Supplementary Materials

A Model Structure Details

This section provides a detailed description of the structure of PriMo, corresponding to the lower part of Figure 1 in the main text. Traditional video action recognition methods typically rely on one-hot encoded labels for supervision, which has limitations in capturing deep semantic information. To fully exploit the rich semantic information embedded in text for action recognition tasks, we extend the successful contrastive learning strategy used in language and image pretraining to video-level action recognition. The core idea is to use text as a supervisory signal by jointly training a video encoder and a text encoder, so that the video and text representations align with each other, thereby improving the performance of action recognition.

Specifically, our model consists of a video encoder E_v and a text encoder E_t. Given the feature representation $\mathbf{F}_v$ obtained through the Cross-Frame Semantic Aggregator and its corresponding category label y, the video encoder transforms $\mathbf{F}_v$ into a video representation $\mathbf{v}$, i.e., $\mathbf{v} = E_v(\mathbf{F}_v)$; the text encoder converts y into a text representation $\mathbf{t}$, i.e., $\mathbf{t} = E_t(y)$. Then, to make the text representation more aligned with the content of the video, Video-specific Prompting P_v leverages the video representation $\mathbf{v}$ to enhance the text representation $\mathbf{t}$, producing a video-specific text representation $\mathbf{t}_v$. Finally, the cosine similarity between the video representation $\mathbf{v}$ and the enhanced text representation $\mathbf{t}_v$ is calculated. The training objective of the model is to maximize the similarity between correctly matched video-text pairs while minimizing the similarity between mismatched pairs, thereby aligning the video and text representations effectively.

The video encoder consists of three components: Patch Embedding, Inter-frame Interaction Transformer, and