Learning Procedure-aware Video Representation from Instructional Videos and Their Narrations

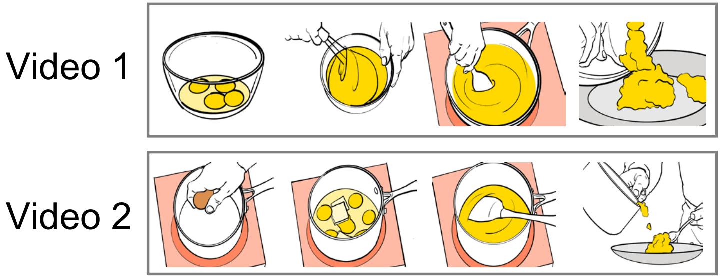
Yiwu Zhong[1], Licheng Yu[2], Yang Bai[2], Shangwen Li[2], Xueting Yan[*2], Yin Li[*1]

(*co-corresponding authors)

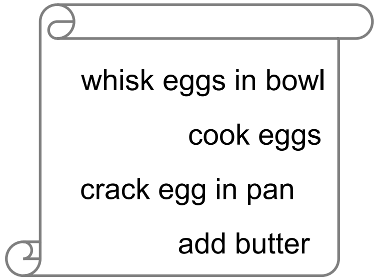
[1]University of Wisconsin-Madison

[2]Meta AI

ProcedureVRL: Procedure-aware Video Representation Learning



Instructional Videos from YouTube (HowTo100M)

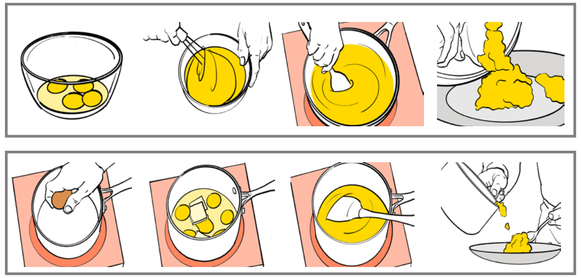


Step Descriptions Extracted from Video Narrations

ProcedureVRL: Procedure-aware Video Representation Learning

Video 1

Video 2



Instructional Videos from YouTube (HowTo100M)

whisk eggs in bowl

cook eggs

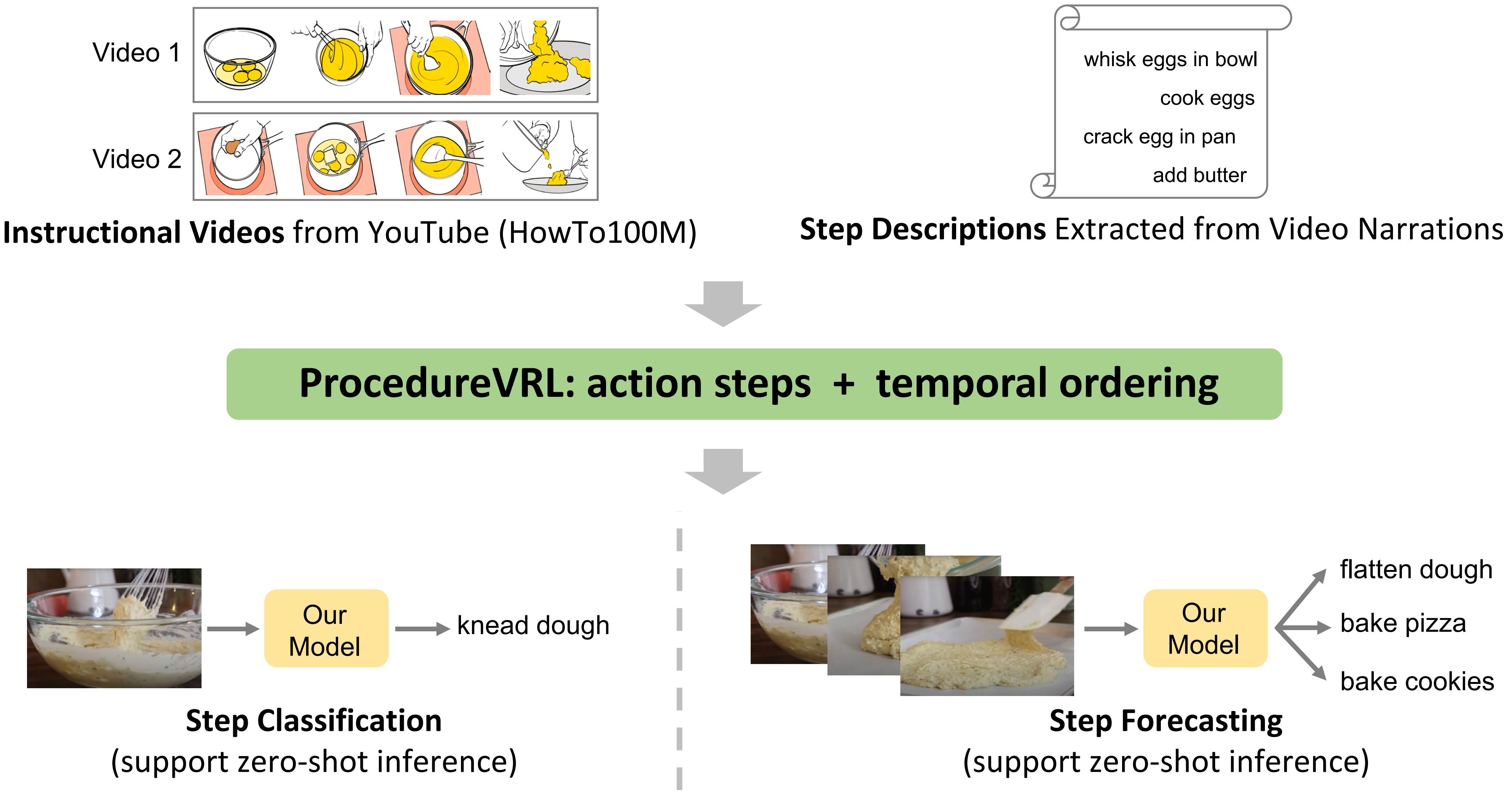
crack egg in pan

add butter

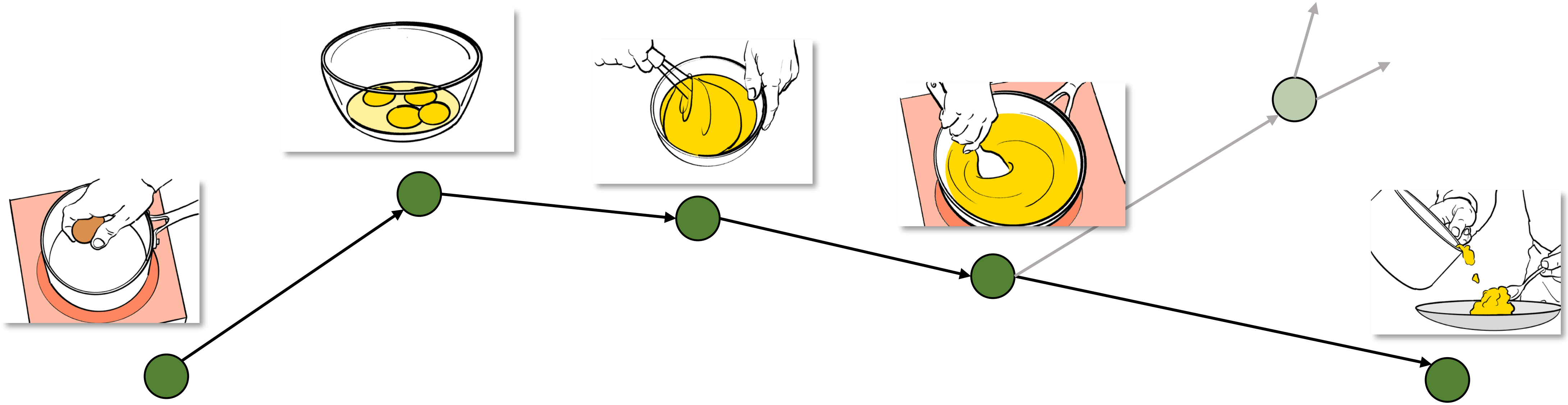
Step Descriptions Extracted from Video Narrations

ProcedureVRL: action steps + temporal ordering

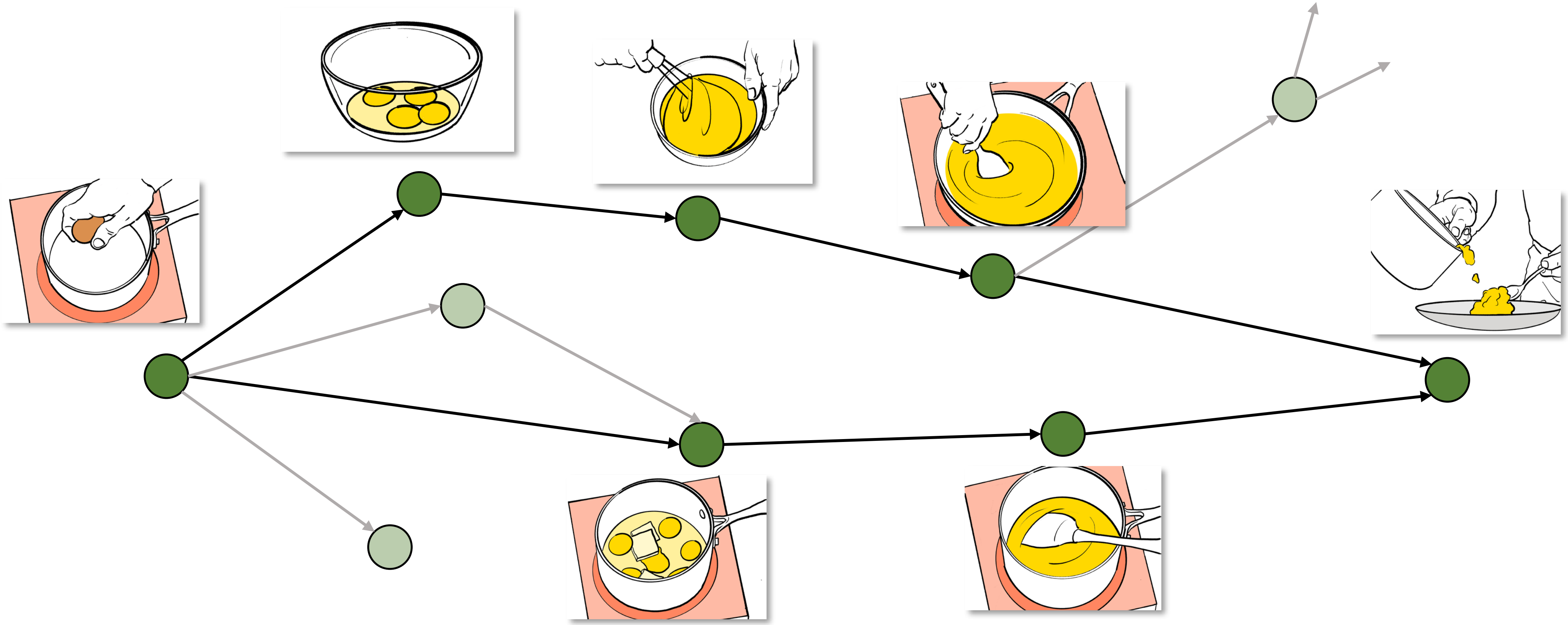
ProcedureVRL: Procedure-aware Video Representation Learning



