

XDen-1K: A Density Field Dataset of Real-World Objects

Jingxuan Zhang^{1,2*}, Tianqi Yu^{1*}, Yatu Zhang^{1,2*}, Jinze Wu¹, Kaixin Yao^{1,2}, Jingyang Liu¹, Yuyao Zhang^{1†}, Jiayuan Gu^{1†}, and Jingyi Yu^{1†}

¹ ShanghaiTech University

{zhangjx12023,zhangyy8,gujy1,yujingyi}@shanghaitech.edu.cn

² Deemos Technology

*Equal contribution. †Corresponding authors.

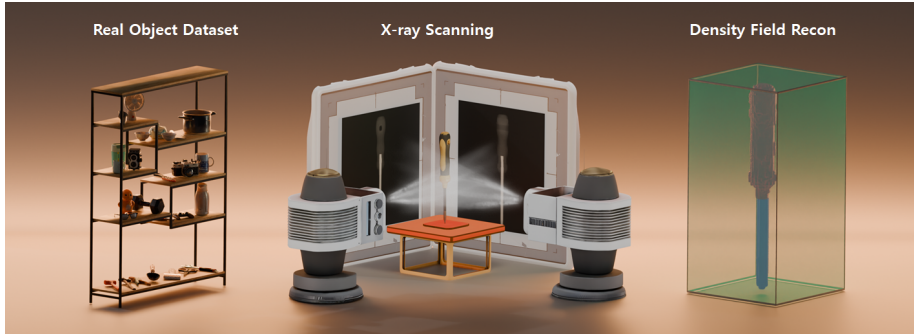


Fig. 1: We present XDen-1K, a multi-modal dataset of real-world objects, featuring paired biplanar X-ray scans and reconstructed density field.

Abstract. A deep understanding of the physical world is essential for robotic manipulation and physically realistic simulation. While current methods, including VLM-based and other learning-based approaches, have shown promise in physical property inference, their evaluation is often hindered by the lack of physically grounded reference data. To address this gap, we introduce XDen-1K, the first large-scale multi-modal dataset that provides physically grounded density field for real-world objects. XDen-1K comprises 1,000 real-world objects spanning 137 categories, with comprehensive data for each object, including a high-resolution, carefully curated 3D geometric model with part-level annotations and paired real-world biplanar X-ray scans. In addition, XDen-1K includes high-fidelity volumetric density field reconstructed from sparse biplanar X-ray views via a novel optimization framework. XDen-1K also provides a benchmark for density estimation and enables X-ray-conditioned volumetric segmentation. Experiments further demonstrate that the center-of-mass prior derived by XDen-1K can improve robotic manipulation performance. By providing real-world X-ray scans and physics-consistent density field, XDen-1K establishes a foundation for advancing physical property inference and embodied AI. Additional

details and dataset resources are available at <https://xden-1k.github.io/> and <https://huggingface.co/datasets/zhangjxx/XDen-1K>.

Keywords: X-ray · Dataset · Density Field

1 Introduction

Objects are more than their appearance: beyond color and texture, they possess internal composition — materials, thickness, voids, reinforcements, and mass distribution — that determine how they function, fail, and can be used. Modeling these physical attributes, coupled with visual inferences of shape and texture, enables more reliable physical reasoning [2, 26, 54], physically-based simulations [10, 23], and, most recently, object and scene reconstructions [47, 52]. For robots, these “hidden” properties are particularly useful for inferring where mass is concentrated [41, 49] (center of mass, moments of inertia), how stiff or compliant parts are, what forces and grasps are safe and effective, how objects will deform or fracture, which tools to use, etc.

Humans infer physical properties by combining visual cues (multi-view geometry, symmetry, material priors) and interactions (e.g. sound, deformation, context) [16, 21, 43, 44]. At surface level, recent success in material-property estimation [19, 32] (e.g. SVBRDF/BSDF, albedo/roughness, microgeometry) is largely attributed to the availability of controlled, large-scale visual datasets. Density, in contrast, is inherently volumetric and, thereby, “invisible” to regular RGB cameras. The latest efforts [1, 6, 49] have focused on labeling parts by associating their appearance with material types. Resulting datasets are therefore synthetic and lack real-world validation. A real-world density dataset will facilitate learning in physically meaningful units and expose models to realistic multi-material mixtures and manufacturing artifacts. It also provides the urgently needed common benchmarks for current and future visual inference tasks.

In this paper, we present **XDen-1K**: an X-ray-imaging-based **density field dataset** of **1,000** real-world objects. XDen-1K is the first of its type, with objects covering sizes from small hardware (e.g. screw, flashlight, key) to medium household goods (e.g. kettle, bicycle helmet, power tool at $\sim 20 - 30$ cm), and to larger items (e.g. cookware, sports equipment, furniture). The most direct brute-force approach to acquiring a ground-truth density field is computed tomography (CT), which was initially developed for medical and industrial inspection. However, high-resolution CT is time-consuming and costly—typical scans take 15–60 minutes and can cost \$100–\$500 per object, so 1,000 objects imply hundreds to thousands of scanner-hours and a six-figure budget. Even with CT, the so-called beam hardening effect will make recovering internal information more difficult.

Density field reconstruction begins with biplanar X-ray imaging of each object and the construction of its 3D mesh. To constrain this highly underdetermined problem, manually annotated object parts are used as structural priors, with each part assumed to be composed of a homogeneous material. Based

on Beer-Lambert law, a differentiable optimization framework is introduced to match rendered X-ray projections with real scans and optimize the linear attenuation coefficient (LAC) of each part, which is then converted to density. Following this pipeline, approximate density field are recovered for 1,000 objects, together with their meshes, part annotations, and biplanar X-ray scans. To further validate the reconstruction, a small-scale CT dataset is collected for evaluation.

