SiPhy: Single-Image Physical Property Reasoning

Anonymous CVPR submission

Abstract

Inferring physical properties such as mass, stiffness, and elasticity from a single image is essential for simulation and embodied AI, yet most existing approaches rely on multi-view reconstruction or physics-based supervision. We introduce SiPhy, a unified framework for single-image physical property reasoning that aligns 3D-aware visual cues with language-based material knowledge. From one RGB image, SiPhy samples pseudo-voxel points, extracts CLIP features, and grounds them to material candidates proposed by an LLM. A part-based contrastive aggregator enforces region consistency, while a heaviness-aware refinement improves thickness and volume estimation for dense objects.

Across ABO-500, MVImgNet-100, and PhysXNet-100, SiPhy achieves state-of-the-art single-image *performance—surpassing multi-view reconstruction methods by improving mass MnRE by up to **93%**, reducing density MAE by **35.5%**, and lowering Young's modulus error by **23.5%**. We further validate SiPhy on real hand–object interaction datasets, demonstrating its potential as a scalable toolbox for physical understanding from single-view imagery. Codes will be released upon acceptance.*

1. Introduction

Humans can effortlessly infer how heavy, rigid, or flexible an object is from a single glance. A metal mug and a foam cup immediately evoke distinct expectations about mass, stiffness, and density even without motion cues or interaction. This ability to perform single-view physical reasoning plays a crucial role in human perception, supporting behaviors such as grasping, tool use, and anticipating object dynamics [15]. This naturally raises the question: *Can an AI system learn to infer an object's physical properties from just a single image?*

Despite its importance, this problem remains largely unsolved. Unlike visual attributes, physical properties are latent and cannot be directly observed. They must be inferred from subtle cues in appearance, 3D geometry, and material semantics. Existing methods often rely on multi-view cues [34, 38] or accurate 3D representation from multi-view image captures (*e.g.*, NeRF [26]). These approaches achieve strong performance but require many input views and heavy optimization. As a result, they are impractical in everyday scenarios where only a single image is available. On the other hand, single-image physical reasoning is fundamental to many downstream applications including sound synthesis [12], virtual editing [16], simulation [10], and embodied reasoning [19].

Prior single-image efforts face two main limitations: 1) Lack of 3D geometric awareness. Most prior works focus on pixel-level material recognition [2, 29], treating the problem purely as a 2D classification task agnostic to the object's geometry. This prevents them from estimating 3D-consistent physical quantities such as volume or mass. 2) Lack of physical grounding beyond appearance. Methods such as Image2Mass [31] directly regress physical quantities from RGB appearance through data-driven learning, but visual appearance alone provides insufficient clues regarding physical behavior. Without explicit material knowledge or physicial knowledge, such models fail to generalize to unseen environments, materials, and object compositions.

To address these challenges, we propose **SiPhy**, a **S**ingle-**i**mage **Phy**sical property reasoning framework that unifies 3D geometry, semantics, and language-based physical knowledge. Our key insight is that core advantages of multi-view physical reasoning, including structured geometry, consistent material inference, and physics-aware aggregation, can be approximated from a single RGB image through deliberate architectural design. Siphy consists of three main components: 1) **3D-Aware Visual Sampling**. We perform geometry-aware 2D sampling that approximates voxel centers when lifted into 3D. This enables coarse but spatially structured reasoning from a single view. 2) **Vision–Language Physical Reasoning**. For each sampled region, CLIP provides visual embeddings while a fine-tuned Large Language Model (LLM) proposes material candidates and associated physical attributes such as density, Young's modulus, and thickness. A mask-aware contrastive module aligns visual patches with material semantics and produces material likelihoods for each pseudo-

voxel. 3) **Physical Property Estimation and Refinement**. The physical property value at each point is obtained as the expectation over material likelihoods, and 3D-level quantities (*e.g.*, total mass) are aggregated over all voxels. A heaviness-aware thickness refinement module further improves accuracy as it is directly correlated and sensitive to mass prediction.

Through this design, SiPhy enables both pixel-level and object-level reasoning from a single image. Across diverse dataset including ABO500 [9], MVImgNet100 [37], and PhysXNet100 [4], SiPhy achieves state-of-the-art performance in mass prediction, material segmentation, density estimation, and Young's modulus against both single-view and multi-view baselines. We further demonstrate strong generalization on real hand-object interaction datasets, highlighting its potential for real-world applications. Our contributions are summarized as follows:

- We introduce **SiPhy**, **the first single-image framework** capable of predicting both 2D- and 3D-level physical properties.
- To unify geometry, semantics, and physical knowledge from a single RGB image, we propose a **3D-aware vision–language physical reasoning pipeline** that integrates CLIP-based visual grounding, LLM-driven material and attribute inference, and geometry-aware voxelization.
- Through comprehensive comparisons across multiple datasets, SiPhy achieves state-of-the-art results on ABO500, MVImgNet100, and PhysXNet100 for mass prediction, material segmentation, density estimation, and Young's modulus. We further validate our model on real hand–object interaction datasets, demonstrating strong generalization and high-quality physical property estimation.

2. Related Work

We begin by discussing **Visual Physical Property Reasoning**, which includes both video-based and image-based approaches. Our work falls into the image-based category. We then broaden our discussion to recent advancement in foundations models' physics reasoning, primarily **LLMs' Physics Reasoning**. This involves understanding intuitive physics, which studies how agents or models reason about physical dynamics and object interactions in simulated environments, and text based LLM understanding, where large language models infer or describe physical properties and relationships purely through language.

2.1. Visual Physical Property Reasoning

Reasoning about physical and material properties from visual observations has long been a core challenge in computer vision. Existing research can broadly be divided into two categories. The first performs reasoning given videos where methods such as [32, 33, 35] infer dynamic physical properties that are closely tied to motion (*e.g.*, mass, friction,) from videos by coupling visual perception with a physics engine or differentiable dynamics simulator. Therefore, these approaches remain restricted to controlled laboratory settings and are difficult to generalize or apply in real-world scenarios.

