KNOCKGS: INTERACTION-GROUNDED CALIBRATION OF PHYSICAL GAUSSIAN REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Physics-integrated 3D Gaussian representations now allow reconstructed deformable objects to be simulated and rendered under explicit material models. Existing pipelines, however, assume that material parameters are known or manually specified, limiting their applicability when these parameters must be inferred from observed object dynamics. We propose **KnockGS**, an interaction-response PhysicalGS framework that estimates the elasticity and density scales of a 3D Gaussian object from its dynamics under a known applied force. We turn the force-induced response into a calibration signal: temporal response features are extracted from the observed dynamics, the two material scales are estimated from those features, and the estimate is then frozen and written back into the same simulator so that it can be tested on an interaction it was never fitted to. We evaluate the framework on both parameter recovery and response-level fidelity. The estimated scales are compared against hidden ground truth, and the re-simulated object is measured against the target using 3D particle trajectories, response-curve statistics, and rendered-frame quality. Across five held-out material targets, our method recovers the scales substantially more accurately than response retrieval, global regression, or a fixed default material, and the frozen estimate remains predictive under interactions that differ in direction and in magnitude. Interaction response therefore carries enough information to calibrate material scales in physically grounded 3D Gaussian representations. Our study is a first step toward interactive PhysicalGS systems that calibrate a Gaussian asset whose rendered appearance and simulated response are consistent. Code has been released: https://anonymous.4open.science/r/ANON-DCE3/.

1 INTRODUCTION

Three-dimensional visual reconstruction can now recover the geometry and appearance of real objects with high fidelity, yet it cannot answer a basic question: what happens to the object when a force is applied to it? When PhysicalGS is to serve as a foundational digital asset, i.e., the underlying representation in robotic manipulation, then its simulated physical behavior and its rendered appearance must agree. Robotic manipulation, Physical AI, digital twins, and interactive simulation all require this predictive capability, motivating a step beyond visual reconstruction toward *3D physical reconstruction*. On the visual side, neural radiance fields and their accelerations have reshaped novel-view synthesis (Mildenhall et al., 2020; Müller et al., 2022; Barron et al., 2022), while 3D Gaussian Splatting (3DGS) provides an explicit representation with real-time rendering (Kerbl et al., 2023). Physics-integrated extensions such as PhysGaussian (Xie et al., 2024) make Gaussian primitives simulatable by associating them with material points evolved by the Material Point Method (MPM) (Sulsky et al., 1994; Stomakhin et al., 2013; Jiang et al., 2015; Hu et al., 2018). Separate visual and physical representations must be converted into one another, kept in correspondence, and rendered consistently; a physics-integrated Gaussian asset avoids all three by unifying appearance and mechanical state. However, such pipelines are predominantly *forward*: the material model and its parameters must be specified before simulation can predict the resulting dynamics.

Gaussians are a natural carrier of physical state. Their particle-like structure maps directly onto MPM dynamics without requiring an explicit mesh, while retaining appearance and mechanical state within a unified representation. Established 3DGS reconstruction pipelines can therefore provide the geometric and visual basis for simulatable assets; what remains unresolved is how to determine

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.

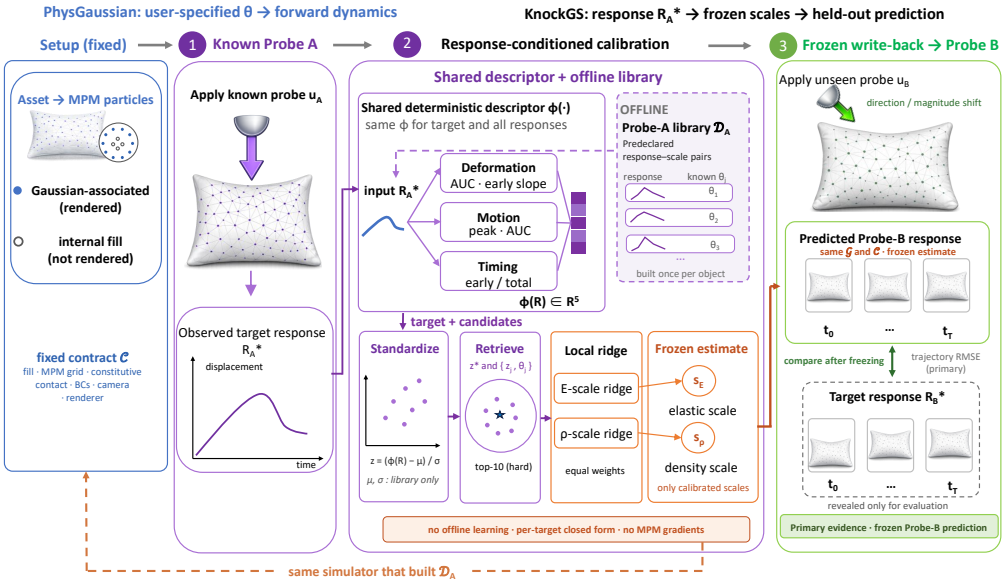


Figure 1: Overview of KnockGS and its frozen-prediction protocol. The setup fixes the Gaussian asset $\mathcal{G}$ and simulator contract $\mathcal{C}$. (1) Known Probe A, u_A, produces target response $R_A^\star$. (2) Descriptor ϕ encodes target and library responses; local ridge estimates continuous elasticity and density scales (s_E, s_ρ). (3) The frozen estimate predicts the response to unseen Probe B, u_B. Target response $R_B^\star$ is revealed only for evaluation, with trajectory RMSE as the primary metric.

their physical parameters. Specifically, visual reconstruction alone does not determine elasticity, density, damping, friction, or internal structure. **Two objects with very different mass and stiffness can fall and appear similarly under gravity.** Objects with nearly identical appearance can exhibit substantially different responses to the same physical interaction. Static appearance and passive observation are therefore fundamentally limited in disambiguating such objects, whereas an *actively applied, known* interaction together with its observed response provides direct physical evidence about the object. Appearance-based physical-property methods (Xu et al., 2025; Shuai et al., 2025; Chopra et al., 2026; Lv et al., 2026) and video-based inverse or dynamics-learning methods (Li et al., 2023; Cai et al., 2024; Zhong et al., 2024; Jiang et al., 2025; Rho et al., 2025; Wang et al., 2026) provide useful alternatives, but they do not directly answer whether a parameter estimate obtained from one *known controlled interaction* predicts a different, unseen interaction.

The target research problem is interaction-grounded physical representation: from an initial observation O_0, a known interaction u_A, and a physically observable response Y_A, infer a representation g_{phys} that predicts the response Y_B to a disjoint interaction u_B. The decisive criterion is therefore not parameter proximity alone, but whether the frozen representation predicts motion, deformation, contact, and ultimately visual response under u_B. Real systems would instantiate Y_A with RGB/RGB-D video, surface tracks, silhouettes, or force/tactile measurements.

This paper studies a controlled version of interaction-grounded physical inference (Fig. 1): can physical parameters inferred from one known interaction, rather than supplied in advance for forward simulation, predict the response to a different, unseen interaction? The unnumbered setup in Fig. 1 fixes the simulatable Gaussian asset (following PhysGaussian (Xie et al., 2024)), MPM solver, particle fill, grid, constitutive family, boundary conditions, and rendering. Within this fixed contract, we estimate only two dimensionless material scales, a Young's modulus scale and a density scale $\boldsymbol{\theta} = (s_E, s_\rho)$. The observed response is simulator-exported and therefore *privileged*, with both parameter and response ground truth known.

We separate calibration from prediction using two interactions. Probe A is a known controlled interaction used to estimate $\boldsymbol{\theta}$; Probe B differs in direction, magnitude, or both and is never used during estimation. The estimated parameters are frozen before Probe B is applied. Thus, Probe A

measures whether the method can recover a physically useful state, while Probe B tests whether that state predicts beyond the interaction it was fitted to.

KnockGS, an interaction-response PhysicalGS, performs this calibration with an object-specific library of Probe-A responses, precomputed once under the same simulator contract. A shared deterministic five-dimensional descriptor encodes both the target and every candidate response; after library-only standardization and hard top-10 retrieval, an equal-weight local ridge fit predicts continuous (s_E, s_ρ) values without differentiating through MPM. The estimate is then frozen and written back into the same simulator. We use held-out Probe-B trajectory RMSE as the primary evidence, with rendered PSNR/SSIM and response-curve statistics as complementary measures.

On held-out Pillow targets, the local estimator outperforms response-nearest, response-kNN, and global ridge given identical evidence, and remains predictive under direction- and magnitude-shifted Probe B. The protocol also succeeds on Ficus and Vasedeck.

Our contributions are:

1. We formulate material-scale inference for physics-integrated Gaussians as a Probe-A-to-Probe-B problem, where the estimate from one known interaction is frozen and judged by its prediction of an unseen one, unlike forward-only PhysGaussian pipelines and passive-video inverse methods.

2. We propose a library-based local estimator that yields continuous (s_E, s_ρ) without per-target simulation or gradients through MPM. Under the same evidence it reduces joint scale error from 2.4% (KNN, global ridge) to 1.1% and lowers held-out Probe-B trajectory error by about $3\times$. On the synthetic-depth track, it also outperforms CMA-ES video optimization with 30 forward calls per target per seed.

3. We show that known probing distinguishes the five evaluated held-out material settings: estimated stiffness-to-density scale ratios span 163%, compared with the targets' 162%. We also analyze the common absolute-scale ambiguity of the passive spring–mass model and characterize the method's failure boundaries.

2 Related Work

2.1 Physics-Integrated Gaussian Representations

PhysGaussian (Xie et al., 2024) demonstrated that 3D Gaussians (Kerbl et al., 2023) can serve as a unified representation for physics-based simulation and rendering, extending neural scene representations (Mildenhall et al., 2020) with MPM-based dynamics (Sulsky et al., 1994; Jiang et al., 2015). Making a Gaussian asset simulatable leaves a separate question: how should its material scales be calibrated from observations? We connect controlled interaction-response probing, particle-level 3D response features, local calibration, and frozen-parameter re-simulation to evaluate this inference step.

2.2 Physical System Identification from Visual Dynamics

The ecosystem can be organized into three adjacent routes. Forward PhysicalGS methods assume material parameters and improve simulation or rendering (Xie et al., 2024; Lin et al., 2025; Cao et al., 2026; Ma et al., 2026). Visual-prior approaches infer or assign attributes from static appearance (Xu et al., 2025; Shuai et al., 2025; Chopra et al., 2026; Lv et al., 2026). Passive-video inverse methods use observed dynamics, often through differentiable optimization or learned prediction (Li et al., 2023; Cai et al., 2024; Zhong et al., 2024; Wang et al., 2026). PhysTwin (Jiang et al., 2025) explicitly takes sparse videos of deformable objects under interaction and evaluates simulation under novel interactions. Our controlled route differs in the information source and validation axis: the excitation is known, repeatable, and actively specified, and parameters estimated from Probe A are frozen before a different Probe B is evaluated. Table 5 in Appendix B summarizes these distinctions.

3 METHOD

Given a fixed PhysicalGS asset $\mathcal{G}$ and simulator contract $\mathcal{C}$, our method uses the observed response to a known *Calibration Probe A* u_A to estimate two dimensionless material scales, $\boldsymbol{\theta} = (s_E, s_\rho)$. The estimate is then frozen, written back into the same PhysicalGS simulator, and used to predict the response to a disjoint *Held-out Prediction Probe B* u_B. The asset, particle fill, MPM solver and grid, constitutive family, boundary conditions, camera, and renderer are built on the PhysGaussian backbone (Xie et al., 2024) and remain fixed within $\mathcal{C}$; only the two scales are calibrated. In the terminology of Sec. 1, the fixed asset $\mathcal{G}$ represents the initial observation O_0, the target response $R_A^\star$ instantiates Y_A, the calibrated asset $(\mathcal{G}, \widehat{\boldsymbol{\theta}})$ under $\mathcal{C}$ instantiates g_{phys}, and $\widehat{R}_B$ is the predicted instantiation of Y_B.

The design is motivated by the high cost of repeated MPM simulation. We precompute a small response library that can be reused for calibration. Sec. 3.2 precomputes an object-specific response library and maps both library responses and target responses to the same deterministic five-dimensional descriptor. This enables response comparison in a compact feature space and allows a target to be matched against precomputed simulations. Figure 1 therefore separates the unnumbered fixed setup from a three-stage per-target closed loop: (1) apply the known Probe A and record the target response; (2) calibrate continuous material scales using the shared descriptor and a local closed-form fit; and (3) freeze the estimate, write it back into the same simulator, and predict the held-out Probe B.

