8613

A SOUND APPROACH: USING LARGE LANGUAGE MODELS TO GENERATE AUDIO DESCRIPTIONS FOR EGOCENTRIC TEXT-AUDIO RETRIEVAL

Andreea-Maria Oncescu[1] João F. Henriques[1] Andrew Zisserman[1] Samuel Albanie[2] A. Sophia Koepke[3]

[1]University of Oxford [2]University of Cambridge [3]University of Tübingen

1. Text-audio retrieval task

Text query: “Water gurgling”

Task: Retrieve matching audio from pool of test samples

Top 3 videos associated with the retrieved soundtrack

2. Introduction

Aim: Text-to-audio retrieval in egocentric setting

Challenge: Lack of labelled audio descriptions

Kinetics class: sanding wood
AudioCaps description: A man speaks as wood is sanded

Approach: Generate audio descriptions with Large Language Models starting from video descriptions.

3. Data and models used

Data:

- EpicMIR[1]: based on EpicKitchens, pairs of verb/s+noun/s and videos
- EgoMCQ[2]: based on Ego4D, contains pairs of (description, 5 clips)
- EpicSounds[3]: based on EpicKitchens, audio class labels and audio

Models:

- LAION-Clap[4]
- WavCaps[5]

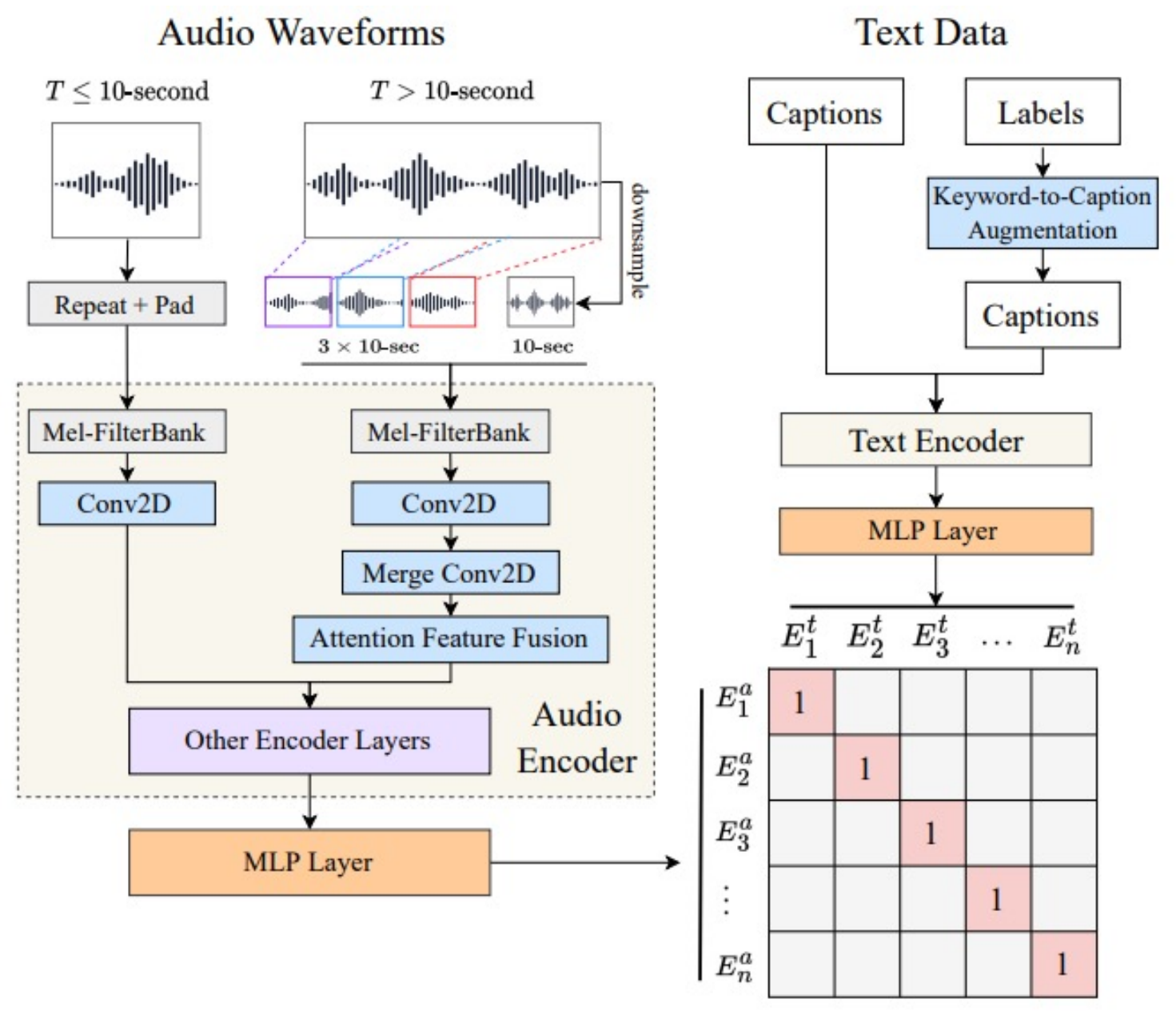
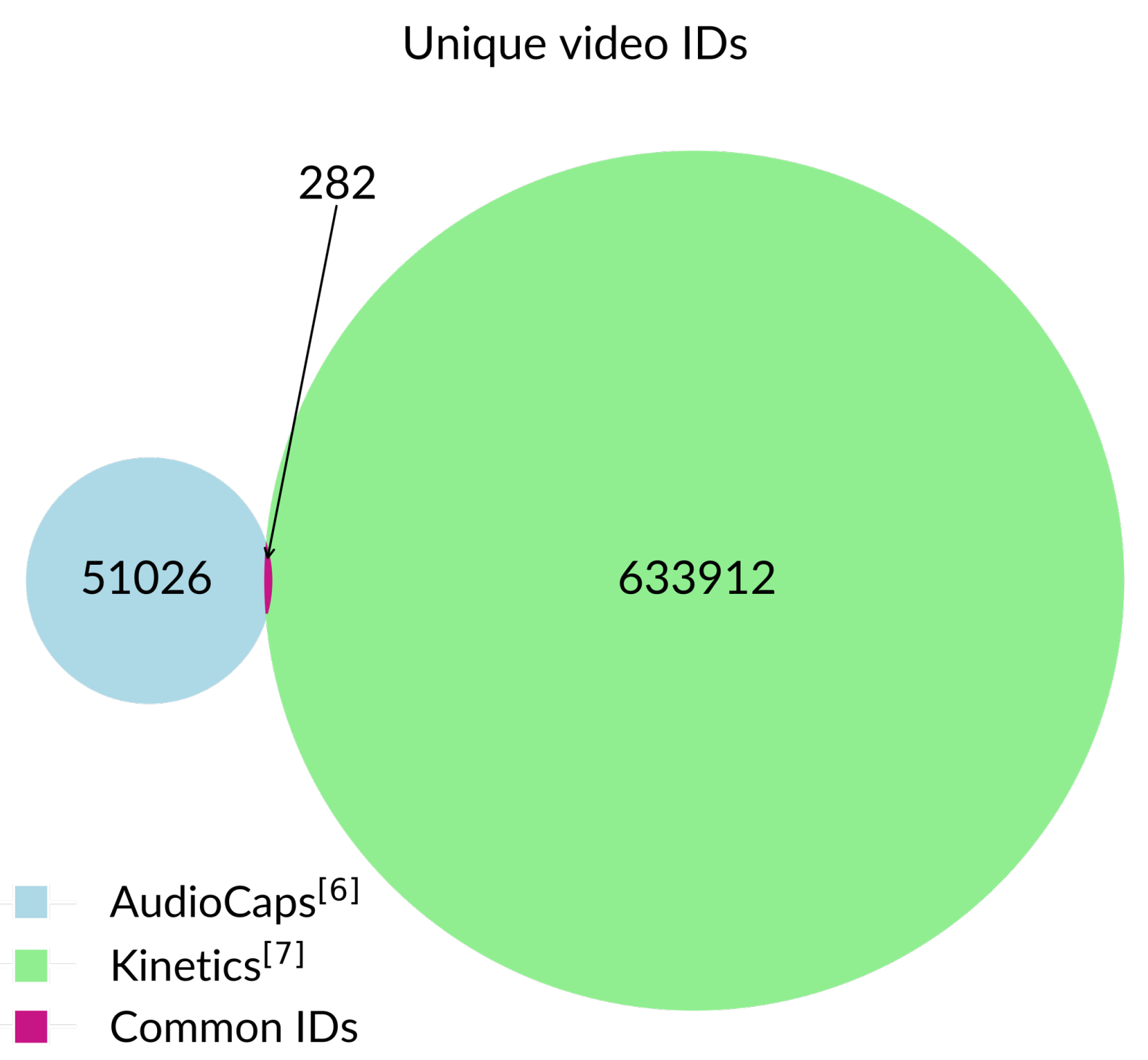


