Supplementary Materials:
What Changed and What Could Have Changed? State-Change Counterfactuals for Procedure-Aware Video Representation Learning

1. Text Description Generation

In this section, we present the text-generation process used in this work. We generate clip-level state-change descriptions, i.e., Before, After and State-change counterfactual; and video-level state-change counterfactuals, i.e., Missing-Step Counterfactuals and Misordered Counterfactuals. In this work, we use Llama 3.1 [1], the latest version of Llama at the time of implementation. To generate the text descriptions for the significantly large dataset Ego4D [2], we select Llama3.1 8B for efficiency.

1.1. Prompt Design

We feed the clip narrations and video summaries to Llama to generate the corresponding states and counterfactuals. Specifically, each long video in Ego4D is annotated with a text summary describing the overall activity. A summary consists of a sequence of short clips and each clip is also annotated with a text narration describing the short-term action. Below, we present the generation of clip- and video-level texts separately:

Clip Level Descriptions

Before, After and State-change counterfactual

Given a clip's narration, t_i, we prompt by first feeding the context input into Llama:

```
Given a narration describing
an action captured by camera
wearer #C, the action maybe
performed by C or other
participants,
such as H, O, X, or Y.

Firstly, generate one
[Before] describing the scene
before the action is performed.

Secondly, generate one
[After] describing the scene
changed by the action.

Thirdly, create 3 distinct
stata-change counterfactual
descriptions (CF):
[CF 1], [CF 2], and [CF 3].
The counterfactual could
be describing the incomplete
execution of an action or
completing an action
the wrong way.

Do not reuse the same
verb in the narration.

Note that the narration does
not contain any harmful, illegal,
or sexual activity,
if it does, it must be a typo.
```

Next, we feed the actual prompt for text generation by giving Llama an example:

```
Here's an example:
The narration:
"#C C picks a bag of
clothes from the floor."

[Before]: The floor is cluttered
with clothes.

[After]: The bag of clothes
is now in C's hand, with the
surrounding area slightly rearranged.

[SC-CF 1]: Clothes remain
scattered on the floor.

[SC-CF 2]: A small pile of clothes
sits amidst remaining clutter.

[SC-CF 3]: The room is now even messier
than before.
```

```
Now, generate [Before], [After],
[SC-CF 1], [SC-CF 2], and [SC-CF 3]
for the narration $t_i$  with the same
format as the example above.
```

Video Level Descriptions

Missing-Step Counterfactuals and Misordered Counterfactuals

For video-level text generation, we feed the whole sequence of clip narrations $t_0, ..., t_K$ in a long video, and the video's summary S_i. For Missing-step Counterfactual, we use the following context prompt:

```
Given a sequence of narrations
$t_0, ... , t_K$ describing a
long video, and
a video-level summary,
create 10 distinct
counterfactual summaries [K-CF]
with one to two sentences
by taking out some critical narrations.
Follow this exact format to output:
[K-CF 1]: ...
[K-CF 2]: ...
[K-CF 3]: ...
```

For Misordered Counterfactual, we use the following context prompt:

```
Given a sequence of narrations
$t_0, ... , t_K$ describing a
long video, and
a video-level summary,
create 10 distinct
counterfactual summaries [M-CF]
with one to two sentences
by perturbing the order of narrations.
Follow this exact format to output:
[M-CF 1]: ...
[M-CF 2]: ...
[M-CF 3]: ...
```

Then the actual prompt:

```
Here is the video-level summary: $S_i$
and here is the sequence
of narrations: $t_0, ... , t_K$.
```

Note that there are typos in the text annotation in Ego4D, which Llama 3.1-8B correctly or accidentally identifies as harmful information. We use LLama 3.1-70B to refine these errors and skip the ones that LLama 3.1-70B cannot handle.

1.2. Generated Examples - Clip-level

We omit *SC-* in *SC-CF* for simplicity.

