HETEROPHYS: HETEROGENEOUS PHYSICAL PROPERTY IDENTIFICATION FROM VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering physical properties from videos is essential for predicting object dynamics beyond observed interactions, yet direct identification of spatially heterogeneous physical fields remains underexplored. We present HeteroPhys, a framework for recovering heterogeneous physical parameter fields from multi-view videos of deformable objects. HeteroPhys extracts Lagrangian motion and local deformation histories and represents material properties in constitutive coordinates separating volumetric and deviatoric responses. A hierarchical conditional diffusion model reconstructs fields from coarse structure to fine spatial variation, while permutation-invariant fusion integrates evidence from multiple interactions. The predicted dense fields are further refined through differentiable MPM optimization. We construct a benchmark with ground-truth heterogeneous fields across elastic and plasticine materials. Experiments show that HeteroPhys generalizes to unseen geometries, benefits from additional interaction evidence, and remains predictive under unseen interaction conditions.

1 INTRODUCTION

Understanding how objects respond to forces and interactions is fundamental to physical reasoning, predictive digital twins, and robotic manipulation. Beyond reconstructing geometry and appearance, an intelligent system must infer intrinsic physical properties such as stiffness, compressibility, and yielding behavior to predict dynamics under unseen conditions. Recovering such properties from visual observations is the goal of vision-based system identification.

Recent methods combine visual reconstruction with physical simulation to identify material properties from videos (Jaques et al., 2020; Jatavallabhula et al., 2021; Ma et al., 2022; Li et al., 2023; Cai et al., 2024). Classical formulations typically recover a compact set of parameters for an entire object, while subsequent work improves geometry-aware optimization (Kaneko, 2024), relaxes predefined material priors (Zhao et al., 2025), or extends inverse physics to more challenging observations (Rho et al., 2026). More recent approaches introduce spatially resolved physical representations (Wada et al., 2026; Kim et al., 2026; Yin et al., 2026). However, identifying dense heterogeneous continuum parameter fields and evaluating them against physical ground truth remain comparatively underexplored.

Dense field identification is substantially more ambiguous than estimating a few global parameters. Under limited excitation, different spatial regions and material responses are constrained unevenly, and distinct parameter fields may produce similar observed trajectories (Holzschuh et al., 2023). Consequently, reproducing observed motion does not necessarily imply recovering the physical properties that generated it. *Matching motion is therefore not equivalent to identifying physics.*

We introduce HETEROPHYS, a framework for identifying heterogeneous physical parameter fields from videos. As illustrated in Figure 1, HETEROPHYS shifts from fitting uniform physical properties from a single physical interaction sequence to recovering spatially varying physical properties from multiple physical interactions for prediction under unseen conditions. HETEROPHYS converts multi-view observations into Lagrangian motion and local deformation histories on a shared particle support and expresses material properties in bulk–shear constitutive coordinates that separate volumetric and deviatoric responses. A hierarchical conditional diffusion model first recovers coarse physical structure and then reconstructs finer spatial variation through residual field prediction. To reduce ambiguity from limited observations, a permutation-invariant fusion module integrates evidence from

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.

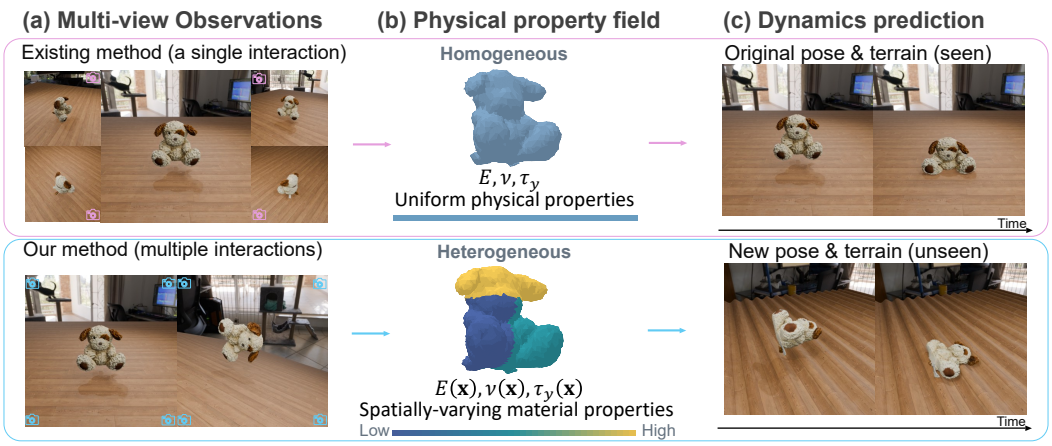

Figure 1: **Comparison with Previous Methods.** Previous visual system-identification settings often infer uniform physical properties from limited interaction evidence and evaluate the future state performance under the same interaction. In contrast, HETEROPHYS integrates complementary evidence from multiple interactions to recover spatially varying physical properties, which are then used to predict dynamics under unseen poses and environments.

one or multiple interactions. At inference time, the diffusion model predicts a dense field proposal, which is projected into the feasible material domain and refined by differentiable MPM against the observed histories before being used under novel interaction conditions.

We evaluate HETEROPHYS directly on heterogeneous elastic and plasticine systems with explicit ground-truth parameter fields spanning homogeneous, piecewise, core–shell, and continuously varying distributions. This enables direct measurement of parameter identification accuracy, analysis of how additional interactions reduce ambiguity, and evaluation under unseen interactions.

Our contributions can be summarized as follows:

- We formulate the problem of physical system identification from multi-view observations under multiple interactions and introduce HETEROPHYS, a diffusion-and-gradient-refinement framework that generates heterogeneous physical property fields with conditional diffusion and further refines them through differentiable MPM simulation and rendering.
- We construct a large-scale dataset with explicit ground-truth heterogeneous physical property fields and diverse interaction conditions, providing a systematic benchmark for physical-field identification from multi-view observations.
- Extensive experiments show that HETEROPHYS accurately recovers spatially varying physical properties, that multiple interactions provide complementary physical evidence that improves field identification, and that the recovered fields generalize to unseen interaction conditions.

2 RELATED WORK

Visual Physical System Identification. Visual system identification estimates physical properties by coupling visual observations with differentiable dynamics (Jaques et al., 2020; Jatavallabhula et al., 2021; Ma et al., 2022). PAC-NeRF (Li et al., 2023) jointly reconstructs geometry and optimizes material parameters using neural radiance fields and MPM, with subsequent work improving Lagrangian particle optimization (Kaneko, 2024). GIC (Cai et al., 2024) introduces motion-factorized 3D Gaussians and a Gaussian-informed continuum for geometry-aware identification, while later methods extend visual inverse physics to spring-based representations (Zhong et al., 2024), monocular observations (Rho et al., 2026), and contact-rich multi-object scenes (Liu et al., 2026). NeuMA (Cao et al., 2024) and MASIV (Zhao et al., 2025) further augment or relax prescribed constitutive models through learned material dynamics. Dynamic Gaussian representations further provide explicit deformation fields that can be used to recover object motion (Cai et al., 2024; Zhao et al., 2025).

In our setting, we construct correspondence-preserving Lagrangian histories on a shared canonical particle support and use them as observations for dense physical-field identification.

Spatially Varying Physical Properties. Recent work has moved beyond homogeneous object-level parameters. PhysTwin (Jiang et al., 2025) reconstructs dense spring-based physical properties through per-instance inverse modeling, while M-PhyGs (Wada et al., 2026) estimates particle-level Young's modulus and density for multi-material objects. EgoPhys (Kim et al., 2026) predicts dense spring-stiffness fields for unseen deformable objects, and PhysCoRe (Yin et al., 2026) infers per-particle elasticity to ground a predictive simulator. Related Gaussian representations also integrate explicit physical dynamics into spatial scene primitives (Xie et al., 2024). We treat the heterogeneous continuum parameter field itself as the identification target, supervise and evaluate it against explicit field-level ground truth, and fuse multiple interactions to reduce ambiguity. The learned conditional inverse model supplies a dense initialization, which bounded differentiable MPM refinement then optimizes against the observed dynamics.

Probabilistic Inverse Physics. Limited observations can admit multiple physically plausible explanations. Score-based methods have therefore been explored for sampling solutions to inverse physics problems (Holzschuh et al., 2023), while video-diffusion priors have been used to infer physical behavior for 3D assets (Zhang et al., 2024; Liu et al., 2024; Huang et al., 2025; Lin et al., 2025). Our diffusion model is trained directly as a conditional inverse model over heterogeneous physical fields, with additional interactions providing complementary evidence to reduce uncertainty. Its samples initialize the bounded differentiable MPM refinement used for the final deployment field.

3 METHOD

The overall input to our method is a collection of multi-view image sequences showing the same object under one or more physical interactions. From these visual observations, our method recovers a spatially varying physical property field in three stages: reconstructing a shared visual and physical particle representation, generating a field with conditional diffusion, and refining it through differentiable simulation and rendering.

Task And Deployment Objective. Let $\mathcal{J}$ denote the nonempty set of observed interactions. For each $j \in \mathcal{J}$, $\mathcal{I}^{(j)}$ is the multi-view image sequence, $\mathcal{M}^{(j)}$ its foreground masks, $\mathcal{K}^{(j)}$ the camera configuration, and $\mathbf{G}^{(j)}$ the interaction configuration. All interactions share the same object and material field but differ in their initial states and scene configurations. The material family χ is fixed for each deployment instance and selects both the constitutive model and field dimension: elastic fields use Young's modulus and Poisson's ratio (E, ν) with $d_\chi = 2$, whereas plastic fields additionally use yield stress τ_y with $d_\chi = 3$. Given the multi-view image sequences, foreground masks, camera configurations, and interaction configurations of all observed interactions, our goal is to identify a single internal material field that consistently explains all observed responses. Dataset supervision and the precise deployment-input convention are detailed in Appendix A.1.

3.1 RECONSTRUCTION AND PARTICLE SUPPORT

Our input contains only visual observations, whereas the target is an internal physical property field that is not visible. To bridge this gap, we first reconstruct the object using dynamic 3D Gaussians (Kerbl et al., 2023), which describe its geometry, appearance, and visible motion and can be rendered from the observed cameras. The reconstructed Gaussians mainly describe the visible surface, whereas the target material field extends throughout the volume and therefore lacks an internal carrier. Given the internal-particle spacing Δx_{mpm}, we jointly denote Gaussian reconstruction and interior sampling by

$$(\mathcal{Q}_{\mathrm{can}}, \mathbf{X}_{\mathrm{can}}) = \mathrm{ConstructParticles}\Big(\{\mathcal{I}^{(j)}, \mathcal{M}^{(j)}, \mathcal{K}^{(j)}\}_{j\in\mathcal{J}}; \Delta x_{\mathrm{mpm}}\Big). \tag{1}$$

Here $\mathrm{ConstructParticles}(\cdot)$ reconstructs a canonical dynamic Gaussian model from the multi-view observations and samples its enclosed volume at spacing Δx_{mpm} to form a shared internal-particle support. Its outputs are the dynamic Gaussian representation $\mathcal{Q}_{\mathrm{can}}$ and the internal-particle coordinates $\mathbf{X}_{\mathrm{can}}$. The Gaussians retain appearance and participate in rendering, whereas the internal particles carry the spatially varying material field used by diffusion and physical simulation.

We bind the internal particles to the reconstructed Gaussian motion. For each interaction, trajectory recovery and physical feature extraction are summarized as

$$\mathbf{H}_{\mathrm{obs}}^{(j)} = \mathrm{ExtractMotion}\Big(\mathcal{I}^{(j)}, \mathcal{M}^{(j)}, \mathcal{K}^{(j)}; \mathcal{Q}_{\mathrm{can}}, \mathbf{X}_{\mathrm{can}}, \mathbf{G}^{(j)}\Big), \qquad j \in \mathcal{J}. \tag{2}$$

Here $\mathrm{ExtractMotion}(\cdot)$ binds the shared internal particles to the reconstructed Gaussian motion for interaction j and extracts their time-varying physical evidence. Its output $\mathbf{H}_{\mathrm{obs}}^{(j)}$ contains internal-particle positions, velocities, and local volumetric and shear changes. Reconstruction, tracking, and local-deformation extraction are detailed in Appendix A.2.

3.2 Conditional Material-Field Diffusion

The reconstructed motion and deformation evidence is inevitably incomplete and noisy because it is inferred from limited visual observations. We therefore first use a conditional diffusion model to estimate the physical properties on the internal particles while representing the uncertainty in the possible solutions. This is important because the inverse problem is ambiguous: different internal material fields can produce similar visible trajectories, especially when observations are limited.

Multi-Interaction Conditioning. At each resolution s, we combine all observed responses into a single condition,

$$\mathbf{C}^{(s)} = \mathrm{Fuse}_s\Big(\{(\mathbf{H}_{\mathrm{obs}}^{(j)}, \mathbf{G}^{(j)})\}_{j\in\mathcal{J}}\Big). \tag{3}$$

Here $\mathrm{Fuse}_s(\cdot)$ encodes each interaction history together with its configuration, aggregates the resulting interaction features in a permutation-invariant manner, and returns the resolution-specific condition $\mathbf{C}^{(s)}$. Different interactions excite different regions and material responses, so their complementary evidence reduces ambiguities that cannot be resolved from a single trajectory.

