AuTO-Pose: Autoregressive Transparent Object 6D Pose Estimation with Ray-Background Context

Anonymous authors
Paper under double-blind review

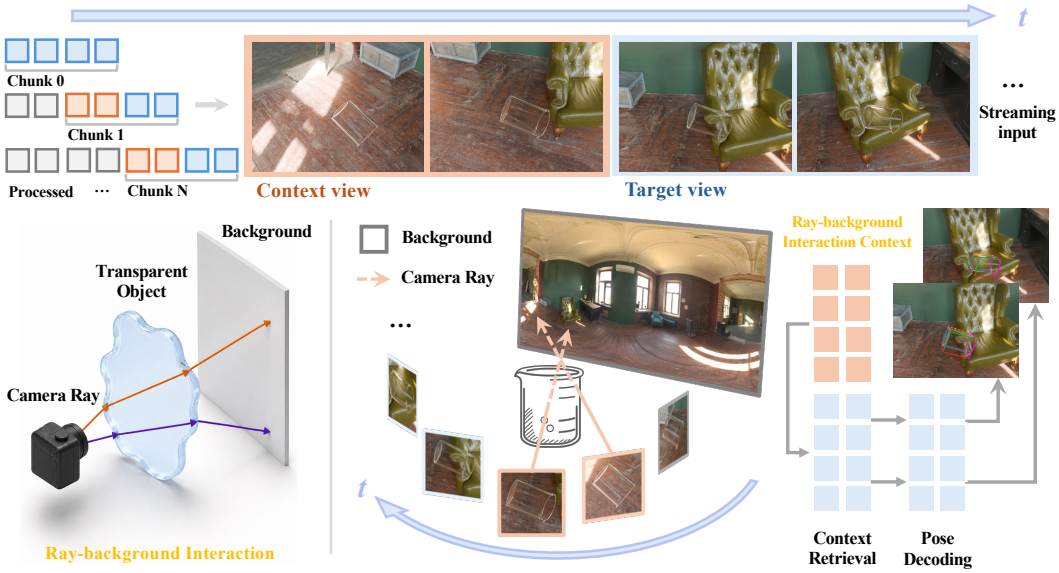

Figure 1: Given a streaming RGB sequence, **AuTO-Pose** processes views in successive chunks, using previously processed views as context to predict the 6D poses of the current target views and then updating the context for the next chunk. The evolving context of ray-background interactions across views reveals geometric cues that are difficult to perceive in any single frame but become apparent through view-dependent changes in the background appearance.

Abstract

We present AuTO-Pose, a novel framework for transparent object 6D pose estimation from RGB videos that causally decodes each target frame window conditioned on previously processed frames. Existing state-of-the-art methods for transparent object pose estimation either adapt opaque object pose estimators or learn cues from isolated transparent observations, largely overlooking temporal context and remaining vulnerable to background-dependent appearance variations. In contrast, AuTO-Pose predicts poses causally from streaming visual observations while accumulating object-centric evidence from previously processed frames. By constructing context from ray-background interaction cues across views, AuTO-Pose exploits how transparent objects refract, reflect, attenuate, and rearrange background evidence over time. This allows the model to resolve ambiguous single-frame cues and generalize better to challenging scenarios with changing backgrounds, weak boundaries, occlusion, and long sequences. We further adopt symmetry-aware supervision to avoid inconsistent training signals for transparent objects with indistinguishable orientations. Extensive experiments on TRansPose, ClearPose, and our self-collected TOS-6D benchmarks show that AuTO-Pose consistently improves transparent object pose accuracy and achieves state-of-the-art performance across benchmark accuracy, sequence generalization, and scenario-level robustness protocols. More details can be found on our project page.

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

1 INTRODUCTION

6D object pose estimation aims to predict an object's 3D rotation and translation relative to the camera and is a fundamental task in 3D computer vision. Accurate object poses provide a geometric bridge between visual observations and downstream applications such as robotic manipulation, AR/VR, and digital-twin inspection. Recent methods have achieved substantial progress on opaque objects through keypoint voting (Peng et al., 2019), dense correspondence regression (Wang et al., 2021; Liu et al., 2025), category-level estimation (Li et al., 2025), and reference-based pose estimation (Liu et al., 2022; Sun et al., 2022; He et al., 2022). These methods recover pose by establishing correspondences between image evidence and object geometry, relying on visual cues that remain stable across viewpoints, such as texture, shading, silhouettes, and local correspondence patterns.

Transparent objects violate this assumption because they often exhibit weak texture and unstable boundaries, while their appearance is strongly influenced by refraction, reflection, attenuation, and the surrounding background. Consequently, identical poses may appear different across backgrounds, whereas distinct poses may produce deceptively similar single-frame evidence. Existing methods address these ambiguities with additional geometric or transparency-specific cues: KeyPose estimates 3D keypoints from multi-view observations (Liu et al., 2020); implicit-rendering methods compare observed images against object renderings (Burde et al., 2024a); and ReFlow6D learns a refraction-guided intermediate representation (Gupta et al., 2024). However, these designs still treat each observation as isolated or compress transparency into a single-frame proxy, leaving the temporal *ray-background interaction context* underexplored. In a continuous sequence, neighboring frames reveal how background evidence seen through the same object varies with viewing rays and poses. Modeling this interaction can disambiguate weak texture and stabilize pose estimates, yet current transparent pose estimators do not explicitly exploit it.

Our key observation is that transparent object pose should be inferred from the evolution of ray-background interactions rather than from independent frame-level appearance. When a transparent object moves relative to the camera or background, each frame records a different set of distorted rays passing through the same object-centric geometry. Although any single frame can be ambiguous, the sequence of interactions forms a pose-sensitive signature: background structures shift coherently through the object, boundaries and highlights change consistently with viewpoint, and dense object coordinates predicted in earlier frames provide anchors for interpreting later observations. This motivates an autoregressive formulation for transparent object 6D pose estimation. Instead of predicting every frame independently, the model processes a streaming RGB sequence in windows, caches object-centric evidence from previously predicted frames, and lets each new target frame query this memory before pose decoding. The resulting representation combines feed-forward visual geometry priors with a persistent memory of ray-background interactions, allowing the pose estimator to recover geometric cues that are only visible across time.

In this paper, we present AuTO-Pose, an autoregressive framework for transparent object 6D pose estimation with ray-background context. Given a streaming RGB sequence, AuTO-Pose extracts frame-level visual tokens with a pretrained visual encoder, exchanges information across context and target views using a lightweight windowed transformer, and decodes dense object coordinates and 6D pose for target frames. The core component is a ray-background interaction context bank: for previously processed frames, the bank stores RoI features together with object-frame coordinate anchors, foreground confidence, image evidence, and temporal embeddings. A target frame retrieves from this bank through cross-attention, enriching its RoI features with geometrically grounded observations collected under earlier rays and backgrounds. This design lets the model use past predictions as structured context while keeping inference causal and bounded over long sequences, and achieves *real-time* throughput. We further handle discrete and continuous object symmetries during supervision, avoiding inconsistent gradients for symmetric transparent objects whose appearance cannot determine a unique rotation.

