WHEN LANGUAGE GAINS ARE NOT SEMANTIC: RETHINKING ATTRIBUTION IN LANGUAGE-SUPERVISED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Language is increasingly used as an auxiliary supervisory signal for representation learning, and improvements over a single-task baseline are often attributed to semantic information in the language targets. That attribution is not identified by the usual comparison: adding language also changes the objective, optimization dynamics, target geometry, and instance-level supervision. We introduce a controlled attribution protocol that separates three claims: whether the complete auxiliary intervention helps, whether correct input–target correspondence helps, and whether structured targets help beyond generic instance supervision. The protocol compares correct language targets with both correspondence-destroying derangements and fixed unstructured instance targets. We evaluate the protocol in radar place representation learning, a setting in which language and camera information are available during training while retrieval remains radar-only. On Boreas, correct captions improve Recall@1 over radar-only training by $+3.18$ points (95% CI $[1.34, 5.02]$). Yet deranged captions improve it by $+2.75$ points and random targets by $+2.17$ points; correct captions exceed these controls by only $+0.43$ ($[-1.66, 2.52]$) and $+1.00$ points ($[-0.33, 2.33]$), respectively. Thus the conventional baseline comparison establishes an auxiliary-intervention effect but neither a correspondence-specific nor structured-target effect. The same attribution failure persists with richer camera-VLM descriptions, a diagnostic-selected caption subset, and a reverse-direction MulRan replication. Direct supervision from frozen camera features likewise fails to establish stable transfer. To audit targets before training, we introduce a stability–collision analysis that tests whether they preserve task-relevant invariances while separating unrelated examples. Increasing caption entropy from 2.91 to 10.19 bits reduces cross-revisit stability from 0.729 to 0.121, exposing a fundamental tension between descriptive richness and invariance. Together, our results show that an improvement over a sensor-only baseline establishes the value of an auxiliary training intervention, not the value of language content. Semantic attribution requires controls that separately test target correspondence and structured target information.

1 INTRODUCTION

Language is increasingly used as an auxiliary supervisory signal for learning representations from images, point clouds, and robotic sensors (Radford et al., 2021; Sautier et al., 2022; Hess et al., 2024; Xue et al., 2023; Zeng et al., 2023). When a language-supervised model outperforms a single-task baseline, the gain is commonly interpreted as evidence that semantic information has shaped the representation. That conclusion is scientifically consequential: it motivates the collection of captions, the use of increasingly capable vision–language models, and claims that language supplies abstractions unavailable to the deployed sensor.

The conventional comparison does not, however, identify a semantic effect. Adding language supervision simultaneously adds an auxiliary loss, a projection head, a target distribution, and an instance-level learning signal. Any of these changes can alter optimization or regularize the representation even when target meaning is irrelevant. The central question is therefore not only whether a

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.

language-supervised intervention improves performance, but which component of that intervention caused the improvement.

We encountered this ambiguity in radar place representation learning. Correct captions produced a statistically resolved gain over radar-only training, a result that would ordinarily support a language-supervision claim. The gain largely survived when every caption was assigned to the wrong observation and when text embeddings were replaced by fixed random instance targets. The same experiment therefore supports two opposite narratives depending on whether negative controls are included. This reversal motivates our broader question: what evidence is required before a representation-learning gain can be attributed to language content?

We distinguish three claims that are often conflated. Correct targets versus a single-task baseline estimate the effect of the complete *auxiliary intervention*. Correct versus deranged targets test whether the input–target *correspondence* matters while preserving the target pool. Correct versus fixed random instance targets test whether *structured target information* helps beyond a generic auxiliary objective. Under our operational criterion, a semantic attribution requires evidence for both the correspondence-specific and structured-target comparisons, not merely an intervention effect.

We use radar place recognition as a controlled stress test. Language and synchronized camera information are available only during training, while the deployed query and database remain radar-only. Repeated traversals provide explicit invariance relationships, geometric ground truth supplies an objective downstream metric, and nuisance changes expose whether a target is stable across observations of the same place. Before training, we audit this necessary target property through cross-revisit stability, between-place collision, and their difference, termed contrast.

Across Boreas and MulRan, deterministic radar captions, camera-VLM descriptions, and direct frozen-camera targets, the result is consistent: the standard sensor-only comparison can suggest a semantic gain that is not separated from correspondence-destroying or unstructured controls. We do not conclude that language supervision is universally ineffective. Instead, we show that a widely used comparison is insufficient to identify why an auxiliary representation-learning intervention succeeds.

Our contributions are:

- we formalize three distinct claims behind language-supervised representation gains—intervention, correspondence, and structured-target effects—and introduce negative controls that operationally separate them;
- we introduce a pre-training target audit based on cross-view stability and between-instance collision, exposing a tension between descriptive richness and task-relevant invariance; and
- across two datasets and multiple target sources, we show that a sensor-only baseline can support a semantic interpretation that disappears under correspondence-destroying and unstructured-target controls.

