EchoGuide: Active Acoustic Guidance for LLM-Based Eating Event Analysis from Egocentric Videos

Vineet Parikh
Cornell University
Ithaca, NY, USA
vap43@cornell.edu

Saif Mahmud
Cornell University
Ithaca, NY, USA
sm2446@cornell.edu

Devansh Agarwal
Cornell University
Ithaca, NY, USA
da398@cornell.edu

Ke Li
Cornell University
Ithaca, NY, USA
kl975@cornell.edu

François Guimbretière
Cornell University
Ithaca, NY, USA
francois@cs.cornell.edu

Cheng Zhang
Cornell University
Ithaca, NY, USA
chengzhang@cornell.edu

ABSTRACT

Self-recording eating behaviors is a step towards a healthy lifestyle recommended by many health professionals. However, the current practice of manually recording eating activities using paper records or smartphone apps is often unsustainable and inaccurate. Smart glasses have emerged as a promising wearable form factor for tracking eating behaviors, but existing systems primarily identify when eating occurs without capturing details of the eating activities (E.g., what is being eaten). In this paper, we present EchoGuide, an application and system pipeline that leverages low-power active acoustic sensing to guide head-mounted cameras to capture egocentric videos, enabling efficient and detailed analysis of eating activities. By combining active acoustic sensing for eating detection with video captioning models and large-scale language models for retrieval augmentation, EchoGuide intelligently clips and analyzes videos to create concise, relevant activity records on eating. We evaluated EchoGuide with 9 participants in naturalistic settings involving eating activities, demonstrating high-quality summarization and significant reductions in video data needed, paving the way for practical, scalable eating activity tracking.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Eating Detection; Acoustic Sensing; Activity Recognition; Foundation Models

ACM Reference Format:
Vineet Parikh, Saif Mahmud, Devansh Agarwal, Ke Li, François Guimbretière, and Cheng Zhang. 2024. EchoGuide: Active Acoustic Guidance for LLM-Based Eating Event Analysis from Egocentric Videos. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers (ISWC '24), October 5–9, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3675095.3676611

1 INTRODUCTION

Self-recording eating behaviors is a step towards a healthy lifestyle recommended by many health professionals. However, the current practice requires users to manually record their eating activities, including when and what they eat, using paper records or smartphone apps. This manual method is often unsustainable and sometimes inaccurate, as users frequently forget to record their activities.

Smart glasses have emerged as a promising wearable form factor for tracking eating behaviors. To alleviate the need for manual recording, various sensing systems based on smart glasses have been developed to distinguish eating behavior from arm movements [41], ambient sound[31] or facial muscle movements[28]. However, most of these systems can only identify when eating occurs but not what is being eaten, which is critical information for interpreting eating behaviors. Conversely, sensing systems such as cameras, which can capture detailed information on eating (e.g., the type of food consumed), have high power consumption, making continuous operation impractical on commodity smart glasses.

In this paper, we explore the research question:

- *Is it possible to use low-power active acoustic sensing on glasses to automatically guide the camera to capture activities, such as eating, in an energy-efficient manner without losing much critical information?*

Active acoustic sensing[33] has been shown as a low-power and powerful sensing modality for tracking and interpret various types of fine-grained body poses on wearables, including facial expressions[18], gaze[17], finger pose[14, 37], body pose[21], tongue gesture[29], silent speech recognition[39, 40], authentication[8, 16] and physiological signal [7, 9]. The latest work ActSonic[22] has shown that using active-acoustic sensing on glasses can recognize over 28 types of everyday activities (including eating) in the wild with 89% F1 score at each second without the need for any training data from a new user. More specifically, it recognizes eating activities with an F1 score of 90% in completely unconstrained environments. However, this sensing modality doesn't capture the full context of a given activity. For activity recording and downstream applications (such as calorie counting or recipe assistance), it's important to understand not only **what action** (e.g., when eating

ISWC '24, October 5–9, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-1059-9/24/10…$15.00
https://doi.org/10.1145/3675095.3676611

happens) is performed given body motion, but also **what objects** (e.g., what is the food being eaten) the action is being performed with.

In this paper, we present the design and implementation of EchoGuide, an application and system pipeline that combines the strengths of active acoustic sensing for action detection, video captioning models for detailed egocentric action understanding, and large-scale language models with retrieval augmentation for conversational QA with action records, to enable efficient and seamless action recording and retrieval within specialized everyday domains such as eating.

By leveraging efficient pre-trained models for action detection via active acoustic sensing from ActSonic[22], EchoGuide can intelligently "clip" the videos to guide the camera and video models, creating an activity record that remains far shorter than naive dense-clip video captioning applications while additionally remaining far more relevant than inflexible sparse clipping methods.

