# POINTER: A Real-Time Framework for Dynamic, Object-Centered Augmented Reality

RON KIBEL, University of California, Santa Barbara, USA
TOBIAS HÖLLERER and MISHA SRA, University of California, Santa Barbara, USA


Fig. 1. The POINTER end-to-end workflow. (1) The user begins at the main inventory screen and initiates a scan. (2) During scanning, the user captures sparse, multi-view images of a novel object to generate semantic and geometric priors. (3) During inference, the system robustly estimates the real-time pose of a target object and anchors interactive digital content directly to it.

Current Extended Reality (XR) systems predominantly rely on spatial anchors that bind digital context to static environmental features, such as walls and floors. While effective for stationary scenes, this assumption breaks down in dynamic contexts where objects are actively manipulated or moved. We present **POINTER**, a real-time framework that anchors digital content to the semantic and geometric identity of unique physical objects. Using a hybrid edge-server architecture, POINTER "retrofits" intelligence onto arbitrary objects via sparse RGB views. By combining Vision Language Models (VLMs) for semantic extraction with a 6 Degrees of Freedom (6DoF) pose estimation pipeline, the system tracks dynamic objects in real-time, allowing them to serve as both mobile canvases for digital overlays and tangible controllers, where rotation and movement drive digital inputs. We demonstrate these capabilities through a series of live scenarios, ranging from dynamic instructional guides to tangible smart home control.

Authors' Contact Information: Ron Kibel, rkibel@ucsb.edu, University of California, Santa Barbara, Santa Barbara, California, USA; Tobias Höllerer, tholleren@ucsb.edu; Misha Sra, sra@ucsb.edu, University of California, Santa Barbara, Santa Barbara, California, USA.


Manuscript submitted to ACM

CCS Concepts: • **Human-centered computing** → **Mixed / augmented reality**; • **Computing methodologies** → *Visual content-based indexing and retrieval*; *Tracking*.

Additional Key Words and Phrases: Object-Centric Interaction, Tangible Augmented Reality, Generalizable Pose Estimation, Real-time tracking

**ACM Reference Format:**
Ron Kibel, Tobias Höllerer, and Misha Sra. 2026. POINTER: A Real-Time Framework for Dynamic, Object-Centered Augmented Reality. In *Proceedings of 31st International Conference on Intelligent User Interfaces (IUI '26)*. ACM, Paphos, Cyprus, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

We are witnessing a fundamental shift in personal computing toward Pervasive Augmented Reality [26]. Driven by the dual maturation of high-fidelity wearable AR hardware and live multimodal AI agents, this evolution envisions computing not as a discrete task, but as an ambient layer embedded within our physical environment. In this future, digital interaction follows the intuitive physics of the real world.

However, current XR systems remain somewhat detached from this type of reality. Built predominantly on "top-down" SLAM (Simultaneous Localization and Mapping) pipelines, modern XR devices excel at understanding *spaces*—the rigid geometry of walls and floors—but remain surprisingly blind to *things* within them. To a modern headset, a book is indistinguishable from the nightstand on which it sits; it is merely a patch of geometry. This approach creates a brittle user experience: the moment a user interacts with the environment, the spatial anchor fails.

To move beyond this configuration, we consider a transition to a "bottom-up" architecture that rebuilds the physical world not as a static stage, but as a dynamic database of indexed entities. In this object-centric model, everyday items possess a persistent digital identity that travels with them. This vision is now technically feasible due to the convergence of two specific AI breakthroughs. First, the arrival of Vision-Language Models [2, 14], and open-vocabulary segmentation [11, 30] has effectively solved the recognition problem, allowing systems to identify arbitrary objects without prior training. Second, advances in real-time 6DoF pose estimation [29] have addressed the dynamic tracking problem, localizing objects in 3D space even as they move through a changing environment.

In this work, we present **POINTER**, a framework that synthesizes semantic intelligence with robust geometric tracking. By coupling the reasoning capabilities of VLMs with state-of-the-art pose estimation, POINTER "retrofits" intelligence onto arbitrary dynamic objects, transforming them from passive matter into active, tracked controllers. We demonstrate how this shift enables a more resilient class of interfaces, from persistent instructional guides to tangible inputs that let users manipulate their smart home.