To sum up, our key contributions are as follows:

- To the best of our knowledge, AuTO-Pose is the first method to introduce an autoregressive paradigm into object 6D pose estimation, enabling causal pose prediction from streaming visual observations rather than independent per-frame inference.
- We introduce ray-background interaction context for transparent object perception. By storing object-anchored interaction records from previous frames and retrieving them during target-

frame decoding, AuTO-Pose directly addresses the ambiguity caused by refraction, reflection, weak texture, and background-dependent appearance.

- Extensive experiments on TRansPose, TOS-6D, and ClearPose show that AuTO-Pose consistently improves transparent object pose accuracy and achieves state-of-the-art performance across benchmark, sequence-generalization, and scenario-level robustness protocols.

2 Related Work

6D Object Pose Estimation 6D object pose estimation encompasses instance-level methods that assume known object instances and leverage instance-specific appearance or geometry for pose estimation (Xiang et al., 2018; Peng et al., 2019; Park et al., 2019; Zakharov et al., 2019; Li et al., 2019; Hodan et al., 2020; Song et al., 2020; Wang et al., 2021; Su et al., 2022; Liu et al., 2025), as well as category-level approaches that learn canonical or context-aware object priors (Wang et al., 2019; Tian et al., 2020; Chen et al., 2020; 2021; Chen & Dou, 2021; Di et al., 2022; Zhang et al., 2024b; Li et al., 2025; 2026b). More recent generalizable methods remove category or CAD-model assumptions and estimate poses for novel objects from reference images or sparse observations (Liu et al., 2022; Sun et al., 2022; He et al., 2022; Shugurov et al., 2022; Labbé et al., 2022; Nguyen et al., 2024; Wen et al., 2024; Lin et al., 2024; Örnek et al., 2024; Moon et al., 2024). While such methods improve flexibility, instance-level pose estimation can exploit object-specific geometry and learned explicit correspondences, which remains advantageous for accurate pose recovery. Our work follows this instance-level setting for transparent objects, where reliable pose cues are entangled with refractive ray-background interactions.

Feed-Forward 3D Reconstruction Feed-forward 3D reconstruction replaces classical SfM/MVS optimization with learned visual geometry priors. DUSt3R (Wang et al., 2024) formulates unposed two-view reconstruction as pointmap regression, while follow-up models improve matching, multi-view reasoning, and transformer-based geometry prediction (Leroy et al., 2024; Wang et al., 2025a;b; Lu et al., 2025). Recent streaming methods further extend this paradigm to online inputs with spatial memory, recurrent states, causal attention, or compact KVcaches (Wang & Agapito, 2025; Chen et al., 2025; Zhuo et al., 2025; Lan et al., 2025; Taher et al., 2025; Chen et al., 2026). Our method adapts this paradigm to object-centric transparent perception by using feed-forward 3D priors as geometric context and integrating refractive ray-background cues for transparency-aware 6D pose estimation.

Transparent Object Reconstruction Reconstructing transparent objects departs from conventional 3D reconstruction, since the observed appearance is often governed by refraction or reflection rather than surface texture. Existing methods either estimate depth or normals on such surfaces, from video-based normal recovery to recent diffusion- and semantics-guided predictors (Yeung et al., 2014; Costanzino et al., 2023; Bai et al., 2024; Xu et al., 2025; Li et al., 2026a), explicitly model refractive view synthesis and neural surfaces (Bemana et al., 2022; Chen et al., 2023; Li et al., 2023; Wang et al., 2023; Deng et al., 2024a;b), or relax the physics with opacity-aware implicit surfaces for thin objects (Wu et al., 2025; Zhang et al., 2024a). Complementary transparent perception work further shows that boundaries and reflections provide strong cues for semantic transparent object segmentation (Vu et al., 2026). Despite this progress, these approaches often rely on controlled capture, masks, known illumination or environment maps, or per-scene optimization. In contrast, we do not aim to reconstruct a full refractive surface; we instead exploit ray-background interactions as pose-sensitive cues for transparency-aware 6D pose estimation.

Transparent Object Pose Estimation For 6D pose estimation, transparency shifts the problem from recovering object geometry to localizing instances whose visual evidence is entangled with the surrounding background. KeyPose estimates poses from multi-view keypoints and introduces the TOD benchmark (Liu et al., 2020), while later methods use geometric fusion, implicit representations, or refraction-guided cues to improve sim-to-real transfer and pose accuracy (Yu et al., 2023; Burde et al., 2024b; Gupta et al., 2024). Datasets such as ClearPose (Chen et al., 2022), TRansPose (Kim et al., 2024), and Trans6D-32K (Yu et al., 2023) further broaden training and evaluation with larger annotations, multispectral sensing, and synthetic-to-real data. Together, these works define transparent pose estimation beyond object settings, but still largely depend on static geometric cues or specialized sensing, leaving ray-object-background interactions underexplored. In contrast, our work incorporates this interaction with streaming feed-forward 3D context for robust and efficient transparent 6D pose estimation.

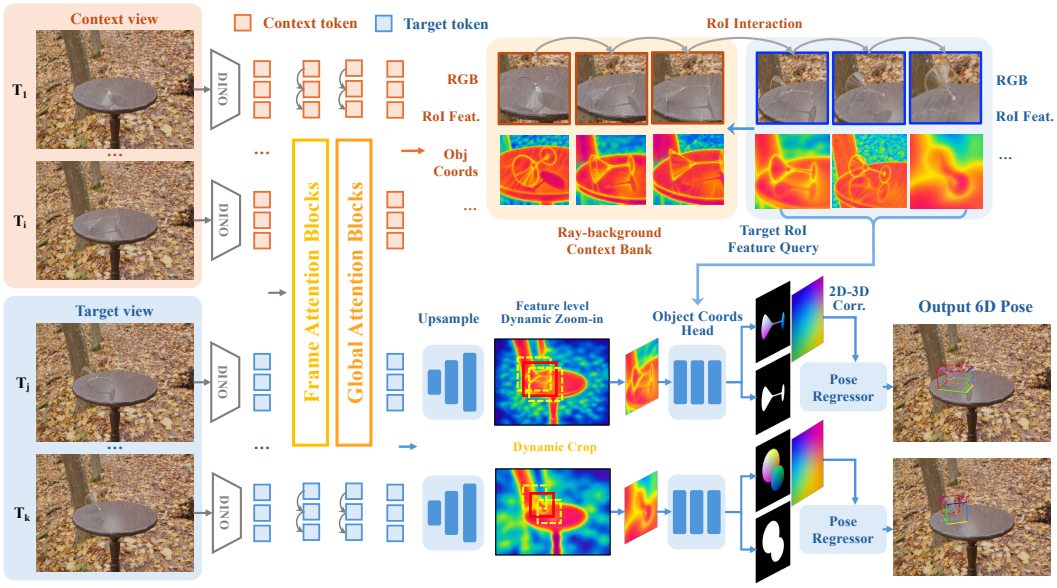

Figure 2: **Overview of AuTO-Pose.** Given a streaming RGB sequence, a DINO encoder extracts frame-level visual tokens, and a windowed autoregressive decoder exchanges information between cached context views and target views. The ray-background interaction context then provides geometrically grounded RoI cues before dense coordinate and 6D pose decoding.

