PixelGen: Rethinking Embedded Cameras for Mixed-Reality

Kunjun Li*, Manoj Gulati*, Dhairya Shah*, Steven Waskito, Shantanu Chakrabarty, Ambuj Varshney
National University of Singapore, Singapore
{kunjun,dhairya,steven.waskito}@u.nus.edu,shantanu.chakrabarty@ncs.com.sg
{manojg,ambujv}@nus.edu.sg

Figure 1: PixelGen argues for decoupling the camera from the mixed-reality headset. The PixelGen system captures a representation of the environment using low-resolution sensors, a monochrome image sensor and environmental sensors. The captured sensor information is then processed through a transformer-based language and image model for generating realistic representations of the environment for visualization on a mixed-reality headset.

Abstract

Mixed-reality headsets offer new ways to perceive our environment. They employ visible spectrum cameras to capture and display the environment on screens in front of the user's eyes. However, these cameras lead to limitations. Firstly, they capture only a partial view of the environment. They are positioned to capture whatever is in front of the user, thus creating blind spots during complete immersion and failing to detect events outside the restricted field of view. Secondly, they capture only visible light fields, ignoring other fields like acoustics and radio that are also present in the environment. Finally, these power-hungry cameras rapidly deplete the mixed-reality headset's battery. We introduce PixelGen to rethink embedded cameras for mixed-reality headsets. PixelGen proposes to decouple cameras from the mixed-reality headset and balance resolution and fidelity to minimize the power consumption. It employs low-resolution, monochrome image sensors and environmental sensors to capture the surroundings around the headset. This approach reduces the system's communication bandwidth and power consumption. A transformer-based language and image model process this information to overcome resolution trade-offs, thus generating a higher-resolution representation of the environment. We present initial experiments that show PixelGen's viability.

CCS Concepts

• **Hardware** → **Sensor devices and platforms**; • **Computer systems organization** → *Embedded systems.*

Keywords

Embedded systems, Networking, Large language Models

ACM Reference Format:
Kunjun Li*, Manoj Gulati*, Dhairya Shah*, Steven Waskito, Shantanu Chakrabarty, Ambuj Varshney. 2024. PixelGen: Rethinking Embedded Cameras for Mixed-Reality. In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), November 18–22, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3636534.3696216

* These authors contributed equally to the work and are co-primary authors.

ACM MobiCom '24, November 18–22, 2024, Washington D.C., DC, USA

ACM ISBN 979-8-4007-0489-5/24/11
https://doi.org/10.1145/3636534.3696216

1 Introduction

Mixed-reality (MR) headsets offer users novel ways to perceive their surroundings. Commodity, off-the-shelf devices like Apple Vision Pro [8], Meta Quest [15] are widely available. They employ various sensors—including high-resolution cameras, acoustic, and motion sensors to track the surroundings. This enables MR modes that project the environment onto a screen before the user's eyes. The cameras also track

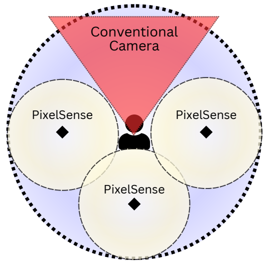


Figure 2: PixelGen is able to address the challenge of limited field of view of cameras and sensors on MR headsets by allowing for decoupled deployment of cameras for a broader coverage of the surrounding area.

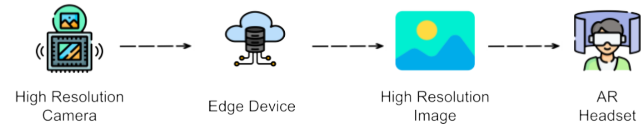
body gestures. Thus, embedded cameras are critical for functionality of MR headsets.

Embedded cameras have attracted significant research interest over decades [4, 19, 23, 26, 27], leading to their ubiquity. These cameras capture visible light-illuminated representations of the environment, generating pixel arrays that form images. Architecturally, embedded cameras follow a pipelined design [4, 18, 19, 23, 27] consisting of sensing and processing stages. A key component is the image sensor, which contains numerous photo sensors to detect visible light. Image sensors produce weak analog signals, which are then processed by analog front ends, usually containing a low-noise amplifier and an automatic gain control block. An analog-to-digital converter then digitizes these signals. Subsequent stages process the digital signals, performing tasks like image compression. The final steps involve further image processing, storage, and wireless communication.

