

SiPhy: Single-Image Physical Property Reasoning

Hoang Le¹, Joonwoo Kwon¹, Elkhann Ismayilzada¹, Yufei Zhang², and Zijun Cui¹ 

¹ Michigan State University, East Lansing MI 48864, USA

² Independent Researcher

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, depth 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 VLM. 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%** (vs. PUGS), reducing density MAE by **35.5%** (vs. NeRF2Physics), and lowering Young’s modulus error by **23.5%**. We further validate SiPhy on real hand-object interaction datasets, demonstrating its potential as a data annotation engine for physical understanding from single-view imagery. Codes are available at <https://github.com/DominoAI-Lab/SiPhy-ECCV-2026>.

Keywords: Physical Reasoning · Vision Language Model · Single-Image Understanding

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 [17]. 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 [36, 41] or accurate

3D representation from multi-view image captures (*e.g.*, NeRF [29]). 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 [14], virtual editing [18], simulation [12], and embodied reasoning [22].

Prior single-image efforts face two main limitations: 1) Lack of 3D geometric awareness. Most prior works focus on pixel-level material recognition [3, 31], 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 [33] 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 physical knowledge, such models fail to generalize to unseen environments, materials, and object compositions.

To address these challenges, we propose **SiPhy**, a **Single-image Physical** 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) **Material Probability Estimation**. For each sampled region (centered with sampled points), CLIP provides visual embeddings while a fine-tuned Vision Language Model (VLM) 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 [11], MVImgNet [40], and PhysXNet100 [5], 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 and depth 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, VLM-driven material and attribute inference, and geometry-aware voxelization.
- Through comprehensive comparisons, 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 its generalization on real hand-object interaction datasets and demonstrate its downstream utility by using SiPhy’s material predictions as physical priors for image-to-audio generation.

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.

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 [34, 35, 37] 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 [3, 31]. 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 [33] 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) [13, 25] and image–text alignment frameworks such as CLIP [9, 24] enhance semantic reasoning and generalization. For example, NeRF2Physics [41] 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 [36] and PUGS [32] adopt 3D Gaussian Splatting [23], 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 unify 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.

LLMs’ Physics Reasoning. Majority of the efforts focus on proposing benchmarks for evaluating LLM’s physics reasoning through visual question answering [2, 4, 7, 8, 26, 42]. These works test agent physics 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 models (LLMs) to predict physical and material properties directly from textual or other symbolic descriptions such as chemical compositions or crystal structures [6, 21, 27, 30]. 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 both pixel-level physical property maps, such as density or Young’s modulus, and global object-level quantities like total mass.

We first introduce the SiPhy VLM, a separately fine-tuned vision-language model that proposes material candidates used by the subsequent physical reasoning pipeline (Sec. 3.1). As shown in Fig. 1, the pipeline consists of three interconnected stages. First, we perform 3D-aware visual sampling to identify spatially distributed 2D points on the object surface, which approximate a coarse geometric structure from a single view (Sec. 3.2). Second, CLIP encodes the patch around each sampled point, and the resulting visual features are aligned with VLM-proposed material candidates to estimate point-wise material probabilities (Sec. 3.3). Finally, we compute the physical properties by integrating localized material likelihoods across the sampled points and apply a heaviness-aware refinement module to stabilize the final object-level predictions (Sec. 3.4).

3.1 SiPhy VLM

To infer part-level materials, SiPhy uses a fine-tuned VLM. For each SAM-derived object part, the VLM predicts K candidate materials, which are encoded

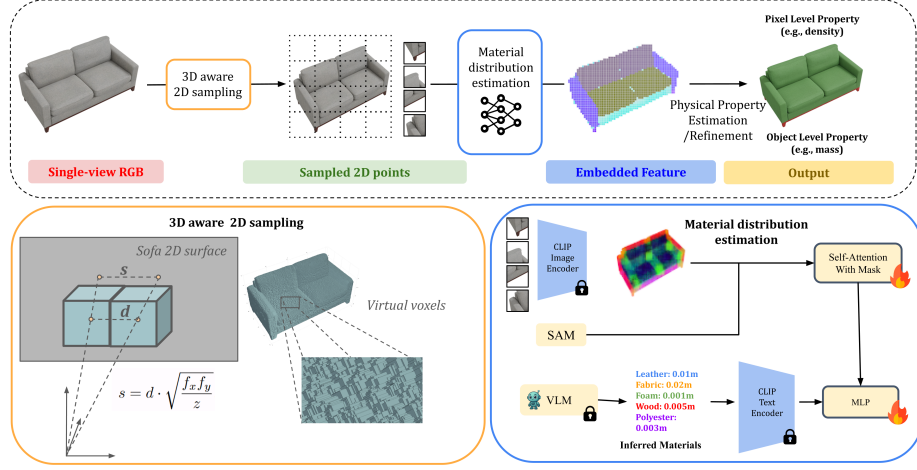


Fig. 1: Overall Architecture of the SiPhy. (1) **3D-aware 2D sampling**: spatially distributed 2D points are sampled from the object surface to approximate coarse geometry from a single view. (2) **Material probability estimation**: CLIP visual features, guided by SAM masks, are aligned with material candidates proposed by a finetuned VLM to produce material probability maps. (3) **Physical property prediction**: localized material probabilities are integrated over the sampled points to infer the properties

by the CLIP text encoder and used by the material probability estimation stage (Sec. 3.3).

