Physically Guided Visual Mass Estimation from a Single RGB Image

Sungjae Lee[1] , **Junhan Jeong**[1] , **Yeonjoo Hong**[1] and **Kwang In Kim**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, South Korea
[2]Department of Electrical Engineering, POSTECH, South Korea
{leeeesj, junhanjeong, yeonjooh, kwangin.kim}@postech.ac.kr

Abstract

Estimating object mass from visual input is challenging because mass depends jointly on geometric volume and material-dependent density, neither of which is directly observable from RGB appearance. Consequently, mass prediction from pixels is ill-posed and therefore benefits from physically meaningful representations to constrain the space of plausible solutions. We propose a physically structured framework for single-image mass estimation that addresses this ambiguity by aligning visual cues with the physical factors governing mass. From a single RGB image, we recover object-centric three-dimensional geometry via monocular depth estimation to inform volume and extract coarse material semantics using a vision-language model to guide density-related reasoning. These geometry, semantic, and appearance representations are fused through an instance-adaptive gating mechanism, and two physically guided latent factors (volume- and density-related) are predicted through separate regression heads under mass-only supervision. Experiments on *image2mass* and *ABO-500* show that the proposed method consistently outperforms state-of-the-art methods.

1 Introduction

Robotic manipulation demands more than geometric perception. It requires an understanding of the physical principles that govern object interactions. Beyond shape and pose, physical properties such as mass, stiffness, and friction directly influence inertial effects, contact stability, and common failure modes. Accordingly, recent work has emphasized physically grounded reasoning, from detecting violations of real-world dynamics [Li *et al.*, 2025] to learning force-related intuition through human contact [Adeniji *et al.*, 2025]. Together, these efforts point to a central challenge: predicting physically relevant object attributes *before* interaction to enable safer and more robust manipulation.

Among these physical properties, *mass* plays a central role. Knowledge of an object's weight enables a system to regulate grasp forces, limit excessive torques, and avoid damage to fragile objects or mechanical components. Traditionally, mass is estimated through physical interaction using force and torque sensing, for example via a small test lift [Han *et al.*, 2020]. While effective, such contact-based approaches delay estimation until interaction has begun and may expose both the system and the object to unsafe forces, while also requiring specialized hardware or repeated exploratory actions that limit their practicality.

To address these challenges, prior work has explored non-contact methods that infer object mass directly from visual input. Existing approaches either regress mass end-to-end from RGB images using large pre-trained models [Trupin *et al.*, 2025], or introduce intermediate representations such as thickness masks and hand-crafted geometric features [Standley *et al.*, 2017]. Some methods explicitly decompose mass into volume and density estimates, reflecting the underlying physical relationship. However, in most existing pipelines, these factors remain tightly entangled within appearance-based RGB representations, making it difficult to isolate the respective contributions of geometry and material. As a result, predictions are often hard to interpret and susceptible to systematic biases when visual cues are ambiguous (e.g., visually similar objects made of different materials).

This limitation arises not only from single-view ambiguity, but also from how visual evidence is organized for physical inference. Because mass depends jointly on geometry and material, neither of which is explicitly represented in RGB, direct mass regression in pixel space is inherently ill-posed. Robust visual mass estimation therefore benefits from representations that make geometric structure and material-related cues more explicit, instead of relying solely on appearance correlations. While retrieval-based approaches can incorporate external metadata, they often fail in realistic settings where objects are novel, unmarked, or modified and identifiers may be unavailable.

Motivated by this perspective, we adopt a physically grounded formulation where mass is expressed as the product of two complementary latent factors: a geometry-related factor and a material-related factor. Rather than entangling these factors within a single appearance-driven predictor, we use geometric and semantic cues to *physically guide* their inference. Specifically, 3D geometry inferred via monocular depth estimation is used to inform volume estimation, while seman-

Extended version: https://arxiv.org/abs/2601.20303

tic material understanding extracted from a vision–language model (VLM) serves as a proxy for density-related reasoning. The resulting image, point cloud, and textual representations are fused using an instance-adaptive gated mechanism that learns the relative importance of each cue for a given object. This design reflects the fact that some objects are more geometry-dominated, whereas others are more strongly influenced by material-dependent density.

Experiments on the *image2mass* [Standley *et al.*, 2017] and *ABO-500* [Zhai *et al.*, 2024] datasets show that our single-image method significantly outperforms existing image-only and depth-augmented baselines.

2 Related Work

Recent work has increasingly incorporated physical reasoning into visual understanding beyond purely geometric or kinematic representations, including dynamics violation detection [Li *et al.*, 2025] and contact-based learning of force-related cues [Adeniji *et al.*, 2025]. More recently, VLMs have enabled high-level reasoning about material properties, surface characteristics, and physical feasibility. For example, *GraspCoT* [Chu *et al.*, 2025] employs VLM-driven inference of material and surface properties to support adaptive grasping under language instructions, while *PhysVLM* [Zhou *et al.*, 2025] models physical reachability to identify unreachable targets and predict task failure. [Guran *et al.*, 2024] construct tree-structured scene representations and infers object attributes, including material and mass-related cues, to support context-aware decision making. In this work, we focus on object mass as a concrete physical attribute and study how it can be inferred from visual input in a physically interpretable manner.