Diffusion Training. The training noise schedule is $\{\bar{\alpha}_t\}_{t=0}^{T}$. At each resolution, we adopt velocity prediction: we progressively corrupt ground-truth material fields on the internal particles and train $\mathbf{V}_{\theta_s}$ to predict the corresponding velocity from the noisy field, the fused condition, and the current time step. The model therefore learns to connect each candidate field with the reconstructed motion that it must explain. To mitigate information loss in visual reconstruction, training mixes reconstructed and ground-truth trajectories to retain clean-trajectory capability while improving robustness; inference uses reconstructed trajectories only. We denote the resulting collection of trained diffusion parameters by θ, with θ_s referring to the parameters at resolution s.

Coarse-To-Fine Generation. Broad response patterns constrain large material regions more reliably than subtle local variations. We therefore perform inference progressively from coarse to fine. We denote the ordered set of these resolutions by $\mathcal{S}$ (with coarse and fine denoting its first and last resolutions, respectively). The coarsest resolution recovers the global layout. At every finer resolution, the recovered parent field is transferred by

$$\widehat{\mathbf{Z}}_0^{(s)} = \mathrm{Interp}_s\Big(\widehat{\mathbf{Z}}_0^{(s^-)}\Big) + \mathbf{Z}_{\tau_0,\mathrm{diff}}^{(s)}, \qquad s \neq \mathrm{coarse}, \tag{4}$$

where s^- is the coarser resolution, $\mathrm{Interp}_s(\cdot)$ interpolates its field to the current resolution similar to Kirillov et al. (2020), and $\mathbf{Z}_{\tau_0,\mathrm{diff}}^{(s)}$ is the newly generated residual.

Inference. We use the reverse-time schedule $0 = \tau_0 < \cdots < \tau_L = T$. At resolution s, $\mathbf{Z}_{\tau_t,\mathrm{diff}}^{(s)}$ denotes the noisy material-field residual at reverse time τ_t, initialized by $\mathbf{Z}_{\tau_L,\mathrm{diff}}^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. At reverse step τ_t, the denoiser $\mathbf{V}_{\theta_s}$ maps the current noisy residual field, fused condition, and time step to the velocity prediction $\widehat{\mathbf{V}}_{\tau_t}^{(s)} = \mathbf{V}_{\theta_s}(\mathbf{Z}_{\tau_t,\mathrm{diff}}^{(s)}, \mathbf{C}^{(s)}, \tau_t)$. Rather than introducing a separate sampler operator, we use the deterministic reverse update (Song et al., 2020):

$$\mathbf{Z}_{\tau_{t-1},\mathrm{diff}}^{(s)} = \left[\sqrt{\bar{\alpha}_{\tau_{t-1}}\bar{\alpha}_{\tau_t}} + \sqrt{(1-\bar{\alpha}_{\tau_{t-1}})(1-\bar{\alpha}_{\tau_t})}\right]\mathbf{Z}_{\tau_t,\mathrm{diff}}^{(s)} + \left[\sqrt{(1-\bar{\alpha}_{\tau_{t-1}})\bar{\alpha}_{\tau_t}} - \sqrt{\bar{\alpha}_{\tau_{t-1}}(1-\bar{\alpha}_{\tau_t})}\right]\widehat{\mathbf{V}}_{\tau_t}^{(s)}. \tag{5}$$

This equation advances the noisy residual field by removing the component indicated by the predicted velocity under the prescribed noise schedule. The finest recovered field is the diffusion prediction,

Algorithm 1 End-to-end physical property field identification

Require: $\{\mathcal{I}^{(j)}, \mathcal{M}^{(j)}, \mathcal{K}^{(j)}, \mathbf{G}^{(j)}\}_{j\in\mathcal{J}}, \chi, \Delta x_{\mathrm{mpm}}, \theta, \mathcal{S}, \{\bar{\alpha}_t\}_{t=0}^{T}, 0=\tau_0<\cdots<\tau_L=T, \mathcal{R}=\{\eta_r\}_{r=0}^{R-1}$

Initialize the χ-specific diffusion model and MPM constitutive law ▷ Select field channels and dimension

// Reconstruction and Particle Support

$(\mathcal{Q}_{\mathrm{can}}, \mathbf{X}_{\mathrm{can}}) \leftarrow \mathrm{ConstructParticles}(\{\mathcal{I}^{(j)}, \mathcal{M}^{(j)}, \mathcal{K}^{(j)}\}_{j\in\mathcal{J}}; \Delta x_{\mathrm{mpm}})$ ▷ Reconstruct shell and sample its interior

for $j \in \mathcal{J}$ **do**

  $\mathbf{H}_{\mathrm{obs}}^{(j)} \leftarrow \mathrm{ExtractMotion}(\mathcal{I}^{(j)}, \mathcal{M}^{(j)}, \mathcal{K}^{(j)}; \mathcal{Q}_{\mathrm{can}}, \mathbf{X}_{\mathrm{can}}, \mathbf{G}^{(j)})$ ▷ Extract particle-wise motion evidence

// Conditional Material-Field Diffusion

for $s \in \mathcal{S}$ from coarse to fine **do** ▷ Resolve global structure before local detail

  $\mathbf{C}^{(s)} \leftarrow \mathrm{Fuse}_s(\{(\mathbf{H}_{\mathrm{obs}}^{(j)}, \mathbf{G}^{(j)})\}_{j\in\mathcal{J}})$ ▷ Fuse complementary interactions

  $\mathbf{Z}_{\tau_L,\mathrm{diff}}^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ ▷ Initialize the residual at this resolution from noise

  for $t = L, \ldots, 1$ **do**

    $\widehat{\mathbf{V}}_{\tau_t}^{(s)} \leftarrow \mathbf{V}_{\theta_s}(\mathbf{Z}_{\tau_t,\mathrm{diff}}^{(s)}, \mathbf{C}^{(s)}, \tau_t)$ ▷ Predict the denoising velocity

    $\mathbf{Z}_{\tau_{t-1},\mathrm{diff}}^{(s)} \leftarrow \left[\sqrt{\bar{\alpha}_{\tau_{t-1}}\bar{\alpha}_{\tau_t}} + \sqrt{(1-\bar{\alpha}_{\tau_{t-1}})(1-\bar{\alpha}_{\tau_t})}\right]\mathbf{Z}_{\tau_t,\mathrm{diff}}^{(s)}$
    $+\left[\sqrt{(1-\bar{\alpha}_{\tau_{t-1}})\bar{\alpha}_{\tau_t}} - \sqrt{\bar{\alpha}_{\tau_{t-1}}(1-\bar{\alpha}_{\tau_t})}\right]\widehat{\mathbf{V}}_{\tau_t}^{(s)}$ ▷ Deterministic reverse update

  $\widehat{\mathbf{Z}}_0^{(s)} \leftarrow \mathrm{Interp}_s(\widehat{\mathbf{Z}}_0^{(s^-)}) + \mathbf{Z}_{\tau_0,\mathrm{diff}}^{(s)}$ ▷ Add the finer residual

// Differentiable MPM Refinement

$\widehat{\mathbf{P}}_0 \leftarrow \widehat{\mathbf{Z}}_0^{(\mathrm{fine})}$ ▷ Initialize from the diffusion prediction

for $r = 0, \ldots, R-1$ **do** ▷ Follow the prescribed MPM schedule

  $\mathcal{L}_r \leftarrow \sum_{j\in\mathcal{J}} \mathcal{L}_{\mathrm{render}}(\widehat{\mathbf{P}}_r; \mathcal{Q}_{\mathrm{can}}, \mathbf{X}_{\mathrm{can}}, \mathcal{I}^{(j)}, \mathcal{M}^{(j)}, \mathcal{K}^{(j)}, \mathbf{G}^{(j)})$ ▷ Simulate, drive Gaussians, and render

  $\widehat{\mathbf{P}}_{r+1} \leftarrow \widehat{\mathbf{P}}_r - \eta_r \nabla_{\widehat{\mathbf{P}}_r}\mathcal{L}_r$ ▷ Backpropagate to the material field

return $\widehat{\mathbf{P}}_R$ ▷ Final field

$\mathbf{P}_{\mathrm{pred}} = \widehat{\mathbf{Z}}_0^{(\mathrm{fine})}$. Training-time history construction, physical coordinates, multi-resolution processing, interaction fusion, denoiser design, and reverse sampling are detailed in Appendices A.3–A.9.

3.3 Differentiable MPM Refinement

We use MPM as the physical simulator because its particle representation naturally handles large deformation, contact, and topology changes. The diffusion output provides a strong prior but may not exactly reproduce the observed motion, so we refine it against the input views. We set $\widehat{\mathbf{P}}_0 = \mathbf{P}_{\mathrm{pred}}$ and use the refinement schedule $\mathcal{R} = \{\eta_r\}_{r=0}^{R-1}$.

Simulation And Rendering. At step r, each observed interaction is simulated with the shared field $\widehat{\mathbf{P}}_r$ and configuration $\mathbf{G}^{(j)}$. The simulated internal particles drive the reconstructed Gaussians, which are rendered using $\mathcal{K}^{(j)}$ and compared with $\mathcal{I}^{(j)}$ and $\mathcal{M}^{(j)}$. The function $\mathcal{L}_{\mathrm{render}}$ denotes the image-space discrepancy obtained by differentiably simulating the internal particles, driving the bound Gaussians, and comparing their renderings with the observed images inside the masks. Summing over interactions gives $\mathcal{L}_r = \sum_{j\in\mathcal{J}} \mathcal{L}_{\mathrm{render}}(\widehat{\mathbf{P}}_r; \mathcal{Q}_{\mathrm{can}}, \mathbf{X}_{\mathrm{can}}, \mathcal{I}^{(j)}, \mathcal{M}^{(j)}, \mathcal{K}^{(j)}, \mathbf{G}^{(j)})$.

Differentiable Update. The reconstruction, particle binding, and appearance remain fixed; only the material field is optimized. Gradients pass from the rendering loss through Gaussian driving and the differentiable MPM rollout to the internal-particle properties, yielding $\widehat{\mathbf{P}}_{r+1} = \widehat{\mathbf{P}}_r - \eta_r \nabla_{\widehat{\mathbf{P}}_r}\mathcal{L}_r$. After the scheduled updates, the final field is $\widehat{\mathbf{P}}_{\mathrm{deploy}} = \widehat{\mathbf{P}}_R$. The differentiable rollout, Gaussian driving, rendering objective, and refinement safeguards are detailed in Appendix A.10.

4 Experiments

We evaluate HeteroPhys from three complementary perspectives: its ability to recover heterogeneous physical property fields on geometrically unseen objects, the effect of multiple interaction observations on identification accuracy and ambiguity, and the generalization of the recovered fields to unseen interaction conditions. We further conduct controlled ablations to analyze the contribution of the major design choices.

4.1 Experimental Setup

Dataset. We construct a benchmark with ground-truth physical fields for two constitutive families, *Elastic* and *Plasticine*. The training set contains 20 object geometries, while evaluation uses eight

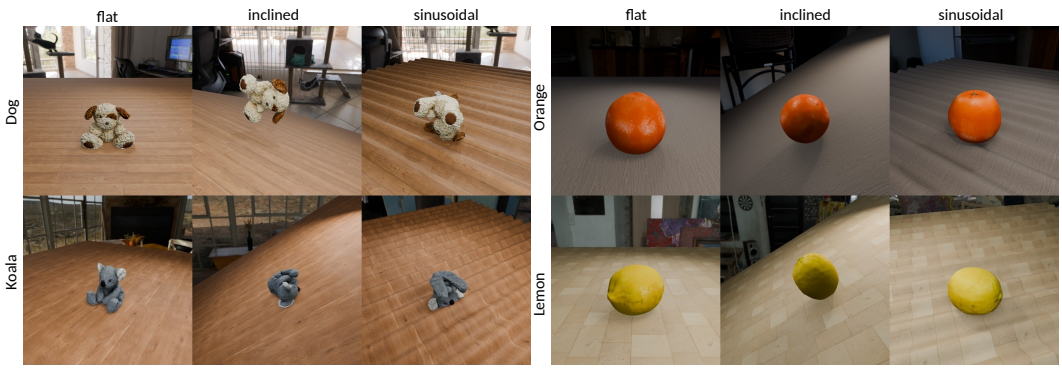

Figure 2: **Overview of the HeteroPhys benchmark.** The four geometrically unseen objects and interaction conditions used for evaluation.

geometrically unseen objects, with four objects per material family. Each evaluation object is instantiated under five distribution types: one homogeneous distribution and four heterogeneous distributions (two-region, three-region, core–shell, and continuous gradient). We generate two independent physical-field realizations for every object–distribution pair, giving $8 \times 5 \times 2 = 80$ evaluation records in total. For every object–field realization, we generate 12 interactions with different initial poses and terrains while keeping the canonical particle identities and physical field unchanged. The diffusion models are trained with a randomly sampled number of observed interactions $N_{\text{obs}} \in \{1, 2, 4\}$. Experiment 1 evaluates the four heterogeneous distributions, while the homogeneous subset is used for the controlled cross-method comparison in Sec. 4.4. Figure 2 provides a summary of the benchmark. benchmark and interaction details are provided in Appendix B.1.

