Audiocards: Structured Metadata Improves Audio Language Models for Sound Design

Sripathi Sridhar[1,2], Prem Seetharaman[2], Oriol Nieto[2], Mark Cartwright[1], Justin Salamon[2]

1 New Jersey Institute of Technology

2 Adobe Research


Stilgar whistle

BBC

SOUND EFFECTS

human whistle short

Showing top **300** of **424** results for **human whistle short**

Sound Mixer | Sort by | Filter (0)

Categories (0) | Duration (0) | Continents (0) | Reset Filters

AUTOPLAY

0:50	Wamba Indigenous Music - Short, sharp high pitched whistle calls using a two tone communication whistle.		Show details
0:14	Human Effects: Comic - One woman, three fairly short screams - 1968 (169A, reprocessed)		Show details
0:16	Steam Roller: Long And Short Whistle - Steam Roller: Long and short whistle. Exterior		Show details
0:05	The Age Of Steam - Guard's whistle.		Show details

BBC

SOUND EFFECTS

human whistle short

Showing top **300** of **424** results for **human whistle short**

Sound Mixer | Sort by | Filter (0)

Categories (0) | Duration (0) | Continents (0) | Reset Filters

AUTOPLAY

0:50	Wamba Indigenous Music - Short, sharp high pitched whistle calls using a two tone communication whistle.		Show details
0:14	Human Effects: Comic - One woman, three fairly short screams - 1968 (169A, reprocessed)		Show details
0:16	Steam Roller: Long And Short Whistle - Steam Roller: Long and short whistle. Exterior		Show details
0:05	The Age Of Steam - Guard's whistle.		Show details

BBC

SOUND EFFECTS

human whistle short

Showing top **300** of **424** results for **human whistle short**

Sound Mixer | Sort by | Filter (0)

Categories (0) | Duration (0) | Continents (0) | Reset Filters

AUTOPLAY

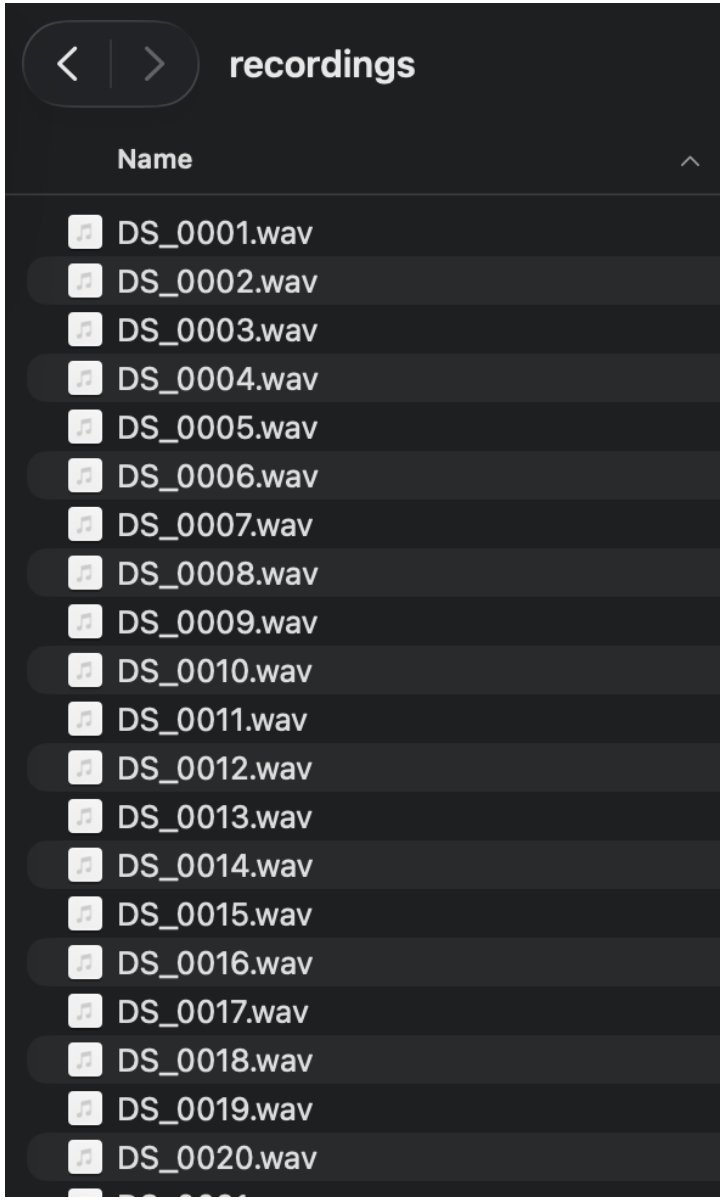
0:50	Wamba Indigenous Music - Short, sharp high pitched whistle calls using a two tone communication whistle.		Show details
0:14	Human Effects: Comic - One woman, three fairly short screams - 1968 (169A, reprocessed)		Show details
0:16	Steam Roller: Long And Short Whistle - Steam Roller: Long and short whistle. Exterior		Show details
0:05	The Age Of Steam - Guard's whistle.		Show details

Motivation

- Sound designers often search large sound effects databases
- However, existing **metadata is often incomplete** or missing, especially for personal sound effects collections

Motivation

- Sound designers often search large sound effects databases
- However, existing **metadata is often incomplete** or missing, especially for personal sound effects collections



BBC

SOUND EFFECTS

human whistle short

Showing top **300** of **424** results for **human whistle short**

Sound Mixer

Sort by

Filter (0)

Categories (0)

Duration (0)

Continents (0)

Reset Filters

AUTOPLAY

0:50 Show details

0:14 Show details

0:16 Show details

0:05 Show details

Text-audio retrieval requires auditioning every result in the absence of metadata

BBC

SOUND EFFECTS

human whistle short

Showing top **300** of **424** results for **human whistle short**

Sound Mixer | Sort by | Filter (0)

Categories (0) | Duration (0) | Continents (0) | Reset Filters

AUTOPLAY

0:50 Show details

0:14 Show details

0:16 Show details

0:05 Show details

Manual annotation is expensive but existing models are ill-suited to sound effects description

Motivation

- Sound designers often search large sound effects databases
- However, existing **metadata is often incomplete** or missing, especially for amateur sound effects collections
- Current **audio understanding models are not well-suited** to such tasks as they are trained on unstructured captions

Goals

In this work, we aim to **bridge this gap** between audio understanding models and sound design workflows by:

- Automatically generating structured attribute fields relevant to sound designers
- Enabling search through large libraries using attributes relevant to sound design
- Organizing large sound effects collections and allowing sound designers to assess search results more quickly

We propose **Audiocards:** structured metadata with fields pertinent to sound design

Car driving on a gravel road

Standard captions

Nouns: BMW, car, vehicle, tires, gravel
Verbs: Driving, moving
Noun-verb pairs: tires on gravel, car driving, vehicle moving
Example visual context: A lone driver getting lost on a gravel road
Adjectives: Isolated, soft, gentle, crunchy, muited
Complementary sounds: Wind rustling through trees, distant bird calls, engine idling, gravel and rocks settling, driver's sighs
UCS Category: Vehicles, ***UCS Subcategory***: Tire
Cause: Car tires rolling on gravel road
Effect: Car slowly coming to a stop, engine idling
Caption in 3 words: Tires on gravel
Caption in 7 words or less: Car driving on a gravel road
Descriptive caption: The soft and gentle sound of a car's tires rolling on a gravel road, creating a crunchy and muted texture, evoking a sense of isolation and calmness.