Single-image visual mass estimation. Estimating object mass from a single RGB image has been studied as a non-contact alternative to force- or interaction-based methods. *Image2mass* [Standley *et al.*, 2017] introduced this problem by decomposing mass estimation into intermediate geometric and material-related predictions, such as thickness and density, which are then combined to obtain mass. Subsequent work has explored end-to-end regression of mass directly from images, including approaches based on large pre-trained visual backbones [Trupin *et al.*, 2025]. While these methods demonstrate that mass can be predicted from appearance to some extent, they typically infer geometric and material cues implicitly within a single latent representation. This entanglement limits interpretability and can lead to systematic errors when visual appearance is ambiguous, for example for objects with similar shapes but different material compositions. Our approach differs in that it explicitly associates distinct sources of visual evidence with geometry-related and material-related inference, rather than entangling them within a single appearance-driven pipeline.

Geometry-aware representations for physical inference. A natural way to reduce ambiguity in single-image mass estimation is to introduce explicit geometric structure. Several works have shown that geometry-aware representations improve physical inference from visual input. In particular, depth maps, voxel grids, and other volumetric representations have been used to refine thickness or volume estimation and improve robustness under real-world imaging conditions [Andrade and Moreno, 2023; Cardoso and Moreno, 2025], demonstrating clear advantages over RGB features alone. Our method follows this perspective by using monocular depth estimation to obtain object-centric 3D geometry, which is explicitly aligned with geometry-related inference rather than treated as an auxiliary cue.

Semantic priors and VLMs. Beyond geometry, semantic understanding provides complementary cues for physical reasoning, particularly for properties governed by material composition. Recent advances in VLMs have shown that large-scale image–text pre-training encodes broad commonsense knowledge about materials and physical attributes. Several works exploit this capability to associate visual observations with physically meaningful concepts without requiring paired supervision for numerical physical quantities [Zhai *et al.*, 2024; Shuai *et al.*, 2025]. However, directly prompting LLMs or VLMs to output numerical physical parameters is often unreliable and susceptible to hallucination [Liao *et al.*, 2024; Chen *et al.*, 2024; Shao *et al.*, 2025]. In contrast, we use VLMs only to extract coarse semantic descriptions of material composition, which serve as informative priors rather than direct numeric estimates. These semantic cues are embedded and fused with visual and geometric representations under mass-only supervision.

3 Physics-Guided Visual Mass Estimation

3.1 Problem Formulation and Core Principle

Given an RGB image I_o depicting an object o, our goal is to estimate its mass m. A physically grounded starting point is the relation

$$m = V \cdot \rho, \tag{1}$$

where V denotes object volume and ρ denotes material-dependent density. Since our model is trained with *mass supervision only*, the predicted quantities $\widehat{V}$ and $\hat{\rho}$ should be interpreted as *physically guided latent factors* rather than metrically calibrated estimates of true volume and density, as any rescaling $(\alpha\widehat{V}, \hat{\rho}/\alpha)$ yields the same mass. Moreover, exploiting this decomposition in a single-view setting is inherently challenging: volume depends on 3D geometry and scale, which are underconstrained in monocular images, while density depends on material composition, which is often visually ambiguous and requires semantic reasoning. As a result, multiple (V, ρ) combinations can explain the same observation, making direct mass prediction ill-posed.

Our central idea is to resolve this ambiguity by explicitly aligning each latent physical factor with a source of visual evidence that is well suited to inform it. We associate the *volume-related factor* with 3D geometric cues recovered via monocular depth estimation, and density-related reasoning with semantic material information inferred by a vision–language model (VLM). Rather than treating these cues as interchangeable inputs, we enforce a structured inductive bias where geometry primarily constrains volume, se-

