QUE**ST**: PERSISTENT **QUE**RIES AS SEMANTIC MONITORS FOR DRIFT **S**UPPRESSION IN LONG-HORIZON **T**RACKING

Mayank Anand, Mohammad Saqlain, Kyan Mahajan, Priya Shukla & G.C Nandi *
Department of Information Technology
Indian Institute of Information Technology Allahabad
Prayagraj, India
{iit2024036,iit2024113,iit2024092,phc2015001,gcnandi}@iiita.ac.in

Andrew Melnik
Department of Mathematics and Computer Science
University of Bremen, Germany
andrew.melnik.papers@gmail.com

ABSTRACT

Tracking points in videos is typically formulated as frame-to-frame correspondence, where each point is matched locally to the next frame. While this works over short horizons, errors accumulate under articulation, occlusion, and viewpoint change, leading to silent semantic drift that existing trackers cannot detect or correct. In this work, we revisit long-horizon tracking from a monitoring perspective and introduce QueST, a monitoring-by-design framework that treats interaction-relevant entities as persistent semantic queries rather than transient point tracks. Instead of local propagation, each query attends globally over spatiotemporal video features at every timestep, providing a stable semantic anchor across time. We further constrain query trajectories with lightweight 3D physical grounding, using geometric plausibility to suppress unbounded drift under occlusion. We evaluate QueST on long-horizon articulated sequences from PartNet-Mobility in SAPIEN and compare against RAFT-3D, CoTracker, and TAP-Net. QueST substantially reduces terminal drift achieving a 67.7% Absolute Point Error (APE) improvement over TAP-Net while better preserving identity over extended horizons. Our results show that embedding semantic monitoring directly into perception enables more reliable long-horizon tracking under distribution shift.

1 INTRODUCTION

Reliable operation of machine learning systems in dynamic, long-horizon environments requires the ability to detect and respond to silent degradation (Quiñonero-Candela et al., 2008; Hendrycks et al., 2018; Koh et al., 2021). In tracking-based perception systems, such degradation often manifests as semantic drift: a tracked entity gradually diverges from its original meaning under distribution shift, without explicit failure signals. This failure mode is particularly dangerous in embodied settings such as robotic manipulation, where incorrect perceptual state can propagate to unsafe actions.

Silent semantic drift in tracking-based perception. Existing tracking pipelines are ill-suited to catch this form of drift. Most rely on Markovian correspondence (Teed & Deng, 2021; Doersch et al., 2023; Karaev et al., 2024), propagating points frame-to-frame using local appearance cues. While effective over short horizons, small correspondence errors inevitably accumulate under articulated motion, occlusion, and viewpoint change. As a result, trackers may continue producing confident predictions even after losing semantic identity (Fig 1B), leaving downstream systems unable to detect that perception has failed.

*Code will be publicly available at https://github.com/AnandMayank/QueST

We argue that this limitation stems from a lack of representation-level monitoring (Fischer et al., 2023; Doersch et al., 2023). Here, representation-level drift refers to shifts in the feature embedding of a tracked entity, where the underlying semantic representation changes even when local pixel correspondences appear consistent (Wang et al., 2024). Correspondence-based trackers do not maintain a persistent notion of what is being tracked, only where a point moves locally. To address this, we introduce QueST, a monitoring-by-design framework that represents interaction-relevant entities as persistent, query-conditioned semantic monitors (Fischer et al., 2023; Carion et al., 2020; Doersch et al., 2023). Rather than propagating pixels, queries attend globally to spatiotemporal video features, enabling continuous assessment of semantic consistency across time (Fig 1C). Unlike Markovian trackers, QueST queries attend globally over the entire spatiotemporal feature volume, enabling identity verification beyond local frame-to-frame propagation.

