VISION MEETS WIFI
PHYSICS-GROUNDED ESTIMATION OF VOLUMETRIC MECHANICAL PROPERTIES

Anonymous authors
Paper under double-blind review

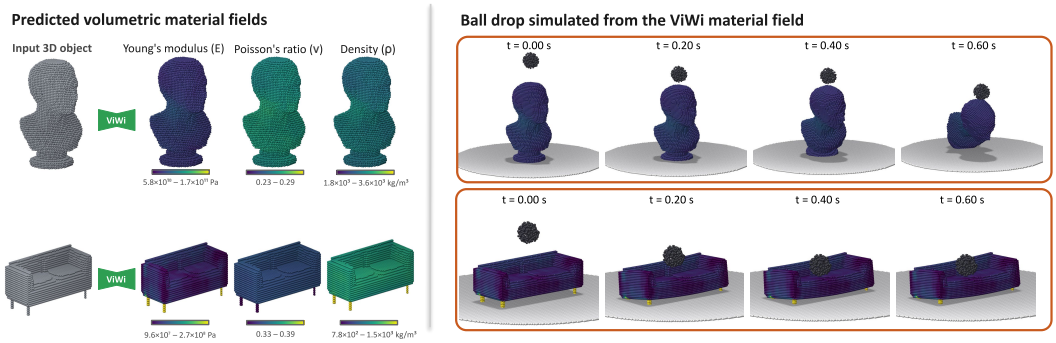


Figure 1: **From volumetric mechanical-property estimation to physics simulation.** ViWi predicts voxel-level Young's modulus, Poisson's ratio, and density from the input 3D object. The predicted material fields are then used to simulate the object's physical response to a ball impact.

ABSTRACT

Estimating volumetric mechanical properties, including Young's modulus, Poisson's ratio, and density at each voxel, is intrinsically ambiguous from vision alone, as visually similar objects may have substantially different material compositions and physical behavior. Existing approaches predict these properties independently across voxels, overlooking the piecewise-constant material structure of real objects and producing noisy or inconsistent estimates for voxels that share the same material, while lacking an explicit mechanism to resolve visual ambiguity. We introduce **ViWi** (**Vi**sion Meets **Wi**Fi), an object-centric framework for volumetric mechanical-property estimation. ViWi represents each object using a compact set of material slots that aggregate evidence from voxels with a shared material identity and produce coherent slot-level property predictions. To complement visual appearance, ViWi incorporates a compact RF descriptor generated through WiFi-band electromagnetic simulation using permittivity and conductivity. The RF descriptor conditions the material slots with global composition cues that may be unavailable from images, while visual features preserve voxel-level spatial localization. On GVM, ViWi improves over the prior state of the art on four of six per-voxel metrics, while its vision-only variant improves all reported mass-estimation metrics on ABO-500. These results demonstrate that combining object-centric material structure with complementary RF evidence enables more accurate and physically coherent volumetric property estimation beyond what is possible from visual appearance alone.

1 INTRODUCTION

Understanding how objects deform, resist external forces, and distribute mass is essential for physically grounded perception. Applications such as robotic manipulation, physics-based simulation, digital twins, augmented and virtual reality, and embodied intelligence require more than an object's geometry or semantic identity; They require spatially resolved estimates of its mechanical properties. In particular, Young's modulus E, Poisson's ratio ν, and density ρ determine an object's stiffness, lateral deformation, and mass distribution, respectively.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Accurate volumetric estimates of these quantities enable downstream systems to anticipate physical behavior, select safe interaction strategies, and construct simulation-ready object representations. However, assigning mechanical properties to 3D assets remains largely manual, requiring material annotations, expert knowledge, or direct physical measurements that are difficult to obtain at scale.

Recent work estimates physical and mechanical properties directly from visual observations (Zhai et al., 2024; Shuai et al., 2025; Lin et al., 2025; Dagli et al., 2025). Most closely related to our setting, VoMP (Dagli et al., 2025) projects multi-view visual features onto a volumetric grid and predicts Young's modulus, Poisson's ratio, and density for individual voxels. Although this formulation substantially reduces the need for manual annotation, it faces two fundamental limitations.

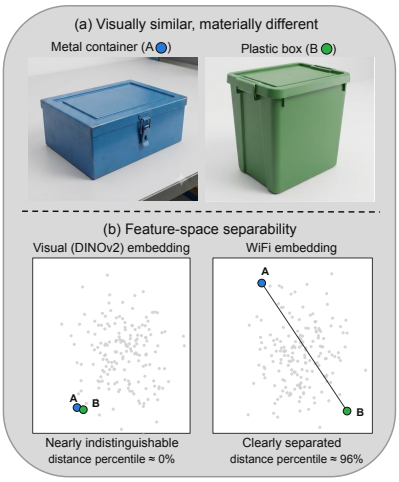


Figure 2: Visual ambiguity and RF separability. **(a)** Illustrative renderings of a metal container (A) and a plastic box (B). **(b)** The pair is nearly indistinguishable visually but clearly separated in the RF embedding.

First, independent voxel-wise prediction overlooks the compositional structure of real objects. Most objects contain only a small number of materials, such as a wooden seat, metal legs, and rubber feet. Voxels made of the same material should therefore exhibit similar mechanical properties and benefit from sharing evidence. Independent regression does not explicitly impose this structure and may produce noisy or inconsistent estimates across voxels belonging to the same material. Second, visual appearance does not uniquely determine material composition. Objects with similar geometry and appearance may be made from materials with substantially different mechanical properties. Figure 2 illustrates this ambiguity using a plastic box and a metal container from the GVM dataset (Dagli et al., 2025). Although the pair is nearly indistinguishable in the DINOv2 feature space, it is clearly separated in the RF embedding: the pairwise-distance percentile is approximately 0% in the visual embedding but 96% in the RF embedding. This example highlights a fundamental limitation of vision-only estimation: similar visual representations may correspond to mechanically distinct materials.

To address these limitations, we introduce **ViWi** (**Vi**sion Meets **Wi**Fi), an object-centric framework that combines visual and radio-frequency information for volumetric mechanical-property estimation. Rather than treating every voxel as an independent prediction problem, ViWi represents an object using a compact set of latent material slots. Voxels are softly assigned to these slots according to both visual compatibility and proximity in a learned material-latent space. Each slot aggregates evidence from the voxels assigned to it and predicts a shared slot-level property prototype, from which voxel-level properties are reconstructed through soft assignments. This formulation promotes consistent predictions across voxels that share a material identity, including spatially disconnected components.

ViWi further incorporates a compact RF descriptor obtained through physics-based electromagnetic simulation. Because RF propagation depends on material-sensitive properties such as permittivity and conductivity, it provides global composition cues that may not be available from appearance alone. The RF descriptor conditions the initial material slots, while the volumetric visual features retain the spatial information required to localize materials. Consequently, RF and visual evidence jointly influence material assignments and property estimation while preserving their complementary roles: RF helps identify which materials may be present, and vision determines where they occur.

Our contributions are summarized as follows:

- We introduce ViWi, the first framework to combine visual observations with RF sensing for volumetric mechanical-property estimation, and construct physics-based RF descriptors for GVM through WiFi-band electromagnetic simulation. The complementary RF descriptor helps distinguish visually similar objects with different material compositions and physical properties.
- We reformulate volumetric property estimation as object-centric material decomposition. Our material-slot formulation groups voxels using visual and material-latent compatibility,

allowing voxels with a shared material identity to aggregate evidence and receive coherent property predictions.

- We design an RF-conditioned material-slot initialization scheme based on bounded, zero-initialized feature-wise linear modulation, enabling stable integration of the global RF descriptor with voxel-level visual features for material grouping and property estimation.
- ViWi improves over the prior state of the art on four of six volumetric mechanical-property metrics on GVM, while its vision-only variant improves all reported mass-estimation metrics on ABO-500. Controlled analyses further show that RF provides the largest gains when visual evidence is ambiguous.

2 Related Work

2.1 Mechanical-Property Estimation from Visual Observations

Estimating mechanical properties from visual observations is challenging because appearance provides only indirect evidence about physical composition. Prior methods use pretrained representations, differentiable simulation, generative priors, or semantic cues to infer physical attributes (Zhai et al., 2024; Shuai et al., 2025; Zhang et al., 2024; Huang et al., 2024; Liu et al., 2024a; Xie et al., 2023; Liu et al., 2024b; Chen et al., 2025; Lin et al., 2025; Le et al., 2025; Zhao et al., 2025; Cao & Kalogerakis, 2025; Cao et al., 2025), but many require per-object optimization, focus on surfaces, or generate physically augmented assets rather than dense mechanical properties throughout an object's volume.

Most closely related, VoMP (Dagli et al., 2025) predicts volumetric Young's modulus, Poisson's ratio, and density independently from multi-view visual features, while AdaVoMP (Dagli et al., 2026) improves spatial resolution using an adaptive voxel structure but remains vision-only. ViWi instead models an object's low-cardinality material composition using material slots adapted from Slot Attention (Locatello et al., 2020), allowing voxels sharing a material identity to aggregate evidence, while RF information helps resolve visual ambiguity.

2.2 RF Sensing and Multimodal Material Perception

RF sensing provides material-sensitive information because electromagnetic propagation depends on object composition. Prior work has used RF for material sensing, activity recognition, and pose estimation (Shang et al., 2024; Lee et al., 2022; Yu et al., 2023; Fan et al., 2024; Knap et al., 2024). Existing vision–RF systems primarily target human sensing, pose estimation, or material classification. In contrast, ViWi uses a physics-based RF descriptor to condition object-centric material slots for dense volumetric mechanical-property estimation, jointly combining global RF composition cues with spatial visual features.

3 Method

3.1 Problem Formulation and Overview

Given multi-view observations of a 3D object, our goal is to estimate its volumetric mechanical-property field. For each occupied voxel i, the model predicts $\mathbf{y}_i = (E_i, \nu_i, \rho_i)$, where E_i, ν_i, and ρ_i denote Young's modulus, Poisson's ratio, and density, respectively.

Rather than predicting each voxel independently, ViWi represents an object using a compact set of latent material slots. Each slot captures a shared material hypothesis, aggregates evidence from the voxels assigned to it, and predicts a slot-level property prototype. Voxel-level predictions are reconstructed through soft assignments to these prototypes.