Based on XDen-1K, our paired X-ray and density field dataset, we introduce three subtasks: benchmarking physical reasoning methods [25, 49, 57], X-ray-conditioned volumetric segmentation and CoM-aware robot manipulation. These three tasks take full advantage of our data, where X-ray provides strong structural and material prior and the density field provides physical information on both static (mass, center of mass) and dynamic (moment of inertia) properties of an object. We commit to **releasing our dataset publicly** to facilitate multi-modal learning and physical reasoning.

Our contributions:

1. We propose the **XDen-1K** dataset, the first large-scale real-world object dataset, coupled with both **biplanar X-ray images** and **density fields**, which serves as a benchmark on related physical reasoning tasks.
2. We develop a differentiable optimization framework that takes part-segmented object geometry and real biplanar X-ray scans as inputs to reconstruct each object’s volumetric density field.
3. We demonstrate the effectiveness of X-ray imaging as an additional modality on 3D volumetric segmentation task.
4. We show that center-of-mass priors derived from XDen-1K’s density information can improve robotic grasping performance.

2 Related Work

Density Inference and Physics-Aware Modeling. A broad line of work seeks to endow 3D vision with physical awareness by predicting material or density-related properties directly from visual input [7, 18, 25, 27, 40, 48, 49, 57]. Optimization-based approaches combine differentiable physics with neural scene representations to fit elastic or material parameters from observed motion, but they are typically evaluated on synthetic or tightly controlled setups, making it unclear how well they transfer to real objects. More recent feed-forward methods instead regress spatially varying material fields from learned 3D features. PIXIE [25], for example, learns a supervised mapping from distilled feature volumes to per-voxel material parameters, while GaussianProperty [49] and PuGS [40] assigns physical attributes to 3D Gaussians using segmentation and vision-language priors. Collectively, these methods demonstrate that appearance encodes rich physical cues, but in the absence of real volumetric ground truth, their predictions remain difficult to rigorously validate. Works like PhysX-3D [6] and PhysX-Anything [7] aim to build physics-ready assets for simulation and also augment large 3D model

collections with absolute scale, material categories, affordances, and kinematic attributes. Datasets like 3DCoMPaT200 [1] contain large amounts of 3D assets and their material composition.

These resources mark an important step toward learning physically grounded representations; yet, they are constructed from synthetic 3D assets and rely on assumed or inferred physical properties rather than measurements. None provide large-scale, real-world volumetric density fields against which both prediction methods and physics-aware generative models can be quantitatively evaluated. In contrast, our XDen-1K dataset is built from calibrated biplanar X-ray measurements, offering empirically grounded density fields for 1,000 real objects and furnishing the missing real-world benchmark for volumetric density estimation.

X-ray and CT imaging. X-rays are high-energy electromagnetic waves characterized by high photon energy and a low probability of interaction with matter. Unlike infrared or visible light, which are easily absorbed or reflected at the surface, X-rays can penetrate most materials and reveal internal structures without causing damage [15]. This strong penetration capability enables the observation and reconstruction of 3D material distributions from 2D projections. Computed Tomography (CT), as one of the most accurate 3D reconstruction techniques, acquires dense angular projections—collectively referred to as a sinogram—by scanning an object from multiple viewpoints. Classical reconstruction algorithms [13, 28, 36, 39] can effectively recover volumetric information from these dense sinograms. Consequently, CT reconstruction plays a crucial role in medical diagnosis, industrial defect detection, and scientific imaging [14, 42].

However, a CT scan often incurs high acquisition cost and long scanning time, while traditional algorithms suffer from severe streak and beam-hardening artifacts in the presence of high-density materials. Considerable research has therefore focused on reducing CT scan time via sparse-view CT reconstruction, which lowers the number of projection angles required for reconstruction. For example, SAX-NeRF [5] and SCOPE [46] use continuum modeling based on Neural Radiance Fields (NeRF) [35], where sparse angular projections are constrained and fused with viewpoints in an implicit space, thereby filling in missing viewpoint information in the learned dense volume representation. Nevertheless, at least dozens of projections are typically required to achieve acceptable quality; otherwise, strong sparse artifacts remain inevitable.

To further reduce acquisition time and radiation dose, biplanar X-ray imaging has emerged as a promising alternative [34]. By simultaneously projecting from two orthogonal viewpoints, this technique partially compensates for the loss of depth information inherent in single-view X-rays. Recent advances have demonstrated its potential in clinical diagnosis [3, 11], providing richer structural cues and improved diagnostic accuracy compared to single-view X-ray imaging. Several studies have explored reconstructing 3D CT-like volumes from biplanar X-rays using either generative models [20, 56] or regression-based mapping networks [17, 24]. These approaches highlight the feasibility of biplanar imaging as a low-cost, low-time, and low-dose substitute for traditional CT. However, they typically rely on large paired datasets of high-quality CT and correspond-

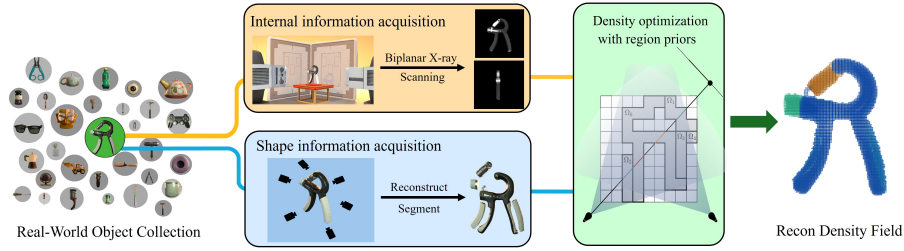


Fig. 2: Overview of XDen-1K generation pipeline. Multi-view RGB images are used to reconstruct a surface mesh, and manually labeled segments provide accurate part-level shape information. Orthogonal biplanar X-rays capture internal attenuation signals. The recovered shape priors are then incorporated into an X-ray optimization process to estimate the volumetric density field.

ing X-rays for supervision. For non-medical objects where such paired data are unavailable, how to effectively incorporate prior knowledge and achieve faithful 3D reconstruction remains a challenging open problem.