The second line of work aims to infer physical properties from static images only (our work is categorized into this line), using either multi-view or single-view setups. Early works tend to work with single-view image, predicting material of objects or real-world scenes [2, 29]. They used CNN or hand-crafted image features. These works often focused on pixel-level property without understanding spatial structure of the object. Due to this, they cannot predict object-level properties such as mass. To enable mass prediction given a single image, image2mass[31] introduced the first such dataset ABO containing 500 objects. A Xception-style framework is introduced to predict mass. However, its data-driven nature limits the generalization of the proposed model.

Recent works advance image-based physical property reasoning through multi-view 3D reconstruction techniques and large foundation models. Multi-view 3D reconstruction enables detailed geometric representations, while large Visual-Language Models (VLMs) [11, 22] and image–text alignment frameworks such as CLIP [8, 21] enhance semantic reasoning and generalization. For example, NeRF2Physics [38] leverages NeRF to represent 3D scenes and injects CLIP features to model physical properties at each spatial location. It virtually voxelizes objects by assigning appropriate thickness and size to the reconstructed 3D points, through which object-level properties such as mass can be estimated. Similarly, GaussianProperty [34] and PUGS [30] adopt 3D Gaussian Splatting [20], where objects are naturally voxelized through Gaussian primitives for efficient physical reasoning. Noticeable limitations of these methods in real-world applications are their tedious and time-consuming process, as well as their reliance on multi-view images, which are often unavailable or costly to obtain. In this work, we unifies visual-language material reasoning and 3D-aware physical inference and propose a single-image physical property reasoning model, SiPhy, that can infer both pixel-level and object-level properties.

2.2. LLMs' Physics Reasoning

Majority of the efforts focus on proposing benchmarks for evaluating LLM's physics reasoning through visual question answering [1, 3, 6, 7, 23, 39]. These works test agent physic understanding through quiz performed in 2D simulator. These are mainly video-based, covering intuitive physics in the form of quiz and yes/no questions.

Previous works also explored adopting large language

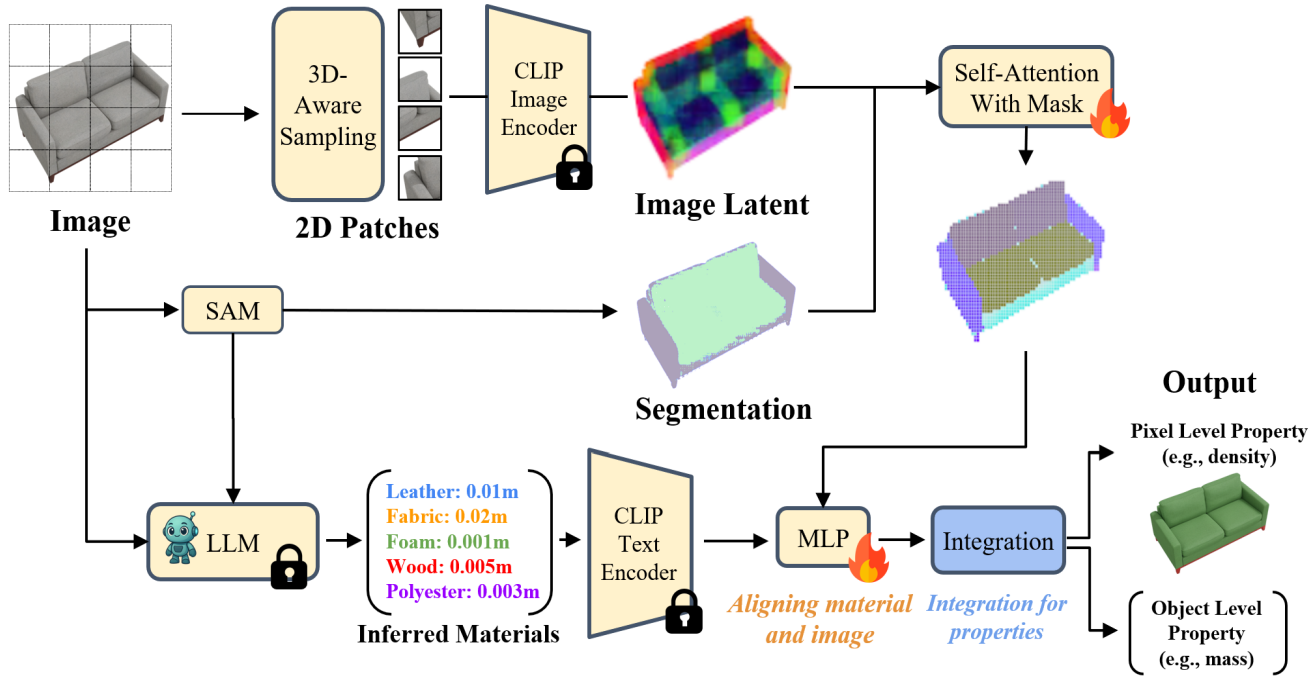


Figure 1. **Overall Architecture of the SiPhy.** Given an image of a object, we first sample 2D points and patches and extract visual features using the CLIP image encoder. The Segment Anything Model (SAM) provides object-part masks that guide the mask-aware self-attention module to aggregate features within material-consistent regions. In parallel, the LLM (e.g., chatGPT) generates material candidates, which are encoded by the CLIP text encoder. A lightweight MLP learns cross-modal similarities between aggregated visual features and material embeddings, forming a similarity table. Finally, object-level properties such as mass are estimated through volume integration over predicted material compositions.

models (LLMs) to predict physical and material properties directly from textual or other symbolic descriptions such as chemical compositions or crystal structures [5, 18, 24, 28]. These works demonstrated that LLMs can regress quantities like material, density, band gap, or elastic modulus, though they remain limited to text inputs. SiPhy follows and expands these works by enabling properties prediction with multi-modality input such as image, text and functional descriptions of objects.

