Foveated Imaging and Perception

Howard Xiao

Abstract—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.

Index Terms—Computational photography, computational imaging, foveated imaging

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 [1], with 400 MP prototypes in development [2]. Achieving such resolution requires pixel sizes below $0.6\,\mu$m 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 for downstream video perception tasks such as object tracking, text recognition, and robotic manipulation, which rely on subtle visual cues like fast motion, small text, and fine textures. But the cost of acquiring and processing such video grows prohibitively with resolution. Bandwidth limits in sensor interfaces, readout, and memory access, along with the quadratic scaling of modern transformer-based perception models, make this especially challenging on edge devices.

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 tradeoffs, 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. 1 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 simultaneous low-resolution full-field-of-view frames and smaller full-resolution regions of interest (ROIs) with programmable locations [3]. Leveraging this capability, we develop a real-time, predictive, task-aware foveation system that determines which pixels to acquire under real-world constraints. We formulate foveated acquisition as a sensor attention policy-learning problem, where past observations guide future measurements. Our system combines a lightweight saliency module with a task-driven policy that steers ROI evolution during readout, yielding adaptive scanpaths that preserve task accuracy while reducing bandwidth.

Our approach is motivated by human vision, which uses eccentricity-dependent retinal acuity and eye movements for bandwidth-efficient sensing. Human gaze fixates the fovea on task-relevant regions, producing scanpaths that vary with scene content and task [4].

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

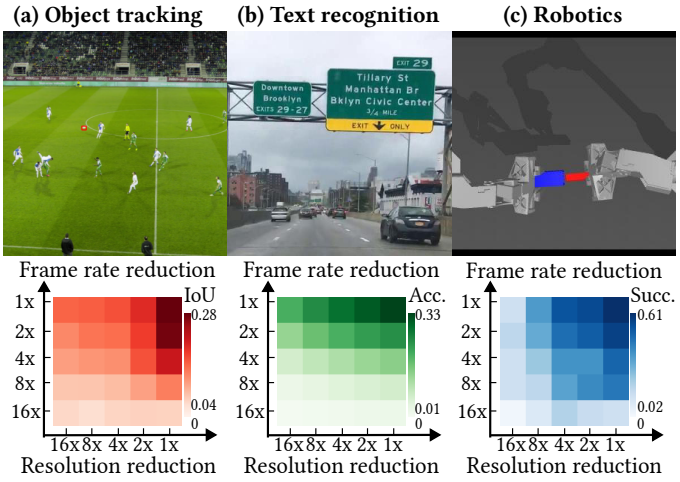

Fig. 1. **Spatio-temporal bandwidth tradeoff in different tasks.** Each column shows an example of a perception task, each benefiting from spatial and temporal detail. The top row depicts example inputs, and the bottom row illustrates how task performance varies with spatial and temporal resolution.

task-agnostic heuristics or saliency cues to identify regions of interest (ROIs) for higher-resolution processing [5], [6], [7], [8]. 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 [9], [10], [11], [12], [13], [14], [15], [16], [17]. Policy-based Recurrent Attention Models (RAM) [18], [19] formulate foveation as a sequential decision-making problem, selecting spatial glimpses conditioned on past observations. The AdaFocus family extends this paradigm to video by learning policies that select task-relevant regions frame by frame from full-resolution inputs [20], [21], [22], [23], [24].

Instead of spatial selection, related approaches also address bandwidth or efficiency constraints by temporally subsampling or selectively processing frames [25], [26], [27], [28], [29], [30], [31], [32], [33].

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 by 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 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 [34], [35], [36], [37], [38], [39], [40], [41]. These ideas have been applied across a range of graphics pipelines, including rasterization and shading, neural rendering, and immersive display systems, enabling significant efficiency gains while maintaining perceptual quality.

Although these graphics systems motivate foveation as a principled tradeoff 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 them.