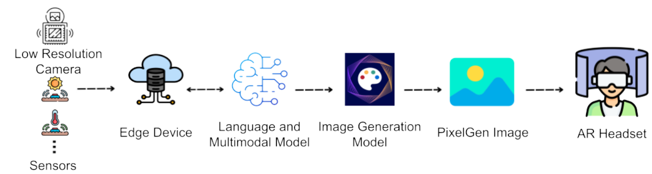
Even though embedded cameras are among the most widely deployed example of an embedded system, their architecture has remained largely unchanged. This presents significant challenges for emerging applications like MR headsets.

Challenges. High power consumption of cameras significantly impacts the MR headset's battery life. Despite advancements in image sensor technology, with some image sensors consuming only microwatts [18], the overall power consumption for cameras remains high. This is due to multiple energy-intensive processing stages required for analog signals from image sensors. Higher-resolution sensors exacerbate this issue by generating even higher bandwidth signals that correspondingly require energy-expensive components such as amplifiers with high gain bandwidth product. Additionally, higher image resolution and frame rates also require proportionally higher communication bandwidth.

Conventional cameras, and their role, is also a limitation in the MR context. While cameras have evolved significantly over millennia, from the camera obscura [3] to the Daguerreotype [12] and modern digital cameras pioneered by Steven Sasson at Kodak [11], they have primarily focused on capturing visible or infrared light. This covers only a limited aspect of our surroundings. Moreover, MR headsets, that project



(a) Conventional camera architecture



(b) PixelGen camera architecture

Figure 3: Conventional cameras capture surroundings illuminated by visible light at high resolution. In contrast, PixelGen captures a broader representation using low-resolution image and environmental sensors thus, reducing power consumption.

the world onto a screen before the user's eyes, could capture and present a broader range of elements, including wind, humidity, acoustic emissions, magnetic and radio fields.

Finally, as we shown in Figure 2, cameras on MR headsets capture only the surroundings facing the user. This is particularly challenging in immersive virtual modes, where users may be unaware of the surrounding events that are not captured by the headset's sensors due to limited field of view, such as someone approaching from behind.

Design. PixelGen is our ongoing effort to tackle the challenges of cameras for MR headsets. This system, in a way, rethinks the conventional architecture of embedded cameras and takes a stab at tackling the tradeoffs between power consumption, bandwidth, sensing resolution, and fidelity. We show the PixelGen platform in Figure 1 contrasting it against conventional camera architecture in Figure 3.

The key component of PixelGen is the PixelSense platform. It comprises a low-power microcontroller, a communication module, and low-resolution image and environmental sensors. We have deliberately made design choices to enable low power consumption of the platform by trading sensing resolution for power consumption. The collected sensor information contains valuable cues about the environment that can be used to fill in missing details in the lower-resolution image, enabling the generation of higher-resolution images or generating novel representations by visualizing information that would be unseen by conventional cameras.

The captured information is then processed through language and image models to generate a higher-resolution representation of the surrounding environment. In particular, we build on the significant recent progress with language and image models [1, 7, 21]. They are made possible due to the emergence of the transformer architecture [36], which has enabled large models to demonstrate impressive emergent capabilities. Today, these models allow us to interact with computers using natural language prompts and help us identify and understand images. Another class of models, called image models, can generate complex images by taking natural language prompts and images as input [17, 25]. PixelGen builds on the emerging capabilities of both language and image models to overcome the tradeoffs made to lower the power consumption of the PixelSense platform.

Finally, the user provides information in natural language or uses an existing template, which enables the PixelGen to process the lower-resolution image and sensor data. This includes generating a high-resolution representation of the surroundings or unique representations. In this work, we demonstrate an end-to-end system that captures low-resolution images with environmental information using PixelSense and generates higher-resolution and artistic representations, which are then presented to the user on an MR headset.

2 Background

Mixed reality and cameras. Previous efforts have explored various aspects of embedded cameras in MR headsets. Rhee et al. investigated cameras enabling 360-degree rendering of surroundings [24]. Koch et al. focused on enhancing depth perception using time-of-flight sensors [9]. Mandl et al. developed methods for coherent rendering by learning camera characteristics to improve lighting, visual coherence, and photo-realistic rendering [14]. Schwandt et al. explored using secondary RGB-D cameras to generate glossy reflections on virtual objects through partial 3D reconstruction of the natural environment [28]. However, all of these approaches rely on power-consuming and conventional camera architectures which are limited to capturing the visible spectrum.

