PEANUT: A Human-AI Collaborative Tool for Annotating Audio-Visual Data

Zheng Zhang*
University of Notre Dame
Notre Dame, Indiana, USA
zzhang37@nd.edu

Zheng Ning*
University of Notre Dame
Notre Dame, Indiana, USA
zning@nd.edu

Chenliang Xu
University of Rochester
Rochester, New York, USA
chenliang.xu@rochester.edu

Yapeng Tian
The University of Texas at Dallas
Richardson, TX, USA
yapeng.tian@utdallas.edu

Toby Jia-Jun Li
University of Notre Dame
Notre Dame, Indiana, USA
toby.j.li@nd.edu
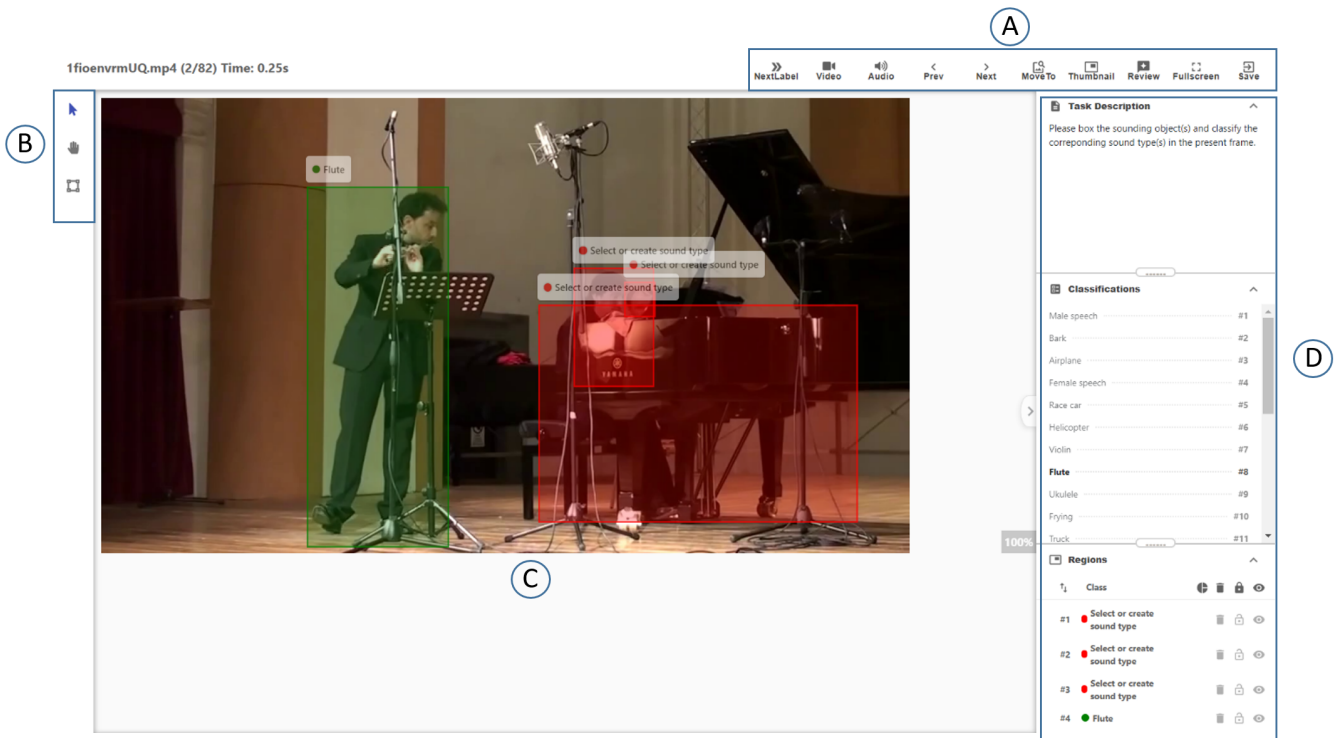
Figure 1: The interface of PEANUT for audio-visual data annotation. A and B: toolbars with editing and annotation assistance functions; C: the annotation workspace; D: the information panel showing meta-information about the annotations on the current frame

*Both authors contributed equally to this work.

ABSTRACT

Audio-visual learning seeks to enhance the computer's multi-modal perception leveraging the correlation between the auditory and visual modalities. Despite their many useful downstream tasks, such as video retrieval, AR/VR, and accessibility, the performance and adoption of existing audio-visual models have been impeded by the availability of high-quality datasets. Annotating audio-visual datasets is laborious, expensive, and time-consuming. To address this challenge, we designed and developed an efficient audio-visual annotation tool called PEANUT. PEANUT's human-AI collaborative pipeline separates the multi-modal task into two single-modal tasks,

and utilizes state-of-the-art object detection and sound-tagging models to reduce the annotators' effort to process each frame and the number of manually-annotated frames needed. A within-subject user study with 20 participants found that PEANUT can significantly accelerate the audio-visual data annotation process while maintaining high annotation accuracy.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Information systems** → *Multimedia information systems.*

KEYWORDS

human-AI collaboration, data annotation, data labeling, audio-visual learning, interactive machine learning

1 INTRODUCTION

Most of our real-world perceptual experiences are specified by multiple cooperating human senses with multi-sensory integration [78]. For example, we can perceive spoken language words and sentiments from lip movements, facial expressions, and speech sounds of the other speakers. To mimic human perception capability, researchers in Artificial Intelligence (AI) community have begun to explore audio-visual machine learning (ML) approaches [7, 10, 48, 49, 85]. As an emerging research field, audio-visual learning has attracted a lot of attention from both the academic community and the industry because of its potential to solve many challenging problems in real-world applications such as video retrieval [70, 122], AR/VR [75, 93], and accessibility [86, 113].

A fundamental task in audio-visual learning is *sounding object localization*, which identifies and localizes sounds to visual objects in videos. This task associates audio data with the corresponding visual data, which represents the critical step for many downstream audio-visual tasks such as audio-visual scene-aware dialogs [2], video sound separation [38, 104], audio spatialization [37, 75], audio-visual captioning [89, 103], and multimodal embodied AI [20, 35].

However, the performance and wide adoption of audio-visual learning have been impeded by the availability of high-quality datasets. For example, commonly used "gold-standard" datasets such as AVE [106] is only weakly supervised for this task (i.e., events are annotated for video segments, but object-level ground-truth labels are not available for frames). Therefore, previous works [7, 10, 48, 49, 85] focus on weakly-supervised or self-supervised methods. These methods not only result in lower model accuracy [98, 106], but also introduce biases in the models. In these methods, existing common label noises will be propagated into learned models, leading to compromised results with ethical issues [105, 119] (e.g., introducing biases and stereotypes presented in datasets into annotation results). Since there are various audible and inaudible visual objects in video frames and even a single sound source can make different sounds with different intensities, the sounding object localization task is hungry for large-scale and high-quality training data to capture the diverse audio-visual correspondences and mitigate data variations.

The lack of high-quality datasets is a direct consequence of the large effort required to create such datasets. Annotating audio-visual data is laborious, expensive, and time-consuming. With the current annotation tools of for audio-visual data (e.g., VIA [30]), annotators need to watch each frame of the video, listen to the corresponding sound, identify the sounding object, draw a bounding box, and indicate the type of sound. Considering that even a short video would require the annotation of hundreds, if not thousands, of frames, this process is highly repetitive and tedious.

Some prior intelligent tools (e.g., [9, 19, 27, 28, 32, 107] have been introduced to provide AI-enabled assistance to users in the data annotation process with promising outcomes to improve annotation efficiency using strategies such as batching [9], rule synthesis [32], and active learning [101]. However, these existing tools are limited to the annotation of data in *one* modality (e.g., text categorization [9, 28], head post recognition [27], handwriting recognition [107], image labeling [12, 19], and video segmentation [88]) while audio-visual data annotation requires the user and its AI assistance to process data from *two* modalities and explicitly connect them together.

In this paper, we present PEANUT[1], a new data annotation tool for improving the efficiency of audio-visual data annotation. To tackle the unique challenge in *multi-modal* data annotation, PEANUT encapsulates a novel human-AI collaborative active learning pipeline where the user validates, revises, and connects the output from multiple single-modal models through a mixed-initiative interface [47].

Instead of using a fixed ML model to pre-label data, PEANUT uses an active learning architecture [24, 99] that allows these partial-automation models to incrementally learn from the user's annotations in real-time to improve model performance, learn about visual and auditory data in new video topics, and adapt to the specific domain of the current video. Several design features are presented to ensure the user's sense of agency and control [110, 111] and alleviate users' overreliance on AI [9], both are notable issues found in human-AI collaboration of data works from previous studies. A within-subjects study with 20 participants showed that PEANUT can significantly accelerate the annotation of audio-visual data (annotate almost 3 times the number of frames compared to the baseline condition) while also achieving high data accuracy.

In summary, this paper presents the following three main contributions:

- A set of interaction mechanisms for incorporating outcomes of *single-modal* ML models into a new human-AI collaborative annotation workflow of *multi-modal* audio-visual data while improving model performance with user annotations in real time using an active learning approach.
- PEANUT, a human-AI collaborative annotation tool that implements these strategies to reduce user efforts and improve efficiency in annotating audio-visual data for sounding object localization.

---

[1]The name PEANUT is an acronym for **P**latform for **E**fficient **A**nnotation with **N**o **U**nnecessary **T**edium.

- A within-subjects user study with 20 participants with diverse annotation and ML expertises showing that PEANUT can improve efficiency in the annotation of audio-visual data while also achieving high data accuracy.

2 BACKGROUND AND RELATED WORK

2.1 Audio-Visual Learning

Audio-visual learning aims to build a multi-sensory perception system that learns from perceived auditory and visual scenes. Mimicking human perception capacity, it can enable a variety of novel applications in many fields, such as multimedia [42, 70, 122], affective computing [72, 81], accessibility [86, 113], and AR/VR [86, 113]. Utilizing and learning from both auditory and visual modalities has attracted significant attention in the AI community.

We have seen great progress in the development of new audio-visual learning problems and applications, such as representation learning [7, 10, 85], audio-visual sound separation [31, 36, 85, 125], sounding object localization [49, 85, 98, 106], audio-visual event localization [67, 106, 118], audio-visual captioning [89, 103], and multimodal embodied AI [20, 35].

2.1.1 Sounding Object Localization. Among audio-visual learning tasks, one important task is the localization of sounding objects. For example, in a symphony concert scenario, the model should identify which instrument is making a particular sound and where the instrument is located in the video. Early work in this area utilized mutual information [45] and canonical correlation analysis (CCA) [55] to localize the sounding visual regions. Recently, deep audio-visual networks have been developed to spatially locate sound sources based on cross-modal embedding similarity [8, 48, 85], audio-guided visual attention [98, 106], audio-visual class activation mapping [87], class-aware object localization map [49], and sounding object visual grounding [104].

These approaches take advantage of the natural synchronization between audio and visual contents and are trained using self-supervised or weakly-supervised methods (i.e., using no or limited annotated training data). Due to the lack of ground truth annotation for training, they tend to make inaccurate predictions. A relevant audio-visual ML task is active speaker detection which focuses on detecting the active speaker from speech, due to the narrower task domain (human speech only) and the availability of datasets such as [56], the state-of-the-art active speaker detection models generally perform better than the domain-general sounding object localization ones [3, 49, 49, 56]. Our work seeks to address this problem by making it easier to annotate large audio-visual datasets.

2.2 Data Annotation for Machine Learning

High-quality data is the foundation of most ML models. The lack of high-quality data has been a long-time bottleneck for many ML tasks [16, 41]. While the undervaluing of data work compared to the lionized work of building novel models and algorithms is common in AI development [95], "data excellence" played a crucial role in the quality of AI systems [95].

