

# Supplementary Material of Sound Bridge: Associating Egocentric and Exocentric Videos via Audio Cues

## 1. Ablation Study on DTW

For video pairs with high synchronization, we found that aligning the two audio sequences before inputting the audio signals into the sound-vision cross-attention module can reduce the impact of sound at different speeds on the model. Since the paired videos may be captured at different times and locations using different recording equipment, the activity sounds in the two videos may vary in speed, affecting the audio representation.

We applied the Dynamic Time Warping (DTW) algorithm [4] to the CharadesEgo dataset, which contains relatively synchronized video pairs. DTW is used to determine the optimal alignment between egocentric and exocentric audio sequences in synchronized videos. In Table 1, we align the sitting by removing the DTW on all datasets. It is evident that our methods still demonstrate a significant superiority over existing methods.

## 2. Prompt Templates

Section 3.2 of the paper utilizes two prompts:  $Prompt_{Text2Sound}$ , which generates audio descriptions from video descriptions, and  $Prompt_{soundRefine}$ , which refines these descriptions by converting them into CC format and filtering out high-frequency words to create unique outputs for each video. Detailed information can be found in Tables 2 and 3.

## 3. More Visualization Examples for Ego-Exo Retrieval and Recognition Tasks

Figures 1 present multiple examples of mutual retrieval between ego and exo perspectives. As shown in the figures, when visual information is similar, audio information effectively distinguishes correctly matched videos from incorrectly matched ones, thereby improving performance in cross-view video association tasks. This demonstrates the effectiveness of our method.

Figure 2 illustrates examples of our model applied to action recognition tasks in exocentric videos. As shown in the figure, exocentric videos often suffer from occlusion of localized actions, making it challenging to accurately identify the ongoing actions. By leveraging the complementary local information provided by egocentric videos, the model successfully recognizes the actions taking place in the video.

Figure 3 presents examples of our model applied to scene recognition tasks in egocentric videos. As shown in the figure, egocentric videos provide only local visual information, making it challenging for the model to distinguish the correct scene label. By leveraging audio information to associate the global visual information provided by exocentric videos, the model successfully identifies the scene occurring in the current video.

Table 1. Retrieval result comparison (top-1 accuracy) of Exo-Ego video alignment task on two benchmarks.

Method	CharadesEgo		
	Eg2Ex	Ex2Eg	Overall
FrozenInTime [1]	6.73	4.96	5.85
LaViLa [7]	58.24	46.93	52.59
EgoVLP [3]	58.27	60.76	59.52
InternVideo [5]	53.55	61.23	57.39
EgoExo [2]	68.32	62.06	65.19
EgoInstructor [6]	62.06	56.38	59.22
SoundBridge w/o DTW	72.58	63.59	68.09
SoundBridge	<b>77.66</b>	<b>71.63</b>	<b>74.65</b>

---

*Prompt<sub>Text2Sound</sub>*

---

You are an expert in audio description of videos, you will generate audio descriptions based on video descriptions.

1. Based on the following video description, generate a single sentence describing the corresponding sounds.
2. Ensure the audio description is concise and accurate.
3. Make the description sound natural and realistic, even if there are no specific sound tags.
4. Avoid generating audio descriptions in the form of "sound of [visual object/action]" and instead use the actual noise or sound that object/action can make. Additionally, seamlessly integrate these audio descriptions into the corresponding visual descriptions to create a more natural and realistic narrative.
5. Only return the audio description sentence without any additional explanations or comments.

To help you better understand my requirements, I am providing the following two reference examples. Video description 1: Hold the knife in your left hand and press the ginger with your right hand. Cut the ginger into slices. Audio description 1: Hold the knife in your left hand and press the ginger with your right hand. Cut the ginger into slices(scraping). Video description 2: Taking a ladle with your left hand, placing cut eggplant into a hot oil pan. Audio description 2: Taking a ladle with your left hand, placing cut eggplant into a hot oil pan (the sound of oil sizzling and the sound of the eggplant being dropped into the pan).

**Input:** "Video description"

**Output:**

---

Table 2. Audio description generation prompt.

---

*Prompt<sub>soundRefine</sub>*

---

You are an expert of closed caption that refines audio descriptions by generating Closed Caption-style sound descriptions.

1. Embed each sound description directly within the original sentence at the most relevant locations, based on the context, in the format [sound source] (sound description).
2. After embedding the sound descriptions, remove high-frequency words or phrases from an audio description, leaving only a single that contains unique Closed Caption-style descriptions of the sounds.
3. Retaining only the essential and unique sound cues, with sound descriptions that make it easy to identify the source and nature of each sound.
4. Only return the Closed Caption-style audio description as a plain text output, without any additional explanation.

