A. Overview of Supplementary Materials

This supplementary material provides technical details for the main paper. We provide the following supplementary derivations, experimental details, and quantitative results in this document:

- Section B: Derivation of clustering algorithm from small-variance asymptotics
- Section C: Full derivation of each step in the GenMatter Gibbs sampler
- Section D: Feature-augmented variant and inference algorithm modifications
- Section E: Human RDK psychophysics experiment
- Section F: Gestalt structure-from-motion experiment
- Section G: Details for all 3D RGB experiments, including first frame visualizations, quantitative results on deformable RGB videos, and technical details about the TAP-Vid-DAVIS benchmark. These videos are meant to capture our model's ability to explain deformable matter in a wide variety of settings.

B. Clustering Algorithm from Small-Variance Asymptotics

We present the technical details of the SVA clustering algorithm and recover a rigid group-centric loss function along with an iterative procedure that minimizes it.

Deriving GenMatter as a Clustering Algorithm

$\boldsymbol{\mu}_\ell^{\mathcal{B}}$ update: In the model, $\boldsymbol{\mu}_\ell^{\mathcal{B}} \sim \mathcal{N}(\boldsymbol{\mu}_k^{\mathcal{H}}, \boldsymbol{\Sigma}_k^{\mathcal{H}})$ and $\mathbf{x}_n^t \sim \mathcal{N}(\boldsymbol{\mu}_\ell^{\mathcal{B}}, \boldsymbol{\Sigma}_\ell^{\mathcal{B}})$. Assume $\boldsymbol{\Sigma}_\ell^{\mathcal{B}} = \epsilon\mathbf{I}$ and $\boldsymbol{\Sigma}_k^{\mathcal{H}} = \eta\mathbf{I}$, where $\epsilon/\eta \to 0$. The negative log-conditional of $\boldsymbol{\mu}_\ell^{\mathcal{B}}$ is:

$$\mathcal{L}(\boldsymbol{\mu}_\ell^{\mathcal{B}}) \propto \frac{1}{\epsilon} \sum_{n \in \mathcal{B}_\ell} \left\| \mathbf{x}_n^t - \boldsymbol{\mu}_\ell^{\mathcal{B}} \right\|^2 + \frac{1}{\eta} \left\| \boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}} \right\|^2$$

As $\epsilon/\eta \to 0$, the first term dominates the posterior, so the minimizer of the loss is:

$$\boldsymbol{\mu}_\ell^{\mathcal{B}} = \arg\min_{\boldsymbol{\mu}} \sum_{n \in \mathcal{B}_\ell} \|\mathbf{x}_n^t - \boldsymbol{\mu}\|^2 = \frac{1}{|\mathcal{B}_\ell|} \sum_{n \in \mathcal{B}_\ell} \mathbf{x}_n^t.$$

$\boldsymbol{\mu}_k^{\mathcal{H}}$ update: Assuming $\boldsymbol{\Sigma}_k^{\mathcal{H}} = \eta\mathbf{I}$ and $\epsilon/\sigma_{\mu^{\mathcal{H}}}^2 \to 0$, the negative log-conditional of $\boldsymbol{\mu}_k^{\mathcal{H}}$ can be approximated by:

$$\mathcal{L}(\boldsymbol{\mu}_k^{\mathcal{H}}) \propto \sum_{\ell \in \mathcal{H}_k} \left\| \boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}} \right\|^2$$

where the minimizer is:

$$\boldsymbol{\mu}_k^{\mathcal{H}} = \arg\min_{\mu} \sum_{n \in \mathcal{B}_\ell} \|\boldsymbol{\mu}_\ell^{\mathcal{B}} - \mu\|^2 = \frac{1}{|\mathcal{H}_k|} \sum_{\ell \in \mathcal{H}_k} \boldsymbol{\mu}_\ell^{\mathcal{B}}$$

$\mathbf{R}_k, \mathbf{t}_k$ update: We restrict n in this step to only index points that are assigned to cluster k. Taking the limit of all noise terms $\sigma \to 0$ collapses out the dependence on $\boldsymbol{\mu}_\ell^{\mathcal{B}}$ and gives a deterministic motion model that only depends on the relative position of $\mathbf{x}_n$ with respect to $\boldsymbol{\mu}_k^{\mathcal{H}}$. Noting that we also collapse out $\boldsymbol{\Sigma}_\ell^{\mathcal{V}}$ and σ_V^2, algebraic manipulation gives that the negative log-conditional of $\mathbf{R}_k, \mathbf{t}_k$ is, with $\mathbf{p}_n = \mathbf{x}_n - \boldsymbol{\mu}_k^{\mathcal{H}}$:

$$\mathcal{L}_k(\mathbf{R}_k, \mathbf{t}_k) = \sum_n \left\| \mathbf{x}_n + \mathbf{v}_n - \left(\mathbf{R}_k \mathbf{p}_n + \boldsymbol{\mu}_k^{\mathcal{H}} + \mathbf{t}_k \right) \right\|^2$$

Letting $\mathbf{q}_n = \mathbf{x}_n + \mathbf{v}_n - \boldsymbol{\mu}_k^{\mathcal{H}}$, the loss becomes:

$$\mathcal{L}_k(\mathbf{R}_k, \mathbf{t}_k) = \sum_n \|\mathbf{q}_n - (\mathbf{R}_k \mathbf{p}_n + \mathbf{t}_k)\|^2 .$$

This expression corresponds to the orthogonal Procrustes problem, which has a standard solution. We first define $\bar{\mathbf{p}} = \frac{1}{N}\sum_n \mathbf{p}_n$ and $\bar{\mathbf{q}} = \frac{1}{N}\sum_n \mathbf{q}_n$ and compute $\tilde{\mathbf{p}}_n = \mathbf{p}_n - \bar{\mathbf{p}}$ and $\tilde{\mathbf{q}}_n = \mathbf{q}_n - \bar{\mathbf{q}}$. We then compute the cross-covariance matrix $\mathbf{S}_k = \sum_n \tilde{\mathbf{q}}_n \tilde{\mathbf{p}}_n^\top$ and its SVD $\mathbf{S}_k = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^\top$. The optimal rotation is $\mathbf{R}_k = \mathbf{U}_k \mathbf{V}_k^\top$ and the optimal translation is $\mathbf{t}_k = \bar{\mathbf{q}} - \mathbf{R}_k \bar{\mathbf{p}}$.

$z_n^{\mathcal{B}}$, $z_\ell^{\mathcal{H}}$ update: Small variance analysis gives the same form of the objective function as the previous step. The negative log-conditional for $z_n^{\mathcal{B}}$, $z_\ell^{\mathcal{H}}$ then becomes, with $\mathbf{p}_n = \mathbf{x}_n - \boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}}$:

$$\mathcal{L}(z_n^{\mathcal{B}}, z_\ell^{\mathcal{H}}) = \sum_n \left\| \mathbf{x}_n + \mathbf{v}_n - \left(\mathbf{R}_{z_\ell^{\mathcal{H}}} \mathbf{p}_n + \boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}} + \mathbf{t}_{z_\ell^{\mathcal{H}}} \right) \right\|^2$$

This is a discrete combinatorial optimization problem that involves searching through particle and cluster assignments.

C. Blocked Gibbs Sampling

We describe the Gibbs sampling approach in greater detail than in the main text. We first independently describe each blocked Gibbs step in Appendix C.1. Then, we describe the procedure of these steps used for initialization in Appendix C.2 and tracking in Appendix C.3.

C.1. Gibbs Update Steps

There are twelve variables of interest, separated at different hierarchical levels as shown:

1. Cluster-level variables:

$$\{\boldsymbol{\mu}_k^{\mathcal{H}},\ \boldsymbol{\Sigma}_k^{\mathcal{H}},\ \mathbf{R}_k,\ \mathbf{t}_k,\ \pi_k^{\mathcal{H}}\}_{k=1}^K$$

2. Particle-level variables:

$$\{\boldsymbol{\mu}_\ell^{\mathcal{B}},\ \boldsymbol{\Sigma}_\ell^{\mathcal{B}},\ \mathbf{v}_\ell,\ \boldsymbol{\Sigma}_\ell^{\mathcal{V}},\ z_\ell^{\mathcal{H}},\ \pi_\ell^{\mathcal{B}}\}_{\ell=1}^L$$

Algorithm 2 Clustering Algorithm for GenMatter via Small-Variance Asymptotics

Input:
  Number of clusters and particles K, L
  Data point positions $\{\mathbf{x}_n, \mathbf{v}_n\}_{n=1}^N$,
Initialize: Assign data points to particles $z_n^{\mathcal{B}}$, particles to clusters $z_\ell^{\mathcal{H}}$
repeat
  for each particle $\ell = 1, \ldots, L$ **do**
    Compute particle mean: $\boldsymbol{\mu}_\ell^{\mathcal{B}} \leftarrow \frac{1}{|\mathcal{B}_\ell|} \sum_{n: z_n^{\mathcal{B}} = \ell} \mathbf{x}_n$
  end for
  for each cluster $k = 1, \ldots, K$ **do**
    Compute cluster mean: $\boldsymbol{\mu}_k^{\mathcal{H}} \leftarrow \frac{1}{|\mathcal{H}_k|} \sum_{\ell: z_\ell^{\mathcal{H}} = k} \boldsymbol{\mu}_\ell^{\mathcal{B}}$
    Collect point pairs $(\mathbf{x}_n, \mathbf{v}_n)$ assigned to cluster k
    Center points: $\mathbf{p}_n \leftarrow \mathbf{x}_n - \boldsymbol{\mu}_k^{\mathcal{H}}$, $\mathbf{q}_n \leftarrow \mathbf{x}_n + \mathbf{v}_n - \boldsymbol{\mu}_k^{\mathcal{H}}$
    Compute cross-covariance: $\mathbf{S}_k \leftarrow \sum_n \mathbf{q}_n \mathbf{p}_n^\top$
    Compute SVD: $\mathbf{S}_k = \mathbf{U}_k \Sigma_k \mathbf{V}_k^\top$
    Set rotation: $\mathbf{R}_k \leftarrow \mathbf{U}_k \mathbf{V}_k^\top$
    Set translation: $\mathbf{t}_k = \bar{\mathbf{q}} - \mathbf{R}_k \bar{\mathbf{p}}$
  end for
  for each data point $n = 1, \ldots, N$ **do**
    Compute motion loss: $\mathcal{L}_n(z_n^{\mathcal{B}}, z_\ell^{\mathcal{H}}) \leftarrow \left\| \mathbf{x}_n + \mathbf{v}_n - \left(\mathbf{R}_{z_\ell^{\mathcal{H}}} \left(\mathbf{x}_n - \boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}} \right) + \boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}} + \mathbf{t}_{z_\ell^{\mathcal{H}}} \right) \right\|^2$
    Update $z_n^{\mathcal{B}}$, $z_\ell^{\mathcal{H}}$ to minimize $\sum_n \mathcal{L}_n(z_n^{\mathcal{B}}, z_\ell^{\mathcal{H}})$
  end for
until assignments converge or objective does not decrease

3. Data point-level variables:

$$\{z_n^{\mathcal{B}}\}_{n=1}^N$$

For each of these variables, we independently describe each of the Gibbs updates.

