

Self-Consistency for LLM-Based Motion Trajectory Generation and Verification

Supplementary Material

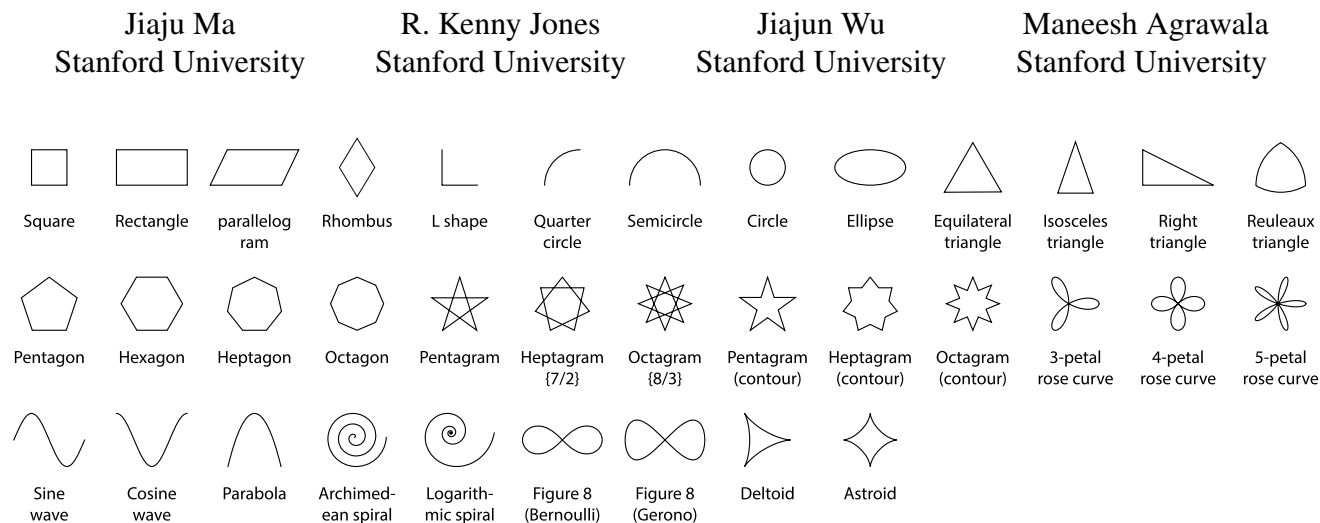


Figure S1. Our benchmark includes 224 prompts of motion trajectories derived from 35 different geometric base shapes. Variations are created by adding specifications to each shape to exercise different warps in our hierarchy of transformations.

Contents

A Motion Trajectory Benchmark Details	1
B Decision Criteria Alternatives	2
C Shape Family Limitations	3
C.1. Shape Families Requiring Multiple Prototypes	3
C.1.1. Extension to Multi-Prototype Shape Families	4
C.2. Shape Families with Anisotropic Similarity	4
D Failure Modes	4
D.1. Decision Criteria Failure Modes	4
D.2. LLM Failure Modes	7
E Method Implementation Details	7
E.1. Computing Warp-Invariant Distance	7
E.2. Sensitivity to DBSCAN Distance Threshold	9
E.3. Generating Motion Trajectories with an LLM	9
E.4. Verifying Motion Trajectories with a VLM	10

A. Motion Trajectory Benchmark Details

Our benchmark consists of 224 motion trajectory prompts describing 35 different basic shapes, as shown in Figure S1.

Each prompt is paired with a ground truth shape family following the definition of Equation 1 in the main paper – one prototype trajectory and a transformation group. Table S1 reports the number of prompts associated with each transformation group. The distribution is determined by the geometric properties of the base shapes (e.g., a circle cannot have anisotropic similarity or affine warps). We discuss and show examples of applying our approach to shape families that do not follow this definition in Section C.1. We illustrate all 224 prompts with their ground truth shape families in our benchmark in Figure S8–S14.

We use a template-based approach to generate the prompts, following MoVer [8] and CLEVR [6]. Each prompt starts with the main sentence `Animate the <SVG element> to move along a path shaped like <shape name>`. We manually pair a prompt to its ground truth shape family consisting of a Lie group of transformations and a prototype trajectory based on the shape properties. For example, we pair “circle” with a circular trajectory and the warp $W_{\text{sim-ref}}$ (similarity with reflections), “ellipse” with an elliptical path and $W_{\text{sim-ani}}$ (anisotropic similarity), and “parallelogram” to a parallelogram path with W_{aff} (affine). We note that, for shapes that are periodic (sine and cosine waves) or

Table S1. Number of prompts per warp in our benchmark.

$W_{\text{rgd-ref}}$	W_{rgd}	W_{sim}	$W_{\text{sim-ref}}$	$W_{\text{sim-ani}}$	W_{aff}	Total
105	44	38	23	12	2	224

self-repeating (Archimedean and logarithmic spirals), the prompts need to specify the number of periods or turns, as changes in these values effectively create different shape families (e.g., a spiral with 2 turns cannot be geometrically warped into a spiral with 3 turns). We add a new sentence after the main sentence for this value: Complete `<periods/turns>` of `<shape name>`.

To introduce prompt variations that exercise all geometric warps in our hierarchy, we add modifiers to a prompt in the form of additional specifications on a shape. The first type of specifications is *shape orientation*, where we require a shape to be traversed in either a clockwise or counterclockwise manner. Adding this specification removes reflections from the warps. We use the following template sentence: Traverse the path in a `<orientation>` manner.

The second type of specifications is *shape size*. For some shapes, their sizes can be fully specified by one parameter (e.g., radius for circles and side length for polygons). Others require two parameters, such as width and height for rectangles and triangles, and horizontal extent and amplitude for sine and cosine waves. Fully specifying a shape’s size reduces the transformation group to either W_{rgd} (rigid) or $W_{\text{rgd-ref}}$ (rigid with reflections). Alternatively, we can specify a shape’s size in terms of ratios (e.g., a rectangle with a 4:3 aspect ratio), and doing so changes the transformation group to either W_{sim} or $W_{\text{sim-ref}}$. We append a prepositional phrase to the main sentence when adding size specifications: a `<shape name>` with `<size param name>` of `<size param value>` (e.g., “a rectangle with a width of 50 px and a height of 80 px.”).

B. Decision Criteria Alternatives

In Section 4 of the main paper we describe how our method is able to recover a shape family from an input prompt. A critical component of this procedure is how to estimate which transformation group W , from our hierarchy of Lie transformation groups (Figure 3, main paper), is most appropriate. We introduce two decision criteria for selecting W , based on underlying assumptions made in the sampled motion trajectory distributions produced by an LLM: Majority-Consensus (Section 4.3.1) and Hierarchical-Consistency (Section 4.3.2). How do we know whether the decision criteria we propose select useful transformation groups? One way to justify our design of these procedures is by comparing their performance against very simple decision criteria alternatives. To this end, we

consider two additional ‘baseline’ decision criteria, which we expect will recover worse shape families compared with Majority-Consensus and Hierarchical-Consistency. *Most-Restrictive* is a criterion that always selects the most restrictive transformation group in our hierarchy (rigid). *Least-Restrictive* is a criterion that always selects the least restrictive transformation group in our hierarchy (affine).