Field-Recovery Metrics. We evaluate the final field $\widehat{\mathbf{P}}$ using the coefficient of determination R^2. Because the reconstruction-injected and simulator-native supports may differ, the recovered field is spatially interpolated to GT object-particle query locations before scoring; fixed scene rows are excluded. Following the bulk–shear parameterization of Appendix A.5, the main metric is computed over $(\log K, \log G)$ for Elastic and $(\log K, \log G, \log \gamma_y)$ for Plasticine.Here, K denotes the bulk modulus, which measures resistance to volumetric deformation, G denotes the shear modulus, which measures resistance to shape deformation, and γ_y denotes the plastic yield strain at the onset of irreversible plastic deformation. Scores are computed independently for each field realization and parameter channel, averaged over realizations of the same object and field distribution, and then macro-averaged across active physical channels. Parameter-wise results are reported both in the bulk–shear parameterization and, after physical decoding, for the corresponding (E, ν, τ_y) parameters in Appendix C.1. We additionally report MAE in $(\log E, \nu, \log \tau_y)$ where applicable. The complete metric definition is provided in Appendix B.2.

Future-Prediction Metrics. For prediction under unseen interactions, we report Chamfer Distance (CD) $\downarrow$, Peak Signal-to-Noise Ratio (PSNR) $\uparrow$, and Structural Similarity Index Measure (SSIM) $\uparrow$. All methods use the same target frames, point-sampling procedure, rendered views, and metric implementations.

4.2 Heterogeneous Physical Field Recovery

Our primary experiment asks whether HeteroPhys recovers the heterogeneous physical property field itself on geometrically unseen objects. This is stricter than evaluating only reconstructed or predicted motion, since distinct parameter fields can produce similar trajectories under limited observations.

Table 1 reports direct field recovery across the four heterogeneous distributions using the bulk–shear R^2 defined in Sec. 4.1. Each entry macro-averages the active physical channels of the corresponding constitutive family. The mean score is 0.7255 across unseen geometries, and performance remains strong across all four heterogeneous distribution types, indicating that the recovered fields are not tied to a single spatial pattern. Complete parameter-wise results in bulk–shear coordinates and after

Table 1: Heterogeneous physical field recovery on **unseen object geometries**.

Object	Two-region	Three-region	Core–shell	Gradient	Mean
Lemon	0.9264	0.8400	0.8458	0.8685	0.8702
Orange	0.6606	0.9452	0.9307	0.8333	0.8425
Peach	0.9073	0.7207	0.6354	0.8133	0.7691
Strawberry	0.7804	0.7267	0.8063	0.8654	0.7947
Bear	0.8673	0.3451	0.8991	0.4543	0.6415
Dog	0.7026	0.6355	0.7237	0.6202	0.6705
Koala	0.1808	0.7943	0.7167	0.6001	0.5730
Shark	0.9658	0.4626	0.4137	0.7289	0.6427
Mean	**0.7489**	**0.6838**	**0.7464**	**0.7230**	**0.7255**

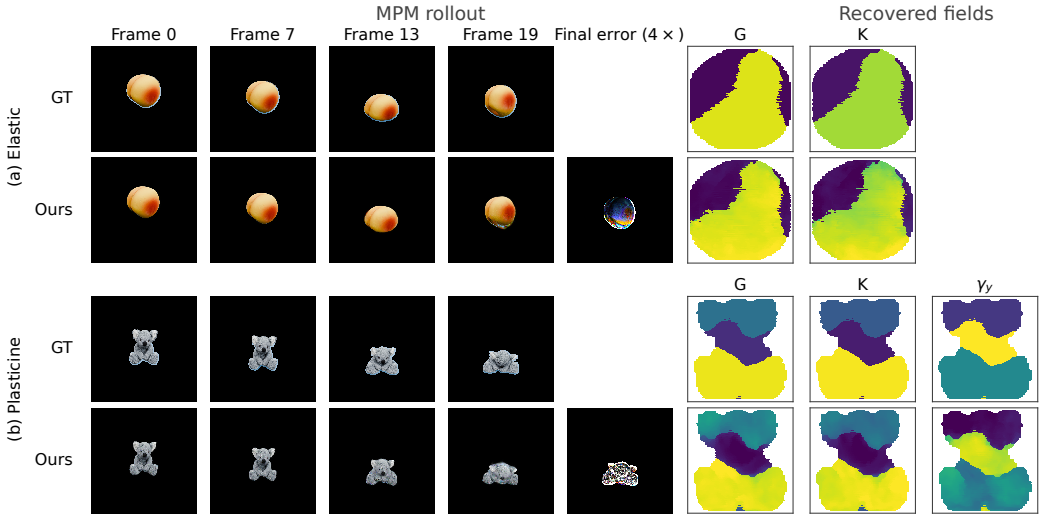

Figure 3: **Heterogeneous physical field recovery on unseen objects.** HETEROPHYS predicts spatial physical parameters in canonical object coordinates, including piecewise heterogeneous fields, internal core–shell variation, and continuously varying gradient fields.

decoding to (E, ν, τ_y) are provided in Appendix C.1. Figure 3 provides direct visualizations of the recovered fields. We include examples from both constitutive families. Ground truth and prediction are visualized using identical physical-value ranges.

4.3 MULTI-INTERACTION EVIDENCE REDUCES PHYSICAL AMBIGUITY

A single observed interaction generally does not uniquely constrain a dense heterogeneous physical field. We therefore evaluate how identification changes as the same object–field realization is observed under $N_{\text{obs}} \in \{1, 2, 4\}$ different interactions. Because the object and physical field remain fixed, this experiment isolates the contribution of additional physical evidence.

Identification accuracy is measured using the same bulk–shear R^2 as in Sec. 4.2. To additionally characterize ambiguity in the learned conditional solution space, we draw S field samples from different diffusion initializations. After channel-wise log transformation and normalization to a common scale for this diagnostic only, we define $D_{\text{field}} = \frac{1}{MC} \sum_{j=1}^{M} \sum_{c=1}^{C} \text{Std}_{s=1,\dots,S}[\widehat{\widetilde{z}}_{j,c}^{(s)}]$.. Lower dispersion indicates that fewer substantially different physical fields remain compatible with the observations. Figure 4 shows how field recovery and conditional sample dispersion change with additional interaction evidence. Table 2 reports the corresponding object-wise results. Increasing N_{obs} from 1 to 4 raises the mean R^2 from 0.0488 to 0.7255 while reducing field dispersion from 0.5657 to 0.2910, confirming that complementary interactions improve accuracy and reduce ambiguity.

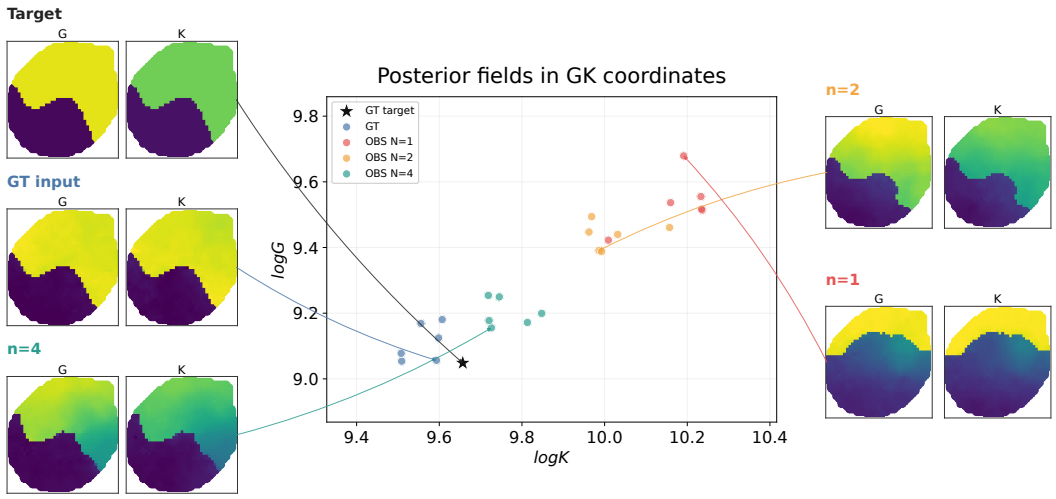

Figure 4: **Additional interaction evidence reduces physical ambiguity.** More observed interactions improve field recovery while reducing dispersion among conditional field samples.

Table 2: Object-wise effect of **multi-interaction evidence**.

	$R^2 \uparrow$			Field Dispersion $\downarrow$		
Object	$N_{\text{obs}} = 1$	$N_{\text{obs}} = 2$	$N_{\text{obs}} = 4$	$N_{\text{obs}} = 1$	$N_{\text{obs}} = 2$	$N_{\text{obs}} = 4$
Lemon	0.2250	0.4347	0.8702	0.4869	0.2551	0.1977
Orange	0.1675	0.3565	0.8425	0.4668	0.4611	0.2147
Peach	0.3303	0.3109	0.7691	0.5047	0.4590	0.2788
Strawberry	0.1544	0.0348	0.7947	0.5128	0.7349	0.2512
Bear	-0.3716	0.2202	0.6415	0.8344	0.5180	0.3462
Dog	-0.0131	0.2653	0.6705	0.5359	0.4641	0.3168
Koala	0.0489	0.2021	0.5730	0.7123	0.6539	0.4732
Shark	-0.1513	-0.5278	0.6427	0.4720	0.6389	0.2493
Mean	**0.0488**	**0.1621**	**0.7255**	**0.5657**	**0.5231**	**0.2910**

4.4 Prediction under Unseen Interactions

Direct field recovery establishes whether the inferred physical properties are numerically consistent with ground truth. We next test whether the recovered fields behave as intrinsic object properties by deploying them under interactions that were not observed during identification.

Cross-Interaction Prediction On Heterogeneous Fields. We first test whether an identified heterogeneous field transfers to unseen physical conditions. The field is inferred from four observed interactions and then kept fixed during target-interaction simulation. We consider three settings: *New Pose*, where the target orientation is unseen; *New Terrain*, where the target terrain is unseen; and *New Pose + New Terrain*, where both are unseen. No target-interaction frames are used for identification or adaptation. Table 3 reports the results. Across all three unseen-interaction settings, the recovered heterogeneous fields maintain mean CD at or below 1.4×10^{-4} and mean SSIM between 0.979 and 0.984, showing stable predictive behavior under changes in pose, terrain, or both.

Comparison With Prior System-Identification Methods. Most existing video-based system-identification methods estimate global physical parameters rather than arbitrary dense heterogeneous fields. For a controlled comparison, we therefore evaluate the homogeneous subset under the most challenging New Pose + New Terrain setting. Table 4 shows that HeteroPhys substantially outperforms prior methods across CD, PSNR, and SSIM for both material families. Using four observations gives the best overall image-space quality while retaining very low geometric error. We therefore perform the cross-method comparison on the homogeneous subset of our benchmark, where all methods have a compatible representation target.

All methods are evaluated on the same held-out *New Pose + New Terrain* interactions, with identified parameters fixed before target simulation. For a controlled observation-matched comparison, PAC-

Table 3: **Prediction under unseen interactions** for heterogeneous physical fields.

Setting	Metric	Lemon	Orange	Peach	Strawberry	Bear	Dog	Koala	Shark	Mean
New Pose	CD ↓	0.00013	0.00014	0.00014	0.00014	0.00013	0.00011	0.00013	0.00009	0.00013
	PSNR ↑	32.70849	34.92677	33.63071	32.81581	29.11582	29.92995	29.07259	32.84033	31.88006
	SSIM ↑	0.98173	0.98675	0.98310	0.96869	0.97483	0.97848	0.97355	0.98785	0.97937
New Terrain	CD ↓	0.00015	0.00015	0.00015	0.00015	0.00013	0.00012	0.00015	0.00010	0.00014
	PSNR ↑	33.45241	34.28873	34.20418	34.18913	30.77858	30.78410	29.35194	34.02240	32.63393
	SSIM ↑	0.98525	0.98377	0.98532	0.98018	0.97993	0.98166	0.97544	0.98919	0.98259
New Pose + Terrain	CD ↓	0.00015	0.00015	0.00015	0.00015	0.00015	0.00012	0.00014	0.00010	0.00014
	PSNR ↑	33.41309	35.40853	34.86339	34.16202	30.25054	30.40162	29.79010	33.91817	32.77593
	SSIM ↑	0.98478	0.98905	0.98783	0.97779	0.97940	0.98140	0.97731	0.99096	0.98356

Table 4: **Future-state prediction** on the homogeneous subset under New Pose + New Terrain. **First**, **second**, and **third** denote the top three ranked results.