To help you better understand my requirements, I am providing the following two reference examples. Audio description 1: Hold a piece of ginger in your left hand, and with your right hand, use a kitchen knife to cut off a small piece of ginger. Then, use the kitchen knife to slice that small piece of ginger into ginger slices. Closed Caption description 1: [hold ginger](soft rustling), [knife](crisp cutting sound), [knife slicing](steady slicing sounds). Audio description 2: A person is undressing and putting the clothes on a shelf. They finish and start playing with a toy. Closed Caption description 2: [undressing](zipper sound), [putting the clothes on a shelf](clothes rustling), [start playing with a toy](toy squeak).").

**Input:** "Audio description"

**Output:**

---

Table 3. Audio description refinement prompt.

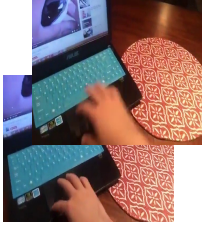
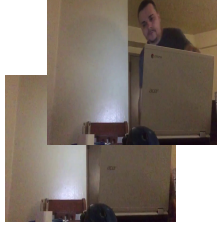
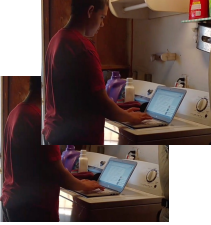
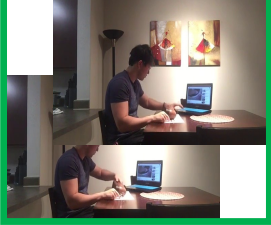
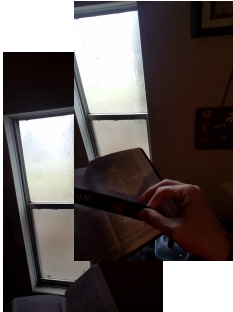
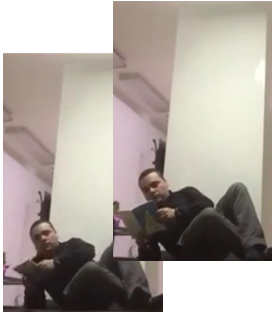

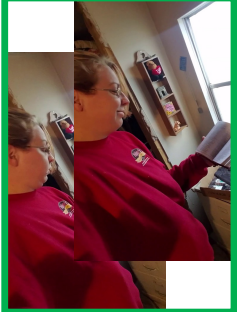
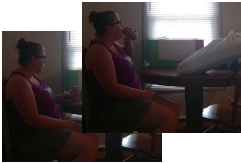
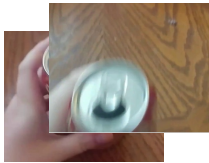
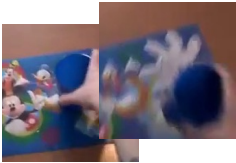
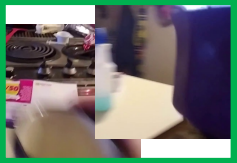
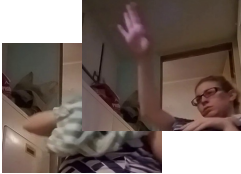
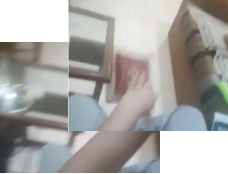

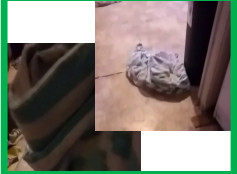
Egocentric View	Exocentric View		
 <p>Query</p> <p>🔊 : The sound from a computer speaker explaining a video about repairing shoes.</p>	 <p>EgoExo</p> <p>🔊 : The sound of touching the laptop.</p>	 <p>EgoInstruct</p> <p>🔊 : The sound of typing on a computer keyboard.</p>	 <p>SoundBridge</p> <p>🔊 : The sound from a computer speaker explaining a video about repairing shoes.</p>
 <p>Query</p> <p>🔊 : The laughter of a person reading a book.</p>	 <p>EgoExo</p> <p>🔊 : The sound of a body lying down.</p>	 <p>EgoInstruct</p> <p>🔊 : The sound from a television behind the person holding a book.</p>	 <p>SoundBridge</p> <p>🔊 : The laughter of a person reading a book.</p>
Exocentric View	Egocentric View		
 <p>Query</p> <p>🔊 : The sound of drinking water.</p>	 <p>EgoExo</p> <p>🔊 : The sounds of drinking water and footsteps.</p>	 <p>EgoInstruct</p> <p>🔊 : The sounds of drinking water and talking on the phone.</p>	 <p>SoundBridge</p> <p>🔊 : The sound of drinking water.</p>
 <p>Query</p> <p>🔊 : The sound of a bag hitting the ground.</p>	 <p>EgoExo</p> <p>🔊 : The sounds of laughter and a pillow hitting the ground.</p>	 <p>EgoInstruct</p> <p>🔊 : The sounds of a camera being placed on a table and shoes hitting the ground.</p>	 <p>SoundBridge</p> <p>🔊 : The sound of a bag hitting the ground.</p>