To reduce ambiguity from visual appearance, ViWi additionally uses an RF descriptor obtained from physics-based electromagnetic simulation. The RF descriptor conditions the material slots, allowing RF and visual evidence to jointly influence material assignments and property estimation.

We build ViWi on the volumetric representation of VoMP (Dagli et al., 2025). Its pretrained geometry backbone aggregates multi-view DINOv2 features on a 64^3 sparse voxel grid and produces $\mathbf{x}_i \in \mathbb{R}^{768}$ for each occupied voxel. We keep the geometry backbone and MatVAE frozen. MatVAE maps

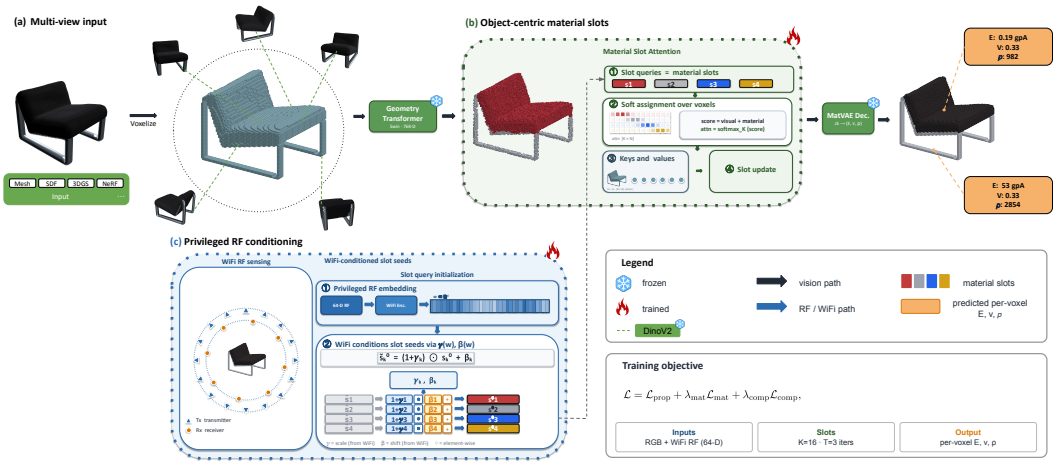


Figure 3: Overview of ViWi. Multi-view RGB features are aggregated on a sparse voxel grid and grouped into material slots using visual and material-latent compatibility. Each slot material latent is decoded by the frozen MatVAE into E, ν, and ρ, and voxel-level properties are reconstructed through soft assignments. A simulated 64-D RF descriptor conditions the initial material-slot states through bounded feature-wise modulation, providing complementary object-level material information.

mechanically valid property triplets to a two-dimensional material-latent space and decodes them back to (E, ν, ρ). Figure 3 provides an overview of ViWi.

3.2 Object-Centric Material Slot Attention

Material-slot representation. Building on Slot Attention (Locatello et al., 2020), ViWi represents each object using K latent material slots. Each slot contains a feature embedding $\mathbf{s}_k \in \mathbb{R}^{d_s}$ that aggregates voxel-level visual evidence and a two-dimensional material-latent seed $\bar{\mathbf{m}}_k \in \mathbb{R}^2$ in the pretrained MatVAE latent space. Together, these two representations allow each slot to capture both visual evidence and a hypothesis about material identity.

Voxel and material representations. The frozen geometry backbone produces voxel features

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top,$$

where $\mathbf{x}_i$ represents the visual information associated with occupied voxel i. Learned projections transform these features into voxel keys $\mathbf{k}_i$ and values $\mathbf{v}_i$, which are used for voxel-to-slot assignment and evidence aggregation, respectively. In parallel, a separate prediction head maps each voxel feature to a vision-based voxel material latent $\widehat{\mathbf{m}}_i \in \mathbb{R}^2$ in the pretrained MatVAE latent space. This latent provides a material-compatibility cue for voxel-to-slot assignment rather than serving as the final voxel-level property prediction.

Material-aware assignment. At each refinement iteration, ViWi computes a soft voxel-to-slot assignment. For voxel i and slot k, the assignment combines visual compatibility with proximity in the material-latent space:

$$\begin{aligned} a_{ik}^{\mathrm{vis}} &= \frac{q(\mathbf{s}_k)^\top \mathbf{k}_i}{\sqrt{d_a}}, \\ a_{ik}^{\mathrm{mat}} &= -\left\|\mathbf{m}_k - \widehat{\mathbf{m}}_i\right\|_2^2, \\ a_{ik} &= \lambda_{\mathrm{vis}} a_{ik}^{\mathrm{vis}} + \lambda_{\mathrm{mat}} a_{ik}^{\mathrm{mat}}, \end{aligned} \tag{1}$$

where $q(\cdot)$ is the slot-query projection, d_a is the attention dimension, and λ_{vis} and λ_{mat} are learned non-negative weights. We normalize the scores across slots:

$$\alpha_{ik} = \frac{\exp(a_{ik})}{\sum_{k'=1}^{K} \exp(a_{ik'})}, \qquad \sum_{k=1}^{K} \alpha_{ik} = 1. \tag{2}$$

The assignment weight α_{ik} represents the degree to which voxel i is assigned to slot k. Normalization across slots creates competition among material hypotheses and encourages the slots to specialize in distinct material identities.

Iterative evidence aggregation. Each slot aggregates the value features of the voxels assigned to it using

$$\mathbf{u}_k = \sum_{i=1}^{N} \alpha_{ik} \mathbf{v}_i. \tag{3}$$

The aggregated feature $\mathbf{u}_k$ updates the slot feature embedding $\mathbf{s}_k$ through a gated recurrent unit followed by a residual feed-forward network. Repeating this process allows the updated slot feature embeddings to refine the subsequent voxel-to-slot assignments, which in turn determine the voxel evidence aggregated by each slot. After the final refinement iteration, the slot material latent is obtained from the updated slot feature embedding through a residual prediction added to its initial material-latent seed.

Slot-level decoding and voxel reconstruction. Each final slot material latent $\mathbf{m}_k \in \mathbb{R}^2$ is mapped by the frozen MatVAE decoder $g_{\mathrm{mat}} : \mathbb{R}^2 \to \mathbb{R}^3$ to a slot-level property prototype,

$$\mathbf{p}_k = g_{\mathrm{mat}}(\mathbf{m}_k) = (E_k, \nu_k, \rho_k) \in \mathbb{R}^3.$$

Collecting all slot predictions gives $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_K]^\top \in \mathbb{R}^{K\times 3}$. The final voxel-level prediction is reconstructed using the soft voxel-to-slot assignment:

$$\widehat{\mathbf{y}}_i = \sum_{k=1}^{K} \alpha_{ik} \mathbf{p}_k = \boldsymbol{\alpha}_i^\top \mathbf{P}, \qquad \boldsymbol{\alpha}_i = [\alpha_{i1}, \ldots, \alpha_{iK}]^\top. \tag{4}$$

This reconstruction promotes consistency among voxels assigned to the same material slot while allowing smooth transitions between material slots through soft assignments.

3.3 Physics-Based RF Representation

To complement visual appearance with material-sensitive information, ViWi constructs a compact RF descriptor for each GVM object through physics-based electromagnetic simulation. Each mechanical material category is mapped to representative electromagnetic parameters, namely relative permittivity ϵ_r and conductivity σ, which determine how RF waves interact with the object.

The simulated propagation paths are aggregated into a fixed-size global descriptor. For each transmitter, we compute four statistics over all receivers: total scattered power, the fraction of receivers reached by a valid path, maximum single-path gain, and RMS delay spread. Stacking these statistics across N_{tx} transmitters gives

$$\mathbf{r} \in \mathbb{R}^D, \qquad D = N_{\mathrm{tx}} \times F, \tag{5}$$

where N_{tx} is the number of transmitters and F is the number of per-transmitter RF statistics. The resulting RF descriptor captures global, material-sensitive information about the object but does not encode the spatial locations of individual materials. We therefore use it to condition the initial material-slot states, while voxel-level visual features provide spatial localization. Detailed electromagnetic simulation settings are provided in the supplementary material.

3.4 RF-CONDITIONED MATERIAL SLOTS

The RF descriptor provides global evidence about material composition but does not encode the spatial locations of individual materials. We therefore use it to condition the initial material-slot states, while voxel-level visual features provide the spatial information required for voxel-to-slot assignment.

The standardized RF descriptor $\mathbf{r}$ is encoded by a multilayer perceptron into an RF embedding $\mathbf{z}_{\mathrm{RF}} \in \mathbb{R}^{d_{\mathrm{RF}}}$. From this embedding, the network predicts feature-wise scale and shift parameters for both the slot feature embedding and the material-latent seed. Following feature-wise linear modulation (FiLM) (Perez et al., 2018), the initial state of slot k is conditioned as

$$\widetilde{\mathbf{s}}_k^0 = (1 + \boldsymbol{\gamma}_k^s) \odot \mathbf{s}_k^0 + \boldsymbol{\beta}_k^s, \qquad \widetilde{\mathbf{m}}_k^0 = (1 + \boldsymbol{\gamma}_k^m) \odot \mathbf{m}_k^0 + \boldsymbol{\beta}_k^m, \tag{6}$$

where the superscript 0 denotes the slot state before the first refinement iteration. Each modulation parameter is obtained from a linear projection of $\mathbf{z}_{\mathrm{RF}}$ followed by $c\tanh(\cdot)$, which bounds its magnitude by c. This prevents the global RF descriptor from overwhelming the spatially resolved visual features. Modulating the material-latent seed $\mathbf{m}_k^0$ directly affects the material-compatibility score a_{ik}^{mat} in Eq. equation 1, while modulating the slot feature embedding $\mathbf{s}_k^0$ affects the visual-compatibility score.

Because the effect of the initial modulation may be attenuated during iterative slot refinement, we additionally apply a bounded RF-dependent correction after the final iteration:

$$\Delta\mathbf{m}_k^{\mathrm{RF}} = c\tanh(\mathbf{W}_r\mathbf{z}_{\mathrm{RF}})\,. \tag{7}$$

This correction is added to the final slot material latent before decoding, providing a direct RF-dependent path to the slot-level property prototype. RF therefore acts only on slot-level quantities, while voxel-level visual features continue to provide the spatial evidence required to localize materials.