The VLM follows a standard LVLM design [1, 28, 39] with a frozen CLIP ViT-L/14 visual encoder [20], a two-layer MLP projector, and Vicuna-7B-v1.5 [10] as the language backbone. For each object part, it takes three visual inputs: the part mask, the cropped part image, and the full object image (Fig. 2). The three CLS features are assigned positional encodings, projected into Vicuna’s token space, and concatenated with the text prompt and GPT-4-generated part description for material prediction.

Training uses two stages. We first train only the projection module on approximately 10k PhysX- and MVImgNet-derived samples, while keeping the visual encoder and LLM frozen. We then fine-tune Vicuna with LoRA and jointly train the projector, with the visual encoder still frozen. We treat material prediction as a classification task: the attention-masked mean-pooled last hidden state of the LLM is passed through a linear classifier and optimized with cross-entropy loss, reducing the output space to K material classes rather than the full LLM vocabulary. At inference, the predicted material names are encoded by CLIP and passed to Sec. 3.3.

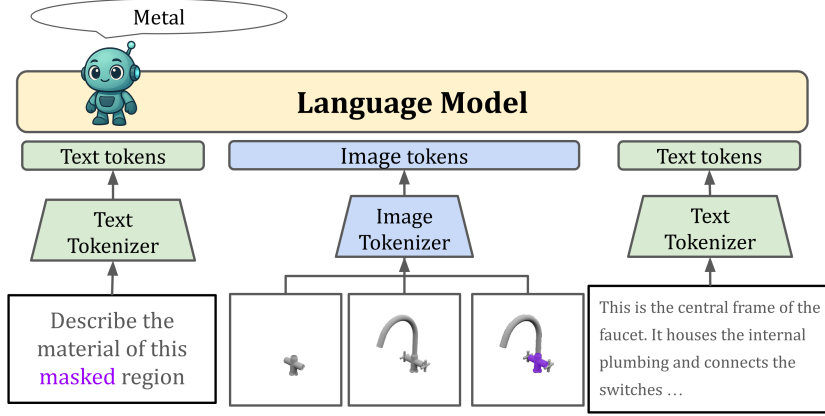


Fig. 2: SiPhy VLM (Sec. 3.1). For each SAM-derived object part, the model receives three visual inputs (part mask, cropped region, full object image) alongside a text prompt and a GPT-4-generated part description. Vicuna-7B-v1.5 jointly processes all inputs and predicts the material category (e.g., “Metal”).

3.2 3D-Aware Visual Sampling

Pixel-level properties can be directly inferred from local visual cues and are closely related to semantic segmentation tasks. In contrast, object-level property prediction, such as total mass, requires a thorough spatial understanding of the object. Such an understanding necessitates 3D voxelization, which involves partitioning the object using a *unit object* representation, such as virtual cubes or Gaussians.

Existing methods adopt two primary voxelization strategies. NeRF2Physics [41] utilizes virtual cubes. We refer to these as *virtual* because the unit cubes do not explicitly exist; rather, they are represented implicitly by their centers’ 2D projections and their 3D dimensions. Alternatively, GaussianProperty [36] and PUGS [32] reconstruct objects using Gaussian Splatting, where Gaussian primitives naturally serve as voxels. In this work, we follow the virtual cube voxelization approach, where each unit is parameterized by its edge length d and its thickness. Notably, thickness is not directly observed but is inferred from the predicted material distribution, for instance, by computing its expectation under the material probability field.

Our objective is to derive the 2D projections of these virtual 3D unit cubes, conditioned on a single RGB image and a predefined side length d (side length of the virtual unit cube). While previous efforts like NeRF2Physics [41] necessitate multi-view reconstruction to facilitate 3D point downsampling and projection, we demonstrate that a structured representation is attainable using only single-view imagery and metric depth estimates. To maintain consistent spatial cover-

age and approximate a pseudo-voxel grid from a single perspective, we determine the adaptive 2D pixel spacing s as follows:

$$s = d \cdot \sqrt{\frac{f_x f_y}{z}} \quad (1)$$

where f_x and f_y denote the camera intrinsics, and z represents the estimated object depth. In scenarios characterized by significant background clutter, we employ a monocular depth model (e.g., Depth Anything [38]) to provide a stable geometric prior. Fig. 1 illustrates how this sampling strategy yields a physically grounded distribution of 2D unit cube projections. Our experiments confirm that a naive fixed-spacing baseline results in suboptimal performance (Tab. 1), primarily due to virtual voxel overlap and redundancy (Fig. 3). Since s depends on estimated depth and camera intrinsics, we analyze sensitivity to perturbations of these quantities in the Supplementary.

