DO VISION-AND-LANGUAGE NAVIGATION AGENTS FOLLOW INSTRUCTIONS—OR JUST THE WORDING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-and-language navigation (VLN) requires agents to follow natural-language route instructions using visual observations. Users can express the same navigation requirements in different words, but aggregate success rates alone do not reveal whether agents consistently solve the same tasks across these descriptions. Our episode-level analysis of seven agents on R2R-CE and RxR-CE reveals a substantial gap beneath nearly unchanged aggregate scores. On each benchmark, results pooled across agents show that more than half of the episodes solved under the original instructions fail under at least one valid rephrasing. Gains on other episodes mask these losses, and for some agents, even capitalization or punctuation changes can reverse the outcome. To examine whether greater instruction diversity can reduce this wording dependence, we fine-tune three agents using only rephrasings of their own training instructions while retaining the existing navigation supervision. The intervention increases reliable success, reduces outcome reversals, and improves performance on the original benchmark instructions. Reversals also decrease on average for rephrasing types held out from fine-tuning. Physical-robot experiments with human rephrasings of ten indoor route instructions reproduce both wording dependence and the intervention's benefits. Without real-world-specific adaptation, the fine-tuned checkpoints increase the fraction of routes completed successfully under all five descriptions from 24.4% to 34.4%, averaged across models. Together, these findings distinguish successful navigation under a particular phrasing from reliable instruction following. VLN evaluation should assess whether success persists across semantically equivalent descriptions, and training should treat instruction diversity as a core consideration. Our code is available here.

1 INTRODUCTION

Vision-and-language navigation (VLN) requires an agent to follow a natural-language route instruction using its visual observations (Anderson et al., 2018; Krantz et al., 2020). As the agent moves, it must relate the instruction's ordered actions, landmarks, and spatial relations to its observations and progress. Users can express the same navigation requirements in different words, so robust instruction following should extend across semantically equivalent descriptions. This raises a basic question: does the success of current VLN agents reflect an understanding of what the instruction requires, or does it depend on familiar wording?

Prior work has documented sensitivity to semantically equivalent prompts in language models (Sclar et al., 2024; Sun et al., 2024) and has examined linguistic sensitivity (Hahn et al., 2023), instruction errors (Taioli et al., 2024), and broader perturbations (Jiang et al., 2026) in VLN. These findings motivate examining not only whether wording changes affect average performance, but whether success persists on the same navigation tasks. Aggregate success rates can remain stable even when the episodes that succeed change. We therefore compare each episode under its original instruction and semantically equivalent rephrasings, tracking switches between success and failure.

We conduct this diagnosis on R2R-CE (Krantz et al., 2020) and RxR-CE (Ku et al., 2020), using controlled rephrasings under fixed evaluation conditions. On each benchmark, aggregate success across seven recent agents remains nearly unchanged even though more than half of the episodes solved under the original instruction fail under at least one valid rephrasing. Improvements on some

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

episodes mask losses on others. Sensitivity also differs across models, and for some agents, even capitalization or punctuation changes can reverse the outcome. These results show that successful navigation can remain tied to a particular phrasing even when the instruction's requirements are unchanged.

These patterns motivate a training hypothesis: limited exposure to alternative phrasings may contribute to wording dependence. We examine this possibility through *rephrasing-only fine-tuning* of three agents, using only rephrasings of their own training instructions while retaining the environments, trajectories, visual observations, and expert actions. The intervention increases reliable success, reduces outcome reversals, and improves performance on the original benchmark instructions. To determine whether this improvement extends beyond the forms used during training, we evaluate rephrasing types held out from fine-tuning and find gains on average. Together, these results show that existing navigation supervision can support more reliable instruction following when paired with more diverse expressions of the same requirements.

To test whether wording dependence and the benefits of the intervention carry over beyond simulation, we deploy the released and rephrasing-trained agents on a quadruped robot. External writers provide four rephrasings of each of ten indoor route instructions in their own words, with independent checks that the intended routes are preserved. Under these human rephrasings, success again depends on how the same route is described. Without real-world-specific adaptation, the rephrasing-trained checkpoints increase the fraction of routes completed successfully under all five descriptions from 24.4% to 34.4%, averaged across models. This replication connects the diagnosis to physical instruction following: wording differences introduced by different people can affect task completion, and the same training intervention can mitigate this dependence.

Together, our findings support a central distinction: successful navigation under a particular phrasing is not the same as robust instruction following. We argue that VLN evaluation should assess whether success persists across semantically equivalent descriptions, and that instruction diversity should be treated as a core training consideration alongside visual and environmental diversity.

Our study makes three contributions:

- **Diagnosis.** We reveal wording dependence hidden by aggregate VLN scores through controlled rephrasings and episode-level analysis.
- **Training intervention.** We show that diversifying instruction wording improves reliability and original-benchmark performance using existing navigation supervision, with gains on held-out rephrasing types.
- **Real-world replication.** We reproduce both wording-dependent success and the intervention's benefits on a physical robot under human rephrasings, without real-world-specific adaptation.

2 Related Work

2.1 Vision-and-Language Navigation

Vision-and-language navigation (VLN) asks an agent to follow natural-language route instructions using visual observations (Anderson et al., 2018). Early formulations represent movement as transitions between viewpoints on a predefined navigation graph, while VLN in continuous environments (VLN-CE) (Krantz et al., 2020) requires agents to execute navigation actions in continuous 3-D space. R2R-CE (Krantz et al., 2020) and RxR-CE (Ku et al., 2020) provide benchmarks for instruction following in this setting, including generalization to unseen environments. NaVid (Zhang et al., 2024b) and NaVILA (Cheng et al., 2024) predict navigation actions from route instructions and visual histories using pretrained vision-language models. StreamVLN (Wei et al., 2025) incorporates past observations through streaming context modeling, while JanusVLN (Zeng et al., 2026) uses memory mechanisms.

2.2 Instruction Sensitivity and Robustness

Studies of language models show that performance can vary under meaning-preserving changes to prompt formatting (Sclar et al., 2024) and instruction phrasing (Sun et al., 2024). For vision-

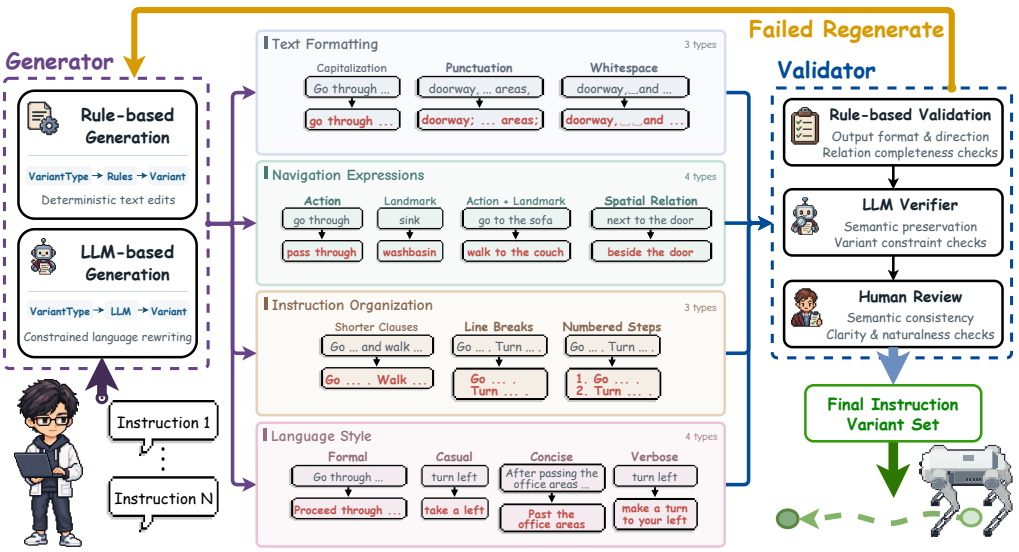


Figure 1: Instruction rephrasing pipeline. Three text formatting types use deterministic rules, and the remaining eleven types use language-model generation. The center panels illustrate the fourteen types across text formatting, instruction organization, navigation expressions, and language style, with rewritten text shown in red. Candidates undergo rule-based checks, LLM verification, and manual review before acceptance.

language reward models, ROBORMBENCH (Jeung et al., 2026) shows that equivalent goal descriptions can lead to inconsistent scores for the same robot trajectory. In robotic manipulation, LIBERO-Para (Kim et al., 2026) evaluates vision-language-action models by independently varying action expressions and object references while preserving the intended task. Its PRIDE metric weights successful executions by the degree of lexical and structural change. Together, these studies show that preserving task meaning does not necessarily preserve model performance.

In VLN, route instructions specify ordered actions, landmarks, and spatial relations that agents must ground as their visual observations change. Diagnostic studies examine agents' reliance on environmental cues (Zhang et al., 2020) and their use of linguistic information (Zhu et al., 2022; Hahn et al., 2023). One way to probe this dependence on language is to perturb the instruction, as in adversarial attacks on navigation agents (Lin et al., 2021). A related line studies errors in the instructions themselves, seeking to detect and localize inaccurate content in route descriptions (Taioli et al., 2024). At a broader multimodal level, NavTrust (Jiang et al., 2026) evaluates embodied navigation under RGB and depth corruptions, together with instruction perturbations.

2.3 Data Augmentation for VLN

Data augmentation has long been used to improve generalization in VLN. Speaker-follower models (Fried et al., 2018) generate synthetic instructions for sampled trajectories to expand training data, while EnvDrop (Tan et al., 2019) combines environmental dropout with back-translation to generate additional instruction–trajectory pairs. More recent work scales synthetic instruction generation across diverse environments to build large navigation training sets (Kamath et al., 2023). These studies establish instruction generation as a means of enriching navigation training.

3 Diagnosing Wording Dependence

To examine whether navigation success depends on an instruction's wording, we vary how a route is described while keeping its navigation requirements fixed. This allows us to test whether success under one phrasing persists under semantically equivalent alternatives. Our criterion is consistent task completion, which allows successful trajectories to differ.

3.1 CONTROLLED EVALUATION

A valid rephrasing preserves the instruction's action sequence, landmark referents and identifying attributes, spatial relations, and stopping conditions. Each type targets a specified aspect of expression within these constraints (Figure 1). *Text formatting* changes capitalization, punctuation, or whitespace while retaining the words and their order. Changes to *navigation expressions* replace phrases for actions, landmarks, and spatial relations while preserving the same actions, referents, and relations. At the level of the whole instruction, *instruction organization* changes how the action sequence is presented, and *language style* varies tone and conciseness without adding or removing navigation requirements.

We implement text formatting through deterministic edits and use type-specific language-model prompts for the other dimensions. For navigation expressions, the prompts restrict changes to the targeted expressions and any necessary local grammatical adjustments. To check that candidates meet the semantic and type constraints, we combine automatic screening, LLM verification, and manual review. Automatic checks validate formatting edits and flag malformed text or potentially altered route information; the verifier then checks semantic preservation and adherence to the requested type. Rejected candidates are regenerated using feedback, and manual review completes verification before accepted variants enter the evaluation sets. Appendix A provides the generation procedure and type coverage.

