SAME GEOMETRY, DIFFERENT SCORES: AUDITING COORDINATE AND SCALE DEPENDENCE IN GEOMETRIC VIDEO EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video generation benchmarks increasingly score geometry: they reconstruct camera paths, depth, and object motion from generated frames and score the reconstruction. The axes and scale of a reconstruction are arbitrary conventions, so rotating the axes or rescaling the scene describes the same scene, and a geometry score should not change. We test whether it does: we rotate or rescale saved reconstructions, compare the scores, and trace every change to the code that causes it; for WorldScore we check instead that a reversed camera path scores lower. The audited component of each benchmark fails its test. On simulator ground truth, the range of PDI-Bench's motion score across 256 sampled rotations has a median of 80.7% of the unrotated score (objects with a nonzero score), because the metric takes medians separately along each axis. Rescaling the scene by 10 changes the final MBench scores of 72% to 83% of 3,218 videos and doubles the gap between two generators, because its frame-pair selector adds a distance to an angle; switching the reconstruction model can, through scale alone, reverse the direction of a score change. In controlled translation tests, WorldScore gives full marks to a camera that moves the opposite way. On Dynamic Replica, a comparison of two pipeline versions reverses once both are scored on the same observations. In the three benchmark components, each failure comes from a few lines of code, and minimal code changes pass the tested relations. These tests are cheap, need no new videos, and should be run before geometric scores are used to compare models.

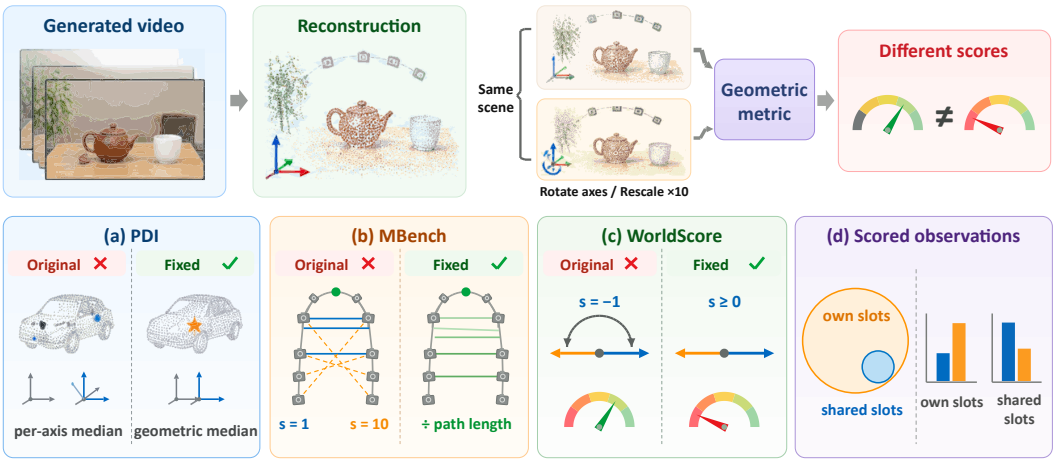


Figure 1: Overview. Rotating or rescaling a reconstruction leaves the scene unchanged, yet PDI and MBench scores change. (a) Per-axis medians make PDI's motion score depend on the axes. (b) MBench's pair selector adds a length to an angle, so rescaling changes the selected pairs. (c) WorldScore's unconstrained scale fit rewards reversed motion with $s = -1$. (d) A comparison can reverse once both sides are scored on the same slots. Each repair passes its test (Section 4).

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.

1 INTRODUCTION

Video generators and interactive world models are now expected to keep scene geometry consistent as the camera moves, objects move, and places are revisited. Recent world models name long-horizon memory and geometric consistency as design goals (Sun et al., 2025; Wu et al., 2026b; Wang et al., 2026b). Benchmarks test this by reconstructing camera paths, depth, and object motion from the generated frames and scoring the reconstruction (Wu et al., 2026a; Zhang et al., 2026; Duan et al., 2025). These geometric scores complement perceptual metrics such as FVD, LPIPS, and VBench (Unterthiner et al., 2018; Zhang et al., 2018; Huang et al., 2024; Liu et al., 2024), and papers use them to compare models.

A reconstruction fixes several quantities by convention rather than by the video. Its world axes are usually anchored to the first camera, and without metric cues its depth and camera translation share an arbitrary scale (Hartley & Zisserman, 2004; Zhang & Scaramuzza, 2018). Rotating those axes, or multiplying depth and translation by a common factor, describes exactly the same scene. A score that measures scene geometry should therefore return the same value under both changes.

This requirement is simple to state but easy to break in code. A median taken separately along each axis, a fixed threshold on depth, or a scale fit without a sign constraint each looks harmless on its own, yet each ties the score to how the reconstruction happens to be written. Audits of metrics for image generation have shown that resizing, feature encoders, and sample size change FID-style scores (Parmar et al., 2022; Kynkäänniemi et al., 2023; Stein et al., 2023; Chong & Forsyth, 2020). Geometric metrics allow a sharper test, because the changes that must leave the score untouched are known exactly, so any score change under them comes from the metric.

We apply this test to three recent benchmarks: the motion component of the Perspective Distortion Index (PDI) in PDI-Bench, the environment-consistency metrics of MBench, and the camera-control component of WorldScore. For each metric we rotate the world frame, rescale the scene, or reverse the camera path, compare the scores, trace every change to the operation that causes it, and test a minimal repair on the same inputs (Figure 1). We also examine a related source of disagreement on Dynamic Replica (Karaev et al., 2023): whether a comparison survives when both sides are scored on the same observations. Our findings, grouped by what the test checks, are:

- **World axes.** On simulator ground truth, PDI's motion score has a median sampled relative range of 80.7% over rotations of the world frame across 1,198 objects with a nonzero score (Figure 2), and on generated videos the gap between two generators varies about 11 times with the choice of axes. Geometric medians make the score identical under every tested rotation.
- **Scale.** MBench chooses frame pairs by adding a camera distance to a rotation angle, so rescaling the scene by 10 changes the final scores of 72% to 83% of its 3,218 videos, lowers generator means by up to 42%, and doubles the gap between two generators. Dividing distances by path length gives every video the same score at every scale from 10^{-3} to 10^{3}.
- **Direction.** WorldScore gives its full camera-control score to a camera that moves in the opposite direction, and across 38,052 perturbed candidates its score rises again once the direction error passes 90 degrees. A nonnegative scale fit no longer rewards reversed motion.
- **Scored observations.** On Dynamic Replica, one version of a reconstruction pipeline looks more accurate when each version is scored on its own observations, and less accurate when both are scored on the same 22,149 observations (Section 4.4).

Each dependence traces to a few lines of code (Table 1), and the checks are cheap enough to serve as unit tests for metrics that score reconstructed geometry.

2 RELATED WORK

Geometric metrics for generated video. PDI-Bench scores perspective realism by reconstructing object trajectories from generated videos and measuring how smoothly they move (Wu et al., 2026a). MBench evaluates the memory of video world models and, in its environment category, checks whether a revisited view is geometrically consistent with the first visit (Zhang et al., 2026). WorldScore evaluates world generation along several axes, including camera control measured against

reference trajectories (Duan et al., 2025). Other geometric metrics detect deformation from residual motion and depth priors (Gu et al., 2025), fit homographies or fundamental matrices between frames (Xue et al., 2026), or evaluate navigation of interactive world models under 6-DoF pose control (Ying et al., 2026). Geometric scores also serve as rewards (Li et al., 2026) and losses (Bai et al., 2025) for training generators, continuing depth-based consistent scene generation (Fridman et al., 2023), and a convention-dependent score used this way passes its dependence on to the model. These works ask what to measure; we ask whether the code measures it independently of conventions, along the dimensions below.

World frame, scale, and direction. Trajectory evaluation for visual odometry already treats the world frame and, for monocular systems, the scale as free, and aligns estimates to the reference before measuring error (Zhang & Scaramuzza, 2018). Its similarity alignment restricts the scale to be nonnegative (Umeyama, 1991). Structure from motion, learned SLAM, and feed-forward reconstruction models inherit the same freedom of frame and scale (Schönberger & Frahm, 2016; Teed & Deng, 2021; Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025; 2026a). Metrics defined purely by relations in image space, such as homographies, do not depend on a reconstruction's frame or scale, so these tests apply once a metric consumes reconstructed cameras, depth, or points.

Scored observations. Point-tracking benchmarks guard against scoring only easy points: the average Jaccard of TAP-Vid counts a visible point marked as occluded as an error (Doersch et al., 2022). The distance-ratio error we examine has no such term, so we ask whether a comparison survives when both sides are scored on the same observations.

Audits of evaluation procedures. Scores such as the Inception score, FID, and KID (Salimans et al., 2016; Heusel et al., 2017; Bińkowski et al., 2018) and their precision and recall variants (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020) depend on choices that users rarely see (Barratt & Sharma, 2018; Borji, 2018). Image resizing and compression shift FID (Parmar et al., 2022), the class content of the feature space shapes what FID rewards (Kynkäänniemi et al., 2023), and the choice of feature encoder changes how well metrics agree with human judgment (Stein et al., 2023). Metamorphic testing checks software by comparing outputs for inputs related in a known way (Xie et al., 2011; Chen et al., 2018), and invariance and directional tests of language models and sanity checks for saliency maps apply similar ideas in machine learning (Ribeiro et al., 2020; Adebayo et al., 2018). Our relations are fixed by geometry: rotation and positive rescaling leave the scene unchanged by construction, so any resulting change in score is an error of the metric, whereas reversing a motion should change it. These audits are instances of the broader gap between a target concept and the procedure that measures it (Jacobs & Wallach, 2021; Bowman & Dahl, 2021); Appendix E compares them with ours by input, relation, observation selection, repair, and evidence.

3 AUDIT METHOD

The audit starts from changes that should not affect a geometric score. Let X be the geometry a metric reads from a reconstruction (camera poses, depth, points, or tracks, with images and intrinsics fixed), $q(X)$ its score, and $O(X)$ its *scored observations*, the object trajectories, frame pairs, or matched points that enter the score, written as a sorted tuple of identifiers. Let G contain the proper rotations of the world frame, applied to all points and poses, and the positive rescalings of depth, points, and camera translation together. A rotation preserves every distance and relative motion, and a rescaling leaves every perspective projection in place, so for every $g \in G$ a score of scene geometry should satisfy

$$|q(gX) - q(X)| \le \tau_{\text{abs}} + \tau_{\text{rel}}|q(X)| \quad \text{and} \quad O(gX) = O(X), \tag{1}$$

where $\tau_{\text{rel}} = 10^{-6}$ and $\tau_{\text{abs}} = 10^{-12}$ absorb floating-point error and cover zero scores. We apply each g to existing reconstructions, so the images and predictions stay fixed and any change in score comes from the scoring code. We test finitely many elements of G: a failure demonstrates a violation, whereas a pass is evidence, not a proof for the whole group. Eq. 1 is a necessary condition for measuring scene geometry, not a full audit of quality; it does not test whether a score is accurate.

The second condition matters because an aggregate can stay the same while its scored observations change, and two scores computed on different observations do not compare like with like. When a

Algorithm 1 Invariance test for a geometric video metric

Input: scoring function q; saved reconstructions $X_1, \ldots, X_n$; tolerances $\tau_{\mathrm{abs}}, \tau_{\mathrm{rel}}$.
Output: for each test, pass or fail, and the first step at which the two runs diverge.

1. Compute $q(X_i)$ and the scored observations $O(X_i)$, such as selected pairs, accepted points, and validity flags.
2. *Rotation.* Draw rotations R uniformly, apply each to all points and camera poses, and require both conditions of Eq. 1 for $g = R$.
3. *Scale.* For s over several orders of magnitude, e.g. $10^{-3}, \ldots, 10^{3}$, multiply depth and camera translation by s, and require the same two conditions.
4. *Direction.* If the metric aligns a predicted trajectory to a reference, reverse the prediction and require a lower score.
5. *Units.* List every constant compared with a length and every sum of terms with different units; scale the constant by s or make the terms dimensionless.
6. On any failure, run both inputs step by step and report the first intermediate quantity that differs.

required measurement is missing, we report no valid score, and ratios with a zero denominator stay undefined.

