Vision-language models for decoding provider attention during neonatal resuscitation

Felipe Parodi[1], Jordan K. Matelsky[2,8], Alejandra Regla-Vargas[3], Elizabeth E. Foglia[6,7], Charis Lim[6,7], Danielle Weinberg[6,7], Konrad P. Kording[1,2], Heidi M. Herrick[6,7,†], Michael L. Platt[1,4,5,†]

[1]Department of Neuroscience, [2]Department of Bioengineering, [3]Department of Sociology,
[4]Department of Marketing, [5]Department of Psychology, University of Pennsylvania,
[6]Division of Neonatology, Department of Pediatrics, University of Pennsylvania Perelman School of Medicine,
[7]Division of Neonatology, Department of Pediatrics, Children's Hospital of Philadelphia,
[8]Johns Hopkins University Applied Physics Laboratory

Correspondence: herrickh@chop.edu; fparodi@pennmedicine.upenn.edu

Abstract

Neonatal resuscitations demand an exceptional level of attentiveness from providers, who must process multiple streams of information simultaneously. Gaze strongly influences decision making; thus, understanding where a provider is looking during neonatal resuscitations could inform provider training, enhance real-time decision support, and improve the design of delivery rooms and neonatal intensive care units (NICUs). Current approaches to quantifying neonatal providers' gaze rely on manual coding or simulations, which limit scalability and utility. Here, we introduce an automated, real-time, deep learning approach capable of decoding provider gaze into semantic classes directly from first-person point-of-view videos recorded during live resuscitations. Combining state-of-the-art, real-time segmentation with vision-language models, our low-shot pipeline attains 91% classification accuracy in identifying gaze targets without training. Upon fine-tuning, the performance of our gaze-guided vision transformer exceeds 98% accuracy in semantic gaze analysis, approaching human-level precision. This system, capable of real-time inference, enables objective quantification of provider attention dynamics during live neonatal resuscitation. Our approach offers a scalable solution that seamlessly integrates with existing infrastructure for data-scarce gaze analysis, thereby offering new opportunities for understanding and refining clinical decision making.

1. Introduction

Neonatal resuscitation is a complex process in which a lead provider is tasked with overseeing the resuscitation progression, monitoring vital signs, and coordinating team response, often within the confines of a bustling delivery room [13, 44, 45]. Even a momentary attentional lapse can escalate the risk of errors and adverse outcomes [50], making it imperative to identify sources of inefficiency and care disruptions [13]. Quantitative assessment of visual attention not only aids in pinpointing sources of inefficiency, but also advances patient care, improves training protocols for medical practitioners, [21, 46, 47], and bolsters real-time decision support [34, 37].

Traditionally, monitoring provider visual attention during resuscitation relied on manual annotations of egocentric videos, primarily captured via head-mounted eye-tracking systems. These tools have been deployed in both simulated [8, 10, 15, 23, 34] and real-world settings [12, 19, 44, 51]. While valuable insights have been gained through these approaches, including discernible patterns of visual attention on infants, monitors, and team members [39, 44], the manual data extraction is time intensive and unscalable. Integration of eye-tracking glasses has partially addressed these gaps, offering a glimpse into physician gaze by providing the location for object fixations and saccades during a given session [19, 38] or for evaluating factors in accessing neonatal equipment [6]. Prior work demonstrates multiple links relating visual search patterns to levels of expertise, dwell times, and eye movements [5, 23, 36, 44]. Developing a fast, robust, automated system capable of performing semantic gaze analysis is, therefore, a priority [33, 48].

[†]Heidi Herrick and Michael Platt contributed equally to this work.

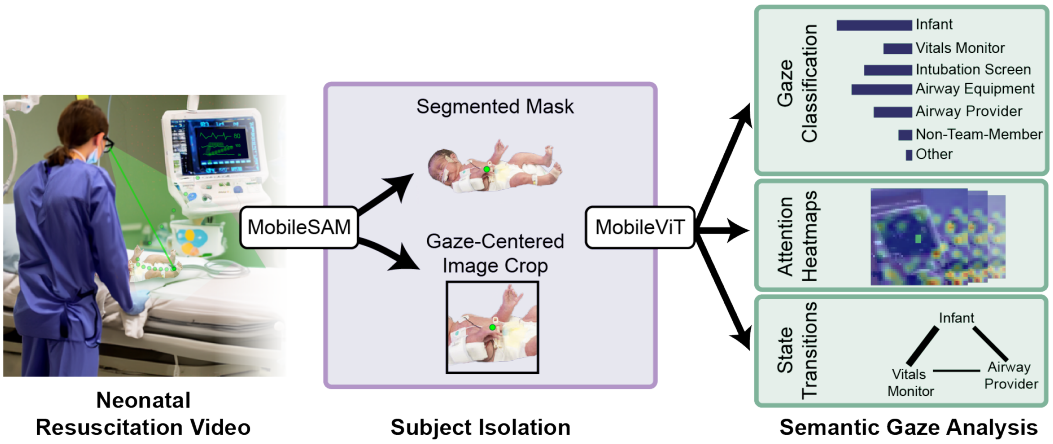

Figure 1. **Approach.** (Left) During resuscitation, physicians must attend to multiple stimuli at once. (Middle) The output of Tobii eye-tracking glasses can be used to isolate the subject with segmentation (top) and cropping (bottom). (Right) Cropped images and object masks are then fed to the model for semantic gaze classification, and prediction scores are aggregated for each target for attention analysis. Note: depicted infant is synthetic.

Such a semantic gaze analysis system should decipher the natural language labels associated with a provider's gaze during complex video scenes, even in situations where data availability is restricted due to privacy considerations in medical settings. Automated gaze analysis would enhance medical education and healthcare, fostering optimal attention strategies and improving the efficacy of neonatal resuscitations by providing nuanced feedback on gaze patterns, optimal team configurations, and task allocations.

Here, we introduce a real-time, data-driven pipeline that automates the analysis of provider visual attention patterns during neonatal resuscitations. Our system first isolates objects of interest using real-time instance segmentation [52] and cropping, which are then jointly classified into various semantic labels by a vision transformer, including MobileViT [24] and CLIP [27], with a top-3 accuracy – the percentage of samples for which the true label is among the top three predicted labels – reaching 98%. Our pipeline, trained on a novel egocentric NICU dataset, integrates outputs from commercial eye trackers and can operate in real time. This approach, validated against human experts, enables an unprecedented level of precision and efficiency in identifying gaze targets among clinically significant regions of interest (ROIs).