## 2 Related Work

### 2.1 Object-Centric Interaction

The ambition to couple digital information with physical matter was formalized in Ishii and Ullmer's seminal *Tangible Bits* [8]. Early implementations like *metaDESK* [27] introduced "phicons"—physical handles for digital data—but were constrained to instrumented tabletops. The shift to Augmented Reality moved these interactions into 3D space, initially via fiducial markers [9, 23], but these markers still break the "seamless" vision of Tangible User Interfaces by requiring users to interact with codes rather than the natural objects themselves.

As tracking technology matured, research shifted toward interacting with unmodified physical objects. Heun et al.'s *Reality Editor* [7] demonstrated the power of "patching" digital functions onto physical switches, though it required manual authoring. Recent work has sought to lower this authoring barrier through "ad-hoc" binding: systems like *Teachable Reality* [18] allow users to define interactions by demonstrating visual states, while *XR-Objects* [5] and *OmniActions* [13] use Vision-Language Models to infer affordances directly from an object's identity. POINTER extends this lineage by synthesizing both semantic understanding (via VLM) with geometric registration to produce a more complete object understanding.

### 2.2 Model-Free Pose Estimation

Improving the fidelity of these tangible interactions requires precise 6DoF object pose estimation. While Model-Based methods often rely on predefined CAD models or category priors, Model-Free approaches, which require only a set of reference images, are essential for open-world, zero-shot generalization to novel objects.

Early model-free methods established 2D-3D correspondences by reconstructing objects from video or dense reference images. For example, *Gen6D* [16] retrieves the nearest reference images from a dense-view database for pose refinement. Moreover, the *OnePose* family [6, 25] uses Structure-from-Motion (SfM) [24] to build point clouds and find correspondences between the reconstruction and the query image. However, these are hindered by the dependency on dense-view captures, which are impractical for spontaneous real-world interaction.

To reduce these requirements, sparse-view works [4, 12, 17, 19, 21] focus on matching a query image to a single or small set of reference images. This improves efficiency and practicality, but reduces robustness. As such, several methods implement pose refinement by representing the object in a more complete space: *LatentFusion* [22] uses a latent space representation, *GS-Pose* [3] uses 3D Gaussians [10], and *GigaPose* [20], *HIPPo* [15], and *SceneComplete* [1] use synthesized meshes generated by image-to-3D diffusion models. Nevertheless, despite their increased robustness, these approaches, alongside foundation models like *FoundationPose* [28], reintroduce real-time latency bottlenecks due to optimization-heavy "render-and-refine" loops or iterative generative synthesis.

To bridge the gap between real-time performance and robust pose estimation, we use *BoxDreamer* [29], as it serves as an ideal middle ground. *BoxDreamer* uses a transformer-based point synthesizer to predict 2D corners and reproject a 3D bounding box, achieving state-of-the-art accuracy in sparse-view settings while maintaining real-time inference rates.

## 3 POINTER

POINTER (Persistent Object-anchored INteractions and Tagging for Enriched Reality) uses a "Scan-and-Play" workflow to transform dynamic everyday physical objects into user interface elements. While the system relies on a heavy computer vision backbone, our primary contribution is the interaction itself. We specifically target mobile devices (iOS) rather than Head-Mounted Displays (HMDs) to maximize interface accessibility. Along the same lines, we lower the barrier to entry by scanning novel objects using only RGB data, without inputs of depth or 3D texture.

The user workflow is divided into two phases: *scanning* and *inference*. To "store" an object, the user initiates a scan by capturing a few photos of the object from various angles. Unlike traditional SLAM, the system performs open-vocabulary discovery, automatically (or manually) identifying the object and its potential affordances. Once registered, the object becomes a persistent anchor.