When Eq. 1 fails, we follow the calculation through the original implementation to the first step at which the transformed and untransformed runs diverge. We then change only that step and rerun the same comparison with the repaired implementation, keeping the matches, missing-value rules, and aggregation of the original code. Negative controls keep the faulty step or alter an unrelated one. WorldScore needs a different relation, because a trajectory and its reversal should remain distinguishable after a positive-scale alignment. Table 1 lists the operations we found, their effect on scores, and the repairs we tested, and Table S19 expands it into a property matrix. Algorithm 1 states the procedure as a test that benchmark maintainers can run on saved reconstructions, without generating any video.

Table 1: Dependencies identified by the audit. Code locations name the responsible operations; the appendix gives source lines and exact changes.

Metric	Convention	Code location	Effect on score	Tested repair	Test of repair
PDI motion	World axes	Per-frame center and three-frame smoother	80.7% median sampled relative range on truth geometry	Geometric medians at both stages	Rotation sweep
MBench environment	Scene scale	Pair selector, lines 82–147; two thresholds	72–83% of 3,218 final scores change at $s = 10$	Path length divides distances and scales thresholds	All videos identical, $s = 10^{-3}$ to 10^{3}
WorldScore camera	Sign of fitted scale	Camera accuracy, line 60	Reversed motion receives full score	Nonnegative scale fit	38,052 turned and noisy candidates

4 RESULTS

4.1 PDI MOTION DEPENDS ON THE ORIENTATION OF THE WORLD FRAME

Two per-axis medians create the dependence. PDI's motion component tracks an object center through the video and scores how smoothly the center moves. The original implementation estimates the center in each frame as the median of the object's reconstructed points, taken separately along each coordinate axis. It then smooths the resulting trajectory with the same per-axis median over three consecutive frames. Speed, acceleration, and the angle between successive velocities are computed afterwards, and none of them depends on the axes. The per-axis medians do depend on them, because each axis can take its median from a different point. Rotating the points changes which points supply the medians and therefore moves the estimated center. A center that rotates together with the data, $c(RX) = Rc(X)$, would not have this problem.

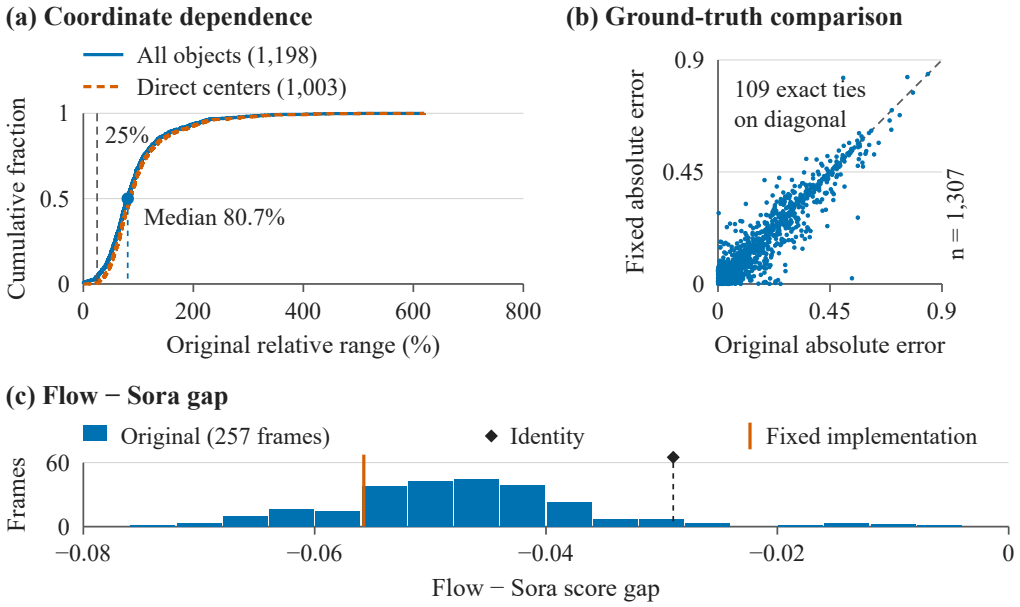


Figure 2: PDI motion score under rotation of the world frame. (a) Sampled relative range (Eq. 2) over 256 rotations for MOVi-E objects with ground-truth geometry, for all objects with a defined range and for objects whose centre is measured directly in every frame. (b) Absolute error against the ground-truth trajectory score for the original and fixed implementations; the fixed implementation is invariant but not closer to the reference. (c) Mean score gap between the Flow and Sora videos in each of 257 coordinate frames under the original implementation; the vertical line marks the fixed implementation, which gives the same gap in every frame.

Simulator ground truth isolates the metric. We scored ground-truth depth, instance masks, and cameras from 98 MOVi-E test records (Greff et al., 2022), which contain 1,307 objects. Each scene was scored in its original frame and in 256 randomly rotated frames, and camera-projection and depth-backprojection checks passed for every record (Appendix A). For each object we measured

$$\text{sampled relative range} = \frac{\max_R q(RX) - \min_R q(RX)}{q(X)}. \tag{2}$$

The ratio is defined for 1,198 objects, and the remaining 109 objects score exactly zero in every frame. The median sampled relative range is 80.7% (quartiles 57.6% and 110.1%; 95% interval $[76.5\%, 83.8\%]$ with records resampled), and 95.7% of objects exceed 25% (Figure 2a). These are sampled ranges: the median grows from 47.5% with 16 rotations to 72.0% with 128 and has not levelled off at 256, so the range over all rotations is at least as large (Appendix A). In score units, rotation moves an object's score by a median of 0.179, close to the spread of scores between objects (interquartile range 0.211), and the unrotated scores are not small (10th percentile 0.116). Restricting the sweep to the 1,003 objects whose centers are measured directly in every frame, which excludes every zero score, raises the median to 84.9%, with 99.4% of them above 25% (Appendix A). Because this input contains no reconstruction error, the dependence comes from the metric itself.

The dependence changes comparisons between generated videos. The same sweep on ten existing reconstructions, six from Flow and Sora videos and four from MOVi-E, gives sampled relative ranges from 12% to 161%. Across the rotated frames, the mean score difference between Flow and Sora ranges from -0.007 to -0.072 (Figure 2c). The sign of the gap never changes, but its size varies about 11 times with nothing except the choice of axes. The repaired implementation gives -0.056 in every frame.

Geometric medians pass the rotation test. The repaired implementation replaces both per-axis medians with the geometric median, computed with modified Weiszfeld iterations (Vardi & Zhang,

2000):

$$c(X) \in \arg\min_{u} \sum_{i} \|x_i - u\|_2. \tag{3}$$

Rotation preserves Euclidean distances, since $\sum_i \|Rx_i - Ru\|_2 = \sum_i \|x_i - u\|_2$ for every u, so the minimizer rotates with the points, and so does every Weiszfeld iterate, including one stopped at the iteration limit (Appendix A). We apply this replacement to the per-frame center and to each three-frame window, and leave the motion formula and the rules for missing centers unchanged. Across all rotated frames, the repaired scores agree to a maximum relative deviation of 2.6×10^{-9} on simulator geometry. The repair makes the score invariant but not more accurate: compared object by object with the ground-truth trajectory, it is no closer than the original implementation (median paired difference 0, 95% interval $[0, 0.0002]$ with records resampled (Cameron & Miller, 2015); Figure 2b).

4.2 MBench depends on reconstruction scale

Joint rescaling leaves projections unchanged. MBench measures environmental consistency from camera trajectories, depth, and image correspondences (Zhang et al., 2026; study-overflow, 2026). Multiplying depth and camera translation by the same positive factor s, with images, intrinsics, rotations, and correspondences fixed, leaves every projection unchanged:

$$\pi\big(K_2\left[R(szK_1^{-1}p) + st\right]\big) = \pi\big(K_2\left[R(zK_1^{-1}p) + t\right]\big), \qquad s > 0, \tag{4}$$

where p is a source pixel with depth z, R and t are the relative rotation and translation, K_1 and K_2 are intrinsics, and π divides by the last homogeneous coordinate.

The frame-pair selector adds a length to an angle. MBench scores each video on up to 20 frame pairs in which the camera returns to an earlier view. To find them, the selector locates the frame where the camera turns back and matches each earlier frame to the later frame that minimizes $d_t + d_r/180$. Here d_t is the distance between the two camera translations and d_r is their relative rotation in degrees. The first term is a length and the second is not, so rescaling the scene changes their relative weight. At large s the selector matches frames mainly by position, and at small s mainly by orientation, so the same video is scored on different frame pairs (Figure 1b).

Two numerical safeguards are also absolute. The reprojection metric keeps a source point only if its depth exceeds 10^{-4}, and the epipolar metric requires the unnormalized fundamental matrix F to have norm above 10^{-12}. These guard against degenerate points and camera pairs rather than define what counts as a consistent revisit, which the metrics judge in pixels. Depth scales with s, and so does F, even though the epipolar geometry $F/\|F\|$ does not change. Scaling both thresholds by s removes this dependence (Table S6 lists the unit of every term and threshold). With the frame pairs held fixed, the pixel statistics and point masks then agree bit for bit at all three tested scales (Appendix B).

Rescaling changes most video scores on the full benchmark. We scored all 3,218 environment videos that MBench provides for six generators, HY-WorldPlay (Sun et al., 2025), Infinite-World (Wu et al., 2026b), LingBot-World (Robbyant Team et al., 2026), Matrix-Game 2.0 and 3.0 (He et al., 2025; Wang et al., 2026b), and Yume-1.5 (Mao et al., 2025), at seven scales from $s = 10^{-3}$ to 10^3, using VGGT reconstructions (Wang et al., 2025). Every scale other than $s = 1$ changes the scores, pixel statistics, or scored frame pairs of 77% to 90% of the videos, and at $s = 10$ the final score itself changes for 2,310 videos (71.8%) in the epipolar metric and 2,664 (82.8%) in the reprojection metric. For HY-WorldPlay, Infinite-World, and Matrix-Game 3.0, the epipolar score itself changes for more than 98% of the videos (Figure S2). The mean epipolar score falls by 8.7% for HY-WorldPlay, 17.5% for Infinite-World, 17.9% for Yume-1.5, and 41.5% for Matrix-Game 3.0 at $s = 10$, all with 95% intervals that exclude zero, and Matrix-Game 3.0 has lost 68% at $s = 10^3$ (Figure 3). Because the falls differ, the gap between HY-WorldPlay and Matrix-Game 3.0 doubles at $s = 10$, from 0.142 (95% interval $[0.115, 0.169]$) to 0.295 ($[0.263, 0.325]$), although the order of the six generators stays the same.