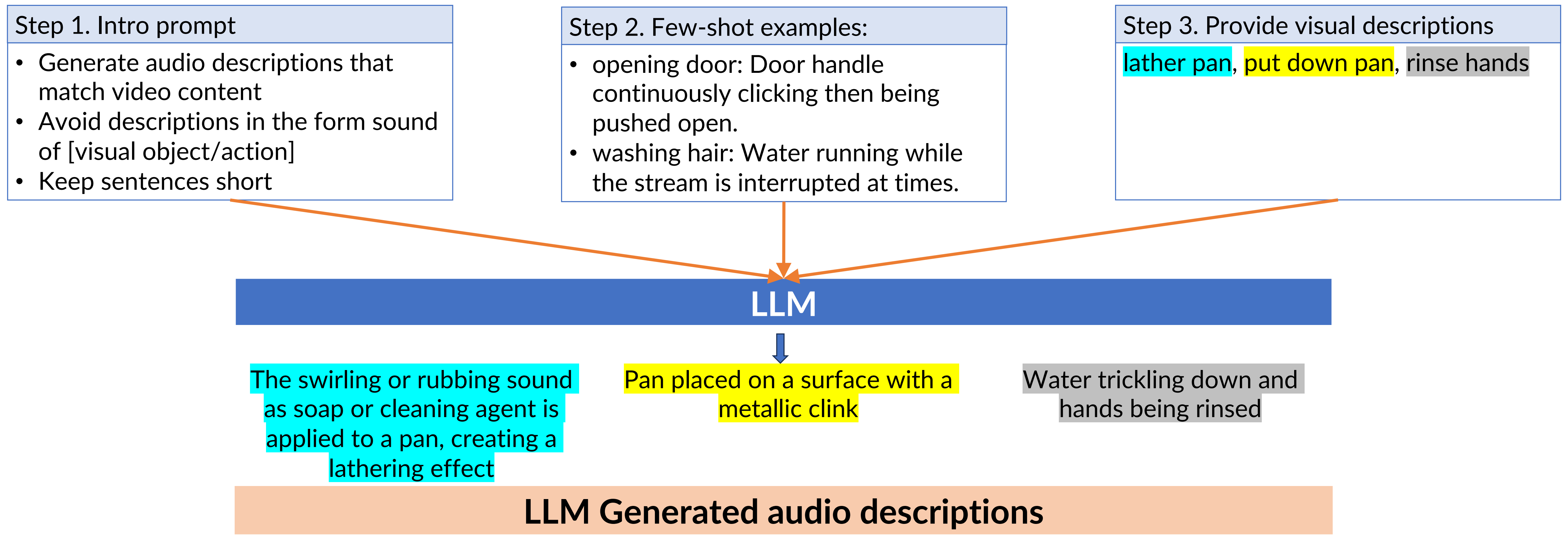
Diagram of LAION-Clap model[6]

Acknowledgements

This work is supported by the EPSRC (VisualAI EP/T028572/1 and DTA Studentship), the Royal Academy of Engineering (RF\201819\18\163), an Isaac Newton Trust Grant and the DFG - EXC 2064/1 - project 390727645. We are grateful to Jaesung Huh, Triantafyllos Afouras and Bernie Huang for their helpful comments and suggestions.

4. Approach





Step 1. Intro prompt

- Generate audio descriptions that match video content
- Avoid descriptions in the form sound of [visual object/action]
- Keep sentences short

Step 2. Few-shot examples:

- opening door: Door handle continuously clicking then being pushed open.
- washing hair: Water running while the stream is interrupted at times.