3. Method

Given a single RGB image $I \in \mathbb{R}^{H\times W\times 3}$, our goal is to infer (1) pixel-level physical property maps (e.g., density or Young's modulus) and (2) object-level quantities such as the total mass. We achieve this through a three-stage framework that approximates the structured geometry, semantic consistency, and physics-aware aggregation typically found in multi-view pipelines (Fig. 1).

3.1. 3D-Aware Visual Sampling

Multi-view methods derive geometric awareness by reconstructing the object in 3D and voxelizing its surface. To approximate this from a single image, we design a 3D-aware 2D sampling strategy that selects N spatially distributed points whose back-projections form a coarse pseudo-voxel grid.

Voxel spacing in 2D and Depth estimation. Let d denote the edge length of a desired voxel in 3D. To maintain consistent spatial coverage, we determine the corresponding 2D spacing s as

$$s = d \cdot \sqrt{\frac{f_x f_y}{z}}, \tag{1}$$

where f_x and f_y are the camera intrinsics and z is the estimated object depth. We estimate z using single-view reconstruction by lifting sampled points into metric coordinates and taking the median depth. For scenes with heavy background clutter, a monocular depth model (e.g., DepthAnything [36]) provides a stable alternative.

Point sampling and patch embedding. Using the spacing s, we sample N non-overlapping 2D points over the object. A square patch of side length s is cropped around each point and embedded using a frozen CLIP ViT-B/16 encoder [17], yielding visual descriptors $\{f_i\}_{i=1}^{N} \in \mathbb{R}^{D}$. This stage provides a spatially structured set of 3D-aware visual tokens that approximate voxel centers for subsequent reasoning. The detailed sampling algorithm is presented in Supplmentary materials.

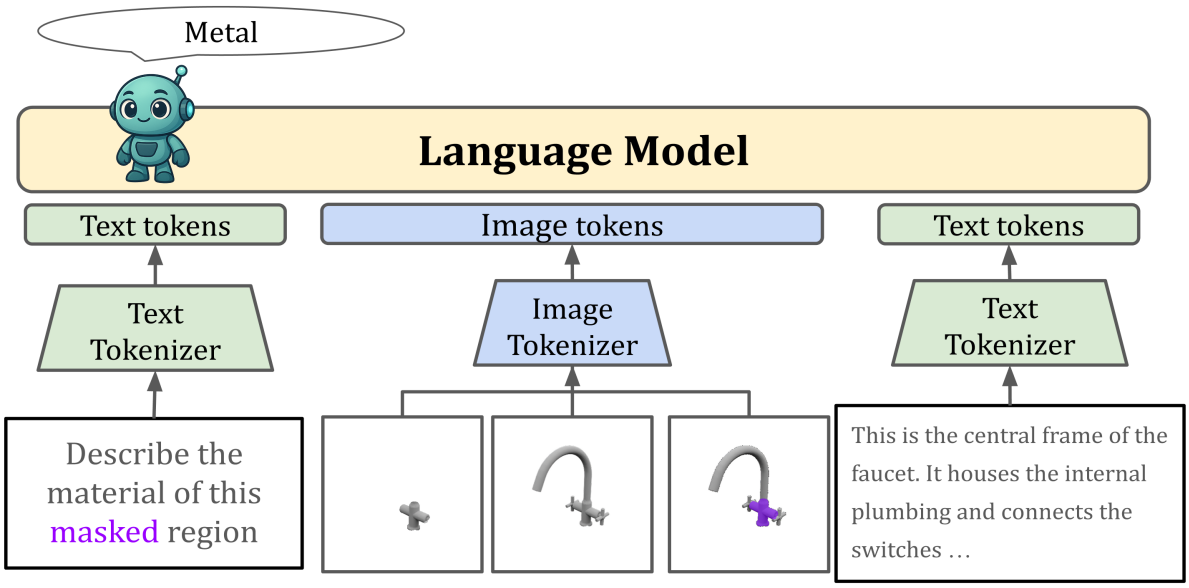


Figure 2. **Material vision–language reasoning module**. Local CLIP features from 2D patches are first aggregated within SAM-derived regions. To overcome the lack of global context, the aggregated embeddings are projected into the LLM token space, where a vision–language model jointly reasons over visual tokens and textual material cues.

3.2. Vision–Language Physical Reasoning

To infer material categories and their associated physical attributes from local appearance cues, SiPhy integrates CLIP-based visual features, LLM-driven material reasoning, and mask-aware contrastive alignment into a unified vision–language module (Fig. 2). This stage produces a material likelihood matrix $P \in \mathbb{R}^{N\times K}$ for the N sampled points.

Material proposal via LLM. Inspired by recent vision–language models such as LLaVA [25], which demonstrate that LLM decoders contain rich world knowledge and reasoning ability, we employ an LLM to propose candidate materials and physical attributes for each object part. Since objects commonly exhibit part-level material consistency, we first obtain part proposals using SAM masks [27]. For each mask region, we compute a visual token by averaging CLIP embeddings of the points within the mask and project this aggregated feature into the LLM token space using a learnable linear projection layer.

We also tokenize high-level object descriptions (either provided by the dataset or extracted using models such as BLIP [22]) and concatenate them with the projected visual tokens. The LLM decoder jointly processes these tokens and outputs a list of K plausible materials along with their associated physical attributes (*e.g.*, density, Young's modulus, thickness).