3 Proposed 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 v with $v^{(k)}$ denoting frame k, then the sensor observation at frame k, $\mathbf{o}^{(k)}$, can be defined as:

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

where we define $\mathcal{D}$ as the spatial downsampling operator with parameters $\phi^{(k)} = \{s_x{}^{(k)}, s_y^{(k)}\}$, $0 < s_x^{(k)}; s_y^{(k)} \leq 1$ represent the spatial pixel resolution reduction factor in x and y directions. Assuming rectangular crops, we denote $\mathcal{C}$ as the frame cropping operator with parameters $\psi^{(k)} = \{x^{(k)}, y^{(k)}, w^{(k)}, h^{(k)}\}$ 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 $\varphi = \{t_s, t_o\}$, 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 $S_\varphi(k) = t_s \cdot k + t_o$ for each k.

In this case, $\mathbf{o}^{(k)}$ is parameterized by sensor attention variables $a^{(k)} = \{\phi^{(k)}, \psi^{(k)}, \varphi\}$ 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. 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. 2). 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

At each frame k, the sensor acquires a low-resolution global context frame $\mathbf{o}_g^{(k)}$ where $\phi_g^{(k)}$ contains downsampling factors strictly less than 1, and $\psi_g^{(k)}$ represents the entire image without cropping. $\mathbf{o}_g^{(k)}$ serves as the sole input to a fast saliency detector. We employ a YOLO-style [42] detector fine-tuned for each downstream task, chosen for its favorable accuracy–latency trade-off. The detector outputs a set of $M^{(k)}$ object hypotheses:

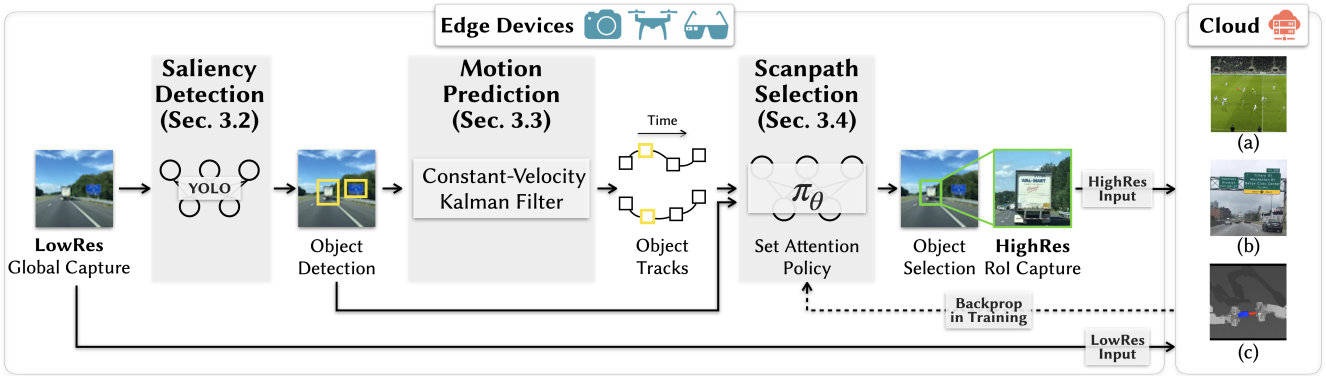

Fig. 2. **Policy-based foveated perception pipeline**. A captured low-resolution frame provides the full-field-of-view global context and is processed to determine salient candidate regions (left). Past observations then guide a per-candidate motion predictor (center left). Our sensor attention policy 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, the parameters of our policy are learned end to end with the task-specific perception model.

,

(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 [43] and estimate object motion. This technique is commonly used in multi-object tracking-by-detection algorithms such as SORT [44] and ByteTrack [45] and is favored for its real-time performance.

Motion patterns in some video perception tasks, such as soccer tracking, can be highly non-linear over long temporal horizons. For this reason, modern multi-object tracking systems often learn motion prediction and appearance modeling end to end. In our setting, however, acquisition-time foveation requires only short-horizon prediction, and decisions are replanned frequently in a receding-horizon manner. Under these conditions, a simple constant-velocity model provides a sufficiently accurate approximation while meeting strict latency constraints.

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:

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

yielding predicted bounding boxes for the next T_p frames. 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 $\psi^{(k)}$.

