Objective

This project is a subsection of the ASSET project, which broadly focuses on frailty. Our work specifically targets dementia-related frailty and aims to develop a real-time visual AI system that captures posture and motion features to support early diagnosis. The project includes four key tasks: collecting dementia-related video data; developing a computer vision-based classification framework using gait, posture, and motion; evaluating the effect of de-identification on model performance; and investigating the role of upper-body movement. We collaborated with Dr. Tianlong Chen's team at UNC.

Current progress/Key finding

We collected several datasets, including video datasets (Dem@Care, REDUCE) and SMPL datasets (Unistra, Gait3D). We prioritized Dem@Care, as it is expert-labeled and rich in motion information. Dem@Care includes DS6 and DS8 subsets; we focused on DS8 because of its gait-relevant activities (Figure 1, Figure 2). As a baseline, we trained a 3D ResNet model on raw DS8 videos with varying frame sequence lengths (20–200), achieving up to 84.61% test accuracy (Table 1). To improve performance, we manually annotated time windows of activities in DS8 (Figure 3), filtered out insignificant frames, and separated each video into gait and non-gait clips. However, training on filtered clips did not yield improvements, with all configurations plateauing at 76.92% (Figure 4), raising concerns about whether the model had truly learned meaningful features.

We used saliency maps to visualize model focus and observed that the 3D ResNet often concentrated on background elements like tables rather than participants (Figure 5). We attempted to address this by cropping frames, generating silhouettes, and applying background blurring. None of these methods improved performance; in fact, the blurred version led to a drop in max accuracy to 61.54%. We also tested two transformer-based models, TimeSformer and VideoMAE, on DS6, which yielded similar results to 3D ResNet (Figure 6). Additionally, we experimented with a video-language model (MotionLLM) fine-tuned on SMPL and video data using prompt engineering, but its performance was poor due to many false positives and negatives.

Given these limitations, we applied data augmentation (rotation, horizontal flip, invert, salt and pepper noise) to cropped DS8 frames. This led to a ~10% improvement in accuracy and more stable training (Figure 7), suggesting data augmentation is a promising direction in low-data settings in this project.

Immediate next step

Moving forward, we plan to expand augmentation techniques (e.g., edge, CLAHE, dropout, cutout, solarize) and continue experimentation with 3D ResNet. To address data scarcity, we also plan to incorporate publicly available Parkinson's Disease datasets and extend our framework to other frailty types.

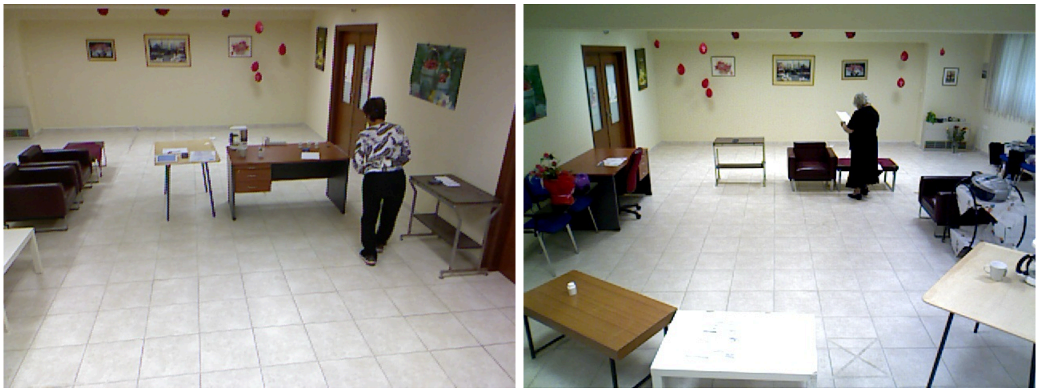
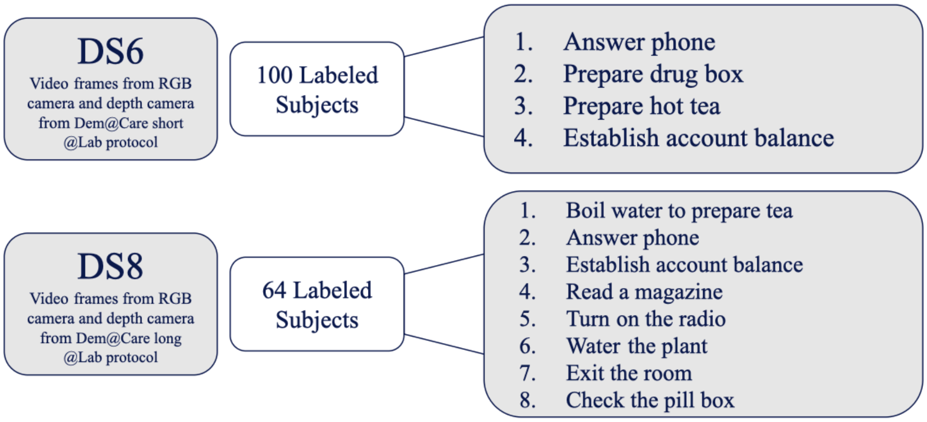


