PhysGS: Bayesian-Inferred Gaussian Splatting for Physical Property Estimation

Supplementary Material

A. Full Bayesian and Uncertainty Formulation

A.1. Observation Model

For completeness, we describe the observation model used in *PhysGS*. Each observation corresponds to a segmented region of the scene and contains semantic (material) and physical (property) information extracted from the vision–language model (VLM). The role of these observations in the Bayesian updates is described in Sec. 3.

Observation tuple. For the m-th segmented region, we define

$$\mathcal{O}_m = \big(c_m,\ p_m,\ \psi_m\big), \tag{15}$$

where c_m is the predicted material class, p_m is the confidence produced by the VLM, and ψ_m is the predicted physical property (e.g., friction, hardness, stiffness, density). Across multiple views, the full set of observations is

$$Z = \{\mathcal{O}_1, \ldots, \mathcal{O}_M\}. \tag{16}$$

The tuple (c_m, p_m, ψ_m) constitutes a noisy measurement of the latent variables $(z_m, \mu_{z_m}, \sigma^2_{z_m})$. In particular, the predicted class c_m serves as a noisy proxy for the true (unobserved) material label z_m, while the VLM estimate ψ_m provides a noisy observation of the underlying material-specific physical property whose distribution is governed by $(\mu_{z_m}, \sigma^2_{z_m})$. These observed quantities supply the confidence-weighted evidence used in the Bayesian updates that follow.

A.2. Dirichlet–Categorical Posterior

The material fusion process follows the Dirichlet–Categorical formulation introduced in Sec. 3.1. The Categorical likelihood and Dirichlet prior correspond to Eqs. (1)–(2).

The posterior Dirichlet parameters update as in Eq. (5):

$$\tilde{\alpha}_i = \alpha_i(0) + \sum_{m:\, c_m = i} \lambda\, p_m. \tag{17}$$

The resulting posterior predictive distribution over material classes is given in Eq. (4).

A.3. Continuous Property Estimation

Continuous physical properties are fused using confidence-weighted running moments as introduced in Sec. 3.2 of the main paper. The accumulators W_i, S_i, and Q_i match Eq. (6) in the main paper.

The posterior mean and variance follow Eq. (7) in the main paper:

$$\mu_i = \frac{S_i}{W_i}, \qquad \sigma_i^2 = \max\!\left(\frac{Q_i}{W_i} - \mu_i^2,\ \epsilon\right). \tag{18}$$

This defines the Gaussian posterior $p(\psi_i \mid Z)$ shown in Eq. (8) of the main paper.

A.4. Mixture Formulation

Marginalizing over discrete material classes using the hierarchical model in Sec. 3.2 leads directly to the mixture distribution shown in Eq. (11), combining material probabilities with the class-conditional Gaussian property estimates.

A.5. Normal–Inverse–Gamma Posterior

We extend our continuous estimator with the Normal–Inverse–Gamma (NIG) prior introduced in Sec. 3.3. The joint prior over (μ_i, σ_i^2) matches Eq. (12).

Given a weighted observation (ψ_m, p_m), the closed-form posterior updates (Eqs. (13)–(14)) are:

$$\tilde{\kappa}_i = \kappa_i + p_m, \tag{19}$$

$$\tilde{\tau}_i = \frac{\kappa_i \tau_i + p_m \psi_m}{\kappa_i + p_m}, \tag{20}$$

$$\tilde{\alpha}_i = \alpha_i + \tfrac{p_m}{2}, \tag{21}$$

$$\tilde{\beta}_i = \beta_i + \frac{p_m \kappa_i (\psi_m - \tau_i)^2}{2(\kappa_i + p_m)}. \tag{22}$$

A.6. Predictive Uncertainty

The decomposition of predictive uncertainty into aleatoric and epistemic components follows Eq. (13). The predictive moments correspond directly to Eq. (14):

$$\mathbb{E}[\sigma_i^2] = \frac{\tilde{\beta}_i}{\tilde{\alpha}_i - 1}, \qquad \mathrm{Var}[\mu_i] = \frac{\mathbb{E}[\sigma_i^2]}{\tilde{\kappa}_i}. \tag{23}$$

Aleatoric uncertainty reflects inherent variability within a material class, while epistemic uncertainty captures uncertainty in the estimated mean due to limited or inconsistent evidence.