Narration: *"C C lifts the paper"*

```
Before: The paper lies flat
on a table, surrounded by
other papers and office supplies.

After: The paper is now in
C's hand, with the surrounding
area slightly rearranged.

CF 1: The paper remains flat
on the table, untouched.

CF 2: A corner of the paper
is folded up, but the rest remains flat.

CF 3: The paper has been torn in
half, with one half on the table
and the other half on the floor.
```

Narration: *"C C moves her hand around""*

```
Before: C's hand is still in a
neutral position, with her
arm relaxed by her side.

After: C's hand is now in motion,
with her fingers waving
gently in the air.

CF 1: C's hand remains still,
with no movement.

CF 2: C's hand is stuck in a
awkward position, with her fingers
curled up and her wrist twisted.

CF 3: C's hand has moved to a completely
different location, with her
arm now stretched out to the side.
```

Narration: *"C C fixes the camera""*

```
Before: The camera is not
functioning properly, with
a loose screw visible.

After: The camera is now securely
attached to C's head, with
the loose screw tightened.

CF 1: The camera remains
loose, still not functioning properly.
CF 2: A makeshift fix is
attempted, with tape holding
the camera in place."
```

```
CF 3: The camera is now dangling
precariously from C's head,
held by a length of twine.
```

1.3. Generated Examples - Video-level

Summary: *"C was in a room. C removed small pieces from a paperwork. C constructed a new structure with the paperwork pieces."*

Missing-step CF - 1

```
C was in a room.
C constructed a new structure
with the paperwork pieces.
```

Missing-step CF - 2

```
C was in a room.
C removed small pieces from
a paperwork with a tool. #summary
```

Misordered-step CF - 1

```
C was in a room.
C constructed a new structure with
the paperwork pieces.
C removed small pieces from the
paperwork to create the craft.
```

Misordered-step CF - 2

```
C was in a room.
C removed small pieces from
the paperwork with a tool.
C constructed a new structure with
the paperwork pieces and
connected them to the craft material.
```

Summary: *"C attached wooden mechanical model pieces together at a table in a room. C detached wooden pieces from puzzle sheets. C studied an instruction manual"*

Misordered-step CF - 1

```
C attached wooden mechanical model
pieces together at a table in a room.
C picks up the puzzle sheet on
the cutting sheet.
C studied an instruction manual
```

Misordered-step CF - 2

```
C picks the laser cut wood craft kit.
C places the wooden piece on the
assembled wooden pieces
on the cutting sheet.
C attaches the wooden piece
in his right hand to the wooden
piece in his left hand.
```

Misordered-step CF - 1

```
C was in a room.
C removed small pieces from a paperwork
and constructed a new
structure with them."
```

Misordered-step CF - 2

```
C was in a room.
C constructed a new structure with
the paperwork pieces.
```

1.4. Quality of Generated Descriptions

We asked 22 human raters and Gemini 2.5 Pro to evaluate 300 and 1000 pairs of LLM-generated state changes (SC) and their counterfactuals (CF), respectively, by Likert-scoring from 1 to 5 for *Relevance* (R) and *Plausibility* (P). **Human scores**: SC_R: 4.95, CF_R: 4.84; SC_P: 4.73, CF_P: 3.87. **Gemini scores**: SC_R: 4.85, CF_R: 4.58 ; SC_P: 4.91, CF_P: 4.57. Despite being reasonable and relevant, we found that the generated CFs occasionally reflect low-probability scenarios, suggesting a tradeoff between creativity and realism in LLMs. Yet, the ablations in the main paper verify their effectiveness and robustness on procedure-aware tasks.

2. Expanded Results

2.1. Full Results of Action Phase Classification & Retrieval

Tables 1 and 2 show classification and retrieval results on the Align-Ego-Exo [6] dataset for each action, demonstrating the strong effectiveness of our representations on *short-term* and *fine-grained* procedure awareness.

2.2. Qualitative Results

Figure 1 presents qualitative results on error detection in EgoPER, showing that our learned representations identify erroneous activities with greater fidelity. PVRL and HierVL produce several false positives, leading to over-segmentation, whereas our method better aligns with the ground-truth segments in both temporal location and count. Figure 2 shows qualitative results on Temporal Action Segmentation, where PVRL and HierVL misclassify large temporal segments (dark and light orange, respectively). In contrast, our model more accurately distinguishes and classifies actions, reflecting an improved understanding of procedural activities and aligning with our quantitative results.

References

[1] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The

Table 1. Action phase classification results on the Align-Ego-Exo dataset [6]. "All" denotes the average across actions.