Figure 1. Example frameS from Dem@Care DS6 (left) and DS8 (right).



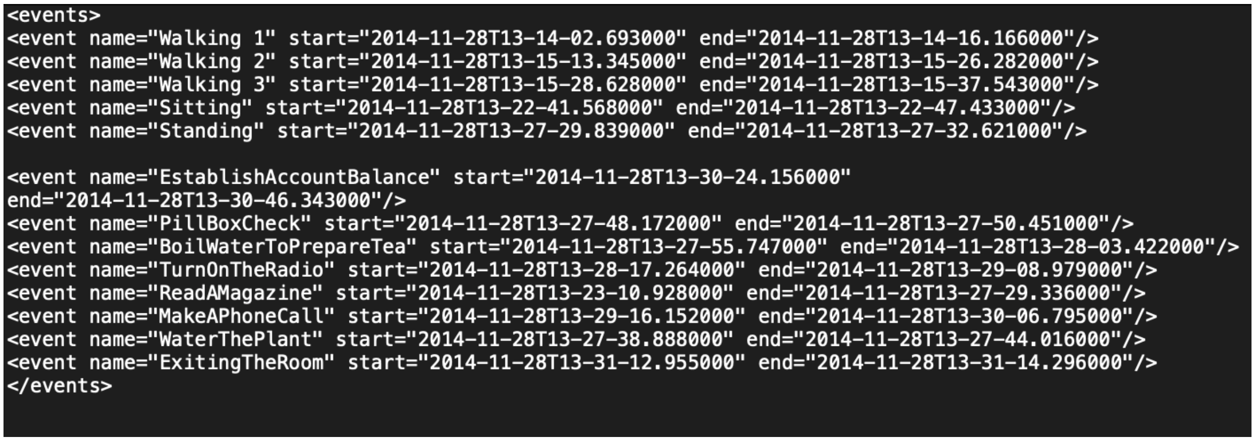
DS6
Video frames from RGB camera and depth camera from Dem@Care short @Lab protocol
100 Labeled Subjects
1. Answer phone
2. Prepare drug box
3. Prepare hot tea
4. Establish account balance
DS8
Video frames from RGB camera and depth camera from Dem@Care long @Lab protocol
64 Labeled Subjects
1. Boil water to prepare tea
2. Answer phone
3. Establish account balance
4. Read a magazine
5. Turn on the radio
6. Water the plant
7. Exit the room
8. Check the pill box

Figure 2. information of Dem@Care DS6 and DS8.

Table 1. Test accuracy of 3D ResNet model using different sample sequence length.