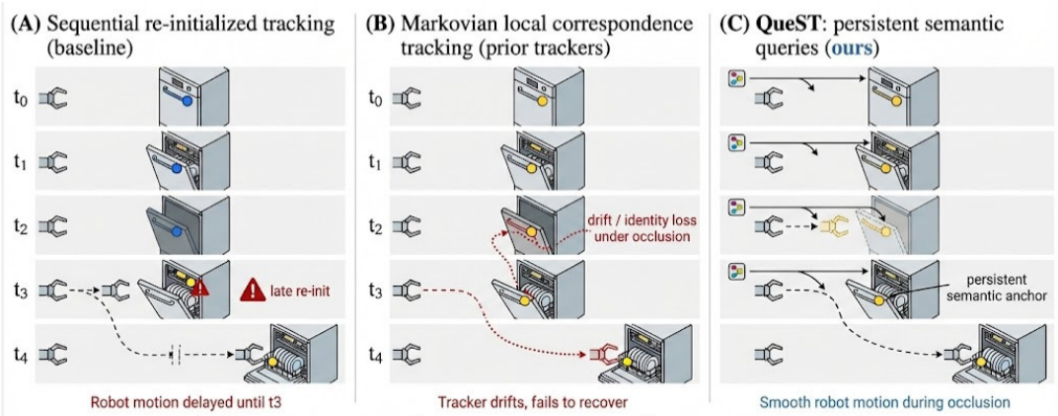


Figure 1: **The Reliability Crisis.** (A) Re-initialization breaks identity; (B) Markovian trackers (e.g., CoTracker) accumulate drift $\sum \epsilon_t$ because they propagate local errors; (C) **QueST** maintains global anchors via persistent learnable queries and 3D physical grounding, suppressing drift.

To adapt when drift begins to emerge, QueST enforces physical consistency by grounding query trajectories in lifted 3D space (Koppula et al., 2024; Xiang et al., 2020). Deviations from plausible geometric structure act as an implicit correction signal, allowing the system to suppress unbounded drift even under prolonged occlusion. This coupling of semantic monitoring with geometric constraints enables stable long-horizon operation without explicit re-initialization.

Our key contributions are as follows:

- We formalize long-horizon tracking as a representation-level drift monitoring problem for embodied perception.
- We introduce QueST, which uses persistent semantic queries to continuously monitor and preserve identity under distribution shift.
- We show this monitoring-by-design approach substantially reduces silent drift in long-horizon articulated scenarios.

2 Problem Formulation

We consider the problem of tracking interaction-relevant points on articulated objects Yu et al. (2024); Guerrier et al. (2025) in video $V = \{I_t\}_{t=1}^{T}$. We formulate this as **query-conditioned interaction tracking**, where a query q specifies a semantic target (e.g., a handle or joint) rather than a specific starting pixel.

The goal is to predict a trajectory $P = \{p_t\}_{t=1}^{T}, p_t \in \mathbb{R}^2$ that satisfies two properties:

1. **Semantic Identity:** The prediction p_t must correspond to the same semantic entity induced by q across all frames, even under occlusion.
2. **Physical Plausibility:** The lifted 3D trajectory $x_t = \Pi^{-1}(p_t, D_t) \in \mathbb{R}^3$ (where D_t is depth) must follow a valid kinematic manifold (e.g., a revolute arc), effectively minimizing drift $\epsilon_{drift} = \|x_t - \mathcal{M}(x_{t-1})\|$.