2. Related Work

Deep learning applications in healthcare. Recent years have seen a proliferation of deep learning applications in healthcare. These include deployment of segmentation in radiology to isolate organs of interest from X-ray images [16, 20, 41], use of image classification algorithms to categorize diseases [14], and implementation of gaze estimation in the operating room to understand surgical decision making [18]. Despite these strides, characterizing physician gaze through deep learning remains challenging, constrained by the limited availability of annotated datasets and by the lack of effective, low-shot models. Recent studies highlight the utility of zero-shot vision-language models like CLIP in clinical settings [1, 25, 27], endorsing the potential of heavily pre-trained deep learning models for gaze analysis in data-scarce healthcare settings.

Among emerging tools in deep learning for health, Vision Transformers (ViT) have ushered in a new era of medical analysis [9]. Various ViT variants have been developed, focusing on enhancing generalizability, reducing latency, and improving data-cost-effectiveness [9, 24, 35] in data-austere environments, such as in predicting COVID-19 from chest X-ray images [26]. The promise of these models invites further research extending them to real-world settings.

Gaze tracking in healthcare. Gaze tracking technology has permeated the healthcare sector, augmenting medical image interpretation and enhancing the diagnostic, treatment, and monitoring processes by providers [20, 37, 43, 51]. This technology has proven useful in high-risk domains such as childbirth and neonatal resuscitation, offering critical insights into the interactions between individuals and their surroundings. There has also been a surge in the adoption of semi-automated provider gaze tracking technology, integrating eye-tracking glasses and multi-modal approaches to quantify providers' attention during medical procedures [31, 32]. These techniques, however, have not yet been extended to real-time operation or semantic gaze analysis. Addressing these limitations, recent efforts have sought to incorporate eye-tracking data into deep learning models to provide an interpretable analysis of visual attention patterns [16, 22, 31, 40, 42]. Despite these promis-

Table 1. Breakdown of the Egocentric Dataset of Infant Resuscitation by physician expertise and image distribution by class and split.

Physician Recording	Length (min:sec)	Train:Val	Airway Equip.	Airway Prov.	Laryng. Screen	Infant	Vitals Monitor	Non Team	Other
Attending_1	02:19	2,379: 567	440	171	382	592	815	206	340
Fellow_8	01:02	1,088: 298	64	21	0	799	498	0	4
Attending_26	01:28	1,175: 274	313	2	356	636	106	0	36
Attending_29	02:11	2,424: 603	446	141	539	1,474	197	16	214
Fellow_30	01:09	1,293: 326	95	36	1164	149	153	0	22
Attending_44	03:31	3,387: 873	840	347	503	2,170	115	39	246
Attending_31	04:01	Held out for testing							
Fellow_56	01:00	Held out for testing							
Fellow_62	05:51	Held out for testing							

ing trends, the transition from controlled experiments, including simulations, to real-world clinical settings remains a formidable challenge.

Semantic gaze analysis. Semantic gaze analysis decodes the objects of gaze fixations into natural language, offering a deeper understanding of observer intent, situational awareness, and high-level decision-making processes. Despite burgeoning interest in this domain, current studies focus on analyzing gaze patterns in simulated environments [23, 49]. A significant gap persists in extending these analyses to real-world clinical settings, especially in high-stakes environments like the delivery room and neonatal intensive care unit.

3. Approach

Overview. Here, we introduce a novel framework that integrates in-situ eye-tracking with state-of-the-art neural networks to generate human-interpretable labels for the target of a physician's gaze during neonatal resuscitation (see Fig. 1 (Left) for a depiction of the eye-tracking gaze estimate). Leveraging the eye-gaze estimate, our pragmatic approach enables real-time gaze characterization in active clinical settings, extending the boundaries of automated and accurate gaze analysis.

Egocentric Dataset of Infant Resuscitation. We recorded nine neonatal resuscitation sessions using Tobii Pro eye-tracking glasses (Tobii Pro, Stockholm, Sweden), which captured high-resolution video at 25 FPS. Six videos contributed to the training dataset, and three were held out for testing (Table 1). Before data collection, eye-tracking calibration was performed for each wearer to ensure accurate gaze estimation. These recordings received expert annotations to serve as benchmarks for evaluating various gaze classification models under zero-shot, few-shot, and fine-tuned conditions. The final EDIR dataset comprises 14,687 unique frames (crop-mask pairs), capturing first-person provider perspectives during neonatal resuscitation in the NICU (Fig. 1), serving as a critical resource for model deployment in an authentic clinical context. The University Institutional Review Board approved this study, and informed consent was obtained from study participants.

Annotations. Before cropping and segmentation, the videos were labeled by annotators using either the Tobii coding software or the DeepEthogram tool [2]. These annotators consisted of one expert neonatologist with experience in neonatal resuscitation and two graduate students trained to identify regions of interest. Inter-rater reliability was quantified using Cohen's Kappa, yielding a coefficient of 0.92, indicating substantial agreement among annotators. Annotators labeled the frame into one or more of seven distinct categories: *Infant*, *Vitals Monitor*, *Video Laryngoscope Screen*, *Airway Equipment*, *Airway Provider*, *Non-Team Member*, and *Other Physical Objects*. Annotators used a two-inch radius around the gaze estimate for context-dependent interpretation, ensuring accurate depiction of gaze focus. The annotated image dataset was then split 80:20 across frames, resulting in 11,746 training frames (not including their segmented pairs), and 2,941 testing frames (not including their segmented pairs), ensuring that frames from each video were distributed across both training and testing sets.

Identifying optimal input resolution. We evaluated several input types, including raw frames and various cropped and masked images, to determine the most effective setup for gaze classification (Table 2). Using OpenCV [3], we centered our crops around the Tobii gaze estimate by identifying the pixel with the highest green intensity. The combination of 128x128 pixel crops and segmentation masks [17, 52, 54] was found to optimize zero-shot classification accuracy, informing the parameters for our Egocentric Dataset.

