DREAMCLOTH: STRUCTURAL COUNTERFACTUAL SCORE FACTORIZATION FOR BEHAVIORAL SYSTEM IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Frozen generative models are increasingly used as differentiable objectives, yet their scores entangle many scene variations while an inverse solver may control only one factor. We introduce *structural counterfactual score factorization*: a differentiable renderer builds a matched 2×2 lattice of scene variants, and a frozen video prior scores all four under one stochastic draw. By construction, the $(+, -, -, +)$ mixed difference cancels the constant and both one-factor main effects of the scores and returns their interaction, without assuming that the network is linear or semantically disentangled. In DreamCloth the two axes are garment appearance and articulated-body motion, and the interaction is distilled through a differentiable codimensional MPM simulator to optimize density, stretching stiffness, thickness and contact friction. We pair this objective with an explicit identification boundary: held-out rollout fidelity supports *behavioral* identification, whereas unique material constants require separate evidence. On four sequences each from ActorsHQ and 4D-DRESS, using multi-view images but no tracked garment supervision, factorized guidance improves held-out geometry and rendering over direct SDS, though it does not reach capture-based methods; on one sequence, removing the fourth corner worsens held-out Chamfer distance from 0.362 to 0.422. Gradient-sign agreement that varies strongly with the prior, boundary-saturated solutions and a 25-initialization landscape study show where latent uniqueness is not supported. The result is an objective-design principle for black-box generative priors in inverse dynamics, together with a way to report what they do and do not identify.

1 INTRODUCTION

Frozen generative models are increasingly repurposed as differentiable objectives: instead of sampling from them, one backpropagates their score into an external renderer, simulator, or structured latent representation (Poole et al., 2023; Wang et al., 2023). This is attractive because the prior is fixed and reusable, but it creates a structural mismatch. A generic score describes all changes that make an observation more likely, whereas an inverse solver often controls only one factor of the scene. If that factor is not directly addressable by text or labels, raw score distillation can spend gradient on appearance, identity, or motion dimensions that the solver cannot and should not change.

This problem becomes sharper in inverse dynamics. A pretrained video model can prefer plausible trajectories, but it is not a calibrated physical sensor. Even if its gradients improve a simulator's held-out behavior, the latent material parameters that produce that behavior may be non-unique, weakly constrained, or pinned to a search boundary. We therefore ask two linked questions:

> Can a structured generator extract a target interaction from a frozen generative objective without retraining it, and what can the resulting signal actually identify about an explicit physical system?

We answer the first question with *structural counterfactual score factorization*. The intervention is performed in the input space of the frozen model: a differentiable renderer constructs matched variants of the same scene with selected factors removed, and the prior evaluates every variant under

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.

the same conditioning, diffusion time, and noise. For two binary factors the four scores form a 2×2 table. Their Möbius/inclusion–exclusion contrast $(+, -, -, +)$ cancels the table's constant and both one-factor main effects exactly; this statement holds for an arbitrary nonlinear black-box score function. The remaining term is the renderer-defined interaction. We do not claim that the video model internally disentangles the corresponding semantics, only that the matched finite difference removes the specified main effects from the four queried values.

We instantiate this idea in DreamCloth, an inverse cloth system built around a differentiable codimensional Material Point Method (MPM) simulator. The two counterfactual axes are garment appearance and articulated-body motion: a static clothed rollout, a garment-free moving body, and a garment-free static body complement the simulated clothed rollout. The resulting interaction defines Factored Score Distillation (FSD), which is differentiated through the simulator; non-smooth contact friction is optimized with simultaneous perturbation stochastic approximation (SPSA; Spall, 1992). This setting is intentionally difficult: cloth motion depends on latent material properties and body–garment contact, many parameter combinations can produce similar drape, and capture-based methods obtain stronger measurements through calibrated multi-view video (Bhat et al., 2003; Zheng et al., 2024). Our goal is not to replace those measurements with a generative model and then call the result metrology. It is to test whether a frozen prior can provide a better-structured inverse signal and to characterize the strength of the identification evidence it provides.

We therefore separate *behavioral* identification from *parameter* identification. Behavioral identification means that the optimized simulator predicts held-out garment dynamics under the evaluation protocol. Parameter identification would additionally require the latent vector to be uniquely and stably determined, preferably against known ground truth. The main benchmark supports the former. Gradient sign agreement, boundary saturation, and an auxiliary 25-initialization geometric-optimizer study diagnose pieces of the latter but do not prove it. This distinction turns negative cases into useful evidence rather than hiding them.

Protocol note. Every experiment in this paper uses multi-view input. Stage 1 builds the scene from the multi-view training frames, as the capture-based baselines do. The difference is what happens next: those baselines fit garment geometry to every frame, which requires tracked, time-synchronized multi-view video, whereas DreamCloth fits no garment geometry at all and takes its dynamics signal from the video prior. The pipeline can also start from a single image with synthesized views, but we do not evaluate that setting here. Test-sequence motion is never used by the inverse objective (Table S4). We use *simulation-consistent* or *behaviorally consistent* for recovered parameters unless a stronger identification statement is explicitly supported.

Our Contributions. We formulate FSD as a four-point Möbius/inclusion–exclusion contrast that cancels the constant and the one-factor main effects of an arbitrary nonlinear black-box score table; the claim is algebraic and assumes no semantic disentanglement. A shared (τ, ϵ) draw preserves every branch's marginal expectation, cancels the explicit residual-noise term in the mixed difference, and creates covariance terms that can reduce variance when matched branches are positively correlated, though we claim no universal variance reduction over vanilla score distillation sampling (SDS; Poole et al., 2023). We distill the factorized signal through a contact-rich codimensional MPM cloth simulator and optimize explicit, re-simulatable material parameters while keeping latent non-identifiability distinct from held-out behavioral success. Empirically, Base SDS $\rightarrow$ four-branch FSD $\rightarrow$ timestep annealing improves held-out fidelity on ActorsHQ and 4D-DRESS; removing the fourth corner worsens CD $0.362 \rightarrow 0.422$ in a matched single-sequence ablation; and gradient sign agreement varies widely across priors and guidance, while boundary saturation and the auxiliary 25-initialization landscape study mark where stronger latent claims are unsupported.

DreamCloth models garments as codimensional thin shells with an anisotropic MPM formulation that separates in-plane, shear, and normal responses. The contribution is the combination of a precise generative objective construction with a calibrated interpretation of what inverse physics evidence it does and does not provide.

2 Background and Related Work

Cloth identification and differentiable physics. Recovering stiffness, damping, friction and density from observed deformation is a classical inverse problem, first approached with video, explicit correspondences and controlled temporal capture (Bhat et al., 2003); because many parameter sets explain a similar static drape, reliable recovery needs additional motion supervision, material priors and stable contact constraints. The Material Point Method (Sulsky et al., 1994; Stomakhin et al., 2013) supplies the forward model, improved for stability and interaction by MLS-MPM (Hu et al., 2018) and extended to 3D Gaussian representations by PhysGaussian (Xie et al., 2024); ChainQueen (Hu et al., 2019) and DiffTaichi (Hu et al., 2020) make its parameters optimizable through contact-rich rollouts.

Diffusion priors for inverse dynamics. Pretrained diffusion models supply score information for differentiable inverse problems (Poole et al., 2023; Wang et al., 2023), and video diffusion models encode temporal statistics of plausible motion (Ho et al., 2022b;a). PhysDreamer (Zhang et al., 2024), Physics3D (Liu et al., 2024) and DreamPhysics (Huang et al., 2025) estimate MPM material properties of 3D Gaussians under video-diffusion supervision, the last adding motion-oriented temporal reweighting against the color bias of vanilla SDS. Two limitations persist: the raw video score still entangles appearance, articulated body motion and target dynamics, which reweighting does not structurally remove — FSD instead subtracts those residuals explicitly under shared noise — and these methods model volumetric visco-elastic or elastoplastic objects rather than codimensional cloth, which buckles into wrinkles instead of compressing, so generic isotropic models wrongly couple in-plane and out-of-plane modes. We therefore adopt codimensional MPM (Jiang et al., 2017) (Sec. 3.4) together with frictional contact for thin, self-colliding surfaces (Li et al., 2021); to our knowledge no prior video-prior distillation method optimizes codimensional cloth physics.

Capture-based avatars, and what is identifiable. PhysAvatar (Zheng et al., 2024) recovers fabric parameters from multi-view video by inverse rendering and inverse physics, and MPMAvatar (Lee et al., 2025) couples 3D Gaussian avatars to a codimensional MPM simulator with anisotropic cloth elastoplasticity (Jiang et al., 2017) and mesh-collider contact robust to SMPL-X self-penetration, whose forward model we adopt. Both rely on tracked surface correspondences or garment trajectories: strong supervision, but it requires the physical garment, calibrated capture and per-garment dynamic observations. DreamCloth replaces that garment tracking with a frozen rectified-flow video prior: it needs multi-view images of the subject but never fits garment geometry to them. We evaluate on ActorsHQ (Işık et al., 2023) and 4D-DRESS (Wang et al., 2024). Because that observation model is a learned distribution over plausible videos rather than a physical law, several material parameter sets may receive similar supervision. This is a real limitation of the supervision, and we analyze it directly: behavioral fidelity and latent parameter uniqueness are distinct, and we diagnose which holds for each sequence (Sec. 4.3); Appendix J compares the methods directly.

3 Methodology

3.1 Problem Setup

Given images of a clothed person, with reference frame I_0 (multi-view images in all our experiments), we estimate an interpretable cloth material field over the garment surface,

$$\boldsymbol{\Theta}(\mathbf{x}) = \big(\rho(\mathbf{x}),\ \kappa_s(\mathbf{x}),\ h(\mathbf{x}),\ \mu(\mathbf{x})\big), \tag{1}$$

assigning density ρ, stretching stiffness κ_s, shell thickness h, and contact friction μ to each material point $\mathbf{x}$ of the garment. Bending behavior is not a free parameter of its own: the normal-compression stiffness κ_n is held at a fixed reference value and the garment's resistance to bending is controlled through h, which sets the particle volumes and hence the thin-shell response (Sec. 3.4). The formulation allows $\boldsymbol{\Theta}(\mathbf{x})$ to vary spatially, for example as per-region constants for garments with mixed fabrics or stiff trims. In this paper we evaluate only homogeneous garments, where the field collapses to a single constant vector $\boldsymbol{\Theta}$; heterogeneous garments would need a larger parameterization and are not tested here. We assume no tracked garment geometry, no garment trajectories, and no per-garment calibration. We reconstruct a body proxy and garment surface, simulate short cloth rollouts with a differentiable MPM solver with contact, render the motion, and optimize $\boldsymbol{\Theta}$ via Factored Score Distillation from a frozen rectified-flow video generation model.

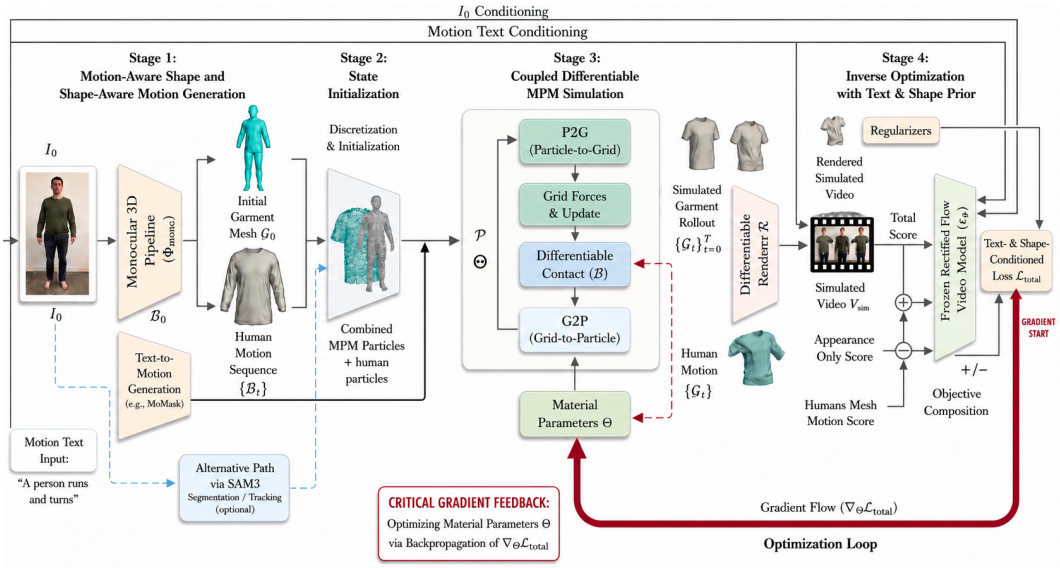

Figure 1: **DreamCloth overview.** From multi-view images (used in all our experiments) or a single image I_0 with synthesized views, we reconstruct a body proxy $\mathcal{B}$ and garment $\mathcal{G}_0$, discretize the garment into MPM particles, simulate short rollouts with differentiable contact, render the clothed rollout together with the static-appearance and body-motion factored branches, and optimize material parameters $\boldsymbol{\Theta}$ via Factored Score Distillation and physics-aware regularizers. Dashed arrows indicate gradient flow through the differentiable simulator and renderer.