3.1 PROBLEM SETTING AND CONTROLLED SCOPE

We study effective material-scale calibration within a declared simulation contract. A PhysicalGS asset $\mathcal{G}$ contains renderable Gaussian kernels and a simulator particle set

$$X^0 = X^0_{\text{GS}} \cup X^0_{\text{fill}}, \tag{1}$$

where Gaussian centers are mapped to Gaussian-associated MPM particles and additional particles fill the interior. The contract $\mathcal{C}$ fixes the asset, coordinate transform, fill construction, MPM grid, constitutive family, contact model, boundary conditions, damping, friction, Poisson ratio, camera, and renderer. The only unknown quantities are

$$\boldsymbol{\theta} = (s_E, s_\rho) \in \Theta, \tag{2}$$

where s_E and s_ρ are effective calibration scales defined relative to the fixed object-specific simulator template.

An interaction protocol u specifies the force vector, contact region, start time, and duration under the fixed support and boundary configuration. The simulator $\mathcal{S}$ produces a rollout under u. At predeclared export times $t \in \mathcal{T}_{\text{out}}$, the main observation operator exports particle positions,

$$R_u(\boldsymbol{\theta}) = \mathcal{O}_{\text{part}}(\mathcal{S}(\mathcal{G}, \mathcal{C}, \boldsymbol{\theta}, u)) = \{\mathbf{x}_i^t\}_{i=1, t\in\mathcal{T}_{\text{out}}}^{N}. \tag{3}$$

The target response to Calibration Probe A is $R_A^\star = R_{u_A}(\boldsymbol{\theta}^\star)$, while $\boldsymbol{\theta}^\star$ remains hidden from every deployable estimator. Given the object-specific Probe-A library $\mathcal{D}_A$, the instantiated objective is

$$\widehat{\boldsymbol{\theta}} = h(R_A^\star; \mathcal{D}_A), \qquad \widehat{R}_B = R_{u_B}(\widehat{\boldsymbol{\theta}}). \tag{4}$$

The held-out target response $R_B^\star$ is used only to evaluate $\widehat{R}_B$ after calibration has been frozen. The main estimator uses privileged simulator-exported state; Sec. H.1 evaluates progressively less privileged synthetic observations.

3.2 OBJECT-SPECIFIC RESPONSE LIBRARY

Offline response bank. The response map $\boldsymbol{\theta} \mapsto R_A(\boldsymbol{\theta})$ is available only through simulation. We therefore evaluate it in advance on a fixed candidate set and reuse those responses for every target governed by the same $\mathcal{C}$ and u_A. For each asset, the resulting library is

$$\mathcal{D}_A = \{(R_A(\boldsymbol{\theta}_j), \boldsymbol{\theta}_j)\}_{j=1}^{J}, \tag{5}$$

where J is the number of candidates. They provide finite coverage of the declared two-scale domain at a reusable offline cost. Appendix D reports the exact support and target split, and Appendix

I measures how candidate count affects accuracy and coverage stability. All held-out targets are excluded from $\mathcal{D}_A$, and neither their hidden scales nor any Probe-B response contributes to library construction. We need a lightweight interpolation mechanism that converts a finite response library into continuous material estimates without additional simulator calls. Hard-neighborhood local ridge is the simplest mechanism we found that satisfies this requirement while outperforming retrieval and global regression under the same evidence.

Shared response descriptor. Raw particle trajectories are high-dimensional and object-dependent, so the same deterministic map compresses every candidate and target response before comparison. Let D_{bbox} be the diagonal of the initial particle bounding box. We first compute normalized RMS displacement and frame-difference velocity curves,

$$r^t = \frac{1}{D_{\mathrm{bbox}}}\sqrt{\frac{1}{N}\sum_i \|\mathbf{x}_i^t - \mathbf{x}_i^0\|_2^2}, \quad v^t = \frac{1}{D_{\mathrm{bbox}}}\sqrt{\frac{1}{N}\sum_i \|\mathbf{x}_i^t - \mathbf{x}_i^{t-1}\|_2^2}. \tag{6}$$

One five-dimensional descriptor then summarizes complementary response-amplitude and timing cues,

$$\phi(R) = \left[\mathrm{AUC}(r),\ s_{\mathrm{early}}(r),\ \max_t v^t,\ \mathrm{AUC}(v),\ \frac{\sum_{t\le t_e} v^t}{\sum_t v^t}\right]^\top, \tag{7}$$

with $t_e = 5$. Its entries measure cumulative deformation, early deformation rate, peak frame-to-frame motion, cumulative motion, and the fraction of motion concentrated early in the rollout, respectively.

3.3 Response-Conditioned Continuous Calibration

Returning the nearest response can only select an existing candidate, so it cannot represent a target whose scales are absent from the finite library. A continuous interpolator is therefore required. Because a single linear map need not approximate the descriptor-to-scale relation across the full domain, we instead assume that it is better approximated within a small response-space neighborhood. A hard neighborhood provides this locality, while weak ridge regularization stabilizes the resulting few-sample fit. This design relaxes the quantization limitation of retrieval and produces a continuous estimate without a new MPM rollout or differentiation through the simulator. Sec. 4.2 evaluates the resulting local closed-form interpolation against response retrieval and global ridge under the same response library and descriptor evidence.

Let $\mathbf{g}_j$ and $\mathbf{g}^\star$ be the standardized descriptors of candidate j and the target response. Euclidean distance is used only to select the hard neighborhood $\mathcal{N}_{10}$ containing the ten nearest candidates. Samples within that neighborhood have equal weight, and we solve

$$\min_{\mathbf{A},\mathbf{b}} \sum_{j\in\mathcal{N}_{10}} \|\mathbf{A}\mathbf{g}_j + \mathbf{b} - \boldsymbol{\theta}_j\|_2^2 + 10^{-3}\|\mathbf{A}\|_F^2, \qquad \widehat{\boldsymbol{\theta}} = \mathbf{A}\mathbf{g}^\star + \mathbf{b}. \tag{8}$$

The intercept is not regularized, and the two material scales are regressed independently. Predictions are clipped to the candidate-set bounds. Inverse-distance KNN is used only if fewer than three neighbors are available or a numerical exception prevents the ridge solve; the main estimator is not a distance-weighted ridge regression. The configuration $k = 10$, $\alpha = 10^{-3}$ is fixed before held-out evaluation.

3.4 Frozen Write-Back and Held-out Prediction

Parameter proximity alone does not establish a useful PhysicalGS representation: an estimate may be numerically close to the hidden scales yet reproduce the wrong motion, or it may reconstruct Probe A without carrying information that transfers to a different interaction. We therefore close the loop in response space. Immediately after observing Calibration Probe A, $\widehat{\boldsymbol{\theta}}$ is frozen and written back into the same simulator that generated $\mathcal{D}_A$. We then distinguish calibration reconstruction from held-out prediction,

$$\widehat{R}_A = R_{u_A}(\widehat{\boldsymbol{\theta}}), \qquad \widehat{R}_B = R_{u_B}(\widehat{\boldsymbol{\theta}}). \tag{9}$$

Probe-A reconstruction is a diagnostic of consistency with the observed response. Held-out Prediction Probe B is the primary evidence: its protocol is predeclared, may change force direction, magnitude,

duration, or contact location, and contributes no frames, trajectories, descriptors, or statistics to estimator design or fitting. Its target response $R_B^\star$ is revealed only for final comparison with $\widehat{R}_B$.

The primary MPM metric compares displacements $\Delta\mathbf{x}_i^t = \mathbf{x}_i^t - \mathbf{x}_i^0$ from each rollout's initial state on a fixed Gaussian-associated particle set $\mathcal{I}_{\mathrm{GS}}$, normalized by the target's initial bounding-box diagonal:

$$\mathrm{RMSE}_{\mathrm{traj}} = \frac{1}{D_{\mathrm{bbox}}} \sqrt{\frac{1}{|\mathcal{I}_{\mathrm{GS}}||\mathcal{T}_{\mathrm{out}}|} \sum_{i\in\mathcal{I}_{\mathrm{GS}}} \sum_{t\in\mathcal{T}_{\mathrm{out}}} \|\Delta\widehat{\mathbf{x}}_i^t - \Delta\mathbf{x}_i^{\star,t}\|_2^2}. \tag{10}$$

Sec. 4 complements this spatiotemporal metric with response-curve and rendered-frame measures. Accordingly, Probe-B evaluation measures cross-interaction prediction within the fixed object-specific simulator contract.

4 Experiments

We test whether local calibration improves parameter recovery under matched Probe-A evidence and whether the frozen estimate predicts unseen Probe B. Additional experiments evaluate probe selection, object-specific repeatability, and a contact-based FEM setting. Observation and numerical diagnostics in Appendices G and H characterize the scope of these results.

4.1 Experimental Contract

The following contract describes the main PhysicalGS/MPM benchmark. Sec. 4.5 separately instantiates the same calibration procedure with a contact-based FEM response library in Isaac Sim.

Input and output. Deployable estimators receive the fixed asset and simulation contract, known Probe A, and its target response. Hidden target scales, Probe-B observations, and cross-target statistics are excluded from calibration. The estimated $(\hat{s}_E, \hat{s}_\rho)$ is frozen before re-simulation and evaluation. The main response input is *privileged* simulator particle state; Appendix H.1 evaluates less privileged synthetic observations.

Table 1: Pillow benchmark under matched Probe-A evidence. Both Probe-B protocols are unseen during estimation and tuning. Trajectory RMSE is dimensionless, normalized by the initial bounding-box diagonal (Eq. 10); lower is better. Parameter-nearest uses hidden target parameters and is an oracle diagnostic.

Method	Joint parameter error (%)	Probe A traj. RMSE	Probe B direction	Probe B magnitude
Fixed default	27.31	3.682×10^{-3}	3.502×10^{-3}	4.407×10^{-3}
Response nearest	6.75	3.722×10^{-4}	3.533×10^{-4}	5.188×10^{-4}
Response KNN	2.37	1.449×10^{-4}	1.412×10^{-4}	2.020×10^{-4}
Global ridge	2.45	1.748×10^{-4}	1.687×10^{-4}	2.466×10^{-4}
Local ridge (ours)	**1.13**	$\mathbf{4.498\times10^{-5}}$	$\mathbf{4.424\times10^{-5}}$	$\mathbf{7.786\times10^{-5}}$
Parameter-nearest oracle	4.45	6.283×10^{-4}	5.984×10^{-4}	7.144×10^{-4}

Baselines. All deployable in-domain methods receive the same Probe-A evidence. Response-nearest, inverse-distance response KNN, global ridge, and our local ridge use the same predeclared object-specific response library and five-dimensional descriptor; fixed default ignores that evidence by construction. They differ in how neighborhood evidence is converted into an estimate, which isolates the contribution of Sec. 3.3. KNN neighborhood size and global-ridge regularization are selected by candidate-only validation. Parameter-nearest reads the hidden target location and is reported only as an oracle diagnostic. Sec. H.2 separately evaluates direct video optimization, PhysGM, and ReconPhys.

Computing resources. The Pillow library costs 54 offline MPM rollouts per object. Target calibration adds no rollout: it retrieves ten neighbors and solves the local ridge systems in closed

form; response evaluation then requires re-simulation. CMA-ES uses 30 forward calls per target per seed, totaling 450 rollouts. Simulation sizes and timing are specified in Appendix D, and optimization settings in Appendix H.2.