Methods	Test Accuracy
Random Guess	58.35%
Seq Length:20	76.92%
Seq Length:30	84.61%
Seq Length:50	84.61%
Seq Length:75	76.92%
Seq Length:100	84.61%
Seq Length:150	69.23%
Seq Length: 200	76.92%

Seq Length:250	76.92%



```
<events>
<event name="Walking 1" start="2014-11-28T13-14-02.693000" end="2014-11-28T13-14-16.166000"/>
<event name="Walking 2" start="2014-11-28T13-15-13.345000" end="2014-11-28T13-15-26.282000"/>
<event name="Walking 3" start="2014-11-28T13-15-28.628000" end="2014-11-28T13-15-37.543000"/>
<event name="Sitting" start="2014-11-28T13-22-41.568000" end="2014-11-28T13-22-47.433000"/>
<event name="Standing" start="2014-11-28T13-27-29.839000" end="2014-11-28T13-27-32.621000"/>

<event name="EstablishAccountBalance" start="2014-11-28T13-30-24.156000"
end="2014-11-28T13-30-46.343000"/>
<event name="PillBoxCheck" start="2014-11-28T13-27-48.172000" end="2014-11-28T13-27-50.451000"/>
<event name="BoilWaterToPrepareTea" start="2014-11-28T13-27-55.747000" end="2014-11-28T13-28-03.422000"/>
<event name="TurnOnTheRadio" start="2014-11-28T13-28-17.264000" end="2014-11-28T13-29-08.979000"/>
<event name="ReadAMagazine" start="2014-11-28T13-23-10.928000" end="2014-11-28T13-27-29.336000"/>
<event name="MakeAPhoneCall" start="2014-11-28T13-29-16.152000" end="2014-11-28T13-30-06.795000"/>
<event name="WaterThePlant" start="2014-11-28T13-27-38.888000" end="2014-11-28T13-27-44.016000"/>
<event name="ExitingTheRoom" start="2014-11-28T13-31-12.955000" end="2014-11-28T13-31-14.296000"/>
</events>
```

Figure 3. Example annotation of different time windows of activities.

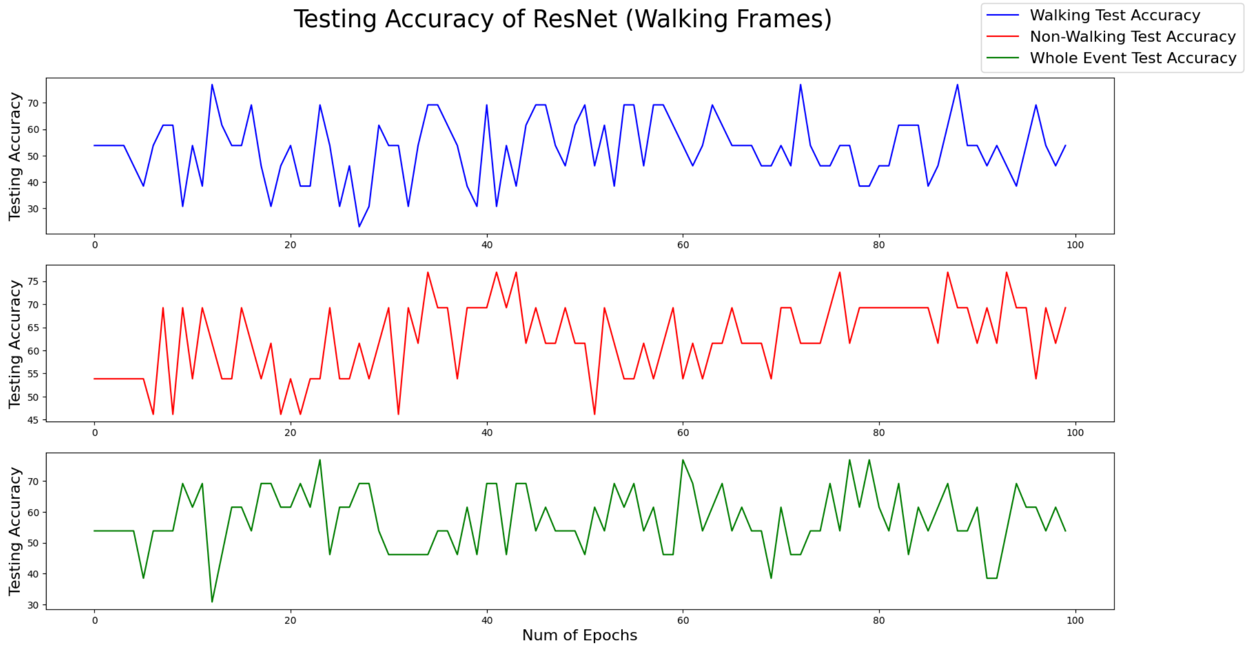


Figure 4. Test accuracy of 3D ResNet trained with gait videos (top), non-gait videos (middle), and full videos (bottom).