Task Procedures

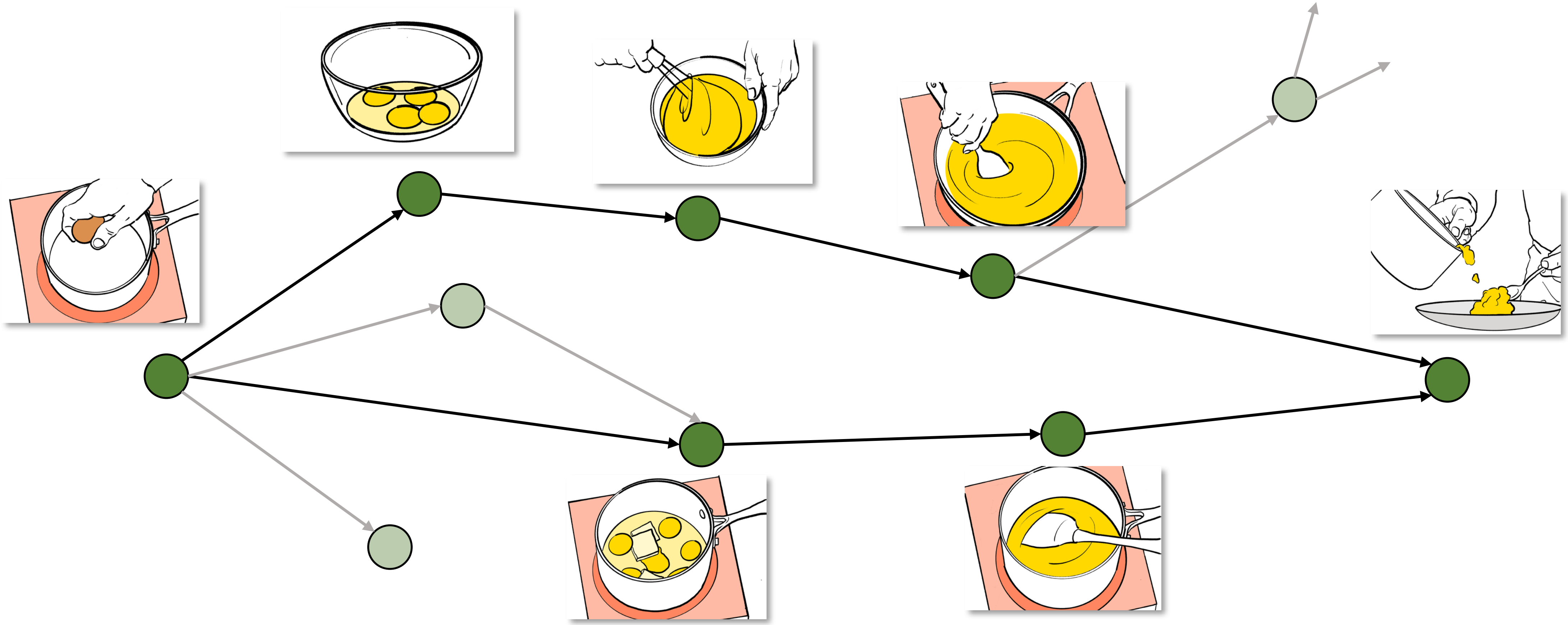


Task Procedures



Various procedures for the same task

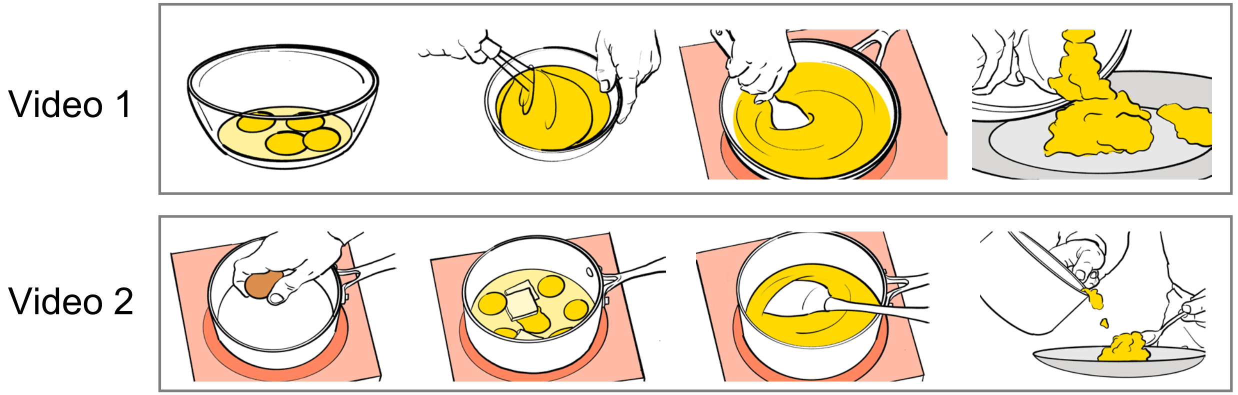
Task Procedures



Task procedure = action steps + temporal ordering

Can we build a vision model that understands **task procedures**?

Our Work: Learning from Videos and Their Narrations



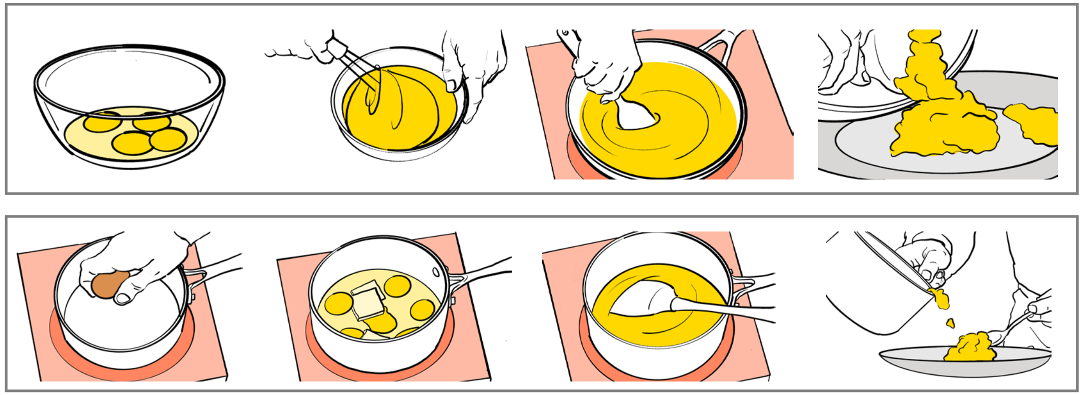
Instructional Videos from YouTube

Task procedure = action steps + temporal ordering

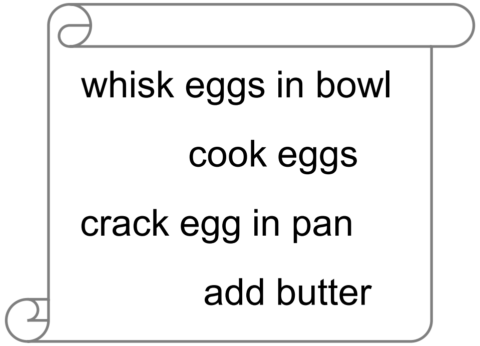
Our Work: Learning from Videos and Their Narrations

Video 1

Video 2



Instructional Videos from YouTube



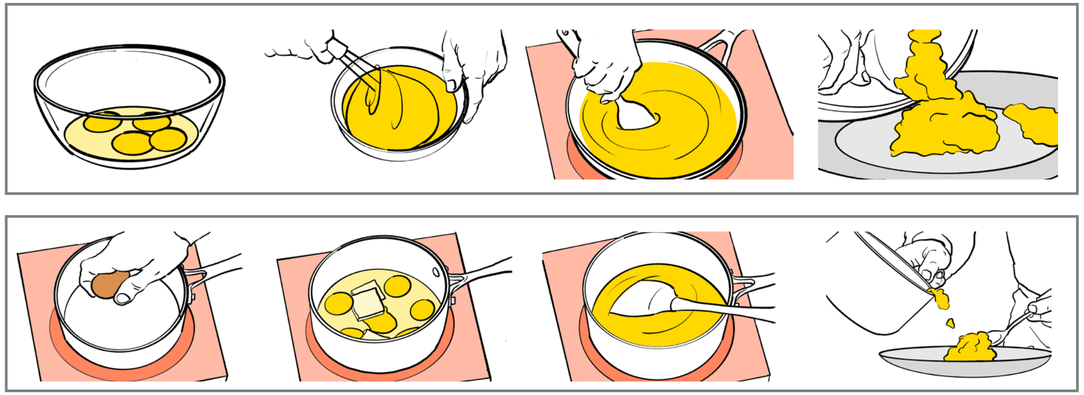
Step Descriptions Extracted from Narrations

Task procedure = action steps + temporal ordering

Our Work: Procedure-aware Video Representation

Video 1

Video 2



Instructional Videos from YouTube

whisk eggs in bowl

cook eggs

crack egg in pan

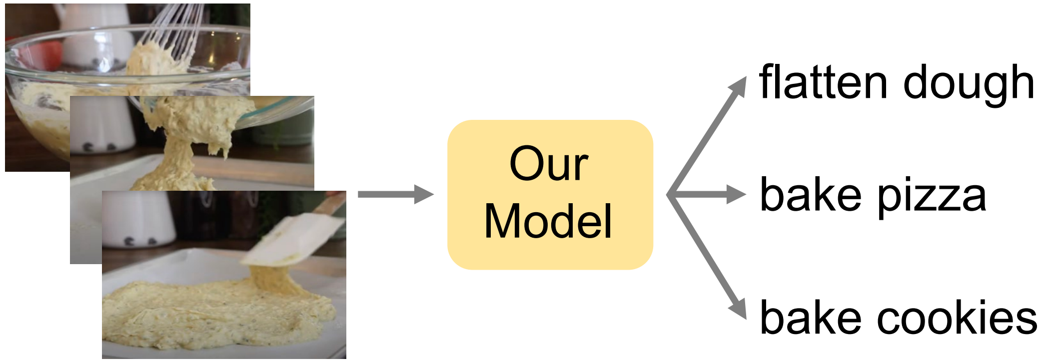
add butter

Step Descriptions Extracted from Narrations

Goal: learning video representation that encodes **action steps** and their **ordering**, **without** using human annotation

