Published: 26 May 2025

RESEARCH-ARTICLE

Zero-Shot Pupil Segmentation with SAM 2: A Case Study of Over 14 Million Images

VIRMARIE MAQUILING, Technical University of Munich, Munich, Bayern, Germany

SEAN ANTHONY BYRNE, Politecnico di Milano, Milan, MI, Italy

DIEDERICK C. NIEHORSTER, Lund University, Lund, Skane, Sweden

MARCO CARMINATI, Politecnico di Milano, Milan, MI, Italy

ENKELEJDA KASNECI, Technical University of Munich, Munich, Bayern, Germany

Open Access Support provided by:

Politecnico di Milano

Technical University of Munich

Lund University

Zero-Shot Pupil Segmentation with SAM 2: A Case Study of Over 14 Million Images

VIRMARIE MAQUILING*, Technical University of Munich, Germany
SEAN ANTHONY BYRNE*, Politecnico di Milano, Italy
DIEDERICK C. NIEHORSTER, Lund University, Sweden
MARCO CARMINATI, Politecnico di Milano, Italy
ENKELEJDA KASNECI, Technical University of Munich, Germany

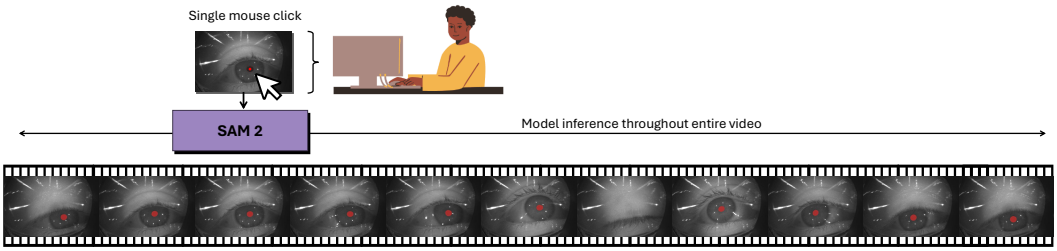

Fig. 1. An illustration demonstrating the data annotation process with SAM 2: the user provides a single point prompt via a mouse click, and SAM 2 automatically handles the rest of the segmentation process. Optionally, the user can refine and add additional prompts in more difficult areas of the video to improve the model's output.

We explore the transformative potential of SAM 2, a vision foundation model, in advancing gaze estimation. SAM 2 addresses key challenges in gaze estimation by significantly reducing annotation time, simplifying deployment, and enhancing segmentation accuracy. Utilizing its zero-shot capabilities with minimal user input—a single click per video—we tested SAM 2 on over 14 million eye images from a diverse range of datasets, including the EDS challenge datasets and Labelled Pupils in the Wild. This is the first application of SAM 2 to the gaze estimation domain. Remarkably, SAM 2 matches the performance of domain-specific models in pupil segmentation, achieving competitive mIOU scores of up to 93% without fine-tuning. We argue that SAM 2 achieves the sought-after standard of domain generalization, with consistent mIOU scores (89.71%-93.74%) across diverse datasets, from virtual reality to "gaze-in-the-wild" scenarios. We provide our code and segmentation masks for these datasets to promote further research.

*Both authors contributed equally to this research.

Authors' Contact Information: Virmarie Maquiling, Technical University of Munich, Human-Centered Technologies for Learning, Munich, Bavaria, Germany, virmarie.maquiling@tum.de; Sean Anthony Byrne, Politecnico di Milano, Dipartimento di Elettronica, Informazione e Bioingegneria, Milan, Italy, seananthony.byrne@polimi.it; Diederick C. Niehorster, Lund University, Lund University Humanities Lab & Department of Psychology, Lund, Sweden, diederick_c.niehorster@humlab.lu.se; Marco Carminati, Politecnico di Milano, DEIB, Milano, MI, Italy, marco1.carminati@polimi.it; Enkelejda Kasneci, Technical University of Munich, Human-Centered Technologies for Learning, Munich, Germany, enkelejda.kasneci@tum.de.

ACM 2577-6193/2025/6-ART23
https://doi.org/10.1145/3729409

CCS Concepts: • **Human-centered computing** → Virtual reality; • **Computing methodologies** → **Image segmentation**; *Neural networks*.

Additional Key Words and Phrases: Eye Tracking, Gaze Estimation, Foundation Models, Methods

ACM Reference Format:
Virmarie Maquiling, Sean Anthony Byrne, Diederick C. Niehorster, Marco Carminati, and Enkelejda Kasneci. 2025. Zero-Shot Pupil Segmentation with SAM 2: A Case Study of Over 14 Million Images. *Proc. ACM Comput. Graph. Interact. Tech.* 8, 2, Article 23 (June 2025), 16 pages. https://doi.org/10.1145/3729409

1 Introduction

The increasing integration of eye tracking into technologies like virtual reality (VR) devices and smart glasses [Byrne et al. 2024a] has amplified the demand for robust gaze estimation systems, where a key task is the accurate localization of the pupil within an image or video [Kim et al. 2019; Maquiling et al. 2024]. Traditional methods for pupil localization—including thresholding and center of mass calculations [Nyström et al. 2023; Peréz et al. 2003; Shortis et al. 1994] and ellipse-fitting algorithms [Santini et al. 2018a,b]—while effective in controlled environments, suffer catastrophic errors in the presence of noise such as occlusions or reflections, limiting their utility in real-world settings [Byrne et al. 2023; Kothari et al. 2022]. To overcome these limitations, deep learning-based approaches have emerged as powerful alternatives, addressing issues plaguing traditional methods like blinks or reflections [Byrne et al. 2024b; Fuhl et al. 2016a, 2017b; Kim et al. 2019] and improving the robustness and accuracy of pupil detection under challenging conditions [Fuhl et al. 2016a]. However, deploying these models requires vast amounts of annotated data and technical expertise. Data annotation can be very costly, requiring significant human labor depending on dataset size and complexity, and training gaze estimation models using supervised machine learning relies heavily on labeled data [Sambasivan et al. 2021; Sun et al. 2017]. To mitigate time and resource constraints, recent work in gaze estimation has striven to achieve domain generalization [Kothari et al. 2022; Nguyen et al. 2024]. This can involve training neural network models on multiple eye image datasets with the intuition that a model would learn a generalized representation of eye features [Kothari et al. 2020]. Domain adaptation techniques have also been explored, such as attempts to broaden the distribution of existing eye image datasets through the inclusion of synthetic eye images. However, these attempts have often found that models trained on synthetic images fail to generalize back to real-world images [Chaudhary et al. 2019; Nair et al. 2020; Nguyen et al. 2024]. While such approaches can reduce the cost and burden associated with capturing and manually labeling large quantities of real human eye data, and in turn promote data privacy [Nguyen et al. 2024], they often involve complex procedures and can be memory intensive restricting these approaches to those who possess large quantities of GPUs.

