nature communications

Article

https://doi.org/10.1038/s41467-025-61458-x

Computational models reveal that intuitive physics underlies visual processing of soft objects

Received: 18 June 2024

Accepted: 19 June 2025

Published online: 09 July 2025

Check for updates

Wenyan Bi[1], Aalap D. Shah[1], Kimberly W. Wong[1], Brian J. Scholl[1,2] & Ilker Yildirim[1,2,3,4]

Computational explorations of human cognition have been especially successful when applied to visual perception. Existing models have primarily focused on rigid objects, emphasizing shape-preserving invariance to changes in viewpoint, lighting, object size, and scene context. Yet many objects in our everyday environments, such as cloths, are soft. This poses both quantitatively greater and qualitatively different challenges for models of perception, due to soft objects' dynamic and high-dimensional internal structure, as in the changing folds and wrinkles of a cloth waving in the wind. Soft object perception is also correspondingly rich, involving distinct properties such as stiffness. Here we explore the ability of different kinds of computational models to capture visual perception of the physical properties of cloths (e.g., their degrees of stiffness) undergoing different naturalistic transformations (e.g., falling vs. waving in the wind). Across visual matching tasks, both the successes and failures of human performance are well explained by Woven: a new model that incorporates physics-based simulations to infer probabilistic representations of cloths. Woven outperforms powerful, performance-equated alternatives, including its ablations and a deep neural network, and suggests that humanlike machine vision may also require representations that transcend image statistics, and involve intuitive physics.

Vision science is arguably one of the most successful areas of cognitive science, because it not only features captivating phenomena and empirical results, but also highly predictive and mechanistically detailed computational models of how visual processing might actually operate. Models of visual recognition, for example, are both highly accurate and highly generalizable—as they can recognize objects despite variation in factors such as image size, viewing angle, lighting, background clutter, and other aspects of viewing conditions[1,2]. At the same time, however, the vast majority of this work has considered only (relatively) rigid objects (e.g., Fig. 1a), whose shapes never change as a function of such variability.

Yet, this is not always true in the real world, where we frequently encounter soft objects, such as cloths, whose shapes can and frequently do change dramatically. Consider how much your T-shirt, for example, changes its shape when worn on your back, vs. hanging on a hook, vs. tossed onto a chair (Fig. 1a). This leads to variability on the retina that is both quantitatively greater and qualitatively different than that due to a moving or rotating rigid object (Fig. 1a)— especially given the intrinsically dynamic and internally complex changes, e.g. when a cloth is waving in the wind. Despite this variability, we have no difficulty recognizing soft objects from moment to moment, and we also readily perceive seemingly higher level properties specific to soft

[1]Department of Psychology, Yale University, New Haven, CT, USA. [2]Wu Tsai Institute, Yale University, New Haven, CT, USA. [3]Department of Statistics & Data Science, Yale University, New Haven, CT, USA. [4]Foundations of Data Science, Yale University, New Haven, CT, USA. e-mail: wenyan.bi@yale.edu; ilker.yildirim@yale.edu

https://doi.org/10.1038/s41467-025-61458-x

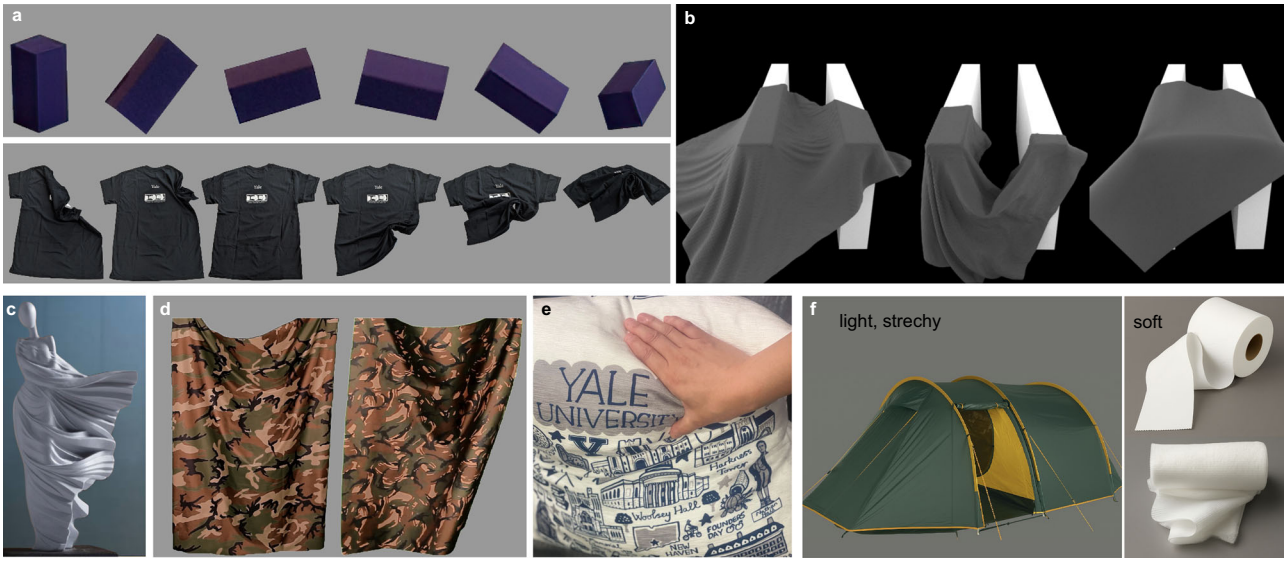

Fig. 1 | Rich percepts of soft objects. a Soft objects cause more variability on the retina than rigid objects, yet, in some ways, they lead to richer percepts, e.g., involving high-level properties such as stiffness. **b** We can tell that the left and middle images are similarly soft and that they may be the identical cloth showing two moments of the same event; moreover, we can tell that the rightmost cloth is different from the other two in terms of its stiffness. **c** Even a static cloth can convey rich dynamic information (a dress blowing in the wind), a perceptual quality the artist skillfully took advantage of. **d** The world of soft objects naturally poses distinct "explaining away" like challenges to perception. For instance, a light cloth (area weight = 1.73 oz/yd^2) in the presence of weak winds (left) can create a similar appearance as that of a heavy cloth (area weight = 11.59 oz/yd^2) in the presence of strong winds (right)[65]. **e** Similarly, when observing a pillow being pressed by an arm, it is ambiguous how much of the deformation is a result of the pillow's softness and how much is due to the force applied. **f** Depending on the task at hand and the intended functionality of the cloth, people selectively process different physical properties, under uncertainty. For example, in the case of a backpacking tent, mass and shear stiffness are the most important—it needs to be lightweight and highly tear-resistant to ensure portability and safety. However, when shopping for toilet paper or tissues, softness becomes a key consideration. Panel (**c**) is reproduced with permission from Fourline Design.

objects—such as whether a garment is silky or stiff. This is true even for an unfamiliar fabric, as in Fig. 1b: we can perceive that the left and middle images likely show different shapes of the same type of cloth, and that they are softer relative to the cloth in the right image. Perception of soft objects also poses additional challenges, with the shape of the cloth also partially reflecting object-external elements in the scene: e.g., we can't help but "fill-in" a blowing wind when looking at the sculpture in Fig. 1c. At the same time, such filling-in leads to naturally and frequently occurring uncertainties. In Fig. 1d, for example, do the two cloths have similar shapes due to comparable levels of stiffness, or due to different wind conditions? And in Fig. 1e, is the indentation due to a relatively soft pillow, or a relatively hard squeeze? The answers to such questions can be consequential when making plans and decisions about soft objects (Fig. 1f).

What computations in the mind and brain transform such unique variability at our retinas, caused by soft objects, into rich and robust percepts? There are two very different possible types of answers, corresponding to very different kinds of computational architectures and mental representations.

One possibility is that soft object perception proceeds in a bottom-up fashion, by selectively processing various properties of the visual input while becoming invariant to other (less informative) properties—in the manner of typical deep neural network (DNN) models. (In essence, this approach is the same one that has proven so successful for modeling rigid object perception.) Existing work makes this approach seem plausible, in part by identifying such informative image and motion features[3] that co-vary with (and in some cases impact) soft object perception— including perception of mass, stiffness[4], and elasticity[5,6]. And beyond these basic cues, recent work has shown the existence of powerful, non-linear feature spaces learned in DNNs that can compute estimates of physical properties of non-rigid materials from raw images[7]. Such models can recapitulate aspects of the stimulus-driven variability in humans' perceptual judgments.

A much different possibility involves transcending raw image properties, and instead building and manipulating representations that correspond to the inferred causes of those properties[8–13]. This approach arises from the observation that the immense variation in raw images may boil down to a small number of causes—a few physical properties and physical interaction rules, that together determine how soft objects move, deform and react to external forces. If an observer has a model of these physical processes (by which scenes with soft objects form and project to images), then this model can be used to "explain" the measurements in visual inputs, in terms of the underlying physical processes that caused them. This "intuitive physics" approach contrasts dramatically with standard bottom-up DNNs, which proceed without explicitly building such intermediate representations.

Which type of approach better characterizes the actual computational basis of human soft object perception? Here, to find out, we implement these perspectives and their variations in multiple performance-calibrated models. To realize the bottom-up approach, we adopt a recent behaviorally validated DNN architecture of physical property estimation[7]. We train this model on clips of cloth animations to infer cloth stiffness and mass across variations in cloth shape, scene configurations, and external forces. This model presents a compelling alternative because it performs well (i.e., as accurately as other models we consider) on objective measures of inferring such physical properties from sensory inputs.

To realize the intuitive physics approach, we embed a simulation-based representation of soft objects within probabilistic computations[8,14]. When conditioned on depth-map observations of animated cloths as inputs, this model infers what we call "structured representations with error bars"—probabilistic representation of a cloth, formalized as a physical system, in the dimensions of its mass and stiffness. The model then performs different perceptual tasks (e.g., visual matching) under this uncertainty. The implementation of this model, which we call Woven, takes advantage of recent advances in

https://doi.org/10.1038/s41467-025-61458-x

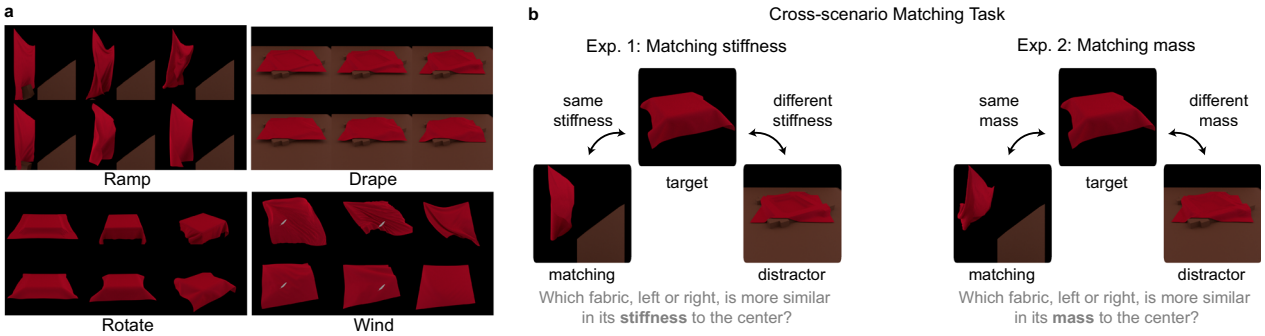

Fig. 2 | Stimuli and tasks. a Snapshot sequences showing example stimuli used in our experiments—animations in which a cloth undergoes a natural transformation in one of four scene configurations (ramp, drape, rotate, and wind). The differences across scene configurations and physical properties introduce substantial variability in image and motion features. In each scenario, the softest cloth is shown in the top row, while the stiffest cloth is shown in the bottom row. All cloths are depicted at the lightest mass. See Supplementary Fig. 2 for examples of cloth with the heaviest mass configuration. **b** Illustration of the cross-scenario matching tasks in Exp. 1 and Exp. 2. A video triad was presented, and the task was to decide which of the two test animations at the bottom had a similar task-relevant physical property as the target animation at the top. The task-relevant physical property was stiffness in Exp. 1 and mass in Exp. 2, which was set to the same value for the target and matching test items. The task-orthogonal physical property (mass in Exp. 1, stiffness in Exp. 2) was assigned at random across the triad. Both the task-relevant and task-orthogonal properties were drawn from a discrete set of predefined values (see Methods). Additionally, the scene configurations of the three animations on a given triad were all different from each other.

probabilistic programming, computer graphics, and computing hardware, providing a rare instance of a sensory-input-computable, probabilistic architecture based on an underlying simulation-based representation in a complex stimulus domain. The result is an account of perception with simulation-based generative models of soft object dynamics at its core.