Using these verified variants, we evaluate the val-unseen splits of R2R-CE (Krantz et al., 2020) and RxR-CE (Ku et al., 2020), containing 1,839 and 1,223 episodes, respectively, in environments absent from benchmark training. Each episode is evaluated under its original instruction and every valid variant, giving up to fifteen phrasings of the same navigation task. For each type, comparisons between original and rephrased instructions use the same episodes with valid variants. These instruction sets are fixed before evaluation and shared across agents and checkpoints.

For each agent, every phrasing of an episode is run from the same initial state, with the checkpoint and inference settings held fixed. Deterministic inference and a unified simulator configuration keep the execution conditions constant across the paired runs.

Success rate (SR) measures the fraction of episodes in which the agent stops within 3 m of the goal. Although SR describes completion under a given phrasing, it does not identify whether the same episodes succeed after rephrasing. We therefore track changes in individual outcomes through *flip rate* (FR). For episode i, let V_i index the original instruction ($j{=}1$) and all valid variants, and let $s_{ij} \in \{0, 1\}$ denote success under phrasing j. Across N evaluation episodes, FR counts both success-to-failure and failure-to-success changes over all valid original–variant pairs.

$$\mathrm{FR} = \frac{\sum_{i=1}^{N} \sum_{j \in V_i \setminus \{1\}} |s_{ij} - s_{i1}|}{\sum_{i=1}^{N} (|V_i| - 1)}. \tag{1}$$

To determine how much success persists across phrasings, we count an episode as reliably successful only when the agent succeeds under its original instruction and every valid variant. The *reliable success rate* (RSR) is the fraction of all N evaluation episodes that satisfy this condition.

$$\mathrm{RSR} = \frac{1}{N} \sum_{i=1}^{N} \prod_{j \in V_i} s_{ij}. \tag{2}$$

Thus, FR measures how frequently rephrasing changes an outcome, while RSR measures successful completion that survives the tested alternatives.

To describe navigation quality, we also report navigation error (NE), the final geodesic distance to the goal, and oracle success rate (OSR), which records whether the agent enters the success region at any point. Success weighted by path length (SPL) captures success adjusted for path efficiency, while normalized dynamic time warping (nDTW) measures agreement with the reference path.

We evaluate the released checkpoints of seven agents on these shared sets: NaVid (Zhang et al., 2024b), Uni-NaVid (Zhang et al., 2024a), NaVILA (Cheng et al., 2024), StreamVLN (Wei et al., 2025), JanusVLN (Zeng et al., 2026), AwareVLN (Guo et al., 2026), and DualVLN (Wei et al., 2026). Model versions and inference configurations are provided in our code release.

Table 1: Performance on original instructions and reliability across rephrasings. SR, NE, SPL, nDTW, and OSR use original instructions. RSR requires success under all valid phrasings, including the original; FR measures outcome reversals over original–variant pairs. NE is in meters; other values are percentages.

	R2R-CE val-unseen							RxR-CE val-unseen						
Model	**SR↑**	**RSR↑**	**FR↓**	NE↓	SPL↑	nDTW↑	OSR↑	**SR↑**	**RSR↑**	**FR↓**	NE↓	SPL↑	nDTW↑	OSR↑
AwareVLN	63.02	21.21	27.05	4.16	54.85	64.66	73.03	65.33	28.29	19.34	4.31	55.72	66.37	72.94
DualVLN	64.06	30.23	16.52	4.12	58.20	70.01	70.47	65.90	27.88	19.65	4.69	56.59	69.05	75.39
JanusVLN	59.05	32.79	13.46	4.97	55.51	65.97	64.00	61.16	29.44	17.02	5.63	54.15	65.36	68.77
StreamVLN	54.43	32.25	11.63	5.00	47.91	63.28	61.45	56.34	26.41	15.93	5.59	48.03	64.16	65.33
NaVILA	54.11	21.15	14.79	5.20	49.55	64.16	61.66	52.66	23.22	15.64	6.07	46.95	61.84	61.65
Uni-NaVid	52.04	28.55	12.35	5.07	48.01	64.64	57.37	58.14	27.88	16.30	5.79	47.19	62.32	67.62
NaVid	41.05	13.87	16.10	5.67	36.07	58.53	48.29	47.59	17.42	19.38	6.77	40.26	55.81	55.85

3.2 Phrasing-Dependent Success

Aggregate scores suggest that navigation performance is largely stable across phrasings. Across the seven released agents and fourteen types, absolute SR changes average 1.01 percentage points on R2R-CE and 1.21 points on RxR-CE, with similarly small mean absolute changes in SPL and nDTW. Individual SR shifts can be larger, reaching 9.03 and 6.21 points, respectively.

This aggregate stability conceals changes in which episodes succeed. Some episodes solved under the original instruction fail after rephrasing, whereas others become successful. Improvements on some episodes therefore offset losses on others, leaving SR nearly unchanged even as individual outcomes reverse. Across valid original–variant pairs, these reversals yield a mean FR of 15.99% on R2R-CE and 17.61% on RxR-CE, averaged across agents.

The extent of this dependence becomes clearer when we ask how many original successes survive every tested phrasing. Pooled across agents within each benchmark, fewer than half do so. On R2R-CE, mean SR is 55.40% and mean RSR is 25.72%, corresponding to 46.43% of the original successes. RxR-CE shows a similar gap, with SR of 58.16% and RSR of 25.79% (Table 1). These measures can even rank agents differently, as AwareVLN has higher SR but lower RSR than StreamVLN on R2R-CE. Success under a particular phrasing therefore overstates how consistently an agent can complete the same navigation task across equivalent instructions.

3.3 Sensitivity Patterns

Sensitivity shows shared trends as well as differences across agents. Concise rephrasings yield the highest FR for every agent on RxR-CE, while sensitivity to text formatting varies markedly across agents on R2R-CE (Figure 2). The same rephrasing type can therefore affect agents to very different degrees.

Capitalization and punctuation edits preserve the instruction's words, word order, and navigation requirements. On R2R-CE, however, punctuation changes produce an FR of 22.51% for NaVILA, exceeding its FR under any of the four language-style types. Even formatting changes can affect whether an agent completes the same navigation task.

These patterns motivate a training hypothesis: limited exposure to alternative phrasings may contribute to the observed wording dependence. If so, a wider range of expressions for the same navigation requirements should help agents maintain success across phrasings.

4 Training Intervention

We test whether greater instruction diversity can make existing navigation supervision support more consistent task completion. Through *rephrasing-only fine-tuning*, we vary the instructions paired with that supervision and assess both cross-phrasing reliability and performance on the original benchmark instructions. Evaluation on held-out rephrasing types tests whether any benefit extends beyond the forms used during fine-tuning.

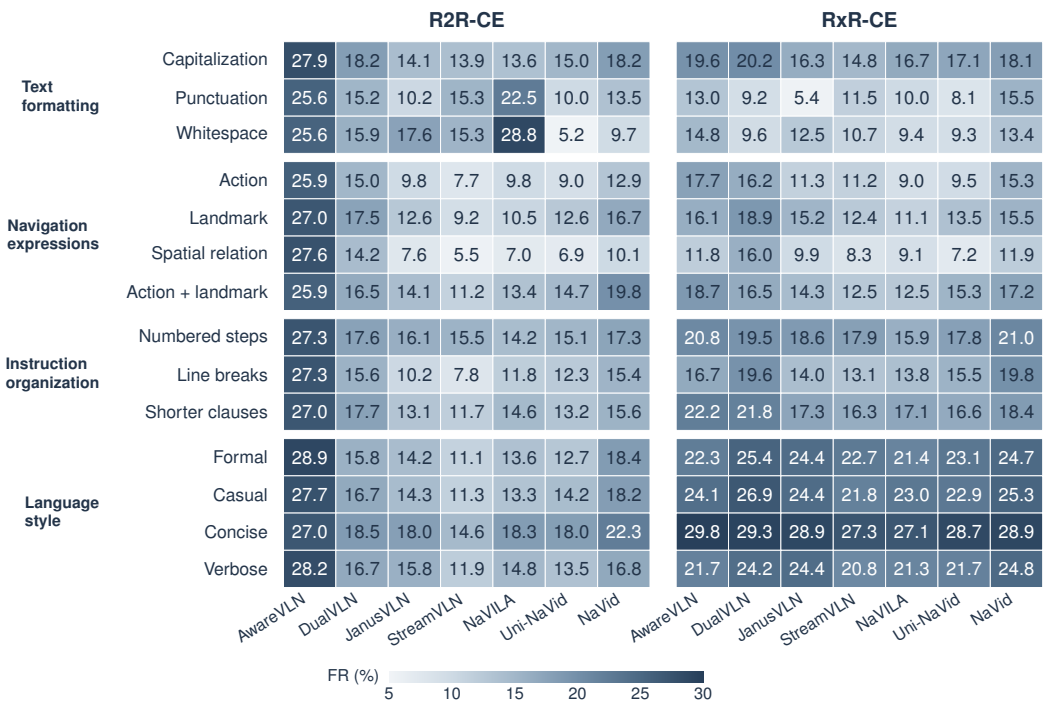


Figure 2: Flip rate by rephrasing type and agent, computed over valid original–variant pairs. Both benchmarks share the same 5–30% color scale.

4.1 Rephrasing-Only Fine-Tuning

For each selected training example, we replace the instruction with a valid rephrasing while preserving the environment, trajectory, visual observations, and expert actions. These examples describe already supervised navigation tasks in different ways, giving the agent exposure to alternative expressions of the same requirements. Only rephrased instructions are used in this fine-tuning stage; the original training instructions are excluded.

We apply this intervention to NaVILA, AwareVLN, and Uni-NaVid, whose training pipelines support instruction replacement with the remaining supervision unchanged. Each model starts from its released checkpoint and uses rephrasings of its own R2R-CE or RxR-CE training instructions. We call the resulting checkpoints *rephrasing-trained*. Released and rephrasing-trained checkpoints are evaluated on the same original instructions and valid variants from §3.1, under the same inference settings.

To test whether gains extend beyond the types used during fine-tuning, we select two types from each of the four dimensions, giving eight training types and six held-out types. The split is fixed before training, with every dimension represented in both subsets. Appendix C provides the type assignment and full results; training-sample construction and model-specific configurations are available in our code release.

4.2 Reliability and Original-Instruction Performance

Rephrasing-only fine-tuning increases RSR and reduces FR for all three models on both benchmarks (Figure 3). RSR gains are larger on R2R-CE, ranging from 6.47 to 17.73 percentage points, compared with 0.82 to 3.27 points on RxR-CE. The reduction in FR also varies, reaching 13.97 points for AwareVLN on R2R-CE but only 0.08 points for NaVILA on RxR-CE. The higher RSR corresponds to successful completion of more episodes under the original instruction and every valid rephrasing.