Foundation models represent a paradigm shift in artificial intelligence, transforming how people interact with, develop, and deploy deep learning models [Bommasani et al. 2021]. They potentially offer a way to sidestep the complexities of domain generalization and domain adaptation by providing out-of-the-box zero-shot methods for eye feature segmentation[Maquiling et al. 2024]. Characterized by vast numbers of trainable parameters and extensive training data, these models demonstrate impressive adaptability to downstream tasks and perform well on data distributions they have not encountered during training [Bommasani et al. 2021; Kirillov et al. 2023]. They have lowered barriers to entry for integrating AI into workflows, simplifying the use of AI-powered tools and handling complex tasks that once required specialized models and custom datasets [Bommasani et al. 2021; Zhou et al. 2023]. Building on these advancements, Maquiling et al. [2024] showcased the potential of zero-shot vision foundation models in annotating eye tracking data by evaluating the Segment Anything Model (SAM) [Kirillov et al. 2023], a vision foundation model released

by Meta AI, on the OpenEDS datasets [Garbin et al. 2019; Palmero et al. 2020]. However, SAM required at least one prompt per image, necessitating manual clicks on every image in the dataset—a time-consuming process, and it was also left unexplored how this new type of model would perform on more challenging gaze in the wild scenarios. Its successor, SAM 2 [Ravi et al. 2024], addresses this limitation by enabling a single prompt to propagate across an entire video, allowing the model to track and segment objects even with occlusions see Fig. 3 for an example. This improvement drastically reduces the need for manual interaction, making the annotation process significantly more efficient.

This paper explores the potential of SAM 2 [Ravi et al. 2024], in advancing gaze estimation research. Importantly, it addresses long-standing challenges in gaze estimation: the need for large annotated datasets, the labor-intensive process of feature annotation, the high barrier of expertise for developing custom models, and the difficulty of domain adaptation where models struggle to generalize across datasets [Byrne et al. 2023; Kim et al. 2019; Kothari et al. 2022]. This feature is particularly relevant in gaze estimation, where models often fail due to variations in differences across participant physiology, recording setups, and environmental lighting conditions [Byrne et al. 2023; Kim et al. 2019; Kothari et al. 2022]. To evaluate SAM 2, we deployed the model across diverse gaze estimation datasets, including VR environments and the world's largest unified public dataset of eye images captured with head-mounted devices [Fuhl et al. 2021]. To evaluate and demonstrate SAM 2's ease of use, we limited the annotation to just one click per video, regardless of its length. A frame where the pupil was clearly visible was selected, and a single point prompt was placed near the center of the pupil. For the OpenEDS2019 dataset [Garbin et al. 2019], which consists of non-sequential eye images, we applied a single prompt to the entire dataset by taking the prompt from the first image in the training set and propagating it to the test and validation sets, covering a total of 152 different participants. We then assess SAM 2's segmentation performance using three key metrics: (1) the intersection-over-union (IoU) of the pupil masks to assess segmentation accuracy; (2) the ratio of frames where the pupil was not successfully tracked (when the pupil was visible) to the total number of frames (referred to as *Pupil Lost*); and (3) the ratio of frames where a blink is correctly detected (i.e., the predicted mask is empty) to the total number of frames where the ground truth pupil mask is empty (referred to as *Blink Detected*). Our annotation process employed SAM 2's smallest model, `SAM2.1_hiera_tiny`, adapting the code released by SAM 2's authors [Ravi et al. 2024] so that it could handle arbitrarily long videos without preprocessing or running into memory issues. This improved version of SAM 2 used for this paper is available from https://github.com/dcnieho/segment-anything-2 while the resulting masks can be downloaded from https://zenodo.org/records/13911636.

2 Related Work

2.1 Foundation Models and the Segment Anything Models

Models such as OpenAI's GPT series [Achiam et al. 2023; Brown 2020; Radford 2018; Radford et al. 2019] and Google's BERT [Kenton and Toutanova 2019], have transformed artificial intelligence by enabling versatile, zero-shot performance across a wide range of downstream tasks [Bommasani et al. 2021]. With large parameter counts and trained on extensive data, these models can adapt effectively to new tasks without fine-tuning, making them broadly applicable beyond the field of Natural Language Processing (NLP) [Bommasani et al. 2021; Zhou et al. 2023]. This cross-domain adaptability has inspired similar advancements in other fields, with foundation models like TimeGPT [Garza et al. 2023] in time series forecasting, as well as DinoBloom [Koch et al. 2024], Nicheformer [Schaar et al. 2024], and others in medical research.

Foundation models have also achieved significant advances in computer vision, particularly in object recognition and segmentation. Notably, the Segment Anything Model (SAM) [Kirillov et al. 2023] represents a breakthrough in zero-shot image segmentation, enabling robust, general-purpose image segmentation across a wide range of visual inputs. SAM is trained on a massive diverse dataset (SA-1B), containing over one billion segmentation masks across 11 million images, making it capable of zero-shot segmentation across a wide range of image types. It operates by embedding both image and prompt inputs, such as points, bounding boxes, or text–and uses these embeddings to output segmentation masks. With its lightweight decoder and promptable structure, SAM can generate accurate masks with minimal input. Building on this, SAM 2 [Ravi et al. 2024] extends SAM's capabilities into video by introducing a memory-augmented architecture, enabling it to track and segment objects over time with higher accuracy and fewer user interventions. Unlike traditional frame-by-frame segmentation, SAM 2's design allows users to provide a single prompt, such as a point or bounding box, to initiate segmentation across an entire video. Users can add additional prompts only where necessary, which the model incorporates as new input to further refine the segmentation in subsequent frames. Trained on the SA-V dataset–which includes over 35 million masks across 50,000 videos, SAM 2 excels at video object segmentation, reducing the need for prompts per frame and operating up to six times faster than SAM for static image tasks. This makes SAM 2 highly suitable for largescale applications. See Fig. 2 for a visualization of the architectural differences between SAM, SAM 2 and traditional specialist models.

Since it was first introduced, the Segment Anything models has been applied in various fields such as medicine [Huang et al. 2024; Ma et al. 2024; Mazurowski et al. 2023; Zhang et al. 2024], remote sensing [Ding et al. 2024; Shankar et al. 2023; Wang et al. 2024], content creation [Psychogyios et al. 2023; Yu et al. 2023] and autonomous driving [Yan et al. 2024; Zhao 2023]. Further, several extensions and improvements of the SAM model have been proposed, such as introducing speed-up alternatives [Zhang et al. 2023a,b; Zhao et al. 2023], extending it to 3D [Cen et al. 2023; Guo et al. 2024; Shen et al. 2023; Yang et al. 2023b], object tracking in video [Cheng et al. 2023; Yang et al. 2023a], finetuning to specific domains [Ma et al. 2024] and so on. Since SAM 2 has been released only very recently, such adaptation have not yet appeared.

2.2 SAM for Pupil Segmentation in Video-Based Eye Tracking

Video-based eye tracking fundamentally depends on accurate segmentation of eye features such as the pupil, sclera, iris, and corneal reflections to enable precise gaze estimation. A widely used approach focuses on localizing the pupil center [Byrne et al. 2023; Fuhl et al. 2016a; Kim et al. 2019]. We propose that methodologies can generally fall into three categories: (1) traditional techniques employing rule-based image processing and thresholding [Byrne et al. 2024b; Santini et al. 2018a,b], (2) deep learning-based models that leverage neural network architectures, and (3) zero-shot foundational models. Traditional methods often underperform in diverse, noisy real-world scenarios. Deep learning methods, while more robust, typically require extensive annotated datasets and are susceptible to domain shifts, which occur when training and testing data distributions differ. In eye tracking, this can occur due to changes in recording device or differences between lighting conditions [Byrne et al. 2023; Kim et al. 2019; Kothari et al. 2022].This paper posits that zero-shot foundational models present a promising third alternative.