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 [20], 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 Supplementary materials.

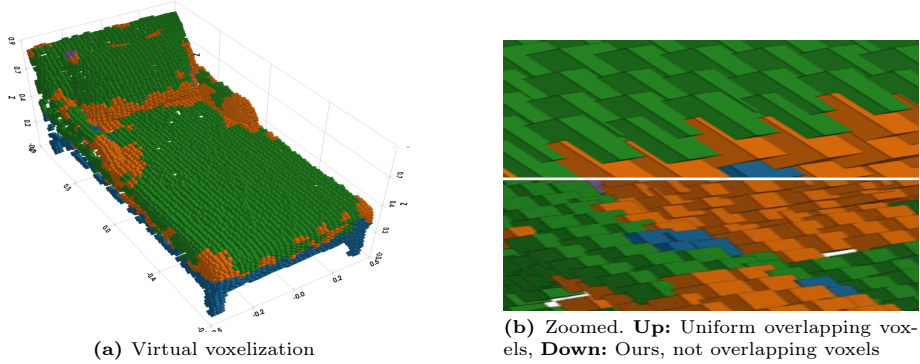


Fig. 3: 3D-aware pseudo-voxelization. (a) Depth-adaptive sampling voxels (colored rectangulars). (b) Zoomed comparison: fixed uniform spacing (top) causes voxel overlap and redundancy; our adaptive spacing (bottom) produces non-overlapping coverage that better approximates a surface voxel grid.

3.3 Material Probability Estimation

To infer material categories and their associated physical attributes from local appearance cues, SiPhy integrates CLIP-based visual features, VLM-driven material reasoning (Sec. 3.1), and mask-aware contrastive alignment. This stage

produces a material likelihood matrix $P \in \mathbb{R}^{N \times K}$ for the N sampled points across K candidate materials.

CLIP visual and text embeddings. Motivated by the sophisticated world knowledge embedded in vision–language models like LLaVA [28], we leverage the fine-tuned VLM (Sec. 3.1) to propose candidate materials and their intrinsic physical properties (*e.g.*, density, Young’s modulus, and thickness) in textual form.

As mentioned above, each sampled patch is encoded via a frozen CLIP ViT-B/16 encoder [20] to extract visual features $\{\mathbf{f}_i\}_{i=1}^N$. Simultaneously, the candidate material names generated by the VLM are projected into the same latent space using the CLIP text encoder, resulting in embeddings $\{\mathbf{t}_k\}_{k=1}^K$. While the cosine similarity between these modalities provides an initial estimation, these likelihoods are subsequently refined through a spatially-aware alignment process to ensure physical consistency. We denote the resulting raw similarity matrix as $S_{\text{CLIP}} \in \mathbb{R}^{N \times K}$, where $[S_{\text{CLIP}}]_{i,k} = \text{sim}(\mathbf{f}_i, \mathbf{t}_k)$.

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:

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

where $\mathbf{F}_{\mathcal{N}(i)} \in \mathbb{R}^{|\mathcal{N}(i)| \times D}$ is the matrix of CLIP features of all sampled points in the same SAM-derived part as i , $W_Q, W_K \in \mathbb{R}^{D \times d_a}$ 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 MLP, producing output $\mathbf{z}_i \in \mathbb{R}^D$ for each sampled point i .

We optimize the similarity matrix S using a supervised contrastive loss [24] defined over part groupings:

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

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_{CLIP} as a regularizer to prevent the module from smoothing embeddings without preserving the material discrimination for each 2D point:

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

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:

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

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

We also tried a naive approach of averaging the visual features of all points in 1 mask, which does not perform well, as shown in Tab. 1.

3.4 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:

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

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 [41] 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.

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 GPT4 to refine the mass based on heaviness of the object. We first use SiPhy’s initial predictions to classify the object as *heavy* or *light*. This classification is then used to adjust thickness prediction, biasing the prediction toward thickness ranges that are physically plausible (Fig. 4).

