Dynamic-0: Dynamic Neural Networks Applied to a Model-free 6D Detection of Unseen Objects Pipeline

Francesca Maria Greco

Department of Mechanical Engineering, ETH Zurich, Zurich, Switzerland
Email: grecof@ethz.ch

***Abstract*—Dynamic-0 (X0) addresses model-free 6D pose estimation for unseen objects—without Computer-Aided Design (CAD) models—while remaining efficient across varied hardware. It uses a dual-agent design: the first agent achieves coarse pose alignment via DINOv3 visual descriptors, Facebook Artificial Intelligence Similarity Search (FAISS)-based retrieval, and Efficient Perspective-n-Point (EPnP) with Super Random Sample Consensus (SuperRANSAC); the second selectively refines this estimate using differentiable 3D Gaussian Splatting (3DGS) rendering, improving geometric precision only where needed to avoid uniform computational overhead. To adapt inference dynamically, X0 integrates multi-scale routing, sparse feature reactivation, and a novelty-aware semantic gating mechanism based on generative density scoring over DINOv3 and MesaNet embeddings. A MesaNet temporal block further stabilizes pose sequences by treating temporal consistency as a locally optimal test-time training problem, solved via conjugate-gradient updates. Evaluated under the Benchmark for 6D Object Pose Estimation (BOP) Challenge 2025 model-free protocol on BOP-H3 datasets (HOPEv2, HANDAL, HOT3D), Dynamic-0 achieves an Average Precision of 0.473 on HOPEv2, with adaptive throughput of 2.27–15.4 Frames Per Second (FPS) and a 73% reduction in temporal jitter (average processing time: 40.98 s/image). Additional results are reported on HANDAL and HOT3D leaderboards, scoring as the first one in terms of accurate precision and average time on the leaderboards of all three datasets of the BOP Challenge 2025. Code is publicly released at the GitHub repository; the official leaderboards are available at the links indicated in the Data Availability section of this paper. These findings suggest that combining self-supervised representations, adaptive computation, generative novelty detection, and temporal optimization offers a practical route toward scalable, hardware-aware 6D pose estimation in real-world, model-free settings.**

***Keywords*—6D pose estimation, model-free detection, zero-shot learning, adaptive inference, MesaNet, high-performance computing**

I. Introduction

6D object pose estimation—the problem of recovering the full three-dimensional position and orientation of a rigid object from one or more visual observations—constitutes a foundational challenge in computer vision with direct relevance to a broad range of downstream applications, including robotic grasping and manipulation planning, autonomous navigation, scene understanding, and spatially stable overlay in augmented and virtual reality systems. Formally, the task demands the simultaneous estimation of a rotation matrix $R \in SO(3)$ and a translation vector $t \in \mathbb{R}^3$ that together describe the transformation from the object coordinate frame to the camera coordinate frame. This inference must frequently be performed under adverse conditions—partial occlusion, photometric clutter, dramatic illumination variation, significant viewpoint change, and sensor noise—making robust pose recovery one of the most demanding perception problems in practice.

Classical approaches to this problem relied predominantly on hand-crafted local feature descriptors, point pair feature voting schemes, and template matching strategies. While competitive on constrained benchmarks, such methods exhibited fundamental brittleness under the variability characteristic of real deployment scenarios. The subsequent emergence of deep learning brought a paradigmatic shift as established in [1]. Over the following years, the field advanced rapidly through the adoption of large-scale visual backbones, self-supervised representation learning, and architecturally integrated geometric refinement strategies [2–4], progressively displacing hand-crafted pipelines in favor of end-to-end learned representations with substantially greater resilience to real-world variability.

Rather than delegating geometric reasoning to a separate, non-differentiable solver, GDRNPP allows gradient signals to propagate through the entire pipeline during training, integrating pose refinement directly into the learning loop. The practical consequence of this design was decisive: GDRNPP dominated the BOP Challenge leaderboard for two consecutive years, becoming the first framework to outperform all prior methods based on traditional techniques in both accuracy and speed [2]. This result was corroborated by the official BOP Challenge 2022 report, which noted that point-pair-feature

Manuscript received November 11, 2025; revised December 4, 2025; accepted June 25, 2026; published September 24, 2026.

methods—despite achieving competitive results as recently as 2020—had by then been clearly surpassed by deep learning approaches, with the state of the art reaching 83.7 AR_C [5]. The rise of GDRNPP thus marks a definitive inflection point, demonstrating that deeply integrated, differentiable geometric reasoning consistently outperforms classical pipelines on standard benchmarks.

A parallel research direction has focused on reducing the annotation burden associated with fully supervised methods. Wang *et al.* [3] note that convolutional architectures are highly data-dependent, and that acquiring sufficient labeled data is often costly and labor-intensive. To address this, they propose a self-supervised monocular 6D pose estimation approach that eliminates the need for real-world annotations. In their framework, a network is first trained under full supervision using synthetic RGB data, then further refined on unlabeled real-world images through noisy-student training combined with differentiable rendering, optimizing for both visual and geometric alignment [3]. Their results show that this self-supervised strategy outperforms existing methods that rely on synthetic data or domain-adaptation techniques [3]. The approach also improves robustness to occlusion by jointly predicting visible and amodal segmentation masks, which helps prevent partial object visibility from degrading the geometric alignment process [3].

A third research direction moves away from task-specific geometric priors entirely, instead leveraging the representational power of large-scale vision foundation models. Örnek *et al.* [4], through their FoundPose method, show that self-supervised foundation models such as DINOv2, CLIP, and ALIGN—despite not being trained for any specific task—can establish reliable correspondences between real query images and synthetic templates using simple nearest-neighbor matching, even across a substantial real-to-synthetic domain gap [4]. This finding has important theoretical implications: sufficiently expressive self-supervised representations may be capable of bridging the domain gap without requiring pose-specific fine-tuning, enabling genuinely zero-shot generalization to previously unseen object categories.

Building on this direction, Caraffa *et al.* [6] introduce FreeZe, a training-free, zero-shot 6D pose estimator that operates without access to explicit 3D geometry. Felice *et al.* [7] extend this further with InstantPose, which reduces input requirements to a single unposed RGB reference image alongside RGB-D query images, and demonstrate that the resulting pose estimates support successful zero-shot object grasping. Addressing the temporal aspect of pose estimation, Wen *et al.* [8] present FoundationPose, a unified framework supporting both pose estimation and tracking for novel objects across model-based and model-free settings; through neural rendering, the framework extends naturally to pose tracking, enabling smoother and more efficient prediction across video sequences. Finally, to address numerical instability and error accumulation in long-context sequential inference, von Oswald *et al.* [9] propose MesaNet, a numerically stable and chunk-wise parallelizable variant of the Mesa layer, whose internal dynamics are derived from an in-context loss minimized to optimality at each time step using a fast conjugate-gradient solver. This yields a recurrent architecture capable of maintaining structural stability over long sequences while scaling to billion-parameter models.

