# From Action Segmentation to Skill Understanding

*Guodong Ding*

*National University of Singapore*

# What People Do in Daily Lives

## Actions

*Short trimmed, isolated actions*

Apply Eye Makeup

Play Dhol

Shaving Beard

Surfing

## Procedures

*a set of actions, structured temporal relationship*

Grind Beans

Tamp Coffee

Extract Espresso

Pour Milk

# Advancing from Actions to Procedures

## Recognizing Actions

*What happened?*


Apply Eye Makeup


Play Dhol


Shaving Beard


Surfing

## Understanding Procedures

*When and in what order?*


Grind Beans


Tamp Coffee


Extract Espresso


Pour Milk

Procedure = Structured Composition of Actions

# Outline

01

How do we learn from procedures?

02

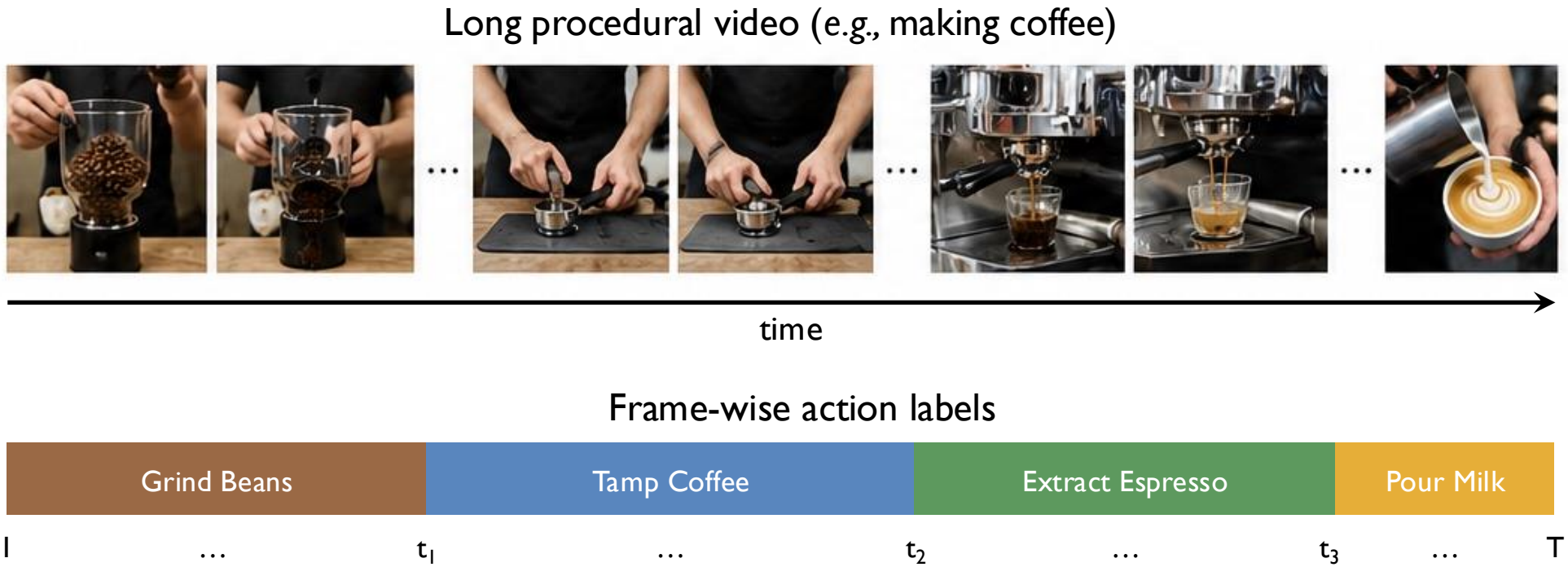
What breaks in the real world?

03

What lies beyond the procedures?

