Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

Four Reference Frames: Angular Diversity and Later View Coverage in Vehicle Tracks

Aaron Hughes[1] · Mason Chapman[2] · Gavin Bell[3]

[1] Spain · [2] Portugal · [3] Austria

Original publication: October 6, 2026 | Article 699

Abstract

Selecting varied reference views sounds useful, but diversity within an earlier segment need not cover the views encountered later. We compared four temporally spaced indices with four farthest-angle indices for 120 vehicle tracks. Selection used the first twenty qualifying frames; evaluation used the next twenty. Mean nearest-reference angular gap changed from 2.9770° to 2.9477°, a small reduction of 0.0293°. Coverage within a fixed 15° tolerance was identical for every track. The continuous and thresholded outcomes therefore gave different practical impressions of the same selection change. The result concerns supplied pose annotations rather than visual identity fidelity.

Annotation-based reference-index selection on public tracking sequences. Angular coverage is a geometric proxy; no identity-preserving video generation, image matching or detection system is evaluated.

Four Reference Frames: Angular Diversity and Later View Coverage in Vehicle Tracks

> Annotation-based reference-index selection on public tracking sequences. Angular coverage is a geometric proxy; no identity-preserving video generation, image matching or detection system is evaluated.

Abstract

Selecting varied reference views sounds useful, but diversity within an earlier segment need not cover the views encountered later. We compared four temporally spaced indices with four farthest-angle indices for 120 vehicle tracks. Selection used the first twenty qualifying frames; evaluation used the next twenty. Mean nearest-reference angular gap changed from 2.9770° to 2.9477°, a small reduction of 0.0293°. Coverage within a fixed 15° tolerance was identical for every track. The continuous and thresholded outcomes therefore gave different practical impressions of the same selection change. The result concerns supplied pose annotations rather than visual identity fidelity.

1. Diversity in the reference pool is not the deployment objective

A reference-selection rule can seek different-looking viewpoints without knowing which view will matter later. That is a useful distinction for any system conditioned on a small set of reference frames. A diverse earlier collection may span its own observed angle range well, while a temporally spaced collection may already contain nearly all of the angles needed in a short future segment.

We ask whether choosing four angularly diverse frame indices improves later observation-angle coverage compared with choosing four indices at uniform temporal positions. The reference budget is held at four, the candidate pool at twenty and the later evaluation segment at twenty. The two rules select from the same source vehicle, and their outcomes are paired within track before averaging.

Meng et al., 2026 study subject preservation through stacked multi-view identity mosaic injection. We do not run its generative system or assume that its reference-selection procedure matches ours. The relevant connection is the interpretation of a multi-view reference budget. Four source frames can be distinct files while supplying almost the same observation angle; conversely, a wider angular span need not improve the particular future views being evaluated.

Our outcome is intentionally narrower than subject identity. It measures the shortest circular distance from a later annotated observation angle to the selected earlier angles. A smaller distance indicates closer geometric coverage under this definition. It does not establish that the reference depicts the needed texture, expression, lighting or identity details, or that a generator would use that reference successfully.

2. The source field and the qualifying track cohort

We use the public KITTI tracking training annotations (Geiger et al., 2012). The official development-kit README distinguishes alpha, the object's observation angle, from rotation_y, its orientation around the camera's vertical axis (Geiger et al., 2026). We use alpha. Substituting rotation_y would define a different variable and would not automatically represent the same camera-relative observation condition.

The documentation also distinguishes tracking truncation flags from fractional truncation values used in another KITTI task. Tracking truncation is an integer field. Our row screen requires flag zero, meaning nontruncated under that convention, and allows occlusion states zero or one. This explicit field

interpretation matters more than treating a familiar column name as though its units were identical across releases or tasks.

Only Car rows with a nonnegative track identifier and alpha inside its documented angular range are candidates. Track identity is the sequence-and-track pair, since numeric identifiers need not refer to the same vehicle across sequences. Qualifying rows are sorted by frame. We split them at every missing or nonconsecutive frame rather than join surviving observations across an intervening gap.