	Method	Elastic					Plasticine				
		Lemon	Orange	Peach	Strawberry	Mean	Bear	Dog	Koala	Shark	Mean
CD ↓	PAC-NeRF (Li et al., 2023)	0.00951	0.03838	**0.01056**	**0.01067**	0.01728	**0.00700**	0.01284	0.01044	0.00826	0.00963
	Spring-Gaus (Zhong et al., 2024)	0.01145	**0.01153**	0.01081	0.01388	**0.01192**	-	-	-	-	-
	GIC (Cai et al., 2024)	**0.00913**	**0.00980**	**0.00960**	**0.00976**	**0.00957**	**0.00538**	**0.00729**	**0.00826**	**0.00436**	**0.00632**
	MASIV (Zhao et al., 2025)	**0.00922**	0.01890	0.01372	0.01269	0.01363	0.00765	**0.01043**	0.01004	**0.00668**	**0.00870**
	HETEROPHYS ($N_{\text{obs}} = 1$)	**0.00014**	**0.00015**	**0.00015**	**0.00015**	**0.00015**	**0.00011**	**0.00012**	**0.00014**	**0.00010**	**0.00012**
	HETEROPHYS ($N_{\text{obs}} = 4$)	**0.00014**	**0.00015**	**0.00015**	**0.00015**	**0.00015**	**0.00011**	**0.00012**	**0.00013**	**0.00010**	**0.00012**
PSNR ↑	PAC-NeRF (Li et al., 2023)	25.97573	20.59523	22.83172	20.73918	22.53547	21.34170	21.59718	23.66508	20.03821	21.66054
	Spring-Gaus (Zhong et al., 2024)	22.85126	22.83038	23.21899	18.86677	21.94185	-	-	-	-	-
	GIC (Cai et al., 2024)	**29.09721**	**25.59151**	**27.64939**	**24.12878**	**26.61672**	**22.39923**	**24.60879**	**25.11257**	**22.86936**	**23.74749**
	MASIV (Zhao et al., 2025)	28.88689	22.27073	23.09122	19.15188	23.35018	21.28437	21.97174	22.69411	19.99248	21.48568
	HETEROPHYS ($N_{\text{obs}} = 1$)	**34.52685**	**34.05656**	**33.13544**	**33.49681**	**33.80392**	**29.89738**	**30.61890**	**29.97076**	**34.15588**	**31.16073**
	HETEROPHYS ($N_{\text{obs}} = 4$)	**34.61661**	**34.40847**	**33.66363**	**35.88994**	**34.64466**	**30.20271**	**30.79593**	**30.93353**	**34.64114**	**31.64333**
SSIM ↑	PAC-NeRF (Li et al., 2023)	0.88880	0.81650	0.86962	0.66330	0.80955	0.66760	0.70794	0.70306	0.75600	0.70865
	Spring-Gaus (Zhong et al., 2024)	0.85645	0.83787	0.87466	0.62770	0.79917	-	-	-	-	-
	GIC (Cai et al., 2024)	**0.92925**	**0.87327**	**0.93001**	**0.74197**	**0.86862**	**0.69673**	**0.75216**	**0.73677**	**0.82547**	**0.75278**
	MASIV (Zhao et al., 2025)	0.92231	0.83601	0.87216	0.65594	0.82161	0.67355	0.72232	0.68724	0.77217	0.71382
	HETEROPHYS ($N_{\text{obs}} = 1$)	**0.99023**	**0.98481**	**0.98247**	**0.97372**	**0.98281**	**0.97925**	**0.97867**	**0.97782**	**0.98926**	**0.98125**
	HETEROPHYS ($N_{\text{obs}} = 4$)	**0.99079**	**0.98627**	**0.98510**	**0.98838**	**0.98764**	**0.98026**	**0.97966**	**0.98279**	**0.99082**	**0.98338**

NeRF, Spring-Gaus, GIC, MASIV, and HETEROPHYS ($N_{\text{obs}} = 1$) use one identification interaction. We additionally report the full HETEROPHYS setting with $N_{\text{obs}} = 4$. PAC-NeRF, GIC, and MASIV are evaluated on both material families, whereas Spring-Gaus is evaluated only on Elastic because its spring-mass formulation does not support Plasticine. Implementation and adaptation details are provided in Appendix B.3.

4.5 Ablation and Design Analysis

We conduct controlled ablations of the major design choices of HETEROPHYS in the appendix. These expesfers to unseen physical conditions. The field is inferred from four observed interactions and then kept fixed during target-interaction simulation. These expeiriments examine the history-source training strategy, volumetric/deviatoric deformation preprocessing, bulk–shear property parameterization, hierarchical multi-resolution diffusion, multi-interaction training and inference, and DDIM sampling. Across these analyses, we vary one component at a time while keeping the data split, optimization protocol, and evaluation setting fixed. The results support the design choices used in the final model and further characterize their behavior under both clean simulator histories and end-to-end video-derived observations. See Appendix A.3, Appendix A.4, Appendix A.5, Appendix A.6, Appendix A.7, and Appendix A.9 for the complete results.

5 Conclusion

We presented HETEROPHYS, a framework for recovering spatially heterogeneous physical-property fields from multi-view videos of one or more collisions. By mapping visual motion and deformation evidence to constitutive coordinates, our hierarchical conditional diffusion model integrates complementary cues across interactions and produces a dense initial property field. We then refine this field through gradient-based optimization with differentiable MPM, enforcing consistency with the observed dynamics. Experiments on elastic and plasticine objects show that HETEROPHYS recovers heterogeneous fields on unseen geometries, that additional collisions improve identification and reduce ambiguity, and that the recovered fields support prediction under unseen poses and terrains. These results suggest that combining learned probabilistic inference with physics-based refinement provides a practical route from visual observations to deployable physical models.

AI USE STATEMENT

Large language models (LLMs) were used for polishing the manuscript language, improving readability, and providing coding assistance. The authors take full responsibility for the entire manuscript and confirm that LLM usage adheres to ethical guidelines, with no plagiarism or academic misconduct.

REPRODUCIBILITY STATEMENT

The final paper will document the problem assumptions and method in Section 3, the evaluation protocol and results in Section 4, and additional implementation details in Appendix A and dataset details in Appendix B.

REFERENCES

Junhao Cai, Yuji Yang, Weihao Yuan, Yisheng He, Zilong Dong, Liefeng Bo, Hui Cheng, and Qifeng Chen. GIC: Gaussian-informed continuum for physical property identification and simulation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Junyi Cao, Shanyan Guan, Yanhao Ge, Wei Li, Xiaokang Yang, and Chao Ma. NeuMA: Neural material adaptor for visual grounding of intrinsic dynamics. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Benjamin Holzschuh, Simona Vegetti, and Nils Thuerey. Solving inverse physics problems with score matching. In *Advances in Neural Information Processing Systems*, volume 36, pp. 61888–61922, 2023.

Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson W. H. Lau. DreamPhysics: Learning physics-based 3d dynamics with video diffusion priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3733–3741, 2025.

Miguel Jaques, Michael Burke, and Timothy Hospedales. Physics-as-inverse-graphics: Unsupervised physical parameter estimation from video. In *International Conference on Learning Representations*, 2020.

Krishna Murthy Jatavallabhula, Miles Macklin, Florian Golemo, Vikram Voleti, Linda Petrini, Martin Weiss, Breandan Considine, Jérôme Parent-Lévesque, Kevin Xie, Kenny Erleben, Liam Paull, Florian Shkurti, Derek Nowrouzezahrai, and Sanja Fidler. gradSim: Differentiable simulation for system identification and visuomotor control. In *International Conference on Learning Representations*, 2021.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7219–7230, 2025.

Takuhiro Kaneko. Improving physics-augmented continuum neural radiance field-based geometry-agnostic system identification with lagrangian particle optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5470–5480, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023.

Hyunjin Kim, Ri-Zhao Qiu, Guangqi Jiang, and Xiaolong Wang. EgoPhys: Learning generalizable physics models of deformable objects from egocentric video. *arXiv preprint arXiv:2606.16202*, 2026.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9796–9805. IEEE, 2020.

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 *International Conference on Learning Representations*, 2023.

Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. OmniPhysGS: 3d constitutive gaussians for general physics-based dynamics generation. In *International Conference on Learning Representations*, 2025.

Chunjiang Liu, Xiaoyuan Wang, Qingran Lin, Albert Xiao, Haoyu Chen, Shizheng Wen, Hao Zhang, Lu Qi, Ming-Hsuan Yang, Laszlo A. Jeni, Min Xu, and Yizhou Zhao. Multi-object system identification from videos. In *International Conference on Learning Representations*, 2026.

Fangfu Liu, Hanyang Wang, Shunyu Yao, Shengjun Zhang, Jie Zhou, and Yueqi Duan. Physics3D: Learning physical properties of 3d gaussians via video diffusion. *arXiv preprint arXiv:2406.04338*, 2024.

Pingchuan Ma, Tao Du, Joshua B. Tenenbaum, Wojciech Matusik, and Chuang Gan. RISP: Rendering-invariant state predictor with differentiable simulation and rendering for cross-domain parameter estimation. In *International Conference on Learning Representations*, 2022.

Daniel Rho, Jun Myeong Choi, Matthew Thornton, Biswadip Dey, and Roni Sengupta. MonoPhysics: Estimating geometry, appearance, and physical parameters from monocular videos. *arXiv preprint arXiv:2605.30320*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Norika Wada, Kohei Yamashita, Ryo Kawahara, and Ko Nishino. M-PhyGs: Multi-material object dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024.

Haocheng Yin, Shuohan Tao, Yongsheng Chen, and Lu Gan. PhysCoRe: Physics-corrected residual world models for material-aware deformable dynamics. *arXiv preprint arXiv:2607.20653*, 2026.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pp. 388–406, 2024.

Yizhou Zhao, Haoyu Chen, Chunjiang Liu, Zhenyang Li, Charles Herrmann, Junhwa Hur, Yinxiao Li, Ming-Hsuan Yang, Bhiksha Raj, and Min Xu. Toward material-agnostic system identification from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5944–5956, 2025.

Licheng Zhong, Hong-Xing Yu, Jiajun Wu, and Yunzhu Li. Reconstruction and simulation of elastic objects with spring-mass 3d gaussians. In *European Conference on Computer Vision*, pp. 407–423, 2024.

A Additional Method Details

All ablation tables in Appendix A report the raw model-inference outputs; no subsequent gradient-based MPM refinement is applied. The overall pipeline of HeteroPhys is visualized in 5.

A.1 Training Records, Supervision, and Inference Inputs

Training records form $\mathcal{D} = \{\mathcal{X}^{(m)}\}_{m=1}^{M}$. Each record contains the material family χ, a ground-truth physical field on the object's internal particles, and J interactions that share the same object and material field while varying pose, terrain, and motion conditions. During synthetic record construction, the object interior is sampled once at the prescribed MPM spacing so that all interactions share the same N internal-particle identities. Clean simulator histories and image-derived histories are stored with this common object-particle ordering and channel layout. Fixed scene geometry and kinematics

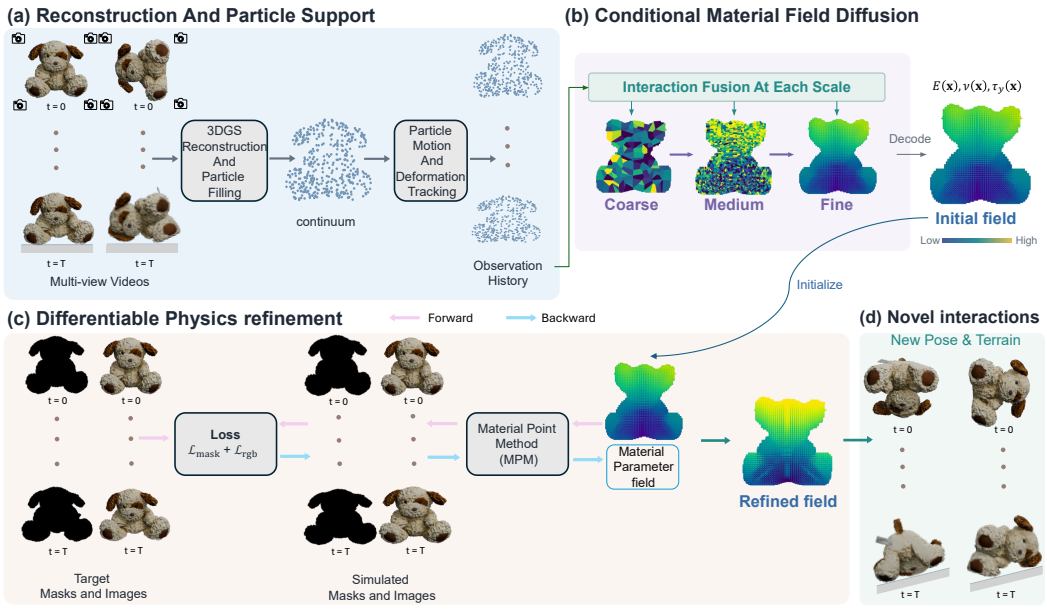

Figure 5: **Overview of the HeteroPhys.** Given one or more calibrated multi-view interaction videos, we reconstruct a shared canonical particle representation and recover correspondence-preserving motion and local deformation histories. The observations are transformed into constitutively structured physical evidence and fused across interactions. A hierarchical residual diffusion model then recovers the heterogeneous physical-property field from coarse structure to fine spatial variation, which can be deployed for simulation under novel interaction conditions.

are supplied by the known interaction configuration and are not prediction targets. Simulator-native particles are retained only as a reference support for quantitative evaluation.

