BenchPRISM: Benchmarking Physical Relationship Understanding In Segmentation Models

Chengze Li
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
leo0912@stanford.edu

Tahmid Jamal
Stanford University
450 Jane Stanford Way, Stanford, CA 94305
tjamal8@stanford.edu

Abstract

We introduce BenchPRISM, a benchmark evaluating segmentation models' ability to identify physically coupled object groups. Our dataset contains 100 images annotated with ground-truth movable segment labels, distinguishing rigid bodies (joint motion under force) from continuous bodies (attachment/support). Evaluating ProMerge, CutLER, and SAM using IoU-based metrics with Hungarian matching, we find ProMerge achieves the highest Average Precision (0.56), while all models demonstrate fundamental limitations in physical reasoning versus visual appearance. The results reveal a significant gap between current segmentation capabilities and the physical understanding needed for robotics applications.

1. Introduction

Modern computer vision has witnessed remarkable advances in image segmentation, with models like the Segment Anything Model (SAM) [6] demonstrating unprecedented zero-shot generalization capabilities across diverse visual domains. SAM and similar foundation models excel at identifying and delineating individual objects based on visual boundaries, achieving impressive performance on traditional segmentation benchmarks. However, these models operate primarily on visual cues—texture, color, and spatial boundaries—without incorporating understanding of the physical world that governs how objects interact and move together in real environments.

And so while the current segmentation models can accurately delineate individual objects within complex scenes, they fail to capture the hierarchical nature of physical object relationships that govern real-world dynamics. Consider a robotic manipulation scenario where a gripper must clear a laboratory workspace: a beaker containing liquid sits within a containment tray, which rests on a mobile cart alongside other scientific instruments. Traditional segmentation approaches would segment the beaker, tray, cart, and instruments as distinct visual entities. However, effective robotic planning requires understanding multiple overlapping movable segments: the beaker can be grasped independently while preserving the liquid, the beaker-tray unit must move together to prevent spillage during transport, the entire instrument cluster moves as one assembly when relocating the cart, and the cart's mobility constraints affect the movement of all supported objects. This multi-scale physical coupling, where objects exhibit different degrees of kinematic constraint depending on manipulation context, represents a critical gap between visual object detection and the physical reasoning required for autonomous robotic systems.

To address this, we construct and release a custom dataset of 100 real-world images, each annotated with ground-truth movable segment labels that reflect context-dependent physical couplings. We benchmark recent segmentation models against this dataset to quantitatively assess their capacity to represent dynamic, relational structures. Through this empirical evaluation, we aim to test the core hypothesis that state-of-the-art visual segmentation models fall short in capturing the physical relationships necessary for robust scene understanding and autonomous manipulation planning. This work establishes a foundation for the development of segmentation models that integrate both visual and physical reasoning, with implications for embodied AI and robotics.

1.1. Problem Statement

We define the problem of movable object group segmentation as the task of identifying and segmenting physically coupled groups of objects that are expected to move together under real-world manipulation. Given a single RGB image as input, the goal is to output a set of instance-level masks, where each mask corresponds to a distinct movable object group. Our annotations are class-agnostic and focus solely on physical groupings, without assigning semantic labels. To evaluate model performance, we compute Inter-

section over Union (IoU) between predicted and ground-truth masks and apply the Hungarian matching algorithm to find an optimal one-to-one correspondence between them. Using this matching, we calculate precision and recall to quantify segmentation accuracy and coverage. All evaluations are conducted on a custom dataset of 100 real-world images containing a diverse set of annotated movable object groups.

2. Related Work

Understanding how objects relate physically within a scene—how they support, attach, or move with one another—has long been a goal in computer vision. Early work such as Roberts' seminal thesis [10] approached this through structured line drawings, proposing that 3D object structure could be inferred from 2D cues. While foundational, its reliance on simplified geometry limited applicability to real-world scenes. Similarly, Biederman's recognition-by-components framework [3] emphasized that humans understand spatial and functional relationships (e.g., cups resting on tables), but neither work operationalized these insights into algorithms for identifying physically actionable groupings.

Later data-driven approaches like the Visual Memex framework [8] related objects through graphs of visual and contextual similarities rather than abstract categories. However, the focus remained on visual retrieval rather than physical reasoning about how scene parts might behave as manipulable units.