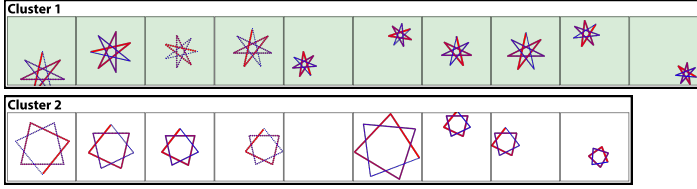
Motion trajectory generation With the same experimental set-up as described in Section 5.2 of the main paper, we compare how the different decision criteria aid in the task of motion trajectory synthesis. We report results of this experiment in Table S2. As a reminder, the metric we track is the “accuracy” of the generations, i.e., whether the production is a member of the ground-truth shape family from our benchmark dataset. For better numerical stability, we compute this accuracy as the ratio of true to false trajectories in the chosen cluster, averaged across tasks in the dataset. Note that for LLM-Direct, we say that the “chosen cluster” includes all trajectories produced from the same prompt.

We find that the two baseline decision criteria do strictly worse compared with the more sophisticated alternatives. Even these simple ways of choosing W do provide some benefit within the self-consistency paradigm in terms of improving the accuracy of the LLM’s generations. For instance, Most-Restrictive is able to increase the accuracy over LLM-Direct by between 3 and 4 absolute percentage points. These results show that even with sub-optimal choices of W , self-consistency can benefit from the greater flexibility afforded in terms of checking whether generations match beyond identity comparisons.

Motion trajectory verification Following the experimental set-up as described in Section 5.3 of the main paper, we compare how these alternative decision criteria perform on the task of motion trajectory verification. We report results of this experiment in Table S3. Hierarchical-Consistency continues to achieve the best F1 score, finding a nice balance between precision and recall. One way to view Least-Restrictive is as a partial implementation of Hierarchical-Consistency; this simplification to never descend the hierarchy results in a steep decrease in precision (15 points) with only moderate increases to recall (3 points). In a similar fashion, Most-Restrictive can also be considered a partial implementation of Majority-Consensus. Always choosing the most restrictive transformation group provides a marginal benefit for precision (.3 points), but it dramatically drops the associated recall (24 points).

Text prompt 1

Animate the blue circle to move along a path shaped like a regular heptagram (with self-intersecting path segments).

Similarity with reflections - $W_{\text{sim-ref}}$ **Text prompt 2**

Animate the blue circle to move along a path shaped like the capital letter M.

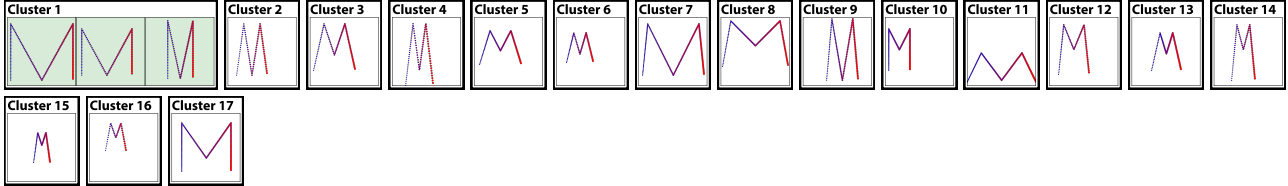
Affine - W_{aff} 

Figure S2. We demonstrate a limitation of our self-consistency approach in terms of the definition of a shape family. In text prompt 1, “heptagram” has two distinct prototypes: heptagram $\{7/2\}$ and heptagram $\{7/3\}$. Under the warp $W_{\text{sim-ref}}$, cluster 1 contains correct $\{7/3\}$ trajectories, while cluster 2 has $\{7/2\}$. While, for generation, our approach would produce a correct interpretation of heptagram, we would mark all heptagram $\{7/2\}$ as false for verification. Similarly, for the letter Ms in text prompt 2, our approach would generate a correct form of the letter, but verify all other letter Ms that are not an affine warp of the shapes in cluster 1 as false.

Table S2. Motion trajectory generation experiment with two additional baseline decision criteria of always selecting the most or least restrictive warps. We observe that they perform strictly worse than the two methods proposed in the main paper.

Method	Decision Criteria	GPT-4.1	GPT-5
LLM-Direct	—	62.1	79.1
Ours	Majority-Consensus	68.0	83.3
Ours	Most-Restrictive	66.5	82.5
Ours	Hierarchical-Consistency	66.7	82.6
Ours	Least-Restrictive	64.6	82.1
Ours	Oracle	68.5	83.5

Table S3. Motion trajectory verification experiment (Table 2 of the main paper) with additional decision criteria of always selecting the most or least restrictive warps.

Method	Decision Criteria	Precision	Recall	F1 Score
GPT-4.1	—	62.0	96.9	75.6
GPT-5	—	74.0	84.7	79.0
Ours	Majority-Consensus	85.8	66.1	74.6
Ours	Most-Restrictive	86.1	42.1	56.5
Ours	Hierarchical-Consistency	80.5	89.0	84.6
Ours	Least-Restrictive	65.0	91.8	76.1
Ours	Oracle	87.9	83.3	85.6

C. Shape Family Limitations

C.1. Shape Families Requiring Multiple Prototypes

In our work, we define a shape family $\mathcal{F}(o, W)$ in Equation 1 of the main paper with one prototype and one geometric warp, and have demonstrated that a wide variety of common shape families (Figure S1) that people use to describe motion trajectories falls under this definition [11].

However, there are shape families that cannot be represented this way. One type of such families requires multiple but finitely many prototypes under one particular warp. For example, as described in the main paper, the shape family “heptagram” needs two distinct prototypes: heptagram $7/2$

and heptagram $7/3$ under the similarity warp (Figure S2). Another type of families either needs an infinite amount of prototypes under one warp, or one prototype with some sophisticated, non-geometric warp. An example of this is the family of any letter of the English alphabet, like the capital letter M (Figure S2). As demonstrated by Hofstadter and McGraw [5], since there are countless ways to produce a trajectory that a human would perceive as an ‘M-shape’, we either need an infinite number of prototypes or a warp beyond the geometric transformations we consider.

In Figure S2, we prompted for the heptagram and capital M shapes. In both cases, while the largest clusters do contain correct trajectories for generation, applying them for verification would result in limited recall, as correct shapes

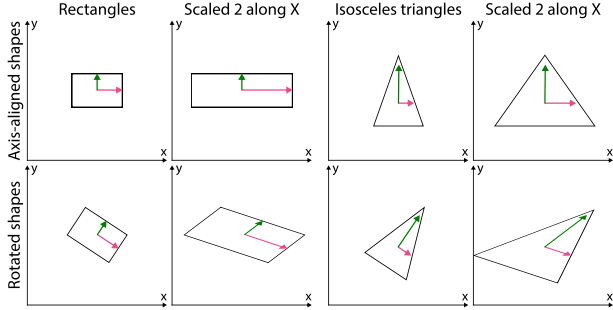


Figure S3. When a shape is aligned with the screen space’s coordinate frame, applying an anisotropic scale of 2 along the x-axis does not introduce skews to the shapes (top). However, once a shape is rotated out of alignment, non-uniform scales effectively skew the shapes out of their respective shape families (bottom).

that cannot be mapped to by the single prototype and the warp would be labeled as false.