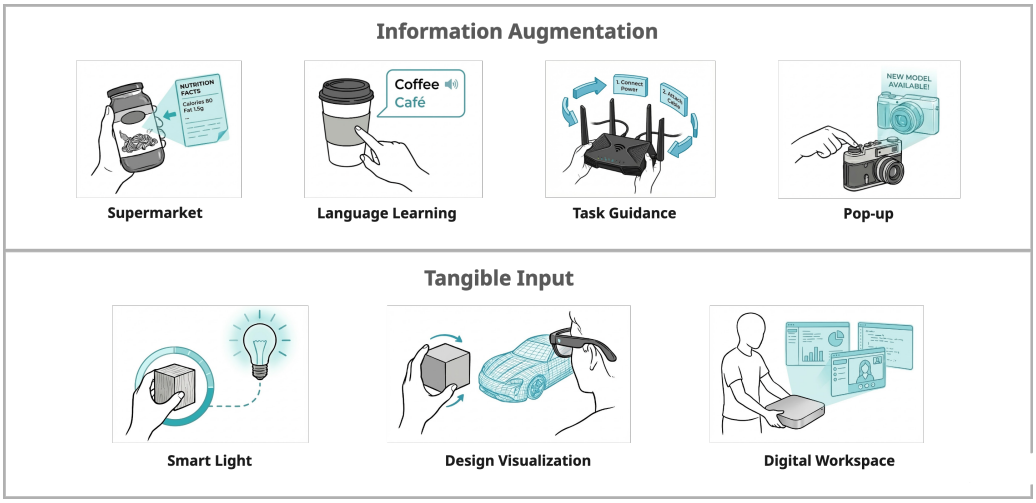


Fig. 2. Overview of the POINTER demonstration scenarios. The top row illustrates **Information Augmentation**, where digital content reveals context inherently tied to the object's identity. *Supermarket* displays nutrition facts, ingredients, and allergens overlaid on food items; *Language Learning* translates object names to a target language; *Task Guidance* overlays step-by-step instructions for complex machinery or electronics; *Pop-up* is an advertising scenario where advertisers attach "new version" advertisements to previous product generations. The bottom row illustrates **Tangible Input**, where the physical object is used as a controller to manipulate external systems. *Smart Light* uses a physical object to control a smart lightbulb; *Design Visualization* is a visual design tool to inspect a virtual model with a physical prop; *Digital Workspace* is a portable office setup that anchors virtual screens to a physical device.

### 3.1 Inference Modes

Once an object is registered, POINTER supports a range of interactions governed by the user's intent. We categorize these into two primary inference modes based on how the user conceptually frames the object.

*3.1.1 Information Augmentation (Object as Canvas).* In this mode, we use the semantic identity of the object to enable situated querying. The physical object retains its identity, but its surface becomes a handle for metadata and instructions. Importantly, pose estimation binds the content based on the user's physical perspective and proximity, making the binding more interactive.

*3.1.2 Tangible Input (Object as Controller).* In this mode, the object is appropriated as an input device for external systems. The object's specific identity is secondary to its physical affordances (shape, weight, graspability). This leverages the natural haptics of the object, which provide the user with passive tactile feedback that screen-based interfaces lack.

## 4 Conclusion

We have presented POINTER, a framework that moves Augmented Reality beyond static spatial anchors to focus on the dynamic objects that populate our lives. By coupling the semantic flexibility of VLMs with the geometric fidelity of pose estimation, we enable a new class of "retrofit" interfaces that grant digital capabilities to physical matter with only a set of sparse RGB images. Our work explores the potential of this object-centric approach to support a class of more resilient and diverse interactions.

## Generative AI Disclosure

The conceptual illustrations presented in Fig. 2 were produced with the assistance of generative AI tools. These images are used strictly for the purpose of visualizing the target application scenarios and do not represent actual system outputs or raw data.

## References


[1] Aditya Agarwal, Gaurav Singh, Bipasha Sen, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2025. SceneComplete: Open-World 3D Scene Completion in Cluttered Real World Environments for Robot Manipulation. arXiv:2410.23643 [cs.RO] https://arxiv.org/abs/2410.23643

[2] Jean-Baptiste Alayrac et al. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35. 23716–23736.

