EV-CLOTH: SEEING CLOTH DEFORMATION BETWEEN RGB FRAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Tracking deforming cloth requires estimating the 3D motion of persistent material points, not merely reconstructing a surface in each image. Between RGB frames, however, visual observers must rely on dynamics predictions: increasing the prediction rate does not provide additional evidence about the actual deformation. We introduce EV-Cloth, an event-assisted observer that uses asynchronous visual measurements to correct cloth states during these observation gaps. Its core module, Action-conditioned, Asynchronous, Association-aware Event correction (A^3E), augments a frozen action-conditioned graph dynamics prior. It associates events with the predicted cloth surface, pools their evidence onto persistent material nodes, and estimates a residual correction to the nominal dynamics update. Corrected positions and displacement history are written back into the recurrent state, informing subsequent prediction and event association, while RGB frames periodically refine geometry. On 24 held-out simulated trajectories with approximately 30-Hz RGB observations and a matched 100-Hz state-update schedule, EV-Cloth reduces inter-frame material tracking error by 36.0% relative to an RGB-only observer and improves 22 of 24 trajectories. Controlled studies examine observed event content, material association, and recurrent state correction. The association-by-recurrence interaction persists across three event representations, while timing controls show that reducing recurrent writeback frequency degrades accuracy even when elapsed event evidence is accumulated. These results support event-assisted recurrent correction for persistent cloth-state estimation between RGB observations.

1 INTRODUCTION

Perceiving cloth dynamics requires more than reconstructing a plausible surface in each image. For deformable-object manipulation, an estimator must also track the changing 3D positions of the same material points over time, providing a persistent state representation for downstream action planning (Huang et al., 2022; Longhini et al., 2025). This distinction makes cloth perception a problem of persistent state estimation, rather than independent geometric reconstruction. Graph-based learned simulators provide dynamics priors for propagating material states (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021), while visually grounded cloth methods connect these predictions to image observations (Longhini et al., 2025; Zhan et al., 2026). Together, these approaches support a prediction-and-correction view of cloth-state estimation.

A fundamental limitation, however, is that prediction frequency and observation frequency are not the same. An RGB-based observer can refine its geometry when a frame arrives, but must rely on its dynamics prior between frames. Increasing the rollout rate produces more intermediate predictions, not more evidence about the actual deformation. Rapid local changes can therefore remain unobserved until the next RGB frame, even when the estimated state is updated frequently. Camera frame rate has also been identified as a limitation in real-world visual grounding of cloth dynamics (Zhan et al., 2026). Event cameras offer a complementary source of asynchronous visual evidence during these observation gaps (Gallego et al., 2022). Building on this sensing opportunity, explored in frame–event perception between conventional image frames (Zhang et al., 2023; Gehrig & Scaramuzza, 2024; Wu et al., 2026a), we investigate how events can support persistent 3D cloth state estimation (Fig. 1).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

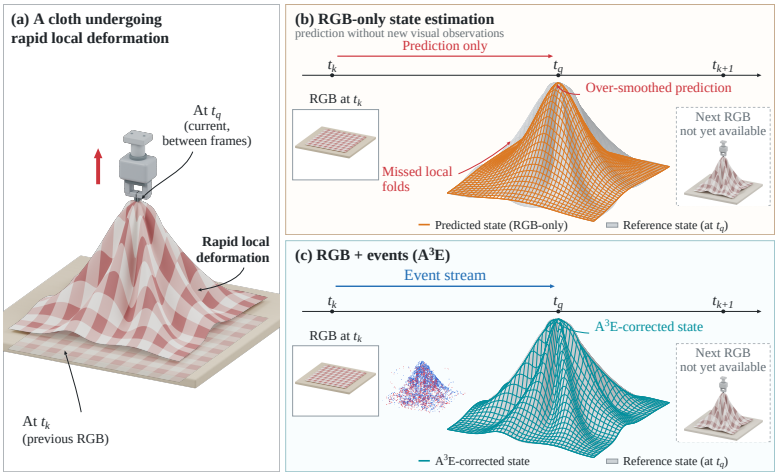

Figure 1: Inter-frame cloth state estimation with RGB and event observations. (a) Rapid local deformation can occur between consecutive RGB frames. (b) During this interval, an RGB-only observer propagates a dynamics prior without new visual evidence and may miss evolving local folds. (c) EV-Cloth uses intervening events to correct the propagated material state through A^{3}E. The illustration is conceptual: all depicted states are synthetic and are not experimental outputs.

The key difficulty is not merely incorporating a higher-rate input, but connecting that input to the state being estimated. Events are image-plane measurements: they do not directly reveal 3D positions or identify the cloth material points that generated them. Their association with material points must instead be inferred through a predicted surface whose projection changes with deformation. Moreover, a correction can either modify only the current output or become part of the state used for subsequent prediction. These choices are coupled: predicted geometry determines where event evidence is assigned, while a correction written back into the state changes the geometry used for future prediction and association. Effective inter-frame estimation therefore requires considering material association and recurrent correction together.

We introduce EV-Cloth, an event-assisted observer built around A^{3}E (Action-conditioned, Asynchronous, Association-aware Event correction). Rather than replacing the dynamics model, A^{3}E augments a frozen action-conditioned graph dynamics prior with observation-driven corrections. It associates incoming events with the predicted cloth surface, pools their evidence onto persistent material nodes, and predicts a centered graph residual to the nominal dynamics update. The corrected positions and displacement history are then written back into the recurrent state, making them available to subsequent dynamics prediction and event association. Incoming RGB frames periodically refine geometry using the same procedure as the RGB-only observer (Fig. 2). This design uses events to inform the evolution of a persistent material state, rather than treating each inter-frame estimate as an isolated output.

We evaluate EV-Cloth on 24 held-out simulated trajectories using approximately 30-Hz RGB observations and a shared 100-Hz state-update schedule. Under this matched schedule, EV-Cloth reduces inter-frame material tracking error by 36.0% relative to the RGB-only observer and improves 22 of the 24 trajectories. To examine the mechanisms behind this improvement, we test observed event content, material-local association, and recurrent writeback on the same RGB30 cohort. A matched two-by-two association-by-recurrence study shows that material-local association improves recurrent estimation; in contrast, the current paired analysis does not resolve a Local/Global difference when corrections affect only the current output. This distinction motivates evaluating association within the recurrent observer, rather than only through output-level corrections. Additional experiments examine the association–recurrence relationship across event representations, separate event evidence from recurrent writeback opportunities, and probe longer RGB gaps and specified action-distribution shifts.

Our contributions are threefold:

- We introduce EV-Cloth to incorporate asynchronous event observations between RGB frames, complementing periodic geometry refinement and dynamics-based state propagation.
- We develop A^3E, which uses predicted geometry to associate image-plane events with persistent material nodes and writes observation-driven residual corrections into the state used for subsequent prediction and association.
- We demonstrate a 36.0% reduction in inter-frame material tracking error on held-out simulated trajectories under matched RGB and state-update schedules. Controlled experiments examine event content, association–recurrence dependence across event representations, and recurrent writeback, with additional stress tests on longer RGB gaps and specified action shifts.

2 Related Work

Cloth state estimation and learned dynamics. Cloth and garment perception has been studied through structured 3D reconstruction and correspondence-aware representations from partial observations (Chi & Song, 2021). Graph-based learned simulators model deformable dynamics over particles and meshes (Sanchez-Gonzalez et al., 2020; Pfaff et al., 2021), while cloth-specific methods infer structured visual states and use learned dynamics for manipulation (Lin et al., 2022; Huang et al., 2022). Action-conditioned cloth tracking has also been used to provide self-supervision for cloth reconstruction (Huang et al., 2023). More recently, Cloth-Splatting couples action-conditioned cloth prediction with a mesh-constrained Gaussian observation model for RGB-based state refinement (Longhini et al., 2025), while CloDS learns cloth dynamics from visually grounded geometry (Zhan et al., 2026). EV-Cloth focuses on the observation gaps between RGB frames, using event measurements to correct the persistent material state propagated by a frozen dynamics prior.

Event-based perception. Event-based frame interpolation reconstructs intermediate images between conventional frames (Tulyakov et al., 2021; 2022). Beyond image reconstruction, hybrid frame–event systems have used intervening events to increase the temporal rate or reduce the latency of task-level perception, including high-frame-rate tracking (Zhang et al., 2023), low-latency object detection between image observations (Gehrig & Scaramuzza, 2024), and anytime inter-frame semantic segmentation (Wu et al., 2026a). For deformable objects, events have also been used for dense deformation measurement (Wu et al., 2026b), hybrid frame–event tracking of deformable linear objects (Hartmann et al., 2026), and non-rigid reconstruction from contour events (Xue et al., 2024). EV-Cloth focuses on causal 3D cloth-state estimation between RGB observations: event evidence is associated with persistent material nodes and recurrently updates the state used by subsequent action-conditioned dynamics.

3D tracking and dynamic reconstruction. Three-dimensional point-tracking methods recover spatial correspondences and trajectories from visual observations, including general-scene and deformable-object tracking approaches (Xiao et al., 2025; Zong et al., 2026). Dynamic reconstruction methods recover time-varying scene geometry and appearance using persistent or deformable scene representations, including Gaussian-based and event-assisted approaches (Luiten et al., 2024; Yang et al., 2024; Wu et al., 2024; Xu et al., 2025; Bai et al., 2026). These methods address related aspects of temporal correspondence and dynamic 3D geometry, but can differ from our causal cloth-state interface in sensing modality, temporal access, correspondence assumptions, or output representation. We therefore report them under explicit observation contracts rather than treating them as same-task baselines; the matched RGB-only cloth observer remains our primary controlled comparison.

3 Problem Formulation

Cloth state and observations. We consider causal 3D cloth-state estimation during manipulation, with particular interest in states between consecutive RGB observations. Following mesh-based cloth representations (Longhini et al., 2025; Zhan et al., 2026), we represent the cloth by a material graph

$$G = (V, E), \tag{1}$$

where V contains N persistent material nodes and E denotes fixed material connectivity. The cloth geometry at time t is

$$X_t = [x_{1,t}, \ldots, x_{N,t}]^\top \in \mathbb{R}^{N\times 3}, \tag{2}$$

where node i refers to the same material location throughout deformation. Thus, the estimation target preserves material correspondence over time rather than independently reconstructing a surface at each observation.