C.1.1. Data point-to-Particle Assignments ($z_{1:N}^{\mathcal{B}}$)

We update each data point's particle assignment $z_n^{\mathcal{B}}$ for $n = 1, \ldots, N$, using the conditional:

$$p(z_n^{\mathcal{B}} = \ell \mid \mathbf{x}_n, \mathbf{v}_n, \text{rest}) \propto \pi^{\mathcal{B}}(\ell) \cdot \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_\ell^{\mathcal{B}}, \boldsymbol{\Sigma}_\ell^{\mathcal{B}}) \cdot \mathcal{N}(\mathbf{v}_n \mid \mathbf{v}_\ell, \boldsymbol{\Sigma}_\ell^{\mathcal{V}})$$

The prior is given by categorical weights $\pi^{\mathcal{B}}$; the likelihood is a product of two Gaussians over position $\mathbf{x}_n$ and velocity $\mathbf{v}_n$. We compute unnormalized log-probabilities $\tilde{p}_{n,\ell}$ for each particle:

$$\tilde{p}_{n,\ell} = \log \pi^{\mathcal{B}}(\ell) + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_\ell^{\mathcal{B}}, \boldsymbol{\Sigma}_\ell^{\mathcal{B}}) + \log \mathcal{N}(\mathbf{v}_n \mid \mathbf{v}_\ell, \boldsymbol{\Sigma}_\ell^{\mathcal{V}})$$

and normalize to obtain the categorical:

$$p(z_n^{\mathcal{B}} = \ell) = \frac{\exp(\tilde{p}_{n,\ell})}{\sum_{\ell'=1}^L \exp(\tilde{p}_{n,\ell'})}$$

from which we sample:

$$z_n^{\mathcal{B}} \sim \text{Categorical}(p(z_n^{\mathcal{B}} = 1), \ldots, p(z_n^{\mathcal{B}} = L))$$

All data points are jointly reassigned in a blocked manner, each selecting the particle that best explains its position and motion, weighted by the prior over particles.

C.1.2. Particle Mixture Weights $\boldsymbol{\pi}^{\mathcal{B}}$

We update the particle mixture weights $\boldsymbol{\pi}^{\mathcal{B}}$ conditioned on data point-to-particle assignments $\{z_n^{\mathcal{B}}\}$. By Dirichlet–Categorical conjugacy, the conditional distribution becomes:

$$\boldsymbol{\pi}^{\mathcal{B}} \mid \{z_n^{\mathcal{B}}\} \sim \text{Dir}(\beta_1 + M_1, \ldots, \beta_L + M_L)$$

where $M_\ell = \#\{n : z_n^{\mathcal{B}} = \ell\}$ counts how many data points are currently assigned to each particle ℓ. This step re-weights the prior particle proportions according to updated data point assignments.

C.1.3. Particle Spatial Means $\boldsymbol{\mu}_\ell^{\mathcal{B}}$

We update each particle center $\boldsymbol{\mu}_\ell^{\mathcal{B}}$ from its Gaussian conditional, combining: (1) a spatial prior from its assigned cluster, (2) position likelihoods from assigned data points, and (3) a velocity constraint derived from rigid motion.

Let $\mathbf{A}_\ell = \mathbf{R}_{z_\ell^{\mathcal{H}}} - \mathbf{I}$ and $\mathbf{b}_\ell = \mathbf{t}_{z_\ell^{\mathcal{H}}} - \mathbf{A}_\ell \boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}}$. Then:

$$\mathbf{v}_\ell \sim \mathcal{N}(\mathbf{A}_\ell \boldsymbol{\mu}_\ell^{\mathcal{B}} + \mathbf{b}_\ell, \sigma_V^2 \mathbf{I})$$

The conditional distribution is a Gaussian-Gaussian conjugate of the form:

$$\boldsymbol{\mu}_\ell^{\mathcal{B}} \,|\boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}}, \boldsymbol{\Sigma}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}}, \mathbf{v}_\ell, \mathbf{t}_{z_\ell^{\mathcal{H}}}, \mathbf{R}_{z_\ell^{\mathcal{H}}}, \sigma_V^2, \{\mathbf{x}_n : z_n^{\mathcal{B}} = \ell\}, \boldsymbol{\Sigma}_\ell^{\mathcal{B}} \sim \mathcal{N}(\mathbf{P}_\ell^{-1}\mathbf{m}_\ell, \mathbf{P}_\ell^{-1})$$

with precision and mean:

$$\mathbf{P}_\ell = (\boldsymbol{\Sigma}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}})^{-1} + N_\ell(\boldsymbol{\Sigma}_\ell^{\mathcal{B}})^{-1} + \frac{1}{\sigma_V^2}\mathbf{A}_\ell^\top \mathbf{A}_\ell$$

$$\mathbf{m}_\ell = (\boldsymbol{\Sigma}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}})^{-1}\boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}} + (\boldsymbol{\Sigma}_\ell^{\mathcal{B}})^{-1}\mathbf{S}_\ell + \frac{1}{\sigma_V^2}\mathbf{A}_\ell^\top(\mathbf{v}_\ell - \mathbf{b}_\ell),$$

where N_ℓ is the number of data points assigned to particle ℓ, and $\mathbf{S}_\ell = \sum_{n:z_n^{\mathcal{B}}=\ell} \mathbf{x}_n$ is the sum of their positions.

C.1.4. Particle Spatial Covariances $\boldsymbol{\Sigma}_\ell^{\mathcal{B}}$

We update each particle's spatial covariance matrix $\boldsymbol{\Sigma}_\ell^{\mathcal{B}}$ using Normal–Inverse-Wishart conjugacy. Let $N_\ell = \#\{n : z_n^{\mathcal{B}} = \ell\}$ be the number of data points assigned to particle ℓ, and define the scatter matrix:

$$\mathbf{S}_\ell = \sum_{n:z_n^{\mathcal{B}}=\ell} (\mathbf{x}_n - \boldsymbol{\mu}_\ell^{\mathcal{B}})(\mathbf{x}_n - \boldsymbol{\mu}_\ell^{\mathcal{B}})^\top$$

Given an Inverse-Wishart prior $\mathcal{W}^{-1}(\boldsymbol{\Psi}^{\mathcal{B}}, \nu^{\mathcal{B}})$, the conditional distribution is:

$$\boldsymbol{\Sigma}_\ell^{\mathcal{B}} \mid \boldsymbol{\mu}_\ell^{\mathcal{B}}, \{\mathbf{x}_n : z_n^{\mathcal{B}} = \ell\} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}'_\ell = \boldsymbol{\Psi}^{\mathcal{B}} + \mathbf{S}_\ell, \nu^{\mathcal{B}} + N_\ell)$$

This update adjusts each particle's spatial uncertainty based on the observed spread of its assigned data points.

C.1.5. Particle Velocity Means $\mathbf{v}_\ell$

We update each particle velocity anchor $\mathbf{v}_\ell$ via a Gaussian conditional distribution combining: (1) a rigid motion prior from its assigned cluster, and (2) velocity observations from assigned data points. Let $\bar{\mathbf{v}}_\ell = \mathbf{t}_{z_\ell^{\mathcal{H}}} + (\mathbf{R}_{z_\ell^{\mathcal{H}}} - \mathbf{I})(\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_{z_\ell^{\mathcal{H}}}^{\mathcal{H}})$ be the prior mean.

Given the set $\{\mathbf{v}_n : z_n^{\mathcal{B}} = \ell\}$ and count $N_\ell = \#\{n : z_n^{\mathcal{B}} = \ell\}$, the conditional is a Gaussian-Gaussian conjugate update:

$$\mathbf{v}_\ell \mid \bar{\mathbf{v}}_\ell, \sigma_V^2, \boldsymbol{\Sigma}_\ell^{\mathcal{V}}, \{\mathbf{v}_n : z_n^{\mathcal{B}} = \ell\} \sim \mathcal{N}(\boldsymbol{\mu}_\ell^v, \boldsymbol{\Sigma}_\ell^v)$$

with:

$$(\boldsymbol{\Sigma}_\ell^v)^{-1} = \frac{1}{\sigma_V^2}\mathbf{I} + N_\ell(\boldsymbol{\Sigma}_\ell^{\mathcal{V}})^{-1}$$

$$\boldsymbol{\mu}_\ell^v = \boldsymbol{\Sigma}_\ell^v \left(\frac{1}{\sigma_V^2}\bar{\mathbf{v}}_\ell + (\boldsymbol{\Sigma}_\ell^{\mathcal{V}})^{-1} \sum_{n:z_n^{\mathcal{B}}=\ell} \mathbf{v}_n \right)$$

This update accounts for the velocity prediction from the cluster's rigid transform along with the empirical data point velocities, with each contribution weighted by its respective uncertainty.