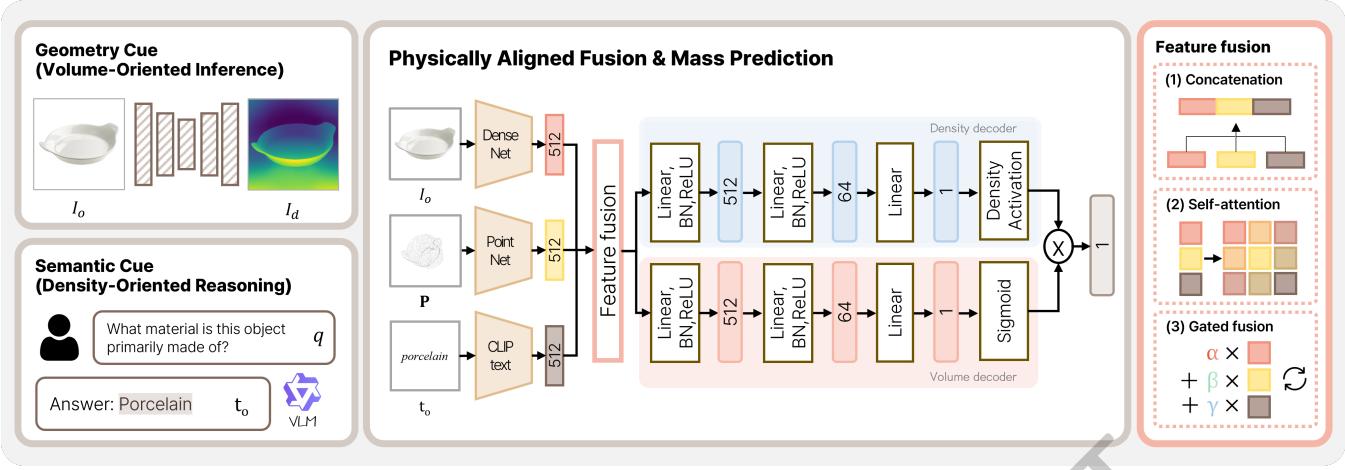

Figure 1: Overview of the proposed physically structured visual mass estimation framework. From a single RGB image, the model infers appearance features, object-centric geometry via monocular depth estimation, and a coarse semantic description of material composition. Geometry and semantic cues are explicitly aligned with volume- and density-related factor inference, respectively, while appearance provides complementary context. These representations are fused using an instance-adaptive mechanism to predict object mass. The framework is trained using mass supervision only and avoids the need for explicit density or volume labels.

mantics constrains density, and appearance provides complementary context. The interaction between these factors is resolved through instance-adaptive fusion, allowing the model to emphasize different cues depending on the object.

This physically structured design yields an interpretable mass estimation framework that reflects the underlying physics in a single-image setting.

3.2 Overview of the Multi-Cue Inference Pipeline

The proposed framework instantiates the above principle using three complementary cues extracted from a single RGB image I_o: an appearance feature, an explicit geometric representation $\mathbf{P} \in \mathbb{R}^{N\times 3}$ obtained by lifting monocular depth into a point cloud, and a semantic representation t_o describing the dominant material. These cues are fused into a joint embedding to predict a volume-related factor $\widehat{V}$ and a density-related factor $\hat{\rho}$, yielding the final mass estimate $\widehat{m} = \widehat{V}\hat{\rho}$. An overview is provided in Figure 1.

3.3 Geometry Cue for Volume-Oriented Inference

Accurate volume-oriented inference benefits from access to 3D geometric structure, which is not explicitly encoded in RGB appearance alone. While an RGB image captures rich information about texture, color, and shading, these cues are only weakly correlated with absolute object size and are often entangled with lighting conditions, viewpoint, and surface reflectance. As a result, objects with similar two-dimensional projections or visual appearance can differ substantially in physical size, leading to large errors in volume and mass estimation when relying directly on appearance cues.

Importantly, this ambiguity arises from inherent single-view uncertainty and from how visual evidence is *represented* for physical inference. Although all cues in our framework are ultimately inferred from a single RGB image, the representation used to organize visual evidence strongly influences which physical properties can be recovered reliably. Monocular depth estimation mitigates this issue by reorganizing appearance cues into an explicit geometric form, enforcing consistency in surface ordering and object extent along the camera ray. This 3D representation captures object shape and relative spatial extent more directly, making it better suited for volume-oriented reasoning.

While depth cameras can provide direct geometric measurements, their deployment is often constrained in real-world settings due to increased hardware cost and failures on reflective or transparent surfaces. Accordingly, we assume the availability of RGB sensors only and rely on monocular depth estimation to recover object-centric geometry.

Specifically, we predict a pixel-wise depth map I_d from the input image I_o following [Cardoso and Moreno, 2025]. We fine-tune the *GLPDepth* model [Kim *et al.*, 2022] on the ShapeNetSem dataset [Chang *et al.*, 2015] to better align the depth estimator with isolated object instances. This step is necessary because off-the-shelf monocular depth models are typically trained on indoor or outdoor datasets rather than object-centric imagery [Yang *et al.*, 2024b; Ke *et al.*, 2024], and consequently often exhibit imprecise object boundaries, background leakage, or inconsistent global depth scales when applied to single-object images. Since ShapeNetSem provides object-level samples with ground-truth depth, fine-tuning on this dataset allows the depth estimator to better match the data distribution encountered in the downstream mass estimation task.

To improve cross-instance scale consistency, the predicted depth values are normalized by the diagonal length of the object's bounding box. The normalized depth map is then converted into a point cloud $\mathbf{P} \in \mathbb{R}^{N\times 3}$, which serves as an explicit geometric representation for subsequent volume-oriented inference and mass estimation.