RGB frames I_k arrive at timestamps t_k^{rgb}. The asynchronous event stream is

$$\mathcal{E} = \{e_\ell\}, \qquad e_\ell = (u_\ell, v_\ell, \tau_\ell, p_\ell), \tag{3}$$

where (u_ℓ, v_ℓ) is the image location, τ_ℓ the timestamp, and p_ℓ the polarity. Let a_τ denote the manipulation action available at time τ. We further assume calibrated camera geometry Π, known material graph G, and an initialized recurrent observer state $\hat{S}_0$, which contains the initial cloth geometry and the motion history required by the dynamics model.

Causal inter-frame estimation. Let $\mathcal{Q}$ denote a prescribed set of query times. At a query time $t \in \mathcal{Q}$, the causal inputs are

$$I_{\le t} = \{I_k \mid t_k^{\mathrm{rgb}} \le t\}, \qquad \mathcal{E}_{\le t} = \{e_\ell \mid \tau_\ell \le t\}, \qquad A_{\le t} = \{a_\tau \mid \tau \le t\}. \tag{4}$$

The task is to estimate the material-node positions

$$\hat{X}_t = \mathcal{O}\Big(\hat{S}_0, G, \Pi; I_{\le t}, \mathcal{E}_{\le t}, A_{\le t}\Big). \tag{5}$$

At query time t, the observer uses only the initialized state, fixed graph and camera geometry, and RGB, event, and action inputs available by t.

Of particular interest are inter-frame queries satisfying

$$t_k^{\mathrm{rgb}} < t < t_{k+1}^{\mathrm{rgb}}, \tag{6}$$

for which the future RGB frame I_{k+1} is unavailable. The estimator propagates the cloth state while incorporating visual evidence that arrives before the query. After initialization, ground-truth cloth states are used only for training or evaluation and never reset the observer. In our implementation, this causal observer is evaluated on a uniform 10-ms recurrent grid indexed by j; we write X_j and $\hat{X}_j$ for the reference and estimated cloth states at recurrent step j, respectively.

4 METHOD

EV-Cloth tracks persistent material state through frozen dynamics, predicted-surface event association, graph residual correction, recurrent writeback, and periodic RGB geometry rebase (Fig. 2).

4.1 FROZEN ACTION-CONDITIONED DYNAMICS PRIOR

Events provide intermediate visual evidence but not a complete 3D cloth transition. We retain the action-conditioned graph dynamics model as a frozen nominal prior and learn an observation-driven correction. Given positions $\hat{X}_j \in \mathbb{R}^{N\times 3}$, displacement history (d_{j-1}, d_j), action $a_j \in \mathbb{R}^3$, picked-node index i_j^{pick}, and grasp indicator g_j, this prior on the fixed material graph $G = (V, E)$ computes

$$\begin{aligned} H_j &= \mathcal{N}_H\Big(\Psi(d_{j-1}, d_j, a_j, i_j^{\mathrm{pick}}, g_j)\Big), \\ \bar{q}_j &= F_\theta(H_j, \Xi_j; G). \end{aligned} \tag{7}$$

$H_j \in \mathbb{R}^{N\times 11}$ is the normalized history/action context shared by the prior and A^{3}E; $\mathcal{N}_H$ is frozen. Ξ_j contains geometry-dependent edge attributes after action/grasp conditioning (Appendix D). Memory consists of positions and displacement history, without a learned hidden state. The Sec. 4.4 update with $c_j = 0$ gives nominal geometry $\bar{X}_{j+1}$ for event association.

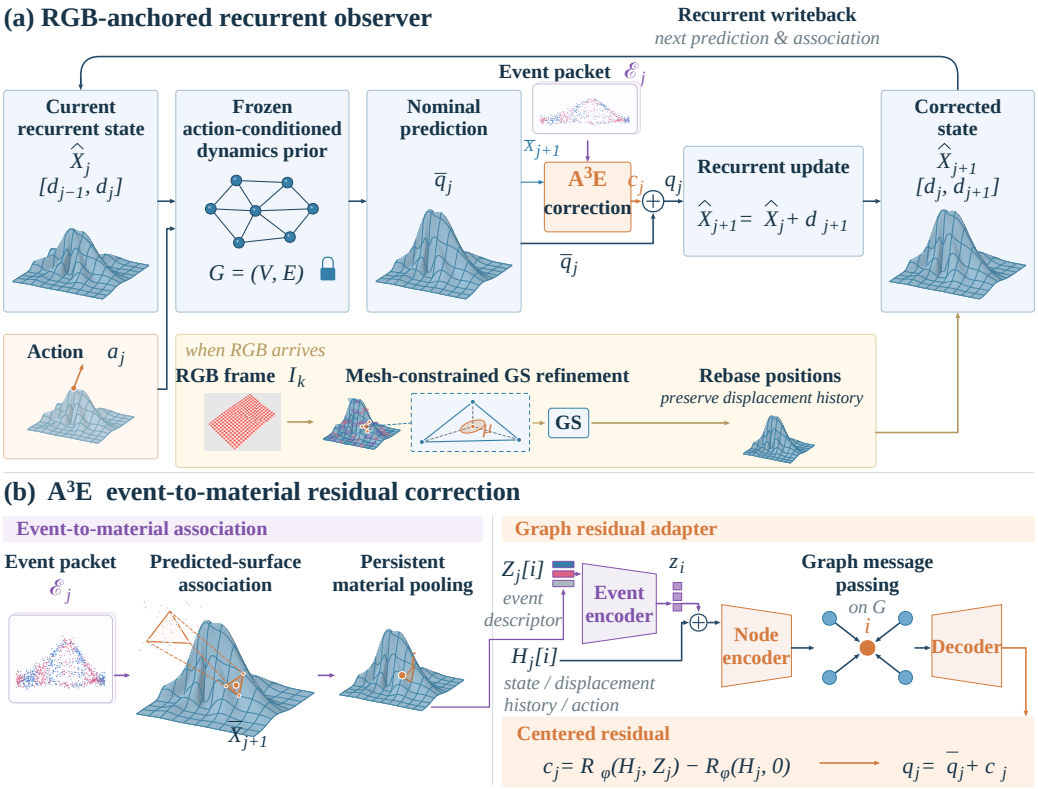

Figure 2: Overview of EV-Cloth and its A^{3}E correction module. (a) The RGB-anchored recurrent observer combines frozen action-conditioned dynamics, event residual correction, recurrent write-back, and RGB-triggered mesh-constrained GS refinement that rebases positions while preserving displacement history. (b) A^{3}E associates the current event packet with the predicted cloth surface, pools evidence onto persistent material nodes, and predicts a centered graph residual correction.

4.2 Event-to-Material Association

The nominal prior predicts cloth motion, but image-plane events lack depth and persistent material identity. Its predicted surface bridges events to material nodes. For each 10-ms transition, the causal packet is $\mathcal{E}_j = \{e_\ell : t_j \le \tau_\ell < t_{j+1}\}$, where $\Delta t_j = t_{j+1} - t_j = 10\,\text{ms}$. A^{3}E updates once per packet while retaining within-packet timestamps. Four temporal bins use interpolated predicted surfaces; $b_\ell \in \{0,1,2,3\}$ indexes a bin and β_b is its normalized midpoint (exact indexing in Appendix D). With fixed sparse-to-full mesh lift L,

$$\widetilde{X}_{j,b} = L(\hat{X}_j) + \beta_b \Big[L(\bar{X}_{j+1}) - L(\hat{X}_j) \Big]. \tag{8}$$

Calibrated projection Π binds event ℓ at image coordinates (u_ℓ, v_ℓ) to this predicted surface:

$$(f_\ell, \lambda_\ell) = \text{ZRasterLookup}\Big(\Pi(\widetilde{X}_{j,b_\ell}), \lfloor u_\ell \rfloor, \lfloor v_\ell \rfloor \Big). \tag{9}$$

The z-buffer returns the front-most face and perspective-correct barycentric coordinates; invalid background or occluded lookups are ignored. Let $\rho_0(v)$ be the fixed frame-0 geodesic Voronoi assignment of full-mesh vertex v to persistent material node i. With event polarity p_ℓ, persistent material pooling is

$$M_{ibp} = \sum_{\ell : b_\ell = b,\, p_\ell = p} \; \sum_{v \in f_\ell} \lambda_{\ell v}\, \mathbf{1}[\rho_0(v) = i], \qquad Z_j[i] = \text{Stats}_{18}(M_i, \ldots). \tag{10}$$

$Z_j[i] \in \mathbb{R}^{18}$ summarizes polarity-aware activity, timing, binding confidence, packet context, and predicted visibility (channels in Appendix D). A naturally empty packet may retain packet and geometry context, unlike the literal zero descriptor used for centering below.

4.3 A^{3}E Graph Residual Correction

The associated material-node evidence is then used to correct the nominal prediction. A^{3}E combines it with shared dynamics context on the material graph to predict per-node residuals to the prior's normalized displacement increment, rather than complete motion from events alone. The event and node encoders compute:

$$z_i = \mathrm{Enc}_e(Z_j[i]), \qquad h_i = \mathrm{Enc}_n([H_j[i], z_i]). \tag{11}$$

A directed message-passing layer propagates these features over G, after which the decoder predicts each node response:

$$\begin{aligned} m_i &= \frac{1}{\max(1, |\mathcal{N}^-(i)|)} \sum_{n:(n,i)\in E} \mathrm{MLP}_m([h_n, h_i]), \\ r_i(H, Z) &= \mathrm{MLP}_d([h_i, m_i]). \end{aligned} \tag{12}$$

Stacking r_i gives $R_\phi(H_j, Z_j) \in \mathbb{R}^{N\times 3}$ (exact widths and counts in Appendix D). Because the adapter also receives H_j, its event-conditioned correction is centered on the same network's literal zero-descriptor response:

$$c_j = R_\phi(H_j, Z_j) - R_\phi(H_j, 0), \qquad q_j = \bar{q}_j + c_j. \tag{13}$$

Both evaluations use the same R_ϕ parameters and H_j; only the event descriptor differs. Literal $Z_j = 0$ gives $c_j = 0$, whereas a naturally empty packet need not produce the zero descriptor.

