Learning Camera Viewpoint Selection in Instructional Videos

Sagnik Majumder
University of Texas at Austin

Instructional videos

Videos made for teaching how to do skilled activities

How to pack a suitcase?

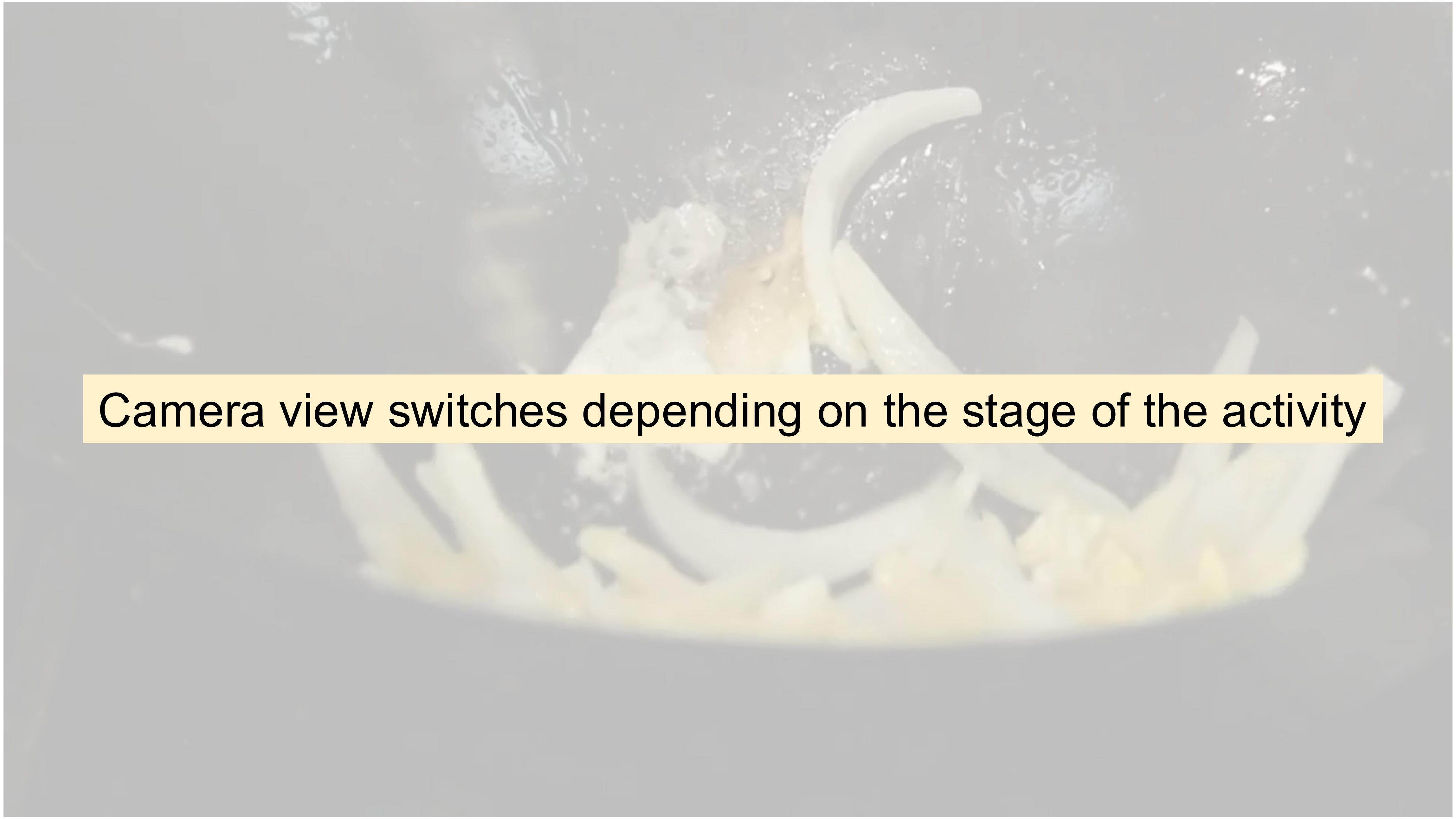
How to apply makeup?

How to hand-make pasta?

How to use a bidet?