C.1.6. Particle Velocity Covariances $\boldsymbol{\Sigma}_\ell^{\mathcal{V}}$

Each particle's velocity covariance $\boldsymbol{\Sigma}_\ell^{\mathcal{V}}$ is inferred using Normal–Inverse-Wishart conjugacy. Let $N_\ell = \#\{n : z_n^{\mathcal{B}} = \ell\}$ be the number of data points assigned to particle ℓ, and define the velocity scatter:

$$\mathbf{T}_\ell = \sum_{n:z_n^{\mathcal{B}}=\ell} (\mathbf{v}_n - \mathbf{v}_\ell)(\mathbf{v}_n - \mathbf{v}_\ell)^\top$$

Given prior $\mathcal{W}^{-1}(\boldsymbol{\Psi}^{\mathcal{V}}, \nu^{\mathcal{V}})$, the conditional distribution is:

$$\boldsymbol{\Sigma}_\ell^{\mathcal{V}} \mid \mathbf{v}_\ell, \{\mathbf{v}_n : z_n^{\mathcal{B}} = \ell\} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}'_\ell = \boldsymbol{\Psi}^{\mathcal{V}} + \mathbf{T}_\ell,\ \nu^{\mathcal{V}} + N_\ell)$$

This update reflects the velocity noise structure within each particle, accounting for spread in assigned data point velocities.

C.1.7. Particle-to-Cluster Assignments ($z_{1:L}^{\mathcal{H}}$)

We update each particle's cluster assignment $z_\ell^{\mathcal{H}}$ for $\ell = 1, \ldots, L$, using the conditional:

$$p(z_\ell^{\mathcal{H}} = k \mid \boldsymbol{\mu}_\ell^{\mathcal{B}}, \mathbf{v}_\ell, \text{rest}) \propto \pi^{\mathcal{H}}(k) \cdot \mathcal{N}(\boldsymbol{\mu}_\ell^{\mathcal{B}} \mid \boldsymbol{\mu}_k^{\mathcal{H}}, \boldsymbol{\Sigma}_k^{\mathcal{H}}) \cdot \mathcal{N}\big(\mathbf{v}_\ell \mid \mathbf{t}_k + (\mathbf{R}_k - \mathbf{I}) \times (\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}}), \sigma_V^2 \mathbf{I}\big)$$

The prior is given by categorical weights $\pi^{\mathcal{H}}$; the likelihood combines a spatial Gaussian over the particle's position $\boldsymbol{\mu}_\ell^{\mathcal{B}}$ and a velocity Gaussian that accounts for rigid-body motion induced by the cluster's rotation $\mathbf{R}_k$ and translation $\mathbf{t}_k$. We compute unnormalized log-probabilities $\tilde{p}_{\ell,k}$ for each cluster:

$$\tilde{p}_{\ell,k} = \log \pi^{\mathcal{H}}(k) + \log \mathcal{N}(\boldsymbol{\mu}_\ell^{\mathcal{B}} \mid \boldsymbol{\mu}_k^{\mathcal{H}}, \boldsymbol{\Sigma}_k^{\mathcal{H}}) + \log \mathcal{N}\big(\mathbf{v}_\ell \mid \mathbf{t}_k + (\mathbf{R}_k - \mathbf{I})(\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}}), \sigma_V^2 \mathbf{I}\big)$$

and normalize to obtain the categorical:

$$p(z_\ell^{\mathcal{H}} = k) = \frac{\exp(\tilde{p}_{\ell,k})}{\sum_{k'=1}^K \exp(\tilde{p}_{\ell,k'})}$$

from which we sample:

$$z_\ell^{\mathcal{H}} \sim \text{Categorical}(p(z_\ell^{\mathcal{H}} = 1), \ldots, p(z_\ell^{\mathcal{H}} = K))$$

This constitutes a blocked Gibbs step, where all particle-to-cluster assignments are jointly updated. Each particle selects the cluster whose spatial and rigid motion parameters best explain its position and velocity.

C.1.8. Cluster Mixture Weights $\boldsymbol{\pi}^{\mathcal{H}}$

We update the cluster mixture weights $\boldsymbol{\pi}^{\mathcal{H}}$ given particle-to-cluster assignments $\{z_\ell^{\mathcal{H}}\}$. Using Dirichlet–Categorical conjugacy, the conditional is:

$$\boldsymbol{\pi}^{\mathcal{H}} \mid \{z_\ell^{\mathcal{H}}\} \sim \mathrm{Dir}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

where $N_k = \#\{\ell : z_\ell^{\mathcal{H}} = k\}$ is the number of particles assigned to cluster k. This step updates the prior cluster proportions based on current assignment counts.

C.1.9. Cluster Spatial Means $\boldsymbol{\mu}_k^{\mathcal{H}}$

We update each cluster center $\boldsymbol{\mu}_k^{\mathcal{H}}$ via a Gaussian conditional that integrates: (1) a Gaussian prior centered at $\mu^{\mathcal{H}}$, (2) assigned particle centers $\boldsymbol{\mu}_\ell^{\mathcal{B}}$, and (3) observed particle velocities corrected by the cluster's affine transform.

Let $\mathbf{A}_k = \mathbf{I} - \mathbf{R}_k$ and $\mathbf{b}_\ell = \mathbf{t}_k - \mathbf{A}_k \boldsymbol{\mu}_\ell^{\mathcal{B}}$. Then the velocity residual is:

$$\mathbf{r}_\ell = \mathbf{v}_\ell - \mathbf{b}_\ell$$

Given the sum of assigned particle means $\mathbf{S}_k = \sum_{\ell : z_\ell^{\mathcal{H}} = k} \boldsymbol{\mu}_\ell^{\mathcal{B}}$, the velocity residual sum $\mathbf{R}_k = \sum_{\ell : z_\ell^{\mathcal{H}} = k} \mathbf{r}_\ell$, and the count $N_k = \#\{\ell : z_\ell^{\mathcal{H}} = k\}$ of particles assigned to cluster k, the conditional is:

$$\boldsymbol{\mu}_k^{\mathcal{H}} \mid \mu^{\mathcal{H}}, \sigma_H^2, \boldsymbol{\Sigma}_k^{\mathcal{H}}, \mathbf{t}_k, \mathbf{R}_k, \sigma_V^2, \mathbf{R}_k, \{\boldsymbol{\mu}_\ell^{\mathcal{B}}, \mathbf{v}_\ell : z_\ell^{\mathcal{H}} = k\} \sim \mathcal{N}(P_k^{-1}\mathbf{m}_k, P_k^{-1})$$

with:

$$P_k = \frac{1}{\sigma_H^2}\mathbf{I} + N_k \left(\boldsymbol{\Sigma}_k^{\mathcal{H}\,-1} + \frac{1}{\sigma_V^2}\mathbf{A}_k^\top \mathbf{A}_k \right)$$

$$\mathbf{m}_k = \frac{1}{\sigma_H^2}\mu^{\mathcal{H}} + \boldsymbol{\Sigma}_k^{\mathcal{H}\,-1}\mathbf{S}_k + \frac{1}{\sigma_V^2}\mathbf{A}_k^\top \mathbf{R}_k$$

This update integrates global priors, spatial evidence from assigned particles, and velocity-based constraints under rigid motion. We parallelize this step by batching cluster-level quantities over K and particle-level inputs over L, with per-cluster residual aggregation. The final blocked multivariate normal update samples new cluster means in parallel from their respective posteriors.

C.1.10. Cluster Spatial Covariances $\boldsymbol{\Sigma}_k^{\mathcal{H}}$

We infer each cluster's spatial covariance $\boldsymbol{\Sigma}_k^{\mathcal{H}}$ using a Normal–Inverse-Wishart update conditioned on its assigned particles. Let $L_k = \#\{\ell : z_\ell^{\mathcal{H}} = k\}$ be the number of particles assigned to cluster k, and define the cluster-centered scatter:

$$\mathbf{S}_k = \sum_{\ell : z_\ell^{\mathcal{H}} = k} (\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}})(\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}})^\top$$

Given the Inverse-Wishart prior $\mathcal{W}^{-1}(\boldsymbol{\Psi}^{\mathcal{H}}, \nu^{\mathcal{H}})$, the conditional becomes, with $\boldsymbol{\Psi}'_k = \boldsymbol{\Psi}^{\mathcal{H}} + \mathbf{S}_k$ and $\nu'_k = \nu^{\mathcal{H}} + L_k$:

$$\boldsymbol{\Sigma}_k^{\mathcal{H}} \mid \boldsymbol{\mu}_k^{\mathcal{H}}, \{\boldsymbol{\mu}_\ell^{\mathcal{B}} : z_\ell^{\mathcal{H}} = k\} \sim \mathcal{W}^{-1}(\boldsymbol{\Psi}'_k, \nu'_k)$$

This posterior captures the spatial extent of each cluster based on the spread of its assigned particle centers.