All RF-conditioning projections are initialized with zero weights and biases. Thus, at the beginning of multimodal training, $\boldsymbol{\gamma} = \boldsymbol{\beta} = \mathbf{0}$ and $\Delta\mathbf{m}_k^{\mathrm{RF}} = \mathbf{0}$, so the RF pathway initially reduces to the identity transformation. The remaining parameters are initialized from the trained vision-only variant, ViWi†, allowing RF conditioning to be learned gradually from the vision-only solution.

3.5 LEARNING OBJECTIVES

Let $\mathbf{y}_i = (E_i, \nu_i, \rho_i)$ and $\widehat{\mathbf{y}}_i$ denote the ground-truth and predicted properties of occupied voxel i. We use an ℓ_1 property loss. We further align the vision-based voxel material latent $\widehat{\mathbf{m}}_i$ with a detached target $\mathbf{m}_i^{\mathrm{enc}}$ from the frozen MatVAE encoder, stabilizing the material-compatibility cue used for voxel-to-slot assignment. In parallel, the RF embedding is supervised with the object-level material composition, encouraging it to encode which materials are present.

For object b in batch $\mathcal{B}$, let $\mathbf{h}_b \in \mathbb{R}^C$ denote its ground-truth material-category distribution over C categories and $\boldsymbol{\ell}_b \in \mathbb{R}^C$ the predicted composition logits. The losses are

$$\begin{aligned}
&\mathcal{L}_{\mathrm{prop}} = \frac{1}{N}\sum_{i=1}^{N} \|\widehat{\mathbf{y}}_i - \mathbf{y}_i\|_1\,, \qquad \mathcal{L}_{\mathrm{mat}} = \frac{1}{N}\sum_{i=1}^{N} \|\widehat{\mathbf{m}}_i - \mathbf{m}_i^{\mathrm{enc}}\|_1\,, \\
&\mathcal{L}_{\mathrm{comp}} = -\frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}}\sum_{c=1}^{C} h_{b,c}\log \mathrm{softmax}(\boldsymbol{\ell}_b)_c, \qquad \mathcal{L} = \mathcal{L}_{\mathrm{prop}} + \lambda_{\mathrm{mat}}\mathcal{L}_{\mathrm{mat}} + \lambda_{\mathrm{comp}}\mathcal{L}_{\mathrm{comp}}.
\end{aligned} \tag{8}$$

4 EXPERIMENTS

We evaluate ViWi on volumetric mechanical-property estimation and assess whether the predicted density fields also improve downstream object-mass estimation. Experiments are conducted on GVM (Dagli et al., 2025) for volumetric estimation of Young's modulus, Poisson's ratio, and density,

Table 1: Volumetric mechanical-property estimation on the GVM hold-out test set. We report ALDE and ALRE for Young's modulus, and ADE and ARE for Poisson's ratio and density, following VoMP (Dagli et al., 2025). Errors are computed per object and then averaged across the test set. VoMP results are reproduced using the authors' pretrained weights to ensure a fair comparison. Lower is better; best results are shown in **bold**.

Method	Young's Modulus (E) [Pa]		Poisson's Ratio (ν)		Density (ρ) [kg/m^3]	
	ALDE (↓)	ALRE (↓)	ADE (↓)	ARE (↓)	ADE (↓)	ARE (↓)
NeRF2Physics - CVPR 2024 (Zhai et al., 2024)	2.8000 (±1.05)	0.1346 (±0.05)	–	–	1432.03 (±964.88)	1.0365 (±0.63)
PUGS - ICRA 2025 (Shuai et al., 2025)	3.3942 (±1.72)	0.1688 (±0.10)	–	–	3568.22 (±2839.13)	3.2429 (±3.56)
Phys4DGen* - ACM MM 2025 (Lin et al., 2025)	4.8967 (±3.17)	0.2227 (±0.14)	0.0407 (±0.04)	0.1467 (±0.18)	1865.57 (±2176.90)	1.4394 (±2.35)
VoMP - ICLR 2026 (Dagli et al., 2025)	0.3952 (±0.30)	0.0427 (±0.038)	0.0245 (±0.007)	0.0842 (±0.029)	153.06 (±174)	0.0898 (±0.071)
AdaVoMP - ICML 2026 (Dagli et al., 2026)	0.3278 (±0.26)	0.0340 (±0.03)	0.0205 (±0.01)	0.0680 (±0.03)	**127.31 (±150.83)**	**0.0842 (±0.07)**
ViWi (Ours)	**0.1797 (±0.21)**	**0.0196 (±0.026)**	**0.0151 (±0.015)**	**0.0515 (±0.052)**	198.22 (±340.85)	0.1034 (±0.11)

and on ABO-500 (Collins et al., 2022) for mass estimation derived from the predicted volumetric density field. Complete dataset descriptions, including the construction of RF descriptors for GVM, together with evaluation metrics, implementation details, training settings, additional ablations, and additional qualitative results are provided in the supplementary material.

4.1 Main Results

Table 1 compares ViWi with prior methods on the GVM benchmark, all evaluated on the same 64^3 grid; for the concurrent AdaVoMP (Dagli et al., 2026) we quote its own reported 64^3 operating point, obtained by halting its coarse-to-fine decoder early. ViWi is best on four of the six metrics. Against VoMP (Dagli et al., 2025), ViWi reduces Young's modulus ALDE/ALRE from $0.3952/0.0427$ to $0.1797/0.0196$, i.e. by 54.5% and 54.1%; this matters because E spans several orders of magnitude and is therefore hard to infer from visual appearance alone. For Poisson's ratio, ADE/ARE fall from $0.0245/0.0842$ to $0.0151/0.0515$ (38.4% and 38.8%).

Table 2: Mass estimation on ABO-500. Results on 100 real test objects. The upper block reports published results, while the lower block reports results reproduced under our fixed evaluation protocol.

Method	**ALDE ↓**	**ADE [kg] ↓**	**ARE ↓**	**MnRE ↑**
Reported in original papers				
NeRF2Physics Zhai et al. (2024)	0.736	12.725	1.040	0.564
PUGS Shuai et al. (2025)	0.661	9.461	0.767	0.576
Phys4DGen* Lin et al. (2025)	0.664	9.961	0.825	0.566
VoMP Dagli et al. (2025)	0.631	8.433	0.887	0.576
Reproduced under our fixed protocol				
VoMP Dagli et al. (2025)	1.170	37.372	3.121	0.395
ViWi† (Ours)	**1.113**	**20.181**	**2.538**	**0.415**

ViWi also outperforms AdaVoMP on E (45.2% lower ALDE, 42.4% lower ALRE) and ν (26.3% lower ADE, 24.3% lower ARE). Both numbers are read at 64^3, yet AdaVoMP enters the comparison with a considerably stronger recipe: a 0.6B-parameter adaptive-voxel model using DINOv3 ViT-H+/16 features and depth-attenuated multi-view averaging, evaluated at 64^3 by halting its coarse-to-fine decoder early. ViWi instead predicts natively on VoMP's fixed 64^3 grid from DINOv2 features with uniform view averaging. Our gains thus stem from material grouping and RF conditioning rather than a stronger backbone or finer discretisation, and the two directions are complementary.

Density is the only property for which ViWi has a higher mean error, with ADE/ARE of $198.22/0.1034$ compared with $153.06/0.0898$ for VoMP. The gap is driven by a small number of outlier objects rather than a systematic performance loss: the two models are effectively tied on the median test object (ADE 113.1 vs. 114.0), the ten worst objects account for 93% of the mean gap, and a paired Wilcoxon signed-rank test on the per-object errors of the 148 test objects finds no significant difference ($p = 0.24$). Density is also the only property whose metric can expose such a tail: Young's modulus is scored in log space, which compresses even order-of-magnitude

Table 3: Material-slot attention ablation on GVM. Comparison of VoMP, ViWi† (Vision Only), and the full ViWi model. Lower is better for all metrics.

Method	Young's Modulus (E)		Poisson's Ratio (ν)		Density (ρ)	
	ALDE	ALRE	ADE	ARE	ADE	ARE
VoMP (Dagli et al., 2025)	0.3952 (±0.3012)	0.0427 (±0.0381)	0.0245 (±0.0072)	0.0842 (±0.0288)	**153.06 (±173.7823)**	**0.0898 (±0.0705)**
ViWi†	**0.1853 (±0.2026)**	**0.0202 (±0.0249)**	**0.0146 (±0.0151)**	**0.0503 (±0.0558)**	238.71 (±304.1573)	0.1243 (±0.1198)

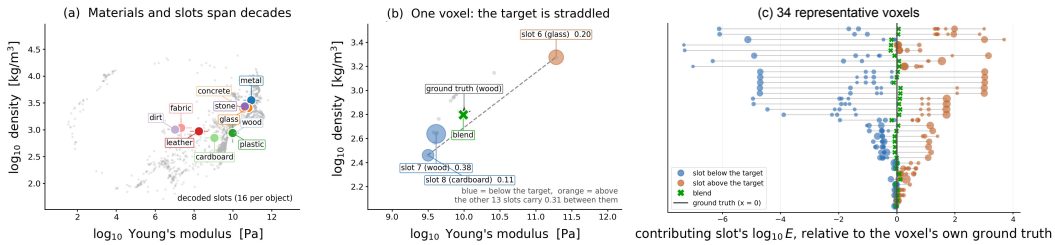


Figure 4: Why soft voxel-to-slot assignment helps. **(a)** Decoded material slots span the broad range of mechanical properties. **(b)** A voxel can combine softer and stiffer slots to match a target not represented by any single slot; circle size is the attention weight. **(c)** This bracketing behavior occurs systematically across voxels, shown for 34 representative ones. **(d)** Replacing soft assignment with straight-through Gumbel one-hot assignment increases overall relative error from 6.42% to 7.72%.

discrepancies, and Poisson's ratio is bounded to $[0, 0.49]$ in GVM, whereas density is scored linearly and spans 102–19,800 kg/m^3.