Source: Youtube

Typical in-the-wild instructional video



Creating a varying-view instructional video

Source: justinodisho.com

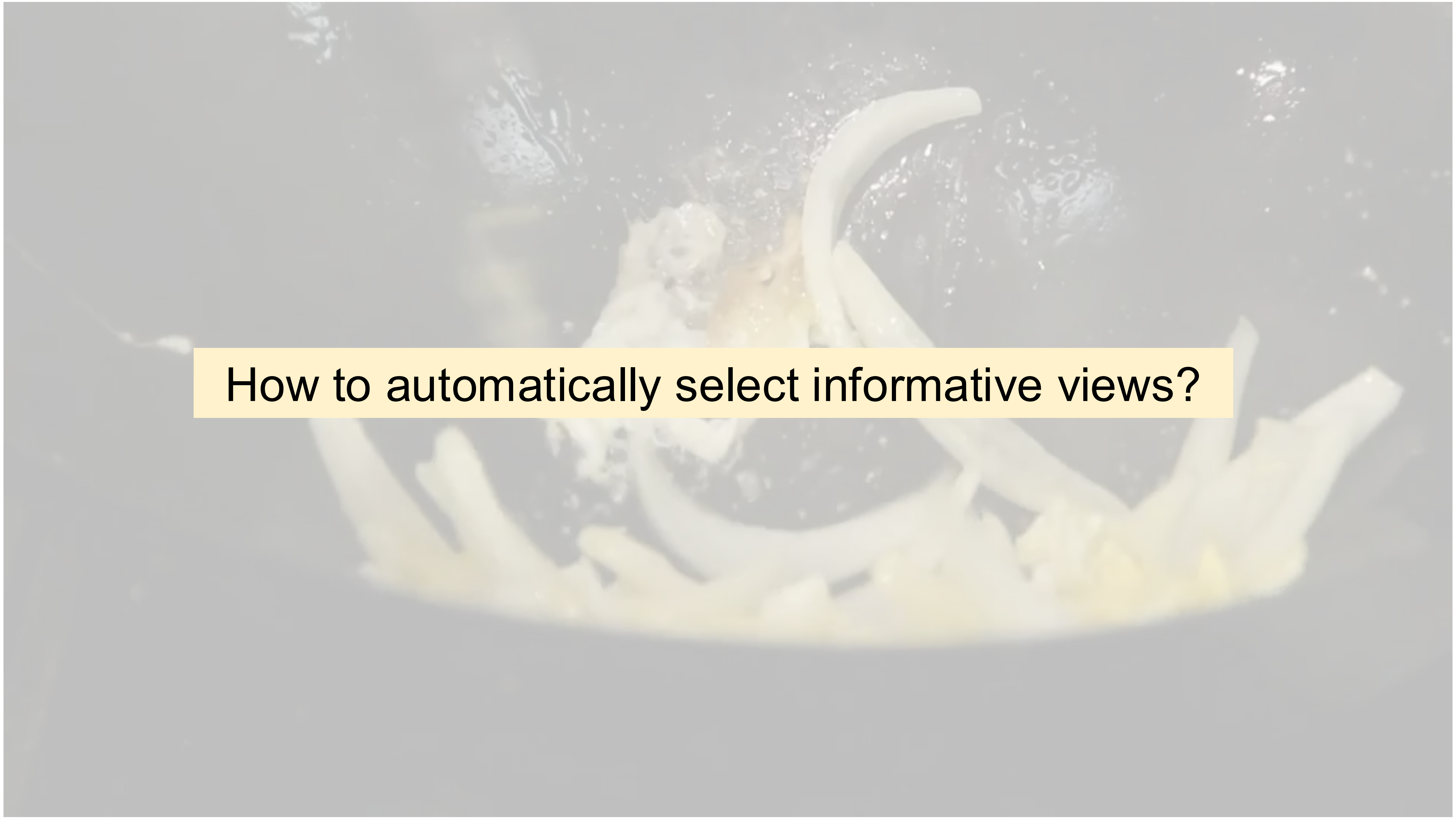
Source: premiumbeat.com

Post-capture editing

Active camerawork

Laborious and time-consuming

View selection in multi-view instructional videos



View selection in multi-view instructional videos

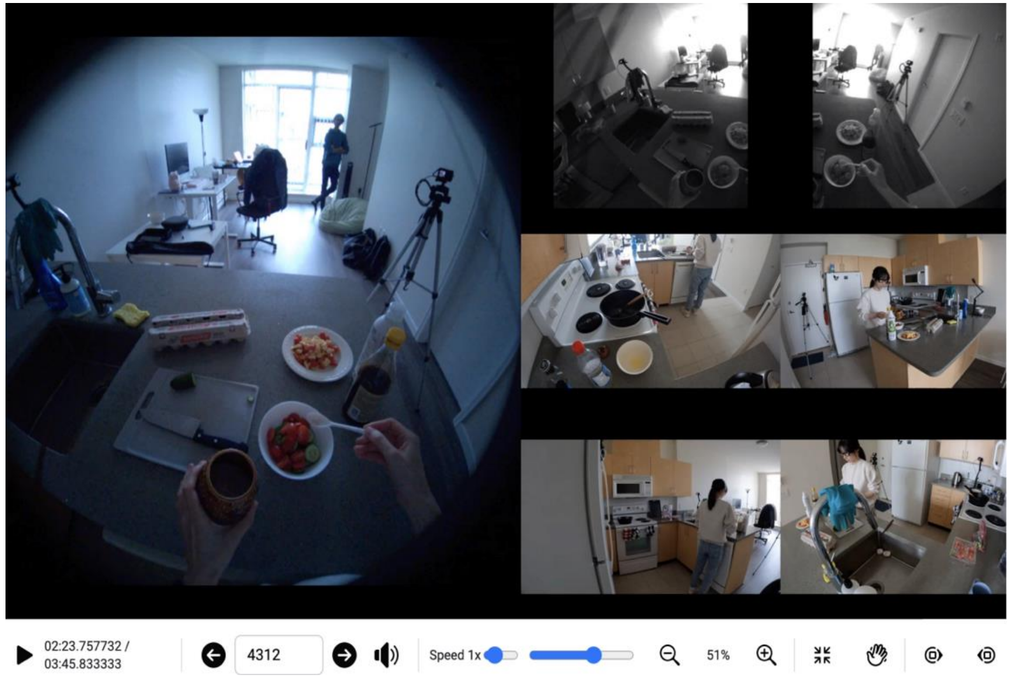
Best view sequence?

Given an instructional video scene captured using multiple cameras, select the sequence of camera views that best illustrates the activity

Narrations

View-independent dense captions describing the instructional activity

Controlled capture (e.g., Ego-Exo4D[1])



The person pours a spoon of salt with their right hand

In-the-wild capture (e.g., HowTo100M [2])

 …  … 

OpenAI Whisper

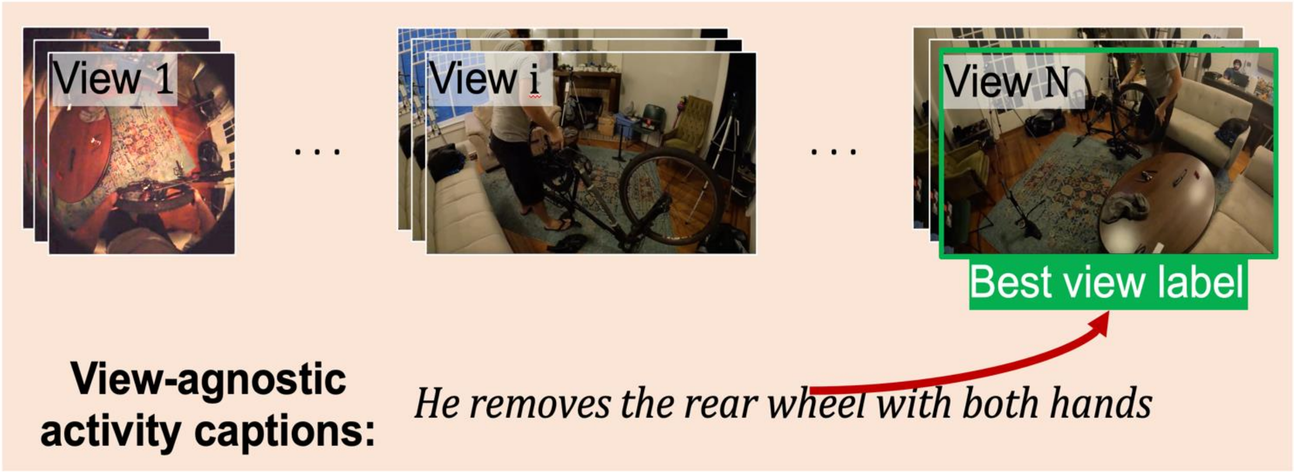
I am adding the sausages to the rice now … *We will use medium heat throughout* … *Stir the rice until you have a uniform mix*

[1] **Ego-Exo4D.** Grauman et al., CVPR 2024.
[2] **HowTo100M.** Miech et al., ICCV 2019.