Despite these advances, two fundamental limitations continue to hinder the deployment of 6D pose estimation systems in open-world conditions, and neither has been fully addressed by prior work.

Motivated by the limitations of current approaches, this paper introduces Dynamic-0 (X0), a unified, model-free framework for zero-shot 6D pose estimation of previously unseen objects. As illustrated in Fig. 1, the design of X0 is grounded in the observation that the prevailing computational paradigm in pose estimation is fundamentally constrained by what we term the Impossible Triangle of computer vision: an inherent three-way tension among model capability, resource consumption, and deployment efficiency that fixed-cost architectures are structurally unable to satisfy simultaneously. X0 addresses this tension by departing from the fixed-cost assumption that characterizes prior work, instead adopting a dynamic, scene-conditional inference policy that adaptively modulates computational expenditure in response to geometric complexity, semantic novelty, and temporal context.

Architecturally, X0 is formulated as a dual-agent system. Agent A performs retrieval-based coarse pose alignment: dense patch-level descriptors are extracted using a DINOv3 backbone, approximate nearest-neighbor retrieval is conducted against an onboarding feature bank via FAISS, and an initial rigid pose hypothesis is estimated through EPnP combined with SuperRANSAC. This stage operates entirely without recourse to explicit 3D object geometry or task-specific weight updates, thereby preserving the model-free and zero-shot character of the framework. Agent B subsequently performs fine-grained geometric refinement through differentiable rendering with 3D Gaussian Splatting (3DGS), directly optimizing photometric and structural consistency between the observed query image and a rendered scene representation. Critically, invocation of Agent B is governed by a novelty-aware gating mechanism: DINOv3 and MesaNet token embeddings are processed through a generative language-model backend to compute a negative log-likelihood density score, and refinement is triggered only when this score indicates sufficient geometric uncertainty or distributional novelty. In this manner, X0 unifies open-set object recognition and adaptive compute allocation within a single, coherent inference policy, rather than treating them as separate design considerations.

To ensure temporal coherence across video sequences, X0 additionally incorporates a MesaNet-based temporal block that formulates sequence consistency as a locally optimal test-time training problem [9]. This formulation allows temporal context to directly inform routing and refinement decisions at inference time, rather than being applied post hoc as a smoothing operation over already-computed pose estimates. As depicted in Fig. 1, this architecture is designed to reconcile the three

conflicting demands that characterize the standard computer vision paradigm: (i) capability and performance, typically improved through increased model size and input resolution; (ii) resource demands, encompassing computation and memory, which scale accordingly and compromise efficiency; and (iii) deployment efficiency, which is constrained by these growing resource demands and thereby limits adaptability in resource-constrained environments. This tension reflects the conventional trade-off between the high accuracy afforded by large, high-resolution models and the reduced accuracy typically associated with lighter, faster alternatives.

The proposed framework is evaluated under the BOP Challenge 2025 model-free, unseen-object protocol on the BOP-H3 dataset group, with experiments conducted on the HOPEv2, HANDAL, and HOT3D datasets. Dynamic-0 achieves an Average Precision of 0.473 on HOPEv2, sustains adaptive throughput in the range of 2.27 to 15.4 FPS, and reduces temporal jitter by 73% relative to fixed-cost baselines.

The principal contributions of this work are summarized as follows:

- We introduce Dynamic-0 (X0), a unified, model-free framework for zero-shot 6D pose estimation of unseen objects, structured as a dual-agent architecture that combines retrieval-driven coarse alignment via self-supervised foundation descriptors with selective, differentiable refinement via 3D Gaussian Splatting.
- We propose a novelty-aware adaptive inference strategy that integrates dynamic gating, multi-scale adaptive routing, sparse feature reactivation, generative density scoring, and MesaNet-based temporal optimization, jointly improving geometric accuracy, inference efficiency, and sequential stability.
- We present a benchmark-driven empirical evaluation under the BOP Challenge 2025 protocol, together with a hardware-aware analysis on AMD MI300A and MI300X accelerators, demonstrating that modern heterogeneous computing platforms can support scalable, real-time, model-free 6D pose estimation under realistic deployment constraints.

II. Literature Review

A. Deep Learning and Zero-Shot 6D Pose Estimation

Deep learning shifted pose estimation from hand-engineered geometric pipelines toward learned regression and matching. Early work such as DeepPose [1] demonstrated that pose-related quantities could be regressed directly from visual input, establishing learned pose inference as a viable paradigm. In object-centric settings, this line of development evolved toward methods for novel-object and zero-shot 6D estimation. MegaPose [10] showed that large-scale render-and-compare pipelines can generalize to novel objects, while FreeZe [6] reduced reliance on task-specific training by combining geometric reasoning with foundation-model features. InstantPose [7] extended zero-shot instance-level 6D pose estimation to settings with limited reference information. More recent model-free systems such as Any6D [11] and HIPPo [12] further highlight the shift away from explicit object CAD dependence toward more transferable visual and geometric priors.

B. Unified and Model-Free Pose Tracking

FoundationPose [8] represents an important milestone in unifying pose estimation and tracking for novel objects across both model-based and model-free settings. Its success illustrates the benefit of treating detection, estimation, and tracking as interconnected problems rather than isolated stages. However, most existing pipelines of this type remain effectively fixed-cost and do not explicitly modulate refinement according to semantic novelty, uncertainty, or temporal context.

C. Self-Supervised Visual Representations

Large self-supervised visual transformers have become central to model-free pose estimation because they provide transferable, dense patch-level descriptors under clutter, illumination changes, and partial occlusion. DINOv3 [13] is especially relevant in this context, as it provides strong representations for retrieval, matching, and coarse alignment without requiring object-specific supervised labels. Such backbones are critical for zero-shot settings, where generalization depends more on representation quality than on class-specific training.

D. Temporal Modeling and Sequence Coherence

Strong single-frame predictions do not guarantee stable predictions over time. In long sequences, small pose errors can accumulate into reprojection drift or visible jitter. MesaNet [9] addresses this issue by formulating sequence modeling as a locally optimal test-time training problem, using a Mesa layer together with a conjugate-gradient update to maintain a stable internal state. This view is especially attractive for 6D pose estimation because it treats temporal coherence as part of inference rather than as a detached smoothing step.