2 Related Work

2.1 Radar place recognition

Radar place recognition spans circularly aligned polar descriptors, learned embeddings, and sequence matching (Kim & Kim, 2018; Kim et al., 2020; Gadd et al., 2020; 2021). Recent systems improve transform robustness, compactness, multimodal alignment, or cross-sensor transfer (Jang et al., 2023; Gadd & Newman, 2024; Kim et al., 2024; Yin et al., 2021; Nayak et al., 2024; Fu et al., 2024; Kim et al., 2025). These methods primarily ask how to improve retrieval. We instead use radar retrieval as a controlled downstream test of what can be attributed to a training-only language signal; detailed benchmark controls appear in Appendix A.

2.2 Cross-modal and language supervision

Learning with privileged information formalizes training-time signals that are absent at deployment; generalized distillation connects this setting to teacher supervision (Lopez-Paz et al., 2016). Large-scale vision–language pretraining further shows that text can organize visual representations

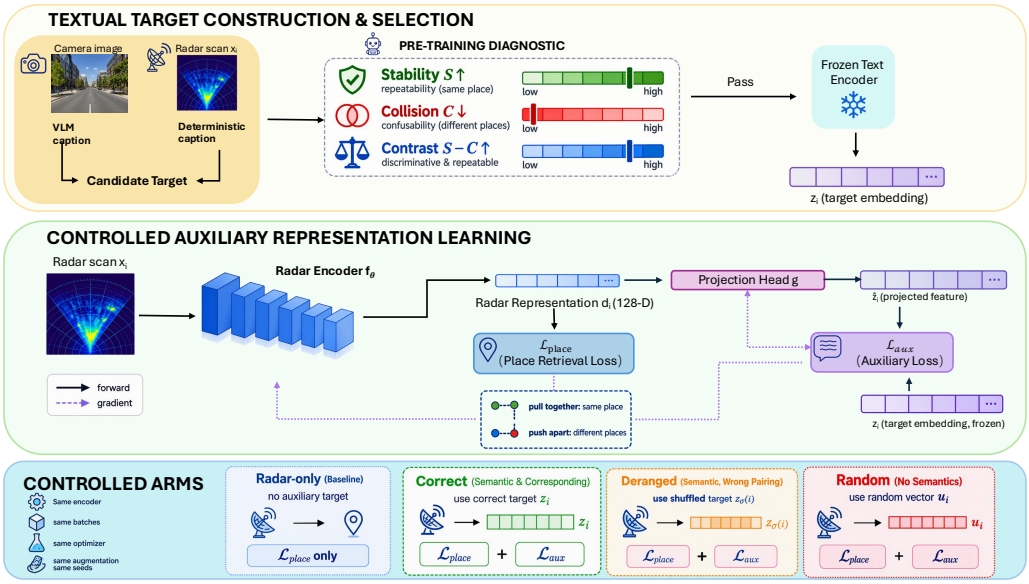


Figure 1: Overview of the proposed attribution protocol. Candidate textual targets are screened for cross-revisit stability and between-place collision, then embedded by a frozen text encoder to supervise radar representation learning. The controlled arms share the encoder, data, optimizer, augmentations, and seeds; only target meaning and target correspondence vary.

(Radford et al., 2021). In autonomous-driving perception, SLidR transfers image features into LiDAR, while LidarCLIP uses image–LiDAR pairs to map point clouds into a frozen language–image space (Sautier et al., 2022; Hess et al., 2024). ULIP and CLIP2 instead align language, images, and point clouds in multimodal feature spaces (Xue et al., 2023; Zeng et al., 2023).

Language can also be the localization query: Text2Pos and Text2Loc retrieve point-cloud submaps from spatial descriptions before fine localization (Kolmet et al., 2022; Xia et al., 2024). Controlled studies further show that language guidance varies with description granularity and distribution shift (Chatterjee et al., 2024). These works ask whether supervision improves performance. An improvement shows that the overall training intervention helped; it does not by itself establish that target meaning or correct cross-modal correspondence caused the gain.

Our setting isolates this distinction. Camera and language information is available only during training, while both the deployed query and database are radar-only. The central question is therefore semantic attribution, not merely supervision effectiveness. Radar-only versus a correct target tests the complete training intervention; a semantic claim additionally requires the correct target to outperform controls that remove target correspondence or target meaning.

2.3 Negative Controls for Semantic Attribution

Auxiliary tasks are known to alter shared representations and optimization even when the main task is unchanged (Caruana, 1997), so they can help through optimization or regularization rather than target meaning. Random-label experiments show that networks can fit arbitrary assignments (Zhang et al., 2017); random auxiliary targets can also induce structured, non-collapsed representations (Lyle et al., 2021). Thus downstream improvement is not evidence that target semantics caused the gain.

What remains missing is a controlled attribution protocol for multimodal robot perception that separates target meaning, target correspondence, and generic auxiliary-objective effects. We introduce two negative controls for that purpose. Deranged captions preserve the empirical caption distribution but destroy observation–target correspondence. Fixed random vectors remove shared semantic and revisit structure while preserving the auxiliary head and continuous instance-level objective. Correct versus deranged tests target correspondence; correct versus random tests structured seman-

tic supervision beyond a generic auxiliary objective. These are attribution controls, not localization methods.