This distinction is important. GT quantities exist because they are available to the synthetic data generator; they are used for supervision, upper-bound diagnostics, and quantitative evaluation. They are not additional deployment inputs. At inference, the model receives calibrated multi-view RGB sequences $\mathcal{I}^{(j)}$, foreground masks $\mathcal{M}^{(j)}$, camera calibration $\mathcal{K}^{(j)}$, and known experimental conditions $\mathbf{G}^{(j)}$. The canonical visual representation, particle support, trajectories, and deformation histories are then reconstructed from these observations.

Each image $\mathbf{I}_{v,f}^{(j)} \in \mathbb{R}^{h\times w\times 3}$ belongs to one of V calibrated views and F frames, and $\mathbf{M}_{v,f}^{(j)} \in \{0,1\}^{h\times w}$ is the corresponding foreground mask. The 28-dimensional condition $\mathbf{G}^{(j)}$ contains a two-entry family code, six density/solver settings, ten gravity/initial-pose values, and ten initial-motion/contact values. The family code is retained for implementation compatibility, while χ is the authoritative selector of the constitutive model and output dimension. It contains no E, ν, τ_y, K, G, or γ_y value. We use bold $\mathbf{G}^{(j)}$ for the scene-level condition; scalar G denotes the shear modulus.

A.2 Observation-History Construction

A.2.1 Canonical reconstruction and support construction

The operator $\mathrm{ConstructParticles}$ consumes the observed multi-view sequences, masks, and calibrations and constructs the shared pair $(\mathcal{Q}_{\mathrm{can}}, \mathbf{X}_{\mathrm{can}})$ used in the main text. In our implementation, the canonical Gaussian appearance is initialized from the calibrated initial views of a designated observed interaction $j_{\mathrm{ref}} \in \mathcal{J}$; the reference is selected from the available inputs and is not an additional observation. Known rigid transforms align the other interactions with this canonical frame, while their complete sequences are used when recovering motion. The cached Gaussian state contains centers, appearance features, opacities, scales, and rotations, and the enclosed volume is sampled at spacing Δx_{mpm} to obtain the internal-particle coordinates $\mathbf{X}_{\mathrm{can}}$.

Adam optimizes each object for 20,000 updates with

$$\mathcal{L}_{\text{ConstructParticles}} = 0.8\|\mathcal{R} - \mathbf{I}\|_1 + 0.2[1 - \text{SSIM}(\mathcal{R}, \mathbf{I})]. \tag{6}$$

We downsample images by a factor of four and sample one calibrated view per update. The representation uses degree-3 spherical harmonics, random-background augmentation, and standard 3DGS densification. Learning rates for centers, scales, opacity, rotation, DC SH, and remaining SH coefficients are respectively 1.6×10^{-4}, 5×10^{-3}, 5×10^{-2}, 10^{-3}, 2.5×10^{-3}, and 1.25×10^{-4}.

The internal-particle identities are fixed across interactions, while the known interaction transforms determine their frame-zero world-space locations. During synthetic data generation, simulator trajectories and ground-truth properties are sampled at these shared object-particle locations to construct clean training histories and supervision. Simulator-native particles are retained only for quantitative evaluation, where the recovered final field is spatially interpolated to their locations. The reference reconstruction reaches PSNR 39.817, SSIM 0.9985, and LPIPS 0.005.

A.2.2 Correspondence-preserving tracking

For each interaction, the dynamic Gaussian front end produces a filled continuum point set at every observed frame. The fitted field is the continuous position–time MLP implemented by `TemporalDeformationField`. Let $\bar{\mathbf{x}}_i = (\mathbf{x}_{i,0} - \mathbf{c})/s$, where $\mathbf{c}$ is the canonical-position mean and $s = \max_i \|\mathbf{x}_{i,0} - \mathbf{c}\|_\infty$. With $\boldsymbol{\omega} = (1, 2, 4)$,

$$\mathbf{x}_i(t) = \mathbf{x}_{i,0} + ts\, D_\phi([\bar{\mathbf{x}}_i, t, \sin(\pi t\boldsymbol{\omega}), \cos(\pi t\boldsymbol{\omega})])\,. \tag{7}$$

$D_\phi : \mathbb{R}^{10} \to \mathbb{R}^3$ is a four-layer MLP with hidden width 128 and SiLU activations; its last layer is zero-initialized. Multiplication by t anchors frame zero exactly and preserves row identity. The field is fitted using a symmetric squared Chamfer objective. When a visual indexed field is available, a smooth-L_1 displacement prior and a frame-zero kNN consistency term regularize the fit without reassigning particle identities.

The default tracking optimization uses 10,000 Adam updates, one nonzero frame per update, and a learning rate decaying from 1.6×10^{-4} to 1.6×10^{-6}. The lowest-validation-Chamfer state is retained. Appearance, Gaussian scale, and Gaussian rotation remain part of the visual reconstruction but are not packed into the physical history.

A.2.3 Local deformation estimation

For every frame-zero particle, we query 32 neighbors and fit a Gaussian-weighted local affine map:

$$\begin{aligned} \mathcal{N}_i &= \text{KNN}_{32}(\mathbf{X}_i), \qquad \widehat{\mathbf{F}}_{i,f} = \text{WLS}_{\mathcal{N}_i}(\mathbf{x}_{\ell,f} \simeq \mathbf{F}\mathbf{X}_\ell + \mathbf{b}), \\ \boldsymbol{\Sigma}^{(j)}_{i,f} &= \sigma^2 \widehat{\mathbf{F}}^{(j)}_{i,f} (\widehat{\mathbf{F}}^{(j)}_{i,f})^\top, \qquad \sigma = \tfrac{1}{2}\Delta x_{\text{mpm}}. \end{aligned} \tag{8}$$

Weights are normalized, coordinates are centered, and a relative 10^{-4} Tikhonov ridge toward $\mathbf{I}_3$ is used. The shared $\sigma = \Delta x_{\text{mpm}}/2$ matches the particle ellipsoid scale used by the extraction and covariance-audit procedure; there is no learned or per-particle σ_i. Tracking and neighborhood covariance fitting are evaluated only on object rows. Fixed-scene history channels are filled directly from the known scene kinematics and retain the shared canonical row order. The packed symmetric order is xx, xy, xz, yy, yz, zz. We treat $\boldsymbol{\Sigma}$ as a deterministic local deformation descriptor derived from tracked positions rather than as an independent observation.

A.2.4 Temporal differencing and packing

Velocities are computed as $\mathbf{V}^{(j)}_{\text{obs}} = \Delta_f \mathbf{X}^{(j)}_{\text{obs}}$. The raw covariance descriptor is converted into the volumetric and deviatoric quantities defined in Appendix A.4. Each final particle–frame row is

$$\mathbf{h}^{(j)}_{i,f} = [\mathbf{x}^{(j)}_{i,f}, \mathbf{v}^{(j)}_{i,f}, \delta^{(j)}_{i,f}, \Delta_f \delta^{(j)}_{i,f}, |\delta^{(j)}_{i,f}|, \mathbf{d}^{(j)}_{i,f}] \in \mathbb{R}^{12}. \tag{9}$$

Stacking particles and frames yields $\mathbf{H}^{(j)}_{\text{obs}} \in \mathbb{R}^{N\times F\times 12}$. Thus ExtractMotion uses $(\mathcal{Q}_{\text{can}}, \mathbf{X}_{\text{can}})$ to recover the interaction-specific trajectories internally and returns the processed history $\mathbf{H}^{(j)}_{\text{obs}}$ appearing in the main text; the intermediate frame-zero state is not an additional output of the operator.

Table 5: History-source ablation, reported separately by material.

Schedule	Input	Elastic		Plasticine	
		$R^2 \uparrow$	$r \uparrow$	$R^2 \uparrow$	$r \uparrow$
GT-only	GT	0.8689	0.9902	0.6100	0.8919
scheduled	GT	0.7151	0.9435	0.5953	0.9192
GT-only	OBS	0.1763	0.8551	0.0498	0.9295
scheduled	OBS	0.6377	0.8601	0.4618	0.8190

A.3 History-Source Selection and Structured Corruption

For each interaction, $c_H^{(j)} \sim \mathrm{Categorical}(0.60, 0.25, 0.15)$ selects among $\mathbf{H}_{\mathrm{gt}}^{(j)}$, $\mathbf{H}_{\mathrm{obs}}^{(j)}$, and a structured corruption $\mathbf{H}_{\mathrm{dam}}^{(j)}$. GT provides clean physical evidence, OBS follows the deployed visual pipeline, and damaged GT exposes the model to controlled observation shift. Damage is applied only to object rows. A dose $d \sim \mathcal{U}(0.3, 1.0)$ multiplies the material-specific OBS–GT envelope, and the corruption composition applies all active damage blocks.

For position, trilinearly interpolated Gaussian grid fields $g_k(\mathbf{x}_i)$ are combined with an AR(1) temporal factor $a_f = 0.99 a_{f-1} + \sqrt{1 - 0.99^2}\, \xi_f$:

$$\Delta \mathbf{x}_{i,f} = c\, d\, [g_1(\mathbf{x}_i), g_2(\mathbf{x}_i), g_3(\mathbf{x}_i)] a_f, \qquad \sqrt{\mathbb{E}_{i,f} \|\Delta \mathbf{x}_{i,f}\|_2^2} = d\, r_x \Delta x_{\mathrm{mpm}}, \tag{10}$$

where the final normalization fixes the realized RMS dose. Velocity receives temporally moving-average-smoothed Gaussian noise, renormalized to RMS dr_v. A positive lognormal factor $\exp(dr_\Sigma \xi_{i,f})$ multiplies all six entries of each particle–frame covariance, thereby preserving positive definiteness. Elastic uses $(r_x, r_v, r_\Sigma, L_x) = (3.40, 2.96\,\mathrm{m/s}, 1.3, 20)$; Plasticine uses $(4.65, 4.05\,\mathrm{m/s}, 2.0, 12)$, with L_x measured in particle spacings.

Training with mixed history sources substantially improves the model's ability to handle OBS inputs encountered in real-world deployment, as reflected by the pronounced R^2 gains for both Elastic and Plasticine. Although performance on clean GT histories decreases to some extent, this trade-off is expected because the model is no longer optimized exclusively for ideal simulator inputs. Given that our target setting relies on video-derived observations, the moderate loss on GT inputs is acceptable in exchange for the considerable improvement in deployment robustness.

A.4 Volumetric and Deviatoric History Preprocessing

The six raw covariance entries mix reference scale, rotation, volume change, and deviatoric stretch. Within $\mathrm{ExtractMotion}$, the raw covariance descriptor is converted before spatial pooling according to

$$\begin{aligned} \mathbf{A}_{i,f} &= \boldsymbol{\Sigma}_{i,0}^{-1/2} \boldsymbol{\Sigma}_{i,f} \boldsymbol{\Sigma}_{i,0}^{-1/2}, \qquad \mathbf{r}_{i,f} = \tfrac{1}{2} \log \boldsymbol{\lambda}(\mathbf{A}_{i,f}), \\ \delta_{i,f} &= \mathbf{1}_3^\top \mathbf{r}_{i,f}, \qquad \mathbf{d}_{i,f} = \mathbf{r}_{i,f} - \tfrac{1}{3} \delta_{i,f} \mathbf{1}_3. \end{aligned} \tag{11}$$

The six covariance channels are replaced by $[\delta, \Delta_f \delta, |\delta|, \mathbf{d}]$. Implementation symmetrizes $\mathbf{A}_{i,f}$, floors eigenvalues at 10^{-10} before inverse-square-root/log operations, and sorts them to fix the deviatoric-slot convention. $\Delta_f \delta$ is a backward difference using the frame duration in $\mathbf{G}^{(j)}$ and is zero at the initial frame.

Table 6 shows that directly using the raw local covariance entries does not provide reliable evidence for field recovery, yielding negative R^2 values for both material families under both GT and OBS histories. In contrast, the physically preprocessed representation consistently achieves positive R^2 and strong correlations. By normalizing the reference configuration and explicitly separating volumetric and deviatoric deformation, the preprocessing exposes constitutively relevant evidence rather than requiring the network to disentangle scale, rotation, volume change, and shear from raw covariance channels. Its consistent advantage on both GT and OBS inputs confirms that this physical structure provides an important inductive bias for heterogeneous-field identification.

Table 6: Raw versus physically preprocessed local-deformation evidence.

Representation	Input	Elastic		Plasticine	
		$R^2 \uparrow$	$r \uparrow$	$R^2 \uparrow$	$r \uparrow$
raw	GT	-0.3628	0.6777	-0.8907	0.5019
preprocessed	GT	0.7151	0.9435	0.5953	0.9192
raw	OBS	-0.8581	0.5902	-1.0003	0.4922
preprocessed	OBS	0.6377	0.8601	0.4618	0.8190

A.5 Bulk–Shear Property Coordinates

For isotropic MPM,

$$K = \frac{E}{3(1-2\nu)}, \qquad G = \frac{E}{2(1+\nu)}, \qquad \gamma_y = \frac{\tau_y}{2G}. \tag{12}$$

The Elastic branch uses the compressible isotropic elastic update defined by K and G. For Plasticine, τ_y enters the dataset-compatible inviscid von Mises return map. Let $\mathbf{e} = \log \mathbf{U}$ denote Hencky strain and $\mathbf{e}_{\mathrm{dev}} = \mathbf{e} - \frac{1}{3}\operatorname{tr}(\mathbf{e})\mathbf{I}$. Yielding occurs when

$$\left\|2G\mathbf{e}_{\mathrm{dev}}^{\mathrm{trial}}\right\|_F > \sqrt{\tfrac{2}{3}}\,\tau_y \quad \Longleftrightarrow \quad \left\|\mathbf{e}_{\mathrm{dev}}^{\mathrm{trial}}\right\|_F > \sqrt{\tfrac{2}{3}}\,\gamma_y. \tag{13}$$

The deviatoric Hencky strain is projected radially to the von Mises surface while the volumetric component is preserved. There is no hardening parameter in this constitutive law.

