

Peer Review

Review of: "Fine-Grained Alignment in Vision-and-Language Navigation through Bayesian Optimization"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper starts from an observation I find convincing: in contrastive training for Vision-and-Language Navigation, language negatives can already be made fine-grained (Airbert's keyword replacement), but vision negatives are still coarse, either shuffled paths or beam-search failures that look nothing like the positive path. The authors propose FGVLN, a min-max scheme in which a Tree-structured Parzen Estimator chooses a binary mask over the frames of a positive trajectory (Eq. 3), the masked frames are overwritten with a replacement frame x_r (Eq. 4), and the resulting hard negatives are appended to the batch for the path-ranking loss of Eq. 2. A target copy of the model scores candidate masks in the inner loop and is synchronised with the online model every J steps (§IV-B, "Delayed Updates").

The evidence offered is of three kinds. Table 1 compares L2 distances between positive and negative trajectory embeddings for Lily and FGVLN on 1,000 sampled pairs. Tables II–IV report navigation results: R2R discriminative (Table II), R2R generative with the recurrent VLN agent (Table III), and REVERIE generative (Table IV). Table V is an ablation over BO iterations, delayed updates, number of fine-grained negatives, and in-domain versus out-domain replacement frames, together with a random-mask control.

2. Whether the conclusion holds on the data

Several things here are genuinely good. The idea of letting a black-box optimiser search for the frames whose replacement most confuses the current model is simple, cheap to bolt onto an existing pipeline, and well motivated by the hard-negative literature the authors cite. The random-mask control (Table V, index 2, FGVLN-Rand) is exactly the control I would have asked for, and it is there. The authors also test

the trained encoder outside the setting it was tuned in, moving it into a generative agent on two benchmarks, which is more than many papers of this kind do.

Now to how far the numbers carry the claims. The direction of every headline comparison is right; the size is small, and the paper does not yet give the reader a way to judge whether it exceeds run-to-run noise.

Size of the gains. In Table II, validation unseen, SR goes from 66.70 (row Lily) to 67.69 (row FGVLN), which is 0.99 percentage points. The text after Table II calls this "a 1.48% improvement in terms of SR"; that is the relative gain, $0.99 / 66.70 = 1.48\%$. For SPL, the same sentence gives 3.12%, while the table shows 0.62 versus 0.64. With two decimals, the SPL difference could be anywhere from about 0.01 to 0.03, so the 3.12% cannot be checked from the table ($0.02 / 0.62$ would give 3.23%). I would state the absolute differences (+0.99 SR, +0.02 SPL) and, if relative numbers are kept, give SPL to three decimals and say that both percentages are relative to Lily.

"Outperforms all the previous models." In Table II, validation unseen, NE is 3.40 for FGVLN against 3.37 for Lily, so on one of the four reported metrics, Lily is slightly better. In Table III, validation unseen, NE is 4.06 for Rec (FGVLN) against 4.01 for Rec (Airbert), and on validation seen, SR is 73 versus 74. In Table IV, OSR is 30.14 for Rec (FGVLN) against 34.51 for Rec (AirBert). None of this undermines the method. It does mean the sentence should say "best SR and SPL" rather than "outperforms all the previous models", and the REVERIE paragraph's closing sentence about validating "the robustness" of the approach asks more of one split and one run than it can give.

Variance and selection. I could not find the number of training runs, seeds, or any spread for Tables II–V. With gains of about one point on R2R, a reader needs at least three seeds per configuration, or a paired bootstrap over episodes, to know whether the sign is stable. A related point: the configuration used everywhere (FGVLN-add-iter) was chosen because it "outperforms all other settings in unseen environments" (SV-D, first paragraph), and Table II then reports results on validation unseen. Selecting on the split you report is a mild but real optimistic bias. A result on the R2R test server, or at least a statement that the selection subset is disjoint from the reported split, would close this.

The ablation (Table V). The trends are informative, and some of the text misreads them. The delayed-update comparison (index 3 versus index 4) gives val_unseen $49.52 \rightarrow 51.02$, which is +1.50 points and is described as "a 3% performance improvement" (relative, $1.50 / 49.52 = 3.03\%$); val_seen in the same pair falls from 60.18 to 57.66, which the text does not mention, and index 4 on val_unseen (51.02) is still below the Lily baseline in index 1 (51.38). The out-domain sentence ("2.6% improvement over the best in-

domain approach and a 1.9% improvement compared to the baseline") matches the table only as relative gains: 52.36 versus 51.02 is +1.34 points, 52.36 versus 51.38 is +0.98 points. Then one factual slip: the text says that "for the seen dataset, the model with three iterations (row 5) performed better" than index 7, but row 5's `val_seen` is 61.25, below index 7's 61.98. The three-iteration model that beats index 7 on `val_seen` is index 6 (63.48). The ablation subset size is not stated, and its baseline (51.38 `val_unseen`, index 1) differs from the full-data Lily figure in Table II (66.70), so I cannot judge how many episodes the 1–5 point differences rest on.