3.2 Pipeline Overview

Our method has four stages: (1) 3D reconstruction of a posed body proxy $\mathcal{B}$ and garment mesh $\mathcal{G}_0$; (2) MPM state initialization; (3) differentiable cloth dynamics with body contact for T steps to obtain $\{\mathcal{G}_t\}_{t=0}^{T}$; and (4) factored-prior inverse optimization of $\boldsymbol{\Theta}$ with temporal/physics regularizers. Figure 1 summarizes the complete pipeline. Algorithm 1 gives the Stage-1 scene construction, and Algorithm S1 in the appendix the simulation-and-optimization loop.

3.3 3D Scene Hypothesis

From the input images we estimate a posed SMPL-X body proxy $\mathcal{B}$, an aligned garment surface $\mathcal{G}_0$, and a camera Π for differentiable rendering, $(\mathcal{B}, \mathcal{G}_0) = \Phi_{\text{mono}}(I_0)$ (Algorithm 1; details in Appendix A). The pipeline accepts either captured multi-view images $\{I_1, \ldots, I_N\}$ or novel views synthesized from a single image by a camera- and pose-controlled video model; all our experiments use captured multi-view images. These drive SMPL-X estimation (Pavlakos et al., 2019; Goel et al., 2023) and a 3D Gaussian Splatting avatar (Kerbl et al., 2023) with unified geometry and texture. Renderings of the avatar are segmented into body, upper/lower garment and dress by a finetuned SegFormer (Xie et al., 2021), or by SAM 3 (Carion et al., 2025) when SegFormer labels are unreliable, and the labels are unprojected into a classified 3D particle volume. The garment particles initialize $\mathcal{G}_0$; the body particles and the SMPL-X surface form the collider $\mathcal{B}$.

3.4 Codimensional MPM Cloth with Body Contact

We follow the codimensional cloth treatment of MPMAvatar (Lee et al., 2025) and the anisotropic elastoplastic model of Jiang et al. (2017). The garment remains a Lagrangian triangle mesh $\mathcal{G}_0$ carrying N_p material points with mass $m_p = \rho V_p$ and volume $V_p \propto h$. Each particle's deformation gradient is built from the two in-plane edge tangents and a normal director. Cloth responds differently to stretch, shear and compression, so we work in the rotation-invariant QR frame $\mathbf{F}_p = \mathbf{Q}_p \mathbf{R}_p$, which separates the in-plane block $\mathbf{R}_{2\times 2}$, the shear pair (r_{13}, r_{23}) and the normal component r_{33}:

$$\Psi(\mathbf{F}) = \Psi_{\text{plane}}(\mathbf{R}_{2\times 2}; \kappa_s, \nu) + \tfrac{\gamma}{2}\left(r_{13}^2 + r_{23}^2\right) + \Psi_n(r_{33}; \kappa_n)\,. \tag{2}$$

Algorithm 1 Stage-1 scene construction

Require: image(s) $\{I_k\}$ (single frame, or a multi-view set)
Ensure: classified particle volume $\mathcal{P}$, body proxy $\mathcal{B}$, garment mesh $\mathcal{G}_0$
1: $(\beta, \theta_0, \mathbf{t}_0) \leftarrow$ SMPL-X fit to $\{I_k\}$; $\mathcal{B} \leftarrow M(\beta, \theta_0, \mathbf{t}_0)$
2: fit a 3D Gaussian avatar to $\{I_k\}$ (multi-view, or novel views synthesized from a single frame)
3: $\{M_k\} \leftarrow$ garment masks from a finetuned SegFormer or text-prompted SAM 3
4: back-project $\{M_k\}$ to the template vertices; per-vertex majority vote $\rightarrow$ 3-class labels
5: topology-snap: assign each connected mesh shell wholly to its dominant class
6: body proxy $\mathcal{B}$ fills the garment–body gap left by the single-layer scan (LBS interior)
7: voxelize the garment shell into particles; set $m_p = \rho V_p$, $V_p \propto h$, $\mathbf{F}_p^0 = \mathbf{I}$
8: **return** $\mathcal{P}$, $\mathcal{B}$, $\mathcal{G}_0$

The first term is a fixed-corotated membrane energy with stretching stiffness $\kappa_s \equiv E$ and Poisson ratio ν, the second penalizes shear against the normal director with weight γ, and the third resists normal compression with stiffness κ_n and switches off under separation. Cloth–cloth friction is a plastic return mapping that projects the shear pair onto a Coulomb cone scaled by the fixed constant c_F and by the friction coefficient μ, so friction is part of the constitutive update rather than a separate collision pass. Of these quantities we optimize ρ, κ_s, h and μ. Since κ_n is fixed, bending is controlled through h: it sets the particle volumes and so scales the membrane and normal contributions of the thin shell differently.

The body collider is the posed SMPL-X mesh $\mathcal{B} = M(\beta, \theta_t, \mathbf{t}_t)$, re-posed at every frame by the driving pose and root translation $\{\theta_t, \mathbf{t}_t\}$; it is the only collider in the simulation. Standard MPM assumes analytic or level-set colliders, which suit a noisy, possibly self-intersecting body mesh poorly, so we never convert the body to a level set. Instead, following MPMAvatar, the body's per-frame surface velocities $\mathbf{v}^B$ and outward normals $\mathbf{n}^B$ are transferred to nearby grid nodes with the same B-spline weights. With $\hat{\mathbf{v}}_i$ the grid velocity at node i after the force update, we remove any inward relative motion,

$$\mathbf{v}_i^{\text{rel}} = \hat{\mathbf{v}}_i - \mathbf{v}_i^B, \qquad v_n = \mathbf{v}_i^{\text{rel}} \cdot \mathbf{n}_i^B, \qquad \mathbf{v}_i^{\text{rel}} \leftarrow \mathbf{v}_i^{\text{rel}} - \min(v_n, 0)\, \mathbf{n}_i^B, \tag{3}$$

and apply Coulomb friction to the tangential part $\mathbf{v}_{i,\parallel}^{\text{rel}} = \mathbf{v}_i^{\text{rel}} - (\mathbf{v}_i^{\text{rel}} \cdot \mathbf{n}_i^B)\, \mathbf{n}_i^B$, scaling it by $\max\Big(0,\, 1 - \mu\, |v_n| / (\|\mathbf{v}_{i,\parallel}^{\text{rel}}\| + \delta)\Big)$ with a small constant δ and the same μ as above. The collider velocity is then added back, $\mathbf{v}_i = \mathbf{v}_i^{\text{rel}} + \mathbf{v}_i^B$.

The grid solve is standard MLS-MPM/APIC (Hu et al., 2018; Jiang et al., 2015) with quadratic B-spline weights, used unmodified from the released MPMAvatar solver and implemented in NVIDIA Warp (Macklin, 2022). It gives the differentiable rollout $\{\mathcal{G}_t\} = \mathcal{S}_{\text{MPM}}(\mathcal{B}, \mathcal{G}_0; \boldsymbol{\Theta})$; every operation, including contact, is smooth almost everywhere, and gradients with respect to $\boldsymbol{\Theta}$ come from reverse-mode differentiation through all T steps. Small fixed probe excitations (gravity settling and a low-amplitude body oscillation) make short rollouts informative. Appendix B gives the material-point representation, the full constitutive model and every per-step transfer, force, damping and advection rule.

3.5 Differentiable Rendering

Each state renders to $\hat{I}_t = \mathcal{R}(\mathcal{B}, \mathcal{G}_t; \Pi)$, giving the clothed rollout $\mathbf{V}_{\text{sim}} = \{\hat{I}_t\}$. We additionally render three factored rollouts under the same camera and lighting: (i) an appearance branch $\mathbf{V}_{\text{app}}$, the clothed avatar held static (frame-0 configuration repeated for all T frames), which exposes the appearance/identity factor of the prior score; (ii) a body-motion branch $\mathbf{V}_{\text{body}}$, the SMPL-X body without the garment executing the same driving motion $\{\theta_t, \mathbf{t}_t\}$, which exposes the human-motion factor; and (iii) a static body-appearance branch $\mathbf{V}_{\text{body,app}}$, the garment-free body held static in its frame-0 configuration, which the inclusion–exclusion correction of Sec. 3.6 restores. All branches share the body identity, camera, lighting, and rendering setup, and differ only in the factor each isolates.

3.6 Structural Counterfactual Score Factorization

The raw video-prior residual carries gradients from appearance and articulated body motion that the simulator parameters do not control. FSD is designed to remove their main effects so that the objective concentrates on the garment dynamics. Whether the result points in a useful direction is an empirical question, examined in Sec. 4.3.

Prior and conditioning. Let $u_\phi(\cdot \mid y, \tau)$ be a frozen rectified-flow (Liu et al., 2023; Lipman et al., 2023) text-to-video model; its weights are never updated. We use Wan2.2-TI2V (5B) (Wan Team et al., 2025) by default and, to see how much the prior matters, also LTX-Video (2B) (HaCohen et al., 2025), Wan2.2-I2V-A14B (Wan Team et al., 2025) and PhyWorld (Kang et al., 2024) (Table S1). The conditioning y is the text embedding of a per-sequence prompt (e.g. "a person in a loose long-sleeve shirt and trousers, turning"), obtained by captioning I_0 once with a VLM. It is held fixed across iterations and across all four branches; otherwise any difference in conditioning would leak into the contrast as a spurious signal.

On the interpolation path $z_\tau(\mathbf{V}) = (1-\tau)\mathbf{V} + \tau\epsilon$, with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and $\tau \in [0, 1]$, the ground-truth velocity is $\epsilon - \mathbf{V}$. The velocity residual

$$\hat{u}_\phi(\mathbf{V}) = u_\phi(z_\tau(\mathbf{V}) \mid y, \tau) - (\epsilon - \mathbf{V}) \tag{4}$$

plays the role of the score residual in ϵ-prediction SDS: it measures how far the rendered video $\mathbf{V}$ is from what the prior expects at noise level τ.

Four-point interaction contrast. Let $F_{am} = \hat{u}_\phi(\mathbf{V}_{am})$ be the residual on the four matched renders, where $a \in \{0, 1\}$ indicates garment appearance and $m \in \{0, 1\}$ articulated motion. Any four values decompose exactly as

$$F_{am} = F_{00} + a\,\alpha_a + m\,\alpha_m + am\,\Gamma, \qquad \Gamma = F_{11} - F_{10} - F_{01} + F_{00}, \tag{5}$$

with $F_{11}, F_{10}, F_{01}, F_{00}$ given by $\mathbf{V}_{\text{sim}}$, $\mathbf{V}_{\text{app}}$, $\mathbf{V}_{\text{body}}$ and $\mathbf{V}_{\text{body,app}}$. FSD uses exactly Γ, so the constant and both main effects cancel for any nonlinear prior, with no assumption about its internal representation. What the algebra does not fix is the meaning of Γ: it is a garment–motion interaction and may contain learned correlations beyond physical cloth deformation, which is why our claims rest on held-out behavior and the diagnostics of Sec. 4.3 (derivation in Appendix C). The resulting residual and gradient are

$$\hat{u}_\phi^{\text{int}} = \hat{u}_\phi(\mathbf{V}_{\text{sim}}) - \hat{u}_\phi(\mathbf{V}_{\text{app}}) - \hat{u}_\phi(\mathbf{V}_{\text{body}}) + \hat{u}_\phi(\mathbf{V}_{\text{body,app}}), \tag{6}$$

$$\nabla_{\boldsymbol{\Theta}} \mathcal{L}_{\text{FSD}} = \mathbb{E}_{\tau,\epsilon}\left[w(\tau)\, \hat{u}_\phi^{\text{int}}\, \frac{\partial \mathbf{V}_{\text{sim}}}{\partial \boldsymbol{\Theta}} \right]. \tag{7}$$

Dropping the fourth branch changes the estimand and leaves an uncancelled $+\epsilon$ term; on s190 it raises held-out CD from 0.362 to 0.422 (Table 2).

