# Condensing Action Segmentation Datasets via Generative Network Inversion

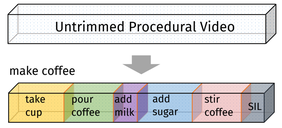
Guodong Ding, Rongyu Chen and Angela Yao
National University of Singapore

CVPR Nashville JUNE 11-15, 2025

## Task & Challenges

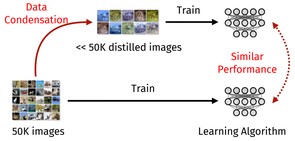
### Temporal Action Segmentation

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




### Dataset Condensation

- Synthesizes a **small set** that **trains models as effectively as** the **full dataset**




### Challenges

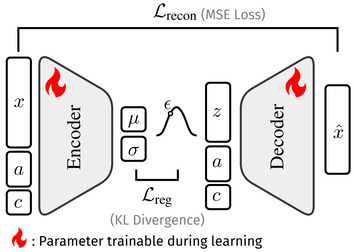
- **Hierarchical redundancy** exists in TAS data, *e.g.,* feature, temporal, and action ordering in the sequence
- How to **restore the actual temporal resolution** if condensed for TAS learning?
- How to ensure **condensed data match the quality of the original**?

## Approach

### 1. Generative *Feature & Temporal* Condensation

**Generative Action Model**

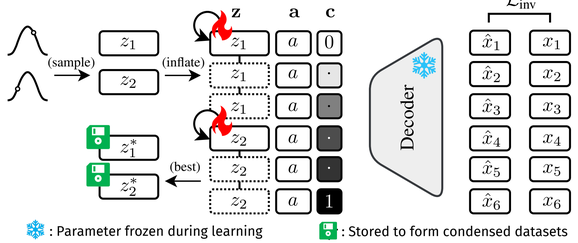
Learn action prior




- frame feature $x$ • action label $a$
- action progress $c_i = (i-1)/(\ell-1)$

**Network Inversion**

Enforce generation resemblance




- Optimize latent codes representing the action segment with:

$$z^* = \arg\min_{z \in \mathbb{R}^d} ||\mathbf{D}(\mathbf{z}, \mathbf{a}, \mathbf{c}) - \mathbf{x}||_2^2,$$

- Multiple codes per segment enable finer-grained and more precise inversion *(Instances Per Segment)*

### 2. Diverse *Sequence* Sampling

Preserve sequence structure and reduce redundancy

- Select sequences that maximize diversity in action orderings until the target count is reached:

$$\mathbf{s}^* = \arg\max_{\mathbf{s}_i \in \mathcal{D} \setminus \mathcal{S}} \min_{\mathbf{s}_j \in \mathcal{S}} \text{Edit}(\mathbf{s}_i, \mathbf{s}_j),$$

$\mathcal{D}$ the full sequence set, $\mathcal{S}$ the sampled set, such that $|\mathcal{S}| = \gamma|\mathcal{D}|$

### 3. Decoding and Training

- Decode full-resolution input for TAS learning using:

$$\hat{\mathbf{x}}^* = \mathbf{D}(\mathbf{z}^*, \mathbf{a}, \mathbf{c})$$

- TAS training with restored video features with standard loss functions:

$$\mathcal{L}_{\text{tas}} = \mathcal{L}_{\text{cls}}(\hat{x}^*, y) + \lambda \cdot \mathcal{L}_{\text{sm}}(\hat{x}^*, y)$$

## Results

### Effectiveness

| | 50Salads [33] | | | | Breakfast [16] | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc | Edit | F1@{10, 25, 50} | Storage | Acc | Edit | F1@{10, 25, 50} | Storage |
| MS-TCN [12] | | | | | | | | |
| Original | 80.6 | 63.1 | 69.9 / 67.4 / 59.0 | 4.5 GB | 67.2 | 60.6 | 50.5 / 46.3 / 36.8 | 28 GB |
| Mean | 69.0 | 42.7 | 50.0 / 46.1 / 37.4 | 7.8 MB | 47.6 | 31.8 | 27.8 / 23.3 / 15.6 | 96 MB |
| Coreset [38] | 61.7 | 43.3 | 49.9 / 46.3 / 35.4 | 7.8 MB | 49.7 | 36.8 | 32.3 / 27.5 / 19.3 | 96 MB |
| TCA [11] | 56.4 | 33.6 | 39.8 / 35.8 / 25.9 | - | 34.2 | 20.7 | 17.9 / 13.8 / 8.4 | - |
| Encoded | 69.0 | 43.6 | 50.6 / 46.0 / 37.4 | 3.9 MB | 37.9 | 49.8 | 40.0 / 32.8 / 19.4 | 44 MB |
| Ours | **74.4** | **59.5** | **65.1 / 61.0 / 50.2** | 3.9 MB | **55.5** | **45.6** | **46.7 / 41.1 / 28.7** | 44 MB |
| Encoded† | 72.1 | 58.2 | 63.2 / 60.0 / 49.3 | 564 MB | 43.4 | 53.2 | 45.8 / 37.4 / 22.8 | 3.4 GB |
| Ours† | **72.8** | **59.8** | **65.2 / 61.3 / 51.3** | 564 MB | **54.1** | **53.3** | **49.8 / 44.3 / 33.1** | 3.4 GB |