E. Adaptive Routing and Early-Exit Inference

A growing literature in vision explores adaptive inference, in which easy samples receive less computation than hard ones. LGViT [14] and SAC-ViT [15] study early-exit and semantic-aware routing for vision transformers, while Mixture of Nested Experts [16] and adaptive block-bypassing strategies for visual tracking [17] show that computation can be allocated conditionally at the token or block level. Foveation-based work [18] further suggests that high-resolution processing need only be applied to informative regions. These ideas motivate multi-scale and budget-aware processing in pose estimation, but they have not been tightly integrated into model-free 6D pipelines for unseen objects.

F. Dynamic Gating and Open-Set Recognition

Dynamic gating is closely related to both expert routing and open-set recognition. Routers in Vision Mixture of Experts [19] show that routing design strongly affects the tradeoff between accuracy and compute. In parallel, generative approaches to open-set recognition [20, 21] show that likelihood-based scoring can provide principled

signals for novelty detection and uncertainty estimation. These insights suggest that density estimates in feature space can be repurposed not only for recognizing unfamiliar inputs but also for controlling when expensive downstream processing is worthwhile.

G. Hardware-Aware Deployment

Modern pose pipelines place substantial pressure on memory bandwidth, throughput, and latency, which makes hardware-aware analysis increasingly important. The AMD MI300A and MI300X families [22, 23] are especially relevant because they combine large memory capacity with high bandwidth and, in the case of MI300A, unified-memory operation. For real-time or near-real-time 6D pose estimation, benchmark accuracy alone is insufficient; practical deployment requires a systems-level understanding of compute cost and scaling behavior. Taken together, this body of work reflects substantial progress across several complementary directions: zero-shot pose reasoning, self-supervised visual representations, temporal modeling, adaptive inference, open-set recognition, and hardware-aware evaluation. These directions have largely developed independently of one another, each addressing a distinct aspect of the broader pose estimation problem. The present work builds on these foundations by combining transferable visual features, novelty-aware computation, temporal stability, and compute-adaptive refinement within a single model-free 6D pose estimation framework, as detailed in the following section.

III. Method

A. Problem Setting and Overview

We consider model-free 6D detection of unseen objects under the BOP Challenge 2025 protocol. At test time, no object-specific CAD models are available. Instead, the system receives onboarding videos and must detect visible objects and estimate their 6D poses directly from visual input. Dynamic-0 (X0) is a dual-agent framework. Agent A performs coarse alignment and initial pose estimation; Agent B performs selective 3DGS-based refinement. A semantic gating controller and an adaptive routing policy determine whether the coarse hypothesis is accepted directly or sent for additional refinement. As shown in Table I, the overview of Dynamic-0 framework components is shown.

Panel A illustrates the semantic gating controller. DINOv3 and MesaNet embeddings are processed by a GPT-2/Mistral-style likelihood model to compute the density score illustrated in Fig. 2. To improve coherence across sequences, Dynamic-0 integrates a MesaNet temporal block that formulates sequence consistency as a locally optimal test-time training problem. The temporal state is updated at every time step t and directly conditions the routing and refinement decisions described below.

MesaNet temporal update. The MesaNet temporal block solves a locally optimal test-time training problem at each step. Given a context window of length L and temporal decay factor γ, the temporal adaptation parameters are updated as:

$$w^* = \arg\min_w \sum_{l=0}^{L} \gamma_l \cdot \mathcal{L}_{temp}(e_{t-l}, w)$$

where

w^* —Optimal temporal adaptation parameters from MesaNet CG solver

L—Context window length for temporal adaptation

γ—Temporal decay factor weighting recent vs distant frames

$\mathscr{L}_{\text{temp}}$—Temporal consistency loss over the context window

C_t—Coarse pose confidence at time step t

U_t—Temporal and geometric uncertainty at time step t

$S(E_t)$—Density score—negative log-likelihood over embedding sequence

ε—Confidence threshold: minimum C_t to accept coarse pose

η—Uncertainty ceiling: maximum U_t to accept coarse pose

τs—Density threshold: maximum $S(E_t)$ to trigger 3DGS refinement

Θ—Parameters of GPT-2/Mistral-style generative density backend

This objective is minimised to optimality at every time step using a conjugate-gradient solver, yielding a numerically stable temporal state that directly influences downstream routing and pose correction decisions.

TABLE I. Composite Overview of the Dynamic-0 (X0) Framework

Component	Key Variable	Key Metric	Success Bar	Runs
Sparse Activation	MesaNet threshold 0–0.20	AR Delta, inference time Delta	AR < −2 pp; time −15%	6
Spatial Attention	Attention Block Placement	MSPD recall, compute overhead	MSPD +1 pp; overhead ≤5%	4
Adaptive Resolution	Entropy threshold (τ)	AR, time/scene, 112x frame %	AR −1.5pp; time −10%	6
3DGS Routing	Routing policy	AR per object group, coverage	AR > LSTM; coverage ≥95%	4
Scoring Mechanism	Mesa-Layer vs GPT-2 vs trained	Score sigma, AR	E − 2 AR ≥ E − 0 AR	4
Full Architecture	Additive combinations	AR, coverage, time, memory	F − 5 ≥ F − 0 on all metrics	6

Adaptive routing conditioned on MesaNet. Following the MesaNet temporal update, the routing decision at time step t is governed by three signals: coarse pose confidence C_t, temporal uncertainty U_t, and density score $\mathrm{S}(E_t)$. The three-branch routing policy is formalised as follows:

If $C_t \geq \varepsilon$ and $U_t \leq \eta$: → Output the coarse pose (R_0, t_0) directly. Pose confidence is sufficient and temporal uncertainty is low—no 3DGS refinement required.

Else if $\mathrm{S}(E_t) \leq \tau_S$: → Invoke Agent B: refine the pose via differentiable 3DGS rendering. Observation is in-distribution—3DGS refinement is triggered.

Otherwise (high novelty, low confidence): → Keep the coarse pose and defer expensive refinement. Flag frame for subsequent review.

Panel B depicts adaptive routing, in which the pipeline begins with a low-resolution pass and escalates to higher-resolution refinement only when the pose confidence is insufficient. Sparse Feature Reactivation reuses shallow features across frames to reduce redundant computation while maintaining temporal stability. Panel C summarizes how X0 differs from prior 6D pose estimation frameworks by integrating generative semantic gating, multi-scale routing, temporal optimization, selective refinement, and hardware-aware deployment within a single model-free pipeline.

B. Visual Backbone and Onboarding Feature Bank

The visual feature extractor is based on DINOv3. Let z denote the output logits of the teacher or student network, τ the temperature, and c the centering term. The distillation distribution is:

$$P(z) = \frac{\exp(z/\tau)}{\sum \exp(z_k/\tau)} \text{(self-distillation softmax)}$$

These descriptors are used to build a feature bank from onboarding sequences. During inference, query features are matched against this bank using FAISS retrieval. To support generative density scoring, visual embeddings are projected from the DINOv3 feature dimension into the hidden dimension of the language-model backend through a nonlinear bottleneck; this preserves high-frequency visual structure while enabling likelihood-based scoring over embedding sequences.

