EGOTRACK: ROBUST 6D OBJECT TRACKING VIA RELIABILITY-AWARE POSE HYPOTHESIS REASONING IN EXTREME EGOCENTRIC VISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust 6D object pose estimation is critical for egocentric and embodied systems operating under challenging conditions such as low light, motion blur, smoke and occlusion. Under such degradation, frame-wise pose estimates can become unreliable: selecting a single estimate may discard useful alternatives, while even the entire candidate set may be uninformative. We introduce *EgoTrack*, which retains multiple pose hypotheses and keeps their influence revisable during trajectory reconstruction. For each frame, a learned policy selects the object template appearance most useful for pose recovery, and a frozen estimator retains its top-K ranked pose hypotheses. Trajectory diffusion then repeatedly retrieves candidate evidence conditioned on the evolving trajectory state and reliability cues; temporal denoising updates the trajectory, which changes subsequent retrieval. A learned NULL token provides an alternative to real-candidate evidence when the local set is uninformative. Experiments show substantial improvements in 6D pose accuracy and trajectory consistency over frame-wise and temporal baselines under severe visual degradation.

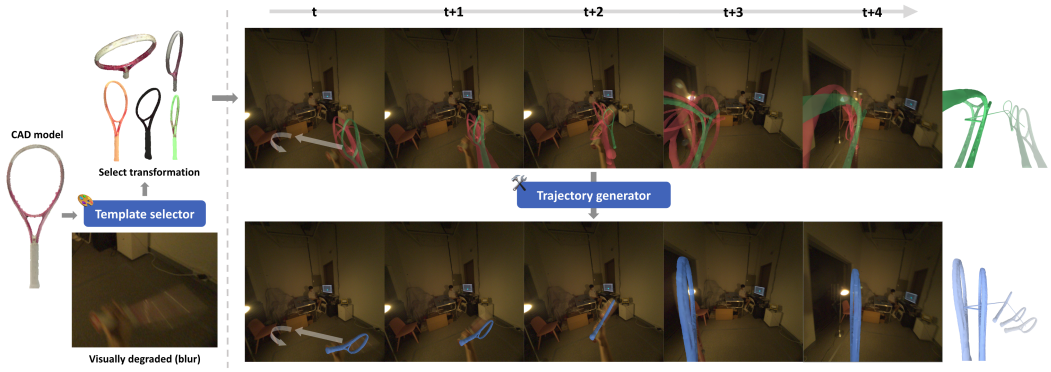

Figure 1: **Recovering an object pose trajectory under severe visual degradation.** For each degraded frame, our template selector chooses an appearance transformation for the CAD templates to facilitate pose estimation. The resulting frame-wise pose hypotheses remain ambiguous, with the top-ranked hypothesis shown in green and the other hypotheses in red. Rather than committing to the top-ranked hypothesis at each frame, our trajectory generator reasons over all candidates across time to recover a coherent 6D pose trajectory (blue).

1 INTRODUCTION

Estimating the 6D pose of objects is a fundamental problem in 3D visual perception, with applications ranging from robotic manipulation to augmented reality and embodied systems (Xiang et al., 2018; Wang et al., 2019; Labbe et al., 2020; Peng et al., 2019; Wang et al., 2021; Su et al., 2022; Lin et al., 2024). While recent generalizable pose estimators reduce the need for object-specific training by leveraging CAD models, rendered views, and learned correspondences, their performance still

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.

depends on reliable visual evidence for correspondence estimation (Labbé et al., 2022; Wen et al., 2024; Nguyen et al., 2024; Lin et al., 2023; Liu et al., 2025). This assumption may not hold under challenging conditions such as extreme illumination, severe motion blur, smoke, rapid camera motion, and occlusion, which can be particularly problematic in egocentric video and were shown by EgoXtreme (Yoon et al., 2026) to substantially degrade 6D object pose estimation performance.

Several strategies have been explored to improve robustness under unreliable visual evidence. Image restoration can produce visually plausible inputs for pose estimation, but may introduce artifacts that reduce pose accuracy (Yoon et al., 2026). Temporal methods propagate or refine pose estimates using local (Nguyen et al., 2025a) or longer-range temporal context (Ponimatkin et al., 2025), while generative pose estimators can model multiple plausible pose hypotheses for a single observation (Zhang et al., 2024). Structured generative models can further incorporate hand–object motion priors (Ye et al., 2026). These approaches establish the value of temporal context and uncertainty-aware inference, but leave a specific interface unresolved under severe degradation: how should a trajectory model use competing frame-wise pose hypotheses when their local ranking may be unreliable, and when the entire local set may be uninformative?

We introduce *EgoTrack*, a two-stage framework that keeps pose evidence revisable during trajectory reconstruction. **Trajectory context can help determine which hypothesis is useful, but that context must itself be inferred from uncertain hypotheses.** *EgoTrack* therefore couples these decisions: candidate evidence informs the trajectory update, and the updated trajectory changes how the same candidates contribute next. First, *task-oriented pose hypothesis generation* selects one appearance-transformation for object templates according to pose-recovery utility, from which a frozen pose estimator produces top-K hypotheses. Second, *state-conditioned candidate-set trajectory diffusion* repeatedly retrieves evidence from these hypotheses using the evolving trajectory state and candidate reliability cues. Temporal denoising updates the trajectory across frames, which changes subsequent candidate retrieval, while a learned NULL token provides an alternative to real-candidate evidence when the local set is unhelpful.

The retained hypotheses thus remain fixed after generation, but *their influence changes as temporal structure emerges*. On EgoXtreme, *EgoTrack* significantly improves 6D pose accuracy and trajectory consistency over the evaluated frame-wise and temporal baselines under extreme visual conditions.

Our contributions are twofold:

- **Task-oriented pose hypothesis generation.** We learn to select one rendered-template appearance per frame according to downstream pose-recovery utility and retain multiple ranked outputs from a frozen pose estimator, avoiding premature commitment to a single frame-wise pose.
- **State-conditioned candidate-set trajectory diffusion.** We keep the retained hypotheses accessible throughout trajectory reconstruction and repeatedly revise their influence using the evolving trajectory state and reliability cues; a learned NULL token provides an explicit alternative to the local candidate set.

2 Related Work

2.1 Generalizable 6D Object Pose Estimation under Challenging Conditions

6D object pose estimation has been extensively studied on benchmarks covering diverse objects, viewpoints, clutter, occlusion, and object symmetries, including LINEMOD and LM-O (Hinterstoisser et al., 2012; Brachmann et al., 2014), T-LESS (Hodan et al., 2017), YCB-Video (Xiang et al., 2018), with the BOP benchmark (Nguyen et al., 2025b). More recent datasets further consider challenging illumination and sensing conditions (Hodan et al., 2018; Tyree et al., 2022; Han et al., 2025), as well as egocentric hand–object interactions with 6D object pose annotations (Kwon et al., 2021; Banerjee et al., 2025). These benchmarks have driven substantial progress in 6D object pose estimation, but are not primarily designed to isolate failures caused by severe degradation of the visual observation itself.

Recent generalizable pose estimators recover unseen-object poses without object-specific training, using object models, rendered views, reference images, or image–object correspondences (Liu et al.,

2022; Sun et al., 2022; Labbé et al., 2022; Wen et al., 2024; Lin et al., 2023; Nguyen et al., 2024; Örnek et al., 2024; Liu et al., 2025). Among them, GigaPose (Nguyen et al., 2024) provides an efficient render-and-match formulation for novel-object pose estimation and serves as our frame-wise pose backbone.

Severe visual degradation, including extreme illumination, motion blur, smoke, and exposure failure, can damage the appearance and local correspondences required for pose recovery, particularly in egocentric video. Image restoration addresses motion blur (Rim et al., 2022), haze (Song et al., 2023), and low-light imaging (Feijoo et al., 2025), but improved human-perceptual quality does not necessarily imply better geometric inference. EgoXtreme (Yoon et al., 2026) was introduced to study this robustness gap and showed substantial degradation in state-of-the-art pose estimation under extreme conditions. It further showed that restoration can fail to recover the geometric evidence needed for pose estimation and can even reduce pose accuracy.

