# Fine-grained Audible Video Description


- *Xuyang Shen[2], *Dong Li[1], *Jinxing Zhou[3], Zhen Qin[2], Bowen He[2], Xiaodong Han[2], Aixuan Li[4]
- Yuchao Dai[4], Lingpeng Kong[5], Meng Wang[3], Qiao Yu[1], ✆Yiran Zhong[1]

- [1]Shanghai AI Lab, [2]OpenNLPLab, [3]Hefei University of Technology
- [4]Northwestern Polytechnical University, [5]The University of Hong Kong



Poster No.225 - WED-AM-225

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# CONTENT


1 Motivation

2 FAVD Task

3 FAVDBench

4 Baseline

5 Future Work

# Motivation

For a human to perceive the world, sight, sound, and language are the most crucial sense modalities.

However, audio is often neglected in today's video understanding and generation.

**When was the last time you watched a mime?**

Charlie Chaplin

# Fine-grained Audible Video Description

**Goal:** provide detailed textual descriptions for the given audible videos, including the appearance and spatial locations of each object, the actions of moving objects, and the sounds in a video.

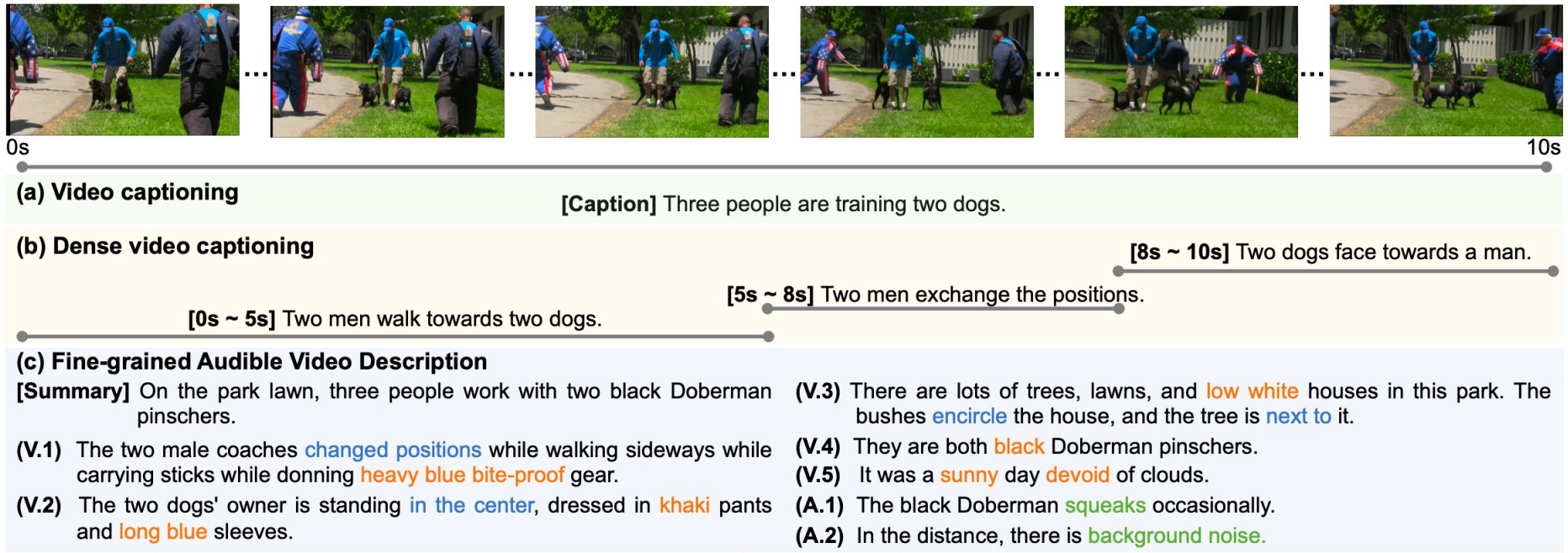
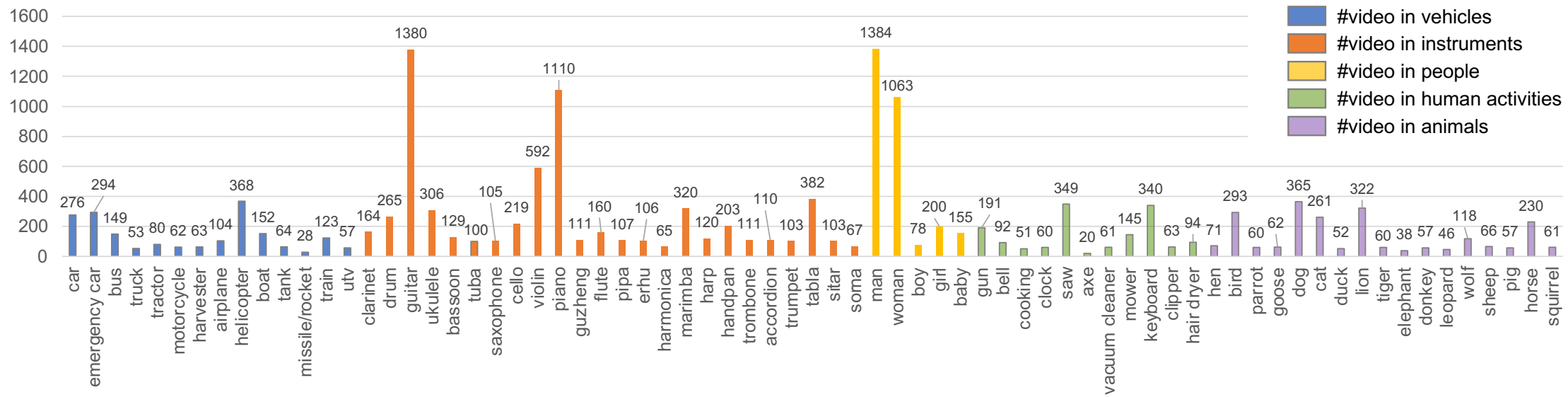



