Policy-based Foveated Imaging and Perception

HOWARD XIAO, Stanford University, USA
JAN ACKERMANN, Stanford University, USA
BOYANG DENG, Stanford University, USA
GORDON WETZSTEIN, Stanford University, USA

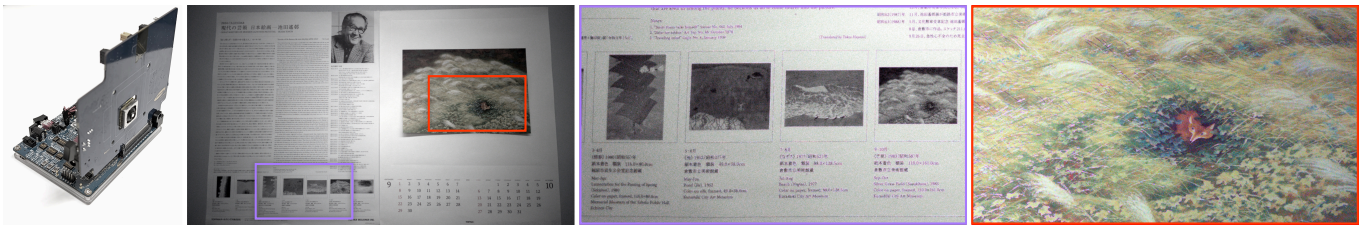

Fig. 1. Emerging sensors, such as our prototype based on Samsung's ISOCELL platform (left), provide hundreds of megapixels of resolution, which then fuel modern perception models in imaging, vision, and robotics applications. The massive amount of raw pixel data, however, is extremely challenging to be read out at high frame rates or processed efficiently with downstream perception networks. To address this challenge, we develop a policy-based foveated imaging framework that operates in real time and decides where and when to sample full-resolution regions of interest (center right and right) in a task-specific manner based on low-resolution full-field-of-view context frames (center left) as well as past observations.

Ultra-high-resolution image sensors offer the potential to capture fine spatial details critical for many visual perception tasks, but acquiring and processing all pixels at full resolution is often infeasible under realistic bandwidth, latency, and power constraints. Existing approaches address this challenge through acquisition strategies such as spatial or temporal downsampling, which irrevocably discard information before task relevance can be assessed. In this work, we introduce a real-time, predictive, and task-aware foveated imaging system that operates directly at image acquisition time. Leveraging emerging dual-stream sensor architectures, our method dynamically allocates limited pixel bandwidth to task-relevant regions of interest while maintaining a low-resolution global context. We formulate foveated acquisition as a sensor attention policy–learning problem, in which past observations guide actions that determine future measurements, closing the perception–acquisition loop. Through extensive simulation across multiple perception tasks, we demonstrate that our approach achieves high task performance under strict pixel budgets and significantly outperforms relevant baselines operating at the same bandwidth. We further validate our system on a 200-megapixel dual-stream sensor, capturing real-world videos under realistic bandwidth and latency constraints, demonstrating the practical feasibility of task-driven, acquisition-time foveated imaging. Our project website is at https://howardxiao.ca/foveated/.

CCS Concepts: • **Computing methodologies → Image and video acquisition**.

Additional Key Words and Phrases: computational photography, computational imaging, foveated imaging

Authors' Contact Information: Howard Xiao, Stanford University, USA; Jan Ackermann, Stanford University, USA; Boyang Deng, Stanford University, USA; Gordon Wetzstein, Stanford University, USA.

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA

ACM ISBN 979-8-4007-2554-8/2026/07
https://doi.org/10.1145/3799902.3811218

ACM Reference Format:
Howard Xiao, Jan Ackermann, Boyang Deng, and Gordon Wetzstein. 2026. Policy-based Foveated Imaging and Perception. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3799902.3811218

1 Introduction

Recent advances in image sensor technology enable the capture of ultra-high-resolution images and videos. Commercial sensors beyond 200 megapixels (MP) are widely available [Choi et al. 2023], with 400 MP prototypes in development [Canon Inc. 2025]. Achieving such resolution requires pixel sizes below 0.6 μm and significantly increases readout bandwidth and downstream processing demands. As a result, under realistic bandwidth, latency, or power constraints, imaging systems cannot afford to acquire, transmit, or process all pixels at full resolution, making selective acquisition essential.

Ultra-high-resolution imagery is increasingly important not only for visual quality, but also for downstream video perception tasks such as object tracking, text recognition, and robotic manipulation. These applications often rely on subtle visual cues—fast-moving objects, small text, or fine-grained surface textures—that are lost at lower resolutions. At the same time, the cost of acquiring and processing ultra-high-resolution video grows prohibitively large as resolution scales. Bandwidth limitations in sensor interfaces, sensor readout, and memory access, along with the quadratic scaling of modern transformer-based perception models with input resolution, further exacerbate this challenge, particularly on edge devices such as augmented-reality glasses, drones, autonomous vehicles, and biomedical systems.

This gap between sensing capability and system constraints raises a fundamental question: *which pixels should be acquired, and when?* Existing systems address this challenge through coarse, task-agnostic

spatio-temporal trade-offs, sacrificing either spatial detail or temporal fidelity. Sensors either downsample spatially—through pixel binning or subsampling—to maintain high frame rates, or reduce temporal resolution to preserve spatial detail. While effective at limiting data transmission, these strategies indiscriminately discard high-frequency information that may be critical for perception. Fig. 2 illustrates this issue for three different downstream applications. Once lost during acquisition, this information cannot be recovered by subsequent processing, often resulting in degraded performance on detail-critical tasks.

Emerging dual-stream sensors with hundreds of millions of pixels support multiple streams of video data to be read out simultaneously, including low-resolution full-field-of-view frames as well as much smaller but full-resolution regions of interest (ROIs) with a dynamically programmable location in the image [Samsung Electronics Co., Ltd. 2025]. Leveraging these emerging hardware capabilities, in this work we develop real-time, predictive, and task-aware foveation algorithms that address the problem of determining which pixels to acquire when, under real-world constraints. For this purpose, we formulate foveated image acquisition as a sensor attention policy-learning problem, in which past observations guide actions that directly shape future measurements, closing the perception–acquisition loop for bandwidth-efficient, task-optimal sensing. Our system uses a lightweight saliency module to propose ROI candidates and a task-driven policy to guide ROI evolution during readout. These real-time decisions minimize acquisition bandwidth while preserving task accuracy. Modeling acquisition as sequential decision making enables adaptive, task-driven scanpaths responsive to changes in both the scene and the task objective.

Our approach is motivated by human vision, which leverages the eccentricity-dependent acuity of the retina and eye movements for bandwidth-efficient sensing. Human vision dynamically moves our gaze to fixate the fovea on the most task-relevant regions of a scene. Candidate regions for fixation are typically called salient and a specific sequence of fixations or eye movements is referred to as a scanpath. Scanpaths can exhibit different characteristics, such as saccading or smooth pursuit behavior, again depending on the type of content or the task at hand [Leigh and Zee 2015].