We further evaluate whether the predicted density fields yield accurate mass estimates on the real-object ABO-500 benchmark. ViWi† denotes our vision-only variant: RF is unavailable here, since our descriptors are simulated from per-voxel electromagnetic properties while ABO-500 provides only a single ground-truth mass. Converting density to mass also requires a volume estimate that prior work leaves unspecified, so published numbers are not directly comparable. We therefore re-evaluate VoMP and ViWi† under one identical protocol. No protocol of the form $m = c\,\rho V$ recovers VoMP's reported ALDE of 0.631 from its released weights (best attainable 1.094), so the gap lies in unpublished evaluation details rather than in the weights. Under our protocol (Table 2), ViWi† improves on all four metrics, reducing ADE by 46.0%.

4.2 Ablation Studies

Where does RF help? Gain versus visual difficulty. We rank the 305 validation and test objects by the overall error of ViWi‡ and split them into four equal-count quartiles (Q_1 easiest to Q_4 hardest). ViWi‡ is ViWi† trained for the same number of steps as ViWi with the RF input withheld, so the gap between them reflects RF rather than extra optimisation. Figure 5 reports the resulting reduction in density error: RF is neutral on the easiest quartile (-0.01 pp) and reaches $+2.87$ pp on the hardest, where 63% of objects improve. RF thus adds complementary material information for density where visual appearance is least reliable, while having little effect where vision already suffices.

Why soft voxel-to-slot assignment helps. ViWi predicts each voxel as a weighted mixture of slots rather than assigning it to a single slot. The slots cover the wide range of materials in GVM (Fig. 4(a)), but a single slot rarely matches a given voxel exactly. Mixing a softer and a stiffer slot allows the prediction to fall between them and land closer to the ground truth than either slot alone (Fig. 4(b)). This is the common case rather than a special one (Fig. 4(c)): for 74.1% of voxels, the true values lie between the contributing slots for all three properties simultaneously; for 83.6%, the mixture outperforms its best single slot; and the error for these voxels is lower (0.041 versus 0.072). Forcing a hard choice with a straight-through Gumbel variant, all else being equal, raises the overall relative error—the mean of the relative errors on $\log_{10} E$, ν, and ρ— from 6.42% to 7.72% at matched sample counts.

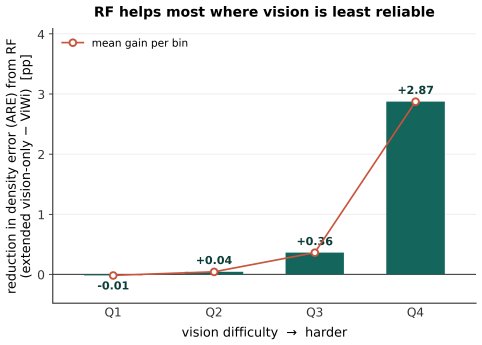


Figure 5: RF gain versus visual difficulty. Objects are split into four quartiles by ViWi‡ error, where ViWi‡ is ViWi† trained to ViWi's step budget without RF. Bars show the mean reduction in density ARE (percentage points); positive is better.

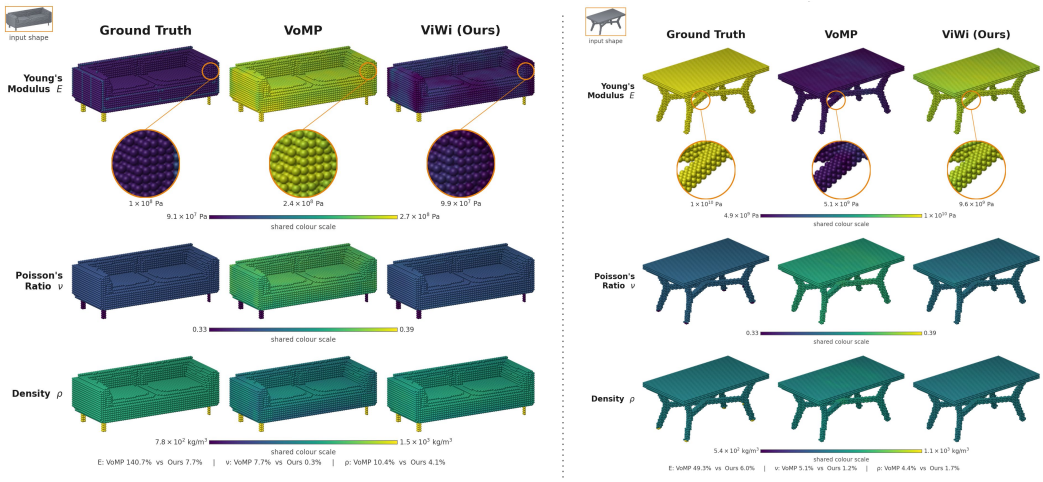


Figure 6: Qualitative comparison of volumetric mechanical-property estimation. Ground truth, VoMP (Dagli et al., 2025), and ViWi predictions for E, ν, and ρ on two representative GVM test objects. Percentages are per-object mean relative errors, with E measured in linear Pa. ViWi more closely matches the ground-truth magnitudes, with reduced local errors in the highlighted regions.

Contribution of material slot attention. To isolate the material-slot formulation, we compare VoMP's voxel-wise regression with ViWi† (Vision Only), which shares the frozen geometry backbone and MatVAE decoder but receives no RF input. As shown in Table 3, material-slot attention substantially improves both metrics for Young's modulus and Poisson's ratio, by approximately 53% and 40%, respectively. Density is the exception, although its gap is largely driven by a few outlier objects as discussed in Sec. 4.1; the full ViWi model recovers much of this loss. These results support grouping voxels through shared material slots rather than predicting each voxel independently.

4.3 Qualitative Results

Figure 6 compares ViWi with VoMP (Dagli et al., 2025) and the ground truth on two representative GVM test objects. VoMP's dominant failure is one of magnitude: it assigns large regions the wrong stiffness level, over-stiffening the sofa and under-stiffening the table, whereas ViWi recovers the correct level for the dominant material of each object. Averaged over all voxels, the relative error in E (in linear Pa) falls from 140.7% to 7.7% on the sofa and from 49.3% to 6.0% on the table, with consistent gains for ν and ρ; the highlighted regions show the same discrepancy at the voxel scale.

4.4 Computational Cost

ViWi adds only 1.88M trainable parameters on top of the frozen VoMP backbone and MatVAE. ViWi† and ViWi are trained for 112k steps on 4 and 8 GPUs, taking 24.5 and 32.6 hours, respectively, at about 7 GB/GPU. As shown in Table 4, ViWi incurs only modest inference-time computation and memory overhead, while RF descriptors are computed once offline.

Table 4: Computational cost at inference.

	VoMP	ViWi†	ViWi
Inference (per object)			
4.3k voxels [ms]	8.4	9.9	10.3
28.8k voxels [ms]	25.4	27.6	27.9
32.8k voxels [ms]	28.2	30.3	30.7
Mean, 32 obj. [ms]	24.3	26.3	26.6
Memory [MB]	758	763	765

5 Conclusion and Limitations

We introduced ViWi, a multimodal framework that combines object-centric material slots with RF evidence for volumetric mechanical-property estimation. ViWi improves four of six metrics on GVM, with strong gains for Young's modulus and Poisson's ratio, while its vision-only variant improves real-object mass estimation on ABO-500. A key limitation is that the RF descriptors are simulated rather than measured; future work will evaluate ViWi with real RF observations under realistic sensing conditions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with brainstorming and refining research ideas, provide feedback on research methodology and experimental design, and support the interpretation of experimental results. We also used generative AI tools for literature retrieval and discovery, drafting parts of the manuscript, and editing and polishing the writing for clarity and readability. We did not use generative AI tools to generate synthetic datasets or to formulate or prove mathematical claims. All AI-assisted research ideas, interpretations, literature suggestions, and manuscript text were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including all text, claims, results, and conclusions.

REPRODUCIBILITY STATEMENT

We provide the implementation details, training settings, evaluation protocols, RF descriptor simulation procedure, and additional ablation studies in the supplementary material. The datasets and evaluation splits used in our experiments are described in the experimental setup, and all reported metrics and model configurations are specified to support reproduction of the results.

REFERENCES

Junyi Cao and Evangelos Kalogerakis. Sophy: Generating simulation-ready objects with physical materials. *arXiv preprint arXiv:2504.12684*, 2025.

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

Boyuan Chen, Hanxiao Jiang, Shaowei Liu, Saurabh Gupta, Yunzhu Li, Hao Zhao, and Shenlong Wang. Physgen3d: Crafting a miniature interactive world from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6178–6189, 2025.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21094–21104. IEEE, 2022.

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

Rishit Dagli, Donglai Xiang, Vismay Modi, Xuning Yang, Gavriel State, David IW Levin, and Maria Shugrina. Adaptive volumetric mechanical property fields invariant to resolution. *arXiv preprint arXiv:2606.18231*, 2026.

Junqiao Fan, Jianfei Yang, Yuecong Xu, and Lihua Xie. Diffusion model is a good pose estimator from 3d rf-vision. *arXiv preprint arXiv:2403.16198*, 2024.

Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson W. H. Lau. Dreamphysics: Learning physics-based 3d dynamics with video diffusion priors. *arXiv preprint arXiv:2406.01476*, 2024.

Pawel Knap, Peter Hardy, Alberto Tamajo, Hwasup Lim, and Hansung Kim. Improving real-time omnidirectional 3d multi-person human pose estimation with people matching and unsupervised 2d–3d lifting. *arXiv preprint arXiv:2403.09437*, 2024.

Long Le, Ryan Lucas, Chen Wang, Chuhao Chen, Dinesh Jayaraman, Eric Eaton, and Lingjie Liu. Pixie: Fast and generalizable supervised learning of 3d physics from pixels. *arXiv preprint arXiv:2508.17437*, 2025.

Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. Hupr: A benchmark for human pose estimation using millimeter wave radar. *arXiv preprint arXiv:2210.12564*, 2022.

Jiajing Lin, Zhenzhong Wang, Dejun Xu, Shu Jiang, Yunpeng Gong, and Min Jiang. Phys4dgen: Physics-compliant 4d generation with multi-material composition perception. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10398–10407, 2025.

Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3d: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024a.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision*, pp. 360–378. Springer, 2024b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *arXiv preprint arXiv:2006.15055*, 2020.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Fei Shang, Haocheng Jiang, Panlong Yang, Dawei Yan, Haohua Du, and Xiang-Yang Li. The field-based model: A new perspective on rf-based material sensing. *arXiv preprint arXiv:2412.05640*, 2024.

