# Sniffing Location Privacy of Video Conference Users Using Free Audio Channels

Long Huang, Chen Wang
*Department of Computer Science, Southern Methodist University, USA*
*Email: huangl@smu.edu, cwang6@smu.edu*

***Abstract*—Since the outbreak of the COVID-19 pandemic, video conferencing apps have been more broadly used to connect geographically distant people for work, school, and social interactions. These apps simulate "in-person" meetings with streamed audio and provide users with full control of their privacy. For instance, users can conveniently disable their microphones whenever they feel the need for privacy following common senses: 1) Audio signals containing semantic or contextual information pose privacy concerns; 2) Microphones are relevant only to acoustic privacy; 3) Meeting participants cannot actively intrude on each other's privacy but only opportunistically exploit accidental privacy leakages or mistakes. This paper investigates the privacy leakages that defy these assumptions. We find that any meeting participant can actively and covertly probe others' location privacy even when the webcam is disabled or virtual backgrounds are used to hide locations. More specifically, the legitimate two-way audio channel of video conferencing facilitates remote acoustic sensing, allowing an attacker to probe the users' physical surroundings and receive location-specific echo signals.**

**However, all video conferencing systems utilize echo cancellation functions to prevent audio feedback, which inherently stops active sensing. To address this challenge, we develop a transformer-based algorithm and leverage the encoders of generative AI to counteract echo cancellation and extract stable location embeddings from severely distorted echo sounds. Furthermore, we propose two types of active acoustic sensing attacks: the *in-channel echo attack*, which breaks through echo cancellation by using carefully crafted signals, and the *off-channel echo attack*, which exploits third-party media sounds (e.g., email notification tones) to evade cancellation. We test these attacks on commercial video conferencing apps, such as Zoom, Teams, and Skype. When using only a single probing sound, our methods achieve $88.3\%$ accuracy in recognizing recurrent places and $88.5\%$ accuracy in identifying the contexts of new (unseen or untagged) places.**

## 1. Introduction

Since the outbreak of the COVID-19 pandemic, video conferencing systems have been more broadly used to connect geographically dispersed people for work, school, and social interactions. For example, in 2020, Zoom reported having over 300 million meeting participants daily [1], and market analysts have forecast the global video conferencing market size to reach $22.5 billion by 2026 [2]. Video conferencing systems, featuring high-definition video and audio transmissions to simulate "in-person" meetings, give users full control of their privacy. A user can turn off the camera or microphone whenever they feel the need for privacy. They take actions based on their own judgments to prevent privacy leakage, but what if their judgments are wrong?

Location is highly sensitive privacy users want to hide over networks. Accessing a device's location data typically requires high-level permissions, and traditional attacks rely on the successful pre-installation of malware. To bypass the challenges of installing malware, recent studies have focused on exploiting side-channel information, aiming for higher success rates and the potential for large-scale damage. However, in client-server-client communications, the server in the middle can easily disrupt such attacks. For example, IP addresses are often linked to users' geolocations [3], [4], [5], and video conferencing servers thus remove end users' IP addresses before forwarding their packets. A recent study found that the recurring timing characteristics of instant messages can expose messenger users' usual locations [6]. This is because the relative network position of a device can result in specific packet paths [7], [8]. However, this threat has not been observed in video conferencing apps. Moreover, the method can only obtain city- or country-level resolution and is vulnerable to the many path and timing changes introduced by the networks, messenger servers, and VPN use. From a video conferencing user's perspective, it is still widely believed that turning off the webcam or using a virtual background can hide locations and that using a mic poses no threat to location privacy. This work investigates location privacy in video conferencing and aims to achieve room-level location inference using the free audio channels.

Acoustic sensing has been extensively studied for localization, from *infrastructure-based* to the more popular *infrastructure-free* approaches. Rather than relying on the beacons of deployed anchor sensors [9], [10], [11], a mobile device can localize itself by recognizing ambient sounds or actively probing the surroundings through acoustic signals [12], [13], [14]. Active acoustic sensing outperforms ambient sounds for room-level localization because it uses at least one speaker-mic pair to form an acoustic signal loop, like a radar system. The received acoustic signals then exhibit high correlations in close locations due to the similar echo paths [12], [13], [14]. However, due to the limited range of sounds, the sensing loops must be formed locally, and no prior work has successfully demonstrated remote active sensing attacks without the need to install malware or place a hidden sensor close to the user [15], [16], [17].

2025 IEEE Symposium on Security and Privacy (SP) | 979-8-3315-2236-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/SP61157.2025.00260

DOI 10.1109/SP61157.2025.00260

We find that video conferencing provides an opportunity for adversaries to remotely close the acoustic sensing loop using its two-way audio channels. However, the system's Acoustic Echo Cancellation (AEC) prevents the audio feedback, inherently breaking the loop again. To address this, this work reduces the impact of echo cancellation to restore the loop, allowing each echo signal to return to the attacker as much as possible. The proposed methods, for the first time, achieve remote active acoustic sensing. Additionally, we find that an adversary can send two types of probing sounds to sense and tag remote locations: using millisecond-level chirp signals to pass through echo cancellation, or using third-party notification tones to evade it.

To analyze the residue echo sounds from the two attacks, we develop a transformer-based algorithm and leverage generative AI encoders to solve two key challenges: 1) counteracting the effects of echo cancellation to enable learning from severely distorted audio streams, and 2) extracting stable location embeddings from complex media sounds, which are significantly suppressed by video conferencing systems as non-human sounds. Specifically, we derive a series of time-frequency images to analyze the returned audio at three stages: the echo sound and the two ambient sound segments before and after the echo. Based on that, our transformer algorithm utilizes self-attention to extract location embeddings based on both the residue echo sound and how the AEC responds to the attack and restores its state afterward. We test both attacks on current video conferencing apps to tag and infer users' locations and place contexts. Defense mechanisms are then proposed.

**Our contributions are summarized as follows:**

- This work investigates the potential of remote active sensing by exploiting the legitimate audio channels of video conferencing. We find that location privacy can be compromised through a device's audio system, even when users disable webcams or use virtual backgrounds to conceal their surroundings.
- We identify two types of probing sounds that could be transmitted over video conferencing to tag users' locations while surviving echo cancellation: The *in-channel echo attack* exploits $ms$-level chirp signals to penetrate echo cancellation after sensing the users' surroundings, while the *off-channel echo attack* triggers third-party media sounds to perform sensing and evade echo cancellation. It is the first work to use media sounds for location sensing.
- We develop a transformer-based algorithm to recognize locations and place contexts from the returned audios of both echo attacks. Specifically, we leverage the self-attention mechanism to not only recognize the surviving echo components but also learn how the AEC functions behave before and after an attack, which are all found to contain location-dependent information.
- Real-world experiments on commercial video call apps show that the two attacks can accurately recognize users' recurrent places with just a single probing sound. For new, untagged locations, we can effectively infer place contexts, such as home rooms, offices, hotels, and vehicles.

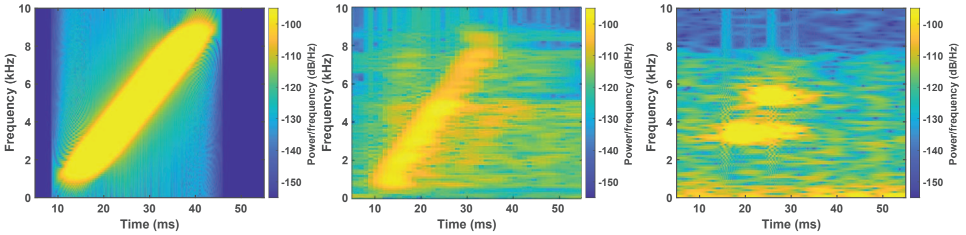
## 2. Related Work

Indoor localization has been an active research area since GPS signals are blocked by the buildings and can not be used for indoors [18]. For example, inertial sensors, accelerometer and gyroscope, are used to estimate the device's displacement and direction [19], [20]. Since inertial sensors suffer from high accumulated errors [21], the more popular ways are using such sensors to count people's walking steps for in-door localization [22], [23]. For the more accurate indoor localization, Radio Frequency (RF) signals, including millimeter waves [24] and WiFi signals [25], have been explored. They rely on existing RF infrastructure or need to deploy RF anchors. Additionally, visual sensors such as RGB camera, Kinect, and LeapMotion can be used for accurate indoor localization, though they are vulnerable to low-light conditions and raise privacy concerns [26].

Acoustic sensing has been extensively explored for indoor localization because of the pervasive availability of speakers and microphones. Related methods can be divided into two categories. Infrastructure-based methods deploy acoustic anchors to estimate the target node's location through acoustic signal transmissions. The time-of-flight (ToF) between the target device and the anchors can be measured to pinpoint the device's location through triangulation with high accuracy [27]. In particular, Cricket [28] achieves an average of $12cm$ localization errors by estimating the ToFs of acoustic signals. The Doppler effect can also be leveraged to estimate the target device's moving direction, which shows a mean angular error within $18$ degrees [29]. Furthermore, the phase shift can be captured to reduce the errors of time synchronization between the target device and the anchors, which is required for obtaining ToFs [30]. Systems with multiple anchors can also calculate the time difference of arrival (TDoA) rather than ToFs to avoid synchronization issues [31].

Infrastructure-free acoustic localizations have become more popular with the advancement of mobile devices, which can be further divided into passive acoustic sensing and active acoustic sensing methods. In passive acoustic localization, the mobile device passively records location-specific ambient sounds to obtain each location's acoustic profile. Location profile models can be created based on machine learning, which later makes localization decisions by comparing the acoustic measurements against the learned profiles [32], [33], [34], [35], [36], [37]. Differently, active acoustic localization sends stimulus signals to sense the surroundings and acquire the location's acoustic signature. For example, RoomSense [38] uses acoustic features of the impulse response to estimate the room locations. EchoTag [14] uses a smartphone to generate high-frequency chirp signals and record the reflections from the surroundings to distinguish the signatures of different indoor locations.

However, existing acoustic localization methods are all local sensing methods, which can not be directly applied in the attacking scenarios we consider. Moreover, they have never considered noise suppression and echo cancellation, the default component of video conferencing systems, which inherently prevents remote active sensing. Different from all



(a) Original Chirp (b) Off-channel Echo (c) In-channel Echo

Figure 1. Two types of echoes of a chirp sound in Zoom calls.

prior works, we explore counteracting the AECs of video conferencing systems to achieve adversarial remote sensing. To achieve the goal, we investigate the design of acoustic signals that can break through or bypass the AECs and develop generative AI encoders to recognize the severely distorted echo signals that survive AECs.