A.7. MMSE Estimate

As shown in Eq. (7), the posterior mean μ_i is the minimum mean-square-error (MMSE) estimator:

$$\hat{\psi}_i = \mu_i. \tag{24}$$

This corresponds to the property value that minimizes expected squared error and is therefore used as the single representative estimate for each material class.

Table 4. Stiffness estimation on MIT Fabric Properties dataset (30 objects). ADE is measured in lbf-in^2. **Bold**: best model.

Method	ADE (↓)	ALDE (↓)	APE (↓)	MnRE (↑)
GPT-4V	0.563	2.380	19.986	0.210
GPT-5	0.126	1.053	2.887	0.452
Ours	**0.040**	**0.725**	**1.338**	**0.553**

A.8. Full Probabilistic Model

The complete hierarchical model underlying *PhysGS* is summarized in Sec. 3 of the main paper and depicted in Fig. 2. For completeness, we restate the probabilistic structure:

$$\boldsymbol{\theta} \sim \text{Dirichlet}(\boldsymbol{\alpha}(0)), \tag{25}$$

$$z_m \sim \text{Categorical}(\boldsymbol{\theta}), \tag{26}$$

$$(\mu_i, \sigma_i^2) \sim \text{NIG}(\tau_i, \kappa_i, \alpha_i, \beta_i), \tag{27}$$

$$\psi_m \sim \mathcal{N}(\mu_{z_m}, \sigma_{z_m}^2). \tag{28}$$

This formulation provides the full probabilistic backbone through which *PhysGS* jointly infers materials, continuous properties, and calibrated uncertainty. First, a Dirichlet prior is placed over the material probabilities $\boldsymbol{\theta}$, reflecting initial uncertainty about the frequency of each material class. Each segmented region then draws a material label z_m from this Categorical distribution. For every material class i, the mean and variance of its physical property are modeled using a Normal–Inverse–Gamma (NIG) prior, which captures both uncertainty in the material's typical property value and its intrinsic variability. Finally, the observed physical property ψ_m for region m is sampled from the Gaussian distribution associated with its material label. Together, this hierarchy defines how materials and their continuous properties jointly generate the observations used in the Bayesian inference procedure.

B. Additional Results

B.1. Stiffness Estimation

Dataset. We employ the MIT Fabric Properties Dataset [2] for evaluating mass prediction, 30 different types of real fabric along with measurements of their material properties. Since these are all videos, we curate an image dataset from this, where all the different fabrics are evaluated for their bending stiffness. While these are video datasets, they are captured from a single view, and thus we evaluate our model on one image per fabric. We pick the first frame of every video.

Metrics. We report the same evaluation metrics for bending stiffness estimation (lbf-in^2) used in evaluation for mass estimation as above: ADE, ALDE, APE, MnRE.