Embedding analysis (Table 1 and Fig. 3). The key column is Fine-grained: mean L2 distance 4.72 for Lily versus 7.64 for FGVLN. Two things make this hard to read as it stands. First, raw L2 distances are not comparable across two differently trained encoders unless their embedding norms are comparable; the same table shows Random Shuffle distances falling from 74.80 (Lily) to 43.25 (Ours), which, read naively, would mean the new encoder separates shuffled negatives worse. A scale-free measure (cosine distance, or each negative distance divided by the encoder's average positive-to-positive distance) would settle it. Second, the σ values are much larger than the means for a non-negative quantity: Fine-grained $\sigma = 95.79$ against $\mu = 4.72$ for Lily, Random Shuffle $\sigma = 200.18$ against $\mu = 43.25$ for Ours. That is possible only with very heavy tails. I would check that the μ and σ columns are not swapped, and report medians and interquartile ranges. For Fig. 3, the projection method (t-SNE, PCA, UMAP) and its settings are not named.

Claims without a matching experiment. Contribution 2 in §I says the encoders show "uniform attentions across sequences" and form "complex decision boundaries". I could not find an attention analysis or a decision-boundary analysis anywhere in the paper as rendered on Qeios. Either the analysis should be added or the contribution reworded to what Table 1 and Fig. 3 actually show.

3. Reproducibility

The paper gives more training detail than many in this area: learning rates (4×10^{-5} for the MLM/MVM stage, 1×10^{-5} for the FGVLN stage and the generative agent), batch sizes (12, 16, and 8), 30 epochs for the FGVLN stage, 300,000 iterations of imitation learning plus A2C for the generative agent, and the hardware (four and eight V100s, one RTX 4080). A code link is given in the abstract. I did not run it and could not verify from the text what it contains.

What I could not find:

- the value of J , the synchronisation period for the target model in §IV-B;

- any constraint on the mask: Eq. 3 lets the optimiser choose any subset of the K frames, and if many frames are replaced the negative stops being fine-grained, so the number of replaced frames per negative (fixed, bounded, or observed distribution) is needed;
- the TPE settings (number of startup trials, candidates per iteration, library and version) and whether a fresh optimiser is fitted per trajectory, per batch, or per epoch;
- how x_r is drawn: §IV-A refers to "Section V-D" for the generation flow, but V-D only says in-domain frames come from the same room and out-domain frames from a different room;
- the extra compute: each inner loop needs R forward passes of the target model per batch, and wall-clock overhead relative to plain Lily fine-tuning is not reported;
- whether baseline rows in Tables II–IV were re-run by the authors or copied from the original papers, and with which checkpoints;
- the size and selection rule of the ablation subset, and seeds anywhere.

4. *What specifically to fix*

1. Text after Table II: replace "1.48% ... 3.12%" with absolute differences (+0.99 SR, +0.02 SPL on val unseen); if relative gains stay, report SPL to three decimals so the 3.12% can be checked (from the rounded table, $0.02 / 0.62$ gives 3.23%).
2. §V-D, iterations paragraph: "the model with three iterations (row 5)" should be index 6 (val_seen 63.48 vs 61.98 for index 7). Row 5 (61.25) is lower than index 7.
3. Report mean and standard deviation over at least three seeds for Lily and FGVLN in Table II, and ideally for indices 1, 2, 6, and 7 of Table V. Without this, the one-point gains cannot be separated from noise.
4. Soften "outperforms all the previous models" (text after Table II) and the REVERIE closing sentence: NE in Tables II and III and OSR in Table IV favour a baseline.
5. Table 1: add a scale-free distance, check the μ/σ columns, and explain why the Random Shuffle distance drops from 74.80 to 43.25 under the new encoder.
6. Contribution 2 in §I. Either add the attention-uniformity analysis it promises or rewrite it.
7. §IV-A, Eq. 3: write the mask as $M \in \{0,1\}^K$ rather than $\mathbb{R}^K$, and state the cap on the number of replaced frames.
8. Give J , the TPE configuration, the ablation subset size, and the per-epoch cost of the inner loop in §V-A.

9. If possible, add an R2R test-server result for the final configuration, since it was selected on validation unseen.
10. Numbering and leftovers from the supplementary version: §V-D refers to "Table 4 of the main manuscript" and "Section 4.2", which should be Table V and §IV-B; the embedding table is "Table 1" while the others use Roman numerals.
11. Table IV, row "Rec (VLN-Bert)[40]": reference [40] is Devlin et al. (BERT), and the recurrent VLN-BERT is [41], which is also attached to the OSCAR row. Please check both citations.
12. §V-A says REVERIE "uses the same data splits as the R2R dataset"; it shares environments but has its own instruction splits, so I would phrase it that way.
13. The Conclusion mentions only R2R. Since the abstract and Table IV include REVERIE, a sentence on it belongs there too.

5. Recommendation

Minor revisions. The method is a clean, sensible contribution: a searched hard-negative generator for the vision side of VLN, with a proper random-mask control and transfer to a second setting and a second benchmark. The gains are modest and currently rest on single runs, the percentages in §V-C and §V-D are relative gains given without the absolute differences, and one row reference in §V-D does not match Table V. Seeds, absolute differences, a scale-free version of Table 1, and the missing hyperparameters would make the paper's claims as solid as its idea.

Declarations

Potential competing interests: No potential competing interests to declare.