3.4 Semantic Cue for Density-Oriented Reasoning

Accurate density-related inference depends primarily on an object's material composition rather than its geometric shape. Objects with similar volumes can differ substantially in mass due to differences in material (e.g., plastic versus metal) making density a critical factor for mass prediction. However, unlike geometry, density is not directly observable from spatial structure and cannot be recovered reliably from pixel-level appearance alone.

As with volume estimation, this challenge arises from representational limitations rather than missing information. Visual appearance provides indirect cues about material, such as texture and surface finish, but these cues are often ambiguous and shared across materials with vastly different densities. Learning a direct mapping from RGB appearance to numerical density is therefore difficult, particularly in the absence of large-scale, accurately annotated datasets. Indeed, obtaining paired visual observations and ground-truth physical properties such as density is costly and rarely feasible at scale, as it requires manual measurement or specialized instrumentation.

Recent work has shown that VLMs encode broad commonsense knowledge about the physical world acquired through large-scale image-text pre-training. Representative examples such as *NeRF2Physics* [Zhai *et al.*, 2024] and *PUGS* [Shuai *et al.*, 2025] demonstrate that VLMs can associate visual appearance with physically relevant concepts without relying on explicitly paired image–property supervision. This capability provides an opportunity to incorporate high-level physical reasoning into visual inference pipelines without requiring direct numerical prediction of physical quantities.

We query the *Qwen2.5-VL* model [Yang *et al.*, 2024a] with the prompt q *"What is this object primarily made of?"* to obtain a coarse semantic description $t_o = \mathrm{VLM}(I_o, q)$ of the object's material composition, which provides a weak but informative cue for density-related inference.

Crucially, we do not interpret this textual output as a direct or precise estimate of density. Instead, it constrains the range of plausible densities by encoding coarse physical differences, thereby stabilizing density-oriented inference and improving interpretability. By operating at a semantic level, this representation avoids the numerical instability and hallucination issues that can arise when language models are tasked with predicting exact physical values.

3.5 Fusion and Mass Prediction

Having obtained geometry- and semantic representations aligned with the volume- and density-related factors, respectively, the remaining challenge is to integrate these heterogeneous cues into a coherent mass prediction. This integration should preserve their distinct physical roles, while remaining flexible enough to accommodate object-dependent variation in which factors dominate mass.

Modality-Specific Encodings

We encode appearance, geometry, and semantic cues using separate modality-specific networks. The RGB image I_o is encoded with *DenseNet-121* [Huang *et al.*, 2017]. The point cloud $\mathbf{P}$ is encoded using *PointNet* [Qi *et al.*, 2017], preserving geometric structure relevant to volume estimation. The semantic description t_o is embedded using a frozen *CLIP* text encoder [Radford *et al.*, 2021]. Each encoder outputs a 512-dimensional feature vector, followed by Layer Normalization. During training, the image and geometry encoders are fine-tuned, while the text encoder remains fixed.

Instance-Adaptive Fusion

Although volume and density arise from different physical factors, their relative influence varies across objects. For example, large objects are often dominated by geometric extent, whereas for smaller objects, differences in physical composition may have a greater effect. To capture this variability, we adopt an instance-adaptive fusion strategy that modulates the contribution of each modality on a per-object basis.

We evaluate three fusion mechanisms. The first **concatenates** modality-specific features into a single 1,536-dimensional representation, serving as a strong yet computationally efficient baseline. The second applies **self-attention** to explicitly model cross-modal interactions. Our primary design employs a **gated fusion** mechanism that predicts scalar weights for each modality and aggregates features through a weighted combination. This formulation allows the model to emphasize geometry, semantics, or appearance depending on which cues are most informative for a given object, while maintaining a clear correspondence between modalities and their physical roles.

Volume and Density Regression

From the fused representation, we predict a volume-related factor $\widehat{V}$ and a density-related factor $\hat{\rho}$ using two dedicated regression heads with identical architectures but distinct output constraints. This reflects the fact that volume and density arise from different physical factors and should be modeled separately rather than conflated in a single predictor.

The density head is implemented as a lightweight multilayer perceptron that outputs a scalar estimate $\hat{\rho}$. We apply the custom activation function of [Cardoso and Moreno, 2025], which matches the empirical density distribution of the *image2mass* dataset (see Section 4) and promotes physically plausible density-factor predictions, stabilizing optimization and suppressing unrealistic values, particularly in early training stages. The volume head applies a Sigmoid activation to ensure non-negative volume estimates.

The final mass prediction is obtained as $\widehat{m} = \widehat{V}\hat{\rho}$ (see Eq 1), and training is supervised solely using ground-truth mass. We employ the Absolute Log Difference Error (ALDE) loss on mass, which compares predictions on a logarithmic scale, reducing sensitivity to outliers while remaining effective across the range of object masses.