C.1.11. Cluster Rotation $\mathbf{R}_k$

We update each cluster's rotation matrix $\mathbf{R}_k$ by evaluating a discrete set of candidate rotations $\{\mathbf{R}^{(j)}\}_{j=1}^{M_r}$ drawn from a spherical cap (e.g., von Mises–Fisher). For each candidate, we compute a probability based on how well the induced rigid motion explains observed particle velocities. Let $\bar{\mathbf{v}}_\ell^{(j)} = \mathbf{t}_k + (\mathbf{R}^{(j)} - \mathbf{I})(\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}})$ be the expected velocity for particle ℓ under candidate j. Then:

$$\log \tilde{q}_j = \sum_{\ell : z_\ell^{\mathcal{H}} = k} \log \mathcal{N}(\mathbf{v}_\ell \mid \bar{\mathbf{v}}_\ell^{(j)}, \sigma_V^2 \mathbf{I})$$

Adding the prior log-probabilities $\log p(\mathbf{R}^{(j)})$, we normalize the log-scores to obtain:

$$q_j = \frac{\exp(\log \tilde{q}_j + \log p(\mathbf{R}^{(j)}))}{\sum_{j'=1}^{M_r} \exp(\log \tilde{q}_{j'} + \log p(\mathbf{R}^{(j')}))}$$

from which we sample:

$$\mathbf{R}_k \sim \mathrm{Categorical}(\{q_j\}_{j=1}^{M_r})$$

This update selects the rotation that best aligns relative particle positions with their observed velocities, conditioned on the current cluster translation $\mathbf{t}_k$, velocity noise σ_V^2, cluster means ($\boldsymbol{\mu}_k^{\mathcal{H}}$) and assigned particle means ($\{\boldsymbol{\mu}_\ell^{\mathcal{B}} : z_\ell^{\mathcal{H}} = k\}$).

C.1.12. Cluster Translation Velocities $\mathbf{t}_k$

We update each cluster's translation velocity $\mathbf{t}_k$ by evaluating a discrete set of candidate translations $\{\mathbf{t}^{(m)}\}_{m=1}^{M_t}$ sampled from an isotropic Gaussian prior $\mathcal{N}(\mathbf{0}, s^2\mathbf{I})$. Each candidate is scored based on how well it explains the observed particle velocities under the current rotation $\mathbf{R}_k$. Let $\bar{\mathbf{v}}_\ell^{(m)} = \mathbf{t}^{(m)} + (\mathbf{R}_k - \mathbf{I})(\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}})$ be the expected velocity for particle ℓ under candidate m. Then:

$$\log \tilde{p}_m = \sum_{\ell : z_\ell^{\mathcal{H}} = k} \log \mathcal{N}(\mathbf{v}_\ell \mid \bar{\mathbf{v}}_\ell^{(m)}, \sigma_V^2 \mathbf{I})$$

We add prior log-probabilities and normalize to form a categorical:

$$p_m = \frac{\exp(\log \tilde{p}_m + \log p(\mathbf{t}^{(m)}))}{\sum_{m'=1}^{M_t} \exp(\log \tilde{p}_{m'} + \log p(\mathbf{t}^{(m')}))}$$

from which we sample:

$$\mathbf{t}_k \sim \mathrm{Categorical}(\{p_m\}_{m=1}^{M_t})$$

This update selects the translation that best explains the observed particle velocities, conditioned on current cluster rotation $\mathbf{R}_k$, velocity noise σ_V^2, cluster center $\boldsymbol{\mu}_k^{\mathcal{H}}$, and assigned particle means $\{\boldsymbol{\mu}_\ell^{\mathcal{B}} : z_\ell^{\mathcal{H}} = k\}$.

C.2. Initialization Procedure

It is well known that MCMC chains are sensitive to the initialization and should be initialized at a high density region. In both the 2D and 3D variants of GenMatter, we use K-Means clustering and a data-driven approach to initialize the MCMC chain for the initial frame ($T = 0$).

C.2.1. K-Means and Data-driven Initialization at T=0

Given the number of particles (L), we use K-means via a K-Means++ initialization to initialize the particle spatial positions ($\boldsymbol{\mu}_\ell^{\mathcal{B}}$). We then use an additional K-means step to initialize the cluster spatial positions ($\boldsymbol{\mu}_k^{\mathcal{H}}$) by treating the particle spatial positions as data points to cluster.

This K-means initialization provides initial values for assignments at both layers ($z_n^{\mathcal{B}}$, $z_\ell^{\mathcal{H}}$). We then use these assignments to initialize the mixture weights at both layers ($\pi^{\mathcal{B}}$, $\pi^{\mathcal{H}}$) by computing the empirical frequencies of each cluster and normalizing: $\pi_\ell^{\mathcal{B}} = \frac{M_\ell}{N}$ and $\pi_k^{\mathcal{H}} = \frac{N_k}{L}$, where M_ℓ is the number of datapoints assigned to particle ℓ and N_k is the number of particles assigned to cluster k. We initialize the velocity mean of each particle $\mathbf{v}_\ell$ by averaging the observed velocities of the datapoints assigned to it:

$$\mathbf{v}_\ell = \frac{1}{M_\ell} \sum_{n: z_n^{\mathcal{B}} = \ell} \mathbf{v}_n.$$

To initialize the covariance matrices, we compute the sample covariance of the relevant residuals for each component:

1. **Particle Spatial Covariance:**

$$\boldsymbol{\Sigma}_\ell^{\mathcal{B}} = \frac{1}{M_\ell - 1} \sum_{n: z_n^{\mathcal{B}} = \ell} (\mathbf{x}_n - \boldsymbol{\mu}_\ell^{\mathcal{B}})(\mathbf{x}_n - \boldsymbol{\mu}_\ell^{\mathcal{B}})^\top.$$

2. **Particle Velocity Covariance:**

$$\boldsymbol{\Sigma}_\ell^{\mathcal{V}} = \frac{1}{M_\ell - 1} \sum_{n: z_n^{\mathcal{B}} = \ell} (\mathbf{v}_n - \mathbf{v}_\ell)(\mathbf{v}_n - \mathbf{v}_\ell)^\top.$$

3. **Cluster Spatial Covariance:**

$$\boldsymbol{\Sigma}_k^{\mathcal{H}} = \frac{1}{N_k - 1} \sum_{\ell: z_\ell^{\mathcal{H}} = k} (\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}})(\boldsymbol{\mu}_\ell^{\mathcal{B}} - \boldsymbol{\mu}_k^{\mathcal{H}})^\top.$$

To initialize each cluster's rigid transform $(\mathbf{R}_k, \mathbf{t}_k)$, we apply the Kabsch algorithm to align assigned particle positions with their next-frame displacements. For cluster k, we collect all datapoints $\mathbf{x}_n$ assigned to particles ℓ with $z_\ell^{\mathcal{H}} = k$ and define their estimated displacements $\mathbf{x}'_n = \mathbf{x}_n + \mathbf{v}_n$. Let $\mathcal{X}_k = \{\mathbf{x}_n\}$ and $\mathcal{X}'_k = \{\mathbf{x}'_n\}$ be the source and target sets.

We compute centroids $\bar{\mathbf{x}}_k = \frac{1}{|\mathcal{X}_k|} \sum \mathbf{x}_n$, $\bar{\mathbf{x}}'_k = \frac{1}{|\mathcal{X}'_k|} \sum \mathbf{x}'_n$, and form centered sets $\tilde{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}}_k$, $\tilde{\mathbf{x}}'_n = \mathbf{x}'_n - \bar{\mathbf{x}}'_k$. The cross-covariance matrix is:

$$\mathbf{H}_k = \sum_n \tilde{\mathbf{x}}_n \tilde{\mathbf{x}}_n'^\top$$

We compute the singular value decomposition $\mathbf{H}_k = \mathbf{U}_k \boldsymbol{\Sigma}_k \mathbf{V}_k^\top$, and define the optimal rotation as:

$$\mathbf{R}_k = \mathbf{V}_k \mathbf{D}_k \mathbf{U}_k^\top$$

where $\mathbf{D}_k$ is defined as:

$$\mathbf{D}_k = \begin{cases} \begin{bmatrix} 1 & 0 \\ 0 & \det(\mathbf{V}_k \mathbf{U}_k^\top) \end{bmatrix} & \text{(2D model)} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \det(\mathbf{V}_k \mathbf{U}_k^\top) \end{bmatrix} & \text{(3D model)} \end{cases}$$

The corresponding translation is:

$$\mathbf{t}_k = \bar{\mathbf{x}}'_k - \mathbf{R}_k \bar{\mathbf{x}}_k$$

This provides an initialization of cluster motion consistent with the observed displacements of assigned particles. The update is applied independently for each cluster $k = 1, \ldots, K$.

C.2.2. Data-Dependent Hyperparameters

We initialize model hyperparameters directly from empirical statistics computed on the initial frame ($T = 0$). The global cluster location prior $\mu^{\mathcal{H}}$ is set to the median datapoint position, while the prior spatial scale $\boldsymbol{\Psi}^{\mathcal{B}}, \boldsymbol{\Psi}^{\mathcal{H}}, \boldsymbol{\Psi}^{\mathcal{V}}$ are initialized using the median initialized particle and cluster covariances length scales.

The degrees of freedom $\nu^{\mathcal{B}}, \nu^{\mathcal{H}}, \nu^{\mathcal{V}}$ are initialized proportionally to the number of datapoints assigned, weighted by particle or cluster weights:

$$\begin{aligned} \nu^{\mathcal{B}} &= \left\lfloor \text{median}(w_\ell^{\mathcal{B}} \cdot N) \right\rfloor, \\ \nu^{\mathcal{H}} &= \left\lfloor \text{median}(w_k^{\mathcal{H}} \cdot N) \right\rfloor, \\ \nu^{\mathcal{V}} &= \left\lfloor \text{median}(w_\ell^{\mathcal{B}} \cdot N) \right\rfloor \end{aligned}$$

where $w_\ell^{\mathcal{B}}$ and $w_k^{\mathcal{H}}$ are the normalized empirical weights of each particle and cluster.

C.3. Tracking Gibbs Procedure

To perform inference over video sequences, we extend our generative particle model into the sequential filtering regime using a structured Markov Chain Monte Carlo (MCMC) procedure. Specifically, we implement a blocked Gibbs sampler that leverages the causal ordering of the variables from the previous frame to initialize each frame and performs bottom-up inference to refine all data point-, particle-, and cluster-level variables. Our approach maintains a tractable posterior approximation at each timestep by propagating forward a subset of latent variables and resampling the remaining ones conditioned on new observations. This sequential per-frame MCMC design supports inference in dynamic scenes where data associations must be re-inferred at every timestep.