Audiocards

Audiocards contain fields pertinent to sound designers

Fields selected from conversations with a sound designer

Fields clearly communicate attributes relevant to sound design in search results

Nouns: BMW, car, vehicle, tires, gravel
Verbs: Driving, moving
Noun-verb pairs: tires on gravel, car driving, vehicle moving
Example visual context: A lone driver getting lost on a gravel road
Adjectives: isolated, soft, gentle, crunchy, muted
Complementary sounds: Wind rustling through trees, distant bird calls, engine idling, gravel and rocks settling, driver's sighs
UCS Category: Vehicles, ***UCS Subcategory***: Tire
Cause: Car tires rolling on gravel road
Effect: Car slowly coming to a stop, engine idling
Caption in 3 words: Tires on gravel
Caption in 7 words or less: Car driving on a gravel road
Descriptive caption: The soft and gentle sound of a car's tires rolling on a gravel road, creating a crunchy and muted texture, evoking a sense of isolation and calmness.

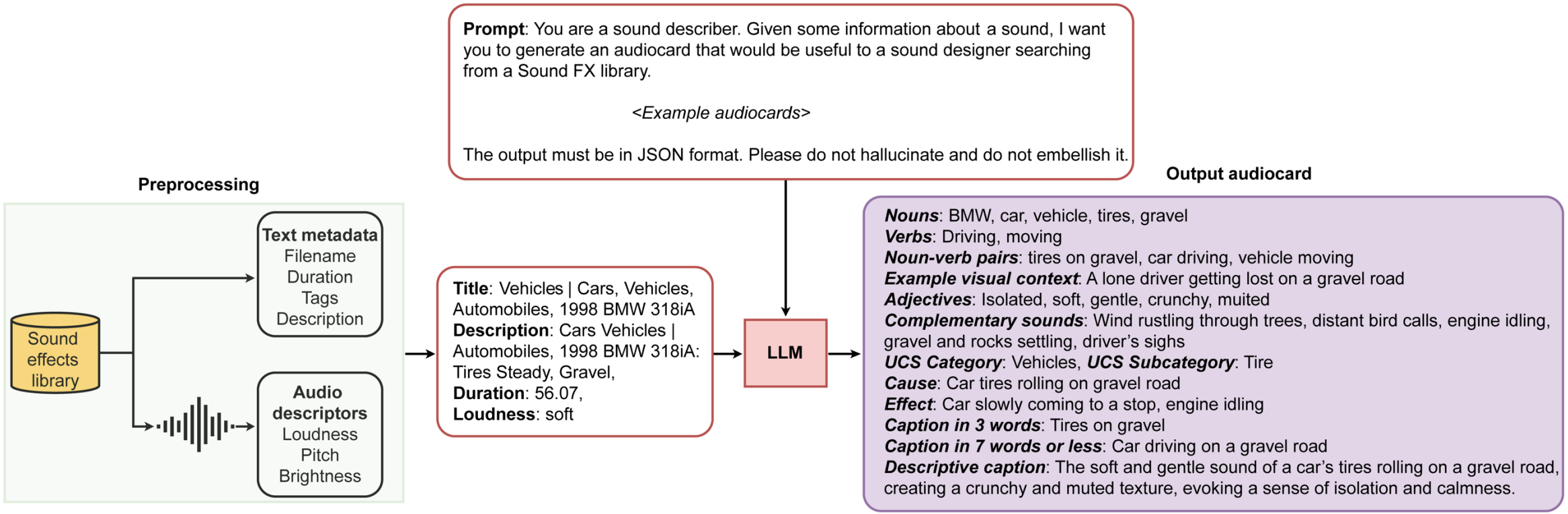
Audiocards can be used to train audio-language models for sound design

Audiocards can be used to train downstream models on tasks such as:

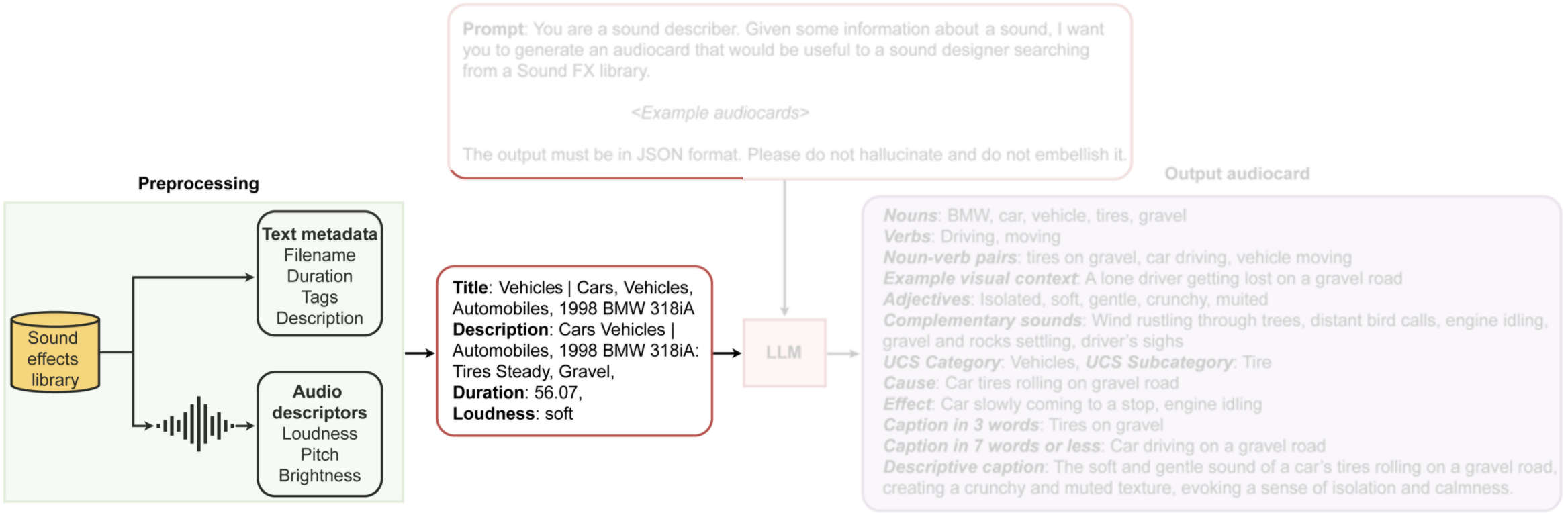
1. Joint text-audio representation learning
2. Sound design captioning
3. Metadata generation