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

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

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 [46]. 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 $\{\hat{b}_i^{(k+\tau)}\}_{\tau=1}^{T_p}$ 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 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 $\psi^{(k+\tau)}$ using the corresponding predicted bounding box at

each timestep. It is important to note that this formulation naturally incorporates receding horizon control [47] that allows the execution of our foveation policy for $T_a < T_p$ future actions before replanning, balancing inference latency and adaptability to changing environment.

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 [48], and (iii) the ability to swap or upgrade components independently guided by downstream perception tasks. In practice, the full pipeline runs comfortably in real time on low-end GPUs and runs in real-time with receding horizon control on CPUs. It could enable acquisition-time deployment on edge devices.

4 Evaluation and Experiments

We evaluate our foveated imaging framework on multiple video perception tasks in simulation. Our experiments ask whether the policy can predict task-relevant ROIs *before* high-resolution measurements are captured, whether this improves downstream perception under strict pixel budgets compared to task-agnostic acquisition, and whether the system is practical on an ultra-high-resolution imaging platform under realistic latency and bandwidth limits.

Unless otherwise specified, all downstream perception models are kept frozen to isolate the effect of acquisition strategy and show that our framework can be layered onto existing models without fine-tuning.

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 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 supplementary 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 [49], which features 1920×1080 high-resolution video clips with fast-moving targets, large camera motion, and frequent occlusions. We use MixFormerV2 [50] 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 [51], which contains 1280×720 outdoor road-scene videos with small and sparsely distributed text regions. We use DeepSolo [52] 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 [53], 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) [53] 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 strategies under the same pixel budget. Spatial downsampling uniformly reduces frame resolution while preserving frame rate, and temporal downsampling reduces frame rate while preserving full spatial resolution. Both allocate pixels uniformly and do not adapt to scene dynamics or task objectives.

4.4 Downstream Video Perception under Limited Pixel Budget

We evaluate whether predictive foveated imaging improves downstream task performance under limited pixel budgets compared to the baselines in Sec. 4.3. For each task, we compare (i) full-resolution inputs, (ii) dual-stream inputs with predicted high-resolution ROIs and downsampled global context (Foveated), (iii) spatial downsampling, and (iv) temporal downsampling, with (ii)–(iv) matched to the same total pixel bandwidth. Table 1 shows that policy-based foveated imaging consistently outperforms task-agnostic baselines and can match full-resolution performance at less than one-eighth the 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).

In contrast, our predictive foveated imaging framework tracks the ball's trajectory despite rapid motion, achieving an IoU of 0.283 at $8\times$ lower bandwidth and effectively matching full-resolution performance (IoU 0.281). An oracle foveation baseline performs even better, indicating that

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

TABLE 1 **Quantitative results of policy-based foveated perception.** We compare downstream task performances 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. **Row 2:** We compare the percentage of distinct text objects correctly detected and transcribed in road scene text recognition task. **Row 3:** We compare the partial and full task success rate over 100 trials of the ALOHA insertion tasks. The GT oracle is not applicable as evaluation happens in open-loop where no GT foveation labels are known. Our approach performs the best among relevant baselines with a comparable bandwidth.

targeted high-resolution measurements can be more informative than uniformly processing all pixels.

The policy also adapts online by changing visual conditioning, enabling pursuit of different objects—including players and referees—within the same video.

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. The gap between full-resolution performance (0.333) and an oracle foveation baseline (0.271) reflects the difficulty of allocating limited resolution when multiple text instances may be present simultaneously.

Robotic manipulation. The Static ALOHA bimanual insertion task requires precise localization of contact regions and timely visual feedback during closed-loop manipulation. Complete success is more challenging than partial success, and 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 while maintaining sufficient temporal feedback, yielding performance close to full-resolution sensing (Table 1). We observe similar trends for full success.

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 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. 3. The predictive attention policy consistently tracks emerging and moving objects, directing 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.