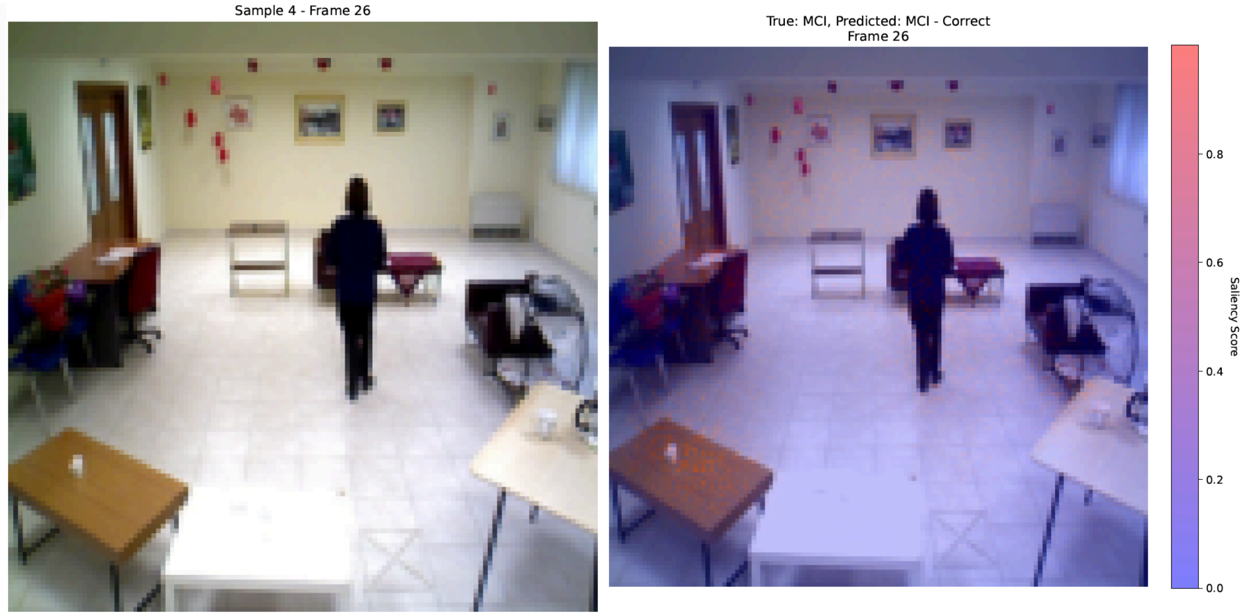


Figure 5. Example visualization of the focus of 3D ResNet.