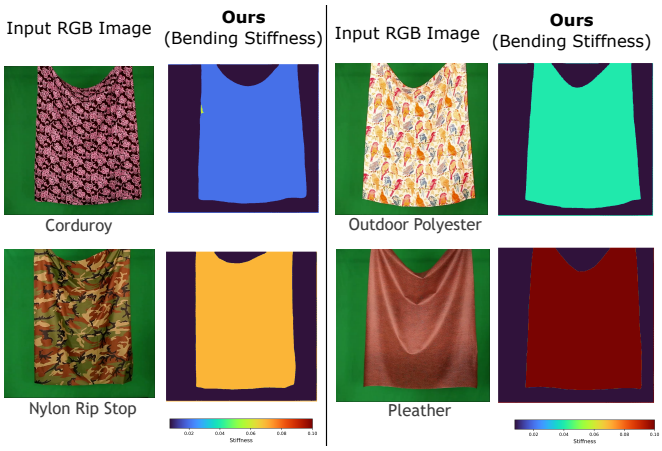

Figure 6. Bending stiffness estimation on real fabric samples from MIT Fabric Properties Dataset. Given an input RGB image, *PhysGS* produces dense stiffness fields that capture material differences across corduroy, nylon ripstop, outdoor polyester, and pleather.

Baselines. We compare our model against several visual and multimodal baselines on the ABO-500 dataset:

- **GPT-4V**: We provide GPT-4V with the image, and ask it to estimate the physical stiffness of the fabric.
- **GPT-5**: Same prompt as GPT-4V.

Quantitative Results. Table 4 reports quantitative results comparing our method against GPT-4V and GPT-5 VLM baselines. Across all metrics, *PhysGS* achieves the strongest performance, reducing ADE by 68.3% compared to GPT-5 and by more than an order of magnitude compared to GPT-4V. Our method also attains the highest MnRE score, indicating substantially improved scale consistency in stiffness estimation. These gains highlight the effectiveness of our Bayesian fusion framework in capturing fine-grained material compliance even in visually ambiguous textile structures.

Qualitative Results. Figure 6 presents qualitative bending stiffness estimation results on real fabric samples from the MIT Fabric Properties dataset. The dataset contains diverse materials with visually similar appearances but substantially different mechanical behavior, making stiffness prediction particularly challenging. Across a variety of textile types, including corduroy, nylon ripstop, outdoor polyester, and pleather, *PhysGS* produces dense stiffness fields that clearly delineate material differences. Each predicted stiffness map exhibits smooth spatial variation and preserves mask-level boundaries, reflecting the underlying compliance characteristics of each fabric.

B.2. Terrain Friction Estimation

Dataset. We evaluate terrain friction prediction using the Terrain Class Friction dataset from [10]. The dataset contains paired RGB images and friction measurements for

seven common indoor and outdoor terrain classes, including carpet, concrete, laminated flooring, rubber, pebbles, rocks, and wood (see Table 5). Following the protocol in [10], we assess prediction accuracy against the mean coefficients of friction obtained from their unimodal Gaussian fits for each terrain class.

Table 5. Mean (μ) and standard deviation (σ) of coefficients of friction reported in [10] for the Terrain Class Friction Dataset.

Terrain Class	μ	σ
Concrete	0.543	0.065
Pebbles	0.428	0.059
Rocks	0.478	0.113
Wood	0.372	0.055
Rubber	0.616	0.048
Carpet	0.583	0.068
Laminated Flooring	0.311	0.045

Metrics. We report the same evaluation metrics for terrain friction estimation (lbf-in^2) used in evaluation for mass estimation as above: ADE, ALDE, APE, MnRE.

Baselines. We compare our model against several visual and multimodal baselines on the ABO-500 dataset:

- **GPT-4V**: We provide GPT-4V with the image, and ask it to estimate the friction of the terrain.
- **GPT-5**: Same prompt as GPT-4V.

Quantitative Results. Table 6 reports quantitative results comparing our method against GPT-4V and GPT-5 VLM baselines. As the dataset consists of single-object, mostly homogeneous surfaces, the benefits of precise part-level segmentation are limited in this setting. Nevertheless, our hierarchical prompting scheme enables both global and local reasoning by guiding the VLM to focus on the dominant surface region while still incorporating contextual cues such as reflectance, roughness, and material structure. Across all four metrics, ADE, ALDE, APE, and MnRE, our method performs on par with or better than GPT-4V and GPT-5.