Instance Segmentation with MobileSAM. To achieve real-time, accurate object segmentation, we utilized MobileSAM, a model recognized for its low latency and robust performance [52]. MobileSAM offers the flexibility to use various inputs for mask generation, including bounding boxes, text, or even pixel coordinates. We used a pretrained

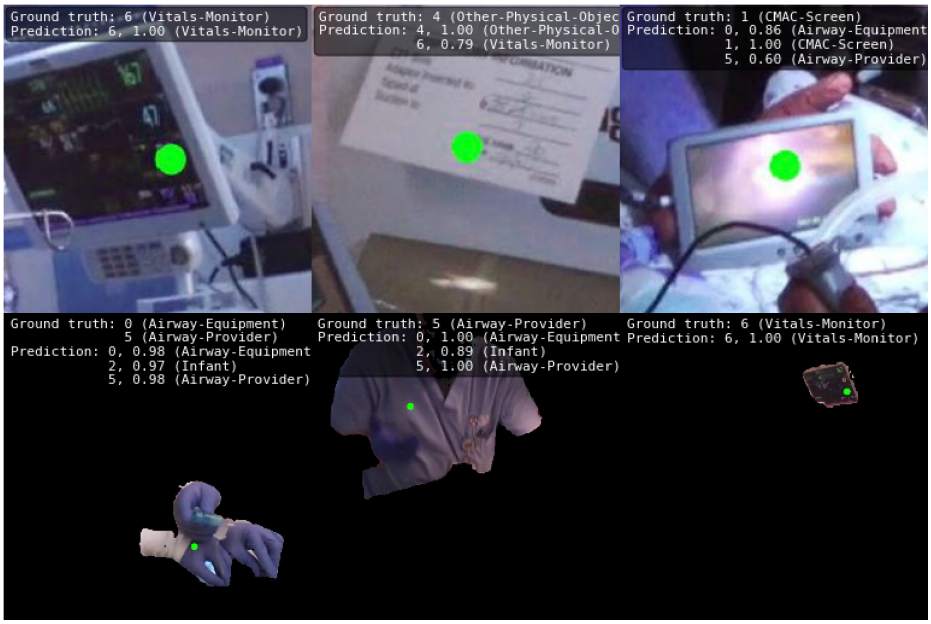

Figure 2. **Sample gaze classification predictions** on cropped (top) and segmented (bottom) testing images. Note: "CMAC-Screen" refers to "Video Laryngoscope Screen."

MobileSAM model to ensure that our gaze-classification pipeline was independent of segmentation accuracy. While this decision simplified our pipeline, it did introduce variability in the quality of the pixel masks, potentially affecting the model's ability to learn specific semantic labels accurately (see Table 5). During segmentation, MobileSAM generated a segmented object mask using the Tobii estimate as input (see Fig. 1 (middle) or Fig. 2 (bottom) for example masks). These masks delineate the physician's focus for a given frame, isolating the region for subsequent analysis.

Low-Shot Semantic Gaze Classification. To address the challenge of gaze classification in the data-scarce neonatal intensive care unit, we leveraged the CLIP (Contrastive Language-Image Pre-training) model [27], a vision-language model adept at aligning image-text representations, making it effective for zero-shot classification. We employed the 'base' vision transformer architecture with 32x32 patches (CLIP-ViT-B-32), which was pre-trained on the LAION-400M image-text dataset [29]. When performing zero-shot classification, provided class labels are embedded using CLIP's heavily pre-trained text encoder, and the similarity between image and text embeddings is computed, ultimately "predicting" the class exhibiting the highest similarity score. We tested the CLIP-ViT-B-32's zero-shot gaze classification capability on our EDIR at different resolutions: the entire frame, a 128x128 pixel crop centered around the Tobii gaze estimate, a 256x256 pixel crop, and the segmentation mask produced with the Tobii gaze estimate as input to MobileSAM.

Following zero-shot gaze classification, we tested how well CLIP could accurately classify a given cropped or segmented frame under low-shot conditions, in which the model would see only a small set of images from the training set and then perform inference. To do this, we relied on Tip-Adapter [53], which enhances CLIP's few-shot ability by creating a feature adapter from a few-shot (16-image) training set to update CLIP's prior encoded knowledge. Akin to our zero-shot experiments, we tested CLIP's low-shot performance on the 128x128 pixel crop, and on the joint crop-mask pair (Table 2).

We evaluated CLIP's performance with Top-1 and Top-3 accuracy on the EDIR testing set (n=2,941). Top-1 accuracy represents the proportion of instances where the true label matches the highest predicted label, whereas Top-3 accuracy accounts for cases where the true label is within the top three predicted labels. This zero- and few-shot paradigm facilitates semantic gaze target prediction without necessitating training on labeled EDIR images, instead harnessing CLIP's generalized knowledge. In subsequent sections, we fine-tune CLIP on EDIR to enhance gaze classification. However, this low-shot evaluation serves as a compelling baseline, showcasing the potential of multi-modal representation learning in data-scarce environments.

Fine-Tuned Semantic Gaze Classification. We next fine-tuned a set of models on the EDIR dataset. Given that zero-shot gaze classification performance was strongest when combining predictions from both the 128x128 pixel crop and segmentation mask inputs, we opted to use this dual input approach for all subsequent few-shot training experiments given an input image size 224x224 px. This al-

Table 2. Semantic gaze prediction under training-free conditions.

Model	Classification	Input	Top-1 Acc (%)	Top-3 Acc (%)
CLIP-ViT-B-32	Zero-Shot	Frame	8.96	38.39
CLIP-ViT-B-32	Zero-Shot	Crop_{128}	36.93	62.22
CLIP-ViT-B-32	Zero-Shot	Crop_{256}	37.92	49.39
CLIP-ViT-B-32	Zero-Shot	Mask	23.15	53.67
CLIP-ViT-B-32	**Zero-Shot**	**Crop_{128} + Mask**	**37.92**	**76.10**
CLIP-ViT-B-32	Zero-Shot	Frame + Crop_{128} + Mask	37.92	56.45
Tip-Adapted-CLIP	**Few-Shot**	**Crop_{128}**	**71.17**	**91.67**
Tip-Adapted-CLIP	Few-Shot	Crop_{128} + Mask	54.55	84.31

lowed us to leverage the strengths of both localized cropping and precise object masking while maintaining consistency across conditions to enable fair comparison between zero-shot and few-shot settings. For fine-tuned gaze classification, we trained three models for both single-label classification, in which there is only one ground truth label per image, and multi-label classification, in which there may be multiple labels per image.