3 METHOD

3.1 OVERVIEW

Given an RGB sequence, our goal is to estimate the 6D pose $\mathbf{T}_t = [\mathbf{R}_t|\mathbf{t}_t] \in SE(3)$ of a transparent object at each frame t. Unlike opaque objects, transparent objects provide weak texture and unstable boundary cues. Their most informative signal comes from *ray-background interaction*: camera rays pass through the object, are distorted by its geometry and pose, and finally project background-dependent patterns. A single frame captures only one such interaction, making pose estimation ambiguous under clutter, weak edges, or changing backgrounds.

To address this, we propose an autoregressive pose estimation framework that accumulates these interactions across nearby frames. The method builds a context bank from historical RoI features that encode past ray-background interactions, and lets each target view query this bank through cross-view attention. The fused target representation therefore combines current-frame evidence with geometric cues observed under previous rays and backgrounds, enabling more stable and accurate transparent object 6D pose tracking. An overview of the framework is shown in Fig. 2.

3.2 AUTOREGRESSIVE OBJECT POSE ESTIMATION

Given a streaming RGB sequence $\{\mathbf{I}_t\}_{t=1}^{N}$, AuTO-Pose employs a windowed autoregressive architecture for object pose estimation. At step k, $\mathcal{C}_k$ denotes the temporally ordered context views decoded at preceding steps, $\mathcal{T}_k$ denotes the target views to be decoded at the current step, and $\mathcal{W}_k = \mathcal{C}_k \cup \mathcal{T}_k$ is the corresponding input window. Each context view $c \in \mathcal{C}_k$ is associated with a record $\hat{\mathbf{Y}}_c$ constructed after that view has been decoded, yielding $\mathcal{Y}_k = \{\hat{\mathbf{Y}}_c\}_{c\in\mathcal{C}_k}$. We denote the predicted 6D pose and dense object-coordinate map for each target view t by $\hat{\mathbf{T}}_t$ and $\hat{\mathbf{X}}_t$, respectively, and let f_θ denote the complete model. The autoregressive update is

$$\left\{\hat{\mathbf{T}}_t, \hat{\mathbf{X}}_t\right\}_{t\in\mathcal{T}_k} = f_\theta(\{\mathbf{I}_v\}_{v\in\mathcal{W}_k}, \mathcal{Y}_k), \quad \mathcal{C}_{k+1} = \mathrm{Last}_K(\mathcal{T}_k), \quad \mathcal{Y}_{k+1} = \left\{\hat{\mathbf{Y}}_t\right\}_{t\in\mathcal{C}_{k+1}}, \tag{1}$$

where $\mathrm{Last}_K(\cdot)$ selects the K temporally latest target views. Thus, the predictions produced at step k provide the contextual evidence used at step $k+1$. The construction and retrieval of these context records are detailed in Sec. 3.3.

The initial window serves as a bootstrap stage with all of its views treated as targets, since no decoded context is yet available,

$$\mathcal{C}_0 = \varnothing, \qquad \mathcal{T}_0 = \mathcal{W}_0, \qquad \mathcal{Y}_0 = \varnothing. \tag{2}$$

Within each window, the context and target views are encoded jointly, whereas information propagates across windows only through records produced by completed autoregressive steps. The resulting formulation is therefore bidirectional within each window and causal across windows. Rather than accumulating context over the entire sequence, each step reconstructs a bounded context from the most recent target views. This design keeps the inference cost independent of the elapsed sequence length and limits the influence of an erroneous context record to subsequent nearby windows. At the bootstrap step, the context pathway contributes no historical evidence; the same all-target configuration is sampled during training to align the training and inference regimes.

To instantiate the model f_θ, a DINOv3 encoder (Siméoni et al., 2025) extracts patch-level visual tokens $\mathbf{F}_v = E_{\mathrm{DINO}}(\mathbf{I}_v)$ for every $v \in \mathcal{W}_k$. The encoder tokens are projected to form the initial decoder representations $\mathbf{G}_v^0$. A lightweight transformer decoder then alternates within-view self-attention and cross-view attention,

$$\{\mathbf{G}_v^l\}_{v\in\mathcal{W}_k} = D^l\Big(\{\mathbf{G}_v^{l-1}\}_{v\in\mathcal{W}_k}\Big), \tag{3}$$

where the context views condition the joint representation and only the target views produce new predictions. The decoded target tokens $\mathbf{G}_t^L$ are upsampled to the input resolution, cropped around the object, and resized to a fixed RoI resolution using feature-level Dynamic Zoom-In (Li et al., 2019; Wang et al., 2021), yielding the target RoI features $\mathbf{H}_t$. These features then query the ray-background interaction context described in Sec. 3.3, and the context-enriched features are passed to the dense coordinate and pose heads.

3.3 Ray-Background Interaction Context

The cross-frame decoder already exchanges token-level information, but leaves the pose head without an explicit, geometrically addressable representation of recent object observations. We therefore maintain a short-horizon ray-background interaction context, consisting of object anchored records gathered from the context views of each autoregressive window. Before pose decoding, the current target frame queries these records to obtain context-enriched RoI target features.