Figure 1. **Comparison of the proposed FAVD task with existing captioning tasks.** (a) Video captioning (VC) uses one sentence to describe the main content of the video. (b) Dense video captioning aims to localize the multiple temporal events and generate corresponding descriptions. Both VC and DVC describe the salient events in videos while losing many details, such as the appearance of objects, spatial relations, and sounds. (c) The proposed FAVD tries to generate a paragraph-level description that contains the caption, named as **Summary**, and the audio-visual descriptions, abbreviated as **A.** and **V.**.

# Fine-grained Audible Video Description

- 5 Major Categories
- 71 Minor Categories
- 11,424 Video Clips
- 24.4 hours
- 2,600,000 Frames
- 40% multiple audio sources



# Fine-grained Audible Video Description



- 5 Major Categories
- 71 Minor Categories
- 11,424 Video Clips
- 24.4 hours
- 2,600,000 Frames
- 40% multiple audio sources

| Dataset | Video | | Annotation | | | | POS tag | | |
|---|---|---|---|---|---|---|---|---|---|
| | #Clip | DUR. (h) | Audio | #Sentence | #Word | Vocabulary | %Adj. | %Noun | %Prep. |
| MSVD [5] | 1,970 | 5.3 | ✘ | 35.5 | 308.3 | 13,010 | 2.6 | 31.8 | 7.7 |
| MSR-VTT [81] | 10,000 | 41.2 | ✘ | 20.0 | 185.7 | 29,316 | 4.8 | 33.9 | 11.5 |
| VATEX [79] | 41,250 | 114.6 | ✘ | 20.0 | 291.8 | 82,654 | 4.4 | 31.8 | 12.4 |
| TVC [29] | 21,793 | 461.3 | ✘ | 5.0 | 67.0 | 57,100 | 2.2 | 36.4 | **12.7** |
| YouCookII [90] | 15,433 | 176.0 | ✘ | 1.0 | 7.9 | 2,583 | 4.1 | **40.3** | 11.6 |
| FAVDBench | 11,424 | 24.4 | ✔ | 12.6 | 218.9 | 73,245 | **13.0** | 30.5 | 12.4 |