View selection in label-scarce settings

View selection in label-scarce settings

Main conference: Wed (10/22) AM, poster #185



LangView

... by using captions for producing best-view pseudo-labels during training

Switch-a-View

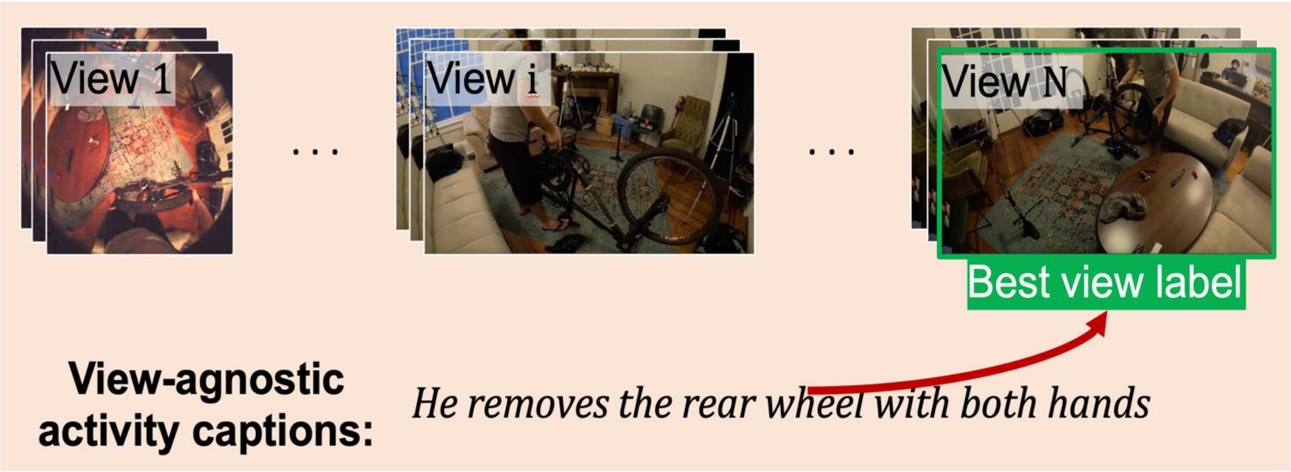
... by learning human view choices from unlabeled but edited in-the-wild how-tos

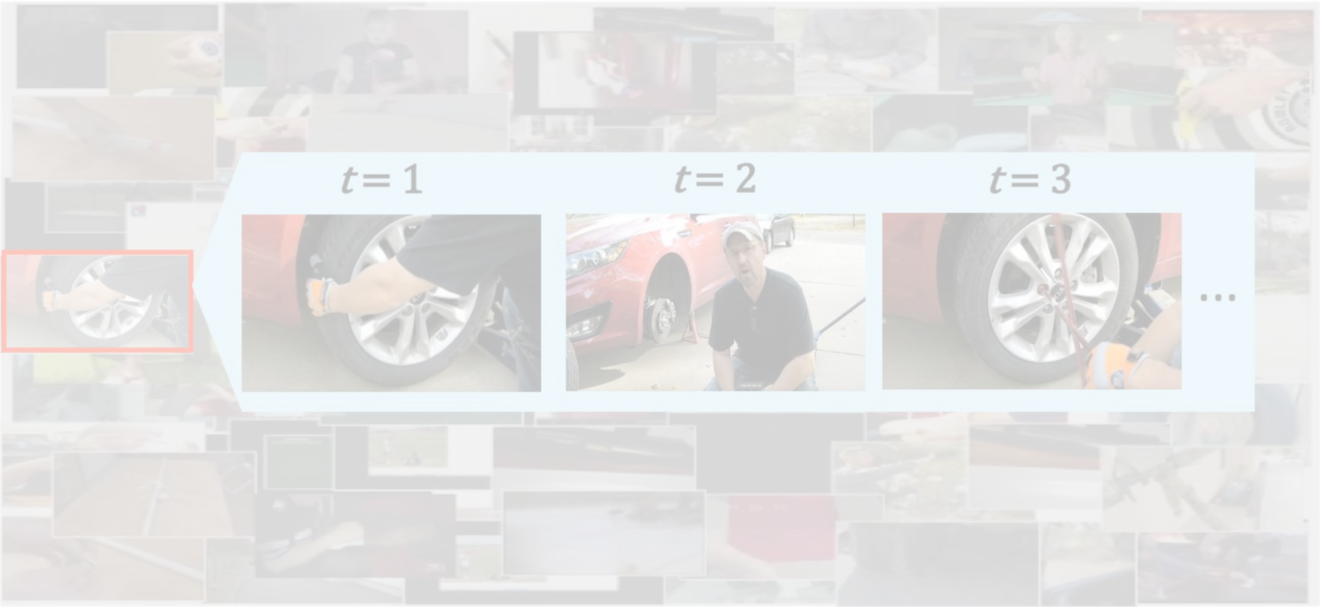
Language for Weakly Supervising View-Selection in Multi-view Instructional Videos
Majumder et al., CVPR 2025 (Highlight).
Project page: https://vision.cs.utexas.edu/projects/which-view-shows-it-best

Switch-a-View: View Selection Learned from Unlabeled In-the-wild Videos
Majumder et al., ICCV 2025.
Project page: https://vision.cs.utexas.edu/projects/switch-a-view/

View selection in label-scarce settings

Main conference: Wed (10/22) AM, poster #185





LangView

… by using captions for producing best-view pseudo-labels during training

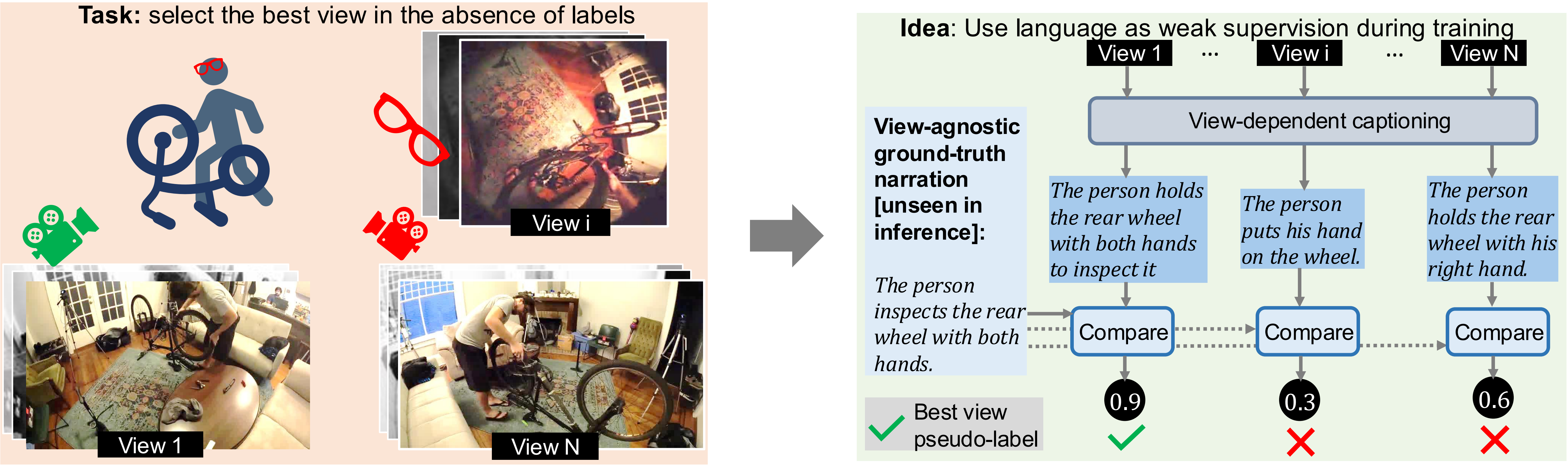
Language for Weakly Supervising View-Selection in Multi-view Instructional Videos
Majumder et al., CVPR 2025 (Highlight).
Project page: https://vision.cs.utexas.edu/projects/which-view-shows-it-best