At each timestep t, we target the posterior over latent variables given the observed data point positions $\mathbf{x}^t_{1:N}$ and velocities $\mathbf{v}^t_{1:N}$:

$$p(\boldsymbol{\mu}^t_{\mathcal{H}}, \boldsymbol{\Sigma}^t_{\mathcal{H}}, \mathbf{R}^t_{\mathcal{H}}, \mathbf{t}^t_{\mathcal{H}}, \boldsymbol{\mu}^t_{\mathcal{B}}, \mathbf{v}^t_{\mathcal{B}}, \boldsymbol{\Sigma}^t_{\mathcal{V}}, z^t_{1:N}, z^t_{1:L}, \boldsymbol{\pi}^t_{\mathcal{B}}, \boldsymbol{\pi}^t_{\mathcal{H}} \mid \mathbf{x}^t_{1:N}, \mathbf{v}^t_{1:N})$$

where $\boldsymbol{\Sigma}_{\mathcal{B}}$ (particle spatial covariances) are held fixed throughout tracking in both 2D and 3D experiments to preserve the spatial extent of the deformable visual matter represented by each particle, and particle-to-cluster assignments $z^t_{1:L}$ are held fixed only in the 3D case to keep consistency with the initial scene segmentation.

Particle Propagation and Initialization Each frame begins by propagating the inferred particle means using their previously inferred velocity vectors:

$$\tilde{\boldsymbol{\mu}}^{\mathcal{B},t}_{\ell} = \boldsymbol{\mu}^{\mathcal{B},t-1}_{\ell} + \mathbf{v}^{t-1}_{\ell}$$

This serves as an initialization for the particle positions in the next frame.

First Assignment: Spatial Anchoring Data points are first assigned to particles based on spatial likelihoods alone:

$$p(z^{\mathcal{B},t}_n = \ell \mid \mathbf{x}^t_n) \propto \pi^{\mathcal{B}}_{\ell} \cdot \mathcal{N}(\mathbf{x}^t_n \mid \tilde{\boldsymbol{\mu}}^{\mathcal{B},t}_{\ell}, \boldsymbol{\Sigma}^{\mathcal{B}}_{\ell})$$

This step is crucial because, in the absence of known correspondences across frames, we cannot assume that data point n at time $t-1$ is the same as datapoint n at time t. Instead, we reinterpret each new frame as an unordered set of observations and rely on spatial proximity to propagated particle means to re-establish associations. By using position alone and excluding any top-down beliefs from velocity or cluster structure, this step provides a stable initialization for the rest of the Gibbs updates. Note that this is a partial version of the full assignment step described in Appendix C.1.1, used here to anchor the initial framewise alignment. After assignments, we update the mixture weights $\boldsymbol{\pi}^{\mathcal{B}}$ by sampling from their conjugate Dirichlet distribution (Appendix C.1.2).

Particle Mean Update After data points have been assigned to particles based on spatial proximity, we update each particle's spatial mean to better reflect this assignment. Specifically, we sample the particle mean from its posterior conditioned on the assigned data points and the expected motion induced by its cluster assignment, as detailed in Appendix C.1.3. Since the assignments in the previous Gibbs step compensate for the absence of point-wise correspondences, this update typically results in small adjustments to the propagated means, ensuring that particles remain anchored to observed data while maintaining temporal coherence with the previous frame.

Second Assignment and Particle Refinement A second data point-to-particle assignment uses both spatial and velocity likelihoods as described in Appendix C.1.1:

$$p(z^{\mathcal{B},t}_n = \ell \mid \mathbf{x}^t_n, \mathbf{v}^t_n) \propto \pi^{\mathcal{B}}_{\ell} \cdot \mathcal{N}(\mathbf{x}^t_n \mid \boldsymbol{\mu}^{\mathcal{B}}_{\ell}, \boldsymbol{\Sigma}^{\mathcal{B}}_{\ell}) \cdot \mathcal{N}(\mathbf{v}^t_n \mid \mathbf{v}_{\ell}, \boldsymbol{\Sigma}^{\mathcal{V}}_{\ell})$$

This step helps resolve ambiguous associations by combining spatial proximity with motion information. The mixture weights $\boldsymbol{\pi}^{\mathcal{B}}$ are updated again based on the refined assignments (Appendix C.1.2).

Each particle's velocity mean $\mathbf{v}_{\ell}$ is updated from its posterior as described in Appendix C.1.5, and the velocity covariance $\boldsymbol{\Sigma}^{\mathcal{V}}_{\ell}$ is resampled as shown in Appendix C.1.6. These updates reflect the motion structure inferred from grouped data point velocities.

Cluster-level Updates Each particle is assigned to a cluster using a joint spatial and velocity likelihood as described in Appendix C.1.7, and the cluster mixture weights $\boldsymbol{\pi}^{\mathcal{H}}$ are resampled using the equation in Appendix C.1.8. Conditioned on these assignments, the cluster mean $\boldsymbol{\mu}^{\mathcal{H}}_k$ and spatial covariance $\boldsymbol{\Sigma}^{\mathcal{H}}_k$ are updated from their conditional distributions (Appendix C.1.9 and C.1.10), and the rigid transform $(\mathbf{R}_k, \mathbf{t}_k)$ is inferred by categorical sampling over candidate rotations and translations (Appendix C.1.11 and C.1.12).

In the 3D experiment, particle-to-cluster assignments $z^t_{1:L}$ are held fixed throughout tracking to stabilize the scene representation's semantic content, which provides a reliable prior over object structure. However, cluster parameters including spatial statistics and rigid transforms are still inferred at each frame to update the spatial localization of the structure given in the original segmentation.

D. Feature-augmented Variant

D.1. Model Modification and Initialization

In the feature-augmented variant of our model, we incorporate image features as additional dimensions of the data points. Following the main text, we define augmented data points $\tilde{\mathbf{x}}_n = [\mathbf{x}_n; \mathbf{f}_n]$ where $\mathbf{f}_n$ are feature vectors extracted from the image. We use the first 10 PCA components of DINO features, where the PCA basis is computed by analyzing all per-pixel features across the entire video. Each particle ℓ is associated with a feature mean $\mathbf{f}_{\ell}$, and the sampling process of the data point features $\mathbf{f}_n$ from the particle features is defined as a Gaussian with variance σ^2_F:

$$\mathbf{f}_n \sim \mathcal{N}(\mathbf{f}_{\ell}, \sigma^2_F \mathbf{I})$$

We only fit our per-particle feature parameter during initialization. We perform the steps described in Appendix C.2.1, followed by computing the initial feature mean

of each particle $\mathbf{f}_\ell$ as the average feature vector of its assigned data points:

$$\mathbf{f}_\ell = \frac{1}{M_\ell} \sum_{n: z_n^{\mathcal{B}} = \ell} \mathbf{f}_n$$

where $M_\ell = \#\{n : z_n^{\mathcal{B}} = \ell\}$ is the number of data points assigned to particle ℓ. This feature mean serves as the representative feature vector for each particle throughout inference.

D.2. Data point-to-Particle Assignments with Feature Likelihood

The main modification to the Gibbs sampler involves the data point-to-particle assignment step, which is modified to include feature similarity. We update each data point's particle assignment $z_n^{\mathcal{B}}$ for $n = 1, \ldots, N$, using the conditional distribution:

$$\begin{aligned} p(z_n^{\mathcal{B}} = \ell \mid \mathbf{x}_n, \mathbf{v}_n, \mathbf{f}_n, \text{rest}) \propto \pi^{\mathcal{B}}(\ell) \cdot \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_\ell^{\mathcal{B}}, \boldsymbol{\Sigma}_\ell^{\mathcal{B}}) \\ \cdot \mathcal{N}(\mathbf{v}_n \mid \mathbf{v}_\ell, \boldsymbol{\Sigma}_\ell^{\mathcal{V}}) \\ \cdot \mathcal{N}(\mathbf{f}_n \mid \mathbf{f}_\ell, \sigma_F^2 \mathbf{I}) \end{aligned}$$

The prior is given by categorical weights $\pi^{\mathcal{B}}$, and the likelihood now consists of three independent Gaussian terms: one for position $\mathbf{x}_n$, one for velocity $\mathbf{v}_n$, and one for features $\mathbf{f}_n$. The feature likelihood uses an isotropic covariance $\sigma_F^2 \mathbf{I}$, which assumes that the features are independent and identically distributed.

We compute unnormalized log-probabilities $\tilde{p}_{n,\ell}$ for each particle:

$$\begin{aligned} \tilde{p}_{n,\ell} = \log \pi^{\mathcal{B}}(\ell) + \log \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_\ell^{\mathcal{B}}, \boldsymbol{\Sigma}_\ell^{\mathcal{B}}) \\ + \log \mathcal{N}(\mathbf{v}_n \mid \mathbf{v}_\ell, \boldsymbol{\Sigma}_\ell^{\mathcal{V}}) + \log \mathcal{N}(\mathbf{f}_n \mid \mathbf{f}_\ell, \sigma_F^2 \mathbf{I}) \end{aligned}$$

and normalize to obtain the categorical conditional distribution:

$$p(z_n^{\mathcal{B}} = \ell) = \frac{\exp(\tilde{p}_{n,\ell})}{\sum_{\ell'=1}^{L} \exp(\tilde{p}_{n,\ell'})}$$

from which we sample:

$$z_n^{\mathcal{B}} \sim \text{Categorical}(p(z_n^{\mathcal{B}} = 1), \ldots, p(z_n^{\mathcal{B}} = L))$$

This update is also a blocked update, executed in a computational manner similar to Appendix C.1.1.