Qualitative Results. Figure 7 presents qualitative friction estimation results on samples from the Terrain Class Friction dataset. Given an input RGB image, *PhysGS* produces smooth and spatially consistent friction fields that align with the visual regions of each surface. The predicted maps clearly distinguish materials such as carpet, wood, and composite flooring, capturing their characteristic friction patterns while preserving coherent region boundaries.

B.3. Outdoor Scene Analysis

Figure 5 shows qualitative results of *PhysGS* applied to outdoor environments with diverse terrain types, vegetation, and natural materials. From a single RGB image, our model predicts material segmentation, friction coefficients, stiffness (Young's modulus) fields, and total uncertainty (aleatoric + epistemic). These results demonstrate the ability of *PhysGS* to extend beyond controlled indoor settings and operate on unstructured outdoor scenes.

Across all examples, the predicted material maps provide reasonable semantic decomposition of natural surfaces such as gravel, grass, bark, mud, water, and leaf litter. The corresponding friction and stiffness fields reflect meaningful physical differences between these materials: solid re-

Table 6. Friction estimation on Terrain Class Friction dataset (30 objects). ADE is measured in lbf-in^2. **Bold**: best model.

Method	ADE (↓)	ALDE (↓)	APE (↓)	MnRE (↑)
GPT-4V	0.129	0.315	0.286	0.747
GPT-5	0.146	0.253	0.291	0.779
Ours	**0.126**	**0.251**	**0.290**	**0.783**

Figure 7. Friction estimation on samples from the Terrain Class Friction dataset. *PhysGS* produces smooth, coherent friction maps that differentiate surfaces such as carpet, wood, and composite flooring directly from RGB input.

Figure 8. Imprecise masks generated by SAM as can be seen in the mask overlay and the part images. This results in less clear material boundaries and higher downstream uncertainty.

gions such as rock, concrete, or bark consistently receive higher stiffness values, whereas deformable surfaces such as mud and grass yield lower estimated moduli. Friction estimates likewise align with expected terrain properties, capturing transitions between slippery, saturated mud and higher-friction vegetation or gravel.

The total uncertainty maps reveal a strong correlation between uncertainty and the quality of SAM-generated segmentations, consistent with the discussion in the limitations section (see Sec. 5). Rows 2 and 3 in Figure 5 contain dense clutter, irregular textures, or ambiguous boundaries (e.g., intertwined vegetation or mud–grass transitions), leading SAM to produce noisier part-level masks. As illustrated explicitly in Figure 8, these mask inaccuracies propagate into the part images and result in less reliable material evidence. In such cases, *PhysGS* assigns noticeably higher total uncertainty, driven by both epistemic uncertainty from inconsistent material cues and aleatoric uncertainty arising from intra-region variability.

Conversely, rows 1 and 4 in Figure 5 contain large, spatially coherent surfaces (e.g., gravel, sky, uniform grass), where SAM produces cleaner segmentations. In these settings, *PhysGS* yields lower uncertainty and more stable physical predictions across the scene. Taken together, these results, supported by both Figures 5 and 8, demonstrate that the Bayesian uncertainty estimates are meaningfully sensitive to segmentation quality and reliably signal when the input evidence is less trustworthy.

C. Additional Experimental Details

C.1. Prompting Details

Figures 9 and 10 show the exact prompting configurations, inspired by [63], used for the MIT Fabric Properties dataset and the RUGD outdoor dataset. In both cases, the VLM is provided with the original RGB image, a segmentation-mask overlay, and an isolated part image. The text prompt directs the model to ignore masked regions and focus only