We evaluate the performance of EchoGuide with 9 participants wearing GoPros and ultrasonic sensors affixed to commodity eyeglasses to collect data about eating in the unconstrained environment of the participant's choice. With customized acoustic data preprocessing, action detection, video captioning, and action retrieval QA pipeline, we efficiently build activity records with significant reductions in record length while maintaining high semantic correlation with densely captured records. We evaluate the system via a semi-in-the-wild user study with 9 participants focused on eating actions. Additionally, we discuss some of the challenges that EchoGuide must address to be deployed further at scale. We summarize the contributions as follows:

- To the best of our knowledge, we are the first to demonstrate the feasibility of leveraging active acoustic sensing on glass frames to guide the highly efficient capture and analysis of egocentric video for eating activity tracking.
- We propose an end-to-end application pipeline enabling seamless and efficient action detection, video captioning, and action retrieval/QA leveraging a combination of active acoustic sensing and egocentric video.
- We evaluated the end-to-end pipeline on eating activities collected in naturalistic settings of 9 participants' choices through a user-independent and session-independent study. Our system provided high-quality summarization (68% average reduction in activity records along with high alignment between reduced and original activity records given eating-focused videos) while significantly reducing the amount of video data needed.

2 RELATED WORK

Multimodal Image/Video Captioning and Summarization: As larger-scale language and multimodal generative "foundation models" [5] have been trained and released, image and video captioning has extended from simply determining the similarity of images/videos to a premade list of captions [19, 25, 27, 32, 35, 36] towards generating captions for new videos based on either finetuning inexpensive smaller-scale Large Language Models with video encoders and captions [43] or leveraging the emergent properties of natively multimodal Large Language models (such as the OpenAI GPT-4 multimodal series [1, 26] and Google Gemini multimodal series [30]), truly "open-world" video captioning becomes more possible especially in a "zero-shot" paradigm without labeled examples. These systems can be applied offline throughout a video to create "activity records": long documents which encode which activities a person might be completing within the course of a video, and which can be efficiently indexed and searched.

Extracting insights from preprocessed "activity records" requires methods which can generate relevant answers to queries that are grounded in specific documents. Recent generative methods, especially in scenarios involving domain-specific information, leverage the Retrieval-Augmented Generation [15] for returning helpful responses given queries and documents containing relevant information.

However, the primary bottleneck when leveraging image/video captioning and summarization systems especially over longer videos is power and compute consumption: wearable cameras such as the GoPro HERO9 [11] do not have sufficient battery life for continuous daily capture, and video-processing models which recognize activities and objects have high compute requirements.

Eating Recognition on Glasses: Various sensing modalities on eyeglasses form factors have been proposed to track eating events. These modalities include EMG electrodes [38], piezoelectric sensing [10, 28], contact microphones [4], microphones and IMUs [24], and sensor fusion [2, 3]. While these systems track eating episodes, they lack the ability to provide detailed information related to eating activities, such as what food a person eats. This limitation is due to the absence of optimized access to an egocentric camera for extended monitoring periods.

3 THE SYSTEM DESIGN OF ECHOGUIDE

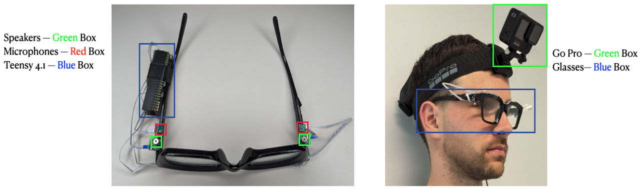

Figure 1: Glasses and GoPro Hardware setup for EchoGuide

In this section, we will present the design of EchoGuide including 1) the hardware prototype we used to collect egocentric acoustic and video data for eating activities; and 2) the software and deep learning pipeline we used to process the acoustic data for eating event segmentation and extract details of eating episodes from the segmented video clips.

3.1 Hardware Prototype

3.1.1 Glasses with active acoustic Sensing. We used a similar hardware prototype design of the glasses as ActSonic[22] as shown in Figure 1. They include two OWR-05049T-38D[1] speakers for chirps

[1] https://www.bitfoic.com/detail/owr05049t38d-14578

and two ICS-43434[2] microphones for receiving signals. The system uses a Teensy 4.1[3] microcontroller to store the transmitted signal and save the received signal on its SD card. Using a similar hardware prototype design will allow us to directly use ActSonic's fully pre-trained deep learning model to identify eating moments in everyday activities without the need of any new training data.

3.1.2 Head-mounted GoPro for egocentric video capture. : To collect egocentric video of user's activities, we used a head-mounted GoPro HERO-9[4], as shown in Figure 1. The data was saved on the SD card within the GoPro.