## 3. Background and System Models

### 3.1. Echo Cancellation in Video Conferencing

Video conferencing systems, such as Zoom, Skype, and Microsoft Teams have grown in popularity since the COVID-19 pandemic, as more people transitioned to work remotely. These audio systems must address the issue of audio feedback–the return of acoustic echoes back to the original human speaker, which degrades call quality and the overall online meeting experience [39]. To eliminate acoustic echoes during calls, traditional AEC models adopt Digital Signal Processing (DSP)-based approaches for echo cancellation. However, such methods assume a relatively constant acoustic environment and rely heavily on background noise estimation. Their performance degrades when dealing with the more complicated or time-varying acoustic environments. In addition to traditional AEC methods, deep learning models have recently been increasingly applied for echo cancellation, which learn from historical observations. In particular, many supervised learning algorithms have shown better performance than the DSP-based AECs [40], [41]. Some studies further combine the traditional DSP-based methods with deep learning algorithms to further improve the noise cancellation performance [41], [42].

According to our experimental evaluations, the popular video conferencing apps all employ AEC algorithms to address their audio feedback. These AECs show different echo cancellation performances and are not open-sourced. To make sure that the audio received at one end should not go back to the sender, the end client's AEC deducts a reference audio from the near-end microphone data before transmitting it. The derivation is based on the received audio and multiple estimated factors, including the echo paths, delays (both circuit and signal propagation), volume levels, and signal distortions. These factors can be learned from short-term observations to address dynamic acoustic environments, such as by using least squares FIR adaptive filters and frequency domain adaptive filters [43], [44].

### 3.2. Acoustic Sensing Under Echo Cancellation

To understand the potential and challenges of acoustic sensing in video conferencing, we use a Zoom app to record a remote laptop device's echo sounds. Specifically, we send a 0-10kHz frequency-sweeping sound to the remote laptop through a Zoom call, which is transmitted in the Zoom system's audio channel and then played by the laptop. For comparison, we use the Media Player app on the same laptop to locally play the same chirp sound during the Zoom call, which is off the Zoom system's audio channel. Figure 1 compares the time-frequency images of the original chirp sound and the Zoom recordings of its echo sounds in the two scenarios (i.e., in-channel echo and off-channel echo). It is worth noting that in both scenarios, we can hardly hear any audio feedback indicating that Zoom does a good job of eliminating user-perceivable echoes to ensure a positive user experience. However, the residues of both echoes can still be observed after signal processing as shown in Figure 1(b) and Figure 1(c). There are two main reasons: 1) It is not easy for AEC to predict the echo sounds to generate the precise reference audio [45], because the physical surroundings cause the played sound to be echoed back with complex direct-path, reflected-path and induced vibration sound beams. There are further impacts from ambient noises and human speech. 2) Current AEC functions care only about user experience and have not considered security. It is thus possible that the residue echo signals after cancellation, though not perceivable to humans, could still be machine-recognizable, which enables unauthorized sensing.

When returning to the sender, both echoes exhibit low signal power and severe distortions, which are the combined effects of the ACE algorithm's two major functions, echo cancellation and noise suppression. The noise suppression function actively detects non-human voice sounds picked up by the microphone and reduces them as noises. As a result, even when an echo sound is not fully eliminated, the resulting signal power is further reduced to avoid disrupting the sender. Furthermore, we observe distinct signal patterns between the two echoes, though they originate from the same chirp. Because the in-channel echo is included in the AEC's reference audio, it experiences more significant cancellation and loses its original frequency-sweeping shape, making it hard to discern any relationship to the original chirp. In comparison, the off-channel echo, which avoids inclusion in the reference audio, better retains the frequency-sweeping shape and preserves more frequency components. However, obtaining this off-channel echo requires manipulating the user's device to play a dedicated sensing sound, which remains challenging without relying on malware. Therefore,

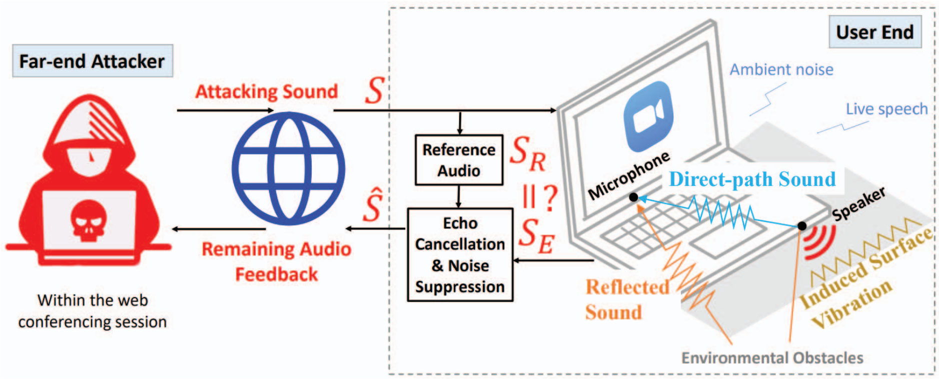


Figure 2. Attacking scenario and echo sensing model.

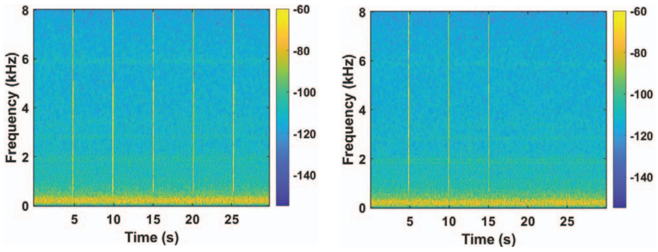
while using both echoes may close the loop for remote acoustic sensing, we still face the challenges of interpreting distorted in-channel echoes and finding practical methods to obtain off-channel echoes. Additionally, both echoes' audio frequencies are limited to $8kHz$ by Zoom, even though laptops support 48kHz to 96kHz acoustic sampling rates. This limitation makes it infeasible to transmit high-resolution signals for sensing or use ultrasonic signals for inaudible attacks over Zoom. Similar observations are obtained with other video conferencing apps, including Teams and Skype.

To further understand the functions of AEC, we play a sequence of five chirp signals (with a 5-second spacing) during a Zoom call. Figure 3 compares the received echoes with and without AEC. In the scenario without AEC (e.g., the regular acoustic sensing scenario), all five chirps are recorded with similar signal patterns, and the frequency sweeping shapes are maintained. But when AEC is applied, the recorded echoes present weaker and weaker signal strengths, and only the first three chirps survive the AEC. The 2nd and 3rd chirps become closer to the ambient noises while the 4th and 5th chirps can not be observed. Even the three survived chirps show different signal patterns, demonstrating the dynamic behaviors of AEC. The reason is that the AEC is implemented based on adaptive filters, which adaptively adjust function parameters according to short-term observations to address dynamic acoustic environments and achieve optimum echo cancellation. The result motivates us to also study how differently AEC responds to the attacking sound and then restores its state to extract more location-related information.

**Echo Sensing Model.** When a user's speaker plays the attacker's audio $S(f)$, the audio feedback to the attacker is the AEC-processed microphone data $\hat{S}(f)$, generally modeled in the frequency domain as

$$\hat{S}(f) = S_H(f) + [S(f)H_E(f) - S(f)H_R(f)] + \alpha N(f), \quad (1)$$

which is the combination of the live human speech $S_H(f)$, the residue echo $S(f)H_E(f) - S(f)H_R(f)$, and the damped ambient noise $\alpha N(f)$. Here, the echo system response $H_E(f)$ describes how the speaker sound is practically echoed back to the microphone by the physical environment as in Figure 2, which results in the echo sound $S(f)H_E(f)$. The reference echo system response $H_R(f)$ is predicted by AEC and is used to generate the reference audio $S(f)H_R(f)$ for echo cancellation. Then, the attacker's aim is simplified as learning the echo system response $H_E(f)$ from the returned audio $\hat{S}(f)$ to decode the physical environment's characteristics while reducing the impacts of human speech, ambient noise and the echo cancellation. Addressing the dynamics of the closed-source AEC and the impact of $H_E(f)$ to enable remote acoustic sensing distinguishes this paper from all prior acoustic works [21], [46], [47], [48], [49], [50], [51], [52], [53], [54], [55], [56].



(a) Without AEC

(b) With AEC

Figure 3. Illustration of AEC's adaptive filters to address audio feedback.

### 3.3. Threat Models

We consider the attacking scenarios when any meeting participant could be an adversary to spy on other users' current whereabouts using the available audio channels of video conferencing. The victim users may already disable the webcam or use virtual backgrounds to hide their locations but need to turn on their microphones to talk. Figure 2 shows the general attack model. A remote attacker creates a long-distance sensing loop by sending one or more probing sounds and receiving the audio feedback (i.e., echo sounds). The probing sounds are played and recorded by the user's own device to complete the location probing. Since video conferencing systems use AEC to prevent all remote participants from hearing echoes, we assume the echo cancellation is not perfect, which leaves residue echoes to return to the attacker carrying sensitive location-specific information.

**Location Privacy Types.** By analyzing the returned echo signals, the adversary can tag meeting participants' lo-

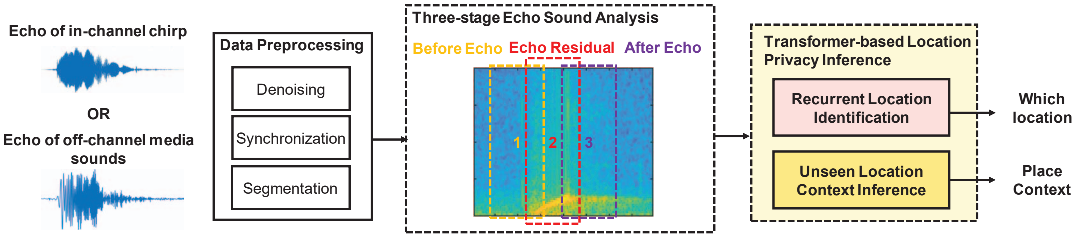

Echo of in-channel chirp
OR
Echo of off-channel media sounds
Data Preprocessing
Denoising
Synchronization
Segmentation
Three-stage Echo Sound Analysis
Before Echo
Echo Residual
After Echo
1
2
3
Transformer-based Location Privacy Inference
Recurrent Location Identification
Unseen Location Context Inference
Which location
Place Context

Figure 4. The flow of our location privacy inference system.

cations with acoustic fingerprints and further achieve unauthorized surveillance. The vulnerable locations include the daily or recurrent places the victim users frequently go to (e.g., office, home rooms and vehicles) and also new places, which have not been tagged by the adversary yet. Considering the vast user base of video conferencing, the attack consequences are enormous. More specifically, based on the attacker's goal and the availability of tagged locations, we consider revealing three types of location privacy.