C.1.1. Extension to Multi-Prototype Shape Families

For shape families requiring a finite number of prototypes under one warp, we experimented with a simple modification to Hierarchical-Consistency that returns all largest clusters whose sizes are within 20% of one another, instead of only the single largest cluster. We tested this modification on four multi-prototype prompts under $W_{\text{sim-ref}}$: heptagram (2 prototypes: $\{7/2\}$, $\{7/3\}$; Figure S2), heptagram outline (2 prototypes: $\{7/2\}$, $\{7/3\}$), octagram outline (2 prototypes: $\{8/2\}$, $\{8/3\}$), and hendecagram (4 prototypes: $\{11/2\}$, $\{11/3\}$, $\{11/4\}$, $\{11/5\}$). With regular Hierarchical-Consistency, the verification F1 score on these multi-prototype prompts is 71.0. After applying the 20% tolerance modification, F1 improves to 88.9. Meanwhile, the modification has a negligible effect on single-prototype prompts, with F1 decreasing only slightly from 84.6 to 84.5, as the tolerance occasionally causes an additional outlier cluster to be selected. This small experiment suggests that we can potentially extend our approach to handle multi-prototype shape families by returning the largest clusters that satisfy certain statistical relationships with one another. Future work should more comprehensively evaluate this on a larger set of multi-prototype prompts and explore other decision criteria for selecting multiple clusters.

C.2. Shape Families with Anisotropic Similarity

For shape families with the anisotropic similarity warp, when a shape is not aligned with the screen coordinate frame (bottom row of Figure S3), a skew can be introduced to distort it out of its family if a non-uniform scaling is applied to it [12]. For symmetric shapes like the ones shown in Figure S3, their lines of symmetry need to align with the screen coordinate axes. For asymmetric shapes with right angles (e.g., right triangles), the lines along the right angles

Table S4. Our decision criteria has different failure modes when not selecting the ground truth warp. Majority-Consensus selects more restrictive distance metrics for 108 prompts (95.6% of the time when it is not matching with the ground truth), while Hierarchical-Consistency is more conservative for 50 prompts (80.6% of the time). Note that both decision criteria choose distance metrics that are on the same level of the hierarchy but different from the ground truth for 1 and 3 prompts respectively.

Decision Criteria	More Restrictive	Match	Less Restrictive
Majority-Consensus	48.2%	49.6%	1.8%
Hierarchical-Consistency	4.0%	72.3%	22.3%

need to be in alignment so that they are preserved under scaling. This means that, in order to properly warp to all other members of the family, we need a prototype that is in alignment with the screen space’s coordinate frame for non-uniform scaling to be applied without skew. We manually ensure this when we are constructing ground truth shape families for our dataset, and assume that the LLM will generate some trajectories in such canonical orientations.

D. Failure Modes

D.1. Decision Criteria Failure Modes

Our method can recover a shape family from an input prompt by using some decision criteria to estimate an appropriate transformation group W . In Section 4.3 of the main paper, we introduce two such decision criteria that analyze the distribution of motion trajectories generated by an LLM. When different sets of assumptions are met, in terms of how the LLM generates motion trajectories, these decision criteria will predict the correct transformation group. For Majority-Consensus the assumptions it makes are (i) the LLM produces correct trajectories more frequently than incorrect ones; (ii) the collection of sampled trajectories diversely covers the modes of variation possible within the transformation group. Hierarchical-Consistency relaxes this second assumption, but introduces a new assumption in that incorrect trajectories produced by the LLM will not cluster with the correct trajectories produced by the LLM under any transformation group in our hierarchy. In cases where these assumptions are broken, it is possible for these decision criteria to fail (i.e. recover a different transformation group compared with the ground-truth annotation).

To demonstrate such cases, we include qualitative examples in Figure S4 and Figure S5. In each figure, we show LLM sampled motion trajectories from a text prompt (top of figure), and how these motion trajectories are clustered under different transformation groups from our hierarchy. The top transformation group we show is affine, the least restrictive, and transformation groups get progressively more restrictive, ending with rigid. For each transformation group,

Text prompt

Animate the blue circle to move along a path shaped like a rhombus

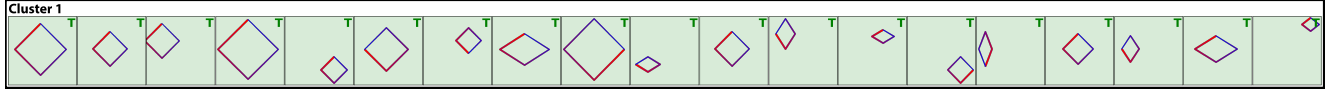
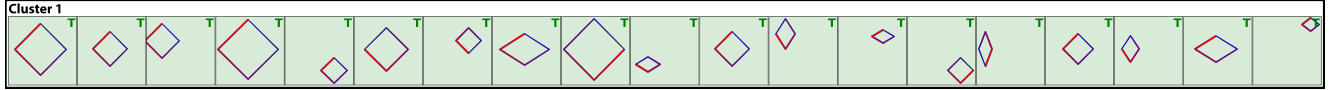
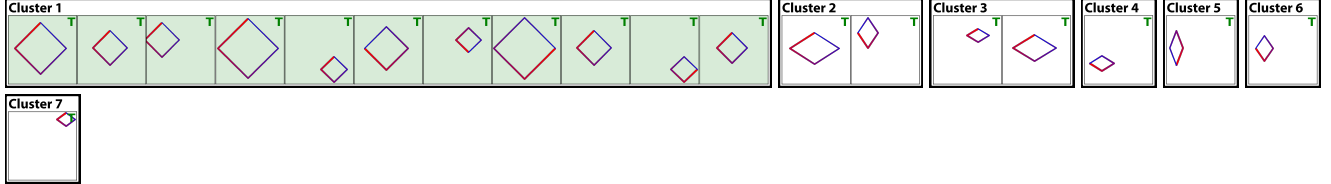
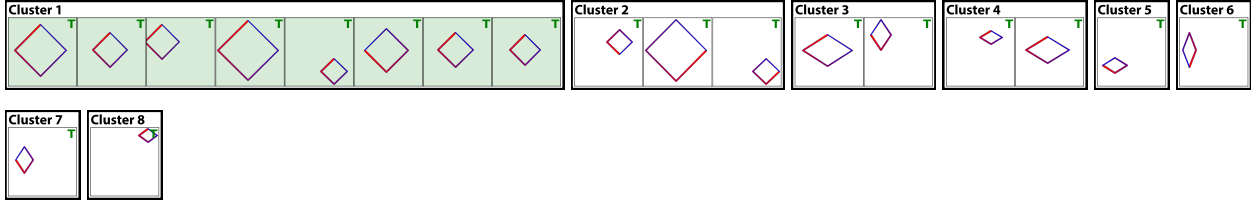
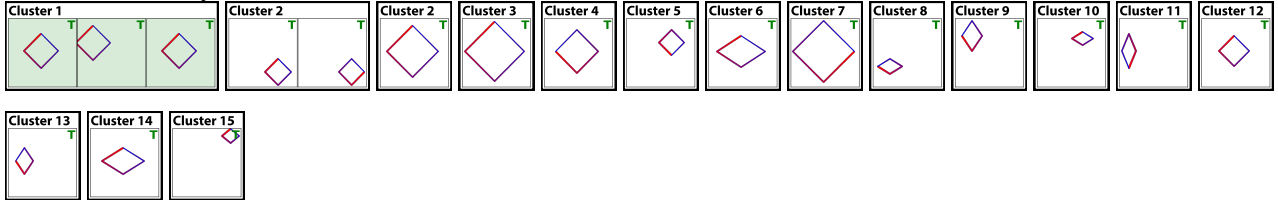
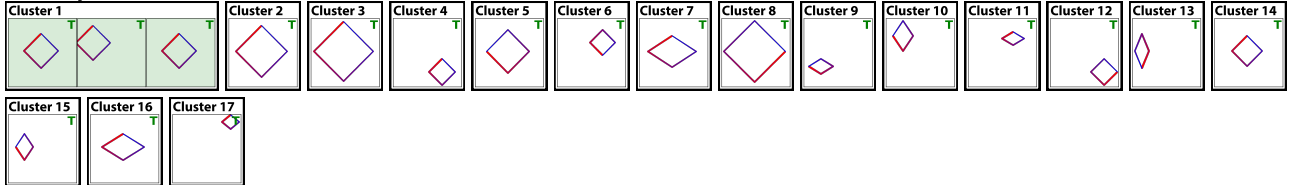
Affine - W_{aff} **Anisotropic similarity - $W_{\text{sim-ani}}$** *Ground truth warp, correctly selected by Hierarchical-Consistency***Similarity with reflections - $W_{\text{sim-ref}}$** *Incorrectly selected by Majority-Consensus***Similarity - W_{sim}** **Rigid with reflections - $W_{\text{rgd-ref}}$** **Rigid - W_{rgd}** 