Historically, feature localization and segmentation of eye images have been cornerstones of video-based eye tracking. Techniques include identifying the pupil center or its outline [Fuhl et al. 2018, 2016b; Santini et al. 2018a,b; Świrski et al. 2012], detecting iris boundaries [Arvacheh and Tizhoosh 2006; Mottalli et al. 2009], and the eye aperture [Fuhl et al. 2017a]. Early methods relied heavily on complex rules, thresholds, and heuristics, while deep learning approaches introduced advancements by reducing reliance on hand-crafted components. However, these models are often optimized

for specific domains—for instance, VR environments [Garbin et al. 2019], controlled laboratory conditions [Byrne et al. 2024b], or challenging in-the-wild scenarios where occlusions and reflections are common [Fuhl et al. 2016a; Kothari et al. 2021]. Notably, many deep learning models for gaze estimation are well-known architectures like U-Net [Ronneberger et al. 2015], Resnet [He et al. 2016] and DenseNet [Huang et al. 2017] with modifications and are then trained on large amounts of eye images, normally from a specific data source (e.g one recording device). Crucially, the training dataset plays a pivotal role in model performance. Variations in participant demographics, recording setups, and lighting conditions significantly impact cross-dataset generalization, leading to performance degradation when models trained on one dataset are tested on another. This challenge, identified as the problems of domain adaptation and domain generalization, has prompted efforts to mitigate it through multi-dataset training [Kothari et al. 2022; Yiu et al. 2019] or synthetic data augmentation [Kim et al. 2019; Nair et al. 2020; Nguyen et al. 2024]. For instance, Yiu et al. [2019] proposed DeepVOG, a fully connected neural network that produces a robust pupil center detection in unseen datasets with minimal accuracy reduction. However, they acknowledge that the network's performance degrades when encountering novel features, such as eyebrows, not present in the training data. These results align with the findings from Nair et al. [2020] which show that CNNs trained for one dataset containing one specific eye tracker can indeed generalize well to previously unseen test data. While these strategies have shown promise, they often require complex training pipelines or access to proprietary data and model weights, which are not always readily available.

In contrast, recent advancements in foundation models have shown potential to address these limitations by enabling effective generalization across datasets without the need for extensive retraining. For example, Maquiling et al. [2024] evaluated the Segment Anything Model (SAM) on OpenEDS datasets [Garbin et al. 2019; Palmero et al. 2020], employing various prompting strategies to segment the sclera, iris, and pupil. Their findings revealed that while explicit prompts—such as combinations of point prompts and bounding boxes—improved sclera and iris segmentation, a single point placed on the pupil sufficed to achieve segmentation performance comparable to specialized models trained on the actual dataset. However, the need to provide individual prompts for each image poses scalability challenges for large datasets such as NVGaze [Kim et al. 2019] or GIW [Kothari et al. 2020], highlighting the necessity for refined pipelines or fine-tuning [Maquiling et al. 2024]. Similarly, Deng et al. [2024] proposed a fully unsupervised pipeline that combines eye priors and image gradients to identify the pupil, iris, and sclera, with SAM refining the segmented boundaries. This approach achieved near-supervised performance levels for pupil and iris segmentation, reaching approximately 90% of benchmark accuracy.

3 Methodology

We evaluated SAM 2's performance on a diverse set of eye tracking datasets to test its generalizability across various domains. These datasets include both VR-based and mobile eye tracking environments, representing controlled and real-world settings. Specifically, we selected four VR-based datasets (including one synthetic) and three mobile eye tracking datasets captured in natural, uncontrolled environments. Two of these datasets are from the OpenEDS challenges [Garbin et al. 2019; Palmero et al. 2020], which focus on creating generalizable and robust semantic segmentations within VR settings. The synthetic NVGaze [Kim et al. 2019] dataset includes its own pupil segmentations for gaze estimation. For the remaining datasets, ground truth segmentations were sourced from TEyeD [Fuhl et al. 2021], the world's largest unified public dataset of eye images taken with head-mounted devices.

Below is a brief description of the datasets used:

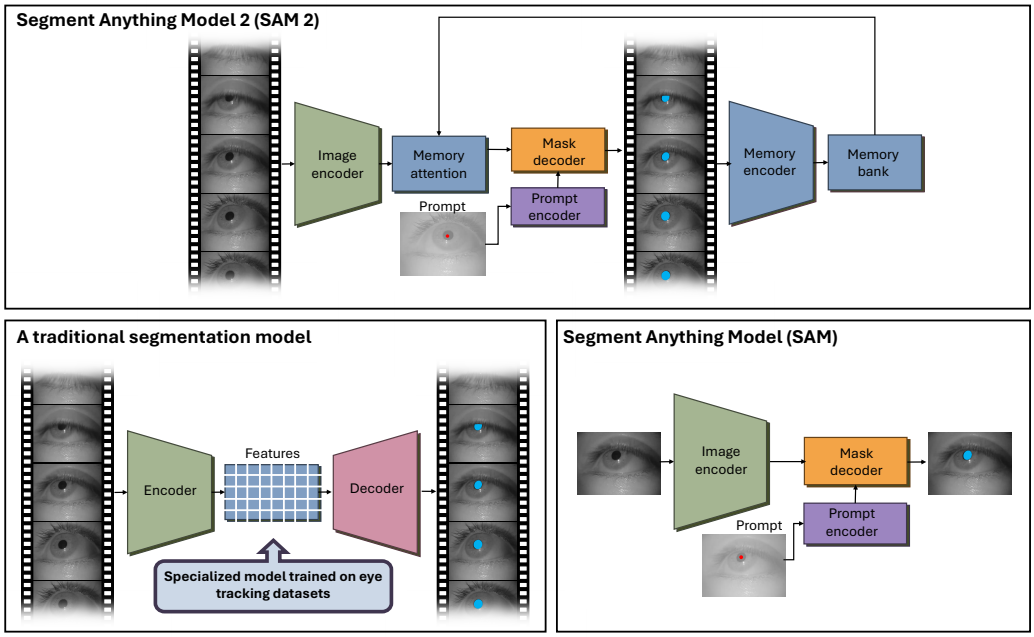

Fig. 2. Comparison between Segment Anything Model 2 (top), a traditional segmentation model trained specifically on eye tracking datasets (left), and Segment Anything Model (right). The sample eye images are taken from the GW dataset [Kothari et al. 2020].