Although both the bottom-up and intuitive-physics approaches can (by design) infer various physical properties from sensory inputs, the key question in the current project is how well these approaches reflect human soft object perception—in terms of both success and failure. We showed human observers short animations of various types of naturalistic cloth transformations—e.g., waving in the wind, vs. draping over a surface—and then used a matching task to test their ability to perceive two key abstract properties, one of which is specific to soft objects (stiffness) and the other of which is more general (mass). We then compared the human results with the performance of both the DNN model and Woven—as well as instances of Woven in which intuitive physics was selectively ablated when performing the particular matching tasks. We find that although the DNN predicts human success (as it must, given that it is calibrated to perform well on ground-truth property extraction), it nevertheless fails dramatically to capture the overall profile of human success and failure—suggesting that this architecture may be a useful AI tool, but a limited explanation for human perception and performance. In stark contrast, we find that humans and Woven both accurately generalize the stiffness of cloths across different scene configurations, but they also both fail to do so when it comes to matching cloth mass. Remarkably, regardless of the overall accuracy of humans in a given task, Woven (unlike the DNN) quantitatively predicts their fine-grained error patterns—thus explaining human perception not only when it is accurate, but also when it performs at chance levels. We also confirm a prediction of Woven, involving the impact of external forces on perceived cloth properties, via additional analysis of our data.

Results

Psychophysical matching of cloth dynamics

To our knowledge, all existing psychological studies targeting object dynamics, e.g., stiffness of elastic objects[15–17], query just one physical property per stimulus set, without addressing the naturally occurring ambiguities caused by the non-queried properties—e.g., querying participants about an object's softness without also querying about the perceived strength of the applied forces on the same object. This approach leaves the tested models under-constrained when human performance is high: Any model that can accurately infer the queried physical property will also highly correlate with trial-level behavioral accuracy performance. In contrast, probing multiple physical properties on the same stimulus set can be thought of as placing multiple constraints on the space of candidate models, facilitating the emergence of more complete accounts of perception. Here, we thus query participants on both stiffness and mass judgments on a shared set of stimuli. As the goal of this study is to innovate in the modeling of soft object perception, we focused on these two particular properties (as opposed to others, such as tensile strength) simply because they have been frequently highlighted in past work, e.g., refs. 4,9,16,18,19.

To do so, we simulated cloth animations under four different scene configurations (Fig. 2a): "ramp" (where a solid box slides and collides into a hanging cloth; see Supplementary Movies 1, 2); "drape" (where a flat, horizontal cloth falls on a rectangular frame on an otherwise flat surface; see Supplementary Movies 3, 4); "rotate" (where a cloth drapes on a table and then spins synchronously with the table's rotation; see Supplementary Movies 5, 6); and "wind" (where a cloth and small feather blows in the wind; see Supplementary Movies 7, 8). The dynamics identity (a pair of stiffness and mass values) of each simulated cloth was drawn from a predefined set of five stiffness and four mass levels, resulting in a video set of 80 animations (for a sample set, see Supplementary Movies 1–8). (Following conventions in simulation engines, we work with inverse-mass values, the reciprocals of the actual mass; thus, a larger mass value indicates a lighter cloth.) See Methods for simulation and rendering settings.

During the experiment, participants matched cloth stiffness (Exp. 1) or mass (Exp. 2) between a target animation and two test animations (referred to as the "matching" and "distractor" items): on each trial, they selected which test item matched the target item on the relevant property (Fig. 2b). All three animations were displayed simultaneously and replayed automatically until participants responded. Critically, the three animations in each trial were from unique scene configurations—so that each trial featured three out of the four scene configurations (Fig. 2a), with each animation reflecting different combinations of multiple external forces, including gravity, torque, rigid-body collision, and wind. In addition to the variation in external forces, the test items also varied in the dynamical property (stiffness or mass) not queried in the task (task-orthogonal property).

https://doi.org/10.1038/s41467-025-61458-x

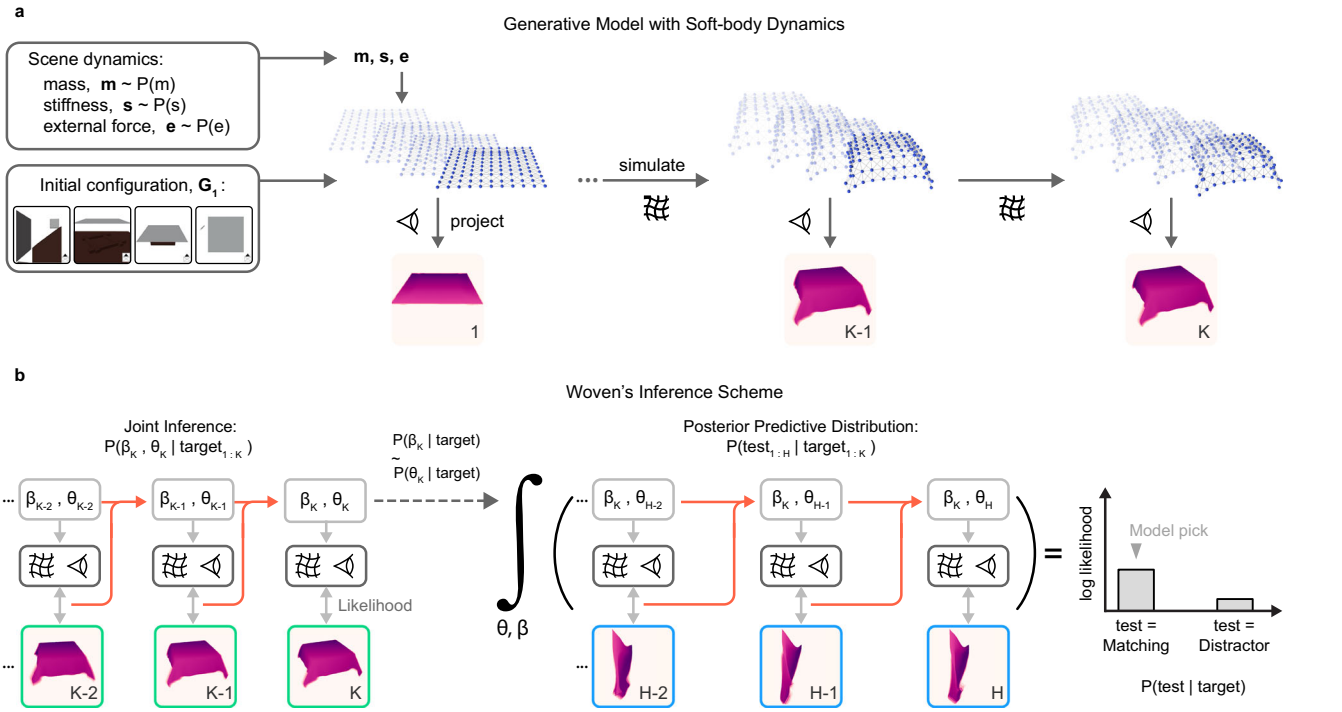

Fig. 3 | Performing the cross-scenario matching tasks with Woven. a Generative model. Woven implements a simulation-based generative model that captures how soft objects move and react to external forces. It defines prior distributions of scene dynamics, including physical properties (m, s) and external forces (e). It then initializes a temporal kernel using these priors and the scene configuration (G_1), including the initial cloth geometry. The simulation unfolds a spring-mass system, whose geometry at each time point $k = 1{:}K$ is projected to sensory features as depth fields. **b** Woven's inference scheme. Woven consists of a sequence of two basic probabilistic computations reflecting the structures of the generative model and the matching tasks. First, Woven jointly estimates a posterior over both the task-relevant β (e.g., stiffness in Exp. 1) and other properties θ (e.g., mass in Exp. 1, and the wind force in the wind scenario for both experiments), conditioned on the depth-field observations of a target animation. It then evaluates the posterior predictive distribution $P(test|target)$, estimating how well the posterior of the target animation explains each test animation—test = matching versus test = distractor—while allowing for the property not queried by the task to be fit further based on its prior and the test animation. We obtain this posterior predictive distribution by summing over the posterior samples of all physical properties (i.e., marginalizing out β and θ). The resulting log-likelihood scores are shown for the matching and distractor test items for an example trial in the stiffness matching task (the same triad in Fig. 2b).

"Woven": a physics-based probabilistic model of soft object perception

We hypothesize that human perception of soft objects arises from probabilistic inference under an internalized generative model of soft object dynamics. We develop a computational account of perception, referred to as "Woven", to implement and test this hypothesis.

At the core of Woven is a simulation-based generative model of how soft objects move and react to external forces, and how the resulting scenes project to sensory features (Fig. 3a). This generative model includes prior distributions over the latent physical properties of cloths—mass $P(m)$ and stiffness $P(s)$—as well as the external force (i.e., wind) strength $P(e)$, which are chosen to be uniform over the (padded) regions between the extreme values these parameters can take in our behavioral experiments. The generative model also takes as input the scene configuration (one of the four that occur in our behavioral experiments), including the initial cloth geometry, G_1. The generative model initializes a temporal kernel using these priors and G_1. At each step k, the temporal kernel unfolds the scene state from G_{k-1} to G_k via approximate "cloth physics", f_ψ, and projects the resulting 3D scene to sensory features, denoted $pred_k$ and termed "depth field", via graphics rendering, f_γ. From G_{k-1} to G_k, each step of the temporal kernel aggregates a small number (four) of actual cloth simulation steps such that the depth field $pred_k$ contains motion information by overlaying the depth images of these actual cloth simulation steps (with a linearly decreasing weight toward the past). We implement f_ψ using a position-based dynamics solver for a planar spring-mass system of particles[20], and f_γ using a depth renderer with perspective projection[21] (see Methods). We note that these implementation choices mostly reflect computational convenience (including utilizing GPU-accelerated simulation for f_ψ and f_γ), rather than detailed theoretical commitments about how these processes might be implemented in the mind and brain.

Woven conditions this generative model on the depth-field observations of an input animation consisting of K steps, $target_{1:K}$, to compute the joint posterior in Eq. (1):

$$P(m, s, e|\text{target}_{1:K}) \propto \mathcal{L}(pred_{1:K}; \text{target}_{1:K}) \cdot P(s) \cdot P(m) \cdot P(e) \cdot \delta_{f_\psi} \cdot \delta_{f_\gamma}, \quad (1)$$

where δ is the delta-dirac function selecting the fixed physics and rendering parameters in f_ψ (e.g., gravity, friction) and f_γ (e.g., viewpoint), $\text{pred}_k = f_\gamma \circ f_\psi(m, s, e; G_{k-1})$, and $\mathcal{L}$ is a likelihood term. The likelihood function $\mathcal{L}$ is a multivariate normal distribution with diagonal covariance, comparing the predicted sensory features with observations at each temporal kernel step.

