STEP: SURGICAL POINT TRACKING FROM SPARSE ENDPOINTS AND PSEUDO-TRAJECTORIES

Anonymous authors
Paper under double-blind review

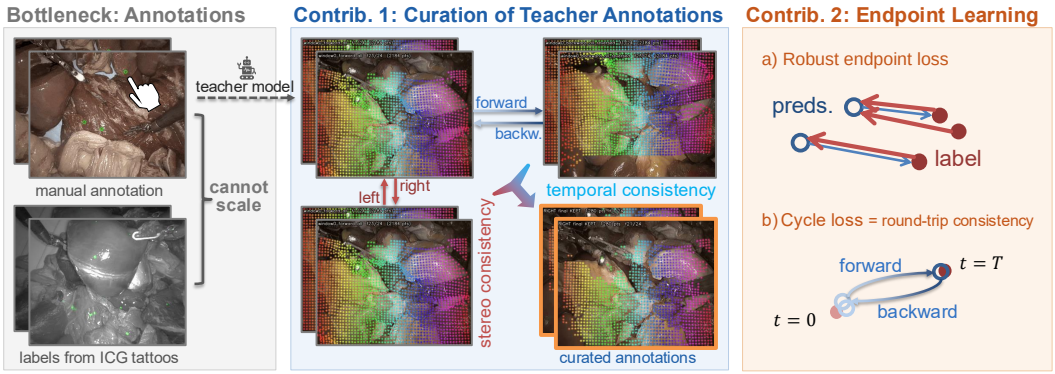

Figure 1: **Scaling supervision for surgical point tracking.** Manual trajectories require repeated annotation. STEP combines curated trajectories and existing physical endpoints to train a causal surgical tracker.

ABSTRACT

Surgical point tracking requires persistent correspondence through tissue deformation and occlusion, yet dense trajectory annotations are costly to obtain. We present STEP, a two-stage framework that learns from surgical video and sparse physical endpoints. First, a frozen offline teacher generates candidate trajectories, and temporal reversibility and stereo consistency select dense supervision for a causal student. Second, a robust physical endpoint loss anchors terminal correspondence, while a round-trip loss constrains temporal consistency. A gradient-free prefix exposes training to accumulated drift, while a final differentiable window bounds activation memory. The trained tracker processes a single RGB stream with fixed weights. In STIR-to-VL-SurgPT transfer, temporal–stereo union selection reduces endpoint error by 7.0% relative to unfiltered trajectories. On a post-hoc diagnostic of 29 VL-SurgPT clips excluded from the endpoint-training manifest, STEP reduces endpoint error from 31.52 to 16.80 pixels relative to dense-label fine-tuning. These results support combining curated trajectories with physical observations to train surgical point trackers.

1 INTRODUCTION

Tracking tissue points through surgical video provides correspondence for motion estimation, deformation modeling, registration, and robotic perception. Tracking becomes difficult when tissue deforms under tool interaction, specular reflections change its appearance, and instruments hide the target. General-purpose trackers learn strong correspondence priors from synthetic and real-world video (Doersch et al., 2023; Karaev et al., 2025; Aydemir et al., 2026); adapting these priors to surgery requires supervision that captures these conditions.

Surgical video offers two complementary sources of supervision. Unlabeled footage captures diverse motion and appearance, but obtaining dense manual trajectories is expensive, especially through deformation and occlusion. Physical markers offer a second resource: STIR and VL-SurgPT record tissue-point positions at the beginning and end of a clip using fluorescent observations (Schmidt et al., 2024a; Zhou et al., 2026). These measurements establish correspondence without tracing every frame. However, they leave the intervening path unobserved, and marker localization, matching, and

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.

switching between fluorescence and white-light imaging can introduce uncertainty. A tracker must learn dense correspondence from video and anchor predictions to sparse measurements.

These signals address different ambiguities. A teacher can produce smooth, reversible tracks that gradually attach to the wrong tissue region, while an accurate endpoint does not specify the route taken through an occlusion. Training only on short windows also avoids the accumulated errors encountered over a complete surgical clip. This motivates combining dense video-derived targets with measured terminal correspondence, and applying endpoint supervision to the student's predictions.

We introduce STEP (Surgical Tracking from Sparse Endpoints and Pseudo-trajectories), which combines these resources in two training stages (Figure 1). The first stage curates dense trajectories from a frozen offline teacher. Temporal reversibility checks whether forward tracking returns to the query under reverse tracking; stereo consistency checks agreement with binocular geometry. The tests accept complementary candidates, so we retain their union on stereo video and use temporal selection on monocular video. The resulting trajectories and visibility masks supervise a causal student throughout each training window.

The second stage uses physical endpoints to refine long-range correspondence. The endpoint loss uses known point identities when available and compares unordered marker sets through robust nearest-neighbor distances otherwise. A round-trip loss adds temporal supervision by requiring tracks to return to their starting locations. To train on long clips with bounded activation memory, the student processes earlier frames without gradients and initializes a final training window from its own predicted positions. Training thus encounters accumulated streaming errors.

Dense teacher trajectories, measured endpoints, and round trips play different roles: they supply local correspondence, anchor terminal positions, and constrain temporal consistency. Both stages update weights before deployment; inference then requires only initial queries and a single RGB video stream. Our main contributions are:

- **Surgical trajectory curation.** Temporal and stereo checks select visibility-aware pseudo-trajectories from surgical video.
- **Endpoint-guided streaming tracking.** Physical endpoint and round-trip objectives train on accumulated prediction error within a bounded gradient window.
- **Empirical findings.** Union selection lowers transfer EPE by 7.0%. On 29 diagnostic clips, STEP lowers Track-On-R EPE by 46.7% versus dense-label fine-tuning.

2 Related Work

General-purpose and surgical tracking. Tracking Any Point estimates the trajectory and visibility of arbitrary queries, with TAP-Vid providing a common evaluation protocol (Doersch et al., 2022). TAPIR refines point correspondences, CoTracker3 reasons jointly over tracks, and MFT composes dense flow over long horizons (Doersch et al., 2023; Karaev et al., 2025; Neoral et al., 2024). Track-On and Track-On-R carry memory across frames for online tracking (Aydemir et al., 2025; 2026). Surgical methods incorporate stereo geometry and deformation (SENDD), adaptive templates (Ada-Tracker), or dense surgical annotations (SurgMotion) (Schmidt et al., 2023; Guo et al., 2024; Zhan et al., 2025); Schmidt et al. (2024b) review this setting. STEP focuses on the supervision needed to adapt pretrained trackers to surgical correspondence. In particular, an offline teacher can use later frames to resolve earlier ambiguities, while a streaming tracker must update its estimate as each frame arrives. STEP transfers the former's predictions into training targets for the latter. Endpoint refinement trains on the student's causal predictions, aligning supervision with streaming inference.