(1) **OpenEDS2019** [Garbin et al. 2019]: Contains 12,759 non-sequential images (400 × 640 pixels) acquired from 152 participants using a VR head-mounted display (HMD) with eye-facing cameras at 200 Hz under controlled lighting. Provides pixel-level annotations for the pupil, iris, and sclera.
(2) **OpenEDS2020** [Palmero et al. 2020]: Features eye-image sequences from 80 participants using a VR HMD at 100 Hz. The Eye Segmentation Dataset includes 200 sequences sampled at 5 Hz totalling to 29,500 images, of which 5% are manually annotated (640 × 400 pixels).
(3) **NVGaze** [Kim et al. 2019]: Comprises two datasets for near-eye gaze estimation under infrared illumination. The real-world dataset includes 264,279 images (640 × 480 pixels) from 14 participants in a VR setting; the synthetic dataset contains 2 million images (1280 × 960 pixels). We evaluated SAM 2 on both.
(4) **Labelled Pupils in the Wild (LPW)** [Tonsen et al. 2016]: Consists of videos from 22 participants recorded in everyday environments using a head-mounted eye tracker at 120 Hz (640 × 480 pixels), covering diverse lighting conditions and natural gaze distributions.
(5) **Gaze-in-Wild (GW)** [Kothari et al. 2020]: Provides naturalistic recordings from 19 participants performing everyday tasks with a mobile eye tracker at 120 Hz (640 × 480 pixels), including eye and head movements, infrared eye images, and scene imagery.
(6) **Dikablis datasets**: A combination of datasets from ElSe [Fuhl et al. 2016b], ExCuSe [Fuhl et al. 2015], PNET [Fuhl et al. 2016a], and a driving study [Kasneci et al. 2014], compiled in TEyeD [Fuhl et al. 2021]. Recorded at 25 Hz (384 × 288 pixels), it features eye recordings from 30 participants.

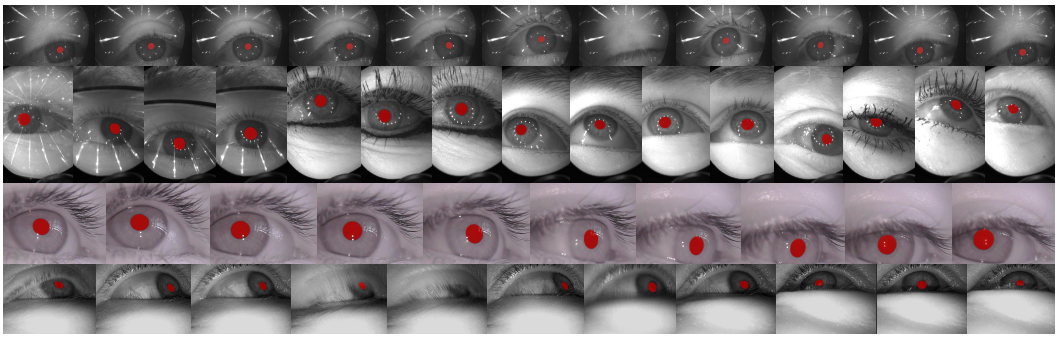

Fig. 3. SAM 2 results on various VR-(first and second rows) and mobile eye tracking (third and last rows) datasets. Images are taken from the OpenEDS2019 [Garbin et al. 2019], OpenEDS2020 [Palmero et al. 2020], LPW [Tonsen et al. 2016], and the Dikablis datasets [Fuhl et al. 2021]. SAM 2 handled occlusions remarkably well as observed in rows 1 and 4, and effectively segmented the pupil across a wide range of datasets, showing its robustness to different eye tracking conditions.

We prompt only one frame for each video or image folder. Maquiling et al. [2024] demonstrated that a single-point prompt is sufficient for accurately segmenting the pupil region. Based on this finding, we decided to limit our approach to using one point prompt on a single frame where the pupil is present. We evaluated SAM 2's performance across three key aspects: pupil segmentation accuracy, pupil tracking consistency (how reliably SAM 2 can track the pupil throughout a video), and blink detection accuracy (how effectively SAM 2 can determine when the pupil is fully occluded (e.g. during a blink event)).

To assess these aspects, we used the following metrics:

- **Intersection-over-union (IoU)**: Defined as the ratio of the overlap area between the predicted segmentation and the ground truth to their total combined area. This metric is calculated only for frames where SAM 2 produces a segmentation in the presence of a pupil in the image.
- **Pupil Lost Rate**: The percentage of frames where the pupil is not successfully tracked.
- **Blink Detection Rate**: The percentage of frames where blink events were correctly identified.

We compare the mean IoU scores with scores achieved by SAM 2 with those reported in prior studies in pupil segmentation, particularly, the best-performing pupil mean IoU by Maquiling et al.[2024] using SAM [Kirillov et al. 2023] with a combination of four point prompts and a bounding box surrounding the pupil, the results from EllSeg-Gen [Kothari et al. 2022] using a model trained on nine publicly available datasets including OpenEDS2019 [Garbin et al. 2019] and NVGaze [Kim et al. 2019], global pupil mean IoU scores from RIT-Eyes [Nair et al. 2020], results from [Kothari et al. 2021] using various models: RITNet [Chaudhary et al. 2019], DenseElNet [Kothari et al. 2021], and DeepVOG [Yiu et al. 2019], results from CondSeg [Jia et al. 2024], as well as the winning leaderboard scores from OpenEDS2019 [Garbin et al. 2019] and OpenEDS2020 [Palmero et al. 2020] which reported averaged mean IoU across all three eye features (pupil, iris, and sclera), and the baseline results from TEyeD [Fuhl et al. 2021] using leave-one-out cross validation for each eye tracker.

4 Results

Table 1 compares the mean IoU of SAM 2 to the reported mean IoU scores from various papers [Garbin et al. 2019; Jia et al. 2024; Kothari et al. 2022, 2021; Maquiling et al. 2024; Nair et al. 2020; Palmero et al. 2020] while Table 2 summarizes SAM 2's pupil lost rate and blink detection rate.