Recent work has bridged this gap using counterfactual and causal reasoning. Visual Jenga [2] studies support dependencies by sequentially removing objects from scenes, while EraseDraw [4] trains models to insert objects by first learning to erase them. Both approaches use intervention to understand structural dependencies, though they focus on synthetic manipulation rather than direct segmentation from unaltered scenes.

More formally, Lopez-Paz et al. [7] proposed techniques for detecting causal relationships in visual data, separating correlation from causal influence. Goyal et al. [5] applied counterfactual visual explanations to identify functional features, while Besserve et al. [1] showed that counterfactuals can uncover modularity in generative models—a property aligning with our goal of identifying coherent, movable object groups.

Despite this progress, widely used segmentation models like SAM [6] and SAM2 [9] remain surface-bound, identifying regions of similar appearance without understanding physical interdependence. They cannot distinguish between a mug and its handle as a single object, or recognize a laptop-tray coupling, requiring manual correction in robotics applications.

Our work proposes a benchmark for movable object group segmentation: identifying scene subsets that are physically coupled under constraints like support and joint motion. By focusing on action-relevant segmentation, we bridge the gap between perceptual grouping and manipulation for real-world interaction.

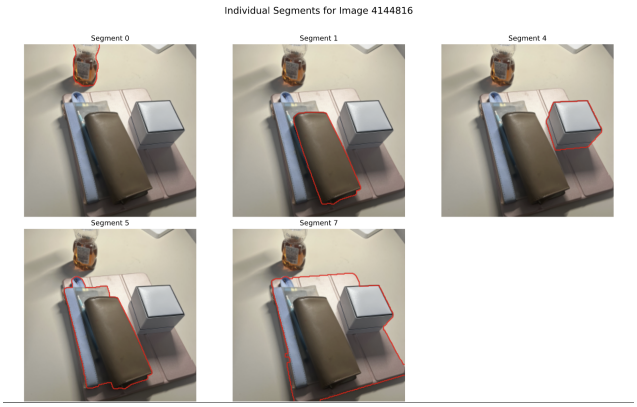

Figure 1. Rigid bodies examples of miscellaneous objects on top of an ipad

3. Dataset

We collected a dataset of 100 images featuring a diverse range of occlusions, lighting conditions, and high-dimensional objects. Each image was cropped and processed using a custom-built software pipeline that utilizes the SAM2 model [9] to get grouped segments that show the physical relationship of the objects. We did this, with the goal of capturing physical dependencies and dynamic relationships within each scene.

During annotation, we define a *movable object* based on two distinct criteria, depending on whether the object is rigid or continuous.

For **rigid bodies**, we define a segment by considering whether an object moves in response to external force applied in arbitrary directions. Specifically, if applying force to one or more objects causes them to move jointly in most cases, they are considered part of the same segment. Otherwise, they are treated as independent segments.

For **continuous and fluid bodies**—such as humans—segmentation is determined based on whether an object is physically attached to or supported in all directions by the continuous body. This criterion captures how such bodies support or constrain other objects within the scene.

For reference, Figure 1 shows physical understanding in rigid bodies and Figure 2 shows in fluid bodies.

During the annotation process, we initially employed the original Segment Anything Model (SAM) to generate base segments for composing our movable object annotations. However, we observed a consistent failure case: SAM frequently segmented visually distinct but physically coupled objects as separate instances, even when clear

Figure 2. Fluid body of the person with a brush attached to its hand

Figure 3. We see an incomplete segment 0, where there are multiple dotted lines/points suggesting failure to distinguish the boundary of a multiple object group segment

human-provided point prompts indicated a unified grouping. This failure was especially prevalent when objects exhibited strong contrast in color, texture, or edge definition, despite being physically interdependent. Figure 3 illustrates one such example, where multiple dotted point prompts intended to guide the model towards a unified movable group instead resulted in an incomplete segment, failing to capture the entire object group as one. Similarly, Figure 4 highlights a mask prediction failure during annotation, where the predicted mask did not align with the true physical grouping, further evidencing the model's inability to physically coupled objects under such conditions.

To mitigate this, we transitioned to using the more recent SAM2 model [9], which demonstrated improved sensitivity to contextual cues. Nonetheless, the same core issue persisted, albeit less frequently. These observations suggest a fundamental limitation in current segmentation models: they rely heavily on local visual features and are not equipped to infer when multiple visually distinct regions should be grouped into a single segment based on physical or functional relationships. This reinforces our central hypothesis and further motivates the need for segmentation models that incorporate physical reasoning beyond boundary-based visual cues.