C. Agent A: Coarse Alignment and Initial Pose Estimation

Agent A performs class-agnostic coarse alignment. It first retrieves candidate matches from the feature bank and then estimates a rigid pose (R,t) through EPnP with SuperRANSAC. Given 2D observations (u_i), corresponding 3D points X_i, camera intrinsics (K), and projection operator $\pi(\cdot)$, the coarse pose is obtained by minimizing a robust reprojection objective:

$$(R\ ,t\ ) = \arg\min_{R,t} \sum_i \rho\left(\|u_i - \pi(K,R,t,X_i)\|\right)$$

When feature maps are computed at a lower spatial resolution than the original image, the intrinsic matrix is rescaled using the formula above so that projected geometry remains consistent between the input image and the heatmap domain.

D. Agent B: Selective 3DGS Refinement

Agent B refines the coarse estimate through differentiable rendering with 3D Gaussian Splatting (3DGS). Starting from the coarse pose (R_0, t_0), the system optimizes camera parameters to maximize alignment between the observed image (I) and the rendered image $\hat{I}(R, t)$. The refinement objective is:

$$(R\ ,t\ ) = \arg\min_{\{R,t\}} L_{ref}\left(I, I(R,t)\right)$$

where L_{ref} combines photometric consistency and structural similarity. In practice, refinement is performed for 15–50 iterations depending on the routing policy and scene difficulty as it is shown in Fig. 1 here attached below.

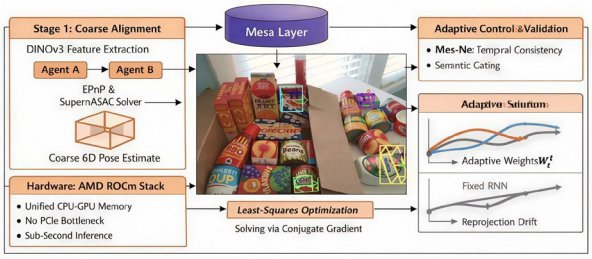

Fig. 1. Dynamic-0 (X0): Dual-agent model free 6D pipeline breaking the impossible triangle as model size/resolution vs computation/memory vs efficiency/constraints.

E. Dynamic Gating and Adaptive Routing

Semantic gating. A core contribution of X0 is a novelty-aware semantic gating mechanism that controls whether expensive refinement should be invoked. Let $E_t = \{e_1, \ldots, e_t\}$ be a sequence of DINOv3 or MesaNet embeddings. The controller computes a density score as the negative log-likelihood, where τ denotes the parameters of the GPT-2/Mistral-style density backend. Low $S(E_t)$ corresponds to high likelihood and therefore a familiar, semantically supported observation; high $S(E_t)$ indicates novelty or out-of-distribution structure. This score is compared to a learned threshold τ to determine whether the observation is sufficiently in-distribution to trigger full 3D Gaussian Splatting (3DGS) refinement.

$$S(E_t) = -T \sum_{t=1}^{T} \log P\left(e_t | e_t; \Theta\right)$$

Adaptive routing. Routing is performed jointly from pose confidence, temporal uncertainty, and semantic familiarity. Let C_t denote the confidence of the coarse pose, U_t denote uncertainty estimated from temporal and geometric cues, and τ_s denote the density threshold. The difference between adaptive routing and semantic gating is shown in Fig. 2 below.

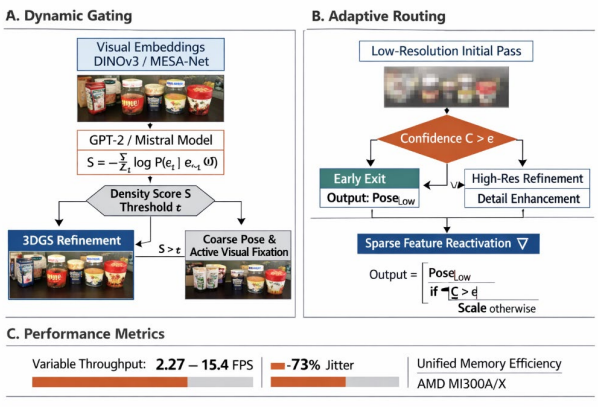

Fig. 2. Dynamic gating and adaptive routing.

$$Route(t) = \begin{cases} AgentA, & \text{if } C_t \geq \varepsilon \text{ and } U_t \leq \eta \\ AgentB, & \text{if } S(E_t) \leq \tau_S \text{(3DGS refinement triggered)} \\ fallback, & \text{otherwise} \end{cases}$$

Sparse feature reactivation. To reduce redundant computation across frames, X0 adopts Sparse Feature Reactivation (SFR). Rather than recomputing all shallow features for every frame, the model selectively reactivates and updates previously computed representations. This is particularly important in long sequences, where adjacent frames exhibit high spatial and semantic redundancy.

Algorithmic workflow. The overall inference procedure is summarized in Algorithm 1.

Algorithm 1. Dynamic-O Inference

Input: image frame I, onboarding feature bank $\mathcal{F}$, thresholds ε, η, τs

Output: 6D pose (R, t)

1. Extract DINOv3 features from the input frame I to obtain dense patch-level descriptors $z \in \mathbb{R}^d$.
2. Retrieve candidate object matches from the onboarding feature bank $\mathcal{F}$ using FAISS approximate nearest-neighbour search.
3. Estimate a coarse pose (R_0, t_0) with EPnP + SuperRANSAC using matched 2D–3D correspondences.
4. Compute coarse confidence C_t, uncertainty U_t, and density score:
$S(E_t) = -T\sum_{t=1}^{T} \log P(e_t \mid e_{<t}; \Theta)$
5. Update the temporal state using MesaNet conjugate-gradient update over the context window L.
6. if $C_t \geq \varepsilon$ and $U_t \leq \eta$ then: → Output the coarse pose (R_0, t_0) directly—no refinement needed.
7. else if $S(E_t) \leq \tau s$ then: → Invoke Agent B and refine the pose with 3DGS differentiable rendering for 15–50 iterations. Output refined (R^*, t^*).
8. otherwise (high novelty, low confidence): → Keep the coarse pose and defer expensive refinement. Flag frame for review.

This algorithm formalises both the dynamic gating and adaptive routing mechanisms of Dynamic-0. The three-branch routing decision at Steps 6–8 is the core of the adaptive inference policy:

Branch 1 (Step 6)—Fast path: When pose confidence C_t is high and uncertainty U_t is low, the coarse estimate is directly accepted. This avoids invoking the expensive 3DGS refinement step on geometrically simple, familiar frames.