CLIP visual and text embeddings. For each sampled point, the s-sized image patch is embedded by the CLIP ViT-B/16 image encoder [17], yielding a set of visual feature vectors $\{\mathbf{f}_i\}_{i=1}^{N}$. The material names predicted by the LLM are tokenized and fed into the CLIP text encoder to generate the corresponding text embeddings $\{\mathbf{t}_k\}_{k=1}^{K}$. The similarity between $\mathbf{f}_i$ and $\mathbf{t}_k$ provides an initial estimate of material likelihood, which is then refined through spatially aware alignment.

Part-based contrastive alignment. Material is generally consistent within the same physical part of an object. However, a direct CLIP similarity between visual and text embeddings does not explicitly enforce this structure. To incorporate *part-level coherence*, we introduce a part-based contrastive alignment module that encourages points within the same part to share similar material predictions, while separating points belonging to different parts.

For each sampled feature $\mathbf{f}_i$, we define the part-aware neighborhood $\mathcal{N}(i)$ as the set of points that fall within the same SAM-derived part mask as i. We then apply self-attention restricted to this part region:

$$\tilde{\mathbf{f}}_i = \mathrm{softmax}\left(\frac{(\mathbf{f}_i W_Q)(\mathbf{F}_{\mathcal{N}(i)} W_K)^\top}{\sqrt{d_a}}\right)\mathbf{F}_{\mathcal{N}(i)} W_V, \quad (2)$$

where $W_Q, W_K \in \mathbb{R}^{D\times d_a}$ and $W_V \in \mathbb{R}^{D\times D}$ are learnable projections and d_a is the attention dimension.

The part-refined features are then concatenated with the

corresponding material text embeddings and passed through a lightweight multilayer perceptron (MLP).

Training objective. We optimize the similarity matrix S using a supervised contrastive loss [21] defined over part groupings:

$$\mathcal{L}_{\text{CL}} = -\frac{1}{N}\sum_{i}\frac{1}{|\mathcal{P}(i)|}\sum_{p\in\mathcal{P}(i)}\log\frac{\exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_p)/\tau)}{\sum_{a\neq i}\exp(\text{sim}(\mathbf{z}_i,\mathbf{z}_a)/\tau)}, \tag{3}$$

where $\mathcal{P}(i)$ denotes the set of samples belonging to the same SAM-derived part as i, τ is a temperature scalar, and $\text{sim}(\mathbf{u},\mathbf{v}) = \mathbf{u}^\top\mathbf{v}$ is the cosine similarity between embeddings.

We use S_{dot} as a regularizer to prevent the module from smoothing embeddings without preserving the material discrimination for each 2D point:

$$\mathcal{L}_{\text{align}} = \frac{1}{NK}\,\mathcal{D}(S_{\text{MLP}},\, S_{\text{dot}})\,, \tag{4}$$

where $\mathcal{D}(\cdot,\cdot)$ denotes the distance between the learned and teacher similarity matrices, instantiated as either ℓ_2 distance $\|S_{\text{MLP}} - S_{\text{CLIP}}\|_F^2$ or cross-entropy divergence.

The overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CL}} + \lambda\,\mathcal{L}_{\text{align}}, \tag{5}$$

where λ balances the alignment strength. This joint formulation encourages intra-part smoothness while keeping the learned similarities consistent with CLIP's embedding geometry.

3.3. Physical Property Estimation and Refinement

Pixel-level property prediction. Given the material likelihoods produced by the vision-language reasoning stage, we compute the physical property at each sampled point as the expectation over material-specific attributes.

Let $P \in \mathbb{R}^{N\times k}$ denote the likelihood matrix for N sampled points and K candidate materials. The predicted property at point i is:

$$\hat{V}_i = \sum_{j=1}^{k} p_{i,j} V_j, \tag{6}$$

where $p_{i,j}$ is the likelihood of material j and V_j is its corresponding attribute (e.g., density or thickness). The probabilities are normalized such that $\sum_{j=1}^{k} p_{i,j} = 1$ for each i. Values for non-sampled pixels are propagated through k-Nearest Neighbor interpolation.

Object-level property prediction. To calculate object level property, such as mass, we follow NeRF2Physics [38] to treat thickness as a property like mass density, and we calculate the expected volume and mass of each point (as if there is a voxel around that point). Eventually, the mass of the object is obtained as the sum of the mass of each point.

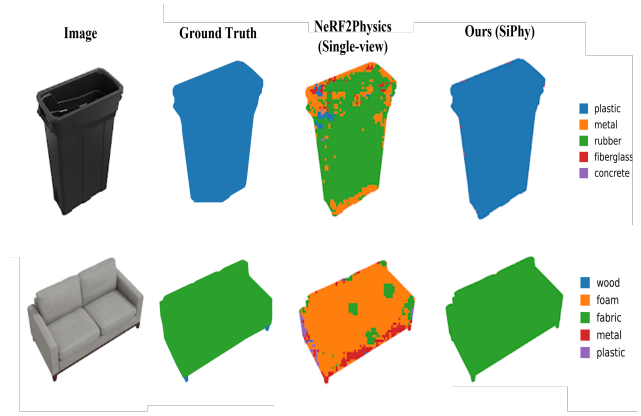
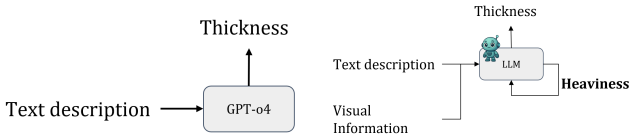


Figure 3. NeRF2Physics lacks spatial awareness. Our part-based attention module enforces that by motivating points within the same masks to show high similarity to same material



(a) NeRF2Physics (b) Ours

Figure 4. NeRF2Physics and our thickness estimation method

Heaviness-aware thickness. We found that the material prediction has weak correlation to mass (proof in Supplementary), which is surprising given the expected value of the mass is the multiplication of mass density and volume. This motivated us to focus on improving thickness prediction. According to Tab. 4, the current thickness method is performing worse on heavy objects. Therefore, we propose a thickness estimation method that are more mass-aware, using heaviness of the object as a prior to refine thickness. We first use SiPhy's initial predictions to classify the object as *heavy* or *light*. This classification is then used to adjust the LLM prompt for thickness, biasing the prediction toward thickness ranges that are physically plausible for the inferred heaviness category.