4. Methods

To evaluate the performance of segmentation models on our custom benchmark for physically movable object groupings, we applied two recent unsupervised segmentation algorithms—CutLER and ProMerge—and compared them to a baseline established using the Segment Anything Model (SAM) [6]. Our approach involved zero-shot evaluation using pretrained models, allowing us to analyze their

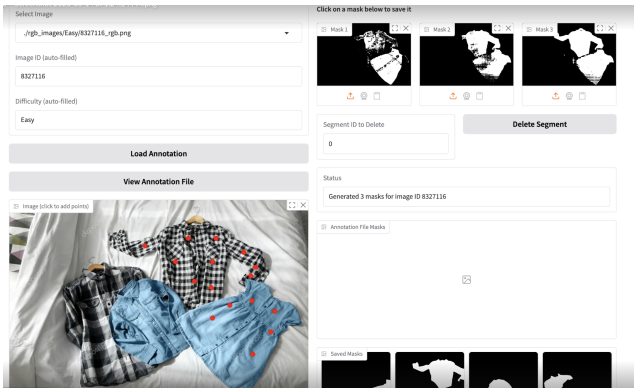

Figure 4. During annotation, our 3 mask predictions failed to capture the correct segment

generalization to physically grounded segmentation tasks without task-specific supervision.

4.1. Dataset and Annotation Protocol

Our dataset consists of RGB images paired with binary segmentation masks, each resized to 256×256 resolution. Masks are stored in `.hdf5` format, with each mask consisting of white pixels (value 1) indicating the foreground segment and black pixels (value 0) representing background. These masks were manually annotated to reflect physically meaningful movable groupings—objects that move together under force due to support or contact constraints. Each image contains one such group as the positive label.

4.2. Model Inference and Processing

We cloned the official repositories of CutLER and ProMerge, both of which provide pretrained weights and inference pipelines. We modified their data loading procedures to handle our fixed image resolution and mask format. Neither model was fine-tuned; we ran inference directly using their default pretrained weights, which were trained on large-scale web or object-centric datasets.

Likewise, for SAM, we used the public checkpoint and evaluated it without modification by directly feeding in our annotated images. For all models, we applied post-processing (if required by the method) to generate predicted binary masks for comparison against the ground truth.

4.3. Segmentation Evaluation Metrics

We evaluated segmentation performance using three standard metrics computed over matched mask pairs: *Average Precision (AP)*, *Average Recall (AR)*, and *Intersection over Union (IoU)*. Given a set of predicted masks $\{\hat{M}_i\}_{i=1}^{N}$ and ground truth masks $\{M_j\}_{j=1}^{M}$, we apply the Hungarian algorithm to compute an optimal one-to-one assignment that maximizes the total IoU across matched pairs.

4.3.1 Metric Definitions

For each matched pair $(\hat{M}_i, M_j)$, we compute the Intersection over Union (IoU) as:

$$\text{IoU}_{ij} = \frac{|\hat{M}_i \cap M_j|}{|\hat{M}_i \cup M_j|}.$$

Let $\mathcal{T} = \{0.50, 0.55, \ldots, 0.90\}$ be a set of IoU thresholds. For each threshold $t \in \mathcal{T}$, we count the number of matched pairs with $\text{IoU}_{ij} \geq t$ and define:

$$\text{Precision}(t) = \frac{\#\text{True Positives at } t}{\#\text{Predicted Masks}} \quad (1)$$

$$\text{Recall}(t) = \frac{\#\text{True Positives at } t}{\#\text{Ground Truth Masks}} \quad (2)$$

We then compute average precision and recall across all thresholds:

$$\text{AP} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{Precision}(t), \qquad \text{AR} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{Recall}(t).$$

Finally, we report the mean IoU over all matched pairs:

$$\text{Mean IoU} = \frac{1}{K} \sum_{k=1}^{K} \text{IoU}_k,$$

where $K = \min(N, M)$ is the number of matched pairs resulting from the Hungarian assignment.

4.3.2 Model Characteristics

CutLER operates by extracting pixel-level embeddings using a vision transformer backbone, followed by spectral clustering and a learned merging step. It is designed to capture both local textures and global object shape, enabling segmentation of arbitrary objects in an unsupervised fashion. The merging module helps reduce oversegmentation and align predictions with true object boundaries.

