The University of Texas at Austin

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]

University of Texas at Austin[1], FAIR, Meta AI[2]

Project page

CVPR SEATTLE, WA JUNE 17-21, 2024

Learning Sounding Actions

Action: C moves a metal cutting machine with hands.

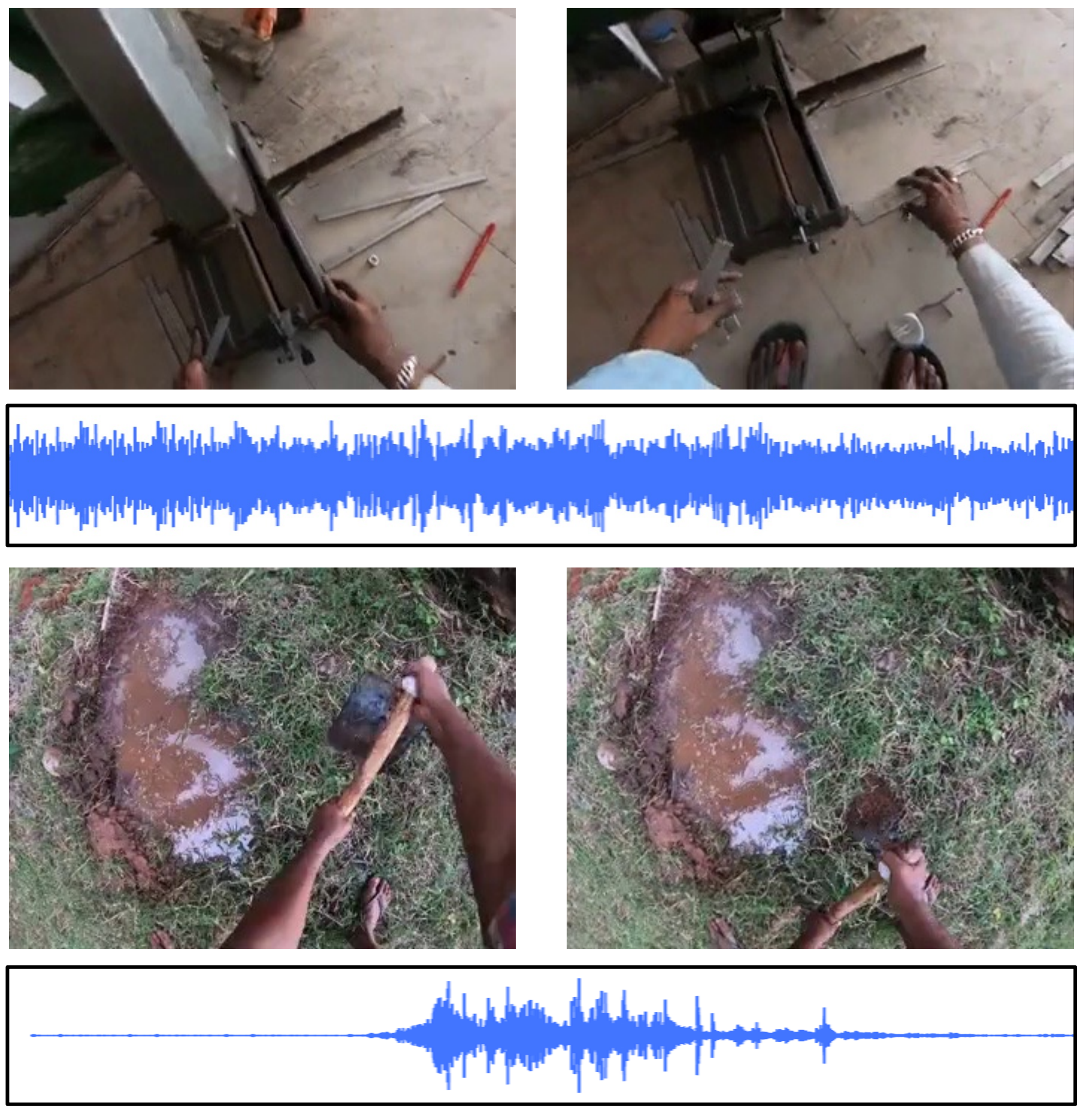
Does the action sound?

NO

Action: C digs the soil with a hoe.

Does the action sound?

YES



Motivation: can we distinguish sounds that are directly caused by human actions from those that are not?