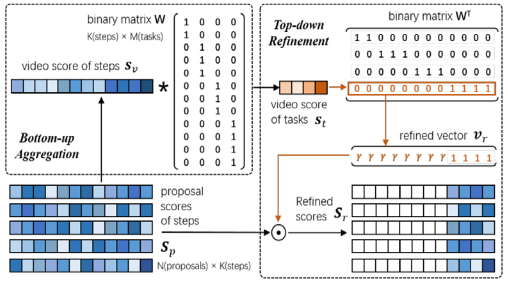
Our Work: Procedure-aware Video Representation

Step Classification
(support zero-shot inference)

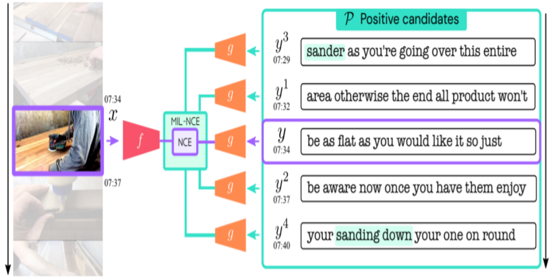


Step Forecasting
(support zero-shot inference)

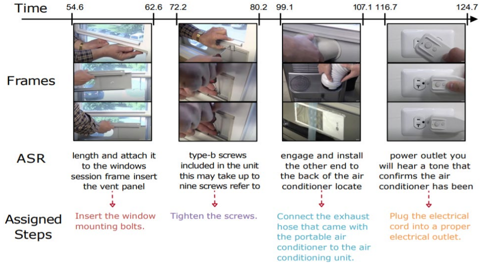
Previous Work for Understanding Procedures



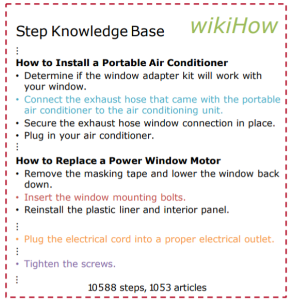
Tang, et al., CVPR 2019



Miech, et al., CVPR 2020



Lin, et al., CVPR 2022



Koupaee, et al., arXiv 2018

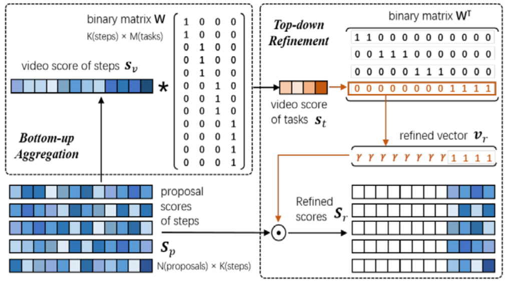
Tang, et al., COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, CVPR 2019

Miech, et al., End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

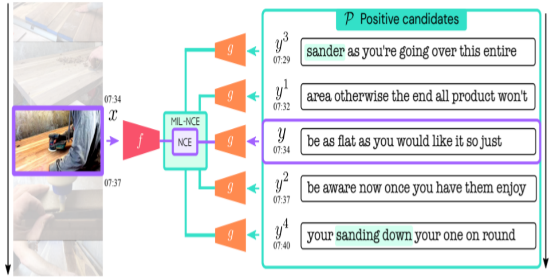
Lin, et al., Learning To Recognize Procedural Activities with Distant Supervision, CVPR 2022

Koupaee, et al., WikiHow: A large scale text summarization dataset, arXiv 2018

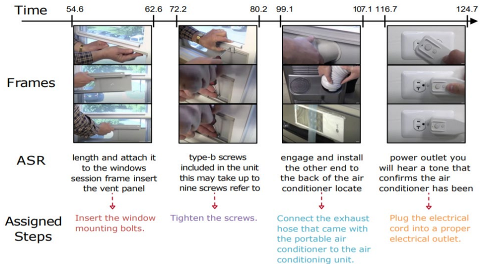
Previous Work for Understanding Procedures



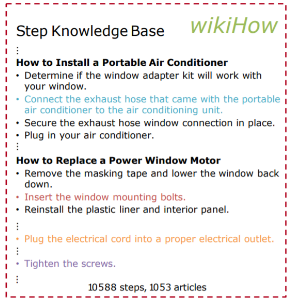
Tang, et al., CVPR 2019



Miech, et al., CVPR 2020



Lin, et al., CVPR 2022



Koupaee, et al., arXiv 2018

- Rely on human annotation → limited step categories
- Rely on procedures summarized by human → fixed steps, fixed ordering

Tang, et al., COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, CVPR 2019
Miech, et al., End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020
Lin, et al., Learning To Recognize Procedural Activities with Distant Supervision, CVPR 2022
Koupaee, et al., WikiHow: A large scale text summarization dataset, arXiv 2018

Key Challenges

- How to obtain the labels of individual video clips? (step concepts)

- How to capture immense variations in procedures? (step ordering)

Our Work: Key Ideas

- How to obtain the labels of individual video clips? (step concepts)
  - ✓ Leverage **image-language model** to align video-step
- How to capture immense variations in procedures? (step ordering)

Open-set understanding

Zero-shot inference

Our Work: Key Ideas

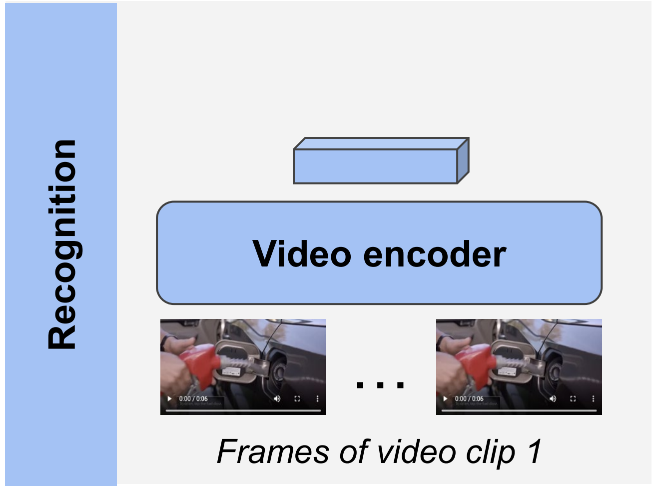
- How to obtain the labels of individual video clips? (step concepts)
  - ✓ Leverage **image-language model** to align video-step
- How to capture immense variations in procedures? (step ordering)
  - ✓ Design a **probabilistic model** to learn variations present in videos

Open-set understanding

Zero-shot inference

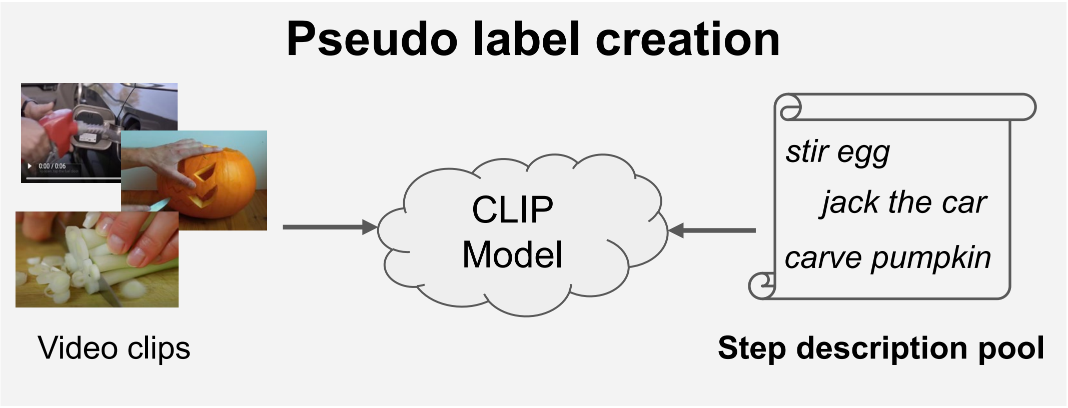
Diverse forecasting

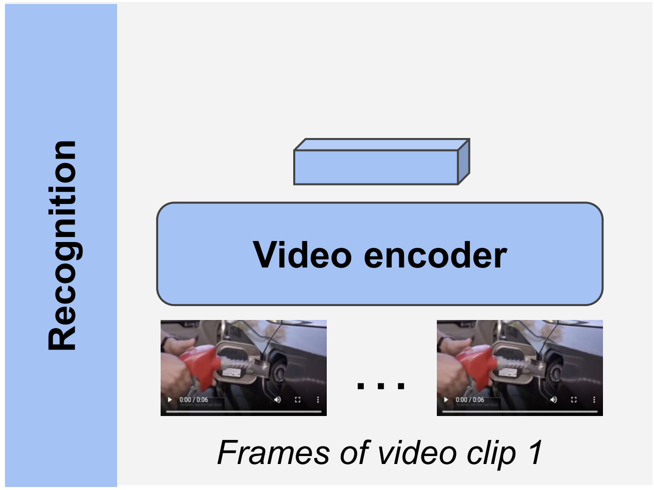
Procedure-aware Video Representation



Miech, et al., HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips, CVPR 2019