# Fine-grained Audible Video Description

## The Baseline Method - AVLFormer

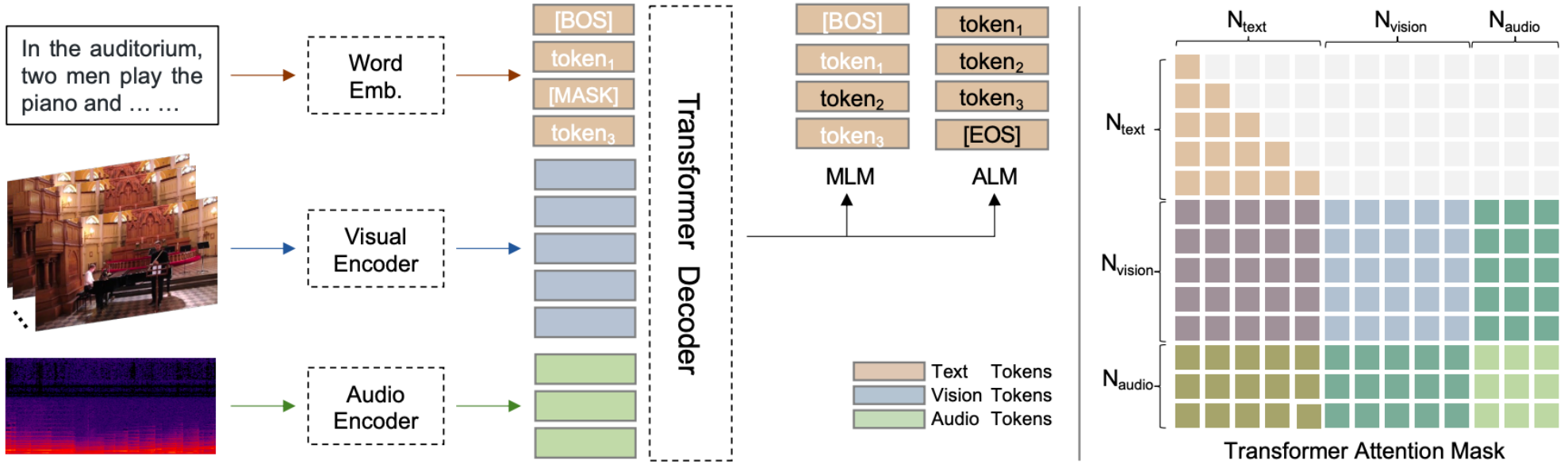



Figure 4. **Overview of AVLFormer.** It consists of a word embedding, a visual encoder, an audio encoder, and a transformer decoder. We adopt the video swin transformer and patchout audio transformer as the visual encoder and audio encoder, respectively. They extract visual and audio features from video frames and audio. Masked language modeling and auto-regressive language modeling are configured in training. The attention mask strategy of AVLFormer is illustrated on the right, where the masked attention is colored in gray. The tokens and attention masks of text, vision, and audio are colored brown, blue, and green, respectively.

# Fine-grained Audible Video Description



## Two new metrices

Existing **BLEU**, **ROUGE**, **Meteor**, and **CIDEr** metrices often concentrate on word-for-word precision by measuring the token similarity between the generated and ground truth texts, whereas we need to measure the similarity at content level.

**EntityScore** measures the extent to which consistently referred words or series of words, known as entities and often manifested as nouns, in the predicted text match those in the annotated text.

$$R(\mathbf{p},\mathbf{r}) = \frac{\#\{\mathbf{p}\cap\mathbf{r}\}}{\#\{\mathbf{r}\}}, C(\mathbf{p},\mathbf{r}) = \frac{\cos(\mathrm{T5}(\mathbf{p}),\mathrm{T5}(\mathbf{r}))+1}{2},$$
$$\mathbb{ES}(\mathbf{p},\mathbf{r}) = \frac{2R(\mathbf{p},\mathbf{r})C(\mathbf{p},\mathbf{r})}{R(\mathbf{p},\mathbf{r})+C(\mathbf{p},\mathbf{r})},$$

**AudioScore** assesses the accuracy of audio descriptions by computing the product of the extracted audio-visual-text unit features, where CLIP is used to extract features for video frames and the corresponding descriptions and PaSST is used for audio waves.

$$\mathbf{e}_a = \mathrm{PaSST}(\mathbf{A}), \mathbf{e}_v = \mathrm{CLIP}(\mathbf{V}), \mathbf{e}_t = \mathrm{CLIP}(\mathbf{T}),$$
$$s = \left(\frac{1}{2}\cos(\mathbf{e}_a,\mathbf{e}_t) + \frac{1}{2}\cos(\mathbf{e}_a,\mathbf{e}_v) + 1\right) \times 0.5,$$
$$\mathbb{AS}(\mathbf{A},\mathbf{V},\mathbf{T}) = \mathbf{f}(s), \mathbf{f}(x) = a\exp(-b\exp(-cx)),$$