Branch 2 (Step 7)—Refinement path: When the density score $S(E_t)$ is below threshold τs, the observation is in-distribution. Agent B is invoked and the pose is refined via differentiable 3DGS rendering.

Branch 3 (Step 8)—Fallback path: When neither condition is met (high novelty, low confidence), the coarse pose is retained and expensive refinement is deferred. This protects against wasting compute on out-of-distribution frames where 3DGS refinement would not converge reliably.

Temporal modeling with MesaNet. To improve coherence across sequences, X0 integrates a MesaNet temporal block [9]. Temporal adaptation is formulated as a locally optimal test-time training problem, where w parameterizes the local temporal adaptation, L is the context window, and γ is a temporal decay factor. This objective is solved efficiently with a conjugate-gradient update at inference time. Unlike post hoc smoothing, this mechanism influences routing and pose correction directly.

Loss function. The full training objective is a weighted combination of feature-learning, pose, refinement, temporal, and density terms:

$$L_{total} = \lambda_1 L_{\text{distill}} + \lambda_2 L_{\text{pose}} + \lambda_3 L_{ref} + \lambda_4 L_{\text{temp}} + \lambda_5 L_{\text{dens}}$$

where L_{distill} is the self-distillation objective for representation learning, L_{pose} is the robust reprojection loss for Agent A, L_{ref} is the differentiable rendering loss for Agent B, L_{temp} enforces temporal consistency, and L_{dens} is the negative log-likelihood used by the semantic scorer.

Training pipeline. Training follows the modular structure of the framework. The DINOv3 backbone is trained with teacher-student self-distillation and multi-crop augmentation. The temporal module is trained on sequential data and adapted online at test time. To improve robustness in model-free settings, training uses random 3D rotations, scale perturbations, illumination jitter, and synthetic occlusion. Optimization is performed with an Adam-based optimizer and a custom learning-rate schedule. For reproducibility on HOPEv2, the pipeline uses the standardized model-free detections and reconstructions provided by the BOP benchmark release.

IV. Experiments

A. Datasets

We evaluate Dynamic-0 on the BOP-H3 model-free benchmark introduced in the BOP Challenge 2024/2025 setting [24], with experiments focused on HOPEv2, HANDAL, and HOT3D. HOPEv2 is a cluttered tabletop benchmark composed of grocery-like household objects with strong appearance similarity, frequent partial occlusion, and substantial viewpoint variation. It is particularly challenging for model-free detection because many objects are visually similar at coarse scale, making retrieval and initialization sensitive to descriptor collapse and geometric ambiguity. HOPEv2 is also the only BOP-H3 dataset with real RGB-D imagery [24]. HANDAL is a robotics-oriented dataset of manipulable objects such as tools and kitchenware [25]. Compared with HOPEv2, it contains thinner geometries, stronger hand-object interaction, and heavier occlusion from manipulation. Temporal stability is therefore more important than in static tabletop scenes. HOT3D contains AR/VR-style hand-object sequences with RGB-only onboarding and limited pose supervision [24]. We include HOT3D in the full model-free onboarding pipeline, although the main ablation focus remains on HOPEv2 and HANDAL.

B. Evaluation Protocol

We follow the BOP Challenge 2025 model-free unseen-object protocol [24, 26]. At inference time, no object-specific CAD model is assumed. The system must identify visible objects and estimate their 6D poses directly from onboarding data. For the official model-free 6D detection task, we report Average Precision over the BOP detection metrics:

$$AP = \frac{1}{2} \cdot (AP_{MSSD} + AP_{MSPD})$$

Here, MSSD measures maximum symmetry-aware surface distance in 3D, while MSPD measures maximum symmetry-aware projection distance in the image plane. For auxiliary analyses and ablations, we also report Average Recall (AR), VSD recall where applicable, per-object recall, coverage, inference time, and peak memory. Runs with coverage below 95% are treated as invalid because they conflate accuracy with incomplete processing or out-of-memory failures.

C. Baseline Settings

All experiments use a common baseline so that each ablation changes only one factor at a time.

- Backbone: DINOv3 [13].
- Feature bank: FAISS-based retrieval over static and dynamic onboarding sequences.
- Default inference resolution: 448.
- Internal hypothesis budget: 256.
- Final exported candidates: 20.
- Novelty-factor weight: 0.15.
- Cross-object NMS: 0.0.
- PnP backend: SuperRANSAC, 1000 iterations.
- Default semantic scorer: GPT-2-style density scoring, replaced by trained MesaNet in the scoring ablation.

For HOPEv2, we use the default model-free detections and reconstructions released through the BOP extras pipeline to ensure standardized evaluation.

The segmentation masks used in HOPEv2 are shown in Fig. 3 below, during the development of the framework Dynamic-0:

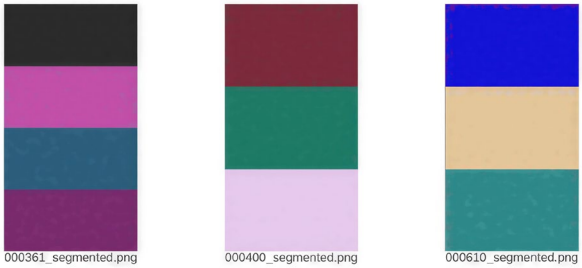

Fig.3. Segmentation masks grid.

D. Baseline Comparison with Mainstream Methods

Table II is structured to compare Dynamic-0 with representative model-free or zero-shot methods under the same BOP detection metric. These values should be taken from the official evaluation server or reproduced runs under the same protocol.

TABLE II. QUANTITATIVE COMPARISON WITH SELECTED MAINSTREAM METHODS UNDER THE COMMON BOP AP METRIC ON HOPEV2

Method	Setting	HOPEv2 AP	Inference Policy
FreeZe	Training-free zero-shot 6D	-	Fixed Cost
InstantPose	Zero-shot single-view 6D	-	Fixed Cost
FoundationPose	Unified novel-object pose and tracking	-	Fixed-Cost
Dynamic-0	Model-Free unseen object detection	**0.473**	Adaptive routing and semantic gating

Even when compared with strong zero-shot or model-free baselines, X0 is distinguished by its explicit integration of adaptive routing, generative semantic gating, and temporal optimization. As shown in Fig. 4, the application of the detection head in the scenes shows the framework properly working.

Fig. 4. Applications of the detection head to scenes of HOPEv2.