Nouns: BMW, car, vehicle, tires, gravel
Verbs: Driving, moving
Noun-verb pairs: tires on gravel, car driving, vehicle moving
Example visual context: A lone driver getting lost on a gravel road
Adjectives: Isolated, soft, gentle, crunchy, muited
Complementary sounds: Wind rustling through trees, distant bird calls, engine idling, gravel and rocks settling, driver's sighs
UCS Category: Vehicles, ***UCS Subcategory***: Tire
Cause: Car tires rolling on gravel road
Effect: Car slowly coming to a stop, engine idling
Caption in 3 words: Tires on gravel
Caption in 7 words or less: Car driving on a gravel road
Descriptive caption: The soft and gentle sound of a car's tires rolling on a gravel road, creating a crunchy and muted texture, evoking a sense of isolation and calmness.

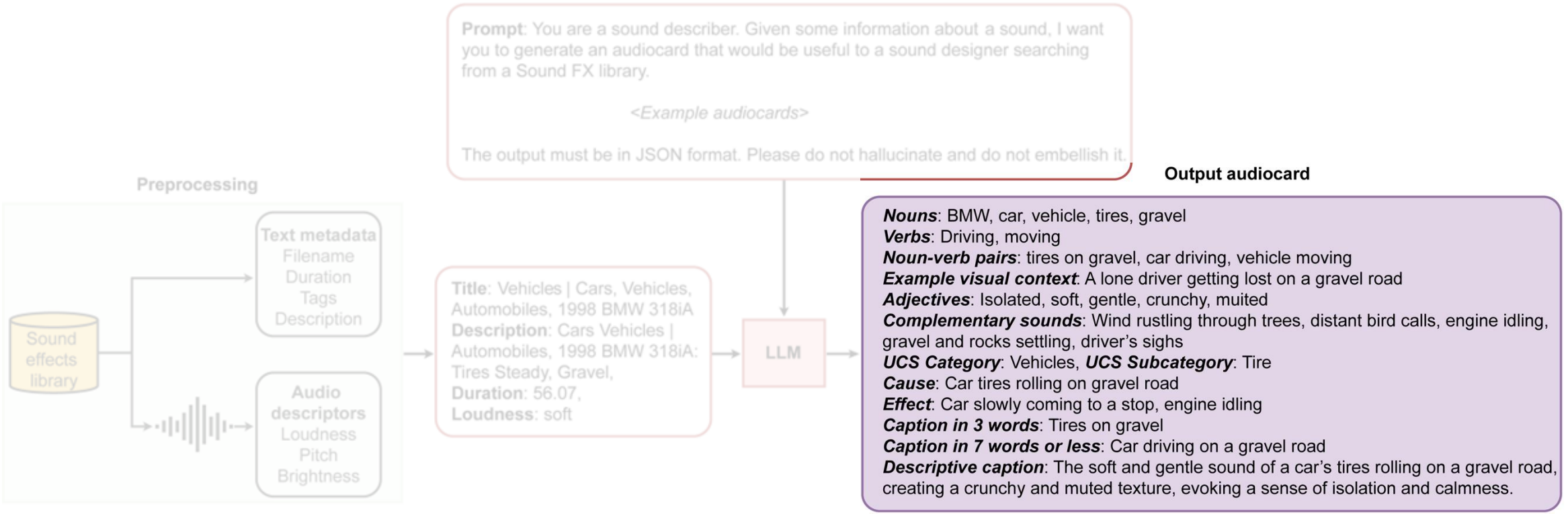
Audiocards are generated from available text metadata and audio descriptors



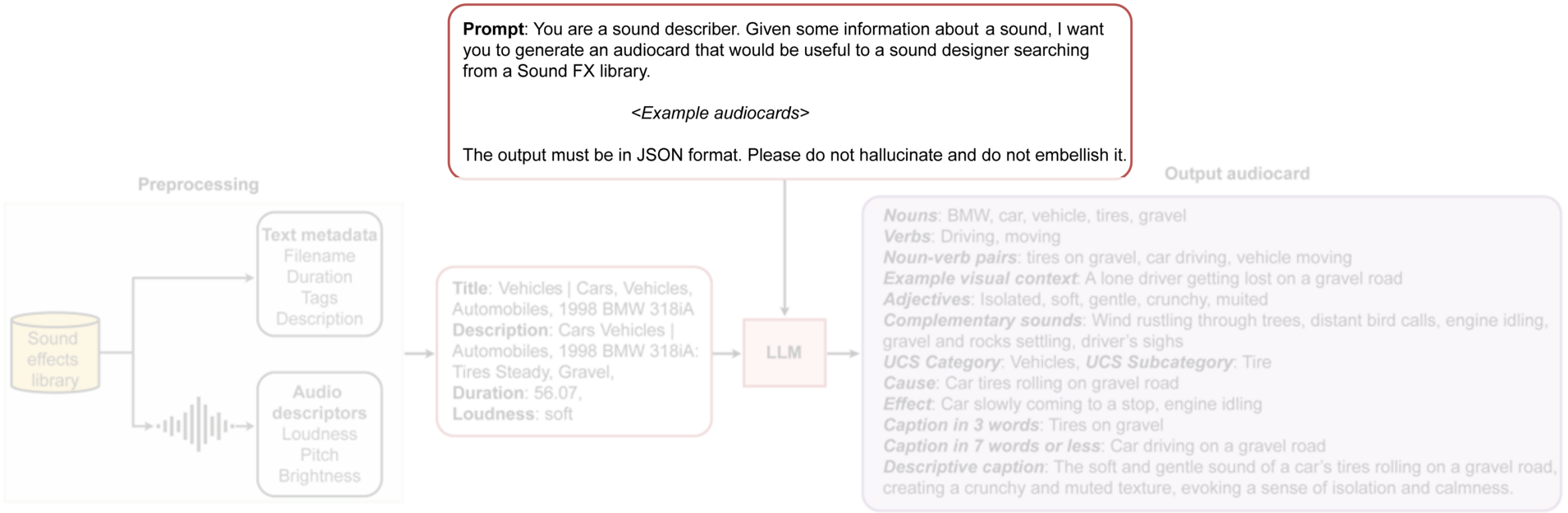
Audiocards are generated from available text metadata and audio descriptors