Figure S4. We demonstrate a case where Hierarchical-Consistency chooses the correct warp but Majority-Consensus fails by selecting a more restrictive warp. The prompt (top) asks for a rhombus-shaped path. We show clustering results of all warps in our hierarchy, with largest clusters under each warp highlighted in green. Each trajectory has a ground truth correctness label on the upper right.

we highlight the largest cluster in green. Each trajectory is labeled with a T/F value in the top-right corner, indicating whether it is a member of our ground-truth shape family.

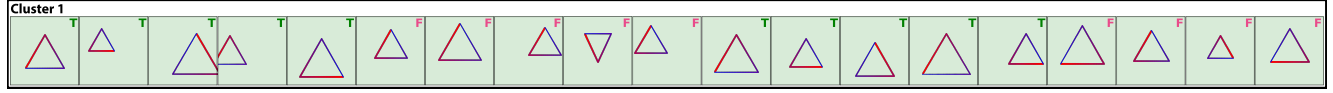
In Figure S4, we demonstrate a case where Hierarchical-Consistency identifies the correct W , while Majority-Consensus fails. For this prompt, all of the LLM generated motion trajectories were correct. For Hierarchical-Consistency, for the affine transformation group it identifies the largest cluster, and observes that this cluster does not lose any members when the transfor-

mation group is made more restrictive with anisotropic similarity, which is the correct transformation group for this shape family. Trying to make W more restrictive (similarity with reflections) would break this cluster, so Hierarchical-Consistency succeeds by stopping at the previous step. In contrast, Majority-Consensus identifies that under the clustering produced by similarity with reflections, over half of the generations would be included in the largest cluster, so it incorrectly would choose an overly restrictive transformation group. This sampling

Text prompt

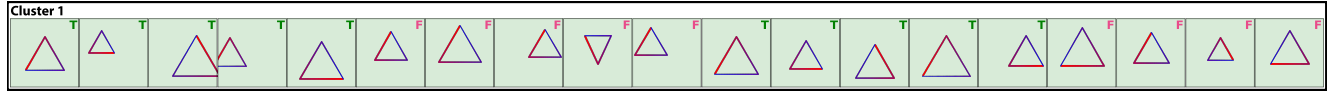
Animate the blue circle to move along a path shaped like an equilateral triangle. Traverse the path in a counterclockwise manner.

Affine - W_{aff}

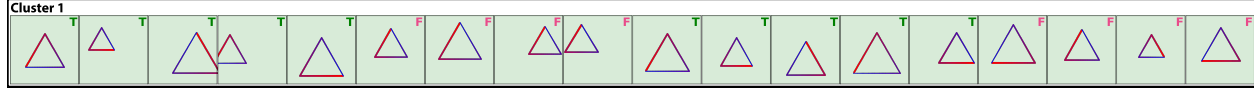


Anisotropic similarity - $W_{sim-ani}$

Incorrectly selected by Hierarchical-Consistency

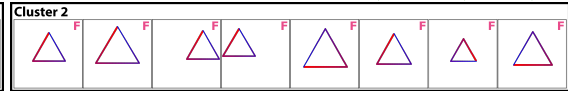


Similarity with reflections - $W_{sim-ref}$

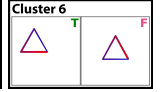
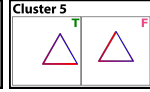
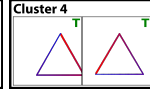
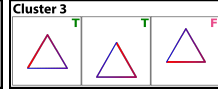
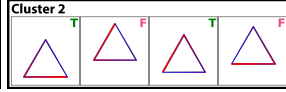
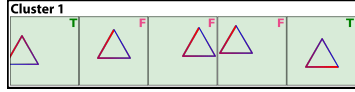


Similarity - W_{sim}

Ground truth warp, correctly selected by Majority-Consensus



Rigid with reflections - $W_{rgd-ref}$



Rigid - W_{rgd}

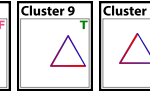
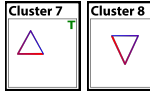
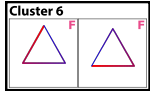
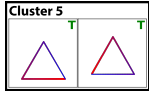
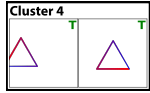
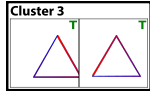
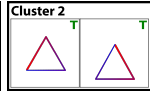
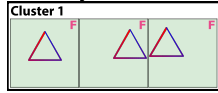


Figure S5. An example case where Majority-Consensus correctly chooses the ground truth warp but Hierarchical-Consistency fails by selecting a less restrictive warp. The prompt (top) asks for a path shaped like an equilateral triangle traversed in a counterclockwise manner. We show clustering results of all warps in our hierarchy.

breaks an assumption made by Majority-Consensus, as the LLM produced trajectories do not provide a balanced enough coverage over the modes of variation under the ground-truth transformation group (anisotropic similarity).

In Figure S5, we demonstrate a case where Majority-Consensus identifies the correct W , while Hierarchical-Consistency fails. In this case, 10 LLM generations were correct and 9 LLM generations were incorrect, as they produced paths that did not traverse in a clockwise manner (note time is indicated as a progression from blue to red in the renders). The correct transforma-

tion group in this case is similarity, as more permissive transformation groups would cause matches that violate the traversal orientation specification. Majority-Consensus starts with the clustering produced by the most restrictive transformation group (rigid), which has a largest cluster of size 3, and progressively considers less restrictive transformation groups until it finds a cluster with at least 10 members, correctly identifying similarity should be used as W . In contrast, Hierarchical-Consistency starts with affine, and once again identifies that all of the LLM productions are grouped into a single cluster. This cluster only loses

members on the transition from anisotropic similarity to similarity with reflections, so Hierarchical-Consistency will incorrectly choose an overly permissive transformation group. In this case, some of the incorrect LLM samples would cluster with correct samples under the affine transformation group, breaking an assumption made by Hierarchical-Consistency.

So, depending on the distribution of the LLM produced motion trajectories, these decision criteria can fail in their task of recovering the correct transformation group. We quantitatively analyze the prevalence of these failure modes in Table S4. For our two proposed criteria, Hierarchical-Consistency and Majority-Consensus, we record for all 224 prompts how the chosen transformation group compares with the ground-truth transformation group in the benchmark dataset. *Match* indicates the criteria chose the correct transformation group, while *more restrictive* indicates the criteria chose a transformation group too low in the hierarchy, and *less restrictive* indicates the criteria chose a transformation group too high in the hierarchy.