This benefit also reaches the original benchmark instructions. Although fine-tuning uses only rephrasings, SR on the original instructions increases by 1.85–4.24 percentage points on R2R-

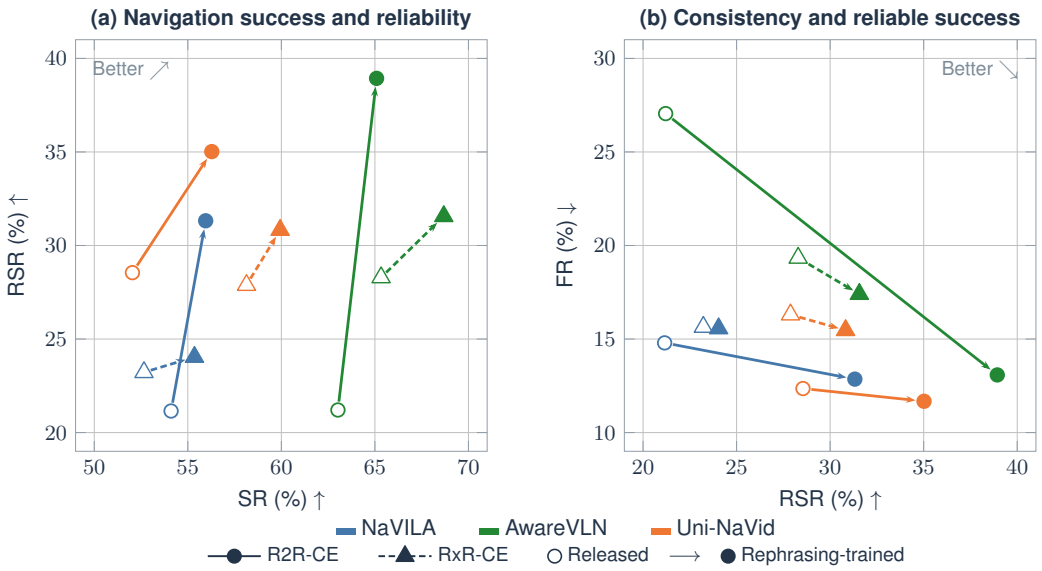


Figure 3: Effects of rephrasing-only fine-tuning on the same evaluation episodes and phrasing sets. **(a)** SR and RSR, with improvement toward the upper right. **(b)** RSR and FR, with improvement toward the lower right. Open and filled markers denote released and rephrasing-trained checkpoints, respectively; arrows connect their measured scores. Colors distinguish models. Circles with solid lines denote R2R-CE, and triangles with dashed lines denote RxR-CE.

CE and 1.80–3.35 points on RxR-CE. SPL and nDTW improve for all three models as well (Appendix C). The intervention therefore improves cross-phrasing reliability alongside performance under the original evaluation conditions, using the agents' existing navigation supervision.

4.3 Unseen Rephrasing Types

The overall improvement leaves open whether the benefit extends to types absent from fine-tuning. We therefore compute FR separately for the eight trained-on types and the six held-out types, pooling valid original–variant pairs within each subset. Released and rephrasing-trained checkpoints are compared on identical subsets.

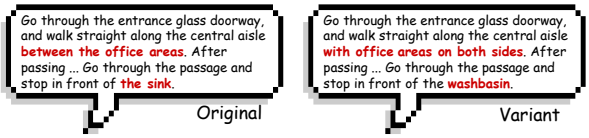
On average, FR decreases on both trained-on and held-out types. On R2R-CE, the mean reduction is 6.04 points on trained-on types and 4.84 points on held-out types; the corresponding reductions on RxR-CE are 0.70 and 1.28 points. The gains vary across agents. AwareVLN and Uni-NaVid reduce FR on held-out types in both benchmarks, whereas NaVILA's held-out FR increases by 0.46 points on R2R-CE and decreases by 0.64 points on RxR-CE (Appendix C).

The reduction in outcome reversals thus extends, on average, to unseen rephrasing types within dimensions represented during fine-tuning. An additional transfer evaluation under erroneous instructions is reported in Appendix D.

5 Real-World Replication

The preceding experiments use constructed rephrasings in simulation. To test whether wording dependence and the benefits of the intervention carry over to physical instruction following, we deploy the released and rephrasing-trained agents on a quadruped robot and evaluate route instructions rephrased by external writers.

Model		SR ↑	RSR ↑	FR ↓
NaVILA	Released	48.7%	23.3%	28.3%
	Rephrasing-trained	54.7%	33.3%	19.2%
AwareVLN	Released	66.7%	43.3%	18.3%
	Rephrasing-trained	70.7%	56.7%	11.7%
Uni-NaVid	Released	22.7%	6.7%	25.8%
	Rephrasing-trained	30.7%	13.3%	21.7%



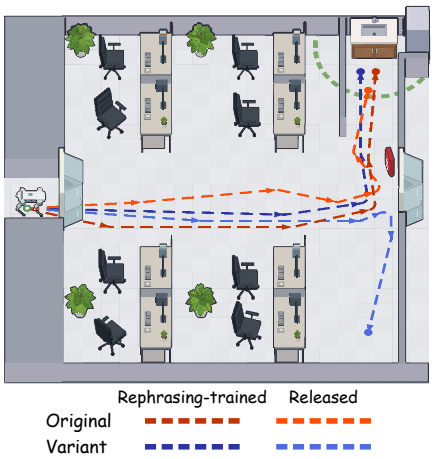


Figure 4: Real-world replication under human rephrasings. **Top left:** SR, RSR, and FR for three agents on 10 routes, averaged over three repetitions. Each route has one reference instruction and four human rephrasings; SR uses all five and RSR requires success under all five. **Bottom left:** A reference instruction and one human rephrasing for an example route. **Right:** AwareVLN trajectories for this instruction pair. The released checkpoint succeeds under one phrasing and fails under the other, while the rephrasing-trained checkpoint succeeds under both.

5.1 Human-Rephrased Navigation

We define 10 indoor routes and write a reference instruction for each in R2R annotation style. Four project-external writers then rephrase each reference instruction in their own words while preserving its navigation requirements. An independent reviewer checks the rephrasings against a floor plan to verify that the intended route is unchanged. Each route is thus described by a reference instruction and four human rephrasings, providing different people's expressions of the same navigation task.

NaVILA, AwareVLN, and Uni-NaVid are deployed in both released and rephrasing-trained configurations under the same inference interface, with no real-world-specific adaptation. Success requires stopping within 1.5 m of the goal. Collisions requiring human intervention, timeouts, and reaching the goal without stopping are counted as failures, so successful completion includes an appropriate stop.

Each route–phrasing–checkpoint combination is run three times. Within each repetition, SR measures success across the reference instruction and all four human rephrasings, while FR compares each rephrasing with the reference. RSR requires success under all five descriptions of a route. We then average each metric across the three repetitions. Appendix E provides the route and execution details.

5.2 Wording Dependence

Across the five phrasings, released agents achieve a mean SR of 46.0%, while RSR is 24.4% and FR is 24.2% (Figure 4, top left). As in simulation, some successes depend on how the same route is described. The trajectories in Figure 4, right, make this difference visible: the released checkpoint succeeds under one of the two descriptions shown and fails under the other, although the route requirements are preserved.

5.3 Intervention Transfer

Under the same deployment conditions, rephrasing-trained checkpoints increase mean SR across the five phrasings to 52.0% and RSR to 34.4%, while reducing FR to 17.5%. The route example illustrates the corresponding change in execution: the rephrasing-trained checkpoint succeeds under both descriptions shown, whereas the released checkpoint succeeds under only one. Improved reliability therefore accompanies improved task completion across the evaluated phrasings.

These improvements span all three agents. RSR increases from 23.3% to 33.3% for NaVILA, from 43.3% to 56.7% for AwareVLN, and from 6.7% to 13.3% for Uni-NaVid. AwareVLN has the strongest released performance and also achieves the largest absolute gain in reliable success. All three agents simultaneously improve SR and reduce FR. Thus, under human rephrasings, the intervention increases successful completion while making outcomes more consistent across different expressions of the same route.

These checkpoints receive no additional training on the real routes or their human rephrasings. Their improvement shows that the benefits of the simulation-based training intervention extend to expressions supplied by people and executed on physical hardware. This replication connects instruction diversity to human–robot navigation, where different ways of expressing the same requirements should support consistent task completion.

6 IMPLICATIONS

For evaluation. Success under a single phrasing provides incomplete evidence of reliable instruction following. Our findings support evaluating whether the same navigation tasks remain successful across equivalent descriptions and reporting SR together with FR and RSR. The three measures distinguish overall completion, changes in which episodes succeed, and success that persists across the tested phrasings. These metrics should be interpreted jointly. An agent that fails under every phrasing has no outcome reversals, so low FR alone does not establish reliable navigation. Conversely, high SR can coexist with frequent reversals when improvements on some episodes offset losses on others. RSR complements these measures by requiring successful completion across the tested descriptions. Comparing agents therefore involves two questions: how many tasks they complete, and how consistently they complete those tasks when the wording changes. This distinction also guides the interpretation of training gains, separating changes in overall task completion from changes in sensitivity to alternative expressions. RSR is specific to this finite set of alternatives, so it should be interpreted with the rephrasing coverage and does not directly estimate success under arbitrary user instructions. The rephrased evaluation sets provide a shared basis for making this comparison across agents.

For training. The intervention shows that existing navigation supervision can support more reliable task completion when paired with diverse expressions of the same requirements. Its gains on original instructions, unseen rephrasing types, and human rephrasings on the robot give instruction diversity a role beyond adaptation to a particular test set. We therefore argue that linguistic diversity deserves attention alongside visual and environmental diversity when developing VLN agents. Improving the range of expressions an agent can follow is part of improving its navigation capability.

Scope of the findings. Our experiments establish behavioral dependence on wording; they do not determine how much semantic understanding a model has internally or identify a single cause of that dependence. The diagnosis covers English instructions on two benchmarks, and we fine-tune three of the seven agents, selected because their pipelines support instruction replacement. The effects of the intervention on other agents and languages remain open. The physical replication uses a limited set of indoor routes and human rephrasings of given reference instructions, so its findings concern that controlled setting rather than unrestricted deployment.

7 CONCLUSION

For the VLN agents studied here, successful navigation does not consistently survive semantically equivalent changes in wording. Similar aggregate scores can conceal different episode outcomes, so completing an instruction in one form provides limited evidence of reliable instruction following. This dependence can be reduced by diversifying the language paired with existing navigation supervision, with benefits extending to unseen rephrasing types and human rephrasings on a physical robot. The results make instruction diversity a concern for both evaluation and training: navigation success should persist when people express the same requirements in different words.

AI USE STATEMENT

We used AI tools to assist with writing and polishing the manuscript, literature search and discovery of related work, and generation of synthetic instruction rephrasings for evaluation and fine-tuning. The generation and verification procedures are described in Appendix A. The authors take responsibility for the final manuscript.

REFERENCES

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 3674–3683. IEEE, 2018.

An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems*, 31, 2018.

Wenxuan Guo, Xiuwei Xu, Yichen Liu, Xiangyu Li, Hang Yin, Huangxing Chen, Wenzhao Zheng, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Awarevln: Reasoning with self-awareness for vision-language navigation. *arXiv preprint arXiv:2605.22816*, 2026.

Meera Hahn, Amit Raj, and James M Rehg. Which way is' right'?: Uncovering limitations of vision-and-language navigation models. In *AAMAS*, pp. 2415–2417, 2023.

Wonje Jeung, Sangyeon Yoon, Hyesoo Hong, Yoonjun Cho, Dongjae Jeon, Bumjun Kim, Jean Oh, Youngjae Yu, and Albert No. Same trajectory, contradictory rewards (ROBORMBENCH): Paraphrase fragility in vision language reward models. *arXiv preprint arXiv:2609.05401*, 2026. URL https://arxiv.org/abs/2609.05401.

Huaide Jiang, Yash Chaudhary, Yuping Wang, Zehao Wang, Raghav Sharma, Manan Mehta, Yang Zhou, Lichao Sun, Zhiwen Fan, Zhengzhong Tu, et al. Navtrust: Benchmarking trustworthiness for embodied navigation. *arXiv preprint arXiv:2603.19229*, 2026.

Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10813–10823. IEEE, 2023.

Chanyoung Kim, Minwoo Kim, Minseok Kang, Hyunwoo Kim, and Dahuin Jung. Libero-para: A diagnostic benchmark and metrics for paraphrase robustness in vla models. *arXiv preprint arXiv:2603.28301*, 2026.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pp. 104–120. Springer, 2020.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4392–4412, 2020.

Bingqian Lin, Yi Zhu, Yanxin Long, Xiaodan Liang, Qixiang Ye, and Liang Lin. Adversarial reinforced instruction attacker for robust vision-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7175–7189, 2021.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Jiuding Sun, Chantal Shaib, and Byron Wallace. Evaluating the zero-shot robustness of instruction-tuned language models. In *International conference on learning representations*, volume 2024, pp. 48103–48141, 2024.

Francesco Taioli, Stefano Rosa, Alberto Castellini, Lorenzo Natale, Alessio Del Bue, Alessandro Farinelli, Marco Cristani, and Yiming Wang. Mind the error! detection and localization of instruction errors in vision-and-language navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 12993–13000. IEEE, 2024.

Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2610–2621, 2019.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025.

Meng Wei, Chenyang Wan, Peng Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, et al. Ground slow, move fast: A dual-system foundation model for generalizable vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 12380–12396, 2026.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 33001–33026, 2026.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224*, 2024a.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024b.

Yubo Zhang, Hao Tan, and Mohit Bansal. Diagnosing the environment bias in vision-and-language navigation. *arXiv preprint arXiv:2005.03086*, 2020.

Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. In *Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 5981–5993, 2022.

A Rephrasing Generation

A.1 Rephrasing Types

The fourteen types in Table 2 implement the four dimensions introduced in §3.1. Every type must preserve the action sequence, landmark referents and identifying attributes, spatial relations, and stopping conditions. Within navigation expressions, three types target action, landmark, and spatial-relation expressions separately, while a fourth combines action and landmark rephrasing. Instruction organization and language style operate on the description as a whole while preserving these same requirements.

On R2R-CE, verification excludes 297 type–episode pairs and retains 25,449 variants across the fourteen types. All 1,839 original instructions are retained; excluding an episode from one rephrasing type does not exclude its variants of other types.

Table 2: Rephrasing types and retained variants on the two val-unseen splits. Counts report variants retained after verification, with exclusions applied separately to each type.

Dimension	Type	R2R-CE	RxR-CE
Text formatting	Capitalization	1,839	1,223
	Punctuation	1,839	1,214
	Whitespace	1,839	1,091
Navigation expressions	Action rephrasing	1,839	1,194
	Landmark rephrasing	1,832	1,183
	Spatial-relation rephrasing	1,629	1,126
	Action and landmark rephrasing	1,832	1,164
Instruction organization	Numbered steps	1,839	1,223
	Line breaks	1,836	1,223
	Shorter clauses	1,813	1,223
Language style	Formal	1,821	1,223
	Casual	1,818	1,223
	Concise	1,834	1,223
	Verbose	1,839	1,223

A.2 Generation and Verification

The three text formatting types use deterministic character edits, and the other eleven types use GPT-5.6 Sol with type-specific prompts and the thinking level set to `high`. For navigation expressions, we first extract and review lists of action, landmark, and spatial-relation expressions from the source instructions. We match each instruction against these lists and select the expressions targeted by the requested type. The generation prompt limits changes to those expressions, permitting only the local grammatical adjustments needed to keep the rewritten instruction well formed.

Automatic checks validate rule-based edits and screen generated rewrites for unchanged or malformed text, direction changes, and incomplete spatial expressions. The LLM verifier also uses GPT-5.6 Sol with the thinking level set to `high` to check whether each candidate preserves the navigation information and follows the requested type. Rejected candidates are regenerated using feedback, with a total budget of $K = 8$ attempts per type–episode pair, including the initial attempt. All variants that pass automated verification undergo manual review for semantic preservation and adherence to the requested type. Any issues identified during this review are corrected manually before the variants are included in the evaluation and fine-tuning sets. For each type, the coverage report distinguishes inapplicable instructions from candidates rejected during generation or verification.

Algorithm 1 summarizes the generation and verification procedure. Target-expression selection applies only to navigation expressions; instruction organization and language style use an empty target set.

Algorithm 1 Controlled instruction rephrasing. Each type–episode pair receives at most one verified variant; pairs without an accepted candidate are excluded for that type only.

Require: Episodes $\mathcal{E} = \{(i, x_i)\}_{i=1}^{N}$; rephrasing types $\mathcal{T} = \mathcal{T}_{\text{rule}} \cup \mathcal{T}_{\text{LLM}}$ with navigation-expression subset $\mathcal{T}_{\text{nav}} \subseteq \mathcal{T}_{\text{LLM}}$; generator G; verifier V; total attempt budget $K = 8$; seed s
Ensure: Verified variant sets $\{\mathcal{V}_t\}_{t\in\mathcal{T}}$

$\mathcal{L} \leftarrow$ ExtractExpressions$(\mathcal{E}; G)$ ▷ action, landmark and spatial-relation lists
$\mathcal{L} \leftarrow$ ReviewExpressions$(\mathcal{L}, \mathcal{E}; V)$
for each type $t \in \mathcal{T}$ **do**
  $\mathcal{V}_t \leftarrow \emptyset$
  for each episode $(i, x_i) \in \mathcal{E}$ **do**
    $F \leftarrow \emptyset$ ▷ feedback from rejected candidates
    for $k = 1, \ldots, K$ **do**
      if $t \in \mathcal{T}_{\text{rule}}$ **then**
        $\mathcal{R} \leftarrow \{r \in$ Edits$(t) : r(x_i) \neq x_i\}$
        if $\mathcal{R} = \emptyset$ **then**
          break ▷ type inapplicable to x_i
        end if
        $r \sim$ Uniform$(\mathcal{R};\ s, i, t, k)$; $\tilde{x} \leftarrow r(x_i)$
      else
        $S \leftarrow \emptyset$
        if $t \in \mathcal{T}_{\text{nav}}$ **then**
          $S \leftarrow$ SelectTargets$\big($Match$(x_i, \mathcal{L}), t;\ s, i, k\big)$
          if S lacks a target required by t **then**
            break ▷ type inapplicable to x_i
          end if
        end if
        $\tilde{x} \leftarrow G(\pi_t, x_i, S, F)$ ▷ type-specific prompt π_t
      end if
      $(ok, F) \leftarrow$ RuleChecks$(x_i, \tilde{x}, t)$
      if ok **and** $t \in \mathcal{T}_{\text{LLM}}$ **then**
        $(ok, F) \leftarrow V(\rho_t, x_i, \tilde{x}, S)$ ▷ type-specific criteria ρ_t
      end if
      if ok **then**
        $\mathcal{V}_t \leftarrow \mathcal{V}_t \cup \{(i, \tilde{x})\}$; **break**
      end if
    end for
  end for
  $\mathcal{V}_t \leftarrow$ ManualReviewAndRepair$(\mathcal{V}_t, \mathcal{E}, t)$ ▷ review all variants; repair violations
end for
return $\{\mathcal{V}_t\}_{t\in\mathcal{T}}$

B Full Results

B.1 Additional Navigation Metrics

Alongside SR, we report oracle success rate (OSR), which measures whether the agent enters the success region at any point, and navigation error (NE), the mean geodesic distance in meters between its final position and the goal. Success weighted by path length (SPL) characterizes path efficiency, while normalized dynamic time warping (nDTW) measures agreement with the reference path. Table 1 gives these metrics on the original instructions together with RSR and FR across valid phrasings, as defined in §3.1.

For nonfinite metric values, we average only the finite records of that metric while retaining episodes in the success and pairing denominators. In the comparisons in §3.2, mean absolute changes are averaged equally across seven released agents and fourteen types. The corresponding mean absolute SPL changes are 0.93 and 1.16 percentage points on R2R-CE and RxR-CE, respectively; for nDTW, they are 0.53 and 0.88 points. The fraction of original successes that remain reliable is computed by pooling reliable-success counts and original-success counts across agents within each benchmark.

B.2 PER-TYPE RESULTS

Tables 3 and 4 give all seven metrics for each rephrasing type, model, and checkpoint. SR, NE, SPL, nDTW, and OSR use the indicated variant. For these tables, RSR requires success under the original instruction and that single variant, and FR counts reversals within that pair; this per-type RSR differs from the all-variant RSR in Table 1. Rephrasing types follow the four dimensions and ordering in Table 2, which reports their retained variant counts.

Table 3: Per-type results on R2R-CE. All values except NE (meters) are percentages. RSR here requires success under the original instruction and the single indicated variant.

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
AwareVLN — Released							
Capitalization	64.44	49.76	27.95	4.09	56.20	65.20	72.59
Punctuation	63.84	50.63	25.61	4.04	55.42	65.07	73.63
Whitespace	62.48	49.97	25.56	4.18	54.29	64.48	73.46
Action rephrasing	63.40	50.24	25.94	4.10	54.98	64.84	72.97
Landmark rephrasing	62.50	49.29	26.97	4.26	54.38	64.44	72.33
Spatial-relation rephrasing	64.03	49.79	27.56	4.05	55.09	64.49	73.24
Action and landmark rephrasing	63.86	50.55	25.87	4.06	54.69	64.64	73.74
Numbered steps	63.78	49.76	27.30	4.02	56.25	66.33	72.76
Line breaks	64.87	50.33	27.29	3.93	56.82	65.80	73.75
Shorter clauses	61.61	48.81	26.97	4.32	53.31	63.25	72.75
Formal	63.59	48.82	28.94	4.11	54.66	64.45	73.09
Casual	63.42	49.39	27.67	4.13	54.44	64.18	73.10
Concise	62.16	49.07	26.99	4.23	53.32	63.57	72.03
Verbose	64.65	49.76	28.17	4.02	55.75	65.05	74.33
DualVLN — Released							
Capitalization	61.66	53.78	18.16	4.29	55.75	68.77	68.68
Punctuation	63.68	56.28	15.17	4.11	57.69	69.67	70.36
Whitespace	62.37	55.25	15.93	4.18	56.59	69.26	69.71
Action rephrasing	63.02	56.06	14.95	4.09	57.34	69.95	69.17
Landmark rephrasing	62.66	54.59	17.52	4.19	56.85	69.58	69.10
Spatial-relation rephrasing	63.41	56.60	14.24	4.14	57.68	69.83	69.18
Action and landmark rephrasing	64.25	55.90	16.48	4.11	58.22	69.54	70.74
Numbered steps	64.93	55.68	17.62	4.04	58.97	70.68	71.02
Line breaks	63.51	55.94	15.63	4.12	57.71	69.87	69.88
Shorter clauses	64.75	55.54	17.65	4.05	58.37	69.33	72.04
Formal	63.76	56.01	15.76	4.18	58.08	69.94	69.30
Casual	63.70	55.50	16.67	4.21	57.77	69.56	70.19
Concise	63.96	54.74	18.48	4.12	58.41	70.04	69.57
Verbose	63.73	55.52	16.75	4.18	58.20	69.68	70.96
JanusVLN — Released							
Capitalization	57.21	51.06	14.14	5.04	53.65	64.81	61.99
Punctuation	58.18	53.51	10.22	5.12	54.48	65.43	62.70
Whitespace	60.74	51.11	17.56	4.78	56.41	66.51	66.78
Action rephrasing	58.89	54.05	9.84	4.93	55.07	65.57	64.00
Landmark rephrasing	59.39	52.89	12.61	4.79	55.46	66.05	64.57
Spatial-relation rephrasing	58.99	55.25	7.55	4.88	55.14	65.90	63.29
Action and landmark rephrasing	58.30	51.58	14.14	4.97	54.68	65.60	63.76
Numbered steps	58.08	50.52	16.10	5.07	54.07	65.24	63.02
Line breaks	60.13	54.47	10.24	4.86	56.02	66.00	64.87
Shorter clauses	58.80	52.34	13.07	5.00	54.78	65.34	63.65
Formal	59.14	52.00	14.22	4.95	55.54	65.91	64.20
Casual	60.62	52.70	14.30	4.73	56.40	66.49	64.47
Concise	58.12	49.56	17.99	4.98	54.64	65.72	62.98
Verbose	59.76	51.50	15.82	4.87	55.79	66.38	64.33