Switch-a-View

… by learning human view choices from unlabeled but edited in-the-wild how-tos

Switch-a-View: View Selection Learned from Unlabeled In-the-wild Videos
Majumder et al., ICCV 2025.
Project page: https://vision.cs.utexas.edu/projects/switch-a-view/

LangView: weakly supervise via language



Task: given a multi-view instructional video, learn to select the best view without manual labels

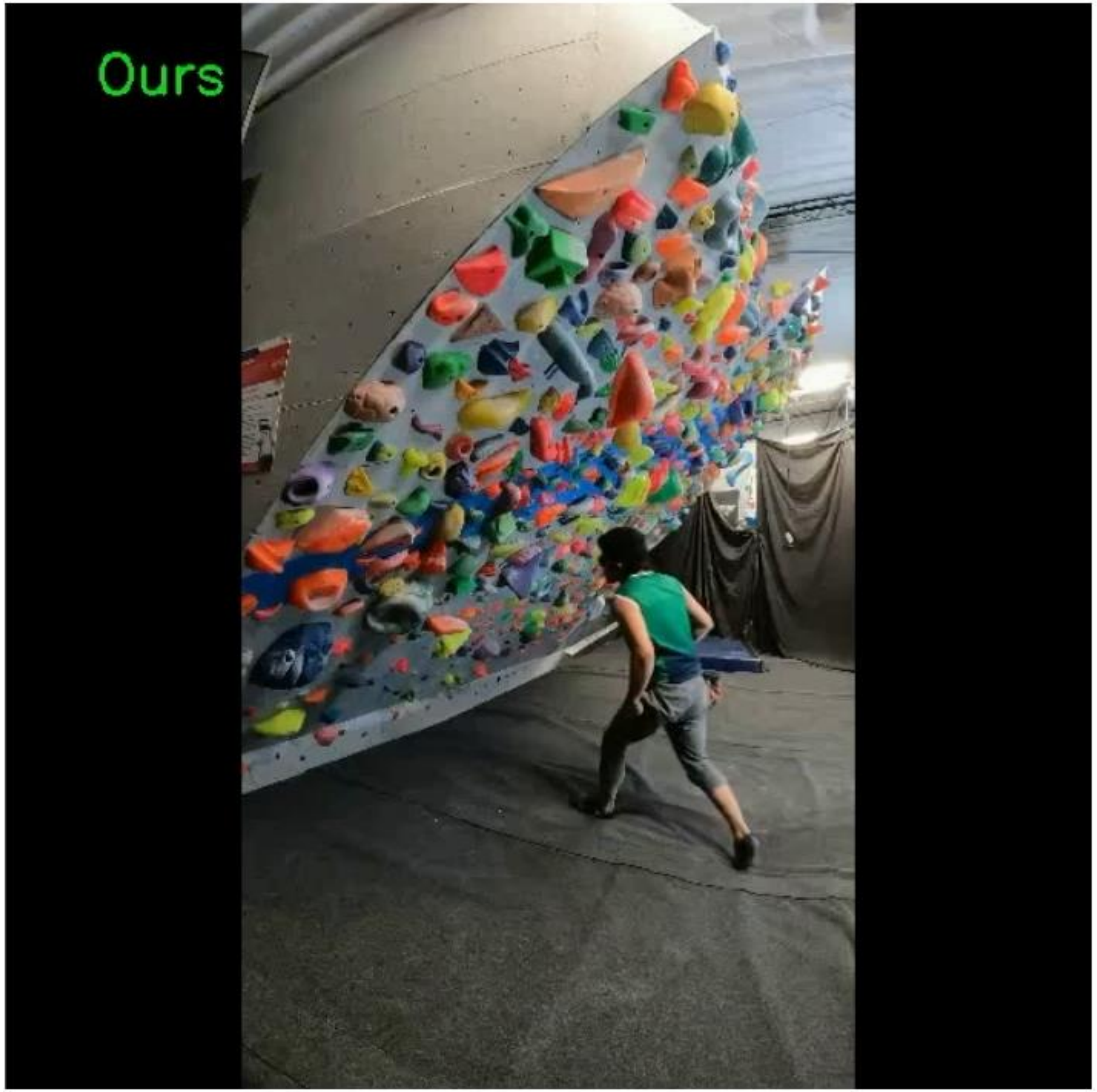
Idea: use video captions (narrations) to provide weak supervision at train time

Method:
- caption each candidate training view separately
- compare predicted captions with view-agnostic ground-truth caption
- choose the view with the most accurate caption as the best view pseudo-label

Majumder et al. CVPR 2025 (Highlight)

Qualitative results: Ego-Exo4D

All-view panel



[1] **Ego-Exo4D.** Grauman et al., CVPR 2024.

Majumder et al. CVPR 2025 (Highlight)

Qualitative results: LEMMA

All-view panel

[1] **LEMMA.** Jia et al., ECCV 2020.

Majumder et al. CVPR 2025 (Highlight)

Automatic evaluation results

		Ego-Exo4D [1]					LEMMA [2]				
		Captioning		*Actions and objects*			*Captioning*		*Actions and objects*		
	Model	CIDEr	METEOR	V-IoU	N-IoU	NC-IoU	CIDEr	METEOR	V-IoU	N-IoU	NC-IoU
Naive heuristics	Ego-only	12.2	47.2	32.2	36.7	30.6	41.7	71.1	38.2	41.3	17.5
	Random	11.5	45.9	30.4	36.6	31.0	30.9	63.1	31.2	33.2	12.8
	Random-exo	11.9	46.0	30.5	37.0	30.9	17.7	51.3	21.6	22.4	6.8
Hand-object interactions and body visibility	Hand-object	12.6	47.4	33.6	36.7	29.6	40.7	72.7	38.5	41.5	17.9
	Body-area	12.9	48.2	32.5	37.2	31.1	42.1	73.8	38.6	41.3	17.6
	Joint-count	12.6	46.6	31.5	29.1	27.7	17.8	51.4	21.7	22.4	6.7
SOTA	Snap angles [3, 4]	12.2	46.7	30.7	35.8	29.1	38.9	70.6	37.1	40.2	17.1
Alternative for using language	Longest-caption	10.7	47.3	30.5	34.6	28.8	32.7	65.4	36.9	37.9	15.3
	Ours	**13.5**	**48.4**	**33.7**	**39.2**	**32.9**	**42.7**	**74.4**	**40.1**	**42.9**	**18.9**