# Understanding Procedures from Videos

## Temporal Action Segmentation




local semantics + temporal structure

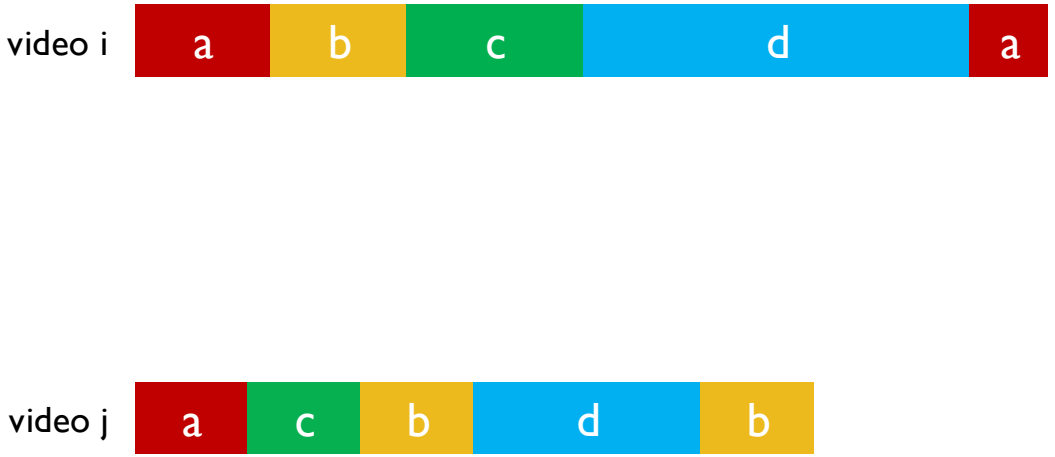
# Long-range Temporal Modeling of Procedures

- Recurrent Neural Networks
  - *Bi-directional GRUs [a],[b]*
- Temporal Convolutional Networks
  - *Encoder-decoder [c],[d]*
  - *Multi-stage TCN [e],[f]*
- Attention & Transformers
  - *Temporal Aggregates [g]*
  - *ASFormer [h]*
  - *UVAST [i]*


[a] Singh et al, CVPR'16
[b] Richard et al, CVPR'17
[c] Lea et al, CVPR'17
[d] Lei et al, CVPR'18
[e] Farha et al, CVPR '19
[f] Singhania et al, TPAMI'22
[g] Sener et al, ECCV'20
[h] Yi et al, BMVC'21
[i] Behrmann, ECCV'22

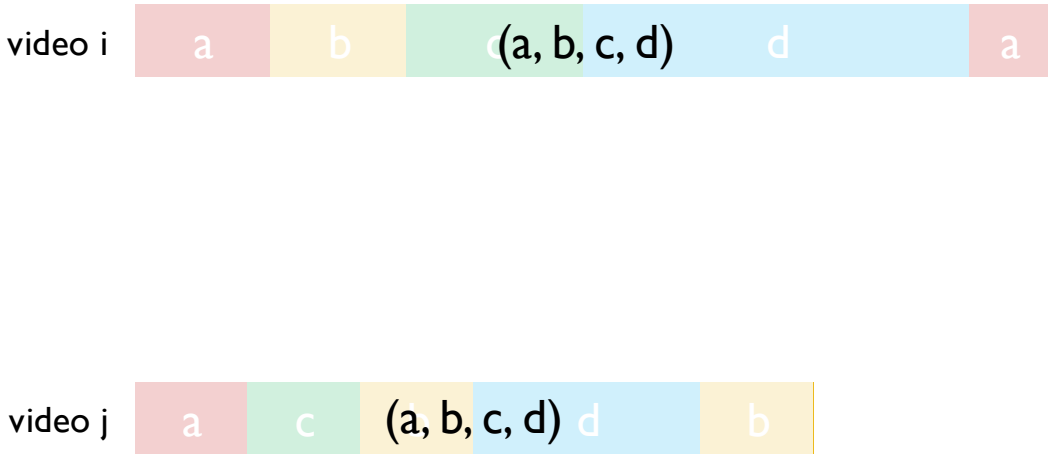
# Learning under Different Supervision Regimes

- Fully-supervised
  - *Frame-wise action labels*



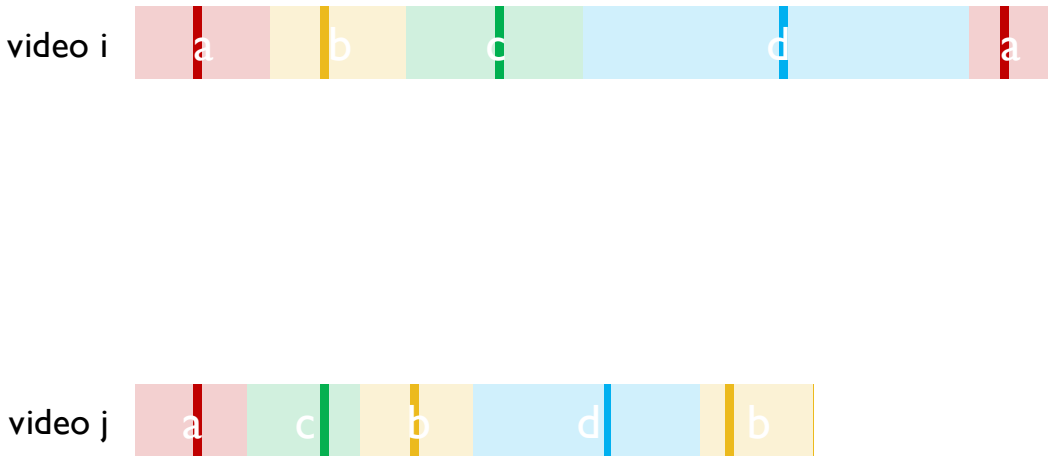
# Learning under Different Supervision Regimes