4. Experiments

4.1. Experimental Settings

Dataset. We evaluate SiPhy on the ABO dataset [9], which provides multi-view posed images, segmentation masks, and ground-truth physical properties such as mass for thousands of everyday objects. Unlike NeRF2Physics [38], which requires multi-view images for 3D reconstruction, our model performs physical property prediction using only a single RGB view. Following prior work, we use the ABO500 subset containing 500 objects, each with 30 uniformly sampled views captured from a viewing hemisphere.

Table 1. Mass prediction evaluation on the ABO-500 test set. "M" and "S" denote multi-view and single-view inputs.

Method	View	ADE (↓)	ALDE (↓)	APE (↓)	MnRE (↑)
NeRF2Physics [38]	M	8.74	0.78	1.06	0.55
PUGS [30]	M	30.30	1.59	7.68	0.30
LlaVa [25]	S	17.33	1.89	1.84	0.31
Image2Mass [31]	S	12.50	1.79	**0.98**	0.31
SiPhy (Ours)	S	**7.78**	**0.74**	1.00	**0.58**

We adopt the same split as NeRF2Physics: 300 training, 100 validation, and 100 test scenes.

Baselines. We compare SiPhy with the following baselines, including prior state-of-the-art:

- **NeRF2Physics** [38] uses NeRF to reconstruct the object, fuses vision-language features to each point with CLIP, and estimates the physical properties of each point using a retrieval-based approach. Then mass is estimated by volumetric integration.
- **PUGS** [30] uses Gaussian Splatting with Shape Aware and Region-Aware Feature Contrastive loss to reconstruct the object. They also fuse vision-language features to each point with CLIP, and propagate based on region-aware features. Mass of each gaussian is calculated from multiplying the expected density and volume of each gaussian.
- **LLaVa** [25] We follow NeRF2Physics procedure to evaluate LlaVa to predict mass of the object given 2D image as input.

Implementation. For part-based contrastive alignment, we apply a local self-attention module restricted to points sharing the same SAM-derived part. Each CLIP image embedding (D=512) is normalized and passed through an attention layer with attention dimension 64. We optimize the supervised contrastive loss using Adam with a learning rate of 1×10^{-4} and temperature $\tau = 0.1$.

For vision–language feature fusion, we use OpenCLIP ViT-B/16 pretrained on DataComp-1B. LLM responses for material attributes, including density and Young's modulus, are generated using GPT-4. We set the number of candidate materials to $K = 5$ and sampling temperature to $T = 0.1$. Captions are generated using Instructional BLIP-2 [11] with Flan-T5-XL.

4.2. Quantitative Comparison to SOTA methods

4.2.1. Mass Evaluation

We evaluate mass prediction using the same metrics as NeRF2Physics [38]: Absolute Difference Error (ADE), Absolute Log Difference Error (ALDE), Absolute Percentage Error (APE), and Minimum Ratio Error (MnRE). MnRE is considered the most reliable metric due to its scale invariance. To ensure a fair and comprehensive comparison, we report results on both the official test split and the full dataset for all models. GaussianProperty [34] does not release mass-estimation code and cannot be evaluated.

As shown in Table 1, SiPhy consistently outperforms all baselines across major metrics. On the primary metric MnRE, our method improves over NeRF2Physics by **5.5%** and over PUGS by **93.3%**, demonstrating strong robustness despite relying only on a single-view input. Our ADE is also substantially lower (7.78 vs. 8.74 for NeRF2Physics and 30.30 for PUGS), indicating improved absolute mass accuracy. These results highlight the advantage of single-view reasoning with structured vision–language physical inference.

Table 2. Material segmentation performance across datasets.

Dataset	Model	mIoU (↑)	M-mIoU (↑)
ABO500 [9]	NeRF2Physics [38]	0.18	0.31
	PUGS [30]	0.22	0.40
	GaussianProperty [34]	**0.29**	**0.49**
	SiPhy (Ours)	0.25	0.42
MVImgNet100 [37]	NeRF2Physics [38]	-	-
	PUGS [30]	-	-
	GaussianProperty [34]	**0.19**	0.23
	SiPhy (Ours)	**0.19**	**0.25**
PhysXNet100 [4]	NeRF2Physics [38]	0.03	0.06
	PUGS [30]	0.04	0.07
	GaussianProperty [34]	**0.14**	0.21
	SiPhy (Ours)	**0.14**	**0.31**

Table 3. Material density and Young's Modulus estimation results on PhysXNet100. "M" and "S" denote multi-view and single-view inputs.

Property	Model	View	MAE (↓)
Density (kg/m^3)	SiPhy (Ours)	S	1315
	NeRF2Physics	M	2044
	PUGS	M	**1297**
Young Modulus (GPa)	SiPhy (Ours)	S	**52**
	NeRF2Physics	M	68
	PUGS	M	68

4.2.2. Material Segmentation Evaluation

As shown in Table 2, SiPhy achieves strong material segmentation performance across all three datasets. On ABO-500, SiPhy improves over NeRF2Physics by **38.9%** mIoU and **35.5%** M-mIoU, and over PUGS by **13.6%** mIoU and **5.0%** M-mIoU. Although GaussianProperty achieves a higher score on ABO-500 due to its multi-view reconstruction, SiPhy surpasses all single-view baselines. On