Yinghao Shuai, Ran Yu, Yuantao Chen, Zijian Jiang, Xiaowei Song, Nan Wang, Jv Zheng, Jianzhu Ma, Meng Yang, Zhicheng Wang, et al. Pugs: Zero-shot physical understanding with gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4478–4485. IEEE, 2025.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023.

Cong Yu, Dongheng Zhang, Zhi Wu, Zhi Lu, Chunyang Xie, Yang Hu, and Yan Chen. Rfpose-ot: Rf-based 3d human pose estimation via optimal transport theory. *arXiv preprint arXiv:2301.13013*, 2023.

Albert J Zhai, Yuan Shen, Emily Y Chen, Gloria X Wang, Xinlei Wang, Sheng Wang, Kaiyu Guan, and Shenlong Wang. Physical property understanding from language-embedded feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28296–28305, 2024.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. *arXiv preprint arXiv:2404.13026*, 2024.

Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. Physsplat: Efficient physics simulation for 3d scenes via mllm-guided gaussian splatting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5242–5252, 2025.

A EXPERIMENTAL SETUP

Datasets. We evaluate ViWi on the same two datasets and under the same evaluation protocol as VoMP (Dagli et al., 2025). The primary evaluation is conducted on GVM, a collection of simulation-ready 3D assets represented on a 64^3 voxel grid. Each object is segmented into material regions, and every occupied voxel is assigned ground-truth Young's modulus E, Poisson's ratio ν, and density ρ through a segment-to-material-to-property annotation pipeline. We use the predefined object-level training, validation, and test splits and report results on the held-out test set. As 18 of the 166 test objects are absent from the public GVM release, we evaluate on the 148 released objects and re-run VoMP on this same subset so that all reported numbers are directly comparable.

We additionally evaluate mass estimation on ABO-500, which contains real product assets with measured object masses. Following VoMP, we use the 100-object test subset only for evaluation and do not train or fine-tune on ABO-500. Since voxel-level mechanical-property annotations are unavailable, RF descriptors cannot be constructed for these objects; we therefore evaluate ViWi†, our vision-only variant. The predicted density field is integrated over the occupied object volume and compared with the measured object mass. This evaluation assesses whether the learned volumetric density fields transfer from simulated assets to real objects.

RF descriptor simulation. Our RF descriptors are simulated with Sionna RT in three stages, all purely additive: the original GVM voxel and feature files are never modified.

(i) Per-voxel electromagnetic properties. Each voxel of an object carries a free-form material string from the GVM annotations. We map every string to one of 12 material categories by exact name match, then by an ordered set of keyword rules (e.g. `chrome`→metal, `pine_bark`→wood), and finally, for uninformative names, by nearest-centroid assignment in $(\log_{10} E,\ \nu,\ \log_{10} \rho)$ space, allowing the mechanical properties to resolve ambiguous material names. Each category is assigned a relative permittivity ε_r and conductivity σ at 2.4 GHz, taken from ITU-R P.2040 where the material is tabulated (wood, glass, concrete, ground) and from the antenna-materials literature otherwise; metal is modeled as a near-perfect conductor. The resulting per-voxel (ε_r, σ) fields are stored in the same row order as the voxel grid and therefore align index-for-index with the mechanical ground truth.

(ii) Geometry. Voxels are grouped by material category and converted to one surface triangle mesh per category. Only *exposed* faces are emitted, i.e. faces whose neighboring voxel is empty, yielding a thin surface shell rather than a union of solid cubes and substantially reducing the triangle count for ray tracing. Each object is centered at the origin and isotropically scaled so that its largest bounding-box dimension equals 1 m, making the sensing geometry comparable across objects with different native scales.

(iii) Propagation. Each category mesh is assigned its own `RadioMaterial`(ε_r, σ), and N_{tx} transmitters are placed on a Fibonacci sphere of radius 1.5 m around the object, with 32 receivers on a concentric sphere at $1.05\times$ that radius. All elements are isotropic and vertically polarized. We trace paths up to depth 5 with specular reflection, diffuse reflection, and refraction enabled, while disabling the line-of-sight path so that the descriptor captures the object's scattering response rather than direct transmitter–receiver coupling. For each transmitter, we aggregate the complex path coefficients over all receivers and paths into four statistics: total scattered power (dB), the fraction of receivers with a valid path, maximum single-path gain (dB), and power-weighted RMS delay spread (ns). Concatenating these statistics over transmitters gives a fixed $4N_{\text{tx}}$-dimensional RF descriptor per object (64 dimensions for $N_{\text{tx}} = 16$), independent of the voxel count N. Transmitters with no valid return are floored at -120 dB. Each descriptor dimension is then z-scored using fixed normalization statistics shared across the train, validation, and test splits.

Evaluation metrics. For volumetric mechanical-property estimation on GVM, we follow VoMP (Dagli et al., 2025). Errors are first computed over the occupied voxels of each object and then averaged across objects to avoid weighting objects by their number of occupied voxels. We report Average Log Displacement Error (ALDE) and Average Log Relative Error (ALRE) for Young's modulus, and Average Displacement Error (ADE) and Average Relative Error (ARE) for Poisson's ratio and density.

For ABO-500, we evaluate the mass obtained from the predicted volumetric density field and report ALDE, ADE, ARE, and minimum ratio error (MnRE), following VoMP. Lower values are better for

ALDE, ADE, and ARE, whereas higher MnRE indicates better agreement between predicted and ground-truth masses.

Implementation details. Unless otherwise specified, ViWi uses $K = 16$ material slots with embedding dimension $d_s = 256$, refined for $T = 3$ slot refinement iterations. The RF descriptor contains four statistics for each of sixteen transmitters, resulting in a 64-dimensional input. We use the modulation bound $c = 0.3$ and set $\lambda_{\text{mat}} = \lambda_{\text{comp}} = 0.1$. The pretrained geometry backbone and MatVAE are kept frozen.

B Additional Ablation Studies

Effect of additional vision-only training (Ablations I–II). Because ViWi is initialized from ViWi†, our vision-only variant, its improvement could partly result from additional optimization rather than RF conditioning. We therefore continue training ViWi† for the same number of steps while keeping RF disabled. Table 5 reports this control on both the full test set (Ablation II) and the visually ambiguous subset (Ablation I), where mean error denotes the average relative error across $\log_{10} E$, ν, and ρ. On the full set, extended vision-only training reduces the mean error from 6.50% to 6.10%, whereas ViWi further reduces it to 5.81%. The same trend is stronger on visually ambiguous objects, where the error decreases from 7.15% to 6.79% with additional vision-only training and to 6.21% with RF. These results show that the gains of ViWi cannot be explained by additional training alone and that RF provides complementary information, particularly when visual evidence is less reliable.

Slot refinement iterations (Ablation III). As shown in Table 5, increasing the number of slot refinement iterations from $T = 3$ to $T = 7$ redistributes accuracy across properties rather than improving it uniformly. Young's modulus and Poisson's ratio degrade (ALDE $0.185 \rightarrow 0.230$, $+24\%$; ADE $0.0146 \rightarrow 0.0167$, $+14\%$), while density improves (ADE $238.7 \rightarrow 201.1$, -16%; ARE $0.124 \rightarrow 0.104$, -17%). Because the mean error averages the relative errors across $\log_{10} E$, ν, and ρ, the density improvement reduces the aggregate error from 6.50% to 6.20%. We retain $T = 3$ as the default because it requires less than half the refinement iterations.

RF noise robustness (Ablation IV). We evaluate robustness by adding zero-mean Gaussian noise to the standardized RF descriptor at inference while keeping the model fixed. As shown in Table 5, performance remains stable under mild perturbations: the mean error changes from 5.82% with clean RF to 5.81% at $\sigma = 0.1$ and 6.02% at $\sigma = 0.5$. Even under the strong perturbation $\sigma = 1.0$, the error increases only to 6.28%. These results show that ViWi degrades gradually as RF corruption increases and is robust to moderate perturbations of the RF descriptor.

Generalization to unseen materials (Ablation V). Because the simulated RF descriptor is derived from material-dependent electromagnetic properties, a possible concern is that the model may simply memorize RF patterns associated with materials seen during training. To test this, we hold out two material categories entirely. Using a per-object membership index derived from the per-voxel material labels, we remove every object that contains metal or rubber in *any* voxel from the training and validation splits, which reduces training from 1255 to 505 objects and validation from 157 to 69. Evaluation is then performed *only* on the 61 of 148 test objects that contain at least one of the two held-out materials; the remaining 87 test objects are not scored in this ablation, so the errors here are not comparable to those of the full-test-split rows above. The holdout is defined at the material-category level rather than by raw material names, so all variants are excluded together (e.g. chrome, steel, brass, and aluminium are all mapped to metal). The model therefore never observes these material categories during training or model selection.

As shown in Table 5, ViWi†, our vision-only variant, obtains a mean error of 20.47%, while RF conditioning, warm-started from the vision-only model, reduces it to 16.18%, an improvement of 4.29 percentage points (21% relative). RF improves both density metrics (ADE $887.0 \rightarrow 757.7$, ARE $0.419 \rightarrow 0.304$) and three of the four metrics for Young's modulus and Poisson's ratio. These results indicate that the RF contribution is not explained solely by memorizing materials observed during training and that the RF descriptor provides transferable material-sensitive information for unseen material categories.

Table 5: Additional ablation studies on GVM. Results for training duration, slot refinement, RF noise robustness, unseen materials, and number of RF transmitters. Metrics are first averaged over occupied voxels within each object and then averaged across objects. Mean error is the average relative error across $\log_{10} E$, ν, and ρ. Lower is better for all metrics.

