SoundingActions: Learning How Actions Sound from Narrated Egocentric Videos

Changan Chen[1] Kumar Ashutosh[1,2] Rohit Girdhar[2] David Harwath[1] Kristen Grauman[1,2]
[1]University of Texas at Austin [2]FAIR, Meta

8. Supplementary

In this supplementary material, we provide additional details about:

1. Supplementary video for qualitative examples (referenced in Sec. 6).
2. Annotation guidelines and interface (referenced in Sec. 5).
3. Additional implementation details (referenced in Sec. 5).
4. Ablations on anchor modality (referenced in Sec. 6).
5. Ablations on hyperparameter (referenced in Sec. 4).
6. Ablations on the time window length (referenced in Sec. 5).
7. Recall @1 for sounding action retrieval (referenced in Sec. 6).
8. More clusters of visual embeddings (referenced in Sec. 6).

8.1. Supplementary Video

In this video, we include examples of Ego4D clips, qualitative examples of sounding action discovery, and examples of sounding action retrieval. Wear headphones to hear the sound.

8.2. Annotation Guidelines and Interface

For annotators, we first tried MTurk, which we found too noisy. To get high-quality annotations, we then hired 8 professional annotators to work on the annotation task. Each instructor received annotation training and read the annotation guidelines before annotating. They are instructed to classify whether the foreground action described by the narration is both visible and audible in the clip. We also provided them with some positive examples and negative examples to start with. Fig. 3 shows the annotation interface.

8.3. Additional Implementation Details

Following the setting of previous work [6], we initialize the video encoder with ViT [5] pretrained on ImageNet [1] that has a latent dimension of 768. We use the "distilbert-base-uncased" transformer from Huggingface as our text encoder, which has a latent dimension of 256. For audio encoder, we use AST [4] that has been initialized with ViT [5] pretrained on ImageNet [1]. For the joint embedding space, we project features of audio, video and text into a latent space with dimension 256. During training, we resize the video to 224 × 224 and use 4 frames per clip. For audio, we use a sample rate of 16000. We extract fbank features from the audio waveform with 128 Mel frequency bins, 10 ms frame shift and hanning windows.

	$V\to A$		$A\to V$		$L\to A$		$A\to L$	
MC3	@5	@10	@5	@10	@5	@10	@5	@10
Audio as anchor	**38.4**	**72.8**	**34.4**	**66.3**	46.2	88.5	**37.5**	**73.8**
Video as anchor	38.1	72.4	31.9	62.5	**46.6**	**88.7**	36.3	70.7
Language as anchor	37.1	70.0	**34.4**	66.0	45.7	84.9	29.6	61.2

Table 1. Ablations on the anchor modality.

	$V\to A$		$A\to V$		$L\to A$		$A\to L$	
α_v	@5	@10	@5	@10	@5	@10	@5	@10
0.1	37.5	71.5	34.0	64.7	46.2	87.9	36.2	71.5
0.25	37.9	72.2	33.9	66.1	**47.4**	87.2	36.1	71.5
0.5	**38.4**	**72.8**	**34.4**	**66.3**	46.2	**88.5**	**37.5**	**73.8**
0.75	37.2	70.7	34.1	65.8	44.2	86.4	36.9	71.4
1.0	37.8	70.6	32.4	62.8	47.3	**88.5**	35.4	70.4
2.0	27.6	52.5	16.4	35.2	43.5	82.0	13.5	24.9

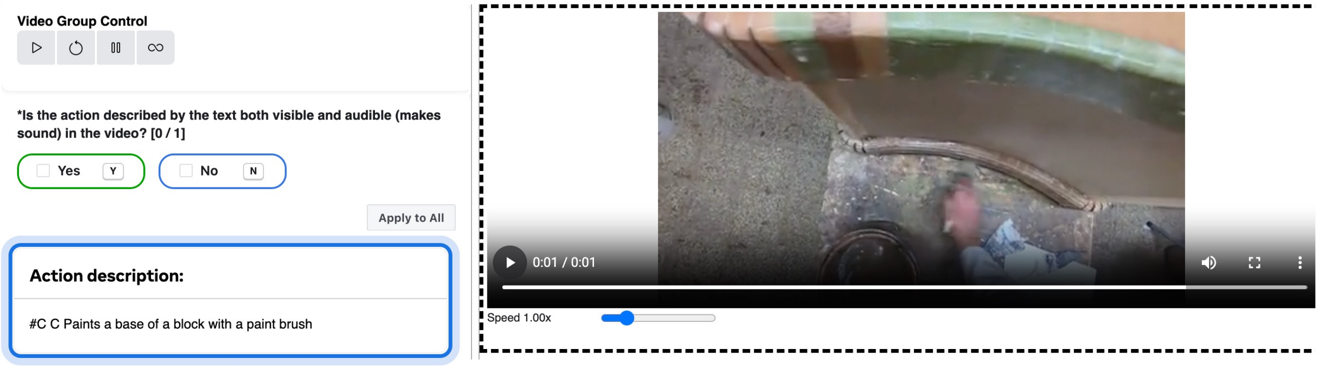
Table 2. Ablations on the hyperparameter.