Exchanging the reconstruction model moves the scale. Reconstructing the same videos with Pi3X (Wang et al., 2026a) gives camera paths 1.4 to 3.1 times as long as VGGT's (median per

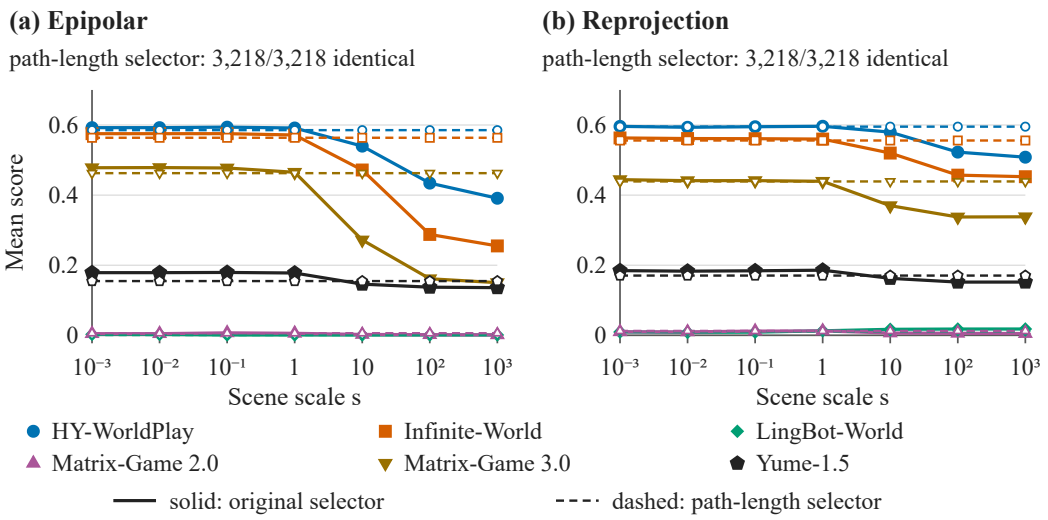


Figure 3: MBench scores when depth and camera translation are multiplied by s, over all 3,218 environment videos (mean over each generator's scorable videos). Solid lines: original frame-pair selector. Dashed lines: selector with the translation distance divided by the path length and thresholds scaled by s; every video receives the same score at all seven scales.

generator), with trajectories that agree after a similarity alignment to within a median of 6% of the path length. Rescaling the Pi3X reconstructions to VGGT's path length changes the original score of 61% to 87% of the videos, so part of the cross-model difference is the scale convention alone. For Matrix-Game 3.0 that part is -0.103 ($[-0.122, -0.084]$) of the epipolar mean, more than the whole cross-model change of -0.068: the original code reports that the score falls with Pi3X, whereas the path-length selector below reports that it rises by 0.032 ($[0.014, 0.049]$) and removes the scale part exactly for every video (Appendix B).

The selector, not the thresholds, drives the change. Crossing fixed pairs (chosen at $s = 1$) or re-selected pairs with original or scaled thresholds separates the two sources: with fixed pairs, the original thresholds change 4 of 3,218 reprojection scores at $s = 0.1$ and no generator mean by more than 10^{-6}, whereas re-selected pairs with scaled thresholds still change 2,864 videos at $s = 10$.

A dimensionless selector passes the scale test. We divide d_t by the path length of the same trajectory, $L = \sum_k \|t_{k+1} - t_k\|$, in both steps of the selector, so that both terms are dimensionless, and keep the scaled thresholds. With these two changes, every one of the 3,218 videos receives bit-for-bit identical scores, frame pairs, and support at all seven scales, in both metrics (Figure 3). Scaling the thresholds by s uses the audit's known factor; tying them to L instead, $z > 10^{-4}L$ and $\|F\| > 10^{-12}L$, needs no s and gives identical results (Appendix B). This repair restores invariance, and at $s = 1$ it stays within 0.023 of every original generator mean. Dividing instead by the largest distance from the first camera or by the root-mean-square spread of the camera positions is equally invariant and keeps the order of the six generators in both metrics.

4.3 WorldScore rewards reversed camera motion

An unconstrained scale fit can absorb the wrong direction. WorldScore compares predicted camera motion with a reference trajectory after fitting a translation scale whose sign the released code leaves unrestricted (Duan et al., 2025). A candidate with camera centers $-C_t$ for reference centers C_t moves the opposite way, yet a fitted scale of $s = -1$ maps it exactly onto the reference and removes both the translation and rotation errors. A change of units is always positive, so a negative scale hides a motion error rather than describing the same motion in other units. For the forward, backward, left, and right commands of the benchmark's camera-construction code, every reversed trajectory indeed receives the full score of one, while restricting the scale to nonnegative

Table 2: WorldScore camera-control score as the candidate's motion turns away from the reference, with noise of 0.1 path extents on the camera centers (median over 1,800 candidates per angle; Static: candidates that do not translate).

Direction error	0°	30°	60°	90°	120°	150°	180°	Static
Negative fitted scale	0%	0%	0%	43%	100%	100%	100%	—
Score, original fit	0.92	0.84	0.74	0.69	0.73	0.82	0.90	0.69
Score, nonnegative fit	0.92	0.84	0.74	0.69	0.69	0.69	0.69	0.69

values lowers these scores to 0, 0, 0.71, and 0.71 and leaves correct and rescaled copies at one (Appendix C).

Past 90 degrees, the score rises as the direction gets worse. The failure does not need an exact reversal: we turned the candidate's motion away from the reference by 0 to 180 degrees, added noise, and included the two orbit commands, for 38,052 candidates. Under the original fit, the median score falls from 0.92 to 0.69 as the error grows to 90 degrees and then climbs back to 0.90 at 180 degrees, because from 120 degrees on every fit takes a negative scale (Table 2). The score is therefore almost uncorrelated with the direction error (Spearman $\rho = -0.03$ without noise), while the nonnegative fit keeps it at its lowest value from 90 degrees on ($\rho = -0.68$). That value is the score of a static camera: without noise, from 120 degrees on the constrained optimum is $s = 0$ and every constrained score equals the static one exactly, so moving the wrong way earns no more than not moving. The constraint only prevents the sign of the scale from absorbing a reversal; it does not make the score decrease with direction error, and beyond 90 degrees it no longer separates one wrong direction from another. In the four cached predictions we examined, the fitted scales were positive, so this defect did not affect their scores.

4.4 The scored observations can decide a comparison

Two versions of one reconstruction pipeline. Dynamic Replica provides a 900-frame sequence of moving articulated objects with ground-truth tracks (Karaev et al., 2023). We compare two versions of a point-tracking reconstruction that differ only in their image tracks. The predicted-track version uses CoTracker3 tracks estimated from the video (Karaev et al., 2024), and the supplied-track version uses the dataset's tracks. Both versions use the same predicted depth, estimated intrinsics, and raster sampler. Each version scores planned slots, where a slot is one query pair at one later frame. For a pair with true distance d_t and reconstructed distance $\widehat{d}_t$ at frame t, the error

$$e_t = \left| \frac{\widehat{d}_t}{\widehat{d}_0} - \frac{d_t}{d_0} \right| \tag{5}$$

measures how well the change in separation is recovered, with the true ratio taken from the dataset's three-dimensional tracks. A slot is scored only when both points have a reconstructed sample and $\widehat{d}_0 > 10^{-6}$; every other slot stays unscored rather than receiving a value. Neither version observes the error on all planned slots, the quantity of interest, so we report three narrower summaries and treat the shared-slot median as the like-for-like comparison: the *own-slot* median over the slots each version scores, the *shared-slot* median over the slots both versions score, and a *frame-weighted* median that weights each scored slot by the inverse of the number of slots scored at its frame.

Scoring the same observations reverses the comparison. In group 2, the predicted-track version scores 22,149 of about 1.81 million planned slots (1.2%), whereas the supplied-track version scores 74.5% (Figure 4a). On its own slots, the predicted-track version has the lower median error, 0.105 against 0.181. On the 22,149 shared slots the order flips, 0.093 for the supplied-track version against 0.105 (difference -0.012, 95% interval $[-0.022, -0.004]$ over query pairs; Figure 4b), and the frame-weighted median also puts the supplied-track version ahead, 0.183 against 0.212, with a wide interval (Table S17). The shared slots are easier than the rest: the supplied-track version's median error is 0.093 on them but 0.181 across all the slots it scores. The ordering thus depends on the support and the summary: the shared-slot medians do not estimate accuracy on all planned slots, but they remove the difference in which slots are compared, and that difference alone reverses the

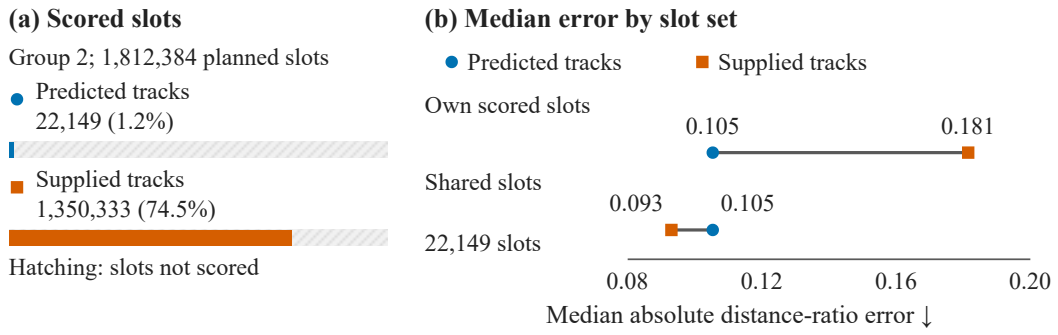


Figure 4: Dynamic Replica, group 2. The two versions of the reconstruction differ only in their image tracks (predicted or supplied by the dataset) and share predicted depth and intrinsics. (a) Fraction of planned pair/frame slots that each version scores. (b) Median error on each version's own scored slots and on the 22,149 slots both versions score; circles mark the predicted-track version and squares the supplied-track version. The saved outputs are identical in both summaries.

result. The predicted-track slots cluster early: 85% of them lie in the first quarter of the frames, where errors are smallest, and within that quarter the shared slots again favor the supplied tracks, 0.075 against 0.087 (Appendix D). In group 0 both versions score most slots (1.19 and 1.28 million, with 1.19 million shared), and the comparison does not reverse (Appendix D).

5 Discussion

The three benchmarks fail at different steps: estimating object centers in PDI, choosing frame pairs and testing numerical validity in MBench, and aligning camera paths in WorldScore. Each step assumes a particular coordinate frame or unit, although the reconstruction fixes neither, so a metric can be wrong even when the reconstruction that feeds it is perfect. Benchmark code can test rotations and positive rescaling on saved inputs without regenerating any video, and the scale test on all 3,218 saved MBench reconstructions runs in minutes on a CPU.

Where else to look. The three failures take forms that are common in geometry code: coordinatewise medians as robust centers of 3-D points, absolute thresholds on depth, baselines, or matrix norms against degenerate cases, and sums of a translation distance and a rotation angle for ranking poses. Each becomes a dependence once the axes or scale are arbitrary, so other metrics that contain them deserve the same tests.

Report what entered the score. A geometric score summarizes only the observations that survive the metric's validity and selection rules. Reporting the planned number of observations, the number actually scored, and, for comparisons, the score on the shared observations would expose reversals like the one on Dynamic Replica.

Scope. We audited one component of each benchmark. On MBench the six generators score far apart, so rescaling changed the gaps between them but not their order. Models that score closer together, such as a new method and its strongest baseline, are more exposed.

6 Conclusion

Geometric video metrics are meant to measure scenes, yet the audited components of PDI-Bench and MBench change their scores when the same scene is rotated or rescaled, and WorldScore rewards reversed camera motion. Each failure comes from a small piece of scoring code, and a minimal repair passes the tested relation, which is necessary for a valid score but not proof of one. Invariance tests and reports of the scored observations should become routine; Appendix F gives reference code.

ETHICS STATEMENT

This work audits evaluation code and involves no human participants or new human annotations. All inputs are public: MOVi-E simulator scenes, the Dynamic Replica dataset, and the generated videos and camera trajectories released with PDI-Bench, MBench, and WorldScore. Our changes to the benchmark implementations alter only how scores are computed. Geometric scores that depend on arbitrary conventions can mislead model selection and published comparisons, and the tests we propose are meant to catch such errors before they propagate.

REPRODUCIBILITY STATEMENT

Section 3 defines the transformations and the invariance criterion. For each metric, Appendices A–C give the code locations, the exact changes, the rotation and scale settings, and the numerical tolerances. Appendix D defines the Dynamic Replica slots and reports both summaries in full, and Appendix G lists where each public input can be obtained. The supplementary material contains the audit code, diffs of every change to the benchmark code, the per-object and per-video result tables, and the figure scripts, with a guide that maps each figure and table to the files that produce it. The figures can be regenerated from these tables without rerunning reconstruction.

AI USE DISCLOSURE

The research idea, the experimental design, and the writing of the paper were led by the authors, and all figures were drawn by the authors. OpenAI Codex and Anthropic Claude Code served as assistants in the research, helping with code implementation, running experiments, data analysis, reference checks, and language editing. The authors reviewed all AI-assisted output and take full responsibility for the manuscript, code, and results.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. GeoVideo: Introducing geometric regularization into video generation model. *arXiv preprint arXiv:2512.03453*, 2025.