Configuration	Young's Modulus (E)		Poisson's Ratio (ν)		Density (ρ)		Mean Error (%)
	ALDE (↓)	ALRE (↓)	ADE (↓)	ARE (↓)	ADE (↓)	ARE (↓)	
Ablation I: Extended Vision-Only training, visually ambiguous objects (80 of 148)							
ViWi†	0.2085 (±0.2036)	0.0232 (±0.0259)	0.0138 (±0.0140)	0.0483 (±0.0533)	244.51 (±234.6938)	0.1430 (±0.1274)	7.15 (±0.0517)
ViWi† (extended training)	0.2132 (±0.2514)	0.0240 (±0.0312)	0.0136 (±0.0150)	0.0468 (±0.0518)	211.83 (±194.4298)	0.1327 (±0.1343)	6.79 (±0.0574)
ViWi	0.2052 (±0.2196)	0.0227 (±0.0270)	0.0148 (±0.0155)	0.0512 (±0.0557)	194.31 (±211.8055)	0.1123 (±0.0995)	6.21 (±0.0474)
Ablation II: Extended Vision-Only training							
ViWi†	0.1853 (±0.2026)	0.0202 (±0.0249)	0.0146 (±0.0151)	0.0503 (±0.0558)	238.71 (±304.1573)	0.1243 (±0.1198)	6.50 (±0.0492)
ViWi† (extended training)	0.1812 (±0.2271)	0.0200 (±0.0276)	0.0145 (±0.0152)	0.0491 (±0.0504)	197.00 (±238.3770)	0.1140 (±0.1228)	6.10 (±0.0516)
ViWi	0.1797 (±0.2121)	0.0196 (±0.0257)	0.0151 (±0.0152)	0.0515 (±0.0516)	198.22 (±340.8454)	0.1034 (±0.1080)	5.81 (±0.0475)
Ablation III: Slot refinement iterations							
ViWi†, $T = 3$	0.1853 (±0.2026)	0.0202 (±0.0249)	0.0146 (±0.0151)	0.0503 (±0.0558)	238.71 (±304.1573)	0.1243 (±0.1198)	6.50 (±0.0492)
ViWi†, $T = 7$	0.2304 (±0.2366)	0.0246 (±0.0271)	0.0167 (±0.0139)	0.0578 (±0.0497)	201.12 (±341.5979)	0.1038 (±0.1138)	6.20 (±0.0458)
Ablation IV: RF noise robustness							
ViWi	0.1797 (±0.2121)	0.0196 (±0.0257)	0.0151 (±0.0152)	0.0515 (±0.0516)	198.22 (±340.8454)	0.1034 (±0.1080)	5.81 (±0.0475)
ViWi, $\sigma = 0.10$	0.1812 (±0.2138)	0.0197 (±0.0257)	0.0151 (±0.0152)	0.0515 (±0.0516)	198.22 (±340.8454)	0.1034 (±0.1080)	5.81 (±0.0475)
ViWi, $\sigma = 0.25$	0.1841 (±0.2091)	0.0200 (±0.0251)	0.0156 (±0.0152)	0.0532 (±0.0519)	208.05 (±391.1836)	0.1045 (±0.1082)	5.92 (±0.0473)
ViWi, $\sigma = 0.50$	0.1838 (±0.2046)	0.0200 (±0.0246)	0.0158 (±0.0151)	0.0536 (±0.0522)	213.47 (±388.0237)	0.1071 (±0.1062)	6.02 (±0.0467)
ViWi, $\sigma = 1.00$	0.1986 (±0.2120)	0.0218 (±0.0259)	0.0154 (±0.0145)	0.0520 (±0.0510)	224.55 (±393.4308)	0.1145 (±0.1062)	6.28 (±0.0463)
Ablation V: Unseen material combination							
ViWi†	0.7002 (±0.5513)	0.0784 (±0.0701)	0.0343 (±0.0340)	0.1164 (±0.1293)	887.01 (±1004.8889)	0.4192 (±0.5558)	20.47 (±0.2030)
ViWi	0.7232 (±0.5160)	0.0756 (±0.0529)	0.0318 (±0.0271)	0.1059 (±0.1018)	757.73 (±838.1707)	0.3040 (±0.3651)	16.18 (±0.1341)
Ablation VI: Number of RF transmitters							
ViWi, $N_{\text{tx}} = 4$	0.1934 (±0.2204)	0.0207 (±0.0258)	0.0157 (±0.0145)	0.0549 (±0.0529)	230.33 (±353.14)	0.1180 (±0.1337)	6.45 (±0.0560)
ViWi, $N_{\text{tx}} = 16$	0.1951 (±0.2403)	0.0213 (±0.0287)	0.0164 (±0.0159)	0.0569 (±0.0547)	190.92 (±246.10)	0.0980 (±0.0995)	5.87 (±0.0452)
Ablation VII: Tx/Rx sphere radius							
ViWi, $r = 1.5\,\text{m}$ (train)	0.1797 (±0.2121)	0.0196 (±0.0257)	0.0151 (±0.0152)	0.0515 (±0.0516)	198.22 (±340.85)	0.1034 (±0.1080)	5.81 (±0.0475)
ViWi, $r = 2.0\,\text{m}$	0.1877 (±0.2193)	0.0206 (±0.0268)	0.0140 (±0.0141)	0.0470 (±0.0483)	198.08 (±297.99)	0.1063 (±0.1092)	5.80 (±0.0479)
ViWi, $r = 4.0\,\text{m}$	0.2067 (±0.2109)	0.0227 (±0.0259)	0.0146 (±0.0149)	0.0489 (±0.0514)	213.06 (±356.55)	0.1150 (±0.1038)	6.22 (±0.0462)
ViWi, $r = 8.0\,\text{m}$	0.2214 (±0.2072)	0.0241 (±0.0251)	0.0160 (±0.0132)	0.0534 (±0.0462)	216.98 (±303.85)	0.1228 (±0.1014)	6.68 (±0.0450)

Number of RF transmitters (Ablation VI). As shown in Table 5, Ablation VI varies the number of transmitters N_{tx} used to construct the RF descriptor. Since each transmitter contributes four statistics, $N_{\text{tx}} = 4$ and 16 correspond to descriptor dimensions of 16 and 64, respectively. Increasing the number of transmitters from 4 to 16 reduces the mean error from 6.31% to 5.81%, with improvements in five of the six reported metrics. This suggests that broader angular coverage provides more informative RF measurements. We therefore use $N_{\text{tx}} = 16$ throughout.

Sensitivity to the sensing distance (Ablation VII). The RF descriptor is simulated with the transmitters and receivers on concentric spheres around the object, whose radius sets how far the antennas are from it. Every object is rescaled to a maximum size of $1\,\text{m}$, so the sensing distance is comparable across objects. The descriptor has the same dimensionality at any radius, so a trained model can directly process a descriptor simulated at a new radius. We therefore re-simulate the test set at three larger radii and evaluate the *same* trained weights on each, without retraining, while keeping the normalization statistics fixed to those of the training radius.

As shown in Table 5, performance remains stable near the training radius: at $2.0\,\text{m}$, the mean error is 5.80%, compared with 5.81% at the training radius of $1.5\,\text{m}$, with slightly better Poisson's ratio

Table 6: Effect of the material-aware assignment score. Comparison of ViWi† with and without the material-compatibility term. All metrics are computed globally over all occupied voxels, rather than per object followed by averaging across objects. Mean error is the average relative error across $\log_{10} E$, ν, and ρ. Lower is better for all metrics.

Configuration	Young's Modulus (E)		Poisson's Ratio (ν)		Density (ρ)		Mean Error (%)
	ALDE	ALRE	ADE	ARE	ADE	ARE	
ViWi†, without material score	0.1969	0.0223	0.0156	0.0543	143.86	0.1061	6.09
ViWi†, with material score	**0.1763**	**0.0199**	**0.0146**	**0.0511**	164.45	**0.0958**	**5.56**

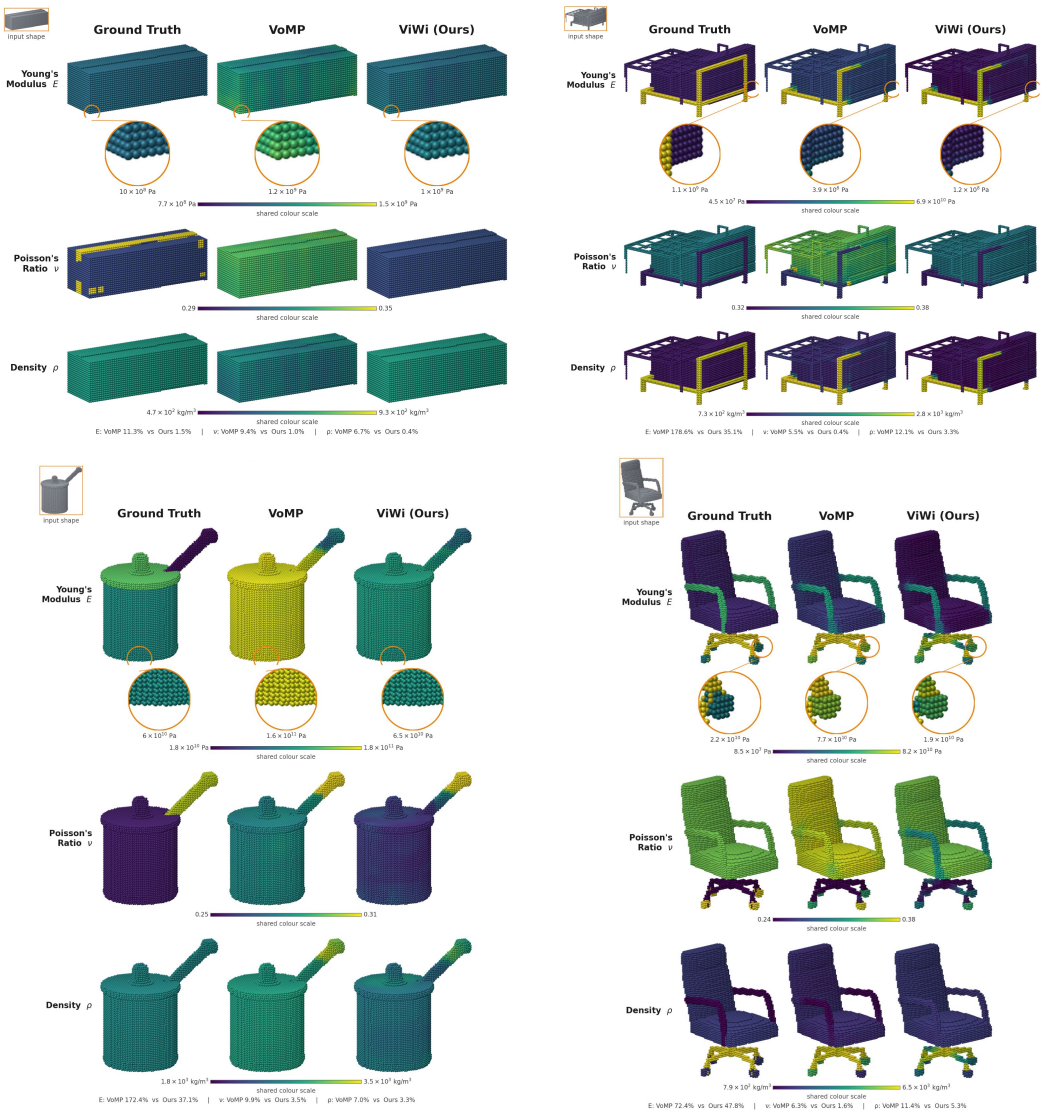


Figure 7: Additional qualitative comparison of volumetric mechanical-property estimation. Ground-truth fields are compared with predictions from VoMP (Dagli et al., 2025) and ViWi for Young's modulus E, Poisson's ratio ν, and density ρ. A shared color scale is used within each property and object. The magnified regions highlight local differences in Young's modulus predictions.

estimation. Accuracy then degrades with distance, reaching 6.22% at $4.0\,\mathrm{m}$ and 6.68% at $8.0\,\mathrm{m}$. As the antennas move farther from the object, the simulator finds no valid propagation path for an increasing fraction of Tx–Rx links, from 16% at $1.5\,\mathrm{m}$ to 96% at $8.0\,\mathrm{m}$. Consequently, the RF descriptor contains progressively less object-dependent information at larger sensing distances. These results show that ViWi is robust to moderate changes in sensing distance around the training configuration, while performance degrades at substantially larger distances.