Despite that many unsupervised, self-supervised, and weakly-supervised models have been found useful in many task domains [29, 126], supervised learning (i.e., models learning from annotated

example input-output pairs) still shows important advantages in model performances and robustness. However, collecting annotated ground truth data is usually a costly and time-consuming process that requires extensive human labor.

There are two types of data annotation—(1) Explicit data annotation, where human annotators use their perceptive and cognitive abilities to categorize and label data for the purpose of creating annotated datasets [9]. This is often a repetitive and tedious process especially because datasets need to be large in order to be effective; (2) Implicit data annotation, where users of a computing system generate useful datasets as a side product of interacting with the system (e.g., users of a recommender system generate useful data when interacting with recommended items). While this approach does not incur additional user efforts, it requires the system to be deployed at a large scale in order to collect sufficient data, which requires significant effort and is not always feasible in all task domains and at all stages of the project. Implicit annotation also faces the "cold start" problem [96]—it still needs a dataset for training the initial model to provide acceptable performance at the beginning before implicitly-annotated data from user interactions come in.

PEANUT is designed to reduce the human effort required in *explicit* data annotation, making the process more efficient while maintaining data accuracy. Meanwhile, PEANUT also uses *implicit* data annotation strategies to improve the performance of its own object detection model in real-time while the user is in the process of explicit interactive data annotation (detail in Section 3.3.4).

2.2.1 Assistance for Explicit Data Annotation. Explicit data annotation is traditionally a fully manual process—a human annotator examines the input data and determines the output result that the ML model should produce using their human knowledge and cognitive abilities [18] (e.g., the commonly used VIA tool [30] for manual audio-visual annotation). Recently, several interactive tools have been developed to assist human annotators with the process [32, 91, 94, 100, 123]. Notably, RULER [32] is an interactive system that synthesizes labeling rules while the human annotator manually assigns labels in textual data (known as the *data programming* process [91]). Like PEANUT, RULER uses a partial-automation approach where an intelligent system helps the human annotator by automating some parts, but not the full end-to-end annotation task. The two systems also share the "explicit+implicit" approach as RULER learns to synthesize new labeling rules while the user manually labels the data. Desmond et al. introduced an AI labeling assistant that uses a semi-supervised learning algorithm to predict the most probable labels for each example in the labeling intents of user natural language inquiries [28]. PEANUT's interfaces for users to verify labels predicted by a model and correct model-generated bounding boxes are similar to prior work in improving object detection models in computer vision [53, 68].

Some ML-enabled annotation assistance tools use off-the-shelf models to pre-label data as suggestions for users. For example, CVAT[2] uses a deep learning model to pre-label the images. Similarly, Ilastik [12] provides pre-labels to support semi-automatic image segmentation using edge detection and watershed models. Although pre-labeling is effective for accelerating the annotation, it risks annotating data with model biases, especially when the

---

[2]https://github.com/opencv/cvat

data is in a new domain previously unseen in the model's training process [120]. The approach used in Peanut has significant differences from these pre-labeling approaches. Instead of performing pre-labeling independent of human annotation, Peanut grounds its annotation suggestions on human-labeled key frames in real-time to balance model performance and human effort for achieving high-quality annotations and ensuring user control of the annotation process at the same time. The key frames are determined in real-time according to video contents and intermediate annotation results (Section 3.4.2). This approach also allows Peanut to learn new topics from the user's few shot annotations.

The problem domain in Peanut is also more complex than the domain in existing AI-enabled annotation support tools for single-modal data (e.g., images [12, 19], text [28, 34, 94, 101]). Peanut works in a task domain with data in multiple modalities (audio and visual). The task explicitly addresses the explicit and implicit relations between data in different frames.

Another type of assistance for explicit data annotation uses the strategy of *batching* (e.g., [9, 27, 107]). The system first puts "similar" data into batches using unsupervised clustering models or pre-trained models (e.g., semantic similarity for NLP tasks) and then asks the human annotator to annotate data by batch. The underlying assumption is that it would be easier and faster to annotate similar data together than to annotate them individually because they are likely to be assigned with the same or similar labels. The batching strategy has been shown to be effective in accelerating the data annotation process [9]. A potential concern with batching is users' overreliance on AI—the human annotator might assign the same label to a batch without carefully examining each data point because "the AI model thought that they were all similar" [9]. Peanut also uses batching—but it was not achieved using an unsupervised clustering model. Instead, Peanut leverages the characteristics of videos so that adjacent frames are often similar to each other. The auditory and visual models used in Peanut detect changes in the scene or sudden movements, which are used to batch frames so that the human annotator only annotates key frames. A guided workflow for human annotation (Section 3.3.2), video playback, and thumbnail preview features (Section 3.3.3) in Peanut alleviate the overreliance issue.

2.2.2 Implicit Data Annotation. Systems that use implicit data annotation strategy include (1) those that collect data from their interactions with users for the purpose of a *different* data task, such as reCAPTCHA [108] that collects user-annotated data for training computer vision models through its interactive process of distinguishing human users from bots for authentication purposes and the Foldit game [25] that collects user-annotated protein structures through an online game; (2) those that collect data from their interactions with users for the purpose of improving the *same* interaction, such as recommender systems that learn about user personal preferences as the user interacts with the recommended items [90] and intelligent agents that learn about tasks while helping users with task automation [61, 63].

Peanut's use of implicit data annotation falls into the latter category. As discussed in Section 3.3.4, the human annotation result for each keyframe is used to fine-tune the visual-sound grounding model, which reduces human effort in annotating the rest of the frames. This strategy is also an example of *active learning* [24, 99], where the system chooses which data the visual-sound grounding model should learn from by querying the user through Peanut's selection of keyframes (Section 3.4.2).

2.3 Human-AI Collaboration in Data Science

Peanut belongs to a fast-growing list of AI-powered interactive tools that assist and augment human capabilities in different sub-tasks in the data science workflow [44, 80, 110, 112]. Besides data annotation, human-AI collaborative tools have also been developed for data wrangling (i.e., cleaning and formatting data to make it suitable for analysis [52]) (e.g., [40]), exploratory data analysis and sensemaking (e.g., [114]), selection of ML models (e.g., [43]), generating new data features (e.g., [33]), testing and debugging ML models (e.g., [115]), and fine-tuning parameters in ML models (e.g., [69]).

The design of Peanut is informed by empirical studies on how data science workers work with data [76, 77] and data workers' perceptions and mental models of human-AI collaborative data science tools [112]. For example, studies [76, 77] reported that the difficulty with finding reliable labels for ground truth is a common problem that data science practitioners encounter. In industry settings, external domain experts often need to be hired [77]. Spreadsheet is commonly used as a tool for labeling—while specialized tools such as CrowdFlower[3] have also been used, they are used to facilitate group collaboration on data annotation [77] with no intelligent automation feature that reduces the workload of the annotation.

More broadly, facilitating effective collaboration between human and intelligent systems has been a long-standing topic since the origin of HCI research in the seminal paper on man-computer symbiosis [64] where computers can "do the routinizable work that must be done to prepare the way for insights" meanwhile human users can leverage their domain expertise to make decisions that computers cannot. The design of Peanut follows this pattern where the system marks potential object candidates and identifies keyframes for users to annotate using single-modal partial-automation models while the user "connects the final dots" with their annotations that finish the end-to-end multi-modal process using human perceptive and cognitive capabilities that ML models do not yet possess.

Later work such as the principles in mixed-initiative interactions [47] identified strategies such as considering uncertainties in user intents, assessing the added-value of automation, providing mechanisms for refining automation results, and maintaining user working memory of interactions. More recently, guidelines in human-AI interaction [4] have been proposed to address challenges that came with the popularity of "black-box-like" data-driven AI models in interaction systems (as opposed to the "traditional" planning-based techniques). These guidelines, principles, and theoretical frameworks have been widely used in the design of human-AI collaborative systems in domains like healthcare [17], creativity support [71], and error repairs in chatbots [62]. Two key human-AI collaboration challenges we specifically address in the design of Peanut are to accommodate the *imperfection* of AI models and to enable the continuous learning of partial-automation models, which we discuss in Section 3.2.

---

[3]https://appen.com/

3 THE PEANUT SYSTEM

3.1 Task: Sounding Object Localization

PEANUT focuses on annotating data for the sounding object localization task. As defined in [50], given a video, sounding object localization aims to semantically correlate each sound to the visual regions containing the sounding source and recognize the category of the sound in each frame. Therefore, in the annotation task, the annotator should first identify all the sounds in the current frame, and for each sound, provide a semantic label (e.g., church bell, dog bark) and associate it to its sounding object (represented as a bounding box) for each and every frame in a video. A frame may contain multiple sounding objects at the same time as well as silent objects that are physically capable of making sounds.

3.2 Design Goals

Informed by results from prior studies on data annotation [9, 73, 91], status quo of relevant ML models, and our experience with data annotation and human-AI collaborative tools, we summarized the following four design goals in our design of an AI-assisted audio-visual data annotation tool to address potential human and AI challenges in sounding object localization task:

DG1: Improving annotation efficiency without compromising accuracy with imperfect AI models. The ultimate goal of the annotation is to collect data to train an end-to-end ML model for the sounding object localization task. Although there are some end-to-end models for this task [49, 85, 98, 106], their performance is limited due to the lack of annotated data, which our work seeks to address. While some off-the-shelf models (e.g., object detectors, audio tagging models) can contribute to the annotation process of audio-visual data, they are often limited in several ways: (1) they often process input data in only one modality; (2) they have limited accuracy (as they are not specifically trained for our domain); and (3) they only assist with a part of the annotation process. In contrast, our annotation task requires: (1) processing multi-modal audio-visual data; (2) achieving high accuracy in annotation; and (3) providing an end-to-end annotation from raw videos to annotated sounding objects in each frame of the video.

However, the limitations of AI models do not preclude their potential to function as assistant to human annotators. Prior works [12, 23, 54] have demonstrated that using AI models to automate parts of data annotation tasks could significantly improve annotation efficiency. Nonetheless, previous research [73] also suggested that without effective human intervention strategy, imperfect AI could result in lower quality in annotations despite the higher efficiency. Therefore, our main motivation is: how can we make annotating audio-visual data more efficient by introducing AI assistance that reduces human effort and cognitive load? Meantime, it is also crucial that improved efficiency does not come at the cost of compromising accuracy. As discussed in [95], data quality issues can cause significant compounding negative effects on downstream tasks, resulting in "data cascades" that harm users and communities.

DG2: Supporting users' agency and mitigating their overreliance on AI models. Prior work [58] find that users tend to give premature cognitive commitment to automation if the assisted task is routine, repetitive, and demanding. In particular, Ashktorab et al. [9] showed

that AI-assisted data labeling tasks conform to these characteristics in which users are inclined to overrely on AI: despite the inaccuracies of the models, human users sometimes perceive *higher* qualities and *higher* capabilities of the models than what they actually are when they observe their high performance in common situations. However, when uncommon situations arise, human users may overrely on AI automation which could threat human discernment and agency in annotation [9]. This poses a challenge in our task given that the frame-wise annotation is highly repetitive and demanding, which may make users less wary of the cross-modal mismatches and inaccurate annotations by the AI. Therefore, our system should make it easy for users to identify potential errors in AI results and ensure that users fulfill their duties of reviewing and validating AI results in the annotation workflow. Furthermore, our system should support user agency in the annotation process, providing the flexibility to control the degree of human involvement based on their perception of AI's accuracy change over the annotation process.

DG3: Minimizing the learning barrier. Data annotation tasks are often conducted by people without significant AI/ML expertise—many users of data annotation tools are either domain experts (e.g., physicians and radiologists who annotate medical imaging data [79]) or laypeople who only annotate data either as a one-off task or only occasionally (e.g., Mechanic Turk workers [73]). Therefore, it would be prohibitive if the tool has a high learning barrier or requires extensive expertise from the users. Ideally, the system should not require users to learn new skills beyond what they would already need if they labeled the data manually.