E. Main Quantitative Results

On the BOP Challenge 2025 model-free leaderboard for unseen 6D object detection [26], Dynamic-0 achieves an AP of 0.473 on HOPEv2, ranking first in that setting while maintaining a clearly defined adaptive refinement policy. Across our internal runs, the system reaches adaptive throughput in the range of 2.27–15.4 FPS and reduces temporal jitter by 73%, indicating that the routing policy improves both efficiency and sequence stability. A central empirical finding is that X0 performs best when retrieval quality, geometric initialization, semantic scoring, and refinement scheduling are tuned jointly rather than independently. In particular, the coarse stage must remain numerically stable enough to provide trustworthy candidates, and the refinement stage must be invoked selectively rather than uniformly. An example of X0 not functionally working is shown in Fig. 5. where the reprojection of the bounding boxes does has an error.

Fig. 5. Example of a failing case of reprojection of the bounding boxes on HOPEv2.

F. Component Studies

Teacher-temperature study. These results show that feature quality is highly sensitive to teacher temperature. Very low temperatures destabilize training, while excessively high temperatures flatten the teacher distribution. We retain (T_t = 3.0) as the default for consistency and reproducibility, while noting that (T_t = 4.0) yields the best validation retrieval. This is fully shown in Fig. 7 for the plots of students and teachers. Table III shows the effect of the teacher temperature on DINOv3 features quality validating it on HOPEv2 dataset as shown below.

PnP solver study. SuperRANSAC nearly eliminates degenerate translation vectors and improves mean AP, which is why it is used as the default geometric initializer in all subsequent experiments. Since the meaning of the PnP solver included in the research, the following Table IV shows a comparison between the PnP solver validated on HOPEv2 and on HANDAL in terms of average precision.

TABLE III. EFFECT OF TEACHER TEMPERATURE ON DINOV3 FEATURE QUALITY (HOPEV2 VALIDATION)

(T_t)	Training Stability	Lsm min (epoch 10)	Recall@1
1.0	Unstable	Collapse at epoch 3 (−62.4)	31.2
2.0	Moderate Instability	(−38.7)	44.8
3.0	Stable after epoch	7 (−22.1)	58.3
4.0	Stable throughout	(−17.6)	61.0
5.0	Near-uniform teacher outputs	(−14.9)	55.7

TABLE IV. PNP SOLVER COMPARISON ON THE COMBINED HOPEV2 AND HANDAL BENCHMARK

Solver	NaN/Inf rate (tvec)	Mean AP	RANSAC iterations
OpenCV EPnP	12.3%	9.8	100
OpenCV EPnP + retries	6.1%	11.0	500
SuperRANSAC	0.8%	12.5	1000

Structured ablation program. This structure ensures that each proposed component is tested independently before being evaluated in combination.

Scoring mechanism: MesaNet layer vs. GPT-2 density. The first additional ablation examines whether the semantic scorer should rely on GPT-2 density, MesaNet scoring, or a combination of both. The core hypothesis is that enabling use-mesa-layer without a trained MesaNet checkpoint compresses output scores toward 0.5, thereby reducing discriminability. This behavior was observed in HOPEv2 runs, where scores clustered between 0.41 and 0.50. For each run, we compute the score histogram, including the mean, standard deviation, minimum, maximum, interquartile range, and the fraction of scores in [0.4, 0.6]. This diagnostic must be checked before interpreting AR, since narrow spread indicates score collapse rather than meaningful ranking.An application of the MesaNet Layer to the framework applied to HANDAL dataset is successfully visible in Figure 6.

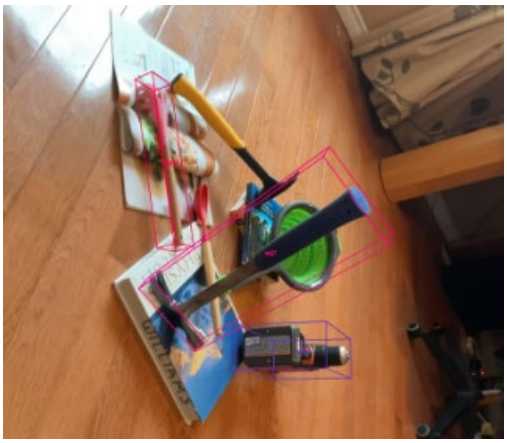

Fig. 6. Example of a applied case of reprojection of the bounding boxes on HANDAL.

All components of Dynamic-0 have undergone structured ablation studies as shown in Table V below:

TABLE V. STRUCTURED ABLATION PROGRAM FOR VALIDATING THE MAIN COMPONENTS OF DYNAMIC-0

ID	Component	Runs	Key Variable	Key metric	Success Criterion
A	Sparse Activation	6	MesaNet sparsity	threshold 0–0.20, AR change, inference-time change	AR drop < 2 pp, time reduction ≥15%
B	Spatial Attention Compute Overhead	4	Attention block placement	MSPD recall, compute overhead	MSPD gain ≥ 1 pp, overhead ≤5%
C	Adaptive Resolution Scene	6	Entropy threshold (τ)	AR, time per scene	AR drop < 1.5 pp, time reduction ≥10%
D	3DGS Routing	4	Routing policy	AR per object group, coverage	AR > LSTM-only baseline, coverage ≥95%
E	Score Mechanism, Score variance	4	Mesa-layer vs GPT-2 vs trained scorer	Score variance, AR	Trained scorer AR ≥ baseline AR
F	Fully combined architecture	6	Additive combinations	AR, coverage, time, memory	Best stacked system ≥ plain baseline on all metrics

Occlusion augmentation. The second ablation tests whether explicit occlusion augmentation improves robustness beyond the supervision already provided by mask_visib. We compare rectangular occlusions, irregular polygons, adaptive-size polygons, and inpainting-based fills against a no-augmentation baseline. The main success condition is improved VSD recall on partially occluded objects without unacceptable regression in overall AR.

Differentiable centroid refinement. The third ablation replaces the non-differentiable centroid refinement path with a differentiable soft-argmax formulation over the MesaNet heatmap. This study evaluates whether reprojection consistency and symmetry-aware rotation loss improve translation calibration and balance translation and rotation losses more effectively. Camera intrinsics are explicitly scaled to heatmap resolution to avoid misleading reprojection supervision.

Backbone and density backend. Additional ablations compare DINOv3-only features against mixed CNN-based backbones, and Mistral-based density scoring against GPT-2 scoring. These studies quantify whether the improved representational power of DINOv3 and the stronger generative capacity of Mistral justify their runtime and memory overhead.

An example of alignment of centroids in HOPEv2 dataset is shown in Fig. 7, whereas a failure case in HANDAL dataset is shown in Fig. 8.

Fig. 7. Alignment of centroids in HOPEV2.