1) *Binary Location Inference*: In many scenarios, the attacker's goal can be just achieved by one-class classification with a binary inference. For example, a supervisor may want to know whether you are at the office during working hours. A thief or spy may want to know whether you are at home.

2) *Multi-location Privacy Inference*: If the target user's multiple places have been tagged before, the adversary can perform a multi-class classification to recognize where the user is whenever they meet online.

3) *Place Context Inference*: For new locations that have not been tagged before, the adversary can check whether they can be distinguished from tagged ones to confirm that they are new locations. Further, the place context can be determined if its physical settings are similar as home rooms, office rooms, lobby, vehicles or hotel rooms. In addition, multiple calls can leak the user's private living patterns, such as staying at different houses every night or living in hotel rooms frequently [57].

**Two Echo Attacks.** The attacking sound needs to meet two requirements: surviving AEC and being non-invasive to raise no suspicion. We identify two types of echoes that meet both requirements and facilitate remote active sensing:

1) *Off-channel echo attack* sends signals off the video conferencing channels to avoid its inclusion in the reference audio by AEC. We find that the adversary can request other communication services to trigger notification sounds during video calls. To understand this threat, we conducted a video conference privacy survey via Survey Monkey and collected 62 anonymous online responses in one year. The respondents are mainly university students. The survey result shows that 71% of respondents have ever heard the notifications of emails and chat messages in video calls. Thus, using such media sounds for malicious sensing is "steganographic".

2) *In-channel echo attack* still sends probing signals via the video conferencing system's audio channel but requires the signals to break through AEC to close a sensing loop. To avoid raising suspicions, the attacker can use millisecond-level acoustic signals that are short enough to be indistinguishable from network noise when heard.

**Launching Strategies.** The attack is designed to last for a sub-second, allowing an adversary to find many opportunities to launch whenever the user unmutes the mic. Even for a vigilant user, who carefully unmutes the mic only when talking, the adversary can exploit the several silent seconds after unmuting and before muting, because people naturally avoid audio streaming margins to ensure their speech can be completely heard by others. Furthermore, we find that when the user talks, echo sounds return with higher energy. This is because video conferencing systems apply acoustic suppression to the silent user ends to eliminate meaningless audio feedback. Accordingly, the user's speech helps boost the attack sound feedback, which can be easily isolated based on their different frequency and time features (e.g., attack at speech pauses).

## 3.4. System Flow

Figure 4 presents the flow of our attack system. The system takes the echo sounds as the input, which can be the resulting echos of in-channel chirp signals or off-channel media sounds. *Data Preprocessing* is first performed to denoise, synchronize, and segment the audio data. Next, the *Three-stage Echo Sound Analysis* derives the time-frequency representations of the audio signal to capture the embedded location information. In particular, we derive a sequence of three time-frequency images to describe the different stages of the audio feedback, including the audio segments before and after the echo sound with an overlap period. While the residual echo carries the location information that fails to be completely removed by AEC, its two adjacent audio segments contain mainly ambient noises. But due to the dynamic effects of AEC, the two ambient noise segments present slightly different acoustic features showing the transition stages when AEC responds to the attacking sound and restores its state. Because AEC derives the reference audio according to specific physical environments, its behaviors carry location information and can be extracted by analyzing the sequence of audio segments.

The derived three-stage time-frequency images are fed into the *Transformer-based Location Privacy Inference* for further analysis. We develop a classification model based on the transformer encoder. The transformer-based model takes each of the three-stage time-frequency images as the input and utilizes the self-attention mechanism to extract the location information from the residue echo and two transition audio segments that capture how the AEC responds to the attacking sound and then restores its state afterward. Specifically, the model learns the acoustic channel's left-side transition from the $1st$ *stage* time-frequency image and the right-side transition from the $3rd$ *stage* time-frequency image. The algorithm extract stable location embeddings from the distorted residual echo signal in the $2nd$ *stage* time-frequency image. The three-stage time-frequency images describe the complete audio channel state under the echo attack, and their transformer outputs are integrated to make the final location inference.

According to our attack models, we train the transformer model to achieve three different goals. Specifically, for recurrent locations, the transformer model is trained to either identify the user's location as one of multiple tagged locations (multi-class) or determine whether the user's location is the one under consideration (binary). If the location recognition result shows a low confidence score, the transformer model would treat the testing location as a new location and recognize its place contexts instead. If the adversary obtains the name of the new location (e.g., by asking the user or when the webcam is enabled), the transformer model can be updated by including the acoustic tags of this new location.

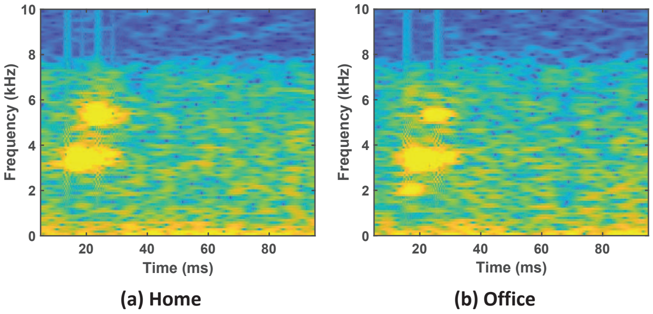
# 4. Approach Design

## 4.1. Sensing Signal Design

The design objective of the probing signal is to minimize the impact of AEC while maximizing privacy inference without arousing suspicion. To achieve this, we propose in-channel echo sensing and off-channel echo sensing.

### 4.1.1. In-channel Chirp Signals.

This is the general way to close a remote active sensing loop using the audio channels of video conferencing. According to the in-channel echo sensing model introduced in Section 3.2, the design of in-channel sensing signal aims to cause the AEC's reference echo system response to fail to match the actual echo system response, thereby maximizing their differences to improve the survival rate of in-channel probing signals. Besides, the signal should also not be noticeable to the user. We find that millisecond-level audible signals in video calls (e.g., $\leq 8kHz$) are indistinguishable from normal noises caused by network transmission (e.g., jitter and packet loss). We thus explore using millisecond-level signals to secretly penetrate AEC for location sensing. Figure 5 presents the initial attempt in Zoom calls with a $25ms$ chirp signal sweeping from $1kHz$ to $8kHz$, when the user is at the office and home, respectively. The signal successfully survives AEC and leaves slightly different time-frequency patterns at the two places, though AEC has significantly disrupted its original frequency-sweeping pattern. We further observe consistent echo patterns when repeating the signal at the same place. When the user talks, after detecting and cropping the millisecond-level echo, we still observe similar patterns but with its SNR enhanced by over $20dB$. This is because the acoustic suppression at the user end is off when user speech is detected, which "boost" the returned attack signal. We next introduce our methods to fine-tune the chirp-like signals to break through the AEC of Zoom. These methods can also be applied to other video conferencing apps, though the resulting attack signals may vary.



(a) Home

(b) Office

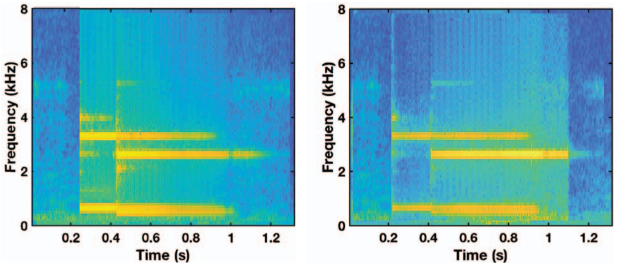
Figure 5. Tagging places with in-channel chirps in Zoom.

**Frequency Selection.** We first test the AEC's echo cancellation effect on the chirp signals of different frequency ranges. In particular, we use a set of chirp signals, which sweep between $0 \sim 1kHz$, $1kHz \sim 2kHz$, $2kHz \sim 3kHz$, $3kHz \sim 4kHz$, $4kHz \sim 5kHz$, $5kHz \sim 6kHz$, $6kHz \sim 7kHz$, and $7kHz \sim 8kHz$, respectively. We then analyze the shapes and SNRs of the recorded chirp echoes. We find that Zoom system's AEC has different impacts on the chirp echoes at different frequency ranges. In particular, the chirp signals within $0 \sim 1kHz$, $4kHz \sim 5kHz$, and $6kHz \sim 7kHz$ survive AEC the best, with most of their frequency components being kept in the residual signals. While the chirp signals at other frequency ranges are significantly canceled by AEC, leaving very limited signal strengths in the residual echo. We thus use the above three frequency ranges to craft the in-channel attack signal to penetrate AEC.

**Duration Selection.** As AEC adopts adaptive filters for echo cancellation, the duration of the chirp signal will have a big impact on the cancellation effect. To study this duration impact, we generate the above in-channel attack signals, which last for $100ms$, $200ms$, $300ms$, and $400ms$, respectively, and play them in Zoom calls. The recorded echo signals are then analyzed. We find that all the signals can penetrate AEC with residue echoes, confirming the effectiveness of the selected frequency ranges in breaking through AEC. Though distorted, the attack signals with a $100ms$ duration maintain relatively stable echo shapes. The signals with longer durations are significantly canceled by AEC at many discrete time points, showing grid-like echo patterns, which are also not consistent. This is because AEC keeps learning from the acoustic channel and adjusts the parameters of adaptive filters based on historical observations for optimal echo cancellation. Longer signals, compared to AEC's observation windows, are canceled more significantly and hard to survive with stable patterns. We thus use $100ms$ as the duration of the in-channel attack signal.

### 4.1.2. Off-channel Media Sounds.

The basic idea of off-channel echo sensing is to exclude the probing sound from being included in AEC's reference audio. To achieve this, the audio channel of video conferencing can not be used. To still manipulate the user's device to play the probing sounds and complete the loudspeaker-to-mic loop for



(a) Home

(b) Office

Figure 6. Tag two places with Gmail alerts in Zoom calls.

sensing, we exploit the popular third-party communication systems to trigger notification tones on the user's device. In particular, many users may keep their email portals open in web browsers or have chat message software (e.g., Teams, Zoom's Team Chat, and WhatsApp) running in the background to enable quick responses when they are at work. Notification tones are then played spontaneously by the device whenever new emails or chats come, which are unpredictable and usually do not raise suspicion. Adversaries thus can exploit such media sounds for acoustic sensing. They can send messages in their own names without exposing the attack, as long as the text content is reasonable. They can also send anonymous messages, as people already get used to the flood of spam messages every day. Further, since video conferencing servers support separately recording/storing each user end's audio rather than mixing them, sending group emails/messages in meetings can amplify the attack and facilitate large-scale privacy data collection.

As the notification tones are transmitted outside of the video conferencing channel, they evade AEC and return to the attacker with more frequencies retained. The off-channel echo sensing can be modeled as

$$\hat{S}(f) = S_L(f) + \beta S(f) H_E(f) + \alpha N(f). \qquad (2)$$

The scalar $\beta$ reflects the echo signal's amplitude and phase changes caused by AEC's noise suppression function. Figure 6 shows the echo spectrograms of the Gmail notification tone when the user is on a Zoom call at home and office, respectively. The media sound echoes retain a significant portion of signal energy when returning to the attacker. Specifically, their SNRs are over 30dB higher than those of in-channel echoes. This experiment confirms that the off-channel transmitted sounds pass through AEC and are not suppressed as noises. Furthermore, the echoes' time-frequency patterns are distinctive between the two places, reflecting their different physical environments. It is worth noting that this paper is the first to demonstrate the use of common media sounds for sensing physical surroundings.

## 4.2. Residue Echo Analysis

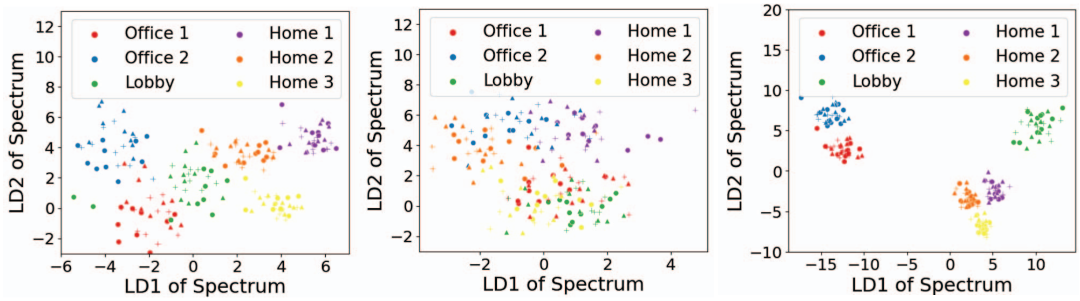
**4.2.1. In-channel Echo.** To decode the physical environment characteristics $H_E(f)$ from the severely distorted audio feedback $\hat{S}_A(f)$ for location recognition, we plan to employ learning algorithms. The algorithms need to detect and recognize recurrent echo residues and further distinguish their minute differences from their already distorted shapes. In particular, we derive two types of time-frequency images to capture the frequency point changes of the echo over time. We then develop generative AI encoders to process such 2D images and build location profiles.

**Three-stage Time-frequency Images.** We derive the short-time Fourier transform and wavelet transform to present the time-frequency images of the echo. Instead of deriving a single time-frequency image of the echo, we apply a sliding window (e.g., with a $100ms$ length and a $50ms$ overlap) to the received audio and derive the time-frequency images covering three stages: the echo and the two audio segments before and after the echo. The $1st$ *stage* contains the ambient noises' audio feedback before the echo and the front half of the echo. The resulting time-frequency image is the left-side transition feature, describing how AEC responds to the in-channel attack. The $2nd$ *stage* includes the major part of echo with the length equalling to that of the in-channel chirp, which captures the sensing information that survives AEC. The $3rd$ *stage* begins with the echo signal tail, which lasts longer than the chirp length, and the audio feedback of ambient noises after the chirp. It measures the right-side transition feature of how the AEC's state is restored after the in-channel attack. The practical ambient noises before and after the in-channel attack should be similar. But because the AEC's function parameters are modified by the attack signal and specific to physical surroundings, the two ambient noise segments' audio feedbacks are slightly different. We thus analyze all three time-frequency images to recognize the user's location.

**4.2.2. Off-channel Echo.** Exploiting common media sounds for sensing is not easy. Compared to the dedicated signals' single frequency and frequency-sweeping patterns, media sounds contain more complicated frequency components and are hard to analyze to capture the unique changes caused by the sensing targets. A recent work demonstrates that the advancement of deep learning has enabled sensing via media sounds [58]. It further captures the strong vibrations resulting from the notification sounds to enable sensing across acoustic and vibration domains. In this work, we derive time-frequency images of the received sound to describe how the original tone signal propagates and is echoed back by physical environments, which is a measurement of the echo system response $H_E(f)$. Similar to the in-channel chirps, three-stage time-frequency images are derived to capture the channel information before, during, and after the tone signal is played. In addition to the spectrogram, we also derive the scalogram, the persistence spectrum, and the Hilbert spectrum for multi-resolution analysis and phase analysis. The derived time-frequency images will be processed by deep learning algorithms to tag and infer the user's room-level locations.

## 4.3. Using Distorted Echo for Location Inference

We conduct 6-month experiments both inside and outside Zoom calls to collect data at different locations, including a lobby, two offices sharing identical layouts, and three rooms having the same floor plans. During this period, there



(a) Ambient sound

(b) Ambient sound in Zoom

(c) Active acoustic chirp

Figure 7. Validation of traditional acoustic localization methods and illustration of the acoustic sensing challenge in Zoom.

are normal object/furniture position changes, in-room device location changes (up to one meter), and human activities. Figure 7 presents the clustering results of the locations using ambient sounds [33] and active sensing with ultrasonic chirps [13]. Spectrum features were derived to capture each location's characteristics and linear discriminant analysis (LDA) was applied for dimension reduction. When outside a Zoom call (i.e., no AEC and 48kHz sampling rate), the effectiveness of traditional ambient sound [33] and active ultrasonic sensing methods [13] are confirmed by Figure 7(a) and (c), while the latter more effectively separates different locations in isolated clusters by using active sensing. However, when in a Zoom call, as shown in Figure 7(b), the ambient sound's performance is largely degraded, where the location clusters are mixed together and hard to differentiate due to AEC and low sampling rates.

This paper aims to recognize the varying residue echoes to identify recurrent locations, and takes one step further to explore the potential of counteracting the effects of AEC. We propose a novel algorithm to capture the unique behaviors of the AEC function at different locations, which exploits the rich temporal information of the received audio before and after a malicious sound is played. Since AEC estimates echo paths based on prior observations, the AEC's behaviors, from responding to the incoming probing sound based on prior observations (e.g., ambient sounds) to updating function parameters and dealing with subsequent sounds (e.g., same ambient sounds), may exhibit location-dependent information. Specifically, we explore the transformer model to examine each audio sample by attending to all other samples in a wider audio clip that encompasses the echo.

## 4.4. Transformer-based Location Inference

We develop a transformer-based model to learn stable location embedding from distorted echo signals. Specifically, the time-frequency images of each returned audio are derived at three stages and fed into the transformer model to recognize separately. Their outputs are integrated based on probabilities for final location inference.

**4.4.1. Model Structure.** We develop the transformer-based location inference model based on the encoder structure of the Vanilla Transformer [59], [60]. Figure 8 shows the architecture of our transformer-based location inference model. The time-frequency image $F \in \mathbb{R}^{\frac{T}{\tau} \times f}$ derived from the received audio $x \in \mathbb{R}^{T}$ with a window length of $\tau$ is taken as the input. Linear transformation is first applied to transform the input into a vector $V \in \mathbb{R}^{\frac{T}{\tau} \times d}$ of the model dimension $d$. Positional encoding then adds the input data samples index information to the vector. The vector with positional information is then fed into the transformer encoder, which consists of $M$ identical structures, where for each structure, a multi-head attention mechanism derives the input vector's self-attentions at each data sample. The transformer encoder learns location information from the time-frequency images and outputs the high-level representation of the location's characteristics $L \in \mathbb{R}^{\frac{T}{\tau} \times d}$, which describes both the spatial and the temporal location information carried by each time-frequency image. The high-level representation is finally passed through a linear transformation and a softmax function to generate the estimated location/context probability.

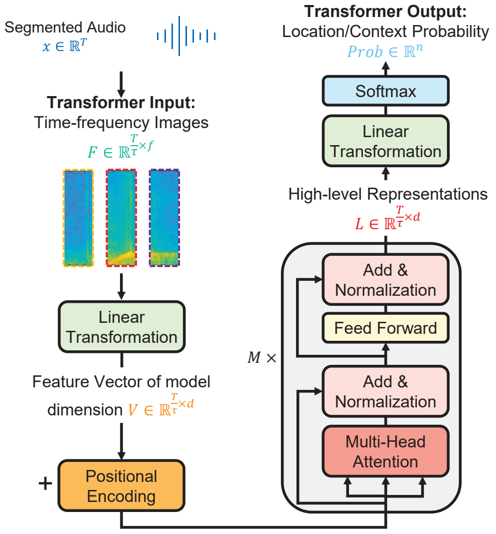
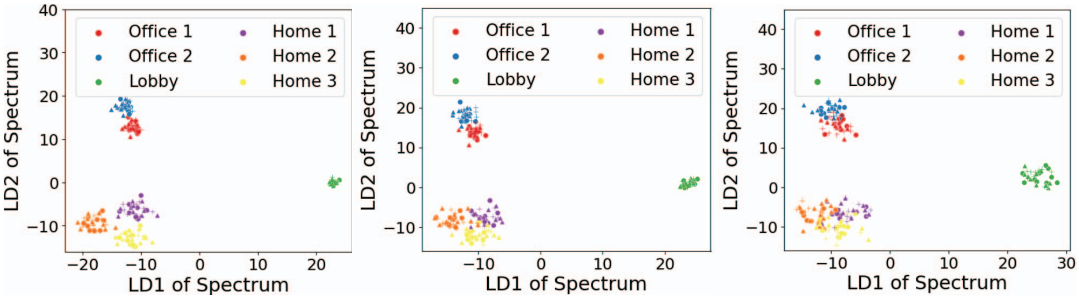


Figure 8. Architecture of the transformer-based location inference model.

Transformers were initially developed for sequence transduction tasks [59], and have been applied for processing time series [61] and images [62]. Compared to CNN and RNN, transformer outperforms by processing all data samples in the input fragment in parallel and enabling each data sample to attend to all other data samples [63], [64], which enables learning the location information with



**(a) Off-channel chirp in Zoom** **(b) Off-channel tone in Zoom** **(c) In-channel chirp in Zoom**

Figure 9. Illustration of our transformer-powered echo sensing attacks: (a) is only feasible with malware installation due to the difficulty in manipulating the user's device to play dedicated sensing signal; (b) and (c) are the proposed two echo attacks that achieve comparable location clustering results as (a).

both spatial and temporal characteristics. The transformer processes the whole time-frequency image at once, where positional encodings are added to make the model aware of the data sample's (time) index in the time-frequency image.

**Linear Transformation and Positional Encoding.** The sub-layers in the transformer model produce the output of the same dimension $d$ to facilitate the add & norm computations. With an input time-frequency image $F \in \mathbb{R}^{\frac{T}{\tau} \times f}$, we first apply linear transformation to convert it into a vector $V \in \mathbb{R}^{\frac{T}{\tau} \times d}$ with dimension $d$, where the parameters of the linear transformation are learnable. Since there is no convolutional layer or recurrent layer in the transformer model, in order to make it aware of the time-frequency image's data samples' index/order information, we add positional encodings to the input time-frequency image. Specifically, sinusoidal encodings are used, which are sine and cosine functions of different frequencies [59].

**Transformer Encoder.** The transformer encoder consists of $M$ identical layers. Each layer contains a multi-head attention sub-layer and a position-wise feed-forward sub-layer. Followed by each sub-layer, there is a residual connection and a layer normalization. The output of each sub-layer has the same dimension $d$, which facilitates the add and norm computations.

When the model is processing the multi-dimensional features (i.e., frequency features) of one data sample (i.e., one time bin), self-attention allows it to attend to all other time bins in the time-frequency image, which enables the model to learn the relationships between the frequency features at different time bins. To calculate self-attentions of the time-frequency image, we first create three vectors, the Query vector, the Key vector, and the Value vector, from the transformer encoder's input vector. The three vectors are created by multiplying the input vector by three matrices that are learnable during training. Then, a score is calculated by taking the dot product of the Query vector and the Key vector. For the frequency features at time bin $t$, its attention scores against the frequency features at all time bins in the time-frequency image are calculated by the dot product of the Query vector for time bins $t$ ($q_t$) and the Key vectors for all time bins ($k_1, k_2, ..., k_T$). The scores are then passed through a softmax function to make itself add up to 1. The softmax score determines how much of each time bin's frequency features could be expressed by that of the currently examined time bin. The value vector at each time bin is then multiplied by the softmax score, and summed up to generate the self-attention for the input time-frequency image at the current time bin.

We use multi-head attention rather than a single attention function to allow the model to focus on different time bins of the input time-frequency image and give the attention layer multiple representation subspaces, which is helpful to capture the location information. In particular, the Query vector, the Key vector, and the Value vector are linearly projected multiple times by different learnable projection matrices. The projected versions of the three vectors are used to perform the attention function in parallel, whose results are finally concatenated and projected to generate the final attention results of the time-frequency image.

**Output.** The transformer encoder outputs the high-level representations $\mathbb{R}^d$ derived for each time bin of the time-frequency image and concatenates them into a single vector $L \in \mathbb{R}^{\frac{T}{\tau} \times d}$, which are the representations of the location information. The obtained location representations are then processed by a linear transformation to generate the prediction vector $\mathbb{R}^n$, where $n$ is the number of classes. In the end, a softmax function is used to estimate the probability distribution for the locations. Depending on the attacking scenario, $n$ can be the number of locations for location identification, or the number of contexts when performing context inference. Figure 9 presents the location clustering results in Zoom calls using our transformer-powered echo sensing methods. Both the off-channel tone and in-channel chirp methods effectively separate different locations. The clustering results are comparable to the traditional active acoustic localization in non-video call scenarios as shown in Figure 7 (a). Also, places of the same context are in closer but isolated clusters, demonstrating the potential of place context inference. Moreover, the off-channel tone and chirp both perform better than the in-channel chirp, confirming that off-channel echoes are slightly less affected by AEC.

**4.4.2. Recurrent Location Identification.** The adversary who has obtained the data from the user's daily locations can train the transformer to perform recurrent location identifica-

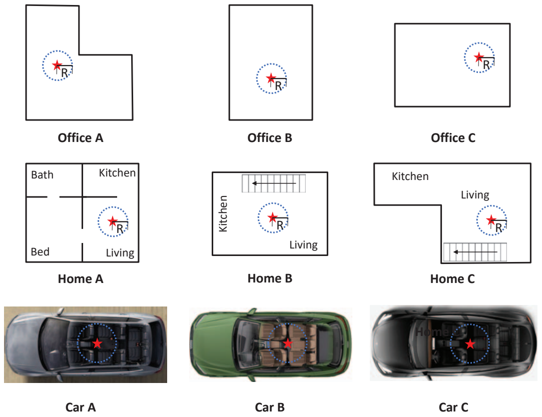


Figure 10. Layouts of partial experimental locations and the illustration of random device placements within a circle of a radius $R$.

tion, which identifies the user's exact location. In particular, the number of classes $n$ in the transformer model is set to be equal to the number of user's daily locations. The returned audio is then classified by the transformer as one of those known locations. If the probability of any known location is lower than a threshold, the echo is considered to be from a new or untagged place. The algorithm can be modified to infer whether or not the echo is from a specific location under consideration with $n = 2$.

**4.4.3. Unseen Location Context Inference.** If the testing echo is determined to be from a new and untagged location by the above algorithm, we propose to infer the place context and add the unseen location to the user's location profile with a partial place context tag. The location tag can be completed later when the attacker obtains the more precise location information, by directly asking the user or capturing data from the user's webcam. The number of classes $n$ is then set to be equal to the number of contexts under consideration. A captured echo is then classified by the transformer into one of these place contexts.

# 5. Performance Evaluation

## 5.1. Experimental Setup

**Devices.** We use a MacBook 13 with MacOS installed to act as the attacker device, where SoundFlower [65] is further installed to create a virtual audio interface for internal recording and playing sounds. By doing so, we can get rid of the background noise on the attacker side, which may otherwise trigger AEC and lead to a more distorted echo. For the victim device, we evaluate four different Windows devices (i.e., a Dell Inspiron laptop, a Dell XPS laptop, a Lenovo Legion laptop, and a Dell Precision desktop) and one MacOS PC (i.e., Macbook 15), while the Dell Inspiron laptop is used for presenting the major results.

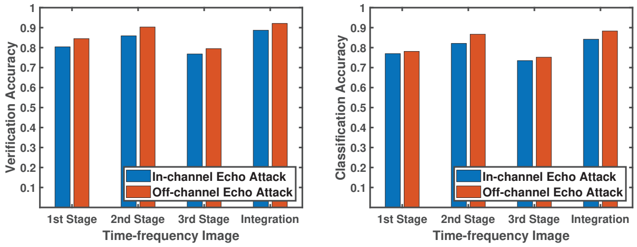
**Video Conferencing Apps.** We test our attacks mainly on Zoom, the most popular video conferencing app nowadays. The noise suppression level of the Zoom on the victim's device is set to the default, while the attacker turns off the noise suppression of the Zoom installed on his device to record a better echo. The output channel of the Zoom app on the victim's device is set to the laptop's loudspeaker, and the input channel is set to the laptop's built-in microphone. The output channel of the Zoom app on the attacker's device is set to one virtual audio channel for internal recording, and the input channel is set to another virtual audio channel for internal playing. The in-channel probing signal is played via the virtual audio channel for internal playing. The off-channel tones are triggered on the victim's device by sending Gmails. We also attack the Skype and Teams platforms.

**Data Collection.** We evaluate the attacks with the victims at 12 different locations, considering five typical place contexts: office, home, vehicle, lobby, and outdoor. Recognizing the victim's specific indoor locations is the attacker's major focus, while all outdoor locations are treated as one type of location (context) without further distinction. We illustrate the layouts of partial typical locations and the device placements in Figure 10. The data is collected in two sessions separated by at least one month to evaluate the long-term performance of the location inference over video conferencing audios, considering the varying states of echo cancellation functions in practical environments, inconsistent device placements and location variations, and normal physical setting changes (e.g., furniture movements). In each data collection session, the victim's device is randomly placed at or around the spot marked by the "star" with up to 1 meter displacement variations (the dotted circles), as shown in Figure 10. For each device-location pair, we repeat both types of echo attacks for 40 times. The victim's devices are also slightly moved to imitate the practical device use. For the Zoom platform, we test all four devices as the victim device. For the Skype and the Teams platforms, we use the Dell Inspiron laptop and repeat the above data collection. A total of 3200 audio clips are collected for each session. We then use the first session data for training and the second session data for testing. For ethical considerations, no human voice was recorded. The experiments were conducted either with no people in proximity or with the people being informed and consented to not speak during our recording. IRB approval was obtained.

## 5.2. Location Privacy Inference Performance

**5.2.1. Recurrent Location Identification.** We first present the recurrent location identification performance. The Dell Inspiron laptop is used as the victim device and both attacks are launched in Zoom meetings.

**Location Verification.** Figure 11(a) presents the performances of using a single probing sound of each echo attack to verify whether the victim is at a specific location under consideration. The in-channel echo attack achieves a location verification accuracy of 88.7% with a TPR of 90.5%, an FPR of 12.8%, and an F1-Score of 89.1%, while the off-channel echo attack achieves slightly higher accuracy, which is 92.1%. The TPR, FPR, and F1-Score achieved by the off-channel echo attack are 95.3%, 10.4%, and 92.6%, respectively. We further study the significances of each echo-sensing attack's three stages. For both attacks, the 2*nd stage* achieves the highest performance, which is 85.9% and 90.3% for the in-channel and off-channel echoes,



(a) Location Verification

(b) Location Classification

Figure 11. Recurrent location identification performance.

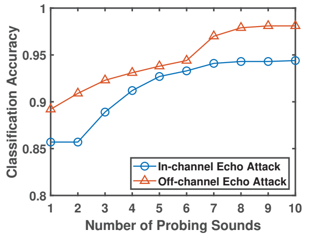


Figure 12. Location inference using multiple probing sounds.

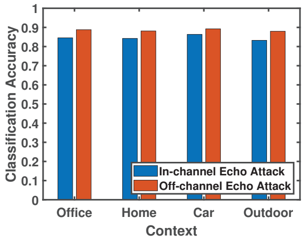


Figure 13. Context inference performance on unseen locations.

respectively. The result reflects our algorithm's effectiveness of recognizing the distorted echo sound, as the $2nd$ *stage* contains the major body of the returned echo sound after cancellation. The $1st$ and $3rd$ *stages* of each echo attack also achieve over $80\%$ location verification accuracy, though they are less accurate than the $2nd$ *stage*. This result confirms the capability of our algorithm to capture how AEC responds to the attacks to reveal additional location information. Our algorithm is designed to integrate all three stages with strong self-attention for achieving high and robust location inference performance over the long term.

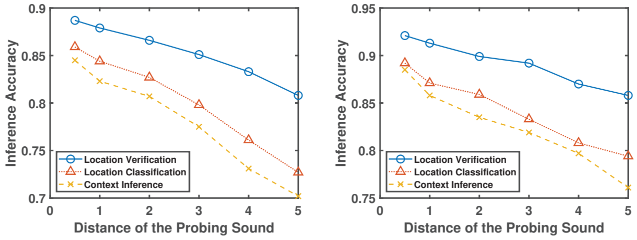
**Location Classification.** Figure 11(b) presents the performance of recognizing the victim's location among a number of tagged locations. This paper uses 12 locations to show the location-distinguishing capability of the proposed methods, but in practical scenarios, the victim's daily places are typically much fewer. When integrating all three stages' information, the in-channel echo attack achieves a location classification accuracy of $84.2\%$ with an F1-Score of 84.8%, while the off-channel echo attack achieves an accuracy of $88.3\%$ with an F1-Score of 89.1%. Detailed performance for each location is presented in Appendix Figure 19 and 20, showing the proposed attacks can even distinguish outdoor places. The result confirms the capability of the proposed methods to tag and probe a number of different locations of the target user in practical scenarios. Moreover, it pushes the limit of the acoustic-based localization to make it still work remotely and under echo cancellation and demonstrates the potential of acoustically distinguishing a large number of places. When looking at the contributions of each stage to location classification, the observations are similar to the above location verification: All three stages of the echo attacks are found to carry sensitive location information, and the $2nd$ *stage* performs the best for both echo attacks.

**Using Multiple Probing Sounds.** The attacker can send more than one probing sound in a meeting session to improve the location inference performance. To fully understand the attacker's capability, we study the impact of the number of probing sounds on location classification, which is presented in Figure 12. We observe that both attacks' location classification performances are significantly improved when multiple probing sounds are used. In particular, when the in-channel chirp is sent $4$ times, the location classification accuracy is over $90\%$. The off-channel tones perform much better with multiple probing attempts. Sending only two emails improves the location classification accuracy to be over $90\%$. Sending $8$ emails increases the location classification accuracy to $98\%$. However, sending too many emails in one meeting session may disturb the meeting and arouse notice. Thus, the off-channel attack has a limited probing number. In comparison, the short in-channel chirps can be transmitted many times whenever the user's mic is unmuted. For example, probing with $7$ in-channel chirps improves the location classification accuracy to $94\%$.

**5.2.2. Unseen Location Context Inference.** For the new and untagged locations, which achieve a low confidence score for all the location tags, we infer the place context instead. For evaluation, we use $7$ locations' data for training and use the other $4$ (new) locations' data for testing. The place contexts include office, home, car and outdoor. As shown in Figure 13, the in-channel attack achieves $84.5\%$, $84.2\%$, $86.3\%$, and $83.2\%$ context inference accuracy with an F1-Score of 85.1%, 84.4%, 86.8%, and 83.5%, for the office, home, car, and outdoor, respectively. The off-channel attack achieves $88.8\%$, $88.1\%$, $89.2\%$, and $87.9\%$ accuracy with an F1-Score of 89.2%, 88.4%, 89.6%, and 88.3%, in inferring these four contexts. The results indicate that even if a location is unseen or untagged before, its context can be recognized through echo sensing with a high accuracy. The location context inference performance can be further improved if multiple probing sounds are transmitted.

**5.2.3. Device Location Variations & Distance Tolerance.** The above results show that our methods can acoustically recognize a user's location even when the user's device is randomly placed in a one-meter circle. This is to imitate the typical scenario when users usually attend video conferences at a relatively fixed spot (e.g., desk or sofa) but with inconsistent device placements. We take one step further to explore the extent of the attack's distance tolerance (i.e., its effective range) by considering the more challenging scenarios when the user completely changes the spot in a room, such as moving from the desk to the bed. Specifically, we collect data at varying distances from the device's original position, ranging from 0.5 meter to 5 meters, only for testing. The training data is still unchanged with device placements within a $0.5m$ radius.

Figure 14 presents the two echo attacks' distance tolerance regarding location verification, location classification, and context inference. Although a longer distance leads to a decreased accuracy as expected, both attacks exhibit strong distance tolerance. In particular, the off-channel attack maintains over 90% location verification accuracy with distances of up to 3 meters, and under the same distance changes, the in-channel attack achieves over 86% accuracy in location



(a) In-channel Echo Attack

(b) Off-channel Echo Attack

Figure 14. Distance tolerance (effective range) of both attacks.

verification. When the distance change is increased to 5 meters, the off-channel and in-channel attacks still achieve 85.8% and 80.8% accuracy, respectively, in location verification. Figure 14 also shows that, for both attacks, location verification accuracy is on average 6% higher than location classification accuracy, and the latter is around 2% higher than context inference accuracy. With a distance change of 5 meters, the location classification accuracy of off-channel and in-channel attacks drops to 79.4% and 72.7%, while their context inference accuracy decreases to 76.1% and 70.2%. The location privacy leakage is still non-negligible.

The strong distance tolerance in location recognition is owing to the high correlations of acoustic echoes exhibited at close spots within the same indoor location, which result from the similar echo paths caused by the same physical environment [12], [13], [14]. In this work, we demonstrate that this phenomenon is still maintained under echo cancellation. Because of the echoes of different spots are correlated, we further find that one-shot or few-shot learning with the echoes of large distance changes can significantly improve the location inference performance. Taking the location classification for instance, by adding a few 5-meter echoes to update our training model, the minimum location classification accuracy increases from 72.7% to 83.5% when the device is randomly placed from 0.5 meter to 5 meters. The more detailed distance tolerance study regarding the device displacement distances in the testing data, the training data and both is shown in Apendix Figure 21.

The distance tolerance study results motivate practical attackers to rely on a two-threshold strategy ($th1 \leq th2$) to send multiple probing sounds and update location profiles. Specifically, if the maximum confidence score of a probing sound is lower than $th2$ but greater than $th1$, the attacker can still believe the user is at one tagged location but may completely change the spot. The attacker can send a $2nd$ or more probing sounds to confirm and use the returned echoes to update that tagged location. If the maximum confidence score of a probing sound is less than $th1$, the attacker would treat it as a new location. The attacker can choose to send more probing sounds to collect information about the location for profiling and then patiently wait for the chance to get the location name to complete the location tag.

**5.2.4. Victim Device Types.** Different devices could have different configurations of speakers and microphones as well as different built-in echo cancellation mechanisms, which could impact the echo-based sensing. We thus examine four laptop devices (i.e., Dell Inspiron, Dell XPS, Lenovo Legion, and MacBook 15) and one desktop (i.e., Dell Precision) as the victim devices and compare the classification accuracy of localizing these devices through Zoom in Figure 15. We can observe that both attacking methods on the five victim devices achieve similar results. In particular, the in-channel echo attack achieves 84.2%, 83.7%, 84.5%, 83.2%, and 84.1% classification accuracy when localizing the Dell Inspiron, Dell XPS, Lenovo Legion and MacBook 15 laptops and the Dell Precision desktop, respectively. The off-channel echo attack performs better on all these devices, achieving 88.3%, 86.8%, 89.1%, 85.9%, and 87.1% location classification accuracy, respectively. The results confirm the capability of both attacks to exploit the user's device for location probing, regardless of the device models.

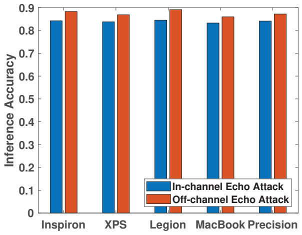


Figure 15. Localizing different victim devices in Zoom calls.

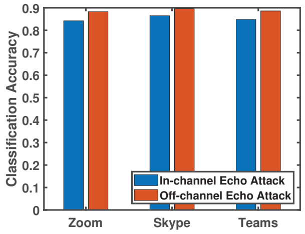


Figure 16. Location classification on different video conferencing apps.

**5.2.5. Different Video Conferencing Apps.** The AECs used by different video conferencing apps may vary, as they are all closed-source. Thus, besides Zoom, we also evaluate the attacks on Skype and Teams, where the Dell Inspiron laptop acts as the victim device. The location classification performances through these three video conferencing apps are compared in Figure 16. We observe that Skype leaks location privacy the most for both the in-channel echo attack and the off-channel echo attack, which achieve a location identification accuracy of 86.5% and 89.6%. The attack performances on Zoom and Teams are similar, which are only slightly lower than Skype. In particular, the location classification accuracy is 84.8% on Teams for the in-channel echo attack and 88.6% for the off-channel echo attack. No significant performance differences are observed among the three video conferencing apps, all of which are vulnerable to our attacks. These results confirm that current AEC implementations used in video conferencing considers only mitigating audio feedback to enhance user experience, while the security considerations are largely overlooked.

**5.2.6. Different Time-frequency Analysis.** We now study using different time-frequency analyses to interpret the in-channel and off-channel echoes. In particular, we compute four types of time-frequency images from each echo, including the short-time Fourier transform (STFT), which has been used by default to present the above results, the scalogram based on the wavelet transform, the persistence spectrum, and the Hilbert spectrum using the Hilbert transform. These 2D images are then fed into our transformer model for location classification. Figure 17 presents the location classification performance of the four time-frequency images for both echo attacks, where the Dell Inspiron laptop and

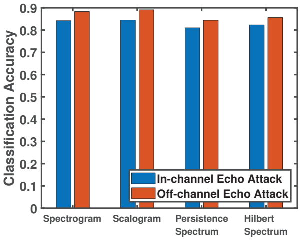


Figure 17. Location classification performance using different time-frequency analysis.

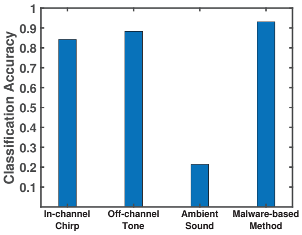


Figure 18. Comparison with ambient sound method (in Zoom) and malware method.

Zoom are used. We find that the scalogram performs the best with $84.5\%$ location classification accuracy for the in-channel echo and $89.1\%$ accuracy for the off-channel echo, which are slightly higher than the spectrogram's results, which are reported above. The reason is that compared to the short-time Fourier transform, the wavelet transform's time-frequency resolution is not fixed but can adjust to the signal's different frequencies, where lower frequency components are represented with finer frequency resolution and coarser time resolution and higher frequency components are represented with coarser frequency resolution and finer time resolution. The persistence spectrum and the Hilbert-Huang transform also achieve good location classification accuracies, though their performances are lower than spectrogram and scalogram. The attacker can further integrates the different types of time-frequency analyses to improve the location inference of each single probing sound.

### 5.3. Compared to Ambient Sounds and Malware

We implement an ambient sound approach and a malware-enabled acoustic sensing approach using our transformer model. Specifically, the ambient sound approach relies on passive acoustic sensing, which needs to exploit video conferencing and is subject to echo cancellation. In contrast, the malware manipulates the user's device to actively play ultrasonic chirps for location probing, which requires no video conferencing and is not affected by echo cancellation. The malware approach represents the state-of-art in infrastructure-free, acoustic-based localization [14].

Figure 18 compares the location classification performances of the above approaches and our two echo sensing attacks, using the Dell Inspiron laptop, Zoom and eight different locations. Each location inference is based on a single probing sound. The ambient sound method achieves an accuracy of $21.4\%$ in classifying the eight locations, which is close to random guessing. Extending the length of the ambient sound segments does not improve the performance. This is because echo cancellation and noise suppression significantly destroy the location information carried by the ambient sounds before they return to the attacker. Differently, the malware approach achieves an accuracy of $93.1\%$ in classifying the eight locations. Because echo cancellation results in the loss of location information in the returned echo sounds, the malware approach sets an upper limit on the performance of our echo sensing methods. Under the same condition, our in-channel and off-channel echo attacks achieve location classification accuracies of $84.2\%$ and $88.3\%$. Though not reaching the limit, our two echo attacks achieve comparable performances to the malware approach without the requirement of malware installation.

## 6. Defense Methods & Discussion

Based on the understanding of the two echo-sensing attacks that exploit video conferencing, we propose the defense approaches. For the in-channel echo attack, we propose to detect suspicious signals that are sent over the audio channels of video conferencing before they are played by users' devices. Since in-channel attack sounds must survive AEC with sufficient echo components, they exhibit patterns as carefully crafted signals unlike human speech or media sounds, which makes the detection of such signals possible. Specifically, the incoming far-end audio will be processed by an observation window and further analyzed by a machine-learning model to determine whether the current audio clip contains suspicious signals. If suspicious signals are detected, the video conferencing app can disable the user's speaker or mic for a short period corresponding to this audio clip to completely prevent acoustic sensing. We implement a transformer-based binary classification model to detect suspicious signals and achieve 98% accuracy in distinguishing normal video conferencing audios (e.g., ambient noises and sounds) from those containing in-channel sensing signals. The detection model can be further enhanced by incorporating a broader range of allowed sounds (e.g., media sounds and human voices) and prevented sounds (e.g., known malicious sensing signals) for training. The client-server-client communications of video conferencing facilitate further locating the malicious sound sources (i.e., meeting participants) for digital forensics.

Since the off-channel echo attack exploits notification sounds that are not transmitted over the audio channel of video conferencing to evade AEC, we propose three methods to prevent the evasion. The first defense method is to disable all applications' notification sounds during online meetings. However, this method requires root permissions to control all other applications and may block urgent notifications that need immediate attention. The second method is to include all device notification sounds into the AEC's reference audio for cancellation. Unlike dedicated sensing signals, the common media sounds under echo cancellation can be hardly recognized for sensing. This defense system can be deployed either at the video conference server or locally on the user's device. The third method aims to fully break the active acoustic sensing loop by including the microphone buffer into the AEC's reference audio derivation, rather than solely relying on the audio channels of video conferencing. This method can prevent all media sounds from leaking sensitive information, not just notification tones.

This paper presents a comprehensive study of acoustic-based localization in typical video conferencing scenarios using laptops and desktops, which provides a foundation for further echo-sensing research. For example, we find that mobile devices and smartphones are also vulnerable to malicious echo sensing in Zoom and Skype calls. However, determining whether reliable location information can be consistently derived from these devices' echoes requires further investigation. This is because echo sensing on mobile devices may be significantly influenced by user hand gripping and device interactions [66], [67]. Furthermore, the way users hold and interact with mobile devices may expose additional sensitive information to malicious echo sensing beyond location privacy, such as sensitive on-screen activities and hand biometrics.

## 7. Conclusion

This work investigates the potential of sniffing the location privacy of video conference users through legitimate audio channels. We identify two new attacks that could survive the video conferencing systems' echo cancellation. Specifically, the in-channel echo attack sends carefully crafted chirp signals through the audio channel to penetrate AEC. The off-channel echo attack triggers the user's device to play notification sounds, which avoids being included in the reference audio of AEC. We utilize both types of sounds to probe the user's surroundings and further develop a transformer algorithm to tag and recognize the user's locations and place contexts from the returned echo sounds. Experiments with commercial video conferencing apps show that the proposed attacks can recognize recurrent locations with $88.3\%$ accuracy and infer the context of unseen locations with $88.5\%$ accuracy, using only one probing sound. The results raise a severe privacy concern since any video conferencing participant could invade each other's location privacy easily without malware installation. We further propose malicious signal detection and end-to-end reference audio derivation to defend against these new attacks.

## 8. Acknowledgment

This work is partially supported by NSF CNS-2450046 and CNS-2440238.

## References

[1] L. Andre, "67 zoom statistics you must know: 2022 market share & data analysis," 2022, https://financesonline.com/zoom-statistics/.

[2] M. R. Report, "Video conferencing market by component (hardware, solutions, and services), application (corporate communication, training and development, and marketing and client engagement), deployment mode, vertical, and region - global forecast to 2026," 2021, https://www.marketsandmarkets.com/Market-Reports/video-conferencing-market-99384414.html.

[3] R. Koodli, "Ip address location privacy and mobile ipv6: Problem statement," Nokia Siemens Networks, Tech. Rep., 2007.

[4] H. Jiang, Y. Liu, and J. N. Matthews, "Ip geolocation estimation using neural networks with stable landmarks," in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2016, pp. 170–175.

[5] S. Zu, X. Luo, S. Liu, Y. Liu, and F. Liu, "City-level ip geolocation algorithm based on pop network topology," *IEEE Access*, vol. 6, pp. 64 867–64 875, 2018.

[6] T. Schnitzler, K. Kohls, E. Bitsikas, and C. Pöpper, "Hope of delivery: Extracting user locations from mobile instant messengers," *arXiv preprint arXiv:2210.10523*, 2022.

[7] K. Kohls and C. Diaz, "{VerLoc}: Verifiable localization in decentralized systems," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 2637–2654.

[8] M. Candela, E. Gregori, V. Luconi, and A. Vecchio, "Using ripe atlas for geolocating ip infrastructure," *IEEE Access*, vol. 7, pp. 48 816–48 829, 2019.

[9] R. Chen, Z. Li, F. Ye, G. Guo, S. Xu, L. Qian, Z. Liu, and L. Huang, "Precise indoor positioning based on acoustic ranging in smartphone," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[10] R. Xi, D. Liu, M. Hou, Y. Li, and J. Li, "Using acoustic signal and image to achieve accurate indoor localization," *Sensors*, vol. 18, no. 8, p. 2566, 2018.

[11] T. Akiyama, M. Sugimoto, and H. Hashizume, "Time-of-arrival-based indoor smartphone localization using light-synchronized acoustic waves," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 100, no. 9, pp. 2001–2012, 2017.

[12] H. Murakami, M. Nakamura, S. Yamasaki, H. Hashizume, and M. Sugimoto, "Smartphone localization using active-passive acoustic sensing," in *2018 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2018, pp. 206–212.

[13] C. Chen, Y. Ren, H. Liu, Y. Chen, and H. Li, "Acoustic-sensing-based location semantics identification using smartphones," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 20 640–20 650, 2022.

[14] Y.-C. Tung and K. G. Shin, "Echotag: Accurate infrastructure-free indoor location tagging with smartphones," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 525–536.

[15] S. Panda, Y. Liu, G. P. Hancke, and U. M. Qureshi, "Behavioral acoustic emanations: Attack and verification of pin entry using keypress sounds," *Sensors*, vol. 20, no. 11, p. 3015, 2020.

[16] J. Yu, L. Lu, Y. Chen, Y. Zhu, and L. Kong, "An indirect eavesdropping attack of keystrokes on touch screen through acoustic sensing," *IEEE Transactions on Mobile Computing*, vol. 20, no. 2, pp. 337–351, 2019.

[17] P. Cheng, I. E. Bagci, U. Roedig, and J. Yan, "Sonarsnoop: Active acoustic side-channel attacks," *International Journal of Information Security*, vol. 19, pp. 213–228, 2020.

[18] L. Zhang, K. Liu, Y. Jiang, X.-Y. Li, Y. Liu, P. Yang, and Z. Li, "Montage: Combine frames with movement continuity for realtime multi-user tracking," *IEEE Transactions on Mobile Computing*, vol. 16, no. 4, pp. 1019–1031, 2016.

[19] A. Parate, M.-C. Chiu, C. Chadowitz, D. Ganesan, and E. Kalogerakis, "Risq: Recognizing smoking gestures with inertial sensors on a wristband," in *Proceedings of the 12th annual international conference on Mobile systems, applications, and services*, 2014, pp. 149–161.

[20] P. Zhou, M. Li, and G. Shen, "Use it free: Instantly knowing your phone attitude," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, 2014, pp. 605–616.

[21] S. Yun, Y.-C. Chen, and L. Qiu, "Turning a mobile device into a mouse in the air," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, 2015, pp. 15–29.

[22] F. Gu, K. Khoshelham, J. Shang, F. Yu, and Z. Wei, "Robust and accurate smartphone-based step counting for indoor localization," *IEEE Sensors Journal*, vol. 17, no. 11, pp. 3453–3460, 2017.

[23] F. Gu, K. Khoshelham, C. Yu, and J. Shang, "Accurate step length estimation for pedestrian dead reckoning localization using stacked autoencoders," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 8, pp. 2705–2713, 2018.

[24] J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev, "Soli: Ubiquitous gesture sensing with millimeter wave radar," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–19, 2016.

[25] M. Abbas, M. Elhamshary, H. Rizk, M. Torki, and M. Youssef, "Wideep: Wifi-based accurate and robust indoor localization system using deep learning," in *2019 IEEE International Conference on Pervasive Computing and Communications (PerCom*. IEEE, 2019, pp. 1–10.

[26] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.

[27] J. Sallai, G. Balogh, M. Maroti, A. Ledeczi, and B. Kusy, "Acoustic ranging in resource-constrained sensor networks." in *International Conference on Wireless Networks*, 2004, p. 467.

[28] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan, "The cricket location-support system," in *Proceedings of the 6th annual international conference on Mobile computing and networking*, 2000, pp. 32–43.

[29] Y. Nishimura, N. Imai, and K. Yoshihara, "A proposal on direction estimation between devices using acoustic waves," in *Mobile and Ubiquitous Systems: Computing, Networking, and Services: 8th International ICST Conference, MobiQuitous 2011, Copenhagen, Denmark, December 6-9, 2011, Revised Selected Papers 8*. Springer, 2012, pp. 25–36.

[30] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proceedings of the 15th annual international conference on mobile systems, applications, and services*, 2017, pp. 15–28.

[31] H. Sundar, T. V. Sreenivas, and C. S. Seelamantula, "Tdoa-based multiple acoustic source localization without association ambiguity," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 1976–1990, 2018.

[32] S. P. Tarzia, P. A. Dinda, R. P. Dick, and G. Memik, "Indoor localization without infrastructure using the acoustic background spectrum," in *Proceedings of the 9th international conference on Mobile systems, applications, and services*, 2011, pp. 155–168.

[33] R. Leonardo, M. Barandas, and H. Gamboa, "A framework for infrastructure-free indoor localization based on pervasive sound analysis," *IEEE Sensors Journal*, vol. 18, no. 10, pp. 4136–4144, 2018.

[34] B. Shrestha, M. Shirvanian, P. Shrestha, and N. Saxena, "The sounds of the phones: Dangers of zero-effort second factor login based on ambient audio," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 908–919.

[35] J. Wendeberg, T. Janson, and C. Schindelhauer, "Self-localization based on ambient signals," *Theoretical Computer Science*, vol. 453, pp. 98–109, 2012.

[36] S. Zhayida, F. Andersson, Y. Kuang, and K. Åström, "An automatic system for microphone self-localization using ambient sound," in *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014, pp. 954–958.

[37] H. Satoh, M. Suzuki, Y. Tahiro, and H. Morikawa, "Ambient sound-based proximity detection with smartphones," in *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*, 2013, pp. 1–2.

[38] M. Rossi, J. Seiter, O. Amft, S. Buchmeier, and G. Tröster, "Roomsense: an indoor positioning system for smartphones using active sound probing," in *Proceedings of the 4th Augmented Human International Conference*, 2013, pp. 89–95.

[39] R. Cutler, A. Saabas, T. Parnamaa, M. Purin, H. Gamper, S. Braun, K. Sørensen, and R. Aichner, "Icassp 2022 acoustic echo cancellation challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 9107–9111.

[40] M. M. Halimeh and W. Kellermann, "Efficient multichannel nonlinear acoustic echo cancellation based on a cooperative strategy," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 461–465.

[41] L. Ma, H. Huang, P. Zhao, and T. Su, "Acoustic echo cancellation by combining adaptive digital filter and recurrent neural network," *arXiv preprint arXiv:2005.09237*, 2020.

[42] H. Zhang, K. Tan, and D. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions." in *Interspeech*, 2019, pp. 4255–4259.

[43] S. Dixit and D. Nagaria, "Lms adaptive filters for noise cancellation: A review," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 7, no. 5, pp. 2520–2529, 2017.

[44] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 7, pp. 2065–2079, 2012.

[45] Z. Support. (2021) Managing audio echo in a meeting. [Online]. Available: https://support.zoom.us/hc/en-us/articles/202050538-Managing-audio-echo-in-a-meeting

[46] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, 2016, pp. 82–94.

[47] L. Lu, J. Liu, J. Yu, Y. Chen, Y. Zhu, X. Xu, and M. Li, "Vpad: Virtual writing tablet for laptops leveraging acoustic signals," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, 2018, pp. 244–251.

[48] L. Lu, J. Liu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, "Enable traditional laptops with virtual writing capability leveraging acoustic signals," *The Computer Journal*, vol. 64, no. 12, pp. 1814–1831, 2021.

[49] Y. Zhang, J. Wang, W. Wang, Z. Wang, and Y. Liu, "Vernier: Accurate and fast acoustic motion tracking using mobile devices," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1709–1717.

[50] X. Xu, J. Yu, Y. Chen, Y. Zhu, L. Kong, and M. Li, "Breathlistener: Fine-grained breathing monitoring in driving environments utilizing acoustic signals," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, 2019, pp. 54–66.

[51] R. Nandakumar, S. Gollakota, and N. Watson, "Contactless sleep apnea detection on smartphones," in *Proceedings of the 13th annual international conference on mobile systems, applications, and services*, 2015, pp. 45–57.

[52] K. Qian, C. Wu, F. Xiao, Y. Zheng, Y. Zhang, Z. Yang, and Y. Liu, "Acousticcardiogram: Monitoring heartbeats using acoustic signals on smart devices," in *IEEE INFOCOM 2018-IEEE conference on computer communications*. IEEE, 2018, pp. 1574–1582.

[53] J. Tan, X. Wang, C.-T. Nguyen, and Y. Shi, "Silentkey: A new authentication framework through ultrasonic-based lip reading," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 1, pp. 1–18, 2018.

[54] L. Lu, J. Yu, Y. Chen, H. Liu, Y. Zhu, Y. Liu, and M. Li, "Lippass: Lip reading-based user authentication on smartphones leveraging acoustic signals," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1466–1474.

[55] L. Zhang, S. Tan, and J. Yang, "Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 57–71.

[56] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, 2018, pp. 321–336.

[57] C. Wang, C. Wang, Y. Chen, L. Xie, and S. Lu, “Smartphone privacy leakage of social relationships and demographics from surrounding access points,” in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 678–688.

[58] L. Huang and C. Wang, “Notification privacy protection via unobtrusive gripping hand verification using media sounds,” in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 491–504.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[60] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, “A transformer-based framework for multivariate time series representation learning,” in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 2114–2124.

[61] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[62] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[63] P. Delgado-Santos, R. Tolosana, R. Guest, F. Deravi, and R. Vera-Rodriguez, “Exploring transformers for behavioural biometrics: A case study in gait recognition,” *arXiv preprint arXiv:2206.01441*, 2022.

[64] G. Stragapede, P. Delgado-Santos, R. Tolosana, R. Vera-Rodriguez, R. Guest, and A. Morales, “Mobile keystroke biometrics using transformers,” in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–6.

[65] M. Ingalls. (2021) Soundflower. [Online]. Available: https://github.com/mattingalls/Soundflower

[66] L. Huang and C. Wang, “Pcr-auth: Solving authentication puzzle challenge with encoded palm contact response,” in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1034–1048.

[67] R. Wang, L. Huang, and C. Wang, “Preventing handheld phone distraction for drivers by sensing the gripping hand,” in *2021 IEEE 18th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. IEEE, 2021, pp. 410–418.

# Appendix A. Location Classification

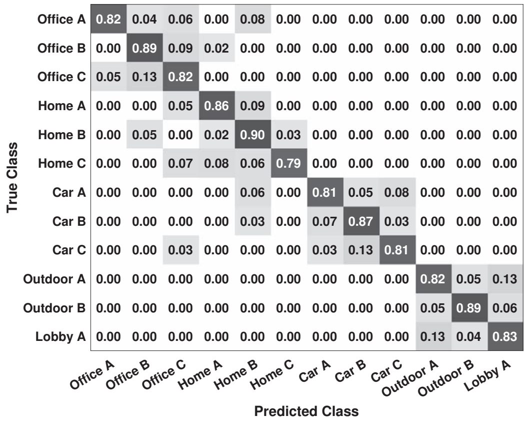


Figure 19. Location classification performance of in-channel attack.

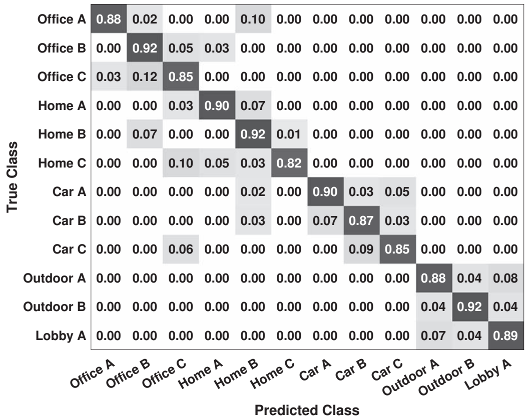
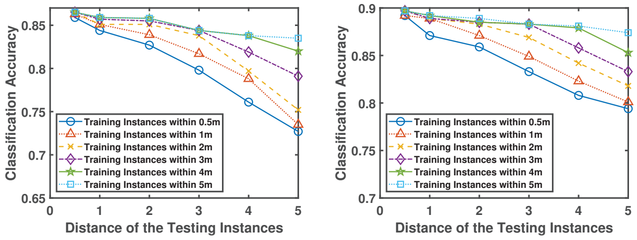


Figure 20. Location classification performance of off-channel attack.

Figure 19 and 20 present the confusion matrices of the location classification performed by the in-channel attack and off-channel attack using single probing sound, which can be further improved with multiple probing sounds. We can observe that even for outdoor spaces, the proposed attacks still achieve an average classification accuracy of 85.5% and 90%, respectively. The classification errors mainly come from the locations with the same context. This also explains why the proposed system achieves better context inference performance, as presented in Section 5.2.2.

# Appendix B. Distance Tolerance and Location Variations

Figure 21 presents the more detailed study of our two attacks’ distance tolerance when the victim device’s displacement variations of different degrees (i.e., distance to the original spot) are included in the testing data, the training data, and both, respectively. The results demonstrate that both the off-channel and the in-channel attacks can effectively recognize the user’s location with up to five-meter displacements, even when they are not in the attacker's location profile. The results also show that the acoustic echoes exhibit high correlations at close spots within the same indoor location, due to the similar echo paths caused by the same physical environment [12], [13], [14]. Furthermore, when the adversary gets a chance to include a few echoes with some displacement changes by sending more probing sounds, the location recognition accuracy can be significantly improved.



(a) In-channel Echo Attack

(b) Off-channel Echo Attack

Figure 21. Location classification performance impacted by the distance of the training and testing instances.

# Appendix C. Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

## C.1. Summary

This paper presents a new attack on location privacy in video conferencing calls, where an adversary uses audio signals to infer a participant's location even if their video is off. By sending low-frequency audio chirps and analyzing the response with a transformer-based model, the attack achieves robust location inference across various environments, conferencing platforms, and device types.

## C.2. Scientific Contributions

- Identifies an Impactful Vulnerability.
- Provides a Valuable Step Forward in an Established Field.

## C.3. Reasons for Acceptance

1) This paper proposes a new and interesting audio chirp-based attack to compromise the location privacy of videoconferencing users. Such an attack which bypasses the Audio Echo Cancellation (AEC) feature employed by most popular videoconference applications has not been done before.
2) The proposed inference framework, which employs a trained transformer-based model that takes as input the time-frequency representation of the audio-response received by the adversary and outputs a probability vector indicating the inferred location/context, is well-designed.
3) A comprehensive evaluation of the proposed attack, under a variety of experimental settings and parameters, has been carried out to demonstrate the robustness of location and context inference performance.

## C.4. Noteworthy Concerns

1) Although some preliminary evaluation of mitigation measures have been presented, they have not been systematically analyzed and evaluated.
2) The approach has not been evaluated in mobile settings (on mobile devices), where location privacy is a more significant concern, and it is not clear if the proposed attack approach will work in such a setting.
3) Survey responses from human subject participants were used to establish the feasibility of off-channel signals for the proposed attack, however, several details related to this data collection experiment involving human subject participants are not present in the paper.