NUS National University of Singapore

Coherent Temporal Synthesis for Incremental Action Segmentation

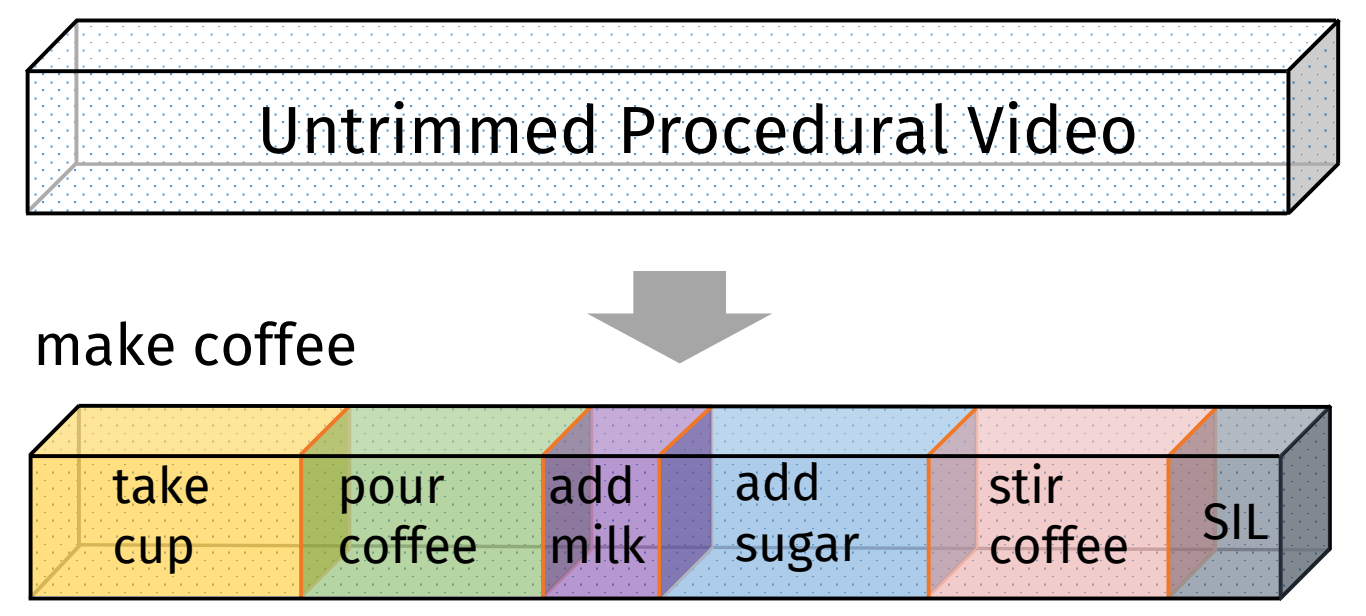
Guodong Ding, Hans Golong, and Angela Yao
National University of Singapore