To perform a given matching task, Woven transforms the posterior in Eq. (1) according to the task structure (Supplementary Fig. 1). Each matching task consists of a relevant physical property β (e.g., stiffness s in Exp. 1) and task-orthogonal physical properties (i.e., properties not queried by the task) θ (e.g., mass m in Exp. 1; Supplementary Fig. 1). Woven estimates how well the inferred task-relevant physical property of the target explains a test animation consisting of H steps, $test_{1:H}$. To do so, we compute the following posterior

predictive distribution in Eq. (2) (dropping the δ notation to avoid clutter).

$$P(\text{test}|\text{target}) \propto \int_{\beta,\theta} \mathcal{L}(pred_{1:H}; test_{1:H}) \cdot P(\beta|\text{target}) \cdot \tilde{P}(\theta|\text{target}) \cdot d\theta d\beta \quad (2)$$

where $P(\beta|\text{target})$ is the posterior of the task-relevant property from Eq. (1), and $\tilde{P}(\theta|\text{target})$ is a mixture of the posterior of the task-orthogonal properties $P(\theta|\text{target})$ and their prior under the generative model $P(\theta)$ (e.g., $P(m)P(e)$ for the stiffness task). The $\tilde{P}(\theta|\text{target})$ term in the posterior predictive distribution allows the task-orthogonal physical properties of the target item to bias the decision boundary: The further the posterior of such a property of the target item from the mean of its prior distribution, the more that property will skew the decision boundary, relative to the task-relevant property. We introduce and test ablation models without this term (see below).

We approximate the distributions in Eqs. (1) and (2) using a sequential Monte Carlo (SMC) algorithm[22]. A schematic of this procedure is shown in Fig. 3b, and its further details are provided in Methods. In our simulations, we find that we can efficiently approximate the quantities in each equation using a small number of SMC particles, which are updated at each time step, forming a sample-based representation of the posterior (and thus, replacing integrals with sums over Monte Carlo samples).

Performing the psychophysical task with Woven. To evaluate Woven's performance, we applied it to the same set of trials as experienced by our participants. On a given trial, Woven was separately simulated for each "target-test" pair (consisting of pairs of target and matching test items, as well as pairs of target and distractor test items), with the total number of simulated chains for each pair equal to the number of times this pair of videos occurred across all participants (with a minimum of three simulated chains per pair). Woven made a decision by comparing the log-likelihood of each test item under the posterior predictive distribution from Eq. (2): $P(\text{test} = \text{matching}|\text{target})$ vs. $P(\text{test} = \text{distractor}|\text{target})$, choosing the one with the higher value. We illustrate Woven's estimates for an example chain on example triads in Supplementary Fig. 3. Woven's hyperparameters were chosen to match the overall average performance of humans in the stiffness matching task (but not the mass matching task). We did so by running a smaller subset of inference chains using different choices for model hyperparameters, including the number of SMC particles and observation noise parameter, to converge to a setting that matched average human accuracy in the stiffness matching task. In additional simulations, we find that our results remain qualitatively the same for nearby choices of these hyperparameters.(see Supplementary Note 1 and Supplementary Tables 1, 2).

Performance-matched baselines: ablation models and a task-optimized DNN. We compare Woven with three alternatives. Critically, similar to Woven, these models were calibrated to match human performance in the stiffness task (Exp. 1). This criterion has two important benefits. First, because these alternative models are similarly performant, they are also compelling baselines when we make comparisons of detailed patterns of behavioral accuracy levels in the stiffness task. Second, because they are not calibrated on the mass task, we can use the overall average performance of these models in the mass task as a test of generalization, in addition to evaluating their detailed patterns of accuracy levels.

The first alternative is a task-optimized convolutional deep neural network (DNN) that lacks the generative model of soft-body dynamics. This model embodies the invariance hypothesis via a powerful bottom-up visual feature hierarchy[1,3,23]. We used a deep neural network architecture, which has been shown to successfully explain human performance[7] (Supplementary Fig. 4a), to regress dynamic properties of cloths from video inputs. We chose this architecture because it not only represents a state-of-the-art approach for processing video input, but also, critically, is specifically designed to match human performance in estimating physical properties of non-rigid materials[7]. Nevertheless, we note that in testing this alternative model, our goal is not to establish whether DNNs, considered as a model class, can account for human soft object perception. (In fact, one can imagine novel DNN architectures, such as those based on graph neural networks, to realize different implementations of Woven parts[24]; see also Discussion). Instead, our goal is to assess whether a performant, previously behaviorally validated instance of the invariance hypothesis, can suffice to explain the variance in human behavior.

Details of the dataset (see Supplementary Fig. 5) and training procedure are available in Methods. To make a decision, we compared the $L1$ distance of the inferred task-relevant physical property between the target and each of the two test items, and chose the closer test item. In the main text, we report the DNN results at the epoch where it performed closest to human-level performance in the stiffness task; we also report results for all epochs (Supplementary Fig. 4c–e) including detailed patterns for the epoch that maximally correlated with behavior, finding qualitatively similar results (Supplementary Fig. 10).

The second and third alternative models are ablations of Woven, referred to as "Woven-No marginalization" and "cue combination", which make decisions without the posterior predictive distribution in Eq. (2). Woven-No Marginalization compares the $L1$ distances between the inferred (based on Eq. (1)) task-relevant physical properties of the target item and each of the two test times (like the DNN model). The cue combination model, on a trial-by-trial basis, makes a decision with a weighted combination of the two physical properties (mass and stiffness), where the weights are inversely proportional to the estimated variance of each of these properties (see Methods). The cue combination model can be viewed as making a decision based on an aggregate representation of the cloth, where the stiffness and mass serve as "inseparable" cues to that aggregate representation. Importantly, both ablation models perform similarly to Woven in overall average accuracy in both the stiffness and mass tasks (Woven-No Marginalization: Supplementary Fig. 8a, e; cue combination: Supplementary Fig. 9a, e), but as we will see, they diverge in their finer-grained accuracy levels.

It is important to note that neither Woven nor these alternatives are designed to receive momentary ground-truth cloth geometry; instead, these models observe sensor-based inputs—i.e., depth fields for Woven, Woven-No Marginalization, and Cue Combination, and RGB images for DNN. And the choice of depth fields in Woven is for the sake of computational convenience, as it is less expensive to project a depth field than an RGB image within the generative model. [In Woven, because the cloths in our animations are generally visible to the observer, we would not expect the posterior distribution (or the posterior predictive distribution, Eqs. (1) and (2)) to substantially differ as a consequence of the choice of the observation space (whether it is images, depth fields, or scene-level information), especially since Woven is provided with the initial scene configuration (G_1), but not with the values of physical properties or external forces. This being said, we view sensor-computable models (relative to scene-computable models) as providing more compelling accounts of visual processing.]

Explaining human stiffness judgments

In Experiment 1, participants ($N = 100$) completed the task of matching cloth stiffness between a target and two test animations (Supplementary Fig. 1a), with each participant performing 56 trials with a uniform distribution of difficulty levels. Difficulty levels were defined by the two factors: The first is the actual differences of the stiffness parameters (i.e., the task-relevant parameters) between the matching

and distractor items (a total of eight levels, shown in the columns of the tile plots in Supplementary Fig. 6a), with small absolute differences leading to more difficult trials. The second involves variation in the task-orthogonal property, which is the difference between the absolute mass differences of the two test items when compared to the target item (a total of seven levels, shown in the rows of Supplementary Fig. 6a), with smaller difference-of-differences leading to easier trials.

To understand what drives the fine-grained patterns of matching performance in humans, we presented the same set of trials to Woven and the DNN model. The detailed accuracy of these models, broken down by the difficulty levels, are shown in Supplementary Fig. 6a). (All models were designed to match humans on average accuracy: Humans: 66%; Woven: 66%; DNN: 66%). These tile plots readily suggest a greater degree of consistency between humans and Woven, relative to the DNN, which we quantify by correlating each model's detailed accuracy levels with those of humans. We find that indeed Woven explains significantly and substantially more variance in human judgments than does the DNN ($r = 0.84$ vs 0.49; $p < 0.001$, using direct bootstrap hypothesis testing; Fig. 4a).

Next, we used multidimensional scaling[25] (see Supplementary Methods for the details) to arrange each of the 80 unique animations that occur across our trials into an embedding space where distances represent subjective perceptual differences. For humans, the optimal 2D embedding (Fig. 4c) shows that stimuli are primarily organized according to stiffness levels (the crescent formation), with smaller distances for similar stiffness values. Additionally, the points deviating from the main dimension imply that other factors, such as mass variation and scene differences, may also play additional roles in shaping the perceptual organization.

We used the same method to also construct embedding spaces for Woven (Fig. 4d) and the DNN (Fig. 4e, see also Supplementary Fig. 7a). We next conducted a Procrustes transformation to align the embedding spaces of the models with the human embedding. The optimized 2D embedding for the two models were qualitatively similar to humans in that they were also primarily organized based on stiffness. However, the embedding space of Woven was substantially more aligned with that of humans than the DNN: Woven's embedding space showed both a greater effect of stiffness (embedding distances induced by stiffness differences were greater for Woven than for the DNN), and the embedding spaces were more dispersed away from the "stiffness crescent", reflecting the effect of stimulus properties other than stiffness (though to a reduced degree, relative to humans). Indeed, quantitatively, we find that Woven explains significantly more variance in the human perceptual space than does the DNN ($r = 0.74$ vs 0.25; $p < 0.001$, using direct bootstrap hypothesis testing; Fig. 4f). This general pattern—wherein Woven better matched the fine-grained aspects of human performance, compared to the DNN—also occurred when considering various subsets of the four different scene configurations alone (Fig. 4 k, l; further detailed in Supplementary Fig. 11).

Explaining human mass judgments

In Experiment 2, another $N = 100$ participants matched cloth mass across target and test animations on the same stimulus set, with each participant performing a total of 54 trials with a uniform distribution of difficulty levels. Mirroring the stiffness task, we compared the accuracy levels as a function of both the mass difference (a total of six levels, shown in the columns of the tile plots in Supplementary Fig. 6b) and the difference-of-differences for the task-orthogonal property (stiffness; a total of nine levels; shown in the rows of the tile plots in Supplementary Fig. 6b).

Surprisingly, and in striking contrast to the stiffness task, humans were at chance on this mass matching task (51%, $p = 0.446$ binomial test). [We note that two previous studies report above-chance human accuracy in a cloth mass estimation task[16,18]. This can be explained by a subtle but critical difference between our experimental setups: instead of comparing cloth mass in the same scenario (i.e., comparing the mass of two cloths waving in the wind), participants in our study were asked to compare cloths across different scenarios (such as a cloth waving in the wind and a cloth draping on the ground).] We first tested whether this substantial reduction in overall average accuracy can be explained by any of the models. (Recall that models were not tuned to match human accuracy in the mass matching task.) Woven's performance, like humans, dropped to chance levels (from 66 to 53%). The source of this reduction in performance seems to be due to the relative difficulty of estimating a posterior over the mass versus stiffness parameter of an observed cloth: Woven-no marginalization model, which excludes the task-relevant marginalization component (Eq. (2)), also performs at chance-level in the mass task Supplementary Fig. 8e. Even though the mass task was also more difficult for the DNN (which is especially salient in the later epochs of training Supplementary Fig. 4d), its performance nevertheless remained higher than that of humans (65%). This shows a better alignment between Woven and humans; but for this alignment to be meaningful, Woven must also explain the detailed error patterns underlying such low accuracy (since of course it is not difficult to make a model that performs at chance).

Critically, despite the chance-level performance of humans and Woven, when comparing their fine-grained accuracy levels (Supplementary Fig. 6b), we found that Woven explains much of the variance in behavior ($r = 0.86$; Fig. 4b). However, this was not the case for the DNN (Supplementary Fig. 6b), which correlated with behavior poorly ($r = 0.25$; $p < 0.001$ pairwise bootstrap comparison to Woven; Fig. 4b). These results were again consistent across the scene configurations, excluding each scene configuration in turn (Fig. 4m, n; see also Supplementary Fig. 12).