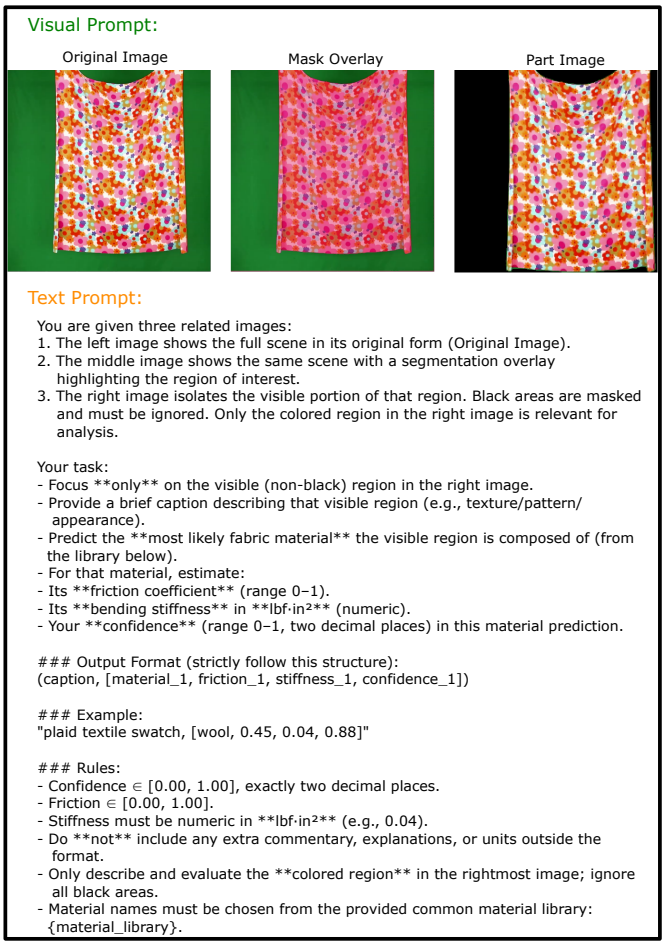

Figure 9. VLM Prompt used to obtain material, friction, and bending stiffness predictions for the MIT Fabric Properties dataset.

on the visible segment, ensuring that predictions are part-specific rather than influenced by the surrounding scene. We also maintain separate indoor and outdoor material libraries so the VLM selects from the most appropriate set of materials for each environment.

For each part, the VLM returns one or more candidate materials with associated physical properties and confidence scores. Each of these candidate predictions is treated as a confidence-weighted observation within our Bayesian framework, allowing PhysGS to fuse evidence across views and produce consistent material and property estimates. Importantly, the distribution of confidence across multiple materials provides a direct signal of semantic ambiguity. When the VLM is uncertain, often due to noisy or imprecise SAM segmentations, the confidence spread increases, which propagates into higher predictive uncertainty in our property fields, consistent with the trends discussed in the limitations section (Sec. 5).

Visual Prompt:

Original Image | Mask Overlay | Part Image

Text Prompt:

You are given three related images:
1. The left image shows the full outdoor scene in its original form (Original Image).
2. The middle image shows the same scene with a segmentation overlay highlighting the region of interest.
3. The right image isolates the visible portion of that region; **black areas are masked and must be ignored.** Only the colored region in the right image is relevant for analysis.

Your task:
- Focus **only** on the visible (non-black) region in the right image.
- Provide a brief caption describing that visible region.
- Predict the **three most likely materials** the visible region could be composed of, in descending order of likelihood.
- For each material, estimate:
- Its **friction coefficient** (0–1 range).
- Its **stiffness** (Young's Modulus in **GPa**).
- Your **confidence** (0–1, two decimal places) in that material prediction.

Output Format (strictly follow this structure):
(caption, [material_1, friction_1, stiffness_1, confidence_1], [material_2, friction_2, stiffness_2, confidence_2], [material_3, friction_3, stiffness_3, confidence_3])

Example:
"rough gravel surface, [gravel, 0.72, 30.0, 0.90], [asphalt, 0.68, 50.0, 0.74], [soil, 0.55, 15.0, 0.61])"

Rules:
- Each confidence value must be between 0.00 and 1.00 (two decimals).
- Each friction value must be between 0.00 and 1.00.
- Stiffness (Young's Modulus) must be given in **GPa**, numeric only (e.g., 50.0).
- Do **not** include any extra commentary, explanations, or text outside the specified format.
- Focus only on the **visible region** in the rightmost image; ignore background and masked (black) regions entirely.
- Material names must be chosen from the provided common material library: {material_library}.