Shared noise. All branches use the same (τ, ϵ), so their residuals are correlated. Since the subtracted branches do not depend on $\boldsymbol{\Theta}$ and have non-zero mean, they change the objective rather than acting as a pure control variate. Sharing the draw keeps each branch's marginal expectation and adds covariance terms that can reduce variance when the branches are positively correlated; we do not claim a provably lower-variance or unbiased estimator. The variance decomposition is in Appendix C.

3.7 Regularization and Optimization

The video prior alone leaves the problem under-constrained, so we add two light regularizers on the simulated rollout. The first suppresses jitter. With $\mathbf{q}_t$ the simulated garment vertex positions at frame t, we penalize acceleration,

$$\mathcal{L}_{\text{temp}} = \frac{1}{T-1} \sum_{t=1}^{T-1} \|\mathbf{q}_{t+1} - 2\mathbf{q}_t + \mathbf{q}_{t-1}\|_2^2, \tag{8}$$

Table 1: **Quantitative comparison on ActorsHQ (Işık et al., 2023) and 4D-DRESS (Wang et al., 2024).** Each model is optimized on 50 multi-view frames and evaluated on unseen test sequences (200 frames for ActorsHQ; 100 for 4D-DRESS). DreamCloth uses the training frames only to build the scene and the body motion; it fits no garment geometry to them. Four-branch FSD is the full estimator of Eq. (6). Masked CD is computed on garment regions segmented by SAM 3 (Carion et al., 2025). Bold and underline denote best and second-best.

	Geometry			Video / Appearance		
Method	CD ($\times 10^3$)↓	Masked CD ($\times 10^3$)↓	F-score↑	LPIPS↓	PSNR↑	SSIM↑
ActorsHQ						
ARAH (Wang et al., 2022)	1.098	2.149	87.8	0.0539	29.17	0.976
TAVA (Li et al., 2022)	0.647	1.160	94.1	0.0500	30.19	<u>0.981</u>
GaussianAvatar (Hu et al., 2024)	0.892	1.794	91.2	0.0431	<u>31.21</u>	<u>0.981</u>
PhysAvatar (Zheng et al., 2024)	0.539	1.097	<u>94.8</u>	0.0343	30.80	0.976
MPMAvatar (Lee et al., 2025)	**0.412**	**0.693**	**97.6**	**0.0323**	**32.64**	**0.982**
DreamCloth–Base SDS	0.606	1.301	90.0	0.0361	27.18	0.955
DreamCloth–Four-branch FSD	0.491	0.947	90.8	0.0351	28.12	0.957
DreamCloth–Four-branch FSD + Time Schedule (Ours)	<u>0.468</u>	<u>0.922</u>	91.0	<u>0.0334</u>	28.79	0.958
4D-DRESS						
PhysAvatar (Zheng et al., 2024)	<u>0.363</u>	0.743	<u>98.5</u>	0.0216	<u>33.86</u>	<u>0.996</u>
MPMAvatar (Lee et al., 2025)	**0.323**	**0.631**	**99.1**	**0.0176**	**34.78**	**0.997**
DreamCloth–Base SDS	0.507	1.004	91.7	0.0206	29.35	0.970
DreamCloth–Four-branch FSD	0.405	0.743	92.2	0.0199	30.15	0.970
DreamCloth–Four-branch FSD + Time Schedule (Ours)	0.375	<u>0.722</u>	92.3	<u>0.0186</u>	30.97	0.971

which is zero for constant-velocity motion, so it removes flicker without pulling the cloth toward rest; its weight is small ($\lambda_{\text{temp}} = 0.02$). The second, $\mathcal{L}_{\text{phys}}$, penalizes interpenetration, contact gap and slip, strain and bending violations, and mesh instability. The full objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{FSD}}\mathcal{L}_{\text{FSD}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}. \tag{9}$$

We update (ρ, κ_s, h) with Adam (Kingma & Ba, 2015), using gradients through the renderer and the MPM rollout, and update μ with SPSA because its gradient is non-smooth (Appendix B). The positive parameters are optimized in log space, and after every step each parameter is clamped to a fixed box in the solver's units: $\rho \in [0.1, 3.0]$, $\kappa_s \in [50, 2000]$ Pa, $h \in [0.8, 1.2]$ and $\mu \in [0.01, 1.0]$ (Table S7). All other constants are the defaults of the released MPMAvatar solver ($\kappa_n = 500$, $\gamma = 500$, $\nu = 0.3$, $c_F = \tan 40^\circ \approx 0.84$, $c_d = 1.1$). Only ν has a conventional physical value; κ_n and γ are reference values, since bending is set through h and we do not try to identify shear. Because the solver damps grid velocities only when $c_d < 1$, there is no explicit damping: all dissipation comes from frictional contact and the return mapping.

4 Experiments

We evaluate the method as an evidence ladder. First, does structural counterfactual factorization improve held-out simulator behavior relative to direct score distillation? Second, does the exact fourth-corner correction matter? Third, what additional evidence—gradient direction, boundary behavior, and landscape concentration—supports or limits claims about the latent material vector? The protocol is summarized in Sec. 1 and Appendix F.

4.1 Experiment 1: Does factorization improve behavioral identification?

ActorsHQ (Işık et al., 2023) contains eight actors, four female and four male, with two motion sequences per actor, yielding 16 sequences in total. We evaluate four actor–sequence pairs from ActorsHQ and four sequences from 4D-DRESS (Wang et al., 2024). Table 1 reports held-out geometric and rendering fidelity under the benchmark protocol, with Base SDS as the direct objective baseline and capture-based methods as stronger-supervision reference points. Removing the appearance and SMPL-X body-motion (Pavlakos et al., 2019) effects with four-branch FSD improves over Base SDS on both datasets (Poole et al., 2023) (ActorsHQ CD $0.606 \to 0.491$; 4D-DRESS $0.507 \to 0.405$), and timestep annealing helps further (0.468 and 0.375). This is our primary result: the form of the video-prior objective materially affects held-out rollout fidelity. Each entry is a single run averaged over four sequences, so these are consistent rather than statistically tested gains, and DreamCloth does not reach the capture-based baselines.

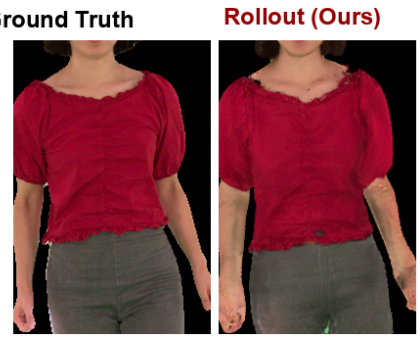

Figure 2: **Held-out 4D-DRESS s191 frame.** Ground truth (left) and DreamCloth (right); more in Figures S5 and S6.

Table 2: **Fourth-corner ablation (4D-DRESS s190).** Same initialization and settings; the three-branch estimator drops the static-body term. Both runs end on a box constraint, so the table compares objectives, not recovered materials.

Estimator	ρ	κ_s (Pa)	h	CD ($\times 10^3$) $\downarrow$
Four-branch FSD	0.563	50	0.881	**0.362**
Three-branch	0.109	535	1.200	0.422

Table 3: **Where do repeated optimizations land?** We rerun an auxiliary geometric optimizer from 25 initializations per sequence and report the spread of the recovered parameters (CV = 100×sd/mean; "Runs" = converged/attempted). s170 and s185 converge to a tight region, s190 spreads along a density–stiffness basin, and s191 piles up at the upper stiffness bound, so its small spread says nothing about identifiability. These are not 25 FSD runs, and no ground-truth material is available.

		ρ		$\kappa_{s,\text{stored}}$			
Sequence / garment	Runs	mean $\pm$ sd	CV (%)	mean $\pm$ sd	CV (%)	range	h
s170 upper	25 / 25	1.675 ± 0.015	0.90	1497 ± 42	2.81	1376–1534	0.975
s185 upper	25 / 25	1.668 ± 0.001	0.06	1498 ± 6	0.40	1489–1509	0.986
s190	25 / 25	1.773 ± 0.114	6.43	1242 ± 133	10.71	789–1358	0.987
s191	25 / 25	0.689 ± 0.002	0.29	2000 ± 0.2	0.01	1998–2000	0.984

Figure 2 zooms in on the garment in a held-out frame of 4D-DRESS s191: the simulated top keeps the puffed sleeves and the drape of the hem seen in the ground truth. Figures S5 and S6 and Table S6 (Appendix G) show the other sequences and full rollouts over time, including the s170 skirt, where the rollout is visibly worse.

4.2 Experiment 2: Does the exact structural correction matter?

The fourth branch is what makes the contrast an exact interaction, so we test whether it matters in practice. On 4D-DRESS s190, with the same initialization and settings, dropping it degrades held-out CD from 0.362 to 0.422 ($\approx$ 17%; Table 2, Appendix C). The supervision is identical in both runs, so the gap reflects the objective rather than the data, although it comes from a single sequence and seed.

4.3 What is identified?

A low held-out error certifies behavioral usefulness, but not that the latent parameter vector is uniquely determined. We use three diagnostics to characterize the gap. Each one supports a narrower claim than full identification; Table S3 (Appendix E) lists what each does and does not show.

Sign agreement (when does behavior carry parameter information?). As a diagnostic—not a proof—we measure how often the sampled FSD gradient agrees in sign with a geometric supervisor direction (Table S1, Appendix C; one initialization per setting). Agreement depends strongly on the prior/guidance pair: it is high for four-branch FSD with Wan2.2-A14B (0.72) and PhyWorld (0.77), near or below chance for Cloth Zoom on the same priors (0.35), and lowest for the Time-Schedule runs that diverged (0.12 and 0.33), so the guidance choice matters as much as the prior. Crucially, sign agreement and held-out error are distinct: LTX-Video attains low error only at the parameter

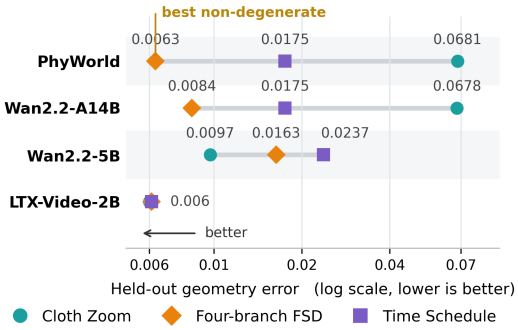

Figure 3: **Guidance strategy versus video prior.** Held-out geometry error on s191 (log scale, lower is better), one initialization per point. Four-branch FSD is best for PhyWorld and Wan2.2-A14B, and Cloth Zoom for Wan2.2-5B. LTX-Video scores lowest overall, but only because its parameters sit on the search bounds, so its result says little about the material.

bounds (boundary saturation, $\kappa_s = 2000$), so rollout quality alone is insufficient evidence of latent identification, and one initialization per setting gives no seed-robust ranking of priors.

Optimizer stability (does repeated optimization concentrate?). Table 3 reruns an auxiliary geometric optimizer from 25 initializations per sequence (not a 25-seed FSD study; Appendix E). Solutions are compact for s170 and s185, spread along a density–stiffness basin for s190, and pile up at the stiffness bound for s191.

A taxonomy of identification. Combining behavioral fit with these diagnostics yields three cases: *behavior + interior stability* (s170/s185), where the geometric proxy concentrates in a narrow region, our strongest evidence of an informative latent; *behavior + broad basin* (s190), behaviorally calibrated but set-valued in the latent vector; and *behavior + boundary saturation* (s191, LTX), a good trajectory with weak evidence about parameter magnitude beyond the box constraint. In short, plausibility to a video model is not physical uniqueness: the prior can guide the simulator to a good rollout while the latent parameters remain non-unique or bounded.

4.4 Video-prior backbones

Across the four priors of Sec. 3.6, recovery quality varies substantially and no single guidance strategy is best (Figure 3). These are single-initialization runs, and two Time-Schedule runs diverged; full results are in Table S1 and Appendix C.

Implementation details. Optimization is per sequence at test time; no network weights are trained. Hardware, optimizer settings and compute are in Appendix H and Table S7; the timestep, view and zoom ablations are in Appendix D.

5 Conclusion, Limitations, and Future Work

DreamCloth rests on two ideas. First, a frozen generative model can be queried with matched scene variants: the four-point mixed difference cancels the nuisance main effects of the scores and leaves an interaction signal that a simulator can optimize. Second, success with such an objective has to be read carefully: better held-out behavior shows that the objective is useful, not that the recovered material constants are unique or physically correct.