Single-Label Gaze Classification. We trained ResNet50 [11], MobileViT [24], and CLIP-ViT-B-32 [27] using the mmpretrain library [7]. For this task, we incorporated several data augmentations into our training set, including horizontal, vertical, and diagonal flipping. We chose the ResNet-50 as a convolutional baseline and chose the MobileViT model due to its lightweight yet robust performance in diverse computer vision tasks [9, 24, 35]. The ResNet-50 and MobileViT were trained using Stochastic Gradient Descent (SGD), with a learning rate of 0.1, momentum of 0.9, and a weight decay of 0.0001 with a MultiStep Learning Rate Scheduler modulated the learning rate. The CLIP model was trained using a linear and then cosine annealing learning rate. For single-label gaze classification, we minimized cross-entropy loss and evaluated the models using Top-1 and Top-3 accuracy (Table 4).

Multi-label Gaze Classification. We next trained the same set of models – ResNet50, MobileViT, and CLIP-ViT-B-32 – on multi-label gaze classification, in which there can be one or more ground truth labels per image. In complex scenes, such as during infant intubation, the neonatologist must navigate multiple stimuli at once, in which case there may be several areas of interest. For multi-label classification, we minimized the binary cross-entropy loss. We held training conditions constant for ResNet-50 and MobileViT, except for the change of loss function and number of ground truth labels per image. For the multi-label classification task, the traditional vision transformer classification head was replaced with a Multi-Label Linear Classification head with a sigmoid activation function and was trained using the AdamW optimizer. Following training, each model was evaluated using Mean average precision (mAP) and F1-score. Mean average precision (mAP) calculates precision and recall over varying thresholds, balancing the impact of false positives and negatives: a higher mAP indicates better overall performance. The F1-score is the harmonic mean of precision and recall, and offers an insight into the balance achieved between the two, especially vital when dealing with class imbalances, as tends to be the case in data-scarce environments.

Table 3. Semantic gaze prediction following supervised training for single- and multi-label classification.

Model	Input	Accuracy (%)	
	Single-Label	*Top-1*	*Top-3*
ResNet50	Crop_{128} + Mask	81.60	96.62
MobileViT	**Crop_{128} + Mask**	**93.02**	**98.74**
CLIP-ViT-B-32	Crop_{128} + Mask	87.54	97.19
	Multi-Label	*mAP*	*F1-Score*
ResNet50	Crop_{128} + Mask	87.72	77.68
MobileViT	**Crop_{128} + Mask**	**96.71**	**91.60**
CLIP-ViT-B-32	Crop_{128} + Mask	92.39	85.70

4. Results

Low-shot and fine-tuned models approach expert-level gaze classification. We assessed the CLIP-ViT-B-32 model's performance in a zero-shot setting, in which the model was not trained on the infant resuscitation dataset, across multiple input types. Using Top-1 and Top-3 accuracy metrics, we found that inputting only the raw frame for classification led to a dismal Top-1 accuracy of 8.96% and a Top-3 accuracy of 38.39%. However, performance increased when we used cropped images at a 128-pixel radius around the Tobii gaze estimate (Crop_{128}), achieving a Top-1 accuracy of 36.93% and a Top-3 of 62.22%. Likewise, Crop_{256} exhibited a similar trend, with a slight uptick in Top-1 to 37.92%, although with a decline in Top-3 accuracy to 49.39%. Incorporating the object segmentation masks, either alone or in conjunction with cropping, dramatically

Table 4. Model training and evaluation.

Task	Model	Pretraining	Batch Size	Epochs	Loss	Evaluation
Single-label	ResNet-50	ImageNet-1k [28]	32	100	Cross-Entropy	Top-1, Top-3
Single-label	MobileViT (s)	ImageNet-1k	128	100	Cross-Entropy	Top-1, Top-3
Single-label	CLIP-ViT-B-32	LAION-400M [29]	128	300	Cross-Entropy	Top-1, Top-3
Multi-label	ResNet-50	ImageNet-1k	32	100	Binary Cross-Entropy	mAP, F1
Multi-label	MobileViT (s)	ImageNet-1k	128	100	Binary Cross-Entropy	mAP, F1
Multi-label	CLIP-ViT-B-32	LAION-400M	128	300	Binary Cross-Entropy	mAP, F1

increased accuracy; specifically, the "Crop_{128} + Mask" configuration reached an impressive Top-3 accuracy of 76.10%. Consequently, the joint input of cropping and segmentation masking yielded the most promising zero-shot gaze classification. During the few-shot learning phase, we assessed the performance of tip-adapted CLIP [53] on the test set using either the cropped image or the crop-mask pair as input. Remarkably, we found that with only 16 "featured" (no training involved) images, the feature adapter boosted CLIP's Top-1 accuracy to 71.17% and Top-3 accuracy to 91.67%.

We next fine-tuned a ResNet50, MobileViT, and CLIP-ViT on the EDIR image dataset under the "Crop_{128} + Mask" resolution setting with limited training data. In the single-label scenario, where each image has only one ground truth label, MobileViT outperformed the other two models, achieving a Top-1 accuracy of 93.02% and a Top-3 accuracy of 98.74%. In comparison, CLIP-ViT-B-32 and ResNet50 yielded Top-1 accuracies of 87.44% and 81.60%, respectively. In the multi-label case, where multiple ground truth labels are possible, MobileViT again excelled, registering an mAP of 96.71% and an F1-score of 91.60%. This finding underscores the model's adeptness at learning from both cropped images and segmentation masks even for myriad ground truth labels. CLIP-ViT-B-32 and ResNet50 followed with mAPs of 92.39% and 87.72%, and F1-scores of 85.70% and 77.68%, respectively. Collectively, these findings endorse the utility of vision transformers for learning from a sparse dataset (e.g., 6 videos, each under 3 min.) to boost the accuracy and dependability of gaze analysis in neonatal care settings (see Fig. 2 for example predictions by MobileViT).

Table 5. Test performance of semantic gaze classification under different segmentation models with MobileViT. Note: the class inferencer was trained on MobileSAM masks.