Continued on next page

R2R-CE per-type results (continued)

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
StreamVLN — Released							
Capitalization	55.25	47.91	13.87	4.96	48.42	63.33	61.94
Punctuation	55.90	47.53	15.28	4.91	49.20	63.99	63.68
Whitespace	57.69	48.40	15.33	4.81	50.71	64.13	63.95
Action rephrasing	54.38	50.57	7.67	5.05	47.94	63.22	61.45
Landmark rephrasing	53.88	49.56	9.22	5.07	47.83	63.13	61.35
Spatial-relation rephrasing	54.57	51.69	5.52	5.05	48.19	63.12	61.39
Action and landmark rephrasing	55.40	49.34	11.19	4.91	49.03	63.46	62.28
Numbered steps	55.03	46.98	15.50	5.02	48.90	63.35	61.56
Line breaks	53.87	50.22	7.84	5.03	47.89	63.46	61.22
Shorter clauses	56.32	49.53	11.69	4.90	49.60	63.87	63.43
Formal	54.91	49.09	11.15	4.99	48.93	63.69	61.83
Casual	55.01	49.06	11.33	5.02	48.74	63.38	61.17
Concise	54.31	47.11	14.56	5.10	48.16	63.21	60.91
Verbose	56.33	49.43	11.91	4.93	50.00	63.83	62.75
NaVILA — Released							
Capitalization	52.58	46.55	13.59	5.37	48.22	63.69	60.58
Punctuation	46.38	38.99	22.51	5.74	41.78	59.25	58.13
Whitespace	45.08	35.18	28.82	5.84	40.47	58.56	53.29
Action rephrasing	54.65	49.48	9.79	5.25	49.97	64.22	62.43
Landmark rephrasing	52.89	48.25	10.53	5.36	48.58	63.99	60.43
Spatial-relation rephrasing	53.53	50.28	7.00	5.27	48.99	64.31	61.02
Action and landmark rephrasing	52.40	46.56	13.43	5.46	48.16	63.21	60.48
Numbered steps	52.31	46.11	14.19	5.32	48.22	63.71	61.07
Line breaks	53.65	48.04	11.76	5.25	49.15	63.94	61.60
Shorter clauses	53.61	46.55	14.62	5.39	48.77	63.10	62.00
Formal	52.11	46.35	13.56	5.53	47.63	62.55	60.35
Casual	52.64	46.70	13.31	5.29	48.26	63.59	61.00
Concise	51.69	43.78	18.27	5.47	47.53	62.84	59.32
Verbose	53.78	46.55	14.79	5.37	49.20	63.93	61.50
Uni-NaVid — Released							
Capitalization	51.77	44.43	14.95	5.15	47.97	64.55	56.77
Punctuation	51.93	46.98	10.01	5.21	47.84	64.10	57.10
Whitespace	52.47	49.65	5.22	5.04	48.50	64.93	57.59
Action rephrasing	52.04	47.53	9.03	5.13	48.17	64.36	57.75
Landmark rephrasing	53.17	46.34	12.55	5.04	49.31	65.04	57.81
Spatial-relation rephrasing	52.55	48.80	6.88	5.02	48.51	65.16	57.52
Action and landmark rephrasing	52.57	44.98	14.68	5.10	48.57	64.53	57.31
Numbered steps	51.60	44.26	15.12	5.05	47.72	64.22	57.37
Line breaks	51.80	45.81	12.25	5.12	48.04	64.39	56.75
Shorter clauses	51.90	45.34	13.18	5.10	47.97	64.26	57.75
Formal	53.16	46.24	12.74	5.03	48.96	64.73	57.83
Casual	51.87	44.88	14.19	5.13	47.91	64.19	56.93
Concise	50.33	42.15	18.05	5.27	46.35	63.56	55.18
Verbose	51.66	45.08	13.54	5.07	47.54	63.98	56.77

Continued on next page

R2R-CE per-type results (continued)

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
NaVid — Released							
Capitalization	41.16	32.03	18.16	5.72	36.33	58.32	48.89
Punctuation	41.22	34.37	13.54	5.69	36.66	58.51	48.72
Whitespace	41.38	36.38	9.68	5.64	36.65	58.61	48.78
Action rephrasing	41.71	34.91	12.94	5.70	36.70	58.44	48.61
Landmark rephrasing	40.12	32.26	16.70	5.82	35.55	58.02	47.87
Spatial-relation rephrasing	42.11	36.53	10.07	5.62	37.28	58.89	48.86
Action and landmark rephrasing	41.48	31.39	19.81	5.70	36.76	58.54	48.85
Numbered steps	41.05	32.41	17.29	5.58	36.04	58.30	47.80
Line breaks	40.03	32.84	15.36	5.70	35.44	58.41	47.33
Shorter clauses	39.99	32.76	15.55	5.78	35.21	57.44	47.82
Formal	41.24	31.96	18.40	5.63	36.42	58.94	49.20
Casual	39.93	31.41	18.21	5.72	35.57	58.63	47.19
Concise	37.24	27.97	22.30	5.94	33.41	57.22	44.33
Verbose	39.70	31.97	16.80	5.74	35.34	58.66	47.31
NaVILA — Rephrasing-trained							
Capitalization	55.14	48.78	13.54	5.34	51.44	64.59	61.01
Punctuation	56.23	50.19	11.80	5.27	52.00	64.74	61.83
Whitespace	56.23	49.37	13.43	5.20	52.27	64.98	62.26
Action rephrasing	56.01	50.73	10.49	5.22	52.28	64.68	61.28
Landmark rephrasing	56.55	51.15	10.15	5.27	52.89	64.99	61.63
Spatial-relation rephrasing	55.43	52.06	7.24	5.33	51.85	64.48	60.47
Action and landmark rephrasing	55.35	48.42	14.52	5.34	51.53	64.39	60.81
Numbered steps	56.50	48.45	15.55	5.18	52.76	65.25	61.77
Line breaks	55.56	50.71	10.19	5.24	52.24	65.17	60.89
Shorter clauses	55.54	49.15	13.18	5.24	51.58	64.37	61.56
Formal	56.67	49.04	14.50	5.23	52.80	64.84	61.94
Casual	56.93	49.23	14.47	5.07	52.91	65.17	62.43
Concise	54.25	46.84	16.58	5.35	50.93	64.08	59.00
Verbose	56.23	49.21	13.76	5.07	52.14	65.08	62.04
AwareVLN — Rephrasing-trained							
Capitalization	62.70	57.31	13.16	3.99	56.86	68.21	69.39
Punctuation	64.98	59.43	11.20	3.79	58.82	69.49	72.05
Whitespace	64.60	59.27	11.15	3.86	58.53	69.00	70.85
Action rephrasing	63.19	58.35	11.58	3.91	57.12	68.63	70.53
Landmark rephrasing	64.19	58.79	11.74	3.97	58.26	68.98	70.36
Spatial-relation rephrasing	63.97	60.04	8.96	3.84	57.92	69.41	70.84
Action and landmark rephrasing	63.16	56.77	14.68	4.04	57.44	68.38	70.36
Numbered steps	64.44	57.86	13.81	3.78	58.39	69.59	71.18
Line breaks	65.03	59.42	11.27	3.85	59.02	69.41	71.13
Shorter clauses	65.53	57.92	14.78	3.85	58.90	68.74	73.58
Formal	64.74	58.76	12.30	3.86	58.43	69.04	71.66
Casual	65.51	58.09	14.47	3.77	59.06	69.42	72.11
Concise	61.23	53.93	18.43	4.12	55.41	67.73	69.08
Verbose	64.17	57.04	15.17	3.88	58.50	69.50	71.29

Continued on next page

R2R-CE per-type results (continued)

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
Uni-NaVid — Rephrasing-trained							
Capitalization	55.79	48.83	14.41	4.98	49.21	64.90	60.74
Punctuation	56.28	50.52	11.53	5.09	49.18	64.83	61.66
Whitespace	56.88	53.62	5.93	4.91	49.73	65.32	61.99
Action rephrasing	56.39	51.88	8.92	5.04	49.34	64.73	61.28
Landmark rephrasing	56.55	50.60	11.57	4.92	49.48	65.10	61.90
Spatial-relation rephrasing	56.91	52.79	7.61	4.97	49.93	65.20	61.94
Action and landmark rephrasing	56.28	49.45	13.59	4.95	49.36	65.13	61.14
Numbered steps	56.39	50.35	11.96	5.13	49.48	64.57	61.39
Line breaks	55.28	50.76	10.02	5.15	48.54	64.23	60.29
Shorter clauses	56.54	50.19	12.47	5.03	49.06	64.34	62.33
Formal	55.96	50.19	11.86	5.01	49.00	64.66	60.90
Casual	55.94	49.56	13.15	5.02	48.92	64.46	60.18
Concise	53.93	46.46	17.34	5.15	46.96	63.91	59.00
Verbose	55.25	49.43	12.67	5.03	48.52	64.43	60.69

Table 4: Per-type results on RxR-CE. All values except NE (meters) are percentages. RSR here requires success under the original instruction and the single indicated variant.