Part-Level Understanding. Decomposing fused 3D reconstructions into semantic parts is still far from solved, largely because reliable part boundaries require reasoning about global 3D geometry rather than local appearance cues alone. A common strategy is therefore to borrow strength from the well-established 2D segmentation methods and project or fuse these predictions back onto the surface [33, 61], but multi-view fusion often introduces cross-view inconsistencies and uneven granularity. Alternatively, 3D-native approaches learn part-sensitive representations directly. Such methods tend to yield more stable and coherent segmentations on surfaces [29, 31, 50], and some go a step further by explicitly generating part decompositions [51, 59]. Moreover, part segmentation is not restricted to mesh surfaces: it has been actively explored in point cloud representation [38, 60, 62], 3D Gaussian representations [8, 53, 55], and radiance-field-based models [9, 22]. Most current method takes RGB cue as the only guidance, which can be ambiguous under texture variation. In this work, we introduce X-ray as an extra condition to guide the segmentation process, yielding more consistent result. Furthermore, the X-ray dataset we provide serves as a valuable benchmark for future multimodal-guided segmentation tasks.

3 Density field reconstruction from biplanar X-ray images

3.1 Preliminary: X-ray Imaging

X-rays possess the ability to penetrate matter. When an X-ray beam passes through an object, its intensity attenuates exponentially according to the Beer-Lambert law. For a ray $\mathbf{r}$ traversing a volumetric object, the detected intensity

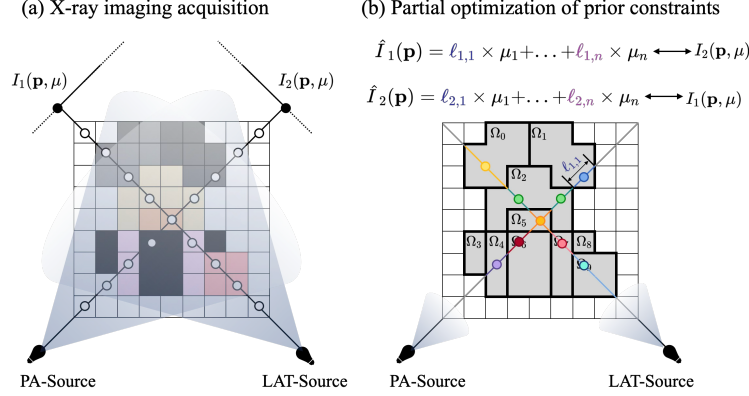


Fig. 3: Differentiable rendering optimization. Region-wise attenuation coefficients are optimized to match captured projections, producing a 3D attenuation field.

I is related to the incident intensity I_0 as

$$I = I_0 e^{-\int_L \mu(x) dx}, \quad (1)$$

where $\mu(x)$ denotes the linear attenuation coefficient (LAC) at spatial location x , which characterizes the probability per unit length that an X-ray photon is absorbed or scattered by the material. Taking the logarithm of both sides yields the projection measurement p :

$$p = -\ln\left(\frac{I}{I_0}\right) = \int_L \mu(x) dx, \quad (2)$$

which corresponds to the line integral of the LAC along the ray trajectory. In practice, the detector measures the transmitted intensity I , and the reference airscan intensity I_0 is pre-calibrated. The logarithmic transformation converts the exponential attenuation into a linear integral form, establishing the foundation of tomographic reconstruction. The LAC value $\mu(x)$ serves as a quantitative descriptor of material composition and density. Physically, the LAC depends on both the density ρ , and the mass attenuation coefficient (MAC) $\frac{\mu}{\rho}(E)$, which is a material-dependent quantity determined by atomic composition and photon energy E :

$$\mu(E) = \rho \cdot \frac{\mu}{\rho}(E). \quad (3)$$

According to reference database XCOM [4] from the National Institute of Standards and Technology (NIST), the MACs of most common materials (e.g., water, polymers, biological tissues) converge to approximately $0.17 \text{ cm}^2/\text{g}$ at a certain energy level (e.g., 100 keV), as detailed in Tab. 1. Consequently, for such materials, μ exhibits an approximately linear relationship with density, enabling direct density estimation from reconstructed LAC values:

$$\rho = \frac{\mu}{(\mu/\rho)} = \frac{\mu}{0.17}. \quad (4)$$

Table 1: Mass attenuation coefficient (MAC), density, and linear attenuation coefficient (LAC) of some representative materials at 100 keV ray energy.

Material	Chemical Formula	MAC (cm ² /g)	Density (g/cm ³)	LAC (cm ⁻¹)
Water	H ₂ O	0.17	1.00	0.17
Air	N ₂ , O ₂	0.15	0.0012	1.8×10^{-4}
Glass	SiO ₂	0.18	2.40	0.432
Plastic (PP)	(C ₃ H ₆) _n	0.18	0.90	0.16
Rubber	(C ₅ H ₈) _n	0.19	1.20	0.22
Wood	C ₆ H ₁₀ O ₅	0.16	0.80	0.12
Aluminum	Al	0.19	2.70	0.51

This principle underpins quantitative X-ray imaging, allowing reconstructed attenuation maps to be interpreted as physically meaningful density fields.

3.2 Reconstruction Algorithms

The **reconstruction objective** is to derive density field over the target object. Inverting Eq. (2) over all rays can recover the 3D spatial distribution of $\mu(x)$, the LAC field and by Eq. (3), the reconstructed LAC field can be converted to density field. While classical algorithms such as Filtered Backprojection (FBP) in CT reconstruction provide accurate 3D reconstructions, its high cost and long acquisition times make it prohibitive for large-scale datasets. In contrast, single-plane X-ray imaging is faster and more economical, yet it completely loses depth information by integrating attenuation coefficients along the projection direction. Consequently, recovering a 3D density field from a single X-ray projection is a severely ill-posed inverse problem. Therefore, we adopt a pragmatic approach that utilizes **two orthogonal, biplanar X-ray projections**, regularized by a strong geometric prior. This strategy leverages the complementary depth cues from the two projections and constrains the solution space through known geometry, thereby making the reconstruction problem tractable.