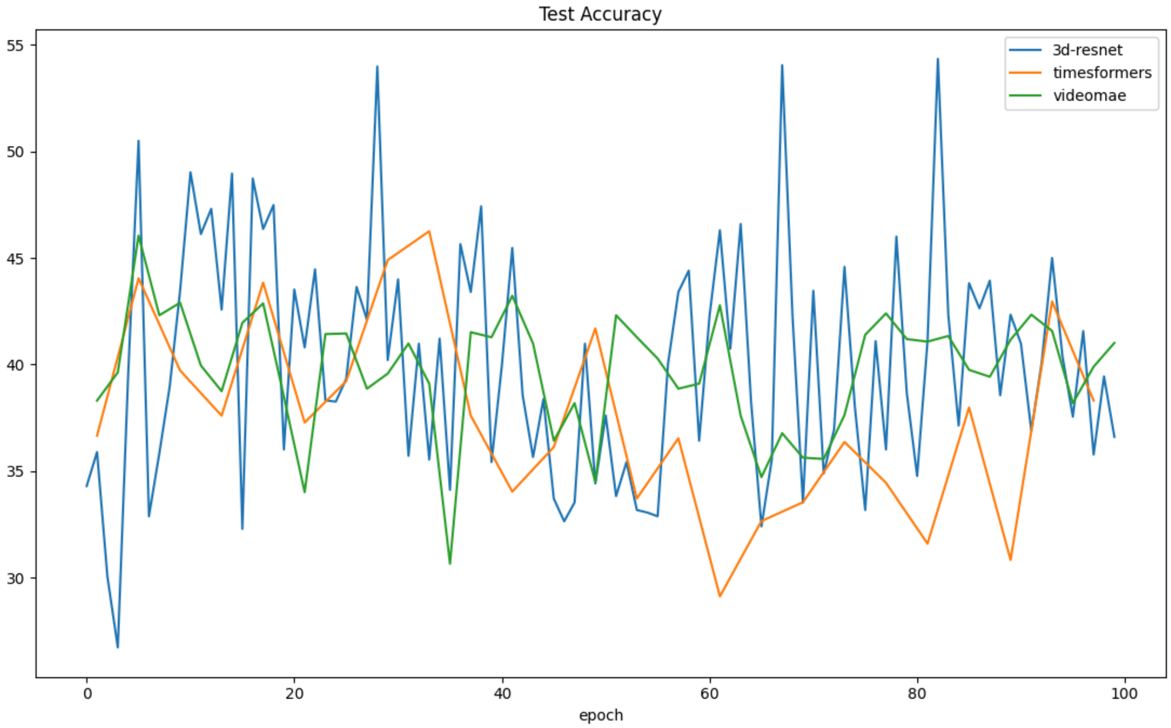


Figure 6. Performance comparison among 3D ResNet, timesformer, and videomae.

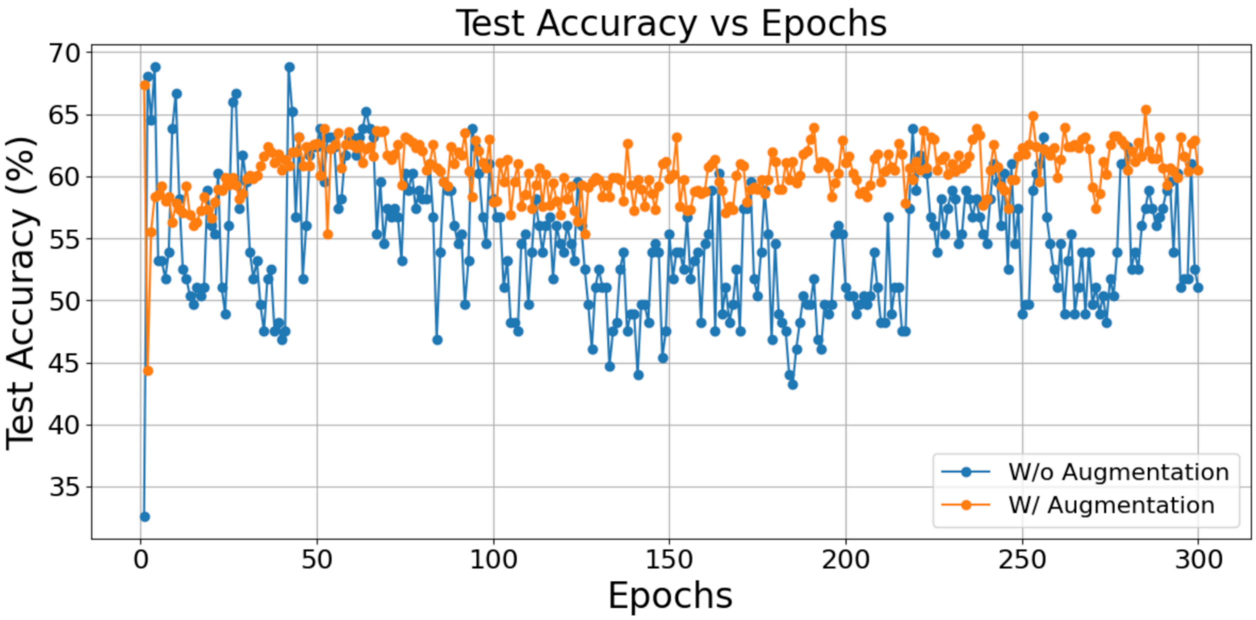


Figure 7. Performance comparison between W/o augmentation and W/ augmentation.