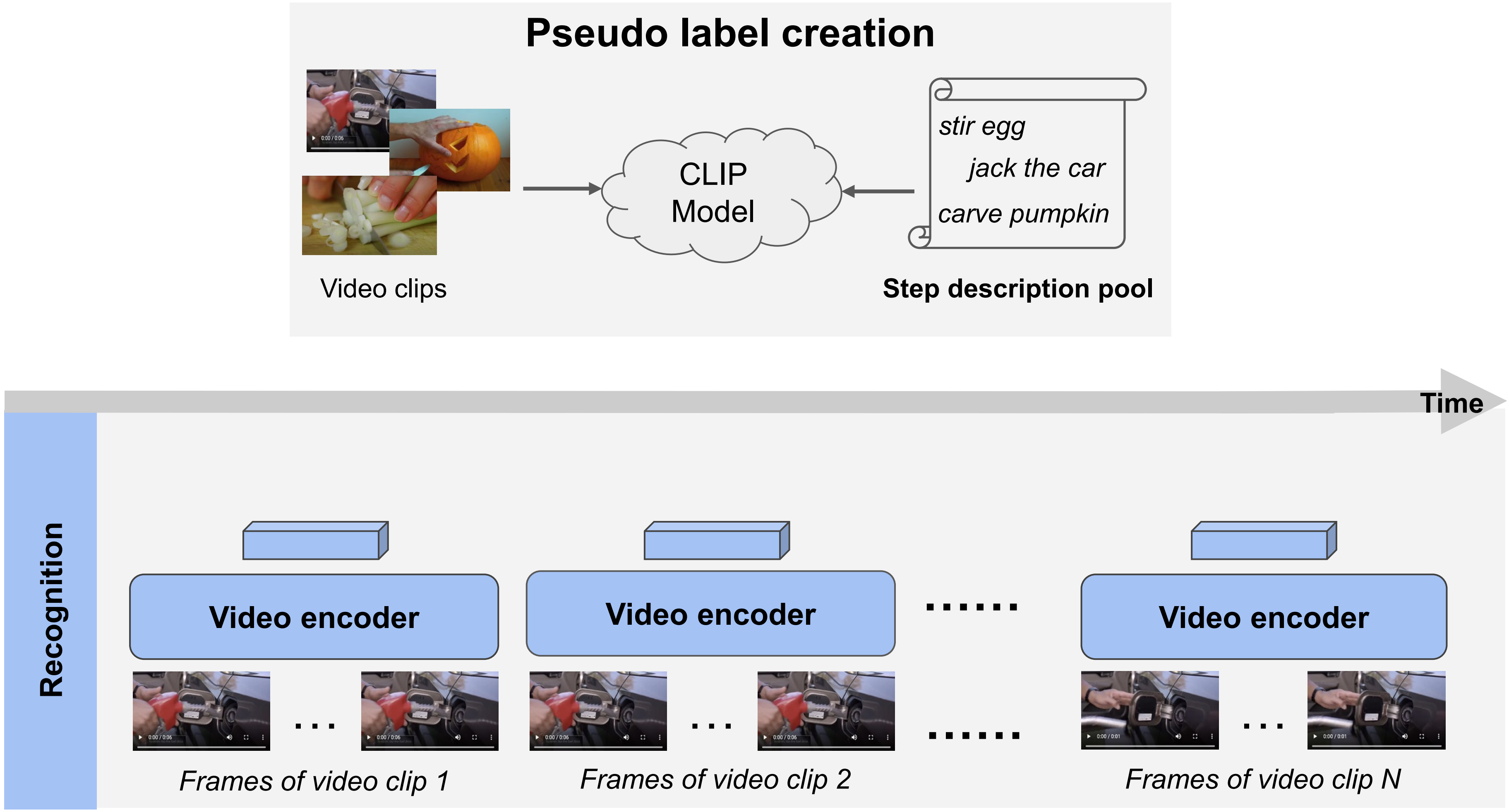
Procedure-aware Video Representation





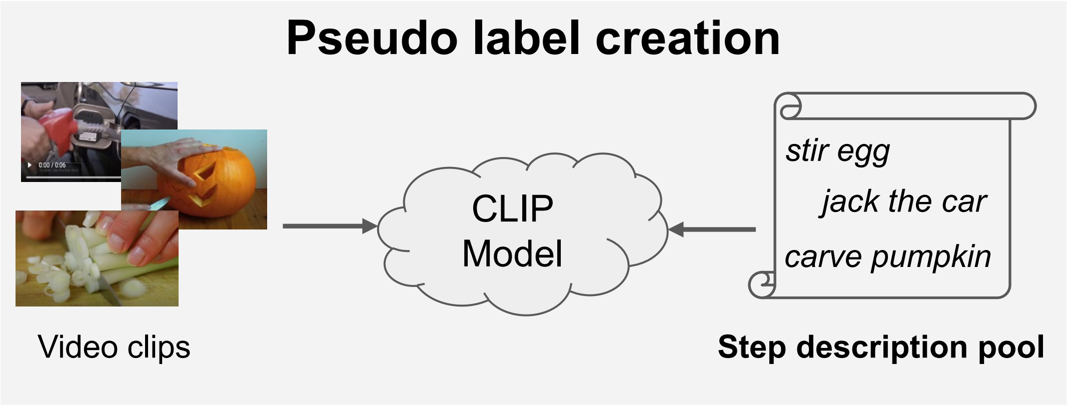
CLIP: creates pseudo labels for **individual** action steps

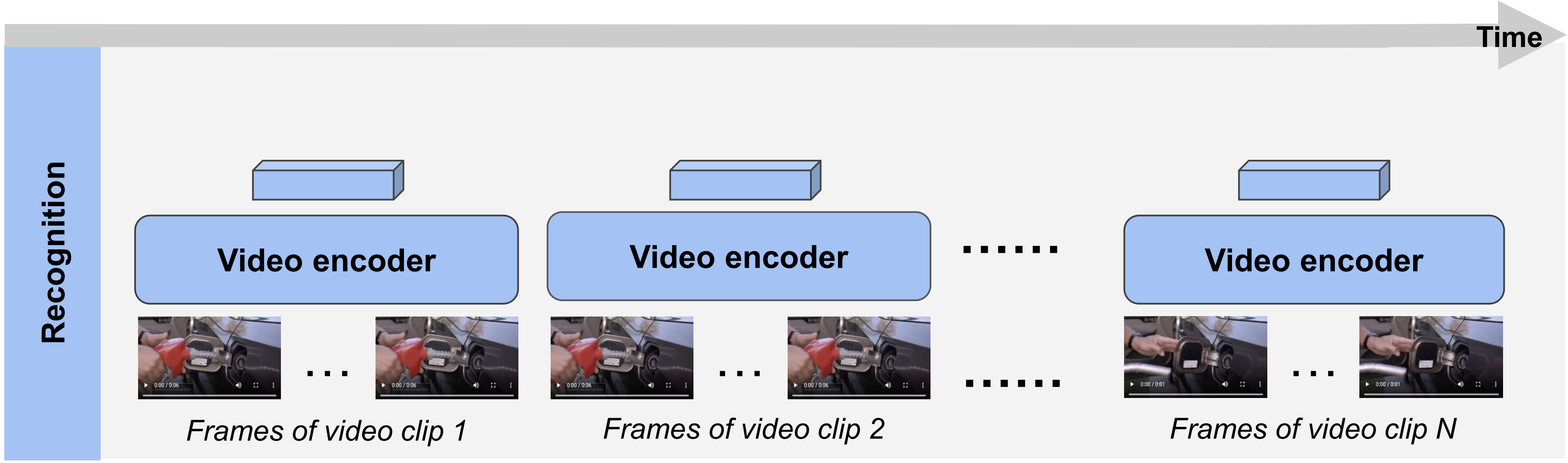
Procedure-aware Video Representation



Video encoder: learns representation for input video clips, **supervised by step description**