Goal: bring audio-visual embeddings closer when sounds are due to the foreground actions and distance them othewise

Our idea: seek semantic agreement between audio, video and language (action descriptions)

Multimodal Contrastive-Consensus Coding (MC3)

Multimodal Contrastive coding

- Treats modalities (i, j) from the same sample (t) as positive

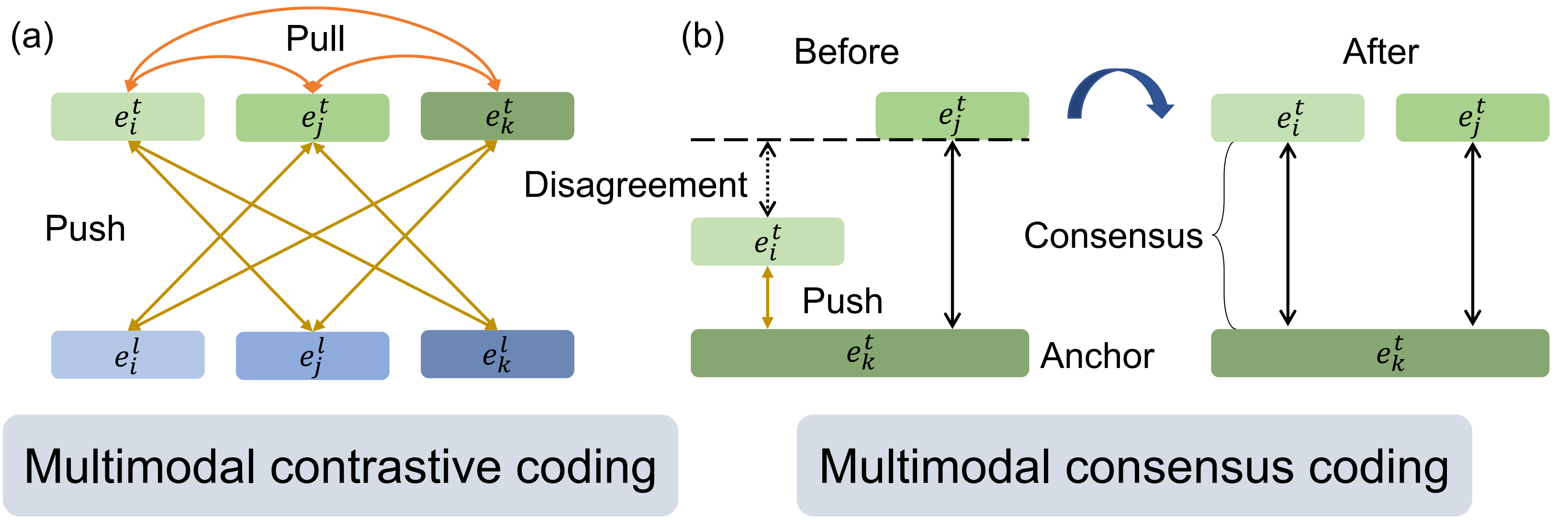
$$L_{contrastive} = \sum_{i,j} \left(-\frac{1}{|B|} \sum_{t \in B} \log \frac{\exp(e_i^t e_j^t / \tau)}{\sum_{l \epsilon B} \exp(e_i^t e_j^l / \tau)}\right)$$

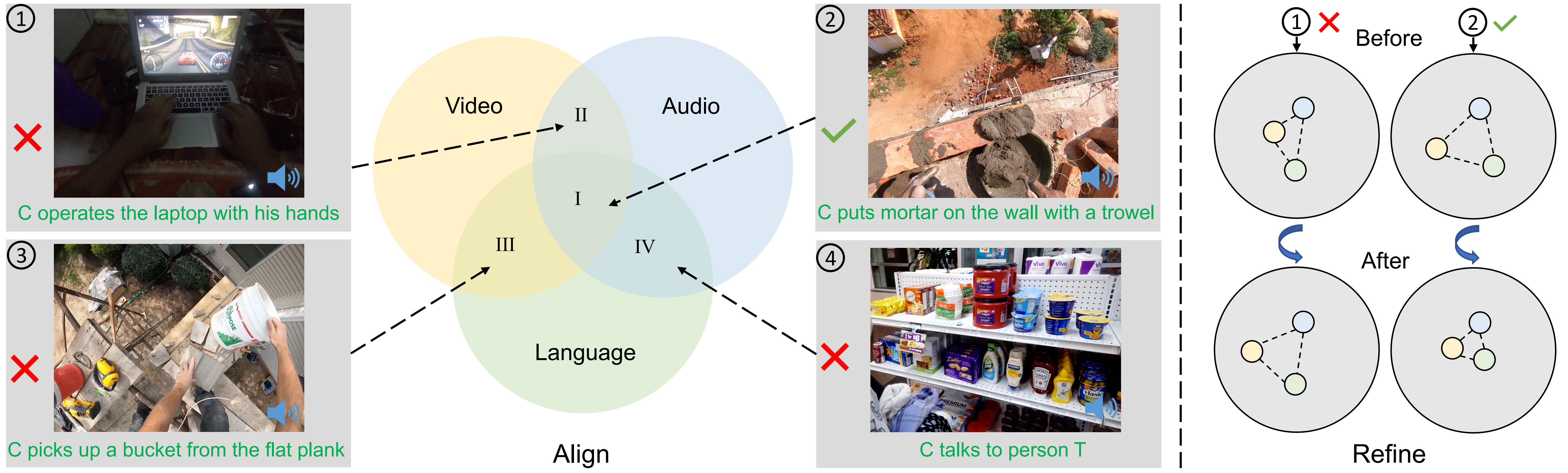
Multimodal Consensus coding

- Compute pairwise similarity score w.r.t. anchor modality a
- Consensus score is the minimum of all pairwise scores