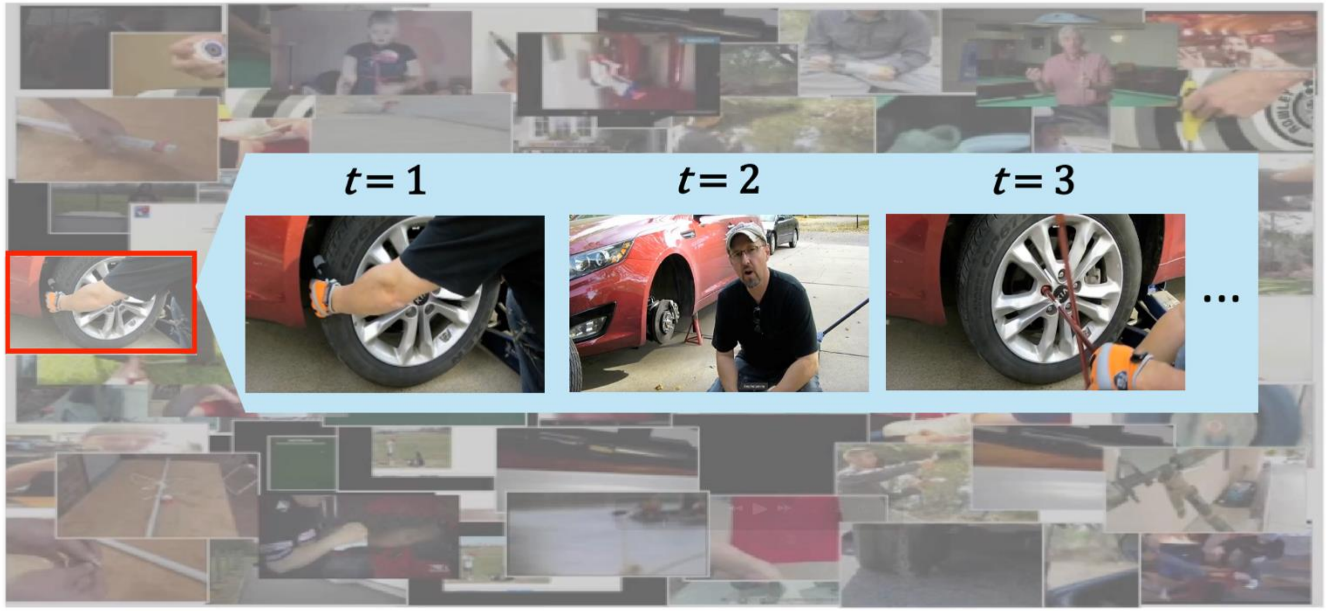
Our model outperforms all baselines on
both Ego-Exo4D and LEMMA datasets across all metrics

[1] **Ego-Exo4D.** Grauman et al., CVPR 2024.
[2] **LEMMA.** Jia et al., ECCV 2020.
[3] **Enhanced 360° viewing via automatic guidance.** Cha et al., ACM Trans. Graph. 2020.
[4] **Snap angle prediction for 360° panoramas.** Xiong and Grauman, ECCV 2018.

Majumder et al. CVPR 2025 (Highlight)

View selection in label-scarce settings

Main conference: Wed (10/22) AM, poster #185



LangView

… by using captions for producing best-view pseudo-labels during training

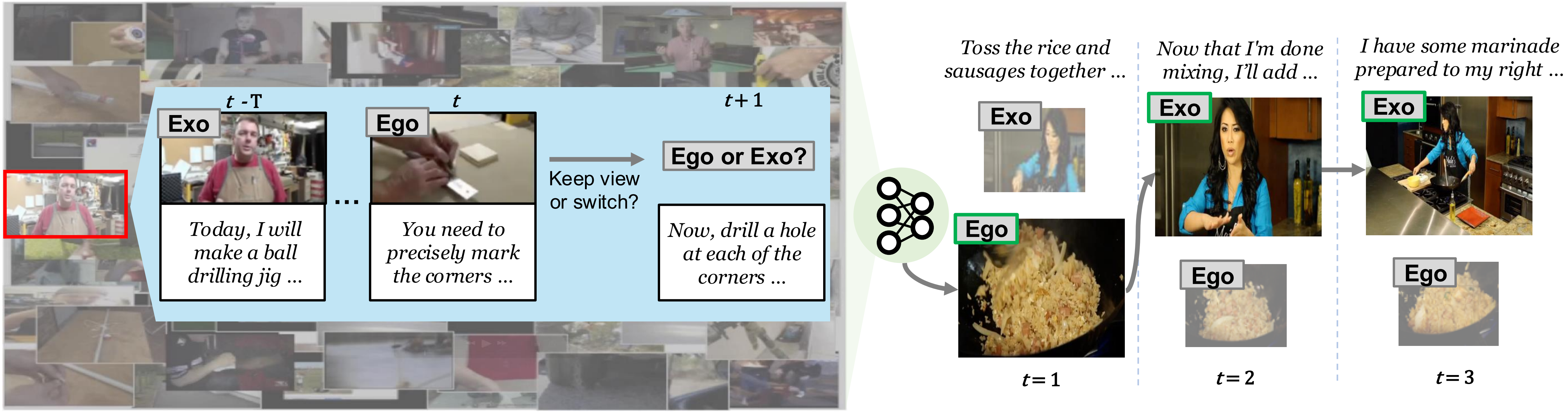
Switch-a-View

… by learning human view choices from unlabeled but edited in-the-wild how-tos

Language for Weakly Supervising View-Selection in Multi-view Instructional Videos
Majumder et al., CVPR 2025 (Highlight).
Project page: https://vision.cs.utexas.edu/projects/which-view-shows-it-best

Switch-a-View: View Selection Learned from Unlabeled In-the-wild Videos
Majumder et al., ICCV 2025.
Project page: https://vision.cs.utexas.edu/projects/switch-a-view/