4.4 Recurrent Update and RGB Rebase

An output-only correction would vanish at the next transition, so A^{3}E writes corrected displacement and position into the physical recurrent state. Let $\mathcal{C}_{a_j}$ action-condition the current picked-node displacement, $\mathcal{B}_{a_j}$ enforce the shared grasp boundary, and $\mathcal{N}_q^{-1}(z) = z \odot \sigma_q + \mu_q$ be the frozen target inverse normalizer. The update is

$$\begin{aligned} d_{j+1} &= \mathcal{B}_{a_j}\left(\mathcal{C}_{a_j}(d_j) + \mathcal{N}_q^{-1}(q_j)\right), \\ \hat{X}_{j+1} &= \hat{X}_j + d_{j+1}. \end{aligned} \tag{14}$$

The state $(\hat{X}_{j+1}, d_j, d_{j+1})$ feeds the next prediction and association; $c_j = 0$ recovers the nominal $\bar{X}_{j+1}$ used in Sec. 4.2.

RGB frames periodically rebase geometry on this propagated state. When frame I_k arrives at boundary $t_j = t_k^{\mathrm{rgb}}$, the incoming event/dynamics transition completes before mesh-constrained rendering-based refinement (Kerbl et al., 2023; Longhini et al., 2025):

$$\hat{X}_j^+ = \mathcal{U}_{\mathrm{rgb}}(\hat{X}_j^-, I_k; \Theta_{\mathrm{app}}), \qquad [d_{j-1}, d_j]^+ = [d_{j-1}, d_j]^-. \tag{15}$$

Appearance is fitted at initialization and held fixed; RGB refinement rebases positions while preserving displacement history (implementation in Appendix D).

4.5 Training Objective

We train A^{3}E through 50-step closed-loop rollouts on its own corrected states using a masked endpoint SmoothL1 objective. Ground-truth positions and displacement history are supplied only at initialization; later steps use the corrected recurrent state. Only the adapter is optimized; the dynamics prior, normalizers, and RGB refinement remain frozen. Appendix A.1 gives the full objective and optimization details.

5 Experiments

We first evaluate whether events improve persistent inter-frame cloth-state estimation under matched RGB observations and update schedules. We then study event content, material association, recurrent correction, and update timing, followed by longer-gap and action-shift stress tests and a qualitative physical study.

5.1 Experimental Setup and Evaluation Protocols

We estimate persistent 3D cloth material states causally on 24 held-out simulated trajectories. RGB arrives at approximately 30 Hz, and both observers advance on the same 100-Hz state-update grid. The matched RGB-only and EV-Cloth observers share initialization, the frozen dynamics prior, actions, RGB observations, and RGB refinement; EV-Cloth additionally uses intervening events for recurrent correction. We report free-node material tracking error (MTE) at queries strictly between RGB frames, averaging seeds within each trajectory and trajectories equally, with paired trajectory-bootstrap intervals. Appendix A gives the full protocol and MTE definition.

5.2 Primary Inter-frame Estimation and Event-Content Controls

On RGB30, EV-Cloth reduces inter-frame MTE from 1.637 to 1.047 mm, a 36.0% reduction, and improves 22 of 24 trajectories. Both observers receive the same RGB observations and operate on the same 100-Hz state-update grid; the gain reflects intervening event observations rather than more frequent predictions. Paired statistics are in Appendix A.

Table 1: Primary RGB30 comparison and event-content controls. Context-only is separately trained; matched-empty removes observed event content at test time from the trained Local recurrent observer. MTE is in mm; lower is better.

Condition	MTE (mm) ↓
RGB-only	1.637
EV-Cloth	**1.047**
Context-only	1.460
Matched-empty	1.180

To test whether observed events contribute beyond packet and predicted-geometry context, we evaluate two controls. Removing observed event content at test time from the trained recurrent observer raises MTE from 1.047 to 1.180 mm; the separately trained Context-only observer reaches 1.460 mm. Thus observed event content contributes beyond contextual features in these comparisons; control definitions are in Appendix B.

Figure 3 shows three held-out inter-frame examples; Appendix C gives an adverse case, and Appendix E reports neighboring-task diagnostics under their respective observation contracts.

5.3 Material Association and Recurrent State Correction

We compare material-local event addressing with global broadcasting to test how association interacts with recurrence. Global retains node-local geometry and visibility; Local and Global share matched training and the frozen prior within each representation. Output-only disables writeback at test time in recurrently trained checkpoints.

Table 2: Association × recurrence across three event representations (MTE in mm). Output-only disables writeback at evaluation time using recurrently trained checkpoints. Each Local/Global pair uses its matched training protocol; the primary purpose of this study is to test whether the association-by-recurrence pattern persists across event representations.

(a) Full18D

	Output-only	Recurrent
Global	1.638	1.532
Local	1.638	**1.047**

(b) Voxel

	Output-only	Recurrent
Global	1.639	1.448
Local	1.638	**1.086**

(c) Time-surface

	Output-only	Recurrent
Global	1.638	1.442
Local	1.637	**1.095**

Global = Global-broadcast; Local = Material-local.

Material-local association is most useful when corrections persist in the recurrent state. For Full18D, it reduces recurrent MTE from 1.532 to 1.047 mm; Local and Global both yield about 1.638 mm with output-only correction. The paired association-by-recurrence interaction is $+0.485$ mm (95%

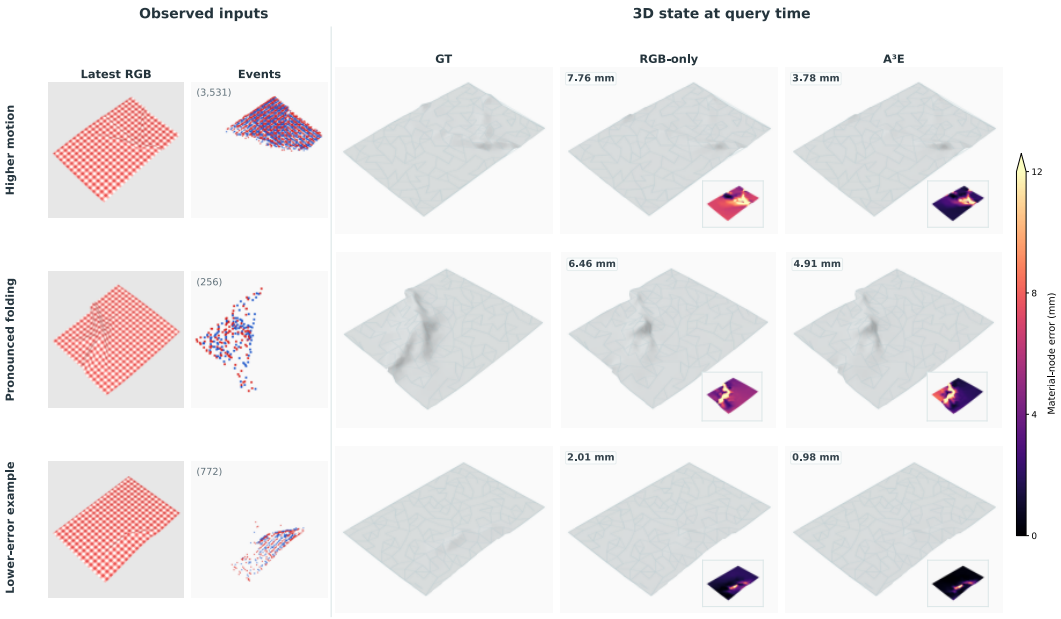

Figure 3: Qualitative inter-frame cloth state estimation. The first two rows show challenging held-out examples under higher motion and pronounced folding; the third is a lower-error held-out example closer to the aggregate evaluation scale. The latest RGB observation and causal events preceding each query appear on the left. GT, RGB-only, and A^{3}E states are rendered from a shared visualization view without per-method alignment; insets show material-node error on a common 0–12 mm scale.

CI $[0.314, 0.678]$). Voxel and Time-surface reproduce positive interactions with paired intervals excluding zero, indicating that the pattern is not specific to Full18D.

Shadow interventions probe which corrected-state pathways carry the benefit forward. Replacing corrected displacement history or dynamics geometry with nominal-shadow counterparts increases error, whereas blocking corrected-state feedback only into future association yields no resolved change. Full paired statistics are in Appendix G.

5.4 Event Evidence and Recurrent Update Opportunities

To distinguish event quantity from recurrent update opportunities, we compare reduced writeback with accumulated elapsed evidence against event thinning at 100-Hz writeback. Dynamics and evaluation remain at 100 Hz.

Table 3: Event evidence and writeback opportunities (MTE in mm). Dynamics and evaluation remain at 100 Hz. Accumulation and thinning alter different aspects of evidence availability and feedback timing, so the rows are not equivalent interventions.

Writeback	Event handling	MTE ↓
100 Hz	Full evidence	**1.047**
50 Hz	Accumulate elapsed-interval evidence	1.200
25 Hz	Accumulate elapsed-interval evidence	1.401
10 Hz	Accumulate elapsed-interval evidence	1.543
100 Hz	Retain 50% of events	1.052
100 Hz	Retain 25% of events	1.067
100 Hz	Retain 10% of events	1.105

With accumulated evidence, error increases monotonically as writeback falls from 100 to 10 Hz, whereas thinning events at 100-Hz writeback causes much smaller degradation. This supports repeated recurrent integration under the tested interventions, but accumulation and thinning alter evidence availability and feedback timing differently and do not isolate writeback frequency alone. At 25 Hz, 95.2% of available evidence is consumed and a final tail remains unconsumed; paired contrasts are in Appendix G.

5.5 Robustness and Scope of Generalization

On a separate 500-ms long-gap protocol, EV-Cloth reduces MTE from 1.570 to 0.942 mm across all 24 trajectories. It remains more accurate than RGB-only under the tested speed and move-hold action shifts, although absolute speed-shift error rises for both methods. These separate stress tests do not establish broad OOD generalization; full results are in Appendices F and G.

5.6 Qualitative Real-World Study

One physical cloth-manipulation recording uses real RGB, event, and robot inputs. Material-point observations are qualitative: without independent 3D material-state ground truth, they establish neither aggregate real-world superiority nor accurate full-3D recovery. Appendix H gives the setup and diagnostics.

6 Discussion and Limitations

Observed event content, material-local association, and recurrent correction jointly support inter-frame estimation. The association benefit is resolved chiefly when corrections persist in the recurrent state; the paired output-only Local/Global comparison remains unresolved. Reduced writeback opportunities degrade accuracy more than event thinning under the tested non-equivalent interventions, consistent with a role for repeated recurrent integration. Shadow interventions further implicate corrected displacement history and dynamics geometry, while blocking corrected-state feedback only into future association yields no resolved additional effect (Appendix G).