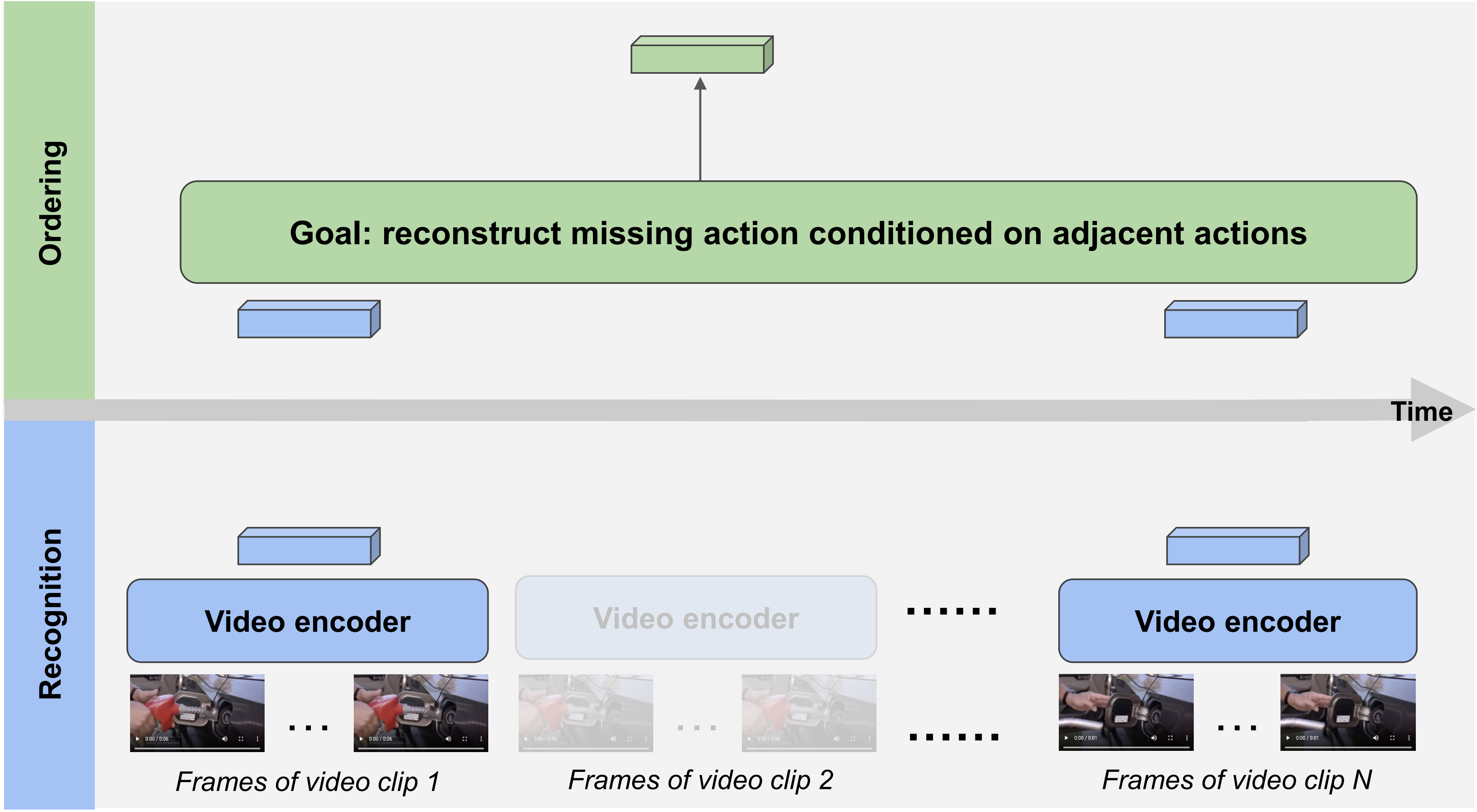
Procedure-aware Video Representation





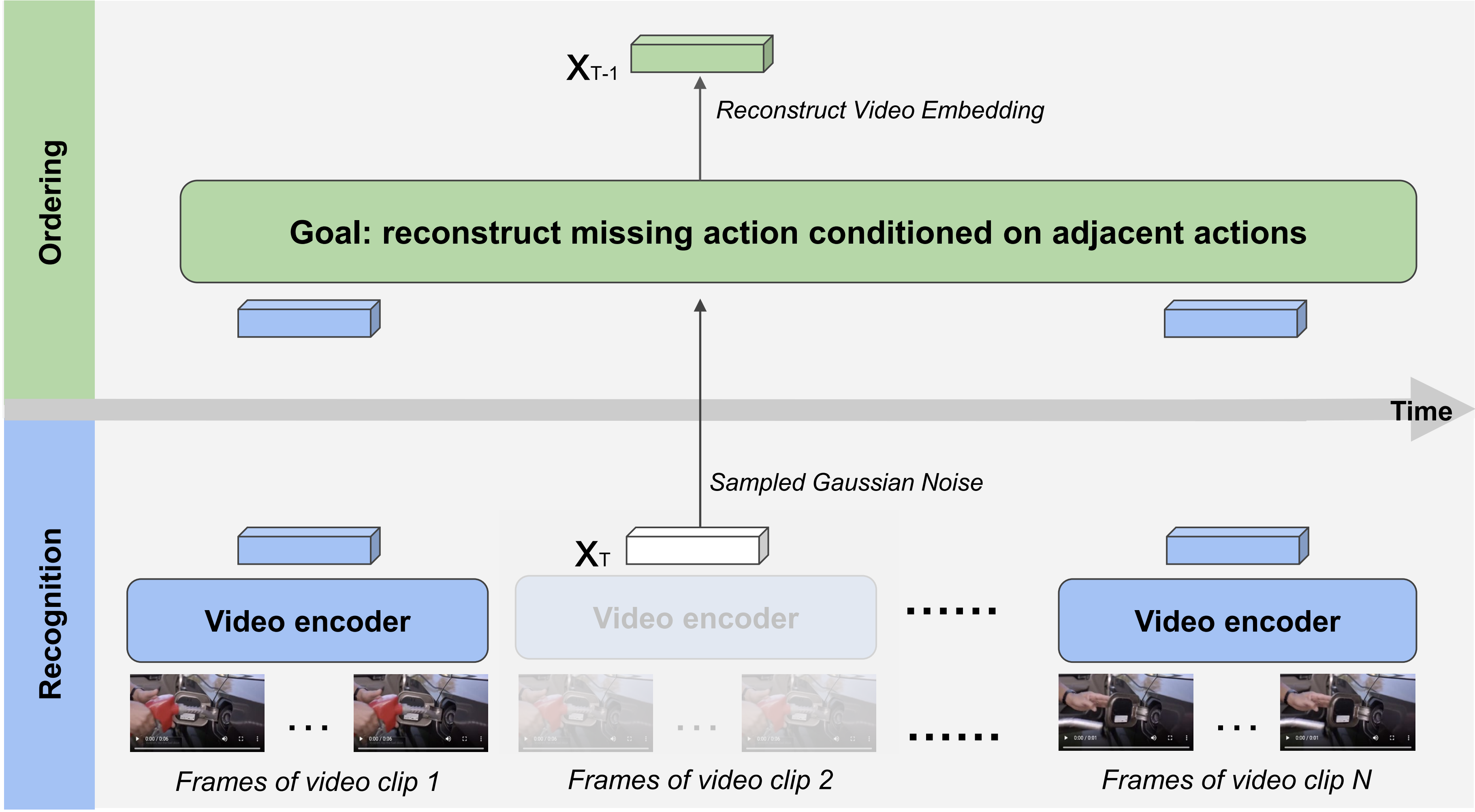
Question: how to learn **step ordering** & capture **procedure variations**?

Procedure-aware Video Representation



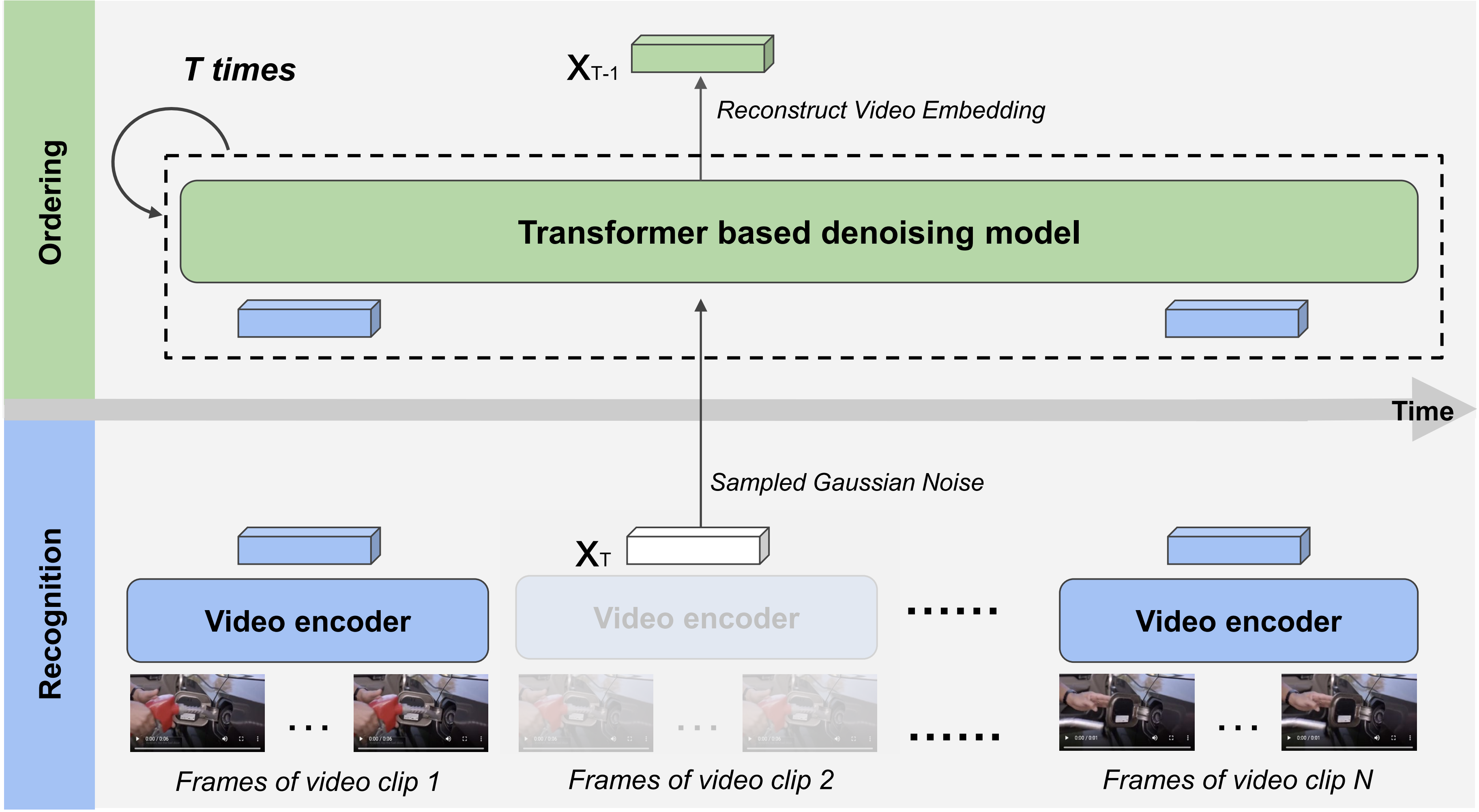
Novelty: learning the **step ordering** provided by videos themselves (self-supervised)

Procedure-aware Video Representation



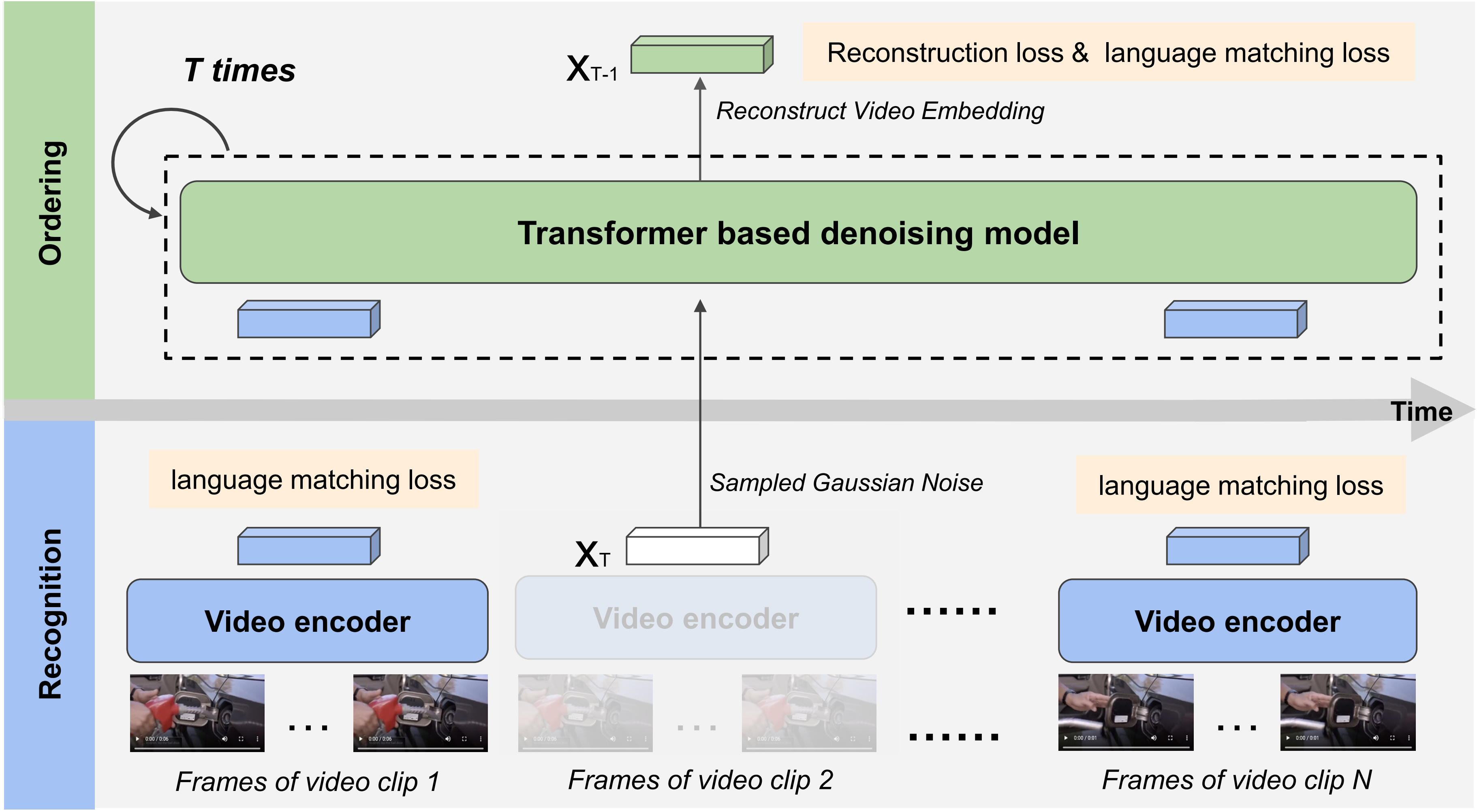
Novelty: design a diffusion model to capture **step ordering** & **variations** in procedures