† indicates compression along feature channels only, preserving the temporal dimension

- Boring baselines (Mean, Coreset) retain ~70% performances on Breakfast, showing significant temporal redundancy
- Ours reduces storage by 500×, yet yielding 83% performance on Breakfast

### Sequence Sampling

- Sampling strategy

| | Sampling | Acc | Edit | F1@{10, 25, 50} | Storage |
|---|---|---|---|---|---|
| GTEA | ✗ | 76.3 | 74.8 | 80.0 / 78.0 / 61.7 | 7.2 MB |
| | Random | 74.0 | 69.3 | 76.2 / 72.8 / 60.4 | 3.6 MB |
| | Ours | 75.2 | 71.9 | 78.3 / 74.6 / 62.7 | 3.6 MB |

- Diversity-based sampling outperforms random sampling

### Instances Per Segment

| IPS ($K$) | Acc | Edit | F1@{10, 25, 50} | Storage | Ratio(%) |
|---|---|---|---|---|---|
| Mean | 47.6 | 31.8 | 27.8 / 23.3 / 15.6 | 96 MB | 0.34 |
| 1 | 52.1 | 32.2 | 28.1 / 23.7 / 16.0 | 11 MB | 0.04 |
| 2 | 52.7 | 38.4 | 34.9 / 30.1 / 21.1 | 22 MB | 0.08 |
| 4 | 52.4 | 45.9 | 40.7 / 35.8 / 26.0 | 45 MB | 0.15 |
| 8 | **55.6** | **52.3** | 47.3 / **42.1 / 31.4** | 91 MB | 0.31 |
| 16 | 54.2 | 51.2 | **47.4** / 41.8 / 31.0 | 182 MB | 0.62 |
| † | 54.1 | 53.3 | 49.8 / 44.3 / 33.1 | 3.4 GB | 12.0 |

- Larger K improves performance by enabling finer-grained segmentation for more precise inversion
- Performance plateaus after K=8, likely limited by cVAE's expressiveness

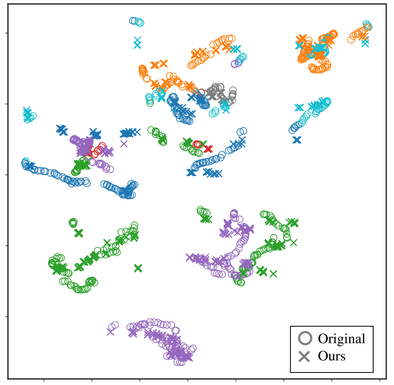
- Sampling capacity

| $\gamma$ | Acc | Edit | F1@{10, 25, 50} | Storage | Ratio(%) |
|---|---|---|---|---|---|
| 0.1 | 45.0 | 44.9 | 40.5 / 35.6 / 23.2 | 1.3 MB | 0.53 |
| 0.2 | 53.1 | 49.1 | 50.4 / 44.6 / 30.6 | 1.9 MB | 0.78 |
| 0.3 | 56.9 | 52.3 | 56.8 / 52.6 / 36.7 | 2.4 MB | 0.98 |
| 0.4 | 73.6 | **72.9** | 77.1 / 74.2 / **63.1** | 2.9 MB | 1.18 |
| 0.5 | **75.2** | 71.9 | **78.3 / 74.6** / 62.7 | 3.6 MB | 1.47 |
| 1 | 76.3 | 74.8 | 80.0 / 78.0 / 61.7 | 7.2 MB | 2.94 |

- Low score with 10% shows that sequence structure is important
- Comparable performance with 50% sequences highlights redundancy

### Visualization

- Original features can be effectively restored




## Takeaways

### Hierarchical Redundancy in TAS

- Feature, temporal and sequence redundancy. TAS DC is promising but underexplored

### Generative Network Inversion

- Generative models can effectively learn action prior
- Network inversion retrieves latent codes that represent the original data

### Sequence Structure

- Informative sequence structure is important for good TAS performance