4 Experiments

4.1 Experimental Setting

Datasets

Experiments are conducted on two benchmarks for visual mass estimation. The *image2mass* dataset [Standley *et al.*, 2017] is a large-scale benchmark constructed from Amazon product listings and contains 141,950 training instances and 1,435 test instances, each annotated with an RGB image, object mass, and bounding box. We also evaluate

on the *Amazon-Berkeley Objects (ABO)* dataset [Collins *et al.*, 2022], which provides calibrated multi-view images and ground-truth mass for real objects. We adopt the *ABO-500* subset proposed by [Zhai *et al.*, 2024], which removes redundant instances across object classes. To ensure a fair comparison under a single-image setting, we follow the view selection protocol of [Zhai *et al.*, 2024] and use one view per object.

Evaluation Metrics

Following [Standley *et al.*, 2017], we report Absolute Log Difference Error (ALDE), Absolute Percentage Error (APE), Minimum Ratio Error (MnRE), and the q-metric (off by a factor of 2). For *ABO-500*, we also report Absolute Difference Error (ADE). See the supplemental material for definitions.

Baselines

We compare against three classes of non-contact, vision-based methods. All learned regression models (RGB, RGB+Depth, and ours) are trained for 25 epochs using Adam with learning rate 1×10^{-4}. VLM-based methods are evaluated in a zero-shot prompting setting.

RGB-based regressors estimate mass directly from a single RGB image. Our *RGB baseline* employs a frozen *ResNet-50* backbone pretrained on ImageNet, followed by three trainable fully connected layers. A Softplus activation is applied to enforce positive-valued predictions. The *image2mass* algorithm [Standley *et al.*, 2017] predicts a thickness mask along with 14 auxiliary geometric features extracted from a single RGB image using an *Xception* backbone [Chollet, 2017], from which object density and volume are inferred to compute the final mass.

RGB+Depth methods incorporate explicit geometric cues by fusing RGB features (*DenseNet*) with pseudo point clouds (*PointNet*) through fully connected layers, following [Cardoso and Moreno, 2025].

VLM-based models include two open-source VLMs, *Qwen2.5-VL* [Yang *et al.*, 2024a] and *LLaVA* [Liu *et al.*, 2023],[1] evaluated under two prompting strategies: *direct* prediction, which outputs mass end-to-end, and *reasoning*, which first infers material properties and conditions mass prediction on them.

4.2 Main Results

Results on *image2mass*

Table 1 summarizes the results. Several clear trends emerge. First, VLM-based approaches perform substantially worse than all learned regression models. Second, purely RGB-based methods achieve limited accuracy, while incorporating geometric information leads to consistent improvements. Third, our proposed multi-cue model achieves the best performance across all metrics.

VLM-based prediction. Both *LLaVA* and *Qwen-VL* perform poorly, especially under direct prompting, exhibiting extremely large APE values and low MnRE and q-scores. These failures often arise from degenerate predictions (e.g., near-zero[2] or unrealistically large masses), indicating that current VLMs are unreliable numerical estimators of physical quantities. Stepwise reasoning improves stability but remains far inferior to learned regressors.

Method	ALDE (↓)	APE (↓)	MnRE (↑)	Q (↑)
VLM-based.				
LLaVA (*direct*)	1.944	63.167	0.302	0.245
LLaVA (*reasoning*)	1.833	54.008	0.361	0.325
Qwen-VL (*direct*)	2.133	46.618	0.402	0.398
Qwen-VL (*reasoning*)	1.074	39.550	0.496	0.478
RGB-based.				
RGB baseline	0.843	1.459	0.517	0.514
image2mass	0.655	1.066	0.579	0.614
RGB+Depth.				
Cardoso and Moreno	0.567	4.886	0.615	0.686
Ours (*concat.*)	<u>0.528</u>	<u>0.681</u>	<u>0.639</u>	<u>0.717</u>
Ours (*self-attn.*)	0.560	0.722	0.627	0.694
Ours (*gated fusion*)	**0.519**	**0.665**	**0.647**	**0.722**

Table 1: Comparison of mass prediction performance on the *image2mass* test set. **Bold**: best; <u>underlined</u>: second-best performance.

Monocular RGB vs. geometry-aware models. The RGB baseline achieves limited accuracy, reflecting the difficulty of inferring mass from appearance representations alone. The *image2mass* model offers only modest gains despite architectural complexity. In contrast, incorporating depth-derived geometry [Cardoso and Moreno, 2025] improves ALDE, MnRE, and q, highlighting the importance of geometric volume estimation.

Our multi-cue model. Our framework, which jointly models geometry and material semantics, consistently outperforms all baselines. Even simple concatenation surpasses prior work, indicating that the gains stem from cue complementarity rather than fusion complexity. Gated fusion yields the best results. The learned modality weights (image: 0.13, geometry: 0.49, text: 0.36) further confirm that geometry and material dominate mass estimation.

Results on *ABO-500*

Table 2 reports results. As on *image2mass*, VLM-based methods perform poorly, reflecting their reliance on unstable numerical prediction rather than physically grounded inference. In contrast, prior learning-based pipelines exhibit more stable behavior by implicitly or explicitly modeling geometric and material factors.