3.2 Data Processing Pipeline

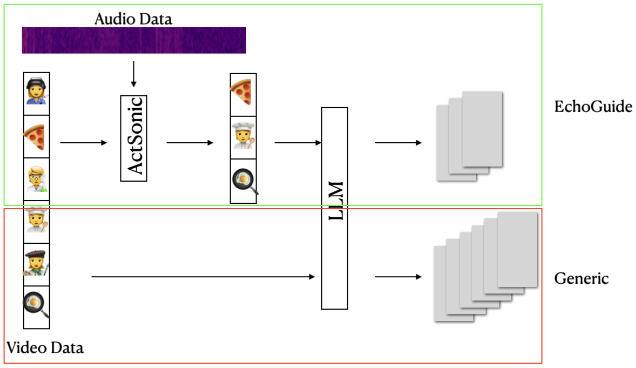

Figure 2: EchoGuide vs Generic LLM Document Generation

The EchoGuide Software and Deep Learning application follows a two-stage pipeline consisting of first processing synchronized egocentric videos and processed active acoustic data into "activity records" representing a history of user activities, and then building a question-answering framework leveraging large language models and Retrieval Augmented Generation for indexing, retrieving, and answering questions grounded in these "activity records". This modular pipeline enables incremental improvement of individual components as they become more capable and is contrasted with a naive dense captioning pipeline which must process the entire video (as shown in Fig 2), with natural language acting as an intermediate step between dense perceptual information and question-answering systems.

3.2.1 Using Active Acoustic Sensing to localize relevant actions and clip videos. We acquired the Resnet-18 model reported in ActSonic[22] via contacting the authors. Our goal is to directly use the pretrained model in ActSonic to determine when an eating activity happens, leveraging the strong user-independent performance of ActSonic for detecting eating activities in in-the-wild settings [22].

We leverage ActSonic's ResNet-18 model as an event detector by splitting the active acoustic differential echo profile (synchronized with the video) into consecutive 2-second windows which can be passed directly into the model. We define a set of "domain-specific classes" within the label space of ActSonic which capture important events for this particular domain, extract class predictions for all sliding windows (essentially treating the class prediction's "timestamp" as the last timestamp of the corresponding window), and construct intervals of events by filtering for "domain-specific classes" and joining equivalent predictions in consecutive windows to create clips without requiring dense captions.

3.2.2 Generating activity records from video and active acoustic sensing. To generate activity records, we take a preprocessed and synchronized dataset containing egocentric videos and acoustic echo/differential profiles from user activities, and apply two modules: a **clipper** to each long untrimmed video/acoustic pair to convert the pair into a series of video clips with possible metadata, and a **captioner** which can take video clips and associated metadata (e.g. timestamp, acoustic classifier label, etc) and generate a "caption" for the clip in EGO-4D format (treating "C" as the camera-wearer) [12], incorporating time metadata as well. We can then join the captions with timestamps to create an "activity record" for the given session. Within EchoGuide(), we primarily focus on proving out the combination of ActSonic as a "clipper" [22] and LaViLa's Narrator (a video-to-GPT2 model fine-tuned on EGO-4D [12], an egocentric vision dataset) as a "captioner" [43].

3.2.3 Answering questions given activity records. We leverage a Retrieval-Augmented Generation [15] framework such as LlamaIndex[20]) for efficiently chunking and embedding a given series of documents (leveraging OpenAI's "text-embedding-ada-002" embedding model) as well as input queries. Given an input query, we run a similarity search on the query embedding vs chunk embeddings (using cosine similarity) and pass the top "k" chunks into the context of a language model (in our case GPT3.5 [6]) to efficiently answer questions about the activity record via a chat/question-and-answer interface.

4 USER STUDY

To collect data for evaluating EchoGuide, we conduct a semi-in-the-wild user study in various naturalistic locations (including participant homes and offices), focusing on capturing natural data of users eating while also performing other activities (such that only parts of each sequence relate to relevant actions). We leverage the activity set proposed in ActSonic [22], which describes a wide collection of everyday activities.

Participants The EchoGuide user study received approval from the Institutional Review Board for Human Participant Research (IRB) at our organization. We recruited 10 participants for a semi-in-the-wild user study at their homes. However, 1 participant's data was lost during the user study. Therefore, we ended up with 9 valid participants in the study, ranging in age from 19 to 34. 6 participants self-reported as male while 3 self-reported as female. We collected basic demographic data and their ratings on the prototype through an IRB-approved questionnaire. The average comfort rating on a Likert scale of 0 to 5 was 2.62.