Step 3. Provide visual descriptions

lather pan, put down pan, rinse hands

LLM

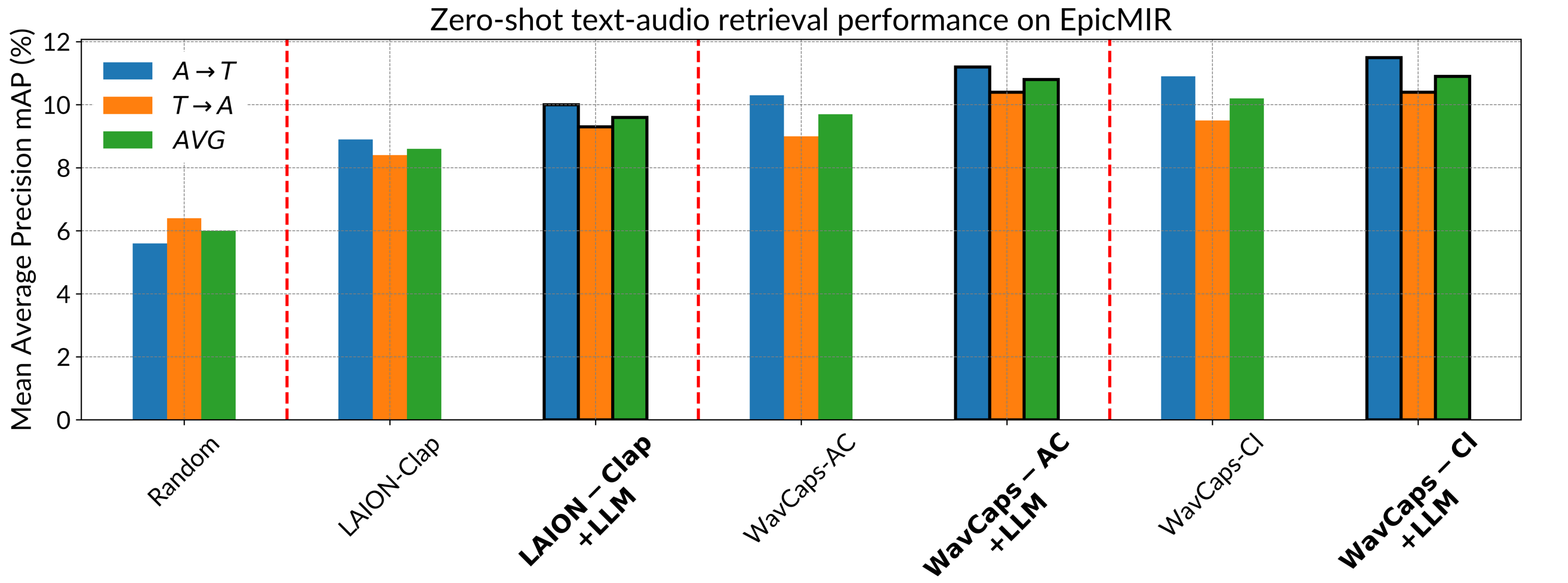
The swirling or rubbing sound as soap or cleaning agent is applied to a pan, creating a lathering effect

Pan placed on a surface with a metallic clink

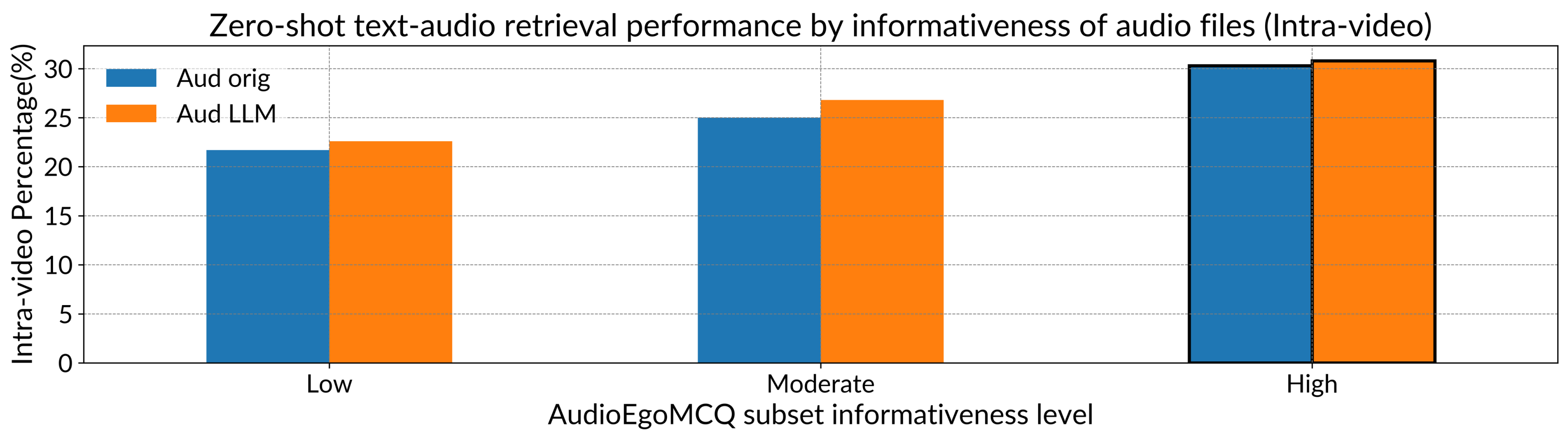
Water trickling down and hands being rinsed