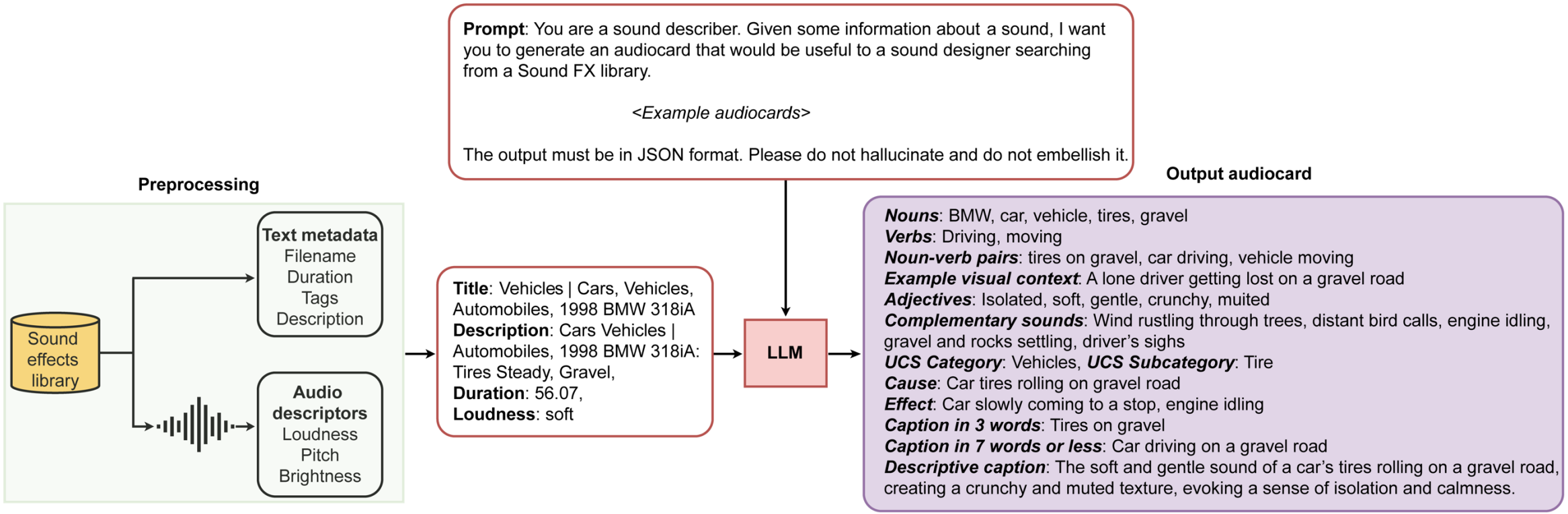
Audiocards are generated from available text metadata and audio descriptors



Audiocards are generated from available text metadata and audio descriptors



Audiocards are generated from available text metadata and audio descriptors



ASFx-eval: Human-verified sound effects evaluation dataset

- Standard captioning/retrieval datasets like Clotho or Audiocaps are unsuitable for evaluating models for sound design
- **ASFx-eval**: We curate a benchmark dataset of 500 manually verified audiocards from Audition Sound Effects with no hallucinations

ASFx-eval dataset on Zenodo

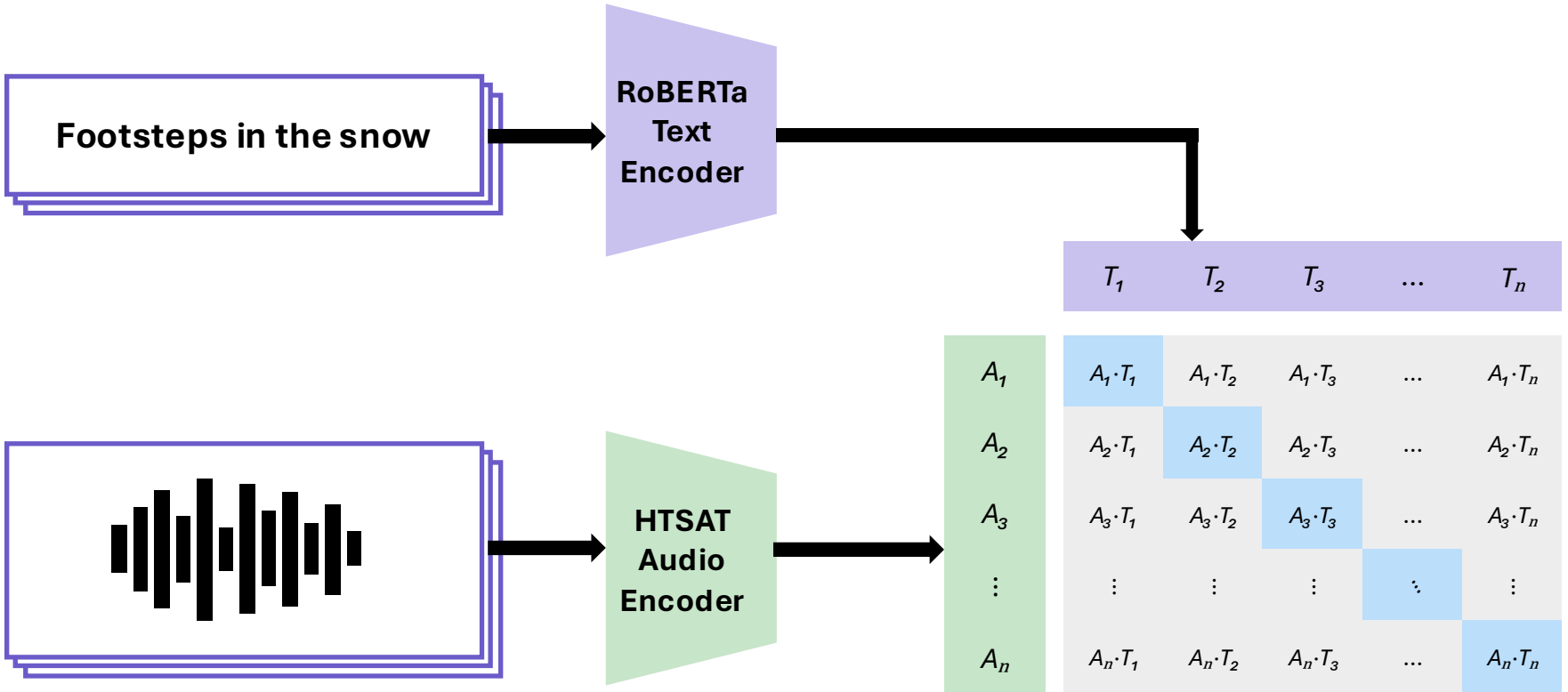
ASFx-eval: Human-verified sound effects evaluation dataset