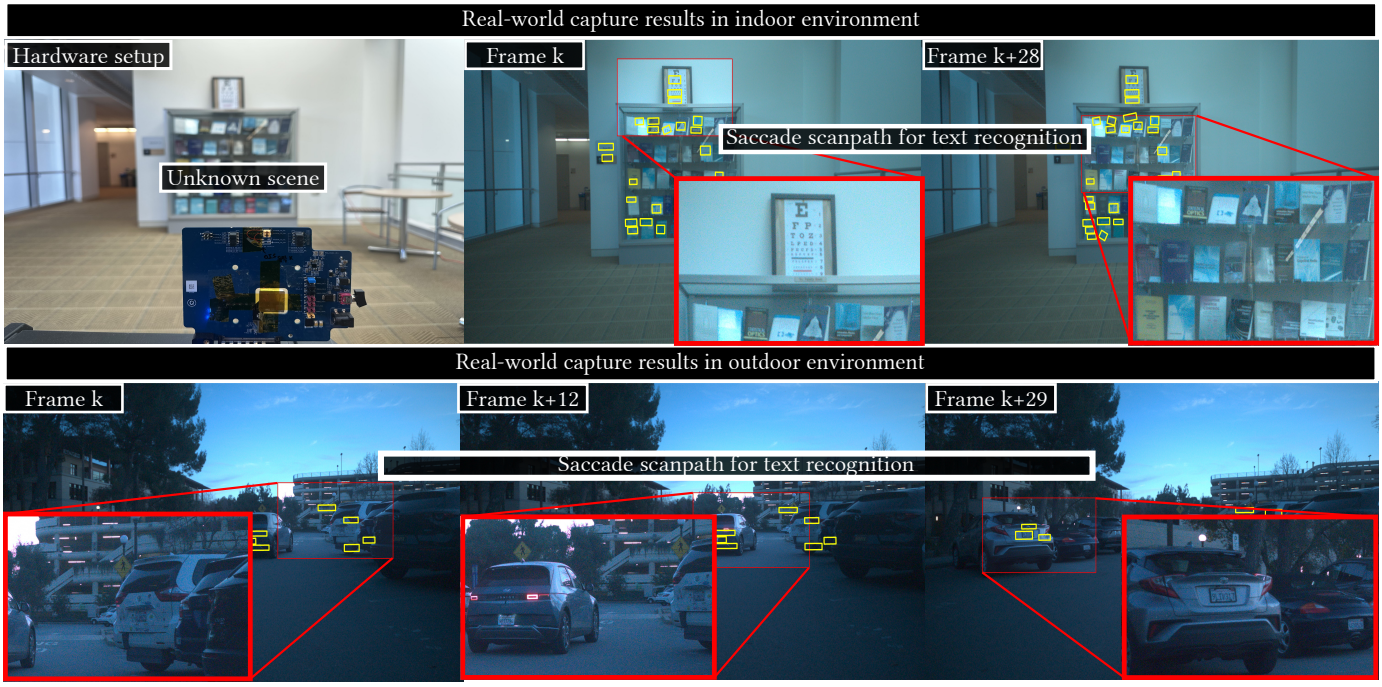

Fig. 3. **Policy-based foveated imaging in real-world captures.** Under realistic bandwidth and acquisition latency constraints, our proposed method runs in real-time on our 200 MP-resolution foveated imaging prototype. We demonstrate expected saccading scanpath for scene text recognition similar to our simulated results in both indoor (top) and outdoor (bottom) environments.

6 Discussion

Our results suggest that predictive, policy-based foveated acquisition is a practical way to operate 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 spanning sensor design, learning-based control, and downstream perception models. While we focus on a specific set of tasks, the framework could extend 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 offers a lightweight solution to conventional sampling trade-offs and points toward more capable task-driven acquisition in computer vision, robotics, and beyond.

References

[1] S. Choi, S. Lee, T. Lee, H. Ji, H. Park, D. Im, D. Lee, J. Kim, S. You, J. Choi *et al.*, "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*, 2023, pp. 22–25.

[2] Canon Inc., "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, 2025.

[3] L. Samsung Electronics Co., "Isocell hp2 — mobile image sensor," 2025, accessed 2025-11-11. [Online]. Available: https://semiconductor.samsung.com/image-sensor/mobile-image-sensor/isocell-hp2/

[4] R. J. Leigh and D. S. Zee, *The neurology of eye movements*. Oxford university press, 2015.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[6] L. A. Remmelzwaal, A. K. Mishra, and G. F. Ellis, "Human eye inspired log-polar pre-processing for neural networks," in *2020 International SAUPEC/RobMech/PRASA Conference*. IEEE, 2020, pp. 1–6.