Among all *single-image* methods, our model achieves the best overall performance. We also report the multi-view *NeRF2Physics* pipeline, which reconstructs full 3D geometry from 30 RGB views, as a reference upper bound. Despite operating on only one view, our method remains competitive with this multi-view system and even outperforms it in terms of ADE and APE.

4.3 Generalization and Robustness

To assess whether the observed performance gains reflect transferable physical reasoning rather than overfitting to the

[1] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

[2] VLMs often produce degenerate outputs such as `Mass=0.0000...`

Method	ADE (↓)	ALDE (↓)	APE (↓)	MnRE (↑)
VLM-based.				
LLaVA (*direct*)	12.083	3.157	2.312	0.280
LLaVA (*reasoning*)	30.588	1.781	12.283	0.368
Qwen-VL (*direct*)	21.871	2.080	15.191	0.430
Qwen-VL (*reasoning*)	76.231	1.336	4.223	0.436
RGB-based.				
RGB baseline	15.431	1.609	14.459	0.362
image2mass	12.496	1.792	0.976	0.341
RGB+Depth.				
Cardoso and Moreno	14.231	1.128	2.117	0.436
Ours (*concat.*)	10.161	1.034	1.639	0.407
Ours (*self-attn.*)	**7.530**	1.019	**0.900**	0.458
Ours (*gated fusion*)	8.321	**0.998**	0.980	**0.472**
Multi-view input.				
NeRF2Physics	8.730	0.771	1.061	0.552

Table 2: Performance comparison on the *ABO-500* test set.

Category	ALDE (↓)	APE (↓)	MnRE (↑)	Q (↑)	Count
Seen	0.485	0.543	0.656	0.747	355
Unseen	0.531	0.708	0.634	0.712	1080
Total	0.519	0.665	0.647	0.722	1435

Table 3: Performance on seen and unseen test categories.

training distribution, we analyze category overlap between the training and test sets. Only 24.7% of the test categories are seen during training, while the remaining 75.3% correspond to entirely novel object categories. Moreover, even within seen categories, test instances differ substantially in appearance, geometry, color, and scale, making generalization nontrivial (see the supplemental material for examples).

Despite this limited overlap and high intra-category variability, our model exhibits strong robustness. As shown in Table 3, performance on unseen categories degrades only moderately relative to seen ones and follows similar trends across all metrics, suggesting that the model is not simply memorizing category-specific statistics but instead learning representations that generalize across object types.

4.4 A Study of Material-Guided Density Factors

A central challenge in visual mass estimation lies in modeling the influence of material-dependent density from visual input. Prior work often relies on VLMs to directly predict numerical density values, but such predictions can be unstable and physically inconsistent, especially when geometry is incomplete in a single-view setting. In contrast, our approach uses VLMs only to infer coarse material semantics (e.g., metal, wood, plastic), which are then interpreted by a learning-based model rather than treated as density values themselves. These semantic cues serve as soft physical priors that guide density estimation while avoiding brittle numeric prediction.

To better understand this design choice, we compare

Method	ALDE (↓)	APE (↓)	MnRE (↑)	Q (↑)
CLIP encoder				
**NeRF2Physics*	1.319	6.686	0.387	0.255
Rule-based	1.301	6.180	0.390	0.249
ULIP encoder				
**NeRF2Physics*	1.323	6.424	0.386	0.324
Rule-based	1.294	6.165	0.392	0.333
Ours	0.519	0.665	0.647	0.722

Table 4: Performance comparison using alternative density representations. * denotes variants operating on a 3D point cloud reconstructed from a single RGB view.

against two alternative density representations. First, we evaluate direct numerical density prediction by adapting *NeRF2Physics* to our single-view setting (denoted **NeRF2Physics* in Table 4), where a reconstructed point cloud is provided in place of its original multi-view NeRF representation. This isolates the effect of VLM-based numeric density prediction under the same geometric constraints as our method.

Unlike the original *NeRF2Physics* pipeline, which relies on full 3D reconstruction from multiple views, our setting uses a point cloud reconstructed from a single RGB image. As a result, point-level features can be sparse and unreliable due to incomplete geometry and limited observations. To mitigate this issue, we also evaluate *ULIP-2* [Xue *et al.*, 2024], a multimodal encoder jointly trained on images, text, and point clouds, which provides richer and more robust geometric–semantic representations.

Second, we consider a rule-based density baseline derived from standard physical tables. Predicted material categories are mapped to canonical density ranges from the literature, and their mean values are used as deterministic density proxies. This enforces physically plausible values but removes the ability to adapt density to object-specific context.

Results are reported in Table 4. Direct VLM-based density prediction via **NeRF2Physics* performs worst, highlighting the difficulty of obtaining reliable numeric density estimates under incomplete single-view geometry. The rule-based approach improves stability but remains limited by coarse category-level averaging. In contrast, our method achieves the best performance by using material semantics as learnable priors and allowing the model to infer how they map to density-related factors in context.