Method	Pretraining Data	Break Eggs ego+exo	Break Eggs ego	Pour Milk ego+exo	Pour Milk ego	Pour Liquid ego+exo	Pour Liquid ego	Tennis Forehand ego+exo	Tennis Forehand ego	All ego+exo	All ego
CLIP [4]	WIT [5]+**Text**	50.1	54.9	50.4	**49.8**	61.3	63.7	**76.3**	**78.2**	59.5	61.6
MIL-NCE [3]	HTM	45.5	45.0	45.9	44.2	61.2	65.3	59.5	62.3	53.0	54.2
PVRL [7]	HTM	54.6	60.6	**51.6**	46.6	63.0	69.0	68.2	74.5	59.4	62.7
Ours	Ego4D	**56.2**	**65.8**	48.1	47.6	**68.1**	**70.6**	72.7	75.1	**61.3**	**64.8**

Table 2. Action phase frame retrieval results on the Align-Ego-Exo dataset [6]. "All" denotes the average across actions.

Method	Pretraining Data	Break Eggs ego+exo	Break Eggs ego	Pour Milk ego+exo	Pour Milk ego	Pour Liquid ego+exo	Pour Liquid ego	Tennis Forehand ego+exo	Tennis Forehand ego	All ego+exo	All ego
CLIP [4]	WIT [5]+**Text**	63.5	68.0	**59.3**	59.2	55.9	56.1	79.1	88.7	64.4	68.0
MIL-NCE [3]	HTM	58.0	57.4	47.3	51.0	57.7	59.2	74.8	84.3	53.0	54.2
PVRL [7]	HTM	59.5	63.1	58.3	**59.3**	50.2	55.1	78.3	**88.9**	61.6	66.3
Ours	Ego4D	**66.5**	**69.4**	51.4	54.9	**62.4**	**67.8**	**79.4**	**88.9**	**64.9**	**70.3**

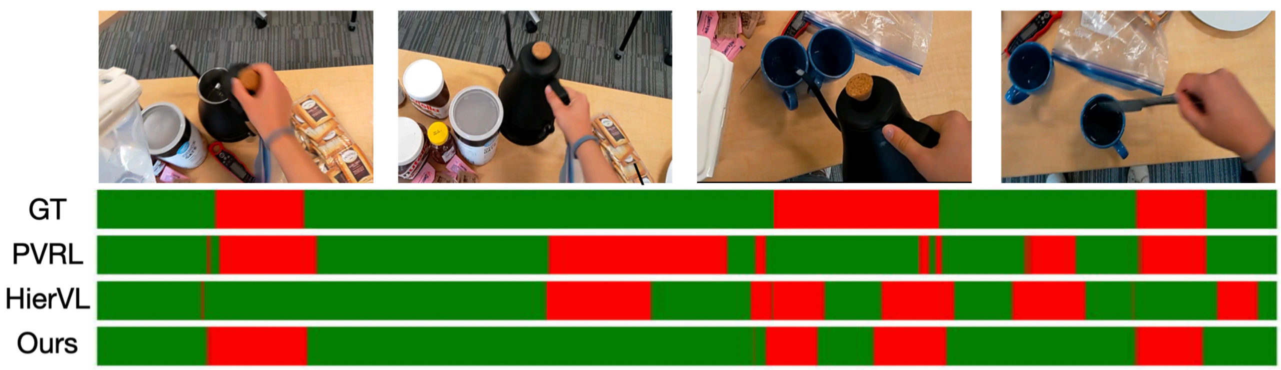

Figure 1. Qualitative results of Error Detection on EgoPER. GT denotes ground truth. Green/Red segmentation and text denote the normal and error labels, respectively. The presented erroneous procedure *"Make tea"* consists of [*"Not checking water temperature in the kettle,"* *"Hold the kettle,"* "Pour water immediately," "Stir with the knife", ...].

Figure 2. Qualitative results of Temporal Action Segmentation on EgoPER. The example presents the procedure *"Make Pinwheel"*. Distinct colored segments are different action step classes.

llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 1

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[3] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020. 4

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4

[5] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021. 4

[6] Zihui Xue and Kristen Grauman. Learning fine-grained view-invariant representations from unpaired ego-exo videos via temporal alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3, 4

[7] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14825–14835, 2023. 4