Analyze (run01, Habitat only)

□□□□□ run01 □□□□□□□□□□□□□□□□□□□□□□□□

0. □□ Gap □□□□□□□□□□□□□□

- Gap A□Viewpoint-Induced Reuse Failure□□□□□□□

-

vis_analyze/reports/run01/figures_pub/fig_gapA_publication.png

- □□□□□□□□□□□“□□□□/□□□□”□□□□□□□□□□□ view-aware reuse□

- Gap B□Instruction-Conditioned Semantic/Compute Mismatch□□□□□□□

-

vis_analyze/reports/run01/figures_fancy/fig_gapB_budget_mismatch_fancy.png

- □□□□□□□□□□□□□ decode/refresh □□□□□□□□□□□□□□□□□□□

□□□□□□

- Gap A □“□□□□□□□□□□□□□”□

- Gap B □“□□□□□□□□□□□□□”□

1. □□□□□□□

- □□□□:

vis_analyze/configs/habitat_dual_system_observation_run01_cfg.py

- □□□□:

vis_analyze/data/raw/run01/step_log_rank0.jsonl

- System2 □□:

vis_analyze/data/raw/run01/s2_log_rank0.jsonl

- episode □□:

vis_analyze/data/raw/run01/episode_log_rank0.jsonl

- □□□□:

vis_analyze/data/eval_output/run01/progress.json

- [illegible]: vis_analyze/reports/run01/run01_analysis_stats.json

[illegible]:

- Episodes: 100
- Step logs: 8178
- S2 calls: 2101

2. [illegible]

2.1 [illegible]run01 baseline[illegible]

- SR = 0.700
- SPL = 0.636
- OS = 0.770
- NE = 3.722
- [illegible] = 81.77[illegible]P50=68.5, P90=117.3[illegible]

2.2 Challenge A: Viewpoint-Induced Reuse Failure[illegible]

- raw patch cosine mean = 0.9360
- aligned patch cosine mean = 0.8263
- delta mean = -0.1098
- delta quantiles: P10=-0.3607, P50=0.0, P90=0.0
- aligned_better_ratio = 0.0

[illegible]:

- [illegible]“alignment”[illegible]
- [illegible] A [illegible]/[illegible]
- [illegible]: “naive yaw-only alignment is insufficient under embodied motion”, [illegible] view-aware [illegible]A2/A3[illegible]

[illegible]:

• □ [fig_gapA_publication.png](vis_analyze/reports/run01/figures_pub/fig_gapA_publication.png) □□

- A □□□ECDF□□□ aligned □□□□□□□□□□□□□□□ raw□

- B □□□delta □□□□□ $\Delta = s_{aligned} - s_{raw}$ □□□□□□□□□

- C □□□|Δyaw| □□□□□□□□□□□□□□□

- D □□□□□□□□□□□□□□□□□□□□□

• □□ Gap A □□□□□□□**□□□□□□□□□□□□□ token**□

□□□□□□□□□□□□□□:

• □□□□□: vis_analyze/reports/run01/gapA_publication_stats.json

• delta_mean = -0.1098□95% CI = [-0.1131, -0.1067]□□□ 0□

• P(delta > 0) = 0.0

• Cohen's d(aligned - raw) = -0.872□□□□□□

□□□□□□□□□□:

• “The negative delta with a non-overlapping 95% confidence interval and large effect size indicates that naïve alignment does not recover reusable correspondence under embodied viewpoint changes.”

2.4 Gap A □□□□□□□□

□□□□□□□□□:

• □: vis_analyze/reports/run01/deep_dive/fig_gapA_deep_dive.png

• □□: vis_analyze/reports/run01/deep_dive/gapA_deep_dive_stats.json

□□□□□□□□□□“□□□□□□□□”:

• □□□□early/mid/late□:

- early mean delta = -0.1453□95%CI [-0.1513, -0.1399]□

- mid mean delta = -0.0960□95%CI [-0.1015, -0.0910]□

- late mean delta = -0.0881□95%CI [-0.0937, -0.0827]□

- □□□□□□□□□□□□□□□□□□□□□□□□□□□

• □□□□□success vs failure□:

- success mean delta = -0.1027

- failure mean delta = -0.1055

- corr(delta, NE) = 0.051□□□□□

- □□□Gap A □“□□□□□□□□”□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□:

• scene □□□□□□□□ 1 □□□□□2azQ1b91cZZ□□□□ run01 □□□□□□□□□

• □□□□□□□□□□“□□□□□□□□□Gap A □□□□”□

• □□□□“□□□□□□□□”□□□□□□ run02□□ scene □□□□□□□□□□□□

2.3 Challenge B + Efficiency: System2 □□□□

• S2 □□□□: 2101

• prompt_len mean = 1903.3□P50=1894, P90=2304□

• gen_len mean = 4.70□P50=5, P90=8□

• total_len mean = 1908.0□P50=1897, P90=2312□

• preprocess_ms mean = 75.55□P50=73.67, P90=92.71□

• generate_ms mean = 372.28□P50=320.07, P90=507.35□

• decode_ms mean = 0.150□P50=0.146, P90=0.162□

□□:

