&B 变体筛选矩阵、tokenizer/seq 敏感性、optimizer/wd/init 对称网格、全局 seed≥7。

附录 E：版本历史

v5.1-v5.9 为未发布的内部里程碑（工作日志性质）；对外发布版本为 v1/v2/v2.1/v3/v4（AiraXiv 2609.0010）与本版 **v6.0**。

版本	日期	性质	主要变更
v1-v3	09-07~09-10	发布	因果修正、公平对照、机制分解、遗忘/适应边界（详见各版内注）

版本	日期	性质	主要变更
v4	09-16	发布	V7 终端形态（稀疏兼容/CPU 端侧/安全框架）、第五次测量偏差修正
v5.0-v5.2	09-16~17	内部	启智 100M 规模验证；v5.1 撤回「+40.1%」伪影（第 6 起偏差）；v5.2 终判 TF 领先 33.7%
v5.3	09-17	内部	V8-A 稳定化解锁证伪（bf16 逐位复现失稳——「快速=自毁」内禀性质）
v5.4-v5.5	09-17	内部	90M 冷启动/流式双判决——适应优势随规模放大，确立混合架构路径
v5.6-v5.7	09-17~18	内部	混合最小原型判决成立；3-seed 复现全判据命中
v5.8-v5.9	09-18	内部	帕累托前沿 + 主干锚定机制；96M 纯 CPU 端侧链路 4/4 判据、step50 协议发现
v6.0	**09-21**	**发布**	**混合架构主线定稿**：①330-354M 外推 3/3 全胜 + 双侧夹逼 + 收窄归因解耦（§4.7，1.76× token 效率）②适应主张诚实重述 + PEFT 3-seed 对照（§7.7，第 7 起偏差）③主干锚定消融转正、h2e4 终端默认 3-seed、16+8 反转定性、gating 阶梯（§7.6）④step50 机制分解（§8.3）⑤章节 8/9 乱序修复、版本历史、D.3 审稿回应（W1/W5/Q1/Q5 已解决）