- Fully-supervised
  - *Frame-wise action labels*
- Weakly-supervised
  - *Transcript / Action sets*



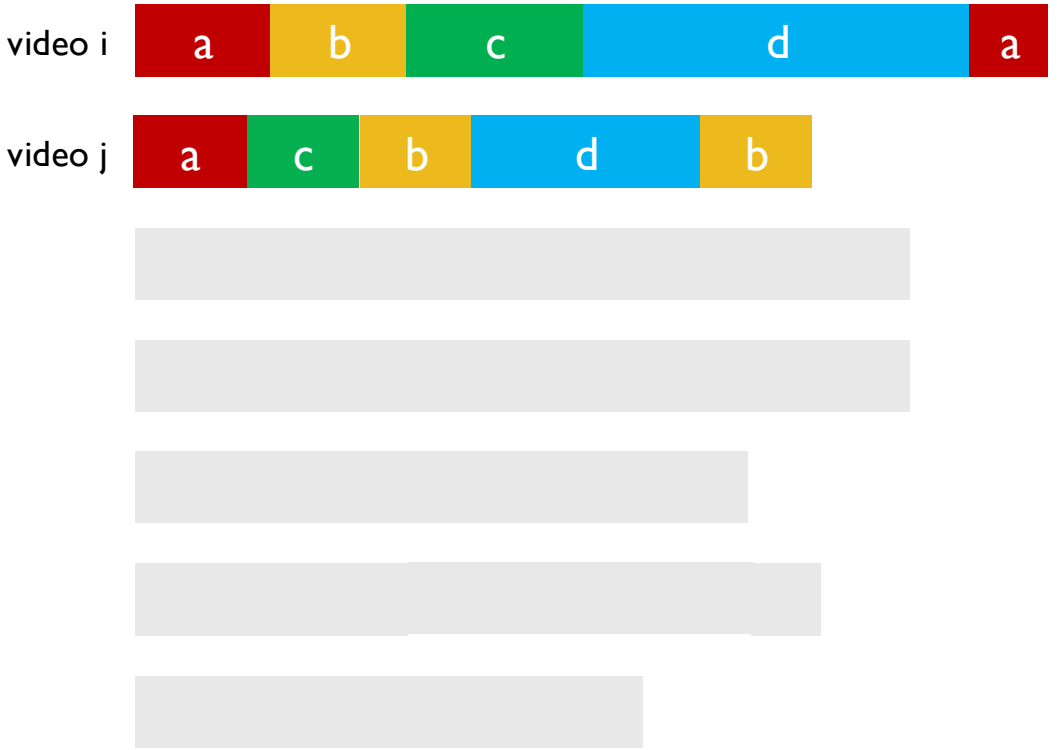
# Learning under Different Supervision Regimes

- Fully-supervised
  - *Frame-wise action labels*
- Weakly-supervised
  - *Transcript / Action sets*
  - *Sparse Timestamps*



# Learning under Different Supervision Regimes

- Fully-supervised
  - *Frame-wise action labels*
- Weakly-supervised
  - *Transcript / Action sets*
  - *Sparse Timestamps*
- Semi-supervised
  - *Limited labeled videos + large unlabeled collection*



# Learning under Different Supervision Regimes

- Fully-supervised
  - *Frame-wise action labels*
- Weakly-supervised
  - *Transcript / Action sets*
  - *Sparse Timestamps*
- Semi-supervised
  - *Limited labeled videos + large unlabeled collection*
- Unsupervised
  - *No action annotation at all*
  - *(all videos belong to the same procedure)*

# What we Learned so far about Procedures

- Procedures organize actions into meaningful structure.
- Long-range temporal modeling is essential for procedural videos.
- Procedural structure can be helpful when supervision is incomplete.

TAS Survey

**G. Ding**, F. Sener and A. Yao. Temporal Action Segmentation: An Analysis of Modern Techniques,. TPAMI 2023.

# Outline

01

How do we learn from procedures?