Many frame-wise pose estimation pipelines ultimately reduce each observation to a single pose prediction before temporal reasoning, while generative approaches such as GenPose (Zhang et al., 2024; Hsiao et al., 2024) explicitly represent multiple plausible pose hypotheses. Under severe degradation, however, the highest-scoring estimate may be incorrect while alternative template matches retain useful geometric evidence. EgoTrack therefore preserves multiple plausible hypotheses generated from an adaptively transformed template bank and defers the resolution to trajectory-level inference.

2.2 6D Object Pose Tracking and Temporal Reasoning

6D object tracking extends frame-wise estimation by exploiting temporal continuity through filtering, pose propagation, registration, correspondence, reconstruction, and iterative refinement (Deng et al., 2019; Wen et al., 2020; Sun et al., 2022; Wen & Bekris, 2021; Wen et al., 2023; 2024; Nguyen et al., 2025a). These approaches are effective when neighboring observations remain reliable and a previous pose provides a useful initialization.

GoTrack (Nguyen et al., 2025a) combines model-to-frame and lightweight frame-to-frame registration for generalizable causal pose tracking, serving as our causal baseline. FreePose (Ponimatkin et al., 2025) exploits video-level correspondence and trajectory refinement to recover temporally consistent 6D object poses, making it a natural baseline for our bidirectional setting. Other recent approaches incorporate structured interaction cues, including hand–object motion priors, to constrain object trajectories (Ye et al., 2026).

These methods demonstrate the benefit of temporal information, but primarily focus on propagating or refining pose observations once they have been obtained. Our setting exposes a different failure mode: under severe visual degradation, the frame-wise observation itself may be ambiguous, confidently incorrect, or unusable. EgoTrack therefore performs temporal reasoning over competing pose hypotheses rather than committing to a single observation beforehand. The evolving trajectory state dynamically associates with these hypotheses, candidate reliability is encoded jointly with pose evidence, and a learned NULL token allows unreliable local candidate evidence to be discounted when none of the candidates is sufficiently trustworthy.

3 Approach

Figure 2 illustrates the full pipeline of *EgoTrack*. Given an RGB sequence $\{I_t\}_{t=1}^{T}$ and the CAD model of a target object, our goal is to estimate its 6D pose trajectory $\hat{\mathbf{P}}_{1:T} = \{\hat{P}_t\}_{t=1}^{T}$. Under severe visual degradation, a single frame may provide ambiguous or misleading pose evidence, making premature commitment to a single pose hypothesis particularly unreliable for tracking.

EgoTrack consists of two stages. **Stage 1** expands frame-wise pose evidence by generating multiple plausible pose hypotheses rather than committing to a single pose estimation. A learned transformation policy selects a template appearance according to expected pose-recovery utility, and a frozen GigaPose (Nguyen et al., 2024) estimator uses the transformed templates to produce K pose hypotheses. Each pose has a corresponding reliability cue.

Stage 2 uses temporal trajectory context to determine which pose hypotheses should influence the final trajectory. At each diffusion step, the current noisy trajectory state queries the frame-wise pose

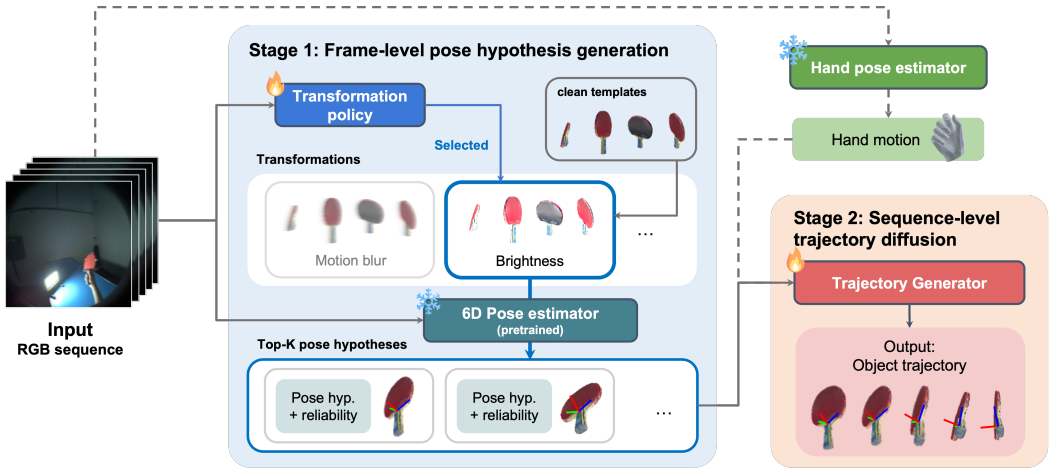

Figure 2: **Overview of the *EgoTrack* pipeline.** Given an RGB sequence and object templates, *EgoTrack* generates multiple frame-wise pose hypotheses and resolves them jointly over time to recover the final 6D object pose trajectory.

candidates through cross-attention, allowing the model to dynamically reweight pose hypotheses as the trajectory is progressively refined. Reliability is encoded as part of each candidate representation, while a learned NULL token allows the model to abstain from the real candidate set when none of the candidates is trustworthy. The resulting evidence is fused with the encoded trajectory state and optional hand information, and processed across the temporal window to predict the clean trajectory for the subsequent denoising update.

3.1 Stage 1: Task-Oriented Pose Hypothesis Generation

Let $\mathcal{T}$ be a bank of clean object templates used by the pretrained 6D pose estimator, and let $\mathcal{A}$ be a finite set of appearance transformations, including variations in brightness, contrast, color, and motion blur. Applying $a \in \mathcal{A}$ to all templates yields an augmented template bank $\mathcal{T}^a = \{a(\tau) \mid \tau \in \mathcal{T}\}$.

Given I_t, the transformation policy assigns each transformation a score that estimates its utility for recovering the pose in that frame. The policy selects the transformation $a_t \in \mathcal{A}$ with the highest score, and the pretrained pose estimator processes I_t using the corresponding template bank $\mathcal{T}^{a_t}$. It generates K pose hypotheses $P_{t,k}$, and the corresponding reliability cue $r_{t,k}$ includes signals such as the policy score and the estimator's template retrieval score. We use $K = 5$ in all main experiments. Details of the policy objective, augmentation-bank construction, and reliability cues are provided in the appendix.

For each frame t, the policy selects a single transformation a_t. Thus, all K hypotheses within the frame share the same transformation identity, while a_t may vary across frames. We define the frame-wise hypothesis set as

$$\mathcal{H}_t = \{(P_{t,k}, a_t, r_{t,k})\}_{k=1}^{K}. \tag{1}$$

During training-data preprocessing, we run the frozen pose estimator over each transformation and train the policy with binary cross-entropy using pose-recovery success labels derived from the resulting pose accuracy; details are provided in the appendix.

3.2 Stage 2: State-Conditioned Candidate-Set Trajectory Diffusion

We formulate trajectory-level hypothesis resolution using conditional denoising diffusion (Ho et al., 2020). Figure 3 illustrates the Stage 2 architecture. Let $\mathbf{x}_{1:T}^{(0)}$ denote the ground truth trajectory used for diffusion, and let $\mathbf{x}_{1:T}^{(s)}$ denote its noisy version at diffusion step s. At each denoising step, the updated trajectory state re-evaluates the candidate set, allowing different pose hypotheses to be emphasized as the trajectory is refined.

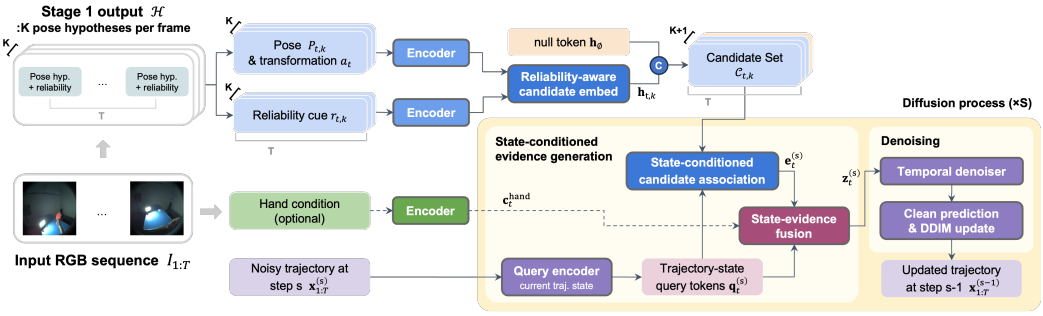

Figure 3: **Stage 2: State-conditioned candidate-set trajectory diffusion.** At each diffusion step, the current noisy trajectory is encoded into frame-wise query tokens that attend to the K pose candidates and a learned NULL token. The retrieved evidence is fused with the trajectory state and optional hand condition, and a temporal denoiser predicts the clean trajectory used for the subsequent denoising update.