Segmentation	**Input**	**Accuracy (%)**	
	Single-Label	*Top-1*	*Top-2*
MobileSAMv2	**Crop_{128} + Mask**	**95.53**	**96.08**
FastSAM-x	Crop_{128} + Mask	90.42	93.34
FastSAM-s	Crop_{128} + Mask	89.14	92.06
ViT-SAM-B	Crop_{128} + Mask	92.97	95.71
ViT-SAM-L	Crop_{128} + Mask	91.51	95.35
ViT-SAM-H	Crop_{128} + Mask	92.06	94.89
No segmentation	Crop_{128} + Mask	90.24	—

Class activation maps visualize the model's attention. After model training, we employed grad-CAM, a technique for generating visual explanations in computer vision, to inspect the class activation maps (CAMs) for each model [30], using testing images. In neural networks, "activation" reveals how specific input regions influence the model's weights; CAMs pinpoint areas deemed crucial by the model for classification. Specifically, grad-CAM computes gradients of the target class score relative to feature map activations, resulting in a localization map that highlights vital regions for prediction. This map is superimposed on the input image, providing a visualization of the model's decision-making rationale (Fig. 3). Notably, both MobileViT and the few-shot trained CLIP models showcased sharply focused heatmaps, signaling exceptional gaze classification precision. In contrast, the ResNet50 and zero-shot CLIP models produced more dispersed activation maps, reflecting diminished performance in this context.

Automated Pipeline Accurately Captures Neonatologist Gaze Dynamics. We next evaluated the prediction capabilities of the best-performing model – the MobileViT – on single-label semantic gaze analysis. We first used the model to run inference on a held-out test video (Fellow_56), whose ground-truth annotations we had. This model yielded a Top-1 accuracy of 95% for this video. For each of the six predicted classes – six because there was no "Non-Team-Member" present in the video – we computed its relative frequency in the ground truth and predicted data. To assess whether the observed and expected frequencies were significantly different on a per-class basis, we performed a z-test for two proportions. Specifically, for each class, we evaluated whether the difference between the observed and expected proportions was statistically significant. We found that all but two of the classes were not significantly different from one another – that is, the predictions of our semantic gaze classification model were statistically equivalent to those in the ground truth dataset ($p > 0.05$; z-statistic: 0.299). When controlling for multiple statistical tests with the Bonferroni correction, in which we di-

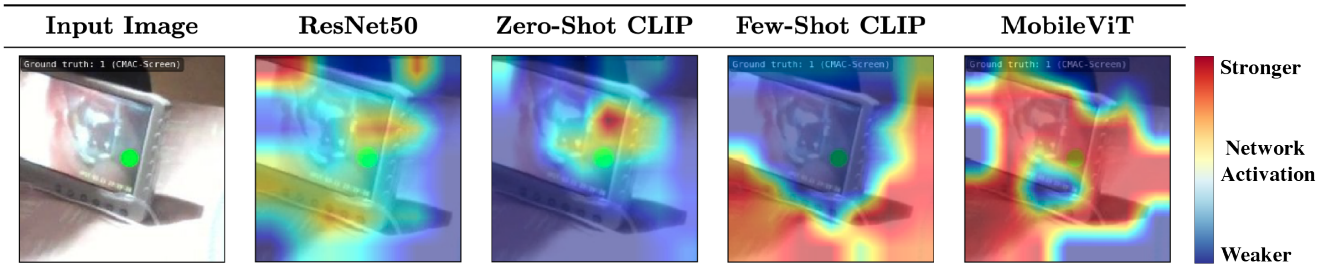

Figure 3. **Model class activation maps with GradCAM on the Laryngoscope Screen.** Each heat map conveys where the model is "looking" in this example image, where each model correctly predicted the class label. Less accurate models, like the ResNet-50, have more diffuse heat maps whereas the higher-performing fine-tuned CLIP (middle) and the MobileViT (far-right) models have heat maps concentrated on the object of attention, the Laryngoscope screen.

vided the threshold for significance by the number of classes tested, we found that the predicted relative frequency of all classes was not statistically different from the ground truth (Figure 4). This suggests that our pipeline can precisely and automatically classify semantic gaze from eye-tracking video alone. To visualize neonatologist visual attention, we computed the transition matrix between classes and plotted the gaze transitions, with scaled nodes and edges for those classes that had greater transitions. Finally, we plotted visual attention throughout resuscitation (Fig 4c).

Real-Time Semantic Gaze Classification Enables Clinical Integration. To assess the computational efficiency of the machine learning models, we conducted inference speed experiments of our three models – ResNet-50, MobileViT, and CLIP-ViT-B-32 – across multiple hardware platforms. The tested hardware configurations included: NVIDIA RTX A6000 GPU; NVIDIA RTX 3080 GPU; AMD Ryzen Threadripper PRO 5975WX CPU; and 11th Gen Intel Core i7-11700K CPU. To capture the model's adaptability to real-time and batch processing scenarios, each model was evaluated under two different batch sizes: a single-instance batch (BS=1) and a batch of eight instances (BS=8). For each batch size and hardware combination, inference speed was measured in frames per second (FPS). To mitigate the effects of background processes, no other significant tasks were executed concurrently. Additionally, the models were allowed a "warm-up" period to ensure that all components were operating at their peak capabilities. Following the completion of 10 repetitions per configuration, we computed the average FPS of each model, which is reported in Table 1. The ResNet50 model led in speed across most configurations, reaching 180.29 FPS and 88.35 FPS on the NVIDIA RTX A6000 and RTX 3080 GPUs, respectively, with single-instance batches (BS = 1). MobileViT also demonstrated efficiency, exceeding the 25 FPS frame rate of the Tobii eye-tracking glasses on three of the four hardware platforms tested. For instance, it clocked 138.39 FPS and 69.39 FPS on the NVIDIA RTX A6000 and RTX 3080 GPUs, respectively. Even with a batch size of eight (BS = 8), MobileViT maintained an average speed of 41.16 FPS on the AMD Ryzen Threadripper PRO 5975WX CPU. Expert human annotators in our study labeled frames at rates ranging from 0.5 to 5 FPS, an order of magnitude slower than even our least efficient models. This highlights the potential of our approach for initial classification tasks. While ResNet50 may boast the highest raw speed, MobileViT offers a balanced profile of speed and adaptability across diverse hardware, making it ideally suited for real-time decision support in clinical settings. Our system is thus well-equipped for real-time semantic gaze classification and visual attention quantification.