Switch-a-View: View Selection Learned from Unlabeled In-the-wild Videos

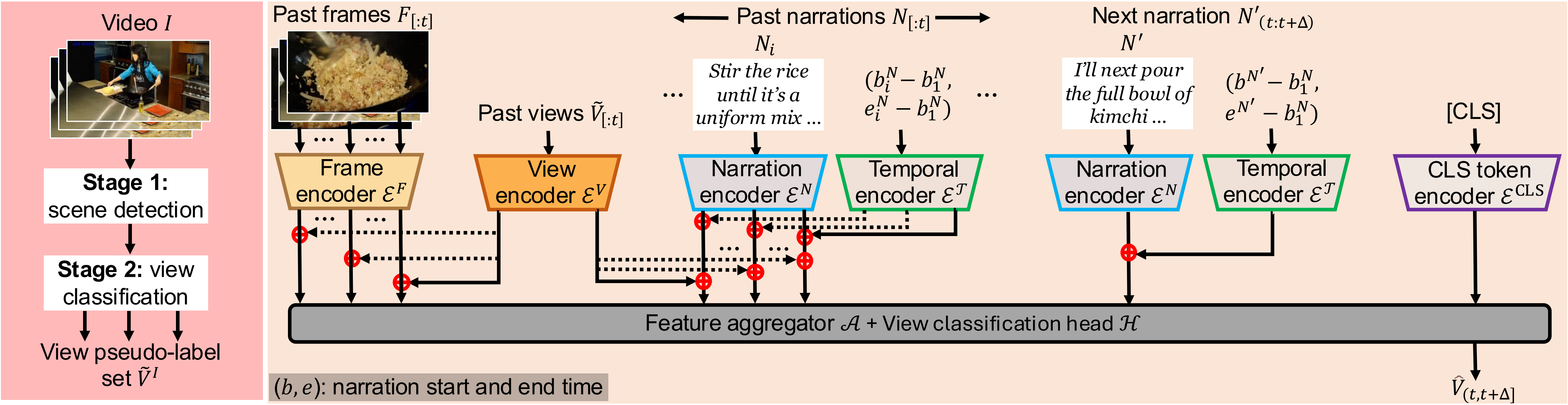


Task: Learn view selection in multi-view instructional videos with **limited** labels

Idea:
- Learn human view choices from large-scale unlabeled in-the-wild videos by solving a weakly-supervised view-switch detection task
- Finetune this model for view selection with limited labels

Majumder et al. ICCV 2025

View-switch detection model



View pseudo-labeler

View-switch detector D

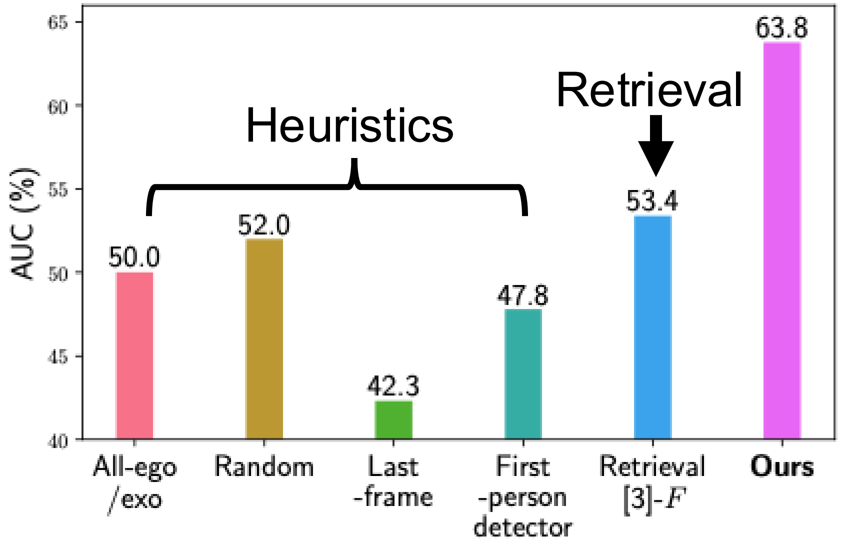
Model components:

1. View pseudo-labeler -- pseudo-labels the dominant view type in a video clip
2. View-switch detector -- given past frames, narrations and view types, and the next narration, predicts the next view type

Training loss: $\mathcal{L}^D = \mathcal{L}_{\text{CE}}(\hat{V}_{(t,t+\Delta]}, \tilde{V}_{(t,t+\Delta]})$

Majumder et al. ICCV 2025

Quantitative results



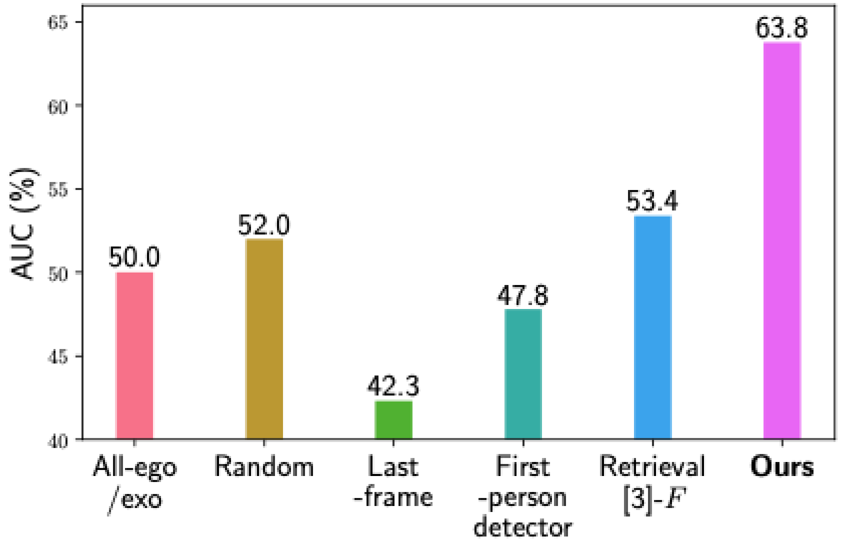
View-switch detection on HT100M

[1] HowTo100M. Miech et al., ICCV 2019.