For each context view $c \in \mathcal{C}_k$, we take its RoI feature map $\mathbf{H}_c \in \mathbb{R}^{G\times G\times d}$ and pair every spatial location $\mathbf{u}$ of the $G \times G$ grid with an object frame anchor read from the dense coordinate prediction of that view,

$$\mathbf{x}_{c,\mathbf{u}} = \hat{\mathbf{X}}_c(\mathbf{u}), \qquad \mathbf{u} = \pi\big(\mathbf{K}_c(\hat{\mathbf{R}}_c\mathbf{x}_{c,\mathbf{u}} + \hat{\mathbf{t}}_c)\big). \tag{4}$$

Since context views are decoded in earlier windows, these anchors are directly available from existing predictions and are expressed in the object frame defined by the cached pose $\hat{\mathbf{T}}_c$. Each entry thus records one ray-background interaction: the appearance observed through the object, bound to the surface point that produced it. Context records and target locations are then embedded into context and query tokens, respectively,

$$\mathbf{e}_{c,\mathbf{u}} = \phi_{\mathrm{ctx}}\big([\mathbf{H}_{c,\mathbf{u}}, \mathbf{x}_{c,\mathbf{u}}, \mathbf{r}_{c,\mathbf{u}}, \mathbf{p}_{c,\mathbf{u}}, m_{c,\mathbf{u}}, \tau_c]\big), \;\; \mathbf{q}_{t,\mathbf{u}} = \phi_{\mathrm{tar}}\big([\mathbf{H}_{t,\mathbf{u}}, \mathbf{r}_{t,\mathbf{u}}, \mathbf{p}_{t,\mathbf{u}}, m_{t,\mathbf{u}}, \tau_t]\big), \tag{5}$$

where $\mathbf{r}$ denotes the RoI RGB embedding, $\mathbf{p}$ the normalized 2D coordinate, m the object mask confidence, and τ a learnable temporal embedding. The query deliberately excludes any pose-derived geometry of the target frame, so geometric information flows only from already-processed views.

Each target query $\mathbf{q}_{t,\mathbf{u}}$ cross-attends to the context tokens $\{\mathbf{e}_{c,\mathbf{v}}\}_{c\in\mathcal{C}_k,\mathbf{v}}$, aggregating evidence $\mathbf{z}_{t,\mathbf{u}}$ over all context views and spatial locations. The retrieved evidence is then fused residually into the target RoI features, $\tilde{\mathbf{H}}_t = \mathbf{H}_t + \psi(\mathbf{z}_t)$, before being passed to the dense coordinate and pose heads. This short-horizon context is refreshed as the window advances and contains only the most recently decoded views. To match this regime during training, context anchors are built from ground-truth poses perturbed with small rotation and translation noise, exposing the model to the imperfect anchors it will encounter at test time. By injecting recent ray-background interactions at the feature level before pose decoding, the context provides the geometric evidence needed to resolve ambiguous single-frame observations of transparent objects.

3.4 SYMMETRY HANDLING

Symmetric objects such as cups, tubes, and bottles induce identical ray-background appearance under multiple valid orientations. We therefore regard an annotated rotation $\mathbf{R}$ as equivalent to $\mathbf{RS}$ for any $\mathbf{S}$ in the object-frame symmetry group $\mathcal{S}_o \subset SO(3)$. For *discrete* symmetries, $\mathcal{S}_o$ is finite and we follow the standard prediction anchored closest pose supervision (Wang et al., 2021; Liu et al., 2025).

For an object with *continuous axial symmetry*, rotating it by any angle ϕ about its unit object-frame symmetry axis $\mathbf{a}$ leaves the observed appearance unchanged. These rotations form the symmetry group $\mathcal{S}_o = \{\exp(\phi[\mathbf{a}]_\times) \mid \phi \in [0, 2\pi)\}$, making the spin angle ϕ intrinsically unobservable. Rather than choosing an arbitrary representative, we make every supervised target invariant to ϕ by retaining only quantities defined on the quotient $SO(3)/\mathcal{S}_o$. **(i) Rotation.** We supervise only the camera-frame axis alignment, $\angle(\hat{\mathbf{R}}\mathbf{a}, \mathbf{R}\mathbf{a})$, leaving the spin unconstrained. **(ii) Dense coordinates.** We discard the azimuth by representing each surface point $\mathbf{X}$ with its axial height $h = \mathbf{a}^\top(\mathbf{X} - \mathbf{o})$ and radial distance $\rho = \|(\mathbf{I} - \mathbf{a}\mathbf{a}^\top)(\mathbf{X} - \mathbf{o})\|$, where $\mathbf{o}$ lies on the symmetry axis. Region targets are likewise binned over (h, ρ). **(iii) Point matching.** We use ADD-S matching on the object's model points, making $\mathcal{L}_{\text{pm}}$ invariant to ϕ. Together, these ϕ-invariant targets ensure consistent supervision across ambiguous spin annotations without penalizing equivalent spin variations as inter-frame motion.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

The input images follow the size configuration of GDR-Net (Wang et al., 2021). Images are resized to 640×480 before encoding. To ensure that each image can be evenly patchified, we adopt the lightweight ViT-S/16 variant of DINOv3 (Siméoni et al., 2025) with a patch size of 16 as the encoder. Both the cross-view decoder and the RoI decoder are implemented as compact two-layer attention modules with a hidden dimension of 256. More details are provided in the supplementary material.

4.2 BENCHMARKS AND PROTOCOLS

We evaluate AuTO-Pose on three complementary protocols covering TRansPose benchmark accuracy, TOS-6D sequence generalization, and ClearPose scenario-level robustness.

Datasets and metrics. We evaluate on three transparent object benchmarks. On **TRansPose** (Kim et al., 2024), we report standard transparent object 6D pose metrics, including AR_{ADD}, the BOP average-recall metrics AR_{MSSD}, AR_{MSPD}, and AR_{BOP}, as well as thresholded 3D bounding box IoU. On our self-collected **TOS-6D** (Transparent Object Sequence 6D Pose) dataset, we use two test sets, *New BG* and *Long Seq*: the former changes the background distribution while preserving object categories, and the latter measures autoregressive stability over long rollouts. We report BOP-style metrics, 3D IoU, and pose-threshold recalls at $5°$2cm, $5°$5cm, and $10°$2cm. Finally, on **ClearPose** (Chen et al., 2022) we assess robustness under challenging scenarios such as novel backgrounds, heavy occlusion, translucent covers, opaque distractors, filled liquid, and non-planar supporting surfaces, reporting Accuracy and $\text{ADD}_{(-\text{S})}$ for each. More details on the dataset configuration and metric definitions are provided in the supplementary material.

Baselines. We use benchmark-specific baselines for each protocol. On TRansPose and TOS-6D, we compare with PVNet (Peng et al., 2019), KeyPose (Liu et al., 2020), GDR-Net (Wang et al., 2021), Reflow6D (Gupta et al., 2024), and GDRNPP (Liu et al., 2025), covering keypoint-based, dense-coordinate, transparent object, and recent high-performance 6D pose estimators. On ClearPose, we compare with the transparent object baselines reported by the benchmark, including Xu et al. (Xu et al., 2020) and FFB6D variants (He et al., 2021), as well as MegaPose (Labbé et al., 2022), IRPose (Burde et al., 2024a), and its finetuned variant IRPose_f.