In general, we find that Hierarchical-Consistency more reliably identifies the correct W compared with Majority-Consensus (72% vs 49%). This likely indicates that the assumptions made by Hierarchical-Consistency are more often met compared with the assumptions made by Majority-Consensus. From the qualitative examples, we can see the LLM struggles to produce sufficiently diverse samples that cover all of the possible modes of variation, so this is likely a major source of error for Majority-Consensus. Beyond checking whether or not these decision criteria recover the exact ground-truth W , we can additionally analyze how these methods tend to fail. Majority-Consensus starts at the bottom of the hierarchy, stopping when a majority clustering is found. As such, Majority-Consensus is overly prone to producing more restrictive transformation groups when it makes mistakes (25 times more likely to have a more restrictive error compared with a less restrictive error). Hierarchical-Consistency starts at the top of the hierarchy, with a stopping criterion when the majority cluster begins to break apart. As such, Hierarchical-Consistency is overly prone to producing transformation groups that are too loose when it makes mistakes (5 times more likely to have a less restrictive error compared with a more restrictive error). These quantitative trends match the qualitative examples.

D.2. LLM Failure Modes

The general self-consistency approach makes an assumption that the LLM is likely to produce the correct response more often than any other incorrect response [14]. We examine cases where this assumption does not hold and present quantitative results in Figure S6. Applying our self-consistency approach with the oracle warps to the trajec-

tries generated by GPT-5 from the 224 prompts in our dataset (same experimental set-up as Section 5.2 of the main paper), we see 16.1% of the prompts with largest clusters that do not contain any correct trajectories. For GPT-4.1 generated trajectories, we have 31.3% (70) such clusters.

We show examples of these failure cases in the bottom half of Figure S6. Text prompt 3 asks for a trajectory in the shape of an outline of a pentagram, and the oracle similarity warp puts the generated trajectories into three clusters. While the LLM (GPT-5) was able to produce the correct trajectory 5 times (Cluster 2), it generates more instances of a regular pentagram shape with the internal path crossovers in both clockwise and counterclockwise fashion. Text prompt 4 asks for counterclockwise reuleaux triangle-shaped trajectories. Under the ground truth rigid transformation warp, the generated trajectories are grouped into three clusters. The LLM (GPT-4.1) similarly generates the correct trajectory 5 times (Cluster 3), but falsely traces the construction lines of a reuleaux triangle for the other 14 times. In both of these cases, the incorrect trajectories formed largest clusters in the form of relative majority with 9 trajectories. As the LLM produces more trajectories outside of the ground truth shape families than inside, the general assumption of self-consistency does not hold and the largest clusters do not contain any correct trajectories.

E. Method Implementation Details

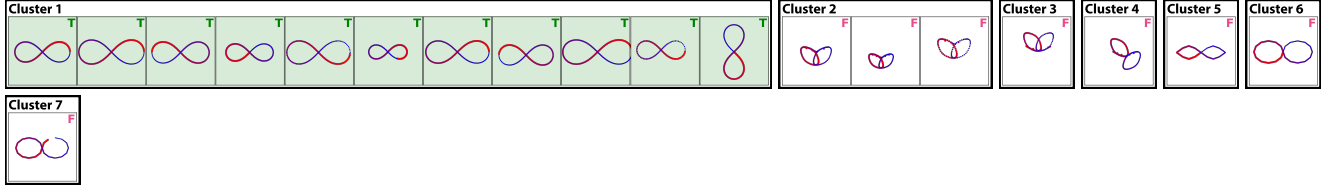
E.1. Computing Warp-Invariant Distance

Algorithm 1 presents our ICP-based approach [1, 2, 4] for computing the warp-invariant distance between a source trajectory R and target trajectory T under a geometric warp W . We begin by resampling both trajectories to have n equally-spaced points along their arc length. This resampling establishes initial point correspondences, where we assume that points at the same index in R and T correspond to one another. The outer loop then randomly selects 3 non-repeated, non-collinear point indices \mathcal{I} from these trajectories. These three points provide sufficient constraints to estimate most geometric transformations in our hierarchy.

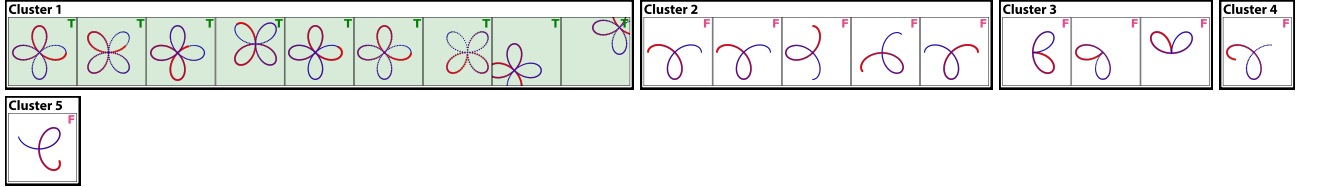
Given the sampled point indices, the inner loop iteratively refines the transformation A . At each iteration, we: (1) estimate a transformation A that aligns the sampled points from R_{trans} to the corresponding points in T , (2) apply this transformation to the source trajectory R , (3) resample the transformed trajectory for a new set of corresponding points, and (4) compute the average point-to-point distance. If the change in distance falls below a convergence threshold ϵ , the inner loop terminates. Otherwise, we map the newly sampled points back to the original R space by undoing the transformation and repeat. This iterative process effectively “nudges” points along trajectory R to better align with T under the transformation A .

Text prompt 1

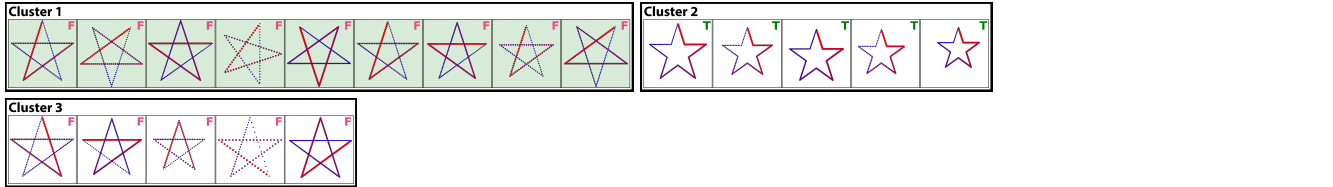
Animate the blue circle to move along a path shaped like a figure 8 shape (lemniscate of Bernoulli).

Similarity with reflections - $W_{sim-ref}$ **Text prompt 2**

Animate the blue circle to move along a path shaped like a 4-petal rose (quadrifolium). The 'a' parameter of the 4-petal rose should be 60.

Rigid with reflections - $W_{rgd-ref}$ **Text prompt 3**

Animate the blue circle to move along a path shaped like a regular pentagram (without the internal crossovers). Only move along the outline of the regular pentagram (no path crossovers). Traverse the path in a counterclockwise manner.

Similarity - W_{sim} **Text prompt 4**

Animate the blue circle to move along a path shaped like a reuleaux triangle with a width of 50. Traverse the path in a counterclockwise manner.

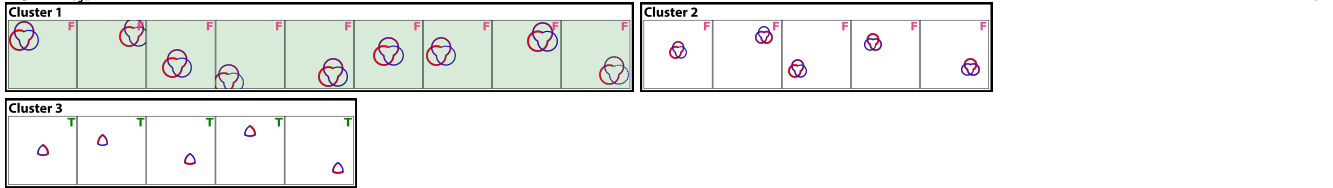
Rigid - W_{rgd} 

Figure S6. More results of applying our self-consistency approach with the oracle warps. The top half shows two success cases where the largest clusters do correspond to the correct shape families (figure 8 and 4-petal rose). The bottom half shows two failure cases where the LLM did not produce more members of the ground truth shape families than non-members (pentagram outline and reuleaux triangle).