Learning from video and physical observations. BootsTAP and CoTracker3 use predictions on unlabeled video to expand training supervision, while Track-On-R uses a learned verifier to select reliable pseudo-labels (Doersch et al., 2024; Karaev et al., 2025; Aydemir et al., 2026). In surgery, STIR and VL-SurgPT provide physically measured correspondence from tissue markings (Schmidt et al., 2024a; Zhou et al., 2026). STEP connects these resources: consistency-selected teacher trajectories provide dense surgical supervision, and physical endpoints refine terminal correspondence. It uses existing benchmark measurements without collecting new manual trajectories.

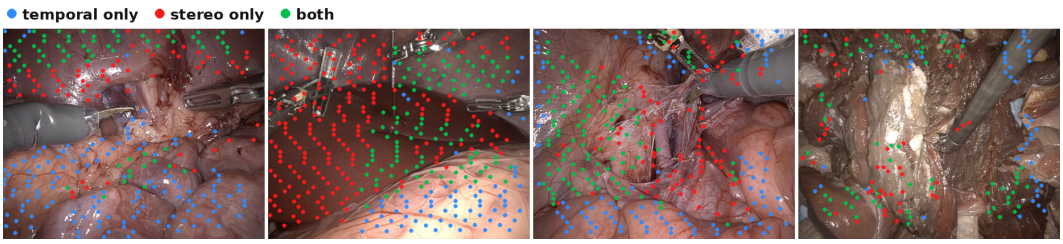

Figure 2: **Temporal and stereo consistency selection.** STIR candidates accepted by temporal checks only (blue), stereo checks only (red), or both (green). Stage 1 retains their union on stereo sequences.

Geometric and temporal consistency. Forward–backward agreement identifies reliable correspondences in unsupervised flow (Meister et al., 2018), while temporal cycles support correspondence learning in TimeCycle and contrastive random walks (Wang et al., 2019; Jabri et al., 2020). Stereo geometry constrains matches across views (Hartley & Zisserman, 2004), with left–right consistency used in depth learning and temporally consistent disparity estimated by Stereo Any Video (Godard et al., 2017; Jing et al., 2025). STEP uses temporal and stereo consistency to select training targets, then combines round-trip supervision with physical endpoints during streaming training.

3 Curated Surgical Supervision

Our curation pipeline converts surgical video into trajectories with per-frame visibility and validity masks. It supports monocular and stereo inputs from STIR, VL-SurgPT, StereoMIS, Hamlyn, and SurgMotion (Schmidt et al., 2024a; Zhou et al., 2026; Hayoz et al., 2023; Hamlyn Centre, Imperial College London; Zhan et al., 2025). The experiments use STIR and VL-SurgPT to study selection and endpoint-guided training; Appendix A describes the source inventories.

3.1 Candidate Trajectories and Consistency Selection

Offline candidate generation. A frozen AllTracker teacher (Harley et al., 2025) predicts trajectories from regular point grids over windows of at most 750 frames. An independently seeded reverse pass supplies correspondences for temporal checking. We retain coordinates and per-frame visibility for each candidate. Visibility is defined by $1 - \text{visibility} \cdot \text{confidence} \leq 0.4$ using the teacher's outputs. The teacher accesses the full window to supervise a causal student.

Temporal consistency selection. We first assess whether a candidate remains consistent under temporal reversal, following the forward–backward principle commonly used in unsupervised flow and correspondence learning (Meister et al., 2018; Wang et al., 2019). For a query q_i at frame t_a, let $F_{a\to b}(q_i)$ denote its predicted location at frame t_b. An independently seeded reverse pass is interpolated at this location to obtain the corresponding return position. We define the temporal round-trip error as

$$e_T(i) = \left\| B_{b\to a}(F_{a\to b}(q_i)) - q_i \right\|_2 . \tag{1}$$

A candidate is included in $\mathcal{T}_T$ if $e_T(i) < \tau_T$ and the correspondence is visible in both passes. We use $\tau_T = 3$ pixels in the teacher coordinate system. Per-frame visibility is retained for selected trajectories, so frames in which the target is occluded do not introduce artificial positional supervision.

Stereo consistency selection. For stereo sequences, we additionally evaluate candidates using cross-view geometry. Calibrated left–right correspondence (Hartley & Zisserman, 2004) is estimated using Stereo Any Video (Jing et al., 2025). Let $d_i(t)$ denote the disparity-induced offset predicted at candidate i and time t, and let $s_i(t)$ denote the offset between its associated trajectories in the two views under the same coordinate convention. Their disagreement is

$$e_S(i,t) = \left\| d_i(t) - s_i(t) \right\|_2 . \tag{2}$$

A stereo correspondence is consistent when $e_S(i,t) < \tau_S$. We associate cross-view tracks within 3 pixels and require disparity agreement within $\tau_S = 8$ pixels over visible frames.

Temporal and stereo consistency provide complementary evidence rather than identical filters. As illustrated in Fig. 2, some candidates satisfy only the temporal criterion (blue), others only the stereo

criterion (red), while a subset satisfies both (green). We therefore retain the union

$$\widetilde{\mathcal{T}} = \mathcal{T}_T \cup \mathcal{T}_S \tag{3}$$

for stereo sequences, rather than requiring their intersection. For monocular sequences, or windows without valid stereo estimates, we use $\widetilde{\mathcal{T}} = \mathcal{T}_T$. The retained per-frame visibility masks determine which point–frame pairs contribute supervision during training.

Training with curated trajectories. The selected trajectories provide pseudo-supervision for Stage 1 (S1). Starting from a released checkpoint, we train the causal student using teacher coordinates $\widetilde{p}_i^t$ only at valid point–frame pairs indicated by w_i^t. The positional component of the objective is

$$\mathcal{L}_{\text{pseudo}} = \frac{\sum_{i,t} w_i^t \, \ell_{\text{track}}(\widehat{p}_i^t, \widetilde{p}_i^t)}{\sum_{i,t} w_i^t}, \tag{4}$$

where ℓ_{track} is the student's native tracking loss. This stage produces initialization θ_1 for endpoint-guided refinement. Section 5.3 evaluates selection through cross-dataset transfer.

The temporal test used here selects fixed teacher targets before optimization. The round-trip objective in Section 4 instead acts on the student's current predictions and updates its parameters. This distinction lets curation exploit an offline teacher's temporal context while endpoint-guided training improves the student that will run causally at deployment.

3.2 Supervision Coverage and Visibility

Each clip-view record stores frame indices, image size, coordinates, per-frame validity, and temporal and stereo acceptance masks. During training, frames and coordinates are resized together; non-finite, out-of-frame, and teacher-occluded positions are excluded from the positional loss. Thus, accepting a trajectory does not imply supervising every point along it.

The generated inventories contain 493,893 candidate trajectories for VL-SurgPT and 1,685,211 for STIR, spanning approximately 103M and 300M point-frames, respectively (Table 6). These inventories precede training selection and split filtering. Each trajectory carries its own validity mask, so increasing coverage does not require treating uncertain or occluded positions as observed coordinates. Together, the source records and masks allow the same candidate pool to support different selection policies without regenerating teacher predictions.