Our work makes the following contributions:

- We introduce a real-time, policy-based, predictive foveated imaging system that dynamically directs sensor attention during image acquisition.
- We demonstrate through extensive simulation that our foveation approach maintains high task performance and significantly outperforms conventional methods in pixel-limited settings across multiple perception tasks.
- We prototype our system using a 200 MP image sensor and capture real-world videos under realistic bandwidth and latency constraints, demonstrating practical feasibility.

2 Related Work

2.1 Foveated Computer Vision

Foveated vision studies how spatial resolution can be allocated non-uniformly across the visual field in order to prioritize task-relevant regions. Early approaches relied on task-agnostic heuristics

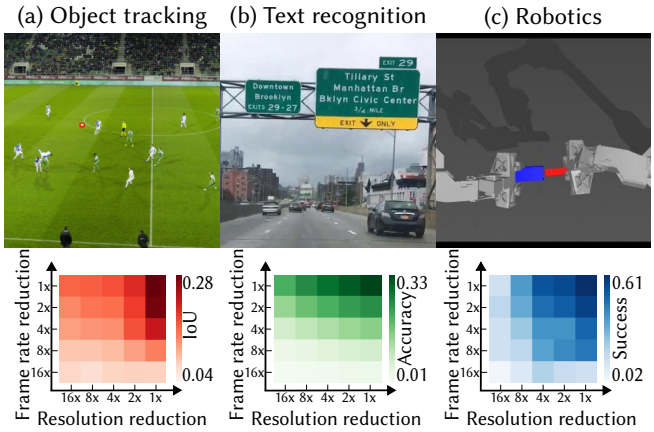

Fig. 2. **Spatio-temporal bandwidth trade-off in different tasks.** Each column shows an example perception task that benefits from both spatial and temporal detail. The top row depicts example inputs, and the bottom row illustrates how task performance varies with spatial and temporal resolution. Task performance is measured by **(a)** intersection-over-union (IoU), **(b)** correct transcription rate (Accuracy), and **(c)** task success rate (Success), consistent with Table 1.

or saliency cues to identify regions of interest (ROIs) for higher-resolution processing [Gomes et al. 2010; Itti et al. 2002; Karpathy et al. 2014; Remmelzwaal et al. 2020]. While such methods approximate aspects of human visual attention, they are not optimized for specific downstream perception objectives.

To incorporate task dependence, a large body of work has explored end-to-end learning of foveated representations jointly with perception tasks [Akbas and Eckstein 2017; Killick et al. 2023; Killick 2025]. Policy-based Recurrent Attention Models (RAM) [Haque et al. 2016; Mnih et al. 2014] formulate foveation as a sequential decision-making problem, selecting spatial glimpses conditioned on past observations. More recent works extend this paradigm to video by learning policies that select task-relevant regions from full-resolution inputs for efficient downstream processing [Shi et al. 2026; Wang et al. 2025].

Instead of spatial selection, related approaches also address bandwidth or efficiency constraints by temporally subsampling or selectively processing frames [Han et al. 2022; Xia et al. 2022].

Our work is closely related in spirit to prior foveated vision approaches, which all post-process high-resolution image and video data, but it differs in a fundamental assumption: because full-resolution frames cannot be efficiently read out from and transferred off the sensor under real-world bandwidth constraints, our method performs foveation at acquisition time, directly determining which measurements are captured.

2.2 Active Vision

Active vision studies how sensing actions can be chosen to improve perception, originally framing sensing as a means to resolve ambiguity and reduce uncertainty [Bajcsy 1988]. Subsequent work explored information-driven and decision-driven viewpoint selection, including next-best-view methods for scene understanding

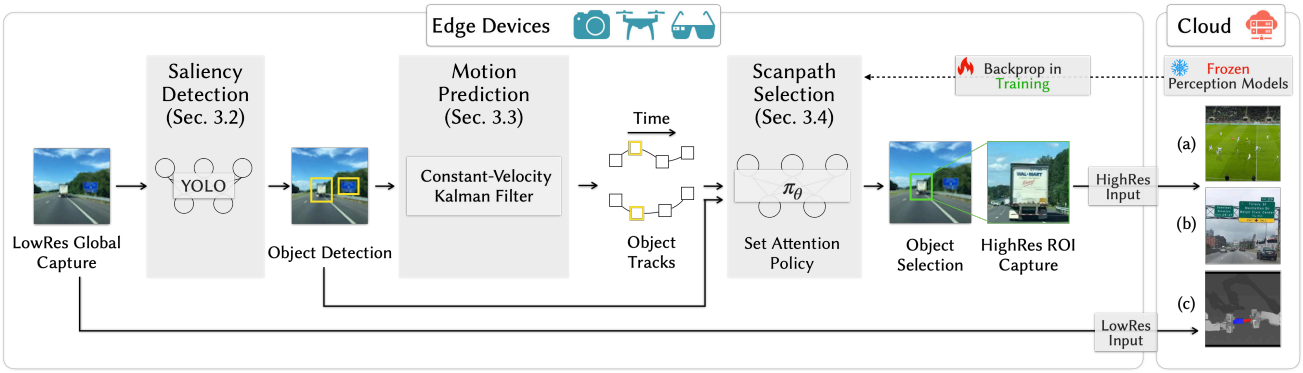

Fig. 3. **Policy-based foveated perception pipeline**. A captured low-resolution frame provides the full-field-of-view global context and is processed to determine salient candidate regions (left). Past observations then guide a per-candidate motion predictor (center left). Our sensor attention policy selects the ROI, which is then read out at full sensor resolution. Both low-resolution context frame and high-resolution ROI are streamed off the edge device and processed by the downstream perception model. At training time, our policy parameters are learned end-to-end with frozen task-specific perception models.

and 3D reconstruction [Connolly 1985; Denzler and Brown 2002; Maver and Bajcsy 2002], as well as active SLAM systems that plan camera or robot motion to gather informative observations [Sim and Roy 2005].

More recently, learning-based active vision has emerged in embodied perception and neural 3D reconstruction, where sensing actions are optimized for downstream objectives such as robotics policies or reconstruction quality [Chaplot et al. 2020; Chen et al. 2021; Chuang et al. 2025; Kerr et al. 2025].

Our approach aligns with active vision in closing the perception–action loop, but operates at a different control level: instead of selecting camera poses or viewpoints, our method actively determines the parameters of the camera's foveation mechanism during runtime.

2.3 Foveated Graphics

Foveation has been widely studied in graphics as a principled way to exploit eccentricity-dependent properties of human vision in order to reduce computation, bandwidth, or power consumption. In foveated rendering and display systems, perceptual models guide level-of-detail and sampling decisions to allocate resources preferentially near the viewer's gaze fixation [Deng et al. 2022; Krajancich et al. 2023; Mohanto et al. 2021; Wang et al. 2023].

Although these graphics systems motivate foveation as a principled trade-off between fidelity and efficiency, they typically operate at rendering or display time; in contrast, our method applies foveation during image acquisition, affecting which data is captured rather than post-processing it. We use "foveated" to refer to spatially selective, variable-resolution acquisition, generalizing the gaze-contingent interpretation common in graphics.