After convergence, we check if the current alignment achieves a lower distance than previous iterations and update the best transformation A_{min} accordingly. The algorithm terminates when the distance falls below a threshold τ or after a maximum number of outer iterations. Our ICP-based implementation is a proof of concept to show the viability of our self-consistency approach for motion trajectories. It would likely be possible to improve its runtime with efficient variants in the future [9, 10].

Transformation estimation. EstimateTransform in line 11 of Algorithm 1 computes A based on the geometric warp type \mathcal{W} . For rigid and similarity transformations (with or without reflections), we use the Kabsch-Umeyama algorithm [7, 13], which computes optimal rotation, translation, and optionally uniform scaling and reflection via

SVD. For affine warps, we directly solve the linear system relating source to target points. For anisotropic similarity, we test both $A = M_{trs} \cdot M_{rot} \cdot M_{scl}$ and $A = M_{trs} \cdot M_{scl} \cdot M_{rot}$ decomposition orders and select the transformation yielding the smaller distance.

Handling closed trajectories Our assumption that points with the same indices after arc-length resampling correspond to one another holds for open trajectories with distinct starting and end points. However, closed trajectories (e.g., circles, squares) may have arbitrary starting points. This means that, if T and R are the same closed trajectories but with distinct starting points, our algorithm will not be able to find an A that produces zero distance between them. To address this, we extend our algorithm with a coarse-to-fine search over different starting points along R . We com-

Algorithm 1 ICP-based Trajectory Distance Metric

```
1: Input: Source trajectory  $R$ , target trajectory  $T$ , and geometric warp type  $\mathcal{W}$ 
2: Output: Minimum distance  $d_{\min}$  between  $R$  and  $T$  with corresponding transformation matrix  $A_{\min}$ 
3:
4:  $T \leftarrow \text{ResampleByArcLength}(T, n)$ 
5:  $d_{\min} \leftarrow \infty, A_{\min} \leftarrow \mathbf{I}$ 
6: for  $i = 1$  to  $M_{\text{outer}}$  do
7:    $R_{\text{trans}} \leftarrow \text{ResampleByArcLength}(R, n)$ 
8:    $d_{\text{prev}} \leftarrow \infty$ 
9:    $\mathcal{I} \leftarrow$  Randomly sample 3 point indices
10:  for  $j = 1$  to  $M_{\text{inner}}$  do
11:     $A \leftarrow \text{EstimateTransform}(R_{\text{trans}}[\mathcal{I}], T[\mathcal{I}], \mathcal{W})$ 
12:     $R_{\text{trans}} \leftarrow A \cdot R$   $\triangleright$  Apply transformation
13:     $R_{\text{trans}} \leftarrow \text{ResampleByArcLength}(R_{\text{trans}}, n)$ 
14:     $d_{\text{avg}} \leftarrow \frac{1}{n} \sum_{k=1}^n \|R_{\text{trans}}[k] - T[k]\|_2$ 
15:    if  $|d_{\text{avg}} - d_{\text{prev}}| < \epsilon$  then
16:      break  $\triangleright$  Inner loop converged
17:    end if
18:     $d_{\text{prev}} \leftarrow d_{\text{avg}}$ 
19:     $R_{\text{trans}} \leftarrow A^{-1} \cdot R_{\text{trans}}$   $\triangleright$  Undo transformation
20:  end for
21:  if  $d_{\text{avg}} < d_{\min}$  then
22:     $A_{\min} \leftarrow A, d_{\min} \leftarrow d_{\text{avg}}$   $\triangleright$  Update alignment
23:  end if
24:  if  $d_{\text{avg}} < \tau$  then
25:    break  $\triangleright$  Outer loop converged
26:  end if
27: end for
28: return  $d_{\min}, A_{\min}$ 
```

Table S5. F1 scores of verification by Hierarchical-Consistency and the average number of DBSCAN clusters across the distance threshold τ .

	$\tau = 0.25$	$\tau = 0.5$	$\tau = 1.0$	$\tau = 2.0$	$\tau = 4.0$	$\tau = 8.0$
F1 score	84.0	84.6	83.8	83.1	80.3	76.8
Clusters (avg \pm std)	4.0 \pm 4.9	3.7 \pm 4.8	3.5 \pm 4.5	3.1 \pm 4.0	2.6 \pm 3.1	2.1 \pm 2.1

pute the distance metric with each point in R taken as the starting point and report the minimum distance across all starting points.

E.2. Sensitivity to DBSCAN Distance Threshold

Equation 2 of the main paper defines that two trajectories in the same shape family should have a distance of 0. We perform DBSCAN clustering with a small threshold τ (Section 4.2 of the main paper) only to account for discretization errors during resampling in our distance metric computation. Our approach is not sensitive to the value of τ . As shown in Table S5, increasing τ by a factor of 32 (from 0.25 to 8.0) results in only a 7.2 percentage point drop in F1 score.

E.3. Generating Motion Trajectories with an LLM

To use an LLM for motion trajectory animation generation, we follow the approach proposed by MoVer [8]; we prompt an LLM with a system message containing overall instructions and the complete documentation of an animation API (GSAP [3]), followed by a user prompt that contains the static SVG code to be animated, the description of a motion trajectory, and additional instructions.

We modify the MoVer approach in two places. First, we updated the animation API with an additional function `setProperty()` that allows changing an element’s properties like positions without adding animation tweens.

```
1 /**
2  * This function immediately sets properties on
3  * an SVG element without animation. Use this
4  * function to initialize or instantly update
5  * element properties such as position, scale,
6  * rotation, opacity, etc. This is useful for
7  * setting up initial states before animating.
8  * @param {object} element - The SVG element
9  * whose properties should be set.
10 * @param {object} properties - An object
11 * containing property-value pairs to set on the
12 * element. Common properties include x, y,
13 * scaleX, scaleY, rotation, opacity,
14 * transformOrigin, and svgOrigin.
15 * @returns {void} - This function does not
16 * return anything.
17 * @example
18 * // Set the translate transform value of the
19 * // element to [100, 200].
20 * // Here, x, y values set the translation of
21 * // the element (displacement from its original
22 * // position at the start of the animation).
23 * setProperty(circle, { x: 100, y: 200 });
24 */
25 function setProperty(element, properties)
```

For the user prompt, we append additional instructions shown below after the motion trajectory prompt. At the end, we instruct the LLM to produce five diverse animations following the diversity prompting strategy [15].

```
1 SVG Code:
2 {{ svg code }}
3
4 Prompt:
5 {{ motion trajectory prompt }}
6
7 ## Additional Instructions
8 - Note that the negative y-axis is the upward
9   direction in the SVG scene.
10 - To start a path animation, use setProperty() to
11   position the element at the beginning of the
12   path. Be mindful of the SVG viewBox and not
13   move the element and the path out of view.
14 - For curved paths, use multiple consecutive very
15   short linear translations to achieve smooth
16   motions.
17 - The entire animation should last between 1 and
18   5 seconds unless otherwise specified.
```

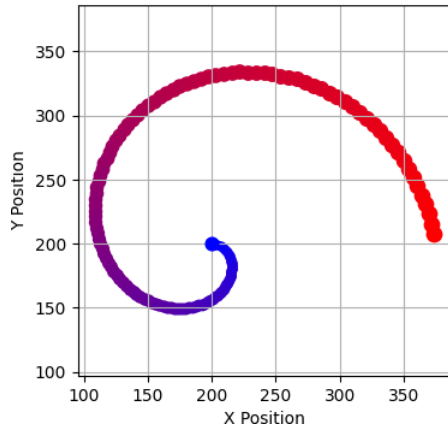


Figure S7. We prompt a VLM with a scatter plot representation of a motion trajectory. Each dot is the center position of the animated element and the color indicates time (transition from blue to red).

```

13 Generate 5 animations to the prompt, each within
    a separate ``javascript `` tag.
14 Each animation must include a numeric probability
    value as a comment within its code block.
15 Please sample at random from the tails of the
    distribution, such that the probability of
    each response is less than 0.10.