On eight sequences, FSD improves held-out rollouts over direct SDS, and on the sequence we tested, removing the fourth corner makes them worse. The limits are just as clear. Every result is a single run; the backbone study uses one initialization per setting and shows that sign agreement depends strongly on the prior and guidance; several solutions sit on parameter bounds; the 25-initialization study probes a geometric proxy rather than repeated FSD runs; the material model has only four garment-wide scalars, with bending set indirectly through thickness; and all evaluations use multi-view input. The next step is therefore stronger evidence rather than stronger claims: synthetic cloth with known parameters, repeated FSD runs, the single-image setting, and other inverse-dynamics problems.

AI USE STATEMENT

We used generative AI tools as a writing assistant: for drafting and editing prose, for improving clarity and phrasing, and for preparing the LaTeX manuscript. We did not use generative AI tools for research ideation, for the design of the method or of the experiments, for implementing the simulator, the estimator or the evaluation pipeline, or for producing or modifying any reported result. The remaining disclosure categories of the ICLR 2027 AI Policy for Authors are not applicable to this work.

All AI-assisted text was read, checked and edited by the authors. Every number, table and figure in this paper was produced by the differentiable simulator and the evaluation pipeline described in Sec. 3.6 and Appendices A–I; none was generated, completed or altered by a generative model. Separately, and as a matter of method rather than of authorship, DreamCloth queries frozen pretrained video generation models as its objective (Sec. 3.6); those models are the object of study, are cited and fully disclosed, and none of their weights is updated at any point. We take responsibility for the final content of this work, including all text, claims and artifacts.

ETHICS STATEMENT

This work raises no concerns that we believe fall under the ICLR Code of Ethics, but three points deserve to be stated explicitly.

Human data. All experiments use two existing, publicly released research datasets of captured humans, ActorsHQ (Işık et al., 2023) and 4D-DRESS (Wang et al., 2024), under the terms set by their authors. We collected no new data, recruited no participants, and performed no capture of our own; DreamCloth is a per-sequence test-time optimization that trains no network weights. No personally identifying attribute is inferred, and no identity information is used by the objective, which operates on rendered garment dynamics.

Potential for misuse. The pipeline reconstructs a clothed human avatar and re-simulates it under new motion, so it shares the dual-use profile of animatable-avatar and human-reenactment research generally: the same machinery that produces simulation-ready garment assets could be used to synthesize video of a real person without their consent. We recover only physical material parameters — density, stretching stiffness, shell thickness and contact friction — which describe fabric rather than a person and carry no identity. We nonetheless consider consent-aware use of the underlying avatar reconstruction to be a precondition for deploying this work outside a research setting.

Inherited bias and research integrity. Our objective queries frozen, publicly available video generation models (Wan Team et al., 2025; HaCohen et al., 2025; Kang et al., 2024), and therefore inherits whatever distributional bias those models carry over body shapes, clothing and appearance; we do not audit or correct for it, and we do not claim the prior is a calibrated physical sensor. On integrity, the central methodological commitment of this paper is to state the limits of its own evidence: we separate behavioral identification from parameter identification, and we report the boundary-saturated solutions, the single-initialization backbone studies, and the broad-basin sequences that argue against the stronger claim rather than omitting them (Sec. 4.3, Appendix E). We declare no competing interests.

REPRODUCIBILITY STATEMENT

The main components of DreamCloth are specified in the paper and the appendix, and the two evaluation datasets and all four video priors are public.

The forward model is given in full: the codimensional MLS-MPM/APIC solver, its material-point representation, the anisotropic constitutive model and the frictional return mapping are restated in Appendix B, with the per-step transfers as Eqs. (S5)–(S10) and the mesh-collider contact treatment alongside them; the SPSA estimator used for the non-smooth friction parameter, including its almost-unbiasedness argument, is Eqs. (S15)–(S18). The objective is given in Appendix C: the exact four-corner interaction contrast (Eq. (S21)), the shared-noise covariance decomposition (Eqs. (S22)–(S23)); the fourth-corner ablation that isolates its effect is Table 2. Scene construction from the input images, including both segmentation branches and the topology-snap step, is Appendix A and Algorithm 1.

Table S4 states the inputs, the driving-motion source, the probe excitation and the evaluation protocol, including the train/test frame split and the evaluation camera. Table S7 lists every fixed constitutive constant, every enforced box constraint, and the optimizer configuration — iteration count, per-parameter learning rates and schedule — so that the optimized values reported in Tables S1 and 3 can be read against their search limits. Table S2 records the diffusion-timestep, camera-view and garment-zoom settings behind the ablations. Appendix G reports the segmentation validation: the training-free promptable branch reproduces the manual ground-truth split exactly on the 4D-DRESS garments (Table S5), so those results do not depend on the finetuned segmenter.

We also state where reproduction will be limited. The backbone comparisons in Table S1 are single-initialization runs, the 25-initialization study in Table 3 probes an auxiliary geometric optimizer rather than repeated stochastic FSD convergence, and several reported solutions lie on a box constraint; these are diagnostics of the inverse landscape, not seed-robust measurements, and we label them as such throughout. Results are therefore expected to reproduce at the level of held-out behavioral fidelity, which is the claim the paper makes, rather than at the level of an exactly recovered latent material vector.

REFERENCES

Kiran S. Bhat, Christopher D. Twigg, Jessica K. Hodgins, Pradeep K. Khosla, Zoran Popović, and Steven M. Seitz. Estimating cloth simulation parameters from video. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, pp. 37–51, 2003.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Dídac Surís, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14783–14794, 2023.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2025.

Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2328–2337, 2023.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8633–8646, 2022b.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards realistic human avatar modeling from a single video via animatable 3D Gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 634–644, 2024.

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 (TOG)*, 37(4):150:1–150:14, 2018.

Yuanming Hu, Jiancheng Liu, Andrew Spielberg, Joshua B. Tenenbaum, William T. Freeman, Jiajun Wu, Daniela Rus, and Wojciech Matusik. ChainQueen: A real-time differentiable physical simulator for soft robotics. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6265–6271, 2019.

Yuanming Hu, Luke Anderson, Tzu-Mao Li, Qi Sun, Nathan Carr, Jonathan Ragan-Kelley, and Frédo Durand. DiffTaichi: Differentiable programming for physical simulation. In *International Conference on Learning Representations (ICLR)*, 2020.

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 (AAAI)*, volume 39, pp. 3733–3741, 2025.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. HumanRF: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):98:1–98:12, 2023.

Chenfanfu Jiang, Craig Schroeder, Andrew Selle, Joseph Teran, and Alexey Stomakhin. The affine particle-in-cell method. *ACM Transactions on Graphics (TOG)*, 34(4):51:1–51:10, 2015.

Chenfanfu Jiang, Theodore Gast, and Joseph Teran. Anisotropic elastoplasticity for cloth, knit and hair frictional contact. *ACM Transactions on Graphics (TOG)*, 36(4):152:1–152:14, 2017.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Changmin Lee, Jihyun Lee, and Tae-Kyun Kim. MPMAvatar: Learning 3D Gaussian avatars with accurate and robust physics-based dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Minchen Li, Danny M. Kaufman, and Chenfanfu Jiang. Codimensional incremental potential contact. *ACM Transactions on Graphics (TOG)*, 40(4):170:1–170:24, 2021.

Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. TAVA: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)*, pp. 419–436, 2022.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Miles Macklin. Warp: A high-performance Python framework for GPU simulation and graphics. https://github.com/nvidia/warp, 2022. NVIDIA GPU Technology Conference (GTC). Accessed: 2026-07-29.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Alexey Stomakhin, Craig Schroeder, Lawrence Chai, Joseph Teran, and Andrew Selle. A material point method for snow simulation. *ACM Transactions on Graphics (TOG)*, 32(4):102:1–102:10, 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.

Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. ARAH: Animatable volume rendering of articulated human SDFs. In *European Conference on Computer Vision (ECCV)*, pp. 1–19, 2022.

Wenbo Wang, Hsuan-I Ho, Chen Guo, Boxiang Rong, Artur Grigorev, Jie Song, Juan Jose Zarate, and Otmar Hilliges. 4D-DRESS: A 4D dataset of real-world human clothing with semantic annotations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 550–560, 2024.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 12077–12090, 2021.

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

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

Yang Zheng, Qingqing Zhao, Guandao Yang, Yifan Wang, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, and Gordon Wetzstein. PhysAvatar: Learning the physics of dressed 3D avatars from visual observations. In *European Conference on Computer Vision (ECCV)*, pp. 262–284, 2024.

REVIEWER'S GUIDE

This supplement provides the material referenced in the main paper and introduces no method-essential content. It is organized to let reviewers audit the complete chain from scene construction to objective mechanics to the strength of the identification evidence:

- Stage-1 scene construction and the single-image pathway: Sec. A, Figure S1 and main-paper Algorithm 1.
- Codimensional cloth simulator: Sec. B, including the c_F versus c_d notation.
- Structural counterfactual factorization, shared-noise analysis, and prior sensitivity: Sec. C, Algorithm S1 and Table S1. The fourth-corner ablation is main-paper Table 2.
- Diffusion timestep, viewpoints, and zoom: Sec. D, Table S2.
- Behavioral and latent identification diagnostics: Sec. E, Table S3. The 25-initialization study is main-paper Table 3.
- Exact input/evaluation protocol: Sec. F, Table S4.
- Segmentation and rendering validation: Sec. G, Tables S5–S6 and Figure S6.
- Implementation details and compute budget: Sec. H.
- Fixed constants, bounds, and optimizer settings: Sec. I, Table S7.
- Comparison with prior inverse-physics and avatar methods: Sec. J, Table S8.

Conventions. Notation follows the main paper. The optimized simulator parameters are density ρ, in-plane stretch stiffness $\kappa_s \equiv E$, shell thickness h, and contact friction μ (shared by cloth–body and cloth–cloth contact; Sec. B). Chamfer distances use the same $\times 10^3$ convention as main-paper Table 1. Bold marks the best result within a comparison block and underlining the second best; mean rows are excluded from that ranking. Unless otherwise noted, supplement experiments use 4D-DRESS, optimize on 50 training frames, and evaluate on 100 held-out test frames. The capture-baseline comparison remains in main-paper Table 1.

Factorization and identification boundary. The four queried renders define an exact 2×2 mixed difference of the queried score table: constant and one-factor main effects cancel algebraically, but the remaining interaction is not assumed to be a pure semantic cloth variable inside the frozen network. Likewise, behavioral identification means that the optimized simulator predicts held-out cloth behavior; parameter identification would additionally require the latent material vector to be uniquely and stably constrained. The fourth-corner and sign-agreement studies diagnose the objective. The 25-initialization study in Sec. E is an auxiliary geometric-optimizer landscape probe, not 25-seed FSD evidence, and boundary saturation counts against a strong latent-identification claim.

A STAGE-1 SCENE CONSTRUCTION

Stage 1 converts the input image(s) into the classified particle volume used to initialize the simulator: a posed body proxy $\mathcal{B}$, a garment mesh $\mathcal{G}_0$, and per-particle body/garment labels. Main-paper Algorithm 1 summarizes the pipeline. The only learned components are SMPL-X body fitting and a 3D Gaussian appearance model.

Figure S1 shows the same pipeline as a diagram: the input views, the reconstructed avatar, and the unprojected per-vertex labels that become the classified particle volume.