Shane Barratt and Rishi Sharma. A Note on the Inception Score. *arXiv preprint arXiv:1801.01973*, 2018. URL https://arxiv.org/abs/1801.01973.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.01401.

Ali Borji. Pros and Cons of GAN Evaluation Measures. *arXiv preprint arXiv:1802.03446*, 2018. URL https://arxiv.org/abs/1802.03446.

Samuel R. Bowman and George Dahl. What Will it Take to Fix Benchmarking in Natural Language Understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. doi: 10.18653/v1/2021.naacl-main.385. URL https://aclanthology.org/2021.naacl-main.385/.

A. Colin Cameron and Douglas L. Miller. A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi: 10.3368/jhr.50.2.317. URL https://doi.org/10.3368/jhr.50.2.317.

Tsong Yueh Chen, Fei-Ching Kuo, Huai Liu, Pak-Lok Poon, Dave Towey, T. H. Tse, and Zhi Quan Zhou. Metamorphic Testing: A Review of Challenges and Opportunities. *ACM Computing Surveys*, 2018. URL https://vuir.vu.edu.au/36919/.

Min Jin Chong and David Forsyth. Effectively Unbiased FID and Inception Score and Where to Find Them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chong_Effectively_Unbiased_FID_and_Inception_Score_and_Where_to_Find_CVPR_2020_paper.html.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A benchmark for tracking any point in a video. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2022.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A Unified Evaluation Benchmark for World Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Duan_WorldScore_A_Unified_Evaluation_Benchmark_for_World_Generation_ICCV_2025_paper.html.

Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. SceneScape: Text-driven consistent scene generation. In *Advances in Neural Information Processing Systems*, 2023.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A Scalable Dataset Generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2203.03570.

Leslie Gu, Junhwa Hur, Charles Herrmann, Fangneng Zhan, Todd Zickler, Deqing Sun, and Hanspeter Pfister. GeCo: Evaluating geometric consistency for video generation via motion and structure. *arXiv preprint arXiv:2512.22274*, 2025.

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. URL https://www.robots.ox.ac.uk/~vgg/hzbook/.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385, 2021. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent Dynamic Depth From Stereo Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. URL https://arxiv.org/abs/2305.02296.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Fréchet Inception Distance. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=4oXTQ6m_ws8.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024. URL https://europe.naverlabs.com/research/publications/grounding-image-matching-in-3d-with-mast3r/.

Zizun Li, Haoyu Guo, Runzhe Teng, Chunhua Shen, and Tong He. Geo-Align: Video generation alignment via metric geometry reward. *arXiv preprint arXiv:2605.23903*, 2026.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. EvalCrafter: Benchmarking and Evaluating Large Video Generation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_EvalCrafter_Benchmarking_and_Evaluating_Large_Video_Generation_Models_CVPR_2024_paper.html.

David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. URL https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. *arXiv preprint arXiv:2512.22096*, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable Fidelity and Diversity Metrics for Generative Models. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://proceedings.mlr.press/v119/naeem20a.html.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Parmar_On_Aliased_Resizing_and_Surprising_Subtleties_in_GAN_Evaluation_CVPR_2022_paper.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing Generative Models via Precision and Recall. In *Advances in Neural Information Processing Systems*, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/f7696a9b362ac5a51c3dc8f098b73923-Abstract.html.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, 2016. URL https://arxiv.org/abs/1606.03498.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0bc795afae289ed465a65a3b4b1f4eb7-Abstract.html.

study-overflow. MBench: Released environment metrics and public-video examples. https://github.com/study-overflow/MBench, 2026. Released implementation, accessed September 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html.

Shinji Umeyama. Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573. URL https://doi.org/10.1109/34.88573.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards Accurate Generative Models of Video: A New Metric & Challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Yehuda Vardi and Cun-Hui Zhang. The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426, 2000. doi: 10.1073/pnas.97.4.1423. URL https://doi.org/10.1073/pnas.97.4.1423.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. In *International Conference on Learning Representations*, 2026a. URL https://github.com/yyfz/Pi3.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Jiaxin Wu, Yihao Pi, Yinling Zhang, Yuheng Li, and Xueyan Zou. Quantitative Video World Model Evaluation for Geometric-Consistency. *arXiv preprint arXiv:2605.15185*, 2026a. URL https://arxiv.org/abs/2605.15185.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. *arXiv preprint arXiv:2602.02393*, 2026b.

Xiaoyuan Xie, Joshua W. K. Ho, Christian Murphy, Gail Kaiser, Baowen Xu, and Tsong Yueh Chen. Testing and validating machine learning classifiers by metamorphic testing. *Journal of Systems and Software*, 84(4):544–558, 2011. doi: 10.1016/j.jss.2010.11.920. URL https://www.cs.columbia.edu/wp-content/uploads/sites/7/2016/08/jss2011.pdf.

Yifei Xue, Yuanchen Fei, Hao Zhang, Chenzhi Nie, Tie Ji, and Yizhen Lao. Illusion or integrity? geometrical consistency metric for AIGC video quality evaluation. *arXiv preprint arXiv:2608.09594*, 2026.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. WBench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Shengjun Zhang, Zhang Zhang, Simin Huang, Zhenyu Tang, Hanyang Wang, Chensheng Dai, Min Chen, Yifan Li, Yuxin Li, Yingjie Chen, Hao Liu, Chen Li, Jing Lyu, and Yueqi Duan. MBench: A Comprehensive Benchmark on Memory Capability for Video World Models. *arXiv preprint arXiv:2606.00793*, 2026. URL https://arxiv.org/abs/2606.00793.

Zichao Zhang and Davide Scaramuzza. A Tutorial on Quantitative Trajectory Evaluation for Visual(-Inertial) Odometry. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018. doi: 10.1109/IROS.2018.8593941. URL https://doi.org/10.1109/IROS.2018.8593941.

A PDI IMPLEMENTATION AND COORDINATE-SWEEP DETAILS

The motion component estimates an object center in each frame from its masked 3-D points and then smooths the resulting trajectory. The per-frame center is `np.median(valid_pts, axis=0)`, a median taken separately along each coordinate. The smoother loops over the three coordinates and applies a three-frame median filter to each one separately. The fixed implementation replaces the per-frame center with the geometric median of the same points, and the smoother with the geometric median of each edge-padded three-frame neighborhood. The speed, acceleration-norm, adjacent-velocity-cosine calculations, masks, frame rates, and missing-center rules remain unchanged.

Modified Weiszfeld iterations start at the mean and account for coincident points with their multiplicities (Vardi & Zhang, 2000). Iterations stop when displacement falls below 10^{-10} times the point cloud's root-mean-square radius, or after 500 iterations. Both the stopping rule and the iteration are rotation-equivariant: the radius is rotation invariant, and each step is a distance-weighted mean of the points, so the returned iterate rotates with the points whether or not the limit is reached. The geometric median is unique unless all points are collinear, and the solver returns the same deterministic iterate in that case as well. We draw 256 rotations uniformly (Haar measure) with SciPy and seed 20260924, and add the original frame for 257 frames in total. Points are stored as float64 row vectors and rotated as $XR^\top$ before center estimation.

The maximum relative deviation compares each fixed score with the fixed score in the original coordinates. The acceptance threshold is 10^{-6}, while constant-zero cases use an exact zero check. A coordinatewise-median negative control fails the rotation relation on the synthetic test. The geometric-median control satisfies the same relation, testing the operation used in the replacement.

A.1 COMPLETE RECONSTRUCTION SWEEP

Table S1: All coordinate-sweep cases. Range is over identity and 256 rotations, divided by the respective identity score. Δ_I is the fixed identity value's percentage change from Original. Fixed deviations are maxima relative to fixed identity.

Case	Original	Range (%)	Fixed	Δ_I (%)	Deviation
Flow 1	0.302542	45.1	0.260620	-13.9	3.43×10^{-13}
Flow 2	0.196085	17.8	0.183630	-6.4	2.80×10^{-14}
Flow 3	0.167343	22.3	0.161755	-3.3	1.23×10^{-13}
Sora 1	0.189773	18.2	0.189111	-0.3	2.17×10^{-14}
Sora 2	0.366364	23.4	0.381234	+4.1	9.46×10^{-15}
Sora 3	0.196833	12.4	0.202916	+3.1	1.21×10^{-13}
MOVi-E 0	0.118650	44.7	0.112694	-5.0	2.07×10^{-14}
MOVi-E 1	0.162487	52.9	0.170480	+4.9	4.71×10^{-14}
MOVi-E 2	0.231828	63.4	0.167462	-27.8	5.20×10^{-14}
MOVi-E 3	0.136152	160.6	0.144119	+5.9	3.91×10^{-14}

Table S2: Metadata-matched Flow-minus-Sora component comparisons. Each original and fixed comparison has 0/256 sign reversals relative to its own identity value.

Prompt label	Original identity	Original rotated range	Fixed identity
red vintage car	0.11277	[0.04107, 0.18091]	0.07151
high-speed train	-0.17028	[-0.21735, -0.12250]	-0.19760
yellow school bus	-0.02949	[-0.05612, -0.00864]	-0.04116

The generated-video comparison retains three Flow and three Sora clips with their existing reconstruction outputs. The other four inputs are MOVi-E reconstructions, using the same rotations and unchanged motion calculation. Sora 2 retains 37 carried centers under the existing missing-center rule. The aggregate Flow-minus-Sora difference ranges from -0.07233 to -0.00667 under the original implementation. The fixed implementation gives a difference of -0.05575 to numerical precision.

A.2 SIMULATOR DATA AND COORDINATE CHECKS

MOVi-E test shards 0–24 contain the 98 records and 1,307 objects used here (Greff et al., 2022). The camera-projection check compares projected true centers with the dataset's image positions. Each record requires median error below 0.5 pixels and a 99th percentile below 1.0 pixel. The depth-backprojection check uses ray distances and normalized inverse-intrinsic rays through pixel centers. It requires at least 95% of masked points inside the true oriented boxes expanded by 5% of their diagonal. Both checks pass in 98/98 records, with no record excluded.

The original initial-mask rule and the carry rule for frames with at most ten valid points remain unchanged. The geometric-median solver reaches its iteration limit in 150,602 of 14,476,810 calls. The reported invariance test includes the resulting terminal iterates without replacing them.

A.3 OBJECT-LEVEL RESULTS AND GROUND-TRUTH TRAJECTORIES

The ground-truth trajectory contains the simulator's true object centers for each frame. The reference score applies the fixed temporal smoother and unchanged motion calculation to that trajectory. The paired error difference subtracts the fixed implementation's absolute reference error from the original implementation's absolute reference error. The confidence interval uses 10,000 whole-record bootstrap replicates with seed 20260924 (Cameron & Miller, 2015). Each replicate retains all objects within each sampled record, including repeated records.

Table S3: MOVi results for all objects and for the objects whose centers are measured directly in every frame. Relative ranges are undefined when the unrotated score is zero; those objects stay in the error comparison.

	All objects	Direct centers
Objects / records	1,307 / 98	1,003 / 98
Original implementation under rotation		
Defined / undefined relative ranges	1,198 / 109	1,003 / 0
Median sampled relative range (quartiles)	80.7% (57.6, 110.1)	84.9% (64.2, 115.3)
Objects above 5%, 10%, 25%	1,185; 1,177; 1,147	1,003; 1,002; 997
Median absolute range (quartiles)	0.179 (0.121, 0.265)	0.179 (0.122, 0.266)
Unrotated score, median (10th percentile)	0.202 (0.116)	0.182 (0.115)
Fixed implementation under rotation		
Maximum relative deviation	2.6×10^{-9}	2.6×10^{-9}
Maximum absolute deviation	4.7×10^{-10}	4.7×10^{-10}
Absolute error against the ground-truth trajectory		
Median, original / fixed	0.094 / 0.086	0.065 / 0.059
Median paired difference (95% interval)	0 (0, 0.0002)	0.0013 (-0.0007, 0.0034)

The direct-center subset contains the 1,003 objects whose center is estimated from their own points in every frame, rather than carried over from an earlier frame. We defined it after the analysis of all objects, and it keeps objects from all 98 records. All 109 objects with constant-zero scores have at least one carried center.