3 Attributing Language-Supervision Gains

3.1 Three claims behind one observed gain

Let B denote a single-task baseline, T training with the correct language-derived target, T_π the same intervention after a fixed-point-free reassignment of targets, and R training with fixed unstructured instance targets. Their downstream scores define three increasingly specific comparisons:

$$\Delta_{\text{int}} = \text{Score}(T) - \text{Score}(B), \tag{1}$$

$$\Delta_{\text{corr}} = \text{Score}(T) - \text{Score}(T_\pi), \tag{2}$$

$$\Delta_{\text{struct}} = \text{Score}(T) - \text{Score}(R). \tag{3}$$

Δ_{int} measures the complete auxiliary intervention and therefore mixes target content with the additional head, loss, and optimization dynamics. Δ_{corr} asks whether the correct input–target assignment matters while preserving the empirical target pool. Δ_{struct} asks whether shared target structure is more useful than an unstructured instance-level objective. Evidence that $\Delta_{\text{int}} > 0$ alone supports an intervention claim; under our operational criterion, attributing the gain to semantic supervision additionally requires evidence in both controlled comparisons. Failure to separate an arm from its controls does not prove that its effect is exactly zero; it bounds the claim supported by the experiment.

3.2 Radar representation learning as a testbed

Given a radar query x_q and database scans $\{x_i\}$ with positions $\{p_i\}$, an encoder f_θ retrieves the nearest descriptor under cosine distance. Recall@1 at radius r is

$$R@1(r) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\left[\|p_q - p_{\pi_q(1)}\|_2 < r\right]. \tag{4}$$

We use 5 m as the primary radius because 25 m retrieval is saturated in the tested repeated-route setting. Traversal-disjoint Boreas splits contain 1414, 696, and 699 distance-sampled train, validation, and test keyframes. Every arm uses the same ImageNet-initialized ResNet-18 (He et al., 2016), 128-D normalized place descriptor, batch-hard triplet objective, optimizer, augmentation, 20-epoch budget, and frozen all-MiniLM-L6-v2 targets (Reimers & Gurevych, 2019). Checkpoints are selected on validation R@1@5m and the test traversal is evaluated once. Results use matched initialization, batches, and triplet sampling over seeds 42–46; intervals are two-sided 95% Student-t intervals over matched-seed differences. Appendix B gives preprocessing, sampling, hyperparameters, and integrity checks.

3.3 Operational negative controls

Let L_{place} denote the triplet objective and z_i an auxiliary target. Language-supervised arms minimize

$$L = L_{\text{place}} + \lambda L_{\text{aux}}(g(f_\theta(x_i)), z_i), \qquad \lambda = 0.1, \tag{5}$$

where g is a projection head. We compare five arms:

1. *radar only*: L_{place};
2. *selected caption*: z_i uses the contrast-selected description of x_i;
3. *deranged selected caption*: $z_{\sigma(i)}$, where $\sigma(i) \neq i$ for every sample;
4. *random instance target*: an independently sampled, fixed 384-D random unit vector per training observation, with one target draw held fixed across all model seeds; and
5. the original, simpler caption vocabulary used to test granularity.

Table 1: Radar-caption target quality. Higher stability and contrast are better; lower collision is better.

Descriptor	Bits	Stability	Collision	Contrast
Original (27 meanings)	4.14	0.444	0.076	0.368
All eight axes	10.19	0.121	0.001	0.120
Greedy: near + far	2.91	**0.729**	0.143	**0.586**

A derangement is essential: a naive shuffle can retain fixed points. We permute within each split and preserve its exact caption pool; coarse semantic classes may still collide at their chance rate because the vocabulary is finite (15.8% of training targets, from a different source observation). The random-instance arm goes further: it preserves the auxiliary head and an instance-level contrastive objective while providing neither caption semantics nor a target shared across revisits. Because the assignment uniquely identifies training observations, we do not call it information-free: it controls generic instance-level auxiliary supervision, not every property of text-embedding geometry. A semantic claim under our operational criterion requires correct captions to outperform both controls beyond seed variation.

3.4 A NECESSARY-CONDITION TARGET AUDIT

For a discrete description function $c(\cdot)$, let P^+ be matched cross-traversal pairs within 5 m and P^- be unrelated pairs separated by at least 50 m. We define

$$S = \Pr[c(x_i) = c(x_j) \mid (i,j) \in P^+], \tag{6}$$

$$C = \Pr[c(x_i) = c(x_j) \mid (i,j) \in P^-], \tag{7}$$

$$D = S - C. \tag{8}$$

Here S is revisit stability, C is collision, and D is contrast. A target intended to encode place invariance must be repeatable at the same place without being ubiquitous elsewhere. Entropy alone is insufficient: adding attributes can lower collision while destroying stability even faster. We use D as a necessary-condition audit rather than a sufficient predictor of downstream performance, greedily adding fields only when they increase it. We also compute a soft version by replacing exact equality with cosine similarity between frozen sentence embeddings.