[7] R. B. Gomes, R. Q. Gardiman, L. E. Leite, B. M. Carvalho, and L. M. Gonçalves, "Towards real time data reduction and feature abstraction for robotics vision," *Robot Vision*, pp. 345–362, 2010.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 2002.

[9] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.

[10] A. Recasens, P. Kellnhofer, S. Stent, W. Matusik, and A. Torralba, "Learning to zoom: a saliency-based sampling layer for neural networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 51–66.

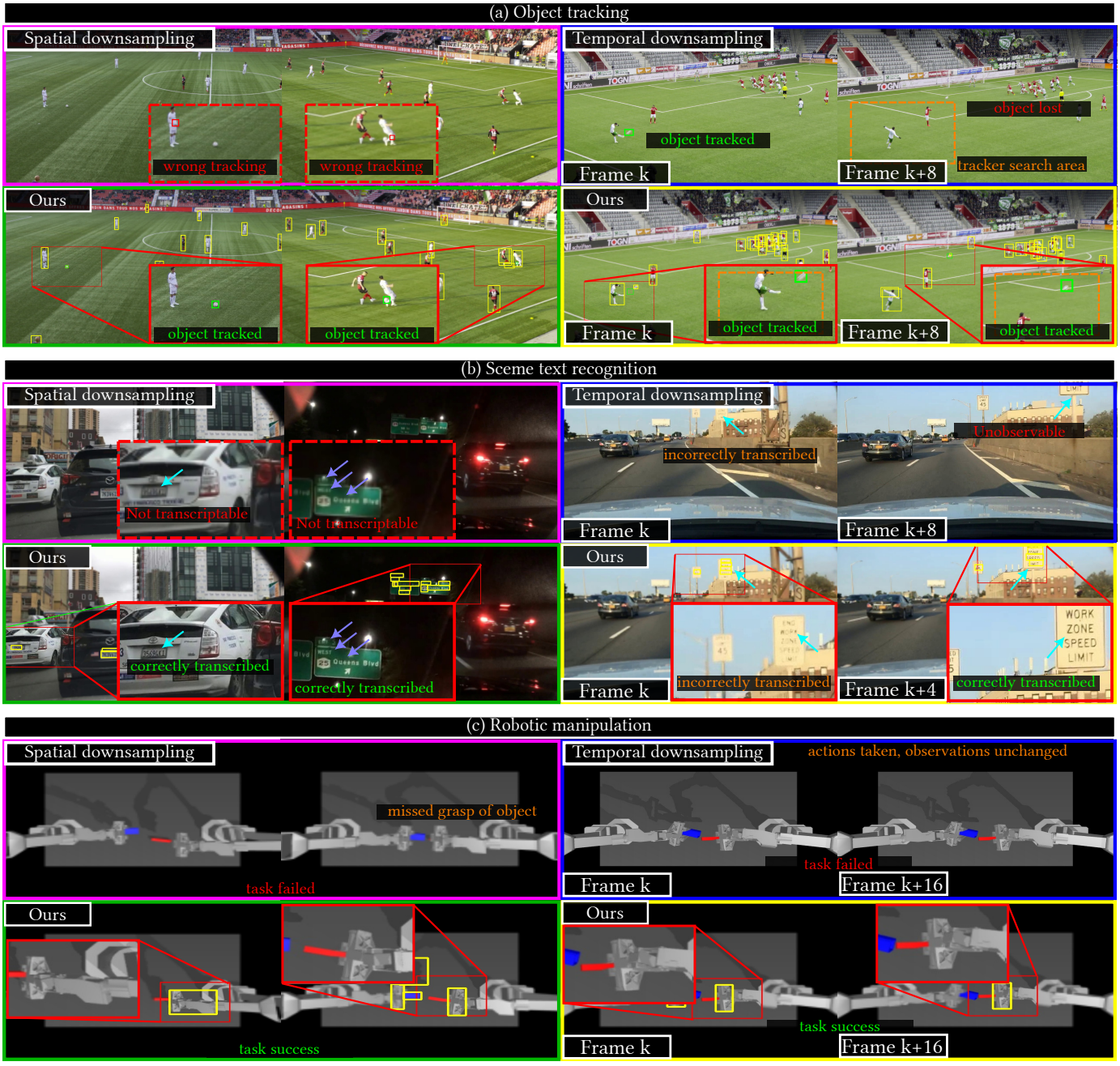

Fig. 4. **Policy-based foveated imaging and perception for simulated video tasks. Row (a):** Our foveated imaging approach correctly allocates higher resolution for pursuing objects of interest in an object tracking task. ROIs from our foveated imaging framework provide fine spatial details required 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 text recognition task compared to naive spatio-temporal downsampling methods where texts are frequently not transcriptable or missed. **Row (c):** In robotics manipulation, our method attends to important regions critical for task success while keeping low latency, ensuring observed state reflects robot actions and significantly improving our performance against task-agnostic baselines.

[11] J. Ba, V. Mnih, and K. Kavukcuoglu, "Multiple object recognition with visual attention. arxiv 2014," *arXiv preprint arXiv:1412.7755*, 2014.

[12] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[13] C. Esteves, C. Allen-Blanchette, X. Zhou, and K. Daniilidis, "Polar transformer networks," *arXiv preprint arXiv:1709.01889*, 2017.

[14] E. Akbas and M. P. Eckstein, "Object detection through search with a foveated visual system," *PLoS computational biology*, vol. 13, no. 10, p. e1005743, 2017.

[15] A. Jonnalagadda, W. Y. Wang, B. Manjunath, and M. P. Eckstein, "Foveater: Foveated transformer for image classification," *arXiv preprint arXiv:2105.14173*, 2021.

[16] G. Killick, P. Henderson, P. Siebert, and G. Aragon-Camarasa, "Foveation in the era of deep learning," *arXiv preprint arXiv:2312.01450*, 2023.