Procedure-aware Video Representation



Diffusion model: captures **step ordering** & **variations** in procedural activities

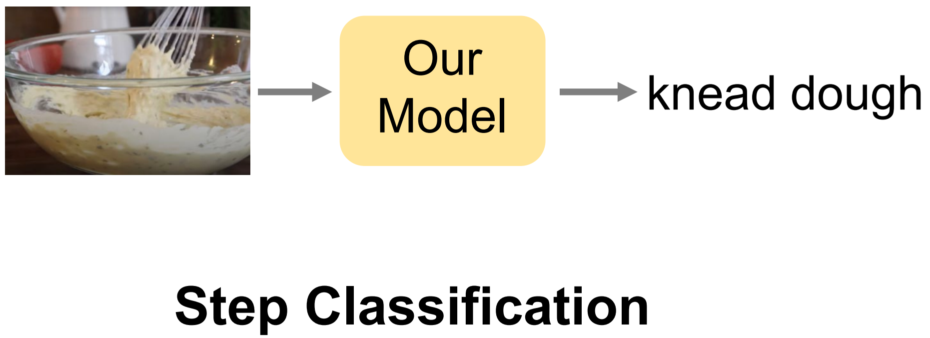
Procedure-aware Video Representation: Training

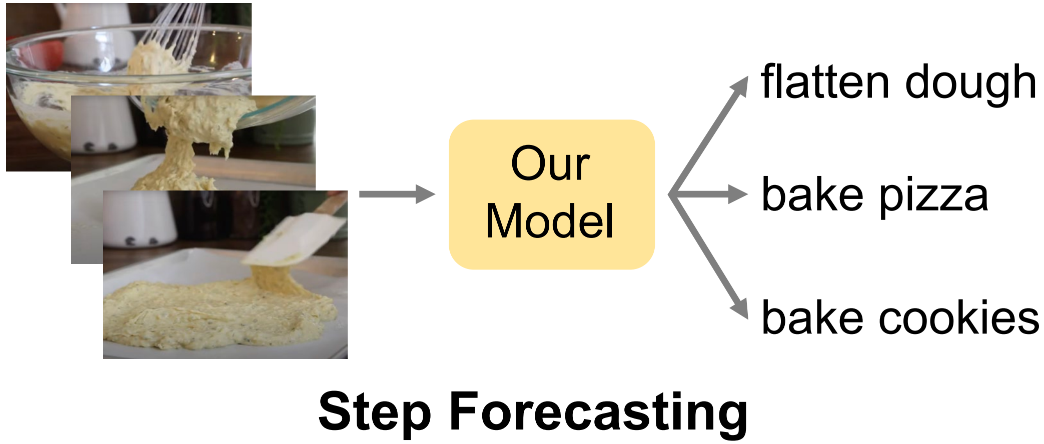


During training, we adopt language matching loss & reconstruction loss

Procedure-aware Video Representation: Inference

Procedure-aware Video Representation: Inference





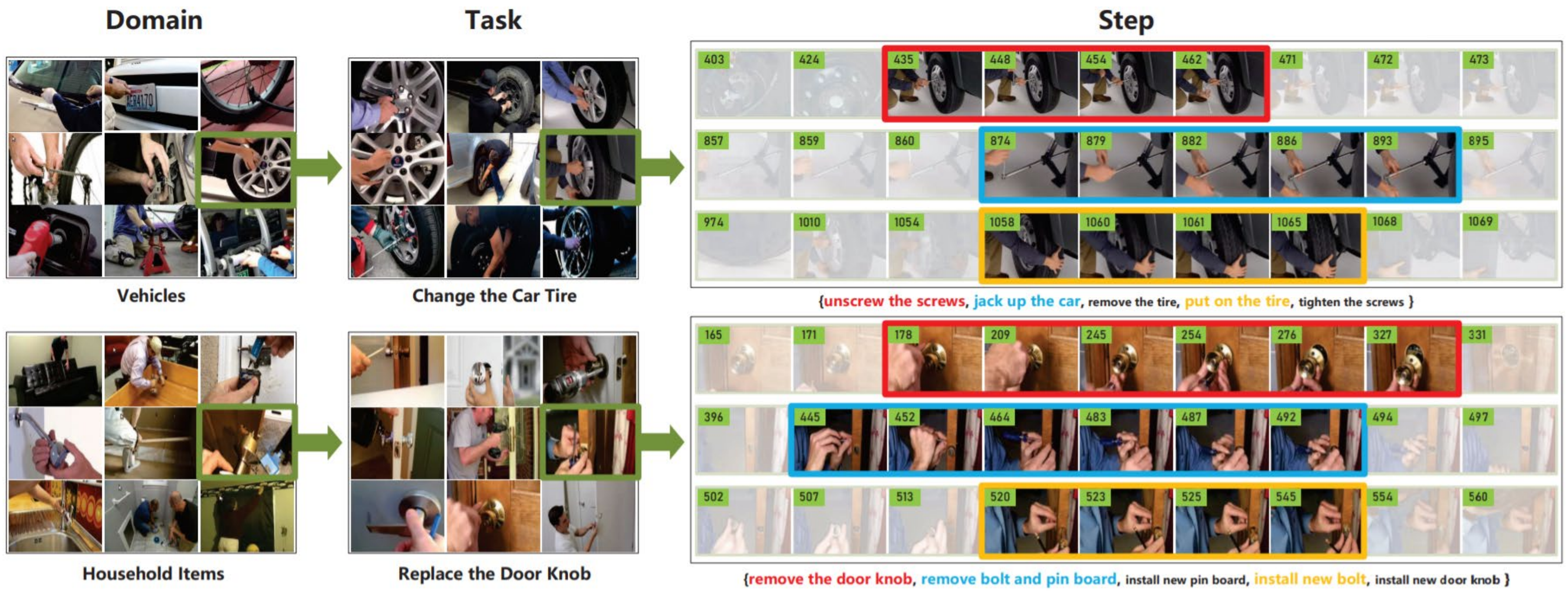
New capabilities:

- **zero-shot inference** for both step classification & forecasting *(first work)*
- **diverse** predictions for step forecasting

Evaluation Benchmark

Evaluation Dataset:

- COIN has 400 hours of videos with step instances annotated (180 tasks, 778 steps)



Tang, et al., COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, CVPR 2019

Evaluation Benchmark

Evaluation Dataset:

- COIN has 400 hours of videos with step instances annotated (180 tasks, 778 steps)

Evaluation tasks

- Step classification: Classify the input short video clip into a step category
- Step forecasting: Forecast next step, given the input video recording previous steps

Tang, et al., COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, CVPR 2019

Evaluation Benchmark

Evaluation Dataset:

- COIN has 400 hours of videos with step instances annotated (180 tasks, 778 steps)

Evaluation tasks

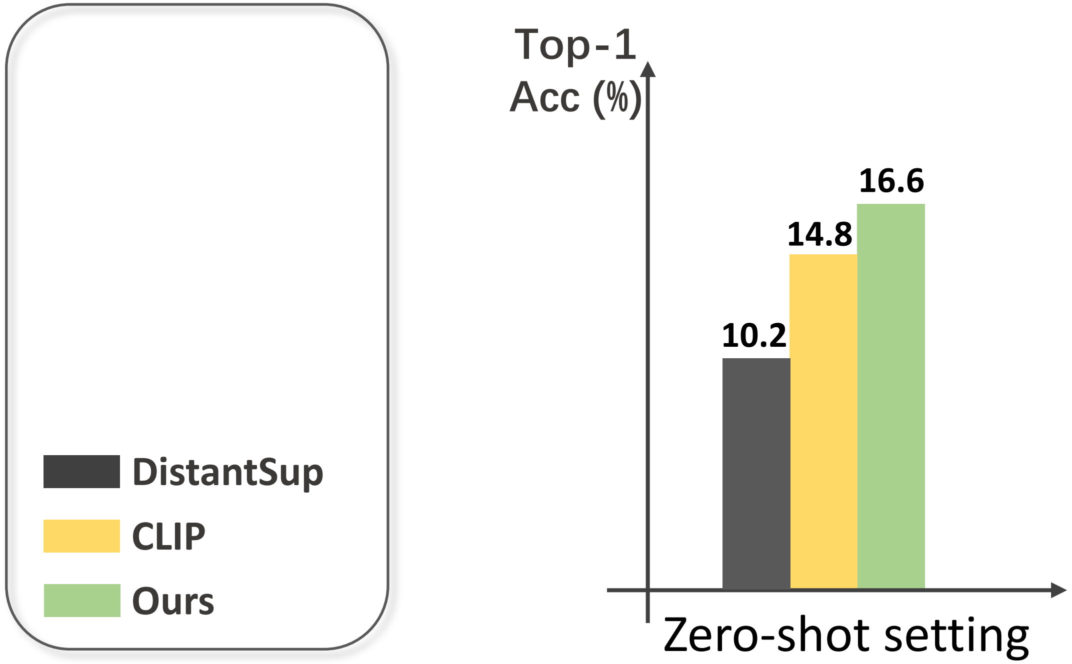
- Step classification: Classify the input short video clip into a step category
- Step forecasting: Forecast next step, given the input video recording previous steps