The process of obtaining the estimated density field is shown in Fig. 3. Given a captured ground-truth biplanar X-ray images I_i and a well-structured segmentation that partitions the image into K disjoint regions $\{\Omega_k\}_{k=1}^K$, each region corresponds to a homogeneous material with a constant linear attenuation coefficient (LAC) μ_k . The volumetric attenuation field $\mu(\mathbf{x})$ is thus modeled as a piecewise-constant function:

$$\mu(\mathbf{x}) = \sum_{k=1}^K \mu_k \mathbf{1}_{\Omega_k}(\mathbf{x}), \quad (5)$$

where $\mathbf{1}_{\Omega_k}(\mathbf{x})$ is the indicator function:

$$\mathbf{1}_{\Omega_k}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \Omega_k, \\ 0, & \text{otherwise.} \end{cases}$$

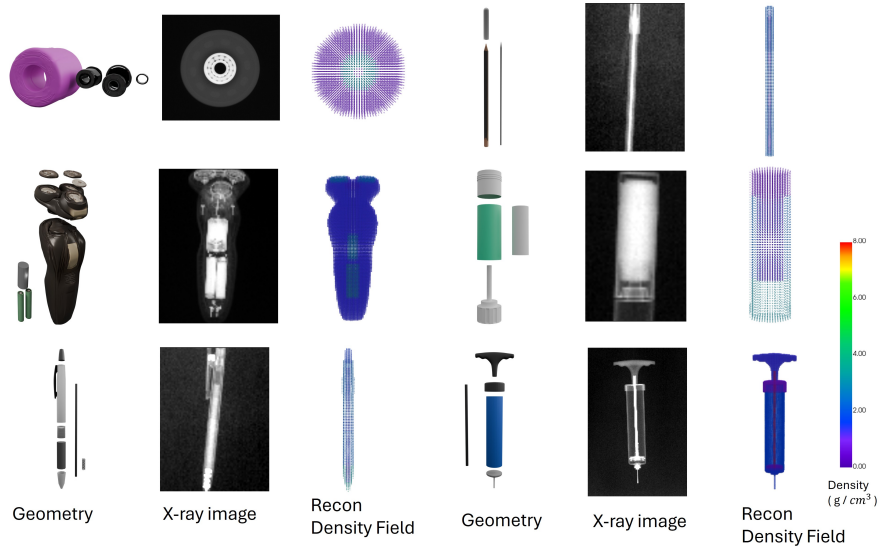


Fig. 5: Gallery of XDen-1K. XDen-1K dataset contains curated geometry, captured biplanar X-ray images, and reconstructed density fields. Some objects are represented by their original meshes, while others include manually modeled internal geometries.

4 XDen-1K dataset

The reconstruction algorithm introduced in Sec. 3.2 enables the recovery of a physically consistent volumetric density field. Building on this capability, a large-scale dataset, XDen-1K, was collected, comprising 1,000 real-world objects. It provides both real captured X-ray observations and corresponding reconstructed 3D density fields.

4.1 Dataset Overview and Preprocessing

To capture the diversity and complexity of objects encountered in everyday environments, the objects covered in XDen-1K span **137** categories, ranging from tools and kitchenware to consumer electronics. As shown in Fig. 4, these categories are relatively evenly distributed across common object families. Frequently appearing objects, such as spoons, screwdrivers, scissors and pliers, are all commonly encountered in daily life. This broad category coverage introduces rich variations in **geometry, size, materials, and internal structure**, enabling the dataset to better reflect real-world object variability.

Each object in the dataset is captured and annotated through a standardized pipeline, as shown in Fig. 2. We first collect calibrated multi-view RGB images to reconstruct a watertight mesh [58] and measure its real-world scale. A human-in-the-loop segmentation [59] is then performed to decompose objects into functionally meaningful components. To perform the density field reconstruction algorithm described in Sec. 3.2, ground-truth X-ray images are re-

Table 2: Mean (μ) and standard deviation (σ) of reprojection errors on real X-rays.

Segmentation prior	LAT- μ	LAT- σ	PA- μ	PA- σ
X-Field prediction	1.121	0.287	0.724	0.268
Current pipeline (manual labels)	0.778	0.191	0.498	0.207
Background noise	0.220	N/A	0.200	N/A

quired as supervision. Therefore, for each object in XDen-1K, we also capture paired biplanar X-ray scans using our X-ray imaging machine. For some objects containing extremely complex interior geometry, we manually model out its interior structure based on the captured X-ray image which clearly reveals its inner structure.

During actual biplanar X-ray capture stage, we use the Eagle Eye biplanar X-ray imaging system developed by TaoImage, with settings of a 100 kV tube voltage, a 450 mA tube current, an exposure time of 280 ms per exposure, and a total exposure of 126 mAs.

Following this standardized acquisition pipeline, XDen-1K provides a comprehensive multi-modal representation for each object. Specifically, for each object, we provide a **part-segmented 3D mesh**, **biplanar X-ray scans**, and a **reconstructed volumetric density field**, as shown in Fig. 5. For validation purposes, a subset of objects is additionally scanned using high-resolution CT, as described in Sec. 4.2.

4.2 Evaluation of XDen-1K

We evaluate the fidelity of the reconstructed density fields from two complementary perspectives: consistency with real X-ray observations and agreement with high-resolution CT references. First, to assess whether the optimized LAC field can faithfully explain the measured X-ray data, we reproject the reconstructed field using Eq. (2) and compare the resulting synthetic X-ray images with the real biplanar acquisitions. We report the Mean Absolute Error (MAE) and Pearson Correlation (Corr) in the image domain. As shown in Tab. 2, our method achieves accurate and stable reprojection results when reliable geometric and structural priors are available, demonstrating the physical consistency of the reconstructed density fields.

For the CT-based evaluation, we note that the reference volumes are not perfect ground truth. Due to beam-hardening artifacts and the relatively low tube voltage, metallic components such as the air pump may suffer from streaking effects and deviate from the assumed LAC-to-density model in Eq. (4). We therefore use the CT scans only as partial volumetric references. As shown in Tab. 3, our reconstructions achieve reasonable agreement with these CT references, suggesting that the optimized density fields capture physically meaningful volumetric structure while remaining primarily constrained by the real X-ray observations.