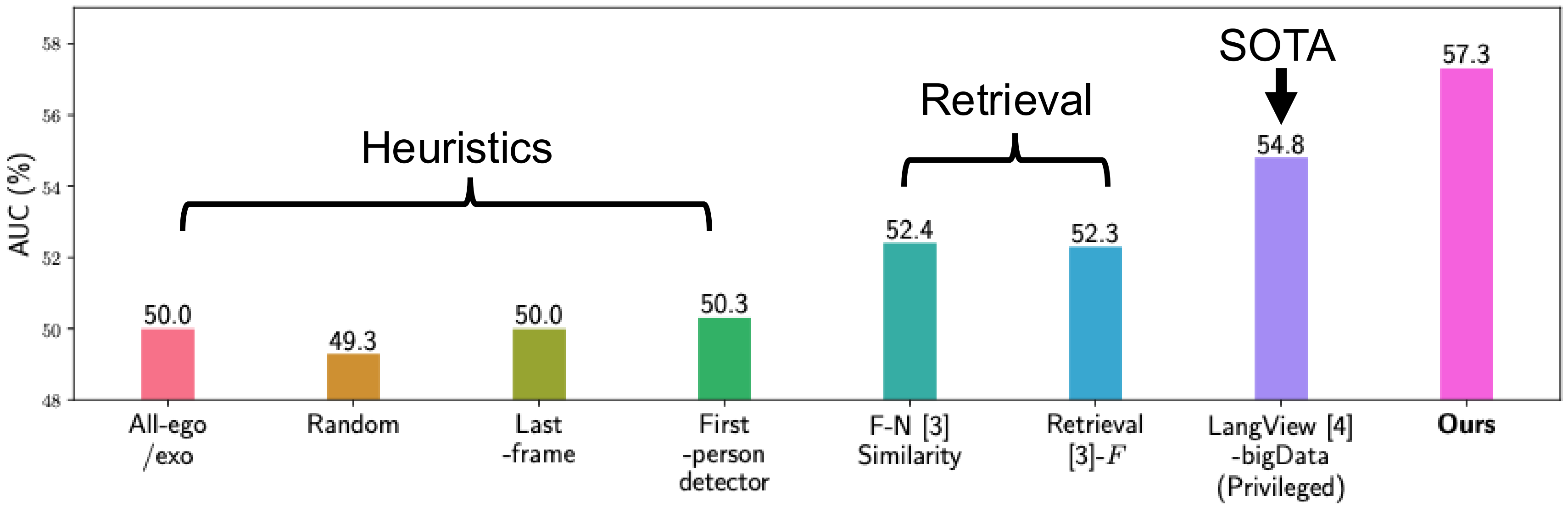
[3] InternVideo2. Wang et al., ECCV 2024.

Majumder et al. ICCV 2025

Quantitative results



View-switch detection on HT100M

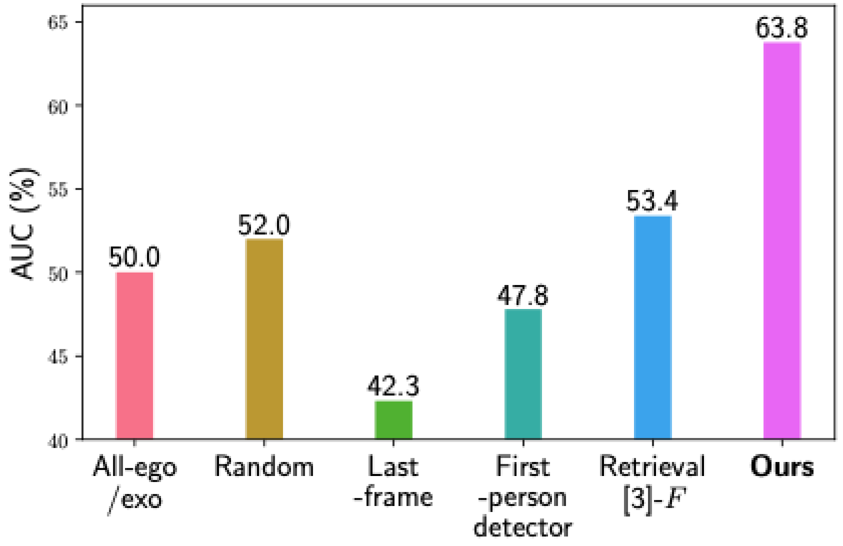


View selection on Ego-Exo4D

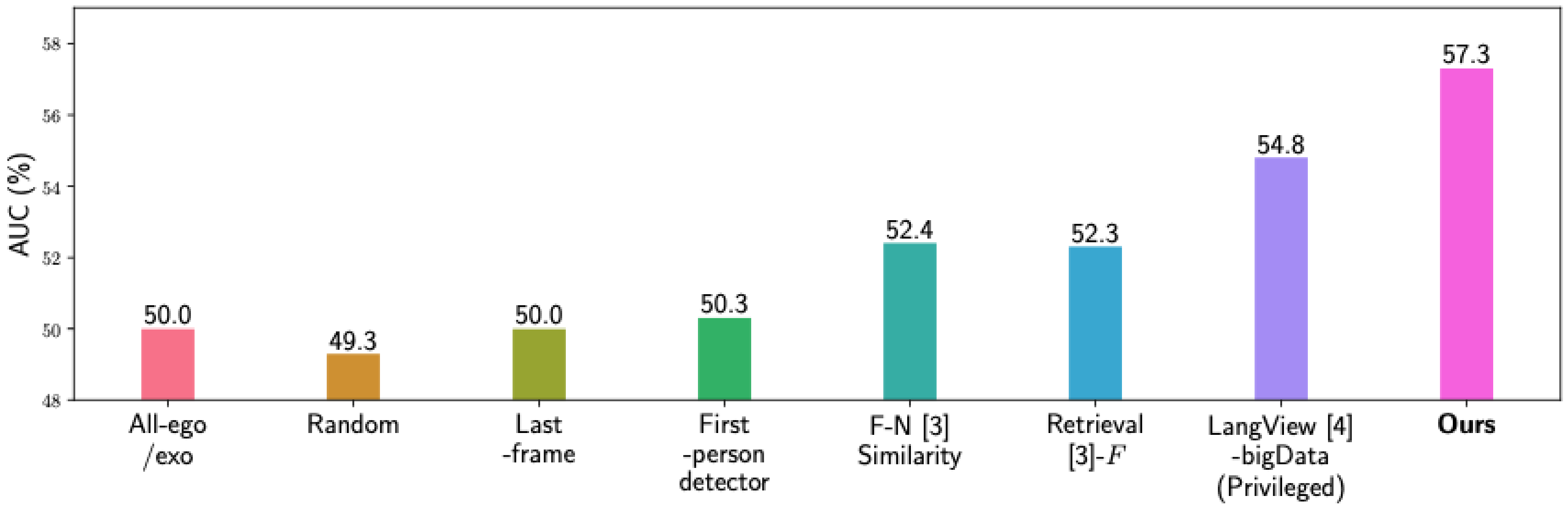
❖ Our model outperforms all baselines on both HowTo100M and Ego-Exo4D on all metrics

[1] HowTo100M. Miech et al., ICCV 2019.
[2] Ego-Exo4D. Grauman et al., CVPR 2024.
[3] InternVideo2. Wang et al., ECCV 2024.
[4] LangView. Majumder et al., CVPR 2025.