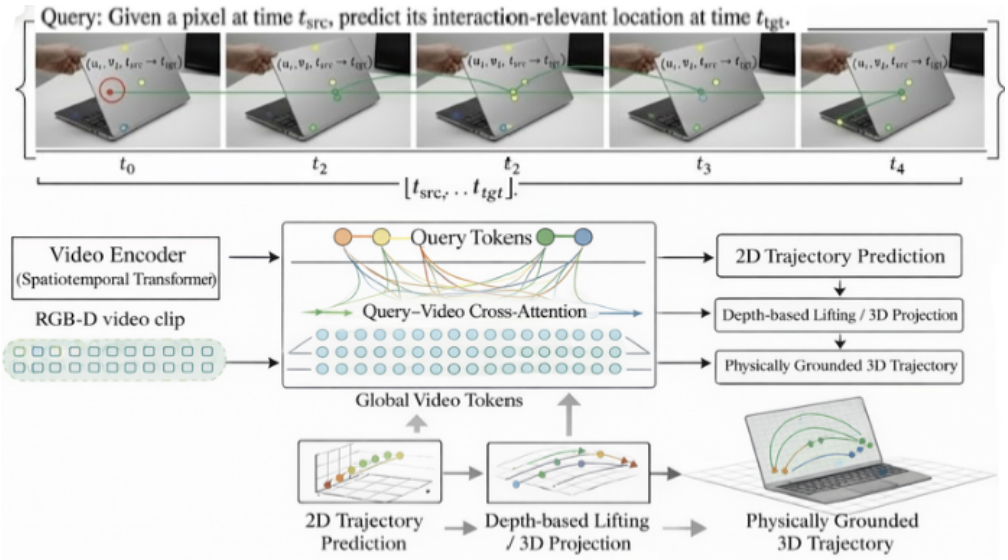


Figure 2: **QueST System Architecture.** Video features $\mathbf{F}_t$ are extracted via a ViT encoder. Persistent learnable queries $\mathbf{Q}$ attend globally across frames to maintain semantic identity. The resulting 2D trajectories are lifted to 3D world coordinates x_t using depth backprojection for physical grounding.

Standard flow-based methods fail this objective because they solve for local pixel affinity between consecutive frames $\arg\max Sim(I_t, I_{t+1})$, which does not enforce long-term semantic or geometric consistency.

3 QueST Framework

QueST replaces recursive point propagation with a query-based transformer architecture that predicts globally consistent trajectories grounded in 3D physics. Following the 'tracking-by-attention' paradigm introduced by DETR Carion et al. (2020) and extended by CoTracker Karaev et al. (2024), we represent tracked entities as learnable queries. See Figure 2 for the complete pipeline.

3.1 QueST-Backbone: Persistent Semantic Monitoring

Video Encoder. We process input frames (resized to 224×224) using a ViT-style encoder. Frames are partitioned into 16×16 patches, resulting in $N = 196$ tokens per frame. These are projected into an embedding dimension $D = 384$ and supplemented with learnable spatial and temporal positional encodings, yielding feature tensor $\mathbf{F} \in \mathbb{R}^{T \times N \times D}$.

Persistent Queries. We maintain a set of $K = 8$ learnable query embeddings $\mathbf{Q} \in \mathbb{R}^{K \times D}$. Each query embedding represents a semantic hypothesis about an interaction-relevant entity (e.g., handle, hinge, or rim) and serves as a persistent anchor used to localize that entity across time. Unlike tracklets in standard trackers, these queries are shared across the entire temporal window and initialized from a learned distribution. They act as "semantic anchors," searching for specific affordance types (e.g., handles, rims) regardless of their screen position.

Global Cross-Attention Decoder. A lightweight transformer decoder (2 layers, 4 attention heads) refines the queries by attending to the video features. At each timestep t, the decoder computes cross-attention between queries $\mathbf{Q}$ and frame features $\mathbf{F}_t$, producing refined embeddings $\tilde{\mathbf{Q}}_t$. A shared Multi-Layer Perceptron (MLP) head then maps $\tilde{\mathbf{Q}}_t$ to 2D coordinates $\hat{p}_{t,k} \in [0, 1]^2$ and confidence scores $c_{t,k}$.

Our backbone architecture and coordinate-conditioned decoder are inspired by the spatiotemporal motion representations in D4RT Zhang et al. (2025).