4 A CONVENTIONAL GAIN THAT FAILS ATTRIBUTION

4.1 TARGET CONSTRUCTION AND INVARIANCE AUDIT

Radar captions encode coarse structural attributes computed from the polar scan, including return density, corridor topology, openness, near/far energy, first-return distance, and angular entropy. Because these captions are a deterministic function of the radar input, they add no Shannon information about place beyond the input. They may nevertheless provide a useful inductive bias; the controls determine whether that bias depends on target meaning.

Table 1 exposes a granularity trade-off. The original 27-meaning vocabulary contains 4.14 bits. An eight-axis vocabulary raises entropy to 10.19 bits and almost eliminates random collision, but its revisit stability collapses to 0.121. Greedy contrast selection instead retains only near- and far-range structure: nine meanings, 2.91 bits, stability 0.729, and the highest contrast of 0.586. This is far below the 11.46 bits required to index 2809 individual keyframes, but place indexing is not our claim; the key observation is that descriptive richness and repeatability conflict.

4.2 FROM INTERVENTION GAIN TO ATTRIBUTION FAILURE

Figure 3(a) gives the decisive experiment. Viewed through the conventional single-baseline comparison, the result appears conclusive: contrast-selected captions rank first, and their matched-seed gain over radar-only is $+3.18$ points (95% CI $[1.34, 5.02]$). This establishes $\Delta_{\text{int}} > 0$ and, without negative controls, would naturally be reported as evidence that language supervision improves

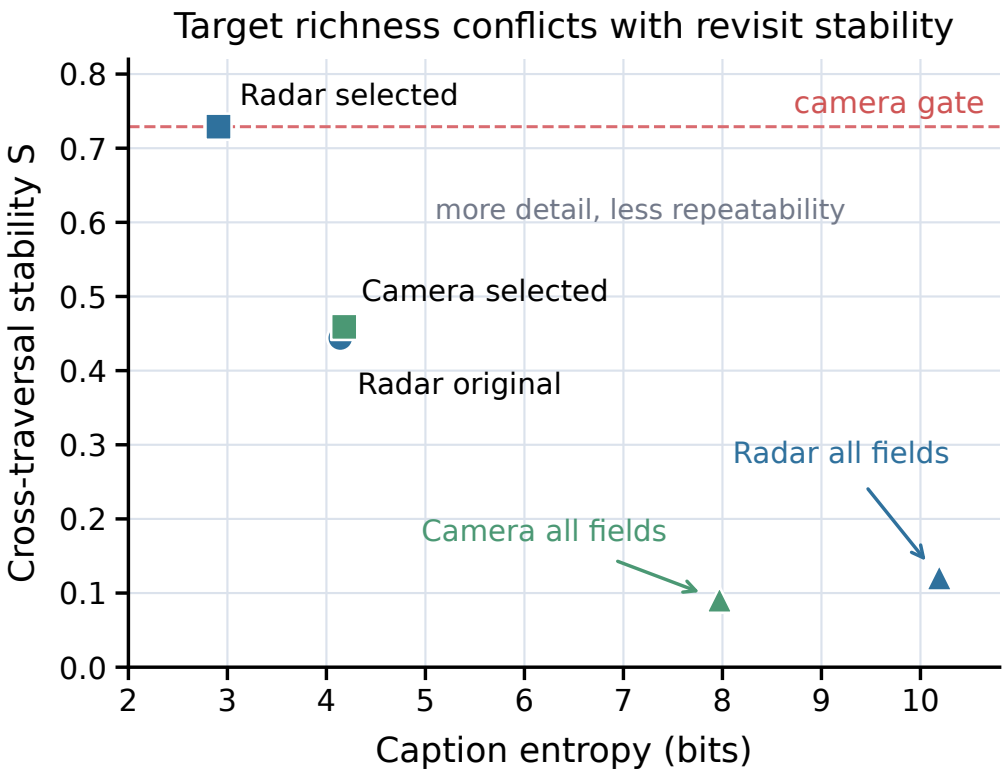


Figure 2: Increasing caption entropy lowers cross-traversal stability for both radar and camera descriptions. Greedy field selection favors coarse, repeatable targets rather than maximal descriptive detail.

the representation. The two non-semantic controls also improve over radar-only: deranged captions by $+2.75$ points (95% CI $[1.45, 4.05]$) and fixed random targets by $+2.17$ points (95% CI $[0.73, 3.62]$). However, the correct captions' gain over deranged captions is only $+0.43$ points (95% CI $[-1.66, 2.52]$), and their gain over fixed random targets is $+1.00$ points (95% CI $[-0.33, 2.33]$). Neither attribution interval excludes zero. The original and deranged-caption arms also have the same mean Recall@1.

The defensible conclusion is therefore stronger than a null result and narrower than a universal claim about language: in this setting, the experiment establishes an auxiliary-intervention effect but neither Δ_{corr} nor Δ_{struct}. The mechanism may combine optimization regularization and instance discrimination. Comparing only the correct-caption and radar-only rows changes the scientific interpretation of exactly the same experiment.