Data Capture Apparatus We captured acoustic data using the sensing system integrated into EchoGuide eyeglasses and recorded egocentric activity video data via the EchoGuide GoPro Hero9 [11] camera mounted on the participants' heads using a lightweight

[2] https://invensense.tdk.com/products/ics-43434/
[3] https://www.pjrc.com/store/teensy41.html
[4] https://gopro.com/en/us/news/hero9-black-announce

body mount from the same manufacturer. The camera's horizontal and vertical field of view was set to 118° and 69° respectively. It recorded egocentric videos at a resolution of 720p and a frame rate of 30 fps.

Study Procedure We conducted a 9-participant semi-in-the-wild user study in unconstrained environments such as participants' homes and offices. The recruited participants were equipped with eyeglasses and a head-mounted camera. We synchronized the acoustic and video data with a clap action performed by the researcher as the two sensors were physically separated. After synchronization, the participants could continue normal activities without interruption if they ate or drank at least one item within the 40-minute window. Upon completing the 40-minute study, the participants returned the devices to the researcher. We had 5 participants collect data at their homes and 4 in their office environments.

5 EVALUATING QUALITY OF EATING ACTIVITY SUMMARIES AND RESPONSES WITH LLMS

5.1 Metrics

To measure the value of leveraging a supervised ultrasonic model to actively guide video captioners toward more efficient action captioning for retrieval, we define two primary metrics for evaluating system quality:

5.1.1 *Answer alignment with dense captioning.*

Given a single question related to the domain, we find semantic similarity between the answer from a RAG QA agent that has indexed an activity record with alternative sampling (e.g. leveraging the ultrasonic modality to filter and caption fewer frames) and the answer from a RAG QA agent that has indexed an activity record with dense video sampling (e.g. captioning the entire video, which can reduce overall efficiency but captures all possible information). Semantic similarity is captured via BERT F-1 scores [42], which captures pairwise cosine similarities (within the range -1 to 1) between BERT output embeddings to capture semantic and contextual information and which shows high correlation with human evaluations on summarization and captioning tasks (closely related to this work). Similarity scores are used to quantify information loss between the densely captioned model and the ActSonic-captioned model.

5.1.2 *Recording reduction compared to dense sampling.*

Different video clipping methods can lead to different “line counts" for an activity record (as each line of an activity record correlates to a clip in the video where a video captioner model was used). We can therefore find the size reduction between EchoGuide/ActSonic records (where clips are extracted using the ultrasonic modality) vs densely-captioned records (where clips are densely extracted at 1-second intervals).

5.2 Evaluation Procedure/Baseline Description

We show per-participant metrics across the two studies across domains.

We focus on the following baselines and report per-participant metrics along with average metrics for both studies across all relevant domains.

- "1-second Dense Captioning with LaViLa" - this baseline densely splits the video into 1-second long clips and uses LaViLa [43] on each clip to caption individual moments in the video.
- "Ultrasonic Action Captioning Without Video" - this baseline uses models trained on the active acoustic sensing modality to generate clips based on whether the ultrasonic classifier (in this case a pre-trained ActSonic [22] model) classifies a particular 1-second clip as within the domain. The caption for this domain is derived from the classifier label (e.g. for a particular label “eating", the caption would be extracted as “C performed the action: eating"). Notably, this method does not need to sample the video at all, but misses vital context which could be useful for understanding the details of the action.
- EchoGuide, using ultrasonic action detection (via a pre-trained ActSonic [22] model) to efficiently clip a video before applying the LaViLa narrator to build an activity record.

5.3 Quantitative Results

We report per-participant metrics in Table 1. We find a relatively large reduction (avg 68%, max 95.9%, min 34.7%) in activity records using active acoustic sensing with relevant domain actions, though reductions are uneven due to the uneven distribution of eating activities (e.g. P06 spent most of the session eating, resulting in a low reduction of the activity record). We found a higher alignment score by combining both ultrasonic and video modalities to capture and record activities when compared to only using the cheaper ultrasonic modality (0.892 avg for EchoGuide vs 0.828 avg for ActSonic, with low alignment values primarily due to a lack of relevant details within the corresponding activity documents, preventing the LLM from giving a detailed response). Notably, these high correlations and significant % reductions are achieved without fine-tuning either the ultrasonic activity clipper or the visual captioning model on new videos, resulting in “session-independent/user-independent" performance metrics. In addition, these results are collected on user study data that is primarily centered around eating activities: if extended to longer “everyday recordings" where eating is comparatively sparse, future iterations of this system could achieve much higher record reduction metrics.