Embedded camera. Early embedded camera designs Cyclops [23], coupled image sensors with complex programmable logic devices (CPLDs) for image processing, easing the burden on microcontrollers. CITRIC [4] combined an image sensor with a powerful microcontroller, large storage, and an 802.15.4-based transceiver for camera-based sensor networks. CMUCam [26, 27] was an open-source embedded camera optimized for near real-time image processing. Although, these platforms consumed significant power, averaging from tens to hundreds of milliwatts, they only managed to capture information from visible spectrum.

Recent efforts have focused on reducing power consumption. WISPCam [19], a battery-free camera based on RFID, harvests radio energy and communicates images using the EPCGen protocol. Naderiparizi et al.[18] designed a camera capable of streaming high-definition images with minimal power consumption using a backscatter mechanism, though limited to a range of a few meters. Camaroptera[6] used the LoRa protocol to transmit low-resolution images over long distances using solar-harvested power, achieving battery-free operation through heavy-duty cycling and infrequent, low-resolution image capture. NeuriCam [37] uses a dual-mode camera system that operates in a low-power grayscale mode and a high-power color mode, using a neural network to enhance video quality by super-resolving low-resolution key frames and colorizing them. However, these systems primarily capture visible spectrum images and often sacrifice resolution or duty cycle to conserve power. While some can capture high-resolution images, their communication range is limited. Furthermore, they do not address MR applications.

3 Design

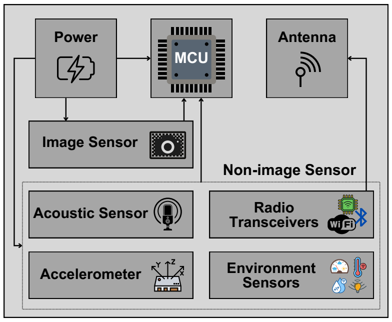
The PixelGen system consists of three components: the PixelSense platform, a MR headset, and an edge device. The PixelSense platform captures information about the surrounding environment and transmits these to the edge device. At the edge device, users can provide context through natural language prompts, or existing prompts are used to contextualize the captured sensor information. These prompts are then processed by a language model to generate relevant context for the image model. Using this information, the image model creates a representation of the surroundings, which is then displayed to the user on the MR headset. We provide an overview of the PixelGen system in the Figure 3(b).

3.1 PixelSense

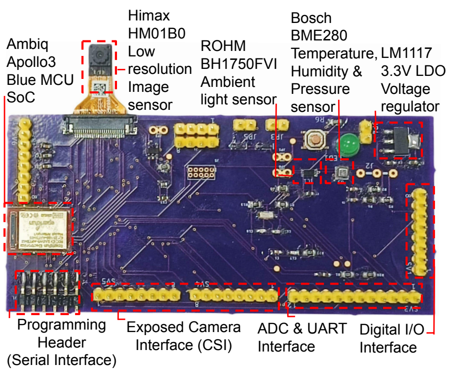
Let us first examine the conventional camera architecture. It employs a pipelined design, beginning with an image sensor that captures light intensity values. These signals are then amplified, stabilized, and digitized for compression or processing. A microcontroller manages the processed image for storage and wireless transmission. While the sensor itself is relatively low-power, high-resolution sensors can consume hundreds of microwatts. Moreover, the processing and transmission stages are energy expensive, contributing to significantly reducing MR headset's battery lifespan, especially during continuous operation. Any increase in the pixel count of the camera further exacerbates the power consumption challenge. We illustrate the architecture in Figure 3(a).

PixelGen architecture. PixelSense adopts an approach that differs from the conventional camera architecture. Rather than using a high-resolution image sensor, it captures a broader environmental information using low-resolution image sensor and environmental sensors. This purposefully helps us to lower the power-consumption of the platform.

PixelSense can also be configured to detect other fields such as radio and magnetic emissions from the surroundings if the application requires. These additional sensor information help serve two purposes: Firstly, they help mitigate the



(a) High-level schematic



(b) Hardware prototype

Figure 4: PixelSense integrates an array of low-resolution, low-power sensors that capture information from the surrounding environment. These trade-offs ensure lower power consumption for the platform.

trade-offs made by using lower-resolution image sensors, allowing the PixelGen to reconstruct higher-resolution representations by taking hints from the captured sensor information. Secondly, they can be used to represent hidden fields and phenomena that conventional cameras cannot capture. We represent a schematic of the PixelSense in the Figure 4(a).