ProMerge, in contrast, focuses on hierarchical merging. It begins with oversegmented superpixels and learns pairwise affinities to merge regions into coherent, semantically meaningful objects. This strategy is particularly suited for objects with internal structure or multiple connected components, making it potentially valuable for our goal of identifying physical groupings.

SAM leverages a promptable vision transformer architecture capable of producing high-quality segmentation masks from points, boxes, or masks. In our experiments, we used SAM in its zero-shot setting without explicit prompts. That is, we fed SAM the same images used for manual annotation and extracted the default mask predictions. While SAM is not trained to identify physically grounded groupings, it provides a useful baseline for how a general-purpose foundation model performs on our benchmark.

Overall, this evaluation pipeline allows us to assess whether unsupervised or promptable segmentation models can approximate human understanding of physically coupled objects, providing insight into the gap between visual segmentation and physically actionable scene understanding.

5. Experimental Results and Anaylsis

Our evaluation of three state-of-the-art segmentation models on the BenchPRISM dataset reveals significant limitations in current approaches when tasked with identifying physically coupled object groups. Table 1 summarizes the quantitative performance of each model across our key metrics: Average Precision (AP), Average Recall (AR), and mean Intersection over Union (IoU).

5.1. Evaluation Table

Table 1. Performance comparison of segmentation models on BenchPRISM. All models were evaluated in zero-shot setting without task-specific fine-tuning.

Method	Mean AP	Mean AR	Mean IoU
ProMerge	**0.5595**	0.4143	0.4680
CutLER	0.1779	**0.4224**	0.4729
SAM (Baseline)	0.3310	0.4152	**0.4755**

5.2. Model Performance Analysis

ProMerge achieved the highest Average Precision (0.5595), demonstrating superior ability to generate accurate positive predictions for physically grounded object groupings. This performance reflects the model's hierarchical region-merging strategy, which begins from oversegmented superpixels and merges them based on learned pairwise affinities. Such an approach appears well-suited for our benchmark, as it enables the model to recover multi-object groupings that are typically fragmented by edge- or saliency-based segmentation methods.

SAM performed moderately well, with an AP of 0.3310, AR of 0.4152, and mean IoU of 0.4755. While its segmentation masks are visually coherent and precisely delineated, its grouping strategy is not guided by physical relationships such as support, attachment, or containment. Instead, SAM tends to predict a large set of independent segments based on visual saliency and objectness priors, resulting in partial alignment with our physically grounded annotations. The recall and IoU scores suggest that SAM covers a reasonable portion of relevant regions, but its AP score indicates difficulty in precisely identifying physically coupled units without introducing redundant or mismatched segments.

CutLER, with an AP of 0.1779, AR of 0.4224, and IoU of 0.4729, demonstrated the lowest precision of the three models. Despite leveraging powerful visual transformer

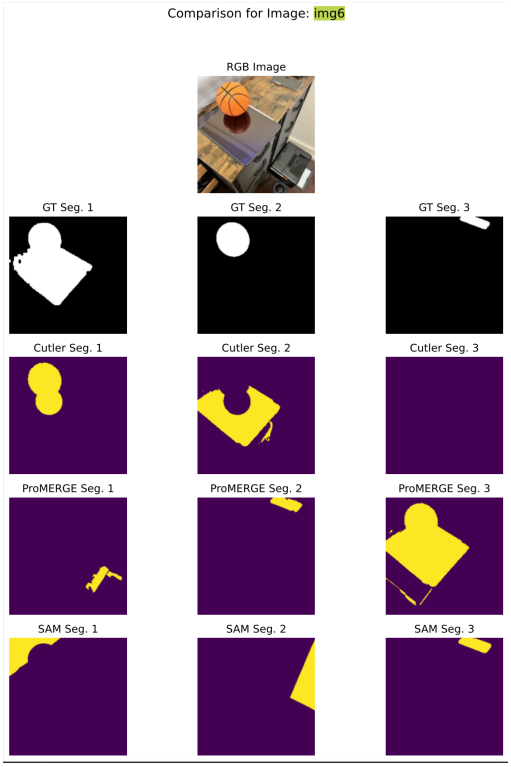

Figure 5. Model segmentation predictions versus ground truth for basketball on iPad. Models fail to recognize the basketball and iPad as a single rigid body that would move together, instead segmenting based on visual boundaries between objects.