DG4: Supporting annotation for diverse video topics. Audio-visual scenarios are intrinsically diverse in terms of the object, sound, and event type involved in videos [7]. Therefore, our annotation tool should be able to provide users with effective AI assistance regardless of the topics in the videos being labeled. To meet this need, models used in the pipeline should be generalizable and not limited to specific domains. While some of these models might be pre-trained to bootstrap the system's performance in cold-start situations for common topics, they should also be able to learn about new object and audio types quickly from the user's few-shot example data in real time.

3.3 System Design

To address the aforementioned design challenges and design goals, we designed and implemented PEANUT, a human-AI collaborative audio-visual annotation tool that seeks to make annotation more efficient using novel interaction strategies, features, and algorithms. Figure 1 shows the main annotation workspace of PEANUT, which consists of three components: (1) a canvas that displays the video frame and its current annotation state (C); (2) top and side toolbars that allow the annotator to perform different operations (A, B); (3) an information panel that summarizes the meta-information about the current object types and displays the operation history (D). The system architecture of PEANUT is shown in Figure 2.

3.3.1 Human annotation. For each frame, PEANUT lowers the user's effort and cognitive load to annotate it by: (1) facilitating the user's access to both global and local audio-visual contexts; (2) inferring
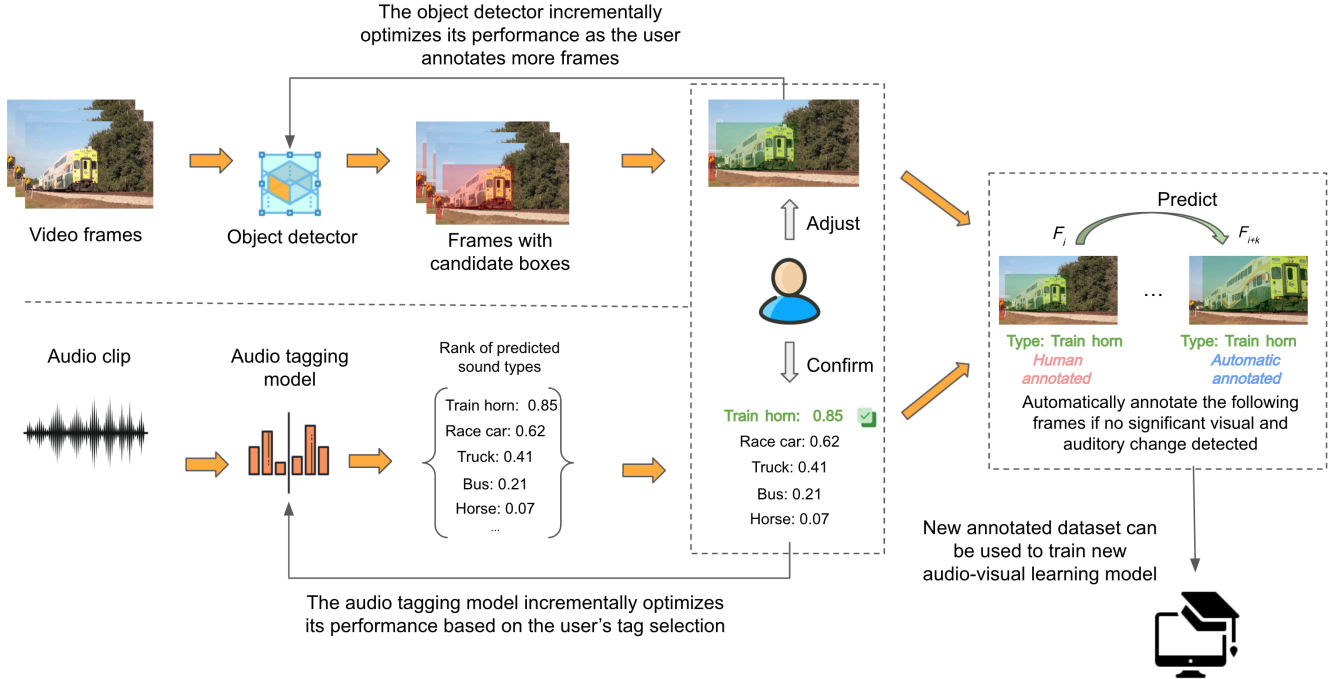
Figure 2: The system architecture of PEANUT. The video frames and the audio clips are processed separately by single-modal partial-automation models before they are presented to the human annotator for the end-to-end result that leverages the annotator's multi-modal perceptive and cognitive capabilities. Both the object detector and the audio tagging model incrementally improve their own performances using an active learning approach.

bounding boxes for potential candidates of sounding objects; and (3) predicting the audio tags for sounds.

When the user moves to a frame that corresponds to the *i*-th millisecond (ms) of the video, PEANUT will auto-play a one-second-long audio clip that contains the soundtrack from *i-500* to *i+500* ms, which gives the annotator a sense of the local audio context. The annotator can replay the local soundtrack by clicking the "Audio" button. In addition, PEANUT allows the annotator to watch the entire video at any time during the annotation process, which helps the annotator disambiguate between the identities of local sound sources with the information of the global context.

PEANUT helps the annotator locate and tag the sounding object. When using PEANUT to annotate a frame, instead of manually recognizing the audio type and drawing the bounding box for its sounding object, the user, in most cases, selects a sound type from a list of predictions made by an audio-tagging model and matches it to one of the visual objects detected by an object detector (red boxes shown in Figure 1). This process can reduce the annotator's cognitive load for locating and labeling the sounding objects (DG1) and demand no domain knowledge from annotators (DG3). If none of the predicted sound types or visual objects is correct, the annotator can manually enter a sound type and draw a new bounding box by clicking the ⌜⌟ button.

3.3.2 Automatic annotation. Besides reducing the efforts required for the user to annotate each frame, PEANUT allows the user to annotate fewer frames. Instead of asking the annotator to label every

frame in the video, PEANUT uses two complementary strategies to automatically infer the annotation result of the remaining frames based on the human annotation of "key frames". The combination of two strategies provides the annotator with the flexibility to adjust the granularity of automatic annotation.

For the first strategy, PEANUT dynamically navigates the annotator to the next keyframe that requires human annotation, from which PEANUT can infer the annotation results of the frames between two human-annotated ones. The annotator can go to the next recommended keyframe by clicking the ❯❯ button. Under the hood, PEANUT uses an *audio-visual-sensitive binary search* algorithm (see Section 3.4.1) to identify the next keyframe that needs human annotation. Note that the recommended key frames may not be in sequential order. For example, after an annotator annotates the 10th and 20th frame consecutively, PEANUT may roll back to the 15th frame if there was a significant visual or auditory change between the 10th frame and the 20th frame according to the human annotation results. When the annotator finishes the annotation of two adjacent recommended keyframes, if both are determined to be continuous in both auditory and visual spaces, PEANUT will automatically interpolate the annotation results of the in-between frames by tracking the movement of known sounding objects, as explained in Section 3.4.1.

In the second strategy, the annotator can choose to automate the annotation of each frame by clicking on the ❯ button. PEANUT will preempt the annotation of the immediate next frame for the annotator to review and confirm. In this way, the annotator can closely
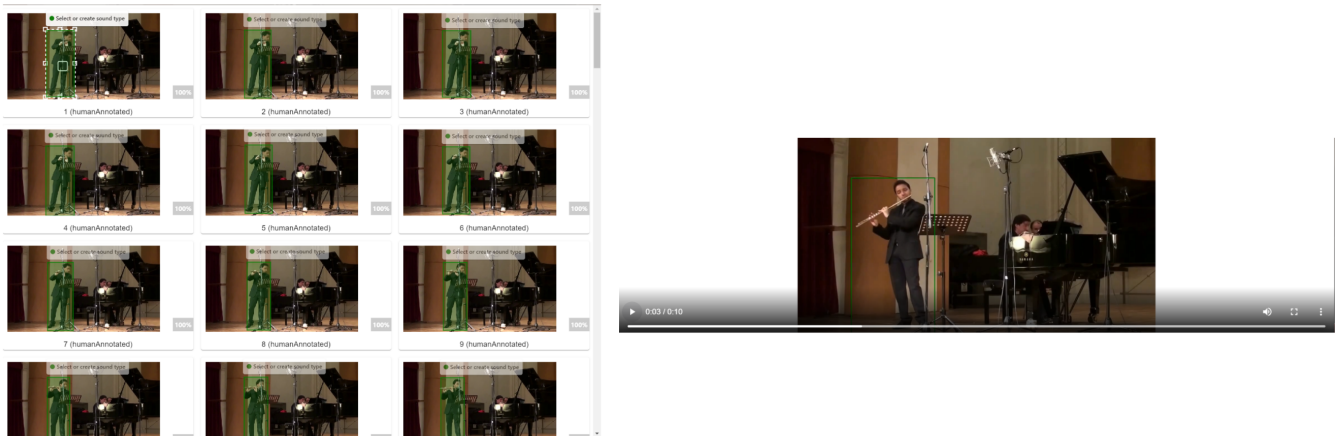
Figure 3: PEANUT provides two interfaces for users to review the annotation result: the frame-by-frame thumbnail review and the annotated video playback preview.

inspect the automatic annotation result to ensure its correctness. The second strategy is especially useful when the auditory or visual context changes quickly in the video.

It's important to note that PEANUT provides users with the choice to select their preferred strategy at any moment during the video annotation, thus allowing dynamic control over the level of human engagement in the annotation process (DG2).

3.3.3 Annotation result review. PEANUT uses two interfaces for the annotator to review the annotation result: *frame-by-frame thumbnail* and *annotated video playback preview* (Figure 3). The *frame-by-frame thumbnail* interface displays the annotation result of each frame in a grid view, enabling the annotator to quickly detect frames with inappropriate annotations. The *annotated video playback preview* interface allows the annotator to examine how the entire video looks after the annotation process is finished. The sound types are also displayed in semi-transparent white boxes above the bounding boxes. The annotator can also go to a specific frame to review and modify the annotation result by clicking the *Move To* button. These two features are designed to assist humans in spotting potential inaccuracies in AI annotations by employing supplementary review strategies, in order to mitigate possible overreliance (DG2).

3.3.4 Active learning. Human-in-the-loop audio-visual data annotation may face two predominant challenges. First, the pre-trained object detector and sound tagging model could only be able to tackle certain objects or sound types and provide little support in those unfamiliar data. Plus, the detection accuracy of object and sound type may be contingent to event scenarios, while audio-visual scenarios are often highly diverse. It is possible that the model has trouble in recognizing a learned type from an unseen scenario. To address these challenges, as shown in Figure 2, PEANUT adopts an active learning strategy to optimize the visual sound grounding network model, object detector and audio tagging model incrementally in real time as the users annotate more data. Because video frames and sounding objects in the same video are usually similar to each other, this active learning strategy allows the model to learn from ground truth data that likely closely resemble input data that it will

process in the future, effectively adapting the model to the domain of the video. In this way, when encountering data type or event scenario unknown to AI model, human annotators can provide the model with a few ground truth annotations to enable it to classify data in the current specific scenario. This strategy also eliminates the need of granting a model with a generalized ability to handle a wide range of diverse scenarios in a single training (DG4), which is formidable for current audio-visual models.