3.2.1 Candidate Evidence Encoding

For each hypothesis in $\mathcal{H}_t$, separate encoders embed the pose $P_{t,k}$ with transformation identity a_t and the corresponding reliability cue $r_{t,k}$. The resulting representations are concatenated and mapped through a learned projection to obtain the candidate token $\mathbf{h}_{t,k}$, which jointly represents the estimated pose, the applied transformation, and its reliability.

For each frame t and diffusion step s, the current pose estimate $\mathbf{x}_t^{(s)}$ and an embedding of diffusion step s are encoded into a query $\mathbf{q}_t^{(s)}$ that attends to the frame-wise pose candidates. We append a shared learned NULL token $\mathbf{h}_\emptyset$ to the K candidate tokens, allowing the model to downweight unreliable frame-wise evidence when none of the real candidates is sufficiently consistent with the current trajectory estimate. The resulting candidate set is $\mathcal{C}_t = \{\mathbf{h}_{t,1}, \ldots, \mathbf{h}_{t,K}, \mathbf{h}_\emptyset\}$. The NULL token is shared, while its attention weight varies with the frame and diffusion state.

3.2.2 State-Conditioned Pose Hypothesis Association

State-conditioned candidate association. The trajectory-state query $\mathbf{q}_t^{(s)}$ attends to the candidate set $\mathcal{C}_t$ to retrieve the candidate evidence adaptively at each denoising step. We compute the state-conditioned observation evidence using cross-attention:

$$\mathbf{e}_t^{(s)} = \sum_{j \in \{1,\ldots,K,\emptyset\}} \text{softmax}_j \left(\frac{(\mathbf{W}_Q \mathbf{q}_t^{(s)})^\top (\mathbf{W}_K \mathbf{h}_{t,j})}{\sqrt{d_k}} \right) \mathbf{W}_V \mathbf{h}_{t,j}. \tag{2}$$

where d_k is the key dimension. Because the association is recomputed at every diffusion step, $\mathbf{e}_t^{(s)}$ can change as the noisy trajectory evolves. The model therefore avoids committing to a fixed pose hypothesis before trajectory inference.

State-Evidence Fusion The observation evidence $\mathbf{e}_t^{(s)}$ summarizes what the candidate set contributes at the current diffusion step, while $\mathbf{q}_t^{(s)}$ retains the current trajectory state. We fuse the trajectory-state query and observation evidence, together with optional hand conditioning, before temporal reasoning.

When hand observations are available, we use hand tracking outputs from Aria Machine Perception Services (MPS) (Engel et al., 2023) to construct the hand pose sequence which is encoded as an additional frame-wise condition $\mathbf{c}_t^{\text{hand}}$ and omitted otherwise. The fused state-evidence representation is

$$\mathbf{z}_t^{(s)} = F_{\text{fuse}} \left(\mathbf{q}_t^{(s)}, \mathbf{e}_t^{(s)}, \mathbf{c}_t^{\text{hand}} \right). \tag{3}$$

3.2.3 INTER-FRAME TRAJECTORY DENOISING

Before temporal reasoning, we add the same diffusion-step embedding used above and a temporal positional encoding to $\mathbf{z}_t^{(s)}$. The resulting frame tokens within the temporal window are processed jointly using a Transformer-based denoiser (Vaswani et al., 2017). The Transformer outputs temporally contextualized frame representations, which are mapped by an output head to a prediction of the clean trajectory $\mathbf{x}_{1:T}^{(0)}$.

The trajectory diffusion objective uses MSE to predict the clean trajectory $\mathbf{x}_{1:T}^{(0)}$ from its noisy state. At inference time, iterative DDIM sampling (Song et al., 2021) updates the noisy trajectory from $\mathbf{x}_{1:T}^{(s)}$ to $\mathbf{x}_{1:T}^{(s-1)}$ and repeatedly recomputes state-conditioned candidate evidence. After the final denoising step, the recovered clean trajectory $\hat{\mathbf{x}}_{1:T}^{(0)}$ is decoded into the estimated 6D pose trajectory $\hat{\mathbf{P}}_{1:T}$. The candidate poses serve as conditioning evidence rather than a discrete output space; consequently, the predicted pose need not coincide with any supplied hypothesis and can be refined beyond the frame-wise candidates using temporal context. Detailed encoder architectures, reliability parameterization, diffusion training, and optimization settings are provided in the appendix.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate *EgoTrack* on EgoXtreme (Yoon et al., 2026), which contains egocentric object pose sequences under illumination changes, motion blur, smoke, and rapid camera motion. We use the predefined training, validation, and test splits, with model selection on validation and final evaluation on the held-out test split. All methods use the same object models, camera calibration, evaluation protocol, and ground-truth object masks.

We report ADD(-S) AUC up to $0.3d$, median translation and rotation errors, and RPE RMSE at $\delta = 10$ frames. AUC and success rates are reported in percentage points, translation in meters, and rotation in degrees; rotation metrics exclude symmetric objects. For Stage 1 oracle analysis, we additionally report pose success at $0.1d$, $0.2d$, and $0.3d$. Unless otherwise stated, *EgoTrack* uses Selected Aug-1 Rank-5 with $K = 5$. Implementation details are provided in the appendix.

Baselines and inference settings. GigaPose (Nguyen et al., 2024) serves as the frame-wise reference, FreePose (Ponimatkin et al., 2025) as the bidirectional baseline, and GoTrack (Nguyen et al., 2025a) as the causal baseline. Our causal variant uses the same architecture with a causal temporal attention mask. We use bidirectional inference as the primary setting and additionally report causal and hand-conditioned variants using Project Aria MPS (Engel et al., 2023).

4.2 COMPARISON WITH 6D OBJECT POSE TRACKING METHODS

Table 1 compares *EgoTrack* with frame-wise and temporal baselines under bidirectional and causal inference. FreePose uses Clean GigaPose Top-1 observations; we also report a privileged variant receiving the GT-best pose among our Stage 1 $K = 5$ candidates at each frame. *EgoTrack* improves AUC and RPE over this privileged-input baseline across all three scenarios.

For causal inference, GoTrack is initialized at each frame from the preceding frame's raw GigaPose estimate rather than its previous refined output. *EgoTrack* improves pose accuracy in Sports and Emergency and reduces RPE across all three scenarios, while remaining comparable to GoTrack in Maintenance AUC. Hand-conditioned variants are also reported.

4.3 ANALYSIS OF DEGRADATION-ROBUST POSE HYPOTHESIS GENERATION

We next isolate Stage 1 from trajectory inference. Because Stage 2 operates on a set of competing pose hypotheses, the purpose of Stage 1 is not simply to improve the top-ranked frame-wise estimate. Instead, it should provide a compact candidate set that contains plausible poses from which the trajectory model can recover.

Table 1: **6D object pose tracking on EgoXtreme.** We report ADD(-S) AUC up to $0.3d$, median translation and rotation error, and RPE RMSE at $\delta = 10$ frames. Rotation metrics exclude symmetric objects. † uses ground truth to select the best pose from the Stage 1 $K = 5$ candidates at each frame and is included only as a privileged reference.