$$c^t = K^{-1}\left(\min_i\left(K_1(e_1^t e_a^t), \ldots, K_n(e_n^t e_a^t)\right)\right)$$

$$L_{consensus} = \frac{1}{|B|} \sum_{t \in B} \sum_{i, i \neq a} \left\|e_i^t e_a^t - c^t\right\|_2$$




1
C operates the laptop with his hands
3
C picks up a bucket from the flat plank
Video
Audio
Language
I
II
III
IV
Align
2
C puts mortar on the wall with a trowel
4
C talks to person T
1
2
Before
After
Refine

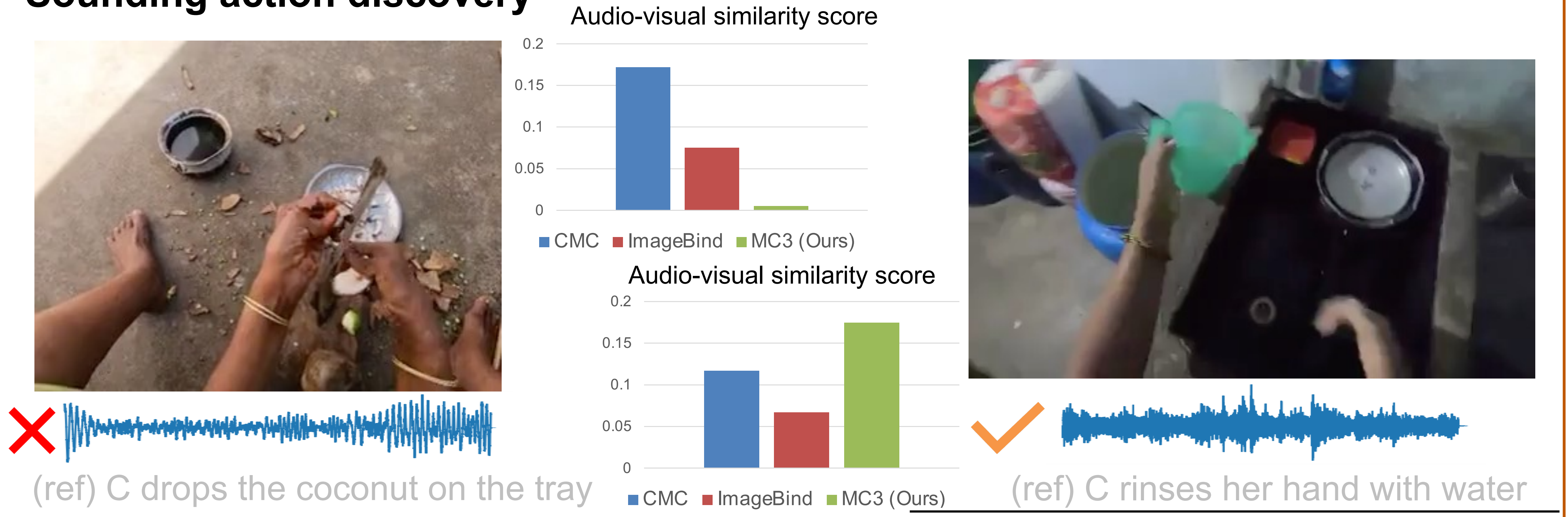
Two-stage training:

- Align stage: pairwise contrastive learning guides modality embeddings to have a good initial alignment ($L_{contrastive}$)
- Refine stage: refine the pairwise aligned embedding with a globally established consensus ($L_{MC3} = L_{contrastive} + L_{consensus}$)

Experiments

- Sample 250K training clips from Ego4D based on narration timestamps
- Annotate 33K clips of whether them containing sounding actions for evaluation

