FOLEYSET: A MULTI-LEVEL HUMAN-ANNOTATED FOLEY SOUND DATASET

Sunshiyu Wang and Alexander Lerch

Music Informatics Group
Georgia Institute of Technology
Atlanta, USA
swang3214@gatech.edu, alexander.lerch@gatech.edu

ABSTRACT

In audiovisual post-production, Foley refers to synchronous sound effects associated with human actions, such as footsteps, cloth rustle, and prop handling, that are recreated to match the on-screen movements and interactions of characters. These sounds are often recorded by professional Foley artists using physical props. This resource-intensive workflow has motivated data-driven research on Foley, including tasks such as classification, retrieval, and generation; however, high-quality annotated Foley datasets for training remain scarce. To address this gap, we present FoleySet, a publicly available Foley dataset of 10,000 audio clips annotated with a two-level Foley taxonomy. This dataset provides a standardized, Creative Commons–licensed resource for data-driven Foley classification, retrieval, and generation.

1. INTRODUCTION

Foley refers to the synchronized reproduction of everyday sound effects for audiovisual media. Examples include footsteps, cloth movement, and object interactions such as doors opening or coins jingling. Unlike purely synthetic sound effects, Foley is often created and recorded in studio settings. In this process, Foley artists use a variety of props to recreate the sounds of real-world actions and material interactions in sync with the visual scene. This makes the production process labor-intensive, costly, and time-consuming. Alternatively, existing Foley sound libraries provide pre-recorded Foley sounds, but require significant time for both finding a fitting sound and manually synchronizing it to the on-screen events. Recent advances in deep-learning-based audio methods have opened up new data-driven directions for automating and supporting Foley-related workflows. However, compared with broader audio research areas, Foley-specific data-driven work remains constrained by the lack of structured, high-quality datasets. Such a dataset is therefore needed to support not only analyzing and synthesizing Foley but also a broader range of data-driven approaches to modern Foley production. Accordingly, we introduce *FoleySet*, a publicly available dataset of 10,000 human-annotated Foley clips organized using a two-level taxonomy.[1] Building on prior research and to clearly define the scope of the dataset, we define Foley as sounds arising from human-related actions—human–material interactions and human-produced sounds—rather than non-human sources such as animal vocalizations.

[1]FoleySet is publicly available at https://zenodo.org/records/20735877.

Several studies have explored sound effect classification, taxonomy construction, and library indexing for film post-production [1, 2, 3, 4]. On the synthesis side, existing deep-learning research on Foley has primarily focused on optimizing model architectures and improving synthesis quality. For example, the DCASE 2023 Task 7 organized a Foley sound synthesis challenge that provided a standardized evaluation framework, baseline systems, and a dataset spanning seven Foley sound classes [5]. Subsequent Foley generation systems, such as T-Foley and MambaFoley, built on the same dataset and further improved performance through advances in model architecture [6, 7]. Foley's strong relation to visual media has also resulted in cross-modal synthesis research, leading to Video-to-Audio models such as AutoFoley and Diff-Foley [8, 9].

Together, these works demonstrate the growing technical progress toward data-driven Foley workflows, while also highlighting the need for Foley-specific datasets and taxonomies. Several public audio datasets contain relevant sound-effect categories, but their taxonomies are typically organized around broad semantic classes rather than the action- and material-level distinctions central to Foley practice [10]. As a result, they do not provide a unified Foley-specific hierarchy or fine-grained annotations for distinguishing subtle differences in action, material, and synchronization context. This mismatch motivates the taxonomy and annotation design of FoleySet.

More specifically, *FoleySet* contains 10,000 audio clips collected from Freesound and annotated using a two-level taxonomy comprising 9 major categories and 73 sub-categories. Alongside each audio file, we collected user-provided metadata from the corresponding Freesound webpage, including tags and textual descriptions [11]. All audio clips were normalized, trimmed to remove leading silence, and segmented into multiple shorter clips when longer than 5 s. Each clip was also assigned a one-shot/multi-shot label to indicate whether it contains a single sound event or repeated occurrences. The released dataset includes only clips distributed under Creative Commons Zero licenses that permit reuse and redistribution. Taken together, FoleySet provides a standardized and reproducible resource for fine-grained Foley research.

The remainder of this paper is structured as follows. Section 2 reviews related audio datasets. Section 3 describes the development of our Foley-specific taxonomy. Section 4 details the dataset construction pipeline. Section 5 presents dataset statistics. Section 6 reports benchmark classification results, and Section 7 concludes the paper.

2. EXISTING DATASETS

This section reviews existing audio datasets that are relevant to FoleySet in terms of content, collection process, or labeling design. Their main characteristics are summarized in Table 1.

Table 1: Comparison of the proposed Foley dataset with existing popular audio datasets.