3.2 Physical Grounding and Objectives

To suppress drift, we lift 2D predictions to 3D world coordinates $x_{t,k}$ using camera intrinsics and depth. We train using a combined objective:

$$\mathcal{L}_{total} = \mathcal{L}_{aff} + \lambda_{smooth}(\mathcal{L}_{vel} + \mathcal{L}_{acc}) + \lambda_{geo}\mathcal{L}_{manifold} \tag{1}$$

where $\mathcal{L}_{aff}$ is the localization error against ground truth. $\mathcal{L}_{vel}$ and $\mathcal{L}_{acc}$ penalize erratic changes in 3D velocity and acceleration, enforcing the prior that object parts follow smooth kinematic paths. This physical grounding Guerrier et al. (2025) acts as a regularizer: if the visual encoder drifts to a background pixel, the resulting 3D trajectory often violates kinematic smoothness, triggering a high loss that corrects the representation during training.

4 Experiments and Conclusion

Evaluation goal. We study whether QueST can detect and suppress silent semantic drift in long-horizon articulated video. Rather than exhaustive benchmarking, we design stress tests that isolate identity preservation under occlusion, articulation, and extended temporal horizons.

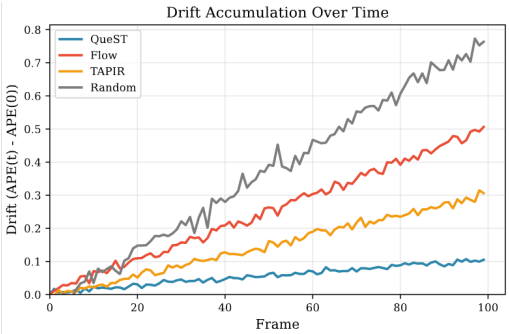


Figure 3: **Drift Analysis.** While Markovian trackers (RAFT-3D, CoTracker) exhibit near-linear error growth, QUEST maintains a bounded error curve via 3D physical grounding.

Setup. We evaluate on long-horizon articulated sequences from PartNet-Mobility Xiang et al. (2020) rendered in SAPIEN, which provide precise 3D ground truth for interaction-relevant regions (e.g., handles and joints). Sequences involve multi-joint articulation, partial occlusion, and viewpoint variation over extended durations ($T \geq 240$ frames). We compare QueST against RAFT, CoTracker, and TAP-Net.

Metrics. We report three drift-aware metrics that match Table 1: (i) **Absolute Point Error (APE)**: average 3D positional error; (ii) **Drift@100**: terminal error at the end of long-horizon sequences; and (iii) **Identity Accuracy**: percentage of frames where semantic identity is preserved.

Results. As shown in Table 1, QueST achieves large reductions in terminal drift and preserves semantic identity where correspondence-based trackers collapse under articulation and occlusion. Figure 3 shows that baselines accumulate near-linear, unbounded drift, while QueST maintains bounded error.

Method	APE ↓	D@100 ↓	Acc ↑
RAFT-3D	0.341	0.472	8.7%
CoTracker	0.276	0.398	19.2%
TAP-Net	0.251	0.372	21.4%
QueST	**0.081**	**0.155**	**86.5%**

Table 1: **Quantitative Comparison.** QueST achieves a 67.7% APE reduction.

Ablation. Removing persistent queries sharply increases identity switches, showing that semantic monitoring is essential. Removing 3D grounding leads to rapid drift under occlusion, confirming the importance of geometric consistency. Together, monitoring (queries) and grounding (geometry) are jointly necessary for reliable long-horizon tracking. We provide detailed ablations and quantitative results in Appendix B.

Conclusion We introduced **QueST**, a monitoring-by-design framework that reframes long-horizon tracking from local correspondence to representation-level semantic monitoring. By representing interaction-relevant entities as persistent semantic queries and constraining them with lightweight 3D grounding, QueST makes semantic drift observable and suppresses it before catastrophic failure. Although evaluated in SAPIEN simulation, the design naturally extends to real-world embodied perception tasks.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

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*, pp. 10061–10072, 2023.

Tobias Fischer, Thomas E Huang, Jiangmiao Pang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15380–15393, 2023.