For positive coordinate bounds $\mathbf{u}_{\chi,\min}, \mathbf{u}_{\chi,\max}$, fixed training center $\mathbf{c}_\chi$, and positive gain $\mathbf{a}_\chi$, calibration is

$$\mathbf{b}_i = 2\frac{\log \mathbf{u}_i - \log \mathbf{u}_{\chi,\min}}{\log(\mathbf{u}_{\chi,\max}/\mathbf{u}_{\chi,\min})} - \mathbf{1}_{d_\chi}, \qquad \mathbf{z}_i = (\mathbf{b}_i - \mathbf{c}_\chi) \odot \mathbf{a}_\chi. \tag{14}$$

A predicted row is inverse-calibrated as

$$\widehat{\mathbf{u}}_i = \exp\!\big(\log \mathbf{u}_{\chi,\min} + \tfrac{1}{2}(\widehat{\mathbf{z}}_i \oslash \mathbf{a}_\chi + \mathbf{c}_\chi + \mathbf{1}_{d_\chi}) \odot \log(\mathbf{u}_{\chi,\max} \oslash \mathbf{u}_{\chi,\min})\big)\,, \tag{15}$$

followed by

$$E = \frac{9KG}{3K+G}, \qquad \nu = \frac{3K-2G}{2(3K+G)}, \qquad \tau_y = 2G\gamma_y. \tag{16}$$

For the reported checkpoint, the fixed channel calibration in equation 14 is

Family	$\mathbf{c}_\chi$	$\mathbf{a}_\chi$
Elastic	$(-0.0790734,\ 0.0180901)$	$(2.0230255,\ 1.4475935)$
Plasticine	$(-0.1508312,\ -0.0823897,\ 0.0717040)$	$(3.1842401,\ 2.4678462,\ 3.0563781)$

The benchmark feasibility boxes are $E \in [2\times10^4, 6\times10^5]$ Pa and $\nu \in [0.15, 0.38]$ for Elastic, and $E \in [1.5\times10^4, 4\times10^5]$ Pa, $\nu \in [0.15, 0.38]$, and $\tau_y \in [5\times10^2, 1.5\times10^4]$ Pa for Plasticine. These material bounds define Ω_χ; the corresponding positive (K, G, γ_y) endpoints define $\mathbf{u}_{\chi,\min}$ and $\mathbf{u}_{\chi,\max}$ analytically. The main-text symbols denote the decoded object-particle field. The inverse calibration is implemented at the output of the finest diffusion stage, so the resulting field is denoted directly by $\mathbf{P}_{\mathrm{pred}} = \widehat{\mathbf{Z}}_0^{(\mathrm{fine})}$ there. Fixed scene properties belong to the known solver configuration and are inserted only when an MPM rollout is assembled; they are neither diffusion outputs nor field-evaluation targets.

Table 7 shows that learning in the constitutively aligned KG coordinates provides more reliable recovery for both materials under both GT and OBS inputs, whereas directly learning $E\nu$ is substantially less stable. This supports using bulk–shear variables as the primary identification coordinates, since they separate the volumetric and deviatoric responses of the constitutive model. Moreover, since $\mathrm{d}\nu = [9G\,\mathrm{d}K - 9K\,\mathrm{d}G]/[2(3K+G)^2]$, small relative errors in the recovered K/G ratio can be amplified after conversion to ν, explaining the lower stability of the decoded conventional parameters.

Table 7: Property-coordinate ablation evaluated in both coordinate systems.

Train coord.	Input	Eval. coord.	Elastic		Plasticine	
			$R^2 \uparrow$	$r \uparrow$	$R^2 \uparrow$	$r \uparrow$
$E\nu$	GT	$E\nu$	-4.0748	0.4115	-8.6696	0.5209
		KG	0.2613	0.5418	0.2330	0.7898
	OBS	$E\nu$	-5.0764	0.3388	-7.1131	0.4842
		KG	0.0567	0.4142	-0.2441	0.7494
KG	GT	$E\nu$	0.2330	0.6850	-0.0629	0.3299
		KG	0.7151	0.9435	0.5953	0.9192
	OBS	$E\nu$	0.4084	0.7474	-0.8887	0.5587
		KG	0.6377	0.8601	0.4618	0.8190

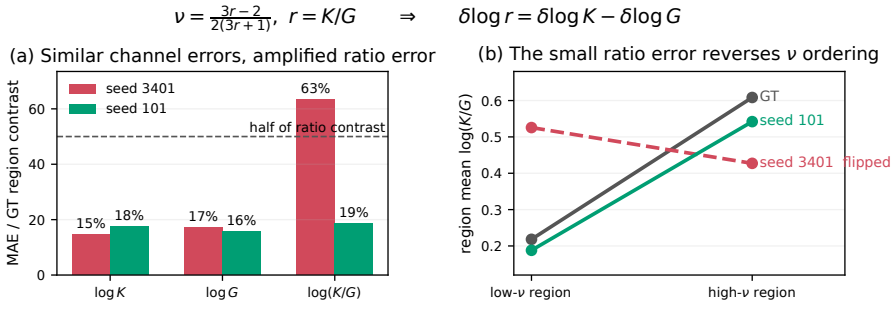

Figure 6: Sensitivity of Poisson's ratio to small errors in bulk–shear coordinates.

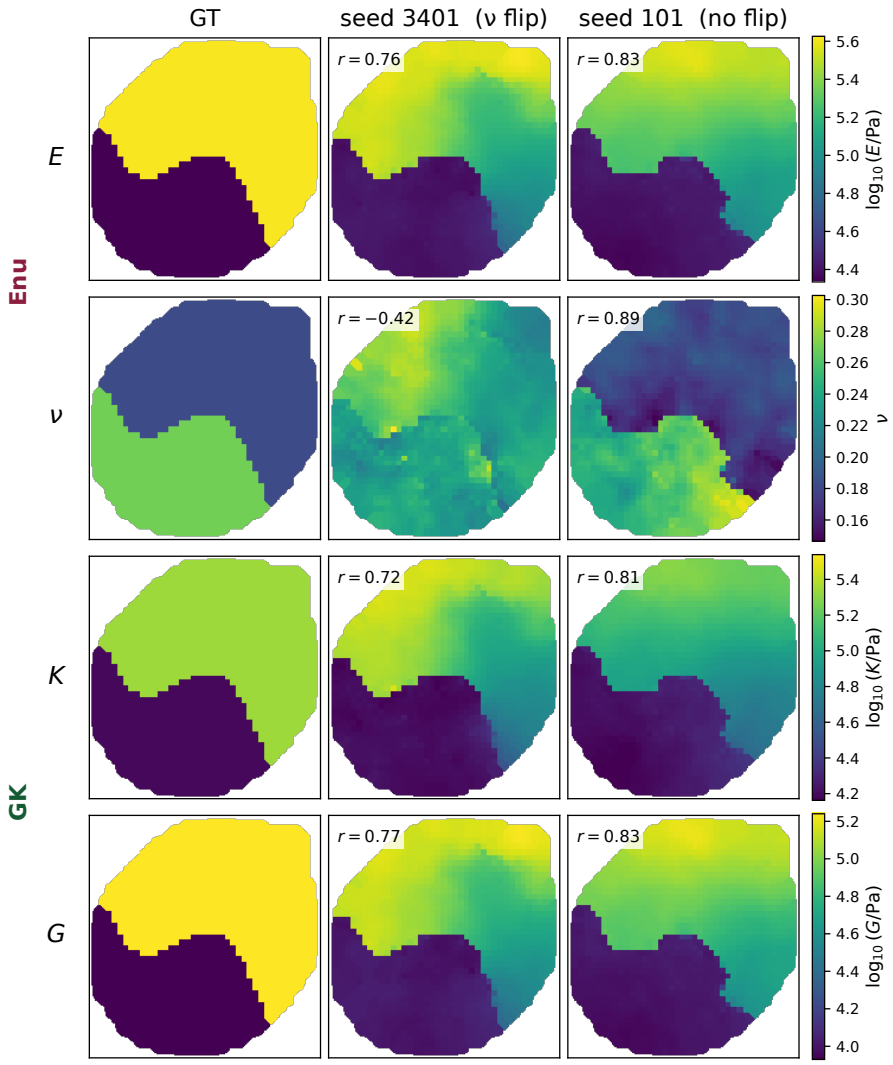

Figure 7: Coordinate-dependent recovery of Elastic fields.

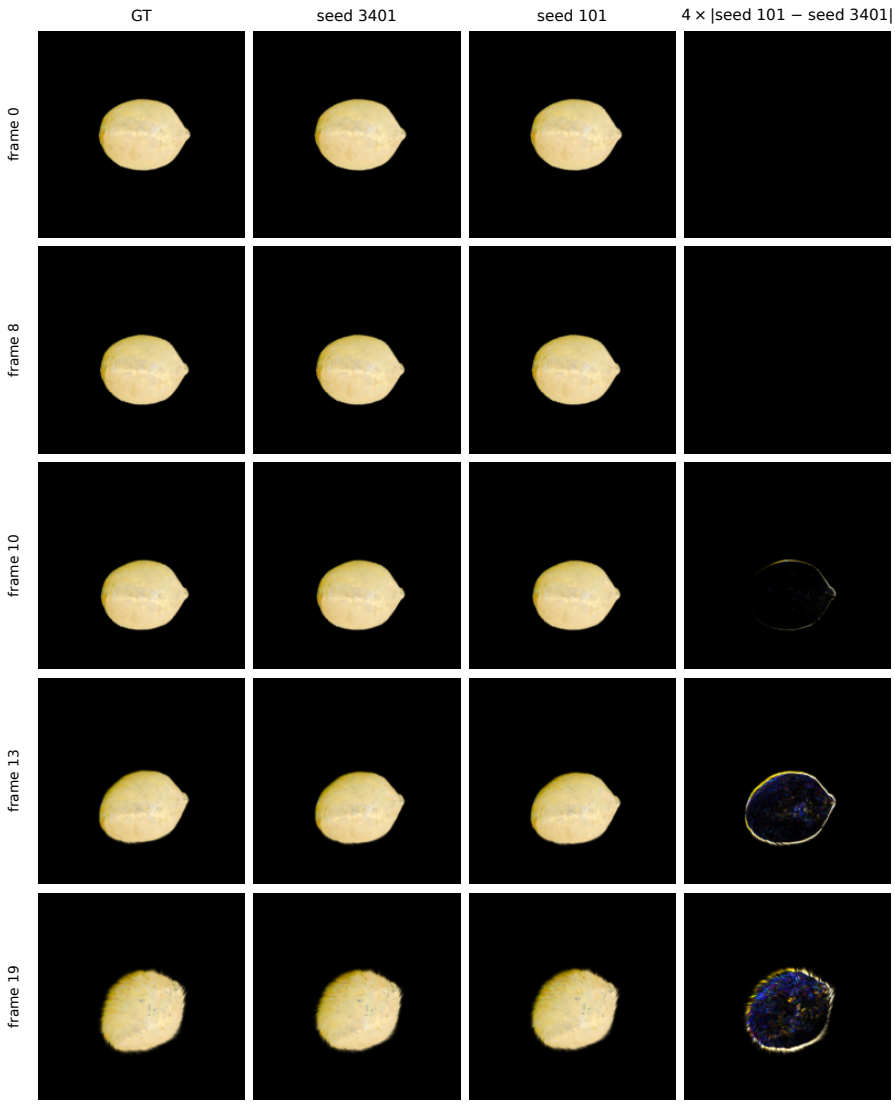

Figure 8: Propagation of material-field errors to MPM rollouts.

A.6 Multi-Resolution Representation and Residual Diffusion

A.6.1 Canonical spatial pooling

Pool_s is the internal deterministic voxel partition used to construct resolution s. Let $\mathbf{p}_i \in \mathbb{R}^3$ denote canonical internal-particle coordinates. With the bounding box of the canonical support,

$$\mathbf{v}_i^{(s)} = \min\left(r_s - 1, \left\lfloor r_s \frac{\mathbf{p}_i - \mathbf{p}_{\min}}{\max(\mathbf{p}_{\max} - \mathbf{p}_{\min}, \varepsilon)} \right\rfloor\right), \qquad g_s(i) = \mathbf{v}_i^{(s)}. \tag{17}$$

We use $r_{\mathrm{coarse}} = 6$, $r_{\mathrm{medium}} = 18$, and singleton groups at the fine scale. If row k corresponds to group $\mathcal{G}_k^{(s)}$, the averaging matrix is

$$(\mathbf{W}_s)_{ki} = \frac{\mathbb{1}[i \in \mathcal{G}_k^{(s)}]}{|\mathcal{G}_k^{(s)}|}, \qquad \mathbf{X}_{\mathrm{can}}^{(s)} = \mathbf{W}_s \mathbf{X}_{\mathrm{can}}, \quad \mathbf{Z}_0^{(s)} = \mathbf{W}_s \mathbf{Z}_0, \quad \mathbf{H}_{\mathrm{obs}}^{(j,s)} = \mathbf{W}_s \mathbf{H}_{\mathrm{obs}}^{(j)}. \tag{18}$$

Targets at every scale are pooled independently from the full-resolution field rather than recursively from a parent prediction.