Table 6. Average inference speed (frames per second) for batch sizes 1 and 8 across hardware, rounded to the nearest integer.

	Model	A6000	RTX 3080	i7-11700K
BS 1	MobileViT	138 FPS	69	11
	ResNet50	180	88	16
	CLIP-ViT-B-32	92	59	24
BS 8	MobileViT	158	87	11
	ResNet50	223	117	19
	CLIP-ViT-B-32	101	68	27

5. Discussion

Conclusions. Here, we report significant progress in semantic gaze analysis with eye-tracking during neonatal resuscitation. Unlike prior work that primarily focused on simulation-based studies or educational applications [15, 21], our method demonstrates highly accurate automated analysis of provider gaze patterns, achieving greater than 93% Top-1 accuracy in live scenarios. This significantly surpasses traditional manual, post-hoc area of interest analysis [44]. By leveraging lightweight, advanced deep learning models, our approach addresses the limitations associated with simulated environments and high data demands, enabling immediate and relevant analysis in critical care settings. Furthermore, our integration of real-time data analy-

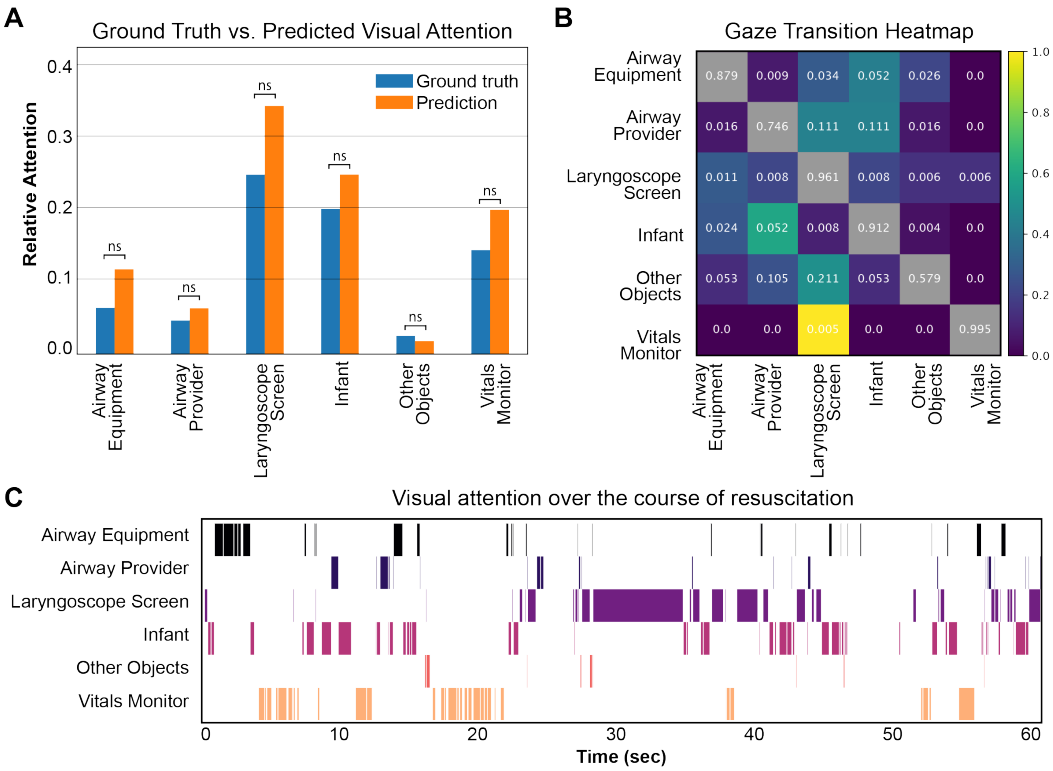

Figure 4. **Automated pipeline captures neonatologist gaze dynamics. (A)** Our multi-modal approach makes predictions as accurately as expert human annotators (displayed: classification comparisons for one test video; ns: not significant). **(B).** Visualizing gaze transitions between areas of interest reveal strong transitions between the *Vitals* $\rightarrow$ *Laryngoscope Screen* and the *Infant* $\leftrightarrow$ *Airway-Provider*, which may be due to their spatial proximity. Transition matrix probabilities are normalized per-row. **(C).** Visualizing gaze dynamics reveals the blocks of *Airway Equipment* early on (i.e., during placement), the *Laryngoscope Screen* block (i.e., during intubation), and the *Vitals Monitor* block (i.e., during patient stabilization).

sis into clinical workflows facilitates the automatic decoding of clinical attention without disrupting operations, effectively bridging the gap in large-scale, data-driven analysis.

Limitations. The primary constraint of this study is the small dataset, which necessitates the acquisition of more extensive data to ascertain the generalizability of our approach. Additionally, due to the specialized nature of our application, there are no publicly available datasets for further validation. Currently, our analysis is limited to the CLIP vision-language model; exploring a broader range of newer Vision Language Models (VLMs) could potentially enhance our approach's efficacy, particularly in zero-shot conditions. Deploying our system in real-time clinical settings also presents several challenges. Precise calibration of eye-tracking devices and robust data security measures are imperative to ensure reliable system performance. Suboptimal placement of glasses and various environmental factors may compromise the accuracy of gaze estimation [4]. Furthermore, improving the transparency and explainability of our models is essential for their adoption in clinical settings [22]. Consequently, extensive validation in diverse clinical environments is necessary to confirm the reliability and applicability of our approach.

Impact. Our system advances beyond traditional, labor-intensive methods like manual video coding by utilizing real-time gaze monitoring to significantly enhance medical training. By automatically quantifying attention dynamics, our approach accelerates existing coding processes and affords real-time decision support. This can not only aid clinical decision-making by pinpointing lapses in situational awareness, but can also enrich patient care through a profound understanding of attention dynamics in critical care environments, laying the groundwork for further exploration to better understand and potentially reduce cognitive load.

6. Acknowledgments

We thank the healthcare providers of the Children's Hospital of Philadelphia (CHOP) for their support and collaboration. This project was supported by the American Academy of Pediatrics Neonatal Resuscitation Program Human Factors Grant (awarded to HMH). HMH is supported by the Agency for Healthcare Research

and Quality career development grant (K08HS029029). MLP is supported by R37-MH109728, R01-MH108627, R01-MH-118203, KA2019-105548, U01MH121260, UM1MH130981, R56MH122819, R56AG071023.