8.4. Ablations on Anchor Modality

To study the importance of the choice of the anchor modality, we experiment with using video or language as the anchor and report the retrieval performance in Table 5. Using video or language as the anchor modality has a similar but slightly lower performance compared to anchoring audio, likely because audio is generally more ambiguous and thus benefits more from being used as the anchor modality.

8.5. Ablations on Hyperparameters

To make scores from different modality pairs comparable, we use $\mathcal{K}_i(x) = ((x+1)/2)^{\alpha_i}$ to adjust the distribution. Since we set audio as the anchor modality, we only need to tune the α_V and α_L. For tuning, we first set α_L to 1 and then perform a grid search of α_V on the validation data. We report

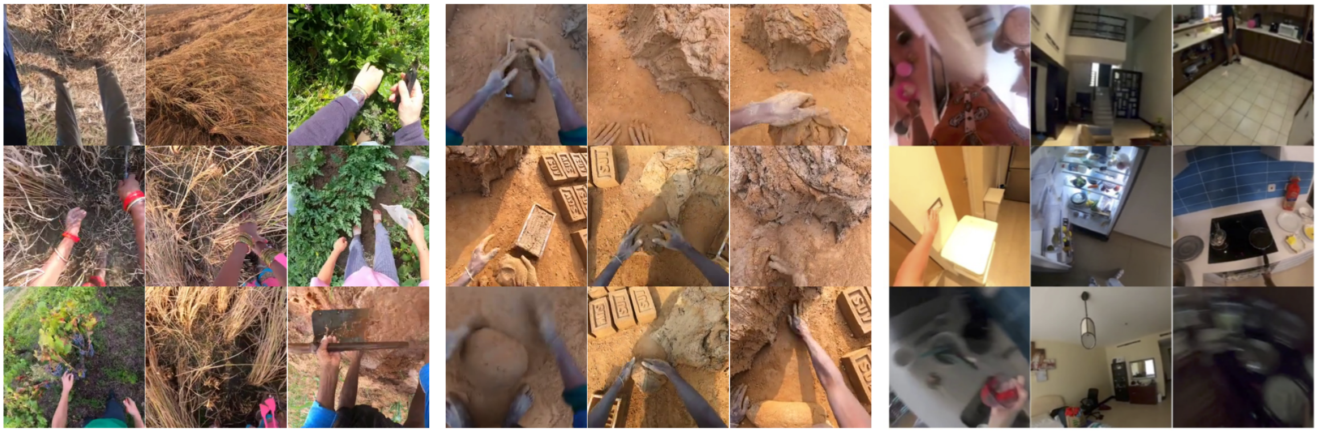


*Is the action described by the text both visible and audible (makes sound) in the video? [0 / 1]

Action description:

#C C Paints a base of a block with a paint brush

Figure 1. Annotation interface.



(a) Actions that make the rustle sound.

(b) Actions that scoop the mud.

(c) Actions that make footstep sound.

Figure 2. More clusters of visual embeddings.

	$V \rightarrow A$		$A \rightarrow V$		$L \rightarrow A$		$A \rightarrow L$	
	@5	@10	@5	@10	@5	@10	@5	@10
0.5 s	26.0	49.0	18.0	36.2	32.0	58.5	20.3	39.9
1.0 s	32.8	62.3	28.7	54.8	42.1	80.1	32.1	62.2
1.5 s	**38.4**	**72.8**	**34.4**	**66.3**	**46.2**	**88.5**	**37.5**	**73.8**
2.0 s	34.3	64.2	30.8	60.7	37.2	71.0	29.8	58.8

Table 3. Ablations on the time window length.

the retrieval performance of all values in Tab. 6. We chose 0.5 as the weight since it achieves the best performance.