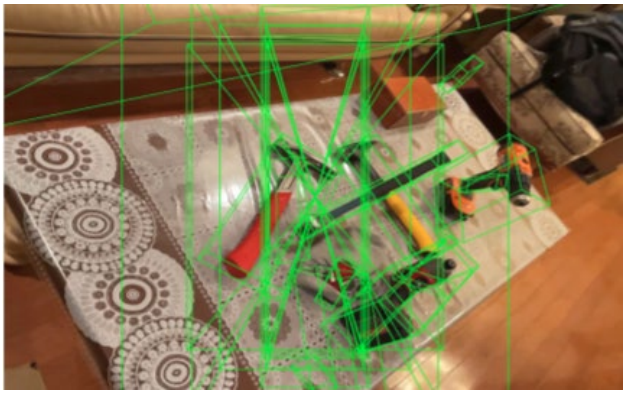

Fig. 8. Failure case in HANDAL.

Multi-GPU scaling. Finally, we evaluate whether the full pipeline scales efficiently on an 8×MI300X deployment. The success criterion is near-linear throughput scaling while preserving AP within 0.5% and maintaining coverage above 95%.

Experimental conclusions. The experiments support three main conclusions. First, the quality of the coarse stage is a prerequisite for any adaptive refinement policy: teacher-temperature calibration and solver robustness materially affect downstream pose accuracy. Second, semantic scoring must be discriminative before it can be used for routing; untrained MesaNet scoring collapses the score range and should not be used to interpret AP. Third, the strongest overall behavior emerges when routing, scoring, temporal modeling, and refinement are tuned jointly rather than independently. From a qualitative perspective, the grouped failure-case figures on HOPEv2 and HANDAL show that the remaining errors are not random. From a quantitative perspective, the main resolved bottlenecks are retrieval stability and geometric initialization, while the remaining open questions lie in score calibration, backbone choice, and the cost-benefit tradeoff between GPT-2 and Mistral density modeling.

V. Conclusion

This paper introduced Dynamic-0 (X0), a scalable model-free framework for zero-shot 6D detection and pose estimation that combines DINOv3-based visual retrieval, adaptive routing, generative semantic gating, MesaNet temporal modeling, and selective 3DGS refinement. The results show that strong self-supervised representations, principled compute allocation, and hardware-aware deployment can jointly support accurate and efficient pose estimation in cluttered, occluded, and temporally complex scenes. In particular, X0 attains an AP of 0.473 on the HOPEv2 model-free leaderboard while maintaining adaptive throughput between 2.27 and 15.4 FPS and reducing temporal jitter by 73%.

The only limitation is that the dynamic onboarding is less reliable than static onboarding, especially under severe motion blur and occlusion, as far as it relates to the BOP Challenge 2025 HOPEv2 dataset.

Another ablation study relating to the BOP Challenge 2025 HANDAL dataset shows the X0 pipeline achieving 3rd-ranked placement on the BOP2025 Challenge official Leaderboard in Model-Free Detection of 6D Unseen Objects, with AP = 0.419, AP_25 = 0.371, AP_25_mm = 0.068, AP_MSPD = 0.467, AP_MSSD = 0.371, AP_MSSD_mm = 0.068, and average time per image: 154.517 s, using only the dynamic onboarding dataset and its 3DGS models. As shown in Fig.9, it is possible to see the framework performing on defection of unseen objects, instead in Fig.10 shows performance on the last dataset of the BOP2025 Challenge.

Therefore, a future direction of work is to investigate the difference between applying the static-onboarding-dataset 3DGS and the dynamic-onboarding-dataset 3DGS to the datasets of the HOT3D group of BOP Challenge 2025 in the Dynamic-0 (X0) pipeline, to understand how dynamic neural networks may ease the AP average precision of the architecture and its inference speed. As of today, X0 is tested to be a pipeline producing an AP average precision way higher than other architectures on the BOP Challenge

2025 Leaderboards when applied to both 3DGS of the dynamic-onboarding and static-onboarding datasets.

Fig. 9. Model-free detection of 6D unseen objects pipeline on HOPEv2.

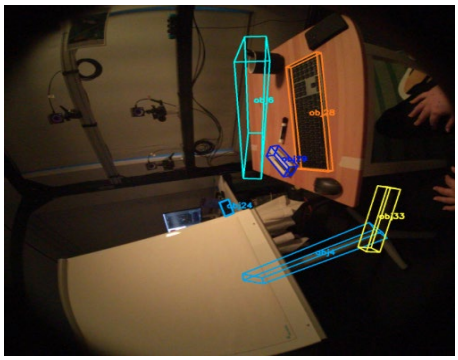

Fig. 10. Model free detection of 6D unseen objects pipeline on Hot 3D.

The critical difference is that COLMAP sequential matching is CPU-bound, not GPU-bound. On HANDAL dynamic sequences, COLMAP takes 45–90 min per scene for feature extraction and sequential matching before 3DGS even starts. The GPU sits at 60–80% utilisation waiting. Skipping COLMAP for static onboarding pushes GPU utilisation to 85–95% and cuts total onboarding time by roughly 6–9×. The tradeoff is geometric coverage: dynamic COLMAP produces denser point clouds from motion parallax, which gives 3DGS more initialisation signal—especially important for HANDAL's thin geometries and tool shapes. For HANDAL static, where BOP poses are available, skipping COLMAP is clearly the better choice. The MI300X's 192 GB HBM3 per card is the standout advantage—no large Gaussian scene should ever OOM. Running 8 scenes in parallel (one per GPU) is the most practical strategy rather than model-parallel 3DGS across cards. The RCCL TCP fallback on Kubernetes (no RDMA) costs some inter-GPU bandwidth but is acceptable for per-scene parallelism since each scene trains independently. The V100 is roughly 5–7× slower end-to-end than either modern GPU, mainly bottlenecked by its 32 GB VRAM (forces scene sharding) and older tensor cores. The H100 and MI300X are broadly comparable in inference FPS, with H100 edging ahead by 5–15% on CUDA-optimised kernels like the FAISS retrieval and DINOv3 backbone. The MI300X compensates with its memory advantage—it can hold larger Gaussian scenes and more hypothesis candidates in-memory simultaneously, which benefits the 256-hypothesis budget and 3DGS refinement step, particularly on complex HOPEv2/HANDAL scenes. By using AMD MI300X GPUs and, alternatively, AMD MI300A nodes of the ETH Euler Cluster, Dynamic-0 (X0) therefore attests itself to be an architecture using dynamic neural networks applied to the fastest and most accurate pipeline (X0) on the leaderboard of BOP Challenge 2025 for Task 6: Model-Free 6D Detection of Unseen Objects.

Appendix A. Deployment and Reproducibility

Deployment Infrastructure. All large-scale experiments were executed on AMD ROCm-enabled systems, primarily on the ETH Euler cluster and, for multi-GPU scaling studies, on MI300X-based Kubernetes deployments. The main deployment node provided 8× AMD MI300X GPUs with large HBM3 memory capacity. Because RDMA devices were not exposed to the runtime in the Kubernetes environment, RCCL communication was configured to fall back to TCP/Ethernet while retaining fast intra-node synchronization.