2.4 Foveated Sensors

Spatially-varying resolution has also been realized at the sensor level through multi-aperture and wide-angle lens designs [Carles et al. 2016; Kuniyoshi et al. 1995] and event-based sensors [Serrano-Gotarredona et al. 2022]. The former provide fixed foveation profiles determined by the optics, while the latter produce an output modality that differs from what downstream perception models expect. Our work targets dual-stream image sensors with programmable ROIs, and develops a sensor attention policy that dynamically allocates high resolution during acquisition.

3 Method

3.1 Problem Formulation

We consider video perception under a strict pixel throughput budget, where an image sensor must dynamically decide *where* and *at what resolution* to acquire visual information in order to maximize downstream task performance. Assume that the full-resolution video is $\mathbf{v}$ with $\mathbf{v}^{(k)}$ denoting frame k, then the sensor observation at frame k, $\mathbf{o}^{(k)}$, can be defined as:

$$\mathbf{o}^{(k)} = \mathcal{D}_{\boldsymbol{\phi}^{(k)}}\left(C_{\boldsymbol{\psi}^{(k)}}\left(\mathbf{v}^{(\mathcal{S}_{\boldsymbol{\varphi}}(k))}\right)\right) \tag{1}$$

where we define $\mathcal{D}$ as the spatial downsampling operator with parameters $\boldsymbol{\phi}^{(k)} = \{s_x{}^{(k)}, s_y^{(k)}\}$, where $0 < s_x^{(k)}, s_y^{(k)} \leq 1$ represent the spatial pixel resolution reduction factors in x and y directions. Assuming rectangular crops, we denote C as the frame cropping operator with parameters $\boldsymbol{\psi}^{(k)} = \{x^{(k)}, y^{(k)}, w^{(k)}, h^{(k)}\}$ with the top-left corner $(x^{(k)}, y^{(k)})$, cropping width $w^{(k)}$, and cropping height $h^{(k)}$ in pixel space. We further define $\mathcal{S}$ as the temporal skipping operator with parameters $\boldsymbol{\varphi} = \{t_s, t_o\}$, where $t_s \in \mathbb{N}, t_s \geq 1$ represents the frame skipping stride and $t_o \in \mathbb{N}, t_o \geq 1$ represents the frame offset. Here t_s, t_o are independent of the frame index k and $\mathcal{S}_{\boldsymbol{\varphi}}(k) = t_s \cdot k + t_o$ for each k.

In this case, $\mathbf{o}^{(k)}$ is parameterized by sensor attention variables $\mathbf{a}^{(k)} = \{\boldsymbol{\phi}^{(k)}, \boldsymbol{\psi}^{(k)}, \boldsymbol{\varphi}\}$ as defined in Eq. (1). Given an observation horizon of the past T_o frames, our goal is to predict a sequence of future sensor attentions over a prediction horizon T_p:

$$\mathbf{a}^{(k:k+T_p)} = \pi_\theta\left(\mathbf{o}^{(k-T_o:k)}, \mathbf{c}\right), \tag{2}$$

where π_θ denotes a task-conditioned sensor attention policy with parameters θ, and $\mathbf{c}$ encodes optional task-specific conditioning,

such as language instructions or visual prompts. The predicted actions directly determine future observations, closing the perception–acquisition loop. In our setting, dual-stream sensors capture at each frame k a low-resolution global context frame $\mathbf{o}_g^{(k)}$ with fixed downsampling $\boldsymbol{\phi}_g^{(k)} < \mathbf{1}$ over the full frame, and a full-resolution ROI crop with $\boldsymbol{\phi}^{(k)} = \mathbf{1}$ and dynamic crop parameters $\boldsymbol{\psi}^{(k)}$. Therefore, the sensor attention reduces to $\mathbf{a}^{(k)} = \{\boldsymbol{\psi}^{(k)}\}$.

Rather than learning π_θ end-to-end from raw pixels, we decompose the problem into three lightweight, interpretable components: (i) a saliency detector, (ii) a motion model, and (iii) a scanpath selection policy (Fig. 3). This modular design enables real-time inference on edge hardware and avoids the instability and latency of monolithic policies.

3.2 Saliency Detection from Low-Resolution Context

We employ a fast, YOLO-style [Redmon et al. 2016] saliency detector fine-tuned for each downstream task, operating only on context frames $\mathbf{o}_g^{(k)}$. This architecture is chosen for its favorable accuracy–latency trade-off. The detector outputs a set of $M^{(k)}$ object hypotheses:

$$\mathcal{B}^{(k)} = \{(\mathbf{b}_i^{(k)}, \mathbf{f}_i^{(k)}, \ell_i^{(k)}, c_i^{(k)})\}_{i=1}^{M^{(k)}}, \tag{3}$$

where $\mathbf{b}_i^{(k)} = (x_i^{(k)}, y_i^{(k)}, w_i^{(k)}, h_i^{(k)})$ is a bounding box in image coordinates, $\mathbf{f}_i^{(k)}$ is a learned object appearance embedding, $\ell_i^{(k)}$ is the predicted class label, and $c_i^{(k)}$ is the detection confidence score. Operating exclusively on low-resolution frames ensures minimal acquisition and compute overhead. Using a detector optimized for real-time multi-object localization allows us to efficiently extract global scene structure and object hypotheses under strict runtime constraints.

3.3 Motion Prediction

To anticipate future object locations at acquisition time, we associate detections across past global frames using the Hungarian matching algorithm [Kuhn 1955] and estimate object motion. This technique is commonly used in multi-object tracking-by-detection algorithms such as SORT [Bewley et al. 2016] and ByteTrack [Zhang et al. 2022] and is favored for its real-time performance. Although motion can be highly non-linear over long horizons in some video perception tasks such as object tracking, our setting requires only short-horizon prediction with frequent receding-horizon replanning, so a simple constant-velocity model provides a sufficiently accurate and low-latency approximation.

For each detected object i, we maintain a state vector $\mathbf{s}_i^{(k)}$ consisting of its bounding box center and velocity. We use a constant-velocity Kalman Filter to propagate this state forward:

$$\hat{\mathbf{b}}_i^{(k+\tau)} = \mathcal{T}_{\mathrm{KF}}\big(\mathbf{s}_i^{(k)}, \tau\big), \quad \tau = 1, \ldots, T_p, \tag{4}$$

yielding predicted bounding boxes for the next T_p frames, which provide the candidate crop parameters $\boldsymbol{\psi}^{(k)}$ in Eq. (1). This explicit motion model enables low-cost temporal extrapolation and allows the scanpath selection policy to reason over predicted object trajectories while relying on frequent replanning to adapt to rapid motion, occlusions, and interaction dynamics.

3.4 Scanpath Selection Policy

The scanpath selection policy predicts *which objects to foveate and when*. Rather than predicting continuous ROI parameters directly, the policy outputs a discrete scanpath over detected objects from the saliency detector, which is later converted into ROI parameters $\boldsymbol{\psi}^{(k)}$.