Implementation. We designed the platform using commodity off-the-shelf components. We use Ambiq Apollo3 microcontroller [16] for its low-power consumption and BLE standard support, enabling communication with the MR headsets without a gateway. We have also included provisions for low-power transmissions using backscatter [33] or tunnel diode based mechanisms [34, 35]. For environmental sensing, we incorporated a Bosch BME280 [30] sensor, and a ROHM BH1750 sensor [29] to estimate light conditions. The platform can capture motion through a 9-axis IMU sensor from ST Microelectronics. The visual information is captured using a Himax HM01B0 [32] low-resolution and monochrome image sensor. We have also provisioned the platform to capture acoustic information. Finally, we have fabricated the platform on a 4-layer FR4 PCB manufactured by OSHPark [22]. Figure 4(b) shows the prototype.

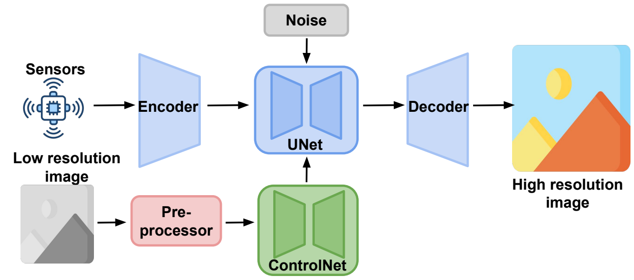


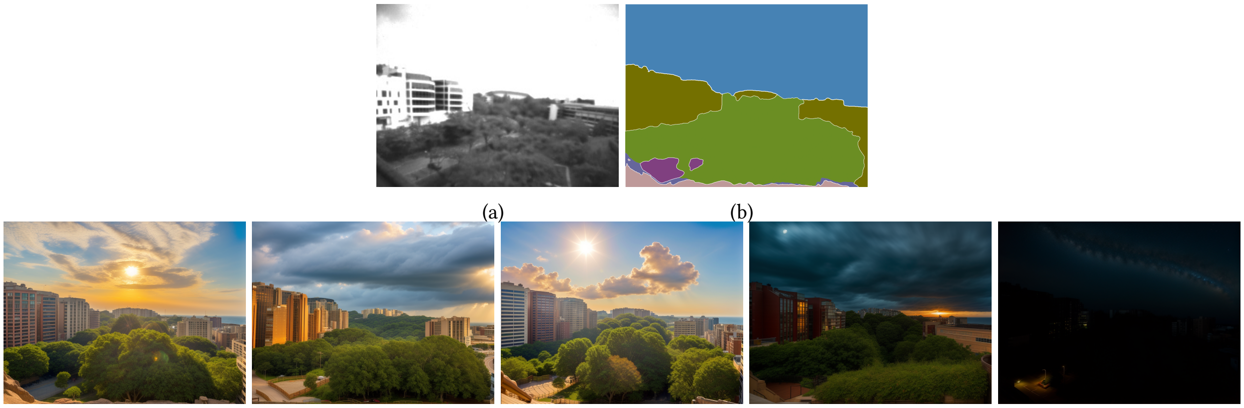
Figure 5: The image model (diffusion architecture) receives input from the language model and the low-resolution images captured by PixelSense. ControlNet extracts spatial information, which is used in the subsequent steps to generate a higher-resolution image or to enhance or add details in the generated image.

Power Consumption. We used a Nordic Power Profiler Kit-2 [20] with a supply voltage of 3.3V. We measured the time and current consumption for stages of image capture. During the initial configuration and calibration of the image sensor parameters, we observed an average current draw of 4 mA over 12s. This setup phase is required only once and consumed 13.2 mW on average, with a total energy consumption of 158.4 mJ. For the image capture phase, frame capture and acquisition takes approximately 2s, drawing 4 mA of current. This results in an energy consumption of 26.4 mJ, with an average power consumption of 13.2 mW. These figures are significantly lower compared to higher-resolution embedded cameras. However, there is potential for further power reduction through optimizations such as lowering the supply voltage, increasing the clock rate, and using state-of-the-art microcontrollers like the Apollo 510 [2].