References

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 2

[2] James P Bohnslav, Nivanthika K Wimalasena, Kelsey J Clausing, Yu Y Dai, David A Yarmolinsky, Tomás Cruz, Adam D Kashlan, M Eugenia Chiappe, Lauren L Orefice, and Clifford J Woolf. DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels. *Elife*, 10:e63377, 2021. Publisher: eLife Sciences Publications Limited. 3

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 3

[4] Mark Browning, Simon Cooper, Robyn Cant, Louise Sparkes, Fiona Bogossian, Brett Williams, Peter O'Meara, Linda Ross, Graham Munro, and Barbara Black. The use and limits of eye-tracking in high-fidelity clinical scenarios: A pilot study. *Int Emerg Nurs*, 25:43–47, 2016. 8

[5] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Therese Eder, Fabian Huettig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–10, New York, NY, USA, 2020. Association for Computing Machinery. 1

[6] Linda Gai Rui Chen and Brenda Hiu Yan Law. Use of eye-tracking to evaluate human factors in accessing neonatal resuscitation equipment and medications for advanced resuscitation: A simulation study. *Frontiers in Pediatrics*, 11, 2023. 1

[7] MMPreTrain Contributors. OpenMMLab's Pre-training Toolbox and Benchmark, 2023. 5

[8] Omar Damji, Patricia Lee-Nobbee, David Borkenhagen, and Adam Cheng. Analysis of eye-tracking behaviours in a pediatric trauma simulation. *CJEM*, 21(1):138–140, 2019. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021. arXiv:2010.11929 [cs]. 2, 5

[10] Aisling A. Garvey and Eugene M. Dempsey. Simulation in Neonatal Resuscitation. *Frontiers in Pediatrics*, 8, 2020. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016. 5

[12] Heidi Herrick, Danielle Weinberg, Charlotte Cecarelli, Claire E. Fishman, Haley Newman, Maria C. den Boer, Tessa Martherus, Trixie A. Katz, Vinay Nadkarni, Arjan B. te Pas, and Elizabeth E. Foglia. Provider visual attention on a respiratory function monitor during neonatal resuscitation. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 105(6):666–668, 2020. Publisher: BMJ Publishing Group Section: Short report. 1

[13] Heidi M. Herrick, Scott Lorch, Jesse Y. Hsu, Kenneth Catchpole, and Elizabeth E. Foglia. Impact of flow disruptions in the delivery room. *Resuscitation*, 150:29–35, 2020. 1

[14] Alexandros Karargyris, Satyananda Kashyap, Ismini Lourentzou, Joy T. Wu, Arjun Sharma, Matthew Tong, Shafiq Abedin, David Beymer, Vandana Mukherjee, Elizabeth A. Krupinski, and Mehdi Moradi. Creation and validation of a chest X-ray dataset with eye-tracking and report dictation for AI development. *Sci Data*, 8(1):92, 2021. Number: 1 Publisher: Nature Publishing Group. 2

[15] Trixie A. Katz, Danielle D. Weinberg, Claire E. Fishman, Vinay Nadkarni, Patrice Tremoulet, Arjan B. Te Pas, Aleksandra Sarcevic, and Elizabeth E. Foglia. Visual attention on a respiratory function monitor during simulated neonatal resuscitation: an eye-tracking study. *Arch Dis Child Fetal Neonatal Ed*, 104(3):F259–F264, 2019. 1, 7

[16] Naji Khosravan, Haydar Celik, Baris Turkbey, Elizabeth C. Jones, Bradford Wood, and Ulas Bagci. A collaborative computer aided diagnosis (C-CAD) system with eye-tracking, sparse attentional model, and deep learning. *Medical Image Analysis*, 51:101–115, 2019. 2

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[18] Chaitanya S. Kulkarni, Shiyu Deng, Tianzi Wang, Jacob Hartman-Kenzler, Laura E. Barnes, Sarah Henrickson Parker, Shawn D. Safford, and Nathan Lau. Scene-dependent, feedforward eye gaze metrics can differentiate technical skill levels of trainees in laparoscopic surgery. *Surg Endosc*, 37(2):1569–1580, 2023. 2

[19] Brenda Hiu Yan Law, Po-Yin Cheung, Michael Wagner, Sylvia van Os, Bin Zheng, and Georg Schmölzer. Analysis of neonatal resuscitation using eye tracking: a pilot study. *Arch Dis Child Fetal Neonatal Ed*, 103(1):F82–F84, 2018. 1

[20] Ayoung Lee, Hyunsoo Chung, Yejin Cho, Jue Lie Kim, Jinju Choi, Eunwoo Lee, Bokyung Kim, Soo-Jeong Cho, and Sang Gyun Kim. Identification of gaze pattern and blind spots by upper gastrointestinal endoscopy using an eye-tracking technique. *Surg Endosc*, 36(4):2574–2581, 2022. 2

[21] Tina A. Leone. Using video to assess and improve patient safety during simulated and actual neonatal resuscitation. *Seminars in Perinatology*, 43(8):151179, 2019. 1, 7

[22] Chong Ma, Lin Zhao, Yuzhong Chen, Lu Zhang, Zhenxiang Xiao, Haixing Dai, David Liu, Zihao Wu, Zhengliang Liu, Sheng Wang, Jiaxing Gao, Changhe Li, Xi Jiang, Tuo Zhang, Qian Wang, Dinggang Shen, Dajiang Zhu, and Tianming Liu. Eye-gaze-guided Vision Transformer for Rectifying Shortcut Learning, 2022. arXiv:2205.12466 [cs]. 2, 8

[23] Ben McNaughten, Caroline Hart, Stephen Gallagher, Carol Junk, Patricia Coulter, Andrew Thompson, and Thomas

Bourke. Clinicians' gaze behaviour in simulated paediatric emergencies. *Arch Dis Child*, 103(12):1146–1149, 2018. 1, 3

[24] Sachin Mehta and Mohammad Rastegari. MobileViT: Lightweight, General-purpose, and Mobile-friendly Vision Transformer, 2022. arXiv:2110.02178 [cs]. 2, 5