Sounding action discovery



(ref) C drops the coconut on the tray

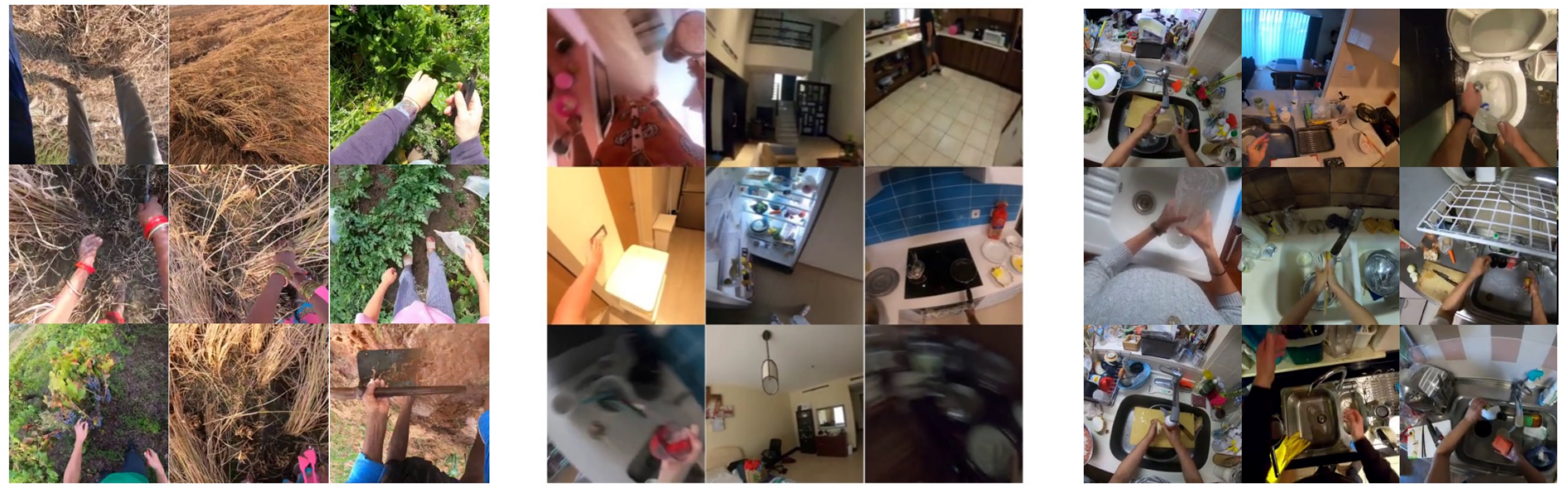
(ref) C rinses her hand with water

- Compute similarity between modality features
- Discover sounding actions better w/ consensus

	🔊	🎥	📄	*AV* ROC	*AV* PR	*AL* ROC	*AL* PR
Random	✗	✗	✗	0.500	0.559	0.500	0.559
CLAP [1]	✓	✗	✓	-	-	0.637	0.695
CM-ACC [2]	✓	✓	✗	0.540	0.590	-	-
CMC [3]	✓	✓	✓	0.550	0.601	0.635	0.693
ImageBind [4]	✓	✓	✓	0.554	0.605	0.642	0.685
MC3	✓	✓	✓	**0.598**	**0.666**	**0.658**	**0.715**

[1] CLAP, Elizalde et al., ICASSP 2023
[2] CM-ACC, Ma et al., ICLR 2021
[3] CMC, Tian et al., ECCV 2020
[4] ImageBind, Girdhar et al., CVPR 2023

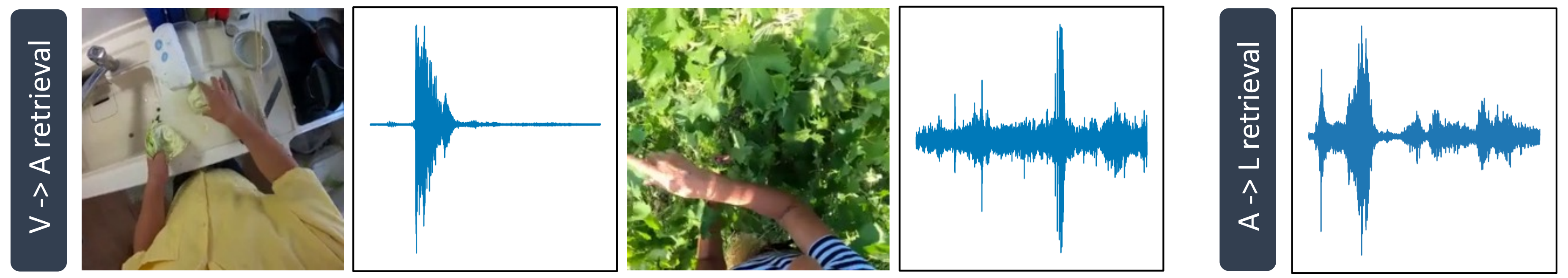
Clustered visual actions



Actions that make rustle sound

Actions that make footstep sound

Actions that make flushing sound



C drops the knife on the chopping board

¹C puts down the knife

C holds the plant upright

¹C holds the plants with the sisal fiber

C moves the clothes on the hanger rail