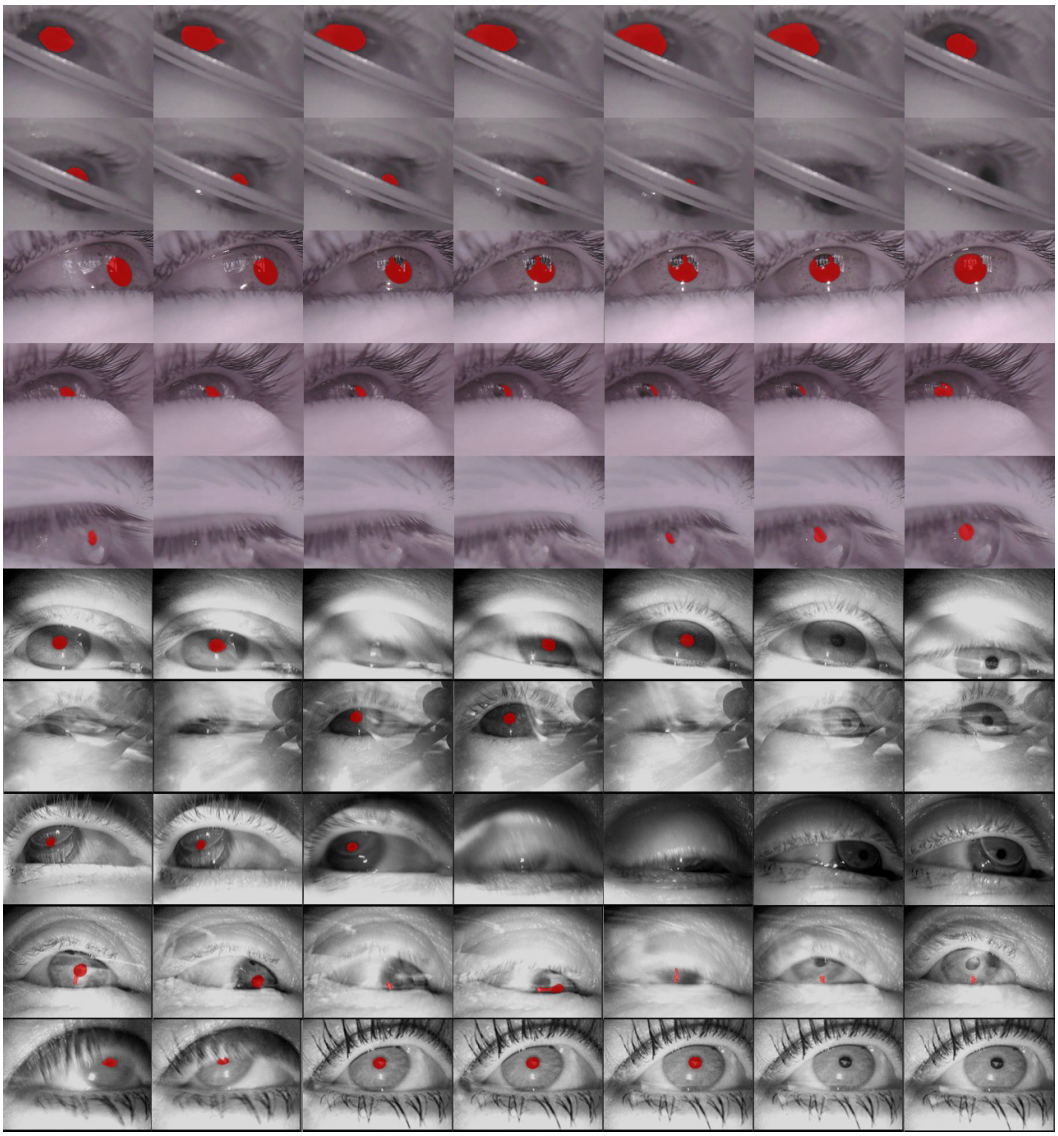

Fig. 4. Noisy examples from mobile eye tracking datasets [Fuhl et al. 2021; Tonsen et al. 2016] where SAM 2 produced low-quality segmentations. Issues include SAM 2 losing track of the pupil (rows 2, 5-8, 10), segmenting the wrong area (row 9), partially segmenting the pupil (rows 3, 4, 5), and oversegmenting the pupil (rows 1, 9).

For instance, on the OpenEDS2019 dataset, SAM 2 achieved a pupil lost rate of 0.75% and a blink detection rate of 94.12% with a mean IoU of 89.97%, competing with specialist models trained on OpenEDS like EllSeg-Gen [Kothari et al. 2022] which reached a mean IoU of 95.6% when trained on multiple datasets including OpenEDS. Similarly, for OpenEDS2020, SAM 2 attained a mean IoU of 92.33%, significantly higher compared to CondSeg [Jia et al. 2024] which reported a mean IoU of 86.80%.

Table 1. Performance of SAM 2 on multiple datasets compared to various published results, including the best-performing scores from Maquiling et al. [2024] on the original SAM [Kirillov et al. 2023] using a combination of bounding box and point prompts for each frame, the all-vs-one scores from EllSeg-Gen [Kothari et al. 2022], global pupil mIoU scores from RIT-Eyes [Nair et al. 2020], results published by Kothari et al. [2021] for RITnet [Chaudhary et al. 2019], DenseElNet [Kothari et al. 2021], and DeepVOG [Yiu et al. 2019], as well as results from CondSeg [Jia et al. 2024]. Leaderboard scores*, which averaged across three eye regions, were also obtained from the OpenEDS2019 [Garbin et al. 2019] and OpenEDS2020 [Palmero et al. 2020] challenge pages, while baseline results from TEyeD were calculated using leave-one-out cross validation for each eye tracker. [Fuhl et al. 2021]. Datasets with a single baseline value are marked with an asterisk (*). As TEyeD provided the ground truth segmentation for NVGaze (real), LPW, GW, and the Dikablis datasets, no baseline mean IoU could be extracted from the original papers.

***Leaderboards: https://eval.ai/web/challenges/challenge-page/353/leaderboard/1002, https://eval.ai/web/challenges/challenge-page/603/leaderboard/1680**

				Mean IoU					
Dataset / Model	SAM 2	SAM	EllSeg-Gen	RIT-Eyes	RITnet	DenseElNet	DeepVOG	CondSeg	OpenEDS
OpenEDS2019	0.8997	0.9330	0.956	0.8926*	0.950	0.954	0.891	0.9091	0.9528
OpenEDS2020	0.9223	0.9334	–	–	–	–	–	0.8680	0.9517
NVGaze (Synthetic)	0.9259	–	0.982	0.8926*	0.932	0.931	0.909	–	–
	SAM 2	TEyeD							
NVGaze (Real)	0.9079	0.65*							
LPW	0.9023	0.65*							
GW	0.9212	0.65*							
Dikablis	0.8835	0.65*							

Table 2. Performance of SAM 2 in detecting Pupil Lost and Blink Detected across multiple datasets.

	Pupil Lost	**Blink Detected**
Dataset / Model	SAM 2	
OpenEDS2019	0.0075	0.9412
OpenEDS2020	0.0104	0.9375
NVGaze (Synthetic)	0.0028	0.9602
NVGaze (Real)	0.0876	0.9889
LPW	0.0940	0.7127
GW	0.0184	0.8513
Dikablis	0.1846	0.9483

On the NVGaze datasets [Kim et al. 2019], SAM 2 performed well on synthetic data with a mean IoU of 92.59% (compared to EllSeg-Gen's 98.2%) and on real data with 90.79%, despite a higher pupil lost rate of 8.76% on the real data. In mobile datasets, SAM 2 achieved mean IoUs of 90.23% on LPW [Tonsen et al. 2016] and 92.12% on GW [Kothari et al. 2020], with higher pupil lost rates due to increased noise and visual obstructions, as evidenced by the Dikablis datasets' pupil lost rate of 18.46% [Fuhl et al. 2021, 2015, 2016a,b; Kasneci et al. 2014].

Overall, SAM 2 [Ravi et al. 2024] performed well on both VR and mobile eye tracking datasets, with VR datasets showing higher performance likely due to more controlled environments. Although SAM 2 slightly underperformed on the nonsequential OpenEDS2019 dataset—composed of individual images rather than continuous video frames—it still demonstrated the ability to generalize across multiple datasets. Notably, SAM 2 outperformed TEyeD's top-performing model, which achieved a mean IoU of just 65% [Fuhl et al. 2021], and delivered competitive results compared

to [Kothari et al. 2022], where a mean IoU of 98.2% was reported on NVGaze's synthetic dataset. It is important to highlight that their model was trained on several datasets, including NVGaze's training set, before being tested on the NVGaze test set. In contrast, SAM 2 achieved these results without any fine-tuning or sacrificing a subset of the datasets for training.

A major advantage of SAM 2 is its minimal human guidance requirement—just a single click per video to indicate that it should segment the pupil—compared to traditional manual annotation that would require thousands of clicks. For example, SAM 2 required only one click for the entire OpenEDS2019 dataset of 12,759 images. Similar reductions were observed across other datasets: 200 clicks (one point prompt per eye-image sequence) for the 29,500 images of the OpenEDS2020 dataset, 66 clicks for LPW's 130,856 images, 54 clicks for NVGaze's 2,264,279 images, 423 clicks for the Dikablis datasets' more than 5.6 million images, and 148 clicks for GW's 6 million images. This significant reduction in human labor highlights SAM 2's efficiency in annotating large datasets with minimal input while achieving high performance in pupil tracking without model fine-tuning or specialized training. Moreover, SAM 2's inference process did not require powerful GPUs, making it feasible for use with low-end hardware. This level of accessibility was not possible before SAM 2 and demonstrates how vision foundation models can democratize large-scale data annotation in eye tracking research.