For VL-SurgPT Stage 1, excluding the 150 evaluation clips leaves 604 source clips, of which 596 yield usable windows. STIR selection experiments share 1,726 clip-view windows from 566 stereo clips and use the same optimization budget. Source-specific preprocessing preserves the correspondence between image content and trajectory coordinates. StereoMIS tracks follow the decoded packet timeline, while Hamlyn uses temporal selection and a first-frame grid check to remove static border candidates. Such regions can appear temporally consistent despite providing no tissue-motion supervision. Appendix A details source inventories and preprocessing.

4 Endpoint-Guided Streaming Tracking

The second stage, endpoint-guided training (S2), combines a physical endpoint loss $\mathcal{L}_{\text{end}}$ and a round-trip consistency loss $\mathcal{L}_{\text{cyc}}$, also called the cycle loss. Given a clip $V = \{I_t\}_{t=1}^{T}$ and query points $\mathcal{P}_1$, we adapt θ_1 using measured endpoints $\mathcal{P}_T$ and forward–backward tracks. Both losses update the tracker before deployment (Figure 3). VL-SurgPT supplies paired endpoint identities; STIR supplies unordered endpoint sets whose cardinalities can differ.

4.1 Streaming Training with a Bounded Gradient Window

Training uses the tracker's causal forward pass. Let h_t denote its recurrent state, including point memory. Each new frame updates the state and predictions:

$$(h_t, \widehat{\mathcal{P}}_t, \widehat{v}_t) = f_\theta(I_t, h_{t-1}; \mathcal{P}_1). \tag{5}$$

An endpoint label may be hundreds of frames from its query, whereas storing activations for the entire clip is expensive. We therefore use a truncated backpropagation-through-time approximation (Werbos,

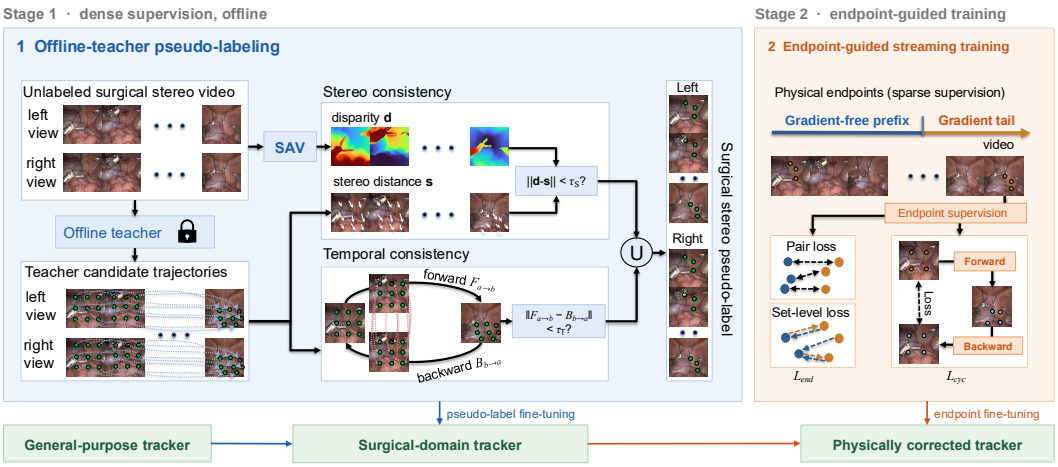

Figure 3: **STEP framework.** (1) Temporal and stereo checks select training trajectories. (2) Endpoint-guided training combines physical endpoint and round-trip losses, with gradients limited to a final window.

1990; Williams & Peng, 1990). We split the clip at $s = T - B$, stream frames $1{:}s$ without gradients, and use the predicted positions to initialize a differentiable B-frame tail:

$$\overline{\mathcal{P}}_s = \mathrm{stopgrad}\big(\mathrm{Stream}_\theta(I_{1:s}, \mathcal{P}_1)\big), \qquad \widehat{\mathcal{P}}_T = G_\theta(I_{s+1:T}; \overline{\mathcal{P}}_s). \tag{6}$$

Here G_θ shares the tracker weights. For Track-On-R, the tail transfers predicted positions and initializes fresh point memory and query descriptors at its boundary. Deployment maintains continuous streaming memory across frames. This truncated training approximation exposes the tail to accumulated drift while limiting stored temporal activations to B frames.

Learning from accumulated error. Initializing the tail from the student's prediction preserves the displacement error accumulated over the prefix. Reinitializing it from a ground-truth point would instead train recovery from an accurate boundary location. Our construction exposes the final window to the former condition and computes endpoint gradients through that window. The prefix therefore supplies realistic starting positions for correction, while the gradient horizon remains independent of the full clip length.

A $K \times K$ support grid supplies context alongside the annotated queries. Only annotated queries contribute to endpoint losses and evaluation.

4.2 Physical Endpoint Loss

Paired endpoints. When identities are available, the prediction for query i is compared directly with its measured endpoint p_i^T:

$$\mathcal{L}_{\mathrm{end}}^{\mathrm{pair}} = \frac{1}{N} \sum_{i=1}^{N} \rho_\delta(\|\widehat{p}_i^T - p_i^T\|_2), \tag{7}$$

where ρ_δ is the Huber robust penalty (Huber, 1964), with $\rho_\delta(d) = d^2/(2\delta)$ for $d \le \delta$ and $\rho_\delta(d) = d - \delta/2$ otherwise. We use $\delta = 8$ pixels in the 640×512 training-coordinate system. The linear tail limits the influence of a widely displaced prediction while preserving a positional correction. Only valid physical endpoints contribute to the average.

Physical observations can contain localization noise, particularly when marker acquisition and white-light frames are not perfectly aligned. The Huber penalty provides quadratic refinement near a measured point and a bounded derivative for large residuals. This retains a correction signal for displaced predictions without letting one endpoint dominate the update.

Unordered endpoints. For observations without persistent start-to-end identities, we use a bidirectional robust nearest-neighbor distance between predicted and observed endpoint sets, analogous to Chamfer-style point-set objectives (Fan et al., 2017):

$$\mathcal{L}_{\mathrm{end}}^{\mathrm{set}} = \frac{1}{|\widehat{\mathcal{P}}_T|} \sum_{\widehat{p} \in \widehat{\mathcal{P}}_T} \rho_\delta\left(\min_{p \in \mathcal{P}_T} \|\widehat{p} - p\|_2\right) + \frac{\beta}{|\mathcal{P}_T|} \sum_{p \in \mathcal{P}_T} \rho_\delta\left(\min_{\widehat{p} \in \widehat{\mathcal{P}}_T} \|p - \widehat{p}\|_2\right). \tag{8}$$

The prediction-to-observation term aligns tracks with measured markers; the reverse term, weighted by $\beta = 0.5$, encourages coverage of the observed set. This accommodates unequal set sizes. When endpoint identities are supplied, Equation 7 preserves them: a permutation of two correctly located markers can have zero set distance while still assigning each query incorrectly.