Rhodri Guerrier, Adam W Harley, and Dima Damen. Pointst3r: Point tracking through 3d grounded correspondence. *arXiv preprint arXiv:2510.26443*, 2025.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Jingshun Huang, Haitao Lin, Tianyu Wang, Yanwei Fu, Xiangyang Xue, and Yi Zhu. Cap-net: A unified network for 6d pose and size estimation of categorical articulated parts from a single rgb-d image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11654–11664, 2025.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *European conference on computer vision*, pp. 18–35. Springer, 2024.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Skanda Koppula, Ignacio Rocco, Yi Yang, Joe Heyward, Joao Carreira, Andrew Zisserman, Gabriel Brostow, and Carl Doersch. Tapvid-3d: A benchmark for tracking any point in 3d. *Advances in Neural Information Processing Systems*, 37:82149–82165, 2024.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2008.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. *ECCV*, 2021.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F Wong, Zhuosheng Zhang, and Rui Wang. Embedding trajectory for out-of-distribution detection in mathematical reasoning. *Advances in Neural Information Processing Systems*, 37:42965–42999, 2024.

Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11097–11107, 2020.

Qiaojun Yu, Junbo Wang, Wenhai Liu, Ce Hao, Liu Liu, Lin Shao, Weiming Wang, and Cewu Lu. Gamma: Generalizable articulation modeling and manipulation for articulated objects. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5419–5426. IEEE, 2024.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K Barral, Raia Hadsell, et al. Efficiently reconstructing dynamic scenes one d4rt at a time. *arXiv preprint arXiv:2512.08924*, 2025.

A APPENDIX

A.1 DATASET PROTOCOLS AND SOURCE

Our dataset is derived from PartNet-Mobility Xiang et al. (2020), utilizing SAPIEN for physics-based rendering. We focus on articulated objects representative of everyday manipulation, including storage furniture (e.g., cabinets with doors/drawers), appliances (e.g., dishwashers), and hinged devices (e.g., laptops). Each object is normalized into a canonical pose Huang et al. (2025). We render synchronized RGB-D sequences from $V = 3$ static camera viewpoints to ensure the model generalizes across camera configurations and does not overfit to a single perspective.

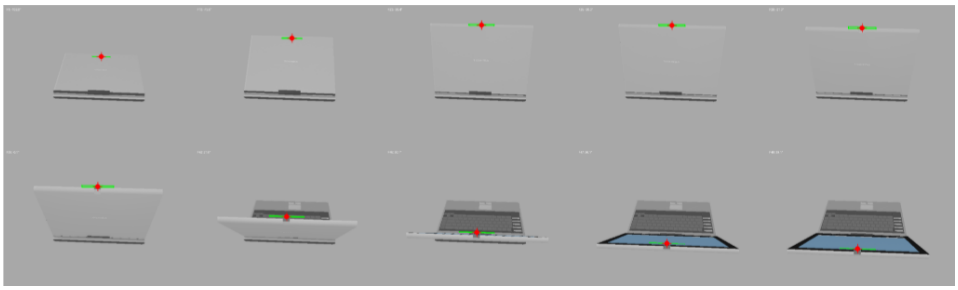


Figure 4: Single-joint interaction sequences (Phase 1) used for short-horizon training, with multi-view RGB-D frames and pixel-level affordance annotations.

Drift Evaluation Protocol. We generate sequences with increasing complexity: (1) **Phase 1 (Single-Joint)** actuates exactly one joint while others remain fixed (Figure 4); (2) **Long-Horizon** ($L \geq 2$) actuates joints sequentially (Figure 5), scaling to 240 frames at Level 4.

Figure 5: Multi-category, multi-step interaction sequences (Level 4) with cumulative joint actuation, designed to evaluate long-horizon temporal consistency under articulated motion and occlusion.

A.2 TRAINING SETUP

We employ a two-stage training process to decouple semantic identity learning from metric depth estimation.

Optimization. We use the **AdamW** optimizer with a learning rate of 1×10^{-4} and weight decay 0.01. Models are trained for up to 50 epochs (patience=15) using sliding temporal windows of $T = 4$ frames.