5 Discussion

The quality of the dataset significantly impacts SAM 2's performance. For instance, datasets with clear eye images, minimal noise and high resolution, such as OpenEDS and NVGaze, yielded higher IoU scores and lower pupil loss rates. However, SAM 2 encountered more difficulty in noisier datasets, like the Dikablis datasets, resulting in more pupil tracking failures.

In terms of human interaction, the effort required is mostly limited to monitoring the quality of the predicted masks and adding additional prompts only when necessary. While we limited our evaluation to a single point prompt per video, SAM 2 supports various other prompt types, such as multiple positive or negative point prompts (where the signs indicate areas that SAM 2 should and should not include in its segmentation) and bounding box prompts, offering flexibility for more complex and noisier datasets that require additional guidance. Below we highlight several practical lessons from conducting our study for researchers interested in implementing SAM 2 for eye tracking data segmentation tasks:

5.1 Practical Lessons Learned

(1) **Significant Time Savings:** SAM 2 significantly reduced the time necessary to annotate entire datasets. The datasets in this study were annotated by two of the authors within a couple of days, demonstrating SAM 2's efficiency in annotating large volumes of data quickly. Importantly, these authors spent most of the time waiting for the model to finish producing its segmentation for the datasets, and only very little time setting up the model and checking its output.

(2) **Reduced Technical Barrier to Entry:** SAM 2 not only saves time in obtaining segmentation masks but also greatly simplifies the process for non-experts. Instead of developing custom pupil segmentation models, users can run SAM 2 with just a few lines of code. This lowers the barrier to entry, enabling more people to develop gaze estimation pipelines.

(3) **Vibrant Open-Source Community:** SAM 2 benefits from an active open-source community (13,500 GitHub stars at the time of writing), which drives innovation through task-specific fine-tuning, pre-trained model sharing, and rapid adaptation across domains such as medical imaging[Zhu et al. 2024] and object detection in remote sensing[Wang et al. 2024]. Unlike specialized eye-segmentation models, which often lack scalability and production-readiness

due to issues such as domain generalization. SAM 2's general-purpose design enables seamless integration into diverse workflows, making it a versatile and approachable alternative for real-world applications.

(4) **Standardizing Results Across Datasets:** SAM 2 enables consistent and comparable results across diverse datasets without requiring separate training for each. Traditional workflows often demand dataset-specific models, which are resource-intensive and limit reproducibility. By leveraging zero-shot capabilities and a universal pre-trained architecture, SAM 2 reduces overfitting risks and improves cross-dataset comparability. For instance, SAM 2 can be directly applied to datasets like NVGaze [Kim et al. 2019], GIW [Kothari et al. 2020], and OpenEDS [Garbin et al. 2019], eliminating the need for retraining and streamlining research workflows.

(5) **Minimal Human Interaction:** The annotation process required minimal human involvement beyond preparing the prompts and finding appropriate frames where the pupil is visible. SAM 2 handled the actual annotation, while the user only needs to perform quality checks and refine prompts when necessary.

(6) **Privacy-preserving Methods:** SAM 2 minimizes the need to collect and transfer sensitive data, aligning with privacy concerns [Zhang and Metaxas 2024]. Traditional eye-tracking models often require high-resolution, person-specific datasets for retraining or validation, which raises privacy risks [Bozkir et al. 2020; Byrne et al. 2024a]. In contrast, SAM 2's pre-trained capabilities eliminate the necessity for large-scale data collection, reducing the handling of sensitive personal information while maintaining performance. This approach ensures that eye-tracking research can adhere to ethical guidelines and protect participant privacy

(7) **No Requirement for Training Data:** SAM 2 eliminates the need for training data, addressing the substantial resource demands of traditional eye-segmentation workflows. Conventional models require extensive datasets that involve recruiting participants, capturing eye-tracking images, and manually annotating features like the pupil, sclera, and iris, all of which are time-consuming and costly. By leveraging SAM 2's pre-trained architecture, researchers can bypass these steps, significantly reducing costs, accelerating research timelines, and minimizing ethical and logistical challenges. This streamlined process enables immediate application to segmentation tasks without sacrificing performance, making eye-tracking research more accessible and efficient.

(8) **Low Hardware Requirements** SAM 2's low GPU requirements make it accessible for researchers with limited computational resources. We used both a high-end NVIDIA A100 (80GB VRAM) and a GeForce RTX 4090 (24 GB VRAM), achieving compute times of up to 40 frames per second (fps) for both. Impressively, SAM 2 also ran on more budget-friendly GPU's such as the RTX 4060 Ti (16 GB VRAM) delivering around 12 fps, and was even functional on laptop-class GPU's, albeit at significantly lower frame rates of just a few fps. However, it is worth noting that inference performance of the model appeared to be limited by CPU and not GPU performance in most of these cases. While models like RITNet [Chaudhary et al. 2019] and DeepVOG [Yiu et al. 2019] offer significantly faster inference speeds, these are task-specific models optimized for real-time pupil tracking. In contrast, SAM 2 is a general-purpose segmentation model that remains functional even on lower-end hardware, making it more accessible for large-scale annotation tasks where real-time inference is not a primary concern.

(9) **Strong Generalization** SAM 2 demonstrated robust performance across a diverse set of eye tracking datasets including VR, mobile, and even synthetic data, despite not being specifically trained on eye tracking data. While domain generalization remains a central goal in gaze estimation, SAM 2 presents a compelling new approach to achieving this objective. Whether

a complete paradigm shift or not, SAM 2 undoubtedly offers a valuable new perspective. Future work could explore the applicability of this approach to other ocular features, such as iris segmentation, and investigate whether fine-tuning on near-eye images leads to further performance gains.

(10) **Dealing with Noisy Data:** While SAM 2 performed well even with occlusions, it may lose track of the pupil, particularly in videos with excessive noise such as those captured "in the wild". For instance, SAM 2's performance on the TEyeD dataset (refer to Figures 3 and 4 for visual examples) demonstrated difficulties in consistently tracking the pupil under such conditions. To address this, further refining of prompts is necessary–this may include adopting a more appropriate prompt strategy, adding prompts on more difficult frames, or focusing on regions where SAM 2 loses track to complete the video sequence–although this approach still involves less human effort compared to traditional annotation processes. Moreover, for more complex eye features, such as the iris and sclera, more advanced prompt strategies may be required. For further insights into effective prompt strategies tailored to specific eye features, refer to the relevant study by Maquiling et al. [2024].

5.2 Open Challenges, Limitations & Future Work

While SAM 2 excelled in pupil segmentation, challenges remain with segmenting less distinct eye features like the iris and sclera. To explore this, we evaluated SAM 2's performance on the OpenEDS2020 dataset, where it achieved an mIoU of 76.53% for the iris and only 7.36% for the sclera with a single box prompt. Fine-tuning SAM 2 on specific eye features, especially under varying conditions like lighting, reflections, and noise, could improve performance by reducing its reliance on ideal conditions and simple prompts.