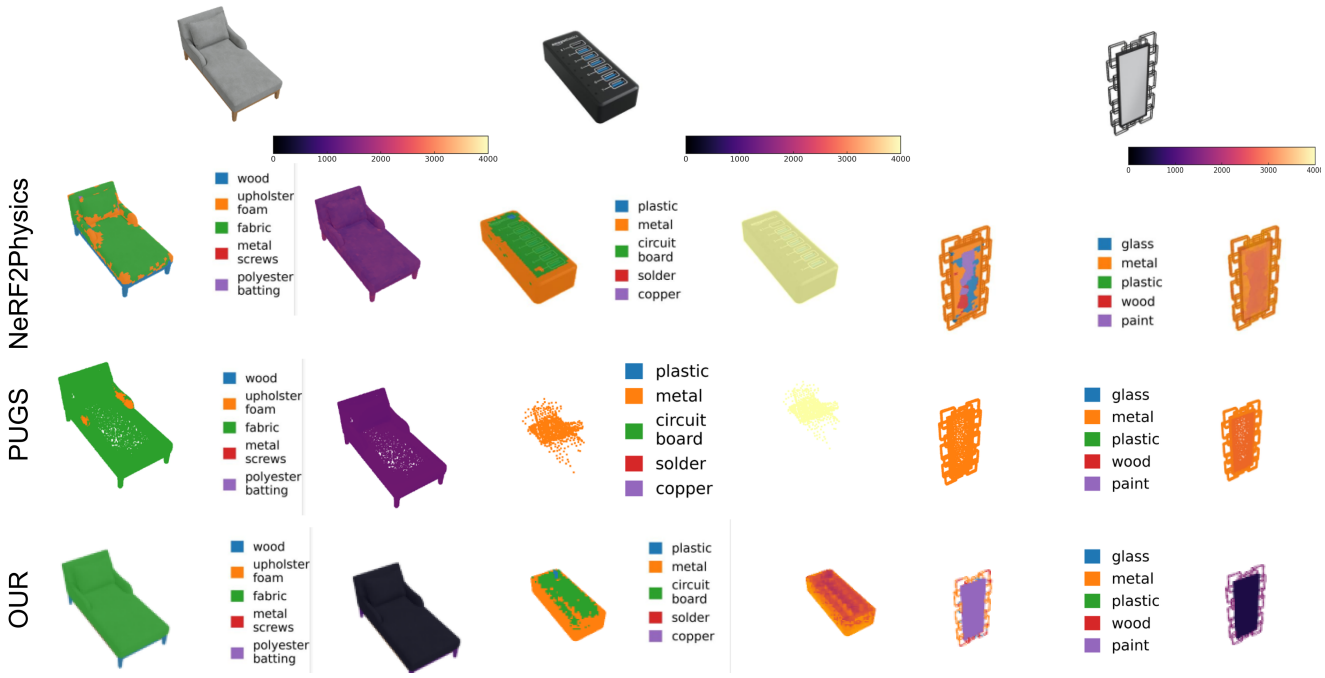


Figure 5. **Qualitative results on ABO500.** SiPhy yields more coherent and complete material segmentation than NeRF2Physics and PUGS, especially on thin structures and fine-grained parts.

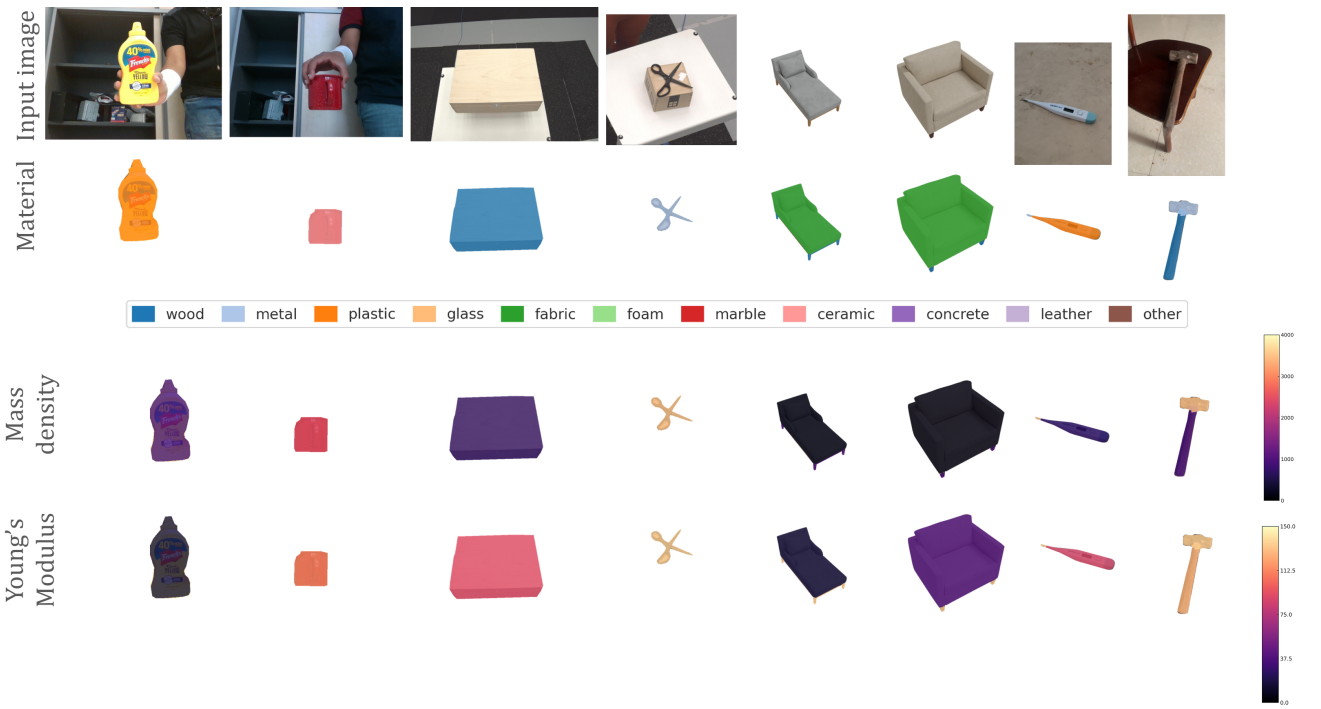


Figure 6. **Qualitative results on HO3D, ARCTIC, ABO500, and MVImgNET.** From a single image, SiPhy predicts material labels, mass density, and Young's modulus with consistent and physically plausible outputs across real and synthetic settings.

