Stage-11 Warp — Wrap-Up & Baseline v4b

Scope. A concise rewrite/roll-up of our Stage-11 experiments to date, consolidating the minimal "always-on" warp doctrine and the budget-GPU baseline (v4b) you froze.

---

0) North Star (Thesis)

Warped space converges better than flat. Keep a small, always-on inward pull so the model lives inside a single-well manifold. Use gates only as *gain control* (never as permission).

---

1) What we built

- **Tap & space.** GPT-2, layer −9 (PCA-2 slice defines "radius" and inward direction). EMA center keeps "inward" meaningful across prompts.
- **Warp step.** Per token, apply a small update:
- $\alpha_t = \alpha_{\min} + (\alpha_0 - \alpha_{\min})\, s_t$ with an optional relative clip ε (not used in v4b).
- **Soft trend gate.** `tr` → `g_tr = σ(k_tr · (tr − τ))` (no hard threshold).
- **Optional Detect (amplifier only).** Matched-filter over a short window + null calibration → `g_det`.
- **Burst shaping.** `s_pre = g_tr × g_det`; apply short `linger` and a `s_latch` floor to suppress flicker → final `s`.
- **Decoding mode.** `--gen_mode stock|geo` controls whether we *decode* under warp (geo) or just *score* under warp while decoding stock.
- **Telemetry fields (per hook line).**
- `tr` (trend), `g_tr`, `g_det`, `s_pre`, `s`, `alpha`.

---

2) Chronology (what we tried → what we learned)

Phase A — Hard-gated warp. Warp applied only if both trend & detect passed hard thresholds. Result: gates rarely opened; `alpha≈0` → no effect.

Phase B — Wobble prompts + instrumentation. Built a prompt pack that induces drift/loops; added per-token sequences and burst metrics to read gate quality.

Phase C — Always-on warp + soft gate. Introduced nonzero `alpha_min`; replaced hard thresholds with sigmoids + short linger. Result: baseline pull on *every* token (`alpha≈0.006–0.015`), with short, targeted bursts (`alpha≈0.02–0.07`) when evidence rises. Matches the thesis.

Phase D — Budget-GPU tuning (v4b). Narrower detect window/sigma, modest latch/linger, slightly higher `τ` and `k_tr` for stability on T4/L4.

---

3) Current baselines

A) v4b — Budget GPU (T4/L4)

Freeze for repro:

```
python3 stage11_ab_eval_v4.py
  --model gpt2 --layer -9
  --prompts wobble_prompts_v1.txt --max_new_tokens 96
  --alpha0 0.05 --alpha_min 0.006
  --trend_tau 0.35 --k_tr 12
  --use_detect 1 --detect_width 24 --detect_sigma 5
  --null_K 32 --null_q 0.92 --k_det 7
  --s_latch 0.30 --linger 2 --ema_center_beta 0.05
  --gen_mode geo --print_every 128 --device cuda
  --out_json ab_results_geo_t4_v4b.json
```

Intent. Always-on curvature with light, verified boosts. Low overhead on T4/L4 while retaining wobble control.

B) v4 — A100 richer runs (reference)

```
--alpha0 0.07 --alpha_min 0.012
--trend_tau 0.30 --k_tr 10
--use_detect 1 --detect_width 32 --detect_sigma 7
--null_K 24 --null_q 0.88 --k_det 9
--linger 3 --s_latch 0.7 --ema_center_beta 0.05
--gen_mode geo
```

---

4) What we observe now

- **Applied everywhere.** `alpha` > `alpha_min` nearly every token; warp influences all prompts.
- **Targeted boosts.** Short bursts where `tr` increases and Detect confirms; logs show `s` spikes aligning with drift/loop spans.
- **String-level behavior.** With `--gen_mode geo`, more graceful exits on wobble prompts and fewer loop/format lapses (qualitative); near-neutral behavior on calm prompts.
- **No denoising yet.** Current wins are from geometry alone.

---

5) Pass/Fail checks (quick sanity)

- **Convergence:** mean per-token radius shrink > 0; end-radius < start-radius per prompt.

- **Warp presence:** `min(alpha_seq) ≥ alpha_min`; applied_rate ≈ 1.0.
- **Burst quality:** mean burst length ≥ 2; adjacency ≥ 0.6 (count tokens with `s ≥ 0.5`).
- **Token economics:** ΔLP inside bursts > 0; outside ≈ 0.
- **Safety (optional):** enable `--eps` (e.g., 0.20–0.25) if you want a relative step clip.

---

6) Risks & mitigations

- **Center drift.** Keep small EMA (`β≈0.05–0.10`); allow a few warmup tokens before measuring trend.
- **Layer sensitivity.** Quick sweep {−6, −9, −12} to find the cleanest well; −9 is working now.
- **Detector reach.** If over/under-firing, adjust `detect_width/sigma` and `null_q`; remember Detect modulates gain only.
- **Harness caveat.** If you decode stock, improvements may be invisible; use `--gen_mode geo` for behavioral diffs.
- **Telemetry integrity.** Deep-copy sequences per row; recompute `steps_applied` from `alpha_seq` when auditing.

---

7) Next experiments (tight loop)

1) **Thesis-pure** (Detect off) on wobble pack → confirm ΔLP & radius metrics. 2) Same config with `--gen_mode geo` → verify string-level wins. 3) **Soft-precision** (Detect on) → ensure boosts concentrate inside short spans. 4) **Layer sweep** (−6/−9/−12) with quick metrics. 5) **Ablations**: EMA off vs on; linger 0 vs 2/3; detect off vs on. 6) **Denoising**: introduce gentle residual clean-up only after we lock baselines.

---

8) Minimal doctrine (final)

- Curvature is *constant* (small always-on pull).
- Evidence scales the *gain* (soft gates); gates never decide *whether* to warp.
- Keep the mechanism simple and observable; add denoise later if/where it pays.

---

9) Appendix — Field meanings (from hook lines)

- `tr`: inward trend (fractional radius shrink; + is "good").
- `g_tr`: sigmoid-scaled trend score.
- `g_det`: detector score (0–1) after windowing + null model.
- `s_pre`: raw soft-gate (`g_tr × g_det`).
- `s`: post-latch/linger soft-gate used for step sizing.
- `alpha`: applied step size.

Baseline check: In calm spans expect `alpha≈alpha_min`. In drift spans, short `s` bursts lift `alpha` briefly and then decay with linger.

Status: *v4b* frozen as the new baseline for budget GPUs; *v4* remains the A100 reference.