4.3 COMPARISON WITH THE STATE OF THE ART

TRansPose benchmark. Table 1 reports the results of AuTO-Pose on the TRansPose dataset, which achieves the best performance across all reported metrics. Compared with the strongest baseline, GDRNPP (Liu et al., 2025), AuTO-Pose achieves consistent gains across pose and 3D-overlap metrics,

Table 1: **Quantitative comparison on TRansPose.** All results are reported in percentages, with higher values indicating better performance. Best results are in **bold**.

Method	$\mathrm{AR}_{\mathrm{ADD}}$	$\mathrm{AR}_{\mathrm{MSSD}}$	$\mathrm{AR}_{\mathrm{MSPD}}$	$\mathrm{AR}_{\mathrm{BOP}}$	$\mathrm{IoU}_{@25}$	$\mathrm{IoU}_{@50}$	$\mathrm{IoU}_{@75}$
PVNet (Peng et al., 2019)	3.50	1.17	2.71	1.94	3.63	1.67	0.38
KeyPose (Liu et al., 2020)	21.94	1.51	4.20	2.85	18.40	4.32	0.26
GDR-Net (Wang et al., 2021)	58.41	19.11	30.36	24.74	58.95	27.37	6.28
Reflow6D (Gupta et al., 2024)	56.17	8.15	9.30	8.72	56.85	20.28	2.17
GDRNPP (Liu et al., 2025)	68.01	25.05	34.15	29.60	70.93	38.55	9.59
AuTO-Pose (**Ours**)	**73.34**	**31.71**	**37.37**	**34.54**	**76.57**	**45.03**	**12.77**

Table 2: **Quantitative comparison on TOS-6D.** *New BG* uses test set with unseen backgrounds to evaluate background generalization, while *Long Seq* uses long videos to evaluate long sequence inference accuracy and stability. Best results are in **bold**.

	New BG subset				**Long Seq subset**			
Method	$5°2\mathrm{cm}$	$5°5\mathrm{cm}$	$10°2\mathrm{cm}$	$\mathrm{IoU}_{@75}$	$5°2\mathrm{cm}$	$5°5\mathrm{cm}$	$10°2\mathrm{cm}$	$\mathrm{IoU}_{@75}$
PVNet (Peng et al., 2019)	2.23	4.91	4.88	8.99	1.36	3.30	3.35	6.44
KeyPose (Liu et al., 2020)	3.12	7.60	6.51	9.89	2.02	4.65	4.22	6.12
GDR-Net (Wang et al., 2021)	75.51	82.21	82.13	88.67	59.75	66.34	67.81	77.26
Reflow6D (Gupta et al., 2024)	81.85	87.83	87.14	91.33	69.58	76.01	76.89	84.53
GDRNPP (Liu et al., 2025)	82.55	87.98	88.82	92.56	69.16	75.53	77.95	85.74
AuTO-Pose (**Ours**)	**89.29**	**91.27**	**92.79**	**95.06**	**77.65**	**80.98**	**82.59**	**88.17**
Method	$\mathrm{ADD}_{(\text{-S})}$	$\mathrm{AR}_{\mathrm{MSSD}}$	$\mathrm{AR}_{\mathrm{MSPD}}$	$\mathrm{AR}_{\mathrm{BOP}}$	$\mathrm{ADD}_{(\text{-S})}$	$\mathrm{AR}_{\mathrm{MSSD}}$	$\mathrm{AR}_{\mathrm{MSPD}}$	$\mathrm{AR}_{\mathrm{BOP}}$
PVNet (Peng et al., 2019)	5.67	1.76	14.46	8.11	3.61	1.17	10.29	5.73
KeyPose (Liu et al., 2020)	4.25	2.57	25.42	13.99	2.54	1.67	17.78	9.72
GDR-Net (Wang et al., 2021)	75.51	63.24	85.85	74.54	62.45	50.49	72.55	61.52
Reflow6D (Gupta et al., 2024)	78.75	66.81	88.82	77.82	69.58	57.35	80.90	69.12
GDRNPP (Liu et al., 2025)	80.31	68.36	90.01	79.19	68.85	56.42	81.03	68.73
AuTO-Pose (**Ours**)	**87.32**	**80.27**	**92.55**	**86.41**	**78.22**	**70.81**	**84.62**	**77.71**

increasing $\mathrm{AR}_{\mathrm{BOP}}$ from 29.60 to 34.54 and $\mathrm{IoU}_{@50}$ from 38.55 to 45.03. These results indicate that modeling ray-background context provides more accurate geometric alignment for transparent objects than single-frame pose regression. Fig. 4 shows qualitative comparisons on TRansPose, where the 3D bounding boxes recovered by AuTO-Pose closely overlap the ground truth, while competing methods drift under the weak and refraction-distorted appearance of transparent objects.

TOS-6D generalization. Table 2 evaluates generalization on our collected TOS-6D dataset. On the New Background (New BG) split, AuTO-Pose outperforms all baselines under both pose-threshold recalls and BOP-style metrics. It improves over GDRNPP (Liu et al., 2025) across all metrics, most notably increasing $\mathrm{AR}_{\mathrm{MSSD}}$ from 68.36 to 80.27. The larger gain on surface-distance recall suggests that the proposed context modeling is especially helpful when transparent object observations change with the background distribution. The Long Sequence (Long Seq.) subset evaluates autoregressive stability over sequences spanning 120–180 frames. Unlike the baselines, which perform single-frame inference, AuTO-Pose predicts poses autoregressively and therefore must contend with the risk of error accumulation over long rollouts. Despite this harder setting, AuTO-Pose surpasses all baselines on every metric and achieves an $\mathrm{AR}_{\mathrm{MSSD}}$ of 70.81, compared with GDRNPP's 56.42. Moreover, its accuracy degrades only modestly relative to the New BG split (e.g., 11.64% on $5°2\mathrm{cm}$), indicating that the proposed context modeling effectively suppresses error accumulation and maintains temporal consistency under long autoregressive rollouts. Fig. 3 provides qualitative comparisons on TOS-6D across distinct background scenes, where the 3D bounding boxes projected from AuTO-Pose's predicted poses align closely with the transparent objects, whereas Reflow6D (Gupta et al., 2024) and GDRNPP exhibit noticeable rotational and translational misalignment.