4.3 Round-Trip Consistency Loss

The endpoint loss anchors the final position. To constrain temporal consistency (Wang et al., 2019; Jabri et al., 2020), we track auxiliary queries c_j forward and then through the reversed clip. Writing $\widehat{c}_j$ for the returned location gives

$$\mathcal{L}_{\text{cyc}} = \frac{1}{M}\sum_{j=1}^{M} \rho_\delta(\|\widehat{c}_j - c_j\|_2), \qquad \mathcal{L}_{\text{stream}} = \mathcal{L}_{\text{end}} + \lambda \mathcal{L}_{\text{cyc}}. \tag{9}$$

The forward traversal runs without gradients. Queries whose terminal positions leave the image are removed, and the return pass retains a gradient tail. Physical endpoints anchor tissue correspondence, while the cycle loss encourages return consistency along the predicted path. The joint Track-On-R recipe accumulates five endpoint-labeled clips and one cycle clip per optimizer step with unit per-clip loss weights, giving an effective expected relative weight $\lambda = 1/5$. Cycle clips contain 96 frames with a 12-frame gradient tail; auxiliary queries use an 8×8 grid.

The cycle objective supplies training queries beyond the sparsely marked tissue points. Its target is the known starting location, so no additional endpoint annotation is required for these auxiliary tracks. Physical endpoints and cycles are optimized together: measurements identify where tracked tissue should end, while round trips test the consistency of the student's predicted correspondence.

Inference. After both training stages, weights remain fixed. Given initial query coordinates, the tracker processes RGB frames in a single forward stream with continuous memory (Equation 5). Only training requires teacher predictions, endpoint labels, stereo inputs, and reverse tracking.

5 Experiments

5.1 Experimental Setup

Training. We adapt released Track-On, CoTracker3, and Track-On-R checkpoints (Aydemir et al., 2025; Karaev et al., 2025; Aydemir et al., 2026). Track-On-R Stage 1 uses eight 48-frame clips per optimizer step and a 32-frame gradient window. Its step-100 checkpoint initializes Stage 2. Both stages use AdamW with a frozen image encoder, peak learning rate 3×10^{-6}, and weight decay 10^{-5} (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). Track-On and CoTracker3 combine the endpoint objective with their native dense trajectory losses; Track-On-R uses endpoint and round-trip losses. Appendix B gives the optimization schedule and window sizes.

Evaluation splits. We evaluate curation by training on STIR and transferring to all 150 VL-SurgPT clips (375 physical endpoints). Endpoint-guided training uses a separate 603-clip manifest that overlaps with 121 of these clips. The remaining 29 clips (73 endpoints) form a post-hoc clip-disjoint diagnostic, used for the main Track-On-R training comparison. It excludes those clips from endpoint training but does not establish procedure-level independence or a prospectively reserved test set. Tables 1 and 5 describe the full 150-clip pool; Table 4 uses the 29-clip diagnostic. Appendix C details the protocols. These comparisons jointly vary label source, density, and placement.

Endpoint metrics. Following threshold-averaged TAP evaluation (Doersch et al., 2022), for N evaluated endpoints with errors e_i, we report

$$\delta_{\text{avg}} = \frac{100}{N|\mathcal{A}|}\sum_{i=1}^{N}\sum_{a\in\mathcal{A}} \mathbf{1}[e_i \le a], \qquad \text{EPE} = \frac{1}{N}\sum_{i=1}^{N} e_i. \tag{10}$$

VL-SurgPT uses paired identities, coordinates scaled to 720×480, and $\mathcal{A} = \{2, 4, 8, 16, 32\}$ pixels. Points are pooled across clips using the same valid point indices for each method. STIR uses 60 clips and 496 predicted endpoints, native 1280×1024 coordinates, and thresholds $\{4, 8, 16, 32, 64\}$; its

Table 1: **VL-SurgPT physical-endpoint comparison.** STEP is compared with released models and manual-trajectory fine-tuning across multiple tracking backbones.

Backbone	Supervision	$\delta_{\text{avg}} \uparrow$	EPE$\downarrow$ (px)
RAFT (Teed & Deng, 2020)	none	33.81	82.16
SEA-RAFT (Wang et al., 2024)	none	33.71	80.17
TAPIR (Doersch et al., 2023)	none	64.80	15.64
Track-On (Aydemir et al., 2025)	none	67.36	11.79
	manual trajectories	73.92	11.05
	STEP (Ours)	**75.89** (+1.97)	**10.70** (−0.35)
CoTracker3 (Karaev et al., 2025)	none	65.87	12.14
	manual trajectories	66.93	13.08
	STEP (Ours)	**67.52** (+0.59)	**11.33** (−1.75)
Track-On-R (Aydemir et al., 2026)	none	72.27	15.27
	manual trajectories	73.17	14.90
	STEP (Ours)	**74.08** (+0.91)	**12.14** (−2.76)

Table 2: **STIR resolution and support-grid controls.** STEP is compared with Track-On-R under matched inference settings.

Setting	Working res.	Grid	$\delta_{60} \uparrow$
Track-On-R	384×512	none	75.89
Track-On-R	384×512	8×8	76.53
Track-On-R	384×512	16×16	77.54
Track-On-R	480×608	none	77.86
Track-On-R	480×608	8×8	78.10
Track-On-R	480×608	16×16	78.35
STEP	480×608	16×16	**79.23**

legacy δ_{60} uses prediction-to-nearest-marker matching. This score measures set proximity, whereas VL-SurgPT measures identity. Intermediate frames and terminal observations are evaluated separately.

Statistical analysis. For the 29-clip diagnostic, we use a nonparametric paired bootstrap (Efron, 1979): we resample clips jointly across methods 10,000 times, pool their endpoint errors, and report percentile 95% intervals. This preserves within-clip dependence. Intervals describe variation over clips for the saved checkpoints; they do not incorporate training-seed variation.

5.2 Physical Endpoint Tracking

Endpoint correspondence. On the 29-clip diagnostic, STEP reduces Track-On-R endpoint error from 31.52 to 16.80 pixels relative to dense-label fine-tuning, a 46.7% reduction. Table 1 complements this comparison with results across three backbones on the full evaluation pool. Endpoint-guided training improves both metrics over dense-label fine-tuning for each backbone: EPE decreases from 11.05 to 10.70 pixels for Track-On, 13.08 to 11.33 for CoTracker3, and 14.90 to 12.14 for Track-On-R. These full-pool results describe adaptation across architectures under the overlap reported in Section 5.

STIR and inference controls. Table 2 separates endpoint training from changes in inference resolution and spatial support. Increasing resolution and adding a support grid improve the released Track-On-R baseline. At the matched 480×608 resolution with a 16×16 grid, STEP reaches 79.23 δ_{60} versus 78.35 for the unadapted model. Its EPE is 12.44 pixels, compared with 16.37 for the released reference configuration. The matched accuracy comparison supports a benefit from endpoint training beyond inference settings. Appendix D.3 evaluates the same predictions in 3D: STEP reaches 74.71 average accuracy versus 51.48 for the static start-cloud control. This measures recovery of tissue displacement relative to leaving points at their initial 3D positions.