Additionally, alternative prompting strategies, such as using bounding boxes or multiple point prompts, may yield better results in challenging cases. A limitation of our study is that we did not explore different prompt strategies, opting for a single point prompt to highlight the simplicity of annotation with SAM 2. While a single prompt proved to be effective for pupil segmentation in many cases, other prompts may improve results in difficult cases or for less well defined features.

Another consideration is the possibility that SAM 2 may have encountered these datasets during training as all the datasets we have used are open to the public. Future work should test SAM 2 on completely novel datasets to validate its generalization capabilities. Additionally, as eye tracking moves to consumer devices, a key challenge will be adapting SAM 2 for low-power hardware like smart glasses [Zhang et al. 2023a], making it crucial to balance performance with reduced computational requirements for real-time applications in VR, AR devices.

5.3 Privacy and Ethics

SAM 2 enhances eye-tracking research by reducing the need for manual annotation, improving efficiency. However passing data through this model still involves handling sensitive biometric information. As self-annotated datasets become more feasible, ensuring data sovereignty and informed consent is critical. Researchers must prioritize transparency and safeguard user data, especially when handling large, potentially sensitive datasets.

More generally, the rise of open foundation models like SAM 2 raises new governance challenges [Kapoor et al. 2024]. It has been documented that in fields such as medical imaging, these foundation models can offer significant advantages over task specific models by accelerating model development, reducing reliance on labeled data, thus providing a further safeguard towards preserving patient privacy. While open-sourcing large medical datasets faces regulatory hurdles, foundation models provide an alternative for knowledge sharing while protecting sensitive information [Zhang and Metaxas 2024].

6 Conclusion

In this study, we assessed the practical segmentation capabilities of the SAM 2 Vision Foundation model. Using SAM 2, we efficiently annotated over 14 million pupil images across multiple datasets with just a few click prompts per dataset, significantly streamlining traditional annotation workflows. Our findings show that foundation models like SAM 2 effectively address key challenges in eye tracking research: data annotation, domain adaptation, and reducing training data requirements. Notably, SAM 2 achieves robust performance without fine-tuning, offering a user-friendly and accurate solution compared to its predecessor, SAM. Its ability to annotate entire datasets with minimal human input makes it suitable practically for large-scale applications. Further, it could be used to standardize gaze estimation across datasets, ensuring fair comparisons. This standardization not only supports reproducibility but also facilitates cross-study evaluations in gaze estimation research.

This work highlights the potential for general-purpose models to benefit other HCI fields where extensive labeled data is needed. As these models advance, we anticipate continued progress in both gaze estimation research and broader human-computer interaction applications.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

Ehsan Mohammadi Arvacheh and Hamid R Tizhoosh. 2006. Iris segmentation: Detecting pupil, limbus and eyelids. In *2006 International Conference on Image Processing*. IEEE, 2453–2456.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

Efe Bozkir, Ali Burak Ünal, Mete Akgün, Enkelejda Kasneci, and Nico Pfeifer. 2020. Privacy preserving gaze estimation using synthetic images via a randomized encoding based framework. In *ACM symposium on eye tracking research and applications*. 1–5.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

Sean Anthony Byrne, Nora Castner, Efe Bozkir, Diederick C Niehorster, and Enkelejda Kasneci. 2024a. From Lenses to Living Rooms: A Policy Brief on Eye Tracking in XR Before the Impending Boom. In *2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE, 90–96.

Sean Anthony Byrne, Virmarie Maquiling, Marcus Nyström, Enkelejda Kasneci, and Diederick C. Niehorster. 2023. LEyes: A Lightweight Framework for Deep Learning-Based Eye Tracking using Synthetic Eye Images. arXiv:2309.06129 [cs.CV] https://arxiv.org/abs/2309.06129

Sean Anthony Byrne, Marcus Nyström, Virmarie Maquiling, Enkelejda Kasneci, and Diederick C Niehorster. 2024b. Precise localization of corneal reflections in eye images using deep learning trained on synthetic data. *Behavior Research Methods* 56, 4 (2024), 3226–3241.

Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. 2023. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems* 36 (2023), 25971–25990.

Aayush K Chaudhary, Rakshit Kothari, Manoj Acharya, Shusil Dangi, Nitinraj Nair, Reynold Bailey, Christopher Kanan, Gabriel Diaz, and Jeff B Pelz. 2019. Ritnet: Real-time semantic segmentation of the eye for gaze tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 3698–3702.

Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. 2023. Segment and track anything. *arXiv preprint arXiv:2305.06558* (2023).

Jiangfan Deng, Zhuang Jia, Zhaoxue Wang, Xiang Long, and Daniel K Du. 2024. Towards Unsupervised Eye-Region Segmentation for Eye Tracking. *arXiv preprint arXiv:2410.06131* (2024).

Lei Ding, Kun Zhu, Daifeng Peng, Hao Tang, Kuiwu Yang, and Lorenzo Bruzzone. 2024. Adapting segment anything model for change detection in VHR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

Wolfgang Fuhl, David Geisler, Thiago Santini, Tobias Appel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2018. CBF: Circular binary features for robust and real-time pupil center detection. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*. 1–6.

Wolfgang Fuhl, Gjergji Kasneci, and Enkelejda Kasneci. 2021. Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. In *2021*

IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, 367–375.

Wolfgang Fuhl, Thomas Kübler, Katrin Sippel, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2015. Excuse: Robust pupil detection in real-world scenarios. In *Computer Analysis of Images and Patterns: 16th International Conference, CAIP 2015, Valletta, Malta, September 2-4, 2015 Proceedings, Part I 16*. Springer, 39–51.

Wolfgang Fuhl, Thiago Santini, and Enkelejda Kasneci. 2017a. Fast and robust eyelid outline and aperture detection in real-world scenarios. In *2017 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 1089–1097.

Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, and Enkelejda Kasneci. 2016a. Pupilnet: Convolutional neural networks for robust pupil detection. *arXiv preprint arXiv:1601.04902* (2016).

Wolfgang Fuhl, Thiago Santini, Gjergji Kasneci, Wolfgang Rosenstiel, and Enkelejda Kasneci. 2017b. Pupilnet v2. 0: Convolutional neural networks for cpu based real time robust pupil detection. *arXiv preprint arXiv:1711.00112* (2017).

Wolfgang Fuhl, Thiago C Santini, Thomas Kübler, and Enkelejda Kasneci. 2016b. Else: Ellipse selection for robust pupil detection in real-world environments. In *Proceedings of the ninth biennial ACM symposium on eye tracking research & applications*. 123–130.

Stephan J Garbin, Yiru Shen, Immo Schuetz, Robert Cavin, Gregory Hughes, and Sachin S Talathi. 2019. Openeds: Open eye dataset. *arXiv preprint arXiv:1905.03702* (2019).

Azul Garza, Cristian Challu, and Max Mergenthaler-Canseco. 2023. TimeGPT-1. *arXiv preprint arXiv:2310.03589* (2023).

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Chengzhuo Tong, Peng Gao, Chunyuan Li, and Pheng-Ann Heng. 2024. Sam2point: Segment any 3d as videos in zero-shot and promptable manners. *arXiv preprint arXiv:2408.16768* (2024).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.

Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. 2024. Segment anything model for medical images? *Medical Image Analysis* 92 (2024), 103061.

Zhuang Jia, Jiangfan Deng, Liying Chi, Xiang Long, and Daniel K Du. 2024. CondSeg: Ellipse Estimation of Pupil and Iris via Conditioned Segmentation. *arXiv preprint arXiv:2408.17231* (2024).