8.6. Ablations on Time Window Length

Narrations are timestamped and the action sound (if any) happens within a time window of the timestamp. For the duration of the time window, we consider a few values (0.5 s, 1.0 s, 1.5 s, and 2.0 s) and report their retrieval performance in Tab. 7. Choosing a 1.5 s window leads to the best performance, which is likely because too short time windows can often miss the action sound while long time windows would introduce noise or other action sounds. However, our model is not super sensitive to the choice of the window length since it also performs well with other lengths.

	$V \rightarrow A$		$A \rightarrow V$		$L \rightarrow A$		$A \rightarrow L$	
	@1	@5	@1	@5	@1	@5	@1	@5
Random	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1
CLAP [2]	-	-	-	-	10.8	**49.8**	6.5	34.0
CM-ACC [7]	7.7	34.6	6.7	30.9	-	-	-	-
CMC [8]	7.6	36.5	**7.6**	33.8	9.7	44.1	6.7	32.8
ImageBind [3]	7.2	32.8	6.1	29.7	8.6	42.6	5.9	30.6
MC3	**7.8**	**38.4**	7.2	**34.4**	**11.3**	46.2	**7.3**	**37.5**

Table 4. Sounding action retrieval. We report *Recall @1 and @5* for different query-retrieval modalities.

8.7. Recall @1 for Sounding Action Retrieval

Due to the space limit in the main, we report Recall @1 for the retrieval experiment in Tab. 8. While the performance gap is small as expected, our model still outperforms baselines on most of the metrics.

8.8. More Clusters of Visual Embeddings

In Sec. 6 of the main, we showed one cluster of visual embeddings. Here we show three more clusters from the same clustering result in Fig. 4. Fig. 4a clusters visual actions that make rustle sounds when interacting with grass/branches, even though some examples have very different backgrounds (yellow vs green). This indicates our model learns to cluster visual actions based on how they sound rather than just how they look. Fig. 4b shows a cluster of visual actions that scoop the mud/dirt. Fig. 4c shows the visual cluster where the walking action produces footsteps. Each cluster has actions with varying degrees of head and hand movement, and our model still captures accurately how actions make sounds despite the movement.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4

[2] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022. 2, 5

[3] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 2, 5

[4] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 1, 4

[5] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. In *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, 2021. 1, 4

[6] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 1, 4

[7] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *ICLR*, 2021. 2, 5

[8] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2, 5

8. Supplementary

In this supplementary material, we provide additional details about:

1. Supplementary video for qualitative examples (referenced in Sec. 6).
2. Annotation guidelines and interface (referenced in Sec. 5).
3. Additional implementation details (referenced in Sec. 5).
4. Ablations on anchor modality (referenced in Sec. 6).
5. Ablations on hyperparameter (referenced in Sec. 4).
6. Ablations on the time window length (referenced in Sec. 5).
7. Recall @1 for sounding action retrieval (referenced in Sec. 6).
8. More clusters of visual embeddings (referenced in Sec. 6).

8.1. Supplementary Video

In this video, we include examples of Ego4D clips, qualitative examples of sounding action discovery, and examples of sounding action retrieval. Wear headphones to hear the sound.

8.2. Annotation Guidelines and Interface

For annotators, we first tried MTurk, which we found too noisy. To get high-quality annotations, we then hired 8 professional annotators to work on the annotation task. Each instructor received annotation training and read the annotation guidelines before annotating. They are instructed to classify whether the foreground action described by the narration is both visible and audible in the clip. We also provided them with some positive examples and negative examples to start with. Fig. 3 shows the annotation interface.

8.3. Additional Implementation Details

Following the setting of previous work [6], we initialize the video encoder with ViT [5] pretrained on ImageNet [1] that has a latent dimension of 768. We use the "distilbert-base-uncased" transformer from Huggingface as our text encoder, which has a latent dimension of 256. For audio encoder, we use AST [4] that has been initialized with ViT [5] pretrained on ImageNet [1]. For the joint embedding space, we project features of audio, video and text into a latent space with dimension 256. During training, we resize the video to 224×224 and use 4 frames per clip. For audio, we use a sample rate of 16000. We extract fbank features from the audio waveform with 128 Mel frequency bins, 10 ms frame shift and hanning windows.