4 Experiments

Dataset. We evaluate on three benchmarks. **ABO-500** [11] contains 500 everyday objects with ground-truth mass; we follow the standard 300/100/100

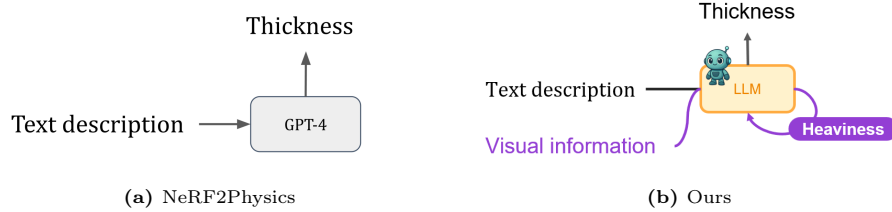


Fig. 4: Heaviness-aware thickness refinement (HAT). (a) NeRF2Physics assigns thickness based on material priors. (b) Our HAT module first classifies the object as *heavy* or *light* and refines thickness prediction toward physically plausible ranges

train/val/test split and use a single RGB view per object. **MVImgNet-100** [40] is a large-scale real-world multi-view dataset; we use 100 randomly sampled objects with a single front-facing view as input. **PhysXNet-100** [5] is a synthetic dataset with dense per-pixel physical property annotations including mass density and Young’s modulus; we randomly sample 100 scenes for evaluation.

Baselines. **NeRF2Physics** [41] uses NeRF to reconstruct the object, fuses CLIP features at each 3D point, and estimates physical properties via retrieval-based aggregation before volumetric mass integration. **PUGS** [32] adopts 3D Gaussian Splatting with shape- and region-aware contrastive losses, propagating CLIP features across Gaussians and computing mass from per-Gaussian density and volume. **GaussianProperty** [36] lifts CLIP features onto Gaussian primitives for property estimation. **LLaVA** [28] predicts mass directly from a single image using a VLM, following the evaluation protocol of NeRF2Physics.

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 [41]	M	8.74	0.78	1.06	0.55
PUGS [32]	M	30.30	1.59	7.68	0.30
LLaVA [28]	S	17.33	1.89	1.84	0.31
Image2Mass [33]	S	12.50	1.79	0.98	0.31
SiPhy (uniform 2D sampling)	S	11.94	1.27	2.61	0.39
SiPhy (average SAM masks)	S	10.23	0.77	1.00	0.54
SiPhy (Ours)	S	7.78	0.74	1.00	0.58

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

Table 2: Material segmentation performance across datasets.

Dataset	Model	mIoU ($\uparrow$)	M-mIoU ($\uparrow$)
ABO500 [11]	NeRF2Physics [41]	0.18	0.31
	PUGS [32]	0.22	0.40
	GaussianProperty [36]	0.29	0.49
	SiPhy (Ours)	0.25	0.42
MVImgNet100 [40]	NeRF2Physics [41]	-	-
	PUGS [32]	-	-
	GaussianProperty [36]	0.19	0.23
	SiPhy (Ours)	0.19	0.25
PhysXNet100 [5]	NeRF2Physics [41]	0.03	0.06
	PUGS [32]	0.04	0.07
	GaussianProperty [36]	0.14	0.21
	SiPhy (Ours)	0.14	0.31

Table 3: Material density and Young’s Modulus estimation on PhysXNet100.

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

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 $T = 0.1$. An ablation on $K \in \{3, 5, 10\}$ is provided in the Supplementary. Captions are generated using Instructional BLIP-2 [13] with Flan-T5-XL.

4.1 Quantitative Comparison to SOTA methods

Mass Evaluation. We evaluate mass prediction using the same metrics as NeRF2Physics [41]: 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 [36] 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.

Material Segmentation Evaluation. As shown in Table 2, SiPhy achieves strong material segmentation across all three datasets. On ABO-500, SiPhy improves over NeRF2Physics by **38.9%** mIoU and over PUGS by **13.6%** mIoU; GaussianProperty scores higher due to its multi-view input. On MVImgNet-100, SiPhy matches GaussianProperty in mIoU and improves M-mIoU by **8.7%**. On PhysXNet-100, SiPhy achieves the largest gains, outperforming all baselines including GaussianProperty (**+47.6%** M-mIoU).