5.3 Effect of Data Curation

We compare six selection policies on the same STIR candidate pool, with identical 300-step optimization budgets and no VL-SurgPT adaptation (Table 3). The temporal–stereo union transfers at 72.91 accuracy and 14.87-pixel EPE, compared with 72.37 and 15.99 for unfiltered candidates. It

Table 3: **Consistency selection.** STIR training; VL-SurgPT evaluation. EPE in pixels.

Selection	Keep (%)	$\delta_{\text{avg}} \uparrow$	EPE↓
Unfiltered	100.0	72.37	15.99
Temporal (T)	57.2	72.16	16.40
Stereo (S)	52.8	72.48	15.96
T ∩ S	47.5	71.04	20.42
Count-matched T	47.5	71.47	17.31
T ∪ S	62.5	**72.91**	**14.87**

Table 4: **Training-stage and loss controls.** 29-clip diagnostic; joint training also changes endpoint exposure and gradient-tail length. EPE is in pixels.

Recipe	$\delta_{\text{avg}} \uparrow$	EPE↓
S1 only	68.77	19.91
S2 endpoint only, no S1	70.96	20.08
S1 + endpoint	70.96	22.26
S1 + cycle only	70.68	23.38
S1 + endpoint + cycle	**71.78**	**16.80**

Table 5: **Occlusion recovery and visibility.** Full-pool endpoints grouped by intermediate visibility. EPE is in pixels; other metrics are percentages.

	Occluded mid-sequence (n=145)			Never occluded (n=230)			Visibility	
Model	$\delta \uparrow$	EPE↓	>100 px↓	$\delta \uparrow$	EPE↓	>100 px↓	Occ. recall↑	False occ.↓
Zero-shot	64.14	28.54	8.97	77.39	6.91	1.30	47.9	2.0
Dense-label FT	64.28 (+0.14)	25.92	9.66	**78.78** (+1.39)	7.95	1.30	49.3	**1.7**
Stage 1 only	64.55 (+0.41)	28.51	8.28	77.74 (+0.35)	**6.53**	**0.87**	49.9	1.9
Stage 2 only	66.76 (+2.62)	24.15	**7.59**	78.17 (+0.78)	7.29	1.30	50.1	1.9
STEP (ours)	**66.76** (+2.62)	**20.83**	**7.59**	78.70 (+1.31)	6.67	1.30	**51.6**	2.1

also exceeds the released Track-On-R initialization's 72.27 accuracy. Relative to unfiltered training, the union improves accuracy by 0.54 points and reduces EPE by 7.0%.

Selection criterion and retained count. The intersection retains 47.5% of candidates in the source-pool probe and transfers at 71.04 accuracy. Randomly reducing temporal candidates to the same per-window count yields 71.47, with EPE 17.31 instead of 20.42 pixels. At this matched count, requiring agreement with both tests does not improve transfer. Complementary coverage is more effective here than requiring both tests. Union-versus-temporal comparisons additionally change the number of retained candidates; they assess the complete selection policy.

Union selection also improves median error from 2.89 to 2.84 pixels and success within 8 pixels from 83.47% to 84.27%. Both union and unfiltered training have 4.8% of endpoints above 100 pixels, indicating improvements beyond the count of extreme failures.

5.4 Training Stages and Loss Design

Curated initialization and physical supervision. Table 4 compares curated-data training (S1) and endpoint-only training (S2). On the diagnostic subset, S1 gives 68.77 accuracy and 19.91-pixel EPE, while S2 without S1 gives 70.96 and 20.08. Following S1 with the endpoint loss also gives 70.96 accuracy, with EPE 22.26. These controls show that curated initialization alone does not guarantee better endpoint accuracy; the strongest result uses the joint training objective.

Cycle-only training after S1 reaches 70.68 accuracy and 23.38-pixel EPE. Compared with S1 alone, its higher thresholded accuracy accompanies a higher mean error, showing that these metrics capture different aspects of the error distribution. The joint configuration improves both. This pattern supports using measured terminal correspondence alongside return consistency to guide training.

Loss combination. From the same S1 initialization, the joint recipe reaches 71.78 accuracy and 16.80-pixel EPE, versus 70.96 and 22.26 for endpoint-only training. The paired EPE-difference interval is $[-14.85, -0.11]$ pixels, whereas the accuracy interval $[-0.56, 2.78]$ includes zero. The joint recipe also increases endpoint exposure and gradient-tail length (Table 7), so this comparison measures their combined effect.

5.5 Occlusion and Error Distribution

Table 5 groups the full-pool endpoints by intermediate visibility. For points occluded mid-sequence, STEP improves accuracy from 64.28 to 66.76 and reduces EPE from 25.92 to 20.83 pixels relative to dense-label fine-tuning. Errors above 100 pixels decrease from 9.66% to 7.59%. For never-occluded

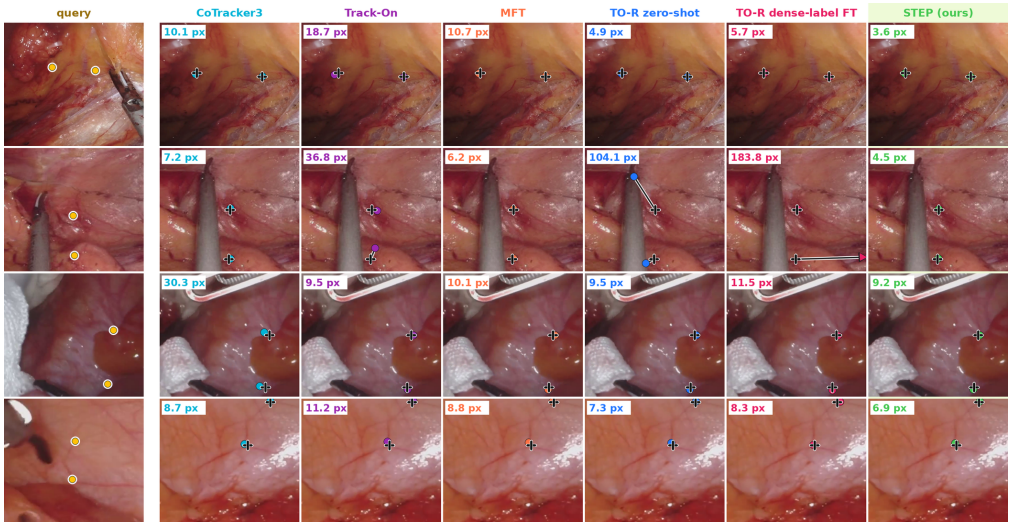

Figure 4: **Qualitative endpoint tracking.** Predictions and physical references (crosses) on four VL-SurgPT sequences; errors are in pixels. The last three columns use the Track-On-R backbone.

points, accuracy is similar (78.70 versus 78.78), with EPE decreasing from 7.95 to 6.67 pixels. The larger gain after occlusion supports endpoint supervision for correspondence recovery.