	EgoXtreme - Sports					EgoXtreme - Maintenance					EgoXtreme - Emergency				
Method	AUC ↑	Trans. ↓	Rot. ↓	RPE-T ↓	RPE-R ↓	AUC ↑	Trans. ↓	Rot. ↓	RPE-T ↓	RPE-R ↓	AUC ↑	Trans. ↓	Rot. ↓	RPE-T ↓	RPE-R ↓
GigaPose	8.33	0.2320	128.20	3.6018	114.28	5.52	0.0694	99.96	0.9705	91.47	4.02	0.1796	145.67	3.0906	92.28
Bidirectional															
FreePose Tracking (GigaPose Input)	6.55	0.2240	122.32	2.3910	79.94	6.12	0.0611	93.01	0.4630	39.24	5.64	0.1667	97.77	1.5038	35.88
FreePose Tracking (K5 Oracle Input)†	12.51	0.1255	115.80	1.9399	80.30	17.38	**0.0207**	81.71	0.0958	38.55	14.08	**0.0638**	86.33	0.5531	34.00
Ours-Bidirectional	47.07	0.0777	14.71	0.2721	26.45	**27.24**	0.0363	13.82	0.0328	20.75	23.29	0.0670	17.06	**0.0668**	**14.42**
Ours-Bidirectional + Hand	**54.73**	**0.0616**	**14.49**	**0.2541**	**21.54**	26.56	0.0366	**13.74**	**0.0313**	**17.99**	**24.22**	0.0672	**15.73**	0.0681	15.48
Causal															
GoTrack	9.09	0.3780	147.75	6.6631	121.47	**26.45**	**0.0363**	87.39	2.2569	102.32	18.48	0.3703	114.74	11.6295	110.98
Ours-Causal	42.40	0.0891	15.54	0.3279	**39.11**	25.71	0.0374	14.09	0.0400	27.45	20.84	0.0752	15.62	0.1012	**21.91**
Ours-Causal + Hand	**48.09**	**0.0715**	**14.56**	**0.3240**	43.44	26.03	0.0373	**13.33**	**0.0387**	**24.93**	**22.66**	**0.0693**	**14.89**	**0.0914**	22.16

Table 2: **Analysis of degradation-robust pose hypothesis generation.** Oracle coverage measures whether at least one generated hypothesis successfully recovers the pose.

				EgoXtreme - Sports			EgoXtreme - Maintenance			EgoXtreme - Emergency		
Candidate generation	# Aug.	# Rank.	# Hyp.	@0.1d ↑	@0.2d ↑	@0.3d ↑	@0.1d ↑	@0.2d ↑	@0.3d ↑	@0.1d ↑	@0.2d ↑	@0.3d ↑
Clean Aug-1 Rank-1	1	1	1	1.44	10.57	23.34	0.47	6.77	23.65	0.39	5.55	14.31
Clean Aug-1 Rank-5	1	5	5	4.55	22.48	43.51	2.59	24.77	54.65	1.69	17.50	30.42
Retrieval Global Top-5	20	5	5	2.70	18.95	37.57	1.50	17.06	46.46	1.30	12.09	23.70
RANSAC Global Top-5	20	5	5	3.43	20.30	38.97	2.06	21.34	51.56	1.27	13.47	25.77
Ours Selected Aug-5 Rank-1	5	1	5	3.19	20.61	40.24	1.87	21.58	54.50	1.89	15.52	28.07
Ours Selected Aug-1 Rank-5	1	5	5	4.39	24.14	44.69	2.59	25.28	58.09	2.24	18.93	31.48
All Aug. Rank-1	20	1	20	6.07	28.87	51.91	3.96	36.16	72.11	3.54	21.79	35.58

Table 2 reports oracle pose coverage for different hypothesis-generation strategies. Here, # Aug. denotes the number of transformed template banks considered, # Rank denotes the pose ranks available from each bank, and # Hyp. denotes the final number of retained pose hypotheses. Oracle coverage measures whether at least one retained hypothesis satisfies the corresponding ADD(-S) threshold.

Using multiple pose ranks from a single template bank already substantially increases recoverability compared with committing to the top-ranked pose. However, the choice of template appearance remains important. Global retrieval- or RANSAC-based selection over the full transformation bank does not consistently yield stronger candidate sets despite searching a substantially larger space.

Our transformation policy instead selects an appearance according to downstream pose-recovery utility. Keeping the top five pose hypotheses from the selected template bank (*Selected Aug-1 Rank-5*) provides strong oracle coverage across the three scenarios while retaining only five hypotheses per frame. This result motivates the Stage 1 configuration used by our final model: the policy first selects a pose-recovery-oriented template appearance, and multiple ranked pose hypotheses from that appearance are preserved for trajectory-level resolution.

4.4 Effect of Candidate-Set Configuration

Oracle coverage alone does not determine the quality of the final trajectory. We therefore evaluate how different candidate-set configurations affect the same Stage 2 trajectory architecture. All variants use the same bidirectional no-hand architecture; only the pose hypotheses provided by Stage 1 are changed.

Table 3 shows that providing too little pose evidence limits trajectory recovery. The strong performance of the single-hypothesis configurations indicates that trajectory-level reconstruction itself provides substantial temporal recovery. Nevertheless, both single-hypothesis settings are generally weaker than configurations that preserve multiple alternatives, showing that preserving multiple hypotheses provides additional gains and supports delaying the frame-wise pose decision. Notably, policy-based template selection does not consistently improve the single-hypothesis setting, suggesting that its benefit in our pipeline emerges in conjunction with preserving multiple alternatives for trajectory reasoning.

At the same time, simply increasing the number of candidates does not monotonically improve performance. Using all transformed-template rank-1 hypotheses produces a substantially larger candidate set but does not improve pose AUC and yields mixed gains in trajectory consistency.

Table 3: **Effect of candidate-set configuration.** All variants use the same Stage 2 architecture and bidirectional no-hand setting; only the candidate set provided by Stage 1 is changed.

				EgoXtreme - Sports			EgoXtreme - Maintenance			EgoXtreme - Emergency		
Candidate generation	# Aug.	# Rank.	# Hyp.	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑	RPE-T ↓	RPE-R ↓
Clean Aug-1 Rank-1	1	1	1	41.95	0.2816	29.36	24.64	0.0334	22.00	17.07	0.0797	18.96
Ours Selected Aug-1 Rank-1	1	1	1	40.84	0.2679	27.01	16.62	0.0372	21.59	18.94	0.0746	19.39
Ours Selected Aug-5 Rank-1	5	1	5	46.38	**0.2609**	28.58	25.63	**0.0322**	**19.66**	**24.24**	0.0715	15.20
Ours Selected Aug-1 Rank-5	1	5	5	**47.07**	0.2721	26.45	**27.24**	0.0328	20.75	23.29	**0.0668**	**14.42**
All Aug. Rank-1	20	1	20	45.86	0.2641	**24.16**	23.02	0.0329	19.71	17.80	0.0733	16.92

Table 4: **Ablation of candidate-set trajectory diffusion.** All variants use the same bidirectional no-hand setting. AUC denotes ADD(-S) AUC up to $0.3d$; RPE-T and RPE-R are RMSE at $\delta = 10$ frames.

			EgoXtreme - Sports			EgoXtreme - Maintenance			EgoXtreme - Emergency		
Variant	State Query	NULL	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑	RPE-T ↓	RPE-R ↓
w/o State Query & NULL	–	–	40.36	0.2801	27.11	23.61	0.0332	20.39	14.94	0.0671	16.75
w/o State Query	–	✓	38.55	0.2721	26.58	21.56	0.0349	20.45	19.13	0.0694	**14.00**
w/o NULL	✓	–	45.31	**0.2587**	**25.07**	23.51	**0.0317**	**19.37**	21.69	**0.0666**	15.37
Full	✓	✓	**47.07**	0.2721	26.45	**27.24**	0.0328	20.75	**23.29**	0.0668	14.42

Among the evaluated configurations, Selected Aug-1 Rank-5 achieves the highest AUC in Sports and Maintenance while remaining competitive in trajectory consistency, and we therefore use it as the default configuration. Together with the Stage 1 oracle analysis, these results suggest that robust tracking benefits from preserving sufficient pose alternatives while constructing an appropriate candidate set for trajectory-level reasoning.

4.5 Ablation Study

Finally, we isolate two key mechanisms for resolving the candidate set during trajectory diffusion. All variants use the same bidirectional no-hand setting and the default *Selected Aug-1 Rank-5* candidate configuration. We ablate the State Query and NULL mechanisms individually and jointly.