3.4 Algorithmic Methods

3.4.1 Recommending the next frame for human annotation. We developed an *audio-visual sensitive binary search* algorithm for PEANUT to decide the next "keyframe" that needs human annotation (illustrated in Figure 4). The details of the algorithm are shown in Algorithm 1. The index of the next key frame is decided by the similarity of the annotation between the left bound frame (*lb*), the current human annotation frame (*cur*), and a stack of right bound frames (*rbs*). The left-bound frame refers to the preceding human-annotated frame that is the closest to the current frame on the timeline. On the contrary, the right-bound frames are those after the current frame, while the annotation similarity between the current frame and the right-bound frames needs to be confirmed. The calculation of the annotation similarity uses a "*predict-select-compare*" strategy. Given the current frame, PEANUT first predicts the sounding objects in the current frame by inheriting the human annotation of *lb*. Then PEANUT compares the user-selected sounding objects and the predicted sounding objects. If these two sets of objects are not the same, this indicates that there is a visual discontinuity between *lb* and *cur* that is not captured by PEANUT. To locate the first frame of this "discontinuity" efficiently, PEANUT adopts binary search and asks the human annotator to label the frame $\lfloor \frac{lb+cur}{2} \rfloor$, meanwhile PEANUT pushes *cur* into *rbs*. On the other hand, if the predicted sounding objects are the same as the user-selected one, PEANUT will further examine whether a right-bound frame exists. If *rbs* is not empty, PEANUT will calculate the similarity between *cur* and *tf*, where *tf* is the frame at the top of
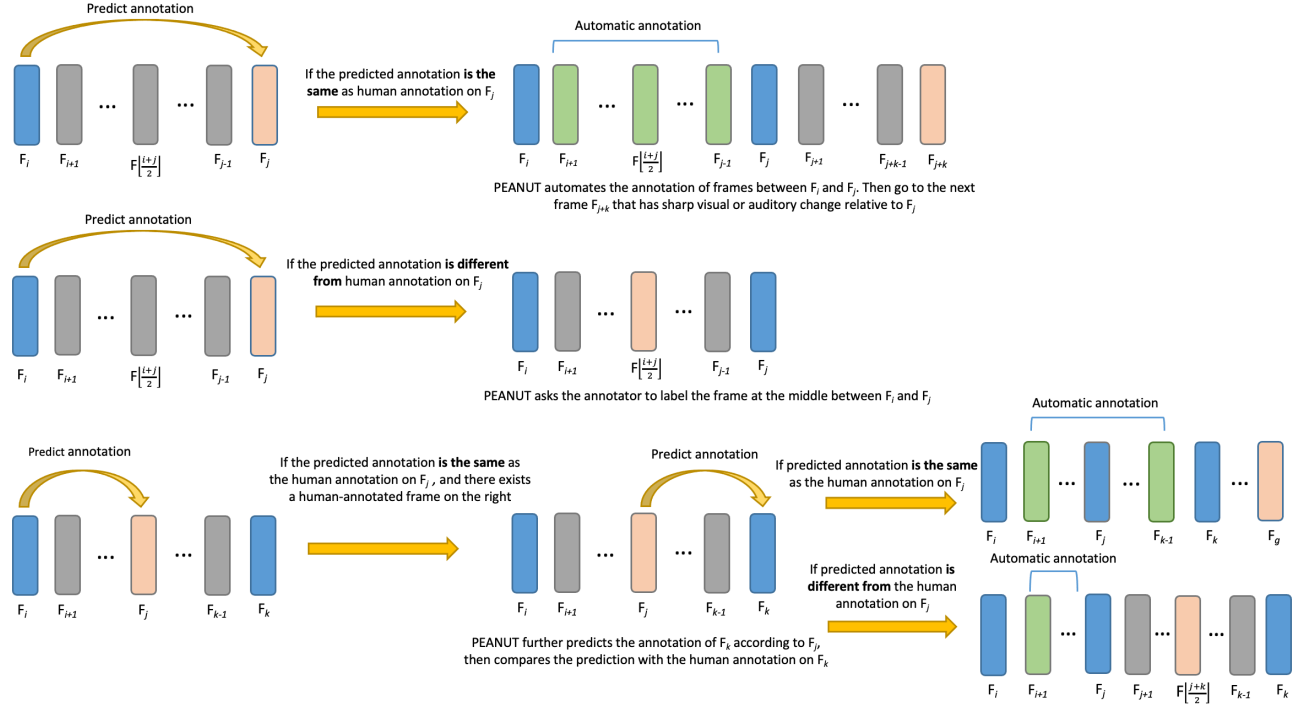
Figure 4: The illustration of the visual-audio sensitive binary search algorithm (Algorithm 1). Blue rectangles represent the frames that have already been annotated by a human annotator, grey rectangles represent the frames that have not been annotated, pink rectangles represent the frames that the human annotator is working on, and green rectangles represent the frames automatically annotated by PEANUT.

rbs. PEANUT will remove *tf* from *rbs* if the similarity holds and automatically interpolate the annotation results for the intermediate frames in between (Algorithm 2). After that, PEANUT continues to compare *cur* and the new frame at the top of *rbs* until *rbs* becomes empty. If the similarity does not hold, PEANUT will ask the human annotator to annotate the frame $\lfloor \frac{cur+tf}{2} \rfloor$.

Otherwise, if the predicted sounding objects are the same as the user-selected ones and no right bound frame exists, PEANUT will move to the first frame that follows the farthest human-annotated frame *hf* and has a significant auditory or visual change compared to *hf* (Algorithm 3). If there is no significant auditory or visual change in the next k frames (we used $k = 10$ in our implementation of PEANUT), PEANUT will ask the human annotator to annotate the frame $hf + k$.

3.4.2 Detecting visual and auditory changes. When a frame has a significant visual or auditory change compared to the last human-annotated frame denoted *src*, PEANUT needs a human annotator to annotate it. A frame denoted *target* is considered to have a significant visual change relative to *src* if (1) the number of detected objects varies between these two frames or (2) a bounding box in *src* does not have a correspondence in *target*, where the correspondence establishes if, given a bounding box *i* on left, there is a bounding box in *target* that overlaps *i* and the overlapping area satisfies the Condition 1, where overlapping parameters α is 0.8 and β is 0.05. The overlapping threshold decreases as the time difference between the target and source frames increases.

$$OverlapArea(B_{highest}, A_i) > [\alpha - (Index_{target} - Index_{src}) \times \beta] \times Area(A_i) \quad (1)$$

The detection of significant auditory changes is based on an audio tagging model. PEANUT uses a state-of-the-art pre-trained Audio Neural Networks [57] to predict the audio tags for each frame. An audio tag describes the possible type (e.g., train horn, race car, truck, as shown in Figure 2) of the sound corresponding to that frame.

3.4.3 Tackling complex audio-visual scenarios. Audio-visual data often inherently possess various complexities. For example, there might be situations where multiple sound-producing objects are active at the same time, making it challenging to discern the specific source of a sound. Besides, some sounds may start at different times and overlap with each other. Additionally, the video may lack a visual indicator of the sounding object, which could complicate the correlation between two modalities. To tackle those complexities, our algorithm focuses on identifying discrepancies or ambiguity in the auditory or visual modality, and asks humans to annotate keyframes. For example, in cases of simultaneous sounds from multiple objects, PEANUT requests human assistance when the audio-tagging model's confidence score is low, indicating auditory uncertainty. Also, the detection of sound changes prompts annotation at the frames with new sound sources. When sounding objects are not visible, PEANUT invokes human intervention at the initial sound frame, predicting the same annotation for subsequent

Name: AudioVisualSensitiveBinarySearch
Input: $LeftBoundFrame, CurrentHumanAnnotatedFrame,$
$GlobalRightBoundStack$
Output: $NextHumanAnnotatedFrame$
$left \leftarrow LeftBoundFrame;$
$right \leftarrow CurrentHumanAnnotatedFrame;$
$grs \leftarrow GlobalRightBoundFrameStack;$
$predObjects \leftarrow$
$PredictSoundingObjectFromPriorAnnotation(left, right);$

if $right == 0$ **or** $(predObjects == SelectedObjects[right]$
  and $grs.length == 0)$ **then**
  | **return** FarthestFrameNeedHumanAnnotation();
else if $predObjects == SelectedObjects[right]$ **and**
  $grs.length > 0)$ **then**
  | $cur \leftarrow right;$
  | $right \leftarrow grs.pop();$
  | **while** $right != Null$ **do**
  | | $predObjects \leftarrow$
  | | $PredictSoundingObjectFromPriorAnnotation(cur, right);$
  | |
  | | **if** $pred\_objects == SelectedObjects[right]$ **then**
  | | | PopulateFrameAnnotation($cur, right$);
  | | | $cur \leftarrow right;$
  | | | $right \leftarrow grs.pop();$
  | | **else**
  | | | $next \leftarrow \lfloor \frac{cur+right}{2} \rfloor;$
  | | | **return** $next;$
  | | **end**
  | **end**
  | **return** FarthestFrameNeedHumanAnnotation();
else
  | $grs.push(right);$
  | $next \leftarrow \lfloor \frac{left+right}{2} \rfloor;$
  | **return** $next;$
end

Algorithm 1: The audio-visual-sensitive binary search algorithm

Name: PopulateFrameAnnotation
Input: $left, right$
Output: $None$
for $i \leftarrow left + 1$ **to** $right$ **do**
  | $predAnnotation \leftarrow$
  | $PredictSoundingObjectFromPriorAnnotation(left, i);$
  |
  | **if** $AduioTags[i].contains(videoTag)$ **then**
  | | $Annotate(i, predAnnotation);$
  | **end**
end

Algorithm 2: Automatically interpolate the annotation results for the frames between two human-annotated frames

Name: FarthestFrameNeedHumanAnnotation
Input: $None$
Output: $next$
$frameIndex \leftarrow getFarthestHumanAnnotatedFrame();$
for $i \leftarrow 1$ **to** k **do**
  | **if** DetectAudioVisualChange(frameIndex, frameIndex+i)
  | $== false$ **then**
  | | $continue$
  | **end**
  | **else**
  | | **return** $frameIndex + i;$
  | **end**
end

Algorithm 3: Return the farthest frame that needs human annotation

frames until significant changes occur (sound cessation or the introduction of a new sound). The "jumpback" (Algorithm 1) solicits human input at the midpoint when successive key frames have inconsistent annotations due to sounding object changing with no visual cue.

Our approach echos prior research [15, 39, 51, 124] in human-AI collaboration for the completion of complex tasks where machine learning models primarily target the automation of repetitive and mundane tasks, resorting to human assistance when the uncertainty score surpasses a threshold.

3.5 Implementation

3.5.1 Web app. The front-end web application of PEANUT is implemented in React based on *react-image-annotate*[4], an open-source framework for the development of image annotation tools and hosted using Python's built-in HTTP server. The back-end server is developed using the Flask framework with a MongoDB database that stores user annotations and log data.

3.5.2 Object detector. PEANUT uses the off-the-shelf Detectron2 [116] library for implementing its object detector. The object detector is built using the Faster R-CNN [92] neural network architecture with R101-FPN feature pyramid networks [65] and is pre-trained on the domain-general MS-COCO dataset [66]. The object detector does not need to be retrained when using PEANUT on videos from a new topic domain. From the input of a video frame in bitmap format, the object detector can identify objects of 80 different types (e.g., "train", "violin", "dog", etc.) and return the corresponding coordinates of the bounding box and the type of each object.

3.5.3 Active visual sound grounding. The Visual Sound Grounding (VSG) network model is trained to identify objects that make sounds among the candidate objects in each video frame. The network is built on top of [104]. The network is first pre-trained on the AVE dataset [106]. Afterward, it is iteratively and incrementally fine-tuned with newly annotated data during the active learning stage. The network takes the feature of the current audio clip and objects that are proposed by the object detector (Section 3.5.2) as inputs and predicts the likelihood that each object is associated with the sound, which allows the model to identify and remove bounding

---
[4]https://github.com/UniversalDataTool/react-image-annotate

boxes that correspond to likely-silent objects from the potential candidates of bounding boxes.