Sayash Kapoor, Rishi Bommasani, Kevin Klyman, Shayne Longpre, Ashwin Ramaswami, Peter Cihon, Aspen Hopkins, Kevin Bankston, Stella Biderman, Miranda Bogen, et al. 2024. Position: On the societal impact of open foundation models. *Proc. Machine Learning Res* 235 (2024), 23082–23104.

Enkelejda Kasneci, Katrin Sippel, Kathrin Aehling, Martin Heister, Wolfgang Rosenstiel, Ulrich Schiefer, and Elena Papageorgiou. 2014. Driving with binocular visual field loss? A study on a supervised on-road parcours with simultaneous eye and head tracking. *PloS one* 9, 2 (2014), e87470.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. Minneapolis, Minnesota, 2.

Joohwan Kim, Michael Stengel, Alexander Majercik, Shalini De Mello, David Dunn, Samuli Laine, Morgan McGuire, and David Luebke. 2019. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

Valentin Koch, Sophia J Wagner, Salome Kazeminia, Ece Sancar, Matthias Hehr, Julia A Schnabel, Tingying Peng, and Carsten Marr. 2024. DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 520–530.

Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports* 10, 1 (2020), 2539.

Rakshit S Kothari, Reynold J Bailey, Christopher Kanan, Jeff B Pelz, and Gabriel J Diaz. 2022. EllSeg-Gen, towards Domain Generalization for head-mounted eyetracking. *Proceedings of the ACM on human-computer interaction* 6, ETRA (2022), 1–17.

Rakshit S Kothari, Aayush K Chaudhary, Reynold J Bailey, Jeff B Pelz, and Gabriel J Diaz. 2021. Ellseg: An ellipse segmentation framework for robust gaze tracking. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2757–2767.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications* 15, 1 (2024), 654.

Virmarie Maquiling, Sean Anthony Byrne, Diederick C Niehorster, Marcus Nyström, and Enkelejda Kasneci. 2024. Zero-shot segmentation of eye features using the segment anything model (sam). *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 7, 2 (2024), 1–16.

Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. 2023. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis* 89 (2023), 102918.

Marcelo Mottalli, Marta Mejail, and Julio Jacobo-Berlles. 2009. Flexible image segmentation and quality assessment for real-time iris recognition. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 1941–1944.

Nitinraj Nair, Rakshit Kothari, Aayush K Chaudhary, Zhizhuo Yang, Gabriel J Diaz, Jeff B Pelz, and Reynold J Bailey. 2020. RIT-Eyes: Rendering of near-eye images for eye-tracking applications. In *ACM symposium on applied perception 2020*. 1–9.

Viet Dung Nguyen, Reynold Bailey, Gabriel J Diaz, Chengyi Ma, Alexander Fix, and Alexander Ororbia. 2024. Deep Domain Adaptation: A Sim2Real Neural Approach for Improving Eye-Tracking Systems. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* 7, 2 (2024), 1–17.

Marcus Nyström, Diederick C Niehorster, Richard Andersson, Roy S Hessels, and Ignace TC Hooge. 2023. The amplitude of small eye movements can be accurately estimated with video-based eye trackers. *Behavior Research Methods* 55, 2 (2023), 657–669.

Cristina Palmero, Abhishek Sharma, Karsten Behrendt, Kapil Krishnakumar, Oleg V Komogortsev, and Sachin S Talathi. 2020. Openeds2020: Open eyes dataset. *arXiv preprint arXiv:2005.03876* (2020).

Antonio Peréz, M Luisa Córdoba, A Garcia, Rafael Méndez, ML Munoz, José Luis Pedraza, and F Sanchez. 2003. A precise eye-gaze detection and tracking system. (2003).

Konstantinos Psychogyios, Helen C Leligou, Filisia Melissari, Stavroula Bourou, Zacharias Anastasakis, and Theodore Zahariadis. 2023. Samstyler: Enhancing visual creativity with neural style transfer and segment anything model (sam). *IEEE Access* (2023).

Alec Radford. 2018. Improving language understanding by generative pre-training. (2018).

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 234–241.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. 2018a. PuRe: Robust pupil detection for real-time pervasive eye tracking. *Computer Vision and Image Understanding* 170 (2018), 40–50.

Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. 2018b. PuReST: Robust pupil tracking for real-time pervasive eye tracking. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*. 1–5.

Anna Christina Schaar, Alejandro Tejada-Lapuerta, Giovanni Palla, Robert Gutgesell, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Mojtaba Bahrami, et al. 2024. Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv* (2024), 2024–04.

Siddharth Shankar, Leigh A Stearns, and CJ van der Veen. 2023. Semantic segmentation of glaciological features across multiple remote sensing platforms with the Segment Anything Model (SAM). *Journal of Glaciology* (2023), 1–10.

Qiuhong Shen, Xingyi Yang, and Xinchao Wang. 2023. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261* (2023).

Mark R Shortis, Timothy A Clarke, and Tim Short. 1994. Comparison of some techniques for the subpixel location of discrete target images. In *Videometrics III*, Vol. 2350. SPIE, 239–250.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 843–852.

Lech Świrski, Andreas Bulling, and Neil Dodgson. 2012. Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the symposium on eye tracking research and applications*. 173–176.

Marc Tonsen, Xucong Zhang, Yusuke Sugano, and Andreas Bulling. 2016. Labelled pupils in the wild: a dataset for studying pupil detection in unconstrained environments. In *Proceedings of the ninth biennial ACM symposium on eye tracking research & applications*. 139–142.

Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. 2024. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems* 36 (2024).

Jun Yan, Pengyu Wang, Danni Wang, Weiquan Huang, Daniel Watzenig, and Huilin Yin. 2024. Segment-anything models achieve zero-shot robustness in autonomous driving. In *2024 IEEE International Automated Vehicle Validation Conference (IAVVC)*. IEEE, 1–8.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. 2023a. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968* (2023).

Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. 2023b. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908* (2023).

Yuk-Hoi Yiu, Moustafa Aboulatta, Theresa Raiser, Leoni Ophey, Virginia L Flanagin, Peter Zu Eulenburg, and Seyed-Ahmad Ahmadi. 2019. DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning. *Journal of neuroscience methods* 324 (2019), 108307.

Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790* (2023).

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. 2023a. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289* (2023).

Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, and Choong Seon Hong. 2023b. Mobilesamv2: Faster segment anything to everything. *arXiv preprint arXiv:2312.09579* (2023).

Shaoting Zhang and Dimitris Metaxas. 2024. On the challenges and perspectives of foundation models for medical image analysis. *Medical image analysis* 91 (2024), 102996.

Yichi Zhang, Zhenrong Shen, and Rushi Jiao. 2024. Segment anything model for medical image segmentation: Current applications and future directions. *Computers in Biology and Medicine* (2024), 108238.

Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. 2023. Fast segment anything. *arXiv preprint arXiv:2306.12156* (2023).

Zihao Zhao. 2023. Enhancing Autonomous Driving with Grounded-Segment Anything Model: Limitations and Mitigations. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*. IEEE, 1258–1265.

Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. 2023. A foundation model for generalizable disease detection from retinal images. *Nature* 622, 7981 (2023), 156–163.

Jiayuan Zhu, Yunli Qi, and Junde Wu. 2024. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874* (2024).