E. Human Psychophysics Experiment

A total of 9 RDKs were created, with each RDK having 3 separate time points and locations where we introduce the red and green dot probes to create a total of 27 stimuli. We recruited a total of 150 human participants through the Prolific platform and all participants were paid at least the local minimum wage for an expected completion time of 4 minutes. The study was designed and conducted under an approved institutional review board (IRB) protocol. All demographic data collected were fully anonymized, and no personally identifying information was provided or collected. All participants were filtered for the following conditions:

1. Fluent in English as the study is conducted in English.
2. Explicitly declared to not have color-blindness, as this study requires each participant to distinguish the red and green probes clearly from the rest of the points in the stimuli.
3. Has normal to corrected vision, as this study requires clear vision of the stimuli.

The instructions as viewed on Prolific for this study can be seen in Figure 7. We used Google Forms to conduct the data collection.

The instructions were repeated in the Google Form and each participant saw two familiarization trials with feedback on the correct answer. Figure 8 shows how these familiarization trials looked to the participant.

F. Gestalt 3D Inference

We provide additional implementation details for the Gestalt 3D structure-from-motion experiments described in the main paper.

Data preprocessing We compute RAFT optical flow and VideoDepthAnything monocular depth on the native 1000×1000 frames, then downsample to 96×96 for inference. We lift 2D pixels to 3D using a pinhole camera model with focal length scaled by 2.0 to enhance depth separation, and compute 3D motion vectors from optical flow.

Initialization GenMatter uses $L = 100$ particles and $K = 5$ clusters. We use the initialization procedure in Appendix C.2.1, with the addition of a coarse segmentation mask proposal containing all data points with flow magnitude above the median flow magnitude, as well as data points within a standard deviation of the median depth. These assumptions are loose and apply to any natural image regime where GenMatter could be run, and are used only to accelerate MCMC burn-in (since a proposal does not change the posterior being approximated). We run 50 Gibbs sweeps on frame 0 to initialize all model parameters.

Per-frame Gibbs schedule For tracking frames $t = 1, \ldots, 4$, we apply 20 Gibbs iterations focused on velocity parameters, followed by 500 full Gibbs sweeps. Particle-to-cluster assignments and particle spatial covariances remain fixed throughout tracking, but data point-to-particle assignments are resampled at each frame.

Figure 9 shows an example of the Gestalt structure-from-motion stimuli with different textures. We visualize the first frame of scene 00000 rendered with seven different texture patterns. The Gestalt experiment uses 20 scenes (00000–00019), each rendered with these seven textures (00, 07,

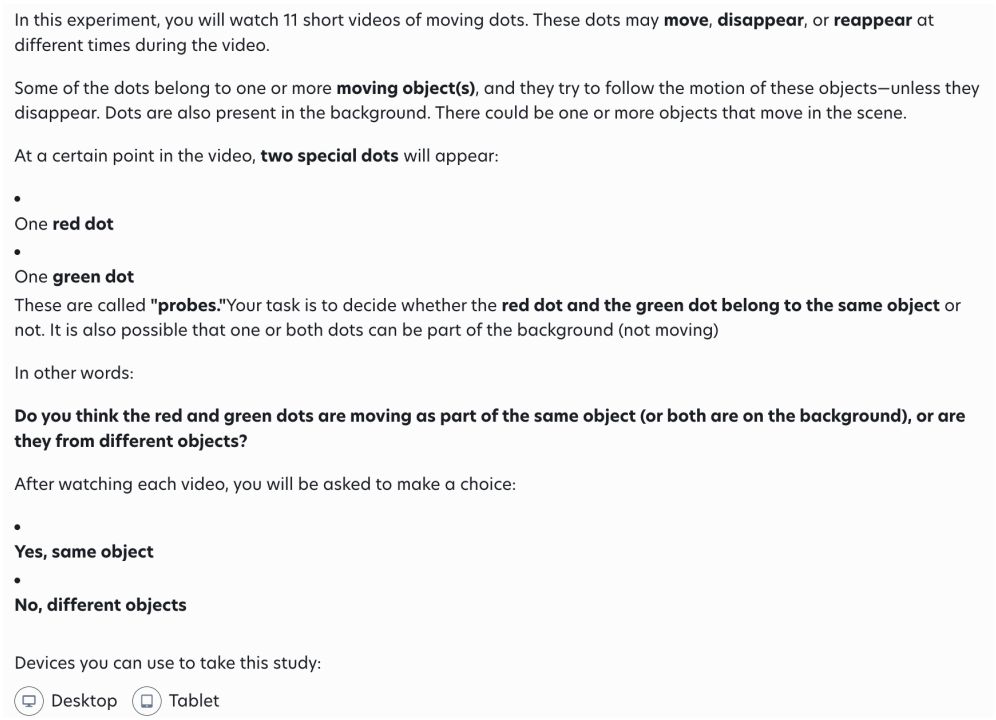


Figure 7. Instructions shown to all participants. This task was allowed to be conducted on either a desktop or a tablet. The 11 videos mentioned refer to the 2 familiarization trials and 9 test trials. Details of compensation is cropped out to preserve anonymity.

13, 16, 21, 22, 25), yielding 140 total stimuli to evaluate structure-from-motion segmentation across diverse visual appearances.

G. RGB 3D Inference

Table 3. **Jaccard Index on Supplementary Videos.** We report the Jaccard index for GenMatter and CoTracker3 on each supplementary video. Best per video is bolded.

Video	GenMatter	CoTracker3
cloth_bag	**0.84**	0.32
gray_jacket	0.91	**0.98**
jello	**0.93**	0.57
manta_ray	0.96	**1.00**
eagle	0.86	**0.89**
ostrich	0.91	**0.97**
purple_jacket	0.81	**0.99**
snake	**0.93**	0.47
whiskey_swirl	0.49	**0.79**
wine_swirl	0.67	**0.97**

G.1. Experimental Details

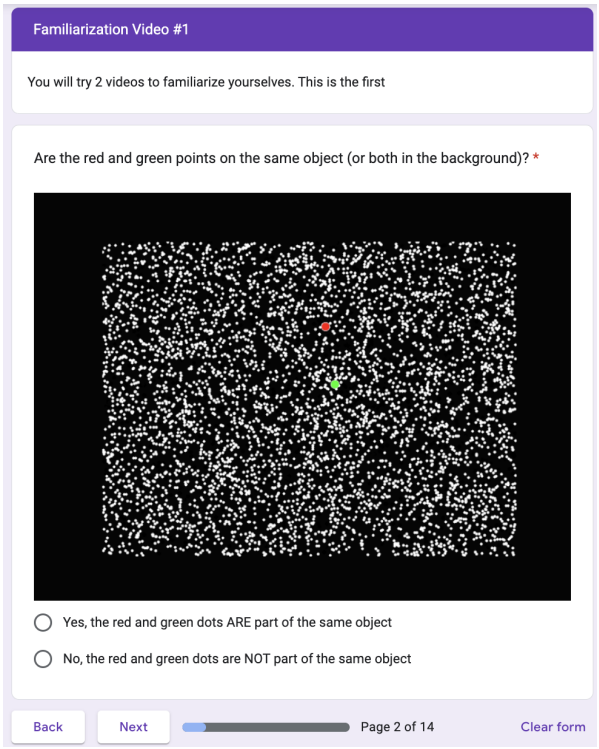
GenMatter Setup We provide additional technical implementation details for our TAP-Vid-DAVIS experiments.

Initialization At frame 0, we perform 30 Gibbs sweeps to initialize all model parameters before sequential tracking begins. Particles are initialized through hierarchical K-means clustering on 3D positions lifted from tracked points using monocular depth estimates. For each particle, DINO features $\mathbf{f}_\ell^{\mathcal{B}}$ are initialized by averaging DINO descriptors over all pixels assigned to that particle. When SAM2 frame-0 masks are available, we adaptively determine the number of clusters K based on the number of components in the mask rather than fixing K. We sample tracked points uniformly across each frame for the whole video.

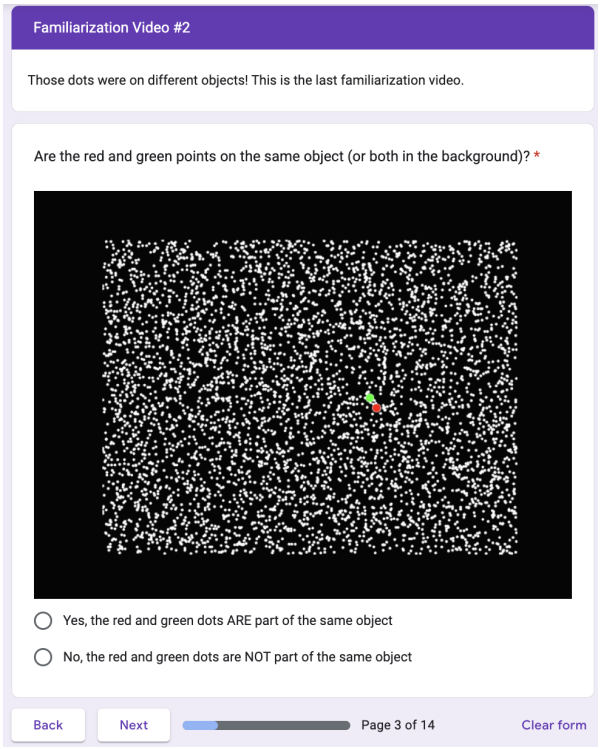
Per-frame Gibbs schedule For each frame $t > 0$, we apply a fixed schedule of blocked Gibbs updates. We apply updates to:

- cluster-level rigid transformations $(\mathbf{R}_k, \mathbf{t}_k)$
- data point-to-particle assignments $z_n^{\mathcal{B}}$, conditioned on position likelihood only (with outliers disabled)
- data point-to-particle assignments $z_n^{\mathcal{B}}$, with full position-velocity-feature likelihood and $p_{\text{outlier}} = 0.1$
- particle spatial means $\boldsymbol{\mu}_\ell^{\mathcal{B}}$ and velocity parameters $(\mathbf{v}_\ell^{\mathcal{B}}, \boldsymbol{\Sigma}_\ell^{\mathcal{V}})$
- particle DINO features $\mathbf{f}_\ell^{\mathcal{B}}$