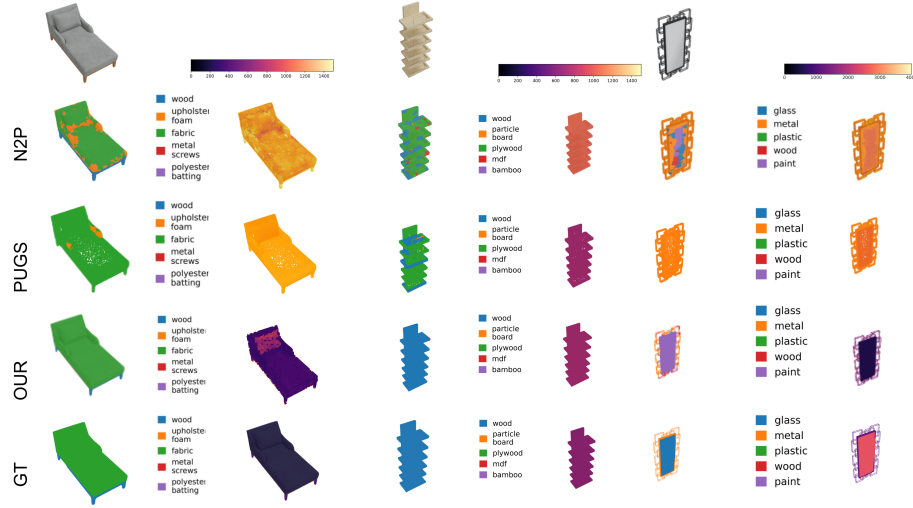


Fig. 5: Qualitative results on ABO500. Material segmentation (left) and mass density (right). SiPhy yields more coherent material segmentation than NeRF2Physics and PUGS, especially on thin structures and fine-grained parts.

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.

4.2 Ablation Study

Effectiveness of heaviness-aware thickness refinement. Table 4 compares SiPhy with and without HAT on the ABO-500 heavy/light splits. Without HAT, mass prediction is substantially worse for heavy objects, with ADE increasing from 2.07 on light objects to 18.86 on heavy objects. HAT improves heavy-object ADE from 18.86 to 15.59 (**17.3%** improvement) and MnRE from 0.57 to 0.65, leading to better overall ABO-500 performance. This suggests that heaviness-

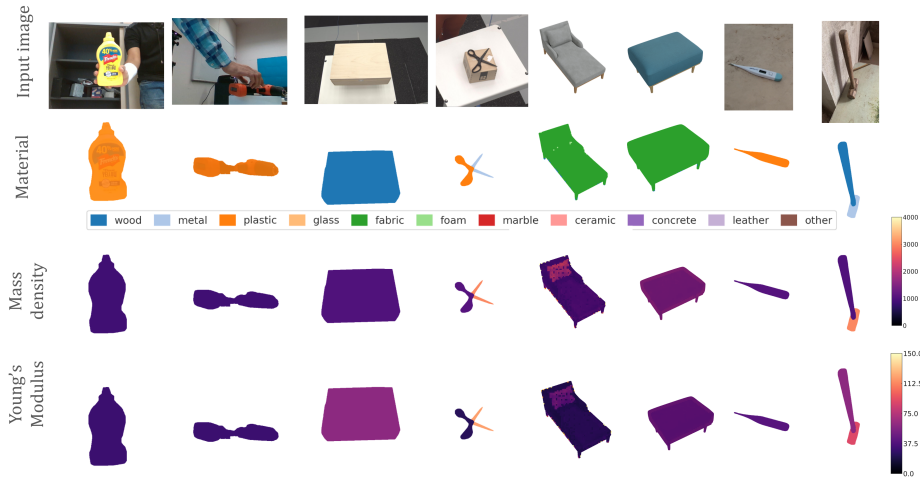


Fig. 6: Qualitative results on HO3D, ARCTIC, ABO500, and MVImgNET. From a single image, SiPhy predicts material labels, mass density, and Young’s modulus with consistent and physically plausible outputs across real and synthetic settings. All property maps use a jet colormap; mass density ranges 0–3000 kg/m³ and Young’s modulus 0–100 GPa. Objects 5 and 6 have markedly lower values and use narrower ranges (density 0–1500 kg/m³, Young’s modulus 0–20 GPa).

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	ABO500	9.84	0.81	0.85	0.53
<i>w/o</i>	ABO500-heavy	18.86	0.64	0.54	0.57
<i>HAT</i>	ABO500-light	2.07	0.73	0.57	0.58
Ours	ABO500	7.78 ^{+2.06}	0.74 ^{+0.07}	1.00 ^{−0.15}	0.58 ^{+0.05}
	ABO500-heavy	15.59 ^{+3.27}	0.48 ^{+0.16}	0.46 ^{+0.08}	0.65 ^{+0.08}
	ABO500-light	3.57 ^{−1.50}	0.85 ^{−0.12}	1.29 ^{−0.72}	0.54 ^{−0.04}

Table 5: Sensitivity analysis under independent perturbations of estimated depth and camera intrinsics on ABO-500.

Setting	ADE (↓)	ALDE (↓)	APE (↓)	MnRE (↑)
Depth × 0.9	8.08	0.77	0.87	0.56
Depth × 1.1	9.14	0.75	1.27	0.56
Intrinsics × 0.9	9.17	0.75	1.28	0.56
Intrinsics × 1.1	8.07	0.76	0.87	0.57

aware thickness refinement mainly benefits dense or bulky objects, where volume errors have a larger impact on mass estimation.

Sensitivity of the spacing formula to depth and intrinsics estimation errors is analyzed in the Supplementary materials.