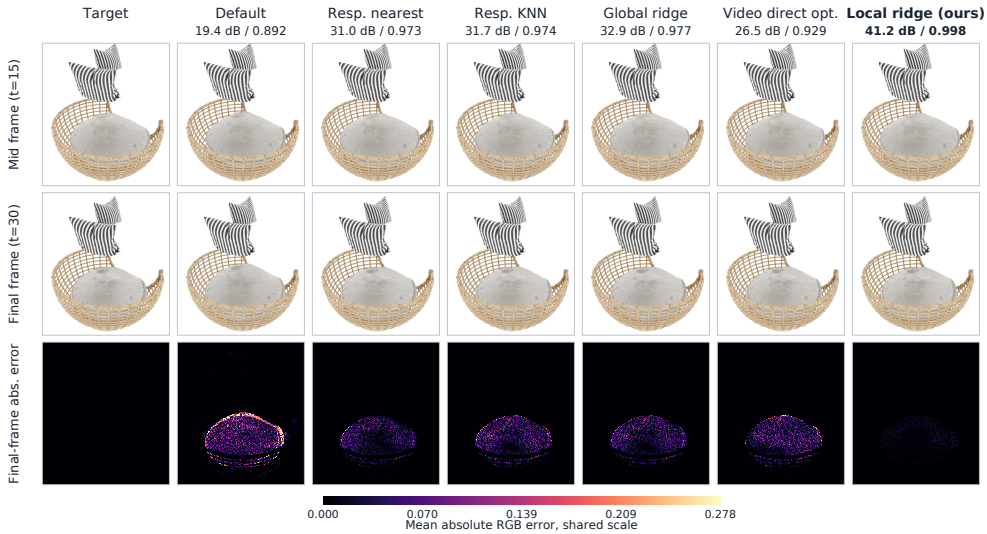


Figure 2: Qualitative held-out Probe-B prediction on Pillow (target 002, $\vec{y}_{standard}$). Rows: exported frames 15 and 30, and final-frame mean absolute RGB error. Error maps share a color scale clipped at the 99.9th percentile of full-frame union-mask errors pooled across predictions. All columns use the same crop. Headers report target-specific sequence-level metrics over all 31 exported frames: foreground PSNR in dB / object-crop SSIM. CMA-ES uses predeclared seed 6401; aggregate direct-optimization results average all three seeds.

Evaluation metrics. We report the joint relative error

$$e_\theta = \frac{1}{2}\left(\frac{|\hat{s}_E - s_E^\star|}{s_E^\star} + \frac{|\hat{s}_\rho - s_\rho^\star|}{s_\rho^\star}\right) \times 100\%, \tag{11}$$

and the Gaussian-associated-particle trajectory RMSE in Eq. (10). Curve RMSE, peak/AUC/final errors, foreground PSNR, and object-crop SSIM are secondary. Trajectories are compared at the same exported times with fixed particle correspondence. The in-domain estimators and MPM evaluations are deterministic under frozen seeds; Appendix E therefore reports every held-out target and mean $\pm$ sample standard deviation across targets, rather than pseudo-replicating frames or particles. The stochastic CMA-ES baseline is additionally reported per target as mean $\pm$ standard deviation over its three predeclared seeds. Completion, leakage, missing-frame, NaN, and duplicate-output audits accompany the experiment packages.

Assets, candidates, and targets. The main benchmark uses a reconstructed pillow/sofa Gaussian asset and a candidate set formed by combining a broad sweep over the declared two-scale domain with a denser sweep that supplies local interpolation support. Removing duplicate parameter pairs yields 54 unique (s_E, s_ρ) candidates. This set was fixed before any held-out evaluation; Appendix I reports post-hoc sensitivity to smaller candidate subsets without retuning the main result. The five held-out scale pairs lie within the covered domain but do not coincide with any library candidate, so the evaluation tests continuous scale estimation rather than exact table lookup. The same five target locations are used for object-specific Ficus and Vasedeck experiments. Candidate and target responses are generated by the same PhysicalGS/MPM implementation. Their parameters are preset by the experimenter and hidden from deployable estimators; they are simulator-defined ground truth, not measured real-material properties. Tables 6–7 in Appendix D give the exact split, material bases, timing, forces, contact regions, and particle populations.

Probe separation. MPM Probe A ($\vec{x}_{standard}$) applies a known 0.1 ms force pulse throughout the selected particle region; for Pillow this includes all physical-state particles, rather than a surface

Table 2: Probe-selection core results on Pillow. Means over five held-out targets; lower is better. Probe-B trajectory RMSE is normalized by the initial bounding-box diagonal and shown in units of 10^{-5}. Each probe is reset independently. Full results and sample standard deviations are in Appendix O.

Probe set	Probes	Joint parameter error (%)	Probe B direction	Probe B magnitude
X (original)	1	1.130	4.424	7.786
Z (selected)	1	0.703	3.551	4.970
$X + Z$ (selected)	2	0.474	**2.552**	**3.559**
$X + Z + L$	3	**0.298**	3.175	3.713

tool impact. Isaac FEM instead uses a contacting rigid probe (Sec. 4.5). Pillow Probe B changes direction ($\vec{y}_{standard}$) or magnitude ($\vec{x}_{strong}$); Vasedeck uses $\vec{z}_{standard}$. Duration and contact-region diagnostics appear in Appendix F. Probe B never enters descriptor design, tuning, retrieval, or parameter estimation.

4.2 Same-Information Calibration and Frozen Probe-B Prediction

Under identical Probe-A evidence, parameter recovery tests the local fit against retrieval and global regression; frozen Probe-B prediction tests whether that estimate remains useful under a change of excitation.

Table 1 and Fig. 3 give the central result. Local ridge achieves 1.13% mean parameter error, improving over response KNN (2.37%), global ridge (2.45%), and response-nearest (6.75%); this establishes better calibration under matched evidence. The primary evidence is the held-out prediction made by the same frozen estimate. On $\vec{y}_{standard}$, its Probe-B trajectory RMSE is 4.42×10^{-5}, versus 1.41×10^{-4} for KNN and 1.69×10^{-4} for global ridge. On $\vec{x}_{strong}$, it obtains 7.79×10^{-5}, versus 2.02×10^{-4} and 2.47×10^{-4}. These results show that scales calibrated from Probe A remain predictive under held-out changes in force direction and magnitude within the same simulation contract.

Fig. 2 makes the same held-out-probe result visible in image space, showing a mid and a final exported frame of target 002 under unseen $\vec{y}_{standard}$; no frame or Probe-B signal is used for fitting. Local ridge reaches 41.20 dB foreground PSNR and 0.998 object-crop SSIM, compared with 32.93 dB/0.977 for global ridge, 31.75 dB/0.974 for response KNN, 26.54 dB/0.929 for the fixed CMA-ES seed shown, and 19.41 dB/0.892 for the default material. Because the deformation is small relative to the static basket, the renders are visually close; the shared-scale error maps localize the mismatch to the deforming seat cushion and show the lowest image-space error for local ridge in this example. The ordering of the remaining methods can differ from the aggregate trajectory ranking.

Across three predeclared alternative target splits, local ridge has lower mean parameter error than the strongest non-oracle method in all three; see Appendix G.

4.3 Probe Selection and Complementary Responses

The calibration interaction also determines which parameter differences become visible. We compare the original X probe with an orthogonal global probe Z and a local probe L, keeping the descriptor and local-ridge settings fixed. Candidate-only leave-one-out validation selects Z at a one-probe budget and $X + Z$ at a two-probe budget, without using target responses or Probe B for selection. Table 2 shows that Z improves both parameter recovery and unseen-response prediction over X, while combining $X + Z$ further reduces both errors. Adding L gives the lowest parameter error, but does not improve Probe-B prediction over $X + Z$: parameter accuracy and response accuracy are complementary tests. This retrospective extension uses the existing five targets and independently reset probes; full subsets, selection diagnostics, and variability are in Appendix O.

Table 3: Isaac Sim FEM validation on five appearance-matched held-out pillows. Entries are means across targets; trajectory RMSE uses all FEM nodes and is reported in mm. Both Probe-B interactions are excluded from estimation. Full sample standard deviations across the five materials ($n = 5$, $\mathrm{ddof} = 1$) are given in Appendix N.

Method	Joint parameter error (%)	Probe B oblique (mm)	Probe B larger pulse (mm)
Fixed default	27.31	0.21246	0.21591
Response nearest	6.98	0.02624	0.02722
Local ridge (ours)	**1.07**	**0.00394**	**0.00427**

4.4 Object-Specific Repeatability

We repeat calibration and frozen prediction on Ficus and Vasedeck, keeping the descriptor and local-ridge settings fixed and building a separate library for each asset. Mean parameter errors are 0.63% and 1.21%, respectively, compared with 1.13% on Pillow. These are independently calibrated object-specific results, not cross-object transfer; Appendix K reports the baselines and held-out responses.

4.5 Contact-Based Calibration in Isaac Sim

We additionally instantiate the calibration protocol in Isaac Sim's PhysX FEM, using a fixed-bottom pillow and a force-driven rigid probe. Five held-out pillows share the same geometry, mesh, and appearance but differ in homogeneous Young's modulus and density. We rebuild a 54-candidate FEM response library and retain the original five-dimensional descriptor and local-ridge estimator. Probe A supplies the only target evidence; the estimates are frozen before testing an oblique force pulse and a larger normal pulse as two unseen Probe-B interactions. All rollouts use the same fixed 3200 Hz configuration.

Table 3 shows a mean joint parameter error of 1.07%, compared with 6.98% for response-nearest and 27.31% for default parameters. All five targets improve over default in both Probe-B tests. Relative to response-nearest, mean prediction error falls by 84.99% and 84.31%, respectively, with lower error on all five targets in each protocol. This evaluates privileged FEM-node prediction using a newly built response library within one solver configuration, not transfer of an MPM library or visual Gaussian reconstruction. Cross-step numerical convergence remains unestablished; Appendix Sec. N gives the protocol, per-target results, and scope.

Offline estimator sensitivity. On the main MPM benchmark, increasing the library from 18 to 54 candidates reduces mean parameter error from 2.078% to 1.130%; all 30 tested (k, α) settings remain below the 2.374% response-KNN baseline. Appendix I gives the full sensitivity analysis; the predeclared main configuration is retained without held-out retuning.

5 Conclusion

We presented KnockGS, a controlled study of response-conditioned calibration for physics-integrated Gaussian assets. An object-specific library and a simple hard-neighborhood local ridge estimator recover two MPM material scales from a known Probe-A response; after write-back, the frozen estimate predicts held-out direction and magnitude probes more accurately than response retrieval, KNN, global ridge, and fixed default materials. The result repeats in object-specific loops on Pillow, Ficus, and Vasedeck, and is supported by Gaussian-associated-particle trajectories and rendered visual fidelity.

Equally important, the stress tests delimit the result. Clean synthetic surface and depth observations retain useful evidence, but ID-free visual tracks and depth perturbations degrade calibration. Fill or grid mismatch and cross-object sharing fail unless the response library is rebuilt under matching conditions. Closing that gap requires stable visual correspondence, uncertainty-aware identification, broader physical states, and validation with measured force and object-level deformation.

AI USE STATEMENT

A generative AI assistant was used during preparation of this version to identify the current ICLR template, convert the LaTeX source, assist with compilation, layout, and formatting checks, and draft the presentation of supplied Isaac Sim experiment records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5460–5469, 2022. doi: 10.1109/CVPR52688.2022.00539.

Junhao Cai, Yuji Yang, Weihao Yuan, Yisheng He, Zilong Dong, Liefeng Bo, Hui Cheng, and Qifeng Chen. Gaussian-informed continuum for physical property identification and simulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yicheng Cao, Zhuo Huang, Yu Yao, Yiming Ying, Daoyi Dong, and Tongliang Liu. i-physgaussian: Implicit physical simulation for 3d gaussian splatting. *arXiv preprint arXiv:2602.17117*, 2026.

Samarth Chopra, Jing Liang, Gershom Seneviratne, and Dinesh Manocha. Physgs: Bayesian-inferred gaussian splatting for physical property estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18980–18990, 2026.

Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Transactions on Graphics*, 37(4):1–14, 2018.

Bei Huang, Yixin Chen, Ruijie Lu, Gang Zeng, Hongbin Zha, Yuru Pei, and Siyuan Huang. Gaussianfluent: Gaussian simulation for dynamic scenes with mixed materials. *arXiv preprint arXiv:2601.09265*, 2026.

Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. The affine particle-in-cell method. *ACM Transactions on Graphics*, 34(4), 2015.

Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7219–7230. IEEE, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Yuanhang Lei, Boming Zhao, Zesong Yang, Xingxuan Li, Tao Cheng, Haocheng Peng, Ru Zhang, yang yang, Siyuan Huang, Yujun Shen, Ruizhen Hu, Hujun Bao, and Zhaopeng Cui. Diffwind: Physics-informed differentiable modeling of wind-driven object dynamics. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 23469–23494, 2026.

Shiqian Li, Ruihong Shen, Junfeng Ni, Chang Pan, Chi Zhang, and Yixin Zhu. Learning physics-grounded 4d dynamics with neural gaussian force fields. In *International Conference on Learning Representations*, volume 2026, pp. 81165–81207, 2026.

Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, Krishna Murthy Jatavallabhula, Ming Lin, Chenfanfu Jiang, and Chuang Gan. PAC-NeRF: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.

Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. OmniPhysGS: 3D constitutive Gaussians for general physics-based dynamics generation. In *Proc. International Conference on Learning Representations (ICLR)*, 2025.

Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. ManiGaussian: Dynamic Gaussian splatting for multi-task robotic manipulation. In *Proc. European Conference on Computer Vision (ECCV)*, 2024.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D Gaussians: Tracking by persistent dynamic view synthesis. In *Proc. International Conference on 3D Vision (3DV)*, 2024.

Chunji Lv, Zequn Chen, Donglin Di, Weinan Zhang, Hao Li, Chen Wei, Yinjie Lei, and Changsheng Li. Physgm: Large physical gaussian model for feed-forward 4d synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29855–29865, 2026.

Yikun Ma, Yiqing Li, Jingwen Ye, Zhongkai Wu, Weidong Zhang, Lin Gao, and Zhi Jin. Fast-physgs: Accelerating physics-based dynamic 3dgs simulation via interior completion and adaptive optimization. *arXiv preprint arXiv:2602.01723*, 2026.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pp. 405–421. Springer, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):1–15, 2022.

M Nomaan Qureshi, Sparsh Garg, Francisco Yandun, David Held, George Kantor, and Abhisesh Silwal. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6502–6509. IEEE, 2025.

Daniel Rho, Jun Myeong Choi, Biswadip Dey, and Roni Sengupta. Projo4d: Progressive joint optimization for sparse-view inverse physics estimation. *arXiv preprint arXiv:2506.05317*, 2025.

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.

Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics*, 32(4):1–10, 2013.

Deborah Sulsky, Zhen Chen, and Howard L. Schreyer. A particle method for history-dependent materials. *Computer Methods in Applied Mechanics and Engineering*, 118(1–2):179–196, 1994.

Boyuan Wang, Xiaofeng Wang, Yongkang Li, Zheng Zhu, Yifan Chang, Angen Ye, Guosheng Zhao, Chaojun Ni, Guan Huang, Yijie Ren, Yueqi Duan, and Xingang Wang. ReconPhys: Reconstruct appearance and physical attributes from single video, 2026. arXiv:2604.07882.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3D Gaussians for generative dynamics. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xinli Xu, Wenhang Ge, Dicong Qiu, ZhiFei Chen, Dongyu Yan, Zhuoyun Liu, Haoyu Zhao, Hanfeng Zhao, Shunsi Zhang, Junwei Liang, et al. Gaussianproperty: Integrating physical properties to 3d gaussians with lmms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7231–7240. IEEE, 2025.

Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. AdaptiGraph: Material-adaptive graph-based neural dynamics for robotic manipulation. In *Proc. Robotics: Science and Systems (RSS)*, 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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5242–5252. IEEE, 2025.

Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3D Gaussians. In *Proc. European Conference on Computer Vision (ECCV)*, 2024.

A Overview

This supplement provides the experimental contracts, metric definitions, target-wise results, and additional analyses supporting the main paper. Table 4 lists its contents in document order. Sections C–E specify the main MPM benchmark; Sections N and O describe the FEM validation and probe-selection extension, respectively.

Table 4: Contents of the supplementary material in document order.

Supplementary section	Contents
Sec. B: Related work	Comparison of physical representations and calibration settings.
Sec. C: Response and metrics	Descriptors, trajectory errors, and image-metric aggregation.
Sec. D: Probe definitions	Assets, material scales, inputs, timing, and A/B separation.
Sec. E: Target-wise results	Per-target accuracy and between-target variability.
Sec. F: Additional probes	Duration and contact-location changes.
Sec. G: Splits and numerics	Alternative target splits and numerical tests.
Sec. H: Diagnostic boundaries	Observation, inverse-route, and discretization diagnostics.
Sec. I: Estimator sensitivity	Library coverage and offline estimator settings.
Sec. J: Observation details	Synthetic observation inputs and their limitations.
Sec. K: Object-specific calibration	Additional assets and independent response libraries.
Sec. L: Alternative routes	External adapters and direct-optimization results.
Sec. M: Reproducibility	Recorded contracts, result packages, and claim scope.
Sec. N: Isaac Sim FEM	Contact-based calibration, baselines, and numerical scope.
Sec. O: Probe selection	Candidate-only selection, all subsets, and sensitivity.
Sec. P: Discussion	Supported conclusions and future directions.

B Comparison and Related Work

Table 5 separates representation, material assignment, interaction evidence, response encoding, and downstream validation. These capabilities answer different questions: representing dynamics does not itself specify how material parameters are obtained, and estimating parameters does not establish prediction under a new excitation. The last two columns connect the calibration signal to the frozen-parameter response test used here.

C Response and Metric Details

Let $\mathcal{I}_{\mathrm{GS}}$ denote the evaluated Gaussian-associated particles with stable identity and $\mathcal{I}_{\mathrm{fill}}$ internal fill particles. The estimator's privileged input may use all exported particles, while the primary trajectory evaluation is restricted to $\mathcal{I}_{\mathrm{GS}}$. For Pillow, a fixed seed of 42 selects 100,000 particles from the 624,324 Gaussian-associated particles. Equation 10 pools squared vector-displacement errors over these particles and all 31 exported frames, takes the square root, and divides by the diagonal of the target's initial all-particle bounding box. Displacements are relative to each rollout's own initial state, with initial correspondence checked before comparison. This dimensionless MPM metric differs from the all-node FEM trajectory error reported in mm in Appendix N. Internal fill particles are excluded from correspondence-based evaluation because their identities can change across runs. Exported states are recorded at the configured output times, not every integration step.

For secondary curve metrics, normalized particle displacement is

$$\tilde{d}_i^t = \|\mathbf{x}_i^t - \mathbf{x}_i^0\|_2 / D_{\mathrm{bbox}}, \tag{12}$$

and the RMS curve is $r^t = (|\mathcal{I}|^{-1} \sum_{i\in\mathcal{I}} (\tilde{d}_i^t)^2)^{1/2}$. We report

$$\mathrm{RMSE}_{\mathrm{curve}} = \sqrt{\frac{1}{|\mathcal{T}_{\mathrm{out}}|} \sum_t (\hat{r}^t - r^{\star,t})^2}, \tag{13}$$

Table 5: Comparison with representative related work in the PhysGaussian ecosystem.

Category	Representative works	GS / particle repr.	Material estimation	Interaction-Response probing	3D response features	Re-simulation fidelity
PhysGaussian base	PhysGaussian (Xie et al., 2024)	✓	✗	✗	✗	✗
Solver / forward enhancement	i-PhysGaussian (Cao et al., 2026), FastPhysGS (Ma et al., 2026), GaussianFluent (Huang et al., 2026)	✓	✗	✗	✗	✗
Multi-material / scene physics	OmniPhysGS (Lin et al., 2025), PhysSplat (Zhao et al., 2025)	✓	(✓)	✗	✗	✗
Visual / VLM property assignment	GaussianProperty (Xu et al., 2025), PUGS (Shuai et al., 2025), PhysGS (Chopra et al., 2026), PhysGM (Lv et al., 2026)	✓	✓	✗	✗	✗
Inverse problem / digital twin	GIC (Cai et al., 2024), Spring-Gaus (Zhong et al., 2024), PhysTwin (Jiang et al., 2025), ReconPhys (Wang et al., 2026), PAC-NeRF (Li et al., 2023)	(✓)	✓	✗	(✓)	(✓)
Dynamics / world models	NGFF (Li et al., 2026), Dynamic 3D Gaussian tracking (Luiten et al., 2024), DiffWind (Lei et al., 2026)	(✓)	✗	✗	(✓)	✗
Robot online adaptation	AdaptiGraph (Zhang et al., 2024), ManiGaussian (Lu et al., 2024), SplatSim (Qureshi et al., 2025)	(✓)	(✓)	✗	✗	✗
Response-conditioned calibration	**KnockGS (ours)**	✓	✓	✓	✓	✓

Note: ✓: satisfied; ✗: not addressed; (✓): partially satisfied

plus $|\max_t \hat{r}^t - \max_t r^{\star,t}|$, the absolute difference of discrete AUC sums, and final-frame error. Curve metrics summarize temporal magnitude but do not replace the spatial trajectory metric.

For the main Pillow evaluation, foreground PSNR is computed over the union of the target and predicted foreground masks, pooling RGB squared error and sample count over all 31 exported frames before conversion to dB. Object-crop SSIM is computed per frame within the bounding box of the same mask union, padded by five pixels, and then averaged across frames. Both target and prediction are rendered by the same Gaussian renderer and camera. They quantify same-simulator visual response fidelity and must not be interpreted as performance on captured real video.

For Ficus, the visual summary and Fig. 10 use a separate frame-averaged protocol over the 30 rendered frames. In each frame, the foreground is the union of pixels with any RGB channel below 250 in either image, dilated with a 7×7 kernel. PSNR is computed from RGB MSE inside this mask; foreground SSIM averages the grayscale SSIM map inside the same mask. Both scores are then averaged over frames. Thus, Ficus foreground SSIM differs from object-crop SSIM, and its mean frame PSNR differs from the pooled-MSE sequence PSNR used above.

D Probe Definitions and Separation

The candidate set comprises 54 predeclared pairs, not a Cartesian grid. Its coordinate levels are $s_E \in \{0.50, 0.55, 0.65, 0.75, 0.85, 0.95, 1.00, 1.05, 1.15\}$ and $s_\rho \in \{0.75, 1.00, 1.10, 1.20, 1.25, 1.35, 1.50, 1.65, 1.75\}$. The five targets are $(0.60, 1.65)$, $(0.70, 1.20)$, $(0.80, 1.60)$, $(0.95, 1.35)$, and $(1.05, 1.10)$; none is one of the 54 candidate pairs. Material scales multiply the frozen object-specific base values in Table 6. Candidate-only standardization and validation never use these targets.

All three objects use an internal MPM substep of 10^{-4} s. Pillow uses $n_{\text{grid}} = 100$ and a 128^3 fill grid with at most four particles per cell and frozen fill seed 60042. Ficus scales both its global material and its configured branch/leaf material region by the same (s_E, s_ρ) pair. Vasedeck uses $n_{\text{grid}} = 120$. Geometry, coordinate transforms, support constraints, gravity, damping, friction, camera, and renderer remain fixed within each object-specific library.

All parameters are estimated using only $\vec{x}_{standard}$. Probe-B responses are generated only after the estimate, descriptor definition, neighborhood size, and regularization have been frozen. The candidate

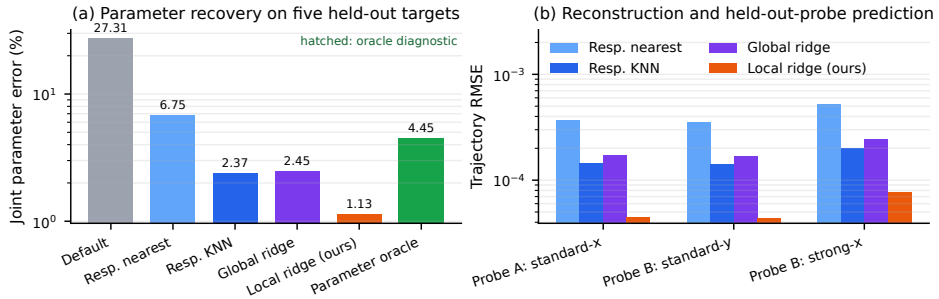


Figure 3: Fair baseline comparison. (a) Local ridge gives the lowest mean joint error among deployable methods; the hatched oracle accesses hidden target parameters. (b) The primary prediction evidence: Probe-A reconstruction and held-out Probe-B trajectories are generated with the same frozen estimates.

Table 6: Frozen object-specific simulation contract. Counts are physical-state particles used by the estimator; Pillow contains 624,324 Gaussian-associated particles and 71,660 internal fill particles. Primary Pillow trajectory evaluation samples at most 100,000 Gaussian-associated particles with seed 42. "Exports" includes the initial state.

Object	Base E	Base ρ	ν	Physical-state population	Exports / horizon
Pillow	10,000	2,000	0.30	695,984 (GS + fixed-seed fill)	31 / 0.60 s
Ficus	2×10^6	200	0.40	171,553 Gaussian-associated	31 / 1.20 s
Vasedeck	10,000	40	0.30	55,364 Gaussian-associated	31 / 0.60 s

library contains Probe-B rollouts for efficient evaluation, but no Probe-B feature participates in the estimator.