State-conditioned candidate association. The *w/o State Query* variant replaces the trajectory-state query used for candidate association with a learned static query. The current trajectory state is still passed to the subsequent state-evidence fusion module, isolating the role of state-conditioned candidate retrieval. Removing the state query reduces AUC, showing that candidate association benefits from adapting to the evolving denoising state.

NULL-based abstention. The *w/o NULL* variant retains the same pose hypotheses and state-conditioned association but removes the option to abstain from the real candidate set. With state-conditioned association, adding NULL improves AUC across all three scenarios, while five of the six RPE measures worsen slightly. This trade-off suggests that NULL primarily benefits absolute pose recovery rather than relative trajectory consistency. Additional qualitative comparisons with and without the NULL token are provided in Appendix D.

Removing both mechanisms also reduces AUC across all three scenarios. Together, these ablations support the combined Stage 2 design: candidate association should adapt to the evolving trajectory state, while the model should also be able to discount unreliable frame-wise evidence.

4.6 Qualitative Results

Figure 4 presents representative tracking results from the Sports and Emergency scenarios of EgoXtreme. The sequences contain challenging frame-wise observations caused by rapid motion and severe illumination degradation. While Clean GigaPose Top-1 estimates become inaccurate or unstable in some frames, *EgoTrack* remains more closely aligned with the ground truth over time. These examples illustrate the benefit of resolving multiple pose hypotheses using temporal context. Additional qualitative results are provided in the Appendix.

We further provide a qualitative real-world demonstration beyond EgoXtreme. Figure 5 shows tracking results for a ping-pong paddle and a fire extinguisher in real-world scenes. For the real-world

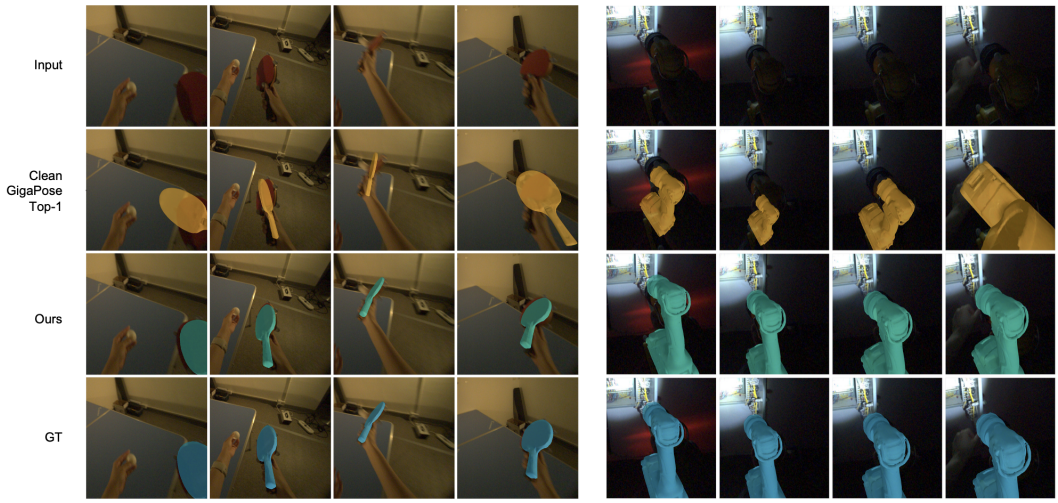

Figure 4: **Qualitative results on EgoXtreme.** Representative sequences from Sports (left) and Emergency (right). Columns show temporally sampled frames, and rows show the RGB input, Clean GigaPose Top-1, *EgoTrack*, and ground truth, respectively. Clean GigaPose Top-1 exhibits substantial pose errors in challenging frames, whereas *EgoTrack* remains more closely aligned with the ground-truth pose over time.

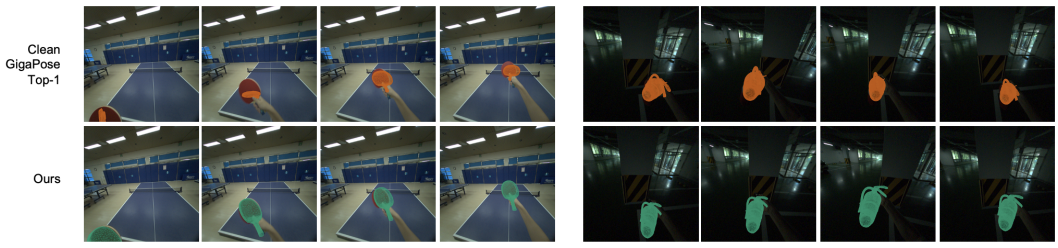

Figure 5: **Qualitative real-world demonstration.** Results on a ping-pong paddle (left) and a fire extinguisher (right). Columns show temporally sampled frames, and rows show GigaPose Top-1 and *EgoTrack*, respectively. *EgoTrack* produces more temporally stable pose estimates than the frame-wise baseline.

demonstration, object masks are obtained using SAM 3 (Carion et al., 2025), which is used only as a segmentation front-end and is not part of *EgoTrack*. Compared with Clean GigaPose Top-1, *EgoTrack* produces more stable and visually consistent pose estimates over time. Although no ground-truth annotations are available, these results suggest that *EgoTrack* remains effective beyond the benchmark setting. Corresponding videos are provided in the supplementary material.

5 Conclusion

We presented *EgoTrack*, a robust 6D object tracking framework that preserves multiple pose hypotheses instead of committing to a single unreliable frame-wise estimate under severe visual degradation. A learned policy selects a template appearance according to pose-recovery utility, while state-conditioned candidate-set trajectory diffusion re-evaluates the resulting pose hypotheses over time using reliability cues and NULL-based abstention. Experiments on EgoXtreme show that the proposed hypothesis generation expands the recoverable pose space and that preserving multiple alternatives improves pose accuracy and trajectory consistency. Overall, our results demonstrate the benefit of delaying frame-wise pose decisions and resolving ambiguity through temporal reasoning.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language polishing, literature retrieval and discovery, research ideation and experimental planning, code development and debugging, and drafting and revising portions of the manuscript. All AI-assisted research ideas, experimental designs, implementations, analyses, and written content were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper, including all claims, results, and artifacts produced with the assistance of generative AI.

REFERENCES

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7061–7071, 2025.

Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pp. 536–551. Springer, 2014.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719.

Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao-blackwellized particle filter for 6d object pose tracking. In *Robotics: Science and Systems (RSS)*, 2019.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023.

Daniel Feijoo, Juan C. Benito, Marcos V. Conde, and Alvaro Garcia. Darkir: Robust low-light image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10879–10889, 2025.

Yegyu Han, Taegyoon Yoon, Dayeon Woo, Sojeong Kim, and Hyung-Sin Kim. Senseshift6d: Multimodal rgb-d benchmarking for robust 6d pose estimation across environment and sensor variations. *arXiv preprint arXiv:2507.05751*, 2025.

Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pp. 548–562. Springer, 2012.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 880–888. IEEE, 2017.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 19–34, 2018.

Tsu-Ching Hsiao, Hao-Wei Chen, Hsuan-Kung Yang, and Chun-Yi Lee. Confronting ambiguity in 6d object pose estimation via score-based diffusion on se(3). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 352–362, 2024.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10138–10148, 2021.

Y. Labbe, J. Carpentier, M. Aubry, and J. Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare. In *CoRL*, 2022.

Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. *arXiv preprint arXiv:2311.15707*, 2023.

Yongliang Lin, Yongzhi Su, Praveen Nathan, Sandeep Inuganti, Yan Di, Martin Sundermeyer, Fabian Manhardt, Didier Stricker, Jason Rambach, and Yu Zhang. Hipose: Hierarchical binary surface encoding and correspondence pruning for rgb-d 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10148–10158, 2024.

Lihua Liu, Jiehong Lin, Zhenxin Liu, and Kui Jia. Picopose: Progressive pixel-to-pixel correspondence learning for novel object pose estimation. *arXiv preprint arXiv:2504.02617*, 2025.

Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *ECCV*, 2022.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Van Nguyen Nguyen, Christian Forster, Bugra Tekin, Sindi Shkodrani, Vincent Lepetit, Cem Keskin, and Tomáš Hodaň. Gotrack: Generic 6dof object pose refinement and tracking. *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2025a.

Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Médéric Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, Stan Birchfield, Jiří Matas, Yann Labbé, Martin Sundermeyer, and Tomáš Hodaň. BOP challenge 2024 on model-based and model-free 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2025b.

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomáš Hodaň. Foundpose: Unseen object pose estimation with foundation features. *European Conference on Computer Vision (ECCV)*, 2024.

Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019.

Georgy Ponimatkin, Martin Cífka, Tomáš Souček, Médéric Fourmy, Yann Labbé, Vladimir Petrik, and Josef Sivic. 6D Object Pose Tracking in Internet Videos for Robotic Manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jaesung Rim, Geonung Kim, Jungeon Kim, Junyong Lee, Seungyong Lee, and Sunghyun Cho. Realistic blur synthesis for learning image deblurring. In *European Conference on Computer Vision (ECCV)*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023.

Jrgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. pp. 573–580, 10 2012. ISBN 978-1-4673-1737-5. doi: 10.1109/IROS.2012.6385773.

Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. *arXiv preprint arXiv:2203.09418*, 2022.

Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. *CVPR*, 2022.

Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 13081–13088. IEEE, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. 2019.

Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16611–16621, June 2021.

B Wen and Kostas E Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2021.

Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E. Bekris. se(3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2020. doi: 10.1109/iros45743.2020.9341314. URL http://dx.doi.org/10.1109/IROS45743.2020.9341314.

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *CVPR*, 2023.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018.

Yufei Ye, Jiaman Li, Ryan Rong, and C. Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *CVPR Findings*, 2026.

Taegyoon Yoon, Yegyu Han, Seojin Ji, Jaewoo Park, Sojeong Kim, Taein Kwon, and Hyung-Sin Kim. Egoxtreme: A dataset for robust object pose estimation in egocentric views under extreme conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

A Implementation Details

A.1 Stage 1: Pose Hypothesis Generation

We use the pretrained GigaPose (Nguyen et al., 2024) model as the frozen frame-wise pose estimator throughout Stage 1. Given an input image, a transformation policy first selects one appearance transformation, which is applied to the full template set. GigaPose then performs template retrieval and pose estimation using the selected template bank, and we retain its top-K ranked pose predictions. Unless otherwise stated, we use $K = 5$.

The transformation bank contains 20 predefined template appearances, including the clean appearance and variations in brightness, contrast, color, and motion blur. The transformations are applied only to the rendered object templates; the observed RGB image is not restored or modified. Thus, the transformations serve as hypothesis-generating operators rather than explicit estimates of the underlying image degradation.

A.2 Transformation Policy

The transformation policy is trained to predict the downstream pose-recovery utility of each template appearance. The transformation policy uses an ImageNet-pretrained ConvNeXt-Tiny backbone with 224×224 input crops and a 512-D hidden representation. We train the policy for 10 epochs with a batch size of 64. We use learning rates of 1×10^{-5} for the backbone and 1×10^{-4} for the prediction head, with weight decay 1×10^{-4}. To construct supervision, we run the frozen pose estimator for every transformation during training-data preprocessing and evaluate its top-ranked pose prediction against the ground-truth pose. A transformation is labeled successful when its top-ranked prediction satisfies the $0.3d$ ADD(-S) pose-success criterion. The policy is trained with binary cross-entropy over these binary pose-recovery labels.

We use two-fold scene-level cross-fitting to generate policy predictions for trajectory training. For each fold, the policy is trained on the other scene fold and applied to the held-out fold, producing out-of-fold policy scores without using labels from that held-out fold for training.

At inference time, the policy evaluates all available transformations from the input observation and selects the highest-scoring transformation a_t. Only the selected template bank is passed to GigaPose, after which its top-K ranked pose hypotheses are retained. This keeps the final hypothesis set compact while avoiding exhaustive pose estimation over all transformed template banks.

A.3 Candidate Representation and Reliability Cues

Each pose hypothesis is encoded using its 6D pose, transformation identity, and a set of reliability cues. The pose is represented by a 9-D state consisting of 3-D translation and the 6-D rotation representation. The transformation identity is retained separately as a categorical index and embedded by the model rather than treated as a continuous reliability value.

In the default configuration, the reliability cue $r_{t,k}$ contains eight inference-time features. Six are derived from the transformation policy: `selector_prob`, `selector_score`, `selector_rank`, `selection_rank`, `selection_score`, and `diversity_penalty`. The remaining two, `retrieval_score` and `ransac_score`, are provided by the frozen pose estimator and capture template-retrieval quality and geometric consistency, respectively. No ground-truth pose error, oracle label, or pose-success indicator is used as a candidate feature.

Translation and rotation are encoded by separate 128-D branches, while the categorical transformation identity is represented by a 32-D embedding. These representations are concatenated and projected to form a 256-D pose token. The reliability cues are normalized over valid candidates, separately encoded using a 128-D hidden representation, and projected to the same 256-D model space. The pose token and reliability embedding are then fused to obtain the candidate token $\mathbf{h}_{t,k}$. Invalid candidates are masked before candidate association. A shared learned NULL token $\mathbf{h}_{\emptyset}$ is appended to the real candidate set and provides an explicit alternative when the local candidate evidence should be discounted.

The reliability encoder additionally predicts candidate-wise translation and rotation uncertainty. We calibrate these uncertainty estimates against the corresponding candidate pose errors using Gaussian negative log-likelihood. Translation uncertainty is calibrated against the 3-D translation residual, while rotation uncertainty is calibrated against the geodesic rotation error. This auxiliary objective calibrates the uncertainty heads while leaving the reliability representation free to support candidate reasoning.

A.4 Candidate-Set Trajectory Diffusion

Stage 2 operates on a temporal sequence of the candidate sets generated by Stage 1. At diffusion step s, the current noisy pose state $\mathbf{x}_t^{(s)}$, together with the diffusion-step embedding, is encoded into the trajectory-state query $\mathbf{q}_t^{(s)}$. This query attends to the K candidate tokens and the shared NULL token to produce the state-conditioned observation evidence $\mathbf{e}_t^{(s)}$.

The trajectory model uses a feature dimension of 256. Candidate association uses eight attention heads. In the default configuration, no additional self-attention is applied among the candidate tokens, and no explicit reliability bias is added to the candidate-attention logits. The query and retrieved candidate evidence are fused into a frame-wise representation. When hand observations are used, the corresponding hand pose feature is included as an additional condition.

The diffusion-step embedding and temporal positional encoding are added to the fused frame tokens. The temporal denoiser consists of six Transformer layers with eight attention heads, a feed-forward expansion ratio of 4, and dropout of 0.1. The resulting frame tokens are processed jointly over the temporal window, and the temporally contextualized representations are mapped to a prediction of the clean trajectory. The primary model uses bidirectional temporal attention, while the causal variant applies a causal attention mask to prevent access to future frames. Because candidate association is recomputed at every diffusion step, the contribution of individual pose hypotheses can change as the estimated trajectory evolves.

A.5 Training and Inference Details

We train the transformation policy and trajectory diffusion model separately, while keeping the pretrained GigaPose estimator frozen. The Stage 1 pose hypotheses used for trajectory training are precomputed.

For trajectory diffusion, we use temporal windows of 64 frames with a frame stride of 1 and a window stride of 16. We use a minimum track length of 8 frames and retain valid tracks shorter than the 64-frame temporal window through our short-track handling. The model is trained for 30 epochs using AdamW with a batch size of 16, a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and gradient clipping with a maximum norm of 1.0.

We use a diffusion process with 1,000 timesteps and a cosine noise schedule. The diffusion reconstruction objective uses MSE to predict the clean trajectory $\mathbf{x}_{1:T}^{(0)}$ from its noisy state, with equal weights for translation and rotation. We additionally apply the candidate uncertainty-calibration objective with a weight of 0.1; no auxiliary ranking loss is used in the final model. For symmetric objects, we align the ground-truth trajectory to a temporally consistent equivalent symmetry branch before constructing the diffusion target.