- Standard captioning/retrieval datasets like Clotho or Audiocaps are unsuitable for evaluating models for sound design
- **ASFx-eval**: We curate a benchmark dataset of 500 manually verified audiocards from Audition Sound Effects with no hallucinations
  - We use this dataset to evaluate sound design captioning and metadata generation

ASFx-eval dataset on Zenodo

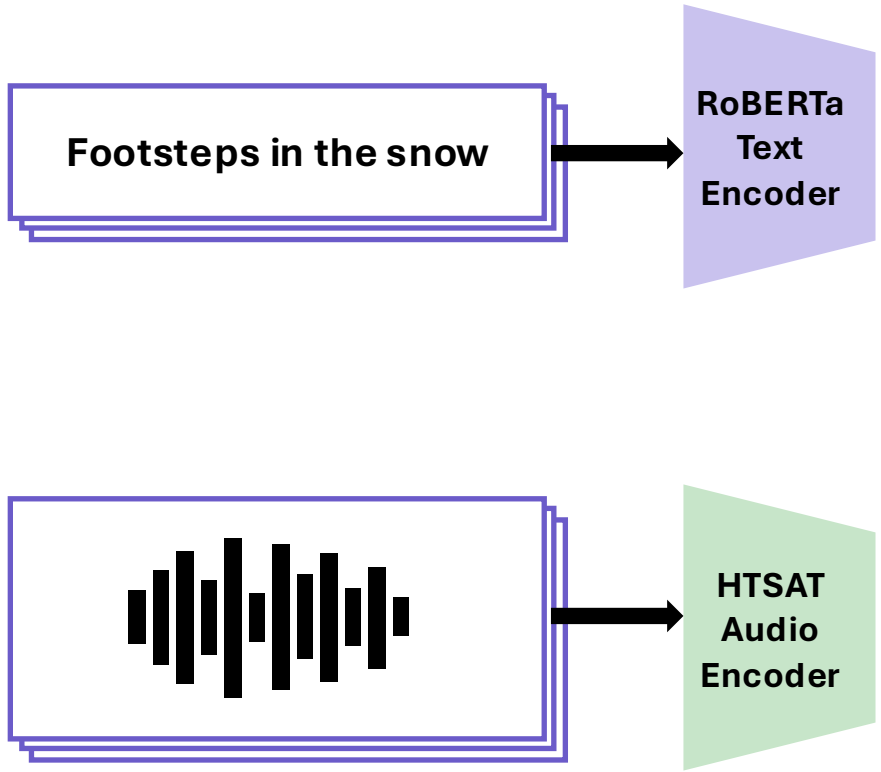
Task 1: Joint text-audio representation learning

Train a contrastive CLAP-style model on (text, audio) pairs



Task 1: Joint text-audio representation learning

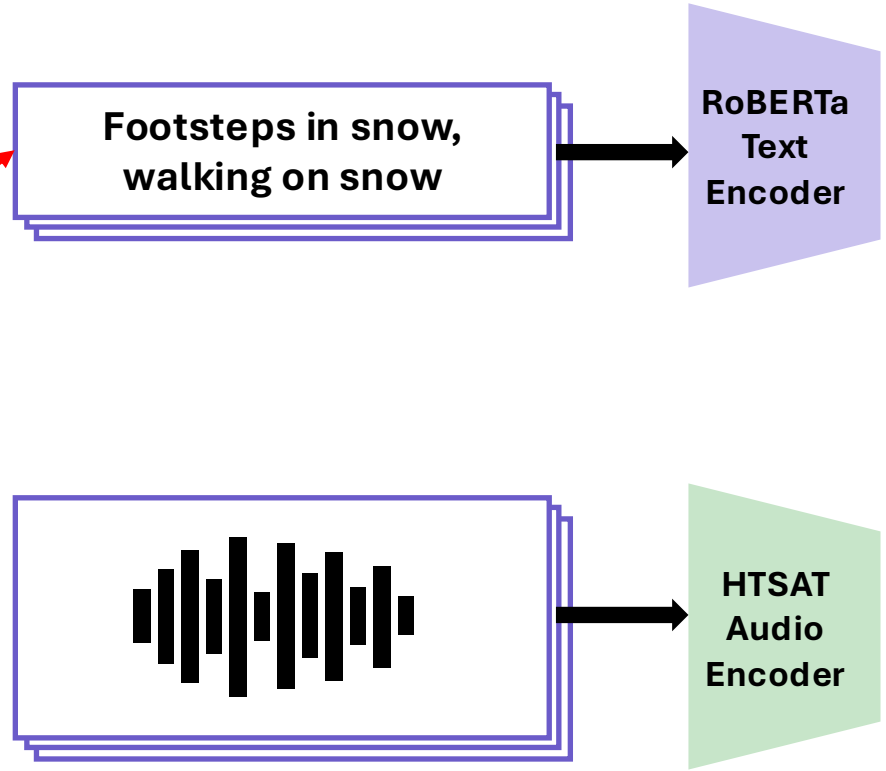
Captions-CLAP: a model trained on **baseline synthetic captions** generated without audiocards from available text metadata



Task 1: Joint text-audio representation learning

Cards-CLAP: Our proposed model trained on randomly sampled **audiocard fields**

Nouns: snow, footstep, weather
Verbs: stepping, walking
Noun-verb pairs: footsteps in snow, walking on snow
Example visual context: A person walking through a snowy
Adjectives: soft, crunchy, cold, natural
Complementary sounds: wind blowing, snowflakes falling,
UCS Category: Footsteps, ***UCS Subcategory***: Human
Cause: A person walking on snow
Effect: A soft, crunchy sound indicating movement through a
Caption in 3 words: Snow footsteps
Caption in 7 words or less: Soft footsteps in the snow
Descriptive caption: A soft and crunchy sound of footsteps
snowy environments.



Does training on audiocards improve downstream text-audio retrieval performance?

Model	Training regime	Dataset	R@10
LAION-CLAP	—	SFx-R	24.85
Captions-CLAP	Baseline captions	SFx-R	73.45
Cards-CLAP	Audiocard fields	SFx-R	**75.40**

Evaluated zero-shot text-audio retrieval using SFx-R: 6k held-out **professional sound effects with human captions**

Wu et al. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP 2023*.

Does training on audiocards improve downstream text-audio retrieval performance?

Model	Training regime	Dataset	R@10
LAION-CLAP	—	SFx-R	24.85
Captions-CLAP	Baseline captions	SFx-R	73.45
Cards-CLAP	Audiocard fields	SFx-R	**75.40**

Training on sound effects data is crucial for downstream retrieval performance

Wu et al. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP 2023*.

Does training on audiocards improve downstream text-audio retrieval performance?

Model	Training regime	Dataset	R@10
LAION-CLAP	—	SFx-R	24.85
Captions-CLAP	Baseline captions	SFx-R	73.45
Cards-CLAP	Audiocard fields	SFx-R	**75.40**

Training on audiocard fields **improves text-audio retrieval on sound effects** data over baseline captions

Wu et al. (2023). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP 2023*.