The two segmentation branches. As described in the Stage-1 scene-construction subsection of the main paper, the default pipeline uses a finetuned SegFormer (Xie et al., 2021), while a promptable segmenter (SAM 3; Carion et al., 2025) serves as an alternative. Following the main paper, we refer to the model as "SAM 3" throughout; the checkpoint used is the released SAM 3.1 model from that family. Since the finetuned segmenter is the only component that could introduce garment-specific supervision into an otherwise test-time pipeline, Sec. G evaluates the promptable branch independently. SAM 3 is applied directly to the raw capture images using generic garment prompts (upper $\in$ {sweater, shirt, top, blouse, t-shirt, jacket, vest}; lower $\in$ {pants, jeans, trousers, leggings,

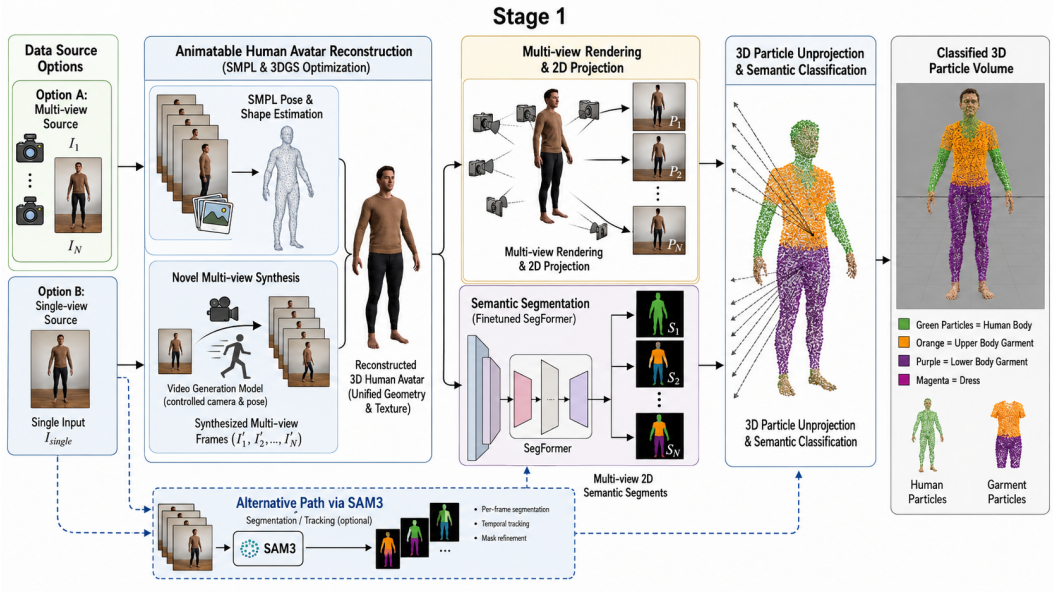

Figure S1: **Stage 1: From images to a classified 3D particle volume.** *Left:* Input consists of captured multi-view images or multi-view frames synthesized from a single image. *Center:* We reconstruct an animatable avatar via SMPL and 3D Gaussian Splatting. The avatar is rendered from multiple views and segmented using a finetuned SegFormer (or optionally SAM 3). *Right:* 2D semantic labels are unprojected and fused into a 3D particle volume, separating the human body from garments for downstream MPM simulation.

shorts, skirt}), without per-outfit prompts or fine-tuning (Figure S2). Tables S5–S6 show that this training-free pipeline exactly reproduces the manual ground-truth split on the 4D-DRESS garments and leaves the per-garment metrics unchanged, so the 4D-DRESS results do not depend on the finetuned segmenter. This check covers the 4D-DRESS garments only.

Back-projection and topology snap. Per-view masks are back-projected onto the template mesh from four camera views and fused by majority vote to produce three-class per-vertex labels. Because the raw votes are noisy and the scan consists of disconnected garment shells, each connected shell is assigned the dominant label of its vertices. This removes isolated labeling errors while preserving the original shell topology, including garment overlap (Figure S3). The recovered segmentation is 100% vertex-identical to the manual ground-truth split for all six garments across the four sequences (Table S5).

Garment–body gap handling. A single-layer scan leaves visible gaps wherever the garment shell does not fully cover the body, most noticeably around the waist during large motions. The body proxy $\mathcal{B}$ is a closed mesh driven by linear blend skinning (LBS) and fills this interior region behind the garment particles, producing a gap-free render (Figure S4). The gap is therefore a rendering and coverage issue rather than a segmentation error, and the garment labels remain unchanged.

B Codimensional MPM Cloth Physics

This section restates the differentiable codimensional MLS-MPM/APIC solver used in Stage 3 of the main paper, following the codimensional formulation of MPMAvatar (Lee et al., 2025) and the anisotropic cloth elastoplasticity model of Jiang et al. (2017). The optimized simulator parameters, $\mathbf{\Theta} = (\rho, \kappa_s, h, \mu)$, are optimized within this physical model. All fixed constants used by the simulator are listed in Table S7.

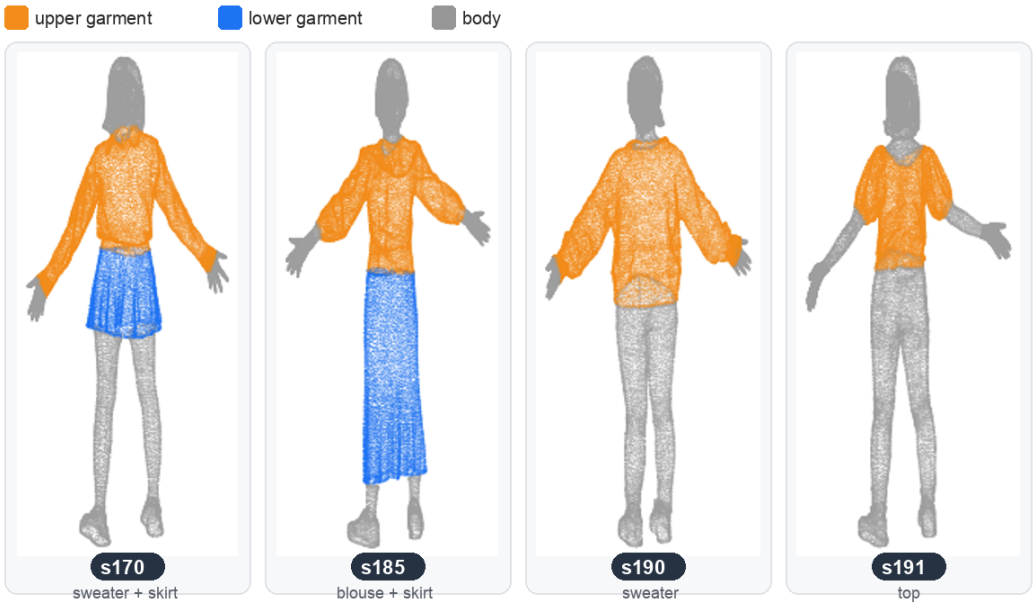

Figure S2: **SAM 3 zero-shot text-prompted garment segmentation** on all four 4D-DRESS sequences, projected onto the classified vertex cloud: the upper garment (orange) and, where present, the lower garment (blue) are separated from the body (grey) using generic text prompts alone, with no fine-tuning. s170/s185 carry two garments; s190/s191 are single-garment.



Figure S3: **SAM 3 followed by topology-snap on the s190 mesh** (two views): red is the recovered cloth shell, grey the body. 18,041 cloth vertices, byte-for-byte identical to the manual GT split.

B.1 Material-Point Representation

The garment is represented as a Lagrangian triangle mesh $\mathcal{G}_0$ carrying N_p material points $\mathcal{P} = \{(\mathbf{x}_p, \mathbf{v}_p, m_p, V_p, \mathbf{F}_p, \mathbf{C}_p)\}$, where $m_p = \rho V_p$, $V_p \propto h$, and the initial state is $\mathbf{v}_p^0 = \mathbf{0}$, $\mathbf{F}_p^0 = \mathbf{I}$, and $\mathbf{C}_p^0 = \mathbf{0}$. Because cloth is a codimensional surface embedded in $\mathbb{R}^3$, the deformation gradient is not directly defined along the normal direction and must be constructed. We form it from the two in-plane edge tangents $\mathbf{f}_1, \mathbf{f}_2$ of the deformed triangle together with the unit normal director $\mathbf{f}_3$, relative to the rest-configuration basis $\mathbf{D}_p$ built from the same two tangents $\mathbf{f}_1^0, \mathbf{f}_2^0$ and unit normal $\mathbf{n}^0$ of the undeformed triangle:

$$\mathbf{F}_p = [\,\mathbf{f}_1\ \mathbf{f}_2\ \mathbf{f}_3\,]\,\mathbf{D}_p^{-1}, \qquad \mathbf{D}_p = [\,\mathbf{f}_1^0\ \mathbf{f}_2^0\ \mathbf{n}^0\,], \tag{S1}$$

so that $\mathbf{F}_p = \mathbf{I}$ in the rest state, matching $\mathbf{F}_p^0 = \mathbf{I}$.

B.2 Anisotropic Constitutive Model

Cloth exhibits different responses to in-plane stretch, in-plane shear, and normal (compression) deformation. We therefore work in the rotation-invariant QR frame $\mathbf{F}_p = \mathbf{Q}_p\mathbf{R}_p$, where $\mathbf{Q}_p \in \mathrm{SO}(3)$ and $\mathbf{R}_p$ is upper triangular. Writing the in-plane block $\mathbf{R}_{2\times 2} = \left(\begin{smallmatrix} r_{11} & r_{12} \\ 0 & r_{22} \end{smallmatrix}\right)$, the shear terms (r_{13}, r_{23}),



Figure S4: **Gap handling on s190 under large motion.** (a) the single-layer garment alone leaves a see-through band at the waist, so torso and legs are visibly disconnected; (b) the LBS-driven body proxy fills the interior; (c) the final classified render is gap-free. The garment segmentation is identical in all three.

and the normal component r_{33}, the strain energy density is

$$\Psi(\mathbf{F}) = \Psi_{\text{plane}}(\mathbf{R}_{2\times 2}; \kappa_s, \nu) + \tfrac{\gamma}{2}\left(r_{13}^2 + r_{23}^2\right) + \Psi_n(r_{33}; \kappa_n)\,, \tag{S2}$$

where Ψ_{plane} is a fixed-corotated membrane energy with in-plane stiffness $\kappa_s \equiv E$ and Poisson ratio ν, γ penalizes shear relative to the normal director, and Ψ_n penalizes normal compression with stiffness κ_n (inactive when $r_{33} \geq 1$). The resulting first Piola–Kirchhoff stress is

$$\mathbf{P}_p = \frac{\partial \Psi}{\partial \mathbf{F}_p}. \tag{S3}$$

Cloth–cloth friction is incorporated through a plastic return mapping on (r_{13}, r_{23}, r_{33}). After the trial elastic update, the shear pair is projected onto the Coulomb cone,

$$\gamma\sqrt{r_{13}^2 + r_{23}^2} \;\leq\; c_F\,\mu\,\left|\Psi_n'(r_{33})\right|, \tag{S4}$$

making friction part of the constitutive update rather than a separate collision step. Here μ is the optimized friction coefficient also used for body contact (Sec. B.4), and $c_F = \tan 40^\circ \approx 0.84$. The optimized simulator parameters are density ρ, in-plane stretch stiffness $\kappa_s \equiv E$, thickness h, and friction μ, while κ_n, γ, ν, and c_F remain fixed. Bending is controlled through h, which determines the particle volume V_p and scales the membrane and normal contributions differently for a thin shell.

B.3 Differentiable MLS-MPM/APIC Transfers

Each particle carries an affine velocity field $\mathbf{v}(\mathbf{x}) = \mathbf{v}_p + \mathbf{C}_p(\mathbf{x} - \mathbf{x}_p)$ (APIC). Using quadratic B-spline weights w_{ip} and grid spacing Δx, each simulation step consists of particle-to-grid (P2G),

grid update, and grid-to-particle (G2P) transfers:

$$m_i = \textstyle\sum_p w_{ip} m_p, \tag{S5}$$

$$(m\mathbf{v})_i = \textstyle\sum_p w_{ip} m_p \left[\mathbf{v}_p + \mathbf{C}_p(\mathbf{x}_i - \mathbf{x}_p)\right], \tag{S6}$$

$$\mathbf{f}_i = -\textstyle\sum_p V_p \mathbf{P}_p \mathbf{F}_p^\top \nabla w_{ip}, \tag{S7}$$

$$\hat{\mathbf{v}}_i = \min(c_d, 1)\left[(m\mathbf{v})_i/m_i + \Delta t\, \mathbf{g} + \Delta t\, \mathbf{f}_i/m_i\right], \tag{S8}$$

$$\mathbf{v}_p \leftarrow \textstyle\sum_i w_{ip}\mathbf{v}_i, \qquad \mathbf{x}_p \leftarrow \mathbf{x}_p + \Delta t\, \mathbf{v}_p, \tag{S9}$$

$$\mathbf{C}_p \leftarrow \frac{4}{\Delta x^2} \textstyle\sum_i w_{ip}\mathbf{v}_i(\mathbf{x}_i - \mathbf{x}_p)^\top. \tag{S10}$$

The rollout $\{\mathcal{G}_t\}_{t=0}^T = \mathcal{S}_{\text{MPM}}(\mathcal{B}, \mathcal{G}_0; \boldsymbol{\Theta})$ is obtained by iterating Eqs. (S5)–(S10) for T steps. The entire simulation is differentiated end to end using reverse-mode automatic differentiation implemented in NVIDIA Warp.