Figure 1. More Ego-Exo retrieval examples. The green box indicates the correct option.

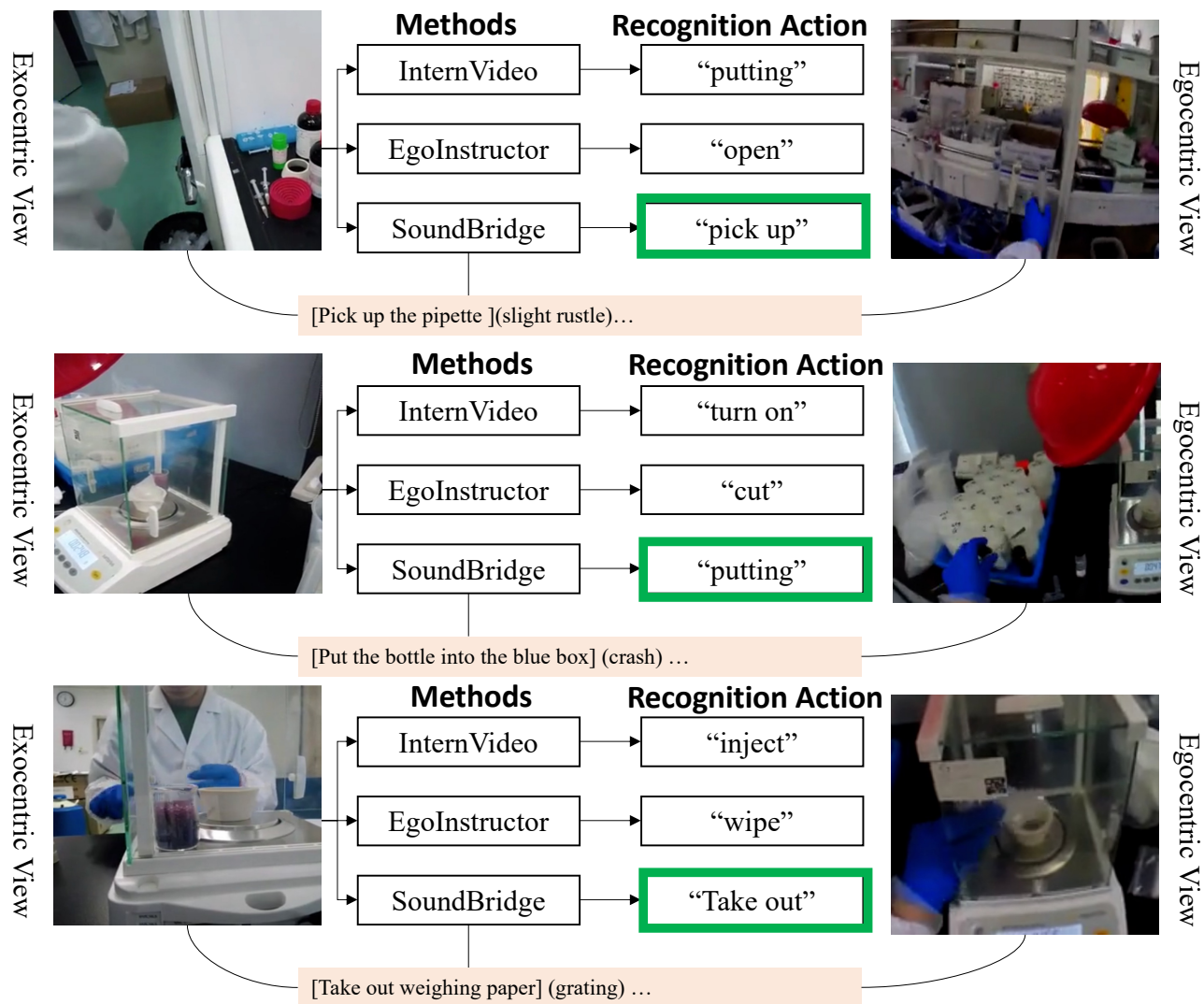


Figure 2. More examples on exocentric action recognition. The green box indicates the correct option.