Object tokens. For each object i, we construct a token $\mathbf{z}_i^{(k)}$ by concatenating three components:

$$\mathbf{z}_i^{(k)} = \big[\ \underbrace{\mathbf{r}_i^{(k)}}_{\text{ROI features}} \ \|\ \underbrace{\mathbf{g}_i^{(k)}}_{\text{global features}} \ \|\ \underbrace{\mathbf{d}_i^{(k)}}_{\text{detection \& motion features}}\ \big]. \tag{5}$$

Here, $\mathbf{r}_i^{(k)}$ encodes the high-resolution visual features of each past-foveated object i using a frozen MobileNetV3-Small visual encoder backbone specifically optimized for edge device performance [Howard et al. 2019]. We employ a separate ROI feature encoder because YOLO-style detectors are not trained to extract fine-grained, high-resolution appearance features suitable for general video perception, particularly for small or texture-sensitive objects. $\mathbf{g}_i^{(k)}$ aggregates low-resolution context features for each object i over the past T_o frames using a temporal 1D convolution network, capturing coarse scene and object context directly from the detector outputs. Finally, $\mathbf{d}_i^{(k)}$ encodes the past bounding box detections, class labels, visibility history, and predicted future boxes $\{\hat{\mathbf{b}}_i^{(k+\tau)}\}_{\tau=1}^{T_p}$ of object i.

Global reasoning and prediction. Given the set of object tokens $\{\mathbf{z}_i^{(k)}\}$ and an optional task conditioning token $\mathbf{c}$, we employ a Set Transformer encoder [Lee et al. 2019] to perform permutation-invariant global object reasoning:

$$\{\tilde{\mathbf{z}}_i^{(k)}\} = \mathrm{SetTransformer}\big(\{\mathbf{z}_i^{(k)}\} \cup \{\mathbf{c}\}\big). \tag{6}$$

Each transformed object token is passed through a lightweight multilayer perceptron (MLP) head to predict object selection logits over the next T_p frames. The logits are then normalized to output a foveation scanpath represented by a categorical distribution over objects at each future timestep $k + \tau$:

$$p_\theta(i \mid k + \tau) = \mathrm{softmax}\big(\mathrm{MLP}(\tilde{\mathbf{z}}_i^{(k)})\big), \quad \tau = 1, \ldots, T_p. \tag{7}$$

The selected object index is mapped to ROI parameters $\boldsymbol{\psi}^{(k+\tau)}$ using the corresponding predicted bounding box at each timestep. It is important to note that this formulation naturally incorporates receding-horizon control [Mayne and Michalska 1988] that allows the execution of our foveation policy for $T_a < T_p$ future actions before replanning, balancing inference latency and adaptability to changing environments.

3.5 Why a Modular Policy?

Our design deliberately separates detection, motion prediction, and foveation scanpath selection. Compared to a possible end-to-end sensor attention policy, our decomposition offers three important advantages: (i) real-time inference with predictable latency, (ii) improved stability and interpretability from component-wise training [Le et al. 2018], and (iii) the ability to swap or upgrade components independently guided by downstream perception tasks. In practice, the full pipeline runs in real time on CPUs and low-end

GPUs with receding-horizon control. It could enable acquisition-time deployment on edge devices. We provide a detailed runtime analysis in the supplemental material.

4 Evaluation and Experiments

We evaluate our foveated imaging framework on multiple video perception tasks in simulation. Our experiments are designed to answer three questions: (1) can the policy predict task-relevant regions of interest (ROIs) *before* high-resolution measurements are captured; (2) does policy-based foveated imaging improve downstream video perception under strict pixel bandwidth constraints compared to task-agnostic acquisition strategies; and (3) can such a predictive foveated imaging system be realized in practice on an ultra-high-resolution imaging platform operating under realistic latency and bandwidth limits.

Unless otherwise specified, all downstream perception models are kept frozen during evaluation to isolate the effect of the acquisition strategy. This design demonstrates that our foveated imaging framework can be layered on top of existing perception models, minimizing the need for fine-tuning or post-training.

4.1 Experimental Protocol

All methods are evaluated under explicitly controlled pixel bandwidth constraints. For a given budget, we ensure that the average pixel throughput over time is identical across all acquisition strategies, including spatial downsampling, temporal downsampling, and our policy-based foveated imaging method. The pixel budgets relative to full resolution are 1/8 for object tracking, 1/8 for scene text recognition, and 1/16 for robotic manipulation; additional results at other pixel budgets are provided in the supplemental material. The same downstream model, dataset split, and evaluation metric are used across acquisition strategies for each task. Additional implementation details, including policy architecture, training procedures, and hyperparameters, are provided in the supplemental material.

4.2 Tasks, Models, and Metrics

We evaluate three video perception tasks with different demands on spatial detail, temporal resolution, and closed-loop responsiveness.

For object tracking, we use the SoccerNet Tracking dataset [Cioppa et al. 2022], which features 1920×1080 high-resolution video clips with fast-moving targets, large camera motion, and frequent occlusions. We use MixFormerV2 [Cui et al. 2023] as the downstream tracker, which outputs a bounding box per frame. Performance is measured using Intersection over Union (IoU) against ground-truth annotations. We evaluate three tracking subjects: the soccer ball, referees, and players.

For scene text recognition, we evaluate on the RoadText-1K dataset [Reddy et al. 2020], which contains 1280×720 outdoor road-scene videos with small and sparsely distributed text regions. We use DeepSolo [Ye et al. 2023] as the downstream model, which performs joint text detection and transcription. Performance is measured using the end-to-end correct transcription rate.

For robotic manipulation, we evaluate on the Static ALOHA dataset [Zhao et al. 2023], which consists of tabletop manipulation tasks that are highly sensitive to spatial detail and temporal feedback. Experiments are conducted in simulation, with frames rendered at 640×480 resolution. We use the pretrained task-specific ALOHA Action Chunking Transformer (ACT) [Zhao et al. 2023] as the downstream model, which predicts action chunks executed by a receding-horizon controller that replans every 15 steps. Following the original benchmark definition, performance is measured by partial and complete task success rates, where partial success corresponds to achieving stable contact between the manipulated objects, and complete success requires correctly inserting one object into the other.

4.3 Acquisition Baselines

We compare our approach against task-agnostic acquisition strategies operating under the same pixel budget. Spatial downsampling uniformly reduces the spatial resolution of the full frame while preserving the original frame rate, trading spatial detail for temporal smoothness. Temporal downsampling reduces the frame rate while maintaining full spatial resolution, preserving fine details at the cost of temporal continuity. Both baselines represent common approaches used in current ultra-high-resolution sensors for video acquisition and perception. They allocate pixels uniformly and do not adapt acquisition decisions based on scene dynamics or task objectives, allowing us to isolate the benefits of our policy-based, task-guided foveated imaging approach. We further include a *GT Oracle* upper bound that bypasses all components in Secs. 3.2–3.4 and directly centers ROIs on the target's ground-truth bounding boxes under the same pixel budget.