What drives this alignment of the chance-level performance in humans and Woven? To understand, we examined the embedding spaces of humans and models. The perceptual embeddings of human judgments revealed stiffness as the predominant dimension, while mass contributes to a wider dispersion of the dots (Fig. 4g). Woven captured this trend quite well, albeit exhibiting a slightly more pronounced impact of mass (Fig. 4h). In contrast, the DNN showed a qualitatively different pattern, highlighting mass as the driving dimension (Fig. 4i, see also Supplementary Fig. 7b). Indeed, quantitatively, Woven's embedding space strongly correlated with humans' perceptual embedding, which was not the case for the DNN ($r = 0.65$ vs 0.00; $p < 0.001$, using direct bootstrap hypothesis testing; Fig. 4j).

Mechanistically, in Woven, this effect of stiffness in the mass task arises from the $\tilde{P}(\theta|\text{target})$ term in the posterior predictive distribution (Eq. (2)), which, as noted before, is a mixture of the prior $P(\theta)$ (i.e., $P(s)$) and the posterior under the target item $P(\theta|\text{target})$ (i.e., $P(s|\text{target})$. The impact of the task-orthogonal property on decision, through the $\tilde{P}(\theta|\text{target})$ term in the posterior predictive distribution, is asymmetrical across the two physical properties. In the mass task, Woven makes accurate inferences about the target's stiffness, which typically diverges considerably from the mean of the prior. This divergence skews the posterior predicted distribution, leading to a bias in Woven's decisions towards the task-orthogonal properties of stiffness in the mass task. This is not the case in the stiffness task, where the inferred mass of the target is often less accurate and remains close to the mean of the prior.

We also compared Woven to its ablations: Woven-no marginalization (excluding the task-relevant marginalization component) and cue combination (which responds by a weighted combination of stiffness and mass, inversely proportional to their reliability), and additionally to the DNN epoch that best correlated with human accuracy levels (in addition to the epoch that matched average human performance; Supplementary Fig. 4d, e). We present the detailed results of these alternative models in Supplementary Figs. 8–10, and provide a summary of the comparisons of these models with Woven in Fig. 5. Even though individual models excel at certain comparisons (although none outperforms Woven), only Woven consistently

https://doi.org/10.1038/s41467-025-61458-x

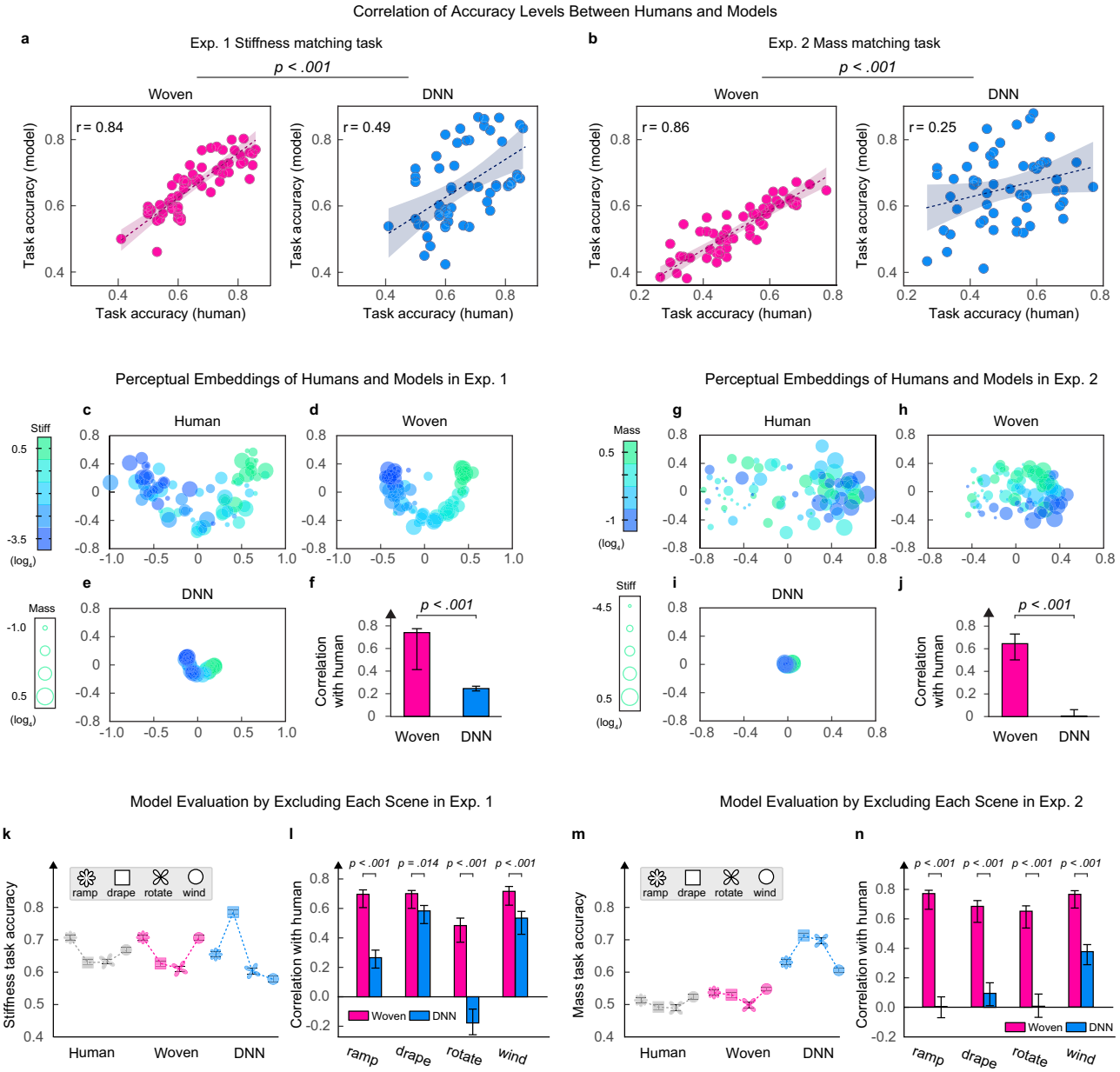

Fig. 4 | Results of the matching tasks. a, **b** Scatterplots showing the correlations of each model with human accuracy across difficulty levels (see Supplementary Fig. 6 for a tile plot of these accuracy levels). Each data point is the average accuracy (of a model or humans) for different combinations of mass and stiffness values in the target and match stimuli. Woven was much more highly correlated, compared to the DNN model, in both the stiffness (**a**) and mass (**b**) tasks. Dotted lines indicate fitted linear regressions with 95% bootstrapped confidence intervals (CIs; shaded regions). **c**–**f** The 2D perceptual space representation of the 80 animations used in the stiffness task, organized by humans (**c**), Woven (**d**), and DNN (**e**). Each dot represents an animation, and closer dots indicate higher perceptual similarity. **f** A bar plot comparing the correlation between model and human perceptual spaces. Woven achieved a significantly higher correlation with humans than the DNN model. **g**–**j**. Same as (**c**–**f**), but for the mass task. **k**, **l** Evaluation of models on subsets of the trials, each excluding one of the four scene configurations in the stiffness task. **k** In the stiffness task, both models are calibrated on the overall average accuracy across all trials, but only Woven recapitulates the finer-grained impact of scene configurations on average accuracy. **l** Woven better correlates with humans' detailed accuracy levels when considering performance in subsets of scene configurations. See also Supplementary Fig. 11. **m**, **n** Evaluation of models on subsets of the trials, each excluding one of the four scene configurations in the mass task. **m** Woven's superior consistency with human performance remains in the mass task. **n** Woven consistently better correlates with humans' detailed accuracy levels when considering performance in subsets of scene configurations. See also Supplementary Fig. 12. In **k**, **m**, data were presented as mean values. In (**f**, **j**, **l**, **n**), data are shown as observed Pearson correlation coefficients. All error bars indicate 95% confidence intervals derived from 1000 bootstrap resamples. Statistical significance is determined using two-sided direct bootstrap hypothesis testing.

explains the human data well across all the different analyses we performed.

Finally, as a more stringent test of models, we compared model and human accuracy at the level of unique individual trials (defined by the unique combinations of stiffness and mass values; 1105 trials in Exp. 1 and 1213 trials in Exp. 2). Remarkably, we find that at this fine-grained comparison to human performance, only Woven is able to consistently explain any variance at all in the data (Fig. 5d, h), lending further support to our hypothesis that simulation-based intuitive physics underlies human perception of soft objects. This also correspondingly represents an especially dramatic and noteworthy failure of the DNN model.

https://doi.org/10.1038/s41467-025-61458-x

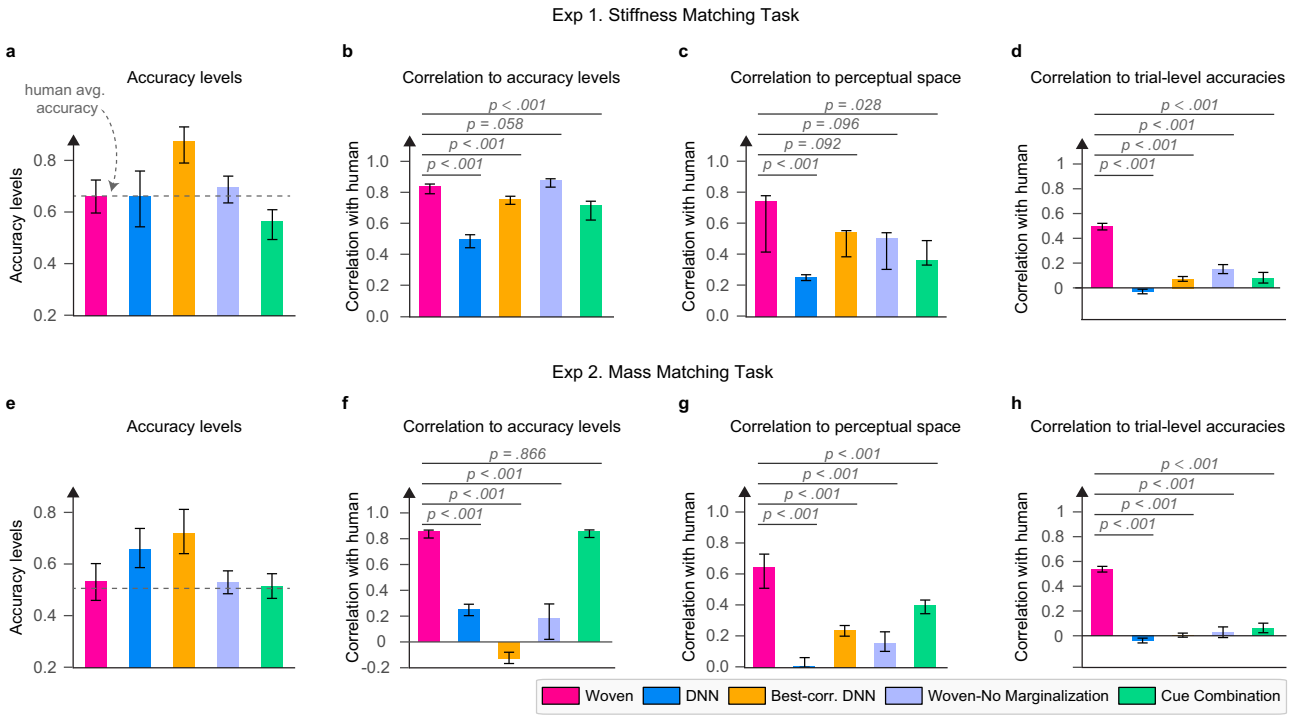

Fig. 5 | Overall comparison of Woven and alternatives in the stiffness and mass tasks. a–d In the stiffness task, the models are evaluated based on their overall average accuracy compared to humans (**a**), and on their correlation with humans in terms of the accuracy levels (**b**) and perceptual embedding spaces (**c**). The models are also evaluated at a finer-grained level of individual trials, defined by the unique combinations of stiffness and mass values (**d**). We plot Woven, DNN, the DNN at the epoch that best correlates with humans ("Best-corr. DNN", which is determined with respect to the stiffness task; the DNN assessed in our primary analyses is also the best-correlating epoch for the mass task), woven-no marginalization, and cue combination. Dashed lines indicate average human accuracy. **e–h** Same as (**a–d**), but for the mass task. Only Woven consistently provides a strong account of behavior across experiments and analyses of performance. In (**a**, **e**), bar heights represent mean values with individual data points overlaid. In all other panels, bar heights indicate observed Pearson correlation coefficients. Error bars indicate 95% confidence intervals derived from 1000 bootstrap resamples. Significance is determined via two-sided direct bootstrap hypothesis testing.