CVPR SEATTLE, WA JUNE 17-21, 2024

Task & Motivation

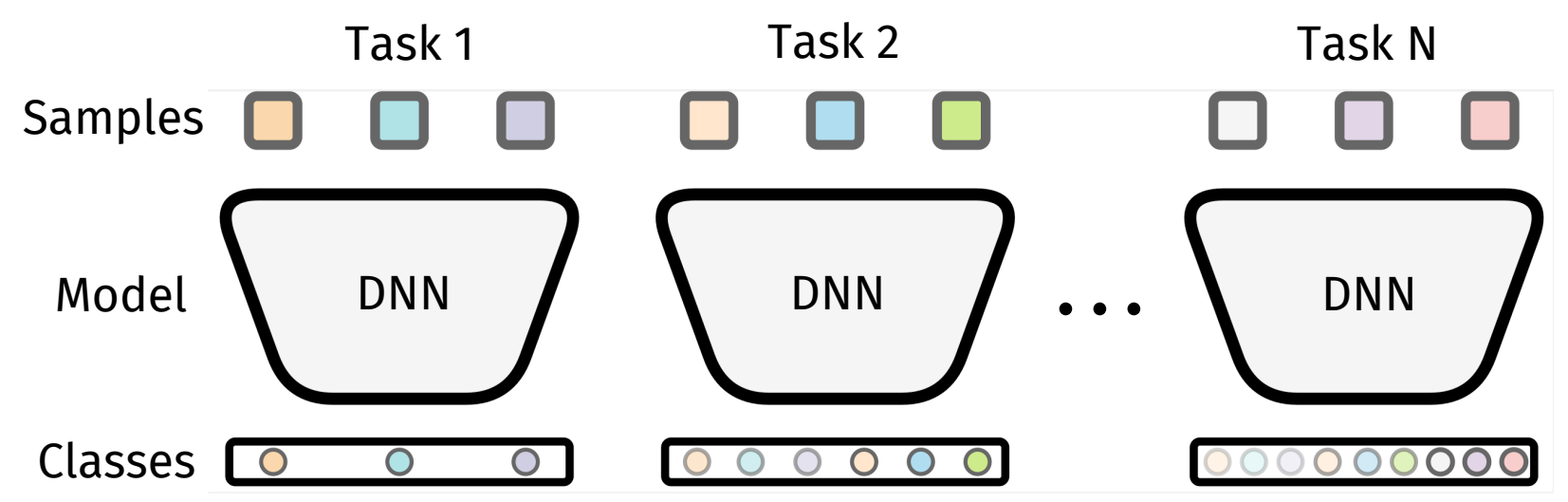
Temporal Action Segmentation

Temporally segment untrimmed procedural videos and assign **frame-wise semantic labels**



Continual Learning

- Each procedural **activity as** a novel **task**
- **Learn new activity** without **catastrophic forgetting**



Data Replay

Re-expose model to **samples of previous tasks**

Challenges

- How can minute-long videos be **stored efficiently** for replay?
- Can **temporal continuity** be maintained in replay videos?

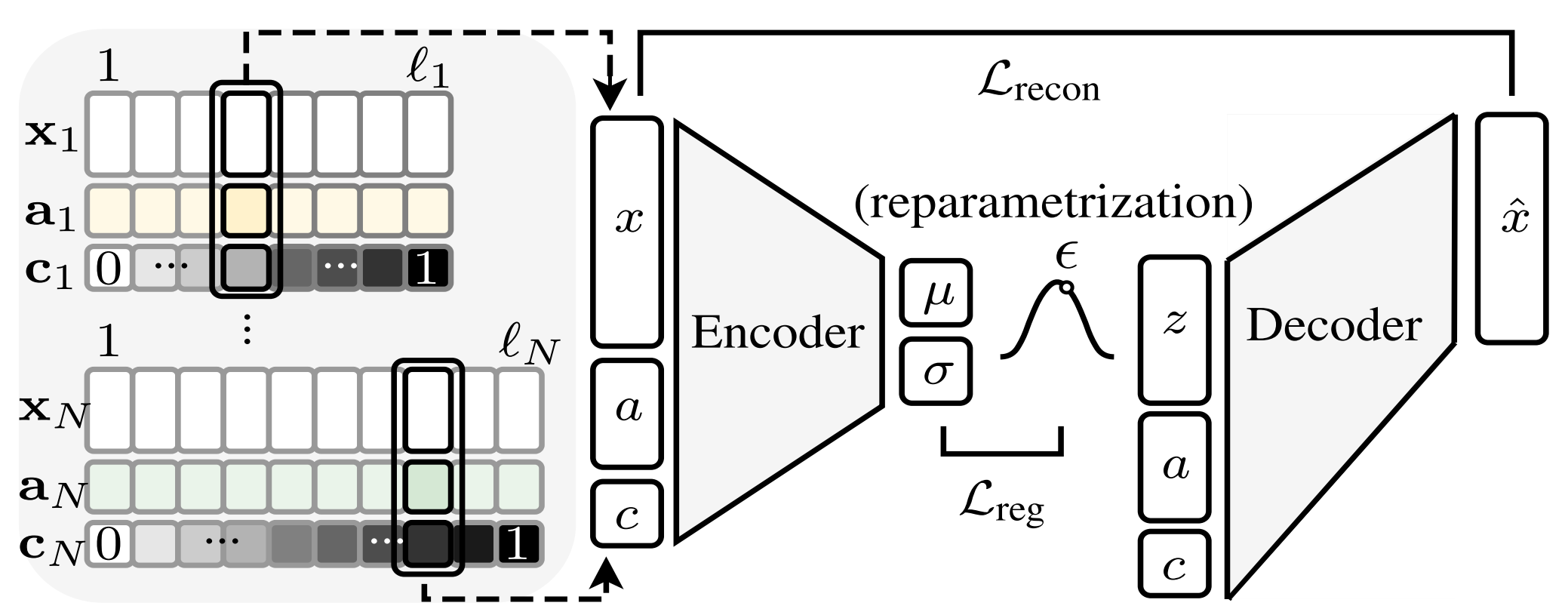
Approach

Temporally Coherent Action Modeling

Model actions with a conditional VAE conditioned on:

- one-hot action label a_n
- temporal coherence c_i

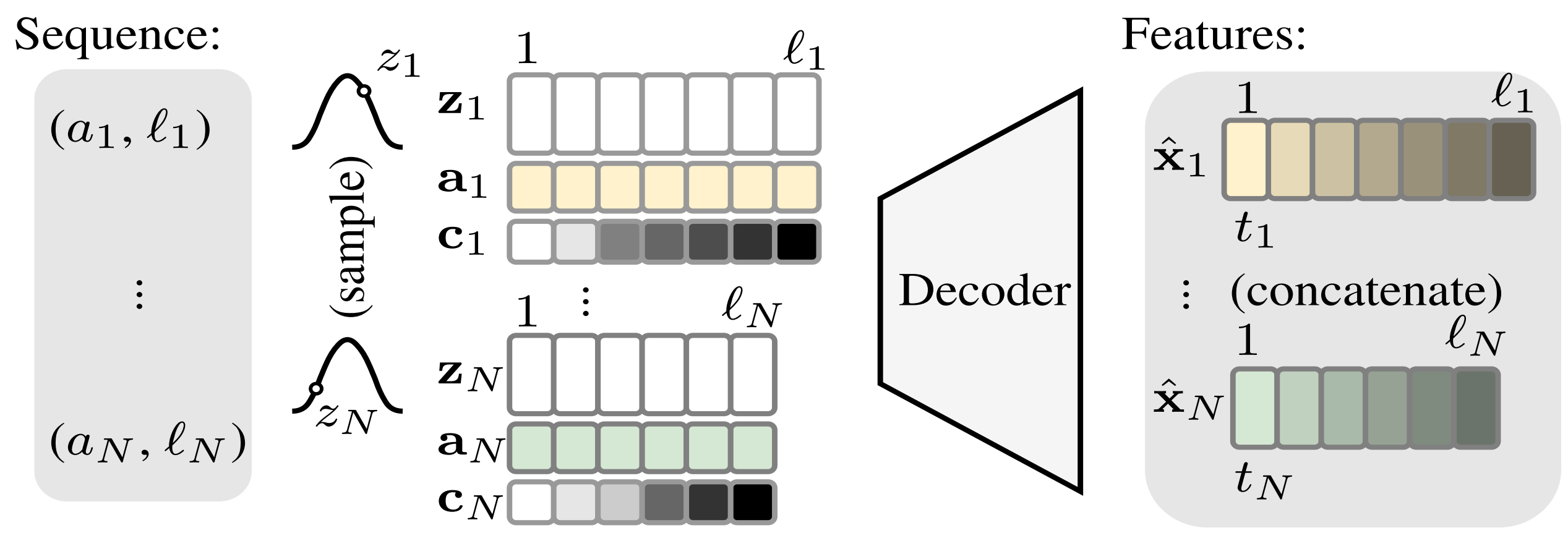
$c_i = (i-1)/(\ell-1)$ and $c_i \in [0, 1]$



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

Coherence variable helps to model **how features evolve as an action progresses**