MVImgNet100, SiPhy matches GaussianProperty in mIoU (0.19) and achieves an **8.7%** improvement in M-mIoU. On PhysXNet100, SiPhy demonstrates the largest gains, outperforming NeRF2Physics by **366%** mIoU and **417%** M-mIoU, and outperforming PUGS by **250%** mIoU and **343%** M-mIoU. Compared to GaussianProperty, SiPhy improves M-mIoU by **47.6%**. These results highlight the effectiveness of our part-based alignment and vision–language physical reasoning in achieving robust single-view material segmentation.

4.2.3. Density and Young's Modulus Evaluation

We further evaluate density and Young's modulus prediction on the PhysXNet100 dataset, which provides per-pixel annotations for both properties. Following prior work, we compute mean absolute error (MAE) between the predicted and ground-truth maps. As shown in Table 3, SiPhy achieves competitive or superior performance compared to multi-view baselines. On density estimation, SiPhy reduces the MAE by **35.5%** compared to NeRF2Physics, despite using only a single-view input. SiPhy performs comparably to PUGS, with only a small 1.4% gap, even though PUGS relies on full multi-view Gaussian reconstruction. For Young's modulus, SiPhy achieves the best performance, reducing MAE by **23.5%** relative to both NeRF2Physics and PUGS. These results demonstrate that our vision–language physical reasoning and part-based alignment generalize effectively to fine-grained material attributes, outperforming reconstruction-heavy multi-view methods under the single-view setting.

4.3. Ablation Study

Worse performance on heavy objects. Table 4 shows that thickness prediction errors increase sharply for heavy objects: without refinement, ADE rises from 2.07 on light objects to 18.86 on heavy ones, with a similar drop in MnRE. Because mass depends on both density and volume, thickness errors have a larger influence on dense or voluminous objects, motivating a heaviness-aware refinement.

Effectiveness of heaviness-aware thickness refinement. Using the object's inferred heaviness as a prior, our refinement reduces ADE for heavy objects from 18.86 to 15.59 (**17.3%** improvement) and increases MnRE from 0.57 to 0.65, with smaller gains on light objects. This indicates that heaviness-aware guidance stabilizes thickness estimation and improves mass prediction accuracy.

Real-world evaluation. On HO3D and ARCTIC, SiPhy produces coherent material segmentation and physically plausible density and stiffness estimates. As shown in Table 5 and Table 6, SiPhy outperforms LLaVA in mass prediction (MnRE 0.67 vs. 0.58 on HO3D; 0.57 vs. 0.53 on ARCTIC) and achieves lower error in density and Young's modulus. On the ABO dataset (Fig. 5), SiPhy also provides smoother and more consistent segmentation than NeRF2Physics and PUGS, demonstrating its robustness as a practical physical-property annotator for real-world robotics.

Table 4. Performance comparison between Our method and NeRF2Physics on ABO500 splits (heavy and light objects), and the effect of the **heaviness-aware thickness (*hat*)** module. Improvements are marked in purple.

Model	Split	ADE (↓)	ALDE (↓)	APE (↓)	MnRE (↑)
Ours *w/o hat*	ABO500	9.84	0.81	0.85	0.53
	ABO500-heavy	18.86	0.64	0.54	0.57
	ABO500-light	2.07	0.73	0.57	0.58
Ours	ABO500	7.78	0.74	1.00	0.58
		+2.06	+0.07	−0.15	+0.05
	ABO500-heavy	15.59	0.48	0.46	0.65
		+3.27	+0.16	+0.08	+0.08
	ABO500-light	3.57	0.85	1.29	0.54
		−1.50	−0.12	−0.72	+0.04

Table 5. Mass estimation on real-world datasets.

Dataset	Model	ADE(↓)	ALDE(↓)	APE(↓)	MnRE(↑)
HO3D [14]	LLaVA [25]	0.24	**0.48**	0.58	0.58
	SiPhy (Ours)	**0.19**	0.50	**0.41**	**0.67**
ARCTIC [13]	LLaVA [25]	**0.21**	**0.60**	0.50	0.53
	SiPhy (Ours)	0.26	0.92	**0.46**	**0.57**

Table 6. Other properties estimation on real-world datasets.

Dataset	Model	Property	ME (↓)
HO3D [14]	LLaVA [25]	Mass density	1308
	SiPhy (Ours)	Mass density	**1192**
ARCTIC [13]	LLaVA [25]	Young's modulus	74
	SiPhy (Ours)	Young's modulus	**62**

5. Conclusion

We presented SiPhy, a single-image framework that unifies geometric cues, semantic understanding, and language-driven physical knowledge to estimate a wide range of physical properties. Despite operating without multi-view supervision, SiPhy achieves strong performance across synthetic and real-world datasets, enabled by part-based contrastive alignment and a heaviness-aware refinement that stabilizes volume estimation for dense objects. While our approach offers a scalable alternative to reconstruction-heavy pipelines, it remains constrained by the quality of single-view geometric priors and the inherent ambiguity of inferring physical attributes from a single RGB image. Future work may incorporate stronger depth priors and view-independent evaluation methods.