Predicting differential perceptual constancy of stiffness and mass to wind strength

Computational models can be especially useful when they not only capture extant human performance, but also inspire new ways of analyzing existing data (or further experiments). Here we demonstrate how Woven fuels just this type of progress, in an analysis of how an individual scenario—namely that involving wind strength (as in Supplementary Movies 7, 8)—is importantly different from the others.

When rating the stiffness and mass of neutral-texture cloth animations, people's mass ratings, unlike their stiffness ratings, vary substantially as a function of the wind strength applied in the animation (illustrated in Fig. 6a)[16]. In other words, participants have less perceptual constancy when estimating cloth mass (vs. stiffness) under unknown wind strengths. Do similar patterns of perceptual constancy emerge in Woven? To address this, we first aggregated inferred estimates of mass, stiffness, and wind strength across all SMC chains executed on the wind scenario animations in our stimuli. We fit bivariate Gaussian distributions to the estimates of mass against wind strength, and stiffness against wind strength (see Supplementary Methods). We found that indeed the inferred wind strengths bias mass inferences significantly more than stiffness inferences (Fig. 6b, $p < 0.001$, direct bootstrap hypothesis testing).

This leads to a prediction of Woven that is readily testable in our experiments: Given Woven's substantial bias in estimating mass (relative to stiffness), we predicted that humans should be more accurate in matching cloth mass in the trials that exclude (vs. include) the wind scenario. And we further predicted that this should not be the case for the stiffness task, at least not to the same degree. To test this prediction, we set up a linear regression model to predict human mass task accuracy using stiffness differences and mass differences as our predictors. We fitted this regression once on the trials excluding the wind scenario, and once on the trials including the wind scenario. Consistent with our prediction, we found that the coefficient of mass differences was significantly greater in the wind-excluding trials than in the wind-including trials. Moreover, as predicted, this was not the case for the stiffness task: humans performed similarly accurately across the wind-excluding and the wind-including trials. Critically, a pattern closely similar to humans emerged in Woven, unlike the alternative models (Fig. 6c, d).

This result also highlights how and why humans may represent mass despite not always being able to infer it accurately—since it turns out that this (in)ability is highly context sensitive for soft objects (see also[18]). And whereas some approaches may attempt to explain mass judgments by appeal to entirely distinct heuristic mechanisms[26,27], the present approach has the advantage of explaining both stiffness and mass judgments via a single integrated framework. And as such, we suggest that the inaccuracy in our mass judgment task may reflect a rational integration of structure and uncertainty.

Discussion

This project contributes to the study of computational vision along three dimensions—corresponding to its stimuli, its results, and its theoretical conclusions—all of which contrast sharply with many of the most popular approaches in the field and point toward new synergistic directions.

Computational vision of soft objects

The current project highlights the importance and challenges associated with studying soft object perception—as a unique and essential

https://doi.org/10.1038/s41467-025-61458-x

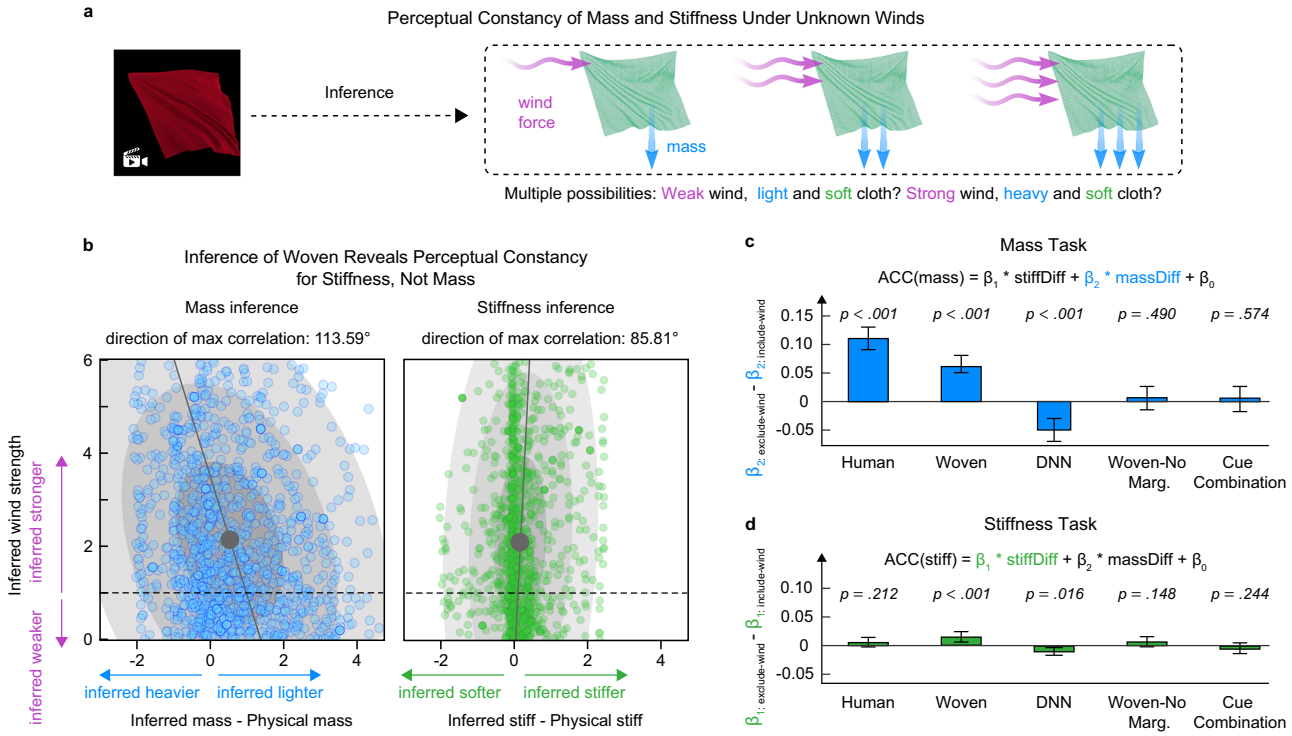

Fig. 6 | Woven predicts differential perceptual constancy to unknown wind strength across mass and stiffness. a Varying the strength of wind blowing at an untextured cloth affects human observers' mass estimates, but not their stiffness estimates. **b** Woven's joint inference of mass, stiffness, and wind strength predicts such differential effects of wind strength on physical properties. When we fit bivariate Gaussian distributions over Woven's inferences (aggregated across all chains we simulated on wind scenario videos), we observe a negative covariance between the mass and wind inference, and nearly no covariance between the stiffness and wind inferences. The shaded ovals indicate probability density. **c** We predict and confirm that humans more accurately match mass in the trials that exclude the wind scenario, relative to those that include the wind scenario. In particular, the effect of the ground-truth mass difference (quantified using the regression model on top, where stiffDiff = $|s_{\text{Match}} - s_{\text{Target}}| - |s_{\text{Distractor}} - s_{\text{Target}}|$ and massDiff = $|m_{\text{Match}} - m_{\text{Distractor}}|$) was significantly greater in the wind-excluding trials than in the wind-including trials. Woven showed qualitatively identical patterns of perceptual constancy as humans, but this effect was reversed for the DNN. **d** Using a similar regression model (where stiffDiff = $|s_{\text{Match}} - s_{\text{Distractor}}|$ and massDiff = $|m_{\text{Match}} - m_{\text{Target}}| - |m_{\text{Distractor}} - m_{\text{Target}}|$), we observed that humans and all models had successful perceptual constancy in their stiffness judgments. Error bars indicate 95% confidence intervals derived from 1000 bootstrap resamples. Significance is determined via two-sided direct bootstrap hypothesis testing.

domain for understanding how human vision works. Most past work that has involved similar approaches has been limited to rigid bodies (e.g., refs. 19,28), has not incorporated any dynamics (e.g., ref. 11), and/or has treated non-rigid stimuli as a constant, pre-configured "nuisance variable" (i.e., a fixed component, like the sliding box in our ramp scenario) while attempting to estimate other properties (such as the shape of a rigid object[13]).

It might initially seem tempting to characterize the jump from rigid to soft objects as relatively modest, or (over-)specialized. But this is not so. In fact, this jump is both substantial and qualitative. The current project has stressed many ways in which soft objects pose exceptionally deep and substantive theoretical challenges for models of perception—due to (1) dramatically greater image-level variability, arising from the high-dimensionality of soft objects, (2) the dramatically greater degrees of intrinsic dynamics of soft objects (insofar as a falling cloth changes vastly more than does a falling block or a pencil), (3) the perception of distinct object properties such as degrees of stiffness, which do not apply to rigid objects, (4) "filling-in" like phenomena of unseen external forces (such as wind), and (5) the naturally and frequently occurring uncertainty due to extrinsic scene elements (as illustrated in Fig. 1).

Our modeling approach is broadly applicable beyond the particular domain we studied here, and could also be applied to the elasticity and softness of non-rigid solids[17,29] and softness, tensile strength and yarn structure of rubber, strings, straps and the strength of knots made with them[30], perception of liquids[31,32], and any multi-material dynamical scene including rigid and non-rigid elements—spanning much of what is in our visual environments.

Explaining human performance and making new predictions about perception

We suspected that solving these unique computational challenges posed by the perception of soft objects might involve physics-based representations in the mind, transcending other popular approaches—and this prediction was confirmed along several dimensions. The key result of this study was that patterns of human performance were only well captured by Woven—a framework with a simulation-based generative model of soft objects at its core. In contrast, multiple competing models—most notably a deep neural network with a powerful bottom-up feature hierarchy—failed to explain human performance in either quantitative or qualitative terms. The details of these analyses were notable (and in part novel) in three related ways, including (1) success vs. error, (2) model-driven predictions, and (3) multiple tasks. First, we tested models' abilities to capture not only patterns of human success, but also patterns of human error. This is critical because models that focus only on success cannot capture human cognition as distinct from generic engineering challenges (and indeed the lack of such a focus has been a common criticism of modeling approaches in general[33,34]). In the current project, we made this focus especially salient by ensuring that all tested models were equated in terms of being able to extract the ground truth—yet only Woven was able to explain how and why humans failed in certain contexts, as when

judging mass as opposed to stiffness (for which the DNN always overperformed). We also emphasize that the magnitude by which Woven outperformed the alternatives (including the DNN) was often not only statistically significant, but relatively massive. This was true throughout the analyses, but was especially noteworthy when analyzing the correlation between the models and human accuracy at the individual trial level: as graphed in Fig. 5d, h, Woven outperformed the other approaches in this context by more than 50%, with no other alternative correlating with human performance by more than 20%. Second, Woven not only explained the baseline patterns of data, but also made a prediction that was then confirmed in additional analyses —that humans' mass estimations (but not their stiffness estimations) were better in scenarios that did not involve wind strength. Third, these results crucially depended on employing multiple potential tasks (here, both stiffness and mass estimation) on the same stimuli set, which provided effective constraints for distinguishing models that are otherwise equally performant.