6 EVALUATING ACTIVITY RECORDS' ABILITY TO ANSWER TARGETED EATING QUESTIONS WITH LARGE IMAGE-LANGUAGE MODELS

6.1 Metrics/Evaluation Procedure

EchoGuide, however, focuses not only on providing general summaries via activity records of an individual's day from video and wearable sensor data, but also on answering targeted questions about these summaries by leveraging the image-text pertaining of large multimodal language models. We evaluate this method via manual review and annotation of the system's answers to three eating questions (“What did C eat/drink? What utensils did C use while eating/drinking? What container did C eat or drink out of?") when configured to use GPT4o [1] to caption images sampled at

	EchoGuide vs Dense	ActSonic vs Dense	% Reduction
P01	0.888	0.854	95.1%
P02	0.897	0.823	53.4%
P03	0.873	0.866	83.7%
P04	0.888	0.862	95.5%
P05	0.906	0.774	55.1%
P06	0.890	0.805	34.7%
P07	0.882	0.785	40%
P08	0.897	0.853	95.9%
P09	0.907	0.833	67.5%

Table 1: EchoGuide metrics across participants in user study, including correlation with dense 1-second clipping (measured as mean BERT F1 score across all sessions per participant) vs ActSonic correlation with 1=second clipping, and % reduction in activity record using active acoustics vs 1-second clipping

	Food type (1fps/0.5fps)	Utensil type	Container type
P01	0/0	0/0	1/0
P02	1/1	1/1	1/1
P03	0/0	1/0	1/0
P04	0/0	0/0	0/0
P05	0/0	1/0	0/0
P06	1/1	1/1	1/1
P07	1/0	1/1	1/1
P08	1/1	1/1	1/1
P09	0/0	1/1	1/1

Table 2: Results of manual evaluation of EchoGuide + GPT4o given 1fps vs 0.5 fps sampling of frames from ActSonic-defined clips (based on zero-shot accuracy). Notation is defined as (X/Y) where X=accuracy at 1fps and Y=accuracy at 0.5fps

two varying FPS levels (1fps and 0.5fps) from clips proposed by ActSonic, and report accuracy metrics showing whether EchoGuide extracts correct values for these questions as compared to manually-determined "ground truth" (taken by watching the reference video and determining which item is present): we've shown accuracy values given 1fps and 0.5fps sampling in Table 2. Accuracy values are defined as a 0/1 binary: 0 represents responses that do not overlap with the ground truth, while 1 represents responses that do completely overlap with the ground truth.

6.2 Quantitative Results and Discussion

In general, we find that while 0.5fps results in a slow reduction in performance for some participants, we can attempt to leverage only a few frames along with metadata information (e.g. classifier outputs) from active acoustic sensors to output useful information, instead of having to process an entire video which could be full of redundant frames. For more ambiguous class types such as "food type" (which may not be easily determinable from appearance alone), we find a relatively low average F1 score (44% for 1fps and 30.5% for 0.5fps) across all participants, whereas for more recognizable/distinctive class type such as utensil type and container type, we find a relatively high average F-1 score (77% for utensils and containers for 1fps, 55% for utensils and containers for 0.5 fps). We find a clear performance drop as FPS is reduced (from 0.55 with 1 fps to 0.47 with 0.5 FPS), due to increased sparsity of frames causing reductions in visual detail for the models. As vision-language models and prompting techniques continue to improve, we expect these numbers will become more accurate over time.

7 DISCUSSION

Further Reduction on video recording when deployed in the wild Our study results showed that EchoGuide helped reduce video recordings by an average of 68% without significantly impacting the quality of summarization for eating activities. However, we want to point out that this percentage of reduction will likely be significantly higher if the system is deployed for full-day recording. In the user study, we only asked participants to collect data for 40 minutes, including the meal. Consequently, the ratio of eating activities in our dataset is significantly higher than it would be in a full day of recording. Therefore, if ActSonic is used to only activate the camera during eating activities in a full-day recording, the data reduction will likely be significantly higher than 68%. Additionally, the frame rate of recording can be further reduced to answer specific questions, saving energy and processing resources. We plan to explore these questions further in future works

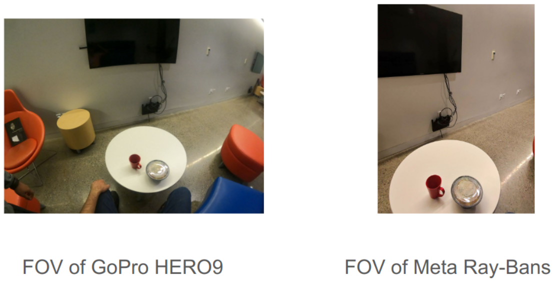

Figure 3: Example showing field of view between GoPro vs Meta Ray-Bans

Comparison to Egocentric Video recorded on glasses The initial hardware system for EchoGuide was not collected using "camera-enabled smartglasses" such as the Meta Ray-Bans [23] due to limitations on recording videos with these off-the-shelf smart glasses. Instead, we used a head-mounted GoPro to easily capture egocentric activity videos. We found the information related to eating captured by smart glasses and our GoPro settings to be highly similar. To help readers understand the difference in images captured by these two devices, we used RayBan Smart glasses and a GoPro HERO9 mounted on the head to capture the same dietary scenario (a drink on the table), as shown in Figure 3. We found that the GoPro has a much wider field of view and can capture more general scene details in the orientation used by EchoGuide compared to the camera on the Meta RayBan smart glasses. However, because most foods are present near the center of the field of view, the difference in view angle between the two cameras did not impact the captured information. Therefore, the results reported in the paper can still be referenced for egocentric video analysis captured on smart glasses.