For each track key, we select its earliest qualifying consecutive segment containing at least forty frames and retain the first forty of that segment. The first twenty form the reference candidate pool and the next twenty form evaluation observations. This yields 120 tracks in eighteen contributing sequences, from 531 car-track keys with at least one row surviving the initial screen.

The design therefore conditions on continued qualifying observation. A vehicle that disappears, becomes heavily occluded or fails the row conditions before forty consecutive frames does not enter this cohort. This retrospective cohort rule must not be described as an online admission policy that would know in advance which vehicles remain observable. It is a selected persistence-and-visibility sample, not all cars in the source benchmark.

Eighteen sequences contribute unevenly: the largest contains 23 retained tracks, while another contributes one. We preserve all eligible tracks and record their source sequence. Counting 2,400 later annotated frames does not create 2,400 independent vehicle histories. The main analysis retains one paired mean outcome per track and addresses sequence grouping in its covariance calculation.

3. Two fixed four-index rules

The temporal rule chooses zero-based candidate indices 0, 6, 13 and 19. It does not inspect candidate angles when choosing those positions. The rule includes the beginning and end of the candidate interval and two intermediate positions, giving an explicit implementation of approximately uniform temporal spacing at this twenty-frame budget.

The angular-diversity rule starts with candidate index zero. At each subsequent step, it computes every unselected candidate's shortest angular distance to its nearest already selected angle and chooses the candidate with the greatest such distance. An earliest-index tie rule makes the procedure deterministic. Selection stops after four distinct frame indices have been chosen.

Farthest-point selection has an established connection to geometric clustering (Gonzalez, 1985). Its theoretical motivation does not guarantee that this four-index heuristic outperforms temporal spacing on later frames. The classical approximation comparison concerns a specified objective on a supplied metric set, not a future distribution that the selection routine never observes. We make no transfer of that guarantee to the held-track outcomes.

The angular rule may select distinct indices whose angles are identical. Once an available angular range is represented, another index can add little new coverage. We retain such selections because the budget is defined in frame indices rather than in distinct angle bins. Claiming four different views solely from four selected files would conceal that potential redundancy.

Angular distance uses the absolute value of the two-argument arctangent of the sine and cosine of an angle difference. It lies between zero and pi and respects wraparound. Circular-data analysis motivates this treatment (Fisher, 1993). Ordinary subtraction would make angles near opposite sides of the stored wrap boundary appear far apart even when their viewing directions are close.

Both policies use the same earlier cohort, and neither reads later angles during selection. The diversity policy does require all twenty supplied candidate angles. These are annotations, not estimates produced by a pose network in the current experiment. Its informational requirement would matter in deployment even though we do not measure the cost of obtaining those angles here.

4. Continuous gap and fixed-tolerance coverage

For each later frame, we calculate the minimum circular distance to the four selected reference angles. Averaging those distances within track gives the primary outcome, expressed in degrees. A second descriptive outcome is the fraction of the twenty later frames whose minimum distance is at most fifteen degrees. That tolerance is a declared operational choice, not an empirically validated threshold for identity preservation.

The mean-gap contrast is diversity minus temporal spacing, so a negative value favors angular selection. Coverage is summarized on the same track cohorts. We also retain the maximum nearest-reference gap inside each earlier candidate pool, which records a selection-side property. That earlier maximum and the later mean are different objectives even before considering the thresholded coverage fraction.

Decision objectives determine the meaning of an apparent improvement (Elkan, 2001). A policy can move a later frame from two degrees away to one degree away without changing whether it meets a fifteen-degree rule. Conversely, a small angular change near the cutoff could change coverage. Reporting both quantities reveals that distinction rather than turning one selected summary into a universal measure of reference utility.

Coordinate geometry supplies the background for camera-relative viewing variables (Hartley and Zisserman, 2004), but the current scalar angle omits much of that geometry. We do not reconstruct object surfaces, account for all three rotational degrees of freedom or assess pixel-level correspondence. Calling this outcome angular coverage preserves its interpretation without implying a complete multi-view reconstruction assessment.