We additionally highlight a key result in our paper that raises a new challenge for DNNs as a model of human vision. To our knowledge, nearly all existing work that reports DNNs as subpar accounts of human perception focuses on scenarios where DNNs simply underperform, in terms of average accuracy, relative to human performance —that is, they focus on domains where DNNs are not yet that good to begin with (e.g., refs. 13,35). Furthermore, in the domains where DNNs perform on par with (or above) humans in average accuracy, the conclusion in the literature is nearly invariably that the DNNs are good accounts, or at least the best available accounts, of human perception. But this type of conclusion could not be farther from what we find: We find that even when DNNs are at human-level performance (or even super-human performance!) respectively in the stiffness and mass tasks, they are still worse accounts of human visual processing relative to Woven. This presents a qualitatively different statement than all papers we know that evaluate DNNs (regardless of whether they conclude DNNs relate or do not relate to human visual perception) and is enabled by studying the perception of cloths. And moreover, we do not just say DNNs are not good accounts; our main contribution is indeed the Woven model.

An intuitive physics core for perception

This study focused on exploring the intuitive physics basis of soft object perception. Using appearance-equated cloths—equating for texture, thickness, size, and yarn-level geometry—allowed us to effectively pursue this question. However, such appearance information can be correlated with physical properties, which can be utilized by the visual system[3]. Recent work, using unsupervised learning and variational autoencoders, has illustrated the utility of these statistical generative models in modeling humans' perception of optical-material properties[36,37]. Woven's simulation-based generative model, and its inference procedure, can be extended with these type of statistical generative models as well as amortized inference procedures[12,38] to account for additional aspects of soft object perception. Future work should also combine these methods to explore how the visual system infers the initial scene configuration from sensory inputs, including whether an object is cloth-like or more rigid (e.g., ref. 39) and whether violations of softness interfere with object persistence[40].

The central theoretical conclusion of this project is captured by its title: Intuitive physics underlies visual processing of soft objects. And the key notion here is intuitive physics. The current ascendant approach in computational vision involves standard deep neural networks, and the ability of the feature hierarchies learned in these networks—without pre-existing explicit representations—to support impressive performance. Such approaches may be able to capture aspects of human visual processing (e.g., refs. 41–45; but cf. refs. 46,47), but the current project suggests that they may struggle with the unique challenges of soft object perception (and likely beyond).

Beyond DNNs, we also tested an alternative possibility we highlighted in the Introduction—that humans may rely on perhaps simpler but perceptually and analytically more explicit image features such as motion[16,48] and deformation statistics[15]. In these additional analyses, we found that these simple features fail to generalize across different scenarios (see Supplementary Note 2 and Supplementary Figs. 13–15), supporting the conclusion that intuitive physics underlies soft object perception. In cognitive science, it has been suggested that the human mind incorporates assumptions based on universal aspects of human experience[49], and principles of physics seem like especially good candidates in this respect. Models of the mind, as well as AI systems, which incorporate intuitive physics —either through the right sorts of inductive biases in new neural network architectures (e.g., object-like or relational structures[24,50,51]) or more explicitly through probabilistic approaches as we did here—may thus not only align better with human performance (as in the current project), but may also integrate computational modeling with existing work that has stressed the importance of intuitive physics in the mind and brain[8,9,14,52–56].

Methods

Psychophysics experiments

Stimuli. We created 80 unique computer-rendered animations of cloth reacting to external forces in four different scenarios: (1) Rotate: a cloth is draped over a square table, and then the table starts moving (at $t = 2s$, spinning around its axis with a time-varying angular velocity $\omega = -0.5 \times (1.03t - 5.28) + 0.29$(rad/s) while the cloth remains on top (due to sufficient friction between the cloth and the table), moving along with the table's rotation; (2) Drape: a cloth falls under gravity, from a fixed height ($h = 1$ m) in a flat and horizontal configuration, onto a wood frame placed on the floor; (3) Ramp: an initially stationary box slides down on a ramp and colliding with a hanging cloth, with a fixed initial position, size (0.46 m × 0.46 m × 0.46 m), and inv-mass $m = 0.45$; (4) Wind: a hanging cloth is blown by an oscillating wind with horizontal (left-right) and vertical (up-down) components. The horizontal component is modeled using a sign-wave function, while the vertical component employs 1D Perlin noise. The wind direction and period could be partially appreciated by the movement of a light feather in the scene. Each scenario consisted of 20 animations, with each animation lasting 6.67 s at a frame rate of 30 frames per second.

The cloth was simulated using Nvidia's FLeX engine v.1.2.0[20,57], modeling the cloth as a 105 × 105 grid of vertices with homogeneous spring stiffness, spanning a dimension of 2.1 m × 2.1 m. The simulations incorporated a broad range of stiffness values ($4^{-3.5}$, $4^{-2.5}$, $4^{-1.5}$, $4^{-0.5}$, $4^{0.5}$) and reciprocal mass (4^{-1}, $4^{-0.5}$, 4^{0}, $4^{0.5}$). Throughout the paper, the reciprocal mass values are referred to as the "mass value", where a higher mass value in fact indicates a lighter cloth. For each animation, the number of solver iterations was set to 80, the number of substep iterations was set to 17, and all other parameters were set to their default values. A total of 200 time steps were simulated, and subsequently rendered using Blender v. 2.9.1. Cycles Render Engine[58] with the same rendering parameters for the four scenarios, except for a difference in the camera angle. Each image was rendered at a resolution of 540 × 540 pixels. These images were then converted into a video format, with a frame rate of 30 frames per second and a duration of 6.67 seconds.

Participants. We recruited $N = 100$ participants from the crowdsourcing platform Prolific to take part in each of the two experiments, with this sample size chosen before data collection began. For the stiffness experiment, participants had an average age of 27.37 years (SD = 8.67 years), with 63 females and 37 males (based on Prolific self-report). For the mass experiment, participants had an average age of 24.76 years (SD = 6.86 years), with 55 females and 45 males. No person participated in more than one experiment. All participants were naive to the purpose of the experiments, and they were required to perform

the experiment using a laptop or desktop computer. Participants provided electronic informed consent prior to the experiment, and received monetary compensation for their participation in a 35-min session. This study was approved by the Yale University Institutional Review Board (IRB).

We could not record the number of participants who did not meet the first criteria, as they were marked as "returned" in Prolific (the online crowd-sourcing platform we used in this study) along with those participants who signed up for the task but did not attempt it at all or quit for other reasons unknown to us. Six participants in the stiffness experiment and four participants in the mass experiment did not meet the second criteria, and were thus replaced by new participants, so we have $N = 100$ in both experiments. These remaining participants had an accuracy of 98.8% on the attention-check trials for the stiffness experiment and 98.4% for the mass experiment.

Procedure and design. Both stiffness and mass experiments were conducted using PsiTurk[59] and followed the same basic procedure. At the beginning of each experiment, participants were presented with 11 pages of instructions that outlined the task they would need to perform, with a requirement to view each page for a minimum of 3 s. These instructions included examples of stimuli displaying cloth in various scenarios, with different mass and stiffness values. Participants were explicitly informed that on each trial, the stiffness and mass values could vary and that they needed to focus on matching either the stiffness value in the stiffness task, or the mass value in the mass task, depending on the specific experiment they were assigned to. Following the instructions, participants completed two example trials, receiving feedback based on their responses.

During the experiment, each trial began with a central cross displayed on the screen, followed by a trial page presenting a cloth triad at the center of the screen. The triad consisted of the target cloth displayed in the top center, and two test videos presented at the lower bottom (as in Supplementary Fig. 1). Participants were instructed to choose the test videos that corresponded to the target cloth in terms of their stiffness values (or the mass values in the mass task). The videos played on a continuous loop until participants chose an option by clicking the appropriate selection button. Notably, these buttons were intentionally hidden and would only become visible after participants had watched the video for a minimum duration of 6 s. After making their selection, participants proceeded to the next trial without receiving any feedback. Importantly, once a choice was submitted, participants were not allowed to change their selection for that particular trial. At any point during the experiment, leaving full-screen mode would pause the experiment, resuming only after returning to full-screen mode.

In total, each participant completed 56 trials for the stiffness experiment and 54 trials for the mass experiment, with the difficulty levels of the trials uniformly distributed. In addition, scenarios were semi-randomly assigned to ensure a nearly equal number of trials for each possible combination of three scenarios within each participant (14 trials per scenario triplet in the stiffness task and 13 or 14 trials per scenario triplet in the mass task.)

The Woven model

Generative model. The generative model incorporates soft-body dynamics—efficient, game-engine style simulation of cloth motion based on a small number of factors, including the cloth's physical properties (stiffness and mass), external forces (wind, gravity, rigid-body collision), and a set of constraints to govern their interactions. These constraints are based on a spring-mass system created by a lattice-graph representation of the planar, flat cloth shape (thus, the name of our model, Woven), as implemented by the FLeX engine (denoted f_ψ). Given the geometry of the animated cloth, the generative model used a depth renderer to project the 3D form of the cloth to a 2D depth image (denoted f_γ).

The generative model places uniform prior distributions over the physical properties of the cloth (i.e., mass and stiffness) and unknown wind forces: mass $m_0 \backsim \text{Uniform}(0.003, 5.0)$; stiffness $s_0 \backsim \text{Uniform}(0.002, 2.5)$; and wind strength $e_0 \backsim \text{Uniform}(0, 6.0)$. These priors cover the entire range of parameters observed in the stimulus set. In addition, the generative model also has a categorical distribution over the four scene configurations, with equal weight. The generative model also has a fixed parameter for friction. Once an initial scene configuration is drawn using these priors, the generative model simulates soft-body dynamics f_ψ, deforming the shape of the cloth over time steps. This results in the 3D mesh of the cloth M_k at each time step k, which is projected to a depth map D_t, using a graphics renderer f_γ with a pinhole camera; the rendering function is implemented using the Open3D library v. 0.13.0.[21]. The generative model has a fixed parameter of viewing angle for each scene configuration.

Finally, the generative model assumes that the mass, stiffness, and wind strength can vary over time with a temporal kernel according to the following Gaussian distributions: $m_k \backsim N(m_{k-1}, \sigma_m)$, $s_k \backsim N(s_{k-1}, \sigma_s)$, and $e_k \backsim N(e_{k-1}, \sigma_e)$; see the next section for how the standard deviations σ_m, σ_s, and σ_e are parameterized. In this work, we incorporate this assumption in the temporal kernel of the generative model primarily for computational convenience to support efficient inference with fewer particles; future work should explore the perceptual persistence of soft objects' physical properties.

Given an observed video of a target animation $\text{target}_{1:K}$ with K depth fields frames, the generative model induces the following posterior that we wish to estimate.

$$P(m_K, s_K, e_K|\text{target}_{1:K}) \propto P(m_1) \cdot P(s_1) \cdot P(e_1) \cdot P(\text{target}_1|pred_1) \cdot \delta_{f_\psi} \cdot \delta_{f_\gamma} \cdot \prod_{k=2}^{K} P(\text{target}_k|pred_k) \cdot P(m_k|m_{k-1}) \cdot P(s_k|s_{k-1}) \cdot P(e_k|e_{k-1}) \quad (3)$$

where $P(\text{target}_k|\text{pred}_k)$ is a likelihood function based on a multivariate Gaussian with diagonal covariance $target_k \backsim N(\text{pred}_k, \sigma_D)$ modeling the observed depth field as a normal distribution centered around the predicted depth field values with an observation noise of σ_D, $p(m_1)$, $p(s_1)$, $p(e_1)$ are the priors described in the main text. We set $\sigma_D = 8$ in our simulations, and the standard deviation for the physical property kernels are defined in the next section. The equation takes advantage of the notational abbreviation of $\text{pred}_k = f_\gamma f_\psi(m_k, s_k, e_k; G_{k-1})$.