8.4. Ablations on Anchor Modality

To study the importance of the choice of the anchor modality, we experiment with using video or language as the anchor and report the retrieval performance in Table 5. Using video or language as the anchor modality has a similar but slightly lower performance compared to anchoring audio, likely because audio is generally more ambiguous and thus benefits more from being used as the anchor modality.

	$V{\to}A$		$A{\to}V$		$L{\to}A$		$A{\to}L$	
MC3	@5	@10	@5	@10	@5	@10	@5	@10
Audio as anchor	**38.4**	**72.8**	**34.4**	**66.3**	46.2	88.5	**37.5**	**73.8**
Video as anchor	38.1	72.4	31.9	62.5	**46.6**	**88.7**	36.3	70.7
Language as anchor	37.1	70.0	**34.4**	66.0	45.7	84.9	29.6	61.2

Table 5. Ablations on the anchor modality.

	$V{\to}A$		$A{\to}V$		$L{\to}A$		$A{\to}L$	
α_v	@5	@10	@5	@10	@5	@10	@5	@10
0.1	37.5	71.5	34.0	64.7	46.2	87.9	36.2	71.5
0.25	37.9	72.2	33.9	66.1	**47.4**	87.2	36.1	71.5
0.5	**38.4**	**72.8**	**34.4**	**66.3**	46.2	**88.5**	**37.5**	**73.8**
0.75	37.2	70.7	34.1	65.8	44.2	86.4	36.9	71.4
1.0	37.8	70.6	32.4	62.8	47.3	**88.5**	35.4	70.4
2.0	27.6	52.5	16.4	35.2	43.5	82.0	13.5	24.9

Table 6. Ablations on the hyperparameter.

	$V{\to}A$		$A{\to}V$		$L{\to}A$		$A{\to}L$	
	@5	@10	@5	@10	@5	@10	@5	@10
0.5 s	26.0	49.0	18.0	36.2	32.0	58.5	20.3	39.9
1.0 s	32.8	62.3	28.7	54.8	42.1	80.1	32.1	62.2
1.5 s	**38.4**	**72.8**	**34.4**	**66.3**	**46.2**	**88.5**	**37.5**	**73.8**
2.0 s	34.3	64.2	30.8	60.7	37.2	71.0	29.8	58.8

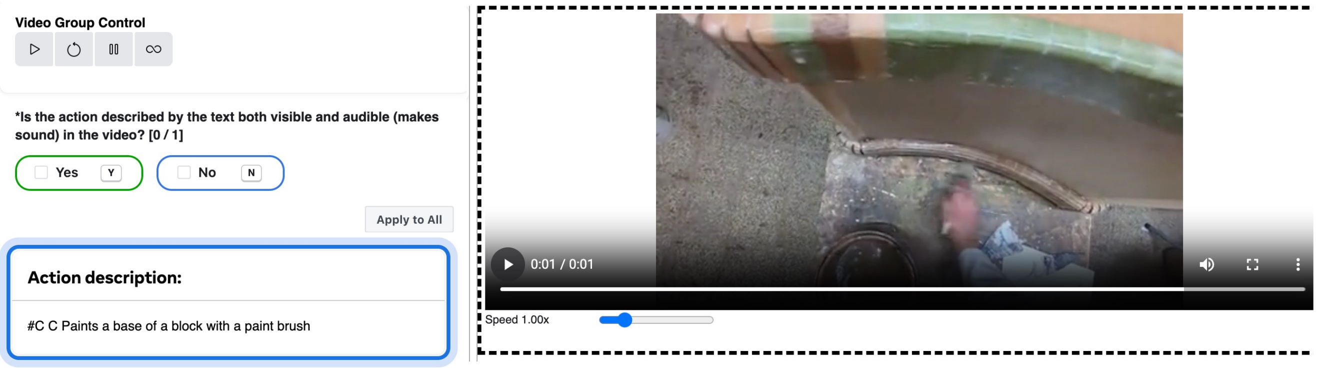
Table 7. Ablations on the time window length.

8.5. Ablations on Hyperparameters

To make scores from different modality pairs comparable, we use $\mathcal{K}_i(x) = ((x+1)/2)^{\alpha_i}$ to adjust the distribution. Since we set audio as the anchor modality, we only need to tune the α_V and α_L. For tuning, we first set α_L to 1 and then perform a grid search of α_V on the validation data. We report the retrieval performance of all values in Tab. 6. We chose 0.5 as the weight since it achieves the best performance.