B.4 Mesh-Collider Contact with the Body

The collider is the posed body mesh $\mathcal{B} = M(\beta, \theta_t, \mathbf{t}_t)$, which is updated every frame using the driving motion $\{\theta_t, \mathbf{t}_t\}$ and is never converted into a level set. Surface velocities $\mathbf{v}^B$ and outward normals $\mathbf{n}^B$ are transferred to nearby grid nodes using the same B-spline weights. Letting $\hat{\mathbf{v}}_i$ denote the post-force grid velocity from Eq. (S8), the relative velocity is projected onto the collider tangent space:

$$\mathbf{v}_i^{\text{rel}} = \hat{\mathbf{v}}_i - \mathbf{v}_i^B, \qquad v_n = \mathbf{v}_i^{\text{rel}} \cdot \mathbf{n}_i^B, \tag{S11}$$

$$\mathbf{v}_i^{\text{rel}} \leftarrow \mathbf{v}_i^{\text{rel}} - \min(v_n, 0)\, \mathbf{n}_i^B, \tag{S12}$$

removing only inward normal motion while preserving tangential sliding. As in the main paper, the tangential remainder $\mathbf{v}_{i,\parallel}^{\text{rel}} = \mathbf{v}_i^{\text{rel}} - (\mathbf{v}_i^{\text{rel}} \cdot \mathbf{n}_i^B)\, \mathbf{n}_i^B$ is then scaled by the Coulomb factor of the same friction coefficient μ as in Eq. (S4), with a small constant δ, and the collider velocity is restored:

$$\mathbf{v}_{i,\parallel}^{\text{rel}} \leftarrow \max\left(0,\, 1 - \frac{\mu\, |v_n|}{\|\mathbf{v}_{i,\parallel}^{\text{rel}}\| + \delta}\right) \mathbf{v}_{i,\parallel}^{\text{rel}}, \tag{S13}$$

$$\mathbf{v}_i = (\mathbf{v}_i^{\text{rel}} \cdot \mathbf{n}_i^B)\, \mathbf{n}_i^B + \mathbf{v}_{i,\parallel}^{\text{rel}} + \mathbf{v}_i^B. \tag{S14}$$

B.5 Parameter Optimization

Given the factored objective (Sec. C), the differentiable parameters $\{\rho, \kappa_s, h\}$ are optimized with Adam using pathwise gradients through the rollout, followed by projection onto the box constraints listed in Table S7. Friction μ is optimized separately because both places it enters, the Coulomb return mapping of Eq. (S4) and the collider's Coulomb clamp of Eq. (S13), are projections that are non-smooth and often inactive. As a result, the pathwise gradient $\partial\mathcal{L}/\partial\mu$ is zero on the sticking branch and discontinuous at the cone boundary, making automatic differentiation unreliable. We therefore estimate $\partial\mathcal{L}/\partial\mu$ using Simultaneous Perturbation Stochastic Approximation (SPSA; Spall, 1992).

SPSA estimator. For parameters $\theta \in \mathbb{R}^p$ and loss $\mathcal{L}$, draw a random perturbation $\Delta = (\Delta_1, \ldots, \Delta_p)$ with Δ_j i.i.d. Rademacher ($\Pr[\Delta_j = \pm 1] = \frac{1}{2}$, so $\mathbb{E}[\Delta_j] = 0$ and $\Delta_j^2 = 1$). With step $c > 0$, the two-sided SPSA gradient estimate is

$$\hat{g}_i(\theta) = \frac{\mathcal{L}(\theta + c\Delta) - \mathcal{L}(\theta - c\Delta)}{2\, c\, \Delta_i}, \qquad i = 1, \ldots, p. \tag{S15}$$

Crucially, Eq. (S15) uses only two loss evaluations to estimate the entire gradient, independent of the dimension p, whereas coordinate-wise finite differences need $2p$.

Almost-unbiasedness. Taylor-expanding both evaluations about θ, with H the Hessian of $\mathcal{L}$ at θ,

$$\mathcal{L}(\theta \pm c\Delta) = \mathcal{L}(\theta) \pm c\, \Delta^\top \nabla\mathcal{L} + \frac{c^2}{2} \Delta^\top H \Delta + O(c^3), \tag{S16}$$

so the even-order terms cancel in the numerator of Eq. (S15) and

$$\begin{aligned}\hat{g}_i(\theta) &= \frac{1}{\Delta_i}\sum_{j=1}^{p}\Delta_j\,\partial_j\mathcal{L} + O(c^2)\\ &= \partial_i\mathcal{L} + \sum_{j\neq i}\frac{\Delta_j}{\Delta_i}\partial_j\mathcal{L} + O(c^2).\end{aligned} \tag{S17}$$

Taking expectations and using independence with $\mathbb{E}[\Delta_j] = 0$ $(j \neq i)$ and $\Delta_i^2 = 1$,

$$\mathbb{E}[\hat{g}_i(\theta) \mid \theta] = \partial_i\mathcal{L} + O(c^2). \tag{S18}$$

That is, SPSA provides an (approximately) unbiased estimate of each partial derivative, with the bias vanishing as $c \to 0$, independently of the number of optimization variables and without requiring differentiation through the non-smooth Coulomb projection. In our implementation, SPSA is applied only to μ, although the two-evaluation cost and Eq. (S18) hold for any parameter dimension p. The perturbation scale c optionally follows a cosine annealing schedule.

Disambiguating the two constants. We distinguish two similarly named constants. The parameter c_F is the fixed cone-scaling constant used in the frictional return mapping (Eq. (S4)), here $\tan 40^\circ \approx 0.84$ (the released MPMAvatar solver's default), whereas c_d scales the grid velocity by $\min(c_d, 1)$ in Eq. (S8), so it damps only when $c_d < 1$; the value $c_d = 1.1$ used in all experiments leaves grid damping inactive. Consequently, all dissipation in the reported rollouts arises from frictional contact and the return mapping. The two constants operate at different stages of the solver and remain fixed throughout the experiments (Table S7).

C Structural Counterfactual Factorization and Prior Sensitivity

Estimator and exact four-corner interaction. For a rendered clip $\mathbf{V}$, the prior is queried at noise level τ on the rectified-flow interpolant $z_\tau(\mathbf{V}) = (1-\tau)\mathbf{V} + \tau\epsilon$, whose target velocity is $\epsilon - \mathbf{V}$. Define the velocity residual $\hat{u}_\phi(\mathbf{V}) = u_\phi(z_\tau(\mathbf{V}) \mid y, \tau) - (\epsilon - \mathbf{V})$. With the four matched renders,

$$\hat{u}_\phi^{\text{int}} = \hat{u}_\phi(\mathbf{V}_{\text{sim}}) - \hat{u}_\phi(\mathbf{V}_{\text{app}}) - \hat{u}_\phi(\mathbf{V}_{\text{body}}) + \hat{u}_\phi(\mathbf{V}_{\text{body,app}}), \tag{S19}$$

$$\nabla_{\boldsymbol{\Theta}}\mathcal{L}_{\text{FSD}} = \mathbb{E}_{\tau,\epsilon}\big[w(\tau)\,\hat{u}_\phi^{\text{int}}\,\partial\mathbf{V}_{\text{sim}}/\partial\boldsymbol{\Theta}\big], \tag{S20}$$

where every branch uses the same (τ, ϵ) draw.

Algorithm S1 states the corresponding optimization loop end to end, including the annealed timestep schedule and the single shared noise draw that all four branches are evaluated under.

To state precisely what is factorized, let q_{am}, with $a, m \in \{0, 1\}$ indexing garment appearance and articulated motion as in the main paper, denote any scalar coordinate of the prior response at the four corners of the renderer-defined 2×2 lattice. Any table of four values has the exact expansion

$$q_{am} = q_{00} + a\,\alpha_a + m\,\alpha_m + am\,\Gamma, \qquad \Gamma = q_{11} - q_{10} - q_{01} + q_{00}. \tag{S21}$$

Thus the $(+,-,-,+)$ mixed difference cancels the constant q_{00} and both one-axis main effects α_a, α_m exactly, for an arbitrary nonlinear video model. No linearity assumption is required. For the system-identification framing, FSD is an objective-design device. The exact algebraic statement concerns the four score evaluations: the $(+,-,-,+)$ contrast returns their two-axis interaction. Whether that interaction is informative about the simulator parameters is an empirical question addressed by held-out behavior, gradient-sign diagnostics, and the fourth-corner ablation.

What the interaction does not guarantee. Equation (S21) is an exact algebraic statement about four evaluations. It does not prove that the remaining interaction is a pure semantic "cloth score," nor that all nuisance information has disappeared. Any learned interaction that is jointly elicited by the two renderer axes may remain. DreamCloth therefore validates the interaction empirically through held-out rollout metrics, sampled gradient-sign agreement, and the fourth-corner ablation rather than assuming semantic disentanglement.

Algorithm S1 Stages 3–4: DreamCloth inverse optimization (rectified-flow FSD). The timestep τ is drawn from an annealed schedule $p_k(\tau)$ that is biased toward mid-noise levels early in optimization and shifts toward clean (low-noise) levels as k increases; the same (τ, ϵ) noises all four branches (simulated, appearance, body-motion, and the static body-appearance correction).

input: image I_0; horizon T; steps K; Δt; renderer $\mathcal{R}$; solver $\mathcal{S}_{\text{MPM}}$; frozen rectified-flow model u_ϕ; conditioning y; range $[\tau_{\min}, \tau_{\max}]$; weights $\lambda_\bullet$
$(\mathcal{B}, \mathcal{G}_0, \Pi) \leftarrow \Phi_{\text{mono}}(I_0)$; $\mathcal{P} \leftarrow \textsc{Discretize}(\mathcal{G}_0)$; $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta}_0$
for $k = 1$ to K **do**
  $\{\mathcal{G}_t\} \leftarrow \mathcal{S}_{\text{MPM}}(\mathcal{B}, \mathcal{P}; \boldsymbol{\Theta}, \Delta t)$
  $\mathbf{V}_{\text{sim}} \leftarrow \mathcal{R}(\{\mathcal{B}_t\}, \{\mathcal{G}_t\})$ ▷ $\mathcal{B}_t$: body posed at frame t
  $\mathbf{V}_{\text{app}} \leftarrow \mathcal{R}(\mathcal{B}_0, \mathcal{G}_0) \times T$ ▷ static clothed avatar
  $\mathbf{V}_{\text{body}} \leftarrow \mathcal{R}(\{\mathcal{B}_t\}, \emptyset)$ ▷ moving garment-free body
  $\mathbf{V}_{\text{body,app}} \leftarrow \mathcal{R}(\mathcal{B}_0, \emptyset) \times T$ ▷ static body-appearance correction
  $s \leftarrow k/K$; $\xi \sim \mathcal{U}(0, 1)$
  $\tau \leftarrow (1 - s)\, F^{-1}_{\text{mid}}(\xi) + s\, F^{-1}_{\text{clean}}(\xi)$ ▷ annealed mid → clean
  $\tau \leftarrow \tau_{\min} + (\tau_{\max} - \tau_{\min})\, \tau$
  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ▷ single shared draw
  for $\mathbf{V} \in \{\mathbf{V}_{\text{sim}}, \mathbf{V}_{\text{app}}, \mathbf{V}_{\text{body}}, \mathbf{V}_{\text{body,app}}\}$ **do**
    $z_\tau(\mathbf{V}) \leftarrow (1 - \tau)\mathbf{V} + \tau\epsilon$
    $\hat{u}_\phi(\mathbf{V}) \leftarrow u_\phi(z_\tau(\mathbf{V}) \mid y, \tau) - (\epsilon - \mathbf{V})$
  end for
  $g_{\text{FSD}} \leftarrow w(\tau)\big[\hat{u}_\phi(\mathbf{V}_{\text{sim}}) - \hat{u}_\phi(\mathbf{V}_{\text{app}}) - \hat{u}_\phi(\mathbf{V}_{\text{body}}) + \hat{u}_\phi(\mathbf{V}_{\text{body,app}})\big]\, \partial \mathbf{V}_{\text{sim}} / \partial \boldsymbol{\Theta}$
  $(\rho, \kappa_s, h) \leftarrow \textsc{Adam}\big(\lambda_{\text{FSD}}\, g_{\text{FSD}} + \nabla \mathcal{L}_{\text{reg}}\big)$ ▷ $\mathcal{L}_{\text{reg}} = \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{phys}}\mathcal{L}_{\text{phys}}$
  $\mu \leftarrow$ two-sided SPSA step on $\mathcal{L}_{\text{total}}$ ▷ Eq. (S15)
  clamp $\boldsymbol{\Theta}$ to the box constraints of Table S7
end for
return $\boldsymbol{\Theta}$

Why the fourth branch is needed. The static garment-free body is the fourth corner required by the exact mixed difference. At the residual level, the coefficients $(+, -, -, +)$ also sum to zero, so with a shared noise draw the explicit ϵ term cancels exactly. Removing $+\hat{u}_\phi(\mathbf{V}_{\text{body,app}})$ changes the estimand to a three-branch contrast and leaves an uncancelled residual-noise contribution. Main-paper Table 2 gives the matched ablation: on s190, removing the fourth corner increases held-out CD from 0.362 to 0.422 (+17%). Because both rows are single-initialization runs and reach a box constraint, the relative degradation is the supported conclusion; the ablation is not evidence of unique material recovery.