Inference using sequential Monte Carlo (SMC). We use the sequential Monte Carlo (SMC) algorithm to estimate the posterior in Eq. (1) in the main text (with the full set of conditional distributions provided in Eq. (4)).

We initialize inference with 20 particles per chain. The standard deviations of the temporal kernels over the physical parameters were modeled as mixtures:

$$\sigma_m = \begin{cases} 0.04, & \text{with Bernoulli}(0.8); \\ 0.8, & \text{otherwise}, \end{cases}$$
$$\sigma_s = \begin{cases} 0.02, & \text{with Bernoulli}(0.8); \\ 0.4, & \text{otherwise} \end{cases}$$

yielding typically small perturbations in the latent physical properties, while allowing for the possibility of occasional bigger jumps. For the wind strength temporal kernel, its standard deviation was set to $\sigma_e = 0.3$.

We implemented this inference procedure, and the rest of our model, using a state-of-the-art probabilistic programming package, `Gen.jl`[60]. The results reported are based on 62 simulated SMC chains for each video in the stimulus set, with a total of $62 \times 80 = 4960$ chains.

Posterior predictive distribution to perform matching tasks. To perform a given matching task, we must evaluate the posterior predictive distribution $P(\text{test}|\text{target})$ for both test = matching and test = distractor, as detailed in Eq. (2) in the main text. To do so, we perform a

nearly identical SMC procedure (but more efficiently using just three particles) as above, but sampling the physical properties, both task-relevant and task-orthogonal, from $P(m_K, s_K, e_K|\text{target}_{1:K})$ while allowing the task-orthogonal properties the full support of their original prior. To marginalize the parameters, we sum over the sample-based representation of the posterior (all particles).

Using the stiffness task as an example, for a given test animation, we initialize an SMC chain by sampling from the posterior of the target animation $P(m_K, s_K, e_K|\text{target}_{1:K})$ and further updating mass and external force (but not stiffness) using their respective priors by processing the test animation. We then sum across the physical properties to obtain $P(test|\text{target})$ as expressed in the following equation.

$$P(\text{test}_{1:H}|\text{target}_{1:K}) \propto \sum_{m,s,e} P(m_1, s_1, e_1|\text{target}_{1:K}) \cdot p(\text{test}_1|pred_1) \cdot \prod_{h=2}^{H} P(\text{test}_h|pred_h) \cdot P(m_h|m_{h-1}) \cdot P(e_h|e_{h-1}) \quad (4)$$

where the test animation consists of H frames and the equation takes advantage of the notational abbreviation $\text{pred}_h = f_\gamma \circ f_\psi(m_h, s_h, e_h; G_{h-1})$. For each test animation, this posterior predictive SMC chain was run as many times as this test animation appeared in a pair with the underlying target animation across all participants. This resulted in 15,467 and 19,002 posterior predictive SMC chain simulations in the stiffness and mass tasks.

Performance-matched DNN model

The DNN model is used to realize the hypothesis that perception can be explained merely as combinations of bottom-up feature hierarchies. We used a powerful DNN architecture designed for inferring physical properties of dynamic stimuli from video inputs[7]. This model was shown to capture a high degree of variance in human perceptual ratings of these properties. To adapt the model for the current task, accounting for the uncertainty arising from the interactions of multiple properties, we adjusted its output layer to jointly estimate stiffness and mass.

We used different strategies to ensure the DNN model was trained with information comparable to the Woven model and that the testing procedure was also consistent. First, the DNN model was trained on a dataset that spanned all four scenarios, as the scene information is also accessible to Woven. Second, the training datasets included cloth animations with stiffness values sampling from the range of [0.002, 2.5] and mass values from the range of [0.003, 5.0], consistent with the Woven's priors for these two parameters. This enabled the DNN to learn all potential parameter combinations that can be recognized by Woven. Third, like Woven, we trained and tested multiple instances of the DNN architecture, treating each instance as a simulated subject. Finally, the reported architecture and results are a result of exploring a wider range of possible ways in which to set up the DNN architecture and training in order to make it competitive with Woven.

These strategies minimize the imbalance of information available to the two models, ensuring that any differences in their performance result from the representations they have learned.

DNN architecture. The DNN consists of three blocks, each with three layers (convolution, ReLU non-linearity, and max-pooling), which is followed by another block with a fully connected layer. The sequence of blocks in the network are connected to gradually fuse information across input frames culminating in a final fully connected linear layer for stiffness and mass read-out. The architecture of this network is visualized in Supplementary Fig. 4a.

Training details. For finetuning, we modified the output layer of this DNN to align with our specific joint-inference task. Specifically, we replaced the final read-out layer with a new fully connected (and randomly initialized) linear layer for regressing stiffness and mass. We then employed a standard regression loss function, MSE (mean squared error), with equal weights of 0.5 assigned to each property. To train the model, we froze the weights and embeddings of the first convolutional block, while finetuning the second block onward.

We conducted a series of experiments to explore the optimized hyperparameters, including data augmentation (flip, shear, reflect, scale, rotate), initial learning rate (1e − 2, 5e − 3, 1e − 3, 5e − 4, 1e − 4), batch size (32, 64, 128), the number of frozen blocks, optimizer (Adam, Sgdm), and the training label scale (linear, log). After semi-systematically evaluating these parameters, we proceeded with training the network for 30 epochs with an initial learning rate of 0.00001, which is down-scaled by a factor of 10 every 10 epochs. We finetuned the network from the second block onward (while keeping the first block frozen at its pretrained weights), with a batch size of 128, the SGDM optimizer, and training labels on a linear scale. These hyperparameters, in comparison to others we tested, yielded better convergence during training and better correlation to behavior across epochs.

Datasets. To create a dataset for finetuning this DNN, we simulated cloth animations within the same four scene configurations of our stimuli, while introducing notable geometric variation. In the drape and rotate scenarios, we altered the initial angle and position of the cloth. In the rotate scene, in addition to the initial angle and position of the cloth, we also introduced changes to the timing at which the table starts rotating. For the wind scene, we manipulated the phase and amplitude of the wind field's sign function, alongside adjusting the threshold for the wind's on-and-off behavior. Lastly, in the ramp scene, we manipulated both the cloth's position and direction of movement for the sliding cube. These manipulations allowed us to generate diverse cloth behaviors under varying initial conditions and scene dynamics (refer to Supplementary Fig. 5 for dataset examples). To ensure dataset quality, we manually reviewed each simulation, removing those with noticeable artifacts such as cloth penetration or falling out of the camera's view. In total, we obtained a total of 1742 wind videos, 1677 ramp videos, 1588 drape videos, and 1315 rotate videos, each consisting of 200 frames and with a resolution of 540 × 540 pixels. Afterwards, we removed the initial and final ten frames that contained less informative data. Following the original study using this DNN, the remaining frames were then segmented into nine consecutive 20-frame clips. Next, we resized the frames in each clip into 64 × 64 and sequentially concatenated them, resulting in a 64 × 1280 image that served as the input to the DNN model. The final dataset comprised 56,898 images depicting cloth with randomly sampled stiffness values from the range of [0.003, 2.5] and mass values from the range of [0.002, 5.0]. This dataset was randomly split into training and validation sets with a split ratio of 0.15.

For the testing dataset, we used the 80 videos from the psychophysics experiment, and performed the identical reprocessing to transform each video into nine image clips, yielding a total of 720 samples in the testing dataset. Importantly, we ensured that identical replicas of the videos in the testing dataset were not included in the training dataset.

Training results. To equate the information available to the models, we trained the DNN using all four scenarios since the relevant scene information is also accessible to Woven. This encourages the DNN to learn the relevant visual features across various scene constraints for inferring stiffness and mass. We trained 62 instances of the same network, each initialized with the weights of a randomly selected pretrained network provided by van Assen et al. (out of 100 such pretrained networks). Additionally, we randomized both the order and train/validation split of the training stimuli across networks.

The validation loss decreased during training (Supplementary Fig. 4b), and correlation to ground-truth physical parameters increased (Supplementary Fig. 4c), suggesting that the DNN efficiently learned useful patterns in the training dataset for physical property inference.

Comparing DNN and human judgments. To compare the DNN's predictions with human performance, we followed a similar procedure to Woven. In particular, for each human participant, we randomly sampled (with replacement) a DNN, and used it to perform the matching task on the same trials experienced by this participant. For each trial, we calculated the L1 difference in the estimated task-relevant physical property for the two pairs: the left pair (comprising the left and target videos) and the right pair (comprising the right and target videos). The DNN's choice was determined by selecting the pair with smaller difference. This process was repeated ten times for each human participant, and the average over these iterations was taken as the DNN's performance in the matching task.

In the stiffness matching task, the average matching accuracy of the DNN model increased and eventually plateaued across training epochs (Supplementary Fig. 4d). When assessing the correlation between DNN and human accuracy levels, we observed a similar trend of increasing and plateauing across training epochs (Supplementary Fig. 4e).

The mass matching task presented a different trend. Although the training enhanced the model's average matching accuracy (Supplementary Fig. 4d), its correlation with human judgments declined over the course of the training epochs (Supplementary Fig. 4e). Interestingly, the DNN's performance indicates that mass perception is more challenging compared to stiffness, as evidenced by its lower correlation with the ground-truth parameters (Supplementary Fig. 4c) and lower matching accuracy in the mass matching task (Supplementary Fig. 4d).

We determined the number of training epochs for our main analysis by choosing the point where the model's average accuracy matched that of humans in the stiffness task. Here, we selected epoch 1 as the representative DNN model for further analysis.

For additional comparisons, we also considered the best-correlating training epochs of the DNN for each experiment separately. Thus, in Exp. 1, we also assessed the DNN's performance at epoch (16) that correlated maximally with behavior in the stiffness task (Supplementary Fig. 10). In Exp. 2, the DNN at epoch 1 was also the best-correlating epoch—the epoch we reported in the main text. In both cases, our results were qualitatively similar to our main results.

Ablation models of Woven

To understand the contribution of the posterior predictive distribution (Eq. (2) in the main text) toward Woven's ability to capture human judgments, we created two ablation models that use heuristic (non-probabilistic) decision rules for performing the matching task. These decision rules are explained in the main text.

To evaluate the ablation models (Woven-no marginalization and cue combination), we used a similar method to that used for the DNN. Specifically, we randomly selected one simulated chain (from the same pool of chains as Woven's posterior predictive distribution draws from) to represent an individual participant and had the simulated chain perform the same set of trials as the assigned human participant. The model's judgment for each trial was determined using L1 distance, with this process averaged across ten iterations per participant, each time randomly selecting a different chain. The detailed results for these ablation models are presented in Supplementary Figs. 8 (Woven-no marginalization) and 9 (cue combination). A noteworthy result is the performance of the Woven-no marginalization model in the mass task: Despite performing at chance levels like Woven, its accuracy levels did not align with humans (Supplementary Fig. 8e–h).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw (anonymized) behavioral data and the data from model simulations are publicly available at https://doi.org/10.17605/OSF.IO/29JND[61].

Code availability

Code implementing the Woven model and a container for full reproducibility are publicly available at https://github.com/CNCLgithub/Woven.git, a Zenodo version is also available at https://doi.org/10.5281/zenodo.15555479[62]. Code for the behavioral experiment is available at https://github.com/CNCLgithub/cloth-intuitive-physics_psiturk.git, a Zenodo version is also available at 10.5281/zenodo.15556559[63]. Analysis code is publicly available at https://github.com/CNCLgithub/cloth-intuitive-physics_analysis.git, a Zenodo version is also available at 10.5281/zenodo.15555453[64].