• System2 □□□□□ generate □□□□□decode □□□□□□

• prompt token □□□gen token □□□□□ Method-C□□□□□□□□□□□□□

• □□□□□□□□ “□□□□□□□□□□□□□ S2 □□□□□□□□□”□

□□□□□□□:

• □

[fig_gapB_budget_mismatch_fancy.png](vis_analyze/reports/run01/figures_fancy/fig_gapB_budget_mismatch_fancy.png) □□

- B1 □□□□ prompt □□□ generate □□□□□□□

- B2 □□□□ generate □□□□□□□□

- B3 □□□□ gen_len □□□□□□□□□

- B4 □□□□ episode □ S2 □□□□□□□□

• □□ Gap B □□□□□□□**□□□□/□□□□□□□□□□□□□□□□**□

3. □□□□□□□□□□□□

Fig.2 □□□□□□□□Challenge A□

□□□:

• □□“□□□□/□□□□□□□□□□□□□□”□

□□□□:

• Fig.2(a): raw_mean □ aligned_mean □□□□□□□□□/□□□□□

• Fig.2(b): delta_mean □□□□□□□□□□□

• Fig.2(c): □□□□□step □□□□ raw/aligned □□

□□□□:

• □□ step_log □ similarity.raw_mean, similarity.aligned_mean, similarity.delta_mean

□□□□:

• “Under the current naïve alignment setting, aligned similarity does not consistently improve over raw similarity, indicating that viewpoint-aware reuse requires stronger geometric/semantic alignment than simple yaw-based warping.”

□□□□□ fancy □□□□□

•

vis_analyze/reports/run01/figures_pub/fig_gapA_publication.png

Fig.3 □□□□□ S2 □□□□Challenge B□

□□□:

• □□ System2 □□□□□□□□□□□□□□□□□□□□□□□

□□□□:

• Fig.3(a): □ episode □ system2_calls □□

• Fig.3(b): prompt_len vs generate_ms □□□

• Fig.3(c): gen_len □□□□□□□□□□□□□□□□□

□□□□:

- progress.json: system2_calls
- s2_log: prompt_len, gen_len, generate_ms, preprocess_ms

□□□□:

- “System2 spends most latency budget in generation while producing short outputs, motivating a stage-aware decode budget controller rather than static max-token settings.”

□□□□□ fancy □□□□□

- vis_analyze/reports/run01/figures_fancy/fig_gapB_budget_mismatch_fancy.png

Fig.5 □□-□□□□□□□□Pareto baseline anchor□

□□□:

- □□□ A/B/C □□□□□□□□□□□□□□

□□□□:

- x □: □□ System2 generate_ms□□□ episode □□□□
- y □: SR / SPL□□□□□□□□

□□ baseline □□:

- generate_ms_mean = 372.28 ms
- SR = 0.700
- SPL = 0.636

4. □□□□□□□□□□□□□□□

Challenge A □□□□□□□

We analyze patch-level cross-frame similarity under embodied navigation dynamics. While adjacent observations maintain high raw similarity on average, naïve yaw-based alignment does not improve matching quality and often decreases aligned similarity. This suggests that simple geometric warping is insufficient for robust token reuse in VLN, where translation, depth variation, and semantic layout changes jointly affect correspondence. Therefore, Challenge A should be addressed with stronger view-aware reuse mechanisms beyond position-wise or weak alignment baselines.

Challenge B □□□□□□□

From 100 validation episodes, System2 is invoked 2101 times, with long prompts (mean 1903 tokens) but short generations (mean 4.7 tokens). Latency is dominated by generation (mean 372 ms) rather than decode overhead. This pattern indicates that static decoding budgets are suboptimal: most calls do not require long autoregressive expansion. The evidence supports an instruction/stage-aware budget controller that adapts generation length and refresh frequency to reduce compute while preserving navigation quality.

5. □□□□□□□□□□□□

- A □□□□ Challenge A□:

- □□□“naive alignment”□□□□□□□

- □□ A2□depth/geometry assisted alignment□□□□□□□□□□□□ run02□

- □□ delta_mean □ aligned_better_ratio □□□□□

- B+C □□□□ Challenge B□:

- □□□ C1□□□□ decode budget□□ B1□□□□□□□□□

- □□□□□□□□□□□ run02 summary□

- □□ generate_ms_mean□SR□SPL□□□□□□□□□

6. □□□□

- □□ Challenge A □□□□□□“□□□□”□□□□□□□□□□□□□□□□
- □□□□□□□□: “baseline failure -> improved design” □□□□□□□ claim□
- □□□□□□ run □□□□□□□□□□□□□□□□