LLM Generated audio descriptions

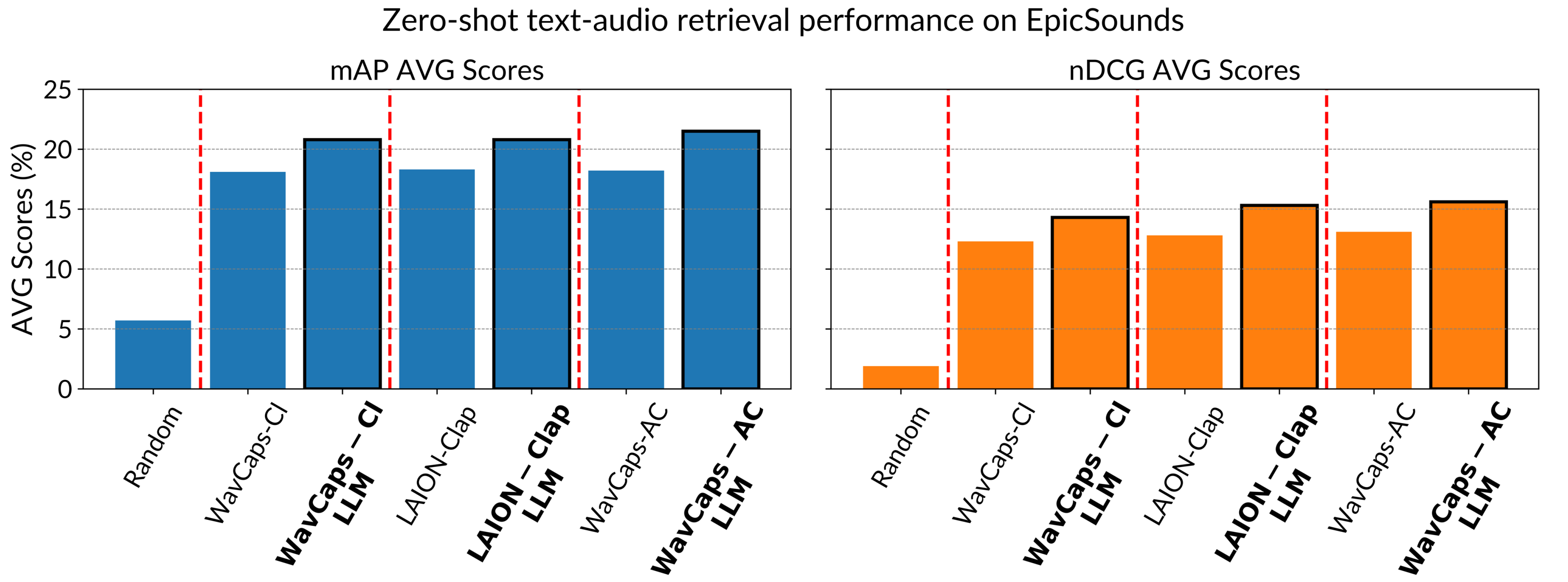
5. Results on EpicMIR: Comparing LLM-generated descriptions to video class labels