Limitations include known initialization, calibrated cameras, fixed topology, and simulation-dominated quantitative evaluation. The grasp-active association study excludes free-flight and non-grasping motion. The primary split is trajectory- and registered-parent-disjoint but does not establish mesh, appearance, physics, or action-family OOD generalization. External comparisons have different observation/correspondence contracts. Fine temporal ordering and low-level lifting choices remain unisolated (Appendix), and 100 Hz denotes the state-update schedule rather than measured end-to-end throughput.

Our quantitative accuracy and mechanism conclusions remain simulation-based. The physical study provides qualitative material-point observations and a secondary diagnostic against a DA3-derived reconstructed 3D pseudo-reference, but no independent full-3D material ground truth. The observations highlight transfer uncertainty under physical sensing, robot execution, cloth dynamics, initialization, calibration, and synchronization. Because these factors are coupled through the recurrent observer, this study does not isolate their individual contributions.

7 Conclusion

We presented EV-Cloth and its A^3E module for persistent cloth-state estimation between RGB frames. Under matched RGB observations and update schedules, A^3E reduces inter-frame MTE by 36.0%, while controlled studies support observed event content, material-local association, and recurrent writeback. These results show that asynchronous events can provide useful intermediate evidence during conventional-frame observation gaps.

AI USE STATEMENT

Generative AI tools, including ChatGPT and Codex, were used to assist with research ideation and execution, including refining hypotheses, providing feedback on experimental design, drafting and debugging research code, and supporting the interpretation of recorded experimental results. These tools were also used for literature retrieval and discovery, as well as for drafting, revising, and polishing portions of the manuscript. Reported numerical results were obtained from executed experiments and saved experimental artifacts rather than generated by language models. The authors reviewed the retained AI-assisted content and take responsibility for the final scientific claims, code, results, citations, and manuscript.

ETHICS STATEMENT

The reported experiments use simulated cloth together with one physical robot–cloth recording. The physical study contains no human-subject measurements. The current evidence does not establish safe physical manipulation or deployment performance.

REPRODUCIBILITY STATEMENT

The appendix specifies protocol identities, statistical units, frozen-model scope, intervention definitions, and observation contracts for the simulation experiments. RGB30 primary results, reused-cohort secondary analyses, sparse-RGB stress, and external comparisons are kept distinct. The real-world study is reported separately as one physical recording evaluated through qualitative material-point behavior and a secondary reconstructed pseudo-reference diagnostic; it is not treated as independent 3D ground-truth validation or evidence of real-time deployment.

REFERENCES

Xiaoyang Bai, Zhenyang Li, Weiwei Xu, Edmund Y. Lam, and Yifan Peng. ERF-GS: Reconstructing fast motion from disjoint event-rgb viewpoints. *arXiv preprint arXiv:2608.08531*, 2026. URL https://arxiv.org/abs/2608.08531.

David Blanco-Mulero, Oriol Barbany, Gokhan Alcan, Adrià Colomé, Carme Torras, and Ville Kyrki. Benchmarking the sim-to-real gap in cloth manipulation. *IEEE Robotics and Automation Letters*, 9(3):2981–2988, 2024. doi: 10.1109/LRA.2024.3360814.

Kaustav Chanda, Aayush Verma, Arpitsinh Vaghela, Yezhou Yang, and Bharatesh Chakravarthi. Event quality score (EQS): Assessing the realism of simulated event camera streams via distance in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5144–5152, 2025. URL https://openaccess.thecvf.com/content/CVPR2025W/EventVision/html/Chanda_Event_Quality_Score_EQS_Assessing_the_Realism_of_Simulated_Event_CVPRW_2025_paper.html.

Cheng Chi and Shuran Song. GarmentNets: Category-level pose estimation for garments via canonical space shape completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3324–3333, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chi_GarmentNets_Category-Level_Pose_Estimation_for_Garments_via_Canonical_Space_Shape_ICCV_2021_paper.html.

Jack Collins, David Howard, and Jürgen Leitner. Quantifying the reality gap in robotic manipulation tasks. In *2019 International Conference on Robotics and Automation*, pp. 6706–6712. IEEE, 2019. doi: 10.1109/ICRA.2019.8793591.

Bardienus P. Duisterhof, Mandi Zhao, Yunchao Yao, Jia-Wei Liu, Jenny Seidenschwarz, Mike Zheng Shou, Deva Ramanan, Shuran Song, Stan Birchfield, Bowen Wen, and Jeffrey Ichnowski. DeformGS: Scene flow in highly deformable scenes for deformable object manipulation. In *Workshop on the Algorithmic Foundations of Robotics*, 2024. URL https://deformgs.github.io/.

Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J. Davison, Jörg Conradt, Kostas Daniilidis, and Davide Scaramuzza. Event-based vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):154–180, 2022. doi: 10.1109/TPAMI.2020.3008413.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629:1034–1040, 2024. doi: 10.1038/s41586-024-07409-w.

Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5633–5643, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Gehrig_End-to-End_Learning_of_Representations_for_Asynchronous_Event-Based_Data_ICCV_2019_paper.html.

Annalena Hartmann, Priyamvada Ajithkumar, Patrick Bründl, and Jörg Franke. MotionDLO: Hybrid event- and frame-based tracking of deformable linear objects. *arXiv preprint arXiv:2608.22398*, 2026. URL https://arxiv.org/abs/2608.22398.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1312–1321, 2021. URL https://openaccess.thecvf.com/content/CVPR2021W/EventVision/html/Hu_v2e_From_Video_Frames_to_Realistic_DVS_Events_CVPRW_2021_paper.html.

Zixuan Huang, Xingyu Lin, and David Held. Mesh-based dynamics with occlusion reasoning for cloth manipulation. In *Proceedings of Robotics: Science and Systems*, 2022. doi: 10.15607/RSS.2022.XVIII.011. URL https://www.roboticsproceedings.org/rss18/p011.html.

Zixuan Huang, Xingyu Lin, and David Held. Self-supervised cloth reconstruction via action-conditioned cloth tracking. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. URL https://arxiv.org/abs/2302.09502.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman. HOTS: A hierarchy of event-based time-surfaces for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1346–1359, 2017. doi: 10.1109/TPAMI.2016.2574707.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. URL https://arxiv.org/abs/2511.10647.

Xingyu Lin, Yufei Wang, Zixuan Huang, and David Held. Learning visible connectivity dynamics for cloth smoothing. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pp. 256–266, 2022. URL https://proceedings.mlr.press/v164/lin22a.html.

Alberta Longhini, Marcel Büsching, Bardienus Pieter Duisterhof, Jens Lundell, Jeffrey Ichnowski, Mårten Björkman, and Danica Kragic. Cloth-splatting: 3d cloth state estimation from rgb supervision. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2845–2865, 2025. URL https://proceedings.mlr.press/v270/longhini25a.html.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *Proceedings of the International Conference on 3D Vision (3DV)*, pp. 800–809, 2024. URL https://dynamic3dgaussians.github.io/.

Miles Macklin, Matthias Müller, Nuttapong Chentanez, and Tae-Yong Kim. Unified particle physics for real-time applications. *ACM Transactions on Graphics*, 33(4):153:1–153:12, 2014. doi: 10.1145/2601097.2601152.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=roNqYL0_XP.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8459–8468, 2020. URL https://proceedings.mlr.press/v119/sanchez-gonzalez20a.html.

Timo Stoffregen, Cedric Scheerlinck, Davide Scaramuzza, Tom Drummond, Nick Barnes, Lindsay Kleeman, and Robert Mahony. Reducing the sim-to-real gap for event cameras. In *Computer Vision – ECCV 2020*, volume 12372 of *Lecture Notes in Computer Science*, pp. 534–549. Springer, 2020. doi: 10.1007/978-3-030-58583-9_32.

Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, and Davide Scaramuzza. Time lens: Event-based video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16155–16164, June 2021.

Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, and Davide Scaramuzza. Time lens++: Event-based frame interpolation with parametric non-linear flow and multi-scale fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17755–17764, June 2022.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20310–20320, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_4D_Gaussian_Splatting_for_Real-Time_Dynamic_Scene_Rendering_CVPR_2024_paper.html.

Xiaoshan Wu, Xiaoyang Lyu, Yifei Yu, Bo Wang, Zhongrui Wang, and Xiaojuan Qi. LiFR-Seg: Anytime high-frame-rate segmentation via event-guided propagation. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/620221d506055afba5ee7ab895277233-Abstract-Conference.html.

Yuliang Wu, Wei Zhai, Yuxin Cui, Tiesong Zhao, Yang Cao, and Zheng-Jun Zha. Event-based visual deformation measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 903–913, June 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_Event-based_Visual_Deformation_Measurement_CVPR_2026_paper.html.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3d point tracking made easy. *arXiv preprint arXiv:2507.12462*, 2025. URL https://arxiv.org/abs/2507.12462.

Wenhao Xu, Wenming Weng, Yueyi Zhang, Ruikang Xu, and Zhiwei Xiong. Event-boosted deformable 3d gaussians for dynamic scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28334–28343, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_Event-boosted_Deformable_3D_Gaussians_for_Dynamic_Scene_Reconstruction_ICCV_2025_paper.html.

Yuxuan Xue, Haolong Li, Stefan Leutenegger, and Jörg Stückler. Event-based non-rigid reconstruction of low-rank parametrized deformations from contours. *International Journal of Computer Vision*, 132:2943–2961, 2024. doi: 10.1007/s11263-024-02011-z. URL https://doi.org/10.1007/s11263-024-02011-z.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20331–20341, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yang_Deformable_3D_Gaussians_for_High-Fidelity_Monocular_Dynamic_Scene_Reconstruction_CVPR_2024_paper.html.

Yuliang Zhan, Jian Li, Wenbing Huang, Yang Liu, and Hao Sun. CloDS: Visual-only unsupervised cloth dynamics learning in unknown conditions. *arXiv preprint arXiv:2602.01844*, 2026. URL https://arxiv.org/abs/2602.01844.

Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, and Xin Yang. Frame-event alignment and fusion network for high frame rate tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9781–9790, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Zhang_Frame-Event_Alignment_and_Fusion_Network_for_High_Frame_Rate_Tracking_CVPR_2023_paper.pdf.

Yeheng Zong, Yizhou Chen, Alexander Bowler, Chia-Tung Yang, and Ram Vasudevan. TrackDeform3D: Markerless and autonomous 3d keypoint tracking and dataset collection for deformable objects. *arXiv preprint arXiv:2603.17068*, 2026. URL https://arxiv.org/abs/2603.17068.