5. The fitted mean contrast

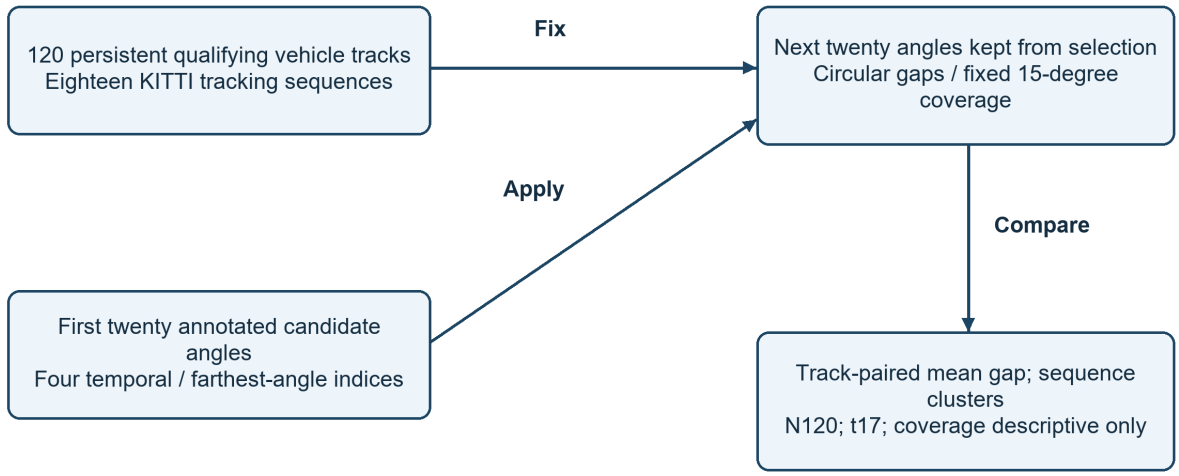

Figure 1. *Two policies choose four earlier frame indices from one twenty-frame candidate pool. Subsequent annotated angles produce a paired track-average gap; fifteen-degree coverage remains a separate descriptive outcome.*

The regression contains an intercept and 120 paired track-average gap changes. There are no controls, sequence fixed effects or covariate-selected subgroups. Standard errors cluster on the eighteen contributing sequences with a finite correction, and confidence intervals use t17. The coefficient is track weighted: sequence clustering changes the covariance estimate, not the weight assigned to each sequence in the point estimate.

Clustered inference concerns a specified dependence structure (Cameron and Miller, 2015), distinct from a variance correction for individually heteroskedastic observations (White, 1980). Shared scene conditions, related vehicles and annotation processes may exceed that working structure. The interval is conditional on the captured cohort, policies and annotation fields, and is not established coverage for future video-reference applications.

Table 1. *Diversity minus temporal-spacing mean nearest-reference angular gap, N=120 tracks, eighteen sequence clusters, finite-corrected covariance and t17 interval. The model contains only an intercept; R-squared is zero up to rounding. Fifteen-degree coverage is reported descriptively because its paired contrast has no variation in this sample.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.029274 (0.009593)	[-0.049513, -0.009035]	120
M1: fit	R-squared = 0.0000	-	-	120

Model labels: M1 = Diverse minus temporal-uniform held mean angle gap.

The temporal policy's mean later angular gap is 2.976978 degrees; the diversity policy's is 2.947704 degrees. The estimated change is -0.029274 degrees, standard error 0.009593 and interval [-0.049513, -0.009035]. The working interval excludes zero, but the effect's small scale remains relevant. It is not sufficient to report only a favorable statistical direction while omitting the size of the gap reduction.

Thirty tracks have a strictly smaller later mean gap under angular selection, thirteen have a larger gap and 77 are unchanged at a tolerance of 1e-12. Those mixed directions are retained. Influence analysis can help identify which tracks drive a mean contrast (Cook, 1977), but we do not remove deteriorating tracks or redefine the population after seeing the result.