Edge-case analysis. We test SiPhy on transparent, highly textured, and reflective objects (Fig. 7). SiPhy handles most transparent and textured cases, while reflective surfaces remain challenging because specular highlights can mislead both SAM segmentation and VLM material prediction.

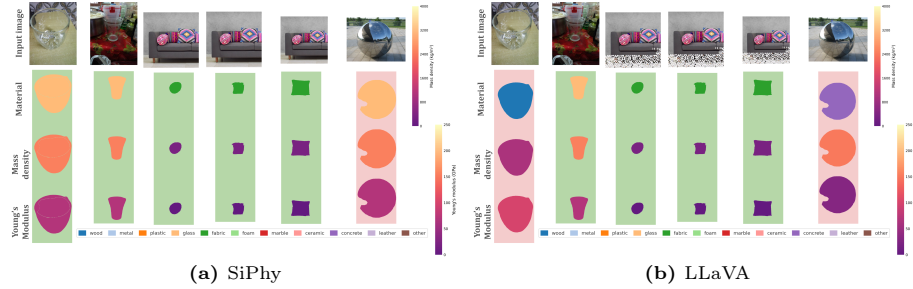


Fig. 7: Edge-case analysis. From left to right: the first 2 objects are **transparent**, the next 3 (pillows, same image) are **highly textured**, and the last is **reflective**. Predictions are bordered with a **green** box if correct and a **red** box if incorrect. SiPhy handles transparent and textured objects well; reflective surfaces remain a failure mode for both methods.

Table 6: Material segmentation performance across datasets.

Dataset	Model	ADE($\downarrow$)	ALDE($\downarrow$)	APE($\downarrow$)	MnRE($\uparrow$)
HO3D [16]	LLaVA [28]	0.24	0.48	0.58	0.58
	SiPhy (Ours)	0.19	0.50	0.41	0.67
ARCTIC [15]	LLaVA [28]	0.21	0.60	0.50	0.53
	SiPhy (Ours)	0.26	0.92	0.46	0.57

Table 7: Material density and Young’s Modulus estimation on PhysXNet100.

Dataset	Model	Property	ME ($\downarrow$)
HO3D [16]	LLaVA [28]	Mass density	1308
	SiPhy (Ours)	Mass density	1192
ARCTIC [15]	LLaVA [28]	Young’s modulus	74
	SiPhy (Ours)	Young’s modulus	62

4.3 Real-world Application

We evaluate two downstream uses of SiPhy. For hand-object interaction, SiPhy can annotate physical properties for datasets that lack such labels. For image-to-audio generation, SiPhy provides material priors for sound synthesis, where material is often more informative than appearance alone.

Hand-object Interaction. HO3D [16] and ARCTIC [15] are real-world hand-object interaction datasets capturing everyday objects featuring partial hand occlusion with varied illumination. We evaluate on 9 objects from HO3D and 10 from ARCTIC, selecting a single front-view RGB frame per object. Ground-truth masses are sourced from manufacturer product specifications. Since multi-view baselines (NeRF2Physics, PUGS) require multi-view captures unavailable in these datasets, we compare against LLaVA as the strongest applicable single-view baseline. As shown in Table 6 and Table 7, 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, confirming its effectiveness as a physical-property annotator for real-world interaction data (Fig. 6).

Image-to-Audio Generation. Object acoustics depend on material, not appearance. When visual texture misleads, appearance-driven systems produce implausible audio; SiPhy’s material prior corrects this bias. We compare ChatGPT-only, Vanilla (VLM caption $\rightarrow$ MakeAnAudio [19]), and Vanilla + SiPhy. As shown in Fig. 8, SiPhy’s material prior corrects appearance-driven errors (e.g., a metal locker misidentified as wood), producing spectrograms that better match the reference. Full results are in the Supplementary.

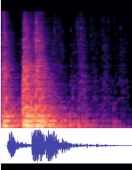
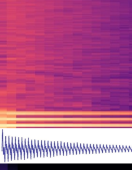
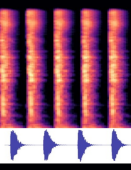
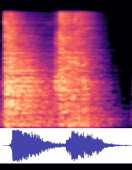
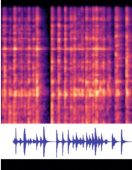
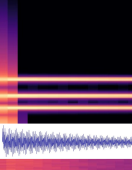
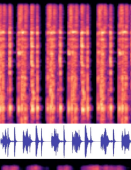
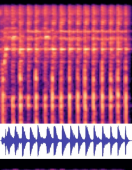
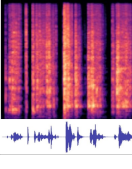
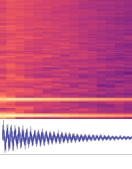
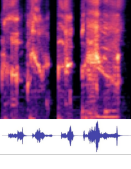
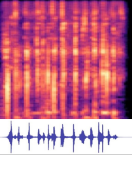
Reference Image	Physical Property	Proxy Reference Audio	ChatGPT	Vanilla (MakeAnAudio)	Vanilla + SiPhy (Ours)	Visual Description
	Metal (Locker)					Sound of slamming a white school locker door.
	Ceramic (Mug)					Sound of tapping a red speckled camping mug.
	Plastic (Ketchup)					Sound of squeezing a red ketchup bottle.