Visibility prediction provides a complementary view: STEP reaches 51.6% occlusion recall with 2.1% false occlusion. Recovery at a visible endpoint and recognition of an intermediate occlusion measure different behaviors. Appendices D.1 and D.2 report intermediate tracking, acquisition conditions, and visibility on the 29-clip diagnostic.

Intermediate correspondence. The annotated path provides another view of supervision. Across 2,293 intermediate point-frames, the three-seed S1+endpoint control improves accuracy by 1.40 points over the released model and by 0.92 over S1, while dense-label fine-tuning scores 1.62 points higher than this endpoint-only control (Appendix D.1). Physical endpoints can therefore improve intermediate tracking as well as terminal correspondence, although direct trajectory labels remain useful for supervising the intervening path.

Efficiency. At 480×608 with a 16×16 support grid, Track-On-R STEP takes 63.3–66.0 ms per frame and 1.21–1.37 GB peak inference memory for 8–64 queries on an A100-80GB. This corresponds to approximately 15–16 frames/s under the same resolution and support-grid setting used for the STIR comparison. For a fixed query count, streaming maintains constant state size. Appendix B reports peak memory usage during training.

Qualitative behavior. Figure 4 illustrates localized errors and larger displacements to different surfaces. STEP stays closer to the physical references in these examples. Avoiding large errors can improve EPE more than thresholded accuracy. On four selected clips, STEP also produces fewer orientation flips in a tracked grid than dense-label fine-tuning (Figure 5). These flips expose local reversals in the predicted motion field that are difficult to see from isolated endpoint markers. The grid and endpoint views assess spatial organization and terminal accuracy together, providing complementary descriptions of tracking behavior.

6 Conclusion

STEP learns surgical correspondence from two complementary resources: dense trajectories curated from video and sparse physical endpoints. Temporal and stereo selection improve cross-dataset transfer, while endpoint-guided streaming training exposes the tracker to accumulated error with bounded activation memory. Combining physical endpoint and round-trip objectives yields lower endpoint error on the VL-SurgPT diagnostic. These findings support using existing surgical video and measured correspondence together, with dense targets providing broad training coverage and physical observations anchoring long-range predictions. The resulting tracker runs on a single RGB stream with fixed weights.

AI USE STATEMENT

Generative AI assisted with code implementation, manuscript editing, figure layout, and consistency checks. The authors developed the research ideas, ran the experiments, verified the reported results, and remain responsible for all content.

ETHICS STATEMENT

This work uses existing surgical datasets under their access terms, with no new patient acquisition. The evaluated trackers are research systems; clinical deployment would require further validation.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 specify the curation pipeline and training objectives. Appendix B reports implementation and efficiency settings; Appendices C and D detail evaluation protocols and additional results. Experiments initialize from released tracker checkpoints and use existing surgical datasets.

REFERENCES

Görkay Aydemir, Xiongyi Cai, Weidi Xie, and Fatma Güney. Track-On: Transformer-based online point tracking with memory. In *International Conference on Learning Representations (ICLR)*, 2025.

Görkay Aydemir, Fatma Güney, and Weidi Xie. Real-world point tracking with verifier-guided pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-vid: A benchmark for tracking any point in a video. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 13610–13626, 2022.

Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10061–10072, 2023.

Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, and Andrew Zisserman. BootsTAP: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2024.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 605–613, 2017.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 270–279, 2017.

Jiaxin Guo, Jiangliu Wang, Zhaoshuo Li, Tongyu Jia, Qi Dou, and Yun-Hui Liu. Ada-tracker: Soft tissue tracking via inter-frame and adaptive-template matching. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15463–15470, 2024. doi: 10.1109/ICRA57147.2024.10611030.

Hamlyn Centre, Imperial College London. Hamlyn centre laparoscopic / endoscopic video datasets. `https://hamlyn.doc.ic.ac.uk/vision/`.

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

Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.

Michel Hayoz, Christopher Hahne, Mathias Gallardo, Daniel Candinas, Thomas Kurmann, Maximilian Allan, and Raphael Sznitman. Learning how to robustly estimate camera pose in endoscopic videos. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 2023. doi: 10.1007/s11548-023-02919-w.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964.

Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-time correspondence as a contrastive random walk. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Junpeng Jing, Weixun Luo, Ye Mao, and Krystian Mikolajczyk. Stereo any video: Temporally consistent stereo matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Nikita Karaev, Iurii 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)*, 2025.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Michal Neoral, Jonáš Šerých, and Jiří Matas. MFT: Long-term tracking of every pixel. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6837–6847, 2024.

Adam Schmidt, Omid Mohareri, Simon DiMaio, and Septimiu E. Salcudean. SENDD: Sparse efficient neural depth and deformation for tissue tracking. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 238–248, 2023. doi: 10.1007/978-3-031-43996-4_23.

Adam Schmidt, Omid Mohareri, Simon DiMaio, and Septimiu E. Salcudean. Surgical tattoos in infrared: A dataset for quantifying tissue tracking and mapping. *IEEE Transactions on Medical Imaging (TMI)*, 2024a.

Adam Schmidt, Omid Mohareri, Simon DiMaio, Michael C. Yip, and Septimiu E. Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 94:103131, 2024b. doi: 10.1016/j.media.2024.103131.

Leslie N. Smith. Cyclical learning rates for training neural networks. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2017.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 2018.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2566–2576, 2019.

Yihan Wang, Lahav Lipson, and Jia Deng. SEA-RAFT: Simple, efficient, accurate RAFT for optical flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Paul J. Werbos. Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

Ronald J. Williams and Jing Peng. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, 2(4):490–501, 1990.

Bohan Zhan, Wang Zhao, Yi Fang, Bo Du, Francisco Vasconcelos, Danail Stoyanov, Daniel S. Elson, and Baoru Huang. Tracking everything in robotic-assisted surgery. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Rulin Zhou, Wenlong He, An Wang, Jianhang Zhang, Xuanhui Zeng, Xi Zhang, Chaowei Zhu, Haijun Hu, and Hongliang Ren. Bridging vision and language for robust context-aware surgical point tracking: The VL-SurgPT dataset and benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

A Data and Curation

Table 6: **Surgical sources and generated trajectory inventory.** STIR counts both views of 566 stereo clips. Pseudo-labels are model-generated; physical labels are existing endpoint observations.

	VL-SurgPT	STIR
Video streams	754	1,132
Frames	181 k	640 k
Acquisition	Monocular	Stereo
Human-annotated trajectories	1,862	0
Generated pseudo-trajectories	493,893	1,685,211
Pseudo-labeled point-frames	103 M	300 M
Selection rule	Temporal	Temporal $\cup$ stereo
Physical supervision	Paired endpoints	Endpoint sets

VL-SurgPT. VL-SurgPT provides monocular surgical video, human-annotated trajectories, and paired physical start/end observations (Zhou et al., 2026). Table 6 summarizes the generated inventory. Temporal selection is used because the source is monocular. Excluding the 150 evaluation clips before Stage 1 window sampling leaves 604 source clips; 596 yield usable training windows.