Table 8: Single-resolution versus multi-resolution cascade.

Structure	Input	Elastic		Plasticine	
		$R^2 \uparrow$	$r \uparrow$	$R^2 \uparrow$	$r \uparrow$
single-fine	GT	0.6586	0.8178	0.4543	0.8186
multi-resolution	GT	0.7151	0.9435	0.5953	0.9192
single-fine	OBS	0.4474	0.6701	0.2935	0.7110
multi-resolution	OBS	0.6377	0.8601	0.4618	0.8190

The multi-resolution cascade outperforms the single-fine model for both material families under both GT and OBS inputs, with particularly clear gains under OBS inputs. This supports the benefit of coarse-to-fine field recovery.

A.6.2 Parent construction and diffusion schedule

For child stage s, draw $c_{\mathrm{par}} \sim \mathrm{Categorical}(\boldsymbol{\pi}_{\mathrm{par}})$. The parent is

$$(\mathbf{Z}_{\mathrm{par}}^{(s^-)}, q^{(s)}) = \begin{cases} (\mathbf{Z}_0^{(s^-)}, 0), & c_{\mathrm{par}} = \mathrm{gt}, \\ (\mathbf{Z}_0^{(s^-)} + \sigma_{\mathrm{par}}\boldsymbol{\zeta}, \sigma_{\mathrm{par}}), & c_{\mathrm{par}} = \mathrm{pert}, \\ (\mathrm{stopgrad}(\widehat{\mathbf{Z}}_0^{(s^-)}), 1), & c_{\mathrm{par}} = \mathrm{pred}, \end{cases} \tag{19}$$

where $\boldsymbol{\pi}_{\mathrm{par}} = (0.15, 0.25, 0.60)$ and $\sigma_{\mathrm{par}} = 0.41$. The scalar $q^{(s)}$ records how the parent was constructed; it is not a confidence estimate.

Parent-to-child interpolation uses at most eight inverse-distance-weighted neighbors:

$$(\mathbf{Z}_{\mathrm{par}}^{(s)})_i = \sum_{p \in \mathcal{N}_i} \omega_{ip} (\mathbf{Z}_{\mathrm{par}}^{(s^-)})_p, \qquad \omega_{ip} = \frac{\max(d_{ip}, \varepsilon)^{-1}}{\sum_{p' \in \mathcal{N}_i} \max(d_{ip'}, \varepsilon)^{-1}}. \tag{20}$$

The base diffusion process has $T = 1000$ steps, linearly spaced β_t from 10^{-4} to 2×10^{-2}, and zero-terminal-SNR rescaling. Timesteps are sampled uniformly and

$$w(t) = \max(1 - \bar{\alpha}_t, 0.1). \tag{21}$$

We use $w_{\mathrm{coarse}} = w_{\mathrm{medium}} = w_{\mathrm{fine}} = 1$.

A.7 Multi-Interaction Fusion

Fuse_s applies shared history, global-condition, and canonical-coordinate encoders to every interaction. Let their width-D outputs be $\mathbf{U}_{H,i}^{(j,s)}$, $\mathbf{U}_{G,i}^{(j,s)}$, and $\mathbf{U}_{X,i}^{(j,s)}$. A residual adapter forms

$$\mathbf{R}_i^{(j,s)} = \mathbf{U}_{H,i}^{(j,s)} + 0.25\, f_{\mathrm{adapt}}([\mathbf{U}_{H,i}^{(j,s)}, \mathbf{U}_{G,i}^{(j,s)}, \mathbf{U}_{X,i}^{(j,s)}]), \qquad a_i^{(j,s)} = f_{\mathrm{score}}(\mathbf{R}_i^{(j,s)}). \tag{22}$$

Scores are normalized only across interactions:

$$\begin{aligned} w_i^{(j,s)} &= \frac{\exp a_i^{(j,s)}}{\sum_{j' \in \mathcal{J}} \exp a_i^{(j',s)}}, \\ \mathbf{U}_i^{(s)} &= \sum_{j \in \mathcal{J}} w_i^{(j,s)} \mathbf{R}_i^{(j,s)}, \qquad \mathbf{E}_i^{(s)} = \sum_{j \in \mathcal{J}} w_i^{(j,s)} \mathbf{E}_i^{(j,s)}, \\ \overline{\mathbf{H}}_i^{(s)} &= \sum_{j \in \mathcal{J}} w_i^{(j,s)} \widetilde{\mathbf{H}}_i^{(j,s)}, \qquad \overline{\mathbf{G}}_i^{(s)} = \sum_{j \in \mathcal{J}} w_i^{(j,s)} \mathbf{G}^{(j)}. \end{aligned} \tag{23}$$

Thus $\overline{\mathbf{H}}^{(s)} \in \mathbb{R}^{N_s \times F \times 12}$ and $\overline{\mathbf{G}}^{(s)} \in \mathbb{R}^{N_s \times 28}$ are particle-wise weighted summaries using the same normalized interaction weights as the fused features. We use $D = 256$ and evidence width $D_e = 64$. The score head is LayerNorm–$256 \to 64 \to 1$, and the adapter uses widths $(3D, D, D)$. The resulting aggregation is permutation invariant to interaction order.

A.8 Denoiser Architecture and Constitutive Evidence Routing

At each resolution, the χ-specific denoiser consumes the noisy residual, fused condition, diffusion time, and, for non-coarse levels, the interpolated parent field. A Point Transformer encoder–decoder models local particle interactions, with a global-attention bottleneck providing object-level context. The elastic model predicts two property channels and the plastic model predicts three, consistently with the dimensions specified in the main text. The three external cascade stages share the same architecture but not parameters.

History evidence is partitioned into volume, deviatoric, position, and velocity groups. Each group has absolute-history and first-frame-relative branches and produces learned temporal summaries. Volume and deviatoric evidence contribute three tokens each; position and velocity contribute six motion tokens in total, giving $L_e = 12$.

Table 9: Condition encoder and denoiser configuration.

Component	Configuration
Feature widths	Shared width $D = 256$; evidence width $D_e = 64$; $L_e = 12$ tokens
History branches	Volume, deviatoric, position, and velocity; absolute and first-frame-relative two-layer MLPs ($c \to 64 \to 64$)
Temporal encoder	Frame fusion to 256; two-layer GRU; two single-head temporal self-attention blocks; learned temporal pooling
Known-condition encoders	28D global condition and canonical coordinates mapped by two-layer width-256 MLPs; four Fourier frequency bands for coordinates
Diffusion condition	Sinusoidal 256D time embedding followed by a two-layer MLP; noisy residual encoded by a two-layer MLP; child stages jointly encode $(\mathbf{Z}^{(s)}_{t,\mathrm{diff}}, \mathbf{Z}^{(s)}_{\mathrm{par}}, q^{(s)})$
Spatial trunk	Point U-Net with six Point Transformer blocks distributed over three encoder levels, the bottleneck, and two decoder levels; skip connections; internal voxel grids 12^3 and 6^3
Spatial attention	20-neighbor kNN graph, eight attention heads, query chunk 512; one eight-head global-attention block at the bottleneck
Interaction fusion	Residual $3D \to D \to D$ adapter and LayerNorm–$256 \to 64 \to 1$ score MLP
Output	Independent normalized MLP condition projection and scalar prediction head for each active physical coordinate
Regularization	SiLU activations; dropout 0

The property-specific routing is

$$\begin{aligned}
\mathbf{B}^{(s)}_K &= [0.05\mathbf{U}^{(s)}_{\mathrm{PT}}, \mathbf{E}^{(s)}_{\mathrm{vol}}, 0.5\mathbf{E}^{(s)}_{\mathrm{dev}}, 0.25\mathbf{E}^{(s)}_{\mathrm{mot}}],\\
\mathbf{B}^{(s)}_G &= [0.05\mathbf{U}^{(s)}_{\mathrm{PT}}, 0.1\mathbf{E}^{(s)}_{\mathrm{vol}}, \mathbf{E}^{(s)}_{\mathrm{dev}}, \mathbf{E}^{(s)}_{\mathrm{mot}}],\\
\mathbf{B}^{(s)}_{\gamma_y} &= [0.05\mathbf{U}^{(s)}_{\mathrm{PT}}, 0.1\mathbf{E}^{(s)}_{\mathrm{vol}}, \mathbf{E}^{(s)}_{\mathrm{dev}}, \mathbf{E}^{(s)}_{\mathrm{mot}}].
\end{aligned} \tag{24}$$

Each branch uses an independent normalization, projection, and prediction head. The coefficients balance feature scales and are not probabilities or confidence values.

A.8.1 Training configuration

Elastic and Plasticine are trained independently with the same optimization protocol.

Table 10: Training protocol for each material-family model.

Item	Configuration
Data split	20 training objects and 4,000 records per family in train dataset: 3,800 training, 100 validation, and 100 reserved records; independent test dataset tree with 8 held-out objects in total (4 per family) and 80 test records in total (40 per family)
Update unit	One property-field record per optimizer update; N_{obs} is sampled uniformly from $\{1, 2, 4\}$
Optimizer	Fused AdamW on CUDA; learning rate 10^{-4}; weight decay 10^{-4}
Numerics	bfloat16 mixed precision; global gradient-norm clipping at 5.0
Budget	50 epochs, i.e., 190,000 optimizer updates per family; one GPU per family-specific model
EMA and selection	Decay 0.999 with a 1,000-update warm-up; select the checkpoint minimizing validation clean-field x_0 MSE under EMA weights

The split unit is the property-field realization, not an individual interaction. All interactions of one realization remain in the same split. Test object identities are absent from the training tree. Their GT

Table 11: OBS-only DDIM step ablation ($N_{\text{obs}} = 4$).

L	Elastic		Plasticine	
	$R^2 \uparrow$	$r \uparrow$	$R^2 \uparrow$	$r \uparrow$
25	0.8640	0.9539	0.7397	0.9125
50	0.8544	0.9512	0.7388	0.9117
100	0.8502	0.9499	0.7353	0.9101
500	0.8456	0.9484	0.7348	0.9097
1000	0.8450	0.9483	0.7346	0.9096

assets, simulator trajectories, and material fields are retained only by the benchmark for quantitative evaluation; HeteroPhys inference reconstructs its canonical representation and observation histories from the multi-view videos.

A.9 DDIM Sampling

Each scale starts from an independent Gaussian state and follows the deterministic reverse schedule $0 = \tau_0 < \cdots < \tau_L = T$, using $L = 25$ steps unless otherwise stated. The completed parent is fixed before sampling the next scale. We use velocity prediction: for a clean residual $\mathbf{Z}_0^{(s)}$ and noise $\boldsymbol{\epsilon}$, the training target is $\mathbf{V}_{\tau_t}^{(s)} = \sqrt{\bar{\alpha}_{\tau_t}}\,\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_{\tau_t}}\,\mathbf{Z}_0^{(s)}$. The following two equations are an equivalent factorization of the explicit reverse update in the main text:

$$\begin{aligned} \widehat{\mathbf{Z}}_{0,\text{diff},t}^{(s)} &= \sqrt{\bar{\alpha}_{\tau_t}}\,\mathbf{Z}_{\tau_t,\text{diff}}^{(s)} - \sqrt{1-\bar{\alpha}_{\tau_t}}\,\widehat{\mathbf{V}}_{\tau_t}^{(s)}, \\ \widehat{\boldsymbol{\epsilon}}_{\tau_t}^{(s)} &= \sqrt{1-\bar{\alpha}_{\tau_t}}\,\mathbf{Z}_{\tau_t,\text{diff}}^{(s)} + \sqrt{\bar{\alpha}_{\tau_t}}\,\widehat{\mathbf{V}}_{\tau_t}^{(s)}. \end{aligned} \tag{25}$$

The deterministic DDIM transition is

$$\mathbf{Z}_{\tau_{t-1},\text{diff}}^{(s)} = \sqrt{\bar{\alpha}_{\tau_{t-1}}}\,\widehat{\mathbf{Z}}_{0,\text{diff},t}^{(s)} + \sqrt{1-\bar{\alpha}_{\tau_{t-1}}}\,\widehat{\boldsymbol{\epsilon}}_{\tau_t}^{(s)}. \tag{26}$$

Deterministic here means deterministic conditional on the initial Gaussian state; changing that state yields another sample from the learned conditional distribution.

The results do not show a meaningful benefit from increasing the number of DDIM steps; instead, all metrics decline slightly as sampling becomes longer. Given the additional inference cost, 25 steps provide the most favorable trade-off.