4 USER EVALUATION

To evaluate Peanut, we conducted a user study[5] with 20 users to compare the efficiency and accuracy of audio-visual data annotation using Peanut with those using a baseline system without intelligent features. The results of the study suggest that Peanut can help users annotate data for the sounding object localization task at a faster speed than in the baseline condition, while also improving the annotation accuracy at the same time. The user study also validated the usability of Peanut and provided insight into user reflections on their experiences using Peanut.

4.1 Participants

For this study, we recruited 20 participants through university mailing lists. 7 were undergraduate students, 8 were Master's students, 4 were doctoral students, and 1 was a high school student. Each participant was compensated with $15 USD for their time.

Our participants had varied levels of prior data annotation experiences and ML backgrounds. 8 participants had no prior experience with data annotation and 12 had annotated data at least once. 5 participants had no ML background, 5 participants had at least a beginner level of ML knowledge (have taken introductory ML courses or had basic ML knowledge), 5 had an intermediate level of ML expertise (have taken advanced ML courses or with equivalent expertise), and 5 identified themselves as experts in ML (experienced researchers or practitioners of ML).

4.2 Study Design

Each user study session lasted around 70 minutes and was conducted remotely on Zoom due to the impact of the COVID-19 global pandemic. Before the beginning of each session, the participant signed the consent form and completed a demographic questionnaire. Participants accessed Peanut using the browser on their own computers and shared their screens with the experimenter. After receiving a 5-minute tutorial on how to use Peanut's interface, each participant completed the annotation tasks under two conditions in random order (see Section 4.2.2) and completed a post-study questionnaire on their perceived usability and usefulness of Peanut. The study session ended with a 10-minute semi-structured interview with the participant in which they reflected on their experience interacting with Peanut. All user study sessions were video recorded.

4.2.1 Dataset. In this study, we trained and evaluated Peanut using the Audio-Visual Event (AVE) dataset [106], which is a widely used benchmark dataset for the audio-visual localization task. The full AVE dataset contains 4,143 video clips from a wide range of topics and domains (e.g., "Church Bell", "Male speech", "Dog Bark") in 28 event categories. Each video clip in the AVE dataset is about 10 seconds long. We re-sampled each video at 8 FPS (a common practice in audio-visual data annotation so there are fewer frames to annotate in each video). For the user study, we used a sample

---

	Average SoC ↓	# of Frames ↑	cIoU ↑
Full Automated	N/A	N/A	0.33
Full Manual	7.73	169.45	0.72
Peanut	5.12	488.85	0.93

Table 1: Statistics of participants' performance in control (Full Manual) and experiment (Peanut) conditions. SoC means second of completion per frame, # of frames means the total number of frames that a participant annotates in the condition, cIoU means consensus intersection over union, which is a well-accepted measure for the accuracy of bounding box annotation. The cIoU accuracy of the fully automated model is also provided as a reference.

of 30 video clips from the AVE dataset. The sampled dataset contains 10 different event categories such as music play, car race, male/female speech to investigate the effectiveness of Peanut on videos with a variety of topics. Each of these 30 video clips was manually annotated by two experts as ground truth data for evaluating the accuracy of user annotation. Two authors, who were experts in audio-visual learning, annotated the ground truth data independently using the baseline Full Manual version of Peanut. We used cIoU (see Section 4.3.3) to measure the inter-annotator agreement. The average cIoU score between the annotation results by two experts is 0.96, suggesting a very high agreement between the two expert annotators.

4.2.2 Conditions. The study used a within-subject design, where each participant performed tasks under two conditions in random order. In the experiment condition (Peanut), the participant used the fully functional Peanut tool to label videos from a split of our sample dataset in 25 minutes. In the control condition (Full Manual), the participant used a baseline version of Peanut with all "intelligent features" (object detector, active visual sound grounding, and annotation interpolation) turned off to label the videos from the other split of our sample dataset in 25 minutes. The control condition reflected the essential practices of the current video or image data annotation tools [30, 109, 121]. The split of the videos between the two conditions and the order of the videos in each condition were randomized in each study session. In Table 1, we also include the accuracy score of a "Fully Automated" model using the pre-trained VSG network as a baseline for annotation accuracy.

4.2.3 Procedure. In the study, participants were asked to annotate audio-visual data for sounding object localization using Peanut and the baseline tool. We randomized the order to control for learning effects. The study procedure consisted of three parts: a 30-minute session with the first interface, a 30-minute session with the second interface, and a 10-minute session for the post-study interview and a post-study questionnaire. In each 30-minute session, the experimenter started with a 5-minute tutorial teaching participants how to use the interface in the condition. Subsequently, participants had 25 minutes to annotate as many video frames as possible using the tool provided. After completing both sessions, each participant filled out a post-study questionnaire. The study session ended with a 10-minute semi-structured interview.

---

[5]The study protocol has been reviewed and approved by the IRB at our institution.

4.3 Results on Annotation Performance

We expect PEANUT to accelerate the annotation task for participants in two ways: (H1) the user can annotate each frame faster because of the model-suggested bounding boxes of detected visual objects and the predicted audio tags; (H2) the user needs to manually annotate fewer frames due to the visual-audio-sensitive binary search process.

As shown in Table 1, we report three statistics. The average second of completion (SoC) validates H1, and the # of Frames validates the combined effect of H1 and H2. cIoU measures the impact of using PEANUT on the accuracy of the annotated data. When reporting those statistics, we also report the standard deviation among all users in the target condition. We will explain each statistics and its results in this section.

4.3.1 Time to completion. We calculated the average seconds of completion (Average SoC) on human-annotated frames. SoC measures the average time participants spent annotating each video frame (not including those automatically annotated by the model). The difference between the two conditions in SoC demonstrates the effectiveness of PEANUT in reducing the effort and cognitive load in the annotation of individual frames, such as displaying the potential candidates of objects and removing silent objects. As shown in Table 1, the average SoC is 5.12 per frame ($SD = 1.87$) in the experiment condition and 7.73 per frame ($SD = 2.23$) in the control condition. The paired t-test showed that there is a significant difference between the average SoC under the two conditions ($p < 0.01$), indicating that participants can annotate a frame faster with PEANUT than with the baseline interface.

4.3.2 The number of annotated frames. We calculated the average number of frames that a participant annotated in a 25-minute session (# of Frames). Note that the count includes the frames automatically annotated by PEANUT in the experiment condition. In addition to capturing the effect of PEANUT features reflected in average SoC, the difference in the average number of frames between two conditions also reflects the effectiveness of automatic annotation in PEANUT—instead of annotating all the video frames as in the baseline condition, participants only need to annotate key frames identified by PEANUT and verify automatic annotation in the experiment condition. As shown in Table 1, the average number of annotated frames is 488.85 ($SD = 167.93$) in the experiment condition, and 169.45 ($SD = 68.26$) in the control condition. The paired t-test showed that there is a significant difference between the average number of frames annotated under the two conditions ($p < 0.001$), indicating that participants can annotate more frames with PEANUT than with the baseline interface in a 25-minute session.

4.3.3 Annotation accuracy. We used *consensus intersection over union* (cIoU) [98] to assess the participants' annotation accuracy in each condition. cIoU is a common metric for quantitatively evaluating the accuracy of bounding box annotations. Given a video frame, cIoU assigns scores to each pixel according to the consensus of multiple expert annotations. In specific, the ground-truth bounding boxes annotated by experts are first converted into binary maps $\{\mathbf{b}_j\}_{j=1}^N$, where N is the number of expert annotators. Then, we calculate a representative score map $\mathbf{g}$ from $\{\mathbf{b}_j\}$ considering the consensus of experts:

$$\mathbf{g} = min(\sum_{j=1}^N \frac{\mathbf{b}_j}{\#consensus}, 1) \qquad (2)$$

where #consensus is a parameter indicating the minimum number of expert annotations to reach agreement. Since we have two expert annotators, we set #consensus=1 by the majority rule in our study. Given this weighted score map $\mathbf{g}$ and participant's annotation α, we define cIoU as:

$$cIoU = \frac{\sum_{i \in A} \mathbf{g}_i}{\sum_i \mathbf{g}_i + \sum_{i \in A-G} 1} \qquad (3)$$

where i indicates the pixel index of the map, $A = \{i | \alpha_i = 1\}$ means the set of pixels that the participant annotates, $G = \{i | g_i > 0\}$ means the set of pixels in the weighted ground truth annotation.

We calculate the cIoU score for each annotated frame of each participant. The average cIoU score for all annotated frames (all by human) in the control condition is 0.72 ($SD = 0.14$), and is 0.93 ($SD = 0.09$) for all annotated frames (by human or by the system) in the experiment condition. A paired sample t-test showed a significant difference between the average cIoU under the two conditions ($p < 0.001$). For comparison, the cIoU of a fully-automated VSG (off-the-shelf pre-trained without any human annotation) on the same sample of videos is 0.33 ($SD = 0.08$). The paired t-test shows that the cIoU for the fully-automated model is significantly lower ($p < 0.001$) than both human-annotated conditions (Full Manual and PEANUT). The result indicates that the use of PEANUT can achieve a high annotation accuracy (and even higher than the full-manual control condition in our experiment).

We suspect two possible reasons for the observed improvement in accuracy in the PEANUT condition compared to the Full Manual condition: (1) The use of object detectors may have improved consistency in selecting the bounding boxes. The annotation of some sounding objects can be inherently ambiguous—for example, when a sounding object is a person playing the violin, should the annotator draw a box for the person (that includes the violin) or only the violin? The labeling of situations like this can be inconsistent in the Full Manual condition (e.g., the violin is selected in some frames, while the person is selected in other frames). In this example situation, the objector detector would consistently go with the person because it prefers larger boxes when choosing from two overlapped ones (which is often consistent with the common best practice of sounding object localization), reducing the bounding box inconsistency in data annotation. (2) Drawing accurate bounding boxes is a challenging task on its own. The bounding boxes drawn manually by users often do not align as accurately with the edges of the object as the model-detected boxes. Moreover, because data annotation with the baseline interface is a highly repetitive and tedious process, the precision of participants' manually created bounding boxes could fluctuate due to fatigue (e.g., not modifying a box when an object has moved "a little bit", but still aligns mostly with the box). We plan to further investigate the factors that contribute to the improvement of accuracy as a future work direction.

4.3.4 The impact of user expertise. Comparing the annotation experience and efficiency, users without prior annotation experience annotated faster than those with prior annotation experience in the

PEANUT condition. The average numbers of frames that participants with and without data annotation experience annotated in the experiment condition were 394.5 ($SD = 95.46$) and 630.4 ($SD = 187.78$) respectively. An unpaired t-test shows that participants without data annotation experience annotated significantly more frames than those with annotation experience ($p = 0.043 < 0.05$). Participants without annotation experience also on average spent less time on each frame ($AVG = 4.35s, SD = 0.75$) than those with annotation experience ($AVG = 6.26s, SD = 0.58$) ($p = 0.031 < 0.05$). There was no significant difference between their average number of frames in the Full Manual condition. The efficiency improvement from using PEANUT (comparing # of Frames between PEANUT and Full Manual conditions) is significant for both groups.

In terms of ML expertise, there was no significant difference in all three metrics (Average SoC, number of Frames, cIoU) among groups of participants with different levels of ML expertise in a one-way ANOVA test. We found no significant difference in annotation accuracy between participants who had no prior annotation experience and those who had prior annotation experience, either.

For the observed difference in efficiency between groups with and without prior annotation experiences, a possible explanation is that users with prior data annotation experience may be more skeptical to automated annotation and thus spent more time double-checking the results. We plan to investigate this phenomenon more closely in future studies.