C.1 Shared-Noise Covariance Decomposition

Sharing (τ, ϵ) preserves each branch's marginal distribution and therefore preserves the expectation of the linear contrast. It also introduces covariance between branches. For one scalar coordinate, write the four residuals in the corner notation of Eq. (S21): $q_{11} = \hat{u}_\phi(\mathbf{V}_{\text{sim}})$, $q_{10} = \hat{u}_\phi(\mathbf{V}_{\text{app}})$, $q_{01} = \hat{u}_\phi(\mathbf{V}_{\text{body}})$, and $q_{00} = \hat{u}_\phi(\mathbf{V}_{\text{body,app}})$. Then

$$\begin{aligned}\text{Var}_{\text{sh}} &:= \text{Var}[q_{11} - q_{10} - q_{01} + q_{00}] \\ &= \text{Var}[q_{11}] + \text{Var}[q_{10}] + \text{Var}[q_{01}] + \text{Var}[q_{00}] \\ &\quad + 2\big(\text{Cov}(q_{10}, q_{01}) + \text{Cov}(q_{11}, q_{00})\big) \\ &\quad - 2\big(\text{Cov}(q_{11}, q_{10}) + \text{Cov}(q_{11}, q_{01}) + \text{Cov}(q_{10}, q_{00}) + \text{Cov}(q_{01}, q_{00})\big).\end{aligned} \tag{S22}$$

With independent draws the cross-covariances vanish, giving $\text{Var}_{\text{ind}} = \text{Var}[q_{11}] + \text{Var}[q_{10}] + \text{Var}[q_{01}] + \text{Var}[q_{00}]$. Hence

$$\begin{aligned}\text{Var}_{\text{sh}} - \text{Var}_{\text{ind}} &= 2\big(\text{Cov}(q_{10}, q_{01}) + \text{Cov}(q_{11}, q_{00})\big) \\ &\quad - 2\big(\text{Cov}(q_{11}, q_{10}) + \text{Cov}(q_{11}, q_{01}) + \text{Cov}(q_{10}, q_{00}) + \text{Cov}(q_{01}, q_{00})\big).\end{aligned} \tag{S23}$$

Shared noise therefore can act as a control-variate construction when the covariance terms in Eq. (S23) have the favorable sign and magnitude; variance reduction is not guaranteed by sharing alone. The present experiments do not include a direct Monte-Carlo estimate of $\text{Var}_{\text{sh}} - \text{Var}_{\text{ind}}$,

Table S1: **Video-prior backbone study.** Matched backbone comparison on held-out 4D-DRESS s191 with the same Sobol initialization, garment split, four-branch estimator, and 500 optimization iterations. Geometry and strain measure held-out cloth error; "Sign agree" is sampled agreement between FSD and geometric gradient directions. Guidance preference is backbone-dependent: Cloth Zoom is best for Wan2.2-TI2V-5B, while four-branch FSD is substantially better for Wan2.2 I2V-A14B and PhyWorld. LTX-Video-2B attains low error only at parameter bounds, so that result is not evidence of reliable latent recovery. All rows are single-initialization comparisons; bold marks the best guidance within each backbone. The fourth-corner ablation is main-paper Table 2.

Backbone	Guidance	Geom. ↓	Strain ↓	Sign agree ↑	ρ	κ_s	h
Wan2.2 I2V-A14B							
	Cloth Zoom	0.0678	0.1988	0.35	1.009	50	1.200
	Four-branch FSD	**0.0084**	**0.0494**	**0.72**	0.536	321	1.008
	Time Schedule	0.0175	0.0771	0.12	–	–	–
LTX-Video-2B							
	Cloth Zoom	**0.0060**	**0.0466**	0.62	0.100	2000	1.003
	Four-branch FSD	0.0061	**0.0466**	**0.68**	0.100	2000	0.999
	Time Schedule	0.0061	**0.0466**	0.60	0.100	2000	0.998
PhyWorld							
	Cloth Zoom	0.0681	0.1996	0.35	1.016	50	1.200
	Four-branch FSD	**0.0063**	**0.0475**	**0.77**	0.561	572	0.996
	Time Schedule	0.0175	0.0771	0.33	–	–	–
Wan2.2-TI2V-5B							
	Cloth Zoom	**0.0097**	**0.0718**	0.60	0.809	50	0.828
	Four-branch FSD	0.0163	0.0829	**0.70**	0.433	63	0.806
	Time Schedule	0.0237	0.0873	0.65	0.477	1672	0.828

so Eqs. (S22)–(S23) motivate the design but should not be read as empirical proof of reduced gradient variance. The exact claim that does hold under shared noise is the cancellation of the explicit residual ϵ term in the four-corner contrast.

Prior sensitivity as an identification diagnostic. The sign-agreement column compares sampled FSD and geometric gradient directions and is therefore a proxy for whether the learned observation objective pushes the simulator in a behaviorally useful direction. Table S1 and main-paper Figure 3 show substantial backbone dependence: four-branch FSD performs well for Wan2.2-A14B and PhyWorld, whereas Cloth Zoom is stronger for Wan2.2-TI2V-5B. Low LTX-Video error coincides with box-bound solutions. Consequently, these data support behavioral usefulness for selected prior/guidance pairings but do not establish unique latent recovery or a universal ordering of priors.

D Diffusion Schedules, View Selection, and Zoom

Table S2 summarizes the diffusion schedule, camera-view, and garment-zoom settings referenced in the main paper. Unless otherwise noted, all experiments are run on all four 4D-DRESS sequences with text conditioning, 500 finite-difference iterations, a fixed rollout horizon, a fixed evaluation resolution, and the same SAM 3 garment-mask protocol. Panels (a) and (b) represent independent comparison blocks. The run identifiers S2-C1/S2-C2 and S3-M1/S3-M2 refer to different configurations and should be interpreted within their respective panels rather than compared directly.

Cloth Zoom protocol. Cloth Zoom magnifies a fixed garment region in each rendered frame before evaluation by the frozen video prior. The crop is defined as the union of the projected ground-truth garment vertices across the clip, expanded by 15%, cached for each camera, and bilinearly resized to the original $H \times W$ resolution. Both cropping and resizing remain differentiable, allowing gradients to propagate to the material parameters. All Cloth Zoom experiments use zoom-only scoring together with the same diffusion timestep and noise as the corresponding comparison branch. Because the crop is fixed and shared across all backbones, it focuses the prior on garment appearance and motion without tracking the model's own prediction or requiring an additional rendering pass. Figure S5 illustrates the crop evaluated by the video prior. Cloth Zoom can help when garment cues are diluted by background, appearance, or body motion, but Table S2(c) shows that the effect

Table S2: **Diffusion schedule, camera-view, and garment-zoom ablations (all four 4D-DRESS sequences).** (a) Plain-SDS schedule and view-coverage ablation, comparing temporal coverage, frame selection, and diffusion-timestep policy in single-run experiments. (b) Matched-view comparison of optimization objectives, with best and second-best results highlighted within each block; DDS denotes Delta Denoising Score (Hertz et al., 2023). (c) Garment-zoom conditioning. Each row replaces Text + Image conditioning with a fixed garment crop, where $\Delta_{Z-1} = \mathrm{CD}_{\mathrm{zoom}} - \mathrm{CD}_{\mathrm{Text+Img}}$ (negative values indicate improvement). Best and second-best results are reported within each sequence block (Mean excluded). All Chamfer distances follow the $\times 10^3$ convention of main-paper Table 1. Every experiment uses text conditioning, 500 finite-difference iterations, the same rollout horizon, rendering resolution, and the SAM 3 garment-mask protocol. Run identifiers S2-C1/S2-C2 (panel a) and S3-M1/S3-M2 (panel b) denote separate experimental configurations and are not pooled.

(a) Timestep-schedule and coverage ablation

Configuration	Views	Slices	Timestep policy	CD ($\times 10^3$) ↓	Masked CD ($\times 10^3$) ↓	F-score ↑	PSNR ↑	SSIM ↑	LPIPS ↓
Transported 1-View Base	1	1	Mid-noise	**0.395**	0.762	92.5	30.78	0.978	0.0161
4-View Batch	4	1	Mid-noise	0.429	0.830	91.8	29.40	0.978	0.0173
S2-C1: 1-View, 4-Slice Coverage	1	4	Mid-noise	0.414	0.800	92.1	29.61	0.978	0.0169
S2-C2: 2-View, 4-Slice Coverage	2	4	Mid-noise	0.420	0.812	91.9	29.35	0.978	0.0172
1-View Uniform Sampling	1	1	Uniform	0.528	1.026	90.1	26.63	0.977	0.0190
1-View Clean Bias	1	1	Low-noise	0.412	0.796	92.2	30.19	0.978	0.0167
1-View Mid-to-Clean Schedule	1	1	Mid-to-clean	0.402	0.757	92.6	31.25	**0.979**	0.0158
1-View Foreground-Masked	1	1	Mid-noise	0.407	**0.717**	**92.7**	**31.28**	**0.979**	**0.0157**

(b) Matched-view comparison across objectives

Objective	View configuration	Views	Slices	Clips/ step	CD ($\times 10^3$) ↓	Masked CD ($\times 10^3$) ↓	F-score ↑	PSNR ↑	SSIM ↑	LPIPS ↓	Control
Matched view-count comparisons											
Plain SDS	S3-M1: 1-View, 4-Slice	1	4	4	0.400	0.772	92.4	30.39	0.978	0.0164	Matched
	S3-M2: 2-View, 4-Slice	2	4	8	0.464	0.899	91.1	28.22	0.978	0.0178	Matched
ISM	1-View	1	1	1	0.402	0.757	92.7	31.33	**0.979**	0.0158	Matched
	4-View	4	1	4	**0.398**	**0.750**	**92.8**	**31.53**	**0.979**	**0.0155**	Matched
Additional view-related configurations											
ISM	2-View, 4-Slice	2	4	8	0.398	0.750	**92.9**	**31.45**	**0.979**	**0.0157**	Slices differ
Four-branch FSD DDS	Transported 1-View	1	1	1	**0.391**	**0.706**	92.6	30.90	0.978	0.0158	Transported
	4-View Batch	4	1	4	0.398	0.719	92.7	31.24	**0.979**	0.0158	Transport differs
	2-View, 4-Slice	2	4	8	0.461	0.836	91.2	28.04	0.978	0.0177	Slices differ

(c) Garment-zoom conditioning (4D-DRESS)

Objective	Conditioning comparison				Garment-zoom test metrics				
	Text+Img CD ($\times 10^3$) ↓	T+I Masked CD ($\times 10^3$) ↓	Zoom CD ($\times 10^3$) ↓	Zoom Masked CD ($\times 10^3$) ↓	Δ_{Z-1}	F-score ↑	PSNR ↑	SSIM ↑	LPIPS ↓
s170, lower garment									
Appearance Ref.	0.380	0.686	0.343	0.601	−0.037	93.2	**30.93**	**0.974**	0.0204
SMPL Motion Ref.	0.341	0.614	0.350	0.614	+0.009	93.1	30.35	**0.974**	0.0209
Paired Deltas	0.353	0.636	**0.330**	**0.578**	−0.023	**93.5**	30.90	**0.974**	**0.0203**
Four-branch FSD	**0.340**	**0.612**	0.337	0.591	−0.003	93.3	30.65	**0.974**	0.0206
ISM Motion Ref.	0.372	0.671	0.377	0.663	+0.005	92.7	29.41	0.973	0.0218
ISM Four-branch FSD	0.385	0.695	0.407	0.717	+0.022	92.3	29.52	0.973	0.0218
Mean	0.362	0.652	0.357	0.627	−0.005	93.0	30.29	0.974	0.0210
s190, upper garment									
Appearance Ref.	0.399	0.721	**0.381**	**0.670**	−0.018	**92.9**	**31.55**	**0.979**	**0.0151**
SMPL Motion Ref.	0.509	0.925	0.407	0.717	−0.102	92.7	31.42	**0.979**	0.0157
Paired Deltas	0.445	0.806	0.389	0.684	−0.056	92.7	31.43	**0.979**	0.0154
Four-branch FSD	0.406	0.734	0.458	0.808	+0.052	91.2	28.03	0.978	0.0176
ISM Motion Ref.	**0.390**	**0.704**	0.401	0.706	+0.011	92.4	30.72	0.978	0.0162
ISM Four-branch FSD	0.392	0.708	0.419	0.738	+0.027	92.5	31.19	**0.979**	0.0162
Mean	0.424	0.766	0.409	0.721	−0.015	92.4	30.72	0.979	0.0160