3.2 Language and Image Model

The PixelSense transmits collected images and sensor data to an edge device. Since PixelSense also supports the BLE standard, this edge device could be a capable MR headset like the Vision Pro, thus eliminating the need for an additional device. Users interact with the system by providing natural language prompts, which guide the language model to create contextual information for processing the sensor data. This information is then used with an image model to represent the surroundings, which is displayed on the MR headset.

Implementation. We implement language model using Meta Llama and OpenAI GPT. As an image model, we use the stable diffusion [31, 38], which is an open-source diffusion model with weights Realistic Vision V5.1 [5]. As an edge device, we selected mini computers that are capable of operating the stable diffusion model locally. We use an XREAL headset connected to the edge device as MR headset.



(a) (b) (c) (d) (e) (f) (g)

Figure 6: (a) A monochrome image (324x244) is captured and input to the OSM, which extracts object locations and shapes, as shown in (b). This information is combined with time-varying sensor data and fed into an image and language model. The model generates realistic images (968x728) as output, corresponding to different light conditions (Lux values): (c) 65535, (d) 3089, (e) 16556, (f) 536, and (g) 0. The segmentation output (background image) remains consistent across these variations. The changing environmental conditions result in corresponding changes in the sensor information and the generated image, resulting in images representing day or night conditions.

Language model. The language model can process input from an existing template, which includes specific instructions for processing sensor data to provide contextual information about the surroundings. Alternatively, the language model can take input from the user using natural language regarding steps for visualizing the surroundings, including instructions for generating higher-resolution images or adding relevant information for processing using image model.

Image model. The contextual information generated by the language model, along with the captured low-resolution image, is used with an image model. These models, trained on vast image collections and fine-tuned with environment-specific details, generate a representation of the surrounding environment. We show the architecture of the image model in the Figure 5. This process occurs in two discrete steps: We process the low-resolution monochrome images using ControlNet, a neural network framework for diffusion models [13]. ControlNet can retrieve additional information from low-resolution images, such as edges, boundaries, and objects. It can be used with various preprocessing models, as demonstrated in our evaluation section. For example, a Canny edge detector can help retain image composition by extracting outlines and edges, while the OneFormer segmentation preprocessor can transfer object location and shape by labeling them with predefined colors.

The second step consists of integrating prompts from the language model and augmenting this with additional control from the original low-resolution images. Using the image model, we generate high-resolution images that maintains the object and background details while enhancing color information and resolution or adding additional sensor fields.

4 Evaluation

Setup. We conducted experiments in both indoor and outdoor environments, but for brevity, we present only the outdoor results. We collected over 126 images and sensor data (temperature, humidity, pressure, and light intensity) in diverse conditions and locations. The captured images have a resolution of 324 x 244 in monochrome format.

We used preprocessing models like the Canny Edge Detection Model (CEDM), Oneformer Segmentation Model (OSM), and the Line Art Control Model (LACM) to extract meaningful information from the captured images. These models helped identify object positions, edges, color palettes, and features like brightness and hues. The captured sensor data was then manually converted into prompts via a language model. These prompts and other extracted information were fed into the image generation model (e.g., stable diffusion) to create high-resolution images of the environment, which were visualized using an XREAL MR headset.

We evaluated the system's ability to generate diverse representations and narrowed the results to three styles: realistic, and Van Gogh-style oil painting. After post-processing, PixelGen regenerates high-resolution images.

Realistic images. We investigate PixelGen's ability to generate realistic visualization of the surroundings. We configure the image model to generate realistic images and preprocess the captured monochrome image with the OSM model. This model extracts the location and shape of objects in the image. Figure 6(a) shows the original input image, and Figure 6(b) shows the segmented output of the OSM model, along with 6(c) the generated image from the image model which was then visualised on a MR headset.

(a) (b) (c)

Figure 7: (a) A monochrome image (324x244) is captured and provided to the LACM, which extracts the essence of the still image, including hues and dark areas, along with the location and shape of objects, as shown in (b). We combine this with an environmental sensor stream (0 Lux, 28.51 °C, 71.37%, 1008.75 hPa, and Wind Velocity: 0.0 m/s) and a prompt generated by the language model. This information is then fed into the image model to generate (c) a realistic Van Gogh Oil Painting (968x728) as output. This artistic representation is then displayed on the MR headset, enabling the user to perceive their surroundings in a unique manner.

