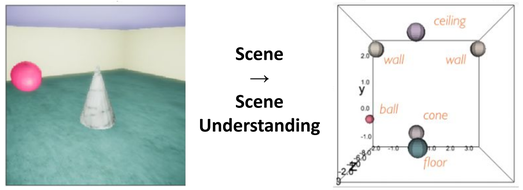
Object Understanding with Visual World Models

Gia Ancone[1], Ashley Xu[1], Rahul Venkatesh[1]
[1]Stanford University

Do World Models Have Object Understanding?

- People understand their surroundings as discrete objects with enduring properties, and even infants can track these objects through occlusion and motion without explicit supervision.
- Neurological Evidence: Conditions like simultanagnosia show the brain binds variables to each object to track state over time.
- AI needs reliable object understanding for manipulation, navigation, and planning.
- Challenge: Supervision for meshes, material labels, and physics is scarce.
- World models learn structure, persistence, and material behavior through prediction of raw sensory inputs.
- This work both advances embodied AI and offers a framework for studying how the brain links perception and action, by modeling objects, materials, and their physical interactions.

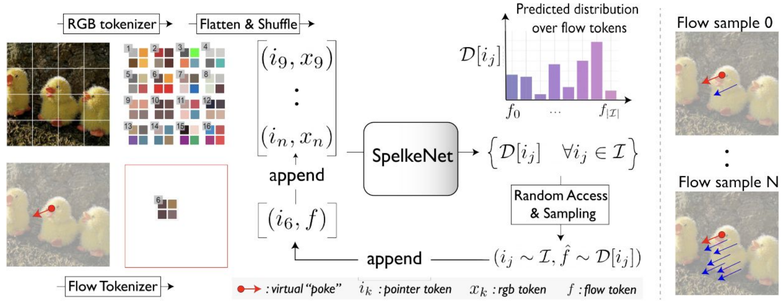


SpelkeNet: A Candidate World Model Built by Stanford Neuro AI Lab

- **LRAS framework:** A probabilistic world model that encodes videos as sequences of locally quantized tokens, using an autoregressive architecture with locality biases for flexible, compositional motion prediction.
- **SpelkeNet** is an instantiation of LRAS trained on large-scale internet video that predicts:
  1. Probability-of-motion map: regions likely to move.
  2. Counterfactual displacements: flow fields from virtual pokes.
  3. Spelke objects: pixel groupings that move as cohesive units.

Key idea: Define object boundaries based on coherent motion across time, rather than on semantic class or visual appearance.

SpelkeNet Architecture



SpelkeBench: A New Dataset We Built for Benchmarking Object Understanding

- Motivation: Conventional segmentations (e.g., SAM, COCO, ADE20K) often split objects into sub-parts, merge multiple movable entities, or include immovable/background regions.
- **SpelkeBench** captures physically grounded objecthood, only entities that move together under external forces, following Psychologist Liz Spelke's principles of cohesion, continuity, and solidity.
- Sources: EntitySeg (high-resolution internet images with dense labels) and OpenX-Embodiment (real-world robot interaction data).
- Annotations:
  - 50 OpenX images manually annotated with Spelke-consistent segments relevant to robot manipulation.
  - 500 EntitySeg images filtered via a 3-stage pipeline:
    1. Remove amorphous background regions ("stuff" like sky/terrain).
    2. Remove functionally immovable objects (e.g., sinks, traffic signs).
    3. Curate diverse, high-quality scenes with only Spelke-consistent regions.
- **Significance**: First benchmark aligned with motion-based segmentation; enables systematic evaluation of whether models discover pixel co-movement.

SpelkeBench Emphasizes Unified, Movable Segments

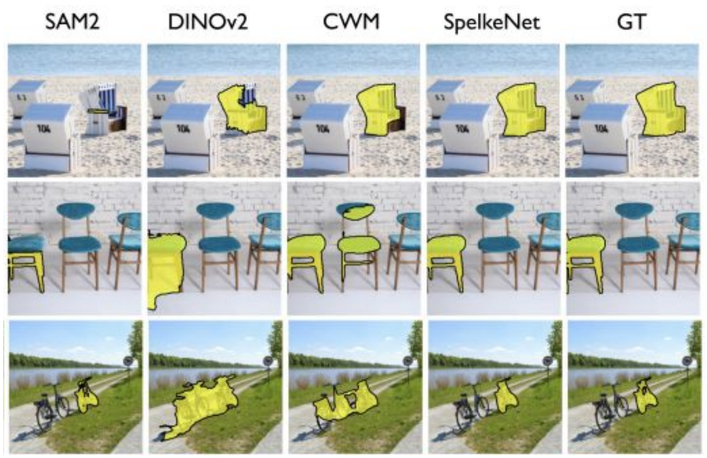
SpelkeNet Outperforms Other Segmentation Models on SpelkeBench

Spelke objects are defined by motion coherence. By using counterfactual probing, **SpelkeNet** identifies object boundaries as statistical aggregates of motion correlation, rather than relying on category-specific labels. This demonstrates that world models can discover objects directly from motion dynamics, aligning more closely with how humans perceive objecthood.