A.4 NUMBER OF ROTATIONS AND UNCERTAINTY

The sampled relative range in Eq. 2 is a maximum minus a minimum over the sampled frames, so it can only grow as rotations are added. Table S4 recomputes it from the saved scores for the first k of the 256 rotations, always including the original frame; the identity scores and ranges reproduce the frozen values exactly. The median is still rising between 128 and 256 rotations, so the reported values describe the sampled range and the range over all rotations is at least as large. The denominator is the unrotated score, which has 5th, 10th, 25th, and 50th percentiles of 0.110, 0.116, 0.137, and 0.202 among the 1,198 defined objects, so the large relative ranges do not come from near-zero denominators. For $k = 256$, a bootstrap over the 98 records (10,000 replicates, seed 20260926, all objects of a drawn record kept) gives 95% intervals of $[76.5\%, 83.8\%]$ for the median relative range of the defined objects and $[81.9\%, 88.3\%]$ for the direct-center objects.

Table S4: Median sampled relative range of the original PDI motion score and share of objects above 25% as a function of the number of sampled rotations k.

k	Defined objects (1,198)		Direct centers (1,003)	
	Median	Above 25%	Median	Above 25%
1	13.4%	25.2%	14.4%	27.0%
4	29.7%	63.0%	30.8%	66.7%
16	47.5%	89.3%	49.8%	94.3%
64	65.4%	93.7%	68.7%	98.2%
128	72.0%	94.8%	75.7%	99.0%
256	80.7%	95.7%	84.9%	99.4%

B MBENCH DETAILS

B.1 SCORING PIPELINE AND CODE LOCATIONS

We use the released environment-consistency code at commit c44dbd9 (study-overflow, 2026) and load its calculation module unchanged. For each video, 32 frames are sampled uniformly and reconstructed once with VGGT-1B (Wang et al., 2025), which gives camera extrinsics, intrinsics, and depth. The selector then chooses up to 20 frame pairs in which the camera returns to an earlier view. It first finds the turnaround frame, the frame that maximizes $d_t + d_r/180$ relative to the first frame, and then matches each earlier frame to the later frame that minimizes the same quantity; if either side of the turnaround has fewer than five frames, it falls back to fixed symmetric pairs. Image correspondences come from scale-invariant feature matches (Lowe, 2004), which do not depend on the reconstruction. The epipolar metric scores the matches against the fundamental matrix of each pair, and the reprojection metric projects source pixels into the target frame with depth and relative pose; each produces a per-video reliability score from the frame-pair errors, and a video for which the metric returns no valid score is unscorable. Rescaling multiplies depth and camera translation by s on the CPU after reconstruction, so every scale uses the same frames, network outputs, and matches. Table S5 lists every line we change.

Table S5: Lines changed in the MBench environment-calculation module.

Source line	Original operation	Tested repair
82–147	Selector criterion $d_t + d_r/180$ (turnaround and matching)	$d_t/L + d_r/180$, with path length $L = \sum_k \|t_{k+1} - t_k\|$
260	Source depth $z > 10^{-4}$	$z > s \cdot 10^{-4}$
161	$\|F\| > 10^{-12}$ before normalization	$\|F\| > s \cdot 10^{-12}$ (and the matching degeneracy check)
271	Target depth $z_{\mathrm{proj}} > 10^{-6}$	Unchanged
218–219	10^{-12} added to squared line norms after normalization	Unchanged (dimensionless)

Table S6 gives the unit of every quantity in the selector and every constant compared with a reconstructed quantity. We separate the terms that define what the metric measures, which decide whether two views count as a revisit and how large an error is, from numerical safeguards, which only exclude degenerate points or camera pairs. Under joint rescaling a length is multiplied by s, while angles, pixels, counts, and normalized quantities are unchanged. A quantity passes the unit check if it is dimensionless or pixel-valued, or if every constant it is compared with scales with it.

B.2 POINT-LEVEL DIAGNOSIS OF THE SOURCE-DEPTH THRESHOLD

The source-depth threshold first appeared in one video whose fixed-pair reprojection error changed from 145.57 to 144.53 pixels at $s = 0.1$ while the number of valid pairs stayed the same. At $s = 0.1$, three backward matches have positive depths below the threshold and are dropped, although they lie in front of both cameras and project inside the target image (Figure S1). Dropping them changes the pixel errors behind the frame-pair median, not the set of valid pairs. Scaling the source-depth

Table S6: Units in the MBench environment metrics. Change at $s = 1$: effect of the tested repair on the unscaled benchmark. $L \approx 0$: behavior when the camera barely translates.

Quantity	Unit	Under s	Tested repair and reason	Change at $s = 1$	$L \approx 0$ or degenerate input
Terms that define the measurement					
Camera distance d_t in the selector	length	$\times s$	d_t/L: a length divided by a length of the same path	generator means within 0.023	$L = 0$ only if the camera never moves; then $d_t = 0$ and the division is skipped; no video has $L < 10^{-9}\times$ median
Rotation angle $d_r/180$	degrees / 180	fixed	none needed	none	unaffected
Pixel errors and their reliability mapping	pixels	fixed	none needed	none	unaffected
Pair count (up to 20) and fallback (five frames)	count	fixed	none needed	none	unaffected
Numerical safeguards					
Source depth $z > 10^{-4}$ (line 260)	length	$\times s$	$s \cdot 10^{-4}$ or $L \cdot 10^{-4}$: the constant must scale with depth	none (identical at $s = 1$)	points at or behind the camera still rejected
$\|F\| > 10^{-12}$ (line 161)	$\propto$ length	$\times s$	$s \cdot 10^{-12}$ or $L \cdot 10^{-12}$: $F = K_2^{-\top}[t]_\times R K_1^{-1}$ is linear in t	none (identical at $s = 1$)	pairs without translation still rejected
Target depth $z_{\text{proj}} > 10^{-6}$ (line 271)	length	$\times s$	left unchanged; changes no output for $10^{-3} \le s \le 10^{3}$	none	could bind below the tested range, where scaling it is the consistent choice
10^{-12} in line norms (lines 218–219)	none	fixed	none needed	none	prevents division by zero

threshold restores identical point masks and pixel statistics at all three scales, whereas scaling only the target-depth threshold leaves the discrepancy in place. The unrounded projection coordinates still differ by at most 2.8×10^{-13} pixels, which the released rounding removes. Synthetic points in front of the camera trigger the same rejection, so the effect does not require a poor reconstruction.

B.3 Full-benchmark results

The full run covers all 547 environment scenes and 3,218 videos; Infinite-World and Yume-1.5 provide 490 and 540 of the scenes, and every video downloaded and passed its size and checksum test. Generator means average the reliability score over the videos that are scorable at that scale, and unscorable videos are counted separately rather than set to zero. Intervals resample scenes 10,000 times and keep all videos of a scene together. Rankings use the scenes that all six generators can score at all three scales (448 epipolar, 412 reprojection). Table S7 lists the scores at each scale. Rescaling by 10 lowers the means of HY-WorldPlay, Infinite-World, Matrix-Game 3.0, and Yume-1.5 in both metrics, with intervals that exclude zero. The scores of LingBot-World and Matrix-Game 2.0 are close to zero at every scale, and their intervals include zero. Rescaling by 0.1 changes 77% of the videos but moves every generator mean by at most 0.012. Across the 15 generator pairs, no difference changes sign with separated intervals at either scale; the only point-estimate change is between LingBot-World and Matrix-Game 2.0 in the reprojection metric at $s = 0.1$, whose scores differ by 0.003.

B.4 Separating thresholds from selection

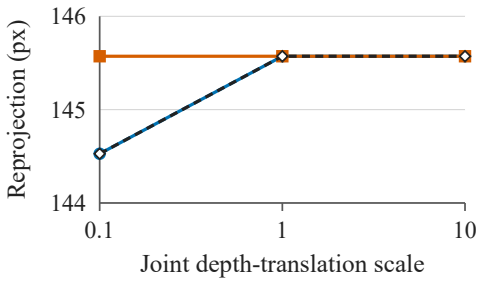
We cross fixed pairs (chosen at $s = 1$) or re-selected pairs with the original or scaled thresholds, which gives four conditions on the same 3,218 videos. The re-selected, original-threshold condition

(a) Three matches cross the absolute source-depth gate

Pair / match	z at scale 1	z at scale 0.1	Included
7/29, bwd #0	8.536e−04	8.536e−05	yes → no
7/29, bwd #3	8.536e−04	8.536e−05	yes → no
8/27, bwd #3	5.695e−04	5.695e−05	yes → no

Source gate: z > 10^−4; frame-pair identities remain fixed.

(b) Matched threshold control and negative control



Scenario	Mask changes
Absolute source gate 144.527 / 145.572 / 145.572 px	3
Source gate scaled 145.572 / 145.572 / 145.572 px	0
Target gate scaled 144.527 / 145.572 / 145.572 px	3

Matched points are compared within the same frame pairs.

Figure S1: An absolute source-depth threshold changes the scored matches. Panel (a) shows the positive depths that cross the original threshold under joint depth and translation rescaling. Panel (b) compares the resulting frame-pair aggregates with the fixed implementation. Scaling the source-depth threshold restores equal reported statistics across scales. Scaling only the target-depth threshold retains the discrepancy and supplies the negative control.

Table S7: MBench scores on all 3,218 environment videos. Means are over each generator's scorable videos at that scale. Δ_{10} is the change from $s = 1$ to $s = 10$ with a scene-resampled 95% interval, and U counts the unscorable videos at $s = 0.1/1/10$. The last column uses the path-length selector of Section 4.2 with scaled thresholds and is the same at every scale.

	Original selector					Fixed
Generator	$s = 0.1$	$s = 1$	$s = 10$	Δ_{10}	U	all s
Epipolar						
HY-WorldPlay	0.594	0.591	0.540	-0.052 $[-0.062, -0.042]$	0/0/0	0.585
Infinite-World	0.575	0.572	0.472	-0.100 $[-0.116, -0.084]$	0/0/0	0.564
LingBot-World	0.001	0.001	0.001	0.000 $[-0.001, 0.000]$	0/0/0	0.001
Matrix-Game 2.0	0.007	0.006	0.002	-0.003 $[-0.008, +0.001]$	15/17/32	0.006
Matrix-Game 3.0	0.477	0.465	0.272	-0.193 $[-0.214, -0.174]$	0/0/0	0.463
Yume-1.5	0.179	0.178	0.146	-0.032 $[-0.043, -0.022]$	1/2/1	0.155
Reprojection						
HY-WorldPlay	0.595	0.596	0.580	-0.017 $[-0.025, -0.009]$	0/0/0	0.595
Infinite-World	0.561	0.560	0.520	-0.040 $[-0.051, -0.029]$	0/0/0	0.556
LingBot-World	0.009	0.013	0.017	$+0.004$ $[0.000, +0.009]$	1/2/9	0.011
Matrix-Game 2.0	0.012	0.012	0.007	-0.006 $[-0.013, 0.000]$	25/29/52	0.012
Matrix-Game 3.0	0.441	0.439	0.370	-0.070 $[-0.084, -0.055]$	0/0/0	0.439
Yume-1.5	0.184	0.186	0.163	-0.023 $[-0.033, -0.014]$	4/4/3	0.170

reproduces the original MBench scores exactly, and fixed pairs with scaled thresholds give bitwise identical results at all three scales. With fixed pairs and original thresholds, 4 reprojection videos change at $s = 0.1$ and none at $s = 10$; with re-selected pairs, 2,487 to 2,489 videos change at $s = 0.1$ and 2,864 at $s = 10$, whether or not the thresholds are scaled. Table S8 splits the change in

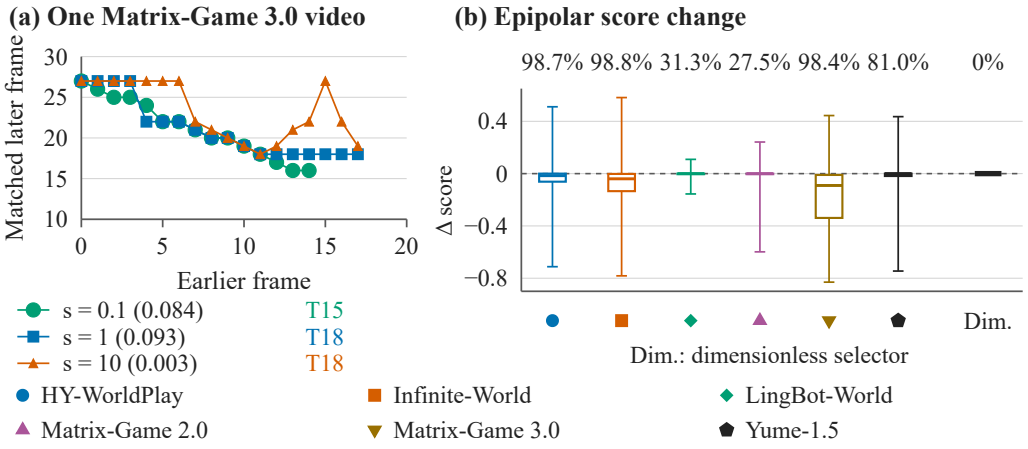


Figure S2: How rescaling changes MBench's frame pairs. (a) Frame pairs that the original selector picks for one Matrix-Game 3.0 video at three scales (the video with the median score change among this generator's changed videos); each point matches an earlier frame to a later frame, and the legend gives the video's epipolar score. (b) Change in each video's epipolar score from $s = 1$ to $s = 10$ under the original selector, by generator; numbers give the fraction of videos whose score changes. The dimensionless selector changes no score.

each generator mean at $s = 10$ into a threshold effect, a selection effect, and their interaction. The threshold effect and the interaction are zero to three decimals for every generator, and the selection effect equals the total change.