[17] Y. Hu, Y. Cheng, A. Lu, Z. Cao, D. Wei, J. Liu, and Z. Li, "Lf-vit: Reducing spatial redundancy in vision transformer for efficient image recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2274–2284.

[18] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[19] A. Haque, A. Alahi, and L. Fei-Fei, "Recurrent attention models for depth-based person identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1229–1238.

[20] Y. Wang, Z. Chen, H. Jiang, S. Song, Y. Han, and G. Huang, "Adaptive focus for efficient video recognition," in *proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 16 249–16 258.

[21] Y. Wang, Y. Yue, Y. Lin, H. Jiang, Z. Lai, V. Kulikov, N. Orlov, H. Shi, and G. Huang, "Adafocus v2: End-to-end training of spatial dynamic networks for video recognition," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 20 030–20 040.

[22] Y. Wang, Y. Yue, X. Xu, A. Hassani, V. Kulikov, N. Orlov, S. Song, H. Shi, and G. Huang, "Adafocusv3: On unified spatial-temporal dynamic video recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 226–243.

[23] Y. Wang, H. Zhang, Y. Yue, S. Song, C. Deng, J. Feng, and G. Huang, "Uni-adafocus: spatial-temporal dynamic computation for video recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[24] Y. Wang, Y. Yue, Y. Yue, H. Wang, H. Jiang, Y. Han, Z. Ni, Y. Pu, M. Shi, R. Lu *et al.*, "Emulating human-like adaptive vision for efficient and flexible machine visual perception," *Nature Machine Intelligence*, pp. 1–19, 2025.

[25] W. Wu, D. He, X. Tan, S. Chen, and S. Wen, "Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6222–6231.

[26] Z. Wu, H. Li, C. Xiong, Y.-G. Jiang, and L. S. Davis, "A dynamic frame selection framework for fast video recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 1699–1711, 2020.

[27] B. Korbar, D. Tran, and L. Torresani, "Scsampler: Sampling salient clips from video for efficient action recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6232–6242.