Manual interior-structure modeling can improve reconstruction quality but introduces additional effort that limits scalability. We therefore evaluate a more

Table 3: Quantitative evaluation of reconstruction quality on XDen-1K. Both density error against CT (MAE) and reprojection error against real X-ray images (LAT/PA) are measured.

	CT	LAT		PA		Fine LAT		Fine PA	
	MAE	MAE	Corr	MAE	Corr	MAE	Corr	MAE	Corr
Glue	0.181	0.298	0.678	0.239	0.616	0.341	0.494	0.381	0.454
Pencil	0.294	0.271	0.492	0.281	0.620	0.211	0.681	0.218	0.546
AirPump	-	1.064	0.185	0.820	0.235	0.741	0.598	0.797	0.323
BallPen	0.710	0.286	0.065	0.258	0.410	0.239	0.382	0.330	0.228
Razor	0.326	1.443	0.489	1.380	0.515	1.218	0.666	1.211	0.668

scalable variant that uses only external geometry and part segmentation as priors, while keeping the optimization procedure unchanged. As shown in Tab. 3, removing manually curated interior priors causes an expected accuracy drop, but the method still maintains reasonable reprojection consistency and volumetric fidelity, demonstrating robustness under weaker priors.

5 Applications on XDen-1K

Building on XDen-1K, we demonstrate three downstream applications: benchmarking existing physics-reasoning methods, **X-ray-conditioned volumetric segmentation**, and **center-of-mass-aware robot manipulation**. We argue that X-ray imaging is a key modality for spatial perception, as it provides rich cues about object geometry, internal structure, and material distribution. When paired with volumetric density field, XDen-1K enables physically grounded object reasoning and supports a broad range of downstream tasks.

5.1 Benchmarking Physical Reasoning Methods

Reasoning about physical properties is fundamental to building physically grounded digital worlds and enabling stable robot manipulation. Among these properties, density is particularly important, as it can be directly converted into mass-related quantities such as total mass, center of mass, and moments of inertia. Given accurate geometry and curated object structures, XDen-1K provides **high-fidelity, physically calibrated** volumetric density field reconstructed by our pipeline, with quantitative validation presented in Sec. 4.2.

This makes XDen-1K a suitable benchmark for supervising and evaluating density estimation methods. We evaluate recent methods, including Pixie [25], GaussianProperty [49], and NeRF2Physics [57], using voxel-wise density error, L1 center-of-mass distance, and mean absolute percentage error (MAPE) of total mass. Results are summarized in Tab. 4. LLM-based methods, NeRF2Physics achieves the best performance in density MAE and center-of-mass distance, while learning-based methods, Pixie obtains the lowest mass MAPE.

Table 4: Comparison of physical reasoning methods in terms of density, center of mass (CoM), and total mass.

Method	Density MAE	CoM distance (cm)	Mass MAPE
NeRF2Physics	1.157	0.523	1.174
GaussianProperty	1.295	0.632	1.247
Pixie	1.228	0.547	0.713

5.2 X-ray-conditioned volumetric segmentation

X-ray images provide strong cues about material composition and internal structure, making them effective conditioning signals for volumetric segmentation. We therefore extend PartField [29] to the volumetric domain and introduce **X-Field**, which conditions on biplanar X-ray images to segment an object’s 3D volume. X-Field is trained on a synthetic dataset built upon PartNeXt [45], with material properties synthesized to match our real-world collection. Experiments on both synthetic and real data show that X-Field consistently outperforms PartField, validating the benefit of X-ray conditioning for volumetric inference.

Conditioning X-ray modality on PartField PartField [29] takes an input point cloud (P), extracts per-point features with PVCNN [30], and lifts them into a triplane representation. A transformer decoder predicts triplane features (L_{tri}), which define a queryable feature field ($F_{\text{tri}}(\cdot)$) in 3D space and can be projected back onto the input points to obtain point-wise features.

We augment this backbone with X-ray conditioning. Given biplanar X-ray images (I_X), an image encoder extracts X-ray features ($L_X = E_X(I_X)$). Since X-ray appearance is strongly scale-dependent, we additionally provide the physical scale (λ) and use FiLM [37] to modulate the encoded X-ray features into a scale-aware latent representation:

$$L_X^\lambda = \text{FiLM}(L_X, \lambda). \quad (8)$$

The resulting scale-aware X-ray features are fused with the triplane features through cross-attention:

$$\tilde{L}_{\text{tri}} = \text{CrossAttention}(L_{\text{tri}}, L_X^\lambda). \quad (9)$$

This produces an X-ray-conditioned volumetric feature field ($\tilde{F}_{\text{tri}}(\cdot)$).

Our goal is to predict a dense volumetric density field rather than only point-wise features on the input surface. We therefore randomly sample query points (P_{query}) within the object volume and query the fused triplane representation at these locations. For each query point ($p \in P_{\text{query}}$), a density MLP head regresses its density value:

$$\hat{D}(p) = f_\theta(\tilde{F}_{\text{tri}}(p)). \quad (10)$$

Training is supervised with an (ℓ_2) loss between the predicted densities and the ground-truth densities from the synthetic volumetric fields:

$$\mathcal{L}_{\text{den}} = \frac{1}{|P_{\text{query}}|} \sum_{p \in P_{\text{query}}} \left\| \hat{D}(p) - D(p) \right\|_2^2. \quad (11)$$

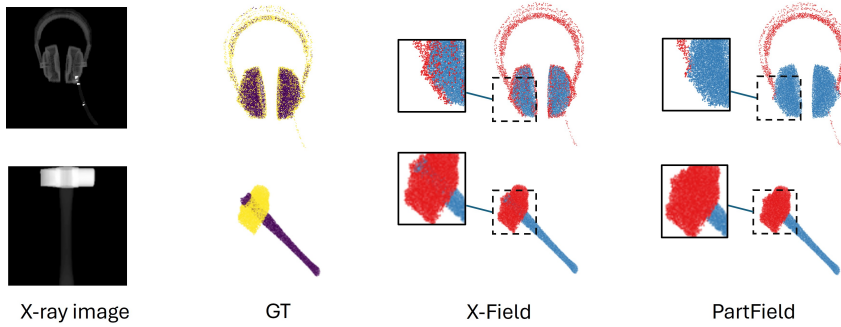


Fig. 6: Comparison of qualitative segmentation results.