ClearPose scenarios. Table 3 reports robustness across challenging transparent object scenarios in ClearPose. AuTO-Pose achieves the highest Accuracy on New BG, Heavy Occ., Opaque Dist.,

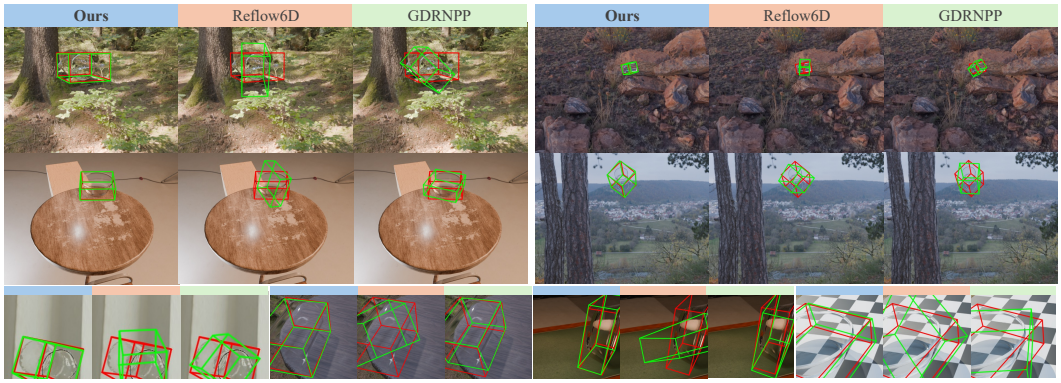

Figure 3: **Qualitative comparison on TOS-6D.** Predicted (green) and ground-truth (red) 3D bounding boxes across distinct background scenes. Our boxes align tightly with the ground truth, while Reflow6D (Gupta et al., 2024) and GDRNPP (Liu et al., 2025) drift in rotation and translation.

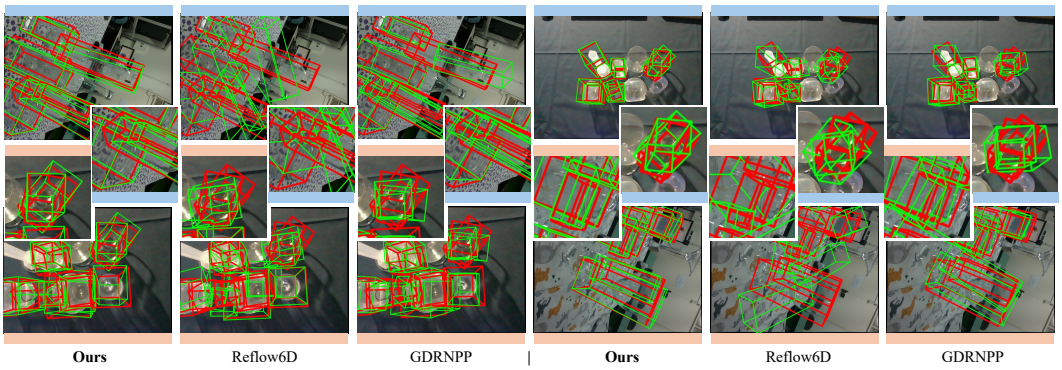

Figure 4: **Qualitative comparison on TRansPose.** Predicted (green) and ground-truth (red) 3D bounding boxes on cluttered real-world scenes, with insets zooming into representative objects. Our poses stay aligned with the ground truth, while Reflow6D (Gupta et al., 2024) and GDRNPP (Liu et al., 2025) show clear box and orientation errors.

and Non Planar, and obtains the best $\text{ADD}_{(\text{-S})}$ on New BG, Heavy Occ., and Non Planar. The improvements are most pronounced under novel backgrounds and heavy occlusion, where cross-frame ray-background evidence provides complementary cues to local RGB appearance. Performance under translucent covers and filled liquids remains lower than specialized baselines, suggesting that strong refraction and altered object contents remain challenging.

4.4 Ablation Studies

We ablate the key design choices of AuTO-Pose on TOS-6D, and report results in Table 4.

Autoregressive modeling. The *w/o autoregressive* variant disables the autoregressive training and inference scheme: instead of a context/target split, all 6 frames in a window are supervised and decoded as target views, so no view conditions on previously predicted frames. Compared with the full model, this variant drops by 2.21% on $\text{ADD}_{(\text{-S})}$ (87.32→85.11), 5.59% on AR_{MSSD} (80.27→74.68), and 3.06% on AR_{BOP} (86.41→83.35). The larger degradation on surface-distance metrics indicates that propagating ray-background evidence from previously predicted frames yields more accurate and temporally consistent pose estimates than treating every frame independently.

Symmetry handling. The *w/o symm. handling* variant removes the continuous and discrete symmetry treatment introduced in Sec. 3.4 and supervises the network with the raw annotated pose. This causes by far the largest drop, with $\text{ADD}_{(\text{-S})}$ falling from 87.32 to 58.74, AR_{BOP} from 86.41 to 53.93, and $5°2\text{cm}$ from 89.29 to 58.95. Without canonicalizing symmetric orientations, visually equivalent pose branches produce contradictory gradients and destabilize training, confirming that explicit symmetry handling is essential for transparent objects whose appearance cannot determine a unique orientation.

Table 3: **Results on ClearPose across six challenging scenarios.** We report Accuracy and $\text{ADD}_{(\text{-S})}$ using the downsampled evaluation following IRPose (Burde et al., 2024a). Results of other methods are taken from their respective papers. Best values are in **bold**.

Testset	Metric	Xu et al.	$\text{FFB}_{r/r}$	$\text{FFB}_{g/c}$	$\text{FFB}_{g/g}$	MegaPose	IRPose	IRPose_f	**Ours**
New BG	Accuracy	50.96	44.26	49.52	59.69	36.92	44.33	43.52	**73.50**
	$\text{ADD}_{(\text{-S})}$	45.23	43.45	47.69	58.22	41.80	45.99	46.24	**65.50**
Heavy Occ.	Accuracy	24.19	14.72	15.16	26.33	37.10	41.45	43.18	**54.27**
	$\text{ADD}_{(\text{-S})}$	22.95	17.87	17.86	31.88	45.48	48.55	49.37	**54.18**
Transluc. Cov.	Accuracy	14.35	5.56	4.53	13.43	35.48	36.34	**40.65**	33.12
	$\text{ADD}_{(\text{-S})}$	14.31	7.60	5.81	17.62	47.88	48.14	**52.04**	40.90
Opaque Dist.	Accuracy	42.63	0.46	1.33	2.35	37.93	43.52	45.16	**46.93**
	$\text{ADD}_{(\text{-S})}$	39.04	0.76	1.55	3.07	46.01	47.85	**48.82**	47.38
Filled Liquid	Accuracy	34.50	7.69	9.06	16.23	44.78	51.21	**51.96**	45.07
	$\text{ADD}_{(\text{-S})}$	32.25	11.15	10.83	18.58	53.73	57.96	**59.56**	47.00
Non Planar	Accuracy	21.02	7.49	7.58	15.57	35.24	42.04	43.31	**52.87**
	$\text{ADD}_{(\text{-S})}$	18.41	7.80	6.73	16.99	41.20	46.12	44.30	**47.49**