Replay Data Generation



- Symbolic action sequences sampling

$$\hat{\mathbf{s}}_b \sim \text{Uniform}(\mathbf{S}_b)$$

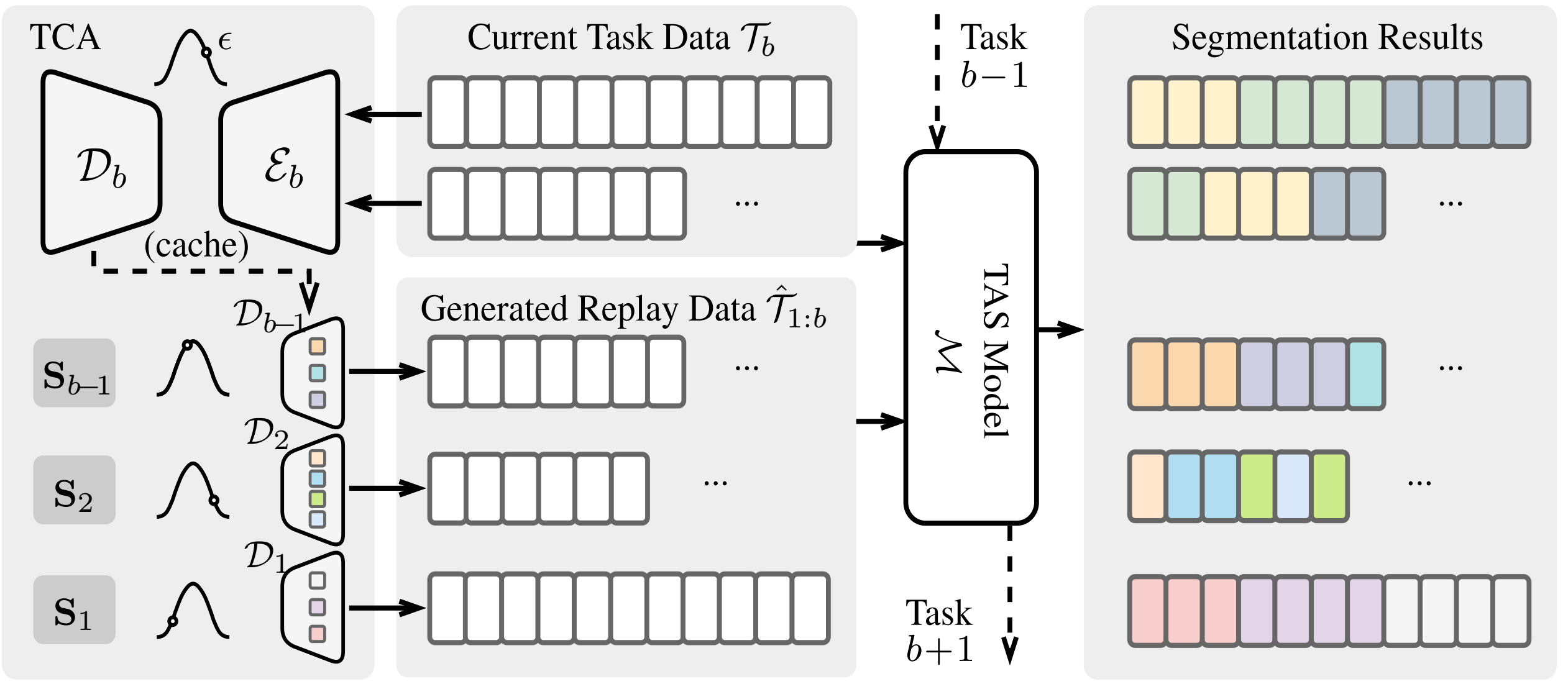
- Generate temporally coherent segments

$$\hat{x}_i = p_\theta(x|z_n, a_n, c_i) \text{ and } i \in [1, ..., \ell]$$

- Concatenate segments to form videos

$$\hat{v} = \text{concat}(\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_{\hat{N}})$$

Incremental Training



Action Segmentation

- Construct replay data with past task generators
- Train segmentation model on new and replay data

Video Replay

- Train generator on incoming data
- Cache generator alongside past task models

Results

Performance

# Tasks		MSTCN					ASFormer				
		Acc	Edit	F1 @ {10, 25, 50}			Acc	Edit	F1 @ {10, 25, 50}		
						Breakfast					
10	Finetune	7.4	7.2	7.5	7.0	5.4	9.9	9.8	10.3	9.4	7.5
	Exemplar	16.1	13.3	13.8	12.5	9.5	12.4	11.2	11.7	10.7	8.5
	Ours	**29.4**	**25.9**	**26.3**	**23.5**	**17.7**	**34.2**	**32.4**	**33.1**	**30.1**	**23.4**
	Original	43.1	41.1	41.2	37.6	29.5	48.1	45.2	45.9	42.4	34.2
5	Finetune	15.4	15.8	16.6	15.8	12.7	15.7	16.1	16.9	15.8	13.2
	Exemplar	32.5	28.9	30.8	28.5	22.9	29.5	27.5	28.7	26.7	22.0
	Ours	**54.5**	**49.4**	**51.1**	**46.9**	**37.7**	**57.2**	**56.8**	**58.3**	**54.0**	**43.6**
	Original	60.4	59.1	60.3	56.1	46.0	65.1	64.2	65.6	61.5	51.0
						YouTube Instructional					
5	Finetune	13.6	2.8	3.6	2.7	0.6	13.9	11.5	11.1	9.8	6.3
	Exemplar	**30.8**	19.7	19.8	16.0	9.3	22.1	18.9	17.7	15.3	10.0
	Ours	30.2	**25.0**	**21.9**	**18.5**	**11.1**	**25.2**	**20.9**	**20.1**	**17.5**	**11.4**
	Original	55.9	39.4	38.1	32.2	19.1	59.2	51.1	45.4	39.1	25.5