Method	Synthetic Data IoU $\uparrow$	Real Data IoU $\uparrow$
PartField	53.185	38.410
X-Field (Ours)	72.645	49.251

Table 5: Comparison of IoU scores on synthetic and real data.

Biplanar X-ray inputs may suffer from missing-view ambiguity due to limited viewing directions. To mitigate this issue, we introduce an auxiliary reconstruction objective: an additional X-ray image is decoded from features associated with the third plane and supervised using an MSE loss against the corresponding ground-truth X-ray image. This auxiliary task regularizes 3D feature learning and encourages more complete volumetric reasoning. Finally, by applying (k)-means clustering to the predicted densities at the query points, the volume is partitioned into regions with similar density values, producing a material-aware volumetric segmentation.

Evaluation on X-ray-conditioned segmentation The volumetric segmentation performance of the original PartField and the proposed X-Field is evaluated on both synthetic and real data. As shown in Fig. 6 and Tab. 5, conditioning on biplanar X-ray images leads to both higher IoU scores and better volumetric segmentations.

Toward Fully Automatic Dataset Construction. Although X-Field shows encouraging real-world performance, Tab. 2 shows that its predicted priors still produce higher reprojection errors than manual labels. We therefore use manually curated labels as ground truth in the current XDen-1K pipeline and present X-Field as a promising downstream application. With more high-quality synthetic labels and X-ray-consistent material properties, X-Field could further improve sim-to-real transfer and accelerate future dataset acquisition.

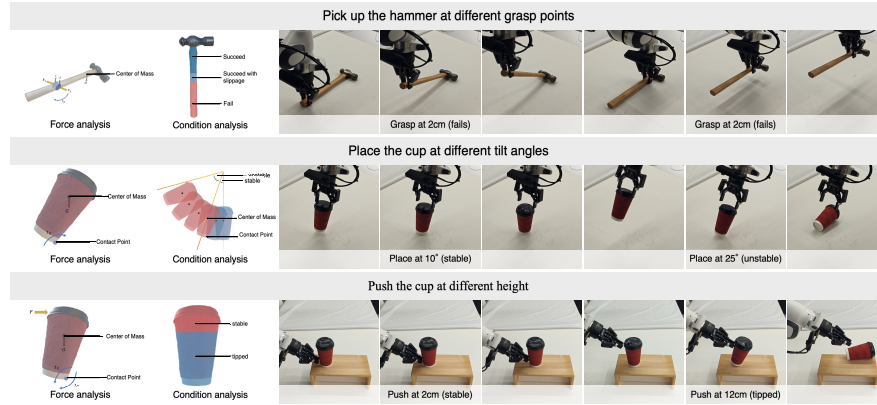


Fig. 7: Effect of the object’s center of mass on manipulation stability under Pick, Place and Push task. Left: force and stability analysis. Right: real-robot executions.

5.3 CoM-aware Robot Manipulation

The perception of an object’s center of mass is fundamental for achieving physically stable and dynamically consistent manipulation. Beyond geometric reasoning, understanding the relationship between the center of mass, external forces, and contact points allows a robot to anticipate rotation, slippage, and tipping during interaction. To study how the center of mass influences manipulation outcomes, we conduct three representative experiments on a Franka Emika Panda robotic arm equipped with a parallel gripper, as shown in Fig. 7.

Pick. Grasping near the CoM produces a small gravitational moment, enabling stable lifting. Increasing the grasp offset enlarges the lever arm and torque, which can overwhelm contact friction, induce rotation, and cause grasp failure.

Place. Stability is governed by whether the CoM projection remains inside the support region. Small tilts create a restoring moment and the object self-rights; larger tilts move the projection beyond the support boundary and result in tipping.

Push. The pushing force creates a moment about the CoM. Forces applied below the CoM favor translation, whereas forces applied above the CoM generate a larger overturning moment and can tip the object.

Beyond qualitative demonstrations, the benefit of CoM reasoning is further evaluated in a downstream grasping pipeline. Density-field prior is used to guide grasp selection in AnyGrasp [12] by selecting the candidate grasp pose whose gripper endpoints are closest to the predicted CoM. Grasp stability is evaluated using grasp success rate and joint torque loads, with particular attention to wrist torque (Joint 7), which reflects the end-effector load during grasp execution. Results are compared across three settings: vanilla AnyGrasp, AnyGrasp guided by the NeRF2Physics CoM, and AnyGrasp guided by the CoM from our reconstructed result. As shown in Tab. 6, accurate CoM guidance from the proposed method substantially improves grasp success rate while reducing both wrist and mean joint torques, indicating more stable and mechanically efficient grasps

Table 6: Grasp success (%) and joint torque ($\text{N} \cdot \text{m}$) under CoM guidance.

	Success Rate $\uparrow$	Joint-7 torque $\downarrow$	Mean torque $\downarrow$
AnyGrasp	27.50	0.521	9.645
+NeRF2Physics CoM	55.00	0.443	9.529
+Ours CoM	77.75	0.435	9.497

These experiments indicate that reasoning about the object’s center of mass plays an important role in improving related robotic manipulation task. Leveraging this physical understanding from our dataset can contribute to more stable behaviors and more reliable interaction with diverse objects in real-world settings.