Task 2: Sound design captioning evaluated on audiocard descriptive captions

Nouns: snow, footstep, weather
Verbs: stepping, walking
Noun-verb pairs: footsteps in snow, walking on snow
Example visual context: A person walking through a snowy landscape, leaving footprints behind
Adjectives: soft, crunchy, cold, natural
Complementary sounds: wind blowing, snowflakes falling, branches cracking, birds chirping
UCS Category: Footsteps, ***UCS Subcategory***: Human
Cause: A person walking on snow
Effect: A soft, crunchy sound indicating movement through a snowy environment
Caption in 3 words: Snow footsteps
Caption in 7 words or less: Soft footsteps in the snow
Descriptive caption: A soft and crunchy sound of footsteps in the snow, ideal for winter scenes or snowy environments.

Task 2: Sound design captioning evaluated on audiocard descriptive captions

We finetune Whisper-medium-v3 in three variants:

Radford et al (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR.

Task 2: Sound design captioning evaluated on audiocard descriptive captions

We finetune Whisper-medium-v3 in three variants:

- **Whisper-Cards (full card)**: Trained on full audiocards

Nouns: snow, footstep, weather
Verbs: stepping, walking
Noun-verb pairs: footsteps in snow, walking on snow
Example visual context: A person walking through a snowy
Adjectives: soft, crunchy, cold, natural
Complementary sounds: wind blowing, snowflakes falling, b
UCS Category: Footsteps, ***UCS Subcategory***: Human
Cause: A person walking on snow
Effect: A soft, crunchy sound indicating movement through a
Caption in 3 words: Snow footsteps
Caption in 7 words or less: Soft footsteps in the snow
Descriptive caption: A soft and crunchy sound of footsteps i
snowy environments.

Radford et al (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR.

Task 2: Sound design captioning evaluated on audiocard descriptive captions

We finetune Whisper-medium-v3 in three variants:

- **Whisper-Cards (full card)**: Trained on full audiocards
- **Whisper-Cards (card caption)**: Trained on audiocard captions only

Nouns: snow, footstep, weather
Verbs: stepping, walking
Noun-verb pairs: footsteps in snow, walking on snow
Example visual context: A person walking through a snowy
Adjectives: soft, crunchy, cold, natural
Complementary sounds: wind blowing, snowflakes falling, b
UCS Category: Footsteps, ***UCS Subcategory***: Human
Cause: A person walking on snow
Effect: A soft, crunchy sound indicating movement through a
Caption in 3 words: Snow footsteps
Caption in 7 words or less: Soft footsteps in the snow
Descriptive caption: A soft and crunchy sound of footsteps i
snowy environments.

Radford et al (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR.

Task 2: Sound design captioning evaluated on audiocard descriptive captions

We finetune Whisper-medium-v3 in three variants:

- **Whisper-Cards (full card)**: Trained on full audiocards
- **Whisper-Cards (card caption)**: Trained on audiocard captions only
- **Whisper-Baseline**: Trained on baseline synthetic captions

Radford et al (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*. PMLR.

Are captions generated in audiocards better than baseline captions?

Model	Training regime	Dataset	SPIDEr	FENSE
Whisper-Baseline	Baseline captions	ASFx-eval	7.98	49.78
Whisper-Cards (card caption)	Audiocard caption	ASFx-eval	**19.36**	**53.40**

We evaluate sound effects captioning on **ASFx-eval**, our manually verified sound effects audiocards dataset

Are captions generated in audiocards better than baseline captions?

Model	Training regime	Dataset	SPIDEr	FENSE
Whisper-Baseline	Baseline captions	ASFx-eval	7.98	49.78
Whisper-Cards (card caption)	Audiocard caption	ASFx-eval	**19.36**	**53.40**

Audiocard captions encode **more semantically useful information** than baseline captions

Are captions generated in audiocards better than LALM captions?

Model	Training regime	SPIDEr	FENSE
Whisper-Cards (card caption)	Audiocard caption	**19.36**	**53.40**
GAMA	---	9.30	45.70
Audio Flamingo 3	---	5.27	32.10
Audio Flamingo 3 (think)	---	9.61	42.61

Our model significantly outperforms SOTA LALMs on sound effects captioning, indicating their lack of domain knowledge

Ghosh et al (2024). *GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities*. EMNLP 2024.
Goel et al. (2025). *Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models*. arXiv preprint.

Does generating full audiocards sacrifice caption quality?

Model	Training regime	SPIDEr	FENSE
Whisper-Baseline	Baseline captions	7.98	49.78
Whisper-Cards (card caption)	Audiocard caption	**19.36**	**53.40**
Whisper-Cards (full card)	Full audiocard	<u>18.61</u>	<u>51.78</u>

No! **Whisper-Cards (full card) can generate full audiocards** while maintaining caption quality