```

E.4. Verifying Motion Trajectories with a VLM

Our experiments use two VLMs (GPT-4.1 and GPT-5) as baseline verification methods, and we provide details here on our prompting approach. We represent the motion trajectory center point time series data of animation as a scatter plot where the color of each dot transitions from blue to red to indicate the passage of time (Figure S7).

We prompt a VLM with a system message asking it to output a true/false label indicating whether a trajectory animation follows a trajectory prompt or not, with instructions on how to read the scatter plot. We then add a user message containing the motion trajectory prompt and the plot.

```

1 You are an expert in evaluating how well motion
  trajectories follow given prompts. Given a
  plot of a motion trajectory, your task is to
  assess whether the trajectory follows the
  prompt (True or False). Blue indicates the
  beginning of the trajectory and red indicates
  the end of the trajectory. The color
  transitions from blue to red as the
  trajectory progresses across time.
2
3 Please output your response as a JSON object with
  the following format:
4 ````json
5 {
6   "<trajectory_id (the index of the trajectory
  file)>": 0 or 1, # 0 means False, 1 means
  True
7 }
8 ````

```




Figure S9. Motion trajectory benchmark: prompt 37–72 with ground truth shape families.

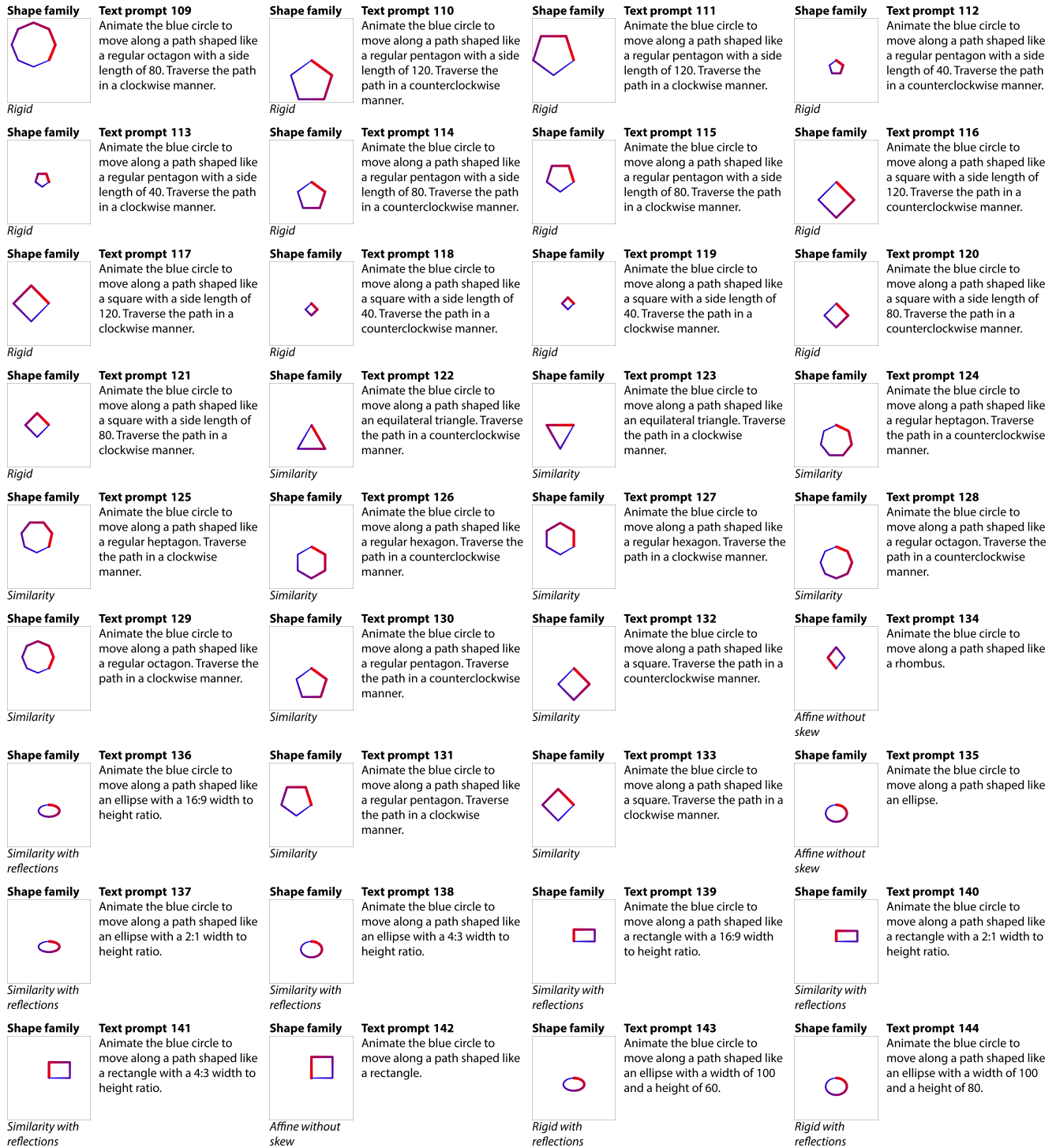


Figure S11. Motion trajectory benchmark: prompt 109–144 with ground truth shape families.




