- Significant improvements over baselines
- Performance gap compared to direct sampling

Diversity vs. Coherence

	SD	FD	TC	Acc	Edit	F1 @ {10, 25, 50}		
Exemplar	✓	✗	✗	27.8	35.6	36.1	31.7	24.3
Ours$_{\text{random}}$	✓	✓	✗	32.9	38.9	40.0	35.6	27.2
Ours$_{\text{static}}$	✓	✗	✗	37.9	42.9	43.8	38.9	29.0
Ours	✓	✓	✓	**41.8**	**45.0**	**47.0**	**41.5**	**32.0**

SD – segment-level diversity
FD – feature diversity
TC – temporal coherence

- Without temporal coherence, static segments work better than segments with high diversity
- Balance of diversity and coherence achieves best performance

Replay Buffer Size

M	Acc	Edit	F1 @ {10, 25, 50}		
30	34.0	39.6	41.0	34.8	24.7
60	35.4	41.2	42.3	36.0	25.6
90	36.2	**42.3**	43.9	**37.3**	**26.8**
120	**38.0**	**42.3**	**44.0**	37.1	26.2

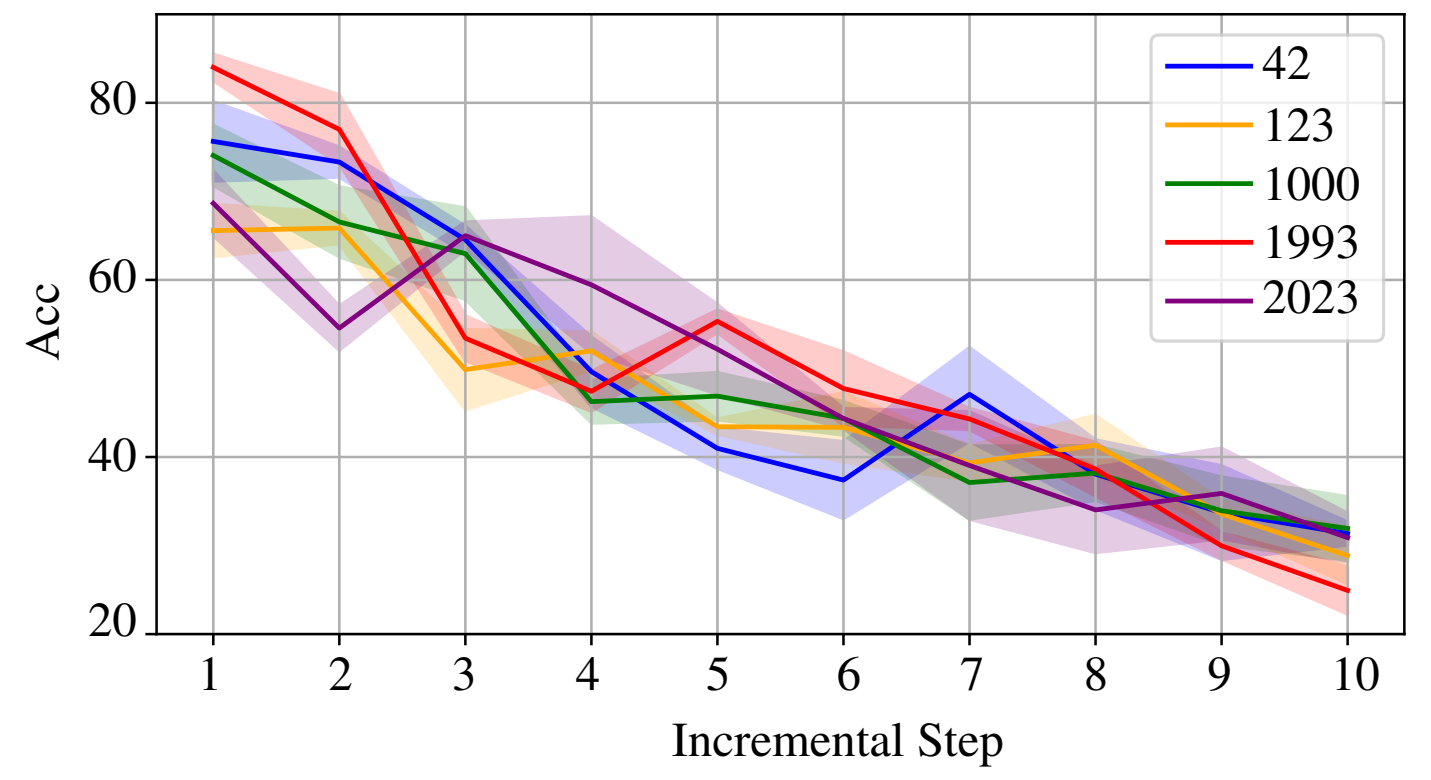
Larger replay size gets better performance