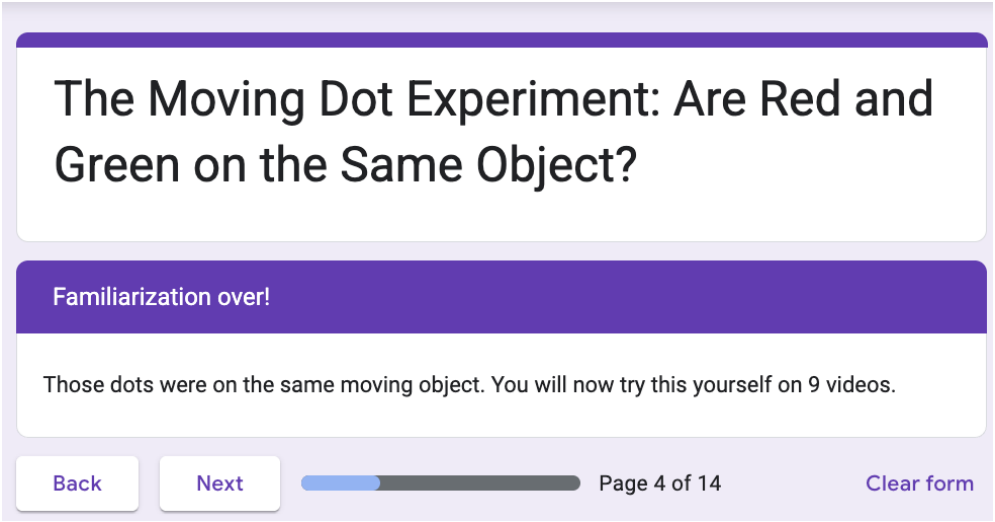
Multiple iterations are performed for spatial and velocity updates to ensure convergence. Mixture weights $\boldsymbol{\pi}^{\mathcal{B}}$ and $\boldsymbol{\pi}^{\mathcal{H}}$ are updated via Dirichlet conditionals after their respective assignment steps.



(a) First familiarization trial



(b) Second familiarization trial with ground truth answer for first familiarization trial revealed.

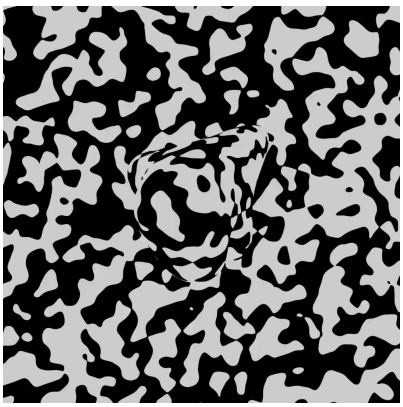


(c) Ground Truth answer for Second Familiarization trial revealed.

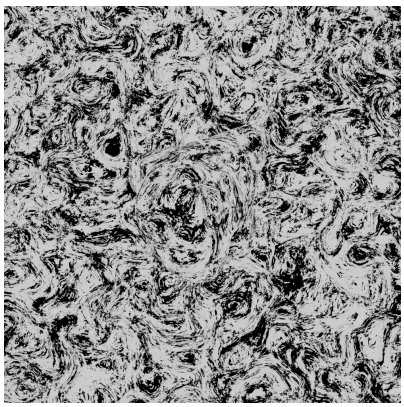
Figure 8. Visual descriptions of the familiarization trials, shown to all 150 participants

Outlier handling During tracking (frames $t > 0$), we enable outliers by including an additional mixture component with weight $p_{\text{outlier}} = 0.1$. The outlier likelihood for a data point with velocity $\mathbf{v}_n$ is modeled as a Gamma distribution on speed $\|\mathbf{v}_n\|$ with shape parameter α and rate parameter β, which accounts for velocity outliers typically arising from unreliable motion estimates at object boundaries.

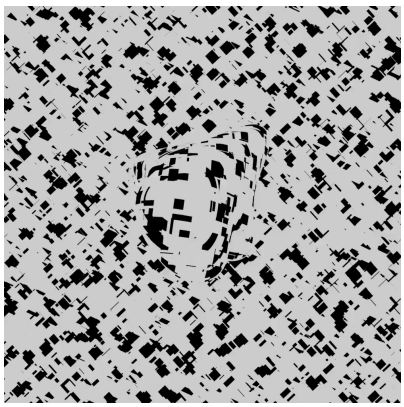
CoTracker3 Setup We run CoTracker3 with its default PyTorch Hub offline mode implementation. We initialize 500 query points in frame 0, matching the particle count used in GenMatter for fair comparison. Query points are randomly sampled uniformly across the first frame. Because we do not use ground-truth segmentation masks during initialization, query points are distributed uniformly across the object and



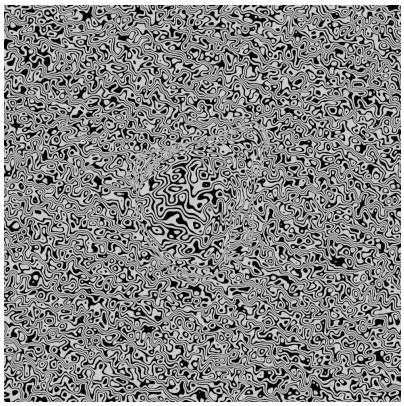
Texture 00



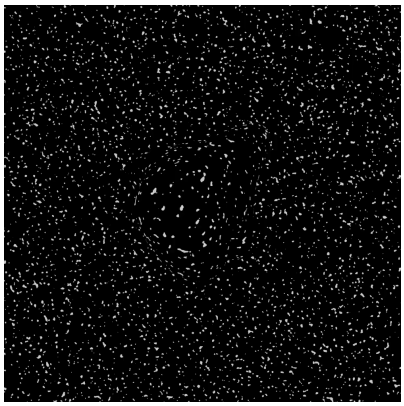
Texture 07



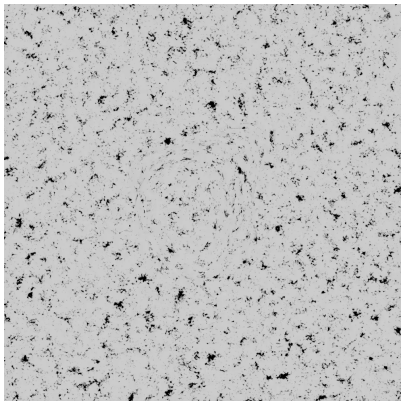
Texture 13



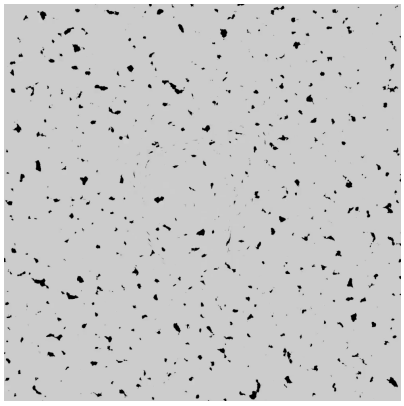
Texture 16



Texture 21



Texture 22



Texture 25

Figure 9. **Example Gestalt Stimuli with Different Textures.** First frame of scene 00000 rendered with seven different texture patterns. The Gestalt experiment uses 20 scenes (00000–00019), each rendered with these seven textures (00, 07, 13, 16, 21, 22, 25), yielding 140 total stimuli to evaluate structure-from-motion segmentation. In the paper, these textures are referenced sequentially as (00, 01, 02, 03, 04, 05, 06).

background regions. As a result, the tracker cannot concentrate query points on the object, making it difficult to share statistical strength across object particles. We evaluate tracking quality using the same particle-based metrics as GenMatter, with the difference that at each frame, we only consider points which CoTracker3 has identified as visible. We first classify particles as object or background based on their frame-0 location relative to the segmentation mask, and we compute per-frame Jaccard by projecting tracked locations onto ground-truth masks at each timestep. The per-

frame Jaccard indices are averaged to obtain the per-video Jaccard index.

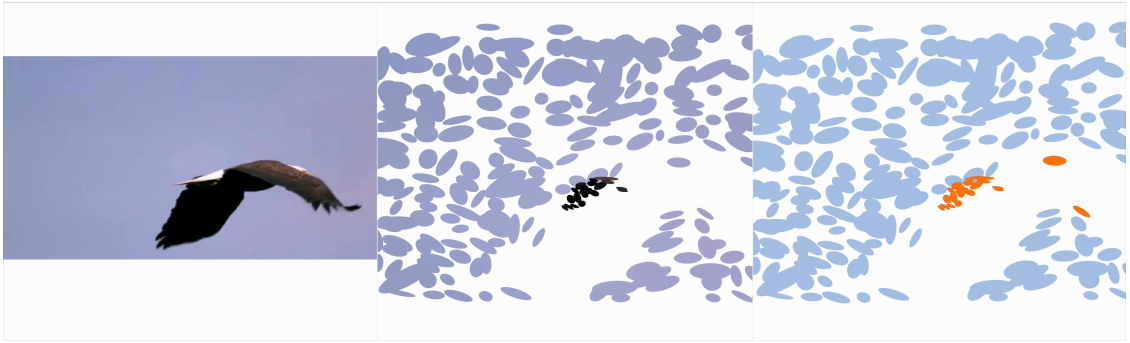
G.2. Supplementary Video Descriptions

We include supplementary videos that visualize the full particle-based inference process across time for the small qualitative deformable dataset introduced in the main paper: `cloth_bag`, `gray_jacket`, `jello`, `manta_ray`, `eagle`, `ostrich`, `purple_jacket`, `snake`, `whiskey_swirl`, and `wine_swirl`. These sequences span a range stuff and things observable in the physical world, including articulated structures, highly deformable solids, and liquids, allowing us to evaluate model performance across the full spectrum of matter interpretable by human vision. The first frame of particle tracking in these videos is shown in Figure 10 , Figure 11, and Figure 12.