5 Testing Alternative Explanations

Radar-derived descriptions cannot add information absent from radar. We therefore test the first alternative explanation—that their failure is caused by an information ceiling—using a stronger source of privileged supervision: synchronized camera images available only during training. Each of the 1414 training radar keyframes is paired with the nearest camera frame under a hard 100 ms gate; all pairs pass, with 13.7 ms median and 21.0 ms 95th-percentile offset. A vision–language model produces canonical structured descriptions over road topology, built form, boundaries, persistent landmarks, spatial layout, and openness. Prompts exclude weather, season, vehicles, color, readable text, and location identity to avoid traversal shortcuts. We use Qwen3-VL-8B-Instruct (Bai et al.,

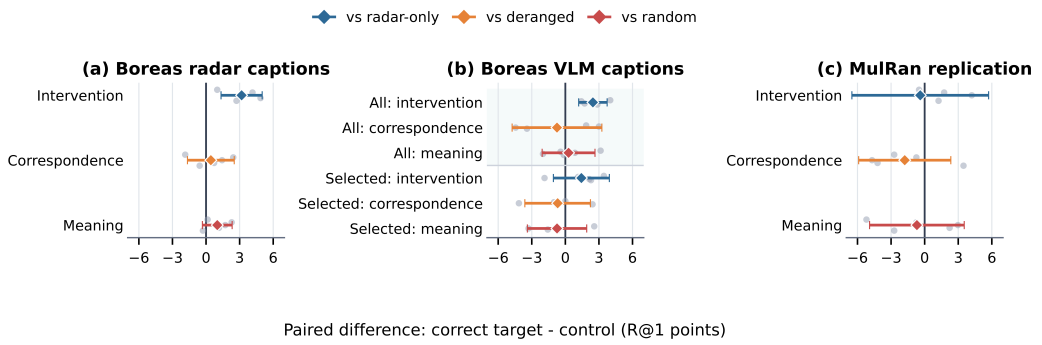


Figure 3: Three attribution levels over seeds 42–46. Circles show matched-seed differences; diamonds and bars show the mean and 95% Student-t interval. Positive values favor correct targets. Blue compares them with radar-only (complete intervention), orange with deranged captions (target correspondence), and red with fixed random instance targets (structured targets beyond a generic auxiliary objective). (a) Boreas radar captions. (b) Boreas VLM captions. (c) MulRan replication. No semantic-attribution interval excludes zero.

Table 2: Cross-traversal quality of camera-VLM and radar-derived captions.

Target	Stability	Collision	Contrast
Radar: near + far	**0.729**	0.143	**0.586**
Camera: topology + built form	0.459	0.085	0.374
Camera: all six fields	0.091	**0.006**	0.085

2025) with deterministic decoding and require a canonical JSON schema; generation is not used at validation or test time.

Before generation and downstream training, we document a gate: the best camera-caption subset must exceed the radar-caption stability of 0.729, yield positive contrast in both traversal directions, and pass a leakage audit. This prevents training and objective tuning after observing weak targets. Table 2 shows that the gate fails. Topology plus built form maximizes contrast, but stability is only 0.459. Using all six fields creates 536 meanings while reducing stability to 0.091. Soft embedding contrast agrees (0.081 for the best subset and 0.048 for all fields). Under the pre-specified protocol this stops a confirmatory camera-language arm: external semantics are not automatically stable place information. We nevertheless conduct two post-gate experiments below as explicitly exploratory falsification tests: the complete target tests the target-coarseness objection, while the diagnostic-optimal subset tests whether field selection repairs attribution. Both field sets are fixed from the pre-training analysis, without downstream selection or training changes.

A stratified visual sanity check of 12 source images finds 10/12 topology and 9/12 built-form labels acceptable, with one additional built-form case ambiguous. This small audit is not an accuracy estimate, but it shows that the gate failure is not explained solely by nonsensical generations; rare junctions and ambiguous building boundaries remain visible sources of label noise.

5.1 Richer language does not rescue attribution

The complete six-axis VLM target contains 536 meanings and 7.97 bits, compared with nine meanings and 2.91 bits for the selected radar near+far target. The diagnostic-optimal topology+built-form subset has 56 meanings and 4.18 bits; its stability/contrast of 0.459/0.374 is substantially higher than the complete target's 0.091/0.085. For each field set, the deranged arm preserves the exact caption pool while changing every frame's selected semantic label; both have zero fixed points and zero same-label assignments. All VLM arms use the same frozen sentence encoder, InfoNCE weight, radar tower, batches, seeds 42–46, and validation-only 5 m checkpoint rule as the primary comparison.

Figure 3(b) demonstrates why a sensor-only baseline is an insufficient attribution test. Correct VLM captions improve over radar-only by $+2.46$ points (95% CI $[1.19, 3.73]$), apparently supporting privileged semantic supervision. However, correct VLM captions trail their derangement by -0.75 points ($[-4.75, 3.26]$), differ from random targets by only $+0.29$ points ($[-2.07, 2.65]$), and are unresolved against the much coarser near+far target (-0.72, $[-3.79, 2.36]$). The diagnostic-optimal subset also fails: correct targets attain $71.30 \pm 2.15\%$, trail their derangement by -0.69 points ($[-3.62, 2.24]$), and differ from random targets by -0.74 points ($[-3.39, 1.91]$). Their $+1.43$-point radar-only difference is unresolved ($[-1.07, 3.93]$). Thus neither greater description entropy nor pre-training field selection produces correspondence-specific evidence. Target richness and semantic attribution are empirically distinct.