TCA Training Data

	$\mathcal{T}(\%)$	Acc	Edit	F1 @ {10, 25, 50}		
Exemplar	-	22.6	34.8	36.0	32.4	25.2
Ours	25	41.7	43.2	46.1	40.9	31.5
	50	42.1	43.3	45.1	40.5	31.5
	75	45.3	45.9	47.8	**43.7**	**34.7**
	100	**47.4**	**46.9**	**48.2**	42.8	33.4

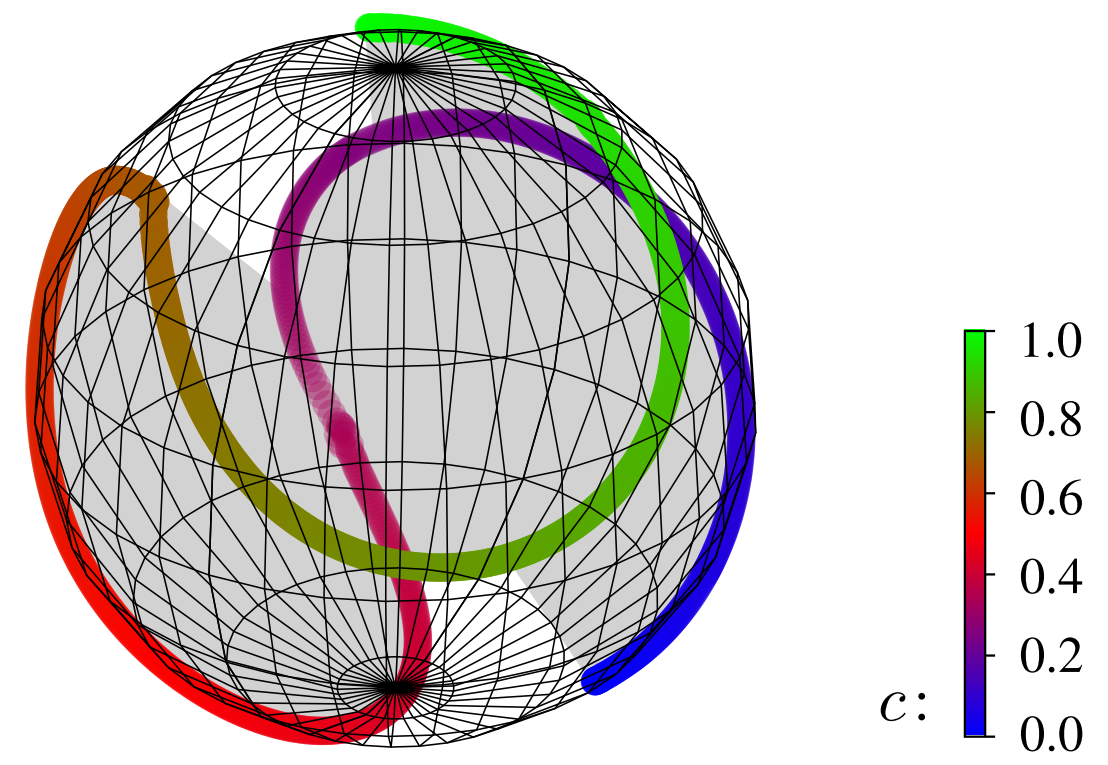
More real data gets better generative ability

Task Sequence Order



- All task sequence orders follow a downward trend
- Task order matters, final performance can differ up to 7%

Segment Visualization



Action segment exhibits continuity in feature space

Takeaways

Temporal Coherence

Temporal coherence of generated training samples is necessary for good performance

Generative Video Replay

Generative data replay approaches work well for incremental video understanding

Incremental Action Segmentation

Incremental learning in videos is underexplored and warrants further exploration