STIR. STIR contains 566 stereo clips, represented as 1,132 camera-view streams (Schmidt et al., 2024a). Temporal reversibility and stereo geometry allow four consistency policies: temporal, stereo, intersection, and union. The policy comparison uses a common pool of 1,726 clip-view windows and 300 optimizer steps per policy. Multiple windows can originate from the same source clip.

Additional sources and preprocessing. The pipeline also generates trajectories for StereoMIS, Hamlyn, and SurgMotion (Hayoz et al., 2023; Hamlyn Centre, Imperial College London; Zhan et al., 2025). StereoMIS trajectories are aligned to the decoded packet timeline so that coordinates correspond to displayed frames. Hamlyn uses temporal selection with a first-frame grid check to remove static border candidates. These steps prevent timing offsets and invalid image regions from entering the training targets. The reported quantitative studies use STIR and VL-SurgPT.

Trajectory generation. STIR and VL-SurgPT candidates are generated by frozen AllTracker (Harley et al., 2025) over windows of at most 750 frames. The teacher occlusion score is $1 - \text{visibility} \cdot \text{confidence}$; scores at most 0.4 are treated as visible. Temporal selection uses a 3-pixel round-trip threshold. On stereo sources, Stereo Any Video (Jing et al., 2025) supplies disparity; track pairing uses a 3-pixel tolerance, with an 8-pixel disparity tolerance on visible frames.

B Training and Efficiency

Stage 1 optimization. The Track-On-R S1 run uses eight 48-frame clips per optimizer step, a 32-frame differentiated window, and 400 optimizer steps. The checkpoint at step 100 initializes Stage 2. The trainable part contains 24.95M parameters; the image encoder is frozen. Optimization uses AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) with weight decay 10^{-5} and a one-cycle schedule (Smith, 2017; Smith & Topin, 2018); the peak learning rate is 3×10^{-6} with 10% warm-up. All STIR selection policies share the released initialization and 300-step budget.

Stage 2 optimization. The endpoint-only Track-On-R control accumulates two endpoint clips per optimizer step, processes streams up to 192 frames, and differentiates the last 16 frames. The joint endpoint+cycle recipe accumulates five endpoint clips and one cycle clip per optimizer step. Endpoint clips can span up to 256 frames and use a 24-frame gradient tail. The cycle clip contains 96 frames, uses a 12-frame return-gradient tail, and seeds an 8×8 auxiliary grid. Per-clip losses have unit weights, so the expected cycle-to-endpoint contribution induced by the sampling schedule is $1/5$. Both recipes use 300 optimizer steps.

Table 7: **Track-On-R endpoint-guided training recipes.** Clips accumulate per optimizer step; all S2 runs use 300 steps. The cycle loss uses a separate 96-frame clip and 12-frame gradient tail.

Losses	Endpoint clips	Cycle clips	Max. span	Grad. tail
Endpoint only	2	0	≤ 192	16
Endpoint + cycle	5	1	≤ 256	24

Training coordinates and losses. The endpoint Huber transition (Huber, 1964) is $\delta = 8$ pixels in the 640×512 training coordinate system. The bidirectional nearest-neighbor loss for unordered STIR endpoints uses $\beta = 0.5$.

Runtime and memory. Runtime is measured on an A100-80GB with raw 1280×1024 frames resized to the indicated working resolution. For fixed query count and memory size, streaming inference maintains constant per-frame state size and $\mathcal{O}(T)$ total computation. At 480×608 with a 16×16 support grid, the joint Track-On-R checkpoint takes 63.3–66.0 ms per frame and 1.21–1.37 GB peak memory for 8–64 queries. Endpoint-only training uses 30.5 GB peak memory with a 16-frame gradient tail. This training measurement applies to the endpoint-only configuration in Table 7.

C Evaluation Protocols

Evaluation populations. The VL-SurgPT evaluation pool contains 150 clips and 375 physical endpoints. The endpoint-training manifest contains 603 clips, including 121 from this pool; the remaining 29 clips and 73 endpoints form the post-hoc clip-disjoint diagnostic. Full-pool analyses describe the existing checkpoints across backbones, visibility groups, and acquisition conditions. The 29-clip subset evaluates clips absent from endpoint training, although procedure independence and prior model-selection exposure are not established. STIR-to-VL-SurgPT curation transfer uses the full pool because those models are trained solely on STIR.

Endpoint scoring. VL-SurgPT predictions and labels are scaled together to 720×480; STIR's legacy image-plane endpoint score uses native 1280×1024 coordinates. Section 5 defines scoring and matching.

STIR 3D protocol. We follow the official STIR 2024 3D evaluation (Schmidt et al., 2024a): each predicted 3D endpoint is matched by nearest neighbour to the ground-truth end cloud, and accuracy is the fraction of matched distances below a threshold, averaged over $\{2, 4, 8, 16, 32\}$ mm. Scoring uses 59 clips and 495 endpoints. The *static* row is the protocol's built-in control: the unchanged ground-truth *start* cloud is the prediction, measuring accuracy without tracking.

Statistical analysis. For the 29-clip diagnostic, we draw 10,000 bootstrap samples with replacement, using the same clip identifiers for every method. Endpoint errors are pooled; the 2.5 and 97.5 percentiles define the interval (Efron, 1979). Resampling clips preserves within-clip dependence. These intervals describe clip variation for fixed checkpoints, rather than training-seed variation.

D Additional Results

D.1 Intermediate Tracking

We evaluate intermediate correspondence on 2,293 annotated point-frames from the 150-clip pool. These samples exclude both the initial query frame and the final physical endpoint, measuring the path between them.

For the three-seed S1+endpoint control without the cycle loss, a paired clip-level bootstrap with 10,000 resamples gives a mean δ_{avg} difference of $+1.40$ points relative to the released initialization (95% interval $[+0.37, +2.59]$), $+0.92$ relative to S1 ($[+0.02, +1.81]$), and -1.62 relative to dense human-trajectory fine-tuning ($[-3.31, -0.02]$). Thus, endpoint-only training improves intermediate correspondence over the released and S1 models, while dense trajectory supervision remains stronger on the annotated path.

For normalized time, let

$$\tau = \frac{k - k_{\text{first}}}{k_{\text{last}} - k_{\text{first}}}, \tag{11}$$

where k_{first} is the first annotated frame after the query and k_{last} is the endpoint. Four bins over $\tau > 0$ contain 378, 483, 513, and 581 point-frames, that is 1,955 of the 2,293 intermediate samples; the remaining 338 lie at $\tau = 0$ and fall outside the binned range. The 375 physical endpoints are scored separately from these intermediate samples.