Average fifteen-degree coverage is 0.968333 for both policies. More strongly, the paired coverage fraction is identical within every retained track. This does not mean every later angle is covered: some tracks have much lower coverage. It means that the selection change produces no difference in the number of covered later frames within each of these track summaries.

A constant-zero difference has no usable variation for the proposed intercept-only coverage regression. We therefore do not manufacture a finite fit statistic or present a zero-width population confidence interval. The unchanged observed count is a fixed-file result. Future coverage could still differ under other trajectories, tolerances, angle errors or reference budgets, none of which is identified by the zero variation here.

The candidate-pool maximum gap averages 1.262708 degrees for temporal selection and 1.286431 for the diversity heuristic. Thus the heuristic does not even lower this particular pool-average maximum summary in the captured sample. Its fixed initial point and greedy decisions are part of the implemented policy. The result reinforces the need to test the actual outcome rather than treating the word diversity as a performance guarantee.

6. What the experiment does not measure about identity

Tracking methods such as SORT, Deep SORT and AB3DMOT address association or state estimation across observations (Bewley et al., 2016; Wojke et al., 2017; Weng et al., 2020). ByteTrack and TrackFormer provide additional tracking formulations (Zhang et al., 2021; Meinhardt et al., 2021), while MOTChallenge supplies a separate benchmarking context (Dendorfer et al., 2021). Our track identifiers are supplied by the annotations. We do not run these trackers or estimate identity switches from the reference-angle gap.

Video assessment likewise includes outcomes absent from this calculation. VBench describes dimensions for evaluating generated video (Huang et al., 2023), and FVD addresses a distributional video comparison (Unterthiner et al., 2018). FID, LPIPS and DISTS concern other feature-based or perceptual comparisons (Heusel et al., 2017; Zhang et al., 2018; Ding et al., 2020). No such score is calculated from an annotation angle, because the required image or video outputs are not evaluated here.

Face identity representations such as FaceNet and ArcFace highlight another missing link (Schroff et al., 2015; Deng et al., 2018). A vehicle's observation angle cannot stand in for an identity-embedding distance, and evidence from these vehicle annotations does not estimate human face recognition. The reference-budget question is shared, but the measured objects and outcomes remain different.

Subject-conditioned generation approaches, including ConsisID, ConsistI2V and OpenS2V-Nexus, examine identity or consistency through their own conditioning and generation mechanisms (Yuan et al., 2024; Ren et al., 2024; Yuan et al., 2025). We use them to situate the broader reference-information problem. Their empirical performance cannot be inferred from which four annotated vehicle-frame indices our heuristic selects.

7. From a small geometric gain to a meaningful follow-up

The important empirical contrast is between a small continuous improvement and unchanged thresholded coverage. Measurement-scale choices determine how that contrast is expressed (Stevens, 1946). Degrees retain information about movement inside the tolerance band; a coverage fraction discards it. Neither quantity alone establishes a useful gain for an application whose actual outcome is subject fidelity in generated frames.

A stronger follow-up would prespecify a reference-selection policy, supply estimated rather than perfect annotated angles where appropriate, and evaluate actual generated or retrieved images. It would retain later observations for evaluation and document the intended identity and temporal outcomes. Validation must reflect that transfer boundary rather than select a convenient source partition after inspecting the results (Roberts et al., 2017).

The current artifacts retain candidate-window locations, selected indices, source angles and track-level outcomes for both policies. Dataset documentation supports interpreting those fields and the persistence screen (Gebru et al., 2021). Reusable provenance links the angle definition to the downloaded official development kit (Wilkinson et al., 2016). The analysis supports a precise, limited statement: four-angle selection modestly improved a later mean geometric gap here, while the fixed coverage criterion recorded no improvement.