Evaluation GenMatter achieves higher average Jaccard (0.83 vs 0.79) on the small qualitative deformable dataset, with strong performance on highly deformable solid matter (`cloth_bag`, `snake`, and `jello` in particular have highly deformable solid matter). It performs weakly on liquid (`wine_swirl` and `whiskey_swirl`) because the appearance of liquid makes it difficult to estimate matter motion, and liquid is particularly unstructured. On the other videos in the set, the performance of both models is similar. This pattern suggests GenMatter's probabilistic particle representation describes highly deformable solid matter better than it describes persistent liquid. The reported Jaccard indices in Table 3 use SAM-generated masks as pseudo-ground-truth, as we do not have ground truth segmentation for these videos. Because GenMatter's initial particle clustering also uses SAM, this evaluation is not as robust as datasets derived from DAVIS. However, visual inspection confirms the SAM-generated masks are accurate, and this evaluation helps us bridge the gap towards more precise evaluation of 3D matter representations.

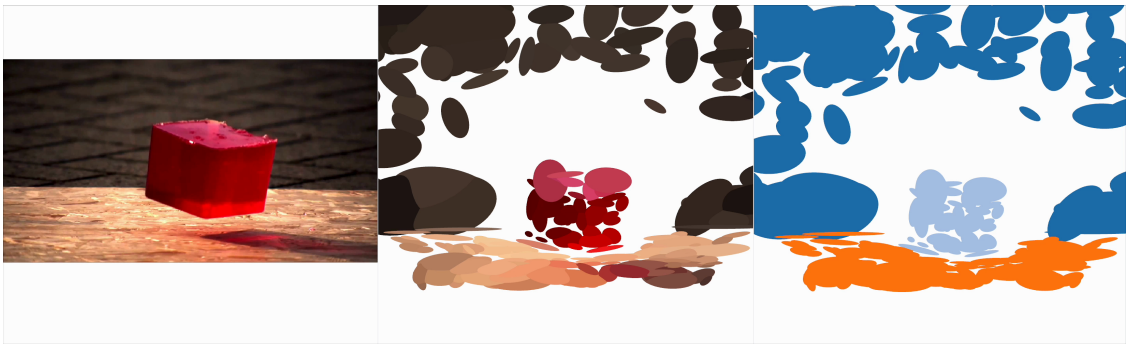
Visualization The supplementary videos apply weight thresholding to particles before rendering. Many particles explain negligible data and have near-zero weights. Our model places little belief in the matter represented by these particles. Removing these low-weight particles improves visual clarity. However, hard thresholding causes flicker at the threshold boundary. Marginal particles will flicker across frames depending on whether their weight exceeds the cut-off. This flicker is a visualization artifact and not model instability.

(a) `cloth_bag`



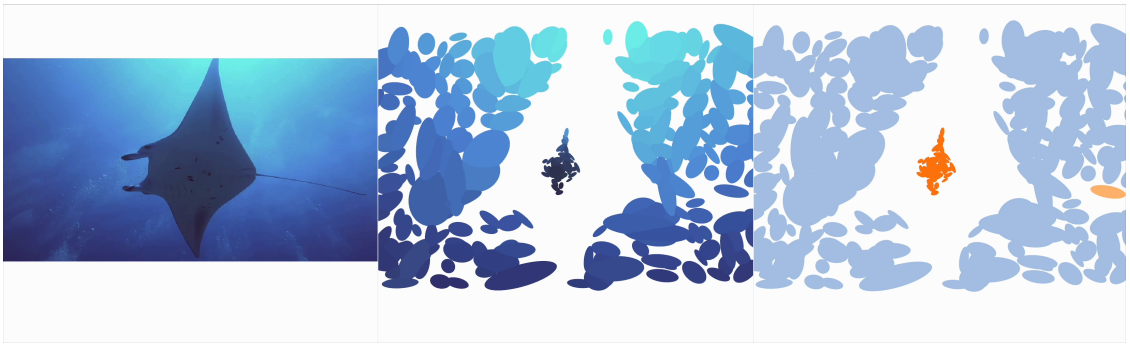
(b) `eagle`

(c) `gray_jacket`

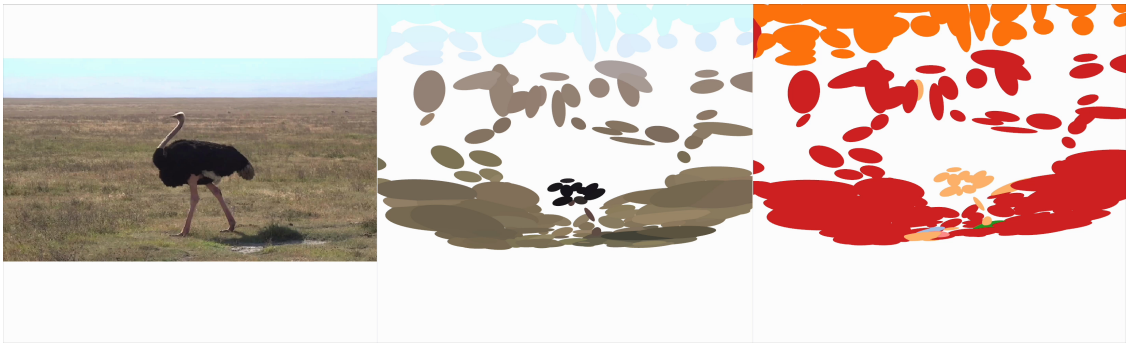


(d) `jello`

Figure 10. First frame visualizations of RGB 3D inference (Part 1). Each image shows the initial particle distribution and segmentation for the respective sequence.



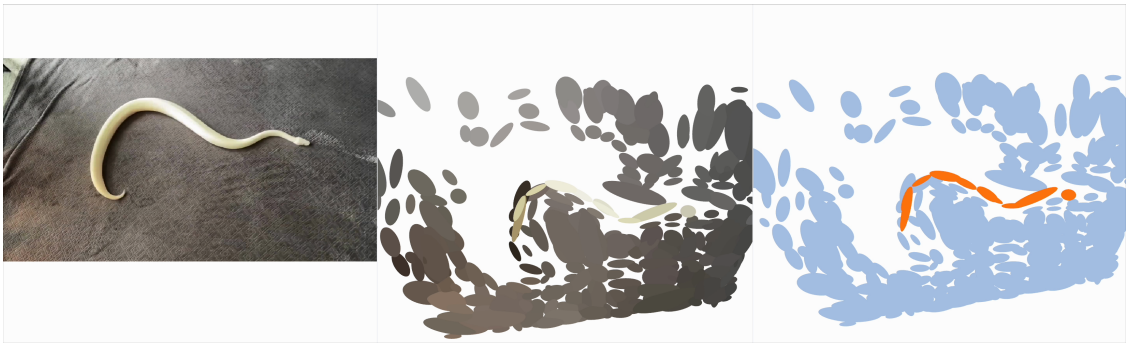
(a) manta_ray



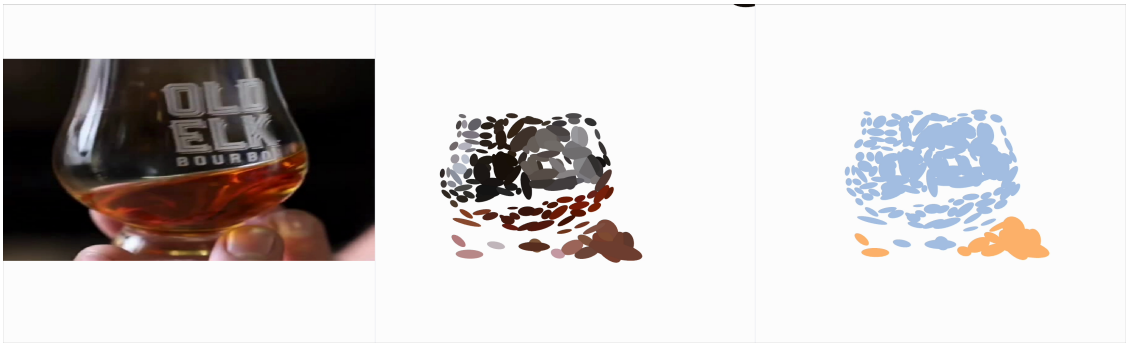
(b) ostrich

(c) purple_jacket

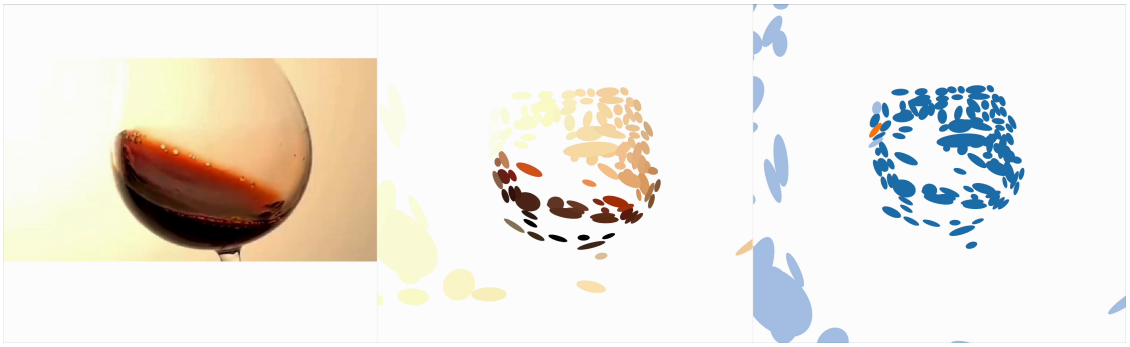
Figure 11. First frame visualizations of RGB 3D inference (Part 2). Each image shows the initial particle distribution and segmentation for the respective sequence.



(a) `snake`



(b) `whiskey_swirl`



(c) `wine_swirl`

Figure 12. First frame visualizations of RGB 3D inference (Part 3). Each image shows the initial particle distribution and segmentation for the respective sequence.