D.2 Acquisition Conditions

We group full-pool endpoints by acquisition condition for the joint Track-On-R model. Physical endpoint δ_{avg}/EPE values are 74.93/14.59 for tissue deformation ($n = 71$), 75.07/13.66 for instrument occlusion ($n = 69$), 73.49/15.31 for camera jitter ($n = 86$), 66.88/8.27 for surface reflection ($n = 64$), and 78.59/8.57 for cauterization smoke ($n = 85$). Across all 375 endpoints, the joint checkpoint gives 74.08/12.14 and 3.73% of errors above 100 pixels.

Visibility on the clip-disjoint diagnostic. On the 29-clip subset, visibility is scored on 59 occluded and 569 visible annotated point-frames. Occlusion recall conditions on occluded observations; false-occlusion rate conditions on visible ones. STEP gives 50.85% recall and 2.99% false occlusion, while S1 gives 62.71% and 3.51%. Endpoint recovery and visibility recall measure different behavior.

D.3 STIR 3D Accuracy

STIR's calibrated stereo observations provide 3D ground truth in addition to image-plane endpoints. We evaluate the predictions from the main STIR experiment with the 3D protocol in Appendix C, comparing STEP with the static start-cloud control (Table 8).

Table 8: **STIR 3D endpoint accuracy.** Official STIR 2024 3D protocol, 59 clips and 495 endpoints, nearest-neighbour matched to the ground-truth end cloud. Higher is better.

Prediction	2 mm	4 mm	8 mm	16 mm	32 mm	Avg↑
Static control (GT start cloud)	15.85	29.55	49.68	71.09	91.22	51.48
STEP	**37.37**	**62.83**	**81.21**	**93.13**	**98.99**	**74.71**

STEP reaches 74.71 average accuracy, compared with 51.48 for the static control. The 23.23-point difference measures the benefit over leaving points at their initial 3D locations. At 2 mm, accuracy increases from 15.85 to 37.37; at 32 mm, both methods accept most endpoints (91.22 and 98.99). The tighter thresholds therefore distinguish correspondence more strongly in this comparison.

E Qualitative Analysis

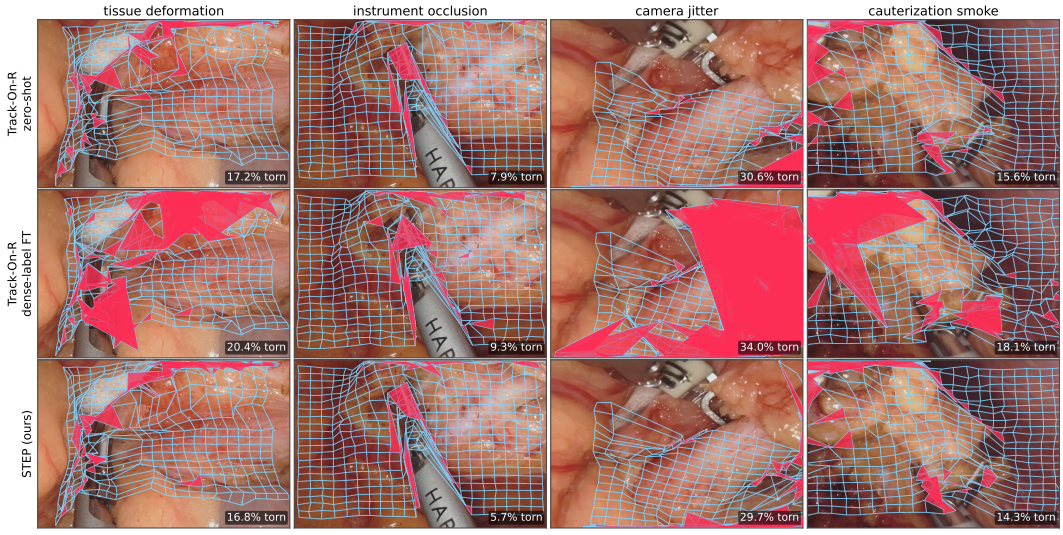

Figure 5: **Motion-field coherence without ground truth.** Track-On-R variants on four VL-SurgPT clips. Red cells mark orientation flips and panels report their rates; compare within columns. Coherent tracks can still drift.

Spatial coherence. Figure 5 tracks a regular grid from the first to the final frame of four selected clips from the 29-clip diagnostic, one per represented VL-SurgPT scenario. All three models use the Track-On-R backbone with different supervision. A grid cell is marked as a tear when its orientation flips relative to the first frame. Dense-label fine-tuning has the highest tear rate in each displayed

scenario and STEP the lowest. Rates depend on grid density and clip length, so comparisons are made within each column. The clips were selected for visual interpretability.

Smooth mesh deformation is a spatial-coherence diagnostic. A field can drift coherently to incorrect positions, while genuine motion discontinuities at instrument boundaries can produce tears. Across the measured clips, tear rate and endpoint error are uncorrelated (Spearman $\rho = -0.09$ over 21 clip-method pairs). Spatial coherence and endpoint accuracy therefore capture complementary properties.

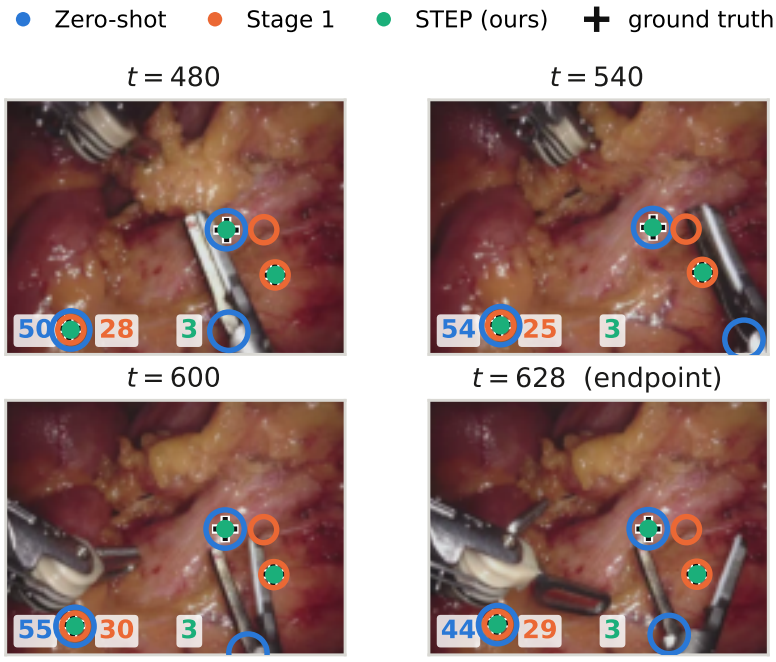

Figure 6: **Additional correspondence examples.** Predictions at selected frames illustrate tracking through the sequence; numbers give errors in pixels.

Endpoint correspondence. Figure 6 complements the endpoint comparison with predictions at selected frames. The examples illustrate how correspondence evolves during tissue motion and tool interaction; the quantitative endpoint and visibility analyses provide the corresponding measured evaluations.