Exploring additional domains for EchoGuide While EchoGuide was adapted to focus primarily on localizing and understanding eating activities from video and acoustics, we also run a separate exploratory study with three participants operating in both the "eating" and "cooking" domains. Each participant engaged in 5 sessions of 4 minutes each within each domain with interventions between sessions to stop and restart data collection, for a total of 40 minutes per participant.

The question for the "eating" domain was ""What did the person C eat or drink?", with "relevant ultrasonic actions" being defined as the set of "eating", "drinking", and "pickup/putdown" (referring to manipulated items) and the question for the "cooking" domain was "What did the person C cook?" with relevant ultrasonic actions being the set of "chopping", "pouring", "stirring", "pickup/putdown", or "walking". We configure LlamaIndex with GPT3.5-turbo and a temperature value of 0, as well as the standard context prompt "You are a chatbot, able to have normal interactions, as well answer questions from the person about what they did today (walking, eating, cooking, etc). Here are the relevant documents for the context: {context_str}. Instruction: Use the previous chat history, or the context above, to interact and help the user. Format responses as a paragraph."

We find a high average % reduction of 87% in record size across both domains by leveraging active acoustic sensing for clipping videos, along with higher correlation with dense captions (0.9 BERT F1 score) while using EchoGuide's multimodal approach over only using active acoustic sensing (0.86 BERT F1 score). We find that combining the video and ultrasonic modalities additionally shows quantitative improvement (with respect to alignment with the dense caption summary of the original video) when compared to only using the ultrasonic modality, while still maintaining high reductions in the activity record. Though more thorough investigation needs to be done to show this system can work across a wider variety of everyday activities, improvements in unseen domains show the relatively task-agonstic nature of the EchoGuide software pipeline.

Improving comfort and reliability of hardware prototype The current hardware prototype leverages a Teensy-based microcontroller on the left side of the eyeglasses which is connected to a phone for power and recording control, along with a Go-Pro head-mounted camera for video capture. The relative weight and complexity of the combined devices were cited in user surveys as the primary reason for the low comfort rating of the prototype (as it weighed more on the head and ears).

Leveraging a lower-power BLE with a LiPo battery (similar to Google Glass) as the primary microcontroller module for active acoustic recording, along with a Flex-PCB that reduces extraneous wiring, can reduce the unwieldy nature of the acoustic system. Developing custom low-power, high-FPS and high-resolution cameras (such as event-based cameras) that are purpose-built for eyeglass frames can also enable seamless video recording without a Go-Pro requirement, reducing the weight of the systems considerably. Building eyeglass frames that can swap lenses in a custom way, or building a system that can be seamlessly applied on any eyeglass, can reduce the likelihood of participants with custom prescriptions being unable to see through the provided eyeglasses.

Reducing software latency to enable real-time applications Currently, EchoGuide processes and asks questions over activity records in an offline fashion, but many users may want to understand their activities in an online fashion (for instance, asking about previous meals while evaluating what food to get at a restaurant). As seen in other concurrent works [40], active acoustic postprocessing could be completed on a smartphone, and with advancements in embedded AI chips and stronger networking modules for more robust cloud access, it may be possible to do end-to-end inference online with both edge-deployed and cloud-deployed models.

Improving overall model flexibility to new situations While Sec 5 and Table 1 show promising results for EchoGuide usage (combining video and ultrasonic modalities) across two distinct domains and procedural styles in everyday activities, further improvements can be made to enhance overall system performance. Collecting and fine-tuning on a larger base dataset of ultrasonic captures of activities can enable more robust, user-independent detection of human body motion, while leveraging steadily more powerful large multimodal models can enable more robust and generalizable video captions that encode more domain-specific or estoeric information.

The EchoGuide pipeline leverages a Large Language Model with Retrieval Augmented Generation to enable document-based question-answering, along with a Large Multimodal Language Model to enable video captioning. Further optimization of system performance could be achieved via careful prompt engineering methods, such as chain-of-thought with few-shot exemplars [13, 34]. We leave this in-depth exploration of prompt engineering methods to future work.