Contribution of the material-aware assignment score. We evaluate the contribution of the material-compatibility term used in voxel-to-slot assignment. In this ablation, we remove only the material score a_{ik}^{mat} from Eq. 1, so assignments are determined solely by visual compatibility. The auxiliary material-latent loss $\mathcal{L}_{\mathrm{mat}}$ remains unchanged, isolating the effect of material compatibility during slot assignment from that of material-latent supervision.

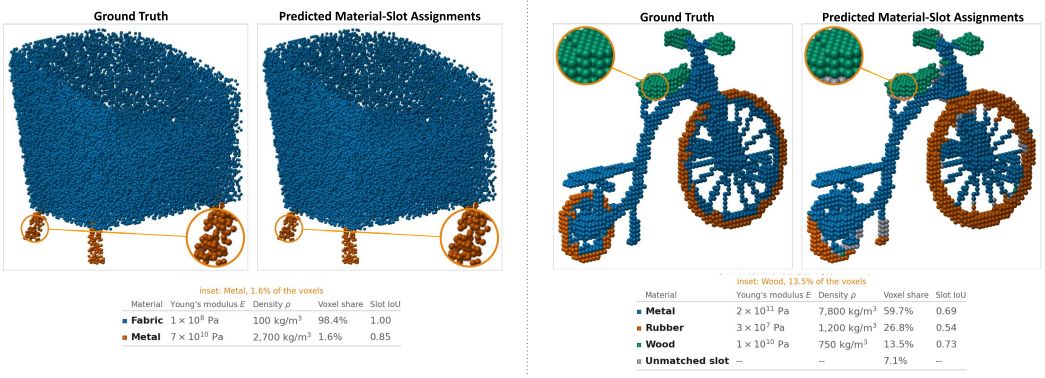


Figure 8: Visualization of learned material-slot assignments. Ground-truth material regions are compared with the material-slot assignments predicted by ViWi for an armchair (left) and a tricycle (right). Colors are matched between ground-truth materials and their corresponding predicted slots. The armchair slots recover the dominant fabric region and the small metal feet, while the tricycle slots separate metal, rubber, and wood components. Insets highlight small material regions. The tables report material properties, voxel shares, and IoU between each ground-truth material region and its matched slot. Gray denotes an unmatched predicted slot.

As shown in Table 6, removing the material score increases the mean relative error from 5.56% to 6.09%. Performance degrades across all three properties: the relative error increases from 1.99% to 2.23% for Young's modulus, from 5.11% to 5.43% for Poisson's ratio, and from 9.58% to 10.61% for density. These results show that visual compatibility alone is insufficient for reliable material grouping. Compatibility in the learned material-latent space helps associate voxels with material slots according to their inferred mechanical identity.

C Additional Qualitative Results

Additional volumetric property visualizations. Figure 7 provides additional qualitative comparisons between ViWi, VoMP (Dagli et al., 2025), and the ground truth on representative GVM objects. Across objects with different geometries and material compositions, ViWi more closely recovers the spatial distributions and magnitudes of Young's modulus, Poisson's ratio, and density. The improvements are particularly visible in regions where VoMP assigns incorrect mechanical-property values over spatially coherent regions of the object. These additional examples further support the ability of ViWi to produce more accurate and spatially consistent volumetric mechanical-property fields.

What the material slots recover. Figure 8 compares the ground-truth material regions with the material-slot assignments produced by ViWi†. Each voxel is colored according to its highest-weight voxel-to-slot assignment, with colors matched to the corresponding ground-truth material regions. For the armchair (left), the dominant slot recovers the fabric shell, which accounts for 98.4% of the occupied voxels, while a second slot isolates the metal feet, which comprise only 1.6% of the object and achieve an IoU of 0.85. Despite their small spatial extent, the metal components differ substantially from the fabric in their mechanical properties.

The tricycle (right) exhibits a more complex material composition. ViWi† separates the metal structure (59.7% of voxels), rubber components (26.8%), and wooden components (13.5%), with slot IoUs of 0.69, 0.54, and 0.73, respectively. Importantly, individual slots can group spatially disconnected components that share a material identity, such as the separate metal feet of the armchair and the wooden components of the tricycle. This behavior is consistent with our formulation, where voxel-to-slot assignment is driven by visual and material-latent compatibility rather than spatial proximity.

The assignments are not intended to provide exact material segmentation. In the tricycle example, 7.1% of the voxels are assigned to a slot that does not match any ground-truth material region, and

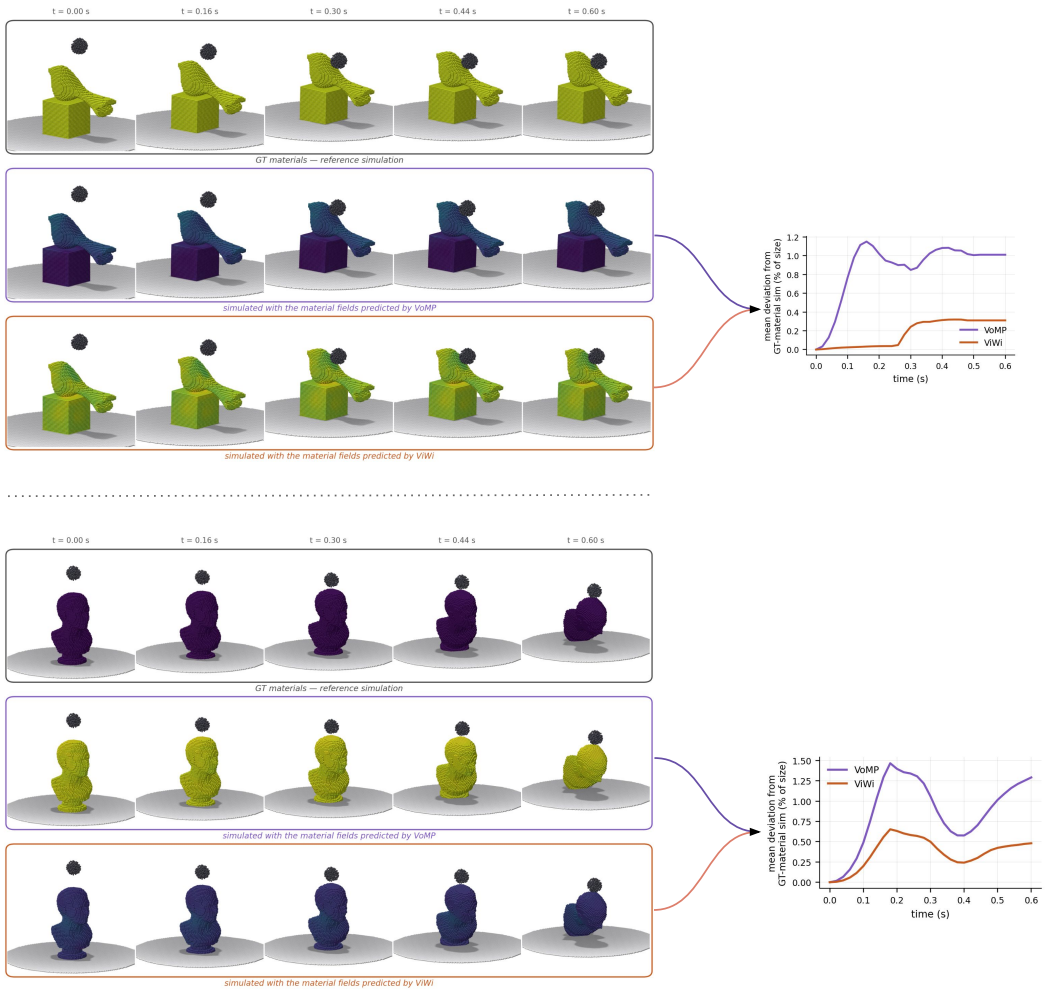


Figure 9: Impact simulation using predicted material fields. Two representative objects are simulated under identical rigid-ball impacts using the ground-truth material field, the field predicted by VoMP (Dagli et al., 2025), and the field predicted by ViWi. For each object, the left panel shows simulation snapshots over time, while the plot on the right reports the trajectory deviation from the ground-truth-driven simulation, normalized by object size. ViWi remains consistently closer to the reference simulation after contact.

some disagreement remains near part boundaries. Nevertheless, the visualizations show that the learned material slots recover meaningful material structure and aggregate voxels with a shared material identity.

Physics simulation with predicted material fields. Accurate estimates of (E, ν, ρ) are most useful when they lead to more accurate physical behavior. We therefore evaluate the predicted volumetric mechanical-property fields inside a simulator using two complementary scenarios: rigid-ball impact and quasi-static compression. In both cases, the same object is simulated three times using the ground-truth mechanical-property field, the field predicted by VoMP (Dagli et al., 2025), and the field predicted by ViWi. All rollouts share the same initial state and the same reduced basis, trained using a uniform material field, so that the mechanical-property field is the only prediction-specific quantity that changes between simulations. We measure the deviation of each prediction-driven trajectory from the ground-truth-driven trajectory using the mean pointwise distance over the simulated object, normalized by the object's spatial extent. Lower values indicate closer agreement with the reference simulation.