02

What breaks in the real world?

03

What lies beyond the procedures?

# From Benchmarks to the Real World

Some benchmark assumptions may not hold in practice.

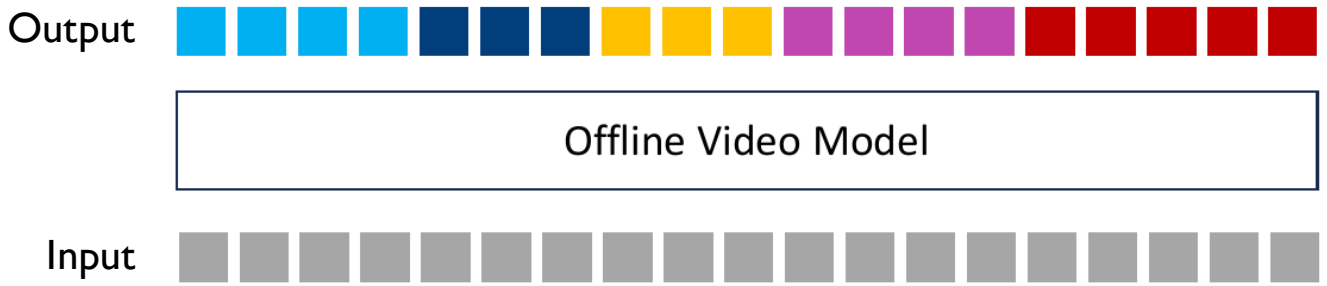
| | Benchmark | Real World | Solution |
|---|---|---|---|
| Data Access | Full video sequence available | Observations arrive continuously | Online procedural understanding |

# Breaking the Full Data Access Assumption

## What makes online understanding difficult?

### Offline

*Access to all frames*
*Complete context:*
*past and future observations*




### Online

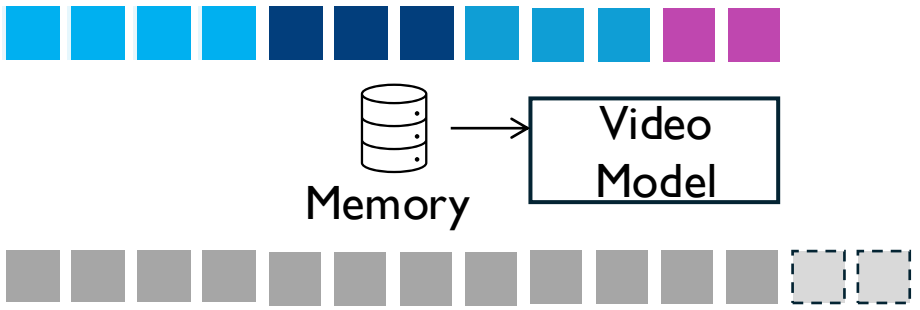
*Causal access to past frames*
*Incomplete context:*
*future is missing.*

# Modeling Context in a Streaming Setup

"Now" only


A sliding window suffices for isolated actions.
Observations beyond the window are discarded.