4.4 Downstream Video Perception under Limited Pixel Budget

We evaluate whether predictive foveated imaging improves downstream task performance under a limited pixel budget compared to task-agnostic baselines in Sec. 4.3. For each task, we compare (i) full-resolution inputs, (ii) dual-stream inputs acquired using predicted high-resolution ROIs with downsampled global context (Foveated), (iii) spatially downsampled inputs, and (iv) temporally downsampled inputs, with (ii), (iii), and (iv) matched to the same total pixel bandwidth. Downstream models are fixed per task. Table 1 summarizes the results. Overall, policy-based foveated imaging consistently outperforms task-agnostic baselines and, in some cases, matches full-resolution performance while using less than one-eighth of the pixel bandwidth.

Object tracking. Objects in SoccerNet Tracking are small and fast-moving, making tracking particularly sensitive to acquisition bandwidth. The soccer ball occupies only a few pixels on average (approximately 15 at full resolution) and, after naive spatial downsampling, falls well below the effective patch size of downstream transformer-based models, leading to severely degraded localization accuracy (IoU 0.122). Temporal downsampling performs slightly better (IoU 0.148), but fails under fast motion: the ball often traverses more than 5% of the field of view in less than 10 frames, making it difficult for search-template-based trackers to reliably establish correspondences (see Figs. 4 and 5).

In contrast, our predictive foveated imaging framework tracks the ball's trajectory despite rapid motion, achieving an IoU of 0.283

Table 1. **Quantitative results of policy-based foveated perception.** We compare downstream task performance of our method against same-pixel-bandwidth downsampling baselines and include full-resolution and oracle performance with ground truth (GT) ROI selections. **Row 1:** We compare downstream soccer ball tracking Intersection-over-Union (IoU) of the baseline methods and our approach. Temporal downsampling IoU is computed only on kept frames. **Row 2:** We compare the percentage of distinct text objects correctly detected and transcribed in the road scene text recognition task. **Row 3:** We compare the partial and complete task success rate over 100 trials of the ALOHA insertion task. The GT Oracle is not applicable because no ground-truth foveation labels are available for this task. Our approach performs the best among relevant baselines with a comparable bandwidth.

Task	Metric	Full-resolution	GT Oracle	Spatial downsampling	Temporal downsampling	**Foveated (Ours)**
Object Tracking	IoU ↑	0.281	0.405	0.122	0.148	**0.283**
Text Recognition	Transcription Rate ↑	0.333	0.271	0.067	0.248	**0.264**
Robotic Manipulation	Success Rate ↑ (Complete | Partial)	0.15 | 0.61	N/A	0.10 | 0.51	0.07 | 0.30	**0.12** | **0.57**

while operating at 8× lower bandwidth and effectively matching full-resolution performance (IoU 0.281). Our method and the GT Oracle both outperform the full-resolution baseline, as passing ROI crops suppresses background distractors and improves tracking despite using fewer pixels, consistent with findings in [Zhu et al. 2018].

Beyond the soccer ball, the policy adapts its foveation behavior online by changing visual conditioning, enabling smooth pursuit of different objects—including players and referees—within the same video. Its causal, acquisition-time operation allows rapid adaptation to occlusions, abrupt motion, and potential identity switches.

Text recognition. Text in RoadText-1K appears only briefly and at varying distances as the ego vehicle moves; text on other vehicles or roadside signs may enter and exit the field of view rapidly and can be difficult to read when small or partially occluded. Under these conditions, spatial downsampling leads to a severe drop in transcription accuracy (0.067), as fine character strokes become unrecognizable. Temporal downsampling performs better (0.248), but remains unreliable because text is often readable only within a narrow temporal window that may be skipped entirely. These failure modes are illustrated in Fig. 4, where spatial downsampling blurs text beyond recognition while temporal subsampling skips the few frames in which text might be legible.

Our foveated approach achieves a transcription rate of 0.264, outperforming both bandwidth-matched baselines by preserving high-resolution detail over text regions while maintaining sufficient temporal coverage. More broadly, RoadText-1K highlights the inherent difficulty of text recognition under bandwidth constraints: multiple text instances may be simultaneously present, and limited pixel budgets require explicit decisions about where to allocate resolution. This is reflected in the gap between full-resolution performance (0.333) and the GT Oracle (0.271), which is close to our result.

Robotic manipulation. The Static ALOHA bimanual insertion task requires high dexterity and tight coordination between perception and control, as successful execution depends on precise localization of contact regions and timely visual feedback during closed-loop manipulation. Complete success is more challenging than partial success, as it requires higher precision across all task-relevant dimensions; accordingly, every instance of complete success also constitutes partial success.

Temporal downsampling severely degrades partial success (from 0.61 to 0.30), as reduced visual feedback causes the controller to overshoot actions without receiving intermediate corrective signals. Spatial downsampling also reduces partial success (to 0.51), though to a slightly lesser extent, reflecting the loss of fine spatial detail needed for accurate alignment between the robot end-effector and the manipulated objects.

In contrast, our foveated imaging framework preserves high-resolution sensing over task-relevant regions—such as the end-effectors and object interaction points—while maintaining sufficient temporal feedback. As a result, our method achieves performance close to that of full-resolution sensing (Table 1). We observe similar trends for complete success.

4.5 Ablation Studies

We conduct ablation studies to isolate the contribution of individual components in our foveated imaging framework. All ablations are evaluated on the three tasks described above. To enable comparison across heterogeneous metrics, we normalize each task's performance relative to its full-resolution performance, preserving relative degradation trends while allowing aggregation across tasks.

System-level ablations. We compare our learned, task-aware foveation policy with simpler alternatives: always-centered ROI selection and deterministic round-robin ROI scanning. Results in Table 2 show that always selecting a centered ROI performs extremely poorly for object tracking and text recognition, as task-relevant content is rarely centered in dynamic scenes. While this strategy performs moderately well for robotic manipulation—where the end-effector often remains near the image center—it fails to generalize across tasks. Round-robin scanning improves over fixed centering by ensuring spatial coverage, but remains substantially inferior to our method, particularly for soccer tracking and manipulation. In contrast, our policy-based approach consistently achieves near- or above-full-resolution normalized performance across all tasks.