E TARGET-WISE RESULTS AND VARIABILITY

Table 8 reports sample standard deviation across the five held-out targets. This quantifies target-to-target variation, not simulator noise: fixed default, retrieval, KNN, and both ridge estimators are deterministic once the library, fill seed, and evaluation sample are frozen. We do not treat 31 frames or up to 100,000 particles as independent replicates.

Table 7: Exact probe values. $\mathbf{f}$ is the force vector per selected particle; contact boxes are reported as center $\mathbf{c}$ / half-extent $\mathbf{h}$. One MPM step is 0.1 ms. Global Pillow and Ficus probes select the full physical-state population. The duration row preserves the nominal force–time integral of $\vec{x}_{standard}$; the two local-contact rows preserve total force and impulse relative to each other.

(a) Force and contact specifications.

Protocol	Object / role	$\mathbf{f}$	Contact $\mathbf{c}/\mathbf{h}$
`standard_x`	Pillow A	$(-0.18, 0, 0)$	global
`standard_y`	Pillow B-direction	$(0, -0.18, 0)$	global
`strong_x`	Pillow B-magnitude	$(-0.36, 0, 0)$	global
`duration_x_fine`	Pillow B-duration	$(-0.0009, 0, 0)$	global
`local_center_x`	Pillow contact ref.	$(-0.18, 0, 0)$	$(1, 1, 1.28)$/ $(0.16, 0.16, 0.16)$
`shifted_ contact_x`	Pillow B-location	$(-0.226245, 0, 0)$	$(1, 1.12, 1.28)$/ $(0.16, 0.16, 0.16)$
`standard_x/y`	Ficus A / B-direction	$(-0.18, 0, 0)$/ $(0, -0.18, 0)$	global
`standard_x/z`	Vasedeck A / B-direction	$(-0.09, 0, 0)$/ $(0, 0, -0.09)$	$(1.05, 0.9, 0.92)$/ $(0.2, 0.14, 0.2)$

(b) Timing, selected population, and exports.

Protocol	Steps / duration	Selected N	Exports
`standard_x`	1 / 0.1 ms	695,984	31 at 20 ms
`standard_y`	1 / 0.1 ms	695,984	31 at 20 ms
`strong_x`	1 / 0.1 ms	695,984	31 at 20 ms
`duration_x_fine`	200 / 20 ms	695,984	121 at 5 ms
`local_center_x`	1 / 0.1 ms	94,657	31 at 20 ms
`shifted_contact_x`	1 / 0.1 ms	75,309	31 at 20 ms
`standard_x/y`	1 / 0.1 ms	171,553	31 at 40 ms
`standard_x/z`	1 / 0.1 ms	4,464	31 at 20 ms

Interpreting the time scales. The actuation duration and export interval describe different parts of the experiment. A one-step, 0.1 ms pulse is shorter than the 20 ms interval between standard exports, so the exported sequence primarily records the subsequent response rather than resolving the applied force history. The known input is specified by the simulator schedule, not reconstructed from these sampled frames. In the duration diagnostic, 5 ms exports instead resolve the 20 ms loading period. Thus, frame count alone is not a measure of temporal resolution; it must be read together with the horizon and actuation schedule.

Table 8: Pillow mean $\pm$ sample standard deviation across five held-out targets. Trajectory columns are in units of 10^{-5}; lower is better.

Method	Joint error (%)	Probe A	B-direction	B-magnitude
Fixed default	27.313 ± 17.577	368.195 ± 248.079	350.172 ± 240.806	440.730 ± 273.547
Response nearest	6.746 ± 2.550	37.219 ± 11.919	35.328 ± 11.008	51.876 ± 14.660
Response KNN	2.374 ± 1.065	14.491 ± 8.842	14.119 ± 8.791	20.203 ± 9.086
Global ridge	2.449 ± 1.594	17.476 ± 11.528	16.874 ± 10.905	24.660 ± 16.290
Local ridge (ours)	$\mathbf{1.130 \pm 0.605}$	$\mathbf{4.498 \pm 1.965}$	$\mathbf{4.424 \pm 1.892}$	$\mathbf{7.786 \pm 3.858}$
Parameter-nearest oracle	4.447 ± 0.628	62.827 ± 24.337	59.837 ± 24.313	71.438 ± 24.884

Table 9: All Pillow held-out targets and methods. e_θ is joint parameter error in percent; trajectory columns are in units of 10^{-5}. No target or method outcome is omitted.

Target (s_E, s_ρ)	Method	e_θ	Probe A	B-direction	B-magnitude
$(0.60, 1.65)$	Fixed default	53.030	715.049	690.177	814.549
	Resp. nearest	7.197	28.205	27.441	40.864
	Resp. KNN	1.417	28.593	28.326	29.045
	Global ridge	1.496	28.680	28.274	29.257
	Local ridge	**0.224**	**1.618**	**1.574**	**1.846**
	Param.-nearest oracle	4.167	91.944	90.996	93.126
$(0.70, 1.20)$	Fixed default	29.762	349.973	325.707	401.730
	Resp. nearest	3.571	56.453	53.032	62.402
	Resp. KNN	2.113	9.326	9.476	16.547
	Global ridge	**0.657**	6.695	6.233	**7.580**
	Local ridge	1.371	**5.592**	**5.647**	10.026
	Param.-nearest oracle	3.571	56.453	53.032	62.402
$(0.80, 1.60)$	Fixed default	31.250	475.247	454.395	577.379
	Resp. nearest	4.687	26.407	24.899	32.067
	Resp. KNN	1.812	6.405	6.076	10.026
	Global ridge	3.354	11.735	12.082	21.260
	Local ridge	**0.993**	**3.457**	**3.576**	**6.191**
	Param.-nearest oracle	4.687	26.407	24.899	32.067
$(0.95, 1.35)$	Fixed default	15.595	248.238	231.909	323.815
	Resp. nearest	8.967	37.628	35.814	58.202
	Resp. KNN	2.353	10.792	10.203	14.869
	Global ridge	2.039	9.118	8.857	15.126
	Local ridge	**1.881**	**6.573**	**6.340**	**11.588**
	Param.-nearest oracle	5.263	75.183	71.076	81.811
$(1.05, 1.10)$	Fixed default	6.926	52.467	48.675	86.176
	Resp. nearest	9.307	37.403	35.456	65.846
	Resp. KNN	4.173	17.339	16.512	30.530
	Global ridge	4.697	31.152	28.924	50.079
	Local ridge	**1.178**	**5.249**	**4.982**	**9.279**
	Param.-nearest oracle	4.545	64.145	59.183	87.783

Reading parameter and response errors together. On $(0.60, 1.65)$, parameter-nearest has lower parameter error than response-nearest (4.167% versus 7.197%), yet higher B-direction error (90.996 versus 27.441 in the table's units). Proximity in parameter space therefore need not rank candidates in the same order as response fidelity. The frozen Probe-B test evaluates this distinction directly.

The per-target table also exposes the primary-protocol exception: on $(0.70, 1.20)$, global ridge has lower parameter and B-magnitude errors, while local ridge remains lower on Probe A and B-direction. The aggregate advantage is therefore a mean result, not dominance on every target and metric.

F Duration and Contact-Shift Diagnostics

The fine duration experiment reuses the five frozen $\vec{x}_{standard}$ estimates and changes only the force time profile. It resolves the 20 ms actuation with 5 ms exports (121 states over 0.6 s), rather than the 20 ms exports used by the main benchmark. The 0.0009-per-particle force is applied for 200 internal steps, giving the same nominal force–time integral as the 0.18 force applied for one step. The two target protocols are measurably distinct (mean target-to-target protocol trajectory RMSE 8.998×10^{-5}).

Table 10: Five-target duration-shift result, mean $\pm$ sample standard deviation; values are $\times 10^{-5}$. Parameters are estimated only from $\vec{x}_{standard}$.

Method	Trajectory RMSE	Curve RMSE
Response KNN	8.224 ± 5.127	7.886 ± 9.484
Global ridge	9.895 ± 6.541	9.882 ± 9.230
Local ridge	$\mathbf{2.503 \pm 1.074}$	$\mathbf{0.978 \pm 0.270}$

For contact location, the existing completed experiment is deliberately reported as a single-target diagnostic, not as five-target evidence. On the predeclared $(0.95, 1.35)$ target, `local_center_x` and `shifted_contact_x` use equal total force and impulse; the latter shifts the contact-box center by 0.12 along y and raises per-particle force to compensate for the smaller selected set. This is a location-only comparison against the local-center reference, not against global $\vec{x}_{standard}$.

Table 11: Single-target contact-location diagnostic; values are $\times 10^{-5}$. The same Probe-A parameter estimate is frozen for both contacts.

Contact	Method	Trajectory RMSE	Curve RMSE
Local center	Global ridge	4.115	3.254
Local center	**Local ridge**	**0.281**	**0.221**
Shifted	Global ridge	4.115	3.255
Shifted	**Local ridge**	**0.281**	**0.221**

Local ridge wins on all five duration targets and on both contact locations for the diagnostic target. The duration result supports a five-target protocol-shift claim; the contact result supports only a controlled representative demonstration and is retained with that limitation to avoid selective overstatement.

G Multiple Splits and Numerical Tests

Local ridge has the lower mean on every split, while the strict target-wise gate passes only two. Probe-B resimulation passes all six split–protocol gates. We retain both facts to distinguish average performance from universal dominance.

The fill and grid experiments separate random realization from a changed numerical model. Random fill seeds preserve the same fill specification and give 1.19% mean error. Changing fill density or n_{grid} changes the response distribution and produces 42.61% and 38.30% error with the original library. Rebuilding matched libraries restores 0.99% and 0.75%, respectively. Equal-total-force normalization alone leaves fill mismatch at 43.30%, so the failure cannot be attributed only to the number of selected particles or total impulse.

Table 12: Three independent five-target splits. The strict gate requires a lower mean than the strongest non-oracle method and at least four of five target-wise non-inferior results.

Split	Local (%)	Strongest non-oracle (%)	Non-inferior	Gate
1	1.39	2.60	3/5	fail
2	1.28	3.47	4/5	pass
3	0.76	4.30	5/5	pass

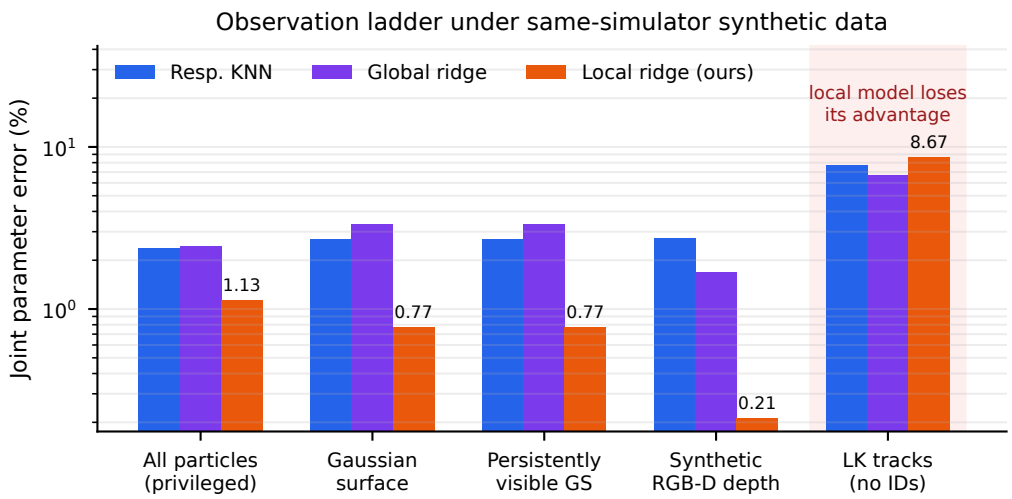


Figure 4: Synthetic observation ladder. Clean surface and depth observations preserve useful calibration evidence under the same simulator, but the local estimator loses its advantage with ID-free LK tracks. The apparently lower clean-depth error is an empirical result for this split and feature construction, not a general claim that less information is intrinsically better.

H Diagnostic Boundaries

H.1 Observation Ladder

The MPM experiments above use simulator-exported particle state, which no camera can provide. This diagnostic addresses the observation component of question (3): by removing one privilege at a time we locate *which* property of the observation the method actually depends on, so that the distance to a deployable visual front end is a measured quantity.