Evaluation Settings:

- Zero-shot setting: directly evaluate the pretrained model
- Fine-tuning setting: fine-tune pre-trained model using human annotation

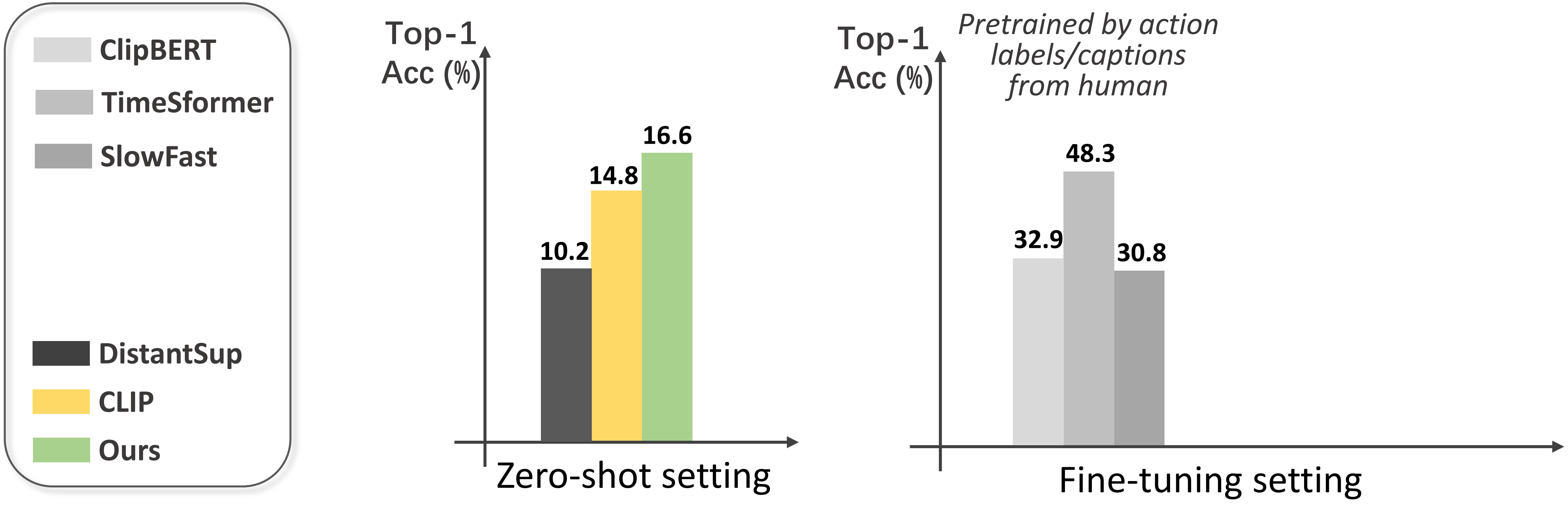
Tang, et al., COIN: A Large-scale Dataset for Comprehensive Instructional Video Analysis, CVPR 2019

Step Classification Results on COIN Benchmark



Feichtenhofer, et al., SlowFast networks for video recognition, ICCV 2019
Miech, et al., End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020
Lei, et al., Less is more: Clipbert for video-and-language learning via sparse sampling, CVPR 2021
Bertasius, et al., Is space-time attention all you need for video understanding, ICML 2021
Xu, et al., VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021
Lin, et al., Learning To Recognize Procedural Activities with Distant Supervision, CVPR 2022

Step Classification Results on COIN Benchmark



Feichtenhofer, et al., SlowFast networks for video recognition, ICCV 2019

Miech, et al., End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

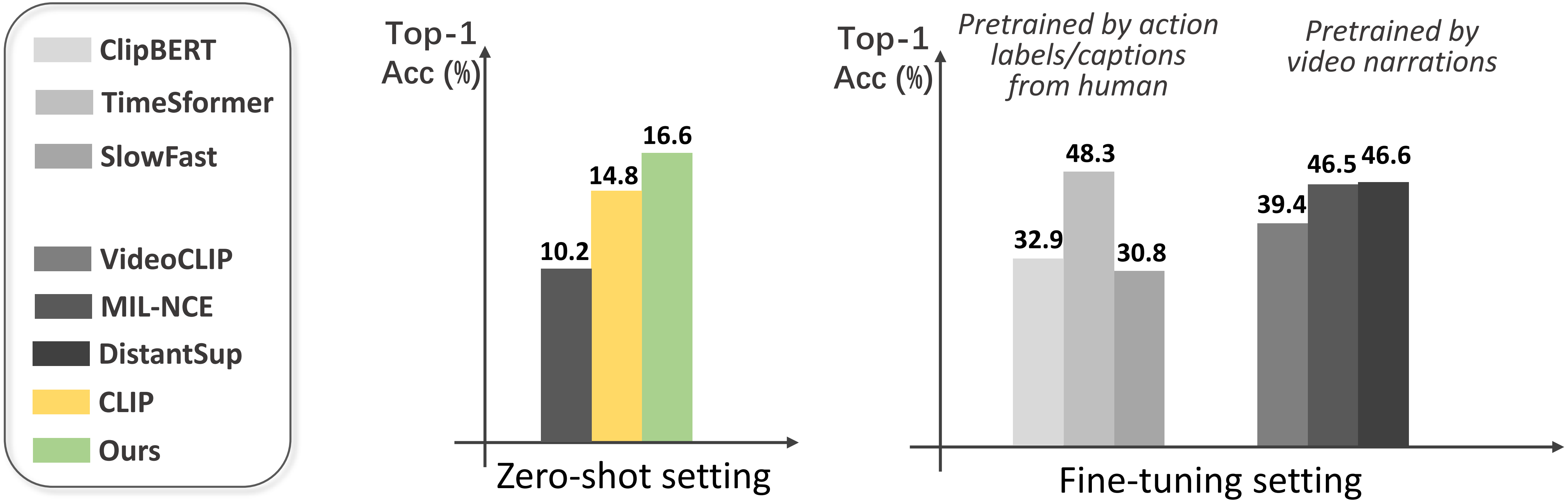
Lei, et al., Less is more: Clipbert for video-and-language learning via sparse sampling, CVPR 2021

Bertasius, et al., Is space-time attention all you need for video understanding, ICML 2021

Xu, et al., VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021

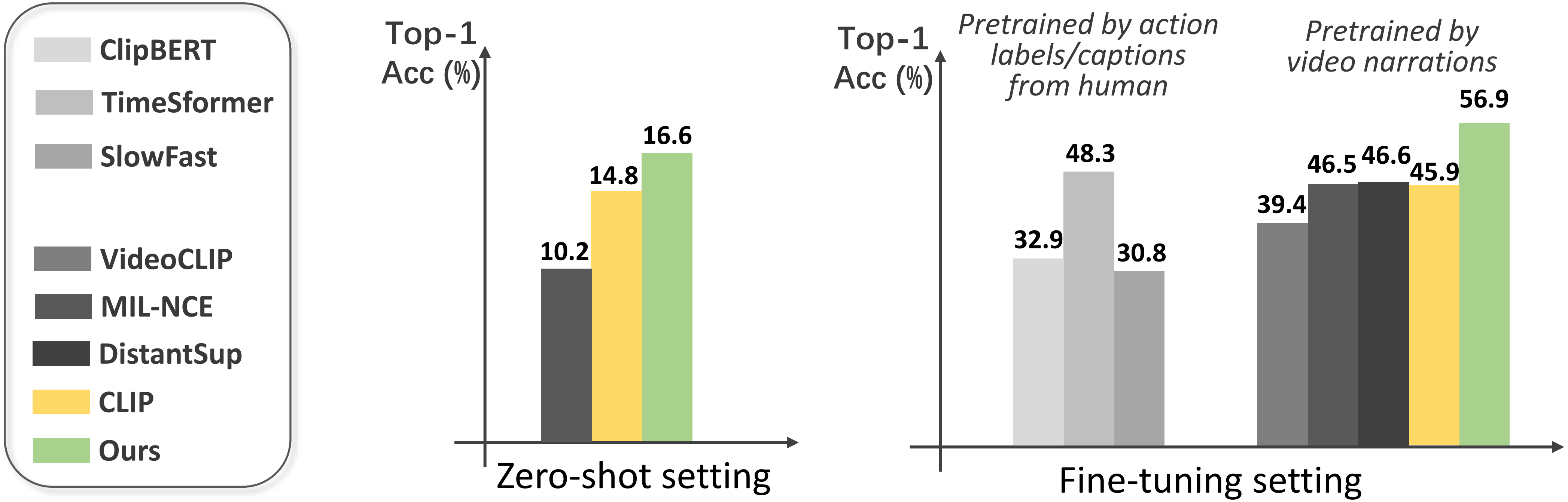
Lin, et al., Learning To Recognize Procedural Activities with Distant Supervision, CVPR 2022

Step Classification Results on COIN Benchmark



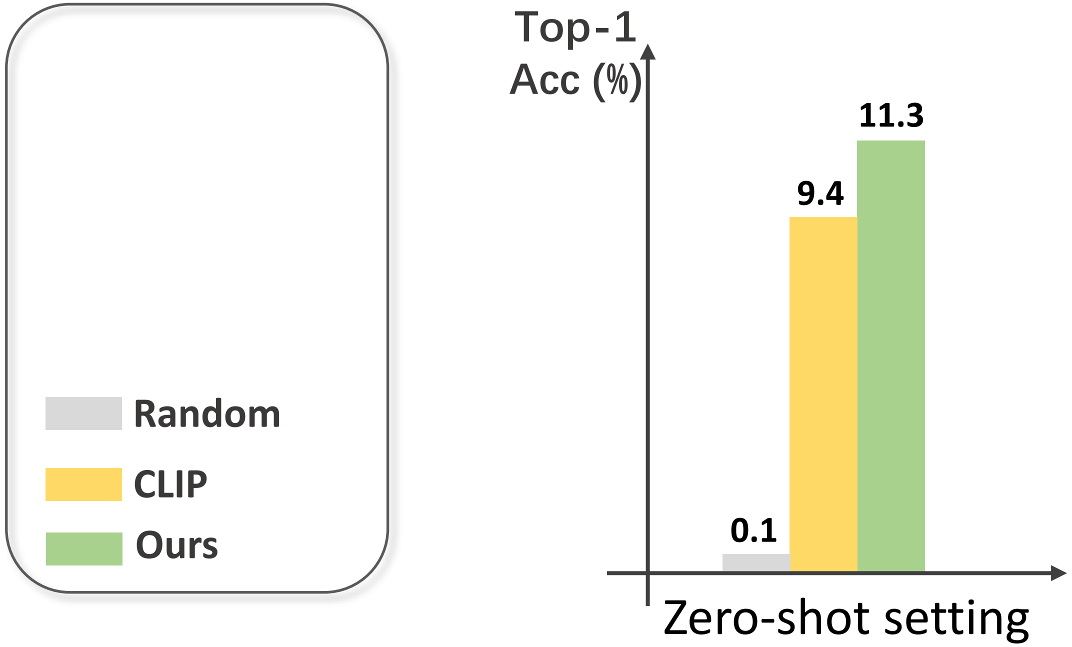
Feichtenhofer, et al., SlowFast networks for video recognition, ICCV 2019
Miech, et al., End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020
Lei, et al., Less is more: Clipbert for video-and-language learning via sparse sampling, CVPR 2021
Bertasius, et al., Is space-time attention all you need for video understanding, ICML 2021
Xu, et al., VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding, EMNLP 2021
Lin, et al., Learning To Recognize Procedural Activities with Distant Supervision, CVPR 2022

