A Mathematically Rigorous Benchmark for AI Spatial Reasoning

The physical world is modeled as a 4D spacetime manifold with three spatial dimensions and one temporal dimension. For the purposes of spatial reasoning benchmarks, we focus on the geometric structure of the spatial slice, which is Euclidean $\mathbb{R}^3$.

One-dimensional (1D) geometry is essentially trivial from the perspective of transformation groups. Up to isomorphism, there are exactly two connected 1-dimensional Lie groups:

1. The real line $(\mathbb{R}, +)$, representing translations along an infinite line.
2. The circle group $S^1 \cong \mathbb{R}/\mathbb{Z} \cong \{z \in \mathbb{C} : |z| = 1\}$, representing periodic motion or rotations in the plane.

These correspond respectively to the only two simply connected and compact connected 1D manifolds, and they underlie the structure of higher-dimensional rotation and translation groups. This prompts any spatial reasoning benchmark to not require 1D reasoning tests.

Two-dimensional (2D) spatial transformations are governed by the Euclidean group $E(2)$, which consists of all isometries of the plane: translations, rotations, and reflections. Rotations in $\mathbb{R}^2$ form the special orthogonal group $\mathrm{SO}(2)$, a one-parameter Lie group isomorphic to the circle S^1; all such rotations occur about a single axis perpendicular to the plane. Consequently, specifying a 2D rigid motion requires a rotation angle $\theta \in [0, 2\pi)$ and a center of rotation $\mathbf{c} \in \mathbb{R}^2$.

While any orientation-preserving isometry in the plane can be decomposed into a translation and a rotation about the origin, orientability introduces a fundamental topological distinction: the full Euclidean group $E(2)$ has two connected components, corresponding to orientation-preserving ($\det = +1$) and orientation-reversing ($\det = -1$) transformations. In particular, two congruent planar shapes (e.g., left and right hands) may be related by a reflection but not by any element of $\mathrm{SO}(2) \rtimes \mathbb{R}^2$. This chirality in 2D is nontrivial: although mirror images are isometric, they are not related by a proper rigid motion. Thus, a complete 2D spatial reasoning benchmark must evaluate not only metric transformations but also invariance (or equivariance) under orientation-reversing maps.

This is achieved in this test by creating an orientable graph where each edge contains only one direction creating a chiral graph that is congruent with its reflection. By gauging the ability to trace paths, compare angles at intersections and understand directionality, the 2D section of the test successfully captures all the complexity of the 2D manifold.

In three dimensions, the rotation group is $\mathrm{SO}(3)$, a non-Abelian, compact, three-dimensional Lie group. A robust and singularity-free parameterization of $\mathrm{SO}(3)$ is provided by the group of unit quaternions $\mathrm{Sp}(1) \cong \mathrm{SU}(2)$, which double-covers $\mathrm{SO}(3)$. Any proper 3D rigid transformation can thus be represented as a translation $\mathbf{t} \in \mathbb{R}^3$ composed with a rotation encoded by a unit quaternion $q \in \mathbb{H}$, $\|q\| = 1$. A benchmark that requires an agent to infer or apply such a transformation, given partial or occluded observations of a 3D object, tests its ability to reason about the full structure of $\mathrm{SE}(3) = \mathrm{SO}(3) \rtimes \mathbb{R}^3$, the 3D special Euclidean group.

Moreover, visual perception in 3D environments is inherently projective: an embodied agent (or model) receives 2D retinal or camera projections of 3D scenes, resulting in occlusion, perspective distortion, and loss of depth information. A rigorous benchmark must therefore include tasks where the agent must infer hidden structure from partial views, effectively solving an inverse problem in the space of 3D poses modulo the projection map $\pi : \mathbb{R}^3 \to \mathbb{R}^2$. Successfully reasoning about occluded geometry demonstrates not only competency in 3D transformation groups but also in amodal completion, a hallmark of robust spatial cognition.

By creating a test that requires the agent to create such a transformation to choose the correct option in the Shepard-Metzler test aswell as accounting for the projective occlusion, it successfully captures the nuances of 3D vision.