[28] Y. Meng, C.-C. Lin, R. Panda, P. Sattigeri, L. Karlinsky, A. Oliva, K. Saenko, and R. Feris, "Ar-net: Adaptive frame resolution for efficient action recognition," in *European conference on computer vision*. Springer, 2020, pp. 86–104.

[29] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei, "End-to-end learning of action detection from frame glimpses in videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2678–2687.

[30] A. Ghodrati, B. E. Bejnordi, and A. Habibian, "Frameexit: Conditional early exiting for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 608–15 618.

[31] B. Xia, W. Wu, H. Wang, R. Su, D. He, H. Yang, X. Fan, and W. Ouyang, "Nsnet: Non-saliency suppression sampler for efficient video recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 705–723.

[32] B. Xia, Z. Wang, W. Wu, H. Wang, and J. Han, "Temporal saliency query network for efficient video recognition," in *European Conference on Computer Vision*. Springer, 2022, pp. 741–759.

[33] X. Sun, R. Panda, C.-F. R. Chen, A. Oliva, R. Feris, and K. Saenko, "Dynamic network quantization for efficient video inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7375–7385.

[34] D. Luebke and B. Hallen, "Perceptually driven simplification for interactive rendering," in *Eurographics Workshop on Rendering Techniques*. Springer, 2001, pp. 223–234.

[35] H. A. Murphy and A. T. Duchowski, "Gaze-contingent level of detail rendering." in *Eurographics (short presentations)*, 2001.

[36] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, "Towards foveated rendering for gaze-tracked virtual reality," *ACM Transactions On Graphics (TOG)*, vol. 35, no. 6, pp. 1–12, 2016.

[37] A. S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, and G. Rufo, "Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–13, 2019.

[38] Q. Sun, F.-C. Huang, J. Kim, L.-Y. Wei, D. Luebke, and A. Kaufman, "Perceptually-guided foveation for light field displays," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–13, 2017.

[39] R. Fan, X. Shi, K. Wang, Q. Ma, and L. Wang, " Scene-aware Foveated Rendering ," *IEEE Transactions on Visualization & Computer Graphics*, vol. 30, no. 11, pp. 7097–7106, Nov. 2024. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/TVCG.2024.3456157

[40] B. Guenter, M. Finch, S. Drucker, D. Tan, and J. Snyder, "Foveated 3d graphics," *ACM transactions on Graphics (tOG)*, vol. 31, no. 6, pp. 1–10, 2012.

[41] B. Krajancich, P. Kellnhofer, and G. Wetzstein, "Towards attention–aware foveated rendering," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023.

[42] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[43] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[44] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. Ieee, 2016, pp. 3464–3468.

[45] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, and X. Wang, "Bytetrack: Multi-object tracking by associating every detection box," in *European conference on computer vision*. Springer, 2022, pp. 1–21.

[46] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[47] D. Q. Mayne and H. Michalska, "Receding horizon control of nonlinear systems," in *Proceedings of the 27th IEEE Conference on Decision and Control*. IEEE, 1988, pp. 464–465.

[48] H. M. Le, N. Jiang, A. Agarwal, M. Dudík, and H. Daumé III, "Hierarchical imitation and reinforcement learning," in *Proceedings of Machine Learning Research*, 2018.

[49] A. Cioppa, S. Giancola, A. Deliege, L. Kang, X. Zhou, Z. Cheng, B. Ghanem, and M. Van Droogenbroeck, "Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3491–3502.

[50] Y. Cui, T. Song, G. Wu, and L. Wang, "Mixformerv2: Efficient fully transformer tracking," *Advances in neural information processing systems*, vol. 36, pp. 58 736–58 751, 2023.

[51] S. Reddy, M. Mathew, L. Gomez, M. Rusinol, D. Karatzas, and C. Jawahar, "Roadtext-1k: Text detection & recognition dataset for driving videos," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 11 074–11 080.

[52] M. Ye, J. Zhang, S. Zhao, J. Liu, T. Liu, B. Du, and D. Tao, "Deepsolo: Let transformer decoder with explicit points solo for text spotting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 348–19 357.

[53] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.