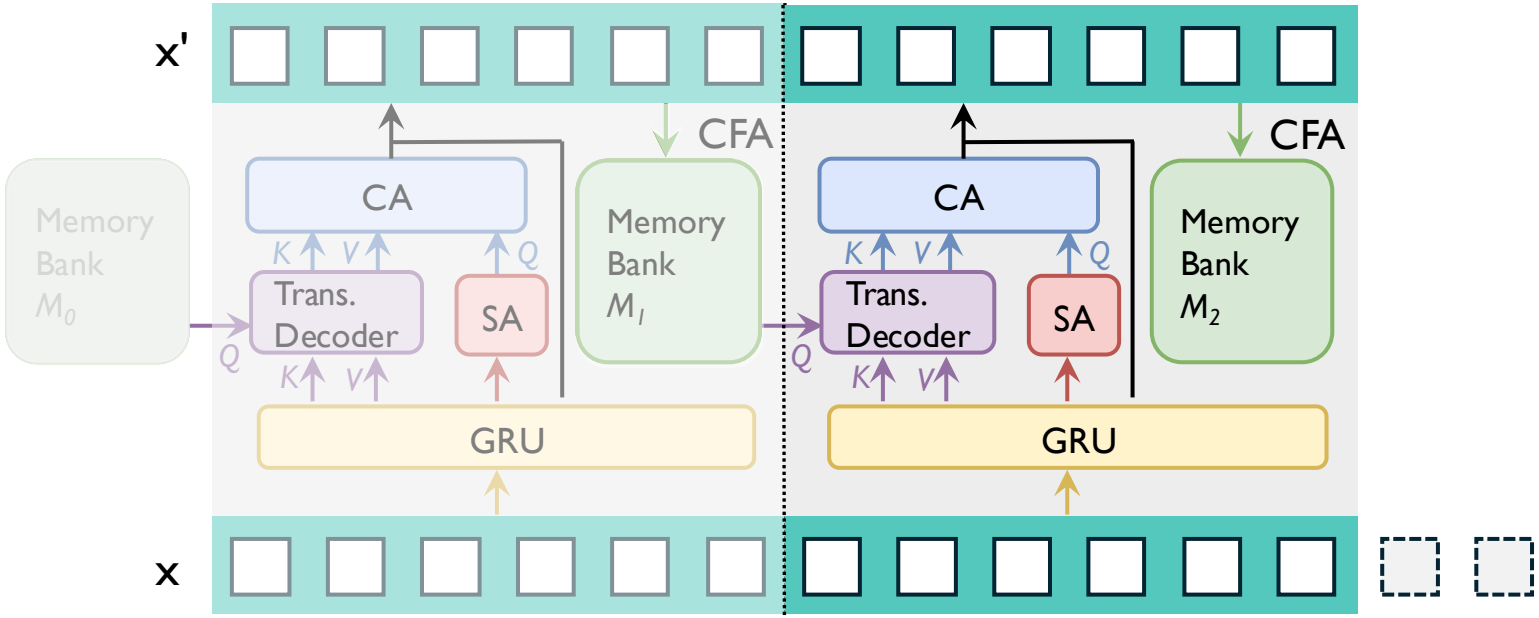
"Now" + Memory




Long-range support for current inference.
Store the past as context in memory.

# Context-aware Feature Augmentation

We incorporate the temporal context into inputs.




Plug-and-play, compatible with different backbones.

Q. Zhong, **G. Ding** and A. Yao. OnlineTAS: An Online Baseline for Temporal Action Segmentation. NeurIPS 2024.

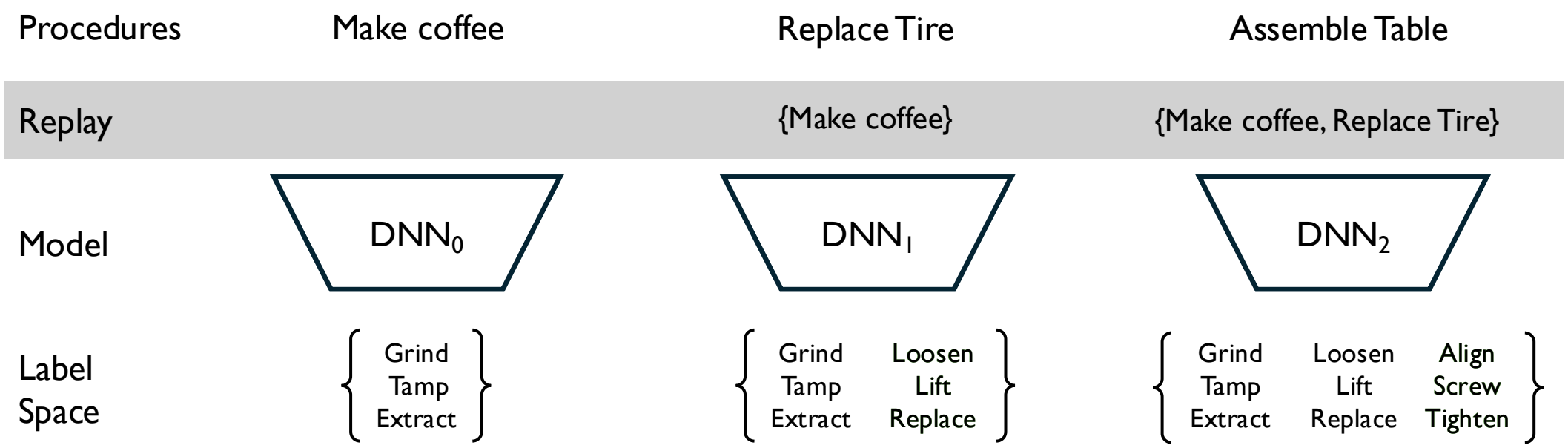
# From Benchmarks to the Real World

Some benchmark assumptions may not hold in practice.

| | Benchmark | Real World | Solution |
|---|---|---|---|
| Data Access | Full video sequence available | Observations arrive continuously | Online procedural understanding |
| Label Space | Fixed procedure set known in advance | New procedures emerge over time | Incremental procedural understanding |