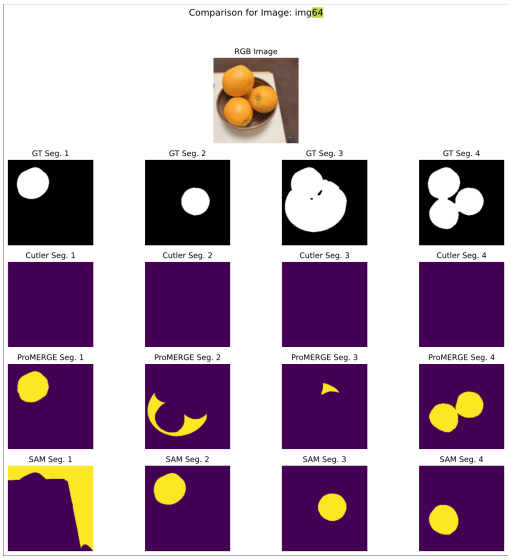

Figure 6. Model segmentation predictions versus ground truth for oranges in bowl. Models over-segment individual oranges rather than recognizing the contained group as a single movable unit when the bowl is manipulated.

embeddings and clustering mechanisms, CutLER struggles to align visual similarity with physical interdependence. Its merging stage mitigates some oversegmentation, but the lack of explicit cues for physical coupling results in undergrouping in scenes with interacting objects. CutLER's relatively high recall and IoU, similar to SAM's, show that it captures relevant regions—but its low AP suggests poor precision in grouping them correctly.

Across all models, the gap between high IoU/AR and lower AP suggests that generating segmentations which overlap with relevant regions is not the core challenge; rather, the main limitation is accurately grouping objects that interact physically into coherent units. ProMerge's learned affinity-based merging provides a partial solution, but all models fall short of capturing the full structure of physical relationships.

5.3. Failure Mode Analysis

The segmentation models demonstrate a fundamental inability to distinguish between visual boundaries and physical coupling relationships, as evidenced by the failure cases in Figure 5 and Figure 6. This visual bias manifests as a consistent pattern where models rely on color, texture, and edge discontinuities rather than understanding which components would move together under applied force.

The most prominent failure mode is undergrouping of physically connected objects. In Figure 5, the models fail to recognize that the toy basketball resting on the iPad constitutes a single rigid body that would move as one unit when the iPad is lifted or tilted. Instead, all three models segment only the ball, its reflection, and the iPad as separate entities, missing the critical support relationship between them. Similarly, in Figure 6, the models incorrectly segment the bowl and the three oranges as individual objects rather than recognizing that the contained oranges form a single movable group with their container. This pattern extends to everyday scenarios such as lids on pots or test tubes in racks, where models consistently produce separate segments for each visually distinct component despite their functional unity.

Containment and support relationships prove particularly challenging for all evaluated models. When objects rest on trays, inside containers, or are otherwise physically dependent on supporting structures, the models segment each item individually rather than recognizing these groupings as manipulation units. This limitation stems from the models' reliance on low-level visual features rather than high-level physical reasoning about mechanical connections, gravitational support, and force transmission.

The relatively low Average Precision scores across all methods (ProMerge: 0.56, CutLER: 0.18, SAM: 0.33) quantitatively reflect this fundamental gap between visual segmentation capabilities and physical understanding. This inability to reason about physical dependencies undermines the models' usefulness for robotics applications, where pre-

dicting object motion and manipulation outcomes depends critically on recognizing physically coupled object groups rather than visually distinct components.

5.4. Implications for Physical Scene Understanding

The overall results highlight a key limitation of current segmentation architectures: their inability to segment scenes based on physical grouping principles. Although models like SAM and CutLER achieve reasonable IoU and recall by identifying salient regions, their low-to-moderate AP scores reveal challenges in grouping visually dissimilar yet physically coupled objects.

ProMerge's relatively higher AP suggests that incorporating learned merging based on region affinity is a promising direction for capturing physically grounded structures. However, it too struggles in cases involving subtle physical cues such as containment, shared motion, or inter-object dependency. For example, nested objects or enclosed groups remain a challenge even for ProMerge, reflecting the need for more explicit modeling of physical relationships.