5.2 Removing language abstraction does not rescue transfer

The second alternative explanation is that language discards useful visual information. We therefore align radar with frozen ImageNet ResNet-18 (He et al., 2016) and DINOv2-S/14 (Oquab et al., 2023) camera features using the same paired data and checkpoint rule (Appendix C). DINOv2 is the stronger camera descriptor: its same-place minus other-place cosine gap is 0.118 versus 0.043 for ResNet-18, and its direct R@1@5m is 17.62% versus 9.41%. Yet ResNet-18 distillation changes radar retrieval by $+1.86$ points (95% CI $[-1.31, 5.03]$), while DINOv2 changes it by only $+0.23$ ($[-3.34, 3.80]$). A target can contain task-relevant information without providing an effective training signal to the deployed modality.

6 Replication Under Viewpoint Reversal

We replicate the learned attribution test on the unsaturated MulRan Riverside route (Kim et al., 2020); Appendix A documents the database-density, calibration, and descriptor audits motivating this choice. Riverside01 and 02 form a 1305-frame training database with cross-traversal positives. The 702 overlapping reverse-direction Riverside03 queries are split using geometry alone: a contiguous 20% segment supplies 140 validation queries, while 404 test queries remain after removing locations within 50 m of validation. The near+far axes are fixed from Boreas, with thresholds fitted on MulRan training only. All four arms share the architecture, optimizer, 20-epoch budget, seeds, auxiliary weight, and circular azimuth augmentation; validation and test use the conventional 25 m radius.

Figure 3(c) shows that the attribution conclusion replicates under a setting in which the generic auxiliary gain itself disappears. Correct captions change radar-only by -0.39 points (95% CI $[-6.50, 5.71]$), trail deranged captions by -1.78 points ($[-5.90, 2.34]$), and trail random targets by -0.69 points ($[-4.92, 3.54]$). Every interval spans zero. Unlike Boreas, MulRan does not establish a generic auxiliary-objective gain either. The auxiliary intervention is dataset-dependent; the inability of the sensor-only comparison to identify a caption-semantic effect is reproduced across both datasets.

7 Discussion

7.1 Implications for representation learning

The main result is an identification failure, not a claim that arbitrary supervision is optimal. A fixed random target can act as an instance identifier, and its geometry differs from text embeddings. It controls instance-level auxiliary supervision, while derangement preserves the caption pool and embedding geometry but destroys correspondence. Agreement between these complementary controls strengthens a non-semantic explanation, without distinguishing regularization from instance-discrimination effects.

This distinction changes what can be learned from a common experimental pattern. All auxiliary arms improve on Boreas; none is separated from radar-only on MulRan. The generic auxiliary effect is dataset-dependent. What replicates is that the sensor-only comparison does not identify whether correspondence or structured language caused an observed improvement—a distinction relevant to any auxiliary language objective.

The target audit exposes a complementary failure: more fields reduce collisions but rapidly damage revisit invariance. Rich VLM captions and the topology+built-form subset both fail controlled attribution, so neither coarse language nor exclusive use of the least stable target explains the result. The audit is therefore a necessary-condition screen, not a sufficient downstream predictor. Likewise, DINOv2 is the stronger direct camera descriptor yet gives the smaller radar gain: target quality is necessary but not sufficient for privileged transfer.

7.2 Limitations

The study covers one route per dataset, one encoder family, one VLM schema and text encoder, and two global camera teachers under a shared InfoNCE design; other targets, dense objectives, or modalities may succeed. MulRan validation and test are spatially separated portions of one reverse traversal. Five matched seeds provide limited power: intervals spanning zero do not prove equivalence or exclude every meaningful effect, and learned-arm uncertainty covers seeds on fixed routes rather than route sampling. Our conclusion concerns what these comparisons identify, not the universal absence of a language benefit.

8 Conclusion

Language-supervised representations can improve without identifying a benefit from language content. Across deterministic captions, camera-VLM descriptions, and a reverse-direction replication, correct targets are not separated from controls that destroy correspondence or shared structure. The broader lesson is methodological: a single-task baseline answers whether an intervention helped, not why. Without attribution controls, the same experiment can support the wrong scientific story.

Reproducibility Statement

Sections 3–7 and Appendices A–C report the data partitions, preprocessing, target construction, negative-control generation, matched seeds, optimization, checkpoint selection, uncertainty estimates, and integrity checks. The supplementary material provides complete configurations, per-seed results, caption-generation provenance, and artifact hashes.

AI Use Statement

Large language models were used for manuscript language polishing. Separately, as part of the experimental method described in Section 5, Qwen3-VL-8B-Instruct was used to generate camera captions for auxiliary training targets; Section 5 describes the generation constraints and a manual sanity check. The authors designed the study, conducted and audited the experiments, verified all numerical results and citations, and take full responsibility for the scientific claims and submitted text.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. doi: 10.1023/A:1007379606734.