Table S8: Threshold and selection effects on the mean score from $s = 1$ to $s = 10$, with scene-resampled 95% intervals. The threshold effect compares fixed pairs with original and scaled thresholds; the selection effect compares re-selected and fixed pairs with scaled thresholds.

Generator	Threshold effect	Selection effect	Interaction
Epipolar			
HY-WorldPlay	0.000 [0.000, 0.000]	−0.052 [−0.062, −0.042]	0.000
Infinite-World	0.000 [0.000, 0.000]	−0.100 [−0.116, −0.084]	0.000
LingBot-World	0.000 [0.000, 0.000]	0.000 [−0.001, 0.000]	0.000
Matrix-Game 2.0	0.000 [0.000, 0.000]	−0.003 [−0.008, +0.001]	0.000
Matrix-Game 3.0	0.000 [0.000, 0.000]	−0.193 [−0.214, −0.174]	0.000
Yume-1.5	0.000 [0.000, 0.000]	−0.032 [−0.043, −0.022]	0.000
Reprojection			
HY-WorldPlay	0.000 [0.000, 0.000]	−0.017 [−0.025, −0.009]	0.000
Infinite-World	0.000 [0.000, 0.000]	−0.040 [−0.051, −0.029]	0.000
LingBot-World	0.000 [0.000, 0.000]	+0.004 [0.000, +0.009]	0.000
Matrix-Game 2.0	0.000 [0.000, 0.000]	−0.006 [−0.013, 0.000]	0.000
Matrix-Game 3.0	0.000 [0.000, 0.000]	−0.070 [−0.084, −0.055]	0.000
Yume-1.5	0.000 [0.000, 0.000]	−0.023 [−0.033, −0.014]	0.000

B.5 Dimensionless selector

The fixed selector divides the translation distance by a length L computed from the same video, in both the turnaround search and the matching step, and runs with the scaled thresholds; every other step is unchanged. We tested three choices of L: the path length $\sum_k \|t_{k+1} - t_k\|$, the largest distance from the first camera, and the root-mean-square spread of the camera positions. If the camera never translates, every choice gives $L = 0$ and every d_t is zero, so the division is skipped without effect; no MBench video has $L = 0$ or L below 10^{-9} times the median. Each choice makes the full per-video output, which contains the score, pixel statistics, selected pairs, match counts,

and scorability, bitwise identical at all seven scales from 10^{-3} to 10^{3} for all 3,218 videos in both metrics (Table S9). Each choice orders the six generators exactly as the original selector does at $s = 1$, with no pairwise sign change. The path-length choice stays closest to the original scores, within 0.023 of every generator mean (Table S11), because the median path length of the VGGT trajectories is 0.81, so dividing by it changes the original weighting only slightly; the original means barely move for $s \leq 1$ and fall once $s > 1$ (Table S12). The three normalizers differ in how much position counts against orientation: the median extent is 0.24 and the median RMS spread 0.082, so they weight position about 3 and 10 times more than the path length does, and their generator means are correspondingly lower (Table S10). Each choice fixes this trade-off independently of the reconstruction's scale; which trade-off best identifies a revisit is a design decision for the benchmark, and we report the path length as one tested repair that stays close to the released scores. The scaled thresholds use the known factor s of each test, so the repair establishes invariance under the tested joint rescalings from 10^{-3} to 10^{3}, not a scorer that needs no knowledge of scale; tying the thresholds to L removes that need (next subsection), and the unscaled target-depth threshold changes no output within this range. The run of all conditions at seven scales takes about 30 minutes on 32 CPU cores.

Table S9: Three dimensionless normalizations of the translation distance, each with scaled thresholds. Identical: videos whose full output is bitwise identical at all seven scales (epipolar/reprojection, of 3,218). τ_b and sign changes compare the order of the six generators with the original selector at $s = 1$. Max $|\Delta|$: largest change of a generator mean from the original at $s = 1$.

Normalizer L	Identical	τ_b	Sign changes	Max $\lvert\Delta\rvert$
$\sum_k \lVert t_{k+1} - t_k \rVert$ (path length)	3,218/3,218	1.00/1.00	0	0.023
$\max_k \lVert t_k - t_0 \rVert$ (extent)	3,218/3,218	1.00/1.00	0	0.171
RMS spread of t_k	3,218/3,218	1.00/1.00	0	0.266

Table S10: Generator means under the original selector and the three normalizers. Original: mean at $s = 1$ and its range (maximum minus minimum) over the seven scales 10^{-3} to 10^{3}. Each normalizer gives the same mean at every scale, so its range is zero and one value describes the whole grid.

	Original selector		Distance divided by		
Generator	$s = 1$	Range	Path length	Extent	RMS spread
Epipolar					
HY-WorldPlay	0.591	0.203	0.585	0.518	0.456
Infinite-World	0.572	0.320	0.564	0.400	0.317
LingBot-World	0.001	0.002	0.001	0.001	0.001
Matrix-Game 2.0	0.006	0.006	0.006	0.002	0.002
Matrix-Game 3.0	0.465	0.328	0.463	0.310	0.199
Yume-1.5	0.178	0.043	0.155	0.143	0.137
Reprojection					
HY-WorldPlay	0.596	0.088	0.595	0.567	0.536
Infinite-World	0.560	0.111	0.556	0.494	0.469
LingBot-World	0.013	0.010	0.011	0.015	0.016
Matrix-Game 2.0	0.012	0.007	0.012	0.012	0.008
Matrix-Game 3.0	0.439	0.107	0.439	0.392	0.353
Yume-1.5	0.186	0.034	0.170	0.157	0.153

B.6 Thresholds from the reconstruction itself

Scaling the two thresholds by s uses the factor that the audit applied, which a scorer never knows. We therefore also tied them to the path length of the trajectory being scored, $z > 10^{-4}L$ and $\lVert F \rVert > 10^{-12}L$ (also in the matching degeneracy check), with the same path-length selector and every other step unchanged; if $L = 0$ the thresholds become zero and the selector division is skipped. On all 3,218 videos this version gives bitwise identical output at all seven scales in both metrics, and at $s = 1$ its output equals that of the s-scaled version for every video, so every generator mean and ranking is unchanged. Table S13 runs four degenerate camera paths through the three

Table S11: Path-length selector minus original selector at $s = 1$: change in each generator mean with scene-resampled 95% intervals.

Generator	Epipolar	Reprojection
HY-WorldPlay	-0.006 $[-0.010, -0.002]$	-0.001 $[-0.004, +0.002]$
Infinite-World	-0.008 $[-0.012, -0.004]$	-0.004 $[-0.009, +0.001]$
LingBot-World	0.000 $[-0.001, 0.000]$	-0.002 $[-0.003, 0.000]$
Matrix-Game 2.0	0.000 $[-0.001, +0.002]$	0.000 $[-0.004, +0.004]$
Matrix-Game 3.0	-0.002 $[-0.011, +0.006]$	0.000 $[-0.006, +0.005]$
Yume-1.5	-0.023 $[-0.033, -0.014]$	-0.015 $[-0.024, -0.008]$

Table S12: Original MBench generator means over seven scales (scorable videos at each scale).

Generator	10^{-3}	10^{-2}	10^{-1}	1	10	10^{2}	10^{3}
Epipolar							
HY-WorldPlay	0.592	0.592	0.594	0.591	0.540	0.434	0.391
Infinite-World	0.575	0.575	0.575	0.572	0.472	0.288	0.255
LingBot-World	0.003	0.003	0.001	0.001	0.001	0.001	0.001
Matrix-Game 2.0	0.005	0.005	0.007	0.006	0.002	0.002	0.001
Matrix-Game 3.0	0.478	0.479	0.477	0.465	0.272	0.161	0.150
Yume-1.5	0.179	0.179	0.179	0.178	0.146	0.137	0.136
Reprojection							
HY-WorldPlay	0.596	0.593	0.595	0.596	0.580	0.523	0.508
Infinite-World	0.563	0.561	0.561	0.560	0.520	0.457	0.452
LingBot-World	0.009	0.008	0.009	0.013	0.017	0.018	0.018
Matrix-Game 2.0	0.010	0.010	0.012	0.012	0.007	0.006	0.005
Matrix-Game 3.0	0.444	0.441	0.441	0.439	0.370	0.338	0.338
Yume-1.5	0.184	0.183	0.184	0.186	0.163	0.152	0.152

implementations. A static camera and pairs without relative translation leave the epipolar metric without valid pairs and the reprojection metric scorable in all three, at every scale. For a real path scaled by 10^{-9} with depth unchanged, the released code's epipolar result depends on the units (0, 2, or 16 valid pairs), whereas the version tied to L keeps all 16 pairs at every scale. Because L measures the path rather than the scene, this version accepts pairs whose baseline is tiny compared with the scene depth; judging degeneracy by baseline relative to depth would be a further design choice, which we did not test. The run takes about 25 minutes on 32 CPU cores.

Table S13: Degenerate camera paths built from one MBench video's frames, depth, and intrinsics. Valid pairs: epipolar pairs that pass the checks at $s = 10^{-3}/1/10^{3}$. Scores at $s = 1$. Original: released code; the other two rows use the path-length selector. Identical: full output bitwise identical over seven scales (epipolar/reprojection). No run raised an exception or produced a non-finite value.