Dataset	Source	Total Dur. (hrs)	Samples	Classes	Annotation Source
ESC-50	Freesound	2.8	2,000	50	Human annotation
UrbanSound8K	Freesound	8.8	8,732	10	Human annotation
FSD50K	Freesound	108.0	51,197	200	Tags + human validation
AudioSet	YouTube	5,833	2,084,320	527	Human-verified labels
DCASE2023 Task 7	US8K / FSD50K / BBC SFX	6.2	5,550	7	Task-specific labels
FoleyBench	Internet video (CC)	$\sim$13	5,000	UCS tax.	Automated pipeline
MINT	AudioCaps	111.2	40,016	4 (11 sub)	Captions + narrative text
6KSFx	Synthetic	8.3	6,000	30	Procedurally generated
FoleySet (ours)	Freesound	9.5	10,000	9 (73 sub)	Human annotation

A closely related open Foley-focused dataset was released for the DCASE 2023 Task 7 "Foley Sound Synthesis" challenge[2]. The dataset contains 5,550 audio excerpts (4,850 for development, 700 for evaluation) drawn from UrbanSound8K[12], FSD50K [10], and the BBC Sound Effects library [13], and is organized into seven categories: `DogBark`, `Footstep`, `GunShot`, `Keyboard`, `MovingMotorVehicle`, `Rain`, and `Sneeze/Cough`. For the challenge, all audio was converted to mono signals with 16-bit and 22,050 Hz, and zero-padded or segmented to a fixed duration of 4 s. The dataset has played an important role in supporting recent research on Foley sound synthesis. However, its broader applicability is limited by its seven-class structure, which does not capture the diversity and granularity needed for fine-grained Foley tasks. Other Foley-related resources address complementary tasks. FoleyBench [14] provides video–audio–caption triplets for evaluating semantic and temporal alignment in video-to-audio generation, MINT [15] focuses on image- and narrative-text–driven Foley audio planning and dubbing, and 6KSFx [16] provides synthetically generated sound effects for procedural-audio research. These datasets differ from FoleySet in modality, source material, annotation design, and primary task.

Beyond the Foley-specific datasets described above, there exist a number of datasets that contain Foley-related classes among others. AudioSet is one of the largest general-purpose audio datasets and among the first to explicitly target broad coverage of everyday sound [17]. It consists of about 2.1 million 10 s clips drawn from YouTube.com and spans 527 classes covering speech, music, and sound events. Its labels are organized using a large-scale hierarchical audio event ontology, and each clip can receive one or more labels that are validated by multiple human raters through majority voting. The underlying ontology of AudioSet comprises 632 categories organized in a hierarchy up to six levels deep. Several top-level branches — notably `Human sounds` and `Sounds of things` — contain classes that overlap with common Foley categories, such as `Footstep`, `Door`, `Glass`, `Hands`, and `Breathing`. However, these Foley-relevant classes are scattered across disparate branches of the ontology rather than consolidated under a unified Foley category, limiting its direct use as a standardized benchmark for fine-grained Foley research. Overall, AudioSet has been widely adopted for tasks such as audio representation learning, and has strongly influenced subsequent audio dataset development. That said, because the original audio is sourced from YouTube, AudioSet is not released as a fully open collection of redistributable audio files. Access to the underlying audio depends on third-party hosting, and clips may become unavailable over time due to deletion, licensing changes, or copyright-related restrictions, which reduces the long-term stability of the dataset.

To address some of AudioSet's limitations, FSD50K was introduced as a sound event dataset based on Freesound audio with distributable waveform files. It contains 51,197 audio clips totaling 108.3 hours of multi-label audio, and is manually annotated using 200 classes drawn from the AudioSet Ontology. These 200 classes are hierarchically organized as a subset of the ontology, including 144 leaf nodes and 56 intermediate nodes. The class selection was driven primarily by the availability of suitable audio material on Freesound, with a focus on sound events produced by physical sound sources and production mechanisms. FSD50K includes classes that overlap with common Foley categories — such as *Footsteps*, *Door*, *Knock*, and *Glass* — but, as it adopts the AudioSet ontology, these classes are not organized under a unified Foley-specific taxonomy. Compared with AudioSet, FSD50K provides a fully open and redistributable benchmark for sound event research, with annotations obtained through a hybrid pipeline in which candidate labels are first inferred from Freesound tags and then confirmed through human validation.

Two other datasets, ESC-50 [18] and UrbanSound8K [12], are also sourced from Freesound, similar to FSD50K. Both focus on more specific sound domains, which aligns with the goal of our work. ESC-50 is a fully human-labeled environmental sound dataset containing 2,000 5 s clips evenly balanced across 50 classes. These classes are organized into five broad categories: `animal sounds`, `natural soundscapes`, `human non-speech sounds`, `interior/ domestic sounds`, and `exterior/ urban noises`. The authors state that the 50 classes were manually selected to provide balanced coverage of major types of environmental sounds. They also note that the dataset design took into account the availability and diversity of Freesound recordings, as well as the usefulness and distinctiveness of each class. Several ESC-50 classes overlap with Foley-related sounds, particularly within the `human non/speech sounds` and `interior/ domestic sounds` categories. However, as an environmental sound dataset, ESC-50 does not ensure that target sounds are isolated from background noise, nor does it indicate whether an event is one-shot or multi-shot, which is often important for Foley sounds.

UrbanSound8K is another dataset sourced from Freesound, designed specifically for urban sound classification and built upon a self-developed taxonomy tailored to this domain. To construct the labeling system, the authors first proposed an urban sound taxonomy consisting of four top-level groups —`human`, `nature`,

[2] `https://dcase.community/challenge2023/task-foley-sound-synthesis`

Figure 1: Overview of the FoleySet construction pipeline.