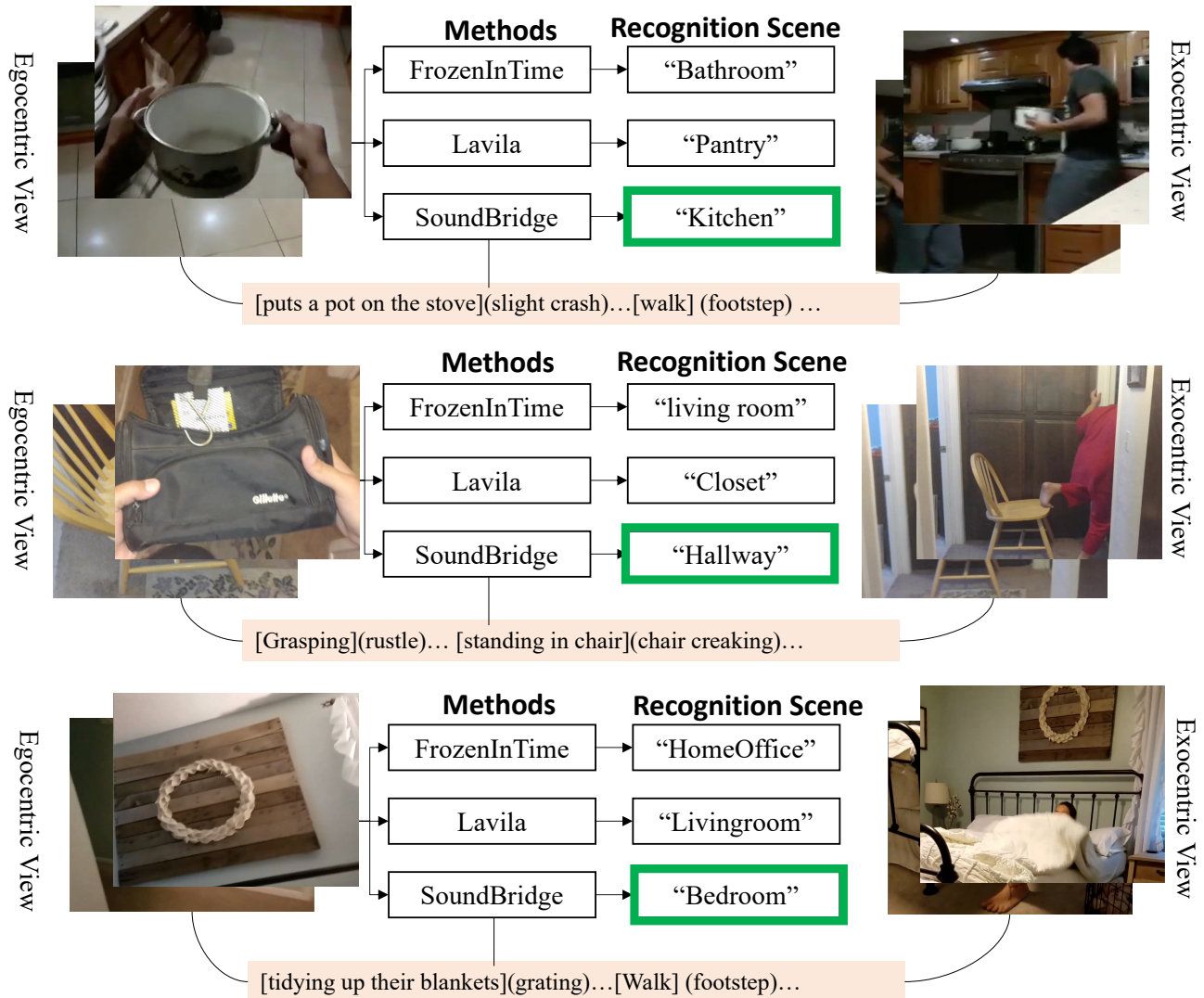


Figure 3. More examples on egocentric scene recognition. The green box indicates the correct option.

## References

- [1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. [1](#)
- [2] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. [1](#)
- [3] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. [1](#)
- [4] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978. [1](#)
- [5] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*, 2024. [1](#)
- [6] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024. [1](#)
- [7] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. [1](#)