Case	Thresholds	Valid pairs	Epipolar	Reprojection	Identical
Static camera, rotating	Original	0/0/0	—	0.604	yes/yes
	Thresholds $\times s$	0/0/0	—	0.604	yes/yes
	Thresholds $\times L$	0/0/0	—	0.604	yes/yes
Path scaled by 10^{-9}	Original	0/2/16	0.799	0.604	no/yes
	Thresholds $\times s$	2/2/2	0.825	0.593	yes/yes
	Thresholds $\times L$	16/16/16	0.743	0.593	yes/yes
Only last two frames move	Original	4/4/4	0.877	0.604	yes/yes
	Thresholds $\times s$	2/2/2	0.840	0.518	yes/yes
	Thresholds $\times L$	2/2/2	0.840	0.518	yes/yes
Zero-baseline pairs	Original	0/0/0	—	0.005	yes/yes
	Thresholds $\times s$	0/0/0	—	0.005	yes/yes
	Thresholds $\times L$	0/0/0	—	0.005	yes/yes

B.7 SECOND RECONSTRUCTION MODEL

We reconstructed the same 3,218 videos from the same 32 frames with Pi3X (Wang et al., 2026a), recovering intrinsics from its predicted rays, and wrote the output into the bundle format that the MBench scorer reads; only the reconstruction network differs. After a similarity alignment of camera centers, the Pi3X and VGGT trajectories differ by a median of 6% of the VGGT path length, and no video exceeds 25%. The pose convention was checked with the MBench reprojection error on 12 videos: the Pi3X error matches VGGT's (median ratio 0.88), and swapping world-to-camera and camera-to-world poses worsens it 1.7 to 35 times in all 7 videos with a usable reconstruction. The pre-specified threshold of 9 of 12 videos was missed because both models fail on the other 5 videos (errors above 135 pixels), where the swap carries no information; we recorded this deviation before reading any score. Reconstruction took about one GPU-hour. Table S14 splits the change from VGGT to Pi3X into the part caused by scale alone, obtained by rescaling each Pi3X reconstruction to VGGT's path length, and the rest. Under the original code the scale part can exceed the whole change and have the opposite sign to the geometric part, as for Matrix-Game 3.0 and Yume-1.5 in the epipolar metric; under the path-length selector it is exactly zero for every video, and the remaining change is the part that does not come from scale; since the trajectories still differ by a median of 6% of the path length after alignment, we do not read it as an improvement in geometry. The epipolar ranking of the six generators is the same for both models; in the reprojection metric one adjacent pair changes order between VGGT and Pi3X whether or not Pi3X is rescaled, so that change does not come from scale.

Table S14: Exchanging VGGT for Pi3X on all 3,218 MBench videos. ρ: median ratio of Pi3X to VGGT path length. Changed: videos whose original score or support differs between Pi3X at its own scale and Pi3X rescaled to VGGT's path length. Original total: Pi3X minus VGGT under the original code; scale part: the share of that change due to scale alone; path-length total: the same change under the path-length selector, whose scale part is exactly zero. Intervals resample scenes 10,000 times.

Generator	ρ	Changed	Original total	Scale part	Path-length total
Epipolar					
HY-WorldPlay	2.58	67%	$+0.035$ $[+0.023, +0.048]$	-0.009 $[-0.014, -0.005]$	$+0.049$ $[+0.037, +0.062]$
Infinite-World	3.08	85%	$+0.027$ $[+0.016, +0.038]$	-0.015 $[-0.023, -0.009]$	$+0.038$ $[+0.027, +0.049]$
LingBot-World	1.41	61%	0.000 $[-0.001, +0.001]$	-0.001 $[-0.002, 0.000]$	$+0.001$ $[-0.001, +0.002]$
Matrix-Game 2.0	1.37	69%	0.000 $[-0.006, +0.007]$	$+0.003$ $[-0.001, +0.007]$	-0.004 $[-0.009, +0.001]$
Matrix-Game 3.0	2.93	87%	-0.068 $[-0.091, -0.043]$	-0.103 $[-0.122, -0.084]$	$+0.032$ $[+0.014, +0.049]$
Yume-1.5	2.68	81%	-0.009 $[-0.019, 0.000]$	-0.017 $[-0.025, -0.010]$	$+0.016$ $[+0.006, +0.026]$
Reprojection					
HY-WorldPlay	2.58	67%	$+0.060$ $[+0.049, +0.072]$	-0.004 $[-0.007, -0.001]$	$+0.066$ $[+0.055, +0.078]$
Infinite-World	3.08	85%	$+0.049$ $[+0.040, +0.059]$	-0.006 $[-0.012, -0.001]$	$+0.054$ $[+0.044, +0.064]$
LingBot-World	1.41	61%	$+0.010$ $[+0.002, +0.019]$	$+0.003$ $[0.000, +0.006]$	$+0.007$ $[-0.001, +0.016]$
Matrix-Game 2.0	1.37	69%	$+0.012$ $[+0.001, +0.024]$	$+0.006$ $[0.000, +0.013]$	$+0.003$ $[-0.007, +0.012]$
Matrix-Game 3.0	2.93	87%	$+0.051$ $[+0.033, +0.070]$	-0.037 $[-0.049, -0.026]$	$+0.083$ $[+0.069, +0.098]$
Yume-1.5	2.68	81%	$+0.001$ $[-0.008, +0.011]$	-0.009 $[-0.017, -0.002]$	$+0.022$ $[+0.012, +0.032]$

C WORLDSCORE SCALE-FIT DETAILS

The camera-accuracy function creates an unrestricted scalar variable at source line 60. The tested change replaces `cp.Variable()` with `cp.Variable(nonneg=True)` at that location. The camera construction, rotation comparison, optimization objective, and score normalization remain unchanged. All four source translation commands use 49 frames and retain the source rotations. The candidate centers are unchanged, multiplied by two, or multiplied by minus one.

All 24 optimization calls return an optimal result from the same solver, and none uses the fallback. For reversed motion the fixed fit returns a scale of zero, the boundary of the nonnegative domain. A zero scale places every fitted camera center at the first one, so a reversed path scores exactly like a candidate that does not translate, and the score then reflects how far the reference itself moves.

With the fixed fit, the reversed forward and backward trajectories have translation error 0.500 and the reversed lateral trajectories 0.250. The score normalization maps these two errors to 0 and 0.707.

Table S15: Complete camera comparison. Correct and rescaled candidates keep near-unit scores in both implementations.

Motion	Candidate	Scale, original	Scale, fixed	Score, original	Score, fixed
Forward	Correct	1	1	1.000	1.000
Forward	Rescaled ($\times 2$)	0.5	0.5	1.000	1.000
Forward	Reversed	-1	0	1.000	0.000
Backward	Correct	1	1	1.000	1.000
Backward	Rescaled ($\times 2$)	0.5	0.5	1.000	1.000
Backward	Reversed	-1	0	1.000	0.000
Left	Correct	1	1	1.000	1.000
Left	Rescaled ($\times 2$)	0.5	0.5	1.000	1.000
Left	Reversed	-1	0	1.000	0.707
Right	Correct	1	1	1.000	1.000
Right	Rescaled ($\times 2$)	0.5	0.5	1.000	1.000
Right	Reversed	-1	0	1.000	0.707

For the four cached predictions, each trajectory is expressed relative to its first camera. The original fits give scales of 11.30, 12.52, 13.71, and 6.23, all positive, and camera scores of 0.994, 0.875, 0.984, and 1.000. The nonnegative fit changes these scores by at most 3.5×10^{-13}.

Stress test. The stress test uses all six camera commands of the benchmark's camera-construction code that translate the camera: push in, pull out, move left, move right, and the two orbits, each over 49 frames. For every reference, a candidate turns each camera displacement from the first camera by an angle $\theta \in \{0^\circ, 30^\circ, \ldots, 180^\circ\}$ about one of two axes perpendicular to the net motion, keeps the camera rotations, scales the displacements by 0.5, 1, or 2, and adds Gaussian noise with standard deviation $\sigma \in \{0, 0.05, 0.1, 0.2\}$ times the reference path extent (50 seeds for each $\sigma > 0$). This gives 38,052 candidates, each scored with both fits; all 76,104 optimizations return an optimal solution from the same solver, and none uses a fallback. Positive amplitude changes move the score by at most 6.8×10^{-8} under the original fit, as expected for a scale-aligned metric. Table S16 gives the results for every noise level. At every level, the share of negative fitted scales jumps from zero below 90° to 100% from 120° on, the original score is lowest at 90° and recovers towards its value at 0° as the motion reverses, and the nonnegative fit stays at its minimum. For both orbits, a 180° turn scores 0.91 or 1.00 under the original fit, depending on the axis, and 0.69 under the nonnegative fit. A static candidate, whose camera centers all stay at the first reference center, scores a median of 0.69 under both fits (1,836 fits with and without noise); without noise, the nonnegative fit gives every candidate turned by 120° or more exactly the static score for all six commands, which separates correct and rescaled motion (score 1.00) from motion that is no better than standing still.

Additional controls. We scored two more candidate families for the same six commands with both fits, after reproducing all 72 noise-free 180° scores of the stress test exactly; all 108 new fits return an optimal solution. A candidate that follows the reference for the first 24 of its 48 steps and then retraces its path receives a fitted scale of 1.00 under both fits and a score of 0.70 for push in and pull out, 0.86 for the lateral moves, and 0.85 for the orbits. The constraint does not act on this candidate, and its score lies between that of a correct candidate and that of a static one, which reflects that the nonnegative fit blocks sign absorption without calibrating the score to partial direction errors. When the reference itself does not translate, the original fit returns a scale of zero and a translating candidate scores 1.00 under both fits, as does a static candidate, so for such references the translation part of the score cannot separate moving from standing still and only the rotations are compared.

D Dynamic Replica observation identities and complete results

Each group contains 64 query points, all 2,016 unordered pairs of them, and the 899 frames after the query frame. Each group therefore has 1,812,384 planned slots. A slot is identified by its group, query pair, and frame index, and the shared slots are the exact intersection of these identities between the two versions. Both versions use the same reference distance ratio and the error in Equation 5.

Table S16: WorldScore stress test: median camera-control score and share of negative fitted scales by direction error and noise level, pooled over six camera commands, two rotation axes, and three amplitudes.

Direction error	0°	30°	60°	90°	120°	150°	180°
Noise 0 path extents, 36 candidates per angle							
Negative scale (%)	0	0	0	17	100	100	100
Original fit	1.00	0.86	0.74	0.69	0.74	0.84	1.00
Nonnegative fit	1.00	0.86	0.74	0.69	0.69	0.69	0.69
Noise 0.05 path extents, 1,800 candidates per angle							
Negative scale (%)	0	0	0	42	100	100	100
Original fit	0.96	0.85	0.74	0.69	0.73	0.84	0.94
Nonnegative fit	0.96	0.85	0.74	0.69	0.69	0.69	0.69
Noise 0.1 path extents, 1,800 candidates per angle							
Negative scale (%)	0	0	0	43	100	100	100
Original fit	0.92	0.84	0.74	0.69	0.73	0.82	0.90
Nonnegative fit	0.92	0.84	0.74	0.69	0.69	0.69	0.69
Noise 0.2 path extents, 1,800 candidates per angle							
Negative scale (%)	0	0	0	45	100	100	100
Original fit	0.84	0.79	0.72	0.69	0.72	0.79	0.83
Nonnegative fit	0.84	0.79	0.72	0.69	0.69	0.69	0.69

Table S17: Dynamic Replica summaries with coverage and uncertainty. Pairs of values are predicted / supplied tracks; coverage is the share of the 1,812,384 planned slots that a version scores. Differences are supplied minus predicted, so a negative value favors the supplied tracks, with 95% intervals from 2,000 paired bootstrap replicates over query pairs.

Group	Summary	Coverage	Median	Median difference	Mean	Mean difference
0	Own slots	65.8% / 70.5%	0.0345 / 0.0359	+0.0014 [+0.0010, +0.0019]	0.0570 / 0.0660	+0.0090 [+0.0057, +0.0131]
0	Shared slots	65.8% / 65.8%	0.0345 / 0.0347	+0.0002 [−0.0001, +0.0006]	0.0570 / 0.0577	+0.0007 [−0.0008, +0.0023]
0	Frame-weighted	—	0.0368 / 0.0384	+0.0016 [+0.0012, +0.0021]	0.0597 / 0.0690	+0.0094 [+0.0058, +0.0138]
2	Own slots	1.2% / 74.5%	0.105 / 0.181	+0.076 [+0.054, +0.094]	0.464 / 0.335	−0.129 [−0.418, +0.074]
2	Shared slots	1.2% / 1.2%	0.105 / 0.093	−0.012 [−0.022, −0.004]	0.464 / 0.200	−0.264 [−0.555, −0.069]
2	Frame-weighted	—	0.212 / 0.183	−0.029 [−0.247, +0.025]	0.772 / 0.334	−0.438 [−1.305, −0.032]

Table S17 reports the three summaries of Section 4.4 with the coverage, median, and mean of each version and paired intervals over query pairs. In group 2 the predicted-track version scores no pair at all in 21 frames; these frames stay in the planned total and receive no weight in the frame-weighted summary, since an unscored frame has no error to weight. The group-2 ordering depends on the estimand: the own-slot median favors the predicted tracks, whereas the shared-slot median and mean and the frame-weighted mean favor the supplied tracks with intervals that exclude zero, and the own-slot mean and frame-weighted median have intervals that include zero. In group 0 the two versions score almost the same slots, the predicted-track version is ahead on every summary, and both shared-slot intervals include zero. The frame-weighted median interval comes from a separate bootstrap with the same design (seed 20260925). Query pairs share query points, so these intervals describe variation within this one sequence and are approximate; they are not population intervals over independent scenes.