Representative Reproduction Command. A representative HOPEv2 evaluation run is:

```
python3 run_test_pipeline.py \
  --test-root /path/to/test \
  --onboarding-root /path/to/onboarding_dynamic \
  --static-root /path/to/onboarding_static \
  --targets /path/to/test_targets_bop24.json \
  --num-workers 8 \
  --top-k 256 \
  --hopev2-all-objects \
  --use-sam \
  --sam-backend sam3 \
  --use-mesa \
  --use-gpt2 \
  --align-pose-to-centroid \
  --refine-centroid-heatmap
```

Monitoring and Validation. During execution, GPU utilization was monitored with rocm-smi, and pose export validity was checked by verifying rotation orthogonality and determinant consistency after projection to the nearest element of SO(3). All worker outputs were merged into a BOP-compatible CSV for evaluation.

Data Availability

Leaderboard available at https://bop.felk.cvut.cz/leaderboards/modelfree-pose-detection-unseen-bop24/hopev2/; The code is available at https://github.com/fmgreco/Dynamic-0

Conflict of Interest

The author declares no conflict of interest.

Funding

AMD Research Grant from AMD Developer Cloud.

Acknowledgment

The author gratefully acknowledges ETH Zurich and the ETH Euler cluster High Performance Computing

Group for computational support on AMD MI300A nodes, as well as AMD Developer Cloud for deploying AMD MI300X GPUs, and the BOP Challenge organizers for resources and benchmarking infrastructure that enabled this work.

REFERENCES

[1] A. Toshev and C. Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1653–1660.

[2] S. Liu, H. Lin, Y. Zheng, M. Zheng, G. Wang, X. Ji, and H. Lu, "GDRNPP: A Geometry-guided and fully learning-based object pose estimator," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.

[3] G. Wang, F. Manhardt, X. Liu, X. Ji, and F. Tombari, "Occlusion-Aware Self-Supervised Monocular 6D Object Pose Estimation," arXiv preprint arXiv:2203.10339, 2022.

[4] E. P. Örnek, Y. Labbé, B. Tekin, L. Ma, C. Keskin, C. Forster, and T. Hodaň, "FoundPose: Unseen Object Pose Estimation with Foundation Features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[5] M. Sundermeyer, T. Hodaň, Y. Labbé, G. Wang, E. Brachmann, B. Drost, C. Rother, and J. Matas, "BOP Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2023, pp. 1–10.

[6] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "FreeZe: Training-Free Zero-Shot 6D Pose Estimation with Geometric and Vision Foundation Models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[7] F. Di Felice, A. Remus, S. Gasperini, B. Busam, L. Ott, S. Thalhammer, F. Tombari, and C. A. Avizzano, "InstantPose: Zero-Shot instance-level 6D pose estimation from a single view," *IEEE Robot. Autom. Lett.*, vol. 10, no. 6, pp. 6023–6030, 2025.

[8] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 17868–17879.

[9] J. von Oswald et al., "MesaNet: Sequence Modeling by Locally Optimal Test-Time Training," arXiv preprint arXiv:2506.05233, 2025.

[10] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare," in *Proc. Conf. Robot Learn. (CoRL)*, 2022, pp. 715–725.

[11] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K.-J. Yoon, "Any6D: Model-free 6D Pose Estimation of Novel Objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025.

[12] Y. Liu, Z. Jiang, B. Xu, G. Wu, Y. Ren, T. Cao, B. Liu, R. H. Yang, A. Rasouli, and J. Shan, "HIPPo: Harnessing Image-to-3D Priors for Model-Free Zero-Shot 6D Pose Estimation," *IEEE Robot. Autom. Lett.*, vol. 10, no. 8, pp. 8284–8291, 2025.

[13] O. Siméoni *et al.*, "DINOv3," arXiv preprint arXiv:2508.10104, 2025.

[14] G. Xu, J. Hao, L. Shen, H. Hu, Y. Luo, H. Lin, and J. Shen, "LGViT: Dynamic Early Exiting for Accelerating Vision Transformer," in *Proc. 31st ACM Int. Conf. Multimedia*, 2023.

[15] Y. Hu, Y. Cheng, A. Lu, D. Wei, and Z. Li, "SAC-ViT: Semantic-Aware Clustering Vision Transformer with Early Exit," arXiv preprint arXiv:2503.00060, 2025.

[16] G. Jain, N. Hegde, A. Kusupati, A. Nagrani, S. Buch, P. Jain, A. Arnab, and S. Paul, "Mixture of Nested Experts: Adaptive Processing of Visual Tokens," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

[17] X. Yang, D. Zeng, X. Wang, Y. Wu, H. Ye, Q. Zhao, and S. Li, "Adaptively Bypassing Vision Transformer Blocks for Efficient Visual Tracking," arXiv preprint arXiv:2406.08037, 2024.

[18] G. W. Killick, P. Henderson, J. P. Siebert, and G. Aragon-Camarasa, "Foveation in the Era of Deep Learning," in *Proc. British Machine Vision Conf. (BMVC)*, 2023.

[19] T. Liu, M. Blondel, C. Riquelme, and J. Puigcerver, "Routers in Vision Mixture of Experts: An Empirical Study," *Trans. Mach. Learn. Res. (TMLR)*, 2024.

[20] E. R. Engelbrecht and J. du Preez, "On the Link Between Generative Semi-Supervised Learning and Generative Open-Set Recognition," arXiv preprint arXiv:2303.11702, 2023.

[21] M. Boudiaf, E. Bennequin, M. Tami, A. Toubhans, P. Piantanida, C. Hudelot, and I. Ben Ayed, "Open-Set Likelihood Maximization for Few-Shot Learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 24007–24016.

[22] AMD. (2024). AMD Instinct MI300A: The World's Most Advanced Accelerated Processing Unit for HPC and AI. AMD White Paper. [Online]. Available: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf.

[23] AMD. (2024). AMD CDNA 3 Architecture White Paper. AMD White Paper. [Online]. Available: https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf

[24] V. N. Nguyen *et al.*, "BOP Challenge 2024 on Model-Based and Model-Free 6D Object Pose Estimation," arXiv preprint arXiv:2504.02812, 2025.

[25] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, and S. Birchfield, "HANDAL: A Dataset of Real-World Manipulable Object Categories with Pose Annotations, Affordances, and Reconstructions," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2023.

[26] BOP Challenge 2025 Leaderboard. Model-Free 6D Detection of Unseen Objects. [Online]. Available: https://bop.felk.cvut.cz/leaderboards/