Step Classification Results on COIN Benchmark



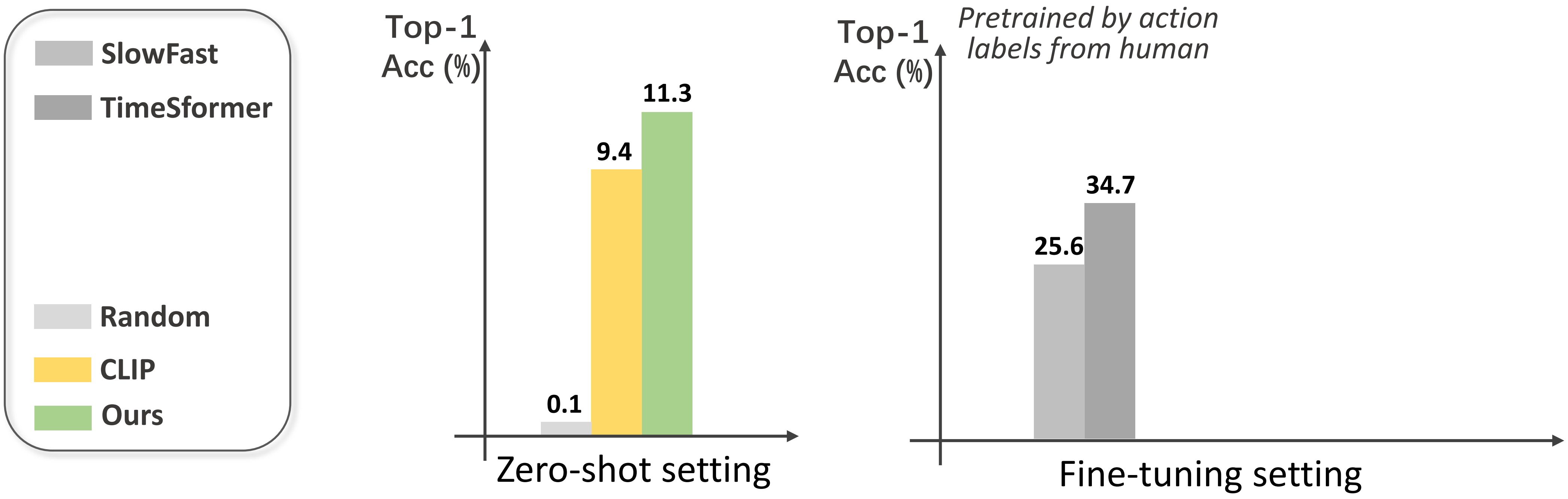
- New **state-of-the-art results** on both zero-shot & fine-tuning settings
- Our **procedure-aware** pretraining learns high-quality video representation

Step Forecasting Results on COIN Benchmark



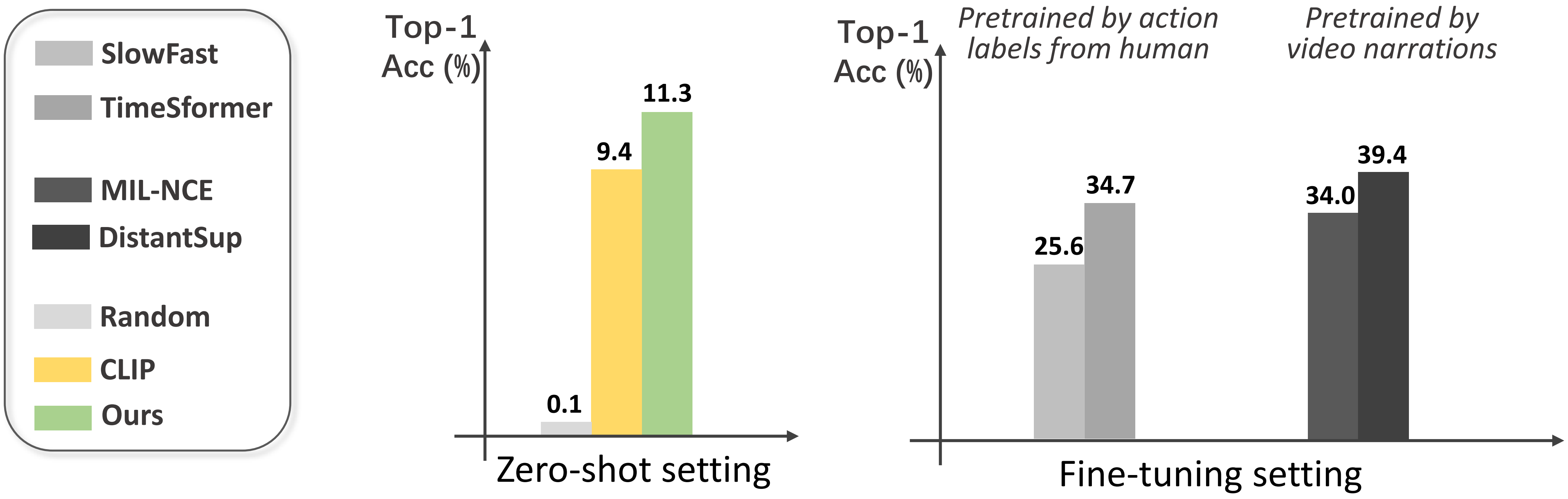
- We're the **first work** that supports **zero-shot forecasting** by learning from unannotated videos

Step Forecasting Results on COIN Benchmark



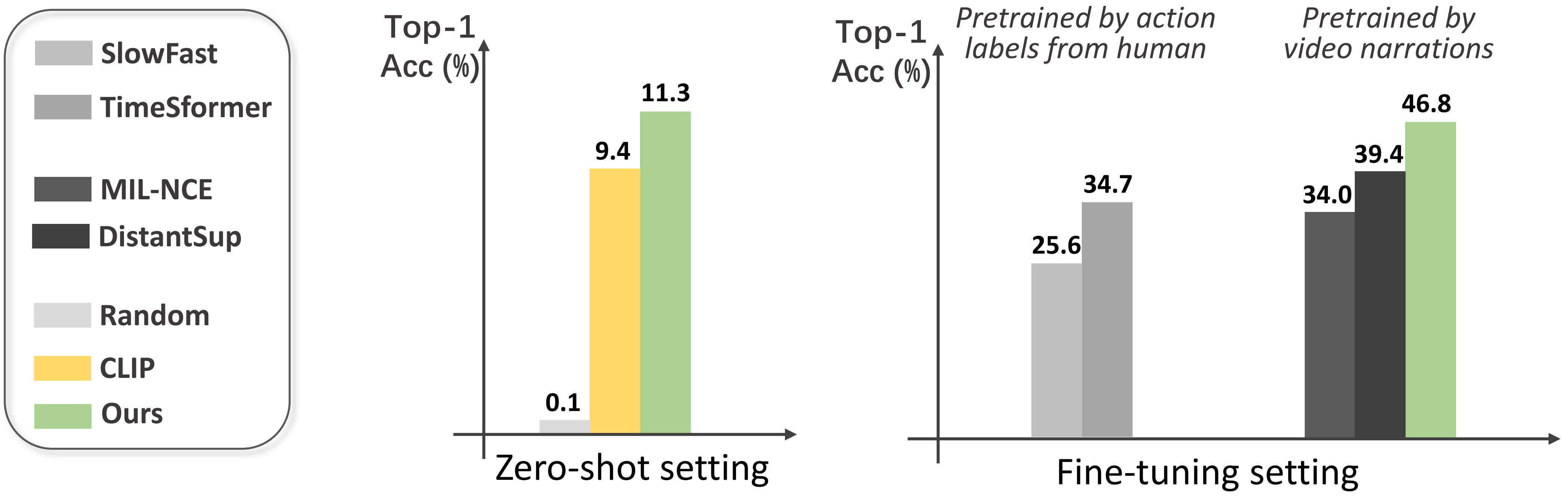
- We're the **first work** that supports **zero-shot forecasting** by learning from unannotated videos

Step Forecasting Results on COIN Benchmark



- We're the **first work** that supports **zero-shot forecasting** by learning from unannotated videos

Step Forecasting Results on COIN Benchmark



- We're the **first work** that supports **zero-shot forecasting** by learning from unannotated videos
- Our learned video representation **largely** facilitates fine-tuning setting

Zero-shot Step Forecasting

Model input: videos

Diverse predictions and generated key frames for next step

flatten the dough

bake pizza

bake cookies

Given an input video, our model outputs diverse predictions for next step

Key Frame Generation

After forecasting, the step description is used for image generation via Stable Diffusion

Rombach, et al., High-Resolution Image Synthesis with Latent Diffusion Models, CVPR 2022

Zero-shot Step Forecasting & Key Frame Generation

Model input: videos

Diverse predictions and generated key frames for next step


flatten the dough


bake pizza


bake cookies


pour some salt to the garlics


put the ingredients into the bowl


prepare seasonings and side dishes

Our model forecasts next step, which is further used for image generation via Stable Diffusion

Conclusion

- ProcedureVRL: learns **procedure-aware** video representation from instructional videos and their narrations, **without** human annotation
- Key technical innovation: joint learning of **video representations of action steps**, as well as **a diffusion model capturing the temporal ordering of the steps**
- Results: **new state of the art** in both step classification and forecasting on instructional video benchmarks, supporting zero-shot forecasting, diverse step prediction, and key frame generation

Code: https://github.com/facebookresearch/ProcedureVRL

Thank you!