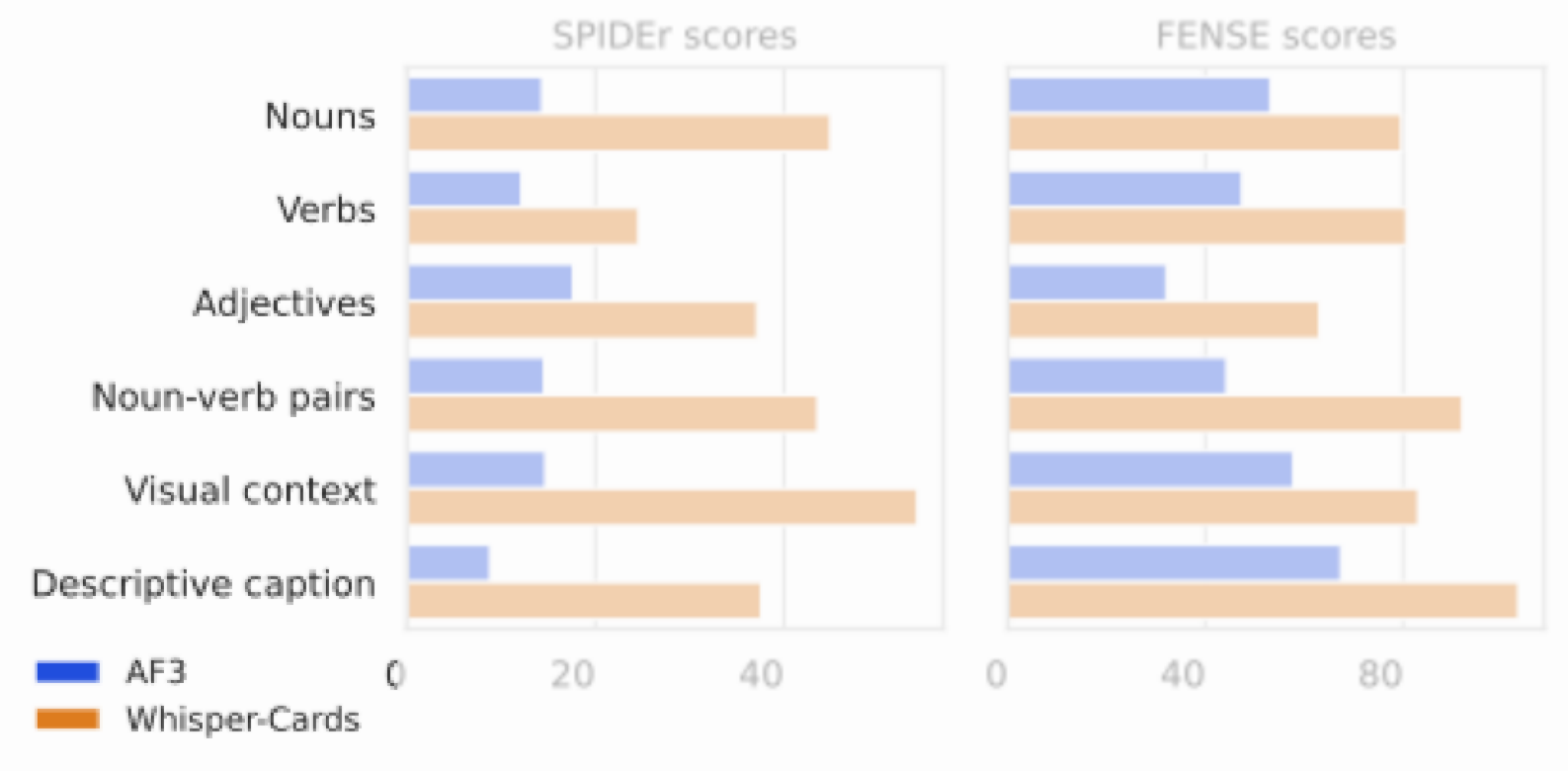
Task 3: Metadata generation

- We propose **audiocard field prediction**, a novel metadata generation task for sound design
- Models generate a full audiocard given an audio file

Nouns: snow, footstep, weather
Verbs: stepping, walking
Noun-verb pairs: footsteps in snow, walking on snow
Example visual context: A person walking through a snowy landscape, leaving footprints behind
Adjectives: soft, crunchy, cold, natural
Complementary sounds: wind blowing, snowflakes falling, branches cracking, birds chirping
UCS Category: Footsteps, ***UCS Subcategory***: Human
Cause: A person walking on snow
Effect: A soft, crunchy sound indicating movement through a snowy environment
Caption in 3 words: Snow footsteps
Caption in 7 words or less: Soft footsteps in the snow
Descriptive caption: A soft and crunchy sound of footsteps in the snow, ideal for winter scenes or snowy environments.

Can LALMs generate audiocard fields?

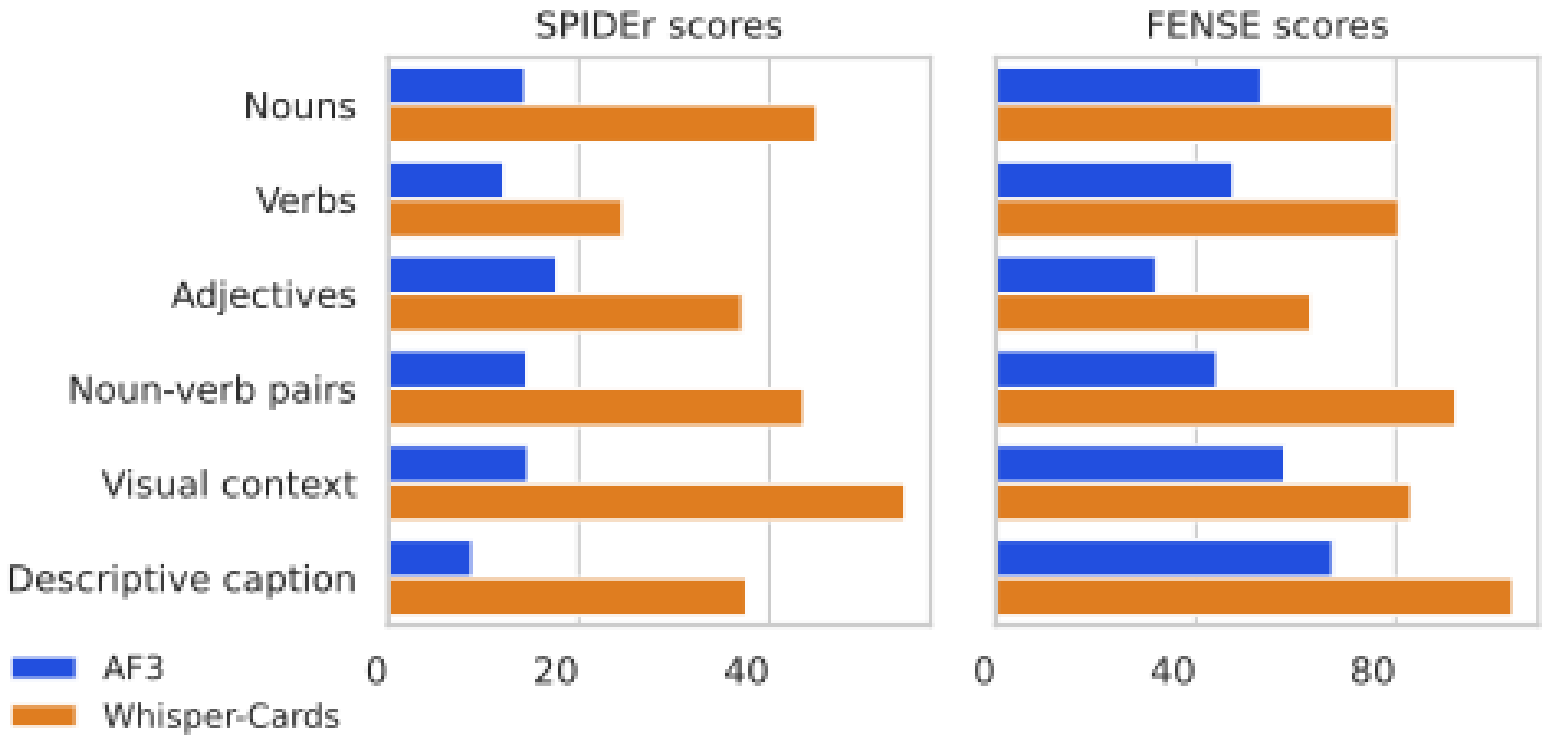
- We benchmark Audio Flamingo 3 (think)
- Models are **evaluated individually on audiocard fields** most relevant to sound effects search



Goel et al. (2025). *Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models*. arXiv preprint.