`mechanical`, and `music`— along with a set of fine-grained and unambiguous leaf categories. This taxonomy was informed in part by sound sources frequently appearing in NYC 311 noise-complaint records. From this broader taxonomy, 10 classes were selected for UrbanSound8K: `air conditioner`, `car horn`, `children playing`, `dog bark`, `drilling`, `engine idling`, `gun shot`, `jackhammer`, `siren`, and `street music`. The resulting dataset contains 8,732 clips, each with a maximum duration of 4 s, for a total of approximately 8.75 hours of audio. To reduce class imbalance, no class contains more than 1,000 clips. UrbanSound8K demonstrates the value of combining domain-specific taxonomy design with publicly sourced audio data, which provide a useful reference for our Foley taxonomy design. Unlike FoleySet, UrbanSound8K targets urban environmental sound recognition and includes only a small number of categories directly related to human-performed Foley events.

Taken together, these prior datasets demonstrate both the usefulness and value of large-scale, broad-coverage audio datasets with weak or machine-assisted labels, useful for training a wide range of audio-related models, and the importance of domain-specific audio collections with carefully designed human annotations. Similar to ESC-50 and UrbanSound8K, the presented dataset FoleySet focuses on a specific sound domain rather than broad audio coverage. This narrow scope enables fine-grained and precise human labeling, targeting Foley-related research tasks that are not well served by large-scale weakly labeled datasets.

3. TAXONOMY

Foley is a commonly used term in audiovisual sound post-production; however, there is no universally accepted definition of its scope, and its boundary with the broader domain of sound effects remains blurred. Existing audio taxonomies therefore show overlaps and inconsistencies in how Foley and other sound categories are labeled. For example, the widely adopted Universal Category System (UCS)[19] includes a Foley category with five subcategories (`Cloth`, `Feet`, `Hands`, `Misc`, and `Props`). However, many signature Foley sounds discussed in Foley-focused research and practice, such as punches, breathing, kissing, or door open/close, are placed under other primary UCS categories [20]. Thus, UCS provides a useful general sound-effects taxonomy but does not offer consistent or exhaustive coverage of Foley sounds.

Given the lack of established standards for structuring and annotating Foley sounds, we propose an operational and reproducible Foley taxonomy and describe its development process in this section. In a first step, we investigate the core characteristics of Foley based on prior research, artist interviews, industry reports, and books. Drawing on these sources [20, 21, 22], we summarize several recurring characteristics of Foley sounds:

(i) Foley is a post-production practice of creating sound effects in synchrony with on-screen actions,
(ii) Foley chiefly relates to the imitation of human-related actions,
(iii) Foley commonly focuses on body-associated events and human interaction with surfaces and objects, and
(iv) Foley often supplies the subtle sonic details audiences expect to accompany visible actions on screen.

Based on these characteristics, we define Foley as sounds arising from human-related actions, including human-driven interactions with materials (e.g., glass clinks or metal impacts) as well as human-produced sounds (e.g., kissing or snoring). For the purposes of FoleySet, we therefore exclude non-human sources such as animal vocalizations (e.g., dog barks or wolf howls), while acknowledging that Foley may be used more broadly in professional production. This working definition provides the conceptual basis for the taxonomy construction process described below.

To ensure the practical relevance of our taxonomy, we draw from reference data collected from commercial Foley sound libraries. We analyzed the taxonomy structures and category distributions of seven commercial Foley libraries [23, 24, 25, 26, 27, 28, 29] and extracted Foley-relevant keywords from each library's metadata: category and sub-category labels, clip names, and descriptions. This allowed us to identify terms commonly used in commercial practice. These common keywords were then normalized into a unified vocabulary by resolving spelling variants and merging synonymous terms, yielding an initial pool of 315 candidate keywords. A subsequent manual review removed terms lacking clear sound or action references, and keywords otherwise falling outside our working definition of Foley. The remaining 119 keywords served as the primary inputs for taxonomy construction. Based on these keywords, we manually refined and consolidated them into 73 sub-categories which are organized into 9 major categories. The resulting two-level taxonomy, as well as the mapping from keyword to sub-category, is summarized in the Appendix in Table **??**.

4. DATASET CONSTRUCTION

FoleySet is constructed through a multi-step pipeline designed to ensure that each audio clip is precisely labeled according to our proposed taxonomy while maintaining high audio quality and a balanced category distribution. All category and subcategory labels were assigned by a single annotator following the predefined taxonomy and annotation criteria. The construction pipeline is shown in Fig. 1.

At the first stage, we collected approximately 23,500 candidate Foley clips from Freesound.org by querying the platform with sub-category names and corresponding keyword-based tags (accessed September 2025). We restricted our search to Creative Commons Zero (CC0)-licensed WAV clips with sampling rates of 44.1 kHz, 48 kHz, or 96 kHz. This restriction allows unrestricted redistribution, while the resulting dataset may reflect the sounds and recording conditions more commonly found in CC0 uploads on Freesound. In addition, because Freesound does not formally