# Fine-grained Audible Video Description

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

## Results


**Reference:** The yacht was being driven by one man while the other man stood stationary. Driving the yacht on the left side is a man with a short hair, sunglasses, and a white shirt. On the right side of the yacht, a man with short black hair, sunglasses, and red jeans is immobile. The two were seated on a blue-and-white yacht with two white engines in the back and a white interior. The beach, lush trees, people, and buildings are to the left of the two, while the sea is in front of them. It's sunny and clear outside. There is a noisy yacht. The video includes the sound of pulsing water.

**PDVC:** A man drives a yacht on the sea. A man with a white hat and a white hat and a white hat is driving on the sea. There are many people on the shore. The sound of waves waves.

**SwinBERT:** On the water a speedboat is traveling quickly. On the water a speedboat with a white body and a white bottom is moving quickly. On the speedboat there are both tall buildings trees and tall buildings. The speedboat is incredibly loud.

**BMT:** On the water, a speedboat is moving. On the water, a white speedboat with a white bottom and a white bottom is moving quickly. On the speedboat, there are a lot of buildings, and trees. The speedboat is moving quickly. The speedboat is moving quickly. The sound of the speedboat 's engine may be heard.

**AVLFormer:** On the water a man pilots a yacht. A man in a blue shirt and sunglasses is operating a white speedboat. On the yacht there is a man wearing glasses and a woman with glasses. On the yacht there are a lot of rocks. The harbor beneath the yacht is covered in several reefs. The boat is moving swiftly across the water creating a buzzing sound as it does so.

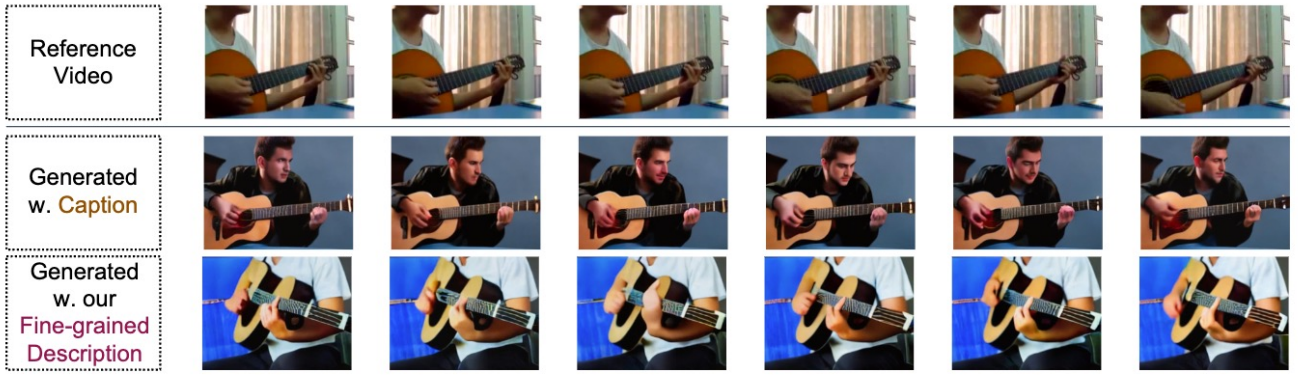
# Fine-grained Audible Video Description



## Video generation


Caption: A lion was lying on the ground outside, roaring incessantly.

Fine-grained Description: A roaring male lion with a yellow and black mane, dark brown fur, and yellow eyes lay on the ground. The lions are surrounded by lush green vegetation. Behind the lion, there are many trees with yellow and dark green leaves.




Caption: A man is singing and playing his guitar while seated at a blue table.

Fine-grained Description: The guitarist is a man with yellow complexion, wearing white short sleeves, and holding the guitar in both hands. The man is holding a brown top guitar with a black bridge that is being played. Behind the left-facing man, beige drapes are hanging. The man is seated in front of a blue table.

# Thanks

Project Page

GitHub

Follow us by

https://www.shlab.org.cn

Shanghai Artificial Intelligence Laboratory