Can LALMs generate audiocard fields?

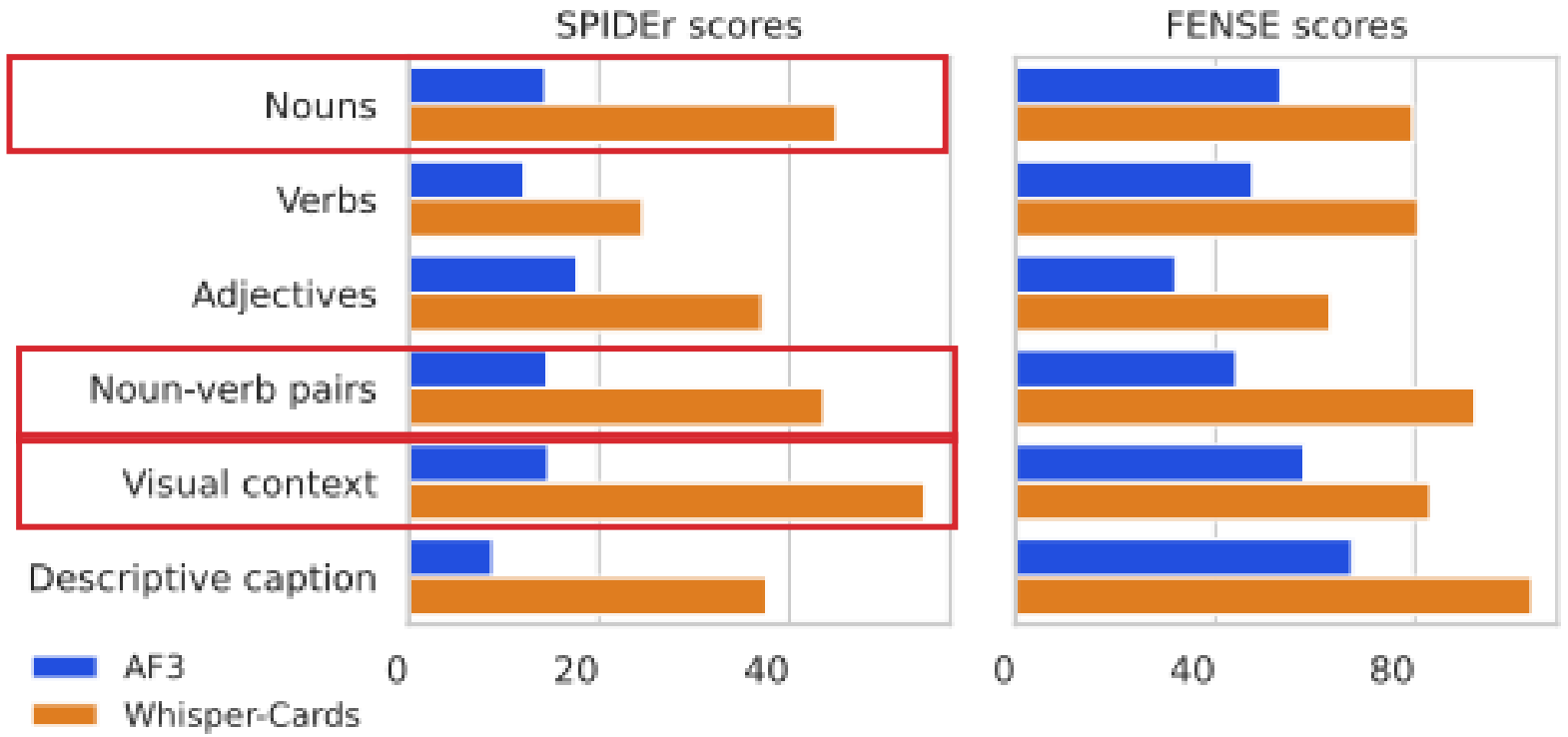
Our model outperforms AF3 across all audiocard fields on metadata generation



Goel et al. (2025). *Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models*. arXiv preprint.

Can LALMs generate audiocard fields?

Our model outperforms AF3 across all audiocard fields on metadata generation



Goel et al. (2025). *Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models*. arXiv preprint.

Example audiocard generated by Whisper-Cards

Nouns: Telephone, phone, ring, electronic
Verbs: Ringing
Noun-verb pairs: Telephone ringing, phone ringing, electronic ring
Example visual context: A modern smartphone on a table, screen lighting up with an incoming call
Adjectives: Electronic, digital, modern, short
UCS Category: Communications, ***UCS Subcategory***: Telephone
Caption in 3 words: Phone ringing
Caption in 7 words or less: Modern electronic phone ring sound
Descriptive caption: A short, electronic phone ring sound, typical of modern telephones and telephones

Limitations and future work

- How do individual audiocard fields contribute to downstream task performance in retrieval and sound effects captioning?
- How does audiocard structure better unlock LLM world knowledge?
- How should audiocards and its downstream models be incorporated into sound design workflows?
- How can structured metadata be adapted for audio understanding tasks in other domains such as music and environmental sound?

Applications of audiocards

- Organization of large sound effects collections
- Structured metadata intermediary for downstream audio captioning modeling

---

TAC: Timestamped Audio Captioning

Sonal Kumar [1 2 *] **Prem Seetharaman** [2 *] **Ke Chen** [2] **Oriol Nieto** [2] **Jiaqi Su** [2]
Zhepei Wang [2] **Rithesh Kumar** [3] **Dinesh Manocha** [1] **Nicholas J. Bryan** [2] **Zeyu Jin** [2] **Justin Salamon** [2]

Conclusion

- We proposed **audiocards**, structured metadata generated from available text metadata and audio descriptors

Conclusion

- We proposed **audiocards**, structured metadata generated from available text metadata and audio descriptors
- We released **ASFx-eval**, a human-verified benchmark dataset of 500 audiocards to encourage future work

Conclusion

- We proposed **audiocards**, structured metadata generated from available text metadata and audio descriptors
- We released **ASFx-eval**, a human-verified benchmark dataset of 500 audiocards to encourage future work
- Our metadata generation and sound effects captioning model trained on audiocards **outperforms SOTA LALMs**

Conclusion

- We proposed **audiocards**, structured metadata generated from available text metadata and audio descriptors
- We released **ASFx-eval**, a human-verified benchmark dataset of 500 audiocards to encourage future work
- Our metadata generation and sound effects captioning model trained on audiocards **outperforms SOTA LALMs**
- Training on audiocards improves captioning and text-audio representations, **enabling the organization of and search through large sound effects libraries**

Resources

ICASSP’26 Paper

ASFx-eval Dataset

Sripathi Sridhar · ss645@njit.edu