[3] Y. Cai and Others. 2024. GS-Pose: Gaussian Splatting for 6DoF Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. 2024. Open-vocabulary object 6D pose estimation. arXiv:2312.00690 [cs.CV] https://arxiv.org/abs/2312.00690

[5] E. Dogan and Others. 2024. XR-Objects: Open-Vocabulary Object-Centric Interactions in Extended Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. ACM.

[6] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. 2023. OnePose++: Keypoint-Free One-Shot Object Pose Estimation without CAD Models. arXiv:2301.07673 [cs.CV] https://arxiv.org/abs/2301.07673

[7] Valentin Heun, James Hobin, and Pattie Maes. 2013. The Reality Editor: Generative Augmented Reality. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM.

[8] Hiroshi Ishii and Brygg Ullmer. 1997. Tangible bits: towards seamless interfaces between people, bits and atoms. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*. ACM, 234–241.

[9] Hirokazu Kato and Mark Billinghurst. 1999. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality (IWAR '99)*. IEEE, 85–94.

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. arXiv:2308.04079 [cs.GR] https://arxiv.org/abs/2308.04079

[11] Alexander Kirillov et al. 2023. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[12] Taeyeop Lee et al. 2025. Any6D: Model-free 6D Pose Estimation of Novel Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Y. Li and Others. 2024. OmniActions: Predicting Physical Follow-up Actions in VR. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM.

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[15] Yibo Liu, Zhaodong Jiang, Binbin Xu, and Jinjun Shan. 2025. HIPPo: Harnessing Image-to-3D Priors for Model-free Zero-shot 6D Pose Estimation. *IEEE Robotics and Automation Letters (RA-L)* (2025).

[16] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. 2022. Gen6D: Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images. In *European Conference on Computer Vision (ECCV)*.

[17] Luqing Luo, Shichu Sun, Jiangang Yang, Linfang Zheng, Jinwei Du, and Jian Liu. 2024. Object Gaussian for Monocular 6D Pose Estimation from Sparse Views. arXiv:2409.02581 [cs.CV] https://arxiv.org/abs/2409.02581

[18] D. Monteiro and Others. 2023. Teachable Reality: Prototyping Tangible AR Interactions with Everyday Objects. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM.

[19] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. 2024. NOPE: Novel Object Pose Estimation from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. 2024. GigaPose: Fast and Robust Novel Object Pose Estimation via One-Step Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Panwang Pan, Zhiwen Fan, Brandon Y. Feng, Peihao Wang, Chenxin Li, and Zhangyang Wang. 2023. Learning to Estimate 6DoF Pose from Limited Data: A Few-Shot, Generalizable Approach using RGB Images. arXiv:2306.07598 [cs.CV] https://arxiv.org/abs/2306.07598

[22] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. 2020. LatentFusion: End-to-End Differentiable Reconstruction and Rendering for Unseen Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Jun Rekimoto and Katashi Nagao. 1995. The world through the computer: Computer augmented interaction with real world environments. In *Proceedings of the 8th Annual ACM Symposium on User Interface Software and Technology (UIST '95)*. 29–36.

[24] Johannes L. Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113. doi:10.1109/CVPR.2016.445

[25] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongwei Zhao, Guofeng Zhang, and Xiaowei Zhou. 2022. OnePose: One-Shot Object Pose Estimation without CAD Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[26] Ryo Suzuki, Mar Gonzalez-Franco, Misha Sra, and David Lindlbauer. 2025. Everyday AR through AI-in-the-Loop. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (Workshop on Everyday AR)*. arXiv:2412.12681.

[27] Brygg Ullmer and Hiroshi Ishii. 1997. The metaDESK: models and prototypes for tangible user interfaces. In *Proceedings of the 10th Annual ACM Symposium on User Interface Software and Technology (UIST '97)*. ACM, 223–232.

[28] Bowen Wen et al. 2024. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[29] Z. Yu and Others. 2025. BoxDreamer: Lifting Objects to 3D with Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '25)*.

[30] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. 2023. FastSAM: Alternative Segment Anything Model for Real-time Applications. *arXiv preprint arXiv:2306.12156* (2023).