A.10 Differentiable MPM Refinement

The final refinement is end-to-end differentiable and uses exactly the quantities summarized by $\mathcal{L}_{\text{render}}$ in the main text: the diffusion field $\widehat{\mathbf{P}}_0$, cached Gaussian state $\mathcal{Q}_{\text{can}}$, internal-particle support $\mathbf{X}_{\text{can}}$, calibrated cameras, observed RGB images and masks, and interaction configurations. Only the object-particle properties are optimized for R steps; the reconstruction, particle support, cameras, and known scene configuration remain constant.

For interaction j, the APIC/MPM solver produces particle states $\mathbf{S}_f^{(j)}(\mathbf{P})$ under the candidate field. The differentiable driving operator transports the canonical Gaussians with the simulated particle motion and local deformation,

$$\mathcal{Q}_f^{(j)}(\mathbf{P}) = \text{DriveGaussians}(\mathcal{Q}_{\text{can}}, \mathbf{S}_f^{(j)}(\mathbf{P}), \mathbf{G}^{(j)}). \tag{27}$$

Gaussian centers and covariance geometry follow the particle state, while reconstructed appearance features and opacities are reused from $\mathcal{Q}_{\text{can}}$. A differentiable Gaussian renderer $\mathcal{R}$ then produces the RGB image and silhouette,

$$(\widehat{\mathcal{I}}_f^{(j)}, \widehat{\mathcal{M}}_f^{(j)}) = \mathcal{R}(\mathcal{Q}_f^{(j)}(\mathbf{P}), \mathcal{K}^{(j)}). \tag{28}$$

The refinement objective is

$$\mathcal{L}_{\text{render}}(\mathbf{P}) = \frac{1}{|\mathcal{J}|(F-1)} \sum_{j\in\mathcal{J}} \sum_{f=1}^{F-1} \big[\lambda_{\text{rgb}}\,\rho(\widehat{\mathcal{I}}_f^{(j)} - \mathcal{I}_f^{(j)}) + \lambda_{\text{mask}}\,\text{BCE}(\widehat{\mathcal{M}}_f^{(j)}, \mathcal{M}_f^{(j)})\big], \tag{29}$$

where ρ is the robust photometric penalty and $\lambda_{\mathrm{rgb}}, \lambda_{\mathrm{mask}}$ are fixed weights. The shared frame $f = 0$ is excluded because it contains no dynamic material evidence.

Crucially, P2G transfer, constitutive updates, contact, G2P transfer, particle-to-Gaussian driving, and rendering all remain in one computation graph. Consequently, $\nabla_{\mathbf{P}} \mathcal{L}_{\mathrm{render}}$ is propagated from the rendered image loss through the Gaussian renderer, the driving operator, and the complete MPM rollout to every optimized object-particle parameter. This is the central role of differentiable MPM in our system.

The refinement follows exactly the main-text schedule $\mathcal{R} = \{\eta_r\}_{r=0}^{R-1}$ and updates only the object-particle field, $\widehat{\mathbf{P}}_{r+1} = \widehat{\mathbf{P}}_r - \eta_r \nabla_{\widehat{\mathbf{P}}_r} \mathcal{L}_r$. Fixed scene properties are read from the known solver configuration when each rollout is assembled and receive no gradient. Non-finite losses or gradients and rollouts leaving the safe MPM stencil reject the current update. After R scheduled steps, $\widehat{\mathbf{P}}_{\mathrm{deploy}} = \widehat{\mathbf{P}}_R$. For field evaluation only, this field is interpolated to ground-truth object-particle locations.

B Additional Experimental Setup

B.1 Benchmark and Interaction Details

Physical-Field Generation. We consider five physical-field distributions for each object. A *homogeneous* field assigns a single physical-parameter vector to all object particles. For *two-region* and *three-region* fields, the canonical object volume is partitioned into two or three spatially coherent regions generated from low-frequency object-relative implicit fields, producing curved interfaces and localized regions rather than fixed planar partitions. A *core–shell* field assigns different physical parameters to an interior core and the surrounding outer region. Finally, the *gradient* setting varies the physical parameters continuously along a randomly sampled object-relative direction. The homogeneous subset is used for cross-method comparison, whereas the four heterogeneous distributions are used for direct field-recovery experiments.

Interaction Generation. Each fixed object–field realization is simulated under 12 interactions formed by four canonical orientations ($0°$, $90°$, $180°$, and $270°$) and three rigid terrains: flat, a $30°$ incline, and a sinusoidal surface. All interactions of the same realization share the same canonical particle identities and physical-property field. Each interaction is a free-fall sequence of 20 frames at 60 fps and is rendered from 11 calibrated cameras. During training, the number of observed interactions is sampled from $N_{\mathrm{obs}} \in \{1, 2, 4\}$.

The benchmark contains 20 training object geometries and eight geometrically unseen evaluation objects, with four evaluation objects per constitutive family. For every evaluation object and each of the five distributions, we generate two independent physical-field realizations. The evaluation set therefore contains $8 \times 5 \times 2 = 80$ object–field records: 16 homogeneous records and 64 records across the four heterogeneous distributions.

B.2 Unified Evaluation Protocol

Heterogeneous-Field Recovery. For physical channel c, let z_c denote the coordinate evaluated at the canonical object-particle query locations; fixed scene rows are excluded. We compute

$$R_c^2 = 1 - \frac{\sum_{j=1}^{M} \left(\widehat{z}_c(\mathbf{q}_j) - z_c^*(\mathbf{q}_j)\right)^2}{\sum_{j=1}^{M} \left(z_c^*(\mathbf{q}_j) - \overline{z_c^*}\right)^2}, \tag{30}$$

where the mean in the denominator is taken over the canonical query locations of the corresponding field realization.

We evaluate Elastic fields in $(\log K, \log G)$ and Plasticine fields in $(\log K, \log G, \log \gamma_y)$. R_c^2 is computed independently for each field realization and parameter channel, and then averaged over realizations of the same object and field distribution. The object-level headline score macro-averages the active physical channels:

$$R^2_{\mathrm{Elastic}} = \frac{1}{2}\left(R^2_{\log K} + R^2_{\log G}\right), \qquad R^2_{\mathrm{Plasticine}} = \frac{1}{3}\left(R^2_{\log K} + R^2_{\log G} + R^2_{\log \gamma_y}\right). \tag{31}$$

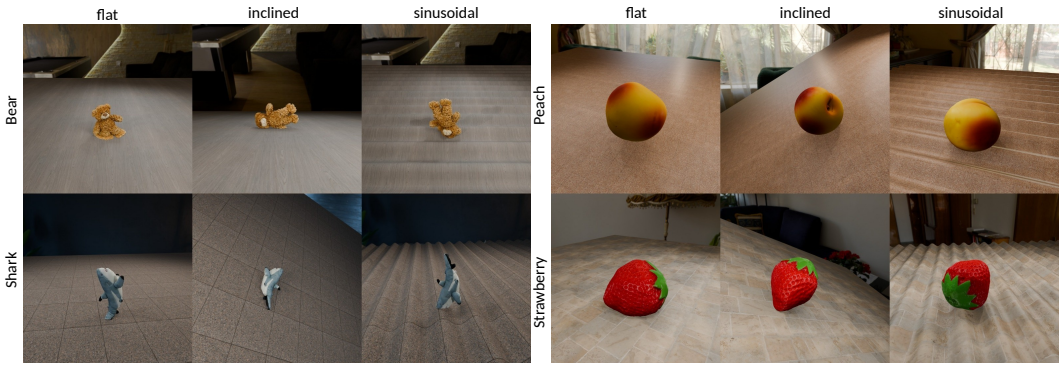

Figure 9: **More overview of the HeteroPhys benchmark.** The four geometrically unseen objects and interaction conditions used for evaluation.

A score of $R^2 = 0$ corresponds to predicting the spatial mean of the corresponding ground-truth channel, while negative values perform worse than this constant-field baseline.

Unless explicitly marked as a diffusion-only diagnostic, headline field-recovery and rollout metrics use the final refined field $\widehat{\mathbf{P}}_{\text{deploy}}$. We additionally retain the raw decode $\mathbf{P}_{\text{raw}}$, the projected diffusion initialization $\widehat{\mathbf{P}}_0 = \Pi_{\Omega_\chi}(\mathbf{P}_{\text{raw}})$, and the fraction of clipped entries as diagnostics of diffusion extrapolation and feasibility projection. Field-recovery metrics spatially interpolate $\widehat{\mathbf{P}}_{\text{deploy}}$ from the reconstruction-injected support to simulator-native GT object-particle query locations; scene particles are never scored.

GT and OBS denote two information regimes. GT histories use clean simulator trajectories; OBS histories pass through rendering, reconstruction, tracking, temporal differencing, and local deformation estimation. GT therefore acts as an upper-bound diagnostic, whereas OBS measures end-to-end deployment performance. Test-time inference uses OBS only.

B.3 Cross-Method Evaluation Protocol on Homogeneous Fields

The cross-method comparison in Table 4 uses the homogeneous subset of our benchmark so that every method is evaluated with a compatible representation target: a single spatially uniform physical parameter vector.

All methods use the same held-out *New Pose + New Terrain* target sequences and the same CD/PSNR/SSIM implementation. The target initial orientation and terrain are absent from the identification observations. Estimated parameters are frozen before target simulation, and target frames are never used for fitting, refinement, interaction selection, or early stopping.

For the observation-matched comparison, all baselines use one fixed identification interaction per realization. We report HeteroPhys under the same $N_{\text{obs}} = 1$ budget and additionally with $N_{\text{obs}} = 4$ interactions to quantify the benefit of multi-interaction evidence. PAC-NeRF, GIC, and MASIV are evaluated on Elastic and Plasticine subsets; Spring-Gaus is evaluated only on Elastic because its spring-mass formulation does not support the Plasticine constitutive family.

For geometric metrics, predicted and GT states are converted to point sets using the same sampling procedure. CD is symmetric and PSNR and SSIM are computed per rendered frame and camera view and then averaged. No post-hoc rigid alignment to the target state is applied.

C Additional Experimental Results

C.1 Parameter-Wise Heterogeneous Field Recovery

Table 1 reports the main object-wise field-recovery score after macro-averaging active physical coordinates. Here we expose the complementary parameter-wise results. Metrics are computed

Table 12: **Parameter-wise R^2 for Elastic heterogeneous fields.**

Coordinate	Parameter	Two-region	Three-region	Core–shell	Gradient	Mean
Bulk–shear	$\log K$	0.8198	0.8070	0.7659	0.8422	0.8087
	$\log G$	0.8176	0.8092	0.8433	0.8480	0.8295
Decoded	$\log E$	0.8209	0.8100	0.8431	0.8549	0.8322
	ν	0.6170	0.3231	0.7170	0.3348	0.4979

Table 13: **Parameter-wise R^2 for Plasticine heterogeneous fields.**

Coordinate	Parameter	Two-region	Three-region	Core–shell	Gradient	Mean
Bulk–shear	$\log K$	0.8412	0.4935	0.6909	0.6914	0.6792
	$\log G$	0.8540	0.3529	0.7016	0.6432	0.6379
	$\log \gamma_y$	0.3422	0.8318	0.6724	0.4681	0.5786
Decoded	$\log E$	0.8674	0.3923	0.7128	0.6737	0.6615
	ν	-0.3546	-0.8562	-0.7164	-1.8223	-0.9374
	$\log \tau_y$	0.2590	0.0850	0.0748	-1.8537	-0.3587

independently for each field realization over canonical query locations, averaged over realizations of the same object and distribution, and finally averaged over held-out objects.

For the bulk–shear parameterization, we report $(\log K, \log G)$ for Elastic and $(\log K, \log G, \log \gamma_y)$ for Plasticine. For transparency, we also physically decode every prediction and report the corresponding $(\log E, \nu)$ or $(\log E, \nu, \log \tau_y)$ parameters.

Error After Physical Decoding. Because R^2 depends on the spatial variance of the ground-truth field, we additionally evaluate the physically decoded parameters using $\mathrm{MAE}(\log E)$, $\mathrm{MAE}(\nu)$, and, for Plasticine, $\mathrm{MAE}(\log \tau_y)$:

$$\mathrm{MAE}_c = \frac{1}{M} \sum_{j=1}^{M} |\widehat{z}_c(\mathbf{q}_j) - z_c^*(\mathbf{q}_j)| \,, \tag{32}$$

with $z_c = \log E$, $z_c = \nu$, or $z_c = \log \tau_y$ as appropriate.

Table 14: **MAE after physical decoding.**

Material	Metric	Two-region	Three-region	Core–shell	Gradient	Mean
Elastic	$\mathrm{MAE}(\log E)$	0.2125	0.2639	0.1600	0.1230	0.1898
	$\mathrm{MAE}(\nu)$	0.0224	0.0260	0.0209	0.0170	0.0216
Plasticine	$\mathrm{MAE}(\log E)$	0.2765	0.4551	0.4126	0.1902	0.3336
	$\mathrm{MAE}(\nu)$	0.0525	0.0573	0.0600	0.0339	0.0509
	$\mathrm{MAE}(\log \tau_y)$	0.4592	0.6614	0.4907	0.4076	0.5047