8 CONCLUSION

In this paper, we present EchoGuide, an innovative application pipeline that combines low-power active acoustic sensing on eyeglasses, egocentric video analysis, and large-scale language models to efficiently detect and analyze eating activities. Our evaluation with 9 participants in naturalistic settings demonstrates that EchoGuide achieves high-quality summarization with a significant reduction in record size while maintaining high semantic correlation with densely-captioned records. As smart glasses become more widespread and equipped with various sensors, multistage pipelines like EchoGuide have the potential to be applied to a broader range of activities and contexts without requiring explicit fine-tuning for individual users.

ACKNOWLEDGEMENT

This project was supported by the National Science Foundation Grant No. 2239569 and partially by the Cornell University IGNITE Innovation Acceleration Program. We also acknowledge and thank Prof. Bharath Hariharan at Cornell CS for help with computing resources.

REFERENCES

[1] [n. d.]. https://openai.com/index/hello-gpt-4o

[2] Abdelkareem Bedri, Diana Li, Rushil Khurana, Kunal Bhuwalka, and Mayank Goel. 2020. FitByte: Automatic Diet Monitoring in Unconstrained Situations Using Multimodal Sensing on Eyeglasses. In *Proceedings of the 2020 CHI Conference*

on Human Factors in Computing Systems (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376869

[3] Abdelkareem Bedri, Yuchen Liang, Sudershan Boovaraghavan, Geoff Kaufman, and Mayank Goel. 2022. FitNibble: A Field Study to Evaluate the Utility and Usability of Automatic Diet Monitoring in Food Journaling Using an Eyeglasses-based Wearable. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (<conf-loc>, <city>Helsinki</city>, <country>Finland</country>, </conf-loc>) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 79–92. https://doi.org/10.1145/3490099.3511154

[4] Shengjie Bi, Tao Wang, Nicole Tobias, Josephine Nordrum, Shang Wang, George Halvorsen, Sougata Sen, Ronald Peterson, Kofi Odame, Kelly Caine, Ryan Halter, Jacob Sorber, and David Kotz. 2018. Auracle: Detecting Eating Episodes with an Ear-mounted Sensor. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 92 (sep 2018), 27 pages. https://doi.org/10.1145/3264902

[5] 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).

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[7] Tao Chen, Yongjie Yang, Xiaoran Fan, Xiuzhen Guo, Jie Xiong, and Longfei Shangguan. 2024. Exploring the Feasibility of Remote Cardiac Auscultation Using Earphones. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 357–372.

[8] Tao Chen, Yongjie Yang, Chonghao Qiu, Xiaoran Fan, Xiuzhen Guo, and Longfei Shangguan. 2024. Enabling Hands-Free Voice Assistant Activation on Earphones. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services* (Minato-ku, Tokyo, Japan) *(MOBISYS '24)*. Association for Computing Machinery, New York, NY, USA, 155–168. https://doi.org/10.1145/3643832.3661890

[9] Xiaoran Fan, David Pearl, Richard Howard, Longfei Shangguan, and Trausti Thormundsson. 2023. Apg: Audioplethysmography for cardiac monitoring in hearables. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[10] Muhammad Farooq and Edward Sazonov. 2016. A novel wearable device for food intake and physical activity recognition. *Sensors* 16, 7 (2016), 1067.

[11] GoPro. 2020. HERO9 Black. https://gopro.com/en/us/shop/cameras/hero9-black/CHDHX-901-master.html. [Online; accessed 12-September-2023].

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

[13] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[14] Chi-Jung Lee, Ruidong Zhang, Devansh Agarwal, Tianhong Catherine Yu, Vipin Gunda, Oliver Lopez, James Kim, Sicheng Yin, Boao Dong, Ke Li, Mose Sakashita, Francois Guimbretiere, and Cheng Zhang. 2024. EchoWrist: Continuous Hand Pose Tracking and Hand-Object Interaction Recognition Using Low-Power Active Acoustic Sensing On a Wristband. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 403, 21 pages. https://doi.org/10.1145/3613904.3642910

[15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[16] Ke Li, Devansh Agarwal, Ruidong Zhang, Vipin Gunda, Tianjun Mo, Saif Mahmud, Boao Chen, François Guimbretière, and Cheng Zhang. 2024. SonicID: User Identification on Smart Glasses with Acoustic Sensing. *arXiv preprint arXiv:2406.08273* (2024).

[17] Ke Li, Ruidong Zhang, Boao Chen, Siyuan Chen, Sicheng Yin, Saif Mahmud, Qikang Liang, Francois Guimbretiere, and Cheng Zhang. 2024. GazeTrak: Exploring Acoustic-based Eye Tracking on a Glass Frame. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) *(ACM MobiCom '24)*. Association for Computing Machinery, New York, NY, USA, 497–512. https://doi.org/10.1145/3636534.3649376

[18] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-Power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. 6, 2, Article 62 (jul 2022), 24 pages. https://doi.org/10.1145/3534621

[19] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. 2022. Egocentric Video-Language Pretraining. *arXiv preprint arXiv:2206.01670* (2022).

[20] Jerry Liu. 2022. *LlamaIndex*. https://doi.org/10.5281/zenodo.1234

[21] Saif Mahmud, Ke Li, Guilin Hu, Hao Chen, Richard Jin, Ruidong Zhang, François Guimbretière, and Cheng Zhang. 2023. PoseSonic: 3D Upper Body Pose Estimation Through Egocentric Acoustic Sensing on Smartglasses. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 111 (sep 2023), 28 pages. https://doi.org/10.1145/3610895

[22] Saif Mahmud, Vineet Parikh, Qikang Liang, Ke Li, Ruidong Zhang, Ashwin Ajit, Vipin Gunda, Devansh Agarwal, François Guimbretière, and Cheng Zhang. 2024. ActSonic: Everyday Activity Recognition on Smart Glasses using Active Acoustic Sensing. *arXiv preprint arXiv:2404.13924* (2024).

[23] Meta. 2023. https://about.fb.com/news/2023/09/new-ray-ban-meta-smart-glasses/

[24] Mark Mirtchouk, Drew Lustig, Alexandra Smith, Ivan Ching, Min Zheng, and Samantha Kleinberg. 2017. Recognizing Eating from Body-Worn Sensors: Combining Free-living and Laboratory Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 85 (sep 2017), 20 pages. https://doi.org/10.1145/3131894

[25] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*. Springer, 1–18.

[26] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[28] Jaemin Shin, Seungjoo Lee, Taesik Gong, Hyungjun Yoon, Hyunchul Roh, Andrea Bianchi, and Sung-Ju Lee. 2022. MyDJ: Sensing Food Intakes with an Attachable on Your Eyeglass Frame. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (, New Orleans, LA, USA,) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 341, 17 pages. https://doi.org/10.1145/3491102.3502041

[29] Rujia Sun, Xiaohe Zhou, Benjamin Steeper, Ruidong Zhang, Sicheng Yin, Ke Li, Shengzhang Wu, Sam Tilsen, Francois Guimbretiere, and Cheng Zhang. 2023. EchoNose: Sensing Mouth, Breathing and Tongue Gestures inside Oral Cavity using a Non-contact Nose Interface. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers* (Cancun, Quintana Roo, Mexico) *(ISWC '23)*. Association for Computing Machinery, New York, NY, USA, 22–26. https://doi.org/10.1145/3594738.3611358

[30] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[31] Edison Thomaz, Cheng Zhang, Irfan Essa, and Gregory D Abowd. 2015. Inferring meal eating activities in real world settings from ambient sounds: A feasibility study. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. 427–431.

[32] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. 2022. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14733–14743.

[33] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW Based Contactless Respiration Detection Using Acoustic Signal. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 170 (jan 2018), 20 pages. https://doi.org/10.1145/3161188

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.

[35] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 450–459.

[36] Michael Wray, Davide Moltisanti, Walterio Mayol-Cuevas, and Dima Damen. 2016. Sembed: Semantic embedding of egocentric action videos. In *European Conference on Computer Vision*. Springer, 532–545.

[37] Tianhong Catherine Yu, Guilin Hu, Ruidong Zhang, Hyunchul Lim, Saif Mahmud, Chi-Jung Lee, Ke Li, Devansh Agarwal, Shuyang Nie, Jinseok Oh, et al. 2024. Ring-a-Pose: A Ring for Continuous Hand Pose Tracking. *arXiv preprint arXiv:2404.12980* (2024).

[38] Rui Zhang and Oliver Amft. 2017. Monitoring chewing and eating in free-living using smart eyeglasses. *IEEE journal of biomedical and health informatics* 22, 1 (2017), 23–32.

[39] Ruidong Zhang, Hao Chen, Devansh Agarwal, Richard Jin, Ke Li, François Guimbretière, and Cheng Zhang. 2023. HPSpeech: Silent Speech Interface for Commodity Headphones. In *Proceedings of the 2023 ACM International Symposium on Wearable Computers*. 60–65.

[40] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-Obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 852, 18 pages. https://doi.org/10.1145/3544548.3580801

[41] Ruidong Zhang, Jihai Zhang, Nitish Gade, Peng Cao, Seyun Kim, Junchi Yan, and Cheng Zhang. 2022. EatingTrak: Detecting Fine-Grained Eating Moments in the Wild Using a Wrist-Mounted IMU. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 214 (sep 2022), 22 pages. https://doi.org/10.1145/3546749

[42] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

[43] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6586–6597.