The main estimator receives all exported particles, which is privileged state. To localize the gap to deployable vision, we progressively replace it with Gaussian surface particles, persistently visible Gaussians, clean synthetic RGB-D depth, and ID-free LK trajectories. Fig. 4 reports the result under the same simulator and target split. Local-ridge error is 1.13% with all particles, 0.77% with Gaussian surface or persistently visible Gaussians, and 0.21% with clean synthetic depth. These values show that the selected descriptors can be recovered from clean surface observations in this controlled setting; they do not establish real RGB-D performance.

The conclusion changes at the final step. With LK tracks and no persistent particle identity, local-ridge error rises to 8.67%, worse than global ridge (6.73%) and KNN (7.76%). The current bottleneck is therefore not simply whether internal fill particles are observed; robust correspondence and observation noise are unresolved. Synthetic-depth perturbations reinforce this boundary: local error grows from 0.21% clean to 2.65% at noise level 0.001, 16.25% at 0.005, and 28.26% at 0.01.

H.2 Alternative Inverse Routes and External Diagnostics

The previous experiments compare estimators that share our library. This diagnostic tests the library route against two families that solve the inverse problem from different information—per-scene optimization against a video, and learned prediction from appearance or passive dynamics—and then distinguishes a structural absolute-scale ambiguity in passive dynamics from the observed discriminative range of the released predictors. Because their evidence and parameter semantics differ, these routes define boundaries rather than an additional same-information ranking.

Direct optimization from one Probe-A video. To test whether the response library is necessary, we optimize (s_E, s_ρ) directly against a single simulator-rendered RGB-D Probe-A video using CMA-ES. The optimizer receives no particle IDs and uses 30 forward calls per target with three predeclared

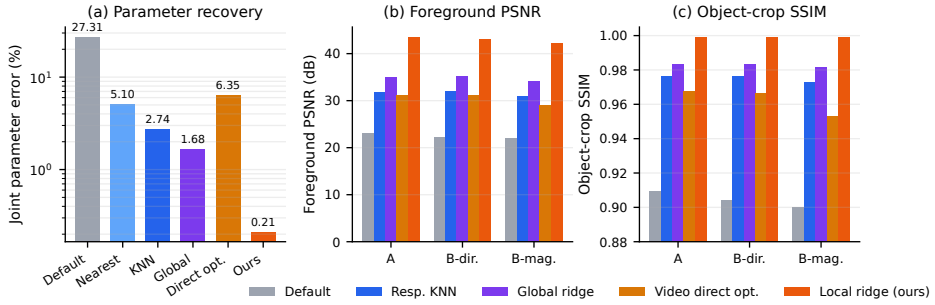


Figure 5: Diagnostic comparison with fixed-budget per-scene video optimization. (a) Parameter recovery; (b) foreground PSNR; (c) object-crop SSIM. Both routes use simulator-generated RGB-D Probe-A evidence, and Probe B is excluded from fitting. The response-library estimator is more accurate under this budget, while requiring an offline object-specific library.

seeds. Parameters are then frozen and evaluated on $\vec{y}_{standard}$ and $\vec{x}_{strong}$. Fig. 5 shows that fixed-budget direct optimization improves substantially over default parameters (6.35% versus 27.31% joint error), but remains behind response-library local calibration (0.21% on the same synthetic-depth observation track). On Probe A, local calibration achieves 43.49 dB PSNR versus 31.10 dB for CMA-ES; the ordering persists on both Probe B protocols. This comparison has complementary computational costs: our method amortizes a 54-simulation object-specific library, whereas direct optimization incurs per-target simulator calls.

Visual-prior and passive-video systems. We additionally run PhysGM and ReconPhys as diagnostic external comparisons. Their native physical parameterizations, training distributions, and solvers differ from our MPM scales, so their outputs are adapted to the same initial asset and common downstream MPM rather than treated as same-information baselines. PhysGM observes four static views of the pillow scene and returns one material prediction for five visually identical targets; its common-adapter Probe-A response error is 11.42× ours and worse than the uncalibrated default on all five targets. ReconPhys features are mapped to our material scales by a candidate-only MPM adapter; this comparison reports a multiplicative parameter error of 20.29%, versus 34.99% for default, using the log-domain metric defined in Appendix L. Its predictions beat default response on $4/5$ targets, but its Probe-A and unseen-Probe-B video errors remain approximately 2.6× and 2.9× ours. These results support the value of known dynamic excitation in this controlled scene.

Passive scale ambiguity and observed discrimination. The passive spring–mass model has a common absolute-scale ambiguity, which is distinct from the empirical accuracy of its predicted parameter ratios. In the spring–mass formulation used by ReconPhys, the released implementation computes spring and damping forces that are linear in k and d, adds gravity as $m\mathbf{g}$, and integrates $\mathbf{v} \leftarrow \mathbf{v} + \mathbf{F}\Delta t/m$; the released configuration resolves ground contact by a mass-independent kinematic projection. The rescaling $(m, k, d) \mapsto (\alpha m, \alpha k, \alpha d)$ therefore leaves every particle trajectory—and hence every rendered frame—unchanged. Under these passive-drop assumptions, absolute mass is not identifiable; the invariance leaves ratios such as k/m unchanged and therefore does not establish whether those ratios can be recovered. This invariance arises from the observation setting and parameterization rather than from an implementation error. A mass-dependent penalty contact model, for example, could break this scaling invariance.

Separately, Fig. 6 measures the empirical discriminative range on our five held-out targets. Panel (a) shows that the true ratio s_E/s_ρ spans 162% and our estimates span 163%, whereas the released ReconPhys predictions span only 4.5% in k/m and PhysGM returns a single prediction because its four static views are identical across targets. Panel (b) adds the per-parameter view: the released outputs sit at the top of the reported training range for mass (5.55–5.88 within $[0.2, 6.0]$) and damping (4.83–4.97 within $[0.1, 5.0]$), and friction is negative on all five targets although the method defines $f \geq 0$. These observations suggest limited target-specific discrimination in this setting, but are not a consequence of the common-scale invariance. A prescribed force with known absolute magnitude

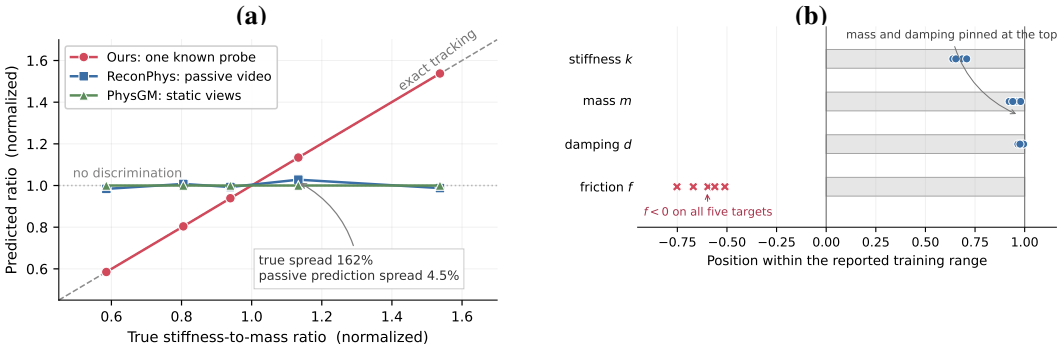


Figure 6: Passive-observation diagnostics on five held-out targets. (a) Discriminative range after per-series mean normalization. (b) Released ReconPhys native outputs relative to its reported training ranges; crosses are inadmissible. These panels diagnose range and saturation under different parameterizations, not like-for-like parameter accuracy.

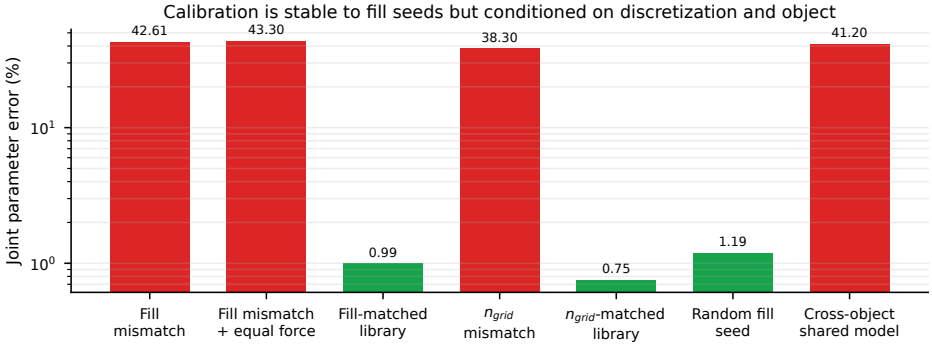


Figure 7: Calibration boundaries for local ridge. Red bars violate the library contract; green bars preserve or rebuild it. Equalizing total force does not fix fill mismatch, while fill- and grid-matched libraries restore accuracy.

breaks that symmetry through the acceleration term $\mathbf{F}_{\text{ext}}/m$; the parameter and held-out response results test the practical information gained from our probes.

H.3 Discretization, Object, and Identifiability Boundaries

Sec. 3.2 constructs each response library under a fixed object-specific simulation contract. To characterize the estimator's dependence on this contract, we vary one component at a time, evaluate the resulting mismatch using the original library, and, where applicable, rebuild a contract-matched library. These controlled comparisons distinguish variations tolerated by the frozen library from those that require library reconstruction.

Fig. 7 summarizes these controlled stress tests. Changing fill density while retaining the original library increases local-ridge error to 42.61%; normalizing the total applied force does not resolve it (43.30%). Rebuilding a fill-matched library restores 0.99%. Changing MPM grid resolution gives 38.30% error, while a grid-matched library restores 0.75%. In contrast, random fill seeds at the same specification remain stable (1.19% mean, 2.45% maximum). Finally, leave-one-object-out use of a shared estimator fails with 38–45% error. The method is therefore robust to stochastic fill realization but conditioned on the object and numerical discretization used to construct the library.

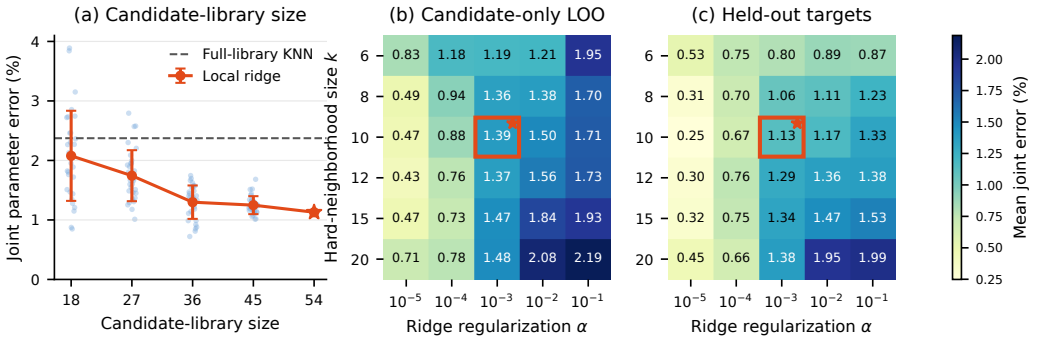


Figure 8: Offline estimator sensitivity; lower is better. (a) Candidate-library size with the local-ridge configuration frozen at $k = 10$, $\alpha = 10^{-3}$. For 18/27/36/45 candidates, dots are the five-target mean from each of 30 predeclared nested random subsets and error bars show standard deviation across subset means; the full 54-candidate library is unique. The dashed line is the full-library response-KNN error. (b) Candidate-only leave-one-out sensitivity over all 54 candidates. (c) Sensitivity on the five held-out targets, used only as a post-hoc diagnostic. The red box and star mark the frozen paper configuration; neither heatmap is used to retune the reported main result.

I Offline Estimator Sensitivity

We test two estimator design choices using only the frozen response-feature tables; no MPM rollout, rendering, or new response generation is performed. The audit covers 2,375 offline predictions: 605 library-size trials, 150 held-out (k, α) trials, and 1,620 candidate-only leave-one-out trials. Every result is finite, no estimator fallback is triggered, and the 54 candidates have zero parameter-pair overlap with the five held-out targets.