References

1. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
2. Bülthoff, H. H. & Yuille, A. Bayesian models for seeing shapes and depth. *Comments Theor. Biol.* **2**, 283–314 (1991).
3. Nishida, S. Image statistics for material perception. *Curr. Opin. Behav. Sci.* **30**, 94–99 (2019).
4. Bi, W., Jin, P., Nienborg, H. & Xiao, B. Estimating mechanical properties of cloth from videos using dense motion trajectories: human psychophysics and machine learning. *J. Vis.* **18**, 12–12 (2018).
5. Paulun, V. C. & Fleming, R. W. Visually inferring elasticity from the motion trajectory of bouncing cubes. *J. Vis.* **20**, 6–6 (2020).
6. Schmid, A. C. & Doerschner, K. Shatter and splatter: the contribution of mechanical and optical properties to the perception of soft and hard breaking materials. *J. Vis.* **18**, 14–14 (2018).
7. van Assen, J. J. R., Nishida, S. & Fleming, R. W. Visual perception of liquids: Insights from deep neural networks. *PLoS Comput. Biol.* **16**, e1008018 (2020).
8. Battaglia, P. W., Hamrick, J. B. & Tenenbaum, J. B. Simulation as an engine of physical scene understanding. *Proc. Natl Acad. Sci. USA* **110**, 18327–18332 (2013).
9. Hamrick, J. B., Battaglia, P. W., Griffiths, T. L. & Tenenbaum, J. B. Inferring mass in complex scenes by mental simulation. *Cognition* **157**, 61–76 (2016).
10. Yuille, A. & Kersten, D. Vision as bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
11. Erdogan, G. & Jacobs, R. A. Visual shape perception as bayesian inference of 3d object-centered shape representations. *Psychol. Rev.* **124**, 740 (2017).
12. Yildirim, I., Belledonne, M., Freiwald, W. & Tenenbaum, J. Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, eaax5979 (2020).
13. Yildirim, I., Siegel, M. H., Soltani, A. A., Chaudhari, S. R. & Tenenbaum, J. B. Perception of 3d shape integrates intuitive physics and analysis-by-synthesis. *Nat. Human Behav.* **8**, 330–335 (2023).
14. Yildirim, I., Siegel, M. & Tenenbaum, J. in *The Cognitive Neurosciences* (eds Mangun, G. R., Poeppel, D. & Gazzangia, M. S.) Ch. 4 (MIT Press, 2020).
15. Paulun, V. C., Schmidt, F., van Assen, J. J. R. & Fleming, R. W. Shape, motion, and optical cues to stiffness of elastic objects. *J. Vis.* **17**, 20–20 (2017).
16. Bi, W. & Xiao, B. Perceptual constancy of mechanical properties of cloth under variation of external forces. In *Proc. ACM Symposium on Applied Perception* 19–23 (ACM, 2016).

17. Kawabe, T. & Nishida, S. Seeing jelly: judging elasticity of a transparent object. In *Proc. Symposium on Applied Perception* 121–128 (ACM, 2016).
18. Bouman, K. L., Xiao, B., Battaglia, P. & Freeman, W. T. Estimating the material properties of fabric from video. *Proc. IEEE International Conference on Computer Vision* 1984–1991 (IEEE Computer Society, 2013).
19. Wu, J., Yildirim, I., Lim, J. J., Freeman, B. & Tenenbaum, J. Galileo: perceiving physical object properties by integrating a physics engine with deep learning. *Adv. Neural Inform. Process. Syst.* **28**, 7–12 (2015).
20. Macklin, M., Müller, M., Chentanez, N. & Kim, T.-Y. Unified particle physics for real-time applications. *ACM Trans. Graph.* **33**, 1–12 (2014).
21. Zhou, Q.-Y., Park, J. & Koltun, V. Open3d: a modern library for 3d data processing. Preprint at arXiv:1801.09847 (2018).
22. Doucet, A. et al. *Sequential Monte Carlo Methods in Practice* Vol. 1 (Springer, 2001).
23. Fleming, R. W. Visual perception of materials and their properties. *Vis. Res.* **94**, 62–75 (2014).
24. Pfaff, T., Fortunato, M., Sanchez-Gonzalez, A. & Battaglia, P. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations* Vol. 8 (2020).
25. Wills, J., Agarwal, S., Kriegman, D. & Belongie, S. Toward a perceptual space for gloss. *ACM Trans. Graph.* **28**, 1–15 (2009).
26. Proffitt, D. R. & Gilden, D. L. Understanding natural dynamics. *J. Exp. Psychol. Hum. Percept. Perform.* **15**, 384 (1989).
27. Gilden, D. L. & Proffitt, D. R. Understanding collision dynamics. *J. Exp. Psychol. Hum. Percept. Perform.* **15**, 372 (1989).
28. Smith, K. et al. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Proc. 33rd International Conference on Neural Information Processing Systems* 8985–8995 (Curran Associates Inc., 2019).
29. Paulun, V. C., Bayer, F. S., Tenenbaum, J. B. & Fleming, R. W. Efficient visual heuristics in the perception of physical object properties. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.24.534031 (2023).
30. Croom, S. & Firestone, C. Looking tight: Visual judgments of knot strength reveal the limits of physical scene understanding. *J. Vis.* **22**, 3448–3448 (2022).
31. Van Assen, J. J. R., Barla, P. & Fleming, R. W. Visual features in the perception of liquids. *Curr. Biol.* **28**, 452–458 (2018).
32. Kawabe, T., Maruya, K., Fleming, R. W. & Nishida, S. Seeing liquids from visual motion. *Vis. Res.* **109**, 125–138 (2015).
33. Marcus, G. F. & Davis, E. How robust are probabilistic models of higher-level cognition? *Psychol. Sci.* **24**, 2351–2360 (2013).
34. Jones, M. & Love, B. C. Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behav. Brain Sci.* **34**, 169–188 (2011).
35. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations* (2018).
36. Storrs, K. R., Anderson, B. L. & Fleming, R. W. Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417 (2021).
37. Liao, C., Sawayama, M. & Xiao, B. Unsupervised learning reveals interpretable latent representations for translucency perception. *PLoS Comput. Biol.* **19**, e1010878 (2023).
38. Dasgupta, I., Schulz, E., Tenenbaum, J. B. & Gershman, S. J. A theory of learning to infer. *Psychol. Rev.* **127**, 412 (2020).
39. Alley, L. M., Schmid, A. C. & Doerschner, K. Visual perception of surprising materials in dynamic scenes. Preprint at *bioRxiv* https://doi.org/10.1101/744458 (2019).
40. Scholl, B. J. Object persistence in philosophy and psychology. *Mind Lang.* **22**, 563–591 (2007).
41. Lindsay, G. W. Convolutional neural networks as a model of the visual system: past, present, and future. *J. Cogn. Neurosci.* **33**, 2017–2031 (2021).
42. Kubilius, J. et al. Brain-like object recognition with high-performing shallow recurrent ANNs. In*Proc. 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., 2019).
43. Zhuang, C. et al. Unsupervised neural network models of the ventral visual stream. *Proc. Natl Acad. Sci. USA* **118**, e2014196118 (2021).
44. Lerer, A., Gross, S. & Fergus, R. Learning physical intuition of block towers by example. In *Proc. 33rd International Conference on International Conference on Machine Learning* 430–438 (JMLR.org, 2016).
45. Conwell, C., Doshi, F. & Alvarez, G. Human-like judgments of stability emerge from purely perceptual features: evidence from supervised and unsupervised deep neural networks. In *2019 Conference on Cognitive Computational Neuroscience* 605–608 (2019).
46. Baker, N., Lu, H., Erlikhman, G. & Kellman, P. J. Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **14**, e1006613 (2018).
47. Bowers, J. S. et al. Deep problems with neural network models of human vision. *Behav. Brain Sci.* **46**, e385 (2023).
48. Bi, W., Jin, P., Nienborg, H. & Xiao, B. Manipulating patterns of dynamic deformation elicits the impression of cloth with varying stiffness. *J. Vis.* **19**, 18–18 (2019).
49. Shepard, R. N. Perceptual-cognitive universals as reflections of the world. *Psychon. Bull. Rev.* **1**, 2–28 (1994).
50. Piloto, L. S., Weinstein, A., Battaglia, P. & Botvinick, M. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nat. Hum. Behav.* **6**, 1257–1267 (2022).
51. Battaglia, P. et al. Interaction networks for learning about objects, relations and physics. *Adv. Neural Inform. Process. Syst.* **29** (2016).
52. Fischer, J., Mikhael, J. G., Tenenbaum, J. B. & Kanwisher, N. Functional neuroanatomy of intuitive physical inference. *Proc. Natl Acad. Sci. USA* **113**, E5072–E5081 (2016).
53. Schwettmann, S., Tenenbaum, J. B. & Kanwisher, N. Invariant representations of mass in the human brain. *Elife* **8**, e46619 (2019).
54. Wong, K. W., Bi, W., Soltani, A. A., Yildirim, I. & Scholl, B. J. Seeing soft materials draped over objects: a case study of intuitive physics in perception, attention, and memory. *Psychol. Sci.* **34**, 111–119 (2023).
55. Ahuja, A., Desrochers, T. M. & Sheinberg, D. L. A role for visual areas in physics simulations. *Cogn. Neuropsychol.* **38**, 425–439 (2021).
56. Bates, C. J., Yildirim, I., Tenenbaum, J. B. & Battaglia, P. Modeling human intuitions about liquid flow with particle-based simulation. *PLoS Comput. Biol.* **15**, e1007210 (2019).
57. NVIDIA Corporation. FleX Physics Simulation Engine, Version 1.2.0. https://github.com/NVIDIAGameWorks/FleX.git (2017).
58. Community, B. O. *Blender – a 3D Modelling and Rendering Package* (Blender Foundation, Stichting Blender Foundation, 2018).
59. Gureckis, T. M. et al. psiturk: an open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* **48**, 829–842 (2016).
60. Cusumano-Towner, M. F., Saad, F. A., Lew, A. K. & Mansinghka, V. K. Gen: a general-purpose probabilistic programming system with programmable inference. In *Proc. 40th ACM SIGPLAN Conference on Programming Language Design and Implementation* 221–236 (ACM, 2019).
61. Bi, W., Shah, A. D., Wong, K. W., Scholl, B. J. & Yildirim, I. Computational models reveal that intuitive physics underlies visual processing of soft objects. *Open Sci. Framework* https://doi.org/10.17605/OSF.IO/29JND (2025).
62. Bi, W., Shah, A. D., Wong, K. W., Scholl, B. J. & Yildirim, I. CNCLgithub/Woven: initial release. *Zenodo* https://doi.org/10.5281/zenodo.15555479 (2025).
63. Bi, W. CNCLgithub/cloth-intuitive-physics_psiturk: initial release. *Zenodo* https://doi.org/10.5281/zenodo.15556559 (2025).

64. Bi, W. CNCLgithub/cloth-intuitive-physics_analysis: initial release. *Zenodo* https://doi.org/10.5281/zenodo.15555453 (2025).
65. Bouman, K. L., Xiao, B., Battaglia, P. & Freeman, W. T. Estimating the material properties of fabric from video. In *2013 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2013).

Acknowledgements

This work was supported by the Air Force Office of Scientific Research (AFOSR) under award number FA9550-22-1-0041 to I.Y. We are grateful for Yale Center for Research and Computing for providing and maintaining a high performance computing cluster (Milgram) utilized by this study. We thank Qi Lin, Mario Belledonne, and the Cognitive and Neural Computation Lab at Yale for comments on a previous version of this manuscript. We also thank Amir A. Soltani for providing a container environment.

Author contributions

W.B., A.D.S., and I.Y. conceived the study, developed the methodology and software. W.B. and A.D.S. collected the data. W.B., A.D.S., and I.Y. performed the analyses and visualized the data with input from K.W.W. and B.J.S. W.B., B.J.S., and I.Y. wrote the manuscript with input from A.D.S. and K.W.W. All authors contributed to the interpretation and editing of the manuscript. B.J.S. and I.Y. supervised the work. I.Y. acquired the funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-61458-x.

Correspondence and requests for materials should be addressed to Wenyan Bi or Ilker Yildirim.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.