Fig. 8: Material-grounded audio generation. Mel-spectrograms for three objects under ChatGPT-only, Vanilla (VLM caption $\rightarrow$ MakeAnAudio), and Vanilla + SiPhy. SiPhy’s material prior corrects appearance-driven errors (e.g., metal locker misidentified as wood), producing spectrograms that better match the reference audio.

5 Conclusion

We presented SiPhy, a single-image framework that unifies geometric cues, semantic understanding, and language-driven physical knowledge to estimate both pixel-level and object-level 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 heaviness-aware refinement for volume estimation. SiPhy offers a scalable alternative to reconstruction-heavy pipelines, but remains constrained by the quality of single-view geometric priors and the inherent ambiguity of inferring physical attributes from one RGB image. Transparent, reflective, and highly textured objects remain challenging because depth estimation and CLIP-based material recognition can become unreliable. HAT currently uses a discrete heavy/light prior; extending it to continuous or compositional thickness estimation is left for future work.

Acknowledgment

This work was supported by Michigan State University. We are grateful to Xiaoming Liu for an insightful discussion that inspired our experiment on generating sounds from material properties.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022), <https://arxiv.org/abs/2204.14198>
2. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., Girshick, R.: Phyre: A new benchmark for physical reasoning (2019), <https://arxiv.org/abs/1908.05656>
3. Bell, S., Upchurch, P., Snaveley, N., Bala, K.: Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). p. – (2015), https://openaccess.thecvf.com/content_cvpr_2015/papers/Bell_Material_Recognition_in_2015_CVPR_paper.pdf, materials in Context Database (MINC)
4. Bordes, F., Garrido, Q., Kao, J.T., Williams, A., Rabbat, M., Dupoux, E.: Intphys 2: Benchmarking intuitive physics understanding in complex synthetic environments (2025), <https://arxiv.org/abs/2506.09849>
5. Cao, Z., Chen, Z., Pan, L., Liu, Z.: Physx-3d: Physical-grounded 3d asset generation. arXiv preprint arXiv:2507.12465 (2025)
6. Chaudhari, A., Guntuboina, C., Huang, H., Farimani, A.B.: Alloybert: Alloy property prediction with large language models (2024), <https://arxiv.org/abs/2403.19783>
7. Chen, Z., Yi, K., Li, Y., Ding, M., Torralba, A., Tenenbaum, J.B., Gan, C.: Comphy: Compositional physical reasoning of objects and events from videos (2022), <https://arxiv.org/abs/2205.01089>
8. Cherian, A., Corcodel, R., Jain, S., Romeres, D.: Llmphy: Complex physical reasoning using large language models and world models (2024), <https://arxiv.org/abs/2411.08027>
9. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 2818–2829. IEEE (Jun 2023). <https://doi.org/10.1109/cvpr52729.2023.00276>, <http://dx.doi.org/10.1109/CVPR52729.2023.00276>
10. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), <https://lmsys.org/blog/2023-03-30-vicuna/>
11. Collins, J., Goel, S., Luthra, A., Xu, L., Deng, K., Zhang, X., Vicente, T.F.Y., Arora, H., Dideriksen, T., Guillaumin, M., Malik, J.: ABO: dataset and benchmarks for real-world 3d object understanding. CoRR **abs/2110.06199** (2021), <https://arxiv.org/abs/2110.06199>
12. Dagli, R., Xiang, D., Modi, V., Loop, C., Tsang, C.F., Chen, A.H., Hu, A., State, G., Levin, D.I., Shugrina, M.: Vomp: Predicting volumetric mechanical property fields. arXiv preprint arXiv:2510.22975 (2025)
13. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)