We observe that PixelGen is able to generate a much higher-resolution image. Furthermore, by incorporating input from the sensor data, it is able to correspondingly represent the environmental conditions in the generated images.

Artistic images. We evaluate PixelGen ability to represent the environment artistically. This capability can be especially relevant for MR headsets, enabling users to observe their surroundings in novel ways. We preprocess captured images through the LACM model. The extracted information from this process is then combined with data from environmental sensors, and a manually provided prompt serves as an instruction to the image model. We show the original image, the output from the LACM, and the subsequently generated images in Figure 7. Notably, the captured image is in monochrome, and the additional input from the environmental sensors provides the image model with context. This enables PixelGen to generate a higher-resolution and artistically rendered version of the surrounding environment.

Image sequence. We investigate PixelGen's ability to generate an image sequence. To achieve this, we trigger a unique prompt for each sensor data input, resulting in varying outputs. We present some key frames from this analysis in Figure 6. The complete video output (combined stream of images) can be viewed at this link [10]. We observe significant variations in sunlight (ambient light intensity) from Figure 6(c) to 6(g), along with changes in cloud density. Notably, the composition of the actual image (similar to the original monochrome image) remains intact. This demonstrates that minor changes in sensor data and prompts are reliably translated into the generated sequence of images. We also present the list of positive prompts generated using GPT-4, which we use to generate these realistic images, along with original environmental sensor data in Table 1.

5 Discussion

Hallucination. Due to lower-resolution images and sensor inputs, providing these to the model sometimes results in hallucinations for some images. This could lead to images generated with objects or features not actually present in the environment. However, fine-tuning the model with a specific dataset could be a potential solution for generating more accurate images and reducing such hallucinations.

Inference on headsets. MR headsets like Apple Vision Pro can run models locally. The image model requires significant computation, which can lead to lower battery lifespan. Techniques such as model pruning, quantization, and knowledge distillation can address this. Model pruning removes redundant parameters, reducing its size and inference time. Quantization lowers memory usage and speeds up processing by using lower precision weights. Knowledge distillation transfers knowledge from a larger model to a smaller one, maintaining performance while enhancing efficiency. They may collectively make it feasible for real-time image generation and minimal latency media streaming on MR headsets.

6 Conclusion

We have presented PixelGen, our ongoing work to rethink embedded cameras for mixed-reality headsets. Through experiments, we demonstrate that the unique convergence of language and image models, combined with a carefully designed low-power embedded camera, can enable us to capture higher-resolution representations or even present the environment uniquely. This work is released as open-source, with design files and other details available at the following repository: https://github.com/weiserlab/PixelGen

Acknowledgments

We thank the anonymous reviewers for their comments. This work was supported primarily through a grant from the NUS-NCS center (A-0008542-02-00), a startup grant (A-8000277-00-00), a MoE Tier 1 Grant (A-8001661-00-00), and an unrestricted gift from Google (A-8002307-00-00) through their Research Scholar Program. All of these grants were administered through the National University of Singapore.

References

[1] Meta AI. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023). https://arxiv.org/abs/2302.13971

[2] Ambiq. 2024. Ambiq. https://ambiq.com/apollo510/ online, accessed 10-July-2024.

[3] Scott Billings. 2023. https://www.hsm.ox.ac.uk/camera-obscura [Online; accessed 13-October-2023].

[4] Phoebus Chen, Parvez Ahammad, Colby Boyer, Shih-I Huang, Leon Lin, Edgar Lobaton, Marci Meingast, Songhwai Oh, Simon Wang, Posu Yan, et al. 2008. CITRIC: A low-bandwidth wireless camera network platform. In *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*. IEEE, 1–10.

[5] CivitAI. 2023. Realistic Vision V5.1. https://civitai.com/models/4201/realistic-vision-v51 online, accessed 10-July-2024.

[6] Harsh Desai, Matteo Nardello, Davide Brunelli, and Brandon Lucia. 2022. Camaroptera: A Long-Range Image Sensor with Local Inference for Remote Sensing Applications. *ACM Trans. Embed. Comput. Syst.* 21, 3, Article 32 (may 2022), 25 pages. https://doi.org/10.1145/3510850

[7] Gemini Team Google. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* (2023). https://arxiv.org/abs/2312.11805

[8] Apple Inc. 2024. Apple Vision Pro. https://www.apple.com/apple-vision-pro/ online, accessed 27-July-2024.