References

1. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
2. Meinhardt, Tim; Kirillov, Alexander; Leal-Taixe, Laura; Feichtenhofer, Christoph. (2021). TrackFormer: Multi-Object Tracking with Transformers. arXiv:2101.02702. https://arxiv.org/abs/2101.02702.
3. Dendorfer, Patrick; Ošep, Aljoša; Milan, Anton; Schindler, Konrad; Cremers, Daniel; Reid, Ian; Roth, Stefan; Leal-Taixé, Laura. (2021). MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. International Journal of Computer Vision 129, 845-881. https://doi.org/10.1007/s11263-020-01393-0.
4. Gonzalez, Teofilo F.. (1985). Clustering to minimize the maximum intercluster distance. Theoretical Computer Science38:293-306. https://doi.org/10.1016/0304-3975(85)90224-5.
5. Heusel, Martin, et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium" *arXiv preprint arXiv:1706.08500*, 2017. https://arxiv.org/abs/1706.08500.
6. Geiger, A.; Lenz, P.; Urtasun, R.. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. 2012 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2012.6248074.
7. Huang, Ziqi, et al. "VBench: Comprehensive Benchmark Suite for Video Generative Models" *arXiv preprint arXiv:2311.17982*, 2023. https://arxiv.org/abs/2311.17982.
8. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
9. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
10. Unterthiner, Thomas, et al. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.
11. Wojke, Nicolai; Bewley, Alex; Paulus, Dietrich. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. arXiv:1703.07402. https://arxiv.org/abs/1703.07402.
12. Fisher, N. I.. (1993). Statistical Analysis of Circular Data. Cambridge University Press. https://doi.org/10.1017/cbo9780511564345.
13. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.
14. Hartley, Richard; Zisserman, Andrew. (2004). Multiple View Geometry in Computer Vision. Cambridge University Press. https://doi.org/10.1017/cbo9780511811685.
15. Geiger, Andreas; Lenz, Philip; Urtasun, Raquel. (2026). The KITTI Vision Benchmark Suite: Tracking Benchmark, development-kit README [Official technical documentation; captured8November2026, captureyearnotoriginalbenchmarkyear]. https://s3.eu-central-1.amazonaws.com/avg-kitti/devkit_tracking.zip.
16. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
17. Yuan, Shenghai; Huang, Jinfa; He, Xianyi; Ge, Yunyuan; Shi, Yujun; Chen, Liuhan; Luo, Jiebo; Yuan, Li. (2024). Identity-Preserving Text-to-Video Generation by Frequency Decomposition. arXiv:2411.17440. https://arxiv.org/abs/2411.17440.
18. Cameron, A. Colin, and Douglas L. Miller. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
19. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
20. Deng, Jiankang, et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.

21. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.

22. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

23. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

24. Zhang, Yifu, et al. "ByteTrack: Multi-Object Tracking by Associating Every Detection Box" *arXiv preprint arXiv:2110.06864*, 2021. https://arxiv.org/abs/2110.06864.

25. Bewley, Alex; Ge, Zongyuan; Ott, Lionel; Ramos, Fabio; Upcroft, Ben. (2016). Simple Online and Realtime Tracking. arXiv:1602.00763. https://arxiv.org/abs/1602.00763.

26. Yuan, Shenghai; He, Xianyi; Deng, Yufan; Ye, Yang; Huang, Jinfa; Lin, Bin; Luo, Jiebo; Yuan, Li. (2025). OpenS2V-Nexus: A Detailed Benchmark and Million-Scale Dataset for Subject-to-Video Generation. arXiv:2505.20292. https://arxiv.org/abs/2505.20292.

27. Weng, Xinshuo, et al. "AB3DMOT: A Baseline for 3D Multi-Object Tracking and New Evaluation Metrics" *arXiv preprint arXiv:2008.08063*, 2020. https://arxiv.org/abs/2008.08063.

28. Ren, Weiming; Yang, Huan; Zhang, Ge; Wei, Cong; Du, Xinrun; Huang, Wenhao; Chen, Wenhu. (2024). ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation. arXiv:2402.04324. https://arxiv.org/abs/2402.04324.

29. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

30. Schroff, Florian, et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.