Stage 1: Perception (The Monitor). The QueST-Backbone maps video features to 2D trajectories. The objective enforces spatial accuracy while strictly penalizing physical inconsistency (drift):

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{aff}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}} + \lambda_{\text{bound}}\mathcal{L}_{\text{bound}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} \tag{2}$$

where $\mathcal{L}_{\text{vel}}$ and $\mathcal{L}_{\text{acc}}$ enforce temporal smoothness, and $\mathcal{L}_{\text{feat}}$ enforces cosine similarity between query embeddings across frames to prevent identity switching. We set $\lambda_{\text{vel}} = 1.0$ and $\lambda_{\text{acc}} = 0.5$.

Stage 2: Flow Prediction (The Adaptation). We freeze the backbone and train the flow head to predict 3D displacement vectors by minimizing the L_1 distance against ground-truth scene flow:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{flow}} \sum_{t,k} \left\| \hat{\mathbf{f}}_{t,k} - \mathbf{f}^*_{t,k} \right\|_1 \tag{3}$$

We did not observe collapse or mode drift across seeds; results are averaged over three runs.

A.3 Inference & Efficiency

At inference time, QueST processes RGB-D videos of arbitrary length without re-initialization. The model runs at approximately **30 FPS** on an NVIDIA RTX 6000 GPU, enabling real-time monitoring applications.

B Extended Quantitative Analysis

B.1 Phase Articulation Complexity on Drift

As shown in Table 2, QueST degrades gracefully as complexity increases from Level 1 to 4, whereas baselines exhibit near-monotonic drift growth. This confirms that persistent semantic queries and 3D grounding are essential for reliable tracking under multi-joint articulation.

Table 2: Performance across increasing articulation complexity.

Method	Manipulation Level (Number of Joints)															
	Level 1 (1)				**Level 2 (2)**				**Level 3 (3)**				**Level 4 (4)**			
	APE↓	D@50↓	D@100↓	Acc↑	APE↓	D@50↓	D@100↓	Acc↑	APE↓	D@50↓	D@100↓	Acc↑	APE↓	D@50↓	D@100↓	Acc↑
QueST (Full)	**0.065**	**0.058**	**0.125**	**0.920**	**0.070**	**0.063**	**0.135**	**0.902**	**0.075**	**0.067**	**0.145**	**0.883**	**0.081**	**0.072**	**0.155**	**0.865**
No Queries	0.088	0.075	0.194	0.754	0.104	0.089	0.229	0.709	0.119	0.103	0.264	0.664	0.135	0.116	0.298	0.619
No 3D	0.081	0.074	0.181	0.718	0.093	0.085	0.208	0.682	0.106	0.097	0.236	0.646	0.118	0.108	0.263	0.611
No Smoothness	0.077	0.078	0.244	0.736	0.094	0.096	0.297	0.685	0.110	0.113	0.351	0.633	0.127	0.130	0.405	0.581
Flow (0.01)	0.113	0.089	0.298	0.496	0.085	0.095	0.136	0.483	0.103	0.108	0.159	0.388	0.107	0.118	0.160	0.581
Flow (0.02)	0.158	0.143	0.342	0.249	0.132	0.126	0.205	0.253	0.197	0.209	0.292	0.157	0.225	0.233	0.333	0.144
Flow (0.05)	0.289	0.297	0.490	0.093	0.312	0.335	0.412	0.059	0.354	0.409	0.421	0.045	0.385	0.451	0.444	0.032

B.2 Temporal Context and Query Capacity

Increasing the temporal window from $T = 2$ to $T = 4$ substantially reduces drift and improves tracking accuracy, while gains diminish at $T = 8$. Varying the number of persistent queries shows that performance improves up to $K = 8$ and then saturates, motivating our default choice of $T = 4, K = 8$. See Table 3 for full results.

Table 3: Temporal window and query capacity ablation.

