TaigiSpeech: A Low-Resource Real-World Speech Intent Dataset and Preliminary Results with Scalable Data Mining In-the-Wild

Kai-Wei Chang[†], [ID] [1], *Yi-Cheng Lin*[*,2], *Huang-Cheng Chou*[*], [ID] [5], *Wenze Ren*[*,2]
Yu-Han Huang[2], *Yun-Shao Tsai*[2], *Chien-Cheng Chen*[2], *Yu Tsao*[3], *Yuan-Fu Liao*[4]
Shrikanth Narayanan [ID] [5], *James Glass*[1], *Hung-yi Lee*[2]

[1] Massachusetts Institute of Technology, USA
[2] National Taiwan University, Taiwan
[3] Academia Sinica, Taiwan
[4] National Yang Ming Chiao Tung University, Taiwan
[5] University of Southern California, USA

kwchang@mit.edu

Abstract

Speech technologies have advanced rapidly and serve diverse populations worldwide. However, many languages remain underrepresented due to limited resources. In this paper, we introduce **TaigiSpeech**, a real-world speech intent dataset in Taiwanese Taigi (aka Taiwanese Hokkien / Southern Min), which is a low-resource and primarily spoken language. The dataset is collected from older adults, comprising 21 speakers with a total of 3k utterances. It is designed for practical intent detection scenarios, including healthcare and home assistant applications. To address the scarcity of labeled data, we explore two data mining strategies with two levels of supervision: keyword match data mining with LLM pseudo labeling via an intermediate language and an audio-visual framework that leverages multimodal cues with minimal textual supervision. This design enables scalable dataset construction for low-resource and unwritten spoken languages. TaigiSpeech will be released under the CC BY 4.0 license to facilitate broad adoption and research on low-resource and unwritten languages. The project website and the dataset can be found on `https://kwchang.org/taigispeech`.
Index Terms: spoken language understanding, low-resource language, intent recognition, Taiwanese Hokkien, database

1. Introduction

More than 7,000 languages are spoken worldwide [1]. However, many of them are not supported by commercial AI services. A substantial portion of these languages lack a universally standardized writing system, leading to inconsistent orthographic resources, or are entirely unwritten [2]. As a result, they remain underrepresented in modern speech and language technologies. *Southern Min (Min-nan)*, commonly referred to as *Hokkien* in Southeast Asia, is one such language. It belongs to the Min branch of Sinitic languages and is spoken by over 45 million people worldwide [2, 3]. It is used across East and Southeast Asia, as well as in diaspora communities in the Americas and beyond [1].

Across regions, Southern Min exhibits substantial variation in phonology, lexicon, and accent. In Taiwan, it plays a particularly important sociolinguistic role [4]. While Mandarin is the dominant language among the general population,

Table 1: *Statistics and intent description of TaigiSpeech dataset.*

| Category	Details
Speakers	21 (Male: 8, Female: 13)
Age Range	54–78 (Mean: 67.9)
Emergency Intents	
1. SOS_CALL	Calls for emergency help.
2. BREATH_EMERG.	Reports breathing difficulty.
3. FALL_HELP	Indicates a fall and requests help.
4. PAIN_GENERAL	Reports physical pain or discomfort.
Non-Emergency Intents / Functional Commands	
5. CALL_CONTACT	Requests to contact a person.
6. LIGHT_ON	Requests to turn on the lights.
7. LIGHT_OFF	Requests to turn off the lights.
8. CANCEL	Cancels a previously triggered alert.
Total Utterances: 3,079	
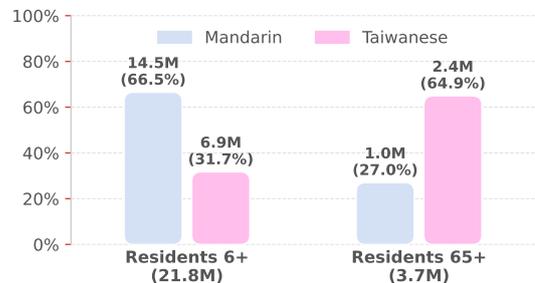
Figure 1: *Primary language use among residents in Taiwan. Elders (aged 65+) have 64.9% reporting Taiwanese as their primary language. Data source: 2020 Population and Housing Census, DGBAS, Taiwan [5].*

Taiwanese Hokkien (often simply referred to as *Taiwanese*) remains widely spoken, especially among older adults. As shown in Fig 1, language usage in Taiwan exhibits clear generational differences. Among residents aged 6 years and above (21.8M), Mandarin is the dominant primary language (66.5%), while Taiwanese is primarily used by 31.7% of the population. However, this pattern reverses among older adults. For residents aged 65 years and above (3.7M), Taiwanese becomes the dominant primary language (64.9%), whereas only 27% primarily use Mandarin [5]. This demographic shift highlights the importance of supporting Taiwanese in elderly-oriented technologies and spoken language systems, especially in home assistance and emergency/healthcare scenarios.

Table 2: *Spoken intent and command corpora across languages and application settings. (# = number of; Spk. = speakers; F = female; M = male; Utt. = utterances; TTS = text-to-speech synthesized audio).*

Corpus	Year	Intent/ Command	# Spk. (Gender)	Age	#Utt.	Hours	Language	Smart Home	Emergency
CEMO (Corpus 2) [6]	2005	Emergency	410 (256F; 154M)	–	–	20	French	–	✓
HIS [7]	2010	Emergency	15 (6F; 9M)	24-43	0.67k	0.14	French	✓	–
CARES [8]	2013	Emergency	40 (19M; 21F)	23–91	–	53.3	Canadian English	–	✓
HIS-Distress Call [9]	2013	20 Distress	10 (7M; 3F)	∼37.2	450	–	French	–	✓
SWEET-HOME [10]	2014	Command+Emergency	71 (Unknown)	19–91	9.2k	33	French	✓	✓
ITAAL [11]	2015	Commands+Distress	20 (10F; 10M)	–	2.4k	1.1	Italian	✓	✓
Speech Commands [12]	2018	30 keywords	1,881 (Unknown)	–	64.7k	18	English	–	–
FSC [13]	2019	31 intents	97 (49F; 48M)	–	30k	19	English	✓	–
VocADom@A4H [14]	2019	7 intents	11 (Unknown)	<28	26k	12	French	✓	–
SLURP [15]	2020	46 actions (18 scenes)	177 (Unknown)	–	72k	58	English	✓	–
SNIPS (synthesized) [16]	2020	7 intents	TTS (Unknown)	–	14k	–	English	✓	–
EMSAssist [17]	2023	46 EMS protocols	6 (1F; 5 M)	–	1.7k	–	English	–	✓
VN-SLU [18]	2024	28 intents	240 (Unknown)	–	17k	20	Vietnamese	✓	–
MDSC [19]	2024	10 wake words + others	46 (15F; 31M)	18–48	18.6k	17	Mandarin	✓	–
TaigiSpeech (ours)	2026	8 intents	21 (13F; 8M)	54–78	3,079	6.1	Taiwanese Hokkien	✓	✓

With rapid population ageing, an increasing proportion of older adults live alone [20, 21, 22]. In such circumstances, immediate access to physical devices, such as phones, may not always be feasible, particularly during sudden medical events [23, 24]. Smart home services and voice-enabled systems, therefore, play a critical role in enabling timely access to medical assistance and community support [25]. For instance, falls are among the most serious health risks faced by older adults [26, 27], with consequences that are especially severe for those living alone. A critical outcome is the so-called *long lie* [28], a prolonged period of immobility after a fall when an individual cannot summon help, which is associated with serious medical complications and increased mortality. Beyond falls, other emergency conditions, such as breathing distress or acute pain, as well as daily functional needs, e.g., controlling lighting or household devices, likewise require reliable hands-free communication/systems [29, 30]. These real-world scenarios underscore the need for robust spoken intent recognition systems to create a safer and more supportive living environment for elderly individuals [31, 32, 33, 34].

However, existing intent recognition datasets predominantly focus on high-resource languages [35] such as English [13], Mandarin [36], and French [14], leaving low-resource languages like Taiwanese Hokkien largely unexplored. Moreover, due to limited linguistic resources, reliable automatic speech recognition (ASR) systems are often unavailable. The lack of domain-specific datasets further constrains the development of practical spoken language understanding (SLU) systems for elderly Taiwanese speakers. To address this gap, we introduce a real-world spoken intent dataset collected from elderly Taiwanese speakers. We recruit participants to record natural speech across diverse healthcare and home-assistant scenarios. The dataset comprises 8 intent categories, including 4 emergency and 4 non-emergency functional intents, forming a pivot Taiwanese Hokkien dataset for healthcare and home-assistant spoken intent recognition.

Furthermore, to address the scarcity of labeled training data, we explore two practical strategies to mine potential intent segments from in-the-wild video sources: (1) *Keyword Match Mining*, which leverages subtitles in a high-resource pivot language (Mandarin), and (2) *Audio-Visual Mining*, which utilizes multimodal representations to discover relevant segments without textual supervision. Together, these approaches provide scalable ways to collect training data and support the development of elderly-oriented intent recognition systems.

To assess the effectiveness of the constructed dataset, we evaluate several baseline models, including lightweight neural networks and self-supervised learning (SSL) speech models for spoken intent recognition. Experimental results show a substantial performance degradation when models trained on mined in-the-wild data are evaluated on real-world elderly recordings, indicating a significant domain mismatch. This finding highlights the challenges of real-world deployment and underscores the necessity of TaigiSpeech as a realistic benchmark for low-resource SLU. Our contributions are summarized as follows:

- **TaigiSpeech Dataset.** We introduce TaigiSpeech, a real-world spoken intent dataset collected from elderly Taiwanese Hokkien speakers in healthcare and home-assistant scenarios. The dataset comprises 8 intent categories (4 emergency and 4 functional intents), covering 21 speakers and over 3,000 utterances. It addresses a critical gap in low-resource spoken language understanding for elderly-oriented applications. The dataset will be publicly released under a CC BY 4.0 license to facilitate future research.

- **Exploration of Data Mining Strategies.** We investigate practical strategies for mining training data from in-the-wild sources, utilizing both cross-lingual keyword matching and audio-visual cues to address the lack of labeled data in low-resource languages.

- **Realistic Benchmark and Empirical Insights.** Through extensive experiments with lightweight and SSL speech models, we observe a substantial performance degradation when models trained on mined in-the-wild data are evaluated on real-world elderly recordings, indicating a significant domain mismatch. This finding highlights the challenges of real-world deployment and underscores the necessity of realistic evaluation benchmarks for low-resource spoken intent recognition.

2. Related Works

2.1. Spoken Language Understanding

Spoken language understanding (SLU) aims to extract semantic meaning from speech signals and serves as a core component in voice assistants and spoken dialogue systems [43]. A central subtask in SLU is *intent recognition*, which identifies the user's underlying intention from spoken utterances and triggers

Table 3: *Overview of Taiwanese Hokkien speech corpora across recording conditions and speech styles.*

Corpus	Recording Condition	Speech Style	Target Task	Hours	Speaker Type
SuiSiann [37]	Studio	Reading speech	TTS	20	1 female
TAT [38]	Studio	Reading speech	ASR	100	92 males, 108 females
TAT-TTS [38]	Studio	Reading speech	TTS	40	2 males, 2 females
CommonVoice [39]*	Crowdsourced	Reading speech	ASR	23.67	Diverse speakers
ML-SUPERB [40]**	TV production	Scripted dialogue	ASR	1.5	Professional Actors
TG-ASR [41]	TV production	Scripted dialogue	ASR	30	Professional Actors
TaigiSpeech	Quiet rooms	Scenario-driven Expressive	Intent Recognition	6.1	Elderly speakers

* CommonVoice refers to the Mozilla Common Voice 24.0 Taiwanese (Minnan) subset.
** Taiwanese Hokkien subset [42] in Multilingual SUPERB (ML-SUPERB)

downstream functionality, such as information retrieval, content delivery, and device control.

Early SLU benchmarks such as ATIS [44] established a narrow-domain benchmark for task-oriented query understanding, while the Speech Commands [12] dataset later provided a widely adopted benchmark for limited-vocabulary spoken command recognition in a keyword-spotting style setting. Subsequent datasets moved closer to practical assistant scenarios. Fluent Speech Commands (FSC) [13] became a standard dataset for end-to-end spoken intent recognition in smart-home settings. SLURP [15] further expanded the scale and semantic diversity of English SLU by covering multi-domain home-assistant scenarios, while SLUE [45] promoted more systematic evaluation on natural speech and pretrained speech representations. In Table 2, we survey representative SLU datasets for smart home and emergency detection scenarios.

On the modeling side, conventional approaches adopt a cascade pipeline that first performs ASR and then applies text-based natural language understanding (NLU) [13]. Although effective in high-resource settings, such pipelines are sensitive to transcription errors and depend heavily on the availability of reliable ASR systems. This limitation motivated the development of end-to-end SLU methods that predict semantic meaning and intent directly from speech signals without an intermediate text representation. More recently, SSL speech models [46] such as HuBERT [47] and WavLM [48] have provided informative speech representations for downstream SLU fine-tuning and can deliver strong performances [49]. Nevertheless, most existing datasets and methods remain centered on high-resource languages. Real-world benchmarks for elderly speakers, emergency scenarios, and low-resource unwritten languages remain scarce, which motivates the proposed TaigiSpeech dataset.

2.2. Command/Smart-home Emergency Scenario Corpora

The deployment of voice assistants in ambient assisted living (AAL) environments has driven the creation of specialized corpora focusing on distress calls and emergency triage. Unlike general smart home/command datasets [12] that primarily capture routine domestic commands (e.g., turning on lights), emergency corpora may capture the unique acoustic characteristics of high-stress situations.

An early and fundamental effort in this domain is the HIS (Health Smart Home) corpus [7] and its targeted distress call subset [9]. Recorded in a fully equipped smart home laboratory in France, this dataset captures normal daily interactions alongside 20 specific distress phrases (e.g., calling for help) uttered by participants. While highly valuable for evaluating keyword spotting and emergency detection in noisy, reverberant home environments, the dataset primarily features younger adults (ages 22–61) and focuses on isolated distress phrases

rather than complex conversational interactions.

Moving beyond isolated keyword spotting toward conversational triage, the CARES (Canadian Adult Regular and Emergency Speech) corpus [8] was developed specifically to train automatic speech recognition interfaces for personal emergency response systems (PERS). It features simulated emergency dialogues, isolated phrases, and spontaneous speech from 40 adult actors. Crucially, the CARES corpus emphasizes the voices of older adults (ages 23-91), providing essential insights into how aging and simulated physiological stress (e.g., shortness of breath, pain) affect speech recognition performance.

Despite these significant contributions, as shown in Table 2, existing emergency corpora exhibit critical limitations. For example, most focus on high-resource languages (French and English) and few target the elders. To address these limitations, our proposed **TaigiSpeech** dataset bridges this gap by combining smart home automation commands with emergency distress scenarios, specifically targeting elderly speakers (ages 54–78) in a low-resource linguistic setting.

2.3. Taiwanese Hokkien Speech Corpus

Hokkien is a major branch of Southern Min (Min-nan) Chinese originating from Fujian, China, with over 45 million speakers worldwide, primarily in Southeast Asia (Singapore, Malaysia, Indonesia, and the Philippines), southern China, and Taiwan. In Taiwan, writing schemes such as Taiwanese Romanization (Tai-lo), Peh-oe-ji (POJ), and other orthographic conventions exist; however, they are not yet universally adopted, making the collection of large-scale annotated speech corpora challenging.

In recent years, efforts have been made to develop Taiwanese Hokkien speech corpora through both controlled recordings and web-crawled resources. Table 3 summarizes representative datasets in the speech processing domain. As shown in the table, most existing corpora rely on read speech (e.g., TAT [38], CommonVoice [39]) or scripted dialogues (e.g., ML-SUPERB [40]) designed for TTS or ASR tasks. Some datasets are sourced from television dramas [42, 41]; however, they often contain background music and post-production effects, as they are originally produced for TV broadcasting.

While valuable for general speech recognition, these datasets lack the natural, scenario-driven expressiveness required for intent recognition in emergency and assistance contexts. TaigiSpeech is the first Taiwanese Hokkien dataset specifically designed for spoken intent recognition in, with a focus on elderly speakers. Collecting speech from elderly speakers is particularly valuable, as Taiwanese Hokkien is often the only fluent language spoken by many older individuals.

It is worth noting that several speech foundation models have been proposed in recent years. Although not specifically trained for Taiwanese Hokkien, models such as Whisper [50]

and Qwen3-ASR [51] have demonstrated the ability to recognize Hokkien and other Min-nan varieties to some extent. We therefore include these models, cascaded with LLM, as baseline systems in our experiments.

3. TaigiSpeech

In this section, we describe the data collection procedure, participant recruitment process, scenario design, and basic statistics of the TaigiSpeech dataset

3.1. Participant Registration and Recording Interface

Figure 2: *Recording app interface. The English translations are shown for illustrative purposes in this paper. All recordings were conducted under the researcher's supervision to assist elderly participants.*

We designed and developed a web-based recording application (Figure 2) that can be accessed on mobile phones, laptops, desktops, and tablets, using either built-in or external microphones (shown in Supplementary A.1). The application supports participant registration, speech recording, playback, and submission.

Participants first register their personal information, including (1) age (2) gender (3) education level (4) native language, (5) Taiwanese Hokkien fluency, and (6) hometown[1]. Participants are allowed to select their preferred session duration (e.g., 30, 60, or 120 minutes, corresponding to 5, 10, or 20 utterances per intent) to reduce tiredness. The interface enables replaying, re-recording, reordering, and undoing recordings via mouse, touch, or keyboard input. We observed that providing multiple input options was essential, particularly for elderly participants who may be less familiar with certain devices, such as

---

[1]Geographic background influences accent variation. Collecting hometown information helps ensure accent diversity and supports future phonetic and sociolinguistic studies.

Table 4: *Intent-level statistics of the TaigiSpeech dataset.*

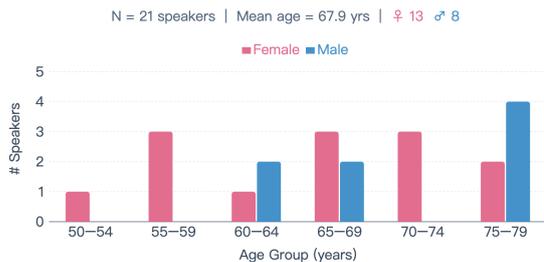| Intent	Count	Duration (s)	Std. (s)	Total (min)
1. SOS_CALL	387	7.78	3.28	50.16
2. FALL_HELP	385	7.77	3.73	49.88
3. BREATH_EMERG	385	8.22	3.30	52.71
4. PAIN_GENERAL	385	7.91	3.73	50.79
5. CALL_CONTACT	385	7.04	5.30	45.16
6. LIGHT_ON	384	5.95	2.81	38.08
7. LIGHT_OFF	384	5.67	2.35	36.28
8. CANCEL_ALERT	384	6.85	3.17	43.85
Overall	**3,079**	**7.15**	**3.66**	**366.91**

Figure 3: *Age distribution of speakers in the TaigiSpeech corpus. Our study focuses on elderly participants.*
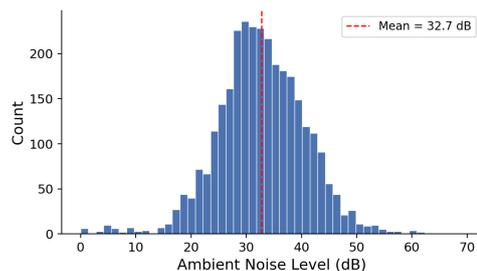

Figure 4: *The distribution of ambient noise level (dB).*

a mouse or touchscreen.

3.2. Scenario Design

TaigiSpeech targets 8 intents, covering both emergency and non-emergency functional speech commands. The selected intents are based on key priorities in geriatric care and common home assistance needs for aging in place.

We identified four critical emergency categories based on eldercare literature and public health statistics: *Fall*, *Pain*, *Breathing*, and *Help*. Falls are the leading cause of injury and accidental death among older adults [26], making their detection a primary safety requirement. Similarly, symptoms such as acute chest pain and shortness of breath are frequent indicators of medical emergencies requiring immediate intervention [52]. The generic *Help* intent was included to capture diverse distress situations not covered by specific medical descriptions.

Complementing these are four non-emergency functional intents: *Light on*, *Light off*, *Call contact*, and *Cancel alert*. These were selected to assist with instrumental activities of daily living, as voice control significantly reduces the physical strain of environmental management for mobility-impaired individuals. This balanced design (4 emergency vs. 4 functional)

Table 5: *Examples of scenario prompts and corresponding elicited spoken utterances in the TaigiSpeech recording protocol. The utterances shown here are English translations of the original Taiwanese Hokkien speech.*

Intent	Scenario Prompt (Imagined Context)	Example Spoken Utterance (English Translation)
Emergency Intents		
1. SOS_CALL	You are at home and notice smoke and flames spreading.	"Ahh! Fire! It's on fire! Help!"
2. FALL_HELP	You slipped in the bathroom and cannot stand up.	"Ah! I fell down... I can't get up. Please help me!"
3. BREATH_EMERG	You suddenly feel chest tightness and difficulty breathing.	"Ah... my chest feels tight. I can't breathe well..."
4. PAIN_GENERAL	You suddenly feel severe stomach pain and cannot move.	"Ahh... my stomach... it hurts so much. I can't move..."
Non-emergency Functional Intents		
5. CALL_CONTACT	You feel unwell and want to call your daughter.	"Call my daughter... ask her to come home."
6. LIGHT_ON	You wake up at night and need more light to walk safely.	"Turn on the bedroom light."
7. LIGHT_OFF	You are going to sleep and want to turn off the lights.	"Turn off the lights."
8. CANCEL_ALERT	You accidentally triggered an alert and want to cancel it.	"No, no! Cancel it. This was a mistake."

enables the evaluation of a system's ability to distinguish critical distress signals from routine commands, a vital feature for reliable home assistant systems.

To elicit natural speech, we design imagined scenarios that allow participants to situate themselves using Google Gemini Pro 2.5 [53] (details, e.g. prompts, are in Supplementary A.2). For example, a prompt may describe a situation such as: "Imagine you have fallen down the stairs and cannot get up." Participants are encouraged to respond freely rather than read fixed scripts to increase linguistic diversity and better reflect real-world usage. Each intent contains 20 distinct scenarios, resulting in 160 prompts (20 x 8) per full recording session. Examples of the imagined scenario, instruction prompt, and the participant's produced utterance are shown in Table 5.

To further enhance realism and immersion, we provide silent video stimuli generated using Google Veo 3 [54] (details, e.g. prompts, are in Supplementary A.3). These 10-second videos are designed to help participants better envision the scenario if needed. Importantly, participants are explicitly instructed that viewing these video clips was optional, and they are not required to follow the video. It serves only as a supportive context. Participants were free to respond in their own words, ensuring that the collected utterances reflect spontaneous and natural speech. Consequently, the resulting data includes not only direct commands but also fragmented sentences, repetitions, and urgent vocalizations typical of genuine distress.

3.3. Recording

The recording process was conducted in strict accordance with protocols approved by the Institutional Review Board (IRB). All recordings were obtained in accordance with ethical guidelines to ensure participant privacy and data protection. During each recording session, at least one trained moderator from the research team was present to introduce the project, explain the procedures, and demonstrate how to use the recording interface.

To ensure the quality and consistency of the collected speech data, the recording interface implemented a pre-recording signal quality control protocol. Before each recording commenced, the system initiated a one-second silent period to measure the *Ambient Noise Level*. The system repeatedly sampled the average frequency-domain energy, calculating a final mean value to represent the background noise. This unitless indicator was logged as metadata for each audio file, providing an objective metric for assessing the recording environment's quietness and enabling data filtering for subsequent analysis.

3.4. Statistics

In total, we collected data from 21 elderly speakers, including 13 females and 8 males. Most speakers completed 160 utterances, while a small number of participants did not reach the full quota. On average, each speaker contributed 146.6 utterances. The average duration per utterance is 7.15 seconds, resulting in a total of approximately 6.3 hours of recorded speech, comprising 3,079 utterances overall. All audios are recorded in mono channel with 48k Hz. The per-intent statistics are provided in Table 4. The age distribution of the participants is shown in Figure 3 and the noise level distribution is presented in Figure 4.

4. Data Mining in-the-wild

For many low-resource and unwritten languages, reliable ASR systems are often unavailable. Also, in most cases, speech data primarily exist in web-based sources (e.g., online videos), where recordings are noisy and loosely structured. These constraints make large-scale supervised data collection impractical, motivating the development of scalable data mining methods. The data mining goal is to automatically discover suitable *speech segments* for downstream intent recognition in the target language (e.g., Taiwanese Hokkien in this paper) from a large source pool.

Inspired by [55, 56], we adopt the Taiwanese Drama dataset as a large-scale source pool for mining candidate speech segments. We select this source because its conversational style and emotional expressiveness more closely align with target smart-home and emergency scenarios than other available corpora. Furthermore, the videos are accompanied by Mandarin subtitles, which provide a practical pivot language and facilitate the design of our mining pipeline. The dataset contains about 7,000 hours of videos. Specifically, we explore two strategies: (1) **Keyword Match Mining**, a stronger supervised approach that leverages Mandarin subtitles and LLM pseudo-labeling filtering; and (2) **Audio-Visual Mining**, a weakly supervised approach that relies on pre-trained multimodal representations with weakly textual supervision.

4.1. Keyword Match Data Mining

Inspired by [56], we adopt Mandarin as an intermediate language for data mining. Although Taiwanese Hokkien does not have a widely adopted written form, Mandarin translations are usually available in the drama subtitles.

As shown on Figure 5 (Left), for each intent $i \in \mathcal{I}$, we manually construct a set of intent-specific keywords $\mathcal{K}_i =$
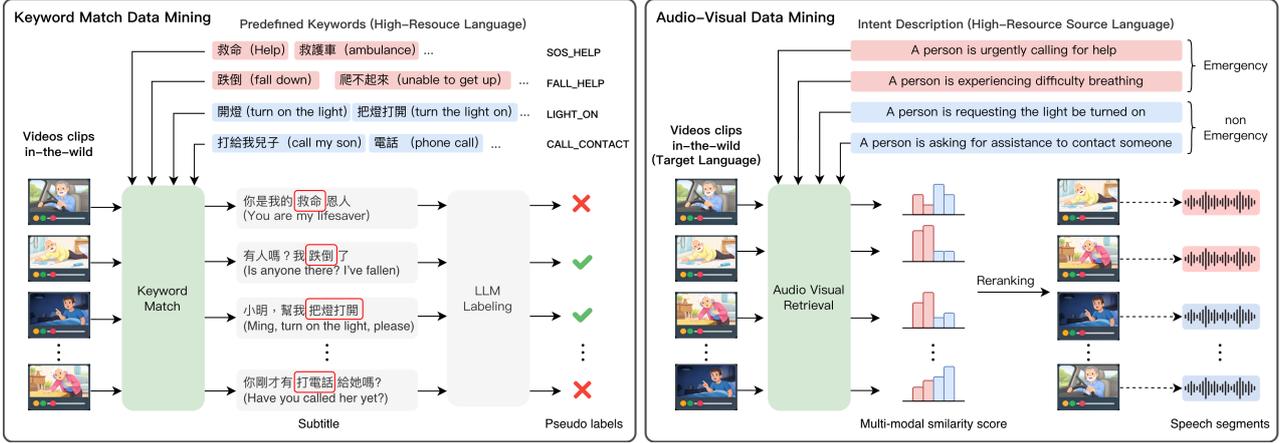
Figure 5: *Overview of the explored data mining methods. Left: Keyword Match Data Mining. Right: Audio-Visual Data Mining*

$\{k_1, k_2, \ldots, k_m\}$. The initial set of keywords is constructed by native speakers proficient in both Taiwanese Hokkien and Mandarin. It is subsequently expanded to incorporate synonyms, commonly used verbal expressions, and context-dependent variations, thereby maximizing lexical coverage. For example, for the SOS_CALL intent, representative Mandarin keywords include *"help"*, *"call the ambulance"*, and *"I need a doctor"*. We first retrieve subtitle sentences that contain at least one keyword from $\mathcal{K}_i$. For each retrieved sentence s_t, we further include its surrounding context (the preceding and following five sentences) to form a context window. An LLM is then employed to perform *pseudo-labeling*, determining whether the contextualized segment semantically corresponds to the target intent i. If labeled as positive, the corresponding speech segment aligned with s_t is extracted.

The statistics of the pseudo-labeled dataset are shown in Table 6. In practice, we use at least 10 keywords per intent for keyword-based mining. In addition, we incorporate approximately 100 daily-life keywords (e.g., *"good morning"*, *"go to bed"* in Mandarin) to collect diverse non-intent samples, which are grouped into the OTHERS category. It is worth noting that although the drama dataset contains approximately 7,000 hours of video, the mined data remain sparse for certain intents. This observation highlights the importance of constructing TaigiSpeech in order to build a robust intent recognition system for Taiwanese Hokkien in home healthcare scenarios.

Keyword Match Mining relies on subtitle translations and the LLM's capability in the intermediate language. Although effective, it may not be applicable to unwritten languages without parallel text. To overcome this limitation, we explore an alternative audio-visual mining strategy that does not require paired textual data and operates under substantially weaker supervision.

4.2. Audio-Visual Data Mining

While text-aligned corpora are scarce, large amounts of audio-visual content are readily available from online videos and daily recordings. Acoustic signals convey paralinguistic information (e.g., emotion, urgency, prosody) that is not tied to a specific written form, and visual context provides complementary semantic grounding independent of language. Recent contrastive multimodal encoders [58] enable cross-modal alignment in a shared embedding space without requiring explicit transcription. We therefore explore an audio-visual mining framework

Table 6: *Statistics (number of sentences) of LLM pseudo labeling in the Keyword Match Mining method.*

Intent	Total (#)	True (#)	False (#)
1.SOS_CALL	4114	2801	1313
2.FALL_HELP	3017	165	2852
3.BREATH_EMERG	614	121	493
4.PAIN_GENERAL	5729	3054	2675
5.CALL_CONTACT	5228	2203	3025
6.LIGHT_OFF	140	85	55
7.LIGHT_ON	157	33	124
8.CANCEL_ALERT	946	55	891
OTHERS	7987	7987	0

based on a pretrained cross-modal retrieval model.

Let $\mathcal{D} = \{v_j\}_{j=1}^N$ denote a pool of N unlabeled video clips collected from available sources. Each clip v_j consists of synchronized audio and visual streams, written as $v_j = (a_j, x_j)$, where a_j is the audio signal and x_j the visual frames. Let $\mathcal{T} = \{t_i\}_{i=1}^K$ denote a set of intent descriptions written in a source language L_s, typically a high-resource language on which the multimodal encoder is pretrained. Our goal is to retrieve speech segments in a low-resource target language L_t that correspond to each intent t_i, without requiring paired annotations in L_t.

4.2.1. Multimodal Retrieval

We adopt a pretrained cross-modal encoder, PE-AV [58], composed of an audio-visual encoder $f_{av}(\cdot)$ and a text encoder $f_{text}(\cdot)$, mapping inputs into a shared d-dimensional embedding space. For each video clip v_j, we compute $\mathbf{z}_j = f_{av}(v_j) \in \mathbb{R}^d$, and for each intent description t_i, we compute $\mathbf{q}_i = f_{text}(t_i) \in \mathbb{R}^d$. We then measure the *multimodal similarity score* using cosine similarity.

We then retrieve the top-M clips according to the similarity scores. The retrieved set is treated as pseudo-labeled data for intent t_i in the target language L_t. This formulation enables weakly supervised cross-lingual data mining: although intents are defined in L_s, the shared embedding space allows retrieval of aligned speech in L_t without transcription or manual labeling.

In this work we found that under the audio-visual mining setting, obtaining fine-grained emergency labels (e.g., distinguishing breath-related emergencies from fall events) remains challenging, possibly due to the cross-lingual setting and weak

Table 7: *Accuracy (%) comparison on the Drama and TaigiSpeech datasets (Taigi) under two data mining strategies. C=5: a 5-class setting (4 emergency intents plus a NON-EMERGENCY category). C=2: binary classification (EMERGENCY vs. NON-EMERGENCY). C=8: the original 8-intent setting in TaigiSpeech. **Taigi-adapt** means models trained on the mined data are further fine-tune on the training set of TaigiSpeech. **Scratch**: Models trained from scratch without pre-trained on mined data. For the Drama test set, the ground-truth labels are obtained via LLM pseudo-labeling, thus evaluating the consistency with the LLM-generated labels.*

Scenario		Keyword Match Mining			Audio-Visual Mining			Scratch
Model	# Params	Drama (C=5)	Taigi (C=5)	Taigi-adapt (C=8)	Drama (C=2)	Taigi (C=2)	Taigi-adapt (C=8)	Taigi (C=8)
MatchboxNet-S	69k	$60.11_{\pm2.51}$	$45.83_{\pm7.97}$	$28.02_{\pm5.68}$	$71.99_{\pm2.30}$	$51.88_{\pm6.67}$	$27.29_{\pm5.73}$	$23.65_{\pm5.99}$
MatchboxNet-M	85k	$64.99_{\pm2.44}$	$48.02_{\pm7.55}$	$33.33_{\pm6.20}$	$74.12_{\pm2.37}$	$52.71_{\pm7.24}$	$25.94_{\pm5.62}$	$19.17_{\pm4.64}$
MatchboxNet-L	119k	$63.86_{\pm2.51}$	$43.44_{\pm8.54}$	$29.38_{\pm6.82}$	$72.70_{\pm2.33}$	$53.75_{\pm7.66}$	$24.79_{\pm5.68}$	$23.65_{\pm5.47}$
HuBERT-base	94M	$89.60_{\pm1.56}$	$67.92_{\pm7.03}$	$90.10_{\pm3.85}$	$79.49_{\pm2.05}$	$54.58_{\pm8.44}$	$86.77_{\pm4.43}$	$84.69_{\pm5.21}$
HuBERT-large	315M	$90.59_{\pm1.49}$	$69.48_{\pm6.82}$	$84.06_{\pm4.79}$	$78.93_{\pm2.05}$	$53.75_{\pm7.97}$	$77.60_{\pm5.05}$	$78.65_{\pm5.73}$
WavLM-base	94M	$87.77_{\pm1.73}$	$71.25_{\pm6.61}$	$88.96_{\pm4.64}$	$78.50_{\pm2.12}$	$\mathbf{54.79_{\pm8.65}}$	$\mathbf{89.06_{\pm5.10}}$	$\mathbf{88.54_{\pm4.58}}$
WavLM-base-plus	94M	$89.67_{\pm1.59}$	$\mathbf{73.23_{\pm6.04}}$	$\mathbf{90.21_{\pm4.12}}$	$80.55_{\pm2.02}$	$54.69_{\pm8.54}$	$87.19_{\pm4.53}$	$86.98_{\pm5.00}$
WavLM-large	315M	$\mathbf{92.36_{\pm1.38}}$	$70.00_{\pm6.93}$	$83.96_{\pm4.95}$	$\mathbf{81.75_{\pm2.05}}$	$52.50_{\pm7.55}$	$64.38_{\pm6.61}$	$77.08_{\pm5.57}$

Confidence intervals are computed using the bootstrapping method [57].

Table 8: *Data statistics under different mining strategies and Taigi adaptation setting.*

	Keyword Mining	AV Mining
# Target Classes	5	2
Train	11.3k	3.6k
Val	1.4k	400
Test (Drama)		1.4k
Test (Taigi)		960
Taigi-adapt (Independent-speaker Split)		
Train	1600 (4 male, 6 female)	
Val	119 (1 male, 4 female)	
Test	960 (3 male, 3 female)	

supervision. We instead aggregate the similarity scores of the four emergency intents and rank videos based on their summed scores to retrieve coarse-grained emergency clips. In the preliminary study, we found the pseudo-labels with this setting are consistent with that of LLM labeling.

5. Experiments

5.1. Experimental Setup

For pseudo-labeling in the Keyword Mining approach, we employ Gemini-3 to generate intermediate annotations. For audio-visual data mining, we adopt the PE-AV large [58] to extract multimodal representations and identify candidate segments without relying on paired textual supervision. Note that the PE-AV model is mainly trained on English videos and text and not trained on the low-resource Taiwanese Hokkien data. Due to computational constraints, rather than directly utilizing the full 7,000 hours of the drama dataset, we conduct a smaller-scale exploration using a dataset derived from keyword-matched queries to ensure that the video clips contain speech. This data pool contains *unlabeled*, mixed emergency and non-emergency (other) segments, totaling about 28k video clips. Future work will focus on improving computational efficiency and conducting more comprehensive large-scale experiments. We select top 2,000 as emergency data and last 2,000 as non-emergency data.

For baseline models, we include both lightweight models (MatchboxNet [59]) and self-supervised learning (SSL) speech models (HuBERT [47] and WavLM [48] variants), covering resource-efficient models as well as large-scale pretrained paradigms. To construct a balanced evaluation benchmark, we

select 6 speakers from TaigiSpeech to form the test set. The set consists of 3 male and 3 female speakers, with ages distributed to maintain balance across gender and age groups. Each speaker contributes 160 utterances, resulting in a total of 960 test samples.

Models are first trained on data mined using the two proposed strategies: **Keyword Match Mining** and **Audio-Visual Mining**. In the Keyword Match Mining setting, we focus on a 5-class classification task (four emergency intents plus a NON-EMERGENCY category). In the Audio-Visual Mining setting, we consider a binary classification task (Emergency vs. Non-Emergency). Models trained on the mined data are evaluated on the drama dataset (denoted as **Drama** [2]) and the real-world recording dataset TaigiSpeech (denoted as **Taigi**). Note that in all scenarios, only the training data differ; the testing data remain the same, and the evaluation protocol depends on the class we focus on.

To simulate a realistic low-resource domain-adaptation scenario and examine domain-mismatch effects, we further fine-tune the models pretrained in the previous stage using TaigiSpeech data of 10 speakers (totaling 1,600 utterances). The adapted models are then evaluated on the Taigi test set. This setting is denoted as **Taigi-adapt**. The data splits used in the experiments can be found in Table 8.

6. Preliminary Results

The main results are presented in Table 7. We summarize the key observations in the following sections.

6.1. Domain mismatch

A clear performance discrepancy is observed between evaluations on the drama dataset and on the real-world TaigiSpeech recordings. This trend consistently appears in both settings: **Drama (C=5)** $\rightarrow$ **Taigi (C=5)** under the Keyword Match Mining scenario, and **Drama (C=2)** $\rightarrow$ **Taigi (C=2)** under the Audio-Visual Mining scenario. In both cases, model performance drops substantially when transferred to the TaigiSpeech test set. For example, under Keyword Match Mining, the top-performing WavLM-large model drops from 92.36% on Drama to 70.00% on TaigiSpeech, while HuBERT-base degrades from 89.60% to 67.92%.

We further observe that current audio-visual mining approaches struggle to achieve non-trivial performance under di-

---

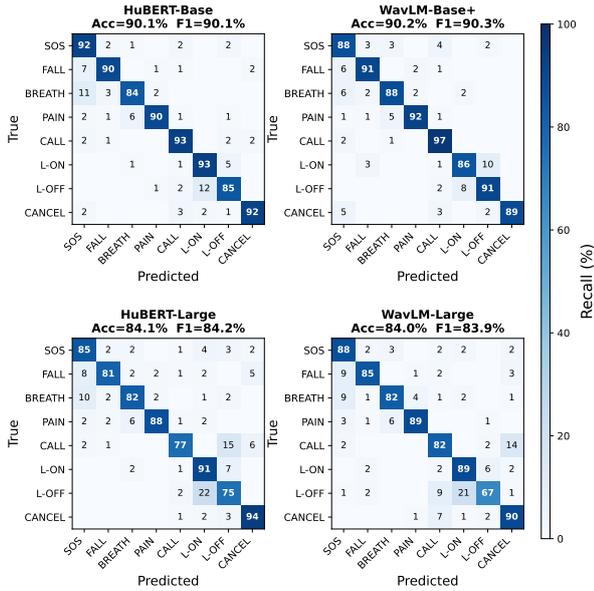[2]Pseudo labels obtained from Gemini-3 are treated as ground truth

Figure 6: *Confusion matrix of SSL speech models under the Taigi-adapt setting in the Keyword Match Mining scenario.*

rect domain transfer (**Taigi (C=2)**), yielding near-random binary classification accuracies across all models. Interestingly, we observed that although the retrieved samples exhibit high consistency with the LLM pseudo labels in the keyword mining stage, and can achieve approximately 70-80% accuracy in drama dataset, this capability does not translate into Taigi.

We hypothesize that under the simplified two-class setting, the model may fail to learn robust semantic meaning, instead relying on superficial patterns that do not generalize across domains. Combined with the substantial domain mismatch between drama speech and elderly Taigi recordings, this leads to near-random performance. Future work may involve fine-tuning the multimodal encoder, exploring alternative training objectives, and evaluating the approach under different class configurations and larger-scale datasets to better understand and improve its generalization ability.

However, under the **Taigi-adapt** setting, where a small amount of in-domain labeled data is available, performance can be substantially recovered. For instance, HuBERT-base and WavLM-base-plus achieve 90.10% and 90.21% accuracy, respectively, under the 8-class Taigi-adapt setting with Keyword Match Mining. Overall, models pre-trained with Keyword Match Mining demonstrate stronger adaptation and perform better in Taigi-adapt compared to Audio-Visual Mining.

Although large-scale unlabeled corpora from drama sources can be effectively leveraged for data mining and pre-training, the significant degradation of real-world elderly recordings reveals the limitation of relying solely on in-the-wild mined data. This finding underscores the importance of TaigiSpeech as a realistic benchmark for evaluating and advancing practical spoken intent recognition systems in Taiwanese Hokkien.

6.2. Model Efficiency

A substantial performance gap is observed when comparing lightweight models (e.g., MatchboxNet) with SSL pre-trained models. This gap highlights a critical deployment challenge in smart home and healthcare scenarios, where edge devices often

Table 9: *Comparison between Cascade ASR+LLM and End-to-End Speech Classification Systems on Taiwanese.*

Model	# Params	Accuracy (%)
Cascade System (ASR + Qwen3-8B)		
Whisper base	74M + 8B	15.00
Whisper small	244M + 8B	15.83
Whisper medium	769M + 8B	42.92
Whisper large-v3	1.55B + 8B	34.38
Qwen3-ASR 0.6B	0.6B + 8B	63.65
Qwen3-ASR 1.7B	1.7B + 8B	74.48
End-to-End Speech Classification		
HuBERT Base	94M	90.10
WavLM Base	94M	88.96
WavLM Base plus	94M	**90.21**

face strict computational constraints and privacy considerations. In such settings, large-scale models may not be deployable locally, and transmitting sensitive speech data to cloud servers is often undesirable or infeasible.

These findings reveal a practical trade-off between model capacity and deployability. Future work will therefore focus on developing more computationally efficient yet competitive architectures, improved training strategies, and more effective data mining approaches to better balance performance and real-world usability.

6.3. Intent Classification Confusion Matrix

Figure 6 presents the confusion matrices of four SSL-based models after continual fine-tuning. Among the four emergency intents, we observe cross-class confusion, as these intents might share similar acoustic characteristics and overlapping vocabulary in spoken Taiwanese. Among the functional intents, LIGHT_ON and LIGHT_OFF are the most commonly confused pair, as their spoken commands differ by only a single word. Since the test set is balanced with 160 samples per class, accuracy and macro F1 are closely aligned across all models.

6.4. Foundation ASR+LLM cascaded methods

Recent studies have shown that large-scale foundation ASR models trained on extensive web data can transcribe or translate Southern Min (Taiwanese Hokkien) to a certain extent. Representative examples include Whisper [50] and Qwen3-ASR [51]. We therefore evaluate a cascaded approach for intent recognition based on these foundation models. Specifically, we first use the ASR model to transcribe TaigiSpeech recordings into text, and then employ a LLM, Qwen3 8B [60], to infer the corresponding intent label from the transcription.

Experimental results indicate that this cascaded ASR+LLM approach achieves reasonable performance. However, it still requires substantially larger model parameters and computational resources, and its performance remains inferior to end-to-end speech classification models such as HuBERT and WavLM. Note that we do not fine-tune the foundation ASR models and the LLM, which remains our future work.

7. Conclusion and Future Work

In this work, we introduce TaigiSpeech, a Taiwanese Hokkien spoken intent dataset designed for elderly users in home-assistant and healthcare scenarios. The dataset comprises 8 intent categories, including four emergency intents

(SOS_CALL, BREATH_EMERGENCY, FALL_HELP, and PAIN_GENERAL) and four non-emergency functional intents (CALL_CONTACT, LIGHT_ON, LIGHT_OFF, and CANCEL_ALERT). It contains recordings from 21 elderly speakers, totaling more than 3,000 utterances. The dataset will be publicly released under the CC BY 4.0 license. In addition, we explore two data mining strategies to collect training data for baseline models. Experimental results reveal a clear performance gap, indicating a severe domain mismatch between mined in-the-wild data and real-world elderly speech. T his finding underscores the necessity of TaigiSpeech as a realistic benchmark for building practical Taiwanese Hokkien home-assistant systems.

In future work, we plan to investigate more efficient and robust data mining strategies and to explore lightweight yet effective speech classification models suitable for real-world deployment. We have been expanding the dataset to include a broader range of speakers, extending beyond elderly participants in Taiwan to include Hokkien (Min-Nan) speakers worldwide. We hope this work will contribute to the development of inclusive spoken language technologies for low-resource languages.

8. Acknowledgments

9. Generative AI Use Disclosure

Generative AI tools were used solely for language editing and minor polishing of the manuscript. All authors take full responsibility for the content of this paper, including its accuracy, originality, and conclusions. No generative AI tools were used to produce substantial portions of the scientific content, analysis, or results.

10. References

[1] W. R. Leben, "Languages of the World," 02 2018. [Online]. Available: https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-349

[2] D. M. Eberhard, G. F. Simons, and A. J. Robinson, Eds., *Ethnologue: Languages of the World*, 29th ed. SIL Global, 2026.

[3] B.-C. Kwok, *Southern Min: Comparative phonology and subgrouping*. Routledge, 2018.

[4] C. C. Kubler, "The influence of Southern Min on the Mandarin of Taiwan," *Anthropological Linguistics*, pp. 156–176, 1985.

[5] A. Directorate General of Budget and T. Statistics (DGBAS), "Language Usage for the Resident Nationals Aged 6 Years and Over - 2020 Population and Housing Census," 2021.

[6] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005, emotion and Brain.

[7] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A Multimodal Corpus Recorded in a Health Smart Home," in *LREC 2010, The International Conference on Language Resources and Evaluation*, Valetta, Malta, May 2010, pp. 99–105.

[8] V. Young and A. Mihailidis, "The CARES corpus: a database of older adult actor simulated emergency dialogue for developing a personal emergency response system," *International Journal of Speech Technology*, vol. 16, no. 1, pp. 55–73, 2013.

[9] A. Fleury, M. Vacher, F. Portet, P. Chahuara, and N. Noury, "A French corpus of audio and multimodal interactions in a health smart home," *Journal on Multimodal User Interfaces*, vol. 7, no. 1, pp. 93–109, 2013.

[10] M. Vacher, B. Lecouteux, P. Chahuara, F. Portet, B. Meillon, and N. Bonnefond, "The Sweet-Home speech and multimodal corpus for home automation interaction," in *LREC 2014*, Reykjavik, Iceland, May 2014, pp. 4499–4506.

[11] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, and F. Piazza, "An integrated system for voice command recognition and emergency detection based on audio signals," *Expert Syst. Appl.*, vol. 42, no. 13, p. 5668–5683, Aug. 2015.

[12] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[13] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech Model Pre-Training for End-to-End Spoken Language Understanding," in *Interspeech 2019*, 2019, pp. 814–818.

[14] F. Portet, S. Caffiau, F. Ringeval, M. Vacher, N. Bonnefond, S. Rossato, B. Lecouteux, and T. Desot, "Context-Aware Voice-Based Interaction in Smart Home - VocADom@A4H Corpus Collection and Empirical Assessment of Its Usefulness," in *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, 2019, pp. 811–818.

[15] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 7252–7262.

[16] C.-W. Huang and Y.-N. Chen, "Learning Asr-Robust Contextualized Embeddings for Spoken Language Understanding," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8009–8013.

[17] L. Jin, T. Liu, A. Haroon, R. Stoleru, M. Middleton, Z. Zhu, and T. Chaspari, "EMSAssist: An End-to-End Mobile Voice Assistant at the Edge for Emergency Medical Services," ser. MobiSys '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 275–288.

[18] T. Tran, K. Le, N. D. Nguyen, M. Vu, H. Ngo, W. Park, and T. T. T. Nguyen, "VN-SLU: A Vietnamese Spoken Language Understanding Dataset," in *Interspeech 2024*, 2024, pp. 1335–1339.

[19] M. Gao, H. Chen, J. Du, X. Xu, H. Guo, H. Bu, J. Yang, M. Li, and C.-H. Lee, "Enhancing Voice Wake-Up for Dysarthria: Mandarin Dysarthria Speech Corpus Release and Customized System Design," in *Interspeech 2024*, 2024, pp. 2465–2469.

[20] H.-Y. Tseng, C.-Y. Lee, C.-S. Wu, I.-C. Wu, H.-Y. Chang, C.-C. Hsu, and C. A. Hsiung, "Examining the role of living alone and loneliness in predicting health-related quality of life: results from the Healthy Aging Longitudinal Study in Taiwan (HALST)," *Quality of Life Research*, vol. 33, no. 4, pp. 1015–1028, 2024.

[21] Y. Lin, W. Chen, H. Yu, and I. Cho, "Spatial Distribution, Characterization, and Policy Opportunities for Taiwan's Solo Elderly: A Big Data Approach," in *Proceedings of DRS2024: Boston*, C. Gray, E. Ciliotta Chehade, P. Hekkert, L. Forlano, P. Ciuccarelli, and P. Lloyd, Eds., Boston, USA, Jun. 2024. [Online]. Available: https://doi.org/10.21606/drs.2024.366

[22] K.-M. Chen, S.-T. Wang, S.-R. Chao, and K. Kasirisir, "Link Between Social Relationships and Vulnerability Among Community-Dwelling Older Adults Living Alone in Taiwan," *Asian Social Work and Policy Review*, vol. 19, no. 1, p. e70000, 2025, e70000 ASWP-May-2024-0051.R1. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/aswp.70000

[23] T. T. Tran, P. W. Chang, J.-M. Yang, T.-H. Chen, C.-T. Su, D. Levin-Zamir, O. Baron-Epel, E. Neter, S. F. Tsai, B. Lo, T. V. Duong, and S.-H. Yang, "Digital health literacy and its determinants among community dwelling elderly people in Taiwan," *DIGITAL HEALTH*, vol. 10, p. 20552076241278926, 2024.

[24] S.-C. Weng, W.-W. Lin, J.-L. Huang, C.-Y. Chao, C.-Y. Hsu, and S.-Y. Lin, "A smartphone model for post-acute care decreases all-cause mortality with improved left ventricular ejection fraction in patients hospitalized with heart failure in Taiwan," *Maturitas*, vol. 197, p. 108269, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0378512225000775

[25] R. Saripalle and R. Patel, "From command to care: a scoping review on utilization of smart speakers by patients and providers," *Mayo Clinic Proceedings: Digital Health*, vol. 2, no. 2, pp. 207–220, 2024.

[26] L. Z. Rubenstein, "Falls in older people: epidemiology, risk factors and strategies for prevention," *Age and ageing*, vol. 35, no. suppl_2, pp. ii37–ii41, 2006.

[27] Y.-C. Tsai, S.-Y. Chen, Y.-M. Chen, E. P.-C. Huang, and F.-P. Lu, "The Growing Burden of Fall-Related Injuries Among Older Adults: A Seven-Year Study from a Tertiary Medical Center in Taiwan," *Annals of Geriatric Medicine and Research*, 2025. [Online]. Available: http://www.e-agmr.org/journal/view.php?number=1240

[28] J. Kubitza, I. T. Schneider, and B. Reuschenbach, "Concept of the term long lie: a scoping review," *European Review of Aging and Physical Activity*, vol. 20, no. 1, p. 16, 2023.

[29] Y.-T. Tsai, C.-L. Fan, C.-L. Lo, and S.-H. Huang, "SmartLohas: A Smart Assistive System for Elder People," in *2017 14th International Symposium on Pervasive Systems, Algorithms and Networks & 2017 11th International Conference on Frontier of Computer Science and Technology & 2017 Third International Symposium of Creative Computing (ISPAN-FCST-ISCC)*, 2017.

[30] R. Chithra, R. Ackshaya, and T. Dhinakaran, "Voice Health Assistant for Elderly with Multi-Language and Offline Support," in *2025 10th International Conference on Smart Structures and Systems (ICSSS)*, 2025, pp. 1–7.

[31] M. Hamill, V. Young, J. Boger, and A. Mihailidis, "Development of an automated speech recognition interface for personal emergency response systems," *Journal of NeuroEngineering and Rehabilitation*, vol. 6, no. 1, p. 26, 2009.

[32] M. Vacher, S. Caffiau, F. Portet, B. Meillon, C. Roux, E. Elias, B. Lecouteux, and P. Chahuara, "Evaluation of a context-aware voice interface for ambient assisted living: qualitative user study vs. quantitative system evaluation," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 7, no. 2, pp. 1–36, 2015.

[33] M. Vacher, F. Aman, S. Rossato, F. Portet, and B. Lecouteux, "Making emergency calls more accessible to older adults through a hands-free speech interface in the house," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 12, no. 2, pp. 1–25, 2019.

[34] K. O'Brien, A. Liggett, V. Ramirez-Zohfeld, P. Sunkara, and L. A. Lindquist, "Voice-controlled intelligent personal assistants to support aging in place," *Journal of the American Geriatrics Society*, vol. 68, no. 1, pp. 176–179, 2020.

[35] B. Lee, I. Calapodescu, M. Gaido, M. Negri, and L. Besacier, "Speech-massive: A multilingual speech dataset for slu and beyond," in *Proc. Interspeech 2024*, 2024, pp. 817–821.

[36] S. Zhu, Z. Zhao, T. Zhao, C. Zong, and K. Yu, "Catslu: The 1st chinese audio-textual spoken language understanding challenge," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 521–525.

[37] Ì-thuân, "Suísiann dataset," https://suisiann-dataset.ithuan.tw/, 2019.

[38] Y.-F. Liao, J. S. Tsay, P. Kang, H.-L. Khoo, L.-K. Tan, L.-C. Chang, U.-G. Iunn, H.-L. Su, T.-G. Thiann, H.-K. Tiun *et al.*, "Taiwanese across taiwan corpus and its applications," in *2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2022, pp. 1–5.

[39] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the twelfth language resources and evaluation conference*, 2020, pp. 4218–4222.

[40] J. Shi, D. Berrebbi, W. Chen, E.-P. Hu, W.-P. Huang, H.-L. Chung, X. Chang, S.-W. Li, A. Mohamed, H.-y. Lee *et al.*, "Ml-superb: Multilingual speech universal performance benchmark," in *Proc. Interspeech 2023*, 2023, pp. 884–888.

[41] C.-Y. Yang, C.-C. Wang, L.-W. Chen, H.-S. Lee, H.-M. Wang, and B. Chen, "Tg-asr: Translation-guided learning with parallel gated cross attention for low-resource automatic speech recognition," *arXiv preprint arXiv:2602.22039*, 2026.

[42] Y.-H. Chou, K. Chang, M.-J. Wu, W. Ou, A. W.-H. Bi, C. Yang, B. Y. Chen, R.-W. Pai, P.-Y. Yeh, J.-P. Chiang *et al.*, "Evaluating self-supervised speech models on a taiwanese hokkien corpus," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–7.

[43] S. Arora, S. Dalmia, P. Denisov, X. Chang, Y. Ueda, Y. Peng, Y. Zhang, S. Kumar, K. Ganesan, B. Yan *et al.*, "Espnet-slu: Advancing spoken language understanding through espnet," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022.

[44] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington, "The atis spoken language systems pilot corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[45] S. Shon, A. Pasad, F. Wu, P. Brusco, Y. Artzi, K. Livescu, and K. J. Han, "Slue: New benchmark tasks for spoken language understanding evaluation on natural speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7927–7931.

[46] A. Mohamed, H.-y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe *et al.*, "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179–1210, 2022.

[47] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[48] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[49] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang *et al.*, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, 2024.

[50] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[51] X. Shi, X. Wang, Z. Guo, Y. Wang, P. Zhang, X. Zhang, Z. Guo, H. Hao, Y. Xi, B. Yang *et al.*, "Qwen3-asr technical report," *arXiv preprint arXiv:2601.21337*, 2026.

[52] M. B. Parshall, R. M. Schwartzstein, L. Adams, R. B. Banzett, H. L. Manning, J. Bourbeau, P. M. Calverley, A. G. Gift, A. Harver, S. C. Lareau, D. A. Mahler, P. M. Meek, D. E. O'Donnell, and American Thoracic Society Committee on Dyspnea, "An official American Thoracic Society statement: update on the mechanisms, assessment, and management of dyspnea," *American journal of respiratory and critical care medicine*, vol. 185, no. 4, p. 435—452, February 2012. [Online]. Available: https://europepmc.org/articles/PMC5448624

[53] G. Team, "Gemini: A Family of Highly Capable Multimodal Models," 2025. [Online]. Available: https://arxiv.org/abs/2312. 11805

[54] G. Team and D. Team, "Veo 3: Advancing High-Fidelity Generative Video Modeling," Google DeepMind, Tech. Rep., 2025, accessed: 2026-02-23. [Online]. Available: https://storage. googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf

[55] Y.-F. Liao, C.-Y. Chang, H.-K. Tiun, H.-L. Su, H.-L. Khoo, J. S. Tsay, L.-K. Tan, P. Kang, T.-g. Thiann, U.-G. Iunn *et al.*, "Formosa speech recognition challenge 2020 and taiwanese across taiwan corpus," in *2020 23rd Conference of the Oriental CO-COSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2020, pp. 65–70.

[56] P.-J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.-A. Chung, P. Tomasello, P.-A. Duquenne, H. Schwenk, H. Gong *et al.*, "Speech-to-speech translation for a real-world unwritten language," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 4969–4983.

[57] L. Ferrer and P. Riera, "Confidence intervals for evaluation in machine learning." [Online]. Available: https://github.com/ luferrer/ConfidenceIntervals

[58] A. Vyas, H.-J. Chang, C.-F. Yang, P.-Y. Huang, L. Gao, J. Richter, S. Chen, M. Le, P. Dollár, C. Feichtenhofer *et al.*, "Pushing the frontier of audiovisual perception with large-scale multimodal correspondence learning," *arXiv:2512.19687*, 2025.

[59] S. Majumdar and B. Ginsburg, "Matchboxnet: 1d time-channel separable convolutional neural network architecture for speech commands recognition," *arXiv preprint arXiv:2004.08531*, 2020.

[60] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.
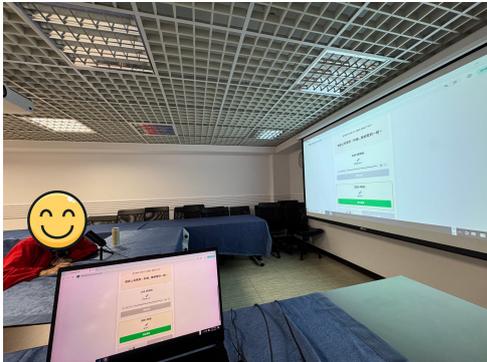
A. Supplementary Materials

A.1. Recording Environment

The recording environment is illustrated by two photographs: one showing the microphone and the other depicting the recording process. At least one researcher was on-site to monitor and control recording quality during every recording session.

(a) *Microphone setup.*

(b) *Recording process.*

Figure 7: *Illustration of the recording environment.*

A.2. Text Prompt Generation

We used the prompt from Google Gemini Pro 2.5 to generate the initial batch of instructions and scenarios. Subsequently, three researchers selected and refined the final 160 sentences. The content prompt used for this task is detailed below:

> **Task:** Generate a JSON object including 20 entries per intent for a Taiwanese spoken language understanding scenario, including intent, description, imagined scenario, and spoken content. Your task is to create a realistic, scenario-based spoken language request from the perspective of someone in need, focusing on emergency or assistance situations. Note that situations are

in-house for smart-home design. Use the provided categories and examples to guide your imagination of the situation and the corresponding vocal expression.

Select Intent includes:

1. SOS_CALL (求救)
2. FALL_HELP (跌倒／起不來)
3. BREATHING_CHEST_EMERG (呼吸／胸悶)
4. PAIN_GENERAL (一般疼痛)
5. CALL_CONTACT (打電話)
6. LIGHT_ON (開燈)
7. LIGHT_OFF (關燈)
8. CANCEL_ALERT (解除警報)

Steps:

1. **Select Intent:** Choose the intent that best matches the emergency or need scenario from the provided list (e.g., SOS_CALL, FALL_HELP).
2. **Describe the Situation:** Write a clear and concise description of the situation causing the need for assistance.
3. **Imagine the Scenario:** Utilize the categories of emotions and physical states to elaborate on the imagined emotional or physical scenario.
4. **Simulate Spoken Content:** Imagine what the person might urgently say aloud to express their need for help accurately and realistically.
5. **Assemble JSON Output:** Combine all information into the specified JSON format.

Output Format & Examples:

```
Output the information as a JSON
object with the following fields,
20 entries per intent (160 in total):
- intent:  The selected intent
indicating the type of situation.
- description:  A brief description
of the situation.
- imagined_scenario:  A narrative of
the scenario including emotional or
physical states.
- spoken_content:  A realistic
utterance that one might say in
this scenario.

Example:
{
  "index":  "0001",
  "intent":  "SOS_CALL",
  "description":  "The house is on
fire (房子失火了)",
  "imagined_scenario":  "Facing an
immediate life threat, such as
being at a fire scene, extreme
panic/terrified (面臨立即的生命威
脅，如火災現場，極度恐慌/魂飛魄散)",
  "spoken_content":  "Help!  Come
save me!  Help!  (救命！快來救我！幫忙
啊！)"
}
```

A.3. Scenario Video Generation

We utilized Google Veo 3 to generate scenario videos to help subjects imagine and be reminded of the specific contexts. Below are the prompts used for generating some of these videos, detailed with side-by-side English and Chinese translations for clarity:

Table 10: *Scenario Video Generation Prompts (Part 1/3).*

Intent & Component	English Prompt	Chinese Translation (中文翻譯)
1. SOS_CALL (求救)		
Video Description (8 seconds)	An elderly Taiwanese man living alone in social housing in Taiwan sits on the living room sofa. Suddenly, his face turns pale, he clutches his chest tightly, and curls up his body. He opens his mouth to gasp rapidly, his eyes full of extreme panic. He struggles to reach out, trying to grab the phone on the table, but lacks strength throughout his body. Finally, his hand drops weakly, and he looks desperately towards the door.	一位在台灣社會住宅的獨居台灣阿公坐在客廳沙發上。突然他臉色慘白，手緊緊揪住胸口，身體蜷縮起來。他張開嘴急促地喘息，眼神中充滿了極度的慌張。他掙扎著伸出手，試圖去拿桌上的手機，但全身無力。最終虛弱地垂下手，絕望地望向門口的方向。
Visual Style	Realistic, with a strong sense of tension and urgency. The camera can shake slightly to simulate the viewer's nervous heartbeat, with cold color tones.	寫實、帶有強烈緊張與急迫感，鏡頭可以微微顫抖，模擬觀看者的緊張心跳，色調偏冷。
Audio	The background only has a faint television sound, quickly covered by rapid, heavy heartbeats and high-frequency tense string music. The sound fluctuates, creating a suffocating sense of life hanging by a thread.	背景只有微弱的電視聲，很快被一陣急促、沉重的心跳聲和高頻的緊張弦樂覆蓋。聲音時強時弱，營造出命懸一線的窒息感。
2. FALL_HELP (跌倒／起不來)		
Video Description (8 seconds)	An elderly Taiwanese person living alone in social housing in Taiwan, wearing loungewear, tries to sit down but accidentally slips from the chair and falls heavily onto the floor. She shows a pained expression, struggles to prop herself up with both hands, and tries twice to stand up. However, her legs are visibly weak due to lack of exercise, and she eventually sits paralyzed and helpless on the spot, her eyes full of frustration and a longing for help.	一位住在台灣社會住宅穿著居家服的獨居台灣長者，想要坐下，不小心從椅子上滑下並重摔在地上。她臉上露出痛苦的表情，掙扎著用雙手撐地，嘗試了兩次想要站起來。但雙腿因缺乏運動明顯無力，最終只能無助地癱坐在原地，眼神充滿了挫敗與求助的渴望。
Visual Style	Realistic, warm home color tones. The focus is on capturing the elder's strenuous but failed attempts to stand, and the subtle facial expression changes from pain to helplessness.	寫實、溫馨的家居色調。重點捕捉長者嘗試站起卻失敗的吃力動作，以及臉上從痛苦轉為無助的細微表情變化。
Audio	Initially, quiet environmental sounds, accompanied by a heavy thud and the sound of clothing friction upon falling. After that, there is only the elder's heavy breathing from exertion and faint Taiwanese/Chinese moaning, with no background music, highlighting her isolated and helpless predicament.	起初是安靜的環境音，在跌倒時伴隨一聲沉重的悶響和衣物摩擦聲。之後只有長者因吃力而發出的沉重呼吸聲和細微的台灣中文呻吟，沒有背景音樂，凸顯其孤立無援的困境。
3. BREATHING_CHEST_EMERG (呼吸／胸悶)		
Video Description (8 seconds)	A close-up shot of the face and upper body of an elderly Taiwanese woman living alone in social housing in Taiwan. She sits quietly, then suddenly furrows her brows and presses her palm firmly against her chest. Her shoulders heave violently, her mouth opens wide, and she gasps for air laboriously like a fish out of water. Her face flushes red from hypoxia, and her eyes are filled with pain and panic.	近距離特寫一位在台灣社會住宅的獨居阿嬤的臉部和上半身。她靜靜坐著，突然眉頭緊鎖，手掌用力按在胸口上。她的肩膀劇烈起伏，嘴巴張得很大，像離開水的魚一樣費力地喘氣。臉上因缺氧而漲紅，眼神中滿是痛苦和恐慌。
Visual Style	Realistic, with a sense of oppression. Uses a shallow depth of field lens to blur the background, focusing entirely on the protagonist's pained expression and rapid movements.	寫實、具有壓迫感。使用淺景深鏡頭，讓背景模糊，所有焦點都集中在主角痛苦的表情和急促的動作上。
Audio	Absolutely no background music. Only amplified, sharp, and labored gasps, along with hissing sounds from the throat, creating an extremely uncomfortable suffocating sensation that directly impacts the viewer's hearing.	完全沒有背景音樂。只有被放大的、尖銳而費力的喘息聲、喉嚨間的嘶嘶聲，營造出令人極度不適的窒息感，直接衝擊觀眾的聽覺。

Table 11: *Scenario Video Generation Prompts (Part 2/3).*

Intent & Component	English Prompt	Chinese Translation (中文翻譯)
4. PAIN_GENERAL (一般疼痛)		
Video Description (8 seconds)	An elderly Taiwanese man living alone in social housing in Taiwan is sitting in the living room watching television. He suddenly stops his actions, closes his eyes tightly, and presses one hand forcefully against his temple. His facial muscles are tense, showing that he is enduring a severe headache. He slowly puts down his book, supports his forehead with his other hand as well, and leans slightly forward.	一位居在台灣社會住宅的獨居阿公正坐在客廳看電視機。他突然停下動作,緊閉雙眼,一隻手用力地按壓著自己的太陽穴。他的臉部肌肉緊繃,顯現出正在忍受劇烈頭痛的樣子。他緩慢地放下書,另一隻手也扶住額頭,身體微微前傾。
Visual Style	Realistic, everyday life. Soft lighting, capturing the tense facial details of the protagonist caused by pain through a close-up shot.	寫實、生活化。光線柔和,透過特寫鏡頭捕捉主角因疼痛而緊繃的臉部細節。
Audio	The background consists of quiet room environmental sounds (e.g., a ticking clock). When the protagonist feels pain, a very faint but piercing high-frequency tinnitus sound effect is added to simulate the feeling of a headache, which then slowly fades.	背景是安靜的房間環境音(如時鐘滴答聲)。當主角感到疼痛時,加入一陣非常細微但刺耳的高頻耳鳴音效,模擬頭痛時的感受,然後慢慢減弱。
5. CALL_CONTACT (打電話)		
Video Description (8 seconds)	An elderly Taiwanese woman living alone in social housing in Taiwan, wearing reading glasses, smiles as she flips through an old photo album. She stops at a family portrait, gently stroking the child's face in the photo with her finger, her eyes filled with longing. Then, she looks up, turns her gaze to the smart speaker in the room, moves her lips slightly, and wears an expectant expression on her face.	一位在台灣社會住宅獨居台灣阿嬤戴著老花眼鏡,微笑著翻閱一本舊相簿。她停在一張全家福照片上,用手指溫柔地撫摸著照片中孩子的臉龐,眼神充滿思念。接著,她抬起頭,將目光轉向房間裡的智慧音箱,嘴唇微動,臉上帶著期待的表情。
Visual Style	Warm, bright, and soft, with warm yellow color tones. Shot with soft lighting to create a warm, nostalgic family atmosphere.	溫馨、明亮且柔和,色調偏暖黃。用柔光拍攝,營造出溫暖、懷舊的親情氛圍。
Audio	Gentle, soothing piano or guitar background music. Accompanied by the paper friction sound of flipping through the photo album.	輕柔、舒緩的鋼琴或吉他背景音樂。伴隨著翻閱相簿的紙張摩擦聲。
6. LIGHT_ON (開燈)		
Video Description (8 seconds)	In the evening, the lighting in the room is dim. An elderly Taiwanese person living alone in social housing in Taiwan squints, leans forward, and struggles immensely to read the small print on a pillbox. He tries several times but still cannot see clearly; he shakes his head in frustration, then looks up at the location of the ceiling light and mouths words as if preparing to speak.	傍晚,房間內光線昏暗。一位在台灣社會住宅的台灣獨居長者眯著眼睛,身體前傾,非常吃力地辨認藥盒上的小字。他試了幾次都看不清楚,困擾地搖了搖頭,然後抬頭望向天花板上電燈的位置,做出一個準備說話的口型。
Visual Style	Realistic, starting from dim natural light; the visual graininess can be slightly heavier to emphasize the lack of light.	寫實,從昏暗的自然光開始,畫面顆粒感可以稍微重一點,以強調光線不足。
Audio	Only very quiet environmental sounds. There can be the slight friction sound of the plastic pillbox, focusing on creating a quiet atmosphere of inconvenience caused by dimness.	只有非常安靜的環境音。可以有藥盒塑膠的輕微摩擦聲,重點是營造一個因昏暗而不便的安靜情境。

Table 12: *Scenario Video Generation Prompts (Part 3/3).*

Intent & Component	English Prompt	Chinese Translation (中文翻譯)
7. LIGHT_OFF (關燈)		
Video Description (8 seconds)	At night, an elderly Taiwanese person living alone in social housing in Taiwan is already lying comfortably in bed, covered with a blanket. He lets out a big yawn, rubs his eyes, and then, with tired eyes, looks up at the ceiling light that is still on, showing obvious sleepiness in his eyes.	夜晚，一位在台灣社會住宅的台灣獨居長者已經舒適地躺在床上，蓋好了棉被。他打了個大大的哈欠，揉了揉眼睛，然後睜著疲憊的雙眼，看著天花板上仍然亮著的燈，眼中流露出明顯的睡意。
Visual Style	Realistic and tranquil. The room has warm lighting, focusing on capturing the protagonist's natural, sleepy actions like yawning and rubbing eyes.	寫實、寧靜。房間內是溫暖的燈光，重點捕捉主角打哈欠、揉眼睛等充滿睡意的自然動作。
Audio	Only faint environmental sounds or distant sounds of insects chirping. The sounds of yawning and blanket friction can be heard clearly, creating a quiet pre-sleep atmosphere.	只有微弱的環境音，或遠處傳來的蟲鳴聲。可以清晰地聽到哈欠聲和棉被摩擦的聲音，營造安靜的睡前氛圍。
8. CANCEL_ALERT (解除警報)		
Video Description (8 seconds)	In the kitchen, the fire detector on the wall is flashing red and emitting a continuous "beep-beep" alarm sound. An elderly Taiwanese woman living alone in social housing in Taiwan calmly takes a towel and fans the detector, wearing a slightly helpless "not again" smile. She casually stands on her tiptoes, presses the button on the detector, and the alarm stops immediately.	廚房裡，牆上的防火災偵測器正閃爍著紅光，並發出持續的「嗶—嗶—」警報聲。一位在台灣社會住宅的台灣獨居阿嬤從容地拿著一條毛巾，對著偵測器搧風，臉上帶著一點「又來了」的無奈笑容。她輕鬆地踮起腳尖，按下了偵測器上的按鈕，警報隨即停止。
Visual Style	Realistic, everyday life. The focus is on the protagonist's calm, accustomed reaction, which sharply contrasts with the piercing alarm sound and flashing red light.	寫實、生活化。重點在於主角鎮定、習以為常的反應，與刺耳的警報聲和閃爍的紅光形成鮮明對比。
Audio	The first half features loud, piercing, and repetitive electronic alarm sound effects. At the exact moment the button is pressed, the alarm stops abruptly, and the world returns to quiet, leaving only the normal operating sound of the range hood and the sizzling sound of something frying in the pan.	前半段是響亮、刺耳、重複的電子警報音效。在按鈕被按下的瞬間，警報聲戛然而止，世界恢復安靜，只剩下抽油煙機的正常運轉聲和鍋裡煎東西的滋滋聲。