Both VLM-based and rule-based methods also struggle with materials that exhibit high intra-class variability (e.g., plastic, wood, and fabric), where a single scalar cannot capture the wide range of possible physical densities. While representing density as a range can partially reflect this uncertainty, downstream mass estimation ultimately requires collapsing the range into a single value, and such averages can deviate substantially from true object densities. These findings further motivate our use of coarse semantic material descriptions combined with learning-based density inference, rather than direct numeric density prediction.

Configuration			ALDE (↓)	APE (↓)	MnRE (↑)	Q (↑)
Image	Density	Volume				
✓	✗	✗	0.779	1.199	0.546	0.570
✗	✓	✗	1.062	2.001	0.437	0.390
✗	✗	✓	0.641	0.887	0.587	0.615
✓	✓	✗	0.742	1.245	0.562	0.589
✓	✗	✓	0.545	0.587	0.633	0.688
✗	✓	✓	0.591	0.733	0.608	0.663
✓	✓	✓	0.519	0.665	0.647	0.722

Table 5: Ablation study evaluating the contribution of individual components to overall mass estimation performance.

4.5 Ablation Studies

Table 5 quantifies the contribution of each cue, where all inputs are ultimately derived from the same RGB image. Among single-cue variants, the geometry/volume-only model performs best, highlighting that explicit 3D structure provides the most reliable signal for mass estimation. In contrast, using material semantics alone yields the weakest performance, reflecting the inherent difficulty of inferring density-related information without geometric context.

Combining density with either appearance or geometry consistently improves results, and jointly modeling both density- and volume-related factors yields a substantial gain, confirming that the two cues provide complementary information even under a single-view setting. Adding appearance further refines the prediction, and the full model integrating Image, Density, and Volume achieves the best performance across all metrics. Overall, these results support our core design choice: explicitly separating and fusing physically distinct cues leads to more accurate and robust mass estimation.

We further test two practical variants: richer material prompts and images with unconstrained backgrounds.

Multi-Material Semantic Descriptions

Objects in everyday scenes are often composed of multiple materials. To test whether richer material cues improve density-oriented inference, we replace the single-material prompt with a multi-material prompt that asks the VLM to enumerate all visible materials, following [Zhai *et al.*, 2024]. As shown in Table 6, this modification does not consistently outperform the single-material variant. We attribute this to three factors: 1) an object's mass is often dominated by a single material [Zhai *et al.*, 2024], 2) our current formulation predicts a single scalar density-related factor, which is insufficient to model mixtures of materials, and 3) longer VLM outputs may be harder to encode reliably with CLIP due to its limited context capacity; long-context text encoders such as Long-CLIP [Zhang *et al.*, 2024] may alleviate this issue. We leave explicit multi-material modeling to future work.

Evaluation Beyond Segmented Benchmarks

Both *image2mass* and *ABO-500* provide segmented object images with uniform backgrounds, which simplifies preprocessing and reduces background-induced ambiguity. To assess performance under less controlled visual conditions, we additionally collect six household objects and capture ten images per object under varying viewpoints. Ground-truth masses are measured manually.

	ALDE (↓)	APE (↓)	MnRE (↑)	Q (↑)
Ours (*single*)	0.519	0.665	0.647	0.722
Ours (*multi*)	0.532	0.651	0.640	0.704

Table 6: Performance comparison between single-material and multi-material semantic descriptions on the *image2mass* test set.

GT	0.014	0.020	0.075
Ours	0.017	0.023	0.072
image2mass	0.065	0.059	0.105
RGB+Depth	0.056	0.036	0.031

Figure 2: Qualitative comparison of our method with *image2mass* [Standley *et al.*, 2017] and the RGB+Depth approach [Cardoso and Moreno, 2025] on household-object images. Values are masses (kg).

Since our implementation (as well as the compared baselines) assumes object-centric inputs, we preprocess each image using zero-shot text-driven segmentation to isolate the target object. Specifically, we use *Lang-SAM*, which integrates *GroundingDINO* [Liu *et al.*, 2025] and *Segment Anything* [Kirillov *et al.*, 2023] (see the supplemental material for details). Representative examples are shown in Figure 2. Across diverse backgrounds, our approach yields the most accurate mass estimates among single-view baselines. Full results are provided in the supplemental material.

5 Conclusions

In this work, we present a physically structured framework for single-image visual mass estimation grounded in the relation $m = V \cdot \rho$, using separate geometry- and material-related latent factors. Instead of relying on implicit appearance correlations or brittle end-to-end regression, we guide inference using modality-specific cues: monocular depth for object-centric geometry (volume) and VLM-derived material semantics as a density-related prior. By operating on semantic embeddings rather than raw numerical outputs, our approach leverages VLM commonsense knowledge while mitigating the brittleness of direct numeric prediction.