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
AwareVLN — Released							
Capitalization	64.19	54.95	19.62	4.55	54.58	65.75	72.20
Punctuation	65.90	59.14	13.01	4.37	56.07	66.12	73.89
Whitespace	65.44	57.93	14.76	4.30	55.78	66.44	73.42
Action rephrasing	67.17	57.37	17.67	4.41	56.63	65.58	73.70
Landmark rephrasing	66.10	57.65	16.15	4.37	55.46	66.03	73.54
Spatial-relation rephrasing	65.45	59.50	11.81	4.59	55.75	65.88	72.38
Action and landmark rephrasing	66.24	56.44	18.73	4.30	56.05	66.29	74.23
Numbered steps	65.66	55.11	20.77	4.38	55.24	66.14	72.53
Line breaks	65.66	57.15	16.68	4.45	56.17	66.79	72.44
Shorter clauses	66.64	54.87	22.24	4.34	55.92	66.12	74.57
Formal	63.12	53.07	22.32	4.55	53.27	64.31	70.89
Casual	63.45	52.33	24.12	4.40	53.35	65.11	70.73
Concise	59.28	47.42	29.76	5.18	48.86	60.60	69.01
Verbose	64.51	54.05	21.75	4.46	54.65	65.08	70.89
DualVLN — Released							
Capitalization	62.06	53.88	20.20	4.90	53.65	67.95	72.53
Punctuation	67.79	62.19	9.23	4.55	58.28	70.63	74.88
Whitespace	65.44	60.86	9.62	4.57	56.94	69.41	76.35
Action rephrasing	63.90	56.78	16.25	4.66	54.37	68.97	72.95
Landmark rephrasing	64.75	55.87	18.93	4.69	54.77	67.75	74.89
Spatial-relation rephrasing	65.72	57.82	15.99	4.64	57.01	68.12	73.27
Action and landmark rephrasing	65.72	57.56	16.49	4.54	56.25	67.90	76.37
Numbered steps	65.90	56.17	19.46	4.74	57.19	68.89	74.49
Line breaks	66.88	56.58	19.62	4.62	57.30	68.33	75.31
Shorter clauses	65.82	54.95	21.83	4.55	56.01	68.61	76.45
Formal	64.84	52.66	25.43	4.79	55.44	67.54	72.12
Casual	64.19	51.59	26.90	4.83	55.17	67.97	75.06
Concise	62.47	49.55	29.27	5.00	54.03	66.47	70.07
Verbose	65.90	53.80	24.20	4.73	56.12	68.95	75.14
JanusVLN — Released							
Capitalization	58.30	51.59	16.27	5.74	51.62	64.42	65.41
Punctuation	61.12	58.40	5.44	5.69	54.85	65.52	66.72
Whitespace	63.24	56.00	12.47	5.32	55.43	66.15	71.59
Action rephrasing	60.64	55.28	11.31	5.61	53.37	64.23	68.84
Landmark rephrasing	61.28	53.68	15.22	5.29	53.49	65.63	69.40
Spatial-relation rephrasing	61.28	56.22	9.86	5.38	53.97	64.55	68.56
Action and landmark rephrasing	60.31	53.52	14.35	5.50	53.09	64.20	69.24
Numbered steps	59.36	50.94	18.64	5.89	52.53	63.40	66.39
Line breaks	64.35	55.76	13.98	5.36	54.92	64.53	70.48
Shorter clauses	60.43	52.17	17.25	5.58	53.62	64.25	67.62
Formal	59.85	48.32	24.37	5.64	52.40	64.29	68.85
Casual	61.90	49.30	24.45	5.41	54.13	66.06	69.26
Concise	56.50	44.40	28.86	6.01	50.21	62.84	63.61
Verbose	62.14	49.47	24.37	5.48	53.46	65.38	69.26

Continued on next page

RxR-CE per-type results (continued)

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
StreamVLN — Released							
Capitalization	55.76	48.65	14.80	5.56	47.50	63.90	64.92
Punctuation	55.93	50.33	11.53	5.53	47.32	63.64	65.16
Whitespace	56.46	51.05	10.72	5.47	48.21	63.76	65.26
Action rephrasing	54.61	49.92	11.22	5.77	46.03	62.92	64.49
Landmark rephrasing	55.71	49.79	12.43	5.45	47.29	64.09	65.26
Spatial-relation rephrasing	56.04	52.04	8.35	5.35	47.83	63.95	65.54
Action and landmark rephrasing	55.84	49.91	12.46	5.49	47.38	63.89	65.12
Numbered steps	56.09	47.26	17.91	5.67	47.27	63.04	65.49
Line breaks	57.15	50.20	13.08	5.37	48.38	64.07	65.25
Shorter clauses	56.75	48.41	16.27	5.42	47.88	64.45	66.97
Formal	54.21	43.91	22.73	5.61	45.99	63.31	64.19
Casual	54.29	44.40	21.83	5.77	45.97	63.49	64.43
Concise	50.12	39.57	27.31	6.31	41.95	60.12	59.69
Verbose	56.17	45.87	20.77	5.50	47.46	64.14	65.49
NaVILA — Released							
Capitalization	51.02	43.50	16.68	6.12	45.03	61.37	59.20
Punctuation	51.98	47.36	9.97	6.20	46.25	61.75	61.20
Whitespace	53.25	48.21	9.44	5.98	47.18	61.99	62.51
Action rephrasing	52.68	48.16	8.96	6.07	46.70	61.71	61.47
Landmark rephrasing	51.23	46.41	11.07	6.22	46.18	61.44	60.86
Spatial-relation rephrasing	52.22	47.87	9.06	6.20	46.42	61.35	60.83
Action and landmark rephrasing	52.41	46.31	12.54	6.06	46.68	61.25	60.74
Numbered steps	51.27	43.99	15.94	6.32	45.59	60.80	59.61
Line breaks	51.92	45.38	13.82	6.19	46.25	60.93	60.67
Shorter clauses	51.92	43.74	17.09	6.18	45.95	60.93	61.16
Formal	50.70	40.96	21.42	6.52	44.51	59.46	59.69
Casual	51.10	40.39	22.98	6.24	45.05	60.29	60.59
Concise	46.85	36.22	27.06	7.05	41.25	58.11	56.26
Verbose	52.17	41.78	21.26	6.19	46.09	60.87	60.75
Uni-NaVid — Released							
Capitalization	55.76	48.41	17.09	6.03	45.98	61.35	65.66
Punctuation	58.90	54.45	8.07	5.77	47.85	62.63	68.04
Whitespace	57.38	53.16	9.26	5.66	46.37	61.68	67.00
Action rephrasing	57.87	53.27	9.55	5.75	46.82	62.26	67.09
Landmark rephrasing	58.50	51.56	13.52	5.69	47.27	62.05	68.89
Spatial-relation rephrasing	58.79	54.88	7.19	5.73	47.43	62.02	68.21
Action and landmark rephrasing	58.85	50.86	15.29	5.54	47.52	62.40	69.24
Numbered steps	55.03	47.67	17.83	6.09	44.96	60.64	65.09
Line breaks	58.79	50.70	15.54	5.77	47.50	61.78	68.68
Shorter clauses	58.22	49.88	16.60	5.71	47.08	61.74	67.54
Formal	57.40	46.20	23.14	5.75	47.51	61.76	67.13
Casual	57.81	46.52	22.89	5.73	47.52	62.53	67.21
Concise	54.13	41.78	28.70	5.87	45.13	60.94	63.12
Verbose	58.63	47.51	21.75	5.73	48.16	62.79	68.68

Continued on next page

RxR-CE per-type results (continued)

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
NaVid — Released							
Capitalization	46.03	37.78	18.07	7.10	38.58	54.80	53.80
Punctuation	46.21	39.13	15.49	6.96	38.38	54.60	54.86
Whitespace	47.30	40.70	13.38	6.94	39.66	55.32	55.73
Action rephrasing	48.74	40.45	15.33	6.85	40.20	55.59	56.20
Landmark rephrasing	47.51	39.81	15.47	6.94	40.14	55.89	55.45
Spatial-relation rephrasing	48.05	41.83	11.90	6.60	40.20	55.71	56.75
Action and landmark rephrasing	47.42	38.92	17.18	6.88	39.88	55.50	55.24
Numbered steps	47.83	37.20	21.01	6.99	39.16	54.24	56.83
Line breaks	47.75	37.78	19.79	6.85	39.94	55.17	56.26
Shorter clauses	44.89	37.04	18.40	7.33	37.16	53.36	53.56
Formal	47.26	35.08	24.69	7.07	39.62	54.74	55.27
Casual	46.61	34.42	25.35	7.10	39.27	55.06	55.27
Concise	43.83	31.23	28.95	7.24	35.46	52.92	51.76
Verbose	47.18	35.00	24.78	7.10	39.16	54.63	55.11
NaVILA — Rephrasing-trained							
Capitalization	53.64	46.44	16.11	5.80	48.45	64.48	59.12
Punctuation	54.12	50.58	7.99	5.75	49.17	64.23	61.12
Whitespace	55.18	49.86	10.27	5.67	49.92	64.25	61.14
Action rephrasing	53.77	49.50	9.72	5.82	49.13	63.56	60.47
Landmark rephrasing	54.35	49.03	11.07	5.91	49.02	63.94	60.86
Spatial-relation rephrasing	54.80	50.36	8.97	5.70	49.52	63.98	60.92
Action and landmark rephrasing	53.52	47.25	13.66	5.76	48.75	63.91	60.57
Numbered steps	54.70	47.10	15.86	5.92	49.27	63.93	60.51
Line breaks	55.27	48.65	13.33	5.68	50.01	64.10	61.41
Shorter clauses	55.11	46.85	16.76	5.60	49.39	64.30	62.47
Formal	54.21	42.19	25.18	6.10	48.31	62.55	61.41
Casual	54.95	44.48	21.34	5.77	49.51	63.87	62.96
Concise	51.92	39.98	27.31	6.25	45.93	61.87	59.53
Verbose	56.17	46.36	18.81	5.73	50.23	64.17	62.71
AwareVLN — Rephrasing-trained							
Capitalization	65.99	58.05	18.56	4.20	57.17	68.54	72.94
Punctuation	68.78	62.52	12.36	3.91	59.32	69.77	74.55
Whitespace	68.01	61.69	13.29	4.01	58.87	69.68	74.61
Action rephrasing	67.25	61.22	13.48	4.28	58.43	69.08	74.04
Landmark rephrasing	67.12	59.76	16.31	4.09	58.48	69.37	73.71
Spatial-relation rephrasing	68.74	62.43	12.70	3.85	59.32	69.82	74.78
Action and landmark rephrasing	68.04	60.57	15.55	3.99	59.06	69.67	73.71
Numbered steps	68.44	59.28	18.56	3.96	58.65	69.34	74.98
Line breaks	67.54	60.67	14.88	3.80	59.04	70.97	74.16
Shorter clauses	68.44	59.53	18.07	3.91	58.59	69.72	74.90
Formal	65.99	56.09	22.49	4.12	57.59	69.16	72.77
Casual	65.82	57.07	20.36	4.10	57.29	69.27	72.77
Concise	61.41	52.17	25.76	4.60	53.06	65.92	69.01
Verbose	68.27	58.38	20.20	3.93	59.69	70.38	75.14

Continued on next page

RxR-CE per-type results (continued)

Type	SR	RSR	FR	NE	SPL	nDTW	OSR
Uni-NaVid — Rephrasing-trained							
Capitalization	58.22	50.53	17.09	5.60	46.53	63.91	64.84
Punctuation	59.72	54.86	9.88	5.55	47.46	63.79	67.63
Whitespace	59.40	55.82	7.70	5.41	47.31	62.93	66.27
Action rephrasing	59.05	55.03	8.88	5.62	47.75	63.25	66.58
Landmark rephrasing	58.50	52.49	13.44	5.49	47.11	63.03	68.30
Spatial-relation rephrasing	60.83	56.22	8.26	5.45	47.97	62.90	67.05
Action and landmark rephrasing	59.71	52.92	13.83	5.31	48.59	63.76	66.75
Numbered steps	57.32	51.51	14.23	5.97	46.58	62.91	65.74
Line breaks	58.38	52.25	13.82	5.63	46.93	63.13	66.15
Shorter clauses	59.28	51.19	16.84	5.47	47.00	63.82	68.27
Formal	57.73	47.75	22.16	5.74	46.31	62.71	64.84
Casual	58.79	48.90	20.93	5.62	46.48	63.72	65.58
Concise	54.87	43.42	27.96	5.84	44.60	61.15	62.39
Verbose	58.63	49.39	19.79	5.62	48.78	63.84	67.21

Table 5: Rephrasing types used for fine-tuning and held out for evaluation within each dimension.