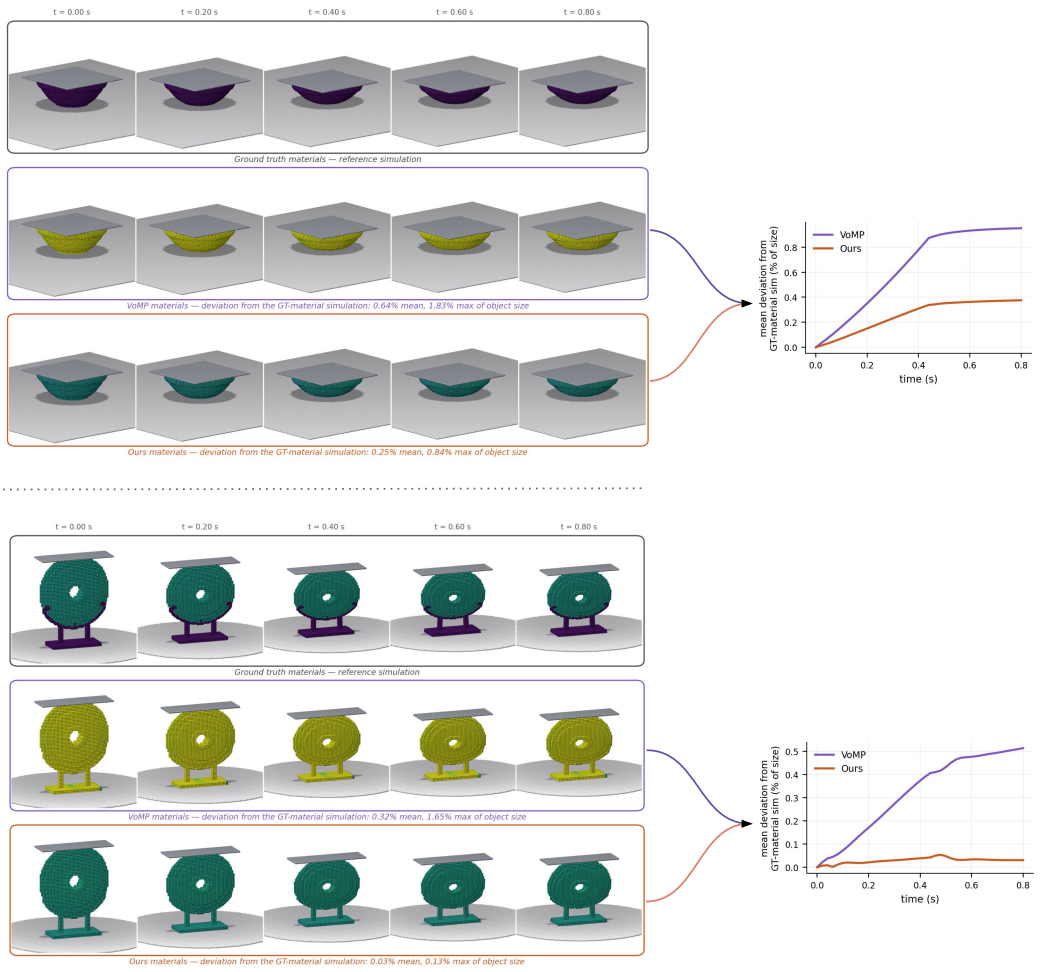


Figure 10: Quasi-static compression using predicted material fields. Two representative objects are compressed by the same prescribed rigid-plate motion using the ground-truth material field, the field predicted by VoMP (Dagli et al., 2025), and the field predicted by ViWi. The filmstrips show the resulting deformation over time, while the plots on the right report trajectory deviation from the ground-truth-driven simulation, normalized by object size. ViWi remains substantially closer to the reference simulation in both examples.

Impact simulation. We first consider a localized dynamic interaction in which a rigid ball is released above the object and allowed to strike it. Figure 9 shows two representative examples. For each object, the filmstrips compare the ground-truth-, VoMP-, and ViWi-driven simulations under identical impact conditions, while the plot on the right reports the deviation from the ground-truth-driven trajectory over time. Before contact, the trajectories remain nearly identical because the object has not yet responded to the applied impact. After contact, the trajectories diverge according to the predicted mechanical-property fields. In both examples, the simulation driven by ViWi remains consistently closer to the reference trajectory than the simulation driven by VoMP. These results show that improved volumetric mechanical-property estimation can translate into more faithful downstream physical dynamics under localized impact.

Quasi-Static Compression. To complement the transient impact test, we further evaluate the predicted mechanical-property fields under global displacement-controlled compression. A rigid plate descends onto each object until a prescribed compression of 30% of the object's height is reached. As in the impact experiment, the scenario is simulated using the ground-truth, VoMP, and ViWi mechanical-property fields under the same initial state and reduced basis. The same prescribed



Figure 11: Qualitative mass estimation on ABO-500. Representative real objects from the ABO-500 test set are shown together with their ground-truth mass (GT) and the mass estimates obtained from VoMP (Dagli et al., 2025) and ViWi†. ViWi† denotes the vision-only material-slot variant used for ABO-500 evaluation. Predicted masses are obtained from the estimated volumetric density fields.

plate motion is applied to all three rollouts, and a common global stiffness scale is applied identically across them, so that the comparison isolates differences in the predicted spatial mechanical-property fields.

Figure 10 shows two representative compression examples. For the bowl, the VoMP-driven simulation reaches a mean trajectory deviation of 0.64% and a peak deviation of 1.83% of object size, whereas the ViWi-driven simulation reduces these values to 0.25% and 0.84%, respectively. For the wheel-and-stand example, the corresponding deviations decrease from $0.32\%/1.65\%$ with VoMP to $0.03\%/0.13\%$ with ViWi. ViWi also more closely recovers the spatial stiffness structure, including the softer support beneath the stiffer wheel.

Together, the impact and compression experiments demonstrate that the mechanical-property fields predicted by ViWi produce physical behavior that more closely follows the ground-truth-driven simulation under both localized dynamic loading and global quasi-static deformation.

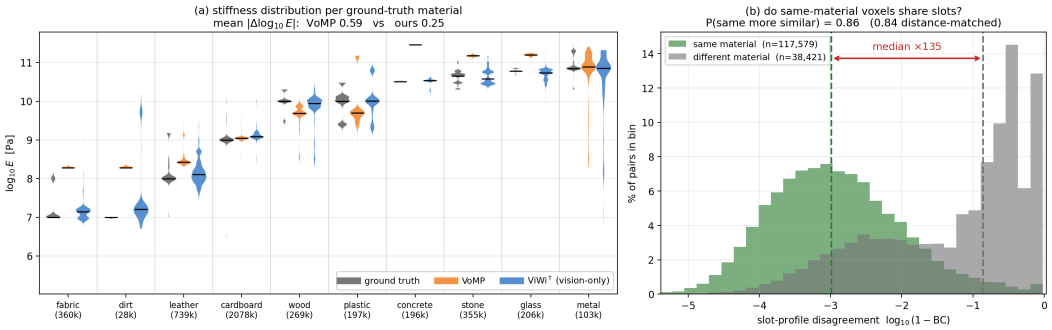


Figure 12: Material consistency of the learned material slots. **(a)** Distribution of $\log_{10} E$ per ground-truth material category. ViWi† tracks the ground truth more closely than the voxel-wise VoMP baseline (mean $|\Delta \log_{10} E|$: $0.59 \rightarrow 0.25$). **(b)** Slot-assignment disagreement $1 - \mathrm{BC}$ for voxel pairs within the same object: same-material pairs share slots far more than different-material pairs (median $\times 135$), and this survives matching on voxel distance ($0.86 \rightarrow 0.84$), so slots capture material rather than proximity.

Qualitative Mass Estimation on ABO-500. Figure 11 provides qualitative examples of mass estimation on the ABO-500 test set. For each real object, we report the ground-truth mass together with the estimates obtained from VoMP and ViWi†, our vision-only material-slot variant. Across objects with different geometries, scales, and appearances, ViWi† often produces mass estimates closer to the ground truth than VoMP. These examples complement the quantitative results in Table 2 and illustrate the generalization of the predicted volumetric density fields from synthetic training data to real-world objects.

What the slots learn. Figure 12 looks inside the model in the vision-only setting, where the slot head is the only difference from VoMP: both use the same frozen encoder and the same MatVAE decoder, and neither uses RF.

Panel (a) groups test voxels by their ground-truth material and plots the predicted stiffness of each group. VoMP pushes soft and very stiff materials toward the middle of the training range, while ViWi† keeps them apart and predicts a tighter range within each material. Averaged over the ten materials with at least $5{,}000$ test voxels, the error in $\log_{10} E$ drops from 0.59 to 0.25; the largest gains are on fabric ($1.08 \rightarrow 0.27$), dirt ($1.34 \rightarrow 0.58$) and concrete ($0.94 \rightarrow 0.05$), and VoMP stays slightly ahead on cardboard and metal.

Panel (b) asks whether the slots group voxels by material, even though they never see a material label. Each voxel has an attention profile over the K slots, which is a distribution, so we compare two voxels with the Bhattacharyya coefficient $\mathrm{BC}(p, q) = \sum_k \sqrt{p_k q_k}$, the standard overlap between two distributions: $\mathrm{BC} = 1$ when the two voxels use exactly the same slots and 0 when they share none. We plot the disagreement $1 - \mathrm{BC}$, so small values mean two voxels were assigned to the same slots. For each of the 78 test objects made of more than one material, we take pairs of voxels from that object and split the pairs by whether the two voxels have the same ground-truth material. Voxels of the same material are assigned to almost the same slots, while voxels of different materials are not: the median disagreement is 1.0×10^{-3} for same-material pairs and 1.38×10^{-1} for different-material pairs, a factor of 135. Comparing one random pair of each type, the same-material pair is the more similar one 86% of the time. This is not an artifact of nearby voxels being smoothed together: repeating the comparison using only pairs that are equally far apart still gives 84%.