At inference time, we initialize the trajectory state from standard Gaussian noise and use DDIM sampling (Song et al., 2021) with 10 denoising steps. At each step, the complete trajectory model is re-evaluated, including trajectory-state encoding, candidate association, state-evidence fusion, and temporal reasoning. The resulting clean-trajectory prediction is used to update $\mathbf{x}_{1:T}^{(s)}$ to $\mathbf{x}_{1:T}^{(s-1)}$.

Because temporal windows overlap, a frame may receive predictions from multiple windows. We average these predictions in the 9-D trajectory-state space for each unique frame and then decode the averaged state into the final 6D pose.

A.6 Data Splits and Evaluation Protocol

EgoXtreme provides predefined training, validation, and test splits, which we use throughout all experiments. Temporal windows are constructed independently within each predefined split, so overlapping windows never cross training, validation, or test boundaries. To construct Stage 1 policy supervision, we exhaustively run the frozen GigaPose estimator over the complete transformation bank on the training split. The resulting top-ranked pose outcomes are used to generate the binary pose-recovery labels for policy training. The training split is divided into two scene-level folds for policy cross-fitting. Each policy is trained on one fold and applied to the held-out fold, producing out-of-fold policy predictions for the entire training split. These out-of-fold predictions are used to construct the Stage 1 pose hypotheses for training the trajectory diffusion model. For validation and test inference, we ensemble the two cross-fitted policy models by averaging their transformation logits and select the highest-scoring transformation from the resulting ensemble prediction. Trajectory-model checkpoint selection is performed on the validation split, and the checkpoint with the best validation performance is evaluated on the held-out test split. The test split is never used for policy training, trajectory-model training, or checkpoint selection.

All methods use ground-truth object masks for a controlled comparison, so the evaluation focuses on pose estimation and tracking rather than differences in object segmentation.

For FreePose (Ponimatkin et al., 2025), we use the same trajectory windows as *EgoTrack*, including the same temporal windowing, multi-instance filtering, and short-track handling. Its coarse pose observations are replaced with Clean GigaPose Top-1 predictions. When FreePose fails to refine a window, we retain the corresponding coarse GigaPose poses rather than dropping the window, so that the evaluated frame set remains unchanged. The privileged FreePose variant differs only in its pose observations: ground truth is used to select the best pose among our Stage 1 $K = 5$ candidates at each frame before passing it to FreePose.

For GoTrack (Nguyen et al., 2025a), we use an adapted causal initialization in which each frame t is initialized from the raw GigaPose estimate at frame $t - 1$, rather than recursively propagating the previous refined GoTrack output. We use this refreshed initialization to avoid compounding drift under severe visual degradation and to keep the causal baseline tied to the same frame-wise GigaPose observations used throughout our evaluation.

Our causal variant uses the same architecture as the bidirectional model with a causal temporal attention mask that prevents access to future frames.

B Evaluation Metrics

We evaluate both absolute pose accuracy and temporal trajectory consistency. For absolute pose accuracy, we use the standard symmetry-aware ADD(-S) metric (Hinterstoisser et al., 2012). ADD is used for non-symmetric objects and ADD-S for symmetric objects. We report the area under the ADD(-S) success curve from 0 to $0.3d$, where d is the object diameter, as well as median translation and rotation errors. Rotation errors are evaluated only on non-symmetric objects.

For temporal consistency, we use relative pose error (RPE), following the standard trajectory evaluation protocol (Sturm et al., 2012). Because EgoXtreme is captured from a moving egocentric camera, object poses are first transformed from the camera coordinate system to the world coordinate system:

$$\mathbf{T}_{WO,t} = \mathbf{T}_{WC,t}\mathbf{T}_{CO,t}, \tag{4}$$

where $\mathbf{T}_{WC,t}$ is obtained from the camera trajectory.

For two frames separated by δ, we compare the predicted and ground-truth relative object motions as

$$\mathbf{E}_{t,\delta} = \left[(\mathbf{T}^{*}_{WO,t})^{-1}\mathbf{T}^{*}_{WO,t+\delta}\right]^{-1}\left[(\hat{\mathbf{T}}_{WO,t})^{-1}\hat{\mathbf{T}}_{WO,t+\delta}\right]. \tag{5}$$

Translation and rotation RPE are computed from the translational and rotational components of $\mathbf{E}_{t,\delta}$, respectively, and reported as RMSE over all valid frame pairs. We use $\delta = 10$ frames in the main experiments. Rotational RPE is computed only for non-symmetric objects.

Table 5: **Extended analysis of degradation-robust pose hypothesis generation.** Oracle coverage measures whether at least one generated hypothesis successfully recovers the pose.

				EgoXtreme - Sports			EgoXtreme - Maintenance			EgoXtreme - Emergency		
Candidate generation	# Aug.	# Rank	# Hyp.	@0.1d ↑	@0.2d ↑	@0.3d ↑	@0.1d ↑	@0.2d ↑	@0.3d ↑	@0.1d ↑	@0.2d ↑	@0.3d ↑
Clean Aug-1 Rank-1	1	1	1	1.44	10.57	23.34	0.47	6.77	23.65	0.39	5.55	14.31
Retrieval Aug-1 Rank-1	1	1	1	1.01	9.19	21.40	0.39	5.39	20.67	0.31	4.50	11.68
RANSAC Aug-1 Rank-1	1	1	1	1.33	10.75	23.95	0.50	7.36	25.87	0.41	5.61	14.15
Ours Selected Aug-1 Rank-1	1	1	1	1.37	11.24	24.43	0.57	7.57	26.63	0.52	7.30	16.53
Clean Aug-1 Rank-5	1	5	5	4.55	22.48	43.51	2.59	24.77	54.65	1.69	17.50	30.42
Ours Selected Aug-1 Rank-5	1	5	5	4.39	24.14	44.69	2.59	25.28	58.09	2.24	18.93	31.48
Random Aug-5 Rank-1	5	1	5	3.35	20.13	39.23	1.45	18.58	49.82	1.52	13.86	26.06
Retrieval Aug-5 Rank-1	5	1	5	2.74	18.46	37.02	1.43	16.75	45.91	1.40	12.40	24.25
RANSAC Aug-5 Rank-1	5	1	5	3.09	18.98	37.45	1.56	18.85	49.04	1.25	12.64	24.74
Retrieval Global Aug. Rank-5	20	5	5	2.70	18.95	37.57	1.50	17.06	46.46	1.30	12.09	23.70
RANSAC Global Aug. Rank-5	20	5	5	3.43	20.30	38.97	2.06	21.34	51.56	1.27	13.47	25.77
Ours Selected Aug-5 Rank-1	5	1	5	3.19	20.61	40.24	1.87	21.58	54.50	1.89	15.52	28.07
Random Aug-5 Rank-5	5	5	25	8.43	36.39	60.76	7.09	46.31	79.61	5.56	28.08	42.03
Retrieval Aug-5 Rank-5	5	5	25	8.05	35.90	60.11	6.92	44.11	76.71	5.22	27.33	41.71
RANSAC Aug-5 Rank-5	5	5	25	8.44	36.46	60.50	7.31	45.25	77.99	5.22	27.74	41.96
Ours Selected Aug-5 Rank-5	5	5	25	8.14	37.28	61.38	7.83	49.78	81.91	6.40	29.44	43.85
All Aug. Rank-1	20	1	20	6.07	28.87	51.91	3.96	36.16	72.11	3.54	21.79	35.58
All Aug. Rank-5	20	5	100	12.55	47.54	69.72	15.69	67.27	90.05	10.56	35.55	50.94

Table 6: **Per-object analysis on EgoXtreme.** AUC denotes ADD(-S) AUC up to $0.3d$. RPE-T and RPE-R are RMSE at $\delta = 10$ frames. Rotation-based metrics are not reported for symmetric objects.