6 Conclusion

We presented **XDen-1K**, a large-scale real-world dataset comprising paired bi-planar X-ray scans and reconstructed 3D density fields. A differentiable X-ray optimization framework is developed to leverage shape and segmentation priors for recovering piecewise-constant densities from two X-ray views, thereby transforming an otherwise ill-posed problem into a tractable reconstruction task. Evaluations against CT references and measured X-ray images show that the proposed method can recover high-fidelity density fields across a wide range of objects. Beyond the dataset and reconstruction framework, XDen-1K introduces three subtasks: benchmarking density estimation methods, X-ray-conditioned volumetric segmentation, and center-of-mass-aware robotic manipulation. The results establish XDen-1K as a physically grounded benchmark and reference resource for physical property estimation, while highlighting the potential of X-ray imaging as a complementary modality for 3D segmentation and object understanding. With more robust segmentation models and standardized X-ray acquisition protocols, the data acquisition pipeline can be further automated and scaled, enabling broader applications in physical reasoning and embodied AI.

7 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant (W2431046), Central Guided Local Science and Technology Foundation of China (YDZX20253100001001), Shanghai Pujiang Program (24PJA080) and by MoE Key Lab of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), the Shanghai Frontiers Science Center of Human-centered Artificial Intelligence. The experiments of this work were supported by the SIST computing Platform and HPC, ShanghaiTech University. We gratefully acknowledge the invaluable discussion and feedback provided by Peijun Xu, Qixuan Zhang and Longwen Zhang from ShanghaiTech University.

References

1. Ahmed, M., Li, X., Prajapati, A., Elhoseiny, M.: 3dcompat200: Language-grounded compositional understanding of parts and materials of 3d shapes (2024)
2. Bakhtin, A., van der Maaten, L., Johnson, J., Gustafson, L., Girshick, R.: Phyre: A new benchmark for physical reasoning. In: NEURIPS (2019)
3. Berg, B.I., Laville, A., Courvoisier, D.S., Rouch, P., Schouman, T.: Experiences with a new biplanar low-dose x-ray device for imaging the facial skeleton: A feasibility study. *PloS one* (2020)
4. Berger, M.J., Hubbell, J.H.: Xcom: Photon cross sections on a personal computer. Tech. rep., National Bureau of Standards, Washington, DC (USA). Center for Radiation ... (1987)
5. Cai, Y., Wang, J., Yuille, A., Zhou, Z., Wang, A.: Structure-aware sparse-view x-ray 3d reconstruction. In: CVPR (2024)
6. Cao, Z., Chen, Z., Pan, L., Liu, Z.: Physx-3d: Physical-grounded 3d asset generation. In: NEURIPS (2025)
7. Cao, Z., Hong, F., Chen, Z., Pan, L., Liu, Z.: Physx-anything: Simulation-ready physical 3d assets from single image. In: CVPR (2026)
8. Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment any 3d gaussians. In: AAAI (2025)
9. Cen, J., Zhou, Z., Fang, J., Yang, C., Shen, W., Xie, L., Jiang, D., Zhang, X., Tian, Q.: Segment anything in 3d with nerfs. In: NEURIPS (2023)
10. Chang, Y., Liu, M., Wang, Z., Chen, P.Y., Grinspun, E.: Lifting the winding number: Precise discontinuities in neural fields for physics simulation. In: SIGGRAPH Conference Papers (2025)
11. Chen, Y., Gao, Y., Fu, X., Chen, Y., Wu, J., Guo, C., Li, X.: Automatic 3d reconstruction of vertebrae from orthogonal bi-planar radiographs. *Scientific Reports* (2024)
12. Fang, H.S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., Liu, W., Xie, Y., Lu, C.: Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics* (2023)
13. Feldkamp, L.A., Davis, L.C., Kress, J.W.: Practical cone-beam algorithm. *Journal of the Optical Society of America A* (1984)
14. Ginat, D.T., Gupta, R.: Advances in computed tomography imaging technology. *Annual review of biomedical engineering* (2014)
15. Glover, J.L., Hudson, L.T.: An objectively-analyzed method for measuring the useful penetration of x-ray imaging systems. *Measurement Science and Technology* (2016)
16. Guo, M., Wang, B., Ma, P., Zhang, T., Owens, C.E., Gan, C., Tenenbaum, J.B., He, K., Matusik, W.: Physically compatible 3d object modeling from a single image. In: NEURIPS (2024)
17. Huang, B., Pei, Y.: Generalizable structure-aware inf: Biplanar-view ct reconstruction via disentangled implicit neural field. In: ACCV (2024)
18. Huang, T., Zeng, Y., Li, H., Zuo, W., Lau, R.W.: Dreamphysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors. In: AAAI (2024)
19. Huang, X., Wang, T., Liu, Z., Wang, Q.: Material anything: Generating materials for any 3d object via diffusion. In: CVPR (2024)
20. Jeong, Y.S., Yoo, H.B., Chun, I.Y.: Dx2ct: Diffusion model for 3d ct reconstruction from bi or mono-planar 2d x-ray (s). In: ICASSP (2025)