14. Dou, Y., Oh, W., Luo, Y., Loquercio, A., Owens, A.: Hearing hands: Generating sounds from physical interactions in 3d scenes. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1795–1804 (2025)
15. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
16. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)
17. Hamrick, J., Battaglia, P., Tenenbaum, J.B.: Internal physics models guide probabilistic judgments about object dynamics. In: Proceedings of the 33rd annual conference of the cognitive science society. vol. 2. Cognitive Science Society (2011)
18. Hsu, H.Y., Lin, C.H., Zhai, A.J., Xia, H., Wang, S.: Autovfx: Physically realistic video editing from natural language instructions. In: 2025 International Conference on 3D Vision (3DV). pp. 769–780. IEEE (2025)
19. Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. arXiv preprint arXiv:2301.12661 (2023)
20. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). <https://doi.org/10.5281/zenodo.5143773>, <https://doi.org/10.5281/zenodo.5143773>, if you use this software, please cite it as below.
21. Jacobs, R., Polak, M.P., Schultz, L.E., Mahdavi, H., Honavar, V., Morgan, D.: Regression with large language models for materials and molecular property prediction (2024), <https://arxiv.org/abs/2409.06080>
22. Jiang, H., Hsu, H.Y., Zhang, K., Yu, H.N., Wang, S., Li, Y.: Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. arXiv preprint arXiv:2503.17973 (2025)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering (2023), <https://arxiv.org/abs/2308.04079>
24. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning (2021), <https://arxiv.org/abs/2004.11362>
25. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023), <https://arxiv.org/abs/2301.12597>
26. Li, S., Wu, K., Zhang, C., Zhu, Y.: I-phyre: Interactive physical reasoning (2024), <https://arxiv.org/abs/2312.03009>
27. Li, Y., Gupta, V., Kilic, M.N.T., Choudhary, K., Wines, D., keng Liao, W., Choudhary, A., Agrawal, A.: 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). <https://doi.org/https://doi.org/10.1039/d4dd00199k>, <https://www.sciencedirect.com/science/article/pii/S2635098X24002341>
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), <https://arxiv.org/abs/2304.08485>
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)

30. Rubungo, A.N., Arnold, C., Rand, B.P., Dieng, A.B.: Llm-prop: Predicting physical and electronic properties of crystalline solids from their text descriptions (2023), <https://arxiv.org/abs/2310.14029>
31. Sharan, L., Liu, C., Rosenholtz, R., Adelson, E.H.: Recognizing materials using perceptually inspired features. *International Journal of Computer Vision* **103**(3), 348–371 (July 2013). <https://doi.org/10.1007/s11263-013-0609-0>
32. Shuai, Y., Yu, R., Chen, Y., Jiang, Z., Song, X., Wang, N., Zheng, J., Ma, J., Yang, M., Wang, Z., Ding, W., Zhao, H.: Pugs: Zero-shot physical understanding with gaussian splatting (2025), <https://arxiv.org/abs/2502.12231>
33. Standley, T., Sener, O., Chen, D., Savarese, S.: image2mass: Estimating the mass of an object from its image. In: Levine, S., Vanhoucke, V., Goldberg, K. (eds.) *Proceedings of the 1st Annual Conference on Robot Learning. Proceedings of Machine Learning Research*, vol. 78, pp. 324–333. PMLR (13–15 Nov 2017), <https://proceedings.mlr.press/v78/standley17a.html>
34. Wu, J., Lu, E., Kohli, P., Freeman, B., Tenenbaum, J.: Learning to see physics via visual de-animation. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/4c56ff4ce4aaf9573aa5dff913df997a-Paper.pdf
35. Wu, J., Yildirim, I., Lim, J.J., Freeman, B., Tenenbaum, J.: Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In: Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*. vol. 28. Curran Associates, Inc. (2015), https://proceedings.neurips.cc/paper_files/paper/2015/file/d09bf41544a3365a46c9077ebb5e35c3-Paper.pdf
36. Xu, X., Ge, W., Qiu, D., Chen, Z., Yan, D., Liu, Z., Zhao, H., Zhao, H., Zhang, S., Liang, J., Chen, Y.C.: Gaussianproperty: Integrating physical properties to 3d gaussians with lms. *arXiv preprint arXiv:2412.11258* (2024)
37. Xu, Z., Wu, J., Zeng, A., Tenenbaum, J.B., Song, S.: Densephysnet: Learning dense physical object representations via multi-step dynamic interactions (2019), <https://arxiv.org/abs/1906.03853>
38. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. *arXiv:2406.09414* (2024)
39. Yu, S., Lin, K., Xiao, A., Duan, J., Soh, H.: Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794* (2024)
40. Yu, X., Xu, M., Zhang, Y., Liu, H., Ye, C., Wu, Y., Yan, Z., Zhu, C., Xiong, Z., Liang, T., et al.: Mvimngnet: A large-scale dataset of multi-view images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9150–9161 (2023)
41. Zhai, A.J., Shen, Y., Chen, E.Y., Wang, G.X., Wang, X., Wang, S., Guan, K., Wang, S.: Physical property understanding from language-embedded feature fields. In: *CVPR* (2024)
42. Zheng, Z., Yan, X., Chen, Z., Wang, J., Lim, Q.Z.E., Tenenbaum, J.B., Gan, C.: Contphy: Continuum physical concept learning and reasoning from videos (2024), <https://arxiv.org/abs/2402.06119>