		Ours			**Selected Aug. Rank-1**			**Stage 1 Oracle**
Object	Sym.	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑
Ping-pong	–	17.28	0.2016	37.89	2.97	2.7881	122.20	10.29
Tennis	✓	58.41	0.5545	–	22.37	2.7004	–	36.18
Bat	✓	54.42	0.4513	–	25.08	2.6947	–	43.02
Golf	–	64.53	0.1422	19.44	2.41	2.8164	112.69	6.73
Hockey	–	58.47	0.1372	16.76	5.25	4.4833	108.75	12.88
Drill	–	26.76	0.0297	17.86	3.88	0.7866	98.14	12.32
Hammer	–	20.20	0.0387	29.80	4.93	0.9172	98.01	12.41
Wrench	–	24.65	0.0343	23.49	5.16	0.8615	113.73	15.61
Saw	–	37.24	0.0325	15.20	6.97	1.0327	94.98	20.38
Extinguisher	–	40.11	0.0587	11.08	9.26	1.7700	87.90	20.48
Flashlight	–	22.50	0.0488	12.04	8.92	3.0089	73.34	22.40
Kit	–	18.66	0.0641	12.47	3.20	0.8094	110.22	9.44

C Additional Quantitative Analysis

C.1 Extended Analysis of Degradation-Robust Pose Hypothesis Generation

Table 5 extends the analysis of degradation-robust pose hypothesis generation reported in the main paper by including the complete set of evaluated candidate-generation configurations. The *All* Aug. setting evaluates all 20 transformations and the top five pose ranks from each transformation, resulting in 100 hypotheses per frame. Overall, larger candidate pools expand the recoverable pose space, while our Selected Aug-1 Rank-5 configuration provides a compact set of five hypotheses with strong coverage and is used in the final model.

C.2 Per-Object Analysis

Table 6 reports per-object results on EgoXtreme. Our method consistently improves over the frame-wise Selected Aug. Rank-1 prediction across all evaluated objects. Rotation-based metrics are omitted for symmetric objects, for which a unique orientation is not defined. Stage 1 Oracle denotes oracle selection among the default frame-wise $K = 5$ pose hypotheses before trajectory diffusion.

C.3 Analysis under Extreme Lighting Conditions

We further analyze robustness with respect to illumination conditions. We group *normal*, *middle*, and *high* illumination as *standard*, and the remaining challenging lighting conditions as *extreme*. As shown in Table 7, *EgoTrack* maintains comparable trajectory accuracy between standard and

Table 7: **Analysis under standard and extreme lighting conditions.** Normal, middle, and high illumination are grouped as *standard*; all remaining challenging illumination conditions are grouped as *extreme*. AUC denotes ADD(-S) AUC up to $0.3d$, and RPE-T and RPE-R are RMSE at $\delta = 10$ frames.

		Ours			Selected Aug. Rank-1			Stage 1 Oracle
Scenario	Lighting	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑	RPE-T ↓	RPE-R ↓	AUC ↑
Sports	Standard	47.13	0.2744	26.82	9.03	3.5417	115.71	18.06
	Extreme	46.90	0.2646	25.30	8.02	2.5173	111.39	17.12
Maintenance	Standard	28.61	0.0333	22.73	5.28	0.9611	101.72	15.42
	Extreme	26.23	0.0325	19.17	6.97	0.9479	91.65	19.06
Emergency	Standard	23.19	0.0664	14.42	6.58	2.9920	96.30	15.56
	Extreme	23.39	0.0672	14.42	3.60	3.5134	99.14	8.37

extreme illumination across all three scenarios. In contrast, the frame-wise Selected Aug. Rank-1 baseline shows larger condition-dependent variation, particularly in the Emergency scenario. This indicates that candidate-set trajectory reasoning remains effective even when individual observations are severely degraded by illumination.

D Additional Qualitative Results

Figure 6 shows additional visualizations on EgoXtreme, especially in maintenance scenario. Figures 7 and 8 illustrate the effect of the learned NULL token. When severe visual degradation makes the frame-wise pose hypotheses unreliable, the denoiser can assign more attention to the NULL token, reducing the influence of unreliable local candidate evidence and relying more on temporal context to recover the pose.

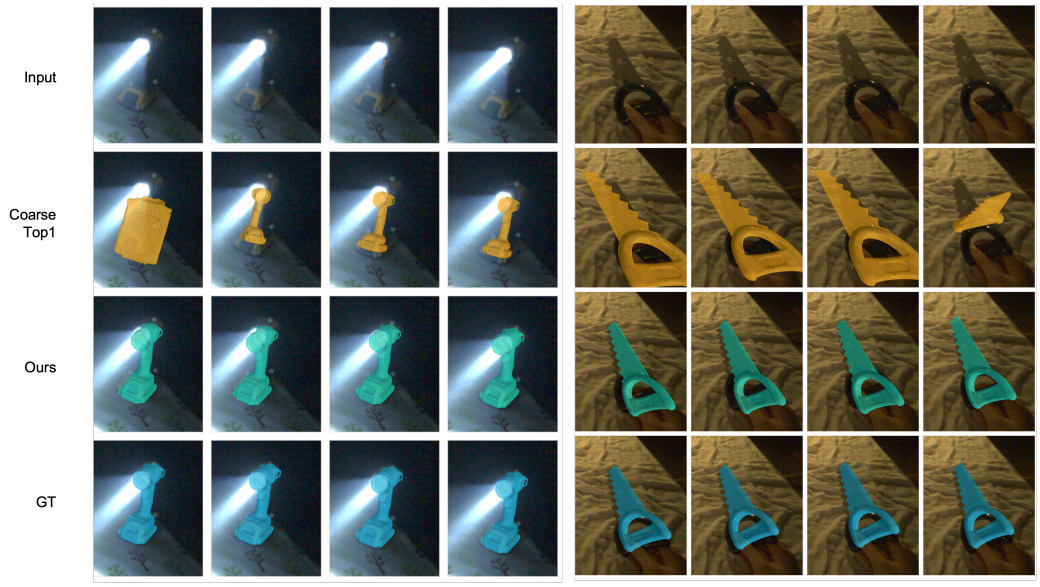

Figure 6: **Additional qualitative results in the Maintenance scenario in EgoXtreme.**

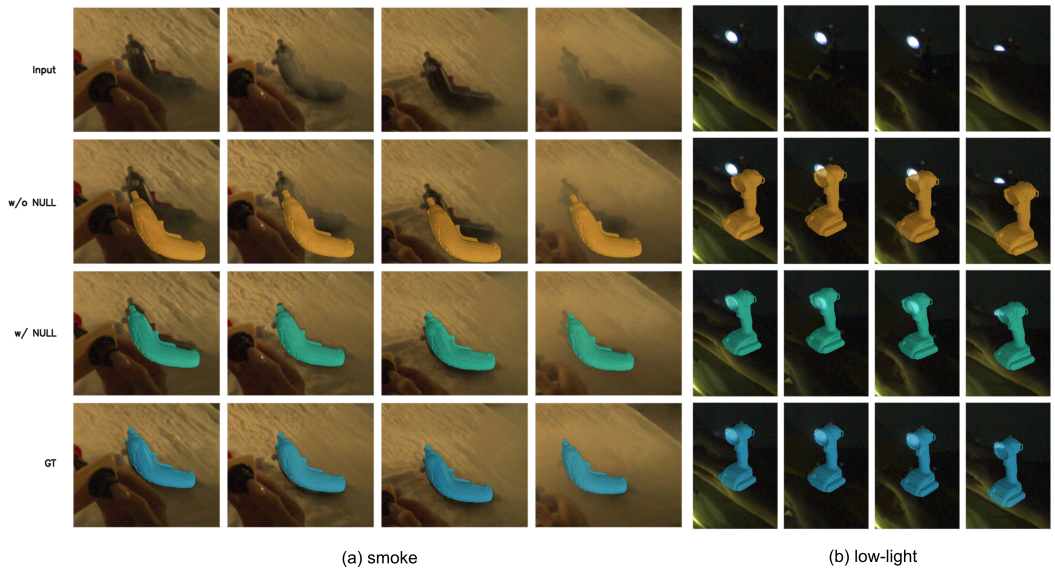

Figure 7: **Qualitative comparison with and without the NULL token in the Maintenance scenario.**



Figure 8: **Qualitative comparison with and without the NULL token in the Emergency scenario.**