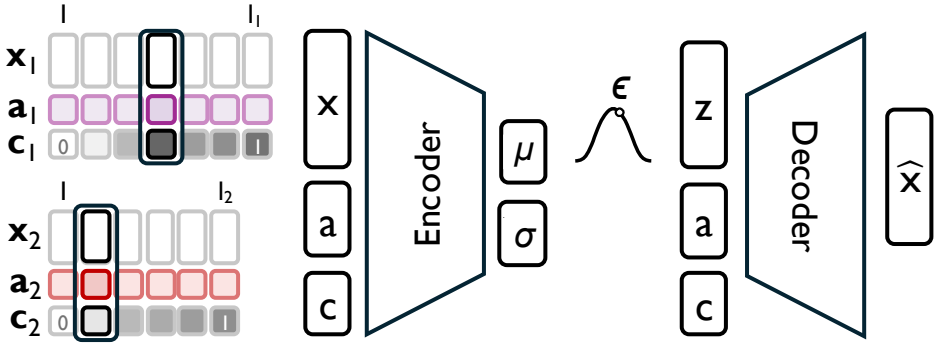
# Breaking the Static Label Space Assumption

What does it mean to handle new procedures incrementally?



# Generative Replay for Incremental Learning

## Action Modeling

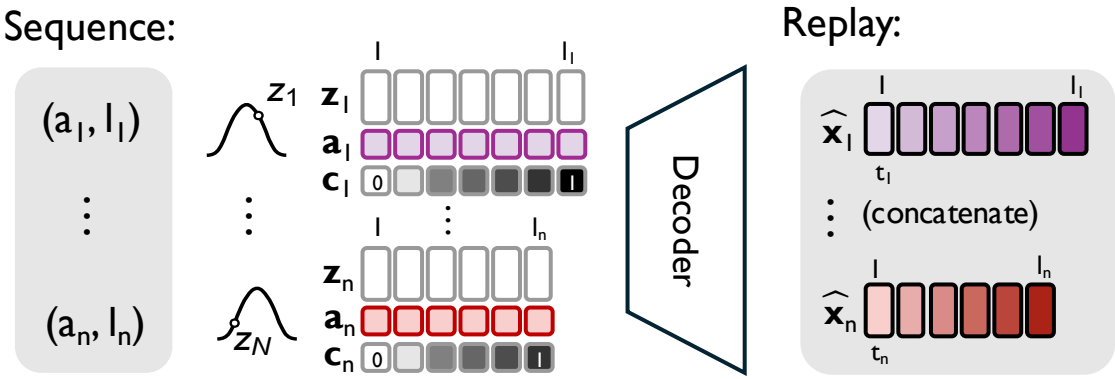



*x: frame feature* a: *action label* c: *coherence variable*

$$\mathcal{L}_{\text{TCA}} = \underbrace{\mathbb{E}_z \log p_\theta(x|z,a,c)}_{\mathcal{L}_{\text{recon}}} - \underbrace{\text{D}_{\text{KL}}(q_\phi(z|x,a,c)||p(z))}_{\mathcal{L}_{\text{reg}}}$$

Learn a generative representation of actions while ensuring temporal coherence

## Replay Generation



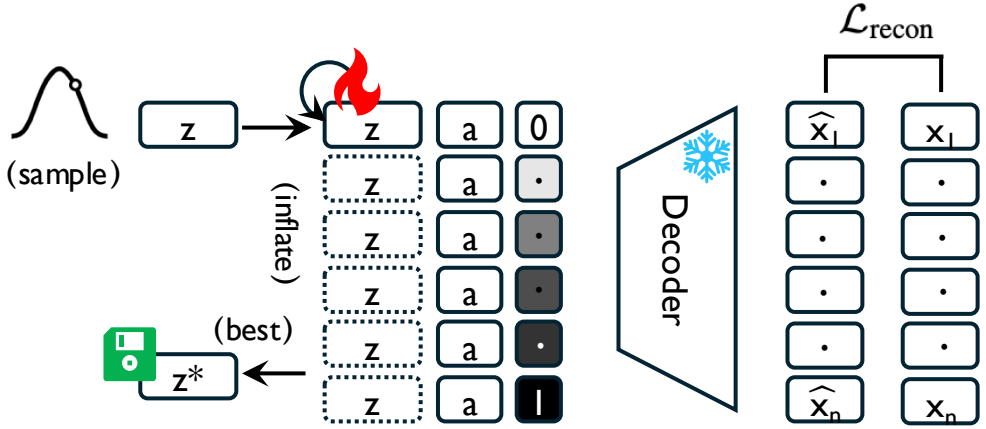

1. Randomly sample a latent code z for one action
2. Repeat z for entire segment
3. Decode segment through the Decoder
4. Repeat for other segments
5. Concatenate in order as a replay video