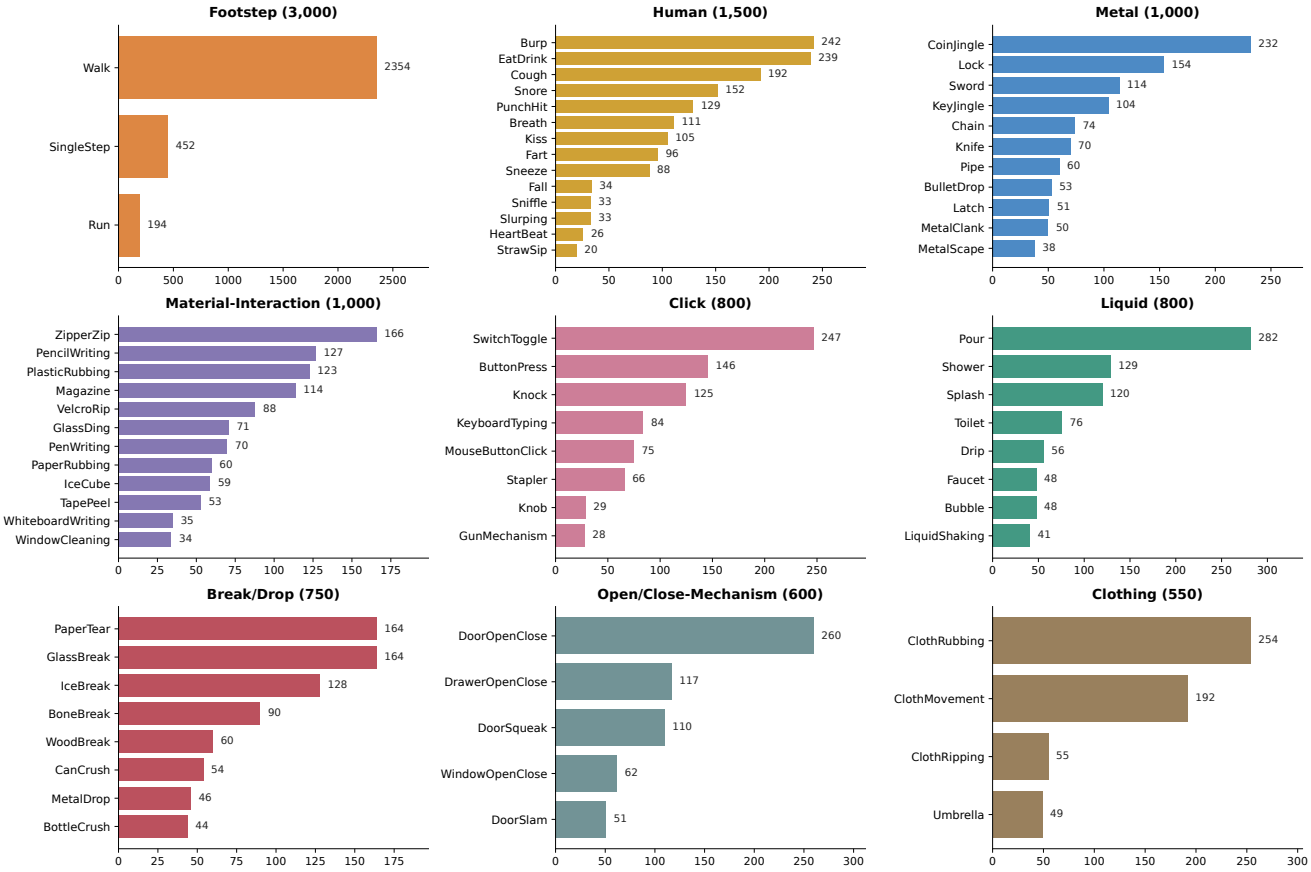

Figure 2: Sub-category distribution within each major-category, sorted by clip count. The number of total clips per category is indicated in parentheses.

review the audio quality of uploaded content and its user-provided tags and descriptions may be subjective or inconsistent, Stage 2 involved an initial manual screening to remove corrupted files, silent or excessively noisy clips, and recordings unrelated to Foley categories (e.g., guitar loops or synth drones). After this step, approximately 18,000 clips remained.

The remaining audio clips were resampled to 44.1 kHz in Stage 3, converted to 16-bit mono, and loudness-normalized to −23 LUFS [30] to ensure comparable levels and consistent format across the dataset. Each file was trimmed to start 100 ms before the onset of the first audible event, providing a clean lead-in while removing unnecessary silence. To keep clip durations comparable, we enforced a maximum clip length of 5 s: recordings longer than this threshold were segmented into multiple clips by cutting at the last silence point before the 5 s limit, with each segment inheriting the original metadata. After segmentation, the pool contained approximately 15,000 clips.

In Stage 4, the remaining clips were manually reviewed and annotated. We set the following target clip counts per major category: Footstep (3,000), Human (1,500), Metal (1,000), Material-Interaction (1,000), Click (800), Liquid (800), Break/Drop (750), Open/Close-Mechanism (600), and Clothing (550). These targets reflect both the relative importance of each category in Foley practice and the availability of suitable source material on Freesound. All targets were met. Candidate clips were manually reviewed and assigned both a major-category and a sub-category label until the target count for each category was reached. In addition, each clip was assigned a one-shot/multi-shot tag to indicate whether it contained a single sound event or multiple sound events. We also retained the original user-provided Freesound tags and textual descriptions as additional metadata for each clip. The raw tags were lightly cleaned through a combination of automatic and manual steps: all tags were lowercased, single-character entries and tags containing digits (e.g., `ntg4`, `zoom-h2`) were removed, and space-separated tag strings were split into individual tokens. Morphological variants were merged into canonical forms (e.g., `footsteps` → `footstep`, `walking` → `walk`, `clothing` → `cloth`), and duplicate tags within the same clip were removed.

Finally, in Stage 5, we shuffled the clip order and renamed all files sequentially (e.g., `00001.wav` to `10000.wav`). The dataset is split into 8,000 training, 1,000 validation, and 1,000 test samples (80/10/10). To prevent data leakage, all segments derived from the same original Freesound recording are assigned to the same partition. Under this constraint, we optimized the split assignment so that the subcategory distribution across partitions matches the overall distribution as closely as possible.