[9] Reinhard Koch, Ingo Schiller, Bogumil Bartczak, Falko Kellner, and Kevin Koeser. 2009. MixIn3D: 3D mixed reality with ToF-camera. In *Dynamic 3D Imaging: DAGM 2009 Workshop, Dyn3D 2009, Jena, Germany, September 9, 2009. Proceedings*. Springer, 126–141.

[10] First Name Last Name. 2023. Video generated using PixelGen system. http://bit.ly/generated_video

[11] Gareth A. Lloyd. 1977. Electronic still Camera. https://patents.google.com/patent/US4131919A/en. [Online; accessed 13-Oct-2023].

[12] Bates Lowry and Isabel Barrett Lowry. 2000. *The Silver Canvas: Daguerreotype Masterpieces from the J. Paul Getty Museum*. Getty Publications.

[13] Maneesh Agrawala Lvmin Zhang, Anyi Rao. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE/CVF, 3836–3847. https://openaccess.thecvf.com/content/ICCV2023/papers/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.pdf

[14] David Mandl, Peter M Roth, Tobias Langlotz, Christoph Ebner, Shohei Mori, Stefanie Zollmann, Peter Mohr, and Denis Kalkofen. 2021. Neural cameras: Learning camera characteristics for coherent mixed reality rendering. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 508–516.

[15] Inc. Meta Platforms. 2024. Meta Quest. https://www.meta.com/quest/ online, accessed 27-July-2024.

[16] Ambiq Micro. 2023. Ambiq Apollo3 Blue. https://ambiq.com/wp-content/uploads/2020/10/Apollo3-Blue-SoC-Datasheet.pdf online, accessed 10-July-2024.

[17] MidJourney. 2024. MidJourney. https://www.midjourney.com/ online, accessed 27-July-2024.

[18] Saman Naderiparizi, Mehrdad Hessar, Vamsi Talla, Shyamnath Gollakota, and Joshua R Smith. 2017. Ultra-low-power wireless streaming cameras. *arXiv preprint arXiv:1707.08718* (2017).

[19] Saman Naderiparizi, Aaron N. Parks, Zerina Kapetanovic, Benjamin Ransford, and Joshua R. Smith. 2015. WISPCam: A battery-free RFID camera. In *2015 IEEE International Conference on RFID (RFID)*. 166–173. https://doi.org/10.1109/RFID.2015.7113088

[20] Nordic. 2024. Nordic. https://www.nordicsemi.com/Products/Development-hardware/Power-Profiler-Kit-2 online, accessed 10-July-2024.

[21] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023). https://arxiv.org/abs/2303.08774

[22] OSHPark. 2023. OSHPark. https://oshpark.com/ online, accessed 10-July-2024.

[23] Mohammad Rahimi, Rick Baer, Obimdinachi I Iroezi, Juan C Garcia, Jay Warrior, Deborah Estrin, and Mani Srivastava. 2005. Cyclops: in situ image sensing and interpretation in wireless sensor networks. In *Proceedings of the 3rd international conference on Embedded networked sensor systems*. 192–204.

[24] Taehyun Rhee, Lohit Petikam, Benjamin Allen, and Andrew Chalmers. 2017. Mr360: Mixed reality rendering for 360 panoramic videos. *IEEE transactions on visualization and computer graphics* 23, 4 (2017), 1379–1388.

[25] Dominik Lorenz Patrick Esser Björn Ommer Robin Rombach, Andreas Blattmann. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752* (2022). https://arxiv.org/abs/2112.10752

[26] Anthony Rowe. 2002. A low cost embedded color vision system. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vol. 1. IEEE, 208–213.

[27] Anthony Rowe, Adam G. Goode, Dhiraj Goel, and Illah Nourbakhsh. 2007. *CMUcam3: An Open Programmable Embedded Vision Sensor*. Technical Report CMU-RI-TR-07-13. Carnegie Mellon University, Pittsburgh, PA.

[28] Tobias Schwandt and Wolfgang Broll. 2016. A single camera image based approach for glossy reflections in mixed reality applications. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 37–43.

[29] ROHM Semiconductor. 2023. ROHM BH1750FVI. https://www.rohm.com/documents/11303/41217/CNA09016_wp.pdf online, accessed 10-July-2024.