Policy input feature ablations. We further analyze the contribution of individual policy inputs by selectively removing feature groups

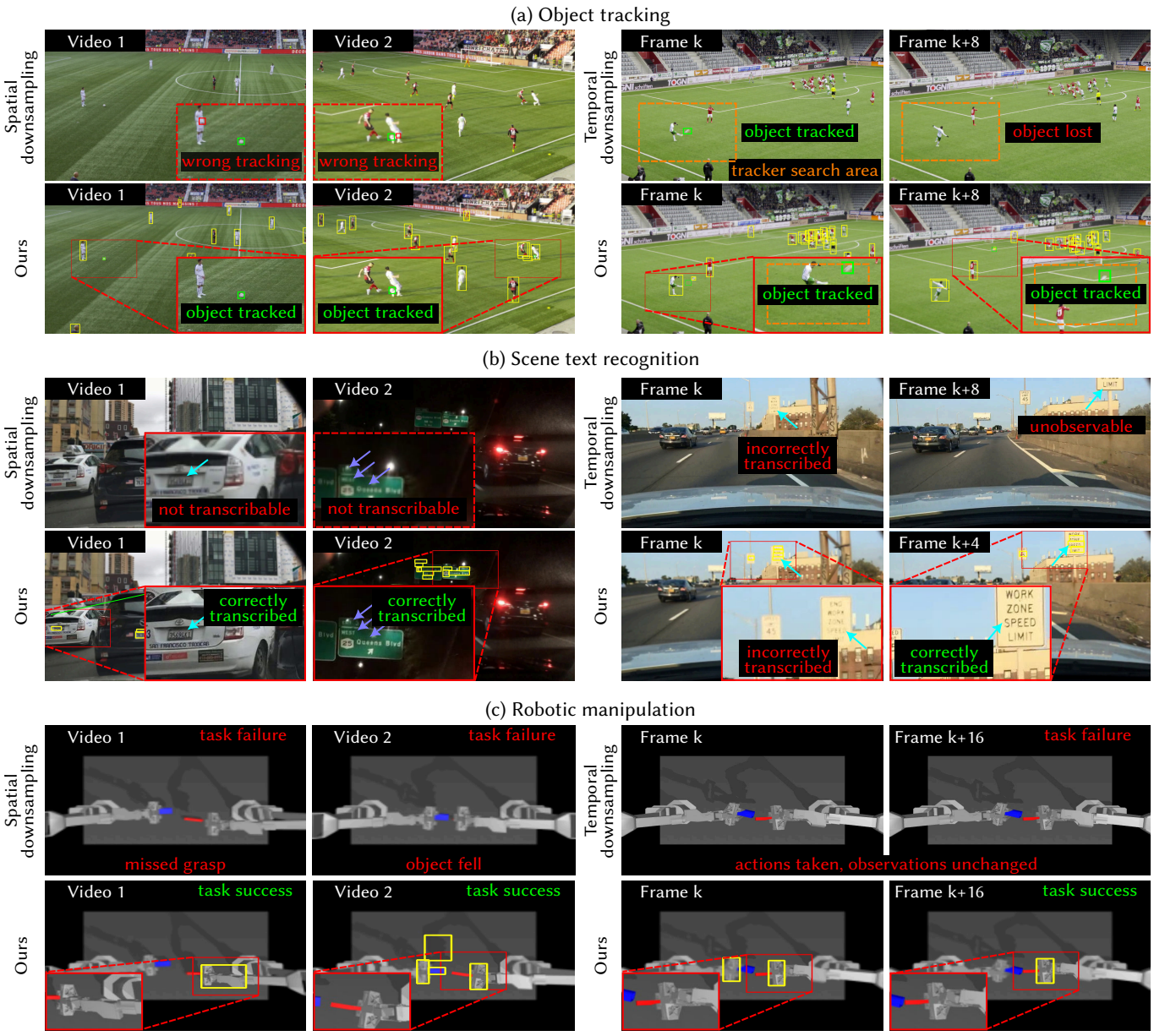

Fig. 4. **Policy-based foveated imaging and perception for simulated video tasks. Row (a):** Our foveated imaging approach correctly allocates higher resolution for pursuing objects of interest in an object tracking task. ROIs from our foveated imaging framework provide fine spatial details required to distinguish similar objects and provide fine temporal details for motion continuity, significantly improving downstream search-based tracker performance compared to task-agnostic spatio-temporal downsampling baselines. **Row (b):** Our method adapts to emerging objects and allocates fine resolution to high-frequency text regions important for the downstream scene text recognition task before frames are captured. Our foveated imaging pipeline improves the transcription rate of the text recognition task compared to naive spatio-temporal downsampling methods where texts are frequently not transcribable or missed. **Row (c):** In robotic manipulation, our method attends to important regions critical for task success while keeping low latency, ensuring the observed state reflects robot actions and significantly improving our performance against task-agnostic baselines. Additional results across more diverse scenes are provided in the supplemental material.

from the scanpath selection policy, including high-resolution ROI features, low-resolution global context features, and motion features derived from short-term trajectory prediction.

Table 3 shows that removing motion features causes a performance drop across all tasks, underscoring the importance of anticipating future object locations when making acquisition-time

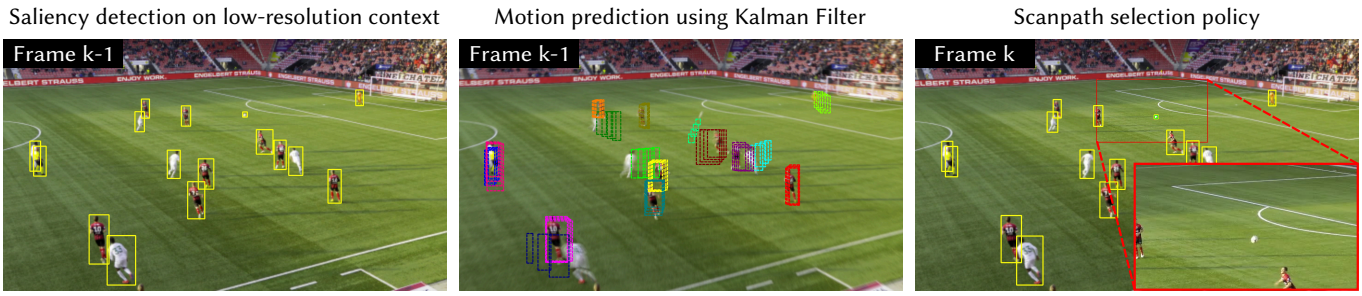

Fig. 5. **Overview of key components of our foveated imaging framework. Left:** At frame $k-1$, multiple task-relevant object locations are proposed by the saliency detector using only low-resolution context. **Middle:** We associate each object detection over the past T_o frames and predict object motion for the future T_p frames using a constant-velocity Kalman Filter. **Right:** Our scanpath selection policy uses high-resolution ROI features, global features, and detection and motion features to predict the future object scanpath at frame k from detected objects in the past. The predicted scanpath allows us to acquire future ROIs with high resolution for downstream perception tasks.

Table 2. **System-level ablation results averaged across all tasks.** Performance is normalized by full-resolution performance per task for cross-comparison. Our dynamic foveation approach outperforms naive ROI selection strategies. Normalized performance here can exceed 100% whenever ROI-based perception outperforms direct processing of the full-resolution image, which contains additional visual clutter in non-task-relevant areas.

Variant	Soccer Tracking	Text Recognition	Robotic Manipulation
Always-centered ROI selection	1.7%	3.6%	82.0%
Round-robin ROI scanning	3.5%	64.0%	72.1%
Ours	**100.7%**	**79.3%**	**93.4%**

Table 3. **Ablation of policy inputs.** Performance is normalized by full-resolution performance per task for cross-comparison. All features are necessary for the best performance; see Eq. (5).