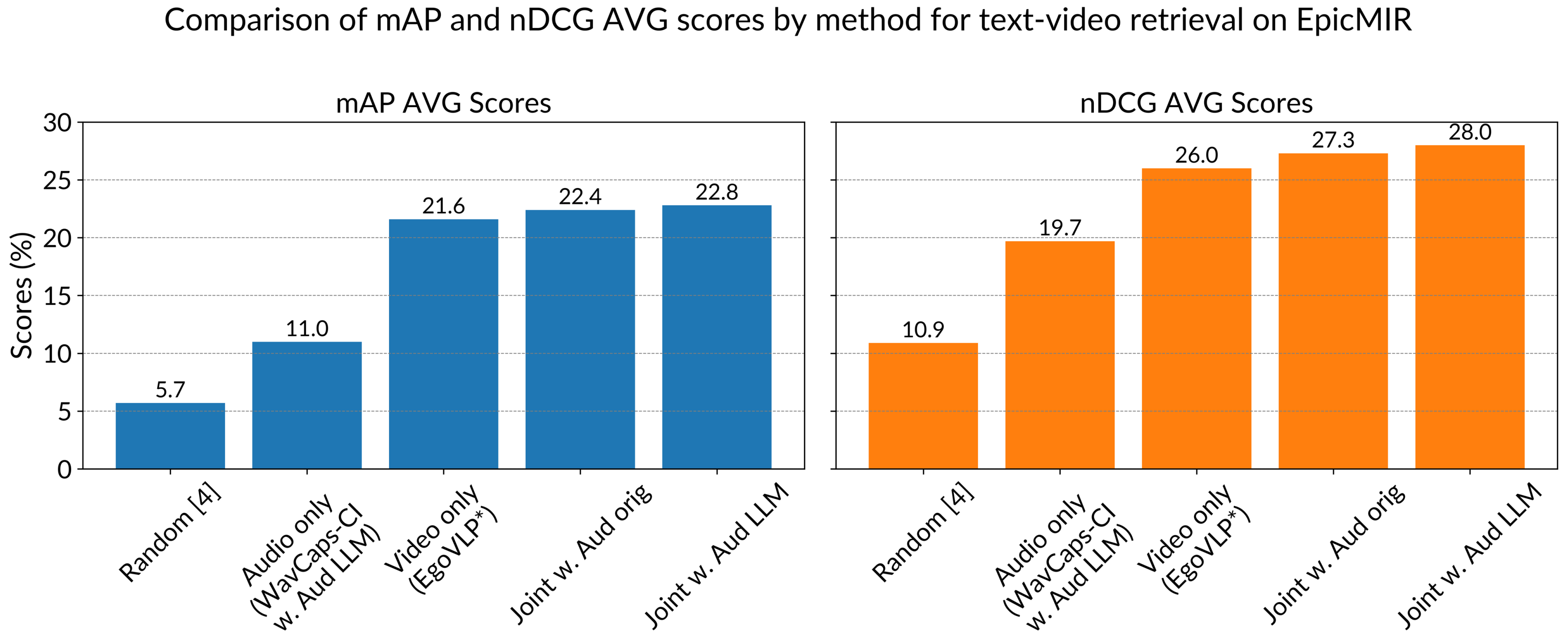
5. Results on AudioEgoMCQ: Text-to-audio retrieval for different subsets (according to audio informativeness)



5. Results on EpicSounds: Comparing LLM-generated descriptions to audio class labels



5. Results on EpicMIR: Benefits of audio for text-to-video retrieval



6. References

[1] D. Damen et al, “Rescaling egocentric vision”, IJCV, 2022
[2] K. Q. Lin et al, “Egocentric videolanguage pretraining”, NeurIPS, 2022
[3] J. Huh et al, “EPIC-SOUNDS: A Large-Scale Dataset of Actions that Sound”, ICASSP, 2023
[4] Y. Wu et al, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation”, ICASSP 2023
[5] X. Mei et al, ”Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research”, 2023
[6] C. D. Kim et al, “Audiocaps: Generating captions for audios in the wild”, Proc.NACCL, 2019
[7] L. Smaira et al., “A short note on the kinetics-700-2020 human action dataset”, 2020

Code (PyTorch): https://github.com/oncescuandreea/audio_egovlp oncescu@robots.ox.ac.uk