5. DATASET STATISTICS

The released dataset contains 10,000 audio clips (44.1 kHz, 16 bit mono, max. length 5 s) totaling approximately 9.5 hours of audio.

Table 2: Overview of the fields provided in the released dataset metadata file.

Field	Description (example)
name	Clip filename in the released split (e.g., 09001.wav).
split	Data split indicator (e.g., train/val/test).
category	Major-category label (9-way; e.g., Liquid).
sub-category	Sub-category label (73-way; e.g., Pour).
Oneshot-Multi	Event density tag (e.g., One-shot vs. Multi-shot).
freesound-id	Source Freesound ID for traceability
url	Link to the original Freesound page.
username	Freesound uploader username for provenance (e.g., Rico_Casazza).
original-tags	Original user-provided Freesound tags.
description	Original Freesound text description.

Clip durations range from 0.3 s to 5 s, with a mean of 3.4 s and a median of 4.3 s. The 10,000 clips originate from 5,739 unique Freesound recordings (1.7 clips per source on average), with 4,128 sources contributing only a single clip. The dataset spans 9 major categories and 73 subcategories. At the major-category level, class sizes range from 550 (Clothing) to 3,000 (Footstep), with an imbalance ratio of approximately 5:1. At the subcategory level, the distribution exhibits a more pronounced long tail: class sizes range from 20 (StrawSip) to 2,354 (Walk), with a mean of 137 and a median of 84 clips per class. Of the 73 subcategories, 16 contain fewer than 50 clips and 41 contain fewer than 100 clips. While these statistics describe class coverage and imbalance, intra-class variation was not explicitly quantified in the current dataset. The original Freesound user-provided tags yield a vocabulary of 4,035 unique tokens across the dataset. In addition, each clip is annotated with a one-shot or multi-shot tag indicating whether it contains a single or multiple sound events; 23.6% of clips are labeled One-shot and 76.4% Multi-shot. The full category and subcategory distributions are shown in Fig. 2.

Each clip's metadata is recorded as a single row in one CSV file containing all 10,000 entries. (Table 2). The file includes the major-category, sub-category, a one-shot/multi-shot event tag, and the original Freesound metadata: user-provided tags, textual description, uploader username, source ID, and URL. The user-provided tags were lightly cleaned and normalized by us; the 50 most frequent tags are shown in Table 3. These fields are intended to support not only supervised classification but also tasks such as text-based audio retrieval, captioning, and tag-based analysis.

6. BENCHMARK RESULTS

We provide benchmark results for two classification tasks:
(i) **Task 1**: 9-class major-category classification and
(ii) **Task 2**: 73-class sub-category classification.
For both tasks, we use PaSST (Patchout Spectrogram Transformer)

Table 3: Top 50 original-tags in FoleySet, ranked by frequency.

#	Tag	Count	#	Tag	Count
1	footstep	2903	26	leather	404
2	walk	2576	27	sneakers	402
3	foley	1949	28	glass	400
4	step	1862	29	soft	398
5	sound	1005	30	click	391
6	running	982	31	movement	386
7	foot	933	32	human	378
8	water	793	33	wooden	364
9	feet	774	34	hit	344
10	wood	738	35	ice	336
11	metal	676	36	forest	335
12	paper	655	37	male	328
13	open	635	38	slow	325
14	door	630	39	slide	323
15	gravel	591	40	writing	310
16	floor	590	41	rustling	307
17	cloth	573	42	concrete	306
18	close	569	43	switch	303
19	shoes	554	44	shoe	301
20	boots	495	45	break	301
21	run	477	46	snow	293
22	dirt	448	47	liquid	289
23	ground	434	48	grass	282
24	crunch	429	49	natural	279
25	fabric	410	50	man	278

[31] as a pre-trained feature extractor, specifically the `hear21passt` package [32], which provides AudioSet-pretrained checkpoints following the HEAR 2021 NeurIPS Challenge API. For the feature extraction, all audio is resampled to 32 kHz and zero-padded to a fixed length of 5 s when shorter than 5 s. The frozen PaSST backbone provides 768-dimensional clip-level embeddings by averaging timestamp-level representations. These embeddings are the input to a linear probe classifier (9-way or 73-way) with class-weighted cross-entropy loss using the AdamW optimizer; each class weight is set to $N/(K \cdot n_k)$, with N the total number of training samples,

Table 4: Per-class results for 9-way major-category classification on the FoleySet test set (N=1000). P = Precision, R = Recall, Sup = Support (number of test samples).

Class	P	R	F1	Sup
Brk/Drp	0.78	0.75	0.77	77
Click	0.82	0.86	0.84	79
Clothing	0.56	0.85	0.68	55
Footstep	0.95	0.86	0.90	296
Human	0.87	0.80	0.83	151
Liquid	0.93	0.90	0.92	79
Mat-Int	0.66	0.63	0.65	101
Metal	0.78	0.87	0.82	103
Op/Cl-Mech	0.80	0.86	0.83	59

Brk/Drp = Break/Drop; Mat-Int = Material-Interaction;
Op/Cl-Mech = Open/Close-Mechanism; Wtd = Weighted.

Table 5: Overall classification results on the Foley test set ($N = 1000$). Acc = Accuracy, P = Precision, R = Recall, F1 = F1-score, and w = weighted average.