8.6. Ablations on Time Window Length

Narrations are timestamped and the action sound (if any) happens within a time window of the timestamp. For the duration of the time window, we consider a few values (0.5 s, 1.0 s, 1.5 s, and 2.0 s) and report their retrieval performance in Tab. 7. Choosing a 1.5 s window leads to the best performance, which is likely because too short time windows can often miss the action sound while long time windows would introduce noise or other action sounds. However, our model



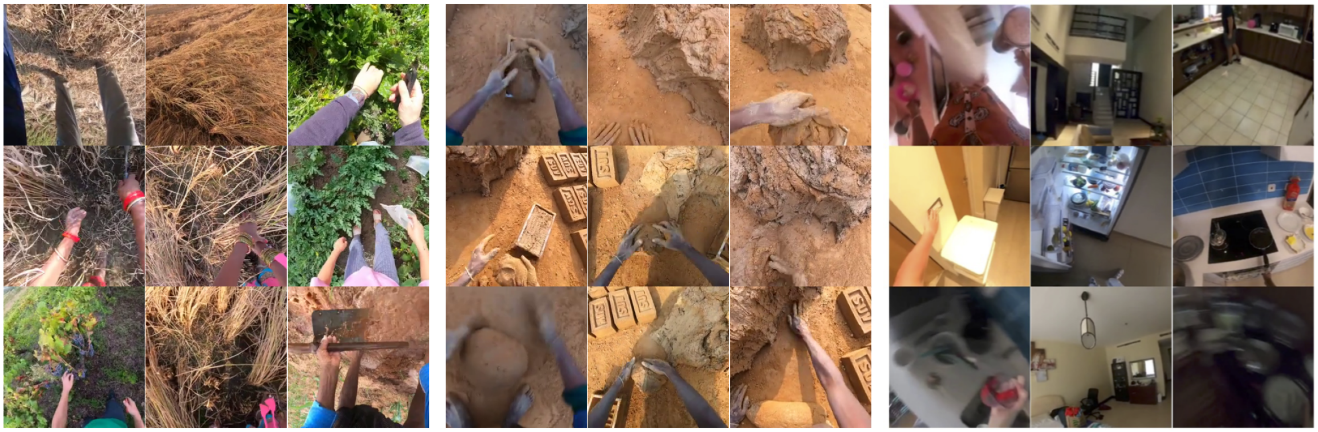
*Is the action described by the text both visible and audible (makes sound) in the video? [0 / 1]

☐ Yes ☐ No

Action description:

#C C Paints a base of a block with a paint brush

Figure 3. Annotation interface.



(a) Actions that make the rustle sound.

(b) Actions that scoop the mud.

(c) Actions that make footstep sound.

Figure 4. More clusters of visual embeddings.

	$V\rightarrow A$		$A\rightarrow V$		$L\rightarrow A$		$A\rightarrow L$	
	@1	@5	@1	@5	@1	@5	@1	@5
Random	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1
CLAP [2]	-	-	-	-	10.8	**49.8**	6.5	34.0
CM-ACC [7]	7.7	34.6	6.7	30.9	-	-	-	-
CMC [8]	7.6	36.5	**7.6**	33.8	9.7	44.1	6.7	32.8
ImageBind [3]	7.2	32.8	6.1	29.7	8.6	42.6	5.9	30.6
MC3	**7.8**	**38.4**	7.2	**34.4**	**11.3**	46.2	**7.3**	**37.5**

Table 8. Sounding action retrieval. We report *Recall @1 and @5* for different query-retrieval modalities.

is not super sensitive to the choice of the window length since it also performs well with other lengths.

8.7. Recall @1 for Sounding Action Retrieval

Due to the space limit in the main, we report Recall @1 for the retrieval experiment in Tab. 8. While the performance gap is small as expected, our model still outperforms baselines on most of the metrics.

8.8. More Clusters of Visual Embeddings

In Sec. 6 of the main, we showed one cluster of visual embeddings. Here we show three more clusters from the same clustering result in Fig. 4. Fig. 4a clusters visual actions that make rustle sounds when interacting with grass/branches, even though some examples have very different backgrounds (yellow vs green). This indicates our model learns to cluster visual actions based on how they sound rather than just how they look. Fig. 4b shows a cluster of visual actions that scoop the mud/dirt. Fig. 4c shows the visual cluster where the walking action produces footsteps. Each cluster has actions with varying degrees of head and hand movement, and our model still captures accurately how actions make sounds despite the movement.