Agneet Chatterjee, Tejas Gokhale, Chitta Baral, and Yezhou Yang. On the robustness of language guidance for low-level vision tasks: Findings from depth estimation. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2794–2803, 2024.

Shaowei Fu, Yifan Duan, Yao Li, Chengzhen Meng, Yingjie Wang, Jianmin Ji, and Yanyong Zhang. CRPlace: Camera–radar fusion with BEV representation for place recognition. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 8421–8427, 2024. doi: 10.1109/IROS58592.2024.10802267.

Matthew Gadd and Paul Newman. Open-radvlad: Fast and robust radar place recognition. In *Proc. IEEE Radar Conf. (RadarConf)*, 2024. doi: 10.1109/RadarConf2458775.2024.10548974.

Matthew Gadd, Daniele De Martini, and Paul Newman. Look around you: Sequence-based radar place recognition with learned rotational invariance. In *Proc. IEEE/ION Position, Location and Navigation Symposium (PLANS)*, 2020.

Matthew Gadd, Daniele De Martini, and Paul Newman. Contrastive learning for unsupervised radar place recognition. *arXiv preprint arXiv:2110.02744*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson. LidarCLIP or: How i learned to talk to point clouds. In *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (WACV)*, pp. 7423–7432, 2024. doi: 10.1109/WACV57701.2024.00727.

Hyesu Jang, Minwoo Jung, and Ayoung Kim. Raplace: Place recognition for imaging radar using radon transform and mutable threshold. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 11194–11201, 2023. doi: 10.1109/IROS55552.2023.10341883.

Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, pp. 4802–4809, 2018. doi: 10.1109/IROS.2018.8593953.

Giseop Kim, Yeong Sang Park, Younghun Cho, Jinyong Jeong, and Ayoung Kim. Mulran: Multimodal range dataset for urban place recognition. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pp. 6246–6253, 2020. doi: 10.1109/ICRA40945.2020.9197298.

Hanjun Kim, Minwoo Jung, Wooseong Yang, and Ayoung Kim. SHeRLoc: Synchronized heterogeneous radar place recognition for cross-modal localization. *arXiv preprint arXiv:2506.15175*, 2025.

Hogyun Kim, Byunghee Choi, Euncheol Choi, and Younggun Cho. ReFeree: Radar-based lightweight and robust localization using feature and free space. *IEEE Robotics and Automation Letters*, 9(12):11042–11049, 2024. doi: 10.1109/LRA.2024.3474554.

Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 6687–6696, 2022.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.

Clare Lyle, Mark Rowland, Georg Ostrovski, and Will Dabney. On the effect of auxiliary tasks on representation dynamics. In *Proc. Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, volume 130, pp. 1–9, 2021.

Abhijeet Nayak, Daniele Cattaneo, and Abhinav Valada. RaLF: Flow-based global and metric radar localization in LiDAR maps. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 8748–8763, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. Conf. Empirical Methods in Natural Language Processing and Int. Joint Conf. Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 9891–9901, 2022.

Yan Xia, Letian Shi, Zifeng Ding, João F. Henriques, and Daniel Cremers. Text2Loc: 3d point cloud localization from natural language. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 14958–14967, 2024. doi: 10.1109/CVPR52733.2024.01417.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1189, 2023. doi: 10.1109/CVPR52729.2023.00120.

Huan Yin, Xuecheng Xu, Yue Wang, and Rong Xiong. Radar-to-lidar: Heterogeneous place recognition via joint learning. *Frontiers in Robotics and AI*, 8, 2021. doi: 10.3389/frobt.2021.661199.

Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. CLIP2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 15244–15253, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *Proc. Int. Conf. Learning Representations (ICLR)*, 2017.

A Benchmark and Sensor Audit

A.1 Repeated-route saturation

At a 25 m acceptance radius, flattened polar-image cosine similarity obtains 95.99% using two database traversals and 99.76% when the database contains six traversals. The latter is not evidence of robustness: database densification reduces the median query-to-nearest-database distance from 1.93 m to 0.83 m. After matching database density, winter-to-summer retrieval remains 97.74%, and even a one-traversal maximum-season-gap protocol reaches 97.99%. These results explain why we use 5 m for learned Boreas comparisons and why density must be controlled before attributing a change to representation quality.

A.2 Cross-sensor calibration

Boreas replaced its scanning radar during collection, changing radial resolution from 0.0596 to 0.04381 m/bin. This creates a controlled sensor shift on the same vehicle and route. We cap each group deterministically at 700 keyframes and compare native polar pixels with a shared metric grid. Table 3 shows that native cross-sensor matching nearly collapses. Metric resampling recovers most performance, showing that the dominant native gap is calibration rather than representation. A material residual gap remains: at 25 m, cross-sensor Recall@1 is 22.14 points below within-sensor for raw cosine and 18.29 points below for Radar Scan Context.

Table 3: Boreas Recall@1 within 25 m under matched database density. Values are within-sensor/cross-sensor percentages.