21. 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. In: ICCV (2025)
22. Kim, C.M., Wu, M., Kerr, J., Tancik, M., Goldberg, K., Kanazawa, A.: Garfield: Group anything with radiance fields. In: CVPR (2024)
23. Kumar, K., Salmond, J., Kularathna, S., Wilkes, C., Tjung, E., Biscontin, G., Soga, K.: Scalable and modular material point method for large-scale simulations. arXiv preprint arXiv:1909.13380 (2019)
24. Kyung, D., Jo, K., Choo, J., Lee, J., Choi, E.: Perspective projection-based 3d ct reconstruction from biplanar x-rays. In: ICASSP (2023)
25. Le, L., Lucas, R., Wang, C., Chen, C., Jayaraman, D., Eaton, E., Liu, L.: Pixie: Fast and generalizable supervised learning of 3d physics from pixels. ArXiv **abs/2508.17437** (2025)
26. Li, S., Wu, K., Zhang, C., Zhu, Y.: I-phyre: Interactive physical reasoning. In: ICLR (2024)
27. Li, X., Qiao, Y.L., Chen, P.Y., Jatavallabhula, K.M., Lin, M., Jiang, C., Gan, C.: Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. arXiv preprint arXiv:2303.05512 (2023)
28. Lin, Y., Samei, E.: An efficient polyenergetic sart (psart) reconstruction algorithm for quantitative myocardial ct perfusion. Medical physics (2014)
29. Liu, M., Uy, M.A., Xiang, D., Su, H., Fidler, S., Sharp, N., Gao, J.: Partfield: Learning 3d feature fields for part segmentation and beyond. In: ICCV (2025)
30. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. In: NEURIPS (2019)
31. Ma, C., Li, Y., Yan, X., Xu, J., Yang, Y., Wang, C., Zhao, Z., Guo, Y., Chen, Z., Guo, C.: P3-sam: Native 3d part segmentation. arXiv preprint arXiv:2509.06784 (2025)
32. Ma, X., Xu, X., Zhang, L., Zhou, K., Wu, H.: Opensvbrdf: A database of measured spatially-varying reflectance. ACM Trans. on Graphics (2023)
33. Ma, Z., Yue, Y., Gkioxari, G.: Find any part in 3d. In: ICCV (2025)
34. Melhem, E., Assi, A., El Rachkidi, R., Ghanem, I.: Eos® biplanar x-ray imaging: concept, developments, benefits, and limitations. Journal of children’s orthopaedics (2016)
35. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM (2021)
36. Nuyts, J., De Man, B., Fessler, J.A., Zbijewski, W., Beekman, F.J.: Modelling the physics in the iterative reconstruction for transmission computed tomography. Physics in Medicine & Biology (2013)
37. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: AAAI (2018)
38. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems (2017)
39. Rodet, T., Noo, F., Defrise, M.: The cone-beam algorithm of feldkamp, davis, and kress preserves oblique line integrals. Medical physics (2004)
40. Shuai, Y., Yu, R., Chen, Y., Jiang, Z., Song, X., Wang, N., Zheng, J., Ma, J., Yang, M., Wang, Z., et al.: Pugs: Zero-shot physical understanding with gaussian splatting. In: ICRA (2025)

41. Slota, G.P., Suh, M.S., Latash, M.L., Zatsiorsky, V.M.: Stability control of grasping objects with different locations of center of mass and rotational inertia. *Journal of motor behavior* (2012)
42. Sun, W., Symes, D.R., Brenner, C.M., Böhnelt, M., Brown, S., Mavrogordato, M.N., Sinclair, I., Salamon, M.: Review of high energy x-ray computed tomography for non-destructive dimensional metrology of large metallic advanced manufactured components. *Reports on Progress in Physics* (2022)
43. Tung, H.Y., Ding, M., Chen, Z., Bear, D., Gan, C., Tenenbaum, J.B., Yamins, D.L., Fan, J.E., Smith, K.A.: Physion++: Evaluating physical scene understanding that requires online inference of different physical properties. In: *NEURIPS* (2023)
44. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vgggt: Visual geometry grounded transformer. In: *CVPR* (2025)
45. Wang, P., He, Y., Lv, X., Zhou, Y., Xu, L., Yu, J., Gu, J.: Partnext: A next-generation dataset for fine-grained and hierarchical 3d part understanding. In: *NEURIPS* (2025)
46. Wu, Q., Feng, R., Wei, H., Yu, J., Zhang, Y.: Self-supervised coordinate projection network for sparse-view computed tomography. *IEEE Transactions on Computational Imaging* (2023)
47. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: *CVPR* (2024)
48. Xie, Z., Liu, Z., Peng, Z., Wu, W., Zhou, B.: Vid2sim: Realistic and interactive simulation from video for urban navigation pp. 1581–1591 (2025)
49. Xu, X., Ge, W., Qiu, D., Chen, Z., Yan, D., Liu, Z., Zhao, H., Zhao, H., Zhang, S., Liang, J., et al.: Gaussianproperty: Integrating physical properties to 3d gaussians with lmms. In: *ICCV* (2025)
50. Yang, Y., Huang, Y., Guo, Y.C., Lu, L., Wu, X., Lam, E.Y., Cao, Y.P., Liu, X.: Sampart3d: Segment any part in 3d objects. *arXiv preprint arXiv:2411.07184* (2024)
51. Yang, Y., Zhou, Y., Guo, Y.C., Zou, Z.X., Huang, Y., Liu, Y.T., Xu, H., Liang, D., Cao, Y.P., Liu, X.: Omnipart: Part-aware 3d generation with semantic decoupling and structural cohesion. In: *SIGGRAPH ASIA* (2025)
52. Yao, K., Zhang, L., Yan, X., Zeng, Y., Zhang, Q., Xu, L., Yang, W., Gu, J., Yu, J.: Cast: Component-aligned 3d scene reconstruction from an rgb image. *ACM Trans. on Graphics* (2025)
53. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: *ECCV* (2024)
54. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: collision events for video representation and reasoning. In: *ICLR* (2020)
55. Ying, H., Yin, Y., Zhang, J., Wang, F., Yu, T., Huang, R., Fang, L.: Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In: *CVPR* (2024)
56. Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y.: X2ct-gan: reconstructing ct from biplanar x-rays with generative adversarial networks. In: *CVPR* (2019)
57. 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)
58. Zhang, L., Wang, Z., Zhang, Q., Qiu, Q., Pang, A., Jiang, H., Yang, W., Xu, L., Yu, J.: Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Trans. on Graphics* (2024)

59. Zhang, L., Zhang, Q., Jiang, H., Bai, Y., Yang, W., Xu, L., Yu, J.: Bang: Dividing 3d assets via generative exploded dynamics. *ACM Trans. on Graphics* (2025)
60. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: *ICCV* (2021)
61. Zhong, Z., Xu, Y., Li, J., Xu, J., Li, Z., Yu, C., Gao, S.: Meshsegmenter: Zero-shot mesh semantic segmentation via texture synthesis. In: *ECCV* (2024)
62. Zhou, Y., Gu, J., Chiang, T.Y., Xiang, F., Su, H.: Point-SAM: Promptable 3d segmentation model for point clouds. In: *ICLR* (2025)