Table 4: **Ablation study on the TOS-6D New BG split.**

Method	$\text{ADD}_{(\text{-S})}$	AR_{MSSD}	AR_{MSPD}	AR_{BOP}	5°2cm	5°5cm	10°2cm	$\text{IoU}_{@75}$
Full Model	87.32	80.27	92.55	86.41	89.29	91.27	92.79	95.06
w/o autoregressive	85.11	74.68	92.02	83.35	87.30	90.86	90.87	93.71
w/o symm. handling	58.74	42.96	74.47	53.93	58.95	67.00	66.48	78.46
w/o context	82.58	74.46	92.82	83.64	87.62	92.10	90.48	93.16
w/o context (single)	80.54	69.67	92.90	81.29	87.04	92.82	90.01	93.27

Cross-view context. The *w/o context* variant removes the context provided across views: the window still contains 3 frames, but they are all treated as target views without any context bank. The frames interact only through the cross-view attention in the window decoder. This reduces $\text{ADD}_{(\text{-S})}$ from 87.32 to 82.58 and AR_{BOP} from 86.41 to 83.64, showing that the context bank improves surface-level alignment by incorporating observations from nearby viewpoints. Pushing this further, the *w/o context (single)* variant keeps only a single frame, removing both the context bank and any cross-frame information exchange, which further isolates the benefit of multi-view modeling.

5 Conclusion and Limitations

Limitations. Although AuTO-Pose achieves strong performance on transparent object pose estimation benchmarks, several limitations remain. Scenarios involving translucent covers or liquid-filled objects can substantially alter the observed ray-background interactions, making the visual evidence less consistent with the object geometry modeled by the context bank. Moreover, our current formulation follows the instance-level setting and assumes known CAD models and object identities. Future work should relax these assumptions to handle novel objects and weaker geometric priors.

Conclusion. We have presented AuTO-Pose, an autoregressive framework for transparent object 6D pose estimation from RGB videos. It stores historical ray-background interactions in an object-centric context bank, allowing each target frame to retrieve geometrically grounded evidence before dense coordinate and pose decoding. With symmetry-aware supervision, this context helps resolve ambiguous and background-dependent single-frame cues and stabilizes learning and inference when appearance does not determine a unique orientation. Experiments on TRansPose, TOS-6D, and ClearPose demonstrate consistent improvements over strong baselines, especially under novel backgrounds, heavy occlusion, and long video sequences. These results suggest that temporally accumulated ray-background context provides an effective cue for robust transparent object pose estimation.

AI USE STATEMENT

The authors independently conducted the research and wrote the complete initial manuscript, including its overall structure and scientific content. Generative AI tools were subsequently used to polish the writing for clarity, concision, and grammar. They were also used to assist with literature retrieval, discovery, and organization during the related-work survey; the authors independently read and verified all cited sources and determined their relevance to this work. In addition, generative AI tools assisted with the presentation of the anonymous project page.

Generative AI tools were not used to formulate the core research ideas or claims, design the proposed method or experiments, generate data or experimental results, or interpret the results. All AI-assisted outputs were reviewed and revised by the authors, who take full responsibility for the final paper and project artifacts.

REFERENCES

Kaixin Bai, Huajian Zeng, Lei Zhang, Yiwen Liu, Hongli Xu, Zhaopeng Chen, and Jianwei Zhang. Cleardepth: enhanced stereo perception of transparent objects for robotic manipulation. *arXiv preprint arXiv:2409.08926*, 2024.

Mojtaba Bemana, Karol Myszkowski, Jeppe Revall Frisvad, Hans-Peter Seidel, and Tobias Ritschel. Eikonal fields for refractive novel-view synthesis. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–9, 2022.

Varun Burde, Artem Moroz, Vit Zeman, and Pavel Burget. Object pose estimation using implicit representation for transparent objects. In *European Conference on Computer Vision*, pp. 226–247. Springer, 2024a.

Varun Burde, Artem Moroz, Vit Zeman, and Pavel Burget. Object pose estimation using implicit representation for transparent objects. In *European Conference on Computer Vision*, pp. 226–247. Springer, 2024b.

Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11970–11979. IEEE, June 2020. doi: 10.1109/CVPR42600.2020.01199. URL https://doi.org/10.1109/CVPR42600.2020.01199.

Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2753–2762. IEEE, October 2021. doi: 10.1109/ICCV48922.2021.00277. URL https://doi.org/10.1109/ICCV48922.2021.00277.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026.

Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1581–1590. IEEE, June 2021. doi: 10.1109/CVPR46437.2021.00163. URL https://doi.org/10.1109/CVPR46437.2021.00163.

Xiaotong Chen, Huijie Zhang, Zeren Yu, Anthony Opipari, and Odest Chadwicke Jenkins. Clearpose: Large-scale transparent object dataset and benchmark. In *European conference on computer vision*, pp. 381–396. Springer, 2022.

Xiaoxue Chen, Junchen Liu, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Nerrf: 3d reconstruction and view synthesis for transparent and specular objects with neural refractive-reflective fields. *arXiv preprint arXiv:2309.13039*, 2023.

Zhuoguang Chen, Minghui Qin, Tianyuan Yuan, Zhe Liu, and Hang Zhao. Long3r: Long sequence streaming 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5273–5284, 2025.

Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, and Luigi Di Stefano. Learning depth estimation for transparent and mirror surfaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9244–9255, 2023.

Weijian Deng, Dylan Campbell, Chunyi Sun, Shubham Kanitkar, Matthew Shaffer, and Stephen Gould. Ray deformation networks for novel view synthesis of refractive objects. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3118–3128, 2024a.

Weijian Deng, Dylan Campbell, Chunyi Sun, Shubham Kanitkar, Matthew E Shaffer, and Stephen Gould. Differentiable neural surface refinement for modeling transparent objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20268–20277, 2024b.

Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6771–6781. IEEE, June 2022. doi: 10.1109/CVPR52688.2022.00666. URL https://doi.org/10.1109/CVPR52688.2022.00666.

Hrishikesh Gupta, Stefan Thalhammer, Jean-Baptiste Weibel, Alexander Haberl, and Markus Vincze. Reflow6d: Refraction-guided transparent object 6d pose estimation via intermediate representation learning. *IEEE Robotics and Automation Letters*, 9(11):9438–9445, 2024.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *Advances in Neural Information Processing Systems*, 2022.

Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3003–3013, 2021.

Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Jeongyun Kim, Myung-Hwan Jeon, Sangwoo Jung, Wooseong Yang, Minwoo Jung, Jaeho Shin, and Ayoung Kim. Transpose: Large-scale multispectral dataset for transparent object. *The International Journal of Robotics Research*, 43(6):731–738, 2024. doi: 10.1177/02783649231213117. URL https://doi.org/10.1177/02783649231213117.