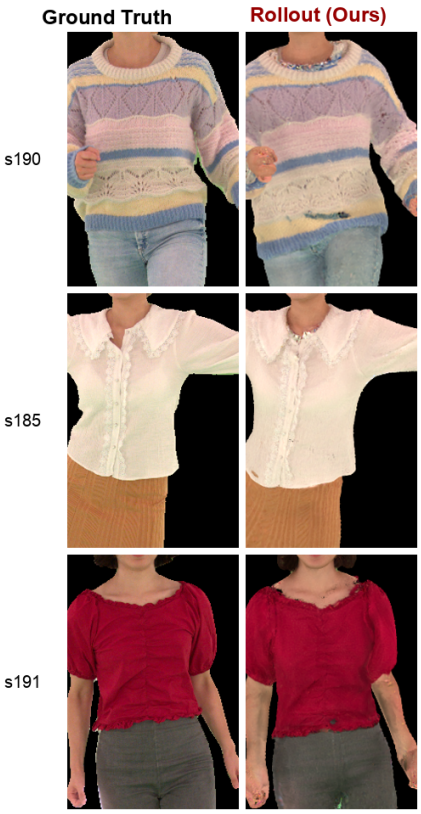

Figure S5: **Zoomed comparison of the garment region** for the ground truth and our optimized rollouts (s190, s185, and s191). At this scale, the optimized rollout preserves garment appearance while reproducing sleeve drape and hem motion. Cloth Zoom supervises exactly this crop: each rendered frame is cropped to the fixed ground-truth garment bounding box (15% padding) and bilinearly resized to $H \times W$ before scoring, encouraging the frozen video prior to focus on garment detail rather than the background.

depends strongly on the backbone and objective. We therefore treat it as a localization intervention rather than a universally superior guidance strategy.

E Behavioral and Latent Identification Diagnostics

The diagnostics in the main paper each support a narrower claim than full identification, and Table S3 sets out what each one does and does not show. The 25-initialization study (main-paper Table 3) separates optimization concentration from parameter identification. An interior, repeatable basin is compatible with identifiability but does not prove it; a solution that concentrates at a bound is especially ambiguous, because the search box itself can produce a small CV. The held-out metrics show that the resulting simulator is behaviorally useful. Establishing latent identification would require repeated full-FSD runs and, ideally, synthetic or calibrated real data with known material parameters. That experiment is not in this paper, so the strongest supported reading is behavioral rather than metrological.

Table S3: **What each diagnostic shows.** Each piece of evidence supports the claim in the middle column and is not used as evidence for the stronger claim on the right.

Evidence	Supports	Does not establish
Held-out CD / rendering	Behavioral usefulness on unseen motion	Unique material constants
Gradient sign agreement	Local directional information in one matched study	Global identifiability or seed robustness
Boundary saturation	Weak information about magnitude beyond search box	Physically correct boundary value
25-init geometric optimizer	Auxiliary landscape concentration / broad basins	25-seed FSD convergence

F EXPERIMENTAL PROTOCOL

Table S4 summarizes the inputs and evaluation procedure used for each component to eliminate ambiguity in the experimental setup. This protocol produces the benchmark results reported in Table 1 of the main paper.

Table S4: **Per-component experimental protocol.** All supplementary experiments follow this protocol on 4D-DRESS (50 training and 100 held-out test frames).

Item	Specification
DreamCloth input (main-paper Table 1)	Stage 1 builds the scene from the multi-view training frames, as the baselines do, but no garment geometry is fitted to them. The pipeline also accepts a single image with synthesized views; that setting is not evaluated here.
Baseline inputs	ARAH, TAVA, GaussianAvatar, PhysAvatar, and MPMAvatar are evaluated using the same rendering and mask protocol. Baseline results are re-run where feasible and otherwise taken from the original papers.
Driving motion $\{\theta_t, \mathbf{t}_t\}$	Estimated only from the training frames; test-sequence motion is never used during optimization.
Rollout horizon T	Fixed simulation length; the video prior scores the rendered clip at the evaluation resolution.
Probe excitation	Gravity settling together with a low-amplitude body oscillation to make short rollouts informative.
Evaluation	Camera 0004. Masked Chamfer distance is computed on SAM 3 garment regions using the train/test splits from the main paper (50 training and 100 test frames for 4D-DRESS; 200 test frames for ActorsHQ).

G STAGE-1 VALIDATION AND QUALITATIVE RESULTS

Split recovery. Table S5 lists the per-garment vertex counts for the training-free promptable branch of Sec. A: it reproduces the manual GT split exactly for all six garments.

Downstream metrics. Because the split is GT-identical, the full reconstruction matches the GT-split pipeline to three decimals (Table S6).

Rollout over time. Figure S6 compares the ground-truth sequence with the DreamCloth rollout across five frames for each subject. The single-garment and upper-garment reconstructions closely follow the ground-truth motion and drape throughout the sequence. The single lower-garment example (s170) shows clear skirt-drape artifacts: the skirt shortens and loses its shape in later frames. This run used a single, under-converged initialization, and we have not established whether better optimization or a richer material model would remove the artifact.

Table S5: **Split recovery.** Generic-prompt SAM 3 masks are back-projected from four cameras, combined by majority vote, and topology-snapped before comparison with the manual GT split. The procedure reproduces the manual split for all six listed garments; every garment is 100% vertex-identical to GT.

Subject	Garment	SAM 3 verts	GT verts	Agree.
s190	upper	18,041	18,041	100%
s191	upper	10,603	10,603	100%
s170	upper	14,741	14,741	100%
s170	lower	4,972	4,972	100%
s185	upper	12,881	12,881	100%
s185	lower	8,459	8,459	100%

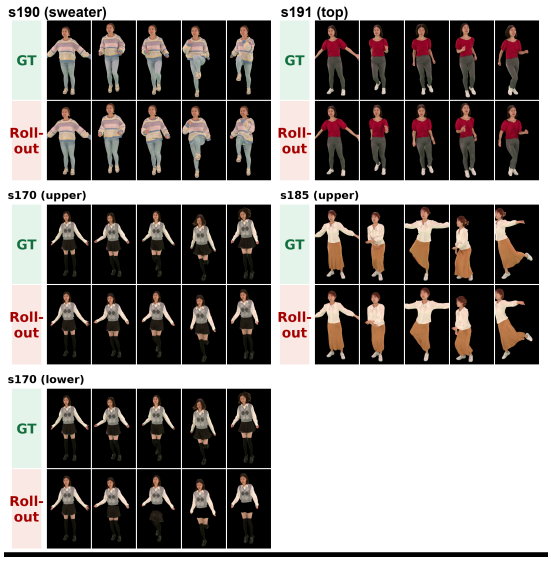

Figure S6: **Ground truth (top) and DreamCloth rollout (bottom) for five of the six garment configurations in Table S5** (four subjects), shown over five frames (left to right). Upper- and single-garment reconstructions closely follow the ground-truth sequence. The single lower-garment example (s170), from an under-converged single initialization, shows clear skirt-drape artifacts. In each single-garment simulation, the selected garment is simulated with the MPM cloth model, while the remaining garment and the body follow the tracked mesh through LBS.

Table S6: **Downstream reconstruction metrics for the training-free SAM 3 branch.** Stage 1 uses the multi-view training frames for consistency with the main-paper protocol. Driving motion is estimated only from the training frames, and test-sequence motion is never used during optimization. Evaluation is performed on the 100 held-out frames from camera 0004 using the fixed rendering protocol and SAM 3 garment masks. Chamfer distances follow the $\times 10^3$ convention of main-paper Table 1. The comparison this table reports is SAM 3 branch versus ground-truth-split branch under an identical protocol: every value agrees with its ground-truth-split counterpart to three decimal places. The rows are per-garment results for the four upper and single garments under the single-garment evaluation of Figure S6 (camera 0004, 100 held-out frames), whereas Table S5 lists six garments across these four sequences. They are therefore not the aggregate benchmark, do not average to the 4D-DRESS row of main-paper Table 1, and should not be used to reconstruct it; that aggregate follows the full main-paper protocol and is not repeated here.

Seq. / garment	CD ($\times 10^3$) ↓	F-score ↑	PSNR ↑	SSIM ↑	LPIPS ↓
s190 upper	0.376	93.19	31.78	0.979	0.0148
s191 upper	0.212	95.05	34.51	0.981	0.0158
s170 upper	0.334	93.49	30.12	0.973	0.0215
s185 upper	0.611	89.89	31.17	0.972	0.0209

H Implementation Details

DreamCloth performs per-sequence test-time material-parameter optimization using a frozen video prior; no network weights are trained. Unless otherwise stated, each sequence is optimized on a single NVIDIA A100-PCIE-40GB GPU for 500 iterations using Adam with a cosine-annealing learning-rate schedule. The base learning rate is $\eta = 10^{-1}$, with parameter-specific rates $\eta_\rho = 10^{-2}$, $\eta_{\kappa_s} = 3 \times 10^{-1}$, and $\eta_h = 10^{-2}$. With the Wan2.2-TI2V (5B) prior, optimization takes approximately 1.4 hours per sequence over the 500 iterations; the faster LTX-Video (2B) prior requires approximately 0.75 hours per sequence. The complete evaluation over eight sequences, four from ActorsHQ and four from 4D-DRESS, requires approximately 11 A100-GPU-hours for one run per sequence. A subset of the ablation experiments was conducted on an NVIDIA L40S GPU. The backbone and the schedule, view and zoom ablations are in Tables S1 and S2.

I Fixed Constants and Optimizer Settings

Table S7 summarizes all fixed constants and enforced parameter bounds used throughout the experiments, allowing the optimized values in Table S1 and main-paper Table 3 to be interpreted relative to the search limits. The values match the implementation details reported in the main paper.

Table S7: **Fixed constants, box constraints, and optimizer settings.** All fixed constants are the defaults of the released MPMAvatar solver. Only ν carries a conventional physical value; κ_n and γ are fixed reference values rather than independently identified quantities. Several optimized values in Table S1 lie on the enforced bounds (including $\kappa_s = 2000$, $\rho = 0.100$, $h = 1.2$, or $\kappa_s = 50$ depending on the row). These cases are therefore boundary-constrained behavioral solutions and should not be interpreted as validated interior material estimates.

Quantity	Value	Role
Optimized parameters (box constraints)		
ρ	$[0.1, 3.0]$	density; log-space, clamped
κ_s	$[50, 2000]$ Pa	in-plane stretch stiffness
h	$[0.8, 1.2]$	shell thickness
μ	$[0.01, 1.0]$	contact friction (cloth–body and cloth–cloth)
Fixed constitutive constants		
κ_n	500	normal-compression stiffness
γ	500	shear vs. normal director
ν	0.3	Poisson ratio
c_F	$\tan 40^\circ \approx 0.84$	Coulomb cone scaling, Eq. (S4)
c_d	1.1	grid damping factor $\min(c_d, 1)$, Eq. (S8); inactive
Optimizer		
Iterations	500	per sequence, test-time only
Optimizer	Adam + cosine anneal	no network weights trained
η	10^{-1}	base learning rate
η_ρ	10^{-2}	density
η_{κ_s}	3×10^{-1}	stiffness
η_h	10^{-2}	thickness
μ update	two-sided SPSA	Eq. (S15); non-smooth branch

J Comparison with Prior Inverse-Physics and Avatar Methods

Table S8 compares DreamCloth with DreamPhysics (Huang et al., 2025), Physics3D (Liu et al., 2024), GaussianAvatar (Hu et al., 2024), PhysAvatar (Zheng et al., 2024) and MPMAvatar (Lee et al., 2025) on four axes: whether the recovered parameters are cloth-specific rather than generic volumetric constants; whether a single captured RGB image suffices, with no extra views or frames; whether tracked per-frame garment geometry, correspondences or trajectories are required; and whether the optimization uses both a static clothed-appearance and a moving garment-free reference.

Table S8: **Comparison with inverse-physics and physics-aware avatar methods** (axes defined in the text). ○: supported by the pipeline but not evaluated here; all DreamCloth results use multi-view input.

Method	Cloth-specific params.	Single-view input	No tracked supervision	Dual reference
DreamPhysics	×	×	✓	×
Physics3D	×	×	✓	×
GaussianAvatar	×	×	✓	×
PhysAvatar	✓	×	×	×
MPMAvatar	✓	×	×	×
DreamCloth (Ours)	✓	○	✓	✓