[30] Bosch Sensortec. 2023. Bosch BME280. https://www.bosch-sensortec.com/products/environmental-sensors/humidity-sensors-bme280/ online, accessed 10-July-2024.

[31] StabilityAI. 2023. Stable Diffusion. https://stablediffusionweb.com/ online, accessed 10-July-2024.

[32] Himax Technologies. 2023. Himax HM01B0. https://www.himax.com.tw/products/cmos-image-sensor/always-on-vision-sensors/hm01b0/ online, accessed 10-July-2024.

[33] Ambuj Varshney, Oliver Harms, Carlos Pérez-Penichet, Christian Rohner, Frederik Hermans, and Thiemo Voigt. 2017. LoRea: A Backscatter Architecture That Achieves a Long Communication Range. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems* (Delft, Netherlands) *(SenSys '17)*. Association for Computing Machinery, New York, NY, USA, Article 18, 14 pages. https://doi.org/10.1145/3131672.3131691

[34] Ambuj Varshney, Andreas Soleiman, and Thiemo Voigt. 2019. Tunnelscatter: Low power communication for sensor tags using tunnel diodes. In *The 25th annual international conference on mobile computing and networking*. 1–17.

[35] Ambuj Varshney, Wenqing Yan, and Prabal Dutta. 2022. Judo: addressing the energy asymmetry of wireless embedded systems through tunnel diode based wireless transmitters. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*. 273–286.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[37] Bandhav Veluri, Ali Saffari, Collin Pernu, Joshua Smith, Michael Taylor, and Shyamnath Gollakota. 2022. NeuriCam: Video Super-Resolution and Colorization Using Key Frames. *arXiv preprint arXiv:2207.12496* (2022).

[38] Wikipedia. 2023. Stable Diffusion. https://en.wikipedia.org/wiki/Stable_Diffusion online, accessed 10-July-2024.

Appendix

Image	Sensor Data	Prompt
c	Ambient Light: 65535 Lux Temperature: 34.51 Celsius Humidity: 63.14 % Pressure: 1006.42 hPa Wind Velocity: 0.4 m/s	Ultra-high-resolution photograph capturing an intensely sunny and hot day, with clear skies, moderate humidity, and a calm atmosphere. The scene is bathed in bright sunlight, creating deep shadows and vivid contrasts. Ideal for 8k UHD, high dynamic range imagery, shot with a Canon EOS-1D X Mark III.
d	Ambient Light: 3089 Lux Temperature: 27.42 Celsius Humidity: 82.03 % Pressure: 1004.32 hPa Wind Velocity: 2.6 m/s	High-resolution photo capturing a warm and overcast day with high humidity, creating a slightly muggy feel. The sky is cloud-covered, diffusing the light and casting a soft glow. A noticeable breeze ripples through the scene. Ideal for 8k UHD, captured with a high dynamic range, using a Leica SL2.
e	Ambient Light: 16556 Lux Temperature: 30.69 Celsius Humidity: 67.07 % Pressure: 1003.70 hPa Wind Velocity: 0.0 m/s	Ultra-high-resolution image of a bright, sunny day with clear skies, warm temperatures, and moderate humidity. The sun casts sharp, defined shadows, and the calm, windless atmosphere creates a serene outdoor scene. Perfect for 8k UHD, high dynamic range photography, shot with a Sony Alpha A9.
f	Ambient Light: 536 Lux Temperature: 27.79 Celsius Humidity: 73.83 % Pressure: 1003.53 hPa Wind Velocity: 1.5 m/s	High-quality image depicting a warm evening with a heavily overcast sky, creating a dim and moody atmosphere. The air is humid and slightly breezy, adding to the sense of an impending nightfall. Ideal for 8k UHD, high dynamic range, captured with a Nikon D850.
g	Ambient Light: 0 Lux Temperature: 26.85 Celsius Humidity: 79.26 % Pressure: 1004.63 hPa Wind Velocity: 0.0 m/s	High-resolution night photograph capturing a warm and very humid environment in complete darkness, conveying a still and heavy night atmosphere. The air is thick with moisture, creating a sense of quiet and solitude. Ideal for 8k UHD, high dynamic range, shot with a Canon EOS R6, excellent for night photography.

Table 1: shows the image index corresponding to the stream of realistic images (Figure 6) along with the actual time-varying environmental sensor data (shown in the second column) and the corresponding positive prompt generated using GPT-4 in the third column, which is fed to the diffusion model to produce output.