[25] Aakash Mishra, Rajat Mittal, Christy Jestin, Kostas Tingos, and Pranav Rajpurkar. Improving Zero-Shot Detection of Low Prevalence Chest Pathologies using Domain Pre-trained Language Models. *arXiv preprint arXiv:2306.08000*, 2023. 2

[26] Sangjoon Park, Gwanghyun Kim, Yujin Oh, Joon Beom Seo, Sang Min Lee, Jin Hwan Kim, Sungjun Moon, Jae-Kwang Lim, and Jong Chul Ye. Vision transformer for covid-19 cxr diagnosis using chest x-ray feature corpus. *arXiv preprint arXiv:2103.07055*, 2021. 2

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. ISSN: 2640-3498. 2, 4, 5

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2015. arXiv:1409.0575 [cs]. 6

[29] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 4, 6

[30] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int J Comput Vis*, 128(2): 336–359, 2020. arXiv:1610.02391 [cs]. 6

[31] J. N. Stember, H. Celik, E. Krupinski, P. D. Chang, S. Mutasa, B. J. Wood, A. Lignelli, G. Moonis, L. H. Schwartz, S. Jambawalikar, and U. Bagci. Eye Tracking for Deep Learning Segmentation Using Convolutional Neural Networks. *Journal of Digital Imaging*, 32(4):597–604, 2019. 2

[32] Joseph N. Stember, Haydar Celik, David Gutman, Nathaniel Swinburne, Robert Young, Sarah Eskreis-Winkler, Andrei Holodny, Sachin Jambawalikar, Bradford J. Wood, Peter D. Chang, Elizabeth Krupinski, and Ulas Bagci. Integrating Eye Tracking and Speech Recognition Accurately Annotates MR Brain Images for Deep Learning: Proof of Principle. *Radiology: Artificial Intelligence*, 3(1):e200047, 2021. Publisher: Radiological Society of North America. 2

[33] Lena Stubbemann, Dominik Dürrschnabel, and Robert Refflinghaus. Neural Networks for Semantic Gaze Analysis in XR Settings. pages 1–11, 2021. 1

[34] Adam Szulewski and Daniel Howes. Combining first-person video and gaze-tracking in medical simulation: a technical feasibility study. *ScientificWorldJournal*, 2014:975752, 2014. 1

[35] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. pages 32–42, 2021. 2, 5

[36] A. van der Gijp, C. J. Ravesloot, H. Jarodzka, M. F. van der Schaaf, I. C. van der Schaaf, J. P. J. van Schaik, and Th. J. ten Cate. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Adv in Health Sci Educ*, 22(3):765–787, 2017. 1

[37] Shyam Visweswaran, Andrew J King, Mohammadamin Tajgardoon, Luca Calzoni, Gilles Clermont, Harry Hochheiser, and Gregory F Cooper. Evaluation of eye tracking for a decision support application. *JAMIA Open*, 4(3):ooab059, 2021. 1, 2

[38] Michael Wagner, Peter Gröpel, Katharina Bibl, Monika Olischar, Marc A. Auerbach, and Isabel T. Gross. Eye-tracking during simulation-based neonatal airway management. *Pediatr Res*, 87(3):518–522, 2020. Number: 3 Publisher: Nature Publishing Group. 1

[39] Michael Wagner, Peter Gröpel, Felix Eibensteiner, Lisa Kessler, Katharina Bibl, Isabel T. Gross, Angelika Berger, and Francesco S. Cardona. Visual attention during pediatric resuscitation with feedback devices: a randomized simulation study. *Pediatr Res*, 91(7):1762–1768, 2022. Number: 7 Publisher: Nature Publishing Group. 1

[40] Bin Wang, Armstrong Aboah, Zheyuan Zhang, and Ulas Bagci. GazeSAM: What You See is What You Segment, 2023. arXiv:2304.13844 [cs]. 2

[41] Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, and Ulas Bagci. GazeGNN: A Gaze-Guided Graph Neural Network for Chest X-ray Classification, 2023. arXiv:2305.18221 [cs]. 2

[42] Sheng Wang, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Follow My Eye: Using Gaze to Supervise Computer-Aided Diagnosis. *IEEE Trans Med Imaging*, 41 (7):1688–1698, 2022. 2

[43] Qiuhong Wei, Huiling Cao, Yuan Shi, Ximing Xu, and Tingyu Li. Machine learning based on eye-tracking data to identify Autism Spectrum Disorder: A systematic review and meta-analysis. *Journal of Biomedical Informatics*, 137: 104254, 2023. 2

[44] Danielle D. Weinberg, Haley Newman, Claire E. Fishman, Trixie A. Katz, Vinay Nadkarni, Heidi M. Herrick, and Elizabeth E. Foglia. Visual attention patterns of team leaders during delivery room resuscitation. *Resuscitation*, 147:21–25, 2020. 1, 7

[45] Gary M. Weiner and Jeanette Zaichkin. *Textbook of Neonatal Resuscitation*. American Academy of Pediatrics. 1

[46] Brett Williams, Andrew Quested, and Simon Cooper. Can eye-tracking technology improve situational awareness in paramedic clinical education? *Open Access Emerg Med*, 5:23–28, 2013. 1

[47] Mark R. Wilson, Samuel J. Vine, Elizabeth Bright, Rich S. W. Masters, David Defriend, and John S. McGrath. Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: a randomized, controlled study. *Surg Endosc*, 25(12):3731–3739, 2011. 1

[48] Julian Wolf, Stephan Hess, David Bachmann, Quentin Lohmeyer, and Mirko Meboldt. Automating Areas of Interest Analysis in Mobile Eye Tracking Experiments based on Machine Learning. *J Eye Mov Res*, 11(6): 10.16910/jemr.11.6.6, 2018. 1

[49] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28, 2014. 3

[50] Nicole K Yamada, Kimberly A Yaeger, and Louis P Halamek. Analysis and classification of errors made by teams during neonatal resuscitation. *Resuscitation*, 96:109–113, 2015. Publisher: Elsevier. 1

[51] Emily C. Zehnder, Georg M. Schmölzer, Michael van Manen, and Brenda H. Y. Law. Using eye-tracking augmented cognitive task analysis to explore healthcare professionals' cognition during neonatal resuscitation. *Resuscitation Plus*, 6:100119, 2021. 1, 2

[52] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 2, 3

[53] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *CoRR*, abs/2111.03930, 2021. 4, 6

[54] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023. 3