For robotics and embodied AI, this limitation has direct consequences. Systems that misidentify coupled objects as separate may fail to plan effective grasps or interactions. For instance, attempting to pick up a tray without accounting for its contents may lead to failure in execution. Similarly, separating a pot and its lid can result in unintended outcomes during manipulation tasks.

These findings underscore the importance of developing segmentation models that integrate physical reasoning. Incorporating cues such as support surfaces, motion correlation, and task-driven grouping can help move beyond appearance-based parsing. BenchPRISM thus serves not only as a benchmark but as a diagnostic tool to guide the design of segmentation models that better reflect real-world physical scene structure.

6. Conclusion

This work introduces BenchPRISM, a novel segmentation benchmark focused on identifying physically movable object groupings—segments defined not purely by visual appearance but by support relationships and collective mobility under force. Our benchmark addresses a critical gap in current segmentation research: the need to move beyond visual coherence and toward physically actionable scene understanding. We evaluated three segmentation models—ProMerge, CutLER, and the Segment Anything Model (SAM)—on 100 real-world images annotated with physically grounded segmentation masks.

ProMerge emerged as the most precise model, achieving the highest Average Precision (0.5595). Its hierarchical merging strategy, which operates by learning pairwise affinities among initially oversegmented superpixels, enables the identification of semantically and physically coherent object clusters. This makes ProMerge especially effective at detecting object groups that may differ in appearance but function as a single movable unit, aligning well with the physical reasoning objectives of BenchPRISM.

SAM performed competitively across all metrics, with a Mean AP of 0.3310, AR of 0.4152, and IoU of 0.4755. These results highlight its ability to generate high-quality object masks and cover a substantial portion of physically relevant regions. However, SAM's segmentation behavior remains primarily guided by visual saliency rather than physical interdependence, which limits its precision in grouping adjacent entities that should be treated as a unit under physical manipulation.

CutLER demonstrated the lowest AP (0.1779) but comparable AR and IoU to SAM, reflecting its capacity to localize plausible regions while struggling to group physically related parts into coherent segments. Its reliance on spectral clustering and visual embedding similarity often leads to undergrouping in scenes with object interaction, support, or containment.

Our analysis shows that all evaluated models rely heavily on appearance-based cues, often failing to capture the underlying physical structure of the scene. Groupings based on containment (e.g., dishes in a drying rack) or support (e.g., tools on a tray) are frequently missed, revealing the limits of purely visual segmentation systems when applied to embodied tasks.

For future work, several directions are promising. First, fine-tuning existing models on BenchPRISM may help bridge the appearance-to-physical gap, revealing the degree to which physical grouping can be learned with supervision. Second, integrating physical priors—such as support detection, object interaction modeling, or contact inference—could better align segmentation with manipulation-relevant units. Third, leveraging synthetic scenes from physics simulators may offer a way to expose models to long-tail interaction scenarios not readily available in natural datasets. Finally, expanding BenchPRISM to include multimodal signals such as depth and tactile feedback could enrich models' understanding of physical constraints.

In summary, BenchPRISM provides a new lens on segmentation evaluation grounded in physical reasoning. Our results indicate that while state-of-the-art models such as SAM and ProMerge are effective in localizing object-like regions, they fall short of reliably grouping objects based on physical coupling. Bridging this gap is essential for perception systems deployed in robotics, assistive technologies, and embodied AI—domains where understanding how objects relate physically is as critical as recognizing what they are.

7. Contributions and Acknowledgements

Tahmid and Leo equally worked on building the annotation code for the dataset, collected the dataset, ran the models for evaluation, and wrote the paper equally.

References

[1] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[2] A. Bhattad, S. Liu, P. Abbeel, T. Darrell, and R. Zhang. Visual jenga: Physical scene understanding via counterfactual inpainting. *arXiv preprint arXiv:2403.21770*, 2024. 2

[3] I. Biederman. On the semantics of a glance at a scene. In *Perceptual Organization*, pages 213–253. Routledge, 1981. 2

[4] A. Canberk, M. Bondarenko, E. Ozguroglu, R. Liu, and C. Vondrick. Erasedraw: Learning to insert objects by erasing them from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[5] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2019. 2

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1, 2, 3

[7] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Schölkopf, and L. Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[8] T. Malisiewicz and A. A. Efros. Beyond categories: The visual memex model for reasoning about object relationships. *Advances in Neural Information Processing Systems*, 22, 2009. 2

[9] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[10] L. G. Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 2