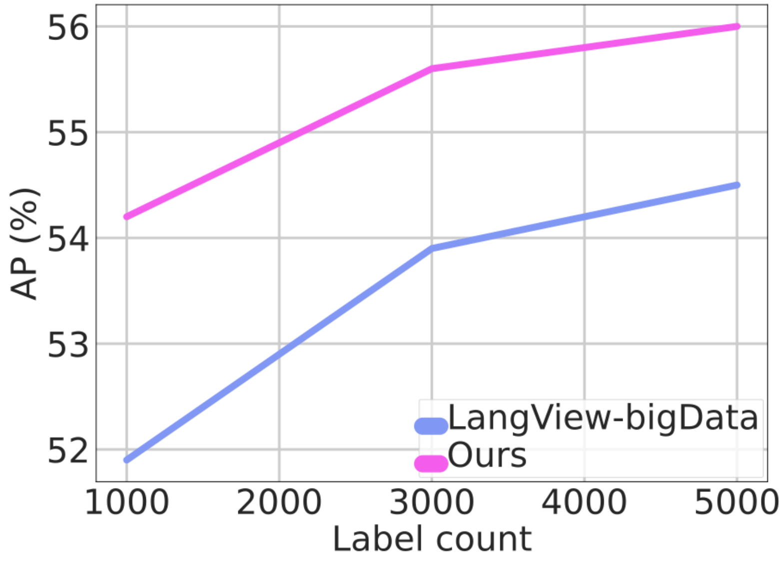
Quantitative results



View-switch detection on HT100M



View selection on Ego-Exo4D



View selection AP vs. label #

- ❖ Our model outperforms all baselines on both HowTo100M and Ego-Exo4D on all metrics
- ❖ The lower the count of best view labels, the higher our margin of improvement is over the baselines

[1] HowTo100M. Miech et al., ICCV 2019.
[2] Ego-Exo4D. Grauman et al., CVPR 2024.
[3] InternVideo2. Wang et al., ECCV 2024.
[4] LangView. Majumder et al., CVPR 2025.

Majumder et al. ICCV 2025

View-switch detection example

Past frames and narrations

Next USEEN frames and SEEN narrations

Majumder et al. ICCV 2025

View selection example

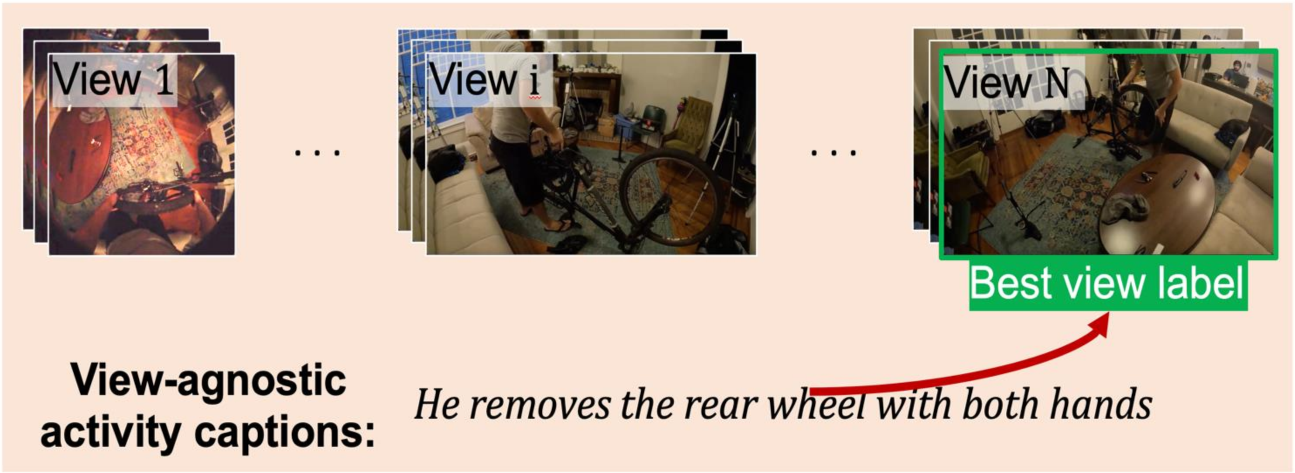
Past frames and narrations

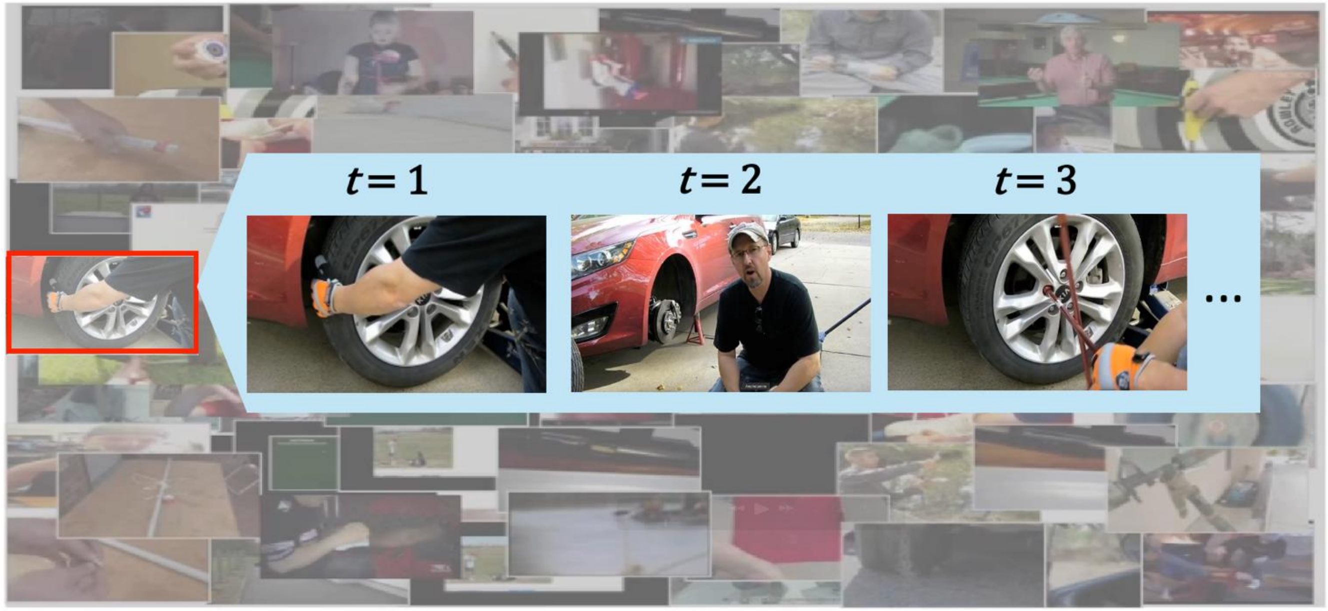
Next narrations and frames from candidate views

Label,
Prediction

View selection in label-scarce settings

Main conference: Wed (10/22) AM, poster #185





LangView

… by using captions for producing best-view pseudo-labels during training

Switch-a-View

… by learning human view choices from unlabeled but edited in-the-wild how-tos

Language for Weakly Supervising View-Selection in Multi-view Instructional Videos
Majumder et al., CVPR 2025 (Highlight).
Project page: https://vision.cs.utexas.edu/projects/which-view-shows-it-best

Switch-a-View: View Selection Learned from Unlabeled In-the-wild Videos
Majumder et al., ICCV 2025.
Project page: https://vision.cs.utexas.edu/projects/switch-a-view/