<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 145 Animate the blue circle to move along a path shaped like an ellipse with a width of 50 and a height of 60.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 146 Animate the blue circle to move along a path shaped like an ellipse with a width of 50 and a height of 80.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 147 Animate the blue circle to move along a path shaped like a rectangle with a width of 100 and a height of 60.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 148 Animate the blue circle to move along a path shaped like a rectangle with a width of 100 and a height of 80.</p>
<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 149 Animate the blue circle to move along a path shaped like a rectangle with a width of 50 and a height of 60.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 150 Animate the blue circle to move along a path shaped like a rectangle with a width of 50 and a height of 80.</p>	<p>Shape family</p>  <p><i>Rigid</i></p>	<p>Text prompt 151 Animate the blue circle to move along a path shaped like a reuleaux triangle with a width of 50. Traverse the path in a counterclockwise manner.</p>	<p>Shape family</p>  <p><i>Rigid</i></p>	<p>Text prompt 152 Animate the blue circle to move along a path shaped like a reuleaux triangle with a width of 50. Traverse the path in a clockwise manner.</p>
<p>Shape family</p>  <p><i>Rigid</i></p>	<p>Text prompt 153 Animate the blue circle to move along a path shaped like a reuleaux triangle with a width of 75. Traverse the path in a counterclockwise manner.</p>	<p>Shape family</p>  <p><i>Rigid</i></p>	<p>Text prompt 154 Animate the blue circle to move along a path shaped like a reuleaux triangle with a width of 75. Traverse the path in a clockwise manner.</p>	<p>Shape family</p>  <p><i>Similarity with reflections</i></p>	<p>Text prompt 155 Animate the blue circle to move along a path shaped like a reuleaux triangle.</p>	<p>Shape family</p>  <p><i>Similarity with reflections</i></p>	<p>Text prompt 156 Animate the blue circle to move along a path shaped like a 3-petal rose (trifolium).</p>
<p>Shape family</p>  <p><i>Similarity with reflections</i></p>	<p>Text prompt 157 Animate the blue circle to move along a path shaped like a 4-petal rose (quadrifolium).</p>	<p>Shape family</p>  <p><i>Similarity with reflections</i></p>	<p>Text prompt 158 Animate the blue circle to move along a path shaped like a 5-petal rose (pentafolium).</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 159 Animate the blue circle to move along a path shaped like a 3-petal rose (trifolium). The 'a' parameter of the 3-petal rose should be 60.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 160 Animate the blue circle to move along a path shaped like a 3-petal rose (trifolium). The 'a' parameter of the 3-petal rose should be 80.</p>
<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 161 Animate the blue circle to move along a path shaped like a 4-petal rose (quadrifolium). The 'a' parameter of the 4-petal rose should be 60.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 162 Animate the blue circle to move along a path shaped like a 4-petal rose (quadrifolium). The 'a' parameter of the 4-petal rose should be 80.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 163 Animate the blue circle to move along a path shaped like a 5-petal rose (pentafolium). The 'a' parameter of the 5-petal rose should be 60.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 164 Animate the blue circle to move along a path shaped like a 5-petal rose (pentafolium). The 'a' parameter of the 5-petal rose should be 80.</p>
<p>Shape family</p>  <p><i>Similarity with reflections</i></p>	<p>Text prompt 165 Animate the blue circle to move along a path shaped like a quarter-circle (a quarter of a circle's circumference).</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 166 Animate the blue circle to move along a path shaped like a quarter-circle (a quarter of a circle's circumference) with a radius of 65.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 167 Animate the blue circle to move along a path shaped like a quarter-circle (a quarter of a circle's circumference) with a radius of 75.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 168 Animate the blue circle to move along a path shaped like a quarter-circle (a quarter of a circle's circumference) with a radius of 85.</p>
<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 169 Animate the blue circle to move along a path shaped like a quarter-circle (a quarter of a circle's circumference) with a radius of 95.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 171 Animate the blue circle to move along a path shaped like a semi-circle (half of a circle's circumference) with a radius of 75.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 173 Animate the blue circle to move along a path shaped like a semi-circle (half of a circle's circumference) with a radius of 95.</p>	<p>Shape family</p>  <p><i>Similarity with reflections</i></p>	<p>Text prompt 174 Animate the blue circle to move along a path shaped like a semi-circle (half of a circle's circumference).</p>
<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 170 Animate the blue circle to move along a path shaped like a semi-circle (half of a circle's circumference) with a radius of 65.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 172 Animate the blue circle to move along a path shaped like a semi-circle (half of a circle's circumference) with a radius of 85.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 175 Animate the blue circle to move along a path shaped like a regular heptagram {7/2} (without the internal crossovers) with an outer radius of 50. Only move along the outline of the regular heptagram {7/2} (no path crossovers).</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 176 Animate the blue circle to move along a path shaped like a regular heptagram {7/2} (without the internal crossovers) with an outer radius of 80. Only move along the outline of the regular heptagram {7/2} (no path crossovers).</p>
<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 177 Animate the blue circle to move along a path shaped like a regular heptagram {7/2} (without the internal crossovers) with each line segment being 50 px long. Only move along the outline of the regular heptagram {7/2} (no path crossovers).</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 178 Animate the blue circle to move along a path shaped like a regular heptagram {7/2} (without the internal crossovers) with each line segment being 80 px long. Only move along the outline of the regular heptagram {7/2} (no path crossovers).</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 179 Animate the blue circle to move along a path shaped like a regular heptagram {7/2} (with self-intersecting path segments) with an outer radius of 50.</p>	<p>Shape family</p>  <p><i>Rigid with reflections</i></p>	<p>Text prompt 180 Animate the blue circle to move along a path shaped like a regular heptagram {7/2} (with self-intersecting path segments) with an outer radius of 80.</p>

Figure S12. Motion trajectory benchmark: prompt 145–180 with ground truth shape families.

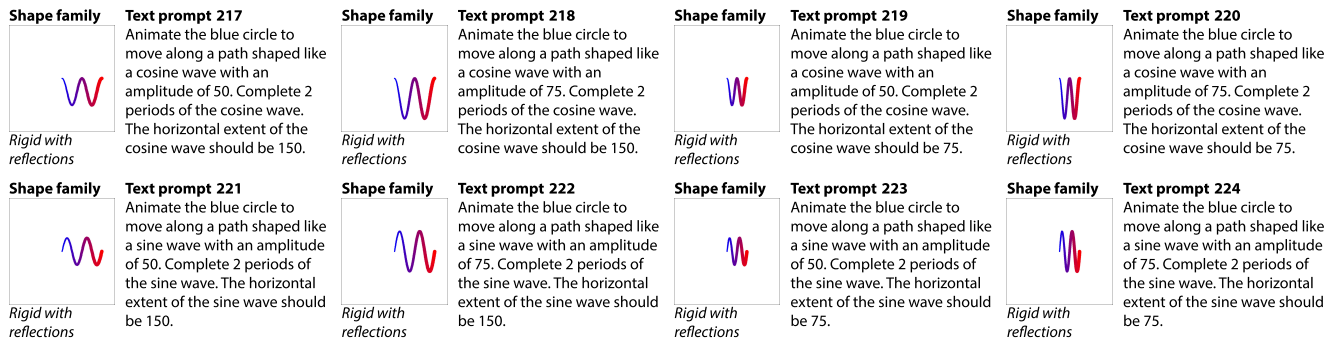


Figure S14. Motion trajectory benchmark: prompt 217–224 with ground truth shape families.

References

- [1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 7
- [2] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 7
- [3] Jack Doyle. Greensock animation platform v3, 2025. 9
- [4] Shaoyi Du, Nanning Zheng, Shihui Ying, and Jianyi Liu. Affine iterative closest point algorithm for point set registration. *Pattern Recognition Letters*, 31(9):791–799, 2010. 7
- [5] Douglas Hofstadter and Gary McGraw. Letter spirit: An emergent model of the perception and creation of alphabetic style. In *Technical Report 68, Center for Research on Concepts and Cognition*. 1993. 3
- [6] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 1
- [7] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 8
- [8] Jiaju Ma and Maneesh Agrawala. MoVer: Motion verification for motion graphics animations. *ACM Trans. Graph.*, 44(4), 2025. 1, 9
- [9] Nicolas Mellado, Dror Aiger, and Niloy J Mitra. Super 4pcs fast global pointcloud registration via smart indexing. In *Computer graphics forum*, pages 205–215. Wiley Online Library, 2014. 8
- [10] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001. 8
- [11] Mathias Sablé-Meyer, Kevin Ellis, Josh Tenenbaum, and Stanislas Dehaene. A language of thought for the mental representation of geometric shapes. *Cognitive Psychology*, 139:101527, 2022. 3
- [12] Carsten Steger. Least-squares estimation of anisotropic similarity transformations from corresponding 2d point sets. *Pattern Recognition Letters*, 33(3):349–355, 2012. 4
- [13] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 8
- [14] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 7
- [15] Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity, 2025. 9