References

[1] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning, 2019. 2

[2] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page –, 2015. Materials in Context Database (MINC). 1, 2

[3] Florian Bordes, Quentin Garrido, Justine T Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments, 2025. 2

[4] Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. *arXiv preprint arXiv:2507.12465*, 2025. 2, 6

[5] Akshat Chaudhari, Chakradhar Guntuboina, Hongshuo Huang, and Amir Barati Farimani. Alloybert: Alloy property prediction with large language models, 2024. 3

[6] Zhenfang Chen, Kexin Yi, Yunzhu Li, Mingyu Ding, Antonio Torralba, Joshua B. Tenenbaum, and Chuang Gan. Comphy: Compositional physical reasoning of objects and events from videos, 2022. 2

[7] Anoop Cherian, Radu Corcodel, Siddarth Jain, and Diego Romeres. Llmphy: Complex physical reasoning using large language models and world models, 2024. 2

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2818–2829. IEEE, 2023. 2

[9] Jasmine Collins, Shubham Goel, Achleshwar Luthra, Leon Xu, Kenan Deng, Xi Zhang, Tomas F. Yago Vicente, Himanshu Arora, Thomas Dideriksen, Matthieu Guillaumin, and Jitendra Malik. ABO: dataset and benchmarks for real-world 3d object understanding. *CoRR*, abs/2110.06199, 2021. 2, 5, 6

[10] Rishit Dagli, Donglai Xiang, Vismay Modi, Charles Loop, Clement Fuji Tsang, Anka He Chen, Anita Hu, Gavriel State, David IW Levin, and Maria Shugrina. Vomp: Predicting volumetric mechanical property fields. *arXiv preprint arXiv:2510.22975*, 2025. 1

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2, 6

[12] Yiming Dou, Wonseok Oh, Yuqing Luo, Antonio Loquercio, and Andrew Owens. Hearing hands: Generating sounds from physical interactions in 3d scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1795–1804, 2025. 1

[13] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[14] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 8

[15] Jessica Hamrick, Peter Battaglia, and Joshua B Tenenbaum. Internal physics models guide probabilistic judgments about object dynamics. In *Proceedings of the 33rd annual conference of the cognitive science society*. Cognitive Science Society, 2011. 1

[16] Hao-Yu Hsu, Chih-Hao Lin, Albert J Zhai, Hongchi Xia, and Shenlong Wang. Autovfx: Physically realistic video editing from natural language instructions. In *2025 International Conference on 3D Vision (3DV)*, pages 769–780. IEEE, 2025. 1

[17] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 3, 4

[18] Ryan Jacobs, Maciej P. Polak, Lane E. Schultz, Hamed Mahdavi, Vasant Honavar, and Dane Morgan. Regression with large language models for materials and molecular property prediction, 2024. 3

[19] Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. *arXiv preprint arXiv:2503.17973*, 2025. 1

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 2

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning, 2021. 2, 5

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2, 4

[23] Shiqian Li, Kewen Wu, Chi Zhang, and Yixin Zhu. I-phyre: Interactive physical reasoning, 2024. 2

[24] Youjia Li, Vishu Gupta, Muhammed Nur Talha Kilic, Kamal Choudhary, Daniel Wines, Wei keng Liao, Alok Choudhary, and Ankit Agrawal. Hybrid-llm-gnn: integrating large language models and graph neural networks for enhanced materials property prediction††electronic supplementary information (esi) available. see doi: https://doi.org/10.1039/d4dd00199k. *Digital Discovery*, 4 (2):376–383, 2024. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 4, 6, 8

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[27] 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. 4

[28] Andre Niyongabo Rubungo, Craig Arnold, Barry P. Rand, and Adji Bousso Dieng. Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions, 2023. 3

[29] Lavanya Sharan, Ce Liu, Ruth Rosenholtz, and Edward H. Adelson. Recognizing materials using perceptually inspired features. *International Journal of Computer Vision*, 103(3): 348–371, 2013. 1, 2

[30] Yinghao Shuai, Ran Yu, Yuantao Chen, Zijian Jiang, Xiaowei Song, Nan Wang, Jv Zheng, Jianzhu Ma, Meng Yang, Zhicheng Wang, Wenbo Ding, and Hao Zhao. Pugs: Zero-shot physical understanding with gaussian splatting, 2025. 2, 6

[31] Trevor Standley, Ozan Sener, Dawn Chen, and Silvio Savarese. image2mass: Estimating the mass of an object from its image. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 324–333. PMLR, 2017. 1, 2, 6

[32] Jiajun Wu, Ilker Yildirim, Joseph J Lim, Bill Freeman, and Josh Tenenbaum. Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 2

[33] Jiajun Wu, Erika Lu, Pushmeet Kohli, Bill Freeman, and Josh Tenenbaum. Learning to see physics via visual deanimation. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[34] Xinli Xu, Wenhang Ge, Dicong Qiu, ZhiFei Chen, Dongyu Yan, Zhuoyun Liu, Haoyu Zhao, Hanfeng Zhao, Shunsi Zhang, Junwei Liang, and Ying-Cong Chen. Gaussianproperty: Integrating physical properties to 3d gaussians with lmms. *arXiv preprint arXiv:2412.11258*, 2024. 1, 2, 6

[35] Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua B. Tenenbaum, and Shuran Song. Densephysnet: Learning dense physical object representations via multi-step dynamic interactions, 2019. 2

[36] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 3

[37] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023. 2, 6

[38] Albert J Zhai, Yuan Shen, Emily Y Chen, Gloria X Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang. Physical property understanding from language-embedded feature fields. In *CVPR*, 2024. 1, 2, 5, 6

[39] Zhicheng Zheng, Xin Yan, Zhenfang Chen, Jingzhou Wang, Qin Zhi Eddie Lim, Joshua B. Tenenbaum, and Chuang Gan. Contphy: Continuum physical concept learning and reasoning from videos, 2024. 2