**G. Ding**, H. Golong and A. Yao. Coherent Temporal Synthesis for Incremental Action Segmentation. CVPR 2024

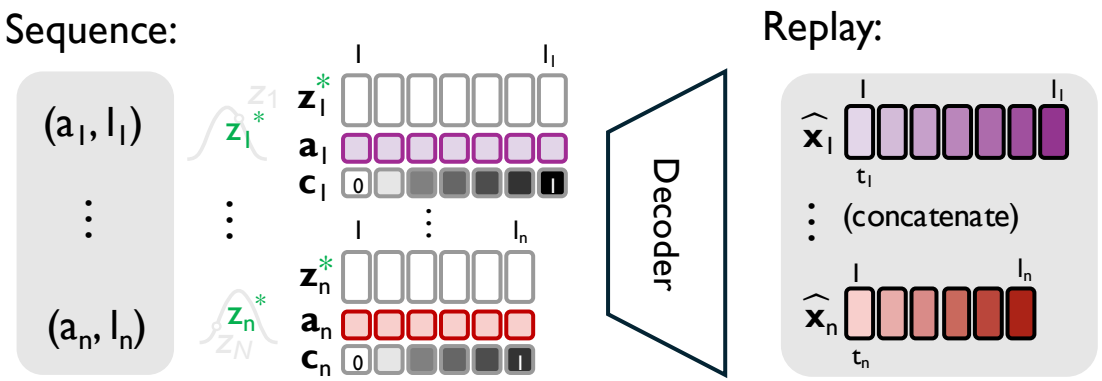
# Improving Replay Fidelity

Randomly sampled latent codes often generate low-fidelity replay.

## Network Inversion




## Enhanced Replay




Representative latent codes produce more effective replay.

**G. Ding**, R. Chen and A. Yao. Condensing Action Segmentation Datasets via Generative Network Inversion. CVPR 2025

# Outline

01

How do we learn from procedures?

02

What breaks in the real world?

**03**

**What lies beyond the procedures?**

# From Procedures to Skills

Understanding of how well a task is performed.

| Make coffee | Replace Tire | Assembly Table |
| --- | --- | --- |
| Procedure completed. | All steps executed. | Procedure followed. |
| But was it done correctly? | But does the worker need assistance? | But is the trainee skilled to work independently? |

Actions and procedures are observable. Skill is a bit latent.

# Mistake Detection as a Window to Skills

Detect the correctness of the ongoing action.






| | | | | | | |
|---|---|---|---|---|---|---|
| Perception | Attach chassis, wheel | Attach base, chassis | Attach cabin, chassis | Detach cabin, chassis | Attach interior, chassis | Attach cabin, chassis |
| Reasoning | correct | correct | mistake | correction | correct | correct |

End-to-end training couples semantic perception and sequential reasoning, reducing interpretability.

We disentangle perception and reasoning. For perception, we train an action recognition model; for reasoning, we propose spatial and temporal beliefs as explicit representation for decision-making.

Sener et al, Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities, CVPR2022.

# Spatial & Temporal Beliefs

- Spatial beliefs

*structural constraints between components, what two parts can be attached*

- attach(roof, cabin), attach(wheels, chassis) Plausible, correct
- attach(roof, wheels) infeasible, mistake

- Temporal beliefs

*sequential constraints of steps, what action(s) must happen before what action*

- attach(interior, chassis) be done before attach(cabin, chassis) * the interior sits within the cabin