Dimension	Used for fine-tuning	Held out
Text formatting	Punctuation; whitespace	Capitalization
Navigation expressions	Action rephrasing; landmark rephrasing	Spatial-relation rephrasing; combined action and landmark rephrasing
Instruction organization	Shorter clauses; line breaks	Numbered steps
Language style	Formal; concise	Casual; verbose

Table 6: Performance before and after rephrasing-only fine-tuning. SR, SPL, and nDTW use the original benchmark instructions; RSR and FR use the original instructions and all valid variants. All values are on a percentage scale.

Benchmark	Model	Checkpoint	SR	RSR	FR	SPL	nDTW
R2R-CE	NaVILA	Released	54.11	21.15	14.79	49.55	64.16
		Rephrasing-trained	55.95	31.32	12.86	52.24	64.91
	AwareVLN	Released	63.02	21.21	27.05	54.85	64.66
		Rephrasing-trained	65.09	38.93	13.08	58.71	69.69
	Uni-NaVid	Released	52.04	28.55	12.35	48.01	64.64
		Rephrasing-trained	56.28	35.02	11.67	49.12	64.82
RxR-CE	NaVILA	Released	52.66	23.22	15.64	46.95	61.84
		Rephrasing-trained	55.36	24.04	15.56	50.12	64.40
	AwareVLN	Released	65.33	28.29	19.34	55.72	66.37
		Rephrasing-trained	68.68	31.56	17.40	59.57	70.61
	Uni-NaVid	Released	58.14	27.88	16.30	47.19	62.32
		Rephrasing-trained	59.93	30.83	15.47	47.41	63.49

C Fine-Tuning Details

C.1 Rephrasing Types

Table 5 lists the fixed split, with two types from each dimension used for fine-tuning and the remaining types held out. This gives eight trained-on types and six held-out types, with every dimension represented in both subsets. The split is fixed before training, and the original training instructions are excluded from fine-tuning.

C.2 Reliability and Original-Instruction Performance

Table 6 reports original-instruction performance together with reliability across the complete set of valid rephrasings. Released and rephrasing-trained checkpoints use the same evaluation episodes, phrasing sets, and inference settings.

C.3 Trained-On and Held-Out Types

Table 7 reports FR changes separately for the eight trained-on types and six held-out types. For each model and subset, we pool the valid original–variant pairs and compare released and rephrasing-trained checkpoints on the same pairs. Mean rows average the three model-specific changes equally.

FR decreases on average in both subsets and benchmarks, with variation across models. NaVILA's FR increases by 0.46 percentage points on held-out types in R2R-CE and by 0.34 points on trained-on types in RxR-CE. AwareVLN and Uni-NaVid reduce FR in both subsets on both benchmarks.

D Transfer to Erroneous Instructions

The main experiments preserve the navigation requirements while varying their expression. Here we examine an additional transfer setting in which part of the instruction is incorrect. This evalua-

Table 7: Changes in FR after rephrasing-only fine-tuning, in percentage points, computed as rephrasing-trained minus released. Negative values indicate fewer outcome reversals. Mean rows average the three model-specific changes.

Benchmark	Model	Trained-on ΔFR	Held-out ΔFR
R2R-CE	NaVILA	-3.71	+0.46
	AwareVLN	-13.98	-13.96
	Uni-NaVid	-0.42	-1.03
	Mean	-6.04	-4.84
RxR-CE	NaVILA	+0.34	-0.64
	AwareVLN	-2.01	-1.84
	Uni-NaVid	-0.45	-1.35
	Mean	-0.70	-1.28

Table 8: SR (%) on R2R-IE-CE val-unseen. Each error column uses the erroneous instructions in its category. Clean reports SR on 1,813 deduplicated control episodes with processed text, which differs from the original instruction strings used in the main R2R-CE evaluation. Released and rephrasing-trained checkpoints share the same evaluation inputs and inference settings.

		Error category					
Model	Checkpoint	Dir.	Room	Obj.	R+O	All	Clean
NaVILA	Released	33.31	35.66	37.25	31.49	27.11	44.02
	Rephrasing-trained	41.14	50.39	49.65	44.98	34.43	56.09
AwareVLN	Released	52.64	56.69	55.53	47.98	36.70	64.98
	Rephrasing-trained	51.34	55.81	59.33	48.56	39.97	65.09
Uni-NaVid	Released	37.66	48.45	48.64	43.02	31.90	52.56
	Rephrasing-trained	38.47	49.32	49.91	43.60	31.53	53.06
Episodes per checkpoint		1,609	1,032	1,581	867	793	1,813

tion asks whether rephrasing-only fine-tuning also benefits task completion under instruction errors, which are absent from the fine-tuning data.

R2R-IE-CE (Taioli et al., 2024) is an instruction-error benchmark derived from the R2R-CE val-unseen split. For each of five error categories—direction, room, object, room+object, and all—it selects eligible instructions and pairs each clean control with an error-injected counterpart. Each pair retains the same environment, starting pose, goal, and reference path, so success under an erroneous instruction is still measured against the original navigation goal.

The *Clean* column reports SR on 1,813 control episodes, deduplicated by original episode ID across the five error categories. These episodes retain the corresponding R2R-CE navigation tasks, while their instruction text is lowercased and spacing is normalized; seven long instructions also have truncated endings. Here, *clean* denotes the absence of injected instruction errors.

For each model, we evaluate released and rephrasing-trained checkpoints on identical clean and erroneous instruction sets, using the same inference settings as in the main experiment.

Rephrasing-only fine-tuning improves SR in 12 of the fifteen model–error-category combinations (Table 8). NaVILA improves across all five categories by 7.31–14.73 percentage points, while AwareVLN and Uni-NaVid improve in three and four categories, respectively. AwareVLN's SR declines by 1.31 points on direction errors and 0.87 points on room errors, and Uni-NaVid's SR declines by 0.38 points in the All category. Clean SR increases by 12.08, 0.11, and 0.50 points for NaVILA, AwareVLN, and Uni-NaVid, respectively. These results show benefits beyond the meaning-preserving rephrasings used for fine-tuning, with transfer varying across models and error categories.

E Real-Robot Evaluation Details

E.1 Robot and Execution Limits

We use a Jueying Lite3 quadruped robot with a camera mounted 1.25 m above the floor, matching the camera height used in our Habitat simulations (Figure 5). The camera image resolution is 512×512 pixels for NaVILA and AwareVLN and 640×480 pixels for Uni-NaVid. Each run is limited to 500 navigation steps, and each request to the model has a 60 s timeout for returning an action. A run is terminated and counted as a failure if it does not complete successfully within the step limit or if an action request times out.

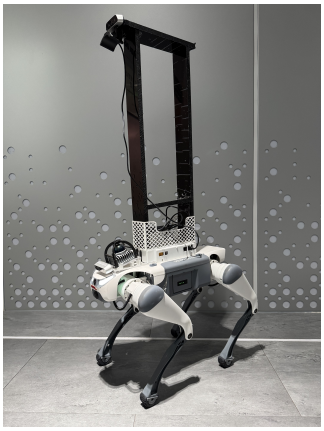


Figure 5: Jueying Lite3 quadruped and camera setup used for real-world evaluation. The camera is mounted 1.25 m above the floor.

E.2 Routes and Instructions

The physical evaluation uses 10 routes in a single-floor indoor environment. Routes span 5–15 m and include turns and landmarks drawn from vocabulary shared with R2R, such as doors, tables, chairs, sofas, and hallways. Each route has a reference instruction in R2R annotation style and four rephrasings by project-external writers. A third party checks the descriptions against a floor plan to verify that they preserve the intended route. The complete layouts and all fifty English instructions appear in §E.4.

E.3 Repeated Execution

Every route–phrasing–checkpoint combination is executed three times under the protocol in §5.1. For each repetition, we group the outcomes by route across its reference instruction and four rephrasings, compute SR, RSR, and FR, and then average the metrics over repetitions. This keeps the reliability criterion tied to the five tested descriptions of each route.

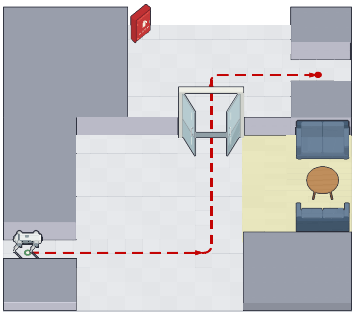
Table 9: Real-world performance under reference instructions and human rephrasings on 10 routes. Metrics are averaged over three repetitions and reported as percentages. SR measures success across all five phrasings. RSR requires success under all five, while FR compares each human rephrasing with its reference instruction.

Model	Checkpoint	SR	RSR	FR
NaVILA	Released	48.7	23.3	28.3
	Rephrasing-trained	54.7	33.3	19.2
AwareVLN	Released	66.7	43.3	18.3
	Rephrasing-trained	70.7	56.7	11.7
Uni-NaVid	Released	22.7	6.7	25.8
	Rephrasing-trained	30.7	13.3	21.7

E.4 Route Layouts and Instructions

The following panels pair all ten route layouts with their reference instructions (Ref.) and four rephrasings (R1–R4). Red dashed lines show the intended routes, and robot icons mark the starts; the layouts are schematic and not to scale. Numbered steps in R2 belong to a single instruction.

Route 01 Corridor and side room



Ref. Walk out of the passage and turn left into the open corridor. Go straight through the glass doorway. Turn right into the room on your right and stop just inside the entrance.

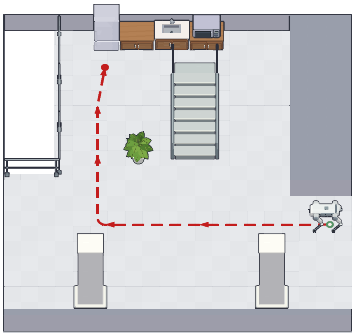
R1 Go out of the passage and turn left into the open corridor. Go straight through the glass doorway. Turn right into the room to your right and stop just inside the entrance.

R2 1. Come out of the passage and take a left into the open corridor. 2. Head straight through the glass doorway, then take a right into the room to your right. 3. Stop just inside the entrance.

R3 Exit the passage and turn left into the open corridor. Go straight through the glass doorway, then right into the room to your right. Stop just inside.

R4 First, exit the passage and turn left into the open corridor. Then, proceed straight through the glass doorway. Finally, turn right into the room to your right and stop just inside the entrance.

Route 02 Plant and refrigerator



Ref. Walk forward along the passage, keeping the plant on your right as you pass it. Once you are past the plant, turn right and walk between the plant and the railing toward the refrigerator. Stop in front of it.

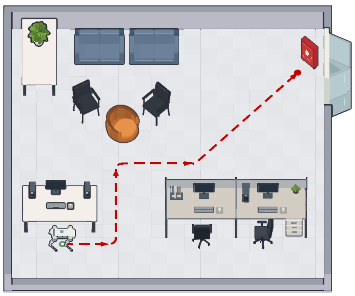
R1 Walk forward along the passage, keeping the plant on your right as you pass it. Once you are past the plant, turn right and walk through the gap separating the plant and the railing toward the fridge. Stop in front of it.