A PRIMARY RGB30 PROTOCOL, TRAINING, AND REPRODUCIBILITY

Primary cohort and evaluation. The primary cohort is `EVCloth_RGB30_FRESH_CONFIRM24_R1`: 24 held-out simulated trajectories, three predetermined A^3E seeds (1234/2024/4321), one deterministic RGB-only stream, 200 persistent material nodes with 199 free nodes, 174 strict inter-frame queries per trajectory, and 250 native 10-ms updates after bootstrap. The target RGB schedule is aligned to legal 10-ms states, yielding 30- or 40-ms RGB intervals and an effective rate of 30.24 Hz. Evaluation uses a known-start initialization: one ground-truth cloth mesh state at t_0 initializes material geometry; RGB observations at the first two bootstrap timestamps determine the initial displacement, which is copied into both displacement-history slots. RGB-only and EV-Cloth use the identical bootstrap, and no later ground-truth geometry enters the recurrent observer. Prediction, state updates, and evaluation remain on the same 100-Hz grid. Sensor events retain their asynchronous timestamps; 100 Hz denotes the packet/state/correction schedule rather than sensor FPS or measured end-to-end throughput.

Queries are averaged within each trajectory and trajectories receive equal weight. The three predetermined A^3E seed metrics are averaged within trajectory before cohort aggregation; the deterministic RGB-only observer is not duplicated across seeds. Paired uncertainty is estimated with 10,000 trajectory-level bootstrap resamples. The primary A^3E–RGB-only contrast is -0.5901 mm with paired 95% CI $[-0.8016, -0.4028]$ mm and 22/2/0 trajectory wins/losses/ties, corresponding to a 36.0% reduction in cohort-mean MTE.

The 24 RGB30 evaluation trajectories are disjoint by registered physical trajectory and registered parent from the frozen dynamics prior's training and development cohorts (including earlier prior-pretraining development cohorts), the normalizer-fitting cohort, and the A^3E adapter-training set. Input and target normalizers are fit on training trajectories only; no RGB30 test result is used for checkpoint selection. Towel meshes are shared across training and evaluation, so this protocol evaluates held-out trajectories rather than mesh-OOD generalization. It does not establish texture-, physics-, or action-family OOD generalization.

Frozen model scope. The RGB30 evaluation loads the fixed step-2,000 S0 dynamics-prior checkpoint. The action-conditioned dynamics prior, its input/target normalizers, and the three primary A^3E checkpoints are frozen throughout the RGB30 primary evaluation and current same-cohort controls. The matched Local/Global association study uses paired adapters trained with the same frozen prior, normalizers, training windows, optimization budget, initialization policy, and fixed final checkpoints.

A.1 RGB30 ADAPTER TRAINING

The primary adapters use the frozen A2RGNS100Hz dynamics prior and its fixed input/target normalizers. The frozen S0 prior and A^3E adapters use the same 24-trajectory training lineage, disjoint from RGB30 test24; the adapter registry contains 1,798 unique $K = 50$ training windows from these trajectories. Three independently initialized adapters use seeds 1234, 2024, and 4321 and the same fixed 2,000-step optimization schedule. Each optimization example rolls the observer closed loop for $K = 50$ native 10-ms transitions: oracle positions and the two-displacement history are provided only at rollout initialization; all later prediction, association, correction, and recurrence steps consume the branch's own corrected state. No intermediate ground-truth state or teacher forcing is used.

For endpoint valid-node mask $\mathcal{F}_K$, excluding the active picked node when grasped, the endpoint position objective normalizes each coordinate error by 0.01 and applies $\mathrm{SmoothL1}_{\beta=1}$:

$$\mathcal{L}(\phi) = \frac{1}{3|\mathcal{F}_K|} \sum_{i \in \mathcal{F}_K} \sum_{\nu=1}^{3} \mathrm{SmoothL1}_{\beta=1}\left(\frac{\hat{X}_{K,i,\nu} - X_{K,i,\nu}}{0.01} \right), \qquad K = 50. \tag{16}$$

Adam uses learning rate 10^{-4}, betas $(0.9, 0.999)$, epsilon 10^{-8}, zero weight decay, and batch size one. The final checkpoint is fixed at step 2,000 for every seed, without DEV-based checkpoint selection. The frozen dynamics prior, its normalizers, and the RGB refinement module are not optimized with the adapter.

Evaluation	Cohort	RGB observations	Role
RGB30 primary	24 held-out trajectories	$\sim$30 Hz	Primary evaluation
RGB30 mechanisms	same 24 trajectories	$\sim$30 Hz	Same-cohort controls
Long-gap stress	separate 24 trajectories	2 Hz	Secondary stress test

Table S1: Evaluation protocols retained in the submission. The RGB30 primary and mechanism analyses share the same held-out cohort and 100-Hz state-update grid. The long-gap experiment is a separate stress test and is not used for a controlled frame-rate comparison with RGB30.

A.2 Simulation, Rendering, and Event Generation

Simulation and actions. We use NVIDIA FleX through PyFlex (Macklin et al., 2014) with an outer step of 0.01 s and eight solver substeps. The dataset contains ten full-resolution meshes from the TOWEL family; the observer tracks 200 selected persistent material nodes and does not replace the full physics/render mesh with a 200-vertex mesh. The PyFlex binding and TOWEL mesh generator are taken from the pinned synthetic-cloth-data codebase; the ten TOWEL meshes used here are generated locally from that generator. Pick and place use distinct particles and a constant-arc-length quadratic Bézier trajectory sampled at 10-ms action intervals.

Rendering. Blender 3.4.1 EEVEE renders one fixed view at 800×800 resolution with a 50° horizontal field of view. Source frames are timestamped exactly at 100 Hz, with no motion blur and no crop or resize.

Events and packetization. Events are generated with v2e (Hu et al., 2021) using the pinned float64 event-time patch and no SuperSloMo. Positive and negative contrast thresholds are both 0.2, threshold sigma is 0.02, cutoff, leak, and shot-noise rates are 0 Hz, and the refractory period is zero. Polarity is stored as 0 for negative and 1 for positive. HDF5 columns are $[t_{\mathrm{us}}, x, y, p]$, where t_{us} is an integer timestamp in microseconds. Events are packetized as $[t_j, t_{j+1})$ in 10-ms intervals, with one A^{3}E update per packet. The RGB30 alignment has legal regular RGB intervals of 30 or 40 ms. Reported metric distances follow the recorded simulator world-unit-to-metre convention; independent physical calibration of that convention is not established.

Full MTE definition. For trajectory r, adapter seed s, strict inter-frame query set $\mathcal{Q}_r$, and free-node set $\mathcal{F}_r$, the metric used in the main text is

$$e_{r,s} = \frac{1000}{|\mathcal{Q}_r||\mathcal{F}_r|} \sum_{q\in\mathcal{Q}_r} \sum_{i\in\mathcal{F}_r} \left\| \hat{X}_{r,s,q,i} - X_{r,q,i} \right\|_2, \quad e_r = \mathrm{mean}_s\, e_{r,s}, \quad \mathrm{MTE} = \frac{1}{R}\sum_{r=1}^{R} e_r. \quad (17)$$

B Extended RGB30 Results

The three seed-level interior means are 1.0354, 1.0106, and 1.0954 mm for seeds 1234, 2024, and 4321. The primary statistical unit remains the trajectory. PRE and POST are reported as separate diagnostics around RGB geometry refinement and are not part of the strict inter-frame metric.

Event content and learned context. The matched association study uses three seeds averaged within each trajectory, 24 equally weighted trajectories, and 10,000 paired trajectory-bootstrap resamples. Queries, nodes, and seed-level cells are not treated as independent statistical units. Local-V2 reproduces the original Full training tensors and losses under the matched training protocol.

Matched-empty uses the Local recurrent checkpoint with observed event-content channels zeroed while retaining the legal packet and geometry context formed from that branch's own predicted state. It therefore tests removal of observed event content from a trained event-assisted observer; it is not a separately trained no-event model. A separately trained RGB30 Context-only capacity/context diagnostic obtains 1.4596 mm, while Full A^{3}E obtains 1.0471 mm (Context-only$-$Full $= +0.4125$ mm, paired 95% CI $[0.1836, 0.6970]$). These controls answer different questions and are not interpreted as an additive decomposition of the 36.0% system-level gain.

Table S2: Full-precision RGB30 primary and phase results. Three frozen A^3E seed metrics are averaged within each of 24 trajectories; trajectories receive equal weight. The primary statistical unit remains the trajectory.

Method	Interior MTE (mm)	PRE	POST
RGB-only	1.637239850303122	1.6795083963212887	1.679611001931333
A^3E	1.0471100607539228	1.0705291672720687	1.0735980714607394

Primary paired statistic	Value
A^3E-RGB MTE (mm)	-0.590129789549199
Relative reduction (%)	36.044186772020076
Paired 95% CI (mm)	$[-0.8015720977642954, -0.40280102776221866]$
Trajectory W/L/T	22/2/0

Phase after RGB	RGB-only	A^3E	A^3E-RGB
10 ms	1.618765	1.036392	-0.582373
20 ms	1.653895	1.057079	-0.596816
30 ms	1.643367	1.049757	-0.593610

Table S3: Full-precision matched RGB30 association-by-recurrence results. Local and Global share the fresh 24-trajectory cohort, three seeds, frozen prior and normalizers, event information, node context, optimizer, training windows, initialization, and recurrent update; only the event address changes. Positive contrasts indicate larger error for the first named condition.

(a) Absolute interior MTE

Variant	MTE (mm)
Local recurrent	1.0471144244978394
Global recurrent	1.5315727687031711
Local output-only	1.6382473106714306
Global output-only	1.6378561810800771
Local matched-empty recurrent	1.1796295330814008

(b) Paired trajectory-bootstrap contrasts

Contrast	Δ MTE (mm)	95% CI (mm)	W/L/T
Global recurrent − Local recurrent	+0.48445834420533179	[0.3141193916658703, 0.6769244309098345]	21/3/0
Global output-only − Local output-only	-0.00039112959135342851	[-0.0026373148949116352, 0.0010934383782543328]	—
Local output-only − Local recurrent	+0.59113288617359105	[0.4046895499256501, 0.8022931178028448]	—
Global output-only − Global recurrent	+0.10628341237690582	[-0.09184678067481453, 0.31542899132490587]	—
Association × recurrence interaction	+0.48484947379668525	[0.3143635069860132, 0.6775334917367555]	—
Matched-empty − Local recurrent	+0.13251510858356144	[0.032692708378975845, 0.2562754310260198]	—

Material association and recurrence. The Local/Global study matches trajectories, frozen prior and normalizers, training windows, optimizer, node context, initialization, event information, and recurrent update while changing how observed event evidence is addressed to material nodes. Material-local association improves recurrent estimation, whereas the paired Local/Global contrast is unresolved in the output-only setting. The corresponding association-by-recurrence interaction is reported in Table S3.

C Representative Adverse Case

Figure S1 documents an individual query where event correction increases error, despite the favorable aggregate trajectory result.

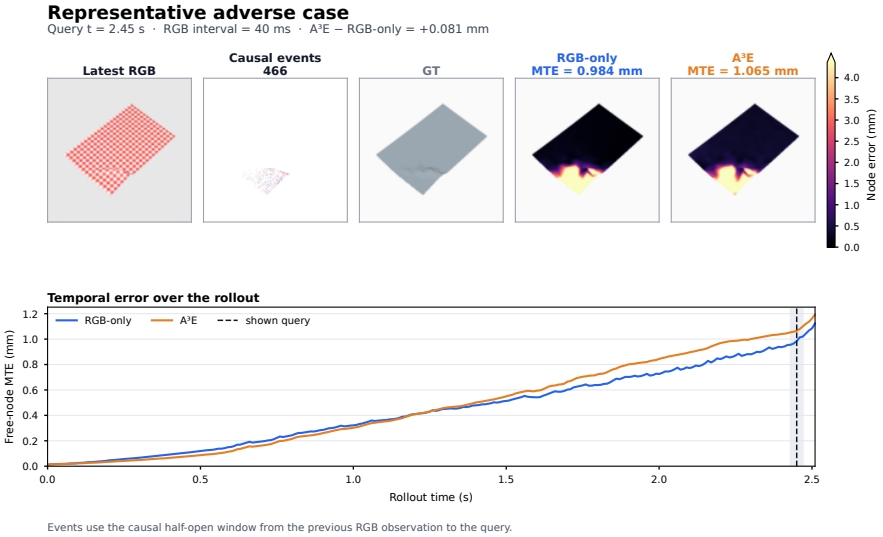

Figure S1: Representative adverse RGB30 case. At the selected query, RGB-only obtains 0.984 mm MTE while A^{3}E obtains 1.065 mm, corresponding to a $+0.081$ mm A^{3}E$-$RGB-only difference. The lower panel shows temporal error over the rollout.

D A3E Implementation and Reproducibility Details

State/action context. The adapter consumes $H_j \in \mathbb{R}^{N\times 11}$ and $Z_j \in \mathbb{R}^{N\times 18}$ and returns an $N\times 3$ normalized displacement-increment response. Before the fixed coordinatewise normalization, the 11 ordered history/action channels are previous displacement xyz, grasp-conditioned broadcast action xyz, action-conditioned current displacement xyz, and a two-channel free/picked one-hot type: $3+3+3+2=11$.

The prior forms $\Xi_j = \Gamma(\hat{X}_j, a_j, i_j^{\text{pick}}, g_j; G)$ after applying the arrived action and grasp condition to the picked node. Each directed edge $(n,i)\in E$ uses $\Xi_{j,ni} = [x_i - x_n, \|x_i - x_n\|_2]$ from these conditioned positions. The action thereby affects H_j, geometry, and the shared grasp boundary without an additional learned input to F_θ.

Event descriptor. For $e_\ell \in \mathcal{E}_j$ with $\Delta t_j = t_{j+1} - t_j$, the exact four-bin association timing and indexing are

$$\alpha_\ell = \frac{\tau_\ell - t_j}{\Delta t_j}, \qquad b_\ell = \lfloor 4\alpha_\ell \rfloor, \qquad \beta_b = \frac{b+1/2}{4}, \quad b\in\{0,1,2,3\}. \tag{18}$$

Let $m_{ibp} = M_{ibp}$, $m_i^p = \sum_b m_{ibp}$, $m_i = \sum_p m_i^p$, and $\eta_{ibp} = m_{ibp}/(\Delta t_j \max(\mathcal{A}_{ib}, \epsilon))$. With $w_{\ell i} = \sum_{v\in f_\ell} \lambda_{\ell v} \mathbf{1}[\rho_0(v) = i]$, the timestamp moments are $\bar{\alpha}_i = m_i^{-1}\sum_\ell w_{\ell i}\alpha_\ell$ and $s^2_{\alpha,i} = m_i^{-1}\sum_\ell w_{\ell i}(\alpha_\ell - \bar{\alpha}_i)^2$, set to zero when $m_i = 0$. Let κ_i be the maximum valid binding confidence and ν_i the predicted visible-area fraction for node i. The ordered channels are:

Alternative event representations. For the representation-robustness study in Sec. 5.3, Voxel uses a local four-bin, polarity-separated log-count event grid from the event-grid representation family (Gehrig et al., 2019). Time-surface uses a local two-polarity-channel exponential-recency representation from the time-surface family (Lagorce et al., 2017); it is not an implementation of the full HOTS hierarchy.

Matched Local and Global association controls. Local-V2 is bitwise equivalent to the primary observer's Full18D construction. Global-V2 aggregates observed event sufficient statistics over material nodes and broadcasts them while retaining the same node-local predicted geometry/visibility context. The prior context H_j, graph, actions, history, frozen prior and normalizers, optimizer, training windows, step-zero initialization, and recurrent update are otherwise unchanged.

Channels	Raw quantity	Units
0–2	Previous displacement $(d_{j-1,x}, d_{j-1,y}, d_{j-1,z})$	world unit
3–5	Grasp-active broadcast $(a_{j,x}, a_{j,y}, a_{j,z})$	world unit
6–8	Action-conditioned current displacement $(\widetilde{d}_{j,x}, \widetilde{d}_{j,y}, \widetilde{d}_{j,z})$	world unit
9	Free-node indicator	binary
10	Picked-node indicator	binary

Table S4: Exact raw channel order of the 11-D prior context before the fixed coordinatewise normalizer $\mathcal{N}_H$.

Channels	Descriptor quantity	Units
0–7	η_{ibp}, four bins $b = 0{:}3$ by polarities $p = 0, 1$	event-equiv./(area s)
8	$m_i^+/\Delta t_j$ (aggregate positive rate)	event-equiv./s
9	$m_i^-/\Delta t_j$ (aggregate negative rate)	event-equiv./s
10	$m_i/\Delta t_j$ (aggregate total rate)	event-equiv./s
11	$\log(1 + m_i/\Delta t_j)$	dimensionless
12	$\bar{\alpha}_i$ (normalized timestamp mean)	dimensionless
13	$s^2_{\alpha,i}$ (normalized timestamp variance)	dimensionless
14	κ_i (binding confidence)	dimensionless
15	Δt_j (packet duration)	s
16	ν_i (predicted visible-area fraction)	fraction
17	m_i (raw bound event mass)	event-equivalent

Table S5: Exact order and definition of the 18-D material-node event descriptor before LayerNorm(18). Polarity 0/1 denotes negative/positive.

Graph residual adapter. The event encoder is LayerNorm(18), Linear(18,64), ReLU, Linear(64,32). Its 32-D output is concatenated with the 11-D context. The node encoder is Linear(43,128), ReLU, Linear(128,128). Graph connectivity supplies directed edge indices. For each edge, the message MLP applies Linear(256,128), ReLU, Linear(128,128) to concatenated sender/receiver features. Incoming messages are summed per target and divided by the target in-degree clamped below by one, yielding mean incoming aggregation. There is one message-passing stage per R_ϕ forward. The decoder applies ReLU, Linear(256,128), ReLU, Linear(128,3) to the concatenated node and mean-message features.

All Linear layers use bias. LayerNorm on the raw 18-D descriptor is the adapter's only normalization layer. There is no dropout, BatchNorm, GRU, independent learned edge-feature encoder, or persistent learned hidden state. The adapter has 108,167 trainable parameters: 3,332 in the event encoder, 22,144 in the node encoder, 49,408 in the message MLP, and 33,283 in the decoder.

Centering and recurrent state. Centering evaluates the same module and parameters twice as $R_\phi(H_j, Z_j) - R_\phi(H_j, 0)$; the reference receives an exact all-zero descriptor before encoding. In RGB30, this correction augments the normalized nominal increment; frozen target inverse normalization then precedes recurrent displacement and position writeback. A naturally empty packet may retain legal packet-duration and predicted-visibility context, unlike the literal zero reference descriptor.

RGB geometry refinement. The paper-facing operator $\mathcal{U}_{\mathrm{rgb}}$ uses the audited rendering-based solver with frozen appearance Θ_{app}. Appearance is fitted at the initial RGB frame (1,500 iterations in the inherited recipe) and then held fixed. The image objective uses the inherited L1/DSSIM terms, with DSSIM implemented as 1-SSIM (Wang et al., 2004; Longhini et al., 2025), and rigidity regularization with coefficients 1.0, 0.1, and 0.3. Geometry refinement permits at most 50 normalized-gradient proposals, with a maximum node step of 10^{-4} recorded world units and five halved backtracking trials. A trial must decrease the image objective and preserve finite, nondegenerate triangles; the picked node is fixed. RGB refinement rebases positions while preserving the stored two-displacement history.

E External Neighboring-Task Diagnostics

We report common-frame tracking and offline reconstruction as neighboring-task diagnostics with different observation and correspondence contracts, rather than a same-task ranking against the primary causal RGB30 observer.

(a) Common-frame tracking

Method	Input	Coverage	3D MTE ↓
Spatial TrackerV2	RGB	80.31%	111.013
Track Deform3D†	RGB-D	13.76%	27.705
RGB-only (ours)	RGB	100%	0.974208
A^{3}E (ours)	RGB+events	100%	0.857035

(b) Offline reconstruction diagnostics—DEV6

Method	Input	Coverage	Adapted MTE ↓
E-D3DGS§	RGB+events	6/6	1.126
Wu 4DGS	RGB	6/6	1.302
RGB 4DGS	RGB	6/6	1.676
ERF-GS	RGB+events	6/6	2.232
DeformGS‡	4-view RGB	6/6	33.404
Dynamic 3D Gaussians‡	4-view RGB	6/6	51.107

Table S6: External tracking and adapted reconstruction comparisons under different observation contracts (mm), not a joint ranking. (a) RGB-frame queries; SpatialTrackerV2 has one native failure. †TrackDeform3D uses privileged RGB-D and sparse material coverage. (b) Offline full-sequence DEV6: IDs bound at initialization, no future GT rematching; not causal S evaluation. §E-D3DGS selects zero dynamic Gaussians on 6/6 sequences: a static fixed-ID diagnostic. ‡DeformGS (Duisterhof et al., 2024) (formerly MD-Splatting) and Dynamic 3D Gaussians (Luiten et al., 2024) use privileged four-view RGB with weak material attachment.

Common-frame tracking uses RGB-frame/POST queries: SpatialTrackerV2 retains native sequence coverage, TrackDeform3D uses RGB-D with sparse material coverage, and our observers use the same full material-node support. The DEV6 reconstruction methods use full-sequence access and bind material IDs at initialization without future GT rematching; the table caption identifies privileged multi-view inputs and the static E-D3DGS diagnostic.

F Sparse-RGB / Long-Gap Stress Test

This separate held-out stress test evaluates five continuous 500-ms RGB gaps with the recurrent state propagated at 100 Hz and no later ground-truth reset. It is reported as a robustness extension of the primary inter-frame result, not as a controlled comparison of RGB frame rates because its held-out cohort and RGB observation schedule differ from RGB30.

Table S7: Final-S full-precision supplied results (mm). Three frozen A^{3}E seeds are averaged within each of 24 equally weighted trajectories. Context-only PRE/POST retain their supplied six-decimal precision; no additional digits or paired CI are inferred.

Method	Interior MTE	PRE	POST
S-RGB-only	1.5695275824611066	2.2299076020201327	1.6904705443295274
S-Context-only	1.411122273071891	2.036898	1.543893
S-A^{3}E	0.9416713648945295	1.3002994448887273	1.017590801937552

Across the 24 held-out stress trajectories, A^{3}E reduces interior MTE from 1.570 to 0.942 mm, a 40.0% reduction, with a paired A^{3}E–RGB-only difference of -0.6279 mm and 95% CI $[-0.7850, -0.4803]$ mm (24/0/0 wins/losses/ties). A separately trained Context-only control obtains 1.411 mm. The long-gap result is used only to show that the event-assisted recurrent observer remains beneficial when the interval between RGB observations is substantially increased.

Table S8: Final-S paired capacity contrasts (mm). Context-only versus RGB crosses zero; A^{3}E versus Context-only favors events.

Contrast	Difference	95% CI; W/L/T
Context-only − RGB	-0.1584053093892153	[-0.31241171147633395, 0.0018836192611975141]; 17/7/0
A^{3}E − Context-only	-0.4694509081773617	[-0.6089987063355901, -0.34333813263960644]; 24/0/0

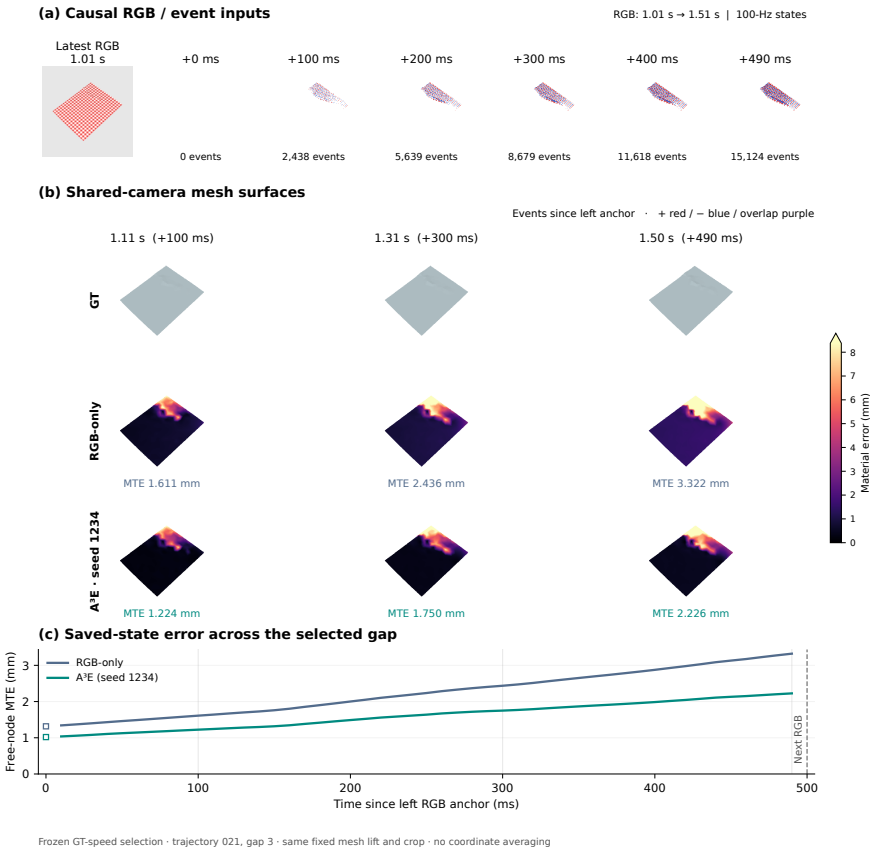

Figure S2: Representative long-gap stress example. The latest RGB observation and causal event evidence are shown together with GT, RGB-only, and A^{3}E states at selected inter-frame query times. This separate stress test is not presented as a controlled frame-rate comparison with RGB30.

G Additional Mechanism and Robustness Results

Recurrent-pathway shadow interventions. On the same primary RGB30 cohort, replacing corrected displacement history by its nominal-shadow counterpart increases MTE by $+0.541$ mm (paired 95% CI $[0.360, 0.749]$). Replacing dynamics geometry increases MTE by $+0.581$ mm ($[0.396, 0.790]$). Association-shadow, which blocks corrected-state feedback into future association while leaving the other recurrent pathways active, changes MTE by -0.000077 mm ($[-0.001242, 0.001285]$). The latter interval includes zero, so it does not resolve an additional accuracy benefit from that specific edge. These interventions are not additive components of the overall recurrent gain.

Specified action-distribution shifts. An exploratory extension uses a frozen Full observer on eight trajectories per condition. The speed-shift action is 4.000×10^{-4} m per 10-ms step, ap-

proximately 1.98 times the audited adapter-training maximum of 2.021×10^{-4} m; the move-hold condition introduces a temporal pattern absent from the audited adapter-training actions. The conditions are not pooled with primary RGB30, and the table is not an independent blind confirmation cohort.

Table S9: Exploratory action shifts, eight trajectories per condition. RGB-only and EV-Cloth MTE are in mm; W/L/T is from EV-Cloth's perspective.

Condition	RGB-only	EV-Cloth	W/L/T
In-distribution control	1.427	**0.665**	8/0/0
Speed shift	5.093	**4.492**	8/0/0
Move-hold temporal shift	3.378	**1.822**	8/0/0

The benefit persists in each tested case, while both observers have markedly larger absolute error under the speed shift than under the in-distribution control. This supports only the specified speed and move-hold conditions, not arbitrary action, mesh, appearance, or physics shifts.

Single extended development rollout. IG01 is one 20.18-s development diagnostic comprising 2,019 temporally linked states, not independent trajectories. Full-rollout MTE changes from 16.98 to 13.77 mm. In the ground-truth-defined major-deformation interval at 8.96–10.97 s, MTE changes from 26.01 to 22.60 mm. The full rollout, rather than a short prefix, supports the reported diagnostic. Its native metric is not pooled with primary RGB30, and substantial residual error remains.

Evidence versus writeback paired contrasts. Relative to the full 100-Hz condition, 50-Hz accumulated-evidence writeback increases MTE by $+0.1531$ mm (paired 95% CI $[0.058, 0.262]$), whereas retaining 50% of events at 100-Hz writeback changes it by $+0.0044$ mm ($[-0.009, 0.020]$). The conditions change distinct aspects of the update process, and the 25-Hz accumulated branch leaves an unconsumed evidence tail (95.2% consumed in the reported audit).

Table S10: Event evidence and writeback opportunities. Dynamics and evaluation remain at 100 Hz. Accumulated-evidence and event-thinning conditions change different aspects of the observation/update process; equal percentages and rates do not denote equivalent inputs. MTE is in mm.

Writeback	Event handling	MTE ↓
100 Hz	Full evidence	**1.047**
50 Hz	Accumulate elapsed-interval evidence	1.200
25 Hz	Accumulate elapsed-interval evidence; 95.2% consumed*	1.401
10 Hz	Accumulate elapsed-interval evidence	1.543
100 Hz	Retain 50% of events	1.052
100 Hz	Retain 25% of events	1.067
100 Hz	Retain 10% of events	1.105

*The 25-Hz branch leaves an unconsumed tail in the final audit. Accumulation also changes evidence availability and feedback timing; this is not a pure frequency-only intervention.

G.1 Representation Accuracy

On the current matched RGB30 fresh24 evaluation, Local-recurrent MTE is 1.047114 mm for Full18D, 1.085794 mm for Voxel, and 1.094541 mm for Time-surface. The direct paired Voxel − Full18D contrast is $+0.038680$ mm (95% CI $[-0.017257, 0.097498]$; Full18D-view W/L/T 14/10/0), leaving that difference unresolved. Time-surface − Full18D is $+0.047427$ mm ($[0.000839, 0.094778]$; 16/8/0), supporting lower Full18D error under this protocol. The association-by-recurrence interactions are $+0.484849$ mm ($[0.314364, 0.677533]$) for Full18D, $+0.360931$ mm ($[0.158687, 0.614764]$) for Voxel, and $+0.346599$ mm ($[0.124490, 0.606550]$) for Time-surface; all intervals are paired 95% CIs.

H REAL-WORLD STUDY

H.1 PHYSICAL SETUP AND EVALUATION CONTRACT

We evaluate one physical SO-101 cloth-manipulation DEV recording by offline causal replay of saved states. The methods share the initial state, robot action, material-node order, RGB projection, and persistent checker correspondence. Dynamics-only has no visual writeback; RGB-only uses the frozen RGB branch and zero event input. The event-assisted real-data variant (E5) uses a dynamic denoised four-bin event provider, hard-face binding, geodesic pooling, a support gate, and a bounded residual. Thus E5 is disclosed as a real-data variant of the simulated event configuration, not an identical transfer. An original-event branch (E1) is retained as a failure control, not a competitive method. The reference comprises 160 persistent image-plane checker tracks with a frozen barycentric material mapping; it is not independently measured 3D ground truth. Saved process repeats are computational repeats of this recording, not independent physical trials.

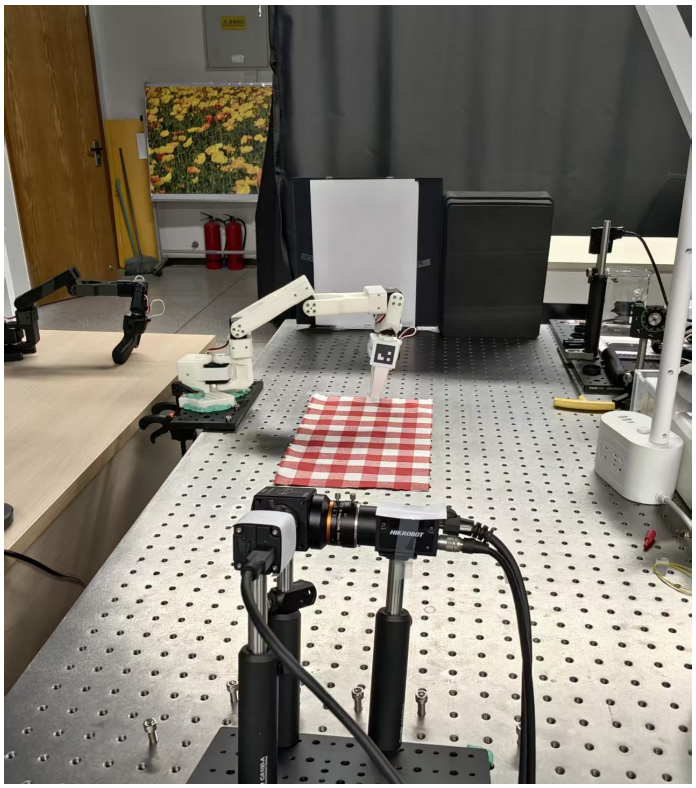

Figure S3: Physical setup for the reported recording. The SO-101 manipulates cloth in the shared RGB/event workspace. The overview documents the robot, cloth, and camera arrangement, rather than independent calibration accuracy.

Table S11: Audited physical configuration. Status terms apply to the reported evaluation contract; provisional and approximate relations are not independently validated calibrations.

Item	Status	Configuration or limitation
Robot and action	PROVISIONAL	SO-101; session-specific robot/world action relation.
RGB intrinsics	CALIBRATED	Frozen K_R, D_R for the canonical RGB view.
Event intrinsics	UNVALIDATED	No independent event intrinsic model used.
RGB/event relation	PROVISIONAL	Historical table-plane 2D map; no accepted 3D stereo extrinsics.
Robot/camera relation	PROVISIONAL	Combined relation inherits provisional robot/world estimate.
Synchronization	APPROXIMATE	RGB/event and robot/RGB offsets include residual timing uncertainty.
Initialization X_0	REUSED	One frozen 200-node initial state.
Checker correspondence	MEASURED	160 persistent 2D IDs and frozen material mapping.
3D evaluation reference	NOT VALIDATED	DA3-derived reconstruction; independent off-plane accuracy unvalidated.

H.2 Qualitative Material-Point Behavior

Figure S4(a) shows a post-hoc selected favorable example on a visibly lifted cloth region. At the three fixed displayed times, the event-assisted real-data variant moves the projected point-65 estimate toward the image-plane checker relative to RGB-only.

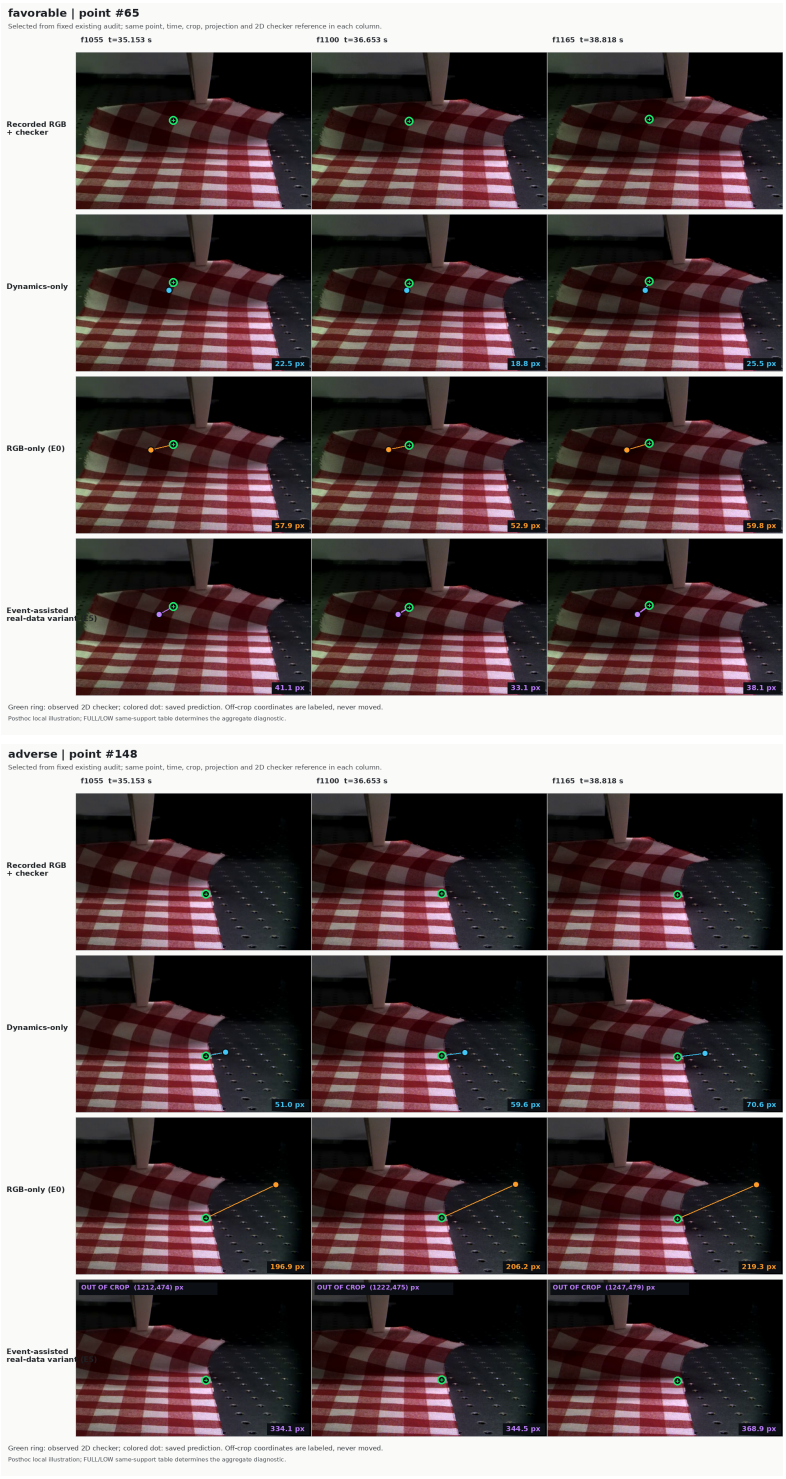

Figure S4: Audited image-plane material-point cases at the same three timestamps. (a) Point 65 is a selected favorable local example; its event-assisted projected point is closer to the checker than RGB-only at these frames. (b) Point 148 was fixed as an adverse/least-gain case before full-metric inspection; the event-assisted projection is outside the displayed crop, with its original full-image coordinates labeled in the panel. Colored dots are saved predictions and green rings are observed 2D checkers. Neither case is an aggregate performance estimate.

Panel (b) shows point 148, fixed in the adverse or least-gain bucket before inspection of the full metrics; at the same times its event-assisted discrepancy is larger. Both panels use frames 1055, 1100, and 1165, at 35.153119, 36.652541, and 38.818373 seconds, with the same material identity, timestamps, projection, crop, and scale across methods within each panel. These local observations do not imply aggregate real-world superiority. A continuous qualitative video is provided in the supplementary artifacts.

The independently audited same-support image-plane checker diagnostic likewise does not establish an aggregate event-assisted improvement; its complete statistics are provided in the supplementary machine-readable artifacts.

H.3 Secondary 3D Pseudo-Reference Diagnostic

The secondary reference follows RGB observations → *Depth Anything 3* DA3-LARGE depth (Lin et al., 2025) → audited calibration and geometric conversion → fixed material correspondence → one shared reconstructed 3D pseudo-reference. The depth-to-metric conversion uses observed supporting table geometry and a fixed initialization correction. The same reconstructed reference, scale, alignment, and visible checker support score all three methods; no method-specific alignment or rematching is applied. The reference is evaluation-only and does not enter recurrent correction after initialization. Common support is 19,656 point-time pairs from the preselected action-defined low-motion segment of the recording, not independent physical recordings.

Table S12: 3D discrepancy against a reconstructed pseudo-reference (mm; 19,656 common point-time pairs). The shared reference derives from RGB observations through DA3-LARGE and audited calibration, not independently measured 3D material ground truth. Discrepancies combine observer error with uncertainty in depth, metric scale, calibration, synchronization, and correspondence; they are not absolute real-world tracking accuracy and are not directly comparable with simulated MTE. Exact values are in the supplementary CSV.

Method	Mean	Median	P95
Dynamics-only	30.595	24.891	70.145
RGB-only	42.030	35.100	100.682
Event-assisted real-data variant	42.878	36.203	93.401

The reconstructed reference is most reliable near the calibrated supporting geometry; independent accuracy for off-plane cloth motion has not been established. Accordingly, this diagnostic is only a secondary geometric check. Because the pseudo-reference uncertainty is not independently quantified, these discrepancies do not establish an ordering in true 3D tracking accuracy.

H.4 Sim-to-Real Considerations

Prior work documents simulation-to-reality discrepancies in event sensing (Stoffregen et al., 2020; Hu et al., 2021; Chanda et al., 2025), robot execution (Collins et al., 2019), and cloth dynamics (Blanco-Mulero et al., 2024). These sources are coupled in EV-Cloth: physical robot motion affects the action passed to the dynamics prior; dynamics mismatch alters predicted geometry; predicted geometry affects event-to-surface association; and visual correction changes the state supplied to later dynamics. Calibration, synchronization, initialization, and pseudo-reference reconstruction add further uncertainty. The present study does not isolate these contributions, so the observed physical discrepancies cannot be assigned to a single component. The scope is one DEV recording without independent physical replication or independently validated off-plane 3D reference accuracy.