Increasing the library from 18 to 54 candidates reduces mean joint error from 2.078% to 1.130%; the standard deviation caused by subset selection contracts from 0.757% at 18 candidates to 0.151% at 45 candidates and vanishes for the unique full library. Thus the larger library improves both accuracy and coverage stability. Even the 18-candidate mean remains below the 2.374% full-library response-KNN baseline, indicating that the local model's advantage is not confined to the densest library.

The frozen $k = 10$, $\alpha = 10^{-3}$ configuration exactly reproduces the reported 1.130% held-out mean. Weak regularization is consistently more accurate in these nearly noise-free synthetic features: candidate-only leave-one-out is lowest at $k = 12$, $\alpha = 10^{-5}$ (0.426%), while the held-out diagnostic is lowest at $k = 10$, $\alpha = 10^{-5}$ (0.249%). We retain the predeclared main configuration rather than retrospectively replacing it. Across all 30 held-out grid cells, mean error remains below the 2.374% response-KNN result, supporting a method-family advantage while also revealing sensitivity to regularization strength.

J Observation Ladder Details

The observation ladder holds the asset, target split, simulator, and Probe A fixed. Its five stages are:

1. all simulator particles (privileged state);
2. Gaussian-associated surface particles;
3. Gaussian-associated particles that remain visible;
4. clean synthetic RGB-D depth projected into 3D;
5. LK optical-flow tracks without persistent simulator IDs.

The first four maintain exact or clean geometric correspondence supplied by the synthetic pipeline. The last is closer to a practical image tracker and breaks that correspondence. Consequently, the ladder identifies a likely bottleneck but is not a substitute for a captured RGB-D benchmark.

K OBJECT-SPECIFIC MATERIAL CALIBRATION.

Fig. 9 repeats calibration independently for three geometrically distinct assets. Local ridge obtains 1.13% error on Pillow, 0.63% on Ficus, and 1.21% on Vasedeck. On the Vasedeck $\vec{x}_{standard} \to \vec{z}_{standard}$ test, its Probe-B trajectory RMSE is 1.37×10^{-5}, versus 3.11×10^{-5} for the strongest non-oracle response-nearest result, and it is non-inferior on all five targets. Ficus yields 0.63% scale error versus 3.71% for KNN/global ridge and 27.31% for the default material. Across the three evaluated assets, the same calibration procedure remains effective when instantiated with an object-specific response library.

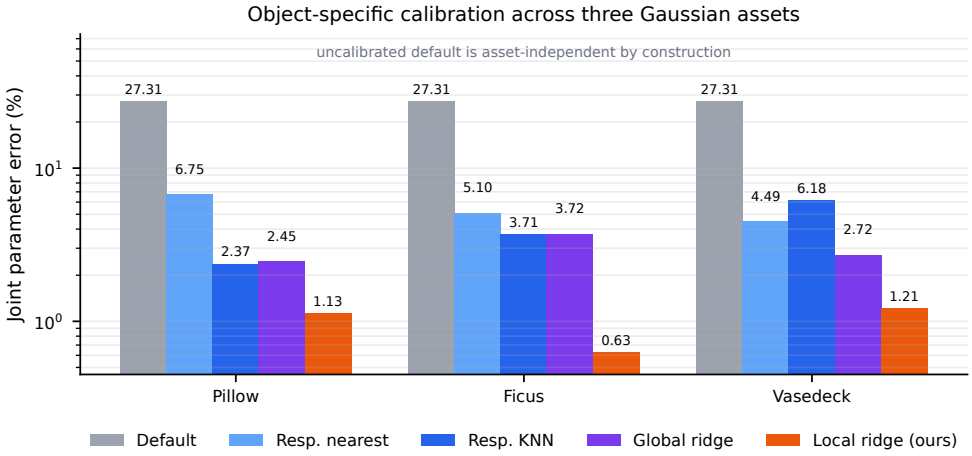


Figure 9: Object-specific material calibration. The same estimator form is used, but each Gaussian asset has its own response library. Local ridge remains the best deployable method on Pillow, Ficus, and Vasedeck.

Rendered response follows the dynamics metrics. Across six pillow protocols, local ridge improves mean foreground PSNR from 55.50 dB (global ridge) to 59.53 dB and object-crop SSIM from 0.9968 to 0.9982. On Ficus, its mean foreground PSNR is 41.87 dB versus 26.80 dB for global ridge. All reported PSNR and SSIM values compare simulator-rendered predictions with simulator-rendered targets and therefore do not measure agreement with real video.

Fig. 10 visualizes the objectively selected median Ficus case: we sort the five held-out targets by local-ridge Probe-B trajectory RMSE and show the middle one, $(s_E, s_\rho) = (1.05, 1.10)$. This rule does not inspect PSNR/SSIM or baseline ordering. Averaged over exported frames, local ridge reaches 41.78 dB/0.997 foreground PSNR/SSIM, versus 22.99 dB/0.846 for global ridge, 22.65 dB/0.775 for response KNN, 21.14 dB/0.722 for response nearest, and 21.33 dB/0.817 for the default material. The residual maps show that the remaining disagreement is concentrated on leaf boundaries and thin branches, where small trajectory errors cause large image-space differences.

L ALTERNATIVE-ROUTE DIAGNOSTICS

The ReconPhys adapter comparison uses a log-domain multiplicative parameter error,

$$e_{\text{mult}} = 100 \left[\exp\left(\sqrt{\frac{1}{2} \sum_{q \in \{E,\rho\}} \log^2 \frac{\hat{s}_q}{s_q^\star}} \right) - 1 \right]. \tag{14}$$

We average this quantity over the five targets. The fixed default $(s_E, s_\rho) = (1, 1)$ gives 34.99% under this metric and 27.31% under the relative-error metric e_θ used in the main table. These percentages use different definitions and are not directly interchangeable.

PhysGM outputs a single parameter set for the five visually identical pillow targets; ReconPhys is calibrated into the available candidate range before common-adapter evaluation. These adaptations

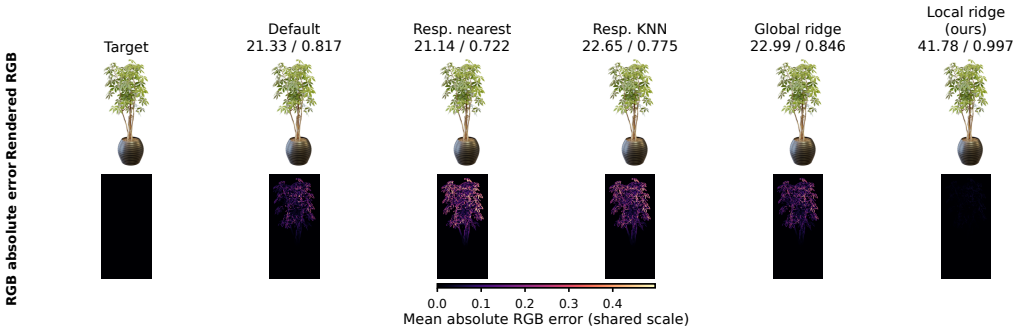


Figure 10: Qualitative held-out Probe-B prediction on the median-error Ficus target ($s_E = 1.05$, $s_\rho = 1.10$, $\vec{y}_{standard}$). Rows show the final rendered frame and its mean absolute RGB error on one shared scale. Headers report target-specific foreground PSNR in dB / foreground SSIM averaged over all 30 rendered frames, using the Ficus protocol in Appendix C. The target is selected only by the median local-ridge trajectory RMSE among all five held-out targets, before inspecting these visual scores. This is an independently calibrated object-specific loop, not cross-object transfer.

Table 13: Diagnostic comparisons with methods using different information or parameter semantics. These rows are not a same-information leaderboard. Response comparisons retain their stated metric: multiplicative factors refer to the corresponding common-adapter response/video error, whereas PSNR entries compare dB values. The ReconPhys parameter entry uses e_{mult} defined above.

Route	Input	Parameter error	Response comparison	Interpretation boundary
PhysGM	four static views	5279.88%	11.42× (Probe A)	predicted Fabric/high stiffness; scene and parameter semantics may be out of distribution
ReconPhys adapter	rendered video features	20.29% (e_{mult})	2.6× A; 2.9× B	improves over default on 4/5 targets; native solver/parameterization differs
CMA-ES direct opt.	synthetic RGB-D Probe A	6.35%	PSNR 31.10 vs. 43.49 dB	30 calls per target, three seeds; fixed-budget, not exhaustive optimization
Local ridge (ours)	synthetic depth descriptor + library	0.21%	1.0×	requires a 54-candidate offline library for the same object/discretization

Table 14: Target-wise CMA-ES direct-video optimization, mean ± sample standard deviation over predeclared seeds 6401, 6402, and 6403. Joint error is in percent and trajectory errors are $\times 10^{-3}$. Each seed uses 30 Probe-A simulator calls; Probe B is evaluation-only.

Target (s_E, s_ρ)	Joint error	Probe A	B-direction	B-magnitude
$(0.95, 1.35)$	7.58 ± 6.80	0.371 ± 0.222	0.357 ± 0.218	0.580 ± 0.437
$(0.60, 1.65)$	8.06 ± 6.40	0.420 ± 0.310	0.433 ± 0.322	0.637 ± 0.471
$(1.05, 1.10)$	5.10 ± 3.74	0.293 ± 0.173	0.277 ± 0.165	0.411 ± 0.225
$(0.80, 1.60)$	8.49 ± 5.62	0.289 ± 0.198	0.291 ± 0.200	0.532 ± 0.365
$(0.70, 1.20)$	2.50 ± 1.93	0.165 ± 0.150	0.157 ± 0.136	0.223 ± 0.170

are useful for testing the proposition that known dynamic excitation carries information missing from appearance or passive input, but they do not establish across-paper superiority. The CMA-ES experiment is closer to the current target protocol because it directly fits Probe-A synthetic video and freezes parameters before Probe B; it therefore serves as the main alternative inverse baseline.

M Reproducibility and Claim Boundaries

The released experiment contracts record the asset, 54 candidates, held-out targets, force definitions, contact regions, export times, particle subset, data split, feature standardization, neighborhood size, regularization, simulator configuration, and metric implementation. Result packages contain per-case CSV/JSON outputs and completion, leakage, missing-frame, NaN, and duplicate audits. No result is interpreted as real-material accuracy because target parameters and responses are generated by the same simulator.

The current evidence supports: (i) same-information local-calibration gain; (ii) object-specific Probe-A-to-B transfer; (iii) same-simulator visual-response fidelity; and (iv) an explicit characterization of observation and numerical-conditioning failures. It does not support real RGB/RGB-D identification, real force-feedback closure, sim-to-real transfer, a cross-object shared estimator, unknown-fill inversion, or a complete physical representation.

N Isaac Sim Contact-Based FEM Validation

Assets and estimator. This experiment rebuilds the response library in Isaac Sim 4.5 (PhysX extension 106.5.7) on an NVIDIA L20. The 54 candidates and five held-out targets share surface geometry, rendering material, and FEM topology (11,203 simulation nodes). Only homogeneous effective E and ρ vary: these assets do not explicitly model different fibers, granular fillings, internal layers, or hidden rigid cores. The reference material is $E_0 = 25{,}000\,\mathrm{Pa}$, $\rho_0 = 60\,\mathrm{kg/m^3}$, so that $s_E = E/E_0$ and $s_\rho = \rho/\rho_0$. The five target pairs in Table 15 are absent from the candidate library. The original five-dimensional response descriptor, candidate-only standardization, hard top-10 neighborhood, and local ridge ($\alpha = 10^{-3}$) are unchanged. Target parameters and all Probe-B responses are withheld during estimation. All 59 Probe-A rollouts and 30 Probe-B comparison rollouts are generated anew under the same implementation; no historical Probe-A records are mixed into this batch. Ten additional response-nearest Probe-B rollouts use the same locked configuration and the candidates selected from Probe A, without refitting. Response-nearest selects by response-feature distance, not target-parameter distance; its selected material is frozen before either Probe-B evaluation.