Ablation Setting	**APE (2D)↓**	**Drift@100↓**	**Tracking Acc↑**
Temporal Window Length T			
$T = 2$	0.142	0.285	0.68
$T = 4$	0.098	0.178	0.78
$\mathbf{T = 8}$	**0.072**	**0.125**	**0.86**
Number of Persistent Queries K			
$K = 4$	0.095	0.168	0.79
$\mathbf{K = 8}$	**0.072**	**0.125**	**0.86**
$K = 16$	0.068	0.118	0.87

B.3 Noise Robustness

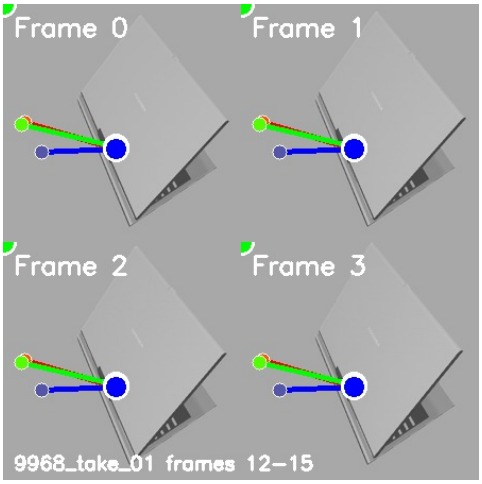
To test reliability under environmental drift, we evaluate QueST under Gaussian noise. Our model maintains $> 96\%$ accuracy at 5% noise levels, whereas correspondence-based baselines (RAFT,

CoTracker) collapse below 40%. This highlights the stability of global semantic queries over local pixel matching.

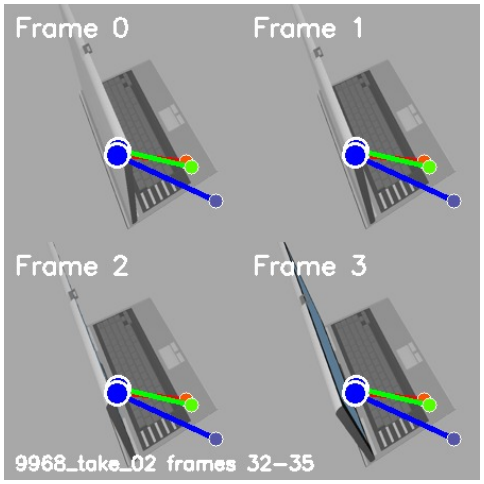
C Adaptation and Semantic Stability

C.1 Query-Conditioned Reasoning

Figure 6a demonstrates that QueST adapts its tracking behavior based on the specific query intent. Given the same input video, different queries (e.g., "Open" vs. "Lift") induce distinct, stable trajectories.



(a) **Opening sequence (Side View):** The blue query remains precisely localized to the lid corner despite rapid rotation.



(b) **Closing sequence (Top View):** The query preserves semantic identity across the closing arc without drifting into the background.

C.2 Hinged Articulation and Viewpoint Drift

The persistent query mechanism acts as a semantic monitor during extreme viewpoint changes. As shown in the hinged laptop sequences (Figures 6a and 6b in the main text), the query preserves identity even as the surface rotates 90°, effectively suppressing the "identity drift" that causes Markovian trackers to fail.

D Failure Case Analysis

In the spirit of analyzing model reliability, we identify two failure modes of the QueST framework:

1. **Extreme Occlusion (>80%):** If a handle is occluded for > 30 frames, the global attention mechanism may drift to a visually similar neighbor. This represents a limit of the current "semantic memory."
2. **Symmetric Ambiguity:** On objects with identical handles (e.g., a bank of lockers), the queries may occasionally switch between equivalent semantic targets. While the tracking remains "accurate" in a general sense, it violates strict identity preservation.

E Broader Implications

Future reliable agents should embed semantic monitoring directly into perception, rather than relying solely on post-hoc drift detectors or reactive retraining. Built-in semantic monitors can serve as an early-warning system for failure, enabling safer and more trustworthy AI systems in robotics, autonomous driving, and industrial automation.