Policy Inputs Removed	Soccer Tracking	Text Recognition	Robotic Manipulation
w/o ROI features	53.3%	78.7%	83.6%
w/o global features	99.2%	72.7%	85.2%
w/o motion features	90.0%	71.8%	90.2%
All features (ours)	**100.7%**	**79.3%**	**93.4%**

decisions. Removing global context features also leads to significant degradation, particularly for text recognition, where multiple candidate regions may be present and scene-level context is required to disambiguate where high-resolution sensing should be allocated. Removing high-resolution ROI appearance features also degrades tracking, text recognition, and manipulation performance, reflecting the importance of fine-grained visual detail.

Overall, these ablations confirm that effective acquisition-time foveation requires the combination of motion cues, global context, and high-resolution local appearance. Removing any of these components degrades performance, whereas their integration enables robust task-aware sensing under strict pixel-bandwidth constraints.

5 Evaluation on a 200 MP Foveated Imaging Prototype

To validate the practical feasibility of predictive foveated imaging, we implement our predictive foveated imaging framework on a hardware prototype built around a 200 MP Samsung ISOCELL HP2 image sensor. The sensor is mounted on a custom control board that supports dual-stream acquisition, enabling simultaneous capture of a low-resolution Full Field-of-View (FFoV) context stream and high-resolution Region-of-Interest (ROI) crops at 30 frames per second. The control board is interfaced with a host system via a Python API, which allows predicted ROI coordinates to be transmitted to the sensor for subsequent frame readout.

In our prototype configuration, the FFoV stream is captured at a resolution of 2040×1148, providing global situational awareness, while each ROI occupies one-quarter of the sensor area and is captured at 4080×2296 resolution, corresponding to a $16\times$ increase in spatial resolution relative to the FFoV stream. The resolution difference between the FFoV and the ROI stream matches our simulation setting across all video perception tasks. This dual-stream setup enables closed-loop, predictive control of sensor readout, allowing high-resolution sensing resources to be dynamically allocated to task-relevant regions during acquisition.

5.1 Hardware and Predictive Acquisition Loop

The prototype implements a predictive dual-stream acquisition loop consisting of four stages. First, in the Low-Resolution Context Capture stage, the sensor reads out an FFoV context frame at 2040×1148 resolution. Second, during Policy Inference, the host controller processes the FFoV frame using our lightweight attention policy to predict the Region-of-Interest (ROI) coordinates for the subsequent frames ($k+\tau$). Third, in the Command Transmission stage, the predicted ROI coordinates are transmitted back to the sensor controller via a Python-to-FPGA interface. Finally, in the High-Resolution ROI Capture stage, the sensor acquires the targeted ROI at 4080×2296 resolution, providing $16\times$ higher spatial detail than the FFoV stream.

Fig. 6. **Policy-based foveated imaging in real-world captures.** Under realistic bandwidth and acquisition latency constraints, our proposed method runs in real time on our 200 MP-resolution foveated imaging prototype. We demonstrate expected smooth-pursuit scanpaths for **(a)** object tracking and saccading scanpaths for **(b)** scene text recognition across diverse scenes and lighting conditions.

This predictive loop ensures that ROI selection decisions are made prior to high-resolution readout, allowing the system to manage sensor bandwidth proactively rather than reactively.

5.2 Real-World Performance and Bandwidth Efficiency

We capture dual-stream Bayer-raw video at 30 fps, with optics manually focused prior to acquisition. Despite the computational overhead of policy inference and bidirectional host–sensor communication, the system maintains a stable end-to-end throughput of 30 fps throughout extended capture sessions.

Qualitative results are shown in Fig. 1 and Fig. 6, where we demonstrate both smooth-pursuit scanpaths for object tracking and saccading scanpaths for scene text recognition. The predictive attention policy consistently directs high-resolution sensing to task-relevant regions that remain indistinguishable in the FFoV context stream. This enables recovery of fine spatial details such as object boundaries and textures under real-world lighting conditions and sensor noise.

Crucially, the system achieves this performance while reading out only 6.25% of the sensor area at full resolution per frame. This demonstrates that predictive foveated imaging provides an effective mechanism for managing the bandwidth of 200 MP-class sensors, preserving task-critical visual information while operating within realistic hardware and interface constraints.

6 Discussion

Our results suggest that predictive, policy-based foveated acquisition is a promising approach for operating ultra-high-resolution image sensors under realistic bandwidth, latency, and power constraints. By explicitly modeling the interaction between sensing and

perception, our framework allocates limited pixel budgets to task-relevant regions *before* high-resolution measurements are captured, preserving downstream performance that would otherwise degrade under conventional spatio-temporal downsampling.

Limitations and Future Work. Despite these advantages, our approach has several limitations. First, the effectiveness of predictive foveation depends on temporal coherence: tasks involving highly stochastic or instantaneous events may reduce the benefit of anticipation. Second, while our attention policy is lightweight, it introduces additional system complexity and must meet strict real-time constraints to be deployed at acquisition time. Third, our current prototype supports a limited number of ROIs per frame; extending the framework to support more flexible or hierarchical foveation patterns remains an interesting direction for future work.

More broadly, our formulation highlights foveated imaging as a systems problem that spans sensor design, learning-based control, and downstream perception models. While we focus on a specific set of video perception tasks, the proposed framework is general and could be extended to other sensing modalities, multi-camera systems, or closed-loop robotic perception pipelines.

Conclusion. Intelligent sensing moves beyond passive capture, enabling systems to decide where and how to sample based on the task at hand. Our policy-based foveation framework breaks conventional sampling trade-offs, providing a lightweight yet powerful solution for high-stakes environments. Inspired by the successes of imitation and reinforcement learning, we believe this paradigm will redefine the boundaries of intelligent data acquisition in computer vision, robotics, and beyond.

Acknowledgments

We thank Samsung for their support with the ISOCELL development kit. Howard Xiao is supported by Stanford Graduate Fellowships (SGF). We thank Hansheng Chen, Ryan Po, Kiyohiro Nakayama, and Zichun Xu for fruitful discussions. Compute resources were provided by the Marlowe cluster at the Stanford University [Kapfer et al. 2025].

References

Emre Akbas and Miguel P Eckstein. 2017. Object detection through search with a foveated visual system. *PLoS computational biology* 13, 10 (2017), e1005743.

Ruzena Bajcsy. 1988. Active perception. *Proc. IEEE* 76, 8 (1988), 966–1005.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 3464–3468.

Canon Inc. 2025. Canon develops CMOS sensor with 410 megapixels, the largest number of pixels ever achieved in a 35 mm full-frame sensor. https://global.canon/en/news/2025/20250122.html.

Guillem Carles, Shouqian Chen, Nicholas Bustin, James Downing, Duncan McCall, Andrew Wood, and Andrew R Harvey. 2016. Multi-aperture foveated imaging. *Optics Letters* 41, 8 (2016), 1869–1872.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. 2020. Learning To Explore Using Active Neural SLAM. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HklXn1BKDH

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*. 14124–14133.