Boundary conditions and actuation. The lowest 10 mm band (785 nodes) is fixed at runtime; there is no ground plane. Each rollout starts from the same construction and reset procedure without an equilibrium-settling phase. Local time zero denotes the recorded post-reset state, not the pre-integration asset. A dynamic 1 kg cylinder of radius 40 mm and height 20 mm is placed at the top of the pillow with zero commanded initial velocity. Let $p(t) = F_{\max}\sin(\pi t/0.3)$ for $0 \le t < 0.3\,\mathrm{s}$ and zero otherwise. The commanded external force on the cylinder is $(rp(t), 0, mg - p(t))$, including gravity compensation; it is not a measured contact force or a prescribed indentation trajectory. Probe A uses $(F_{\max}, r) = (5\,\mathrm{N}, 0)$; the oblique Probe B uses $(5\,\mathrm{N}, 0.5)$, and the larger normal Probe B uses $(7.5\,\mathrm{N}, 0)$. Thus the oblique test adds a lateral component while retaining the normal component, rather than rotating a force of constant total magnitude. The force is sampled at each substep midpoint. The TGS solver uses 32 iterations and a fixed 3200 Hz step rate. Poisson's ratio, friction, elasticity damping, and damping scale are held at 0.3, 0.6, 0.02, and 0.03, respectively. Automatic contact offsets are retained consistently and recorded from the runtime.

Evaluation and aggregation. Each 0.6 s rollout exports 31 states at 0.02 s intervals. We compare all corresponding native FEM nodes, including the fixed nodes, at the recorded times without event alignment or rigid-motion removal:

$$\mathrm{RMSE}_{\mathrm{FEM}} = \sqrt{\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\mathbf{x}_{t,i}^{\mathrm{pred}} - \mathbf{x}_{t,i}^{\mathrm{GT}}\right\|_2^2}. \tag{15}$$

Here $T = 31$ and $N = 11{,}203$; errors are converted to mm. Ground truth is a rollout with the preset target material in the same FEM simulator. The single Probe-A estimate for each target is frozen and reused in both Probe-B protocols. Each target–protocol–method condition has one rollout; standard deviations below are sample standard deviations across the five materials ($n = 5$, $\mathrm{ddof} = 1$), not

Table 15: FEM parameter recovery from Probe A alone. E is in kPa and ρ in kg/m^3. Joint error averages the two relative parameter errors. All displayed estimates use the original local-ridge estimator. The final two rows report means and sample SDs across five materials ($\mathrm{ddof} = 1$).

Target	$E^\star$	$\hat{E}$	$\rho^\star$	$\hat{\rho}$	E error (%)	ρ error (%)	Joint error (%)
H001	15.00	15.190	99	99.536	1.266	0.541	0.903
H002	17.50	17.667	72	72.634	0.957	0.880	0.918
H003	20.00	20.148	96	96.744	0.742	0.776	0.759
H004	23.75	24.307	81	82.650	2.347	2.037	2.192
H005	26.25	26.440	66	66.291	0.722	0.440	0.581
Mean					1.207	0.935	1.071
Sample SD					0.674	0.641	0.641

Table 16: Frozen Probe-B prediction in FEM: per-target full-node trajectory RMSE (mm). Each panel reports all five targets, their arithmetic mean, and sample SD ($n = 5$, $\mathrm{ddof} = 1$). SD describes across-material variation, not a confidence interval.

Oblique pulse			
Target	Default	Response nearest	Local ridge
H001	0.468374	0.031386	0.007059
H002	0.203547	0.023797	0.003372
H003	0.263283	0.017790	0.002199
H004	0.110400	0.028777	0.005327
H005	0.016707	0.029431	0.001730
Mean	0.212462	0.026236	0.003937
Sample SD	0.170946	0.005488	0.002230
Larger normal pulse			
Target	Default	Response nearest	Local ridge
H001	0.476414	0.032134	0.007610
H002	0.209198	0.024949	0.003238
H003	0.265468	0.017925	0.002534
H004	0.111388	0.030680	0.005697
H005	0.017061	0.030420	0.002278
Mean	0.215906	0.027221	0.004271
Sample SD	0.173803	0.005872	0.002304

repeat-run uncertainty or confidence intervals. This convention applies to both parameter errors and trajectory errors throughout the FEM results.

Table 15 gives mean relative errors of 1.207% for E and 0.935% for ρ. Local ridge improves joint error over response-nearest on each held-out target. Joint parameter errors (mean $\pm$ sample SD, %) are 27.3126 ± 17.5771 for default, 6.9828 ± 1.5829 for response-nearest, and 1.0706 ± 0.6412 for local ridge. Table 16 shows lower local-ridge error than both baselines on all five targets in each unseen interaction. Using $1 - \overline{\mathrm{RMSE}}_{\mathrm{local}}/\overline{\mathrm{RMSE}}_{\mathrm{baseline}}$, the reductions are 98.15% and 98.02% relative to default, and 84.99% and 84.31% relative to response-nearest, for oblique and larger normal pulses. These are ratios of across-target means, not means of per-target percentages. Figure 11 shows the temporal errors for all five targets, including the recovery interval after the pulse ends.

Contact checks and numerical scope. Before library construction, dynamic-probe motion, collision-surface proximity, and a sustained reconstructed dynamics residual support actual contact; the residual is an engineering diagnostic, not a force-sensor measurement. Two independent C038 resets differ by 0.000755 mm trajectory RMSE (0.214% of its response RMS) at 3200 Hz. However, cross-step initialization and actual-input alignment at 3200/6400 Hz did not meet the predeclared budgets, and the separate $\leq 2\%$ numerical-convergence criterion has not passed. This experiment therefore measures calibration and prediction accuracy relative to a fixed numerical configuration;

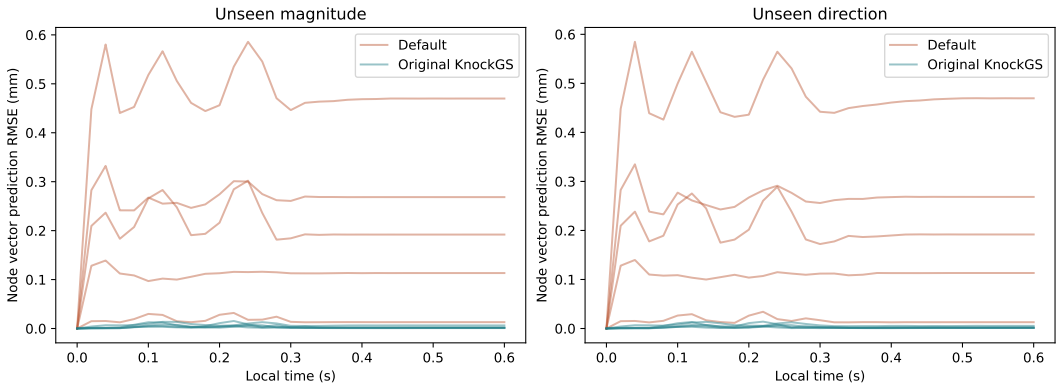


Figure 11: FEM-node prediction error over time for all five held-out materials. Left: larger normal pulse; right: oblique pulse. Each line is one target, with default and calibrated predictions shown in different colors. At each time, the plotted value is the vector RMSE over all nodes; Table 16 aggregates squared errors over the whole trajectory. Local time is measured from the recorded post-reset state. The curves are individual-target values, not means or uncertainty bands; across-material sample SDs are reported in Table 16.

its micrometer-scale errors do not establish physical accuracy at that scale. Reset-induced motion is included in the recorded responses, with no equilibrium claim. The evidence concerns effective homogeneous material parameters under privileged node access; it does not demonstrate a camera-based GS front end, explicit internal filling identification, robot grasping, or zero-shot cross-solver transfer.

O Probe Selection and Complementary Responses

O.1 Protocol and Candidate-Only Selection

The calibration pool consists of X = `standard_x`, Z = `standard_z`, and L = `dynamic_local_x` on the same Pillow asset, 54 candidates, and five held-out targets as the main benchmark. Each interaction starts from the same initial object state. The global probes X and Z match force magnitude and duration; L changes the loaded region and aggregate force. A budget counts interactions, not equal impulse or energy. Probe B remains `standard_y` and `strong_x` throughout.

For each probe subset, we concatenate its standardized five-dimensional descriptor blocks and apply the original unweighted local ridge with ten neighbors, regularization 10^{-3}, and an unregularized intercept. In each of 54 leave-one-out folds, standardization, neighborhood retrieval, and parameter clipping use only the remaining 53 candidates. The subset with lowest mean candidate error is fixed separately for each budget before evaluating target predictions. This selects Z, $X + Z$, and $X + Z + L$ at budgets one, two, and three. No target response or Probe-B score enters selection. The study reuses existing held-out targets and verified cached responses; it is a retrospective extension, not a new prospective blind test or an online, state-dependent next-contact policy.

O.2 Complete Subset Results

Table 17 reports every nonempty subset. Parameter variability is the sample standard deviation across five materials ($n = 5$, $\mathrm{ddof} = 1$), not a confidence interval over random seeds. Trajectory errors use the main benchmark's fixed particle subset and initial-bounding-box normalization. $X + Z$ gives the best mean prediction on both unseen probes, whereas $X + Z + L$ gives the best parameter recovery. The local probe alone performs poorly; additional interactions are useful when their responses constrain complementary parameter directions.

Table 17: All calibration-probe subsets. Target results are mean ± sample SD across five targets (ddof = 1). Probe-B columns report normalized trajectory RMSE in units of 10^{-5}. Candidate LOO uses only library candidates.

Subset	Candidate LOO error (%)	Target parameter error (%)	B direction $\times 10^{-5}$	B magnitude $\times 10^{-5}$
X	1.385	1.130 ± 0.605	4.424 ± 1.892	7.786 ± 3.858
Z	0.860	0.703 ± 0.179	3.551 ± 1.092	4.970 ± 0.939
L	13.595	3.222 ± 2.711	14.231 ± 9.298	23.159 ± 16.631
$X + Z$	**0.602**	0.474 ± 0.156	$\mathbf{2.552 \pm 0.862}$	$\mathbf{3.559 \pm 0.980}$
$X + L$	1.711	0.798 ± 0.303	5.590 ± 4.835	7.560 ± 4.222
$Z + L$	1.093	0.501 ± 0.268	3.659 ± 2.639	4.557 ± 2.466
$X + Z + L$	0.713	$\mathbf{0.298 \pm 0.060}$	3.175 ± 2.811	3.713 ± 2.618

Budget-matched policies. The fixed order is $X \to Z \to L$. Candidate-selected and fixed-order policies therefore coincide at budgets two and three; selection improves the one-probe choice, rather than every budget. The uniform-random comparator is the exact expectation over all subsets of a given size. Its mean parameter errors are 1.685%, 0.591%, and 0.298% at budgets one through three, compared with 0.703%, 0.474%, and 0.298% for candidate-only selection.

O.3 Local Sensitivity Diagnostic

We fit a response Jacobian in $(\log s_E, \log s_\rho)$ around each candidate using its 12 nearest other candidates. A shared feature scale permits comparison across probes. The median smallest singular values for X, Z, and L are 1.497, 3.937, and 0.243; their median condition numbers are 6.009, 2.550, and 24.767. Thus Z better separates the two local parameter directions under this diagnostic, consistent with candidate-only validation. The corresponding values for $X + Z$ are 4.827 and 2.676. Checks with 8 and 16 neighbors accompany the experiment records. These are local geometric sensitivity measures, not Fisher information under a calibrated noise model or a proof of global identifiability.

Evidence accounting. The extension computes 378 candidate leave-one-out predictions, 1,134 local Jacobian fits, and 35 target predictions. It reuses 177 feature traces and 70 full Probe-B outputs after source, parameter, frame-count, and numerical checks; it adds no MPM simulation runs. The main single-X result and its frozen estimates remain unchanged.

P Discussion

The supported claim is controlled, object-specific, discretization-conditioned calibration of two MPM material scales from a known interaction, followed by unseen-interaction prediction in the same simulator family. The experiments do not establish real RGB/RGB-D input, real force feedback, measured material ground truth, sim-to-real transfer, a cross-object shared estimator, fill-unknown inversion, full multi-parameter material identification, damage prediction, or active-probe optimality. The observation ladder uses synthetic observations, the external routes use different native parameterizations, and the passive-identifiability result applies to the analyzed spring–mass/contact formulation; none should be read as a broader real-world or like-for-like accuracy claim.

The measured failures point to concrete future work. The degradation with ID-free tracks motivates robust visual correspondence, while sensitivity to synthetic-depth perturbations calls for uncertainty-aware estimation. Descriptor ambiguity under a fixed Probe A motivates active interaction selection and richer response representations, and the current two-scale state should be extended to richer physical parameterizations. Finally, closing the simulation-to-reality gap requires validation with measured interactions and real object deformation. These are future directions rather than capabilities of the current system.