	SAM2	DINOv1-B/8	DINOv2-L/14	DINOv2-G/14	CWM	**SpelkeNet**
AR	0.4816	0.2708	0.2524	0.2254	0.3271	**0.5411**
mIoU	0.6225	0.4990	0.4931	0.4553	0.4807	**0.6811**

Quantitative evaluation. AR measures how likely ground-truth segments are detected, while mIoU measures boundary precision.

- **SpelkeNet:** Sharp and physically grounded segments aligned with motion coherence.
- **SAM2:** Segments textures, patterns, or immovable regions (appearance bias).
- **DINO:** Merges nearby same-category objects (semantic bias).
- **CWM:** Produces diffuse, blurry boundaries from noisy reconstructions.



Qualitatively Comparing SpelkeBench Results

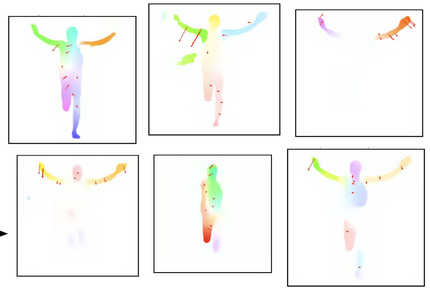
Formulating How Material Properties Can Be Extracted from SpelkeNet

- We hypothesize that world models can capture microscopic material properties by analyzing local responses to force.
- To do this, we isolate patches and apply PCA to flows from interventions, then compute the **deformation gradient** (Jacobian) for each motion mode.
- From these gradients, we derive strain measures such as the **Green Strain Tensor,** which reveal elasticity, stiffness, and flexibility, enabling inference of internal material deformation.

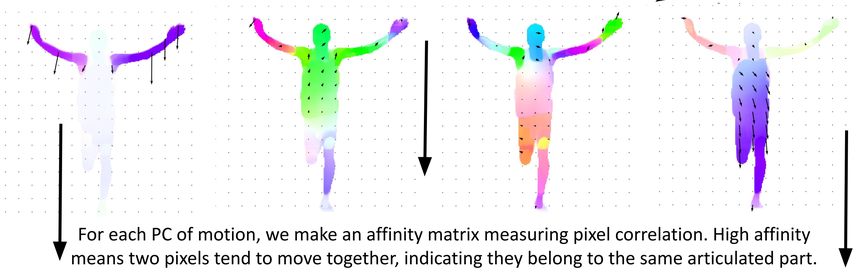
Extracting Richer Object Affordances from SpelkeNet

- We aim to move beyond boundaries and study the macroscopic motion constraints of objects.
- Building on SpelkeNet, we show that world models can be used to reveal object affordances: how articulated bodies move and what actions they support.
- The schematic below illustrates our approach for extracting these affordances from large-scale flow predictions.

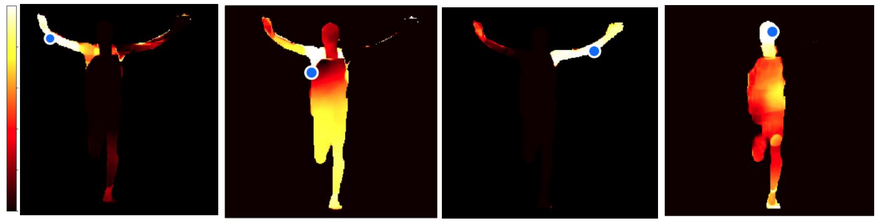
SpelkeNet generates thousands of counterfactual flows under interventions, with random fixed-point constraints to expose diverse articulated motions.



Principal Component Analysis (PCA) over predicted flow fields reveals dominant motion directions across interventions.



For each PC of motion, we make an affinity matrix measuring pixel correlation. High affinity means two pixels tend to move together, indicating they belong to the same articulated part.



Slices from the affinity matrices of different principal components reveal structured co-movement patterns within objects.

Proposed Next Steps for Understanding Macroscopic Motion Constraints

- Apply clustering to the affinity matrices to segment pixels into articulated parts that move coherently.
- PCA on Parts: Within each part, perform PCA again to extract low-dimensional motion representations.
- Motion Axes Constraints: These part-level PCs reveal dominant motion axes (e.g., hinge rotations, linear translations) and constraints specific to each articulated component.

Outcome: Provides a structured description of how individual parts move, complementing whole-object affordances with articulated detail.

Conclusion and Future Work

Our work shows that world models can acquire object understanding across three levels:

1. **Discovering object boundaries** with SpelkeNet.
2. **Extracting macro-level affordances** of articulated bodies using large-scale flow generation, PCA, and affinity-based clustering.
3. **Probing micro-level material properties** with patch-based PCA and strain analysis.

Future work includes developing world models that explicitly capture local deformation and elasticity, improving segmentation and scaling affordance analysis to more complex objects and scenes, and applying these representations in robotics, where knowledge of motion axes and material behavior is critical for manipulating articulated mechanisms and deformable objects.