Sungsoo Choi, Seungjoon Lee, Taeheon Lee, Hochul Ji, Haeyong Park, Dongmo Im, Dongchul Lee, Jinyoung Kim, Sungyong You, Jaeho Choi, et al. 2023. World smallest 200Mp CMOS image sensor with 0.56 μm pixel equipped with novel deep trench isolation structure for better sensitivity and higher CG. In *Proceedings of the Int'l Image Sensor Workshop (IISW), Crieff, UK*. 22–25.

Ian Chuang, Jinyu Zou, Andrew Lee, Dechen Gao, and Iman Soltani. 2025. Look, focus, act: Efficient and robust robot learning via human gaze and foveated vision transformers. *arXiv preprint arXiv:2507.15833* (2025).

Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. 2022. SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3491–3502.

C. I. Connolly. 1985. The determination of next best views. In *Proceedings. 1985 IEEE international conference on robotics and automation*, Vol. 2. IEEE, 432–435.

Yutao Cui, Tianhui Song, Gangshan Wu, and Limin Wang. 2023. Mixformerv2: Efficient fully transformer tracking. *Advances in neural information processing systems* 36 (2023), 58736–58751.

Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. 2022. FoV-NeRF: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3854–3864.

Joachim Denzler and Christopher M Brown. 2002. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on pattern analysis and machine intelligence* 24, 2 (2002), 145–157.

Rafael B Gomes, Renato Q Gardiman, Luiz EC Leite, Bruno M Carvalho, and Luiz MG Gonçalves. 2010. Towards real time data reduction and feature abstraction for robotics vision. *Robot Vision* (2010), 345–362.

Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. 2022. Dynamic neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2022), 7436–7456.

Albert Haque, Alexandre Alahi, and Li Fei-Fei. 2016. Recurrent attention models for depth-based person identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1229–1238.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1314–1324.

Laurent Itti, Christof Koch, and Ernst Niebur. 2002. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20, 11 (2002), 1254–1259.

Chris Kapfer, Katie Stine, Balasubramanian Narasimhan, Christof Mentzel, and Emmanuel Candès. 2025. Marlowe: Stanford's GPU-based computational instrument. doi:10.5281/zenodo.14751899

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

Justin Kerr, Kush Hari, Ethan Weber, Chung Min Kim, Brent Yi, Tyler Bonnen, Ken Goldberg, and Angjoo Kanazawa. 2025. Eye, Robot: Learning to Look to Act with a BC-RL Perception-Action Loop. *arXiv preprint arXiv:2506.10968* (2025).

George Killick, Paul Henderson, Paul Siebert, and Gerardo Aragon-Camarasa. 2023. Foveation in the era of deep learning. *arXiv preprint arXiv:2312.01450* (2023).

George William Killick. 2025. *Image Classification with Foveated Neural Networks*. Ph. D. Dissertation. University of Glasgow. https://theses.gla.ac.uk/85208/.

Brooke Krajancich, Petr Kellnhofer, and Gordon Wetzstein. 2023. Towards attention–aware foveated rendering. *ACM Transactions on Graphics (TOG)* 42, 4 (2023), 1–10.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2 (1955), 83–97.

Yasuo Kuniyoshi, Nobuyuki Kita, Kazuhide Sugimoto, Shin Nakamura, and Takashi Suehiro. 1995. A foveated wide angle lens for active vision. In *Proceedings of 1995 IEEE International Conference on Robotics and Automation*, Vol. 3. IEEE, 2982–2988.

Hoang M. Le, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and Hal Daumé III. 2018. Hierarchical Imitation and Reinforcement Learning. In *Proceedings of Machine Learning Research*.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, 3744–3753.

R John Leigh and David S Zee. 2015. *The neurology of eye movements*. Oxford university press.

Jasna Maver and Ruzena Bajcsy. 2002. Occlusions as a guide for planning the next view. *IEEE transactions on pattern analysis and machine intelligence* 15, 5 (2002), 417–433.

David Q Mayne and Hannah Michalska. 1988. Receding horizon control of nonlinear systems. In *Proceedings of the 27th IEEE Conference on Decision and Control*. IEEE, 464–465.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. *Advances in neural information processing systems* 27 (2014).

Biplab Mohanto, Abir Tanvir Islam, Enrico Gobbetti, and Oliver Staadt. 2021. An integrative view of foveated rendering. *Computers & Graphics* 101 (2021), 74–98.

Sangeeth Reddy, Minesh Mathew, Lluis Gomez, Marçal Rusinol, Dimosthenis Karatzas, and CV Jawahar. 2020. Roadtext-1k: Text detection & recognition dataset for driving videos. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11074–11080.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

Leendert A Remmelzwaal, Amit Kumar Mishra, and George FR Ellis. 2020. Human eye inspired log-polar pre-processing for neural networks. In *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 1–6.

Samsung Electronics Co., Ltd. 2025. ISOCELL HP2 | Mobile Image Sensor. https://semiconductor.samsung.com/image-sensor/mobile-image-sensor/isocell-hp2/ Accessed 2025-11-11.

Teresa Serrano-Gotarredona, Farnaz Faramarzi, and Bernabé Linares-Barranco. 2022. Electronically foveated dynamic vision sensor. In *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*. IEEE, 1–6.

Baifeng Shi, Stephanie Fu, Long Lian, Hanrong Ye, David Eigen, Aaron Reite, Boyi Li, Jan Kautz, Song Han, David M Chan, Pavlo Molchanov, Trevor Darrell, and Hongxu Yin. 2026. Attend Before Attention: Efficient and Scalable Video Understanding via Autoregressive Gazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Robert Sim and Nicholas Roy. 2005. Global a-optimal robot exploration in slam. In *Proceedings of the 2005 IEEE international conference on robotics and automation*. IEEE, 661–666.

Lili Wang, Xuehuai Shi, and Yi Liu. 2023. Foveated rendering: A state-of-the-art survey. *Computational visual media* 9, 2 (2023), 195–228.

Yulin Wang, Yang Yue, Yang Yue, Huanqian Wang, Haojun Jiang, Yizeng Han, Zanlin Ni, Yifan Pu, Minglei Shi, Rui Lu, et al. 2025. Emulating human-like adaptive vision for efficient and flexible machine visual perception. *Nature Machine Intelligence* (2025), 1–19.

Boyang Xia, Wenhao Wu, Haoran Wang, Rui Su, Dongliang He, Haosen Yang, Xiaoran Fan, and Wanli Ouyang. 2022. Nsnet: Non-saliency suppression sampler for efficient video recognition. In *European Conference on Computer Vision*. Springer, 705–723.

Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, and Dacheng Tao. 2023. DeepSolo: Let Transformer Decoder with Explicit Points Solo for Text Spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19348–19357.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2022. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*. Springer, 1–21.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705* (2023).

Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. 2018. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 101–117.