Figure 10. VLM prompt used to obtain material, friction, and stiffness predictions for the RUGD dataset. By predicting multiple plausible materials with associated confidences, this prompting strategy enables *PhysGS* to estimate the total uncertainty for each mask.

C.2. Baseline Details

To benchmark *PhysGS* against existing vision–language models, we evaluate GPT-4V and GPT-5 on the MIT Fabric Properties and Terrain Class Friction datasets using a simplified prompting strategy tailored for fair comparison (see Figure 11). For each image, the VLM receives only the raw RGB frame and is instructed to (1) describe the dominant visible region, (2) predict the most likely material based solely on visual appearance, and (3) estimate a friction coefficient, stiffness value, and confidence score.

This baseline prompt does not include segmentation cues or part-based isolation, and therefore tests each VLM's ability to infer material and physical properties directly from appearance alone. The resulting predictions serve as a reference for evaluating the gains provided by our part-aware prompting, used in *PhysGS*.

Visual Prompt:

Text Prompt:

You are given an image of a scene or object.

Your task:
- Observe the image and identify the **most dominant material** visible.
- Provide a short caption describing the visible region (e.g., color, texture, or surface characteristics).
- Predict the **most likely material** based on visual appearance alone.
- Estimate:
- Its **friction coefficient** (range 0–1).
- Its **stiffness** (in appropriate units depending on the material type, e.g., GPa for solids or lbf·in² for fabrics).
- Your **confidence** (range 0–1, two decimal places) in this material prediction.

Output Format (strictly follow this structure):
(caption, [material_1, friction_1, stiffness_1, confidence_1])

Example:
"shiny metallic surface, [metal, 0.25, 200.0, 0.93]"

Rules:
- Confidence $\in$ [0.00, 1.00], exactly two decimal places.
- Friction $\in$ [0.00, 1.00].
- Stiffness must be numeric (e.g., 0.03 or 50.0).
- Do **not** include any commentary, units, or text outside the specified format.
- Base your answer entirely on the **visual cues** in the provided image (color, gloss, pattern, texture, reflectance).

Figure 11. Baseline VLM Prompt used to obtain material, friction, and bending stiffness predictions for the MIT Fabric Properties dataset.

C.3. Computational Overhead

Table 7. Runtime and peak VRAM usage for pipeline on one ABO-500 scene (30 images).

Module	Time (s)	VRAM (GB)
SAM	537.4	6.45
GPT-5 (server)	2386.3	–
3DGS	533.3	12.0
Bayesian inference	0.0002	$\sim$0
Mass estimation	0.6	0.57
Total	**3457.6**	**12.0**

Table 7 summarizes the runtime and peak VRAM usage for a single ABO-500 scene with 30 images. The overall runtime is dominated by SAM, the remotely executed VLM, and 3DGS. In contrast, peak VRAM consumption is primarily driven by SAM and 3DGS, with other components contributing negligibly.

C.4. Calibration Error

Table 8. Expected Calibration Error (ECE) for all datasets

Dataset	ECE ↓	Samples
Friction and Hardness (Shore hardness)	0.256	31
Friction and Hardness (Friction)	0.208	6
MIT Fabric Properties	0.135	30
Terrain Class Friction	0.293	7
Weighted Average	**0.207**	**74**

We quantitatively evaluate calibration (see Table 8) using Expected Calibration Error (ECE), defined as $\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$. Across datasets (with per-point ground truth labels), ECE ranges from 0.135–0.293 with a weighted average of 0.207, indicating moderate overconfidence consistent with prior VLM behavior. While imperfectly calibrated, confidence remains informative across properties and domains.