Difficulty strata. Table S18 stratifies group 2 by frames since the query and by the true initial pair separation. The predicted-track coverage falls from 4.1% in the first quarter of the frames to at most 0.3% later, so 18,737 of its 22,149 scored slots lie in the first quarter. In that quarter the own-slot medians favor the predicted tracks and the shared-slot medians favor the supplied tracks, with an interval that excludes zero; the later quarters contain 655 to 1,580 shared slots each, too few for a precise comparison. Intervals come from 2,000 bootstrap replicates over query pairs.

Table S18: Dynamic Replica group 2 by difficulty stratum. Pairs of values are predicted / supplied tracks. The difference is supplied minus predicted on the shared slots, with a 95% interval over query pairs.

Stratum	Coverage	Own-slot median	Shared slots	Shared-slot median	Difference
Frames since query, quartiles					
Q1	4.1% / 76.7%	0.087 / 0.153	18,737	0.087 / 0.075	$-0.013\ [-0.017, -0.009]$
Q2	0.3% / 76.1%	0.195 / 0.205	1,177	0.195 / 0.270	$+0.075\ [-0.020, +0.252]$
Q3	0.1% / 71.4%	0.213 / 0.172	655	0.213 / 0.221	$+0.008\ [-0.045, +0.283]$
Q4	0.3% / 73.8%	0.197 / 0.200	1,580	0.197 / 0.218	$+0.021\ [-0.031, +0.103]$
Initial separation, quartiles					
Q1	1.6% / 74.2%	0.230 / 0.320	7,139	0.230 / 0.250	$+0.021\ [-0.034, +0.076]$
Q2	1.3% / 64.6%	0.094 / 0.197	5,899	0.094 / 0.072	$-0.022\ [-0.041, -0.009]$
Q3	1.2% / 76.1%	0.078 / 0.136	5,591	0.078 / 0.070	$-0.008\ [-0.015, +0.005]$
Q4	0.8% / 83.0%	0.061 / 0.153	3,520	0.061 / 0.052	$-0.009\ [-0.013, -0.005]$

E Property matrix and relation to prior audits

Table S19 collects, for each audited component, what was transformed, on which inputs, with which tolerance, and what the original and repaired implementations returned. It is meant as a regression record: a maintainer who changes one of these components can rerun the same rows.

Table S20 places the audit next to earlier work that checks evaluation procedures through relations the output should satisfy. The geometric benchmarks discussed in Section 2 propose metrics; the audit does not propose a metric but tests three existing ones against relations fixed by the geometry of reconstruction.

Table S19: Property matrix of the audited components: expected relation, transformed input, observation identity, tolerance, original and repaired values, and open limitations.

	PDI motion	MBench epipolar and reprojection	WorldScore camera
Expected relation	score unchanged under rotation of the world frame	score, frame pairs, and support unchanged under joint rescaling of depth and translation	a reversed camera path scores lower than the correct one; positive rescaling leaves the score unchanged
Transformed input	256 Haar-random rotations of the world frame and the identity	joint rescaling, $s = 10^{-3}, \ldots, 10^{3}$ (seven values)	positive rescaling of the candidate ($\times 0.5$, 1, 2); reversal and turns of 0° to 180°
Input support	98 MOVi-E records, 1,307 objects (1,198 with a nonzero score); ten reconstructions	3,218 videos from 547 scenes and six generators; VGGT and Pi3X reconstructions	six translating commands; 38,052 perturbed candidates; four cached predictions
Observation identity	same objects and frames by construction (masks, point counts, and carried centers do not depend on rotation)	selected pairs, match counts, and scorability compared bitwise	every frame is scored; no selection
Score tolerance	$\tau_{\mathrm{abs}} + \tau_{\mathrm{rel}}\lvert q\rvert$, $\tau_{\mathrm{rel}} = 10^{-6}$; exact zero for constant-zero scores	bitwise identity of the full per-video output	scale: same tolerance; direction: reversal must lower the score
Original value	median sampled relative range 80.7% (95% interval 76.5% to 83.8%)	final scores change for 71.8% (epipolar) and 82.8% (reprojection) of videos at $s = 10$; outputs or support change for 77–90% at every $s \neq 1$	reversed motion scores 1.00 for all four commands; median 0.90 at 180° with noise
Repaired value	maximum relative deviation 2.6×10^{-9} over all rotations	3,218 of 3,218 videos bitwise identical at all seven scales	reversed motion scores 0 or 0.71; the median stays at the static score from 90° on
Change at the original frame or scale	identity score changes by -27.8% to $+5.9\%$ on ten reconstructions; error against truth unchanged (median paired difference 0)	generator means change by at most 0.023; thresholds identical at $s = 1$	none while the fitted scale is positive (at most 3.5×10^{-13} on the cached predictions)
Degenerate input	109 constant-zero objects: relative range undefined and reported separately; iteration limit in 1.0% of solver calls, iterate still rotates with the points	$L = 0$ only for a camera that never moves (division skipped); unscorable videos counted separately, never set to zero	reversed motion: scale 0, scores as a static camera; static reference: translation part cannot separate moving from static candidates
Open limitations	rotation only; the repair is invariant but no closer to the ground truth; a finite sample of rotations	thresholds tied to L accept pairs whose baseline is tiny relative to depth; target-depth threshold unscaled, so invariance is shown for $10^{-3} \leq s \leq 10^{3}$; the choice of L is a design decision	beyond 90° the repaired score no longer separates one wrong direction from another
Backend and precision	NumPy float64 on CPU; rotations from SciPy with a fixed seed	VGGT-1B once per video on GPU; rescaling and scoring on CPU, projections agree to 2.8×10^{-13} pixels before rounding	CVXPY with the Clarabel solver on CPU; every fit optimal, no fallback

Table S20: Relation to earlier audits of evaluation procedures and to relation-based testing.

Work	Input representation	Relation or group tested	Observation selection	Repair	Evidence
Clean-FID (Parmar et al., 2022)	images and their preprocessing	sensitivity to the resizing implementation and to compression	not addressed	standardized preprocessing	score shifts across libraries
ImageNet classes in FID (Kynkäänniemi et al., 2023)	Inception features	matching class statistics without improving images	not addressed	not the focus	class-histogram matching
Flaws of generative metrics (Stein et al., 2023)	feature encoders	agreement with human judgment	not addressed	DINOv2 encoder	human study
CheckList (Ribeiro et al., 2020)	text	invariance and directional tests of models	not addressed	not the focus	test suites
Sanity checks (Adebayo et al., 2018)	images and model weights	explanations should change under randomization	not addressed	not the focus	randomized models
Metamorphic testing (Chen et al., 2018)	program inputs	known input–output relations	not addressed	not the focus	software testing
Trajectory evaluation (Zhang & Scaramuzza, 2018)	camera poses	frame and monocular scale removed by alignment	not addressed	alignment before error	evaluation protocol
This audit	reconstructed cameras, depth, points, and tracks	rotation, positive rescaling, reversal	identifiers of scored observations	local code changes, retested	simulator truth; 3,218 videos; two reconstruction models

F Reference implementation of the invariance tests

The following code implements Algorithm 1 for a scoring function `q` that takes a reconstruction and returns its score together with its scored observations. A reconstruction is a dictionary with world points `points` of shape $(N, 3)$, per-frame `depth`, and world-to-camera poses $x_{\mathrm{cam}} = R_{wc}x + t_{wc}$ stored as `R_wc` of shape $(T, 3, 3)$ and `t_wc` of shape $(T, 3)$. Rotating the world by R replaces x by Rx and R_{wc} by $R_{wc}R^{\top}$ and leaves t_{wc} and depth unchanged; rescaling by s multiplies points, depth, and t_{wc} by s. Both changes leave every camera-frame coordinate, and hence every projection, unchanged up to the factor s in depth. The function `changed` applies the tolerance of Section 3, $\tau_{\mathrm{abs}} + \tau_{\mathrm{rel}}|q|$, so a zero score must stay within τ_{abs} of zero. Every function body has a single level of indentation, and `check_reversal` reflects the camera centers about the first one.

```
import numpy as np
from scipy.spatial.transform import Rotation

def rotate(rec, R):
    out = {**rec, "points": rec["points"] @ R.T}
    out["R_wc"] = rec["R_wc"] @ R.T
    return out

def rescale(rec, s):
    keys = ("points", "depth", "t_wc")
    return {**rec, **{k: s * rec[k] for k in keys}}

def changed(score, obs, score0, obs0, rtol, atol):
    close = np.isclose(score, score0, rtol=rtol, atol=atol)
    return not close or obs != obs0

def check_invariance(q, rec, n_rot=16, scales=(1e-3, 0.1, 10, 1e3),
                     rtol=1e-6, atol=1e-12, seed=0):
    score0, obs0 = q(rec)
    Rs = Rotation.random(n_rot, random_state=seed).as_matrix()
    runs = [("rotation", R, q(rotate(rec, R))) for R in Rs]
    runs += [("scale", s, q(rescale(rec, s))) for s in scales]
    return [(kind, g, score) for kind, g, (score, obs) in runs
            if changed(score, obs, score0, obs0, rtol, atol)]

def check_reversal(q_align, reference, centers):
    flipped = 2 * centers[:1] - centers
    return q_align(reference, flipped) < q_align(reference, centers)
```

`check_invariance` returns the list of failing transformations, which is empty for an invariant metric. The scored observations should be returned as plain, comparable values, for example a sorted tuple of frame-pair indices or a tuple of accepted point indices, so that `obs != obs0` detects a change in what was scored even when the score itself happens to agree. The file `tests/invariance_tests.py` in the supplementary material contains this code together with a self-test: a rotation- and scale-invariant spread statistic passes every check, whereas a coordinatewise median of the same points fails all 16 rotations. Running `python tests/invariance_tests.py` with NumPy and SciPy prints `PASS: invariant metric passes; coordinatewise median fails 16/16 rotations`. Text extracted from this PDF loses the indentation of the listing; the supplementary file is the reference copy. A search of the scoring code for numeric literals that are compared with depths, distances, or matrix norms finds the candidates for the unit step of Algorithm 1; in MBench this search finds both absolute thresholds, and the selector is found by looking for sums of a distance and an angle.

G Data, code, and compute

Data. MOVi-E test shards 0–24 are available through the Kubric release and include depth, instance masks, cameras, and object trajectories (Greff et al., 2022). Dynamic Replica provides the video sequence, image tracks, and three-dimensional tracks through its public dataset release (Karaev et al., 2023). The MBench environment videos of the six generators are distributed

with the benchmark (Zhang et al., 2026), and the PDI-Bench Flow and Sora videos and the WorldScore camera trajectories come from the respective releases (Wu et al., 2026a; Duan et al., 2025).

Code. PDI-Bench, MBench, and WorldScore provide their scoring implementations through the repositories linked by their publications (Wu et al., 2026a; study-overflow, 2026; Duan et al., 2025). All experiments run these implementations with only the local changes listed in Appendices A–C. The supplementary material contains the audit scripts, diffs of our changes to each benchmark, the transformation settings, the per-object, per-video, and per-candidate result tables, and the figure scripts, together with instructions for obtaining the public inputs. Model weights and third-party data remain available from their providers under their distribution terms.

Compute. Reconstructing the 3,218 MBench videos with VGGT-1B takes roughly 40 A100 GPU-hours, including download, with a peak of 10.2 GiB of GPU memory per process. The MBench scoring experiments run on the CPU on these saved reconstructions: the four threshold-by-selection conditions take 6.2 CPU-hours and the dimensionless selector 2.7 CPU-hours. No experiment trains a model.