R2 1. Head along the passage and go past the plant, keeping it to your right. 2. Once you have passed it, take a right and head between the plant and the railing toward the fridge. 3. Stop in front of it.

R3 Follow the passage past the plant, keeping it to your right. Turn right after it and take the gap between the plant and railing toward the fridge. Stop in front of the fridge.

R4 First, proceed forward along the passage, keeping the plant on your right as you pass it. Then, after you have passed the plant, turn right and proceed through the gap separating the plant and the railing toward the refrigerator. Finally, stop in front of it.

Route 03 Tables, sofa, and fire cabinet



Ref. Walk forward beside the nearest white table and turn left around its end, keeping the table on your left. Walk toward the brown sofa and turn right before reaching it. Continue past the sofa, then bear left toward the red fire equipment cabinet. Stop in front of the cabinet.

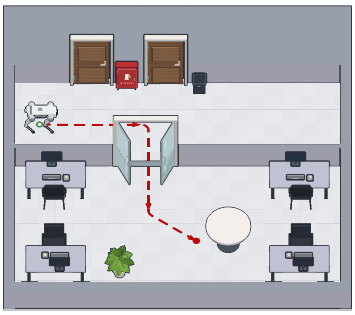
R1 Walk forward beside the nearest white table and turn left around its end, keeping the table to your left. Walk toward the brown couch and turn right before reaching it. Continue past the couch, then bear left toward the red fire equipment cabinet. Stop in front of the cabinet.

R2 1. Go forward alongside the nearest white table, then take a left around its end and keep the table to your left. 2. Head toward the brown couch and take a right before you reach it. 3. Go past the couch, then angle left toward the red fire equipment cabinet and stop in front of it.

R3 Follow the nearest white table and turn left around its end, keeping it to your left. Head toward the brown couch, turning right before it. Pass the couch, then angle left toward the red fire equipment cabinet. Stop in front of the cabinet.

R4 First, proceed forward beside the nearest white table and turn left around its end, keeping the table to your left. Then, proceed toward the brown sofa and turn right before reaching it. Next, continue past the sofa, then angle left toward the red fire equipment cabinet. Finally, stop in front of the cabinet.

Route 04 Corridor and round table



Ref. Walk along the corridor past the first closed door on your left. Turn right through the glass doorway. Bear left toward the white round table. Stop beside the table on the side nearest the doorway you entered through.

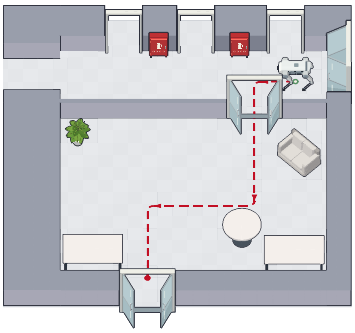
R1 Go along the corridor past the first closed door to your left. Turn right through the glass doorway. Bear left toward the white round table. Stop beside the table on the side nearest the doorway you entered through.

R2 1. Head along the corridor and go past the first closed door to your left. 2. Take a right through the glass doorway, then angle left toward the white round table. 3. Stop beside the table on the side closest to the doorway you came through.

R3 Follow the corridor past the first closed door to your left. Turn right through the glass doorway, then angle left toward the white round table. Stop beside it on the side closest to your entry doorway.

R4 First, proceed along the corridor past the first closed door to your left. Then, turn right through the glass doorway. Next, angle left toward the white round table. Finally, stop beside the table on the side nearest the doorway you entered through.

Route 05 Passage and interior doorway



Ref. Walk forward along the passage. Pass the first doorway on your right, then turn left through the glass doorway into the room. Walk past the sofa on your left toward the white round table. Turn right before reaching the table and continue until you are level with the other glass doorway on your left. Turn left toward that doorway and stop just before it, remaining inside the room.

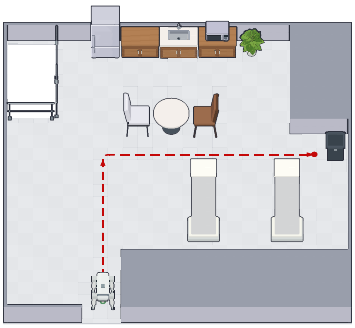
R1 Walk forward along the passage. Pass the first doorway to your right, then turn left through the glass doorway into the room. Walk past the couch on your left toward the white round table. Turn right before reaching the table and continue until you are level with the other glass doorway on your left. Turn left toward that doorway and stop just before it, remaining inside the room.

R2 1. Head forward along the passage. 2. Go past the first doorway to your right, then take a left through the glass doorway into the room. 3. Head past the couch to your left toward the white round table. 4. Take a right before you reach the table and keep going until you are level with the other glass doorway to your left. 5. Take a left toward it and stop just before the doorway, still inside the room.

R3 Follow the passage past the first doorway to your right, then turn left through the glass doorway. Pass the couch to your left toward the white round table. Turn right before the table and continue until level with the other glass doorway to your left. Turn left toward it and stop just before it, inside the room.

R4 First, proceed forward along the passage. Then, pass the first doorway to your right and turn left through the glass doorway into the room. Next, proceed past the sofa on your left toward the white round table. Next, turn right before reaching the table and continue until you are level with the other glass doorway on your left. Finally, turn left toward that doorway and stop just before it, remaining inside the room.

Route 06 Round table, pillars, and bin



Ref. Walk forward toward the round table with a white chair and a brown chair beside it. Turn right before reaching the table. Continue with the two white pillars on your right. Stop in front of the black trash bin.

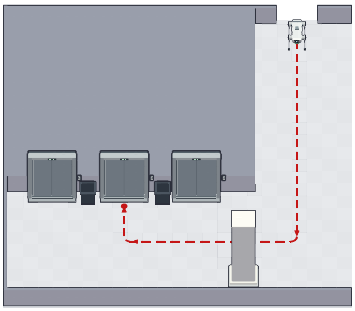
R1 Walk forward toward the round table with a white chair and a brown chair beside it. Turn right before reaching the table. Continue with the two white pillars to your right. Stop in front of the black garbage bin.

R2 1. Head toward the round table with a white chair and a brown chair beside it. 2. Take a right before you get to the table, then keep going with the two white pillars to your right. 3. Stop in front of the black garbage bin.

R3 Approach the round table beside the white and brown chairs. Turn right before it, then continue with both white pillars to your right. Stop in front of the black garbage bin.

R4 First, proceed forward toward the round table with a white chair and a brown chair beside it. Then, turn right before reaching the table. Next, continue with the two white pillars to your right. Finally, stop in front of the black trash bin.

Route 07 Elevator lobby



Ref. Walk along the passage into the elevator lobby. Turn right and continue with the white pillar on your left. Pass the first elevator on your right, then turn right toward the second elevator. Stop in front of its doors.

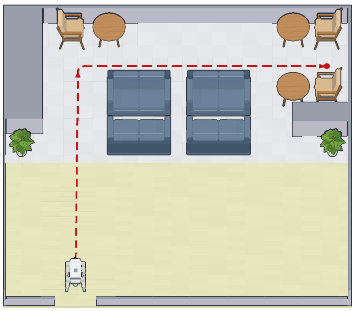
R1 Walk along the corridor into the elevator lobby. Turn right and continue with the white pillar to your left. Pass the first elevator on your right, then turn right toward the second elevator. Stop in front of its doors.

R2 1. Head along the passage into the elevator lobby. 2. Take a right and keep going with the white pillar to your left. 3. Go past the first elevator to your right, then take another right toward the second one. 4. Stop in front of its doors.

R3 Follow the passage into the elevator lobby. Turn right with the white pillar to your left. Pass the first elevator to your right, then turn right toward the second. Stop in front of its doors.

R4 First, proceed along the passage into the elevator lobby. Then, turn right and continue with the white pillar to your left. Next, pass the first elevator on your right, then turn right toward the second elevator. Finally, stop in front of its doors.

Route 08 Sofas and wooden tables



Ref. Walk down the corridor, keeping the group of four sofas on your right. Once you are past the sofas, turn right and continue toward the two round wooden tables ahead. Walk between these two tables and stop once you have passed through the gap.

R1 Go down the corridor, keeping the group of four sofas on your right. Once you are past the sofas, turn right and continue toward the two round wooden tables ahead. Walk through the space separating these two tables and stop once you have passed through the gap.

R2 1. Head down the corridor with the group of four couches to your right. 2. Once you have gone past the couches, take a right and head toward the two round wooden tables ahead. 3. Go between the two tables and stop once you are through the gap.

R3 Follow the corridor with the four couches to your right. After passing them, turn right toward the two round wooden tables ahead. Go through the gap between them and stop after clearing it.

R4 First, proceed along the corridor, keeping the group of four sofas on your right. Then, after you have passed the sofas, turn right and continue toward the two round wooden tables ahead. Finally, proceed through the space separating these two tables and stop once you have passed through the gap.

Route 09 Office aisle and sink



Ref. Go through the entrance glass doorway, and walk straight along the central aisle between the office areas. After passing the office areas, turn left into the side passage before reaching the glass doorway ahead. Go through the passage and stop in front of the sink.

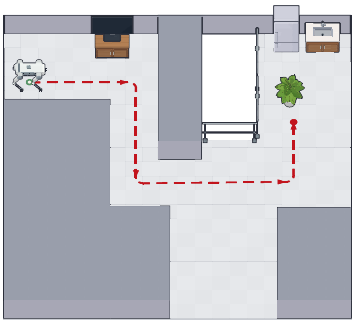
R1 Go through the entrance glass doorway, and walk straight along the central aisle with office areas on both sides. After passing the office areas, turn left into the side passage before reaching the glass doorway ahead. Go through the passage and stop in front of the washbasin.

R2 1. Come through the entrance glass doorway, and head straight along the central aisle with office areas on both sides. 2. Once you are past the office areas, take a left into the side passage before you reach the glass doorway ahead. 3. Go through the passage and stop in front of the washbasin.

R3 Enter through the glass doorway and go straight along the central aisle, with office areas on both sides. After passing them, turn left into the side passage before the glass doorway ahead. Follow this passage through and stop in front of the washbasin.

R4 First, proceed through the entrance glass doorway, and proceed straight along the central aisle with office areas on both sides. Then, after passing the office areas, turn left into the side passage before reaching the glass doorway ahead. Finally, proceed through the passage and stop in front of the sink.

Route 10 Television, railing, and plant



Ref. Walk forward past the television on your left, then turn right. Continue to the end of the wall on your left and turn left. Walk forward with the railing on your left. Once you are past the end of the railing, turn left toward the plant and stop beside it.

R1 Walk forward past the TV on your left, then turn right. Continue to the end of the wall on your left and turn left. Walk forward with the railing to your left. Once you are past the end of the railing, turn left toward the plant and stop beside it.

R2 1. Head forward past the TV to your left and take a right. 2. Keep going to the end of the wall to your left, then take a left. 3. Go ahead with the railing to your left. 4. Once you are past its end, take another left toward the plant and stop beside it.

R3 Pass the TV to your left and turn right. At the end of the wall to your left, turn left. Continue with the railing to your left. After its end, turn left toward the plant and stop beside it.

R4 First, proceed forward past the television on your left, then turn right. Then, continue to the end of the wall on your left and turn left. Next, proceed forward with the railing to your left. Finally, after you have passed the end of the railing, turn left toward the plant and stop beside it.