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. *arXiv preprint arXiv:2212.06870*, 2022.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pp. 71–91. Springer, 2024.

Mingwei Li, Hehe Fan, and Yi Yang. Transnormal: Dense visual semantics for diffusion-based transparent object normal estimation, 2026a. URL https://arxiv.org/abs/2602.00839.

Weihang Li, Hongli XU, Junwen Huang, Hyunjun Jung, Peter KT Yu, Nassir Navab, and Benjamin Busam. Gce-pose: Global context enhancement for category-level object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 27154–27165, 2025.

Weihang Li, Lorenzo Garattoni, Fabien Despinoy, Nassir Navab, and Benjamin Busam. Object pose transformer: Unifying unseen object pose estimation. *arXiv preprint arXiv:2603.23370*, 2026b.

Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zongcheng Li, Xiaoxiao Long, Yusen Wang, Tuo Cao, Wenping Wang, Fei Luo, and Chunxia Xiao. Neto:neural reconstruction of transparent objects with self-occlusion aware refraction-tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18547–18557, October 2023.

Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27906–27916. IEEE, June 2024. doi: 10.1109/CVPR52733.2024.02636. URL https://doi.org/10.1109/CVPR52733.2024.02636.

Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11602–11610, 2020.

Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Gu Wang, Jiwen Tang, Zhigang Li, and Xiangyang Ji. Gdrnpp: A geometry-guided and fully learning-based object pose estimator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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 *European Conference on Computer Vision*, pp. 298–315. Springer, 2022.

Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation for dynamic videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22820–22830, 2025. doi: 10.1109/CVPR52734.2025.02125.

Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. GenFlow: Generalizable recurrent flow for 6D pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10039–10049. IEEE, June 2024. doi: 10.1109/CVPR52733.2024.00957. URL https://doi.org/10.1109/CVPR52733.2024.00957.

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 (CVPR)*, pp. 9903–9913. IEEE, June 2024. doi: 10.1109/CVPR52733.2024.00945. URL https://doi.org/10.1109/CVPR52733.2024.00945.

Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. FoundPose: Unseen object pose estimation with foundation features. In *Computer Vision – ECCV 2024*, pp. 163–182. Springer Nature Switzerland, October 2024. doi: 10.1007/978-3-031-73347-5_10. URL https://doi.org/10.1007/978-3-031-73347-5_10.

Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7667–7676. IEEE, October 2019. doi: 10.1109/ICCV.2019.00776. URL https://doi.org/10.1109/ICCV.2019.00776.

Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4561–4570, 2019.

Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. OSOP: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6825–6834. IEEE, June 2022. doi: 10.1109/CVPR52688.2022.00671. URL https://doi.org/10.1109/CVPR52688.2022.00671.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 428–437. IEEE, June 2020. doi: 10.1109/CVPR42600.2020.00051. URL https://doi.org/10.1109/CVPR42600.2020.00051.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6728–6738. IEEE, June 2022. doi: 10.1109/CVPR52688.2022.00662. URL https://doi.org/10.1109/CVPR52688.2022.00662.

Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

Marwan Taher, Ignacio Alzugaray, Kirill Mazur, Xin Kong, and Andrew J Davison. Kv-tracker: Real-time pose tracking with transformers. *arXiv preprint arXiv:2512.22581*, 2025.

Meng Tian, Marcelo H. Ang, Jr., and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision – ECCV 2020*, pp. 530–546. Springer International Publishing, 2020. doi: 10.1007/978-3-030-58589-1_32. URL https://doi.org/10.1007/978-3-030-58589-1_32.

Tuan-Anh Vu, Hai Nguyen-Truong, Ziqiang Zheng, Binh-Son Hua, Qing Guo, Ivor W Tsang, and Sai-Kit Yeung. Power of boundary and reflection: Semantic transparent object segmentation using pyramid vision transformer with transparent cues. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3360–3369, 2026.

Dongqing Wang, Tong Zhang, and Sabine Süsstrunk. Nemto: Neural environment matting for novel view and relighting synthesis of transparent objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 317–327, October 2023.

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.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2637–2646. IEEE, June 2019. doi: 10.1109/CVPR.2019.00275. URL https://doi.org/10.1109/CVPR.2019.00275.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10510–10522, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20697–20709, 2024.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, June 2024. doi: 10.1109/CVPR52733.2024.01692. URL https://doi.org/10.1109/CVPR52733.2024.01692.

Tianhao Wu, Hanxue Liang, Fangcheng Zhong, Gernot Riegler, Shimon Vainer, Jiankang Deng, and Cengiz Oztireli. αsurf: Implicit surface reconstruction for semi-transparent and thin objects with decoupled geometry and opacity. In *2025 International Conference on 3D Vision (3DV)*, pp. 44–63. IEEE, 2025.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems XIV*. Robotics: Science and Systems Foundation, June 2018. doi: 10.15607/RSS.2018.XIV.019. URL https://doi.org/10.15607/RSS.2018.XIV.019.

Chi Xu, Jiale Chen, Mengyang Yao, Jun Zhou, Lijun Zhang, and Yi Liu. 6dof pose estimation of transparent object from a single rgb-d image. *Sensors*, 20(23):6790, 2020.

Shaocong Xu, Songlin Wei, Qizhe Wei, Zheng Geng, Hong Li, Licheng Shen, Qianpu Sun, Shu Han, Bin Ma, Bohan Li, et al. Diffusion knows transparency: Repurposing video diffusion for transparent object depth and normal estimation. *arXiv preprint arXiv:2512.23705*, 2025.

Sai-Kit Yeung, Tai-Pang Wu, Chi-Keung Tang, Tony F Chan, and Stanley J Osher. Normal estimation of a transparent object using a video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4):890–897, 2014.

Haixin Yu, Shoujie Li, Houde Liu, Chongkun Xia, Wenbo Ding, and Bin Liang. Tgf-net: Sim2real transparent object 6d pose estimation based on geometric fusion. *IEEE Robotics and Automation Letters*, 8(6):3868–3875, 2023. doi: 10.1109/LRA.2023.3268041.

Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1941–1950. IEEE, October 2019. doi: 10.1109/ICCV.2019.00203. URL https://doi.org/10.1109/ICCV.2019.00203.

Haoran Zhang, Junkai Deng, Xuhui Chen, Fei Hou, Wencheng Wang, Hong Qin, Chen Qian, and Ying He. From transparent to opaque: Rethinking neural implicit surfaces with α-neus. *Advances in Neural Information Processing Systems*, 37:28078–28102, 2024a.

Mengchen Zhang, Tong Wu, Tai Wang, Tengfei Wang, Ziwei Liu, and Dahua Lin. Omni6d: Large-vocabulary 3d object dataset for category-level 6d object pose estimation, 2024b. URL https://arxiv.org/abs/2409.18261.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.