Our instance-adaptive fusion mechanism further improves robustness by dynamically weighting geometry, appearance, and material cues in an object-dependent manner. Experiments on *image2mass* and *ABO-500* demonstrate that our approach achieves state-of-the-art accuracy among single-view methods and remains competitive with multi-view pipelines that rely on full 3D reconstruction.

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No. RS-2022-II220290, Visual Intelligence for Space-Time Understanding and Generation, and No. RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)) and by the National Research Foundation of Korea (NRF) grant (No. RS-2024-00337559), all funded by the Korean government (MSIT).

References

[Adeniji *et al.*, 2025] Ademi Adeniji, Zhuoran Chen, Vincent Liu, Venkatesh Pattabiraman, Raunaq Bhirangi, Siddhant Haldar, Pieter Abbeel, and Lerrel Pinto. Feel the force: Contact-driven learning from humans. *arXiv:2506.01944*, 2025.

[Andrade and Moreno, 2023] João Martinho Lopes Andrade and Plinio Moreno. Improving the estimation of object mass from images. In *IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 2023.

[Cardoso and Moreno, 2025] Ricardo Cardoso and Plinio Moreno. Estimating object physical properties from RGB-D vision and depth robot sensors using deep learning. In *Iberian Conference on Pattern Recognition and Image Analysis*, 2025.

[Chang *et al.*, 2015] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv:1512.03012*, 2015.

[Chen *et al.*, 2024] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[Chu *et al.*, 2025] Xiaomeng Chu, Jiajun Deng, Guoliang You, Wei Liu, Xingchen Li, Jianmin Ji, and Yanyong Zhang. GraspCoT: Integrating physical property reasoning for 6-DoF grasping under flexible language instructions. In *ICCV*, 2025.

[Collins *et al.*, 2022] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F. Yago Vicente, Thomas Dideriksen, Himanshu Arora, Matthieu Guillaumin, and Jitendra Malik. ABO: Dataset and benchmarks for real-world 3D object understanding. In *CVPR*, 2022.

[Guran *et al.*, 2024] Nurhan Bulus Guran, Hanchi Ren, Jingjing Deng, and Xianghua Xie. Task-oriented robotic manipulation with vision language models. *arXiv:2410.15863*, 2024.

[Han *et al.*, 2020] Yuanfeng Han, Ruixin Li, and Gregory S. Chirikjian. Can I lift it? Humanoid robot reasoning about the feasibility of lifting a heavy box with unknown physical properties. In *IROS*, 2020.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[Ke *et al.*, 2024] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024.

[Kim *et al.*, 2022] Doyeon Kim, Woonghyun Ka, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv:2201.07436*, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *CVPR*, 2023.

[Li *et al.*, 2025] Wenqiao Li, Yao Gu, Xintao Chen, Xiaohao Xu, Ming Hu, Xiaonan Huang, and Yingna Wu. Towards visual discrimination and reasoning of real-world physical dynamics: Physics-grounded anomaly detection. In *CVPR*, 2025.

[Liao *et al.*, 2024] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *EMNLP*, 2024.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[Liu *et al.*, 2025] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *ECCV*, 2025.

[Qi *et al.*, 2017] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Shao *et al.*, 2025] Jiandong Shao, Yao Lu, and Jianfei Yang. Benford's curse: Tracing digit bias to numerical hallucination in LLMs. *NeurIPS*, 2025.

[Shuai *et al.*, 2025] Yinghao Shuai, Ran Yu, Yuantao Chen, Zijian Jiang, Xiaowei Song, Nan Wang, Jv Zheng, Jianzhu Ma, Meng Yang, Zhicheng Wang, Wenbo Ding, and Hao Zhao. PUGS: Zero-shot physical understanding with Gaussian splatting. *arXiv:2502.12231*, 2025.

[Standley *et al.*, 2017] Trevor Standley, Ozan Sener, Dawn Chen, and Silvio Savarese. image2mass: Estimating the mass of an object from its image. In *CoRL*, 2017.

[Trupin *et al.*, 2025] Noah Trupin, Zixing Wang, and Ahmed H. Qureshi. Dynamic robot tool use with vision language models. *arXiv:2505.01399*, 2025.

[Xue *et al.*, 2024] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. ULIP-2: Towards scalable multimodal pre-training for 3D understanding. In *CVPR*, 2024.

[Yang *et al.*, 2024a] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv:2412.15115*, 2024.

[Yang *et al.*, 2024b] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *NeurIPS*, 2024.

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

[Zhang *et al.*, 2024] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-CLIP: Unlocking the long-text capability of CLIP. In *ECCV*, 2024.

[Zhou *et al.*, 2025] Weijie Zhou, Manli Tao, Chaoyang Zhao, Haiyun Guo, Honghui Dong, Ming Tang, and Jinqiao Wang. PhysVLM: Enabling visual language models to understand robotic physical reachability. In *CVPR*, 2025.