Descriptor	Native pixels	Metric grid
Raw cosine	95.14 / 5.86	92.57 / 70.43
Radar Scan Context	95.14 / 2.71	90.43 / 72.14

A.3 MulRan descriptor audit

Riverside01 is the database; Riverside02 and the reverse-direction Riverside03 are queries. Radar scans are synchronized to the provided global poses (median offset below 2.6 ms), sampled every 10 m, and restricted to query keyframes with a database reference within 25 m. This retains 95.3% and 96.8% of the two query routes. Raw cosine uses resized, unnormalized polar intensities. Radar Scan Context follows the released MulRan evaluator with 40 rings, 120 sectors, exhaustive circular alignment, and mean sector-wise cosine similarity. Confidence intervals use 10,000 circular moving-block resamples of ten adjacent keyframes (approximately 100 m), preserving local route correlation.

Unlike Boreas, MulRan is not saturated by direct matching (Table 4). Same-direction R@1 at 25 m reaches 74.80% for raw cosine and 72.42% for Radar Scan Context. Under reverse travel, all descriptors degrade sharply; circularly aligned Radar Scan Context is strongest but reaches only 10.40%. A paired moving-block bootstrap does not separate raw cosine and Radar Scan Context under same-direction travel (raw minus RSC: $+2.38$ points, 95% CI $[-0.95, 6.02]$). Under reverse travel, RSC is reliably stronger (raw minus RSC: -5.13 points, 95% CI $[-9.69, -1.14]$).

Table 4: MulRan Riverside R@1 within 25 m using Riverside01 as database.

Descriptor	Riverside02	Riverside03 (reverse)
Raw cosine	**74.80**	5.27
Azimuth profile	60.70	2.85
Radar Scan Context	72.42	**10.40**

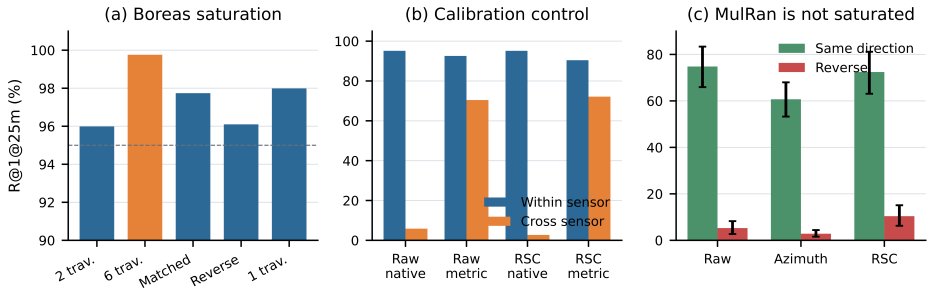


Figure 4: Benchmark audit. (a) Boreas remains near ceiling across repeated-route protocols, and increasing database traversals makes it easier. (b) A shared metric grid removes most of an apparent Boreas cross-sensor collapse, although a residual gap remains. (c) MulRan Riverside is not saturated; reverse travel is substantially harder despite circular alignment. Error bars are 95% moving-block bootstrap intervals.

B Boreas Training and Integrity Details

Keyframes are sampled every 10 m of travel rather than at fixed time intervals, preventing stops from filling the database with near duplicates. The training traversals are 26 November and 1 December 2020, validation is 18 December 2020, and test is 15 January 2021. A query and database entry are never drawn from the same traversal. The first 2.5 m of each radar scan is masked; intensities are clipped at per-scan 1st/99th percentiles and min–max normalized; and polar images are area-resampled to 128×128 without Cartesian projection.

The radar tower produces a 128-D normalized place descriptor and uses a batch-hard triplet loss. AdamW uses learning rate and weight decay 10^{-4}, batch size 16, and triplet margin 0.2. Positives lie within 10 m, negatives beyond 50 m, and cross-traversal positives are preferred. The auxiliary loss weight is 0.1 and temperature is 0.07. Augmentation uses circular azimuth shifts up to 3% of a revolution, $\pm 10\%$ intensity scaling, and Gaussian noise with standard deviation 0.01.

The split builder enforces traversal disjointness, and a separate leakage audit checks frame identifiers, timestamps, paths, and coordinate overlap. Derangements have no fixed points or cross-split mappings; chance semantic collisions are recorded. Frozen text features are generated from training captions only. All arms use one code path and differ only in target assignment and the presence of the auxiliary loss. Regression tests cover self-exclusion from triplet negatives, radar decoding, calibration-aware resampling, retrieval masks, and MulRan pose synchronization.

C Direct Camera-Feature Controls

We use 512-D ImageNet ResNet-18 features and 384-D DINOv2-S/14 CLS tokens. Each arm uses the synchronized training pairs, projection dimension, symmetric InfoNCE loss with weight 0.1, model seeds, and 5 m checkpoint rule of the caption arms. Camera input is absent at validation and test time.

Table 5: Frozen camera-feature controls. Target contrast is same-place minus other-place cosine; direct and trained columns are R@1@5m percentages.

Camera target	Contrast	Direct	Trained
ResNet-18	0.043	9.41	71.73 ± 1.63
DINOv2-S/14	**0.118**	**17.62**	70.10 ± 1.80
Radar only	–	–	69.87 ± 1.16