	9-way major-cat.	73-way sub-cat.
Acc	0.82	0.64
$\mathbf{P}_{\text{macro}}$	0.79	0.60
$\mathbf{R}_{\text{macro}}$	0.82	0.59
$\mathbf{F1}_{\text{macro}}$	0.80	0.56
$\mathbf{P}_{\text{w}}$	0.84	0.71
$\mathbf{R}_{\text{w}}$	0.82	0.64
$\mathbf{F1}_{\text{w}}$	0.83	0.64

K the number of classes, and n_k the count of class k. We monitor validation accuracy for early stopping (patience: 10 epochs) and select the best checkpoint based on validation performance. The model is then evaluated on the held-out test set.

We evaluate performance using overall accuracy (micro-avg) and F1 scores (macro-averaged and weighted). Both tasks are evaluated on the same held-out test set ($N = 1000$). Overall results are listed in Table 5, with per major-class breakdowns in Table 4 and the corresponding confusion matrix in Fig. 3 and Fig. 4.

6.1. Major-Category Classification

On the 9-way major-category test set, the baseline achieves overall 0.82 accuracy, 0.80 macro-F1, and 0.83 weighted-F1. Per-class results show strong performance for categories with distinctive acoustic signatures (e.g., `Footstep`: F1=0.90; `Liquid`: F1=0.92), while comparatively lower F1 scores are observed for `Material-Interaction` (F1=0.65) and `Clothing` (F1=0.68), suggesting that these categories may be more difficult to classify because they are acoustically less homogeneous and exhibit broader intra-class variation. The row-normalized confusion matrix (Fig. 3) further shows strong diagonal dominance overall, with `Liquid` achieving the highest class-conditional accuracy (0.90) and `Material-Interaction` the lowest (0.63). Most `Break/Drop` errors are predicted as `Material-Interaction` (0.13), reflecting similar transient impact cues. `Material-Inter action` shows the broadest confusion, most often mapped to `Clothing` (0.11), `Break/Drop` (0.09), and `Metal` (0.07). Near-symmetric confusions can also be observed between `Clothing` and `Footstep` (0.07/0.06) and between `Metal` and `Open/Close-Mechanism` (0.06/0.05).

6.2. Sub-Category Classification

For Task 2, performance decreases to 0.64 accuracy, 0.56 macro-F1, and 0.64 weighted-F1 on the 73-way sub-category test set (full per-class results are provided in the Appendix in Table **??**). This unsurprising drop relative to Exp. 1 reflects the substantially greater difficulty of fine-grained Foley recognition with many classes. Several sub-categories with acoustically distinctive characteristics still achieve strong results. For example, `Fart`, `HeartBeat`, `Toilet`, and `WhiteboardWriting` all achieve an F1 score of 1.00, while `Kiss` (F1 = 0.96), `Knock` (F1 = 0.96), and `Bubble` (F1 = 0.91) also perform well. In contrast, several sub-categories are particularly difficult to recognize, including zero-recall classes such as `Drip` (support = 5), `Latch` (5), `MetalDrop` (4), `StrawSip` (2), `WindowOpenClose` (6), and `WoodBreak` (6), as well as

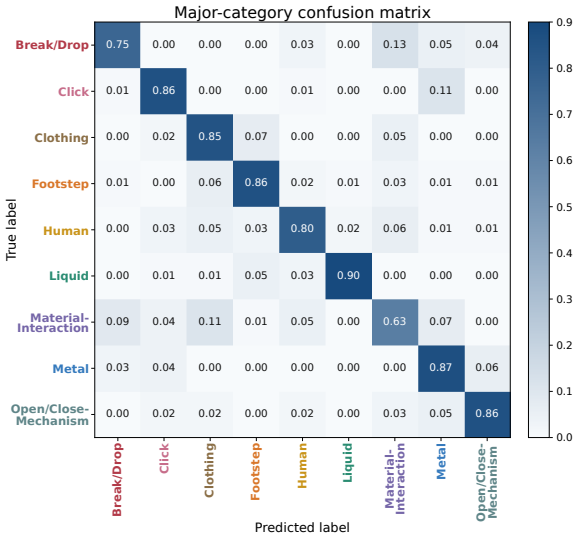

Figure 3: Confusion matrix on the major-category test set of FoleySet.

low-performing classes such as `IceBreak` (support = 13, F1 = 0.09) and `PaperTear` (support = 17, F1 = 0.21). Taken together, these cases suggest that weak sub-category performance is influenced by both limited learning resources and strong acoustic overlap between related classes. This interpretation is visualized by the sub-category confusion matrix Fig. 4, where errors often cluster within acoustically related regions rather than being distributed uniformly across the taxonomy. For example, `IceBreak` shows only a small recall (about 0.08), with strong confusion with nearby break-related classes, while `PaperTear` shows a similarly weak recall (about 0.12) together with visible off-diagonal confusion toward related tearing-like classes. A similar pattern appears among mechanism-related categories. For example, `DoorOpenClose` (R = 0.54), `DrawerOpenClose` (R = 0.45), and `WindowOpenClose` (R = 0.00) are often confused with other nearby mechanism classes in the confusion matrix, suggesting that the model captures the broader open/close sound family more reliably than the fine-grained distinctions between specific sources. We also observe cases of high precision but low recall, such as Chain (P = 1.00, R = 0.25) and `PaperTear` (P = 1.00, R = 0.12), indicating that the classifier is overly conservative for some rare classes: it correctly identifies a small number of highly prototypical examples, but assigns the majority of instances to acoustically similar neighboring categories.

Overall, these results provide a reproducible baseline for major and sub-category recognition under our taxonomy, while highlighting that Foley sub-category classification is challenged not only by limited learning resources, but also by fine-grained acoustic overlap and the difficulty of separating closely related classes with broad intra-class variation.