Mistake occurs if order is violated

**G. Ding**, F. Sener, S. Ma and A. Yao. Spatial and temporal beliefs for mistake detection in assembly tasks. CVIU 2025

# Online Building of Spatial & Temporal Beliefs

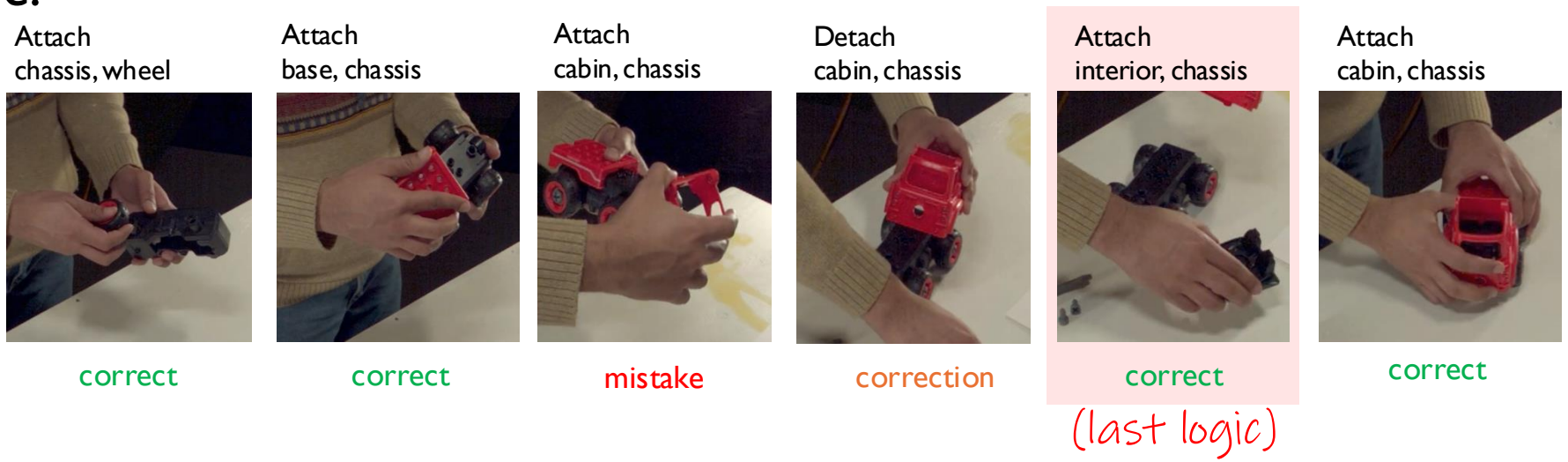
## Basic logics

- A correct action verifies the spatial relationship.
- Later actions don't form temporal links to previous correct actions.
- An error action's dependencies fall between its fix and correct execution.

## A running example:

Spatial Beliefs
- (chassis, wheel)
- (base, chassis)
- (interior, chassis)
- (cabin, chassis)

Temporal Beliefs
- (cabin, chassis) must be after (interior, chassis)




**G. Ding**, F. Sener, S. Ma and A. Yao. Spatial and temporal beliefs for mistake detection in assembly tasks. CVIU 2025

# Using Beliefs for Mistake Detection

How to detect mistakes

- Predict action labels for video segments with recognition model
- Check the rules in both spatial and temporal beliefs
- If satisfied, then correct; otherwise, a mistake

These belief sets are

- explicit, easy to comprehend
- can be hard to generalize
- The performance of rule-based reasoner heavily relies on the accuracy of perception module (action recognition model)

# Tasks towards Skill Understanding

- Mistake Detection
  - *Assembly101 [a]*
  - *PREGO [b]*
- Skill Assessment
  - *Skill Ranking [c]*
  - *ProSkill [d]*
  - *SkillSight [e]*
- Struggle Localization
  - *EvoStruggle [f]*

[a] Sener et al, CVPR'22
[b] Flaborea et al, CVPR'24.
[c] Doughty et al, CVPR'18
[d] Mazzamuto et al, WACV'26
[e] We et al, CVPR'26
[f] Feng et al, ICPR'26

# Open Problems & Outlook

- How should we represent procedural skills?
  - *Action- or procedure-level?*
- How do we evaluate procedural intelligence?
  - *Task success?*
  - *Execution quality & error recovery?*
  - *Generalization?*
- What procedural skill benchmarks are missing?
  - Mistake and error recovery
  - Skill progression

# ~~Outline~~ Takeaways

01

Procedures require long-horizon temporal modeling.

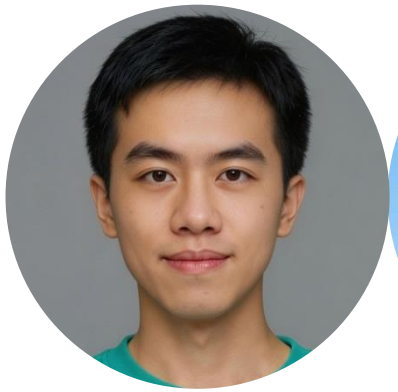
02

Real-world deployment demands streaming and adaptation.

03

Skills require reasoning beyond actions and procedures.

# Acknowledgements



Rongyu Chen


Qing Zhong


Hans Golong


Shugao Ma


Fadime Sener


Angela Yao


# Thanks!