4.4 Results on User Behaviors and Experiences

4.4.1 Usage statistics. We analyzed the interaction logs and screen recordings of participants in the experiment (PEANUT) condition to understand how participants interacted with PEANUT's AI assistance, especially on the ratio between manual vs. automated annotation and how often they edit the automated annotation results. As we see in Table 1, each participant, on average, annotated 488.85 frames in a session. Among them, an average of 37.2 (7.6%) frames are manually annotated by the participant, 421.7 (86.2%) are automatically annotated without any user modification, 29.9 (6.1%) are first automatically annotated by the model and then modified by the user. In each session, a participant on average resized/moved the model-predicted bounding boxes of visual objects in 12.6 frames (2.6%), created new bounding boxes for visual objects in 14.2 frames (2.9%), and edited the model-predicted audio tags in 3.1 frames (0.6%).

Among the annotations made in the PEANUT condition, the average cIoU of the 13.8% human annotated frames is 0.81 ($SD = 0.12$) and average cIoU of the 86.2% automated annotated frames is 0.96 ($SD = 0.07$). This result indicates that with a small portion of user annotation on mostly "key frames" identified by our algorithm, the model can get very accurate at automated annotations (compared to the average cIoU of 0.33 in the Full Automated condition), indicating the effectiveness of PEANUT's active learning pipeline.

4.4.2 Post-study questionnaire. In a post-study questionnaire, we asked each participant to rate seven statements about the usability, usefulness, and user experience of PEANUT on a 7-point Likert scale from "strongly agree" to "strongly disagree". The results are shown in Figure 5. Specifically, PEANUT scored on average 4.9 ($SD = 1.92$) on "*PEANUT is easy-to-use*", 5.7 ($SD = 2.10$) on "*I can learn*

to use PEANUT easily", 5.6 ($SD = 2.19$) on "*I found the feature of recommending candidate boxes is useful*", 5.15 ($SD = 2.08$) on "*I found the feature of navigating to next frame to label is useful*", 5.15 ($SD = 2.16$) on "*I found the features of reviewing annotation result are useful*", 5.5 ($SD = 2.16$) "*I found the feature of playing audio corresponding to a frame is useful*", and 4.8 ($SD = 2.11$) on "*I enjoy using PEANUT*". The results indicate that our participants generally found PEANUT easy to use, and the design features are useful for their annotations.

When a participant rated a statement lower than "agree", the experimenter would take a note and ask the participant about the specific issues they had encountered and solicit their suggestions for addressing these issues in the interview. We will report on these findings in Section 4.4.3 below.

4.4.3 User experiences, challenges, and feedback. In the interview, we discussed with the participants about their post-study questionnaire responses, the difficulties they encountered when using PEANUT, the different user experiences of annotating data in two conditions, and their suggestions for the design of PEANUT. The leading author first coded all the interview transcripts independently and discussed the codebook with another author to reach a consensus. The unified codebook was then used by the other author to code all the interview transcripts independently again to validate the result. The overall inter-rater reliability is 0.72.

Most of the reported difficulties originated from the default positions and sizes of the canvas and the annotation box. Due to an implementation bug, the video frames would be shown in a smaller window in the upper left corner and users "*had to drag and resize them every time*" (P13). The panel for predicted bounding boxes may also appear outside the frame sometimes, so the user has to drag it back to the current view. Both of these implementation issues can be easily fixed in new versions of PEANUT. Besides, after carefully watching the video multiple times, participants still sometimes had trouble recognizing the sound type or locating the source due to "*the background noise, ambiguity of the sound, and existence of multiple objects of a similar type*" (P15). Lastly, a few participants found the difference between the *Next* button and *Next Label* confusing, especially in videos that had rapid visual or auditory changes, for which PEANUT tends to conservatively recommend the immediate next frame when participants clicked the *Next Label* button, which could confuse the user (P13).

When asked to describe the pros and cons of the AI-assisted PEANUT system compared to the baseline, the 20 participants mentioned that PEANUT significantly accelerated their annotation process and reduced their workload. Participants thought "*PEANUT saved their time from doing repetitive and tedious manual labeling, especially when frames varied little once a time*" (P6). With the assistance of PEANUT, they only need to "*focus on a handful of key frames*" (P10) or "*keep clicking the Next button to oversee the automatic annotation frame-by-frame*" (P12). In addition, some participants thought that the systems recommended more accurate bounding boxes than what they can manually create, thus they did not need to "*struggle with creating precise boxes manually*" (P13). Compared to the baseline tool, PEANUT "*makes the annotation process much less exhausting*" (P13). On the weakness of PEANUT, participants thought
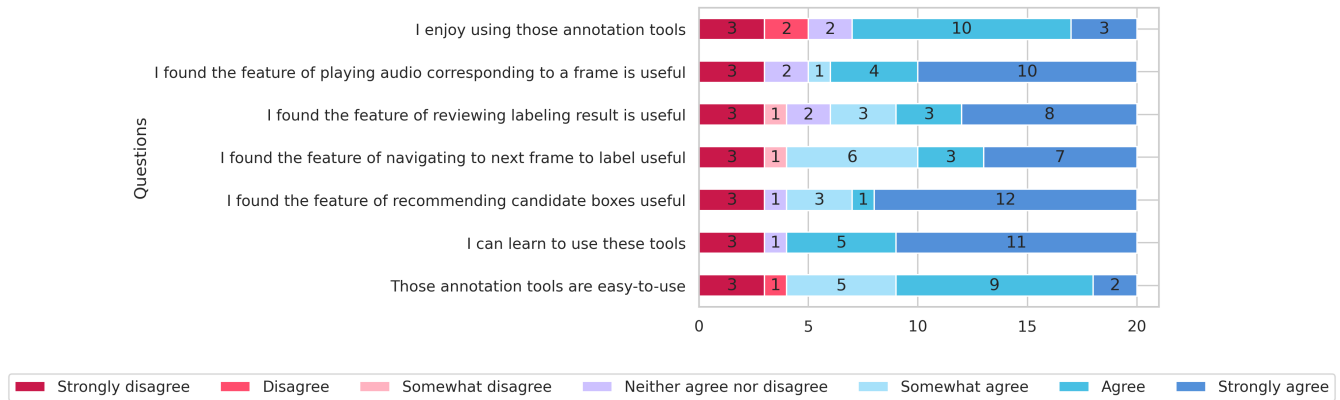
Figure 5: Results of the post-study questionnaire

that with the *Next Label* button they may easily overtrust the system from the beginning, and "*merely realized the wrong or missing annotation when reviewing back at the end*" (P4).

With regard to human agency, participants thought the system "*allow them to jump in and take control back at any time when they feel the AI model start going wrong*" (P16). P9 suggested that "*the system should provide more information or clear warning to them so that they can better notice the issue and decide when to get in*". Moreover, as to their reliance on AI models, some participants said they "*tend to trust AI more when the sounding object was obvious in the video and AI correctly labelled them for consecutive frames*" (P10). In contrast, they "*would choose to do it on their own when sounding object changes a lot*" (P5).

For the design of PEANUT. P4 suggested enabling keyboard shortcuts for frequently used operations such as drawing, deletion, and frame switch. Furthermore, some participants thought "*the algorithm behind the Next Label button should be more robust*" (P15) especially "*when the scene contains complex objects or has a large variation between the frames*" (P12).

5 DISCUSSION AND IMPLICATIONS

In this section, we discuss several novel interaction strategies that we used in the design of PEANUT, the lessons we learned in the user study, and their implications for human-AI collaboration.

5.1 Connecting and Unifying Different Data Modalities

The design of PEANUT introduces a strategy for human-AI collaboration in multi-modal tasks that demand high perceptive and cognitive overhead to process and associate input from different modalities. PEANUT uses two single-modal models that can provide partial automation. The human user then contributes by (1) validating the partial automation result; (2) associating the partial automation result of one model from one input modality with that of a different model from a different input modality to achieve the multi-modal end goal. Many participants found this partial automation approach effective for reducing cognitive load and fatigue when annotating a large number of similar frames. For example, P8 said "*I almost fell asleep

when using the first interface (baseline). In comparison, the second one (PEANUT) allowed me to focus on those (frames) worth my effort so that I did not need to do a tedious, repeated annotation for every picture*".

This new interaction strategy represents the adoption and development of classic theories in multi-modal interfaces [83, 84] in a new domain. Instead of an interface that processes user input data in multiple (usually complementary) modalities and tries to better understand the input from one modality using the input from another modality (mutual disambiguation [82]), PEANUT works in the reversed direction, utilizing humans' perceptive and cognitive capabilities of understanding inputs from one modality using inputs from another modality in order to ground auditory data to visual data. The strategy discussed in this paper and design implications from the study could translate to other multi-modal data annotation tasks like natural language visual grounding [5] visual question answering [6]. Besides, we expect this interaction strategy to be useful in other application domains beyond data annotation, such as helping users with visual or hearing impairments better understand videos as discussed in Section 6.

5.2 Minimizing the Overhead Cost of AI in Human-AI Collaboration

When we design the user workflow of PEANUT, an important goal is that the use of PEANUT should not introduce additional burdens or learning barriers compared to what the user would experience if they manually annotated the same data. The user does not need to learn a new skill or make any configuration of PEANUT before starting to use it. Anything the user does in PEANUT is either a sub-process of what they would need to do in manual annotation (e.g., choosing the sounding object from several candidate boxes instead of identifying the sounding object in the video and drawing a box for it) or the same process but less repetitive (e.g., annotating for only the key frames instead of all the frames). This is different from many other human-AI collaboration scenarios, where the use of AI incurs significant overhead, requiring justifying the use of AI by assessing whether the benefit exceeds the cost.

To achieve this goal, we used a strategy of keeping the user's original workflow as much as possible. For example, PEANUT can identify the next key frame that the user can label (Section 3.3.2).

This intelligent feature does not require any additional input from the user, as it relies on only the user's annotation result for the current frame and the current state of the frame sequence. The user's interaction with the system also remains the same as in a manual labeling tool—they click on the "Next" button to work on a different frame after finishing annotating the current one with the only difference being that the "next" frame is no longer necessarily the $(n+1)^{th}$ frame after the nth frame. The user retains the ability to freely move along the sequence of frames, edit previously annotated frames, or annotate frames out of the "recommended" order as they wish. Peanut will automatically track the next recommended frame to be annotated without requiring user intervention, allowing users to retain control and agency in the human-AI collaborative data annotation process. The participants confirmed the effectiveness of human-AI collaboration in the interview. For example, P5 said that "*I felt it is smooth to work with the algorithm behind Peanut. It takes care of many simple annotations that had to be done by myself in the previous fully manual version*".

Another common type of "cost" in human-AI collaboration that we address in Peanut is the degraded accuracy due to users' overreliance on AI, which has been identified as a key issue in AI-assisted data annotation [9]. When the audio-visual sound grounding model identifies candidate objects and removes likely silent objects from the candidates, we still expect the user to maintain their attention and edit the results if needed. To address this challenge, Peanut provides video playback and thumbnail preview features (Section 3.3.3) that allow the user to quickly validate the result and identify annotation issues if there are any. In the user study, these features indeed helped participants locate incorrect automatic annotations and consequently calibrate their trust in AI automation. For example, P14 said that "*I totally trusted the system to label the frames at the beginning. However, as I looked at the review of the annotation result, I realized the system could make mistakes and I have to go back to improve the annotation*".

5.3 The Role of Partial Automation in Pursuit of Full Automation

The traditional workflow of data annotation for ML models represents an approach that goes from *full manual* efforts directly to *full automation*—human users go through a fully manual process to create a dataset, which is then used to train a fully-automated ML model. In contrast, Peanut's approach highlights the role of incrementally-trained partial-automation models that can bridge the two ends in pursuit of full automation.

In Peanut's model, as the user is annotating the data in a manual process, partial-automation models are incrementally trained with the user's incomplete annotation of each frame. Thank to the characteristic of video data that frames in the same video are usually similar to each other, domain-general partial-automation models can quickly adapt to the specific domain of the video in a *few-shot* fashion. The active learning process guided by the key frame selections (Section 3.4.2) in Peanut incrementally improves the performance of partial-automation models as the user annotates more frames, reducing human efforts to reach the data size and data quality required for training an end-to-end fully-automated model. During the interview, P4 commented: "*One feature I can imagine is that, with*

my annotation on a few frames at the beginning of a video, the system can learn to label the following frames or even other videos with similar content and do the remaining annotation on behalf of me".

We expect that such an approach can be useful for a variety of other human-in-the-loop ML applications. An example is the human/ML hybrid sensing approach such as Zensor [59] where sensors can switch between crowd intelligence and ML to adapt to environmental changes. However, unlike Zensor which toggles between either full automation or full manual for the primary sensing task and uses human annotation results as a validation method, Peanut's approach allows for more flexible partial-automation states in between, taking advantage of the partially annotated data to accelerate the annotation before a fully automated model is ready.

5.4 Mitigating Biases in AI-Assisted Data Annotation

While the issue of biases is less prominent in task domain of Peanut since unlike many other domains vulnerable to subjective bias (e.g., hate speech detection [74]), the result of sounding object localization does come with an objective truth, mitigating the biases of AI and human annotators in this task is still an important consideration. Presumably, the addition of AI assistance in data annotation could introduce or amplify two kinds of biases in annotation.

First, AI models in Peanut may introduce intrinsic biases to annotation results [14, 26, 60]. These biases originated from the dataset on which these models are pre-trained on. To address them, Peanut identifies key frames with previously unseen objects and proactively requests human annotation in these key frames. On those key frames, the role of AI is to *assist* with human annotation by suggesting label candidates for user to use rather than attempting to fully automating their annotations using the pre-trained model. The design allows users to take control and accountability of annotations on keyframes and offsets AI biases with human judgement.

In addition to introducing biases themselves, AI models may also amplify human errors when performing automated annotation based on previous human annotations [1, 13, 21]. For example, when there are multiple guitars in a key frame in the scene, the human annotator may have difficulty in identifying which guitar is currently making the sound and select a wrong guitar as the sound source. In this case, the object detector in Peanut will propagate this wrong annotation to automated annotations in subsequent frames. The playback reviews (Section 3.3.3) could be useful to mitigate this issue by providing a global context that facilitates users to identify errors in continuous scenes.

We also expect to introduce other bias-reducing strategies in the future version of Peanut. For example, we may implement assistance functions to support human decision-making of the sound source at key frames when the human annotator is unsure. For example, Peanut may leverage the model-inferred depth and direction information of the sound to indicate the likelihood of each visual region in the frame containing the sound source object.

6 LIMITATIONS & FUTURE WORK

6.1 Incorporating Speech and Natural Language Models

The current version of PEANUT only supports the use of auditory and visual models, but not natural language understanding (NLU) models that process the content of speech in videos. Speech is one of the most ubiquitous sound sources in videos. Human speech usually contains important semantic information relevant to the surrounding audio and visual scenes in physical environments. Moreover, unlike other sounds (*e.g., traffic, rain, dog barking*), speech can be transcribed into text using automatic speech recognition [46]. Benefitting from advances in NLU [11, 22, 97, 102], we can automatically extract structured and abstracted semantic content from transcribed text. This could provide opportunities for us to incorporate NLU models with PEANUT to facilitate more powerful multimodal data collection to support multidisciplinary research across audio, visual, and language.

6.2 Expanding to Audio-Visual Tasks beyond Sounding Object Localization

We implemented and evaluated the current version of PEANUT in the context of the sounding object localization task. With its support for user interaction with both audio and visual modalities, PEANUT can be easily expanded to a range of multi-modal audio-visual tasks, such as audio-visual event localization [106, 118], audio-visual video parsing [105, 117], and audio-visual video captioning [89, 103].

Audio-visual event localization aims to temporally localize audio-visual events[6] and recognize the categories of events. Toward more unified multi-sensory perception, audio-visual video parsing aims to recognize event categories bind to sensory modalities and find temporal boundaries of when such an event starts and ends. To train models for addressing these two tasks, two datasets: AVE [106] and LLP [105] have been collected, respectively. Due to the lack of efficient annotation tools, only second-level temporal boundaries with the corresponding categories are fully manually annotated in these datasets. We believe that more precise frame-level annotations will enable more accurate audio-visual ML models and facilitate the development of future research.

In addition to facilitating existing tasks, our system has the potential to help researchers investigate new problems. For example, by collecting temporal boundary, object box, and category annotations, we can formulate a new space-time audio-visual parsing task that aims to perform spatio-temporal multi-modal analysis over videos to predict temporal boundaries of audio, visual, and audio-visual events, their associated semantic categories, and spatially localized sounding objects.

6.3 Release and Deployment

We plan to release the PEANUT tool for public use. We are also currently planning a large-scale deployment to complete the annotation of all 4,143 video clips in the AVE dataset [106] for sounding object localization. The annotation result on the full AVE dataset

---

will allow us to train a new *supervised* sounding object localization model with the dataset, compare its performance with the current state-of-art model, and illustrate PEANUT's effectiveness in improving the performance of ML models by enabling the creation of better-quality annotated datasets.

7 CONCLUSION

In this paper, we presented PEANUT, a human-AI collaborative audio-visual annotation tool for improving the data annotation efficiency of the sounding object localization task. A controlled user study of PEANUT demonstrated that a human-AI collaborative approach with several new mixed-initiative partial-automation strategies can enable human annotators to perform the data annotation task faster while maintaining high accuracy. Our findings provide design implications for AI assistance in data annotation as well as human-AI collaboration tools for working with multi-modal data.

REFERENCES

[1] Saad Bin Ahmed, Saif Ali Athyaab, and Shaik Abdul Muqtadeer. 2021. Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach. *2021 6th International Conference on Inventive Computation Technologies (ICICT)* (2021), 557–563.
[2] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7558–7567.
[3] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. 2021. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 265–274.
[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233
[5] Hazan Anayurt, Sezai Artun Ozyegin, Ulfet Cetin, Utku Aktas, and Sinan Kalkan. 2019. Searching for Ambiguous Objects in Videos using Relational Referring Expressions. In *Proceedings of the British Machine Vision Conference (BMVC).*
[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision.* 2425–2433.
[7] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision.* 609–617.
[8] Relja Arandjelovic and Andrew Zisserman. 2018. Objects that sound. In *ECCV.*
[9] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T. Wolf, Evelyn Duesterwald, Casey Dugan, Werner Geyer, and Darrell Reimer. 2021. AI-Assisted Human Labeling: Batching for Efficiency without Overreliance. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 89 (April 2021), 27 pages. https://doi.org/10.1145/3449163
[10] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems* 29 (2016), 892–900.
[11] Jonathan Berant and Percy Liang. 2014. Semantic parsing via paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 1415–1425.

---

[6]Audio-visual events are synchronized video segments in which the sound sources are visible and their sounds are audible.

[12] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, et al. 2019. Ilastik: interactive machine learning for (bio) image analysis. *Nature methods* 16, 12 (2019), 1226–1232.

[13] Shruti Bhargava and David Forsyth. 2019. Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models. *ArXiv* abs/1912.00578 (2019).

[14] William Blanzeisky and Padraig Cunningham. 2021. Algorithmic Factors Influencing Bias in Machine Learning. In *PKDD/ECML Workshops*.

[15] José Bobes-Bascarán, Eduardo Mosqueira-Rey, and David Alonso-Ríos. 2021. Improving medical data annotation including humans in the machine learning loop. *Engineering Proceedings* 7, 1 (2021), 39.

[16] Anthony Brew, Derek Greene, and Pádraig Cunningham. 2010. The interaction between supervised learning and crowdsourcing. In *NIPS workshop on computational social science and the wisdom of crowds*.

[17] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300234

[18] Steve Cassidy and Thomas Schmidt. 2017. *Tools for multimodal annotation*. Springer, Springer Nature, United States, 209–227. https://doi.org/10.1007/978-94-024-0881-2_7

[19] Chia-Ming Chang, Chia-Hsien Lee, and Takeo Igarashi. 2021. Spatial Labeling: Leveraging Spatial Layout for Improving Label Quality in Non-Expert Image Annotation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 306, 12 pages. https://doi.org/10.1145/3411764.3445165

[20] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. 2020. Learning to set waypoints for audio-visual navigation. *arXiv preprint arXiv:2008.09622* (2020).

[21] Yunliang Chen and Jungseock Joo. 2021. Understanding and Mitigating Annotation Bias in Facial Expression Recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 14960–14971.

[22] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 167–176.

[23] Youngwon Choi, Marlena Garcia, Steven S Raman, Dieter R Enzmann, and Matthew S Brown. 2022. AI-human interactive pipeline with feedback to accelerate medical image annotation. In *Medical Imaging 2022: Computer-Aided Diagnosis*, Vol. 12033. SPIE, 741–747.

[24] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research* 4 (1996), 129–145.

[25] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

[26] Yehuda Dar, Vidya Muthukumar, and Richard Baraniuk. 2021. A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning. *ArXiv* abs/2109.02355 (2021).

[27] Meltem Demirkus, James J Clark, and Tal Arbel. 2014. Robust semi-automatic head pose labeling for real-world face video sequences. *Multimedia Tools and Applications* 70, 1 (2014), 495–523.

[28] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, and Qian Pan. 2021. *Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface*. Association for Computing Machinery, New York, NY, USA, 392–401. https://doi.org/10.1145/3397481.3450698

[29] Carl Doersch, Abhinav Gupta, and Alexei A Efros. 2015. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*. 1422–1430.

[30] Abhishek Dutta and Andrew Zisserman. 2019. The VIA Annotation Software for Images, Audio and Video. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2276–2279. https://doi.org/10.1145/3343031.3350535

[31] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *TOG* (2018).

[32] Sara Evensen, Chang Ge, and Cagatay Demiralp. 2020. Ruler: Data Programming by Demonstration for Document Labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics,

Online, 1996–2005. https://doi.org/10.18653/v1/2020.findings-emnlp.181

[33] Sainyam Galhotra, Udayan Khurana, Oktie Hassanzadeh, Kavitha Srinivas, Horst Samulowitz, and Miao Qi. 2019. Automated Feature Enhancement for Predictive Modeling using External Knowledge. In *2019 International Conference on Data Mining Workshops (ICDMW)*. 1094–1097. https://doi.org/10.1109/ICDMW.2019.00161

[34] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhan, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis. *arXiv preprint arXiv:2304.07366* (2023).

[35] Ruohan Gao, Yen-Yu Chang, Shivani Mall, Li Fei-Fei, and Jiajun Wu. 2021. ObjectFolder: A Dataset of Objects with Implicit Visual, Auditory, and Tactile Representations. *arXiv preprint arXiv:2109.07991* (2021).

[36] Ruohan Gao, Rogerio Feris, and Kristen Grauman. 2018. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 35–53.

[37] Ruohan Gao and Kristen Grauman. 2019. 2.5D Visual Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Ruohan Gao and Kristen Grauman. 2019. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3879–3888.

[39] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L Glassman, and Toby Jia-Jun Li. 2023. Patat: Human-ai collaborative qualitative coding with explainable interactive rule synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[40] Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, and Jeffrey Heer. 2011. Proactive Wrangling: Mixed-Initiative End-User Programming of Data Transformation Scripts. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 65–74. https://doi.org/10.1145/2047196.2047205

[41] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[42] Mohamed Hamroun, Karim Tamine, and Benoît Crespin. 2021. Multimodal Video Indexing (MVI): A New Method Based on Machine Learning and Semi-Automatic Annotation on Large Video Collections. *International Journal of Image and Graphics* (2021), 2250022.

[43] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A Survey of the State-of-the-Art. *Knowledge-Based Systems* 212 (2021), 106622.

[44] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[45] John R Hershey and Javier R Movellan. 2000. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*. 813–819.

[46] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* 29, 6 (2012), 82–97.

[47] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. ACM, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[48] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep Multimodal Clustering for Unsupervised Audiovisual Learning. In *CVPR*.

[49] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. 2020. Discriminative Sounding Objects Localization via Self-supervised Audiovisual Matching. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 10077–10087. https://proceedings.neurips.cc/paper/2020/file/7288251b27c8f0e73f4d7f483b06a785-Paper.pdf

[50] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. 2021. Class-aware sounding objects localization via audiovisual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[51] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business horizons* 61, 4 (2018), 577–586.

[52] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926. https://doi.org/10.1109/TVCG.2012.219

[53] Prannay Kaul, Weidi Xie, and Andrew Zisserman. 2022. Label, verify, correct: A simple few shot object detection method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14237–14247.

[54] Kenneth L Kehl, Wenxin Xu, Alexander Gusev, Ziad Bakouny, Toni K Choueiri, Irbaz Bin Riaz, Haitham Elmarakeby, Eliezer M Van Allen, and Deborah Schrag. 2021. Artificial intelligence-aided clinical annotation of a large multi-cancer genomic dataset. *Nature communications* 12, 1 (2021), 7304.

[55] Einat Kidron, Yoav Y Schechner, and Michael Elad. 2005. Pixels that sound. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 88–95.

[56] You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. 2021. Look Who's Talking: Active Speaker Detection in the Wild. In *Proc. Interspeech 2021*. 3675–3679. https://doi.org/10.21437/Interspeech.2021-2041

[57] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

[58] Ellen J Langer. 1989. Minding matters: The consequences of mindlessness–mindfulness. In *Advances in experimental social psychology*. Vol. 22. Elsevier, 137–173.

[59] Gierad Laput, Walter S. Lasecki, Jason Wiese, Robert Xiao, Jeffrey P. Bigham, and Chris Harrison. 2015. Zensors: Adaptive, Rapidly Deployable, Human-Intelligent Sensor Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1935–1944. https://doi.org/10.1145/2702123.2702416

[60] Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. 2020. Mitigating Gender Bias in Machine Learning Data Sets. *ArXiv* abs/2005.06898 (2020).

[61] Toby Jia-Jun Li, Amos Azaria, and Brad A. Myers. 2017. SUGILITE: Creating Multimodal Smartphone Automation by Demonstration. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 6038–6049. https://doi.org/10.1145/3025453.3025483

[62] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M. Mitchell, and Brad A. Myers. 2020. Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (UIST 2020)*. ACM. https://doi.org/10.1145/3379337.3415820

[63] Toby Jia-Jun Li, Marissa Radensky, Justin Jia, Kirielle Singarajah, Tom M. Mitchell, and Brad A. Myers. 2019. PUMICE: A Multi-Modal Agent that Learns Concepts and Conditionals from Natural Language and Demonstrations. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology (UIST 2019)*. ACM. https://doi.org/10.1145/3332165.3347899

[64] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[65] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[66] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[67] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. 2019. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2002–2006.

[68] Minzhe Liu, Li Du, Yuan Du, Ruofan Guo, and Xiaoliang Chen. 2020. Faster Human-Machine Collaboration Bounding Box Annotation Framework Based on Active Learning. (2020).

[69] Sijia Liu, Parikshit Ram, Deepak Vijaykeerthy, Djallel Bouneffouf, Gregory Bramble, Horst Samulowitz, Dakuo Wang, Andrew Conn, and Alexander Gray. 2020. An ADMM based framework for automl pipeline configuration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4892–4899.

[70] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487* (2019).

[71] Ryan Louie, Andy Coenen, Cheng Zhi Huang, Michael Terry, and Carrie J. Cai. 2020. Novice-AI Music Co-Creation via AI-Steering Tools for Deep Generative Models. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376739

[72] Yaxiong Ma, Yixue Hao, Min Chen, Jincai Chen, Ping Lu, and Andrej Košir. 2019. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion* 46 (2019), 184–192.

[73] Tanushree Mitra, C.J. Hutto, and Eric Gilbert. 2015. Comparing Person- and Process-Centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1345–1354. https://doi.org/10.1145/2702123.2702553

[74] Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems* (2020), 1–16.

[75] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. 2018. Self-supervised generation of spatial audio for 360 video. *arXiv preprint arXiv:1809.02587* (2018).

[76] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300356

[77] Michael Muller, Christine T. Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, and Casey Dugan. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 94, 16 pages. https://doi.org/10.1145/3411764.3445402

[78] Micah M Murray and Mark T Wallace. 2011. The neural bases of multisensory processes. (2011).

[79] Mariana Neves and Ulf Leser. 2014. A survey on annotation tools for the biomedical literature. *Briefings in bioinformatics* 15, 2 (2014), 327–340.

[80] Zheng Ning, Zheng Zhang, Tianyi Sun, Yuan Tian, Tianyi Zhang, and Toby Jia-Jun Li. 2023. An empirical study of model errors and user error discovery and repair strategies in natural language database queries. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 633–649.

[81] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. 2017. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing* 10, 1 (2017), 60–75.

[82] Sharon Oviatt. 1999. Mutual disambiguation of recognition errors in a multi-modal architecture. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 576–583.

[83] Sharon Oviatt. 1999. Ten Myths of Multimodal Interaction. *Commun. ACM* 42, 11 (Nov. 1999), 74–81. https://doi.org/10.1145/319382.319398

[84] Sharon Oviatt and Philip Cohen. 2000. Perceptual User Interfaces: Multimodal Interfaces That Process What Comes Naturally. *Commun. ACM* 43, 3 (March 2000), 45–53. https://doi.org/10.1145/330534.330538

[85] Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 631–648.

[86] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 747–759.

[87] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. 2020. Multiple Sound Sources Localization from Coarse to Fine. In *ECCV*.

[88] Nan Qiao, Yuyin Sun, Chongyu Liu, Lu Xia, Jiajia Luo, K. Zhang, and Cheng-Hao Kuo. 2022. Human-in-the-Loop Video Semantic Segmentation Auto-Annotation. (2022).

[89] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. 2019. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8908–8917.

[90] Al Mamunur Rashid, Istvan Albert, Dan Cosley, Shyong K. Lam, Sean M. McNee, Joseph A. Konstan, and John Riedl. 2002. Getting to Know You: Learning New User Preferences in Recommender Systems. In *Proceedings of the 7th International Conference on Intelligent User Interfaces* (San Francisco, California, USA) *(IUI '02)*. Association for Computing Machinery, New York, NY, USA, 127–134. https://doi.org/10.1145/502716.502737

[91] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. 269.

[92] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015), 91–99.

[93] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. 2021. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 41–50.

[94] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[95] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to Do the Model Work, Not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 39, 15 pages. https://doi.org/10.1145/3411764.3445518

[96] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 253–260.

[97] Michael Schmitz, Stephen Soderland, Robert Bart, Oren Etzioni, et al. 2012. Open language learning for information extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. 523–534.

[98] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4358–4366.

[99] Burr Settles. 2012. Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018 arXiv:https://doi.org/10.2200/S00429ED1V01Y201207AIM018

[100] Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, Dakuo Wang, et al. 2022. Label Sleuth: From Unlabeled Text to a Classifier in a Few Hours. *arXiv preprint arXiv:2208.01483* (2022).

[101] Eyal Shnarch, Alon Halfon, Ariel Gera, Marina Danilevsky, Yannis Katsis, Leshem Choshen, Martin Santillan Cooper, Dina Epelboim, Zheng Zhang, Dakuo Wang, Lucy Yip, Liat Ein-Dor, Lena Dankin, Ilya Shnayderman, Ranit Aharonov, Yunyao Li, Naftali Liberman, Philip Levin Slesarev, Gwilym Newton, Shila Ofek-Koifman, Noam Slonim, and Yoav Katz. 2022. Label Sleuth: From Unlabeled Text to a Classifier in a Few Hours. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics. https://arxiv.org/abs/2208.01483

[102] Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. *arXiv preprint arXiv:1805.02473* (2018).

[103] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. 2018. An attempt towards interpretable audio-visual video captioning. *arXiv preprint arXiv:1812.02872* (2018).

[104] Yapeng Tian, Di Hu, and Chenliang Xu. 2021. Cyclic Co-Learning of Sounding Object Visual Grounding and Sound Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2745–2754.

[105] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. 2020. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 436–454.

[106] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.

[107] Szilárd Vajda, Yves Rangoni, and Hubert Cecotti. 2015. Semi-Automatic Ground Truth Generation Using Unsupervised Clustering and Limited Manual Labeling. *Pattern Recogn. Lett.* 58, C (June 2015), 23–28. https://doi.org/10.1016/j.patrec.2015.02.001

[108] Luis Von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. recaptcha: Human-based character recognition via web security measures. *Science* 321, 5895 (2008), 1465–1468.

[109] Kentaro Wada. 2016. labelme: Image Polygonal Annotation with Python. https://github.com/wkentaro/labelme.

[110] Dakuo Wang, Josh Andres, Justin D. Weisz, Erick Oduor, and Casey Dugan. 2021. AutoDS: Towards Human-Centered Automation of Data Science. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama,

Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 79, 12 pages. https://doi.org/10.1145/3411764.3445526

[111] Dakuo Wang, Pattie Maes, Xiangshi Ren, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2021. *Designing AI to Work WITH or FOR People?* ACM, New York, NY, USA. https://doi.org/10.1145/3411763.3450394

[112] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 211 (Nov. 2019), 24 pages. https://doi.org/10.1145/3359313

[113] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[114] Kanit Wongsuphasawat, Zening Qu, Dominik Moritz, Riley Chang, Felix Ouk, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2017. Voyager 2: Augmenting Visual Analysis with Partial View Specifications. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2648–2659. https://doi.org/10.1145/3025453.3025768

[115] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

[116] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

[117] Yu Wu and Yi Yang. 2021. Exploring Heterogeneous Clues for Weakly-Supervised Audio-Visual Video Parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1326–1335.

[118] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. 2019. Dual Attention Matching for Audio-Visual Event Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[119] Dean Wyatte. 2019. De-biasing Weakly Supervised Learning by Regularizing Prediction Entropy. (2019).

[120] Zhujun Xiao, Yanzi Zhu, Yuxin Chen, Ben Y. Zhao, Junchen Jiang, and Haitao Zheng. 2018. Addressing Training Bias via Automated Image Annotation. *arXiv: Computer Vision and Pattern Recognition* (2018).

[121] Xtract.io. 2020. *Xtract.io video annotation tool.* https://www.xtract.io/lp/image-annotation-tool

[122] Donghuo Zeng, Yi Yu, and Keizo Oyama. 2018. Audio-visual embedding for cross-modal music video retrieval through supervised deep CCA. In *2018 IEEE International Symposium on Multimedia (ISM)*. IEEE, 143–150.

[123] Yu Zhang, Yun Wang, Haidong Zhang, Bin Zhu, Siming Chen, and Dongmei Zhang. 2022. OneLabeler: A Flexible System for Building Data Labeling Tools. In *CHI Conference on Human Factors in Computing Systems*. 1–22.

[124] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. *arXiv preprint arXiv:2304.07810* (2023).

[125] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The sound of pixels. In *ECCV*.

[126] Zhi-Hua Zhou. 2018. A brief introduction to weakly supervised learning. *National science review* 5, 1 (2018), 44–53.