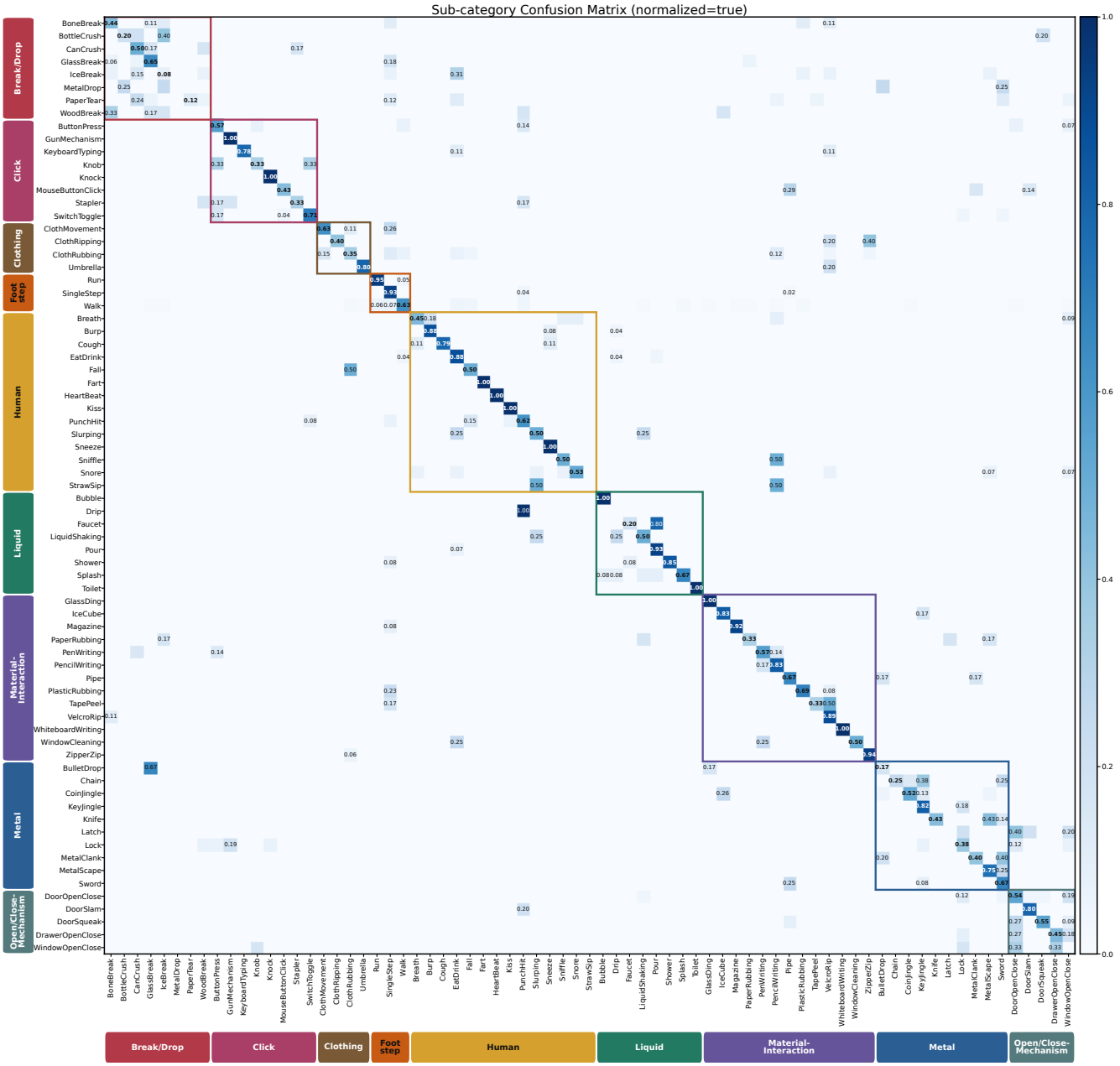

Figure 4: Confusion matrix on the sub-category test set of FoleySet.

7. CONCLUSION

We presented FoleySet[3], a novel human-annotated Foley dataset built upon a consistent and data-informed taxonomy. The work has two main contributions. First, it addresses the need for a Foley-related dataset in an important yet underexplored subdomain of the broader field of sound effects, providing a resource that facilitates future Foley research while also offering value for other data-driven audio tasks. Through workflow-derived annotations and carefully curated audio clips, FoleySet is intended to serve as a practical resource for developing Foley and broader sound-effect models with potential relevance to real-world production workflows. Second, it introduces a novel Foley-specific taxonomy with a systematic and well-documented design process, rationale, and structure, providing both a basis for future maintenance, refinement, adaptation, and extension and a reference for other researchers developing taxonomies for specialized audio domains. While the current release provides a strong foundation, its annotations were produced by a single annotator and its subcategory distribution remains long-tailed. Building on this work, future efforts will explore independent annotation validation, broader coverage of underrepresented categories, and evaluation across additional Foley-related tasks.

[3]The complete keyword-to-category mapping and full per-class subcategory results are available in the project repository.

8. REFERENCES

[1] G. G. Peeters and J. D. Reiss, "A deep learning approach to sound classification for film audio post-production," in *Proc. 148th AES Convention*, no. 10322, 2020.

[2] D. Moffat, D. Ronan, and J. D. Reiss, "Unsupervised taxonomy of sound effects," in *Proc. 20th Intl. Conf. on Digital Audio Effects (DAFx)*, Edinburgh, UK, 2017.

[3] A. B. Ma and A. Lerch, "Representation learning for the automatic indexing of sound effects libraries," in *Proc. 23rd Int. Society for Music Information Retrieval Conf. (ISMIR)*, Bengaluru, India, 2022, pp. 387–394.

[4] D. C.-E. Lin, A. Germanidis, C. Valenzuela, Y. Shi, and N. Martelaro, "Soundify: Matching sound effects to video," in *Proc. ACM Symp. User Interface Software and Technology (UIST)*, San Francisco, CA, Oct. 29–Nov. 1, 2023.

[5] K. Choi, J. Im, L. M. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the dcase 2023 challenge," in *Proc. Workshop Detection and Classification of Acoustic Scenes and Events (DCASE)*, Tampere, Finland, Sept. 21–22, 2023.

[6] Y. Chung, J. Lee, and J. Nam, "T-FOLEY: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Seoul, Korea, 2024.

[7] M. F. Colombo, F. Ronchini, L. Comanducci, and F. Antonacci, "MambaFoley: Foley sound generation using selective state-space models," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025, pp. 1–5.

[8] S. Ghose and J. J. Prevost, "Autofoley: Artificial synthesis of synchronized sound tracks for silent videos with deep learning," *IEEE Trans. on Multimedia*, vol. 23, pp. 1895–1907, 2021.

[9] S. Luo, C. Yan, C. Hu, and H. Zhao, "Diff-Foley: Synchronized video-to-audio synthesis with latent diffusion models," in *Advances in Neural Information Processing Systems 36 (NeurIPS)*, New Orleans, LA, Dec. 10–16, 2023.

[10] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[11] Freesound. Freesound. Accessed: Sep. 2025. Available: https://freesound.org.

[12] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. of the 22nd ACM Intl. Conf. on Multimedia (MM)*, Orlando, FL, 2014.

[13] BBC. BBC Sound Effects. Accessed: Sep.–Nov. 2025. Available: https://sound-effects.bbcrewind.co.uk/.

[14] S. Dixit, K. Saito, Z. Zhong, Y. Mitsufuji, and C. Donahue, "Foleybench: A benchmark for video-to-audio models," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 4–8 2026.

[15] R. Fu, S. Shi, H. Guo, T. Wang, C. Qiang, Z. Wen, J. Tao, X. Qi, Y. Lu, X. Wang, Z. Wang, Y. Liu, X. Liu, S. Zhang, and G. Li, "Mint: a multi-modal image and narrative text dubbing dataset for foley audio content planning and generation," *arXiv preprint arXiv:2406.10591*, 2024.

[16] N. Garcia and J. Reiss, "6ksfx synth dataset," *arXiv preprint arXiv:2501.17198*, 2025.

[17] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2017.

[18] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proc. of the 23rd ACM Conf. on Multimedia (MM)*, 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2733373.2806390

[19] T. Nielsen, J. Drury, K. Paquin *et al.*, "Universal Category System (UCS) category list," https://universalcategorysystem.com/, 2024, version 8.2.1, accessed March 10, 2026.

[20] M. Lewis, "Ventriloquial acts: Critical reflections on the art of foley," *The New Soundtrack*, vol. 5, no. 2, pp. 103–120, 2015.

[21] S. Trento and A. de Götzen, "Foley sounds vs. real sounds," in *Proc. Sound and Music Computing Conference (SMC)*, 2011.

[22] L. F. Donaldson, "The work of an invisible body: The contribution of foley artists to on-screen effort," *Alphaville: Journal of Film and Screen Media*, no. 7, 2014.

[23] Sound Ideas, "Hollywood foley fx," Sound Ideas product page, 2025, accessed: Sep. 2025. [Online]. Available: https://sound-ideas.com/products/hollywood-foley-fx

[24] ——, "Foley sound effects," Sound Ideas product page, 2025, accessed: Sep. 2025. [Online]. Available: https://sound-ideas.com/products/foley-sound-effects

[25] ——, "Art of foley sound effects library," Sound Ideas product page, 2025, accessed: Sep. 2025. [Online]. Available: https://sound-ideas.com/products/art-of-foley-sound-effects-library

[26] ——, "Hd – foley sound effects," Sound Ideas product page, 2025, accessed: Sep. 2025. [Online]. Available: https://sound-ideas.com/products/hd-foley-sound-effects

[27] ——, "Ultimate foley sound effects collection," Sound Ideas product page, 2025, accessed: Sep. 2025. [Online]. Available: https://sound-ideas.com/products/ultimate-foley-sound-effects-collection

[28] ——, "Fight foley sound effects," Sound Ideas product page, 2025, accessed: Sep. 2025. [Online]. Available: https://sound-ideas.com/products/fight-foley-sound-effects

[29] ——, "Foley footsteps sound effects library," Sound Ideas product page, 2025, accessed: Sep. 2025. [Online]. Available: https://sound-ideas.com/products/foley-footsteps-sound-effects-library

[30] European Broadcasting Union, "EBU R 128: Loudness normalisation and permitted maximum level of audio signals," European Broadcasting Union (EBU), Geneva, Switzerland, Tech. Rep. R 128, 2011.

[31] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Proc. Interspeech*, 2022, pp. 2753–2757.

[32] K. Koutini, "hear21passt: PaSST pre-trained models," https://github.com/kkoutini/PaSST, 2022, accessed: Feb. 2, 2026.