LPA-CWM: A LEARNED PHYSICAL ADJUDICATOR FOR MOTION REASONING WITH COUNTERFACTUAL WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Counterfactual world models (CWM) extract motion from pretrained video predictors by comparing factual and intervened predictions, but uniform aggregation weights responses equally without explicitly incorporating physical priors. Our key insight is to incorporate physical priors into candidate reliability learning, motivating **LPA-CWM** with a lightweight Learned Physical Adjudicator (LPA). Trained on dense MOVi-F trajectories, the 3.0M-parameter LPA compares visual context and response structure across an unordered candidate set to predict relative weights; windowed localization and one paired re-evaluation recover motion with the CWM frozen. Existing video-level benchmarks do not directly assess motion correspondence, where low localization error can conceal missing trajectory segments. We introduce Completeness-aware Motion Correspondence (CMC), a ground-truth-anchored protocol jointly measuring localization, completeness, visibility, and continuity, counting missing predictions as failures on visible dynamic points. Across DAVIS, Kinetics, and RoboTAP, LPA-CWM improves all main CMC measures over Uniform CWM, with relative gains of 18.1%–60.0% in average Dynamic Correspondence Accuracy (DCA_{avg}), and improves TAP-Vid First tracking accuracy (overview: `https://LPA-CWM.github.io`).

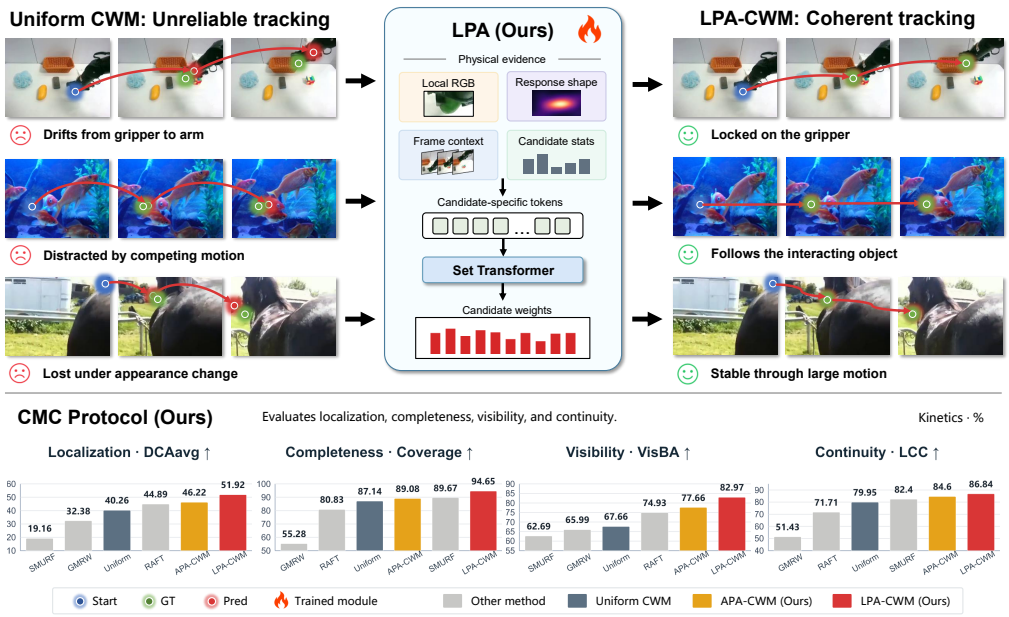


Figure 1: **Overview of our work.** Uniform CWM can suffer from three characteristic failures in motion reasoning: drifting from moving objects, being distracted by competing motion, and losing track under appearance changes. Our Learned Physical Adjudicator (LPA) identifies reliable counterfactual motion evidence, enabling LPA-CWM to recover more coherent trajectories. We further introduce the CMC protocol to evaluate motion correspondence across localization, completeness, visibility, and continuity. Compared with external methods, Uniform CWM, and APA-CWM, LPA-CWM performs strongly across all four dimensions.

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.

1 INTRODUCTION

World models learn predictive representations of visual dynamics (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020). Counterfactual World Models (CWM) extract motion from these representations by comparing predictions with and without localized interventions (Bear et al., 2023; Venkatesh et al., 2024; Stojanov et al., 2025). Different target-frame masks can produce inconsistent responses for the same query: some localize the motion well, while others are affected by background content, competing motion, or visual ambiguity. Uniform aggregation treats all responses as equally reliable and can dilute useful evidence. This limitation partly stems from insufficient use of physical priors, motivating physically informed aggregation and comprehensive trajectory evaluation.

To incorporate physical priors into CWM inference, we formulate multi-mask aggregation as candidate reliability estimation. We first derive APA-CWM, an **Analytic Physical Adjudicator (APA)**, from the Normalized Spatiotemporal Gradient (NSG) (Zhang et al., 2025). APA reads the NSG-derived spatiotemporal consistency around each candidate response peak and converts these local scores into closed-form reliability weights. It provides a training-free physical reference for testing how far an analytic consistency prior can resolve candidate ambiguity. We then propose LPA-CWM, whose **Learned Physical Adjudicator (LPA)** combines local visual evidence, response morphology, shared frame-pair context, and candidate statistics. A Transformer without candidate-index positional embeddings jointly compares the unordered candidate set and predicts relative weights.

We train only the 3.0M-parameter LPA on dense Kubric MOVi-F trajectories (Greff et al., 2022) to learn physical motion priors. For each query, dense forward motion provides the oracle target endpoint; candidate endpoint errors define relative reliability targets. Training combines distribution matching, weighted-coordinate localization, and best-candidate ranking. This gives direct supervision for which CWM responses are trustworthy while leaving the pretrained CWM unchanged. At inference, LPA weights the initial responses, followed by windowed localization and one paired CWM re-evaluation to obtain the final motion.

Existing benchmarks assess video quality, temporal consistency, physical and commonsense faithfulness, controllability, embodied utility, and counterfactual consistency (Huang et al., 2024; Zheng et al., 2025; Duan et al., 2025; Shang et al., 2026; Begiristain et al., 2026). They broaden evaluation beyond visual plausibility, but do not directly measure point correspondence recovered through CWM interventions. In motion recovery, low error on returned predictions can hide missing trajectory segments, while smooth trajectories may still follow incorrect motion.

We therefore introduce **Completeness-aware Motion Correspondence (CMC)**, a ground-truth-anchored protocol for annotated dynamic point tracks. CMC evaluates localization, completeness, visibility, and temporal continuity. Dynamic Correspondence Accuracy requires visible predictions within a distance threshold and counts missing predictions as failures on ground-truth-visible points. Track-level success further requires sufficient coverage and bounded localization error, while visibility and continuity diagnostics expose interrupted trajectories. CMC complements TAP-Vid First (Doersch et al., 2022) by explicitly measuring completeness on moving tracks.

We evaluate on DAVIS, Kinetics, and RoboTAP (Pont-Tuset et al., 2017; Kay et al., 2017; Vecerik et al., 2024) against six flow, correspondence, and point-tracking baselines (Stone et al., 2021; Shrivastava & Owens, 2024; Jiang et al., 2024; Karaev et al., 2025; Teed & Deng, 2020; Wang et al., 2024). Across the three datasets, LPA-CWM improves all main CMC measures over Uniform CWM, with relative $\mathrm{DCA}_{\mathrm{avg}}$ gains of 18.1%–60.0%. Dedicated trackers remain stronger on some localization or visibility measures, while LPA-CWM leads in completeness and continuity and achieves the best results on all main CMC measures on Kinetics. Within the CWM family, LPA-CWM also leads all TAP-Vid First measures across the three datasets.

In summary, our main contributions are:

- We identify a key principle for CWM inference: physical priors should guide candidate reliability estimation and counterfactual response aggregation.
- We propose LPA-CWM, which uses a 3.0M-parameter adjudicator to predict candidate weights before windowed localization and one paired CWM re-evaluation.
- We introduce CMC for joint evaluation of localization, completeness, visibility, and continuity, demonstrating 18.1%–60.0% relative $\mathrm{DCA}_{\mathrm{avg}}$ gains over Uniform CWM.

2 Related Work

World models and counterfactual motion. Video world models learn predictive representations through feature prediction, as in V-JEPA (Bardes et al., 2024), or action-controllable generation, as in Genie (Bruce et al., 2024). CWM extracts correspondence by comparing factual and locally intervened predictions (Bear et al., 2023), and later work applies this interface to physical dynamics (Venkatesh et al., 2024). Opt-CWM learns query-conditioned perturbations for flow and occlusion (Stojanov et al., 2025); KL-tracing measures the response of generative video models through KL divergence (Kim et al., 2025); and Point Prompting propagates colored markers through video diffusion models (Shrivastava et al., 2026). LPA-CWM focuses on learning the relative reliability of the multiple responses already produced by a frozen CWM.

Diffusion models as structural priors. Diffusion features also support visual correspondence. SD4Match adapts Stable Diffusion features with image-pair-conditioned prompts (Li et al., 2024), while Track4Gen jointly trains video generation and point correspondence to reduce appearance drift (Jeong et al., 2025). NSG relates a diffusion score field to observed inter-frame changes as a spatiotemporal consistency signal (Zhang et al., 2025). APA-CWM uses this signal for analytic candidate weighting, while LPA-CWM learns the weights from trajectory supervision.

Flow, correspondence, and point tracking. Our external baselines cover complementary motion interfaces. RAFT uses all-pairs correlations with iterative flow refinement (Teed & Deng, 2020), and SEA-RAFT improves the framework with probabilistic training and a more efficient design (Wang et al., 2024). SMURF learns optical flow through full-image warping and multi-frame self-supervision (Stone et al., 2021). GMRW learns global correspondence with contrastive random walks (Shrivastava & Owens, 2024), whereas DODUO combines semantic priors with self-supervised flow (Jiang et al., 2024). CoTracker3 is a dedicated point tracker trained with pseudo-labels from real videos (Karaev et al., 2025); AllTracker extends dense long-range tracking with multi-frame flow and temporal attention (Harley et al., 2025).

Evaluation of motion consistency. General video and world-model benchmarks cover visual quality, temporal consistency, physical faithfulness, controllability, embodied utility, and counterfactual consistency (Huang et al., 2024; Zheng et al., 2025; Duan et al., 2025; Shang et al., 2026; Begiristain et al., 2026). Motion-specific evaluation includes FVMD for point-track motion distributions (Liu et al., 2024), Physics-IQ for prediction against real physical experiments (Motamed et al., 2026), and TAPVid-3D for long-range 3D correspondence and visibility (Koppula et al., 2024). ITTO further diagnoses point trackers along motion-complexity and reappearance axes, exposing failures under fast motion and repeated occlusion (Demler et al., 2025). In contrast, CMC evaluates the quality of recovered 2D dynamic trajectories themselves, jointly measuring localization, completeness, visibility, and continuity while counting missing predictions on visible points as failures.

3 Method

Given a source–target frame pair and a query point, our method first uses a frozen counterfactual world model (CWM) to produce multiple motion candidates. We then compare two weighting schemes: an analytic rule, APA-CWM, and a learned adjudicator, LPA-CWM. Both variants combine the candidate response maps before applying the same localization and re-evaluation procedure. Figure 2 summarizes the method.

3.1 Counterfactual Motion Candidates from a Frozen World Model

Let $\mathbf{I}_t, \mathbf{I}_{t+1} \in [0,1]^{256\times256\times3}$ denote a source–target RGB frame pair, and let $\mathbf{p}_t \in \Omega$ be a query in the source frame, where Ω is the 256×256 grid. We build on counterfactual world models (Bear et al., 2023; Venkatesh et al., 2024; Stojanov et al., 2025), with an RGB predictor Ψ^{RGB} and a query-conditioned intervention generator $\boldsymbol{\delta}_\theta$. Both Ψ^{RGB} and $\boldsymbol{\delta}_\theta$ are pretrained and remain fixed.

The generator produces the intervention $\boldsymbol{\delta}_t = \boldsymbol{\delta}_\theta(\mathbf{I}_t, \mathbf{p}_t) \in \mathbb{R}^{256\times256\times3}$. For each target-frame mask $\mathcal{M}_m$, CWM measures how this intervention changes the target prediction:

$$\boldsymbol{\Delta}_m = \Psi^{\mathrm{RGB}}(\mathbf{I}_t + \boldsymbol{\delta}_t, \mathcal{M}_m(\mathbf{I}_{t+1})) - \Psi^{\mathrm{RGB}}(\mathbf{I}_t, \mathcal{M}_m(\mathbf{I}_{t+1}))\,. \tag{1}$$

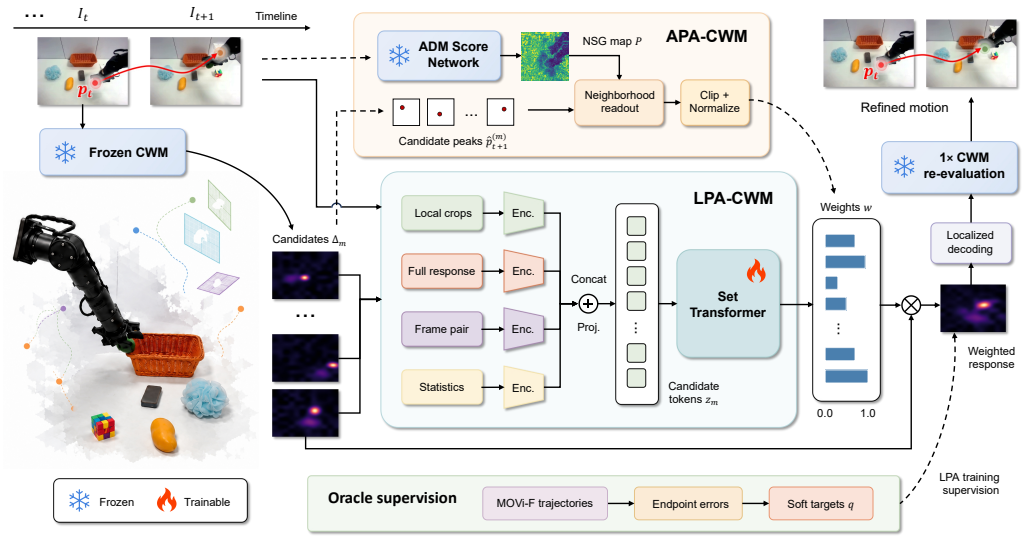


Figure 2: **Overview of APA-CWM and LPA-CWM.** Given a source–target RGB frame pair $(\mathbf{I}_t, \mathbf{I}_{t+1})$ and a query point $\mathbf{p}_t$, a frozen CWM generates motion candidates $\{\boldsymbol{\Delta}_m\}$. APA-CWM derives candidate weights from the NSG-derived map $\mathcal{P}$ around candidate peaks $\widehat{\mathbf{p}}_{t+1}^{(m)}$. LPA-CWM instead encodes local crops, full responses, the frame pair, and response statistics; the resulting tokens $\mathbf{z}_m$ are processed by a trainable Set Transformer to predict weights. The weighted response undergoes windowed localization and one CWM re-evaluation to obtain refined motion. During LPA training, endpoint errors from MOVi-F trajectories are converted into soft targets $\mathbf{q}$.

Thus, $\boldsymbol{\Delta}_m(\mathbf{u})$ measures how much the prediction at target pixel $\mathbf{u}$ changes when the queried source region is perturbed. A large response indicates a strong dependency between that target location and the queried source region. Repeating this process for M masks produces

$$\mathcal{C}_t = \left\{ \boldsymbol{\Delta}_m, \widehat{\mathbf{p}}_{t+1}^{(m)} \right\}_{m=1}^{M}, \qquad \widehat{\mathbf{p}}_{t+1}^{(m)} = \arg\max_{\mathbf{u}\in\Omega} \|\boldsymbol{\Delta}_m(\mathbf{u})\|_1 . \tag{2}$$

Here, $\mathbf{u}$ indexes target-frame pixels and the ℓ_1 norm reduces the three RGB channels. The peak $\widehat{\mathbf{p}}_{t+1}^{(m)}$ is the target location with the strongest response under mask m, and therefore serves as that mask's proposed endpoint for the query $\mathbf{p}_t$. Because different masks expose different target contexts, their response maps and candidate endpoints can disagree. Uniform CWM averages all responses equally, whereas APA-CWM and LPA-CWM estimate candidate-specific weights.

3.2 Analytic Candidate Weighting

APA-CWM provides a closed-form reference based on the Normalized Spatiotemporal Gradient (NSG) (Zhang et al., 2025). Let $\mathbf{x}_t, \mathbf{x}_{t+1} \in [-1, 1]^{224\times224\times3}$ denote the frame pair $(\mathbf{I}_t, \mathbf{I}_{t+1})$ after preprocessing for a frozen Ablated Diffusion Model (ADM) (Dhariwal & Nichol, 2021). The notation distinguishes these ADM inputs from the 256×256 RGB frames used by CWM. Let $\mathbf{s}_\phi$ be the frozen ADM score network, t_s its diffusion time, and $\mathbf{u}$ a location on the ADM grid.

Following NSG, we compute a **spatial score** $\mathcal{P}$ that compares the local ADM score magnitude with its alignment to the observed frame change:

$$\mathcal{P}(\mathbf{u}) = \frac{\|\mathbf{s}_\phi(\mathbf{x}_t, t_s)(\mathbf{u})\|_2}{|\langle \mathbf{s}_\phi(\mathbf{x}_t, t_s)(\mathbf{u}), \mathbf{x}_{t+1}(\mathbf{u}) - \mathbf{x}_t(\mathbf{u})\rangle| + 10^{-6}}. \tag{3}$$

The numerator is the local score magnitude, while the denominator is the absolute score–change inner product plus a 10^{-6} numerical stabilizer. The norm and inner product are computed over RGB channels. APA uses the resulting $\mathcal{P}$ map as an NSG-derived heuristic for candidate weighting, rather than a calibrated probability of correct correspondence.

We bilinearly resize $\mathcal{P}$ to the CWM response grid and keep the same notation for the aligned map. For each candidate, APA averages this score in the radius-six neighborhood $\mathcal{N}_6$ around its response

peak and then normalizes the candidate scores:

$$r_m^{\mathrm{APA}} = \frac{1}{\left|\mathcal{N}_6(\widehat{\mathbf{p}}_{t+1}^{(m)})\right|} \sum_{\mathbf{u}\in\mathcal{N}_6(\widehat{\mathbf{p}}_{t+1}^{(m)})} \mathcal{P}(\mathbf{u}), \qquad w_m^{\mathrm{APA}} = \frac{\mathrm{clip}(r_m^{\mathrm{APA}}, 10^{-3}, 10^3)}{\sum_j \mathrm{clip}(r_j^{\mathrm{APA}}, 10^{-3}, 10^3)}. \tag{4}$$

Here, r_m^{APA} is the local NSG score for candidate m. Clipping limits extreme values, and normalization converts these scores into nonnegative candidate weights w_m^{APA} that sum to one. APA thus provides a training-free way to favor candidates supported by the NSG cue. Since this cue is not directly supervised by endpoint accuracy, we next learn candidate reliability from dense motion trajectories.

3.3 Learned Physical Adjudicator

Candidate representation and set reasoning. LPA represents each candidate using four sources of information. **(1) Local visual evidence:** the local branch receives a 10-channel 32×32 tensor containing source and target RGB crops, their absolute RGB difference, and a crop of the candidate response. These inputs correspond to 64×64 regions centered at $\mathbf{p}_t$ and $\widehat{\mathbf{p}}_{t+1}^{(m)}$. Before cropping, the RGB response is reduced and normalized as $\widetilde{d}_m(\mathbf{u}) = \|\boldsymbol{\Delta}_m(\mathbf{u})\|_1/(\max_{\mathbf{v}} \|\boldsymbol{\Delta}_m(\mathbf{v})\|_1 + 10^{-8})$. **(2) Full response:** a second branch encodes the complete $\widetilde{d}_m$ map to retain the global shape and ambiguity of the response. **(3) Frame context:** a third branch encodes the full frame pair to provide scene-level visual information shared by all candidates. **(4) Candidate statistics:** the final branch encodes 16 scalar descriptors covering displacement, response concentration, ambiguity, photometric change, and edge strength. The exact descriptors and encoder architectures are given in Appendix A.

The four encoded features are combined and projected to a 256-dimensional candidate token $\mathbf{z}_m$. We denote the Transformer encoder (Lee et al., 2019) without candidate-index positional embeddings by SetEnc. It compares all candidate tokens, after which a shared scoring head h_{score} produces one score per candidate:

$$[\widetilde{\mathbf{z}}_1, \ldots, \widetilde{\mathbf{z}}_M] = \mathrm{SetEnc}\left([\mathbf{z}_1, \ldots, \mathbf{z}_M]\right), \qquad w_m^{\mathrm{LPA}} = \frac{\exp(h_{\mathrm{score}}(\widetilde{\mathbf{z}}_m))}{\sum_j \exp(h_{\mathrm{score}}(\widetilde{\mathbf{z}}_j))}. \tag{5}$$

The softmax produces the candidate-weight vector $\mathbf{w}^{\mathrm{LPA}} = (w_1^{\mathrm{LPA}}, \ldots, w_M^{\mathrm{LPA}})$, whose nonnegative entries sum to one. Without candidate-index positional embeddings, permuting the input candidates only permutes their output weights, while the weighted response remains unchanged. We also randomize candidate order during training.

Dense-trajectory supervision. We train LPA on candidate sets of size $M = 10$ generated from adjacent MOVi-F frames (Greff et al., 2022); training and validation videos are disjoint. MOVi-F provides a dense ground-truth forward motion field $\mathbf{F}_{t\to t+1}$, where $\mathbf{F}_{t\to t+1}(\mathbf{p}_t)$ is the displacement of $\mathbf{p}_t$ from frame t to $t+1$. The corresponding **ground-truth endpoint** is $\mathbf{p}_{t+1}^* = \mathbf{p}_t + \mathbf{F}_{t\to t+1}(\mathbf{p}_t)$.

We use this endpoint to measure how reliable each candidate is. Specifically, e_m is the endpoint error of candidate m, and these errors are converted into a **soft reliability target** $\mathbf{q} = (q_1, \ldots, q_M)$:

$$e_m = \left\|\widehat{\mathbf{p}}_{t+1}^{(m)} - \mathbf{p}_{t+1}^*\right\|_2, \qquad q_m = \frac{\exp(-e_m/\tau_q)}{\sum_j \exp(-e_j/\tau_q)}, \qquad \tau_q = 4. \tag{6}$$

Candidates closer to the ground-truth endpoint receive larger q_m. Unlike a hard label that selects only the single best candidate, $\mathbf{q}$ retains the relative quality of all candidates. The temperature τ_q controls how strongly the target distribution favors candidates with lower endpoint error.

Training objective. We train LPA with the following objective:

$$\mathcal{L} = D_{\mathrm{KL}}\left(\mathbf{q}\|\mathbf{w}^{\mathrm{LPA}}\right) + \lambda_1 \,\mathrm{SmoothL1}\left(\sum_m w_m^{\mathrm{LPA}} \widehat{\mathbf{p}}_{t+1}^{(m)}, \mathbf{p}_{t+1}^*\right) + \lambda_2 \,\mathrm{CE}\Big(\mathbf{a}, \arg\min_m e_m\Big), \tag{7}$$

where $\mathbf{a} = (h_{\mathrm{score}}(\widetilde{\mathbf{z}}_1), \ldots, h_{\mathrm{score}}(\widetilde{\mathbf{z}}_M)) \in \mathbb{R}^M$ is the vector of pre-softmax scores. The KL term matches the predicted weights to the soft reliability targets $\mathbf{q}$. The SmoothL1 term encourages the weighted candidate endpoint toward the ground-truth endpoint, while the cross-entropy term encourages the lowest-error candidate to receive the highest score. Here we set $\lambda_1 = 2$ and $\lambda_2 = 0.2$.

The weighted candidate endpoint in the SmoothL1 term is used only as a differentiable training signal. During inference, LPA does not directly average candidate peak coordinates; instead, its predicted weights are applied to the full CWM response maps, which are then decoded as described next. MOVi-F correspondence filtering and further training details are given in Appendix A.

3.4 Weighted Aggregation and Localized Decoding

Weighted response aggregation. Let $\mathbf{w} = (w_1, \dots, w_M)$ collect the aggregation weights, with $w_m = 1/M$ for Uniform CWM, $w_m = w_m^{\mathrm{APA}}$ for APA-CWM, and $w_m = w_m^{\mathrm{LPA}}$ for LPA-CWM. We combine the full candidate responses and reduce the three RGB channels to a scalar localization map:

$$\overline{\boldsymbol{\Delta}}_w = \sum_{m=1}^{M} w_m \boldsymbol{\Delta}_m, \qquad \overline{d}_w(\mathbf{u}) = \frac{1}{3} \left\| \overline{\boldsymbol{\Delta}}_w(\mathbf{u}) \right\|_1, \qquad w_m \geq 0, \quad \sum_m w_m = 1. \tag{8}$$

Here, $\overline{\boldsymbol{\Delta}}_w$ is the weighted CWM response and $\overline{d}_w(\mathbf{u})$ measures its response strength at pixel $\mathbf{u}$. Larger values indicate locations more strongly supported by the aggregated candidates.

Windowed localization and re-evaluation. To locate the endpoint, we turn $\overline{d}_w$ into a normalized spatial distribution $\pi^{(0)}(\mathbf{u}) \propto \exp(\beta_\pi \overline{d}_w(\mathbf{u}))$, with $\beta_\pi = 200$, normalized over all $\mathbf{u} \in \Omega$. Thus, $\pi^{(0)}(\mathbf{u})$ gives the relative support for pixel $\mathbf{u}$ as the endpoint, while $\pi^{(k)}$ denotes this distribution after k windowing updates. Standard localization directly selects the peak of $\pi^{(0)}$; our windowed localization instead applies two updates to suppress distant secondary peaks.

At each update, we compute the spatial mean $\boldsymbol{\mu}^{(k)}$ and root-mean-square spatial spread $\sigma^{(k)}$ of the current distribution, then apply a Gaussian window centered at $\boldsymbol{\mu}^{(k)}$:

$$\begin{cases} \boldsymbol{\mu}^{(k)} = \sum_{\mathbf{u}} \pi^{(k)}(\mathbf{u})\mathbf{u}, \\ \sigma^{(k)} = \left[\sum_{\mathbf{u}} \pi^{(k)}(\mathbf{u}) \left\| \mathbf{u} - \boldsymbol{\mu}^{(k)} \right\|_2^2 \right]^{1/2}, \\ \pi^{(k+1)}(\mathbf{u}) \propto \pi^{(k)}(\mathbf{u}) \exp\left[-\frac{\left\| \mathbf{u} - \boldsymbol{\mu}^{(k)} \right\|_2^2}{2[3\max(\sigma^{(k)}, 0.5)]^2} \right]. \end{cases} \tag{9}$$

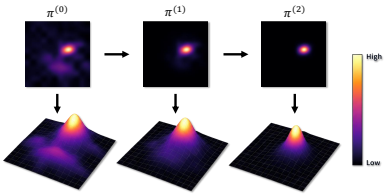


Figure 3: **Windowed localization progressively suppresses distant modes.** Top: flattened maps from $\pi^{(0)}$ to $\pi^{(2)}$. Bottom: the corresponding 3D views.

For $k = 0, 1$, we renormalize $\pi^{(k+1)}$ over Ω after applying the window. This reweighting downweights pixels farther from the current mean. The window width is $3\max(\sigma^{(k)}, 0.5)$ pixels, with the lower bound preventing window collapse. The discrete argmax of $\pi^{(2)}$ gives the initial endpoint.

Finally, this initial endpoint determines the target-mask center for one fresh, full-field CWM evaluation, while the source query remains $\mathbf{p}_t$. The new M responses are averaged uniformly and decoded to obtain the refined endpoint $\widehat{\mathbf{p}}_{t+1}$; APA and LPA are not rerun, and their initial weights are not reused. The pairwise motion is $\widehat{\boldsymbol{\phi}}_t = \widehat{\mathbf{p}}_{t+1} - \mathbf{p}_t$. In evaluation, each target is predicted from the fixed first-visible query without recursion. Visibility comes from the final paired CWM response-strength signal; LPA has no separate visibility head. Further details are provided in Appendix A.

4 Experiments

Our experiments mainly address three questions: **(1)** Does learned candidate weighting improve motion recovered from a frozen CWM? **(2)** How does LPA-CWM compare with analytic weighting and external motion baselines under CMC and standard point-tracking evaluation? **(3)** Which components and weighting behaviors contribute to the improvement?

4.1 Experimental Setup

Models and data. Uniform CWM, APA-CWM, and LPA-CWM use the same frozen predictor, intervention generator, evaluation queries, and mask-sampling protocol. They differ in initial candidate weighting, while paired re-evaluation is performed separately at the endpoint produced by

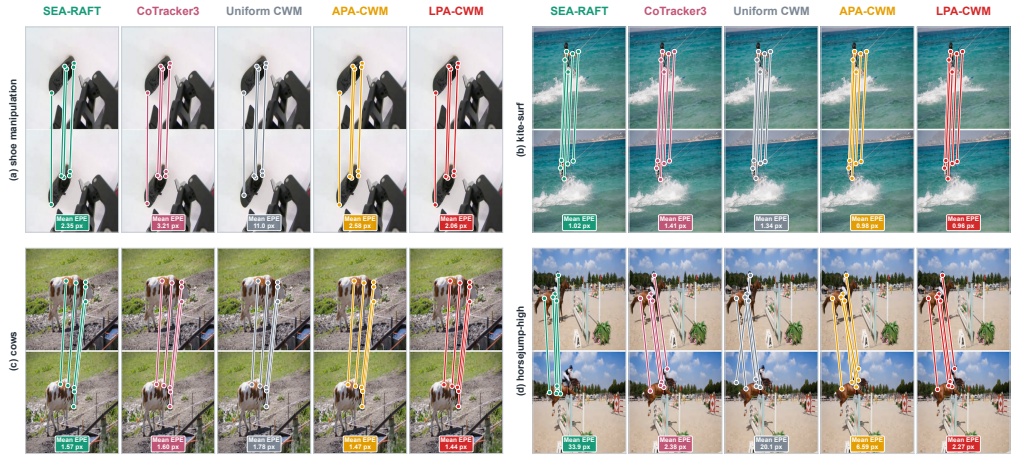


Figure 4: **Qualitative motion correspondence.** Five methods are compared on identical frames and visible queries across manipulation, human-action, and animal-motion scenes. Badges report mean EPE in pixels. LPA-CWM yields accurate, coherent correspondences across the displayed cases.

each variant. We fix $M = 10$, use one paired re-evaluation ($R = 1$), set $\beta_\pi = 200$, and apply two windowed localization updates. Only the 3.0M-parameter LPA is trained on MOVi-F trajectories, with checkpoint selection performed on held-out MOVi-F videos. We evaluate on three datasets—DAVIS (Pont-Tuset et al., 2017), Kinetics (Kay et al., 2017), and RoboTAP (Vecerik et al., 2024)—covering 90 videos and approximately 1,300 annotated point trajectories in total. All methods use the same evaluation subsets and queries; further protocol details are provided in Appendix B.

Evaluation and comparisons. We evaluate motion recovery using our CMC protocol and the standard TAP-Vid First protocol. For TAP-Vid First, we report Average Jaccard (AJ), the average fraction of points within the benchmark distance thresholds ($< \delta^x_{\text{avg}}$), mean and median endpoint distance (AD and MD), and occlusion accuracy (OA). AJ jointly reflects localization and visibility, while AD and MD summarize localization error. External baselines include SMURF, GMRW, DODUO, CoTracker3, RAFT, and SEA-RAFT, covering optical flow, correspondence, and dedicated point tracking. All methods receive identical evaluation frames and queries, and no baseline uses post-hoc smoothing or repair. Appendix D describes the flow and correspondence adapters.

4.2 Completeness-aware Motion Correspondence

CMC evaluates post-query motion on ground-truth dynamic tracks, counting missing predictions on visible points as failures. We retain tracks with at least 4 pixels of cumulative post-query displacement at 256×256 resolution. CMC uses four dimensions. **1. Localization** uses DCA_{avg} (DCA), averaging visible-and-correct predictions over $\{1, 2, 4, 8, 16\}$-pixel thresholds. **2. Completeness** uses Coverage (Cov.) and Complete@0.8 (C@.8), **3. visibility** by visibility balanced accuracy (VisBA) and occlusion F1 (OF1), and **4. continuity** by LCC, the longest uninterrupted predicted-visible run. Figure 5 shows one metric per dimension; Table 1 reports the full set. Complete Track Success (CTS) measures track-level success by requiring at least 80% coverage and mean endpoint error (EPE) at most 4 pixels on jointly visible points. Definitions, threshold sweeps, and fragmentation diagnostics appear in Appendix C.

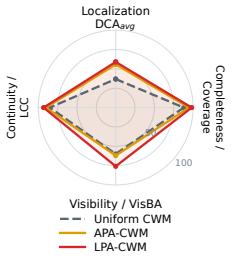


Figure 5: **CMC summary on DAVIS.** Higher is better.

Table 1 reports CMC results for all methods on three datasets. Within the CWM family, LPA-CWM leads all seven CMC measures across the three datasets. Dedicated trackers such as CoTracker3 remain stronger on some localization or visibility measures, while LPA-CWM leads in completeness and continuity. Notably, on Kinetics, LPA-CWM achieves the best result among all evaluated methods on every main CMC measure. Relative to Uniform CWM, learned adjudication improves DCA_{avg} by 22.11, 11.66, and 7.69 percentage points on DAVIS, Kinetics, and RoboTAP, respectively; relative to APA-CWM, the corresponding gains are 3.19, 5.70, and 3.35 points.

Table 1: **Cross-dataset evaluation on DAVIS, Kinetics, and RoboTAP.** LPA is trained exclusively on MOVi-F and directly evaluated on all three datasets without adaptation. CMC metrics are percentages. For TAP-Vid First, AJ, $< \delta^x_{\text{avg}}$, and OA are percentages, while AD and MD are pixels. Bold marks the best result among the three CWM variants. Stronger external results are shaded separately.

Dataset	Method	CMC (Ours)							TAP-Vid First				
		Cov.↑	C@.8↑	DCA↑	CTS↑	VisBA↑	OF1↑	LCC↑	AJ↑	$< \delta^x_{\text{avg}}$ ↑	AD↓	MD↓	OA↑
DAVIS	SMURF	91.98	85.54	11.25	2.67	56.64	14.92	83.29	7.39	14.34	32.34	30.06	78.14
	GMRW	37.02	20.62	23.92	17.00	63.24	45.22	35.46	21.85	41.55	20.59	13.38	50.77
	DODUO	31.44	15.04	22.25	14.42	58.94	45.83	28.44	26.15	52.49	8.79	2.90	49.46
	CoTracker3	92.31	85.33	71.81	70.50	76.43	58.97	81.01	62.90	76.81	4.20	2.01	86.11
	RAFT	68.74	49.58	49.00	31.77	70.16	41.66	62.02	37.08	58.79	15.83	5.60	69.86
	SEA-RAFT	57.76	31.25	43.01	27.43	65.78	39.56	54.07	40.72	51.84	19.17	7.78	65.22
	Uniform CWM	88.10	80.96	36.87	21.46	59.91	33.91	85.20	24.31	40.21	13.10	3.98	76.56
	APA-CWM (Ours)	94.74	89.98	55.79	47.91	62.07	38.48	90.13	41.73	57.35	10.27	2.94	83.47
	LPA-CWM (Ours)	**98.21**	**95.63**	**58.98**	**57.54**	**75.89**	**48.14**	**93.18**	**45.29**	**63.92**	**6.73**	**2.07**	**84.18**
Kinetics	SMURF	89.67	81.89	19.16	14.17	62.69	35.11	82.40	17.28	29.43	27.67	20.56	70.48
	GMRW	55.28	36.25	32.38	30.42	65.99	46.34	51.43	29.83	48.48	17.87	12.13	53.52
	DODUO	67.04	49.87	40.21	39.87	78.37	58.39	59.83	36.89	46.32	13.80	4.15	67.69
	CoTracker3	87.06	79.67	50.15	46.61	78.00	58.56	80.41	50.86	63.21	9.85	3.60	84.80
	RAFT	80.83	69.24	44.89	43.94	74.93	60.72	71.71	37.65	47.67	18.20	4.00	80.52
	SEA-RAFT	78.59	65.73	43.36	40.71	72.67	51.96	70.65	37.88	47.65	18.61	3.71	79.45
	Uniform CWM	87.14	77.11	40.26	34.54	67.66	44.97	79.95	35.44	46.62	15.26	7.44	85.99
	APA-CWM (Ours)	89.08	83.78	46.22	37.10	77.66	59.22	84.60	35.66	47.40	12.15	4.01	87.19
	LPA-CWM (Ours)	**94.65**	**89.97**	**51.92**	**58.25**	**82.97**	**62.67**	**86.84**	**41.91**	**53.12**	**10.87**	**3.45**	**88.96**
RoboTAP	SMURF	90.33	81.38	21.13	11.67	57.17	28.66	78.28	27.03	42.56	23.89	15.91	75.68
	GMRW	30.31	10.89	18.41	7.01	61.48	35.21	28.76	20.83	34.73	23.49	13.76	47.69
	DODUO	52.31	30.82	31.93	18.73	73.28	53.09	47.15	40.71	56.52	9.86	3.85	69.92
	CoTracker3	81.02	73.92	51.40	43.05	73.11	59.35	75.21	56.58	71.79	11.21	1.96	80.14
	RAFT	76.69	54.71	49.10	41.27	70.60	44.27	69.46	45.57	57.69	13.06	3.21	72.00
	SEA-RAFT	75.58	62.14	42.45	24.68	65.23	37.86	64.72	47.38	61.93	12.04	2.61	77.62
	Uniform CWM	85.92	79.06	42.57	30.55	72.12	50.77	76.89	38.26	51.01	8.68	3.52	80.85
	APA-CWM (Ours)	91.10	81.78	46.91	42.34	75.63	57.44	78.71	43.17	56.90	7.88	2.88	83.24
	LPA-CWM (Ours)	**92.00**	**86.88**	**50.26**	**54.22**	**78.03**	**61.00**	**85.49**	**47.42**	**58.35**	**7.74**	**2.75**	**86.45**

4.3 TAP-Vid First and Qualitative Results

TAP-Vid First queries each point at its first visible frame and evaluates the subsequent trajectory (Doersch et al., 2022). Within the CWM family, LPA-CWM achieves the best result on all main measures across the three datasets (Table 1), while CoTracker3 remains stronger on several conventional tracking measures, particularly on DAVIS. Together with CMC, these results show that learned candidate weighting improves both tracking accuracy and motion completeness.

Figure 4 shows four qualitative comparisons spanning manipulation, human-action, and animal-motion scenes. LPA-CWM gives the lowest displayed mean EPE in all four cases and remains stable under large motion and appearance changes. Additional examples are shown in Appendix Figure F.1.

4.4 Ablation Studies

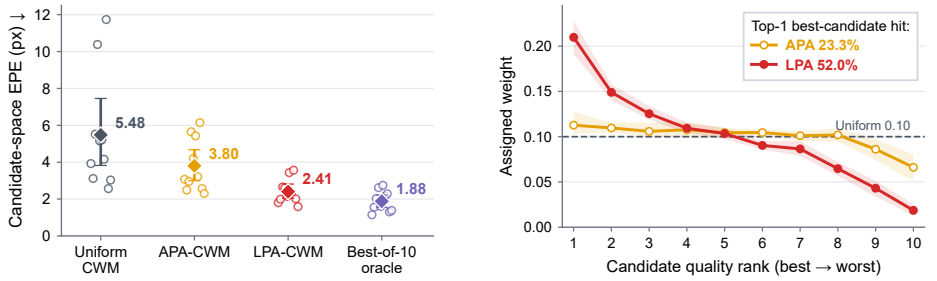
Mask count and re-evaluation. We select M and R on 40 fixed MOVi-F query groups, sweeping $M \in \{5, 10, 20\}$ and $R \in \{0, 1, 2\}$. The best accuracy–cost setting is $M = 10, R = 1$; $M = 20$ increases latency without improving accuracy, while $R = 2$ does not further reduce EPE.

Architecture, objectives, and decoding. Table 2 isolates the main design choices under matched inference settings. Removing candidate-set interaction or visual cues consistently reduces performance, while coordinate or best-candidate supervision improves over KL matching alone. The full objective performs best on all reported measures, and windowed localization improves every weighting rule. Appendix E provides the full sweeps, timing results, and detailed controls.

Candidate weighting. Figure 6 shows that LPA reduces candidate-space EPE toward the best-of-10 oracle, favors better candidates, and more than doubles APA's top-1 hit rate. The rank-wise weight distribution shows that LPA assigns greater weight to more accurate candidates.

Table 2: **Ablations on DAVIS under CMC and TAP-Vid First.** All configurations use $M = 10$ and $R = 1$. Architecture and objective variants use windowed localization; "Std." and "Win." denote standard and windowed localization. Metric conventions follow Table 1; best values within each block are bold, enabling direct comparison across design choices.

Configuration	CMC (Ours)							TAP-Vid First				
	Cov.↑	C@.8↑	DCA↑	CTS↑	VisBA↑	OF1↑	LCC↑	AJ↑	$< \delta^x_{\text{avg}}$ ↑	AD↓	MD↓	OA↑
Architecture ablation												
Uniform CWM (reference)	88.10	80.96	36.87	21.46	59.91	33.91	85.20	24.31	40.21	13.10	3.98	76.56
w/o candidate-set interaction	96.58	91.93	51.86	45.55	63.16	38.73	90.30	41.52	56.82	8.99	2.66	82.75
Response + scalar cues only	95.99	90.31	49.75	38.24	61.62	39.77	90.85	41.89	57.20	8.46	3.26	83.61
Full LPA-CWM	**98.21**	**95.63**	**58.98**	**57.54**	**75.89**	**48.14**	**93.18**	**45.29**	**63.92**	**6.73**	**2.07**	**84.18**
Objective ablation												
KL only	95.73	91.78	49.64	36.99	60.47	35.03	90.77	41.56	57.48	9.09	2.80	82.84
KL + coordinate supervision	96.60	94.33	52.13	49.42	62.00	39.07	91.59	42.51	58.21	8.11	2.12	83.34
KL + best-candidate supervision	97.89	95.00	51.27	48.75	62.34	38.50	92.89	42.92	57.96	7.95	2.16	83.73
Full objective	**98.21**	**95.63**	**58.98**	**57.54**	**75.89**	**48.14**	**93.18**	**45.29**	**63.92**	**6.73**	**2.07**	**84.18**
Aggregation and localization												
Uniform + Std.	86.93	77.00	34.54	17.71	57.14	31.43	84.05	23.25	39.46	13.61	4.26	76.18
APA + Std.	89.49	83.46	50.78	44.80	61.30	34.11	89.78	40.54	56.19	10.54	3.12	83.29
LPA + Std.	95.02	88.79	55.86	48.33	67.33	39.62	91.71	42.80	58.47	7.17	2.85	83.85
Uniform + Win.	88.10	80.96	36.87	21.46	59.91	33.91	85.20	24.31	40.21	13.10	3.98	76.56
APA + Win.	94.74	89.98	55.79	47.91	62.07	38.48	90.13	41.73	57.35	10.27	2.94	83.47
LPA + Win.	**98.21**	**95.63**	**58.98**	**57.54**	**75.89**	**48.14**	**93.18**	**45.29**	**63.92**	**6.73**	**2.07**	**84.18**



(a) Candidate-space weighted-coordinate EPE. (b) Weight versus ground-truth candidate rank.

Figure 6: **Candidate weighting on MOVi-F validation videos. (a)** Candidate-space weighted-coordinate EPE with video-macro means. **(b)** Weight versus ground-truth candidate rank; the inset reports top-1 hit rates. Error bars and shading show 95% video-cluster bootstrap intervals.

5 Conclusion and Limitations

We presented LPA-CWM for motion reasoning with a frozen counterfactual world model. LPA learns relative reliability across competing CWM candidates from dense trajectory supervision, while APA-CWM provides a training-free analytic reference. We also introduced CMC to evaluate localization, completeness, visibility, and continuity on dynamic tracks. Across DAVIS, Kinetics, and RoboTAP, LPA-CWM consistently improves over Uniform CWM and APA-CWM under both CMC and TAP-Vid First, with ablations confirming the contributions of candidate interaction, trajectory supervision, and windowed localization.

Three limitations remain. First, LPA can only reweight motion evidence already present in the frozen candidate set; missing or systematically incorrect motion requires improved candidate generation or adaptation of the underlying CWM. Second, LPA is trained on synthetic MOVi-F trajectories, while the current quantitative evaluation spans real-world and embodied datasets within one CWM family. Extending trajectory supervision to real-world data and evaluating across broader datasets and additional world-model architectures will better establish generalization. Third, multi-mask inference and paired re-evaluation increase computation. More efficient candidate selection, shared computation across masks, or fewer CWM evaluations could reduce this cost.

AI USE STATEMENT

Generative AI tools were used as research assistants for manuscript drafting and language editing, literature and baseline discovery, research ideation and experimental planning, code preparation and debugging, and the organization and interpretation of experimental results. They also assisted with preparing supporting documentation. All AI-assisted code and experimental procedures were reviewed by the authors, who executed the experiments and verified the reported numerical results against the corresponding logs and evaluation artifacts. References, bibliographic information, and related-work claims were checked by the authors. The authors take full responsibility for the final manuscript, scientific claims, experimental results, code, and associated artifacts.

REPRODUCIBILITY STATEMENT

Section 3 describes counterfactual candidate construction, analytic and learned candidate weighting, the training objectives, and the localization and paired re-evaluation procedure. Appendix A specifies the LPA architecture, candidate features, MOVi-F correspondence filtering, and optimization and inference settings. LPA is trained exclusively on MOVi-F with disjoint training and validation videos, while the CWM predictor and intervention generator remain frozen. Checkpoint selection uses held-out MOVi-F videos, and evaluation on DAVIS, Kinetics, and RoboTAP is performed without dataset-specific adaptation.

Appendices B–D document the evaluation datasets, TAP-Vid First and CMC protocols, metric definitions, and the coordinate and visibility adapters used for external baselines. All methods use the same evaluation subsets and queries, while the CWM variants use the same frozen candidate generator and mask-sampling protocol. Appendix E details the controlled ablations, mask-count and re-evaluation sweeps, candidate-weight analysis, and latency measurement protocol. An overview of the method and results is available on the project page.[1]

ETHICS STATEMENT

This work studies the reliability of motion correspondence recovered from frozen counterfactual world models. We use synthetic MOVi-F trajectories for LPA training and existing DAVIS, Kinetics, and RoboTAP benchmarks for evaluation. Real-world videos may depict people, so downstream applications should respect privacy and avoid non-consensual monitoring. LPA reweights existing motion candidates and cannot guarantee correct trajectories when the underlying candidates are unreliable. Likewise, high completeness or continuity alone does not establish accurate correspondence or physical validity. Applications in safety-critical robotic systems therefore require independent validation and appropriate safeguards. The use and redistribution of datasets, pretrained models, and derived artifacts should comply with their respective licenses and usage conditions.

REFERENCES

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Daniel M. Bear, Kevin T. Feigelis, Honglin Chen, Wanhee Lee, Rahul Mysore Venkatesh, Klemen Kotar, Alex Durango, and Daniel L. K. Yamins. Unifying (machine) vision via counterfactual world modeling, 2023. URL https://arxiv.org/abs/2306.01828.

León Begiristain, Olaf Dünkel, and Adam Kortylewski. CRONOS: Benchmarking counterfactual physical consistency in video models, 2026. URL https://arxiv.org/abs/2605.23699.

[1] https://LPA-CWM.github.io

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Ilona Demler, Saumya Chauhan, and Georgia Gkioxari. Is this tracker on? A benchmark protocol for dynamic tracking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 19446–19455, 2025. doi: 10.52202/085713-0552. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/18093dfe68516361d5b6239d33e045b1-Abstract-Datasets_and_Benchmarks_Track.html.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A benchmark for tracking any point in a video. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13610–13626, 2022. doi: 10.52202/068431-0989. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/58168e8a92994655d6da3939e7cc0918-Abstract.html.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. WorldScore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27713–27724, October 2025. doi: 10.1109/ICCV51701.2025.02573.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3749–3761, June 2022.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2451–2463. Curran Associates, Inc., 2018. URL https://papers.nips.cc/paper/7512-recurrent-world-models-facilitate-policy-evolution.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Adam W. Harley, Yang You, Xinglong Sun, Yang Zheng, Nikhil Raghuraman, Yunqi Gu, Sheldon Liang, Wen-Hsuan Chu, Achal Dave, Suya You, Rares Ambrus, Katerina Fragkiadaki, and Leonidas Guibas. AllTracker: Efficient dense point tracking at high resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5253–5262, October 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin,

Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818, 2024. doi: 10.1109/CVPR52733.2024.02060.

Hyeonho Jeong, Chun-Hao P. Huang, Jong Chul Ye, Niloy J. Mitra, and Duygu Ceylan. Track4Gen: Teaching video diffusion models to track points improves video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7276–7287, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jeong_Track4Gen_Teaching_Video_Diffusion_Models_to_Track_Points_Improves_Video_CVPR_2025_paper.html.

Zhenyu Jiang, Hanwen Jiang, and Yuke Zhu. Doduo: Learning dense visual correspondence from unsupervised semantic-aware flow. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12420–12427, 2024. doi: 10.1109/ICRA57147.2024.10611587. URL https://ut-austin-rpl.github.io/Doduo/.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, October 2025.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset, 2017. URL https://arxiv.org/abs/1705.06950.

Seungwoo Kim, Khai Loong Aw, Klemen Kotar, Cristobal Eyzaguirre, Wanhee Lee, Yunong Liu, Jared Watrous, Stefan Stojanov, Juan Carlos Niebles, Jiajun Wu, and Daniel L. K. Yamins. Taming generative video models for zero-shot optical flow extraction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4335. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bcc97a47e54a9bec240cf5e5aadfd8fb-Abstract-Conference.html.

Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, João Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. TAPVid-3D: A benchmark for tracking any point in 3D. In *Advances in Neural Information Processing Systems*, volume 37, pp. 82149–82165, 2024. doi: 10.52202/079017-2611. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9566607d423f8c32a2d5ce09a8b62232-Abstract-Datasets_and_Benchmarks_Track.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Xinghui Li, Jingyi Lu, Kai Han, and Victor Adrian Prisacariu. SD4Match: Learning to prompt stable diffusion model for semantic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27558–27568, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_SD4Match_Learning_to_Prompt_Stable_Diffusion_Model_for_Semantic_Matching_CVPR_2024_paper.html.

Jiahe Liu, Youran Qu, Qi Yan, Xiaohui Zeng, Lele Wang, and Renjie Liao. Fréchet video motion distance: A metric for evaluating motion consistency in videos. In *First Workshop on Controllable Video Generation at ICML 2024*, 2024. URL https://openreview.net/forum?id=tTZ2eAhK9D.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958, March 2026. doi: 10.1109/WACV61042.2026.00099.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS challenge on video object segmentation. *arXiv:1704.00675*, 2017. URL https://arxiv.org/abs/1704.00675.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Xin Wang, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226, 2026. ISSN 2666-6510. doi: 10.1016/j.aiopen.2026.07.002. URL https://www.sciencedirect.com/science/article/pii/S2666651026000136.

Ayush Shrivastava and Andrew Owens. Self-supervised any-point tracking by contrastive random walks. In *Computer Vision – ECCV 2024*, volume 15072 of *Lecture Notes in Computer Science*, pp. 267–284. Springer, 2024. doi: 10.1007/978-3-031-72630-9_16.

Ayush Shrivastava, Sanyam Mehta, Daniel Geng, and Andrew Owens. Point prompting: Counterfactual tracking with video diffusion models. In *International Conference on Learning Representations*, 2026.

Stefan Stojanov, David Wendt, Seungwoo Kim, Rahul Venkatesh, Kevin Feigelis, Klemen Kotar, Khai Loong Aw, Jiajun Wu, and Daniel L. K. Yamins. Self-supervised learning of motion concepts by optimizing counterfactuals. In *Advances in Neural Information Processing Systems*, volume 38, pp. 31090–31115, 2025. doi: 10.52202/085713-0921. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/277f69e1ef71224fd604f87175822f61-Abstract-Conference.html.

Austin Stone, Daniel Maurer, Alper Ayvaci, Anelia Angelova, and Rico Jonschkowski. SMURF: Self-teaching multi-frame unsupervised RAFT with full-image warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3887–3896, June 2021.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020*, volume 12347 of *Lecture Notes in Computer Science*, pp. 402–419. Springer, 2020. doi: 10.1007/978-3-030-58536-5_24.

Mel Vecerik, Carl Doersch, Yi Yang, Todor Davchev, Yusuf Aytar, Guangyao Zhou, Raia Hadsell, Lourdes Agapito, and Jon Scholz. RoboTAP: Tracking arbitrary points for few-shot visual imitation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5397–5403, 2024. doi: 10.1109/ICRA57147.2024.10611409. URL https://ieeexplore.ieee.org/document/10611409/.

Rahul Venkatesh, Honglin Chen, Kevin Feigelis, Daniel M. Bear, Khaled Jedoui, Klemen Kotar, Felix J. Binder, Wanhee Lee, Sherry Liu, Kevin A. Smith, Judith E. Fan, and Daniel L. K. Yamins. Understanding physical dynamics with counterfactual world modeling. In *Computer Vision – ECCV 2024*, volume 15082 of *Lecture Notes in Computer Science*, pp. 368–387. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72691-0_21.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, efficient, accurate RAFT for optical flow. In *Computer Vision – ECCV 2024*, volume 15065 of *Lecture Notes in Computer Science*, pp. 36–54. Springer, 2024.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for AI-generated video detection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 193477–193524, 2025. doi: 10.52202/085713-5810. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ff7c9c90030f1eb3cbf5c81b6fbd9a05-Abstract-Conference.html.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL https://arxiv.org/abs/2503.21755.

A Implementation Details

Frozen candidate generator. The counterfactual predictor and intervention generator remain fixed during LPA training and inference. CWM operates internally at 256×256 resolution and uses a 32×32 patch-grid mask, where each mask element corresponds to one 8×8-pixel image patch; approximately 89.9% of the target-frame patch grid is hidden. For clarity, Section 3 uses the canonical 256×256 CWM grid for notation; in the implementation, responses and candidate coordinates are mapped to the evaluator grid $H \times W$ before LPA input construction. Its candidate endpoint is obtained from the peak of the channel-reduced absolute RGB response. Averaging rather than summing over RGB channels changes only a constant factor and therefore does not change the peak location.

LPA architecture. The complete LPA module contains 3,000,257 trainable parameters. The main text uses height–width–channel notation, whereas the PyTorch implementation stores image-like tensors in channel-first format. The local candidate encoder therefore receives a $10 \times 32 \times 32$ tensor and applies four 3×3 convolutional blocks with stride 2 and padding 1, using channel dimensions $10 \to 32 \to 64 \to 128 \to 192$. Each block uses GroupNorm and GELU, followed by global average pooling to produce a 192-dimensional local feature.

Before global encoding, the complete normalized single-channel response is bilinearly resized from the evaluator grid $H \times W$ to 32×32, and the concatenated source–target RGB frames are bilinearly resized to 64×64, both with `align_corners=False`. The global response encoder therefore receives a $1 \times 32 \times 32$ tensor and uses four stride-2 convolutional blocks with channel dimensions $1 \to 16 \to 32 \to 64 \to 96$, with 3×3 kernels, padding 1, GroupNorm, and GELU, followed by global average pooling to produce a 96-dimensional feature. The shared frame-pair encoder receives the resulting $6 \times 64 \times 64$ tensor and uses four analogous blocks with channel dimensions $6 \to 32 \to 64 \to 128 \to 192$, producing a 192-dimensional shared context feature.

The scalar encoder maps the 16-dimensional candidate descriptor through an MLP with dimensions $16 \to 64 \to 64$, followed by GELU and LayerNorm, to obtain a 64-dimensional feature. The four branches are concatenated into a 544-dimensional representation and projected by $\mathrm{Linear}(544, 256)$, LayerNorm, and GELU to form a 256-dimensional candidate token. The set encoder contains four pre-normalized Transformer layers with eight attention heads, a 512-dimensional feed-forward block, dropout 0.1, and GELU activation. No candidate-index positional embedding is used. The shared scoring head applies $\mathrm{LayerNorm}(256)$, $\mathrm{Linear}(256, 128)$, GELU, and $\mathrm{Linear}(128, 1)$, followed by a softmax over the candidate dimension.

Candidate statistics. The scalar branch receives the 16-dimensional descriptor listed in Table A.1. These statistics complement the learned visual and response encoders with compact information about displacement, response concentration, ambiguity, photometric change, and local edge structure.

Table A.1: **Candidate scalar descriptor used by LPA.** The support column indicates the spatial region used to compute each statistic.

Dim.	Statistic	Definition	Support
1	Normalized vertical displacement	Candidate y displacement normalized by image height	Global
2	Normalized horizontal displacement	Candidate x displacement normalized by image width	Global
3	Endpoint response	Response magnitude at the candidate endpoint	Single pixel
4	Global response mean	Mean response magnitude	Full response
5	Global response standard deviation	Population standard deviation of response magnitude	Full response
6	Global response median	Median response magnitude	Full response
7	Local peak ratio (7×7)	Endpoint response divided by local mean response	7×7
8	Local peak ratio (15×15)	Endpoint response divided by local mean response	15×15
9	Peak-to-median ratio	Endpoint response divided by global response median	Full response
10	Peak-to-second-peak ratio	Endpoint response divided by the strongest response outside an 11×11 exclusion region	Global + local exclusion
11	Local response mass ratio	Local response mass divided by total response mass	7×7
12	Local response mass ratio	Local response mass divided by total response mass	15×15
13	Local response mass ratio	Local response mass divided by total response mass	31×31
14	Normalized spatial entropy	Entropy of the normalized spatial response distribution	Full response
15	Local photometric change	Mean frame difference around the candidate endpoint	7×7
16	Local source-edge strength	Mean source-frame edge magnitude around the candidate endpoint	7×7

Dimensions 3–10 first undergo signed-log compression, using $x \mapsto \operatorname{sign}(x)\log(1+|x|)$. All 16 dimensions are then standardized using population means and standard deviations computed over all candidate rows in the MOVi-F training split; the same normalization statistics are stored with the checkpoint and reused at inference. Candidate-centered scalar statistics use the rounded and image-clipped candidate endpoint. Local windows are clipped to valid image support without padding, and local means are normalized by the number of valid pixels.

Local patch sampling. Each local visual patch is centered at the source query or candidate endpoint, covers a continuous 64×64 spatial extent on the evaluator grid, and is bilinearly sampled onto a 32×32 grid with `align_corners=True`. If the sampling region extends beyond the image domain, out-of-bounds values use border replication; candidate centers are neither discarded nor recentered.

MOVi-F training correspondence filtering. Training samples are drawn from adjacent frames in the official MOVi-F training split using the native 128×128 dense ground-truth forward and backward flow fields. For a native-resolution source point $\mathbf{p}$, let

$$\mathbf{p}' = \mathbf{p} + \mathbf{F}_{t\to t+1}(\mathbf{p}).$$

The backward field is read at the rounded target coordinate, without bilinear interpolation, and the forward–backward consistency error is

$$e_{\mathrm{fb}}(\mathbf{p}) = \left\|\mathbf{F}_{t\to t+1}(\mathbf{p}) + \mathbf{B}_{t+1\to t}\left(\operatorname{round}(\mathbf{p}')\right)\right\|_2 .$$

Here, $\mathbf{F}_{t\to t+1}$ and $\mathbf{B}_{t+1\to t}$ denote the ground-truth forward and backward flow fields at the native 128×128 resolution, respectively.

A correspondence is retained only when $e_{\mathrm{fb}} \le 2.0$ pixels at the native 128×128 resolution and both the source and warped target are at least four pixels from the image boundary. Out-of-frame targets, points too close to the boundary, and non-finite consistency errors are discarded rather than clamped back into the valid domain. The sampling procedure additionally requires a minimum motion magnitude of 0.25 pixels at the native resolution. Accepted frames and coordinates are subsequently mapped to the common 256×256 training scale.

LPA training. MOVi-F is partitioned at the source-video level before frame-pair and query sampling, and each training sample retains the complete $M = 10$ candidate set. Candidate order is randomly permuted during training. AdamW uses a learning rate of 3×10^{-4}, batch size 16, and a cosine schedule. Checkpoint selection uses candidate-space weighted-coordinate EPE on held-out MOVi-F validation videos. This criterion corresponds to the weighted-coordinate supervision in Section 3.3 and is distinct from the final deployed endpoint EPE reported in the M–R configuration study.

Analytic precursor. APA-CWM uses a frozen Ablated Diffusion Model (ADM) score network with approximately 552M parameters, evaluated in FP16 at 224×224 resolution and diffusion time $t_s = 0.006$. The score map P is computed once per frame pair using the same numerical stabilizer 10^{-6} as in the main text and bilinearly resampled to the aligned CWM response grid. APA averages the aligned statistic within a radius-six neighborhood around each candidate response peak, clips the candidate values to $[10^{-3}, 10^{3}]$, and applies L1 normalization across candidates. The score map is used only by APA-CWM; LPA-CWM does not consume the diffusion score map.

Counterfactual decoding. For each frame pair and query, M denotes the number of mask-conditioned candidate responses in the initial CWM set, whereas R denotes the number of paired re-evaluation passes after the initial localization. The two hyperparameters therefore control different stages of inference.

When $R = 0$, only the initial candidate set is weighted and localized, with no additional CWM pass. For the default $R = 1$, the decoder-specific initial endpoint conditions one coarse-conditioned paired re-evaluation. The source location remains the original query, while the initial endpoint determines the center of the heuristic masking applied to the target frame. At 256×256 resolution, both source and target retain complete spatial support. A fresh set of M masks is sampled, and the frozen CWM recomputes M mask-conditioned responses. APA and LPA are not rerun, the initial candidate

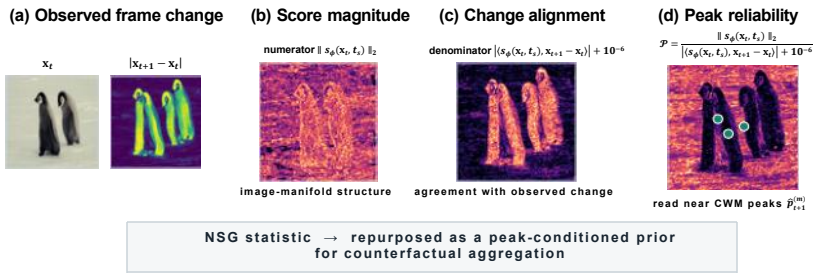


Figure A.1: Evidence chain for the NSG-derived analytic precursor on a real frame pair. The reference frame and observed change define the temporal evidence; score magnitude and score–change alignment produce the analytic reliability map P. APA-CWM averages this map within a local neighborhood around CWM candidate peaks $\hat{p}_{t+1}^{(m)}$. LPA-CWM does not use the diffusion score map; it learns candidate reliability from the features described in Section 3.3. Heatmaps are independently normalized for visualization.

weights are not reused, and the fresh responses are uniformly aggregated. The discrete peak of the resulting response distribution gives the refined endpoint. Thus, the default $R = 1$ stage is a full-field coarse-conditioned paired re-evaluation rather than a local crop refinement.

For $R > 1$, later passes may become progressively local. The crop side follows $s_r = s_1 \eta^{r-1}$ with fixed $\eta = 0.75$; η only controls crop size and is not learned or used in candidate weighting. For non-square inputs, the first square crop uses the shorter image dimension and is shifted to remain within the valid region rather than padded. Here, $r = 1, \ldots, R$ indexes the re-evaluation pass, s_r is its crop side length, and s_1 is the first-pass crop side length, equal to 256 pixels for our square inputs.

Standard and windowed localization. Both decoders start from the same aggregated response distribution using the softmax inverse-temperature parameter $\beta_\pi = 200$. Standard localization applies no Gaussian-window update and passes the discrete argmax of the unwindowed response distribution to the subsequent R stage. Windowed localization performs two Gaussian-window reweighting steps. At each step, the continuous distribution mean and spatial standard deviation determine the window center and scale, with width $3\max(\sigma, 0.5)$; the distribution is renormalized after each update. The coordinate passed to the R stage is the discrete argmax after the second update, not the continuous expectation. The two window updates operate within a fixed response field and are distinct from the paired CWM re-evaluation passes.

Trajectory construction. Under the reported TAP-Vid First and CMC evaluations, the first-visible point remains the source query. For each evaluated target frame, CWM independently processes the corresponding query–target pair, and the complete trajectory is assembled from these direct predictions. Predictions at different target times are not recursively propagated.

CWM visibility. LPA does not predict visibility. Let Δ_m^{final} denote the three-channel response of mask m from the final paired CWM pass. We compute

$$s_m = \max_{\mathbf{u}} \left\| \Delta_m^{\text{final}}(\mathbf{u}) \right\|_1, \qquad s = \frac{1}{M} \sum_{m=1}^{M} s_m, \qquad \hat{v} = \mathbb{1}[s \geq 0.05].$$

We retain the fixed response-strength threshold of 0.05 used by the original CWM evaluator; it is not tuned on any evaluation dataset. This frozen CWM visibility rule is shared by Uniform CWM, APA-CWM, and LPA-CWM. The optional forward–backward occlusion refinement is disabled in all reported CWM experiments.

B Datasets and Evaluation Protocols

Evaluation datasets. We evaluate TAP-Vid DAVIS and Kinetics (Doersch et al., 2022) and Robo-TAP (Vecerik et al., 2024), spanning real-world video and embodied manipulation. DAVIS emphasizes

deformation, occlusion, distractors, and fast object motion, whereas Kinetics contains diverse in-the-wild actions and camera motion. RoboTAP extends the evaluation to robot-manipulation settings. Across the three datasets, our evaluation covers 90 videos and nearly 1,300 annotated point trajectories in total. The same fixed evaluation subsets and query tracks are used for all methods, and no model parameter is adapted to any evaluation dataset.

Evaluation protocol. Following the TAP-Vid First protocol (Doersch et al., 2022), each query is introduced at its first visible frame and evaluated on subsequent frames. Ground-truth coordinates and visibility are taken directly from the corresponding point-track annotations. The same query sets are used for Uniform CWM, APA-CWM, LPA-CWM, and all external baselines, and missing predictions on annotated visible points are retained as failures rather than removed from evaluation. The CWM variants use the same locked videos, sampled frames, first-visible queries, target frames, frozen candidate generator, and mask-sampling protocol. Initial candidate weighting is method specific. Because different weighting can produce different initial endpoints, the paired re-evaluation is computed separately for each variant at its own endpoint.

TAP-Vid metrics. We report Average Jaccard, points within the benchmark distance thresholds, mean (AD) and median (MD) endpoint distance, and occlusion accuracy (OA). Average Jaccard (AJ) jointly reflects localization and visibility, while the distance metrics separately summarize typical and tail localization errors. CMC uses the same ground-truth tracks but evaluates motion correspondence from the complementary perspectives of localization, completeness, visibility, and temporal continuity, as detailed in Appendix C.

C Completeness-aware Motion Correspondence

We introduce *Completeness-aware Motion Correspondence* (CMC) to evaluate motion correspondence on ground-truth dynamic tracks while explicitly penalizing incomplete predictions. Dynamic tracks are selected from ground truth, and missing visible-point predictions remain in the accuracy denominator. The following metrics are computed separately for each video and then averaged with equal video weight.

For all three datasets, ground-truth coordinates and visibility labels are read from the corresponding point-track annotations. We use TAP-Vid DAVIS and Kinetics tracks for the two real-world video benchmarks and RoboTAP annotations for RoboTAP. The same ground-truth trajectories are used by all compared methods.

Let q_i be the query frame of trajectory i, $x_{i,t}^{\mathrm{gt}}$ and $\hat{x}_{i,t}$ the ground-truth and predicted coordinates, and $v_{i,t}^{\mathrm{gt}}, \hat{v}_{i,t} \in \{0,1\}$ their visibility labels. For completeness and localization, we evaluate post-query, ground-truth-visible frames:

$$E_i = \{t : t > q_i,\ v_{i,t}^{\mathrm{gt}} = 1\}.$$

The post-query ground-truth path length is

$$L_{i,>q}^{\mathrm{gt}} = \sum_{t \geq q_i} \mathbf{1}[v_{i,t}^{\mathrm{gt}} = v_{i,t+1}^{\mathrm{gt}} = 1] \|x_{i,t+1}^{\mathrm{gt}} - x_{i,t}^{\mathrm{gt}}\|_2.$$

The dynamic set $\mathcal{D} = \{i : L_{i,>q}^{\mathrm{gt}} \geq 4\}$ is defined on a common 256-pixel coordinate scale and is independent of all predictions.

Completeness. Per-track visibility coverage and its thresholded counterpart are

$$C_i = \frac{\sum_{t \in E_i} \mathbf{1}[\hat{v}_{i,t} = 1]}{|E_i|}, \qquad \text{Coverage} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} C_i,$$

$$\text{Complete@0.8} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathbf{1}[C_i \geq 0.8].$$

These are completeness diagnostics, not localization-accuracy measures.

Dynamic Correspondence Accuracy. With $e_{i,t} = \|\hat{x}_{i,t} - x_{i,t}^{\mathrm{gt}}\|_2$, we define

$$\mathrm{DCA}_\tau = \frac{\sum_{i\in\mathcal{D}} \sum_{t\in E_i} \mathbf{1}[\hat{v}_{i,t} = 1]\mathbf{1}[e_{i,t} < \tau]}{\sum_{i\in\mathcal{D}} |E_i|}.$$

Thus a prediction marked invisible contributes zero rather than disappearing from the denominator. The primary scalar summary is

$$\mathrm{DCA}_{\mathrm{avg}} = \frac{1}{5} \sum_{\tau\in\{1,2,4,8,16\}} \mathrm{DCA}_\tau.$$

Complete Track Success. Let $\bar{e}_i^{\mathrm{joint}}$ be the mean EPE over frames in E_i that are also predicted visible. We set $\bar{e}_i^{\mathrm{joint}} = +\infty$ when no jointly visible frame exists. The track-level criterion is

$$\mathrm{CTS}(\tau, c) = \frac{1}{|\mathcal{D}|} \sum_{i\in\mathcal{D}} \mathbf{1}[C_i \geq c]\mathbf{1}[\bar{e}_i^{\mathrm{joint}} \leq \tau].$$

We report CTS@(4,0.8) and CTS@(8,0.8). CTS is secondary to DCA because it still allows up to 20% of a track to be absent.

Visibility and continuity. Visibility metrics use all post-query frames of the selected dynamic tracks, including ground-truth-occluded frames. Visibility balanced accuracy is

$$\mathrm{VisBA} = \tfrac{1}{2}(\mathrm{TPR}_{\mathrm{visible}} + \mathrm{TNR}_{\mathrm{occluded}}),$$

and occlusion F1 is computed from post-query occlusion labels.

For each track, longest continuous coverage (LCC) is the longest predicted-visible run within any ground-truth-visible segment, divided by $|E_i|$. Fragmentation counts the number of $1 \to 0 \to 1$ visibility recoveries within such segments; Table C.2 reports the mean count.

Table C.1: **Full threshold-wise CMC accuracy results on DAVIS, Kinetics, and RoboTAP.** All entries use the same 0–100 scale as Table 1. Cov. and C@.8 denote Coverage and Complete@0.8; CTS4 and CTS8 denote CTS@(4,0.8) and CTS@(8,0.8), respectively. Results follow the presentation convention of Table 1.

Dataset	Method	Cov.↑	C@.8↑	DCA@1↑	DCA@2↑	DCA@4↑	DCA@8↑	DCA@16↑	DCA↑	CTS4↑	CTS8↑
DAVIS	SMURF	91.98	85.54	1.72	4.11	8.72	14.07	27.62	11.25	2.67	3.33
	GMRW	37.02	20.62	11.47	21.20	26.10	28.57	32.25	23.92	17.00	17.00
	DODUO	31.44	15.04	6.76	18.23	25.89	29.59	30.79	22.25	14.42	15.04
	CoTracker3	92.31	85.33	40.55	64.40	80.49	85.20	88.41	71.81	70.50	74.82
	RAFT	68.74	49.58	29.93	46.92	53.02	56.58	58.53	49.00	31.77	34.59
	SEA-RAFT	57.76	31.25	27.93	40.57	47.35	49.01	50.18	43.01	27.43	30.54
	Uniform CWM	88.10	80.96	6.75	19.30	39.85	52.99	65.46	36.87	21.46	37.96
	APA-CWM (Ours)	94.74	89.98	**19.31**	41.71	61.91	73.94	82.05	55.79	47.91	58.14
	LPA-CWM (Ours)	**98.21**	**95.63**	17.87	**44.73**	**67.49**	**80.13**	**84.67**	**58.98**	**57.54**	**66.79**
Kinetics	SMURF	89.67	81.89	4.61	9.36	16.87	24.29	40.65	19.16	14.17	14.79
	GMRW	55.28	36.25	11.23	24.66	36.71	41.66	47.66	32.38	30.42	34.38
	DODUO	67.04	49.87	6.95	22.30	46.45	59.92	65.42	40.21	39.87	48.62
	CoTracker3	87.06	79.67	13.04	32.29	54.89	71.99	78.56	50.15	46.61	60.89
	RAFT	80.83	69.24	10.02	26.02	49.49	67.38	71.55	44.89	43.94	60.75
	SEA-RAFT	78.59	65.73	9.49	25.19	47.29	65.01	69.80	43.36	40.71	56.55
	Uniform CWM	87.14	77.11	4.47	20.93	44.69	60.91	70.31	40.26	34.54	50.58
	APA-CWM (Ours)	89.08	83.78	8.44	22.86	47.03	72.16	80.61	46.22	37.10	64.66
	LPA-CWM (Ours)	**94.65**	**89.97**	**9.58**	**35.22**	**61.37**	**72.47**	**80.97**	**51.92**	**58.25**	**67.04**
RoboTAP	SMURF	90.33	81.38	7.53	16.12	22.03	26.82	33.15	21.13	11.67	16.25
	GMRW	30.31	10.89	8.83	15.88	20.65	23.22	23.46	18.41	7.01	8.35
	DODUO	52.31	30.82	8.51	23.62	34.39	44.86	48.28	31.93	18.73	29.57
	CoTracker3	81.02	73.92	26.55	44.04	56.76	61.22	68.43	51.40	43.05	49.70
	RAFT	76.69	54.71	15.95	40.72	58.73	62.72	67.37	49.10	41.27	46.64
	SEA-RAFT	75.58	62.14	16.03	36.57	49.31	52.86	57.48	42.45	24.68	25.31
	Uniform CWM	85.92	79.06	**10.07**	22.22	43.06	63.79	73.71	42.57	30.55	55.11
	APA-CWM (Ours)	91.10	81.78	6.33	21.73	51.54	75.91	**79.05**	46.91	42.34	59.53
	LPA-CWM (Ours)	**92.00**	**86.88**	6.47	**29.41**	**59.94**	**76.56**	78.93	**50.26**	**54.22**	**66.00**

Table C.2: **Visibility and continuity diagnostics for CMC on DAVIS, Kinetics, and RoboTAP.** OF1 denotes occlusion F1. VisBA, OF1, and LCC are percentages; Frag. is the mean fragmentation count $\times 100$. Results follow the presentation convention of Table 1.

Method	DAVIS				Kinetics				RoboTAP			
	VisBA↑	OF1↑	LCC↑	Frag.↓	VisBA↑	OF1↑	LCC↑	Frag.↓	VisBA↑	OF1↑	LCC↑	Frag.↓
SMURF	56.64	14.92	83.29	21.54	62.69	35.11	82.40	9.66	57.17	28.66	78.28	16.35
GMRW	63.24	45.22	35.46	8.46	65.99	46.34	51.43	7.16	61.48	35.21	28.76	7.59
DODUO	58.94	45.83	28.44	12.88	78.37	58.39	59.83	16.82	73.28	53.09	47.15	17.33
CoTracker3	76.43	58.97	81.01	1.25	78.00	58.56	80.41	8.41	73.11	59.35	75.21	5.05
RAFT	70.16	41.66	62.02	7.62	74.93	60.72	71.71	9.38	70.60	44.27	69.46	22.24
SEA-RAFT	65.78	39.56	54.07	9.38	72.67	51.96	70.65	10.91	65.23	37.86	64.72	24.58
Uniform CWM	59.91	33.91	85.20	3.57	67.66	44.97	79.95	10.55	72.12	50.77	76.89	11.29
APA-CWM (Ours)	62.07	38.48	90.13	2.86	77.66	59.22	84.60	4.03	75.63	57.44	78.71	7.29
LPA-CWM (Ours)	**75.89**	**48.14**	**93.18**	**0.63**	**82.97**	**62.67**	**86.84**	**2.78**	**78.03**	**61.00**	**85.49**	**2.59**

Threshold sensitivity. Figure C.1 evaluates whether the CMC conclusions depend on the default dynamic-motion and coverage thresholds. We vary the ground-truth motion threshold $L_{\text{motion}} \in \{2, 4, 8\}$ pixels and report DCA_{avg}. LPA-CWM changes only from 59.42 to 58.30 across this range and remains above APA-CWM and Uniform CWM at every setting. We then fix $L_{\text{motion}} = 4$ pixels and vary the coverage threshold $c \in \{0.6, 0.7, 0.8, 0.9\}$ for CTS@(4,c). LPA-CWM remains best throughout, decreasing moderately from 58.79 to 56.29 as the coverage requirement becomes stricter. The method ordering is therefore stable around the default $L_{\text{motion}} = 4$ pixels and $c = 0.8$.

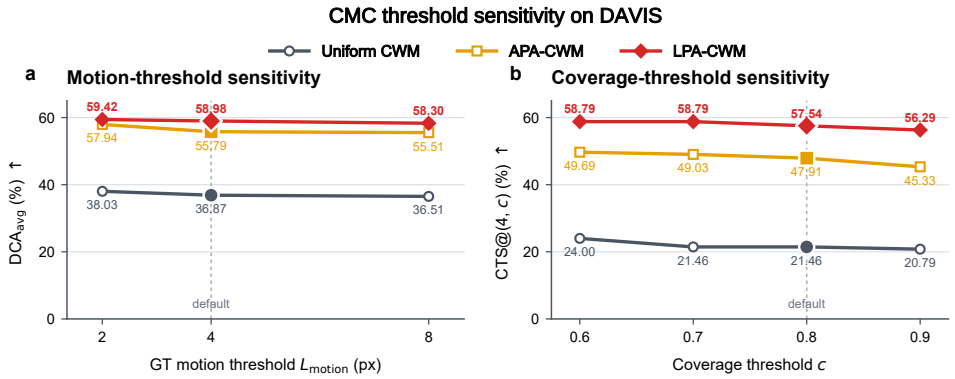


Figure C.1: **CMC threshold sensitivity on DAVIS. (a)** DCA_{avg} as the dynamic-track motion threshold varies over $\{2, 4, 8\}$ pixels. **(b)** CTS@(4,c) as the coverage threshold varies over $\{0.6, 0.7, 0.8, 0.9\}$ with $L_{\text{motion}} = 4$ pixels. Dashed vertical lines indicate the default settings.

D Adapting External Estimators

Temporal and coordinate adaptation. For the reported TAP-Vid First and CMC evaluations, the CWM variants and the pairwise estimators RAFT, SEA-RAFT, DODUO, and SMURF use a direct query-to-target construction. For a query introduced at frame q, each target $t > q$ is predicted independently from the fixed first-visible query and the corresponding pair (I_q, I_t). The four pairwise baselines use their released inference implementations and checkpoints within this adapter. CoTracker3 instead uses its official offline point-tracking interface, which jointly processes the full video and the first-visible query, while GMRW uses its released adjacent-frame tracking utility. Thus, all methods are evaluated on the same query identities and target frames, while their native temporal inference mechanisms remain method specific.

Visibility. Table D.1 summarizes the visibility rules used by the external estimators. CoTracker3 uses the Boolean visibility returned by its official offline predictor, which thresholds the native

visibility score at 0.9 and forces the annotated query point to be visible. No additional forward–backward consistency test is applied.

Table D.1: **Visibility adaptation for external estimators.** Thresholds are defined on the corresponding evaluator grid.

Method	Visibility source	Threshold	Sampling
RAFT	Direct query-to-target forward–backward consistency	6 px	Forward at integer query; backward bilinear
SEA-RAFT	Direct query-to-target forward–backward consistency	6 px	Forward at integer query; backward bilinear
DODUO	Direct query-to-target forward–backward consistency	6 px	Forward at integer query; backward bilinear
CoTracker3	Native visibility from official offline predictor	0.9	Native sparse-query trajectory; no additional FB test
SMURF	Native occlusion map	0.5	Query-location lookup
GMRW	Official tracking utility with adjacent-frame forward–backward consistency	3 px	Bilinear at propagated positions

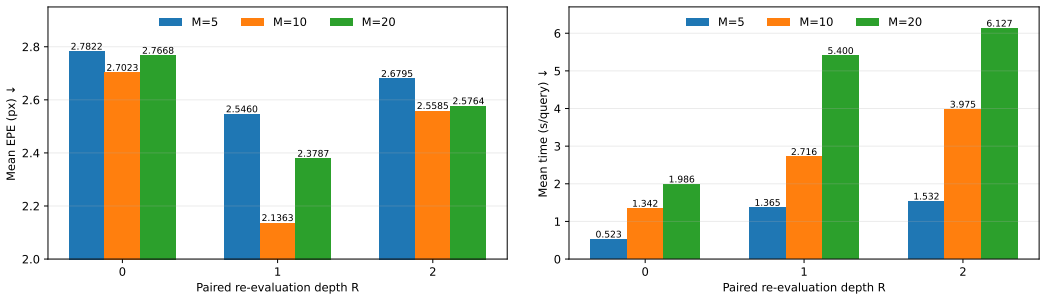
For the flow and correspondence estimators, out-of-frame or non-finite predictions are treated as invisible by the corresponding adapters. CoTracker3 retains the native visibility returned by the official predictor without an additional out-of-frame, non-finite, or forward–backward filter. These evaluation-time visibility rules are distinct from the ground-truth forward–backward filtering used to construct MOVi-F training samples in Appendix A.

Aggregation. All metrics are first computed from the same per-video query set and then aggregated with equal video weight. No point or video is selected using a method's predicted error, coverage, or visibility. Missing predictions enter DCA as failures.

E Ablation Protocols

Joint M–R sensitivity. The audit uses 40 query groups from 10 held-out MOVi-F videos, shared across configurations. We evaluate $(M, R) \in \{5, 10, 20\} \times \{0, 1, 2\}$. Masks are nested prefixes of the same ordered 20-mask set, avoiding independent draws for smaller M. All runs use full LPA-CWM on identical videos, frame pairs, and queries. We report final endpoint EPE after localization and R paired re-evaluations. The sweep selects $M = 10, R = 1$ for the main comparison.

Latency measurement. Figure E.1 uses a single NVIDIA GeForce RTX 5090 with batch size 1, where each timed sample contains one adjacent frame pair and one query group. CUDA is synchronized immediately before and after the complete model call. Timing includes the factual frozen-CWM forward pass, M mask-conditioned counterfactual evaluations, LPA feature extraction and set processing, candidate aggregation, localization, the specified R paired CWM re-evaluation passes, and final endpoint extraction. Disk I/O, CPU preprocessing, host-to-device transfer, and model or checkpoint initialization are excluded. We therefore report end-to-end per-query latency under this protocol rather than throughput-optimized peak speed.



(a) Final deployed endpoint EPE **(b) End-to-end inference time**

Figure E.1: **Joint sensitivity to mask count and paired re-evaluation depth.** Full LPA-CWM results on 40 fixed MOVi-F query groups. **(a)** Final deployed endpoint EPE. **(b)** End-to-end inference time per query. $R = 0$ omits paired re-evaluation; $M = 10, R = 1$ is the deployed setting.

Candidate-weight analysis. Candidates are ranked within each complete ten-candidate group by ground-truth endpoint EPE. The top-1 oracle hit rate measures how often the highest-weight candidate is the minimum-EPE candidate. Effective support is $\exp(-\sum_m w_m \log w_m)$: it equals 1 for one-hot weights and 10 for uniform weights. Figure 6 reports video-macro summaries. This analysis uses the MOVi-F validation split used for checkpoint selection and serves as mechanism evidence; real-video tracking is evaluated separately.

EPE definitions across analyses. Figure E.1 evaluates 40 fixed query groups using the final deployed endpoint after localization and the requested R re-evaluation passes; the default $M = 10, R = 1$ setting obtains 2.1363 pixels. Figure 6 instead evaluates 640 complete $M = 10$ candidate groups using the candidate-space weighted-coordinate surrogate before response-level localization and paired re-evaluation, where LPA obtains 2.41 pixels. These values are not directly comparable because they use different validation subsets and endpoint definitions.

Inference-stage ablation. All six configurations use the same frozen CWM, $M = 10$, $R = 1$, evaluation seed, DAVIS batches, frame pairs, queries, and initial masks. The rows are *Uniform + Std.*, *APA + Std.*, *Uniform + Win.*, *APA + Win.*, *LPA + Std.*, and *LPA + Win.* "Std." uses the discrete peak of the unwindowed aggregated response distribution as the initial endpoint. "Win." applies two Gaussian-window reweighting updates and uses the discrete peak of the final distribution. Uniform, APA, and LPA respectively denote equal, analytic, and learned weights for the initial candidate set. All six rows then use the same one-pass coarse-conditioned paired re-evaluation after their decoder-specific initial endpoint. The second pass does not rerun APA or LPA, does not reuse the initial weights, and uniformly aggregates the freshly generated responses. This 3×2 design therefore separates initial candidate weighting from localization.

Learned architecture and objective ablations. Architecture variants use the complete three-term objective, the same windowed localization, and the same $R = 1$ paired re-evaluation. "w/o candidate-set interaction" removes the permutation-equivariant set Transformer, while "response + scalar cues only" removes the local RGB and shared frame-pair raw encoders but retains response maps, 16 scalar statistics, and candidate-set interaction.

Objective variants retain the full LPA architecture and compare "KL only", "KL + coordinate supervision", "KL + best-candidate supervision", and "Full objective". Table 2 reports every row under both CMC and TAP-Vid First on the same DAVIS evaluation set, queries, and initial mask protocol. The CMC columns summarize completeness, localization, visibility, and continuity, while the TAP-Vid First columns report conventional localization and occlusion measures. The inference procedure is fixed except for the named architecture or objective component.

F Additional Qualitative Results

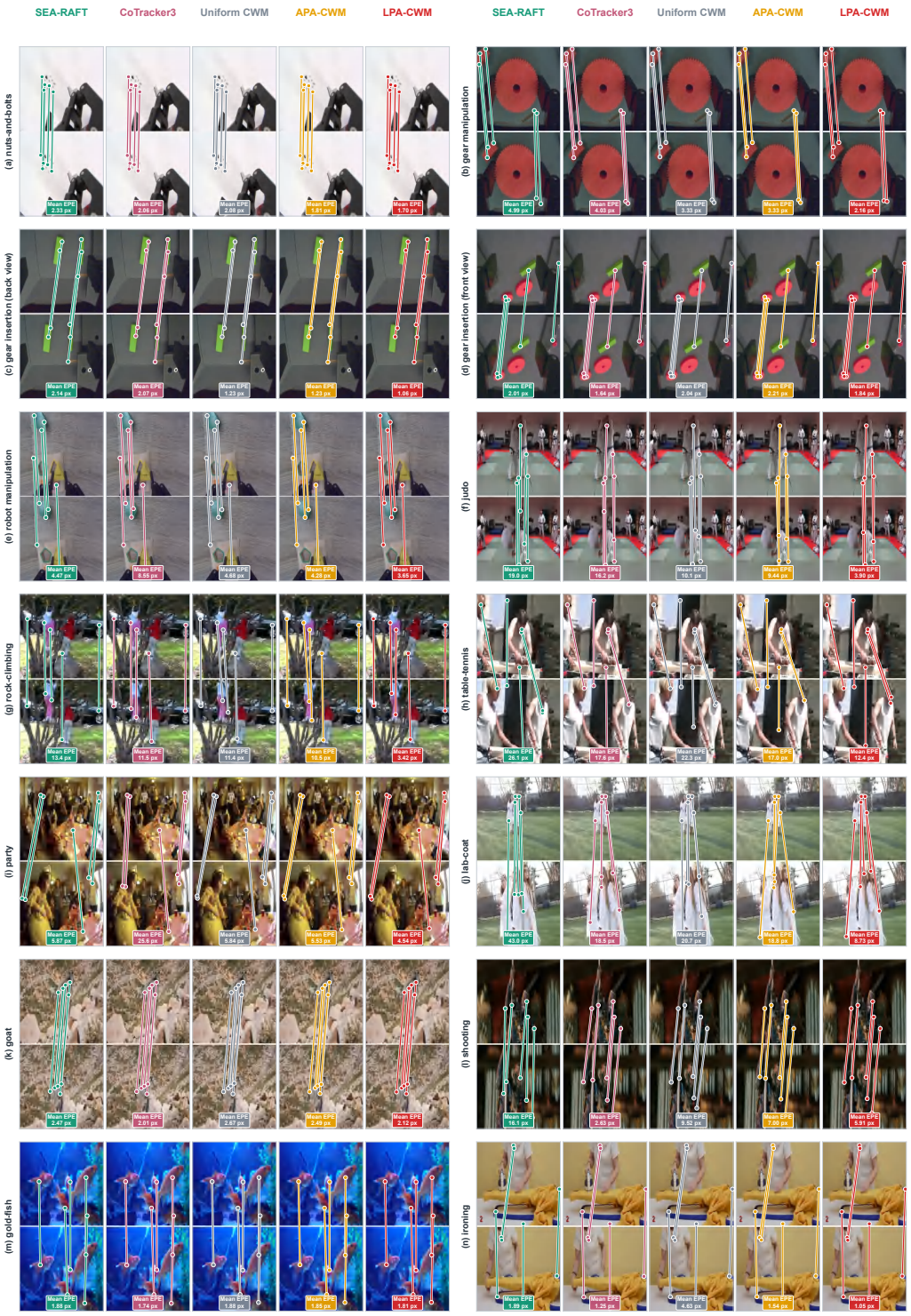


Figure F.1: Additional qualitative results on fourteen cases distinct from Figure 4. Columns show SEA-RAFT, CoTracker3, Uniform CWM, APA-CWM, and LPA-CWM in this order. Every method receives the same frame pair and the same five query identities. A method uses one color throughout; continuous lines connect source queries to target predictions, and badges report mean EPE over the displayed queries.