# ZERO-SHOT OBJECT UNDERSTANDING WITH A PHYSICALLY CONTROLLABLE WORLD MODEL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Humans acquire intuitive notions of objects directly from visual experience: they perceive objects as bounded physical units that move together and learn properties such as 3D structure and materials. Existing AI models fail to develop these capabilities in a self-supervised way. Even when provided with expensive large-scale supervision, the models fail to generalize to tasks beyond specific domains where annotation is feasible, and fall short of developing a holistic physical understanding of objects and their properties. To address these gaps, we introduce **PhyWM**, a **Phy**sically controllable **W**orld **M**odel that takes the state of the world represented by RGB patches (appearance) and provides a natural interface for physical control through flow patches (dynamics)—allowing causal queries such as how the rest of the scene would evolve if a region were set into motion. **PhyWM** can be trained in a self-supervised manner on video datasets and simple zero-shot inference strategies applied to it, unlocks diverse object understanding. **PhyWM** discovers objects by virtually poking different parts of an image and observing which pixels move together. Having discovered object boundaries, **PhyWM** can manipulate them in 3D, by specifying multiple virtual pokes on the object. Finally, we show that **PhyWM** can be used for various forms of physical reasoning such as identifying material properties and understanding inter-object physical relationships. **PhyWM** outperforms both task-specific models and other generative world models on physical object discovery and 3D object understanding. With it's self-supervised pretraining objective and rich physically controllable interface, **PhyWM** emerges as a universal object-understanding model.

## 1 INTRODUCTION

As work in developmental psychology has shown, children in their early months of life already possess notions of objecthood, segmenting the visual world into bounded units that move and interact as cohesive wholes, and acquiring an intuitive understanding of object properties such as their 3D shape, materials and relationship with other objects in the scene Spelke (1990).

Yet in AI, there are still no models that learn all aspects of object understanding directly from raw videos, as humans do. Evaluations on physical object discovery benchmarks Zhang et al. (2023); Ji et al., suggest that current models Ravi et al. (2024), often do not group pixels based on physical notions (such as what moves together) and realigning their outputs with new annotated data can prove expensive. On the other hand, specialized methods for 3D interaction exhibit geometric and appearance inconsistencies in complex scenes Chen et al. (2025). Other capabilities central to object understanding, such as extracting physical material properties such as deformability from input images, remain unsupported by existing methods Tung et al. (2023).

To address the need for better models of object understanding, we introduce **PhyWM**, a physically controllable world model. Formally, **PhyWM** is a probabilistic graphical model (PGM) that accepts the state of the world and defines for every other location, a distribution over possible states. The world state is represented by RGB patches and we provide a natural interface for physical control using flow patches—allowing us to ask causal questions such as the effects of motion applied at a specific location. We implement this PGM using a novel autoregressive sequence modeling architecture, where world states are expressed as sequences of pointer–content token pairs: pointers specify spatiotemporal locations and contents specify locally observed values. This allows us to construct sequences in arbitrary order, enabling the learning of causal dependencies in both time and space unlike standard raster-scan autoregressive vision models which are constrained by fixed serialization Sun et al. (2024).

With this generic world model, we can define multiple inference pathways to pull out rich understanding, beginning with discovering "physical objects" which are defined here as regions that move together when interacted with. To realize this query, we extract two intermediate representations from **PhyWM**: (1) a motion affordance map, indicating regions that are likely to move when external forces are applied, and (2) an expected displacement map, predicting how the rest of the scene would move in response to a virtual poke. We sample poke locations from high-affordance regions

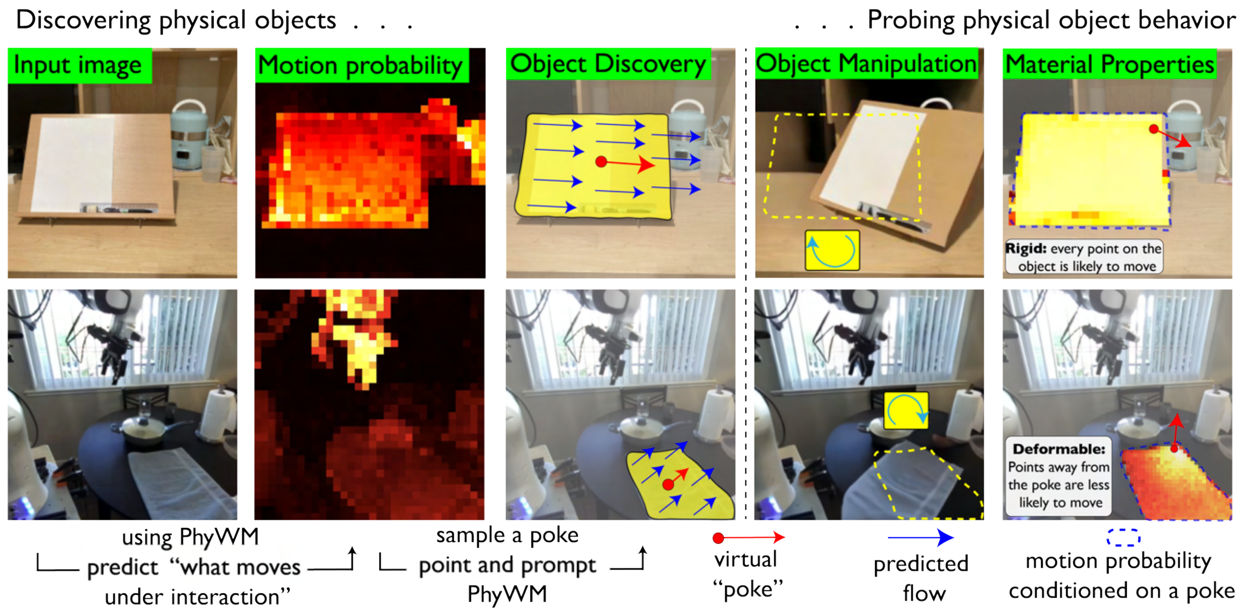
Discovering physical objects . . .
. . . Probing physical object behavior
Input image
Motion probability
Object Discovery
Object Manipulation
Material Properties
Rigid: every point on the object is likely to move
Deformable: Points away from the poke are less likely to move
using PhyWM predict "what moves under interaction"
sample a poke point and prompt PhyWM
virtual "poke"
predicted flow
motion probability conditioned on a poke

**Figure 1: Overview of PhyWM's capabilities.** Our model first predicts a probability of motion map, indicating regions likely to undergo movement independent of camera motion—i.e. candidate movable objects. We sample a point from this map and apply a virtual poke. Conditioned on this intervention, our model completes the flow field. From this, we extract a grouping of pixels, or a "segment" corresponding to an entity that would move as a cohesive whole under the application of external forces (i.e., a physical object). On the right, we illustrate how PhyWM performs 3D object edits by using the segments it discovers to precisely define the physically movable region we desire to manipulate—ensuring that edits are applied to groups of pixels that would move together in the real world as opposed to segments defined based on appearance or semantics. In the last column we show that motion probability maps can be used to explain material properties. Pokes applied on a rigid object yield nearly uniform motion probability across their extent, while deformable objects like cloth show highly localized responses near the poke point.

(as shown in Figure 1) and perform a "statistical counterfactual probing" procedure on PhyWM to isolate objects—a model analog of the physical act of "poking" multiple times at a location in a static image and observing correlated motion patterns in the expected displacement maps predicted by the model. We find that PhyWM outperforms state-of-the-art segmentation models like SAM Ravi et al. (2024) on physical object discovery.

Having discovered these object boundaries, PhyWM can manipulate objects in 3D by specifying flows on an object that represent a desired transformation and predicting the resulting scene appearance while accounting for occlusion, lighting, and shadows. On the 3DEditBench benchmark Lee et al. (2025), PhyWM achieves state-of-the-art results in 3D object manipulation.

Finally, structures extracted from PhyWM support various physical reasoning applications. Motion affordance maps can be used to probe material properties, distinguishing rigid objects, which move uniformly under pokes, from deformable ones, which respond only locally. Expected displacement maps expose inter-object relationships, for example, when an object at the base of a supportive structure is poked, the induced displacements include the supported objects.

## 2 Related Works

**Object discovery models:** Supervised models SAM2 Ravi et al. (2024) produce visually coherent segments but are misaligned with the physical notion of what moves together. Moreover, they depend on expensive annotations Kroemer et al. (2021). Self-supervised methods like CutLER Wang et al. (2023) and ProMerge Li and Shin (2024) group attention maps in pretrained visual encoders Oquab et al. (2023) to discover objects, but are considerably worse than supervised methods. Latent-slot approaches (e.g., Slot Attention Locatello et al. (2020)) encourage object-level decomposition by routing representations through a fixed number of slots, but do not scale beyond simple synthetic datasets. Methods such as EISEN Chen et al. (2022) learn segments from motion, but generalize poorly to natural settings. In contrast, PhyWM defines a powerful physically controllable world model that discovers entities via virtual interactions, yielding physically meaningful object boundaries.

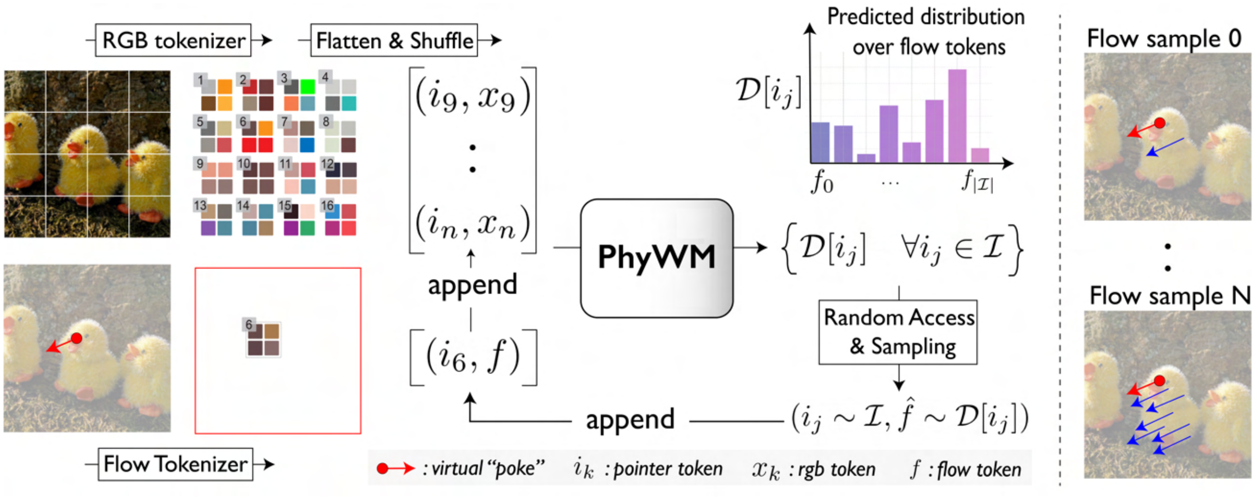

**Figure 2: Using PhyWM for discovering physically co-moving regions.** The **left** panel illustrates the application of **PhyWM** to the task of optical flow completion for physical object discovery. The input consists of a tokenized RGB image and a sparse *virtual poke* indicated by a flow token, $f$. Each token is paired with a pointer token indicating its spatial location, forming a 1D sequence of (pointer, content) pairs. The model accepts this sequence and predicts a categorical distribution $\mathcal{D}[i_j]$ over the flow token vocabulary for *every* spatial location $i_j$ in the image. The **right** panel shows that autoregressively sampling from these distributions yields a complete flow field in pixel space.

**Object manipulation models:** Drag-based image editing methods Wu et al. (2024) perform object manipulation by specifying object transforms as 2D motion vectors. Another class of models Pandey et al. (2024); Gu et al. (2025) use depth-conditioned diffusion models to generate the edited image. Most of these methods are known to exhibit poor performance on complex real-world scenes Chen et al. (2025). Moreover, they rely on spatial segmentation masks obtained by off-the-shelf methods to define the set of pixels to be edited. In contrast, **PhyWM** performs both object manipulation and segmentation using a unified model, and its robust control interface and generative capabilities help achieve superior performance on standard editing benchmarks.

**Emergent object understanding in visual world models.** Visual world models such as CWM Venkatesh et al. (2024) simulate object movement using RGB patch motion counterfactuals and then compute the optical flow between the intervention outcome and the original image to identify segments of pixels that consistently move together. However, as CWM is deterministic, it tends to produce blurry predictions that lead to poor segment quality. In contrast **PhyWM**'s autoregressive generative modeling capability, helps estimate correlated motion statistics over diverse plausible futures, thereby recovering more accurate co-moving segments. Some recent world models add motion-based control to diffusion models Gillman et al. (2025); Chen et al. (2025) to perform various tasks like force prompting, novel view synthesis and object manipulation. However, we show in this paper that they fail to generalize to complex scenes, making it hard to use them as a generic method for object understanding.

# 3 Methods

## 3.1 World model architecture and training details.

In this paper, we build **PhyWM** ($\Phi$), a world model that achieves physical control over scenes, by exposing a prompt interface in *flow space*. This provides a natural, local action space, supporting virtual-interaction driven causal queries, helping us extract rich scene understanding.

**Formulating the world model as a probabilistic graphical model (PGM).** We build $\Phi$ as a PGM (Koller and Friedman (2009)) over spatiotemporal pointer-indexed variables whose nodes can be (i) *RGB patches*, (ii) *flow patches*, or (iii) *global camera tokens*. In its simplest form, $\Phi$ takes a partially specified state of the world as a datum $\mathbf{Z} : S \to \mathcal{V}$, where, $S$ is a subset of all spatio temporal pointers, $\mathcal{I}$ and $\mathcal{V}$ is the set of all content values. For any unobserved pointer $i \in \mathcal{I} \setminus \mathrm{dom}(\mathbf{Z})$, the world model returns the conditional marginal distribution ($\mathcal{D}[i]$) over $\mathcal{V}$, providing an estimate of the state of the world at every other location:

$$\Phi : (\mathbf{Z},\, i \notin \mathrm{dom}(\mathbf{Z})) \longmapsto \{\mathcal{D}\big[(i, v) \mid \mathbf{Z}\big] \;:\; v \in \mathcal{V}\}. \tag{1}$$

**Implementing the PGM as a sequence model.** We implement the PGM as a GPT style transformer Radford et al. (2018) that operates on *pointer* tokens (selecting spatiotemporal locations) and *content* tokens (encoding the value at

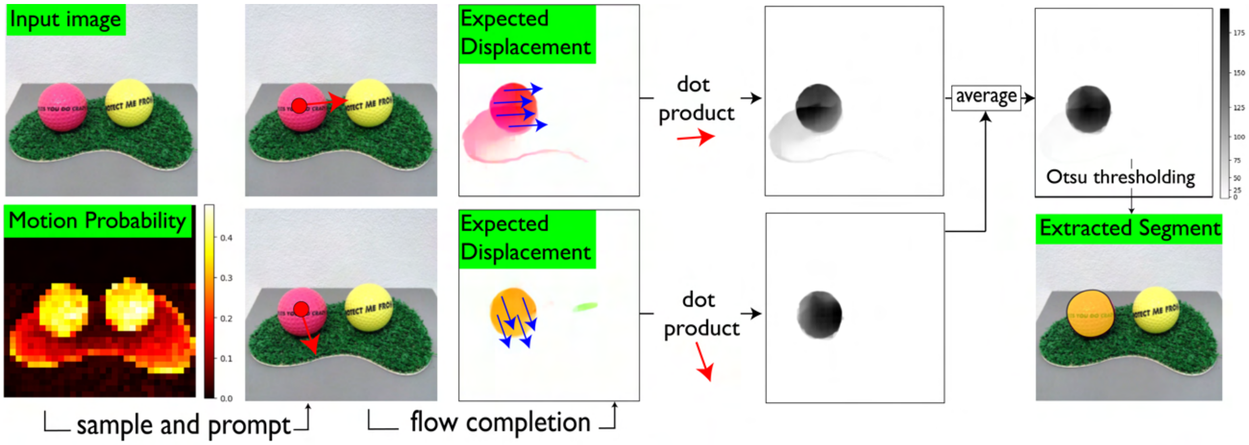

**Figure 3: Physical object discovery using statistical counterfactual probing**. Multiple virtual pokes are applied at a location sampled on the $p_{\text{motion}}$ map that indicates which regions are likely to move under the application of external forces. The average dot product of the poke vector with the expected displacement maps isolates the physical object.

that location). We use learned *local* patch quantizers for both modalities: RGB content tokens $\mathcal{X}$ with RGB pointer tokens $\mathcal{I}^{\text{rgb}}$, flow content tokens $\mathcal{F}$ with flow pointer tokens $\mathcal{I}^{\text{flow}}$; locality enables *distributional queries* about the local state of the world unlike global VQGAN-style tokens Li et al. (2023) (see supplementary Section A.2 for details). Let $\mathbf{z} = [(i_k, v_k), ...]$ denote any valid serialization of $\mathbf{Z}$ into an interleaved pointer–content sequence, where, $i_k \in \mathcal{I} = \mathcal{I}^{\text{rgb}} \cup \mathcal{I}^{\text{flow}}$, $v_k \in \mathcal{V} = \mathcal{X} \cup \mathcal{F}$. Unlike raster-scan autoregressive models Sun et al. (2024), the use of pointer tokens decouples fixed serialization, letting the model capture causal dependencies *both* across space and across time. The PGM conditional in equation 1 is realized as next-token prediction—for each spatiotemporal location, distributional predictions over the content vocabulary are queried by appending a pointer at the end of the sequence:

$$\{\mathcal{D}[i_k] = \Phi\big(\mathbf{z} \oplus [\, i_k \,]\big),\ \forall i_k \in \mathcal{I}\} \;\equiv\; \{\mathcal{D}\big[(i_k, v) \mid \mathbf{Z}\big] : v \in \mathcal{V}\} \tag{2}$$

where $\mathcal{D}[i_k]$ is the categorical distribution over content values queried at pointer $i_k$. We train with standard cross-entropy on content tokens exactly as an LLM, and at inference, we *roll out* by iteratively choosing undecoded pointers and sampling their content—where the type of pointer token we choose (i.e. from $\mathcal{I}^{\text{rgb}}$ or $\mathcal{I}^{\text{flow}}$) decides the modality we decode. We illustrate this procedure in Figure 2.

**Sequence design for training and inference.** In this paper, we consider a simple two-frame, forward-in-time instantiation of the PGM: a natural way to study scene understanding, by modeling plausible future states of a scene. Our sequence begins with the initial frame, represented as pointer–RGB token pairs $\mathbf{x}^0$, followed by a camera motion token $c$, a sequence of pointer–flow token pairs $\mathbf{f}$ and future frame tokens $\mathbf{x}^1$. This results in a forward model trained on the sequence $\mathbf{z} = \mathbf{x}^0 \oplus [c] \oplus \mathbf{f} \oplus \mathbf{x}^1$, which naturally supports useful inference pathways:

- $\mathbf{f} = \Phi(\mathbf{x}^0)$: predicting plausible future motions of an input image.
- $\mathbf{f} = \Phi(\mathbf{x}^0 \oplus [c] \oplus \mathbf{f}^{\text{sparse}})$: spatially completing the flow, given a sparse poke $\mathbf{f}_{\text{sparse}} = [(i, f^{sp})]$.
- $\mathbf{x}^1 = \Phi(\mathbf{x}^0 \oplus [c] \oplus \mathbf{f})$: generating appearance changes resulting from dynamics, $\mathbf{f}$, and viewpoint changes, $c$.

### 3.2 DISCOVERING PHYSICAL OBJECTS WITH PhyWM

**Probing the world model to understand the effect of interactions.** To inject physical control, we probe the model with a virtual poke, $\mathbf{f}^{\text{sparse}}$. Here, as we are only interested in understanding the effect of interactions, we append a zero camera pose token to discount pixel motion due to camera movement. As shown in Figure. 2, $\Phi$ autoregressively rolls out flows by decoding using flow pointer tokens, producing a spatially completed flow field denoted as $\hat{\mathbf{f}}_t \sim \Phi(\mathbf{x}^0 \oplus [c = 0] \oplus \mathbf{f}^{\text{sparse}}; \text{seed} = t)$. A robust estimate of the most likely flow at each location is obtained by computing a statistical average over multiple stochastic generations of the model, producing the **expected displacement field**:

$$\mathbb{E}_{\text{disp}} = \frac{1}{T}\sum_{r=1}^{T} \hat{\mathbf{f}}_t \tag{3}$$

**Statistical counterfactual probing for discovering physical objects given point prompts.** In our framework, *physical objects* are defined through statistical counterfactual probing: sets of pixels that move together under diverse virtual

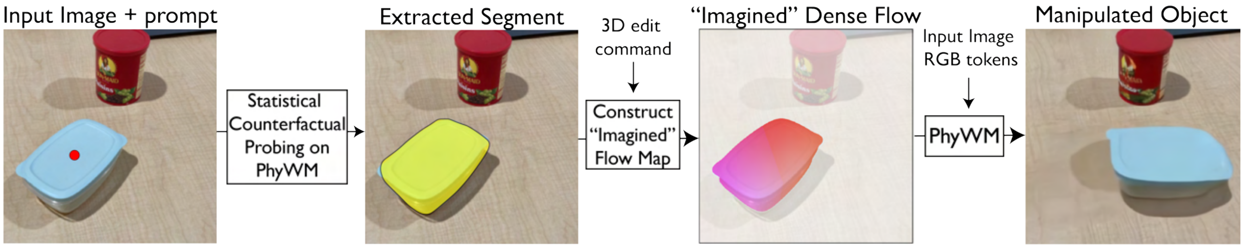

**Figure 4: Physical object manipulation with PhyWM:** Given an input image, a point prompt, and a desired 3D transformation, PhyWM first extracts the corresponding object segment. We then construct a flow field on the object representing the specified transformation and have PhyWM manipulate it by rolling out RGB tokens. Thus, both object discovery and 3D manipulation are achieved within a single unified model.

interactions, as revealed in motion completions generated by PhyWM. To identify such objects, we apply $R$ virtual pokes $f^{(r)}$ at a candidate location $i_k$. For each poke, the model produces an expected displacement field $\mathbb{E}_{\text{disp}}^{(r)}$ given the input sequence $\mathbf{z} = \mathbf{x} \oplus [c = 0] \oplus [i_k, f^{(r)}]$. To extract *physical objects*, we then compute the **expected motion correlation**, as shown in Figure 3, by averaging the dot product between each poke vector and its corresponding displacement field.

$$\bar{\text{dot}} = \frac{1}{R} \sum_{r=1}^{R} \left\langle f^{(r)}, \mathbb{E}_{\text{disp}}^{(r)} \right\rangle . \tag{4}$$

**Automatic segment discovery with motion affordance maps**. In some practical settings such as robotics, we may not have a point prompt to begin with. To discover objects automatically in these cases, we extract candidate locations from the world model where virtual pokes can be applied—regions that are likely to move under external forces (i.e., *movable entities*). We refer to this notion as the probability of motion affordance map, denoted $p_{\text{motion}}$. To compute $p_{\text{motion}}$, we define a set of flow tokens that correspond to motion greater than some threshold $\tau$, and then sum their estimated probabilities. $\mathcal{F}_{\text{motion}} = \{f_j \in \mathcal{F} \mid \|\mathbf{v}_j\|_2 > \tau\}$ , where $\tau$ is a threshold, and $\mathbf{v}_j$ is a 2D flow vector that maps to flow token $f_j$. Supplementary Section A.6 describes how this mapping is done.

Next, given a sequence of RGB tokens $\mathbf{x}$, as we are only interested in finding regions that are likely to move under external forces, we concatenate the sequence with a token indicating zero camera motion to discount it (i.e. $\mathbf{z} = \mathbf{x} \oplus [c = 0]$) and obtain the predicted flow token distributions. Using these distributions, the **probability of motion** at each spatial location $i_k$, is computed by summing over the token set $\mathcal{F}_{\text{motion}}$:

$$p_{\text{motion}}[i_k] = \sum_{f_j \in \mathcal{F}_{\text{motion}}} \mathcal{D}[i_k, f_j] \tag{5}$$

$p_{\text{motion}}$ is thus a 2D heatmap of the regions likely to move under external forces, as illustrated in Figure 3. For discovering objects, we first sample a location that is likely to move: $k$ such that $p_{\text{motion}}(k) > \tau_p$, and then apply statistical counterfactual probing to extract objects. We describe more detailed algorithms in the supplementary Section C.2.

### 3.3 Using PhyWM for physically manipulating discovered objects

We describe here how PhyWM can be used for 3D object manipulation. The pipeline begins with an image, a point prompt and a desired 3D transformation as shown in Figure 4. PhyWM first extracts the corresponding physical object using statistical counterfactual probing, and manipulation is then achieved by specifying a dense flow field, $\mathbf{f}$, representing the 3D transformation (see supplementary Section C.3 for more details). The model then autoregressively rolls out future RGB tokens $\hat{\mathbf{x}}_t \sim \Phi(\mathbf{x}^0 \oplus \mathbf{f}; \text{seed} = t)$, rendering one plausible manipulated outcome. Because the extracted segments correspond to a group of pixels that move together, the resulting edits remain physically plausible—for example, rotating a chair moves the entire chair, rather than incorrectly rotating only a subpart as might occur with a semantic mask from SAM.

### 3.4 From structure extractions to physical reasoning with PhyWM

Structure extractions from PhyWM can be used for various forms of physical scene understanding. When conditioned on a poke, $p_{\text{motion}}$ maps enable reasoning about fundamental object attributes like material properties. Pokes applied on a rigid object yield nearly uniform motion probabilities across their extent, while deformable objects like cloth or plastic covers show highly localized responses near the poke point. Expected displacement maps can provide valuable guidance about how objects might move if interacted with, revealing inter-object physical relationships. For instance, when an object at the base of a stacked structure is virtually poked, the displacement maps include every entity it supports—providing a direct handle on support relationships.

**Table 1: Quantitative evaluation of point-prompted segmentation accuracy across models on SpelkeBench.** We report Average Recall (AR) and mean Intersection over Union (mIoU) for various segmentation methods. PhyWM outperforms both the self-supervised baselines like CWM and supervised baselines like SAM2.

| | MaskFormer | SAM2 | Slot Attention | DINOv2 | CutLer | ProMerge | Force Prompting | Perception AsControl | EISEN | CWM | PhyWM |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AR | 0.439 | 0.482 | 0.115 | 0.225 | 0.321 | 0.342 | 0.051 | 0.071 | 0.158 | 0.327 | **0.541** |
| mIoU | 0.506 | 0.623 | 0.253 | 0.455 | 0.423 | 0.431 | 0.107 | 0.119 | 0.334 | 0.481 | **0.681** |

# 4 Results

## 4.1 Point-prompted segmentation

**Task & metrics** Here, we consider point-prompted physical object discovery: given a point prompt, the goal is to recover the region that would move together if a virtual force were applied at that point. We use our SpelkeBench dataset for evaluation. For each point prompt, we use 8 virtual pokes to perform statistical probing (Section 3.2) with 3 rollouts used to compute $\mathbb{E}_{\text{disp}}$. Segment boundary precision is measured using mean intersection over union (mIoU), while Average Recall (AR) measures detection accuracy as the fraction of segments with IoU $\geq \tau$, averaged over a range of thresholds in [0.50, 0.99].

**Benchmark.** To benchmark physical object discovery, we introduce SpelkeBench, a curated set of images with ground-truth annotations of physical objects defined as pixel groups that move together under virtual pokes, unlike existing datasets such as COCO Lin et al. (2015), which emphasize semantic or instance labels. We curate a dataset from two complementary sources: the EntitySeg benchmark Qi et al. (2023) and the OpenX-Embodiment robotics dataset Collaboration et al. (2023). While EntitySeg is designed for high-resolution internet imagery with dense segmentation annotations, OpenX consists of real-world, egocentric robot interactions. In these datasets, we filter out segments which do not align with our definition of a physical object, and annotate segments when needed. We show some comaprisons between SpelkeBench and other datasets in Figure 6 and more details on our dataset collection procedure are provided in supplementary Section B.2.

**Results.** PhyWM obtains state-of-the-art results on both Average Recall (AR) and mean IoU (mIoU) metrics on SpelkeBench, as shown in Table 1 and qualitatively in Figure 5. We find that self-supervised methods leveraging DINO's Oquab et al. (2023) features—such as CutLer Wang et al. (2023) and Promerge Li and Shin (2024)—tend to merge multiple instances of the same category, since the contrastive learning objective encourages similarity between their representations. Supervised methods like SAM2 Ravi et al. (2024) often segment object subparts that do not move independently, whereas PhyWM 's segments better align with the notion of physical objects as units that move together. For SAM2, which produces multiple masks per prompt, we use the most confident mask. We also attempt to apply our statistical probing procedure to other world models such as CWM Venkatesh et al. (2024), which performs physical interactions through RGB patch motion counterfactuals, and methods like Force Prompting Gillman et al. (2025) and Perception-as-Control Chen et al. (2025), which do so with 2D drag vectors. While drag-based methods

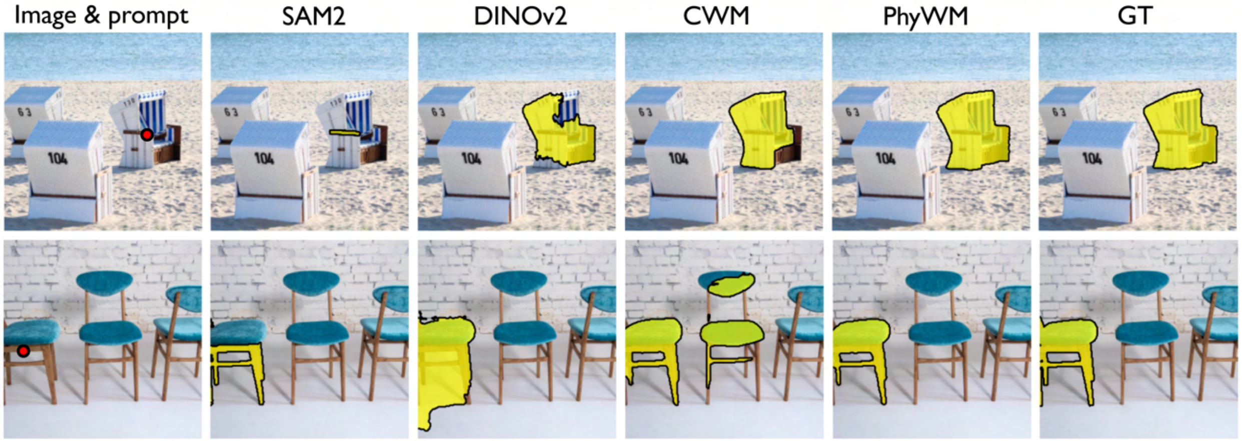

**Figure 5: Qualitative results for point-promoted segmentation across models.** PhyWM yields sharper segments, better aligned with the physical definition of grouping pixels based on co-movement.

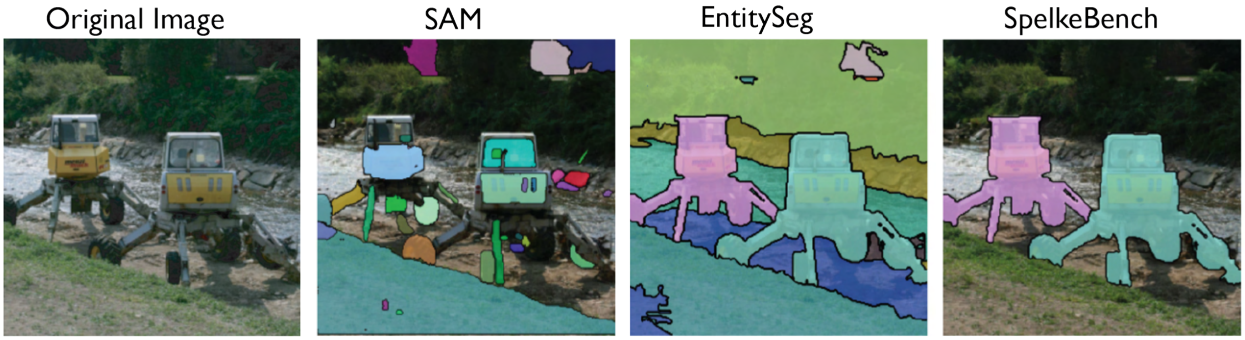

**Figure 6: Benchmarking physical object discovery: comparing our physical segments in `SpelkeBench` with conventional segmentation definitions.** SAM Ravi et al. (2024) produces fine-grained segments but often includes regions that do not typically move independently when forces are applied—such as logos on bottles, shadows, or sub-parts of objects like camera lenses—reflecting its focus on visual distinctiveness over physical structure. Entity segmentation Qi et al. (2023) more closely approximates a physical notion of what moves together but still includes non-movable elements like walls, streets, and fixed shelves. Our **`SpelkeBench`** benchmark contains segments that correspond to physically grounded entities defined by correlated motion in response to applied forces.

generalize very poorly to the complex scenes in **`SpelkeBench`**, CWM actually performs reasonably well, emerging as the strongest self-supervised baseline. Nonetheless, **`PhyWM`** remains superior, as it offers robust physical control over the scene. Finally, baselines such as EISEN Chen et al. (2022), which learn segments from motion, and Adaptive Slot Attention Fan et al. (2024), do not generalize beyond the simple datasets they were trained on. See Supplementary Section F.2 for more qualitative results and analysis of baseline failure modes.

**Ablations.** We also show in supplementary Section E, ablations that study the effect of number of pokes in statistical probing, number of generation seeds used to compute $\mathbb{E}_{\text{disp}}$, number of parameters in the world model, and the importance of flow tokens for enabling physical control in $\Phi$.

### 4.2 AUTOMATIC SEGMENTATION

**Task & Evaluation Metrics.** So far, we have quantified how well our method discovers physical objects given point prompts. However, as we discussed in Section 3.2, it is often desirable to automatically discover every physical object in the scene without any point prompts. To compute the metrics we first match the set of predicted segments to ground-truth segment, by running Hungarian matching Kuhn (1955) on the pairwise IoU cost-matrix. We introduce two additional metrics for this task. Average Precision (AP) measures the fraction of predicted segments that end up being matched and detected (i.e., those that are in fact Spelke objects) across multiple IoU thresholds in the range $\tau = (0.5, 0.99)$. The F1-Score balances AP and AR by computing their harmonic mean. A model that predicts only a few high quality segments may achieve high precision but low recall as it may miss many segments, while a model that over-segments may boost recall at the cost of precision.

**Table 2: Quantitative evaluation of unprompted automatic segmentation across models on `SpelkeBench`.** We find that **`PhyWM`** obtains superior performance compared to existing methods.

| | SAM2 | CutLER | ProMerge | **`PhyWM`** |
|---|---|---|---|---|
| AP | 0.11 | 0.41 | **0.42** | 0.35 |
| AR | 0.62 | 0.32 | 0.34 | **0.46** |
| mIoU | 0.68 | 0.42 | 0.43 | **0.57** |
| F1-score | 0.17 | 0.34 | 0.36 | **0.38** |

**Results.** Overall, we find that **`PhyWM`** outperforms other self-supervised methods such as CutLER Wang et al. (2023) and ProMerge Li and Shin (2024) and supervised methods like Ravi et al. (2024) (see Table 2). ProMerge slightly exceeds **`PhyWM`** in AP due to its tendency to predict fewer segments than those in the GT which align well with ground truth and thus boost precision at the cost of lower recall. CutLER and ProMerge merge objects in the scene belonging to the same category, whereas **`PhyWM`** 's poke-based method harmoniously separates these objects. Compared to supervised methods like SAM2 Ravi et al. (2024), **`PhyWM`** achieves a higher F1-score. Qualitative results shown in Figure 7 suggest that SAM often over-segments scenes based on texture similarity, producing many non-physical segments which lead to poor precision and reduced Interpretability for downstream physical reasoning. More results are provided in supplementary Section F.3.

### 4.3 3D OBJECT MANIPULATION WITH **`PhyWM`**

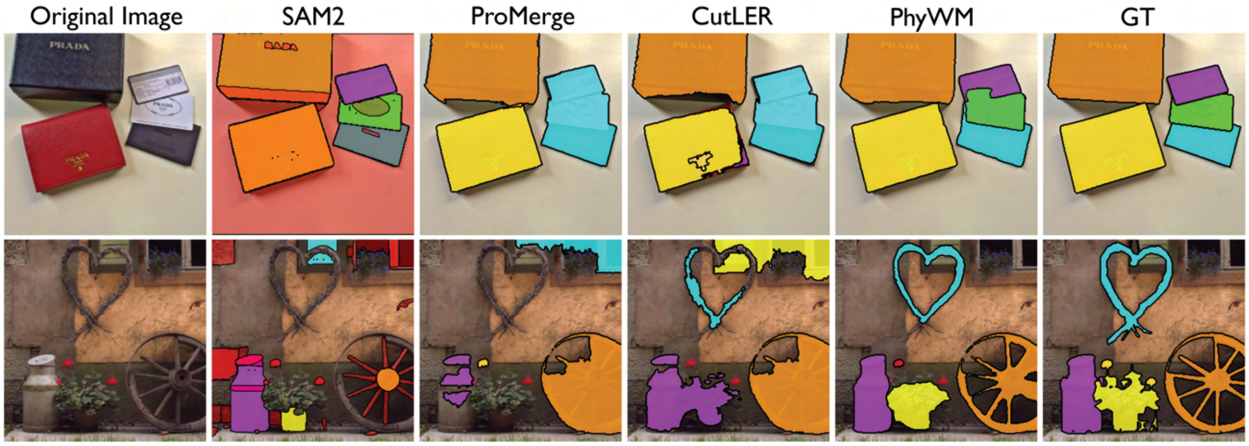

**Figure 7: Qualitative results for auto segmentation on `SpelkeBench`.** `PhyWM` produces a set of physical objects in the scene, where each segment aligns with the co-movement principle. Methods like SAM2 over-segment scenes based on texture similarity, grouping parts that may not move independently, limiting their downstream physical reasoning applications. Both CutLER and ProMerge merge objects that belong to the same category type.

**Task & Dataset.** We consider the task of 3D object manipulation, where a user clicks a point on an object and provides an edit prompt specifying a 3D transformation. To evaluate performance of models, we use `3DEditBench`, recently introduced in Lee et al. (2025). It comprises a diverse range of object types undergoing physical changes such as rotations, translations, and inter-object occlusions. For measuring performance, in addition to standard metrics like PSNR, SSIM, and LPIPS that capture image quality, we include the Edit Adherence (EA) metric introduced in prior work Pandey et al. (2024) which measures physical edit accuracy by computing the IoU between ground truth segment and predicted segment in the edited image.

**Results.** We find that `PhyWM` achieves state-of-the-art object manipulation performance (see Table 3) with both segments from SAM as well as those from `PhyWM` itself. Most existing methods fail on complex scenes in `3DEditBench` as shown in Figure 8. More importantly, the segments extracted from `PhyWM` prove useful in general not just for improving `PhyWM`'s results, but actually consistently outperforms SAM, yielding physically grounded segments that improve realism when used across diverse image editing models (PerceptionAsControl (PasC) Shi et al. (2024), DiffusionHandles (DH) Pandey et al. (2024), Diffusion-as-Shader (DasS) Gu et al. (2025) (see Table 3). In contrast, SAM-generated masks capture only sub-parts of objects, resulting in fragmented or implausible edits (see Figure 8). Additional qualitative results are shown in supplementary Section F.4.

**Table 3: Quantitative evaluation of manipulation quality across segmentation methods and editing pipelines.** Lower ↓ is better, higher ↑ is better. `PhyWM` emerges as the best performing manipulation model (indicated by an underline) and also the best-performing segmentation method (indicated by bold).

| Method | Segment | LPIPS ↓ | SSIM ↑ | EA ↑ |
|---|---|---|---|---|
| PasC | `PhyWM` | **0.195** | **0.672** | **0.679** |
| | SAM2 | 0.241 | 0.658 | 0.536 |
| DH | `PhyWM` | **0.364** | **0.555** | **0.576** |
| | SAM2 | 0.419 | 0.526 | 0.495 |
| DasS | `PhyWM` | **0.194** | **0.707** | **0.640** |
| | SAM2 | 0.253 | 0.682 | 0.503 |
| `PhyWM` | `PhyWM` | <u>**0.161**</u> | <u>**0.736**</u> | <u>**0.776**</u> |
| | SAM2 | 0.183 | 0.720 | 0.633 |

### 4.4 Using structure extractions from `PhyWM` for physical reasoning

As discussed in Section 3.4, `PhyWM`'s structure extractions like probability of motion maps ($p_{\text{motion}}$), expected displacement maps ($\mathbb{E}_{\text{disp}}$) and physical object segments exhibit emergent properties that can be leveraged for physical reasoning. Here, we present qualitative results in Figure 9 supporting this claim. We show a) how poke-conditioned $p_{\text{motion}}$ maps encode information about material type and b) how physical support hierarchies within a scene can be understood using object segments computed using $\mathbb{E}_{\text{disp}}$.

## 5 Conclusion & Future Work

Our work demonstrates a recipe to build a generic self-supervised, physically promptable visual world model and define simple procedures to extract various forms of rich object understanding in a zero shot manner. First, we show that physical objects defined as a collection of stuff that moves together, can be discovered by prompting the model with

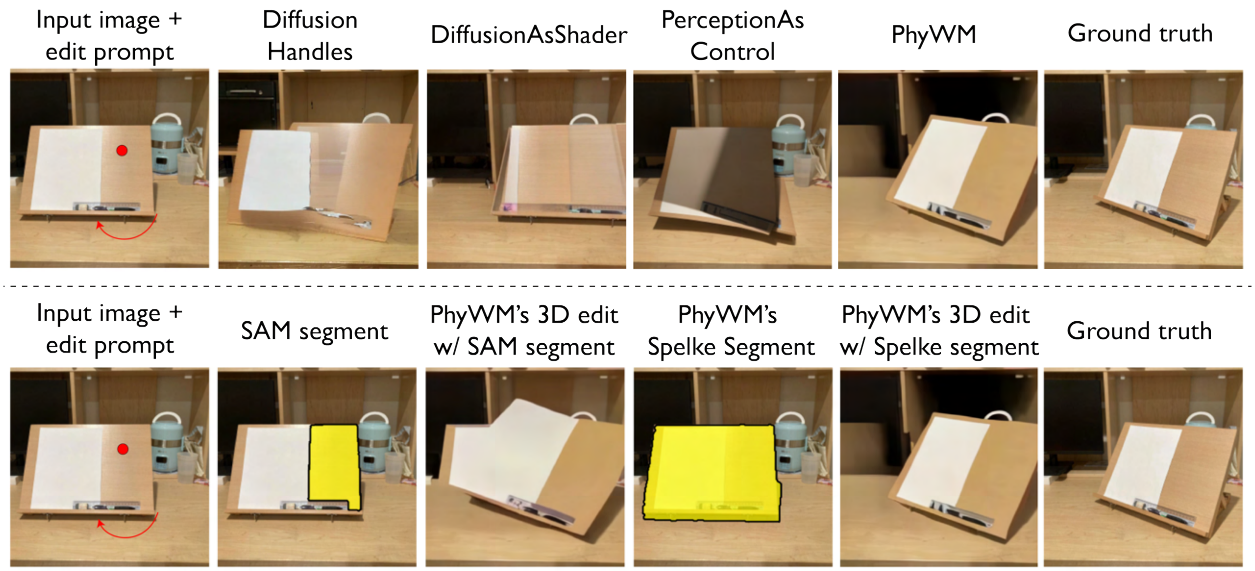

**Figure 8: Physical object manipulation with `PhyWM`.** On the **top** we show compare different object manipulation techniques. We find that most existing methods fail on complex scenes in **`3DEditBench`**, while **`PhyWM`** obtains impressive editing quality. On the **bottom** we show qualitative comparisons of scene edits using SAM masks versus **`PhyWM`** segments. Each row shows the original image, the user click location, and the resulting edited image using segments from different methods. **`PhyWM`**'s physical object segments consistently lead to more plausible manipulation.

virtual interactions and measuring motion correlation patterns in the model's responses. **`PhyWM`** obtains segments better aligned with this physical notion compared to state-of-the-art models like SAM. Having discovered these objects, we show that the same world model can be used for probing complex object behaviors, such as manipulating objects in 3D, understanding material properties and physical support hierarchies in visual scenes. Though our focus in this paper has been on human-centric macroscopic physical scenes, the underlying philosophy of using predictive models to uncover causal and structural patterns through probing could open new avenues for data-driven discovery in other domains where humans have less direct intuition about the nature of objecthood. For example, in medical imaging, a model trained on time-lapse microscopy might help identify cohesive intra-cellular structures, while in astrophysics, models trained on galaxy motions could be probed to discover gravitationally bound systems.

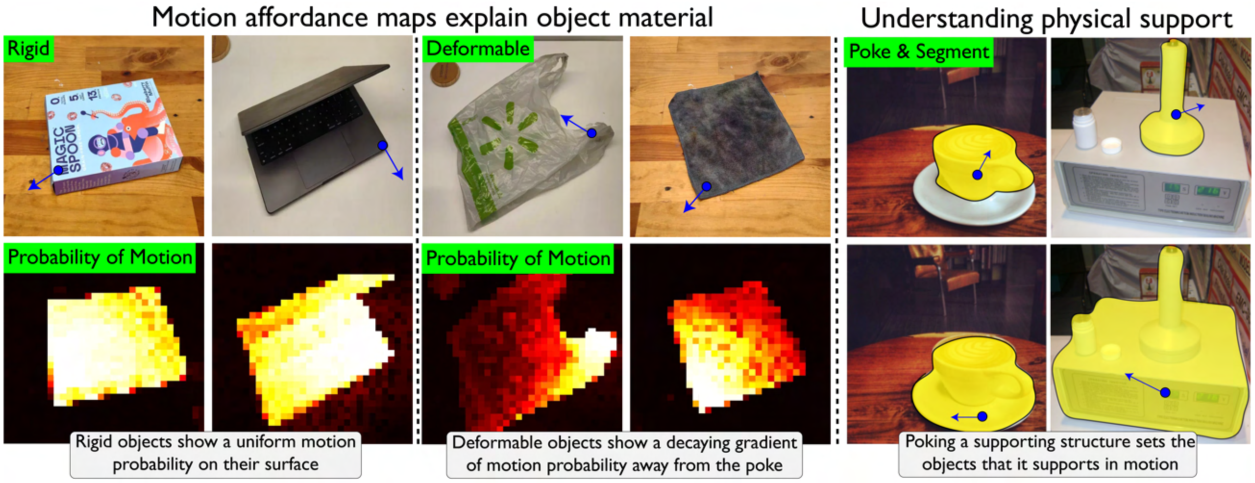
Motion affordance maps explain object material
Understanding physical support
Rigid
Deformable
Poke & Segment
Probability of Motion
Probability of Motion
Rigid objects show a uniform motion probability on their surface
Deformable objects show a decaying gradient of motion probability away from the poke
Poking a supporting structure sets the objects that it supports in motion

**Figure 9: Using structure extractions from `PhyWM` for physical reasoning.** On the **left:** we show how the $p_{\text{motion}}$ map reveals material properties—rigid objects show uniform motion probability across the object while deformable objects concentrate motion near the poke location. And on the **right:** we show that **`PhyWM`** implicitly captures the physical support hierarchy within a scene—when a virtual poke is applied to an object, the extracted segment includes not only the directly contacted object but also any objects it is physically supporting.

## REFERENCES

Elizabeth S Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990.

Chaoning Zhang, Yu Qiao, Shehbaz Tariq, Sheng Zheng, Chenshuang Zhang, Chenghao Li, Hyundong Shin, and Choong Seon Hong. Understanding segment anything model: Sam is biased towards texture rather than shape. *arXiv preprint arXiv:2311.11465*, 2023.

W Ji, J Li, Q Bi, W Li, and L Cheng. Segment anything is not always perfect: An investigation of sam on different real-world applications. arxiv 2023. *arXiv preprint arXiv:2304.05750*.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Yingjie Chen, Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Perception-as-control: Fine-grained controllable image animation with 3d-aware motion representation. *arXiv preprint arXiv:2501.05020*, 2025.

Hsiao-Yu Tung, Mingyu Ding, Zhenfang Chen, Daniel Bear, Chuang Gan, Josh Tenenbaum, Dan Yamins, Judith Fan, and Kevin Smith. Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. *Advances in Neural Information Processing Systems*, 36:67048–67068, 2023.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation, 2024. URL https://arxiv.org/abs/2406.06525.

Wanhee Lee, Klemen Kotar, Rahul Mysore Venkatesh, Jared Watrous, Honglin Chen, Khai Loong Aw, and Daniel LK Yamins. 3d scene understanding through local random access sequence modeling. *arXiv preprint arXiv:2504.03875*, 2025.

Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *Journal of machine learning research*, 22(30):1–82, 2021.

Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023.

Dylan Li and Gyungin Shin. Promerge: Prompt and merge for unsupervised instance segmentation. In *European Conference on Computer Vision (ECCV)*, 2024.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in neural information processing systems*, 33:11525–11538, 2020.

Honglin Chen, Rahul Venkatesh, Yoni Friedman, Jiajun Wu, Joshua B Tenenbaum, Daniel LK Yamins, and Daniel M Bear. Unsupervised segmentation in real-world images via spelke object inference. In *European Conference on Computer Vision*, pages 719–735. Springer, 2022.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation, 2024.

Karran Pandey, Paul Guerrero, Metheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. Diffusion handles: Enabling 3d edits for diffusion models by lifting activations to 3d. *CVPR*, 2024.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. Diffusion as shader: 3d-aware video diffusion for versatile video generation control, 2025. URL https://arxiv.org/abs/2501.03847.

Rahul Venkatesh, Honglin Chen, Kevin Feigelis, Daniel M Bear, Khaled Jedoui, Klemen Kotar, Felix Binder, Wanhee Lee, Sherry Liu, Kevin A Smith, et al. Understanding physical dynamics with counterfactual world modeling. In *European Conference on Computer Vision*, pages 368–387. Springer, 2024.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. *arXiv preprint arXiv:2505.19386*, 2025.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. In *International Conference on Computer Vision (ICCV)*, October 2023.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi ”Jim” Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick ”Tree” Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

Ke Fan, Zechen Bai, Tianjun Xiao, Tong He, Max Horn, Yanwei Fu, Francesco Locatello, and Zheng Zhang. Adaptive slot attention: Object discovery with dynamic slot number. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23062–23071, 2024.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Yujun Shi, Jun Hao Liew, Hanshu Yan, Vincent Y. F. Tan, and Jiashi Feng. Lightningdrag: Lightning fast and accurate drag-based image editing emerging from videos, 2024. URL https://arxiv.org/abs/2405.13722.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Vighnesh Birodkar, Agrim Gupta, Xiuye Gu, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2308.11417. Dataset: ScanNet++.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://arxiv.org/abs/2109.00512. Dataset: CO3D.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM SIGGRAPH Conference Proceedings*, 2018. URL https://arxiv.org/abs/1805.09817. Dataset: RealEstate-10K.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. URL https://arxiv.org/abs/1705.06950. Dataset: Kinetics-400.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz MuellerFreitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. URL https://20bn.com/datasets/something-something/v2. Dataset: 20BN–Something–Something V2.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, efficient, accurate raft for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2405.14793.

Daniel M. Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel L. K. Yamins. Unifying (machine) vision via counterfactual world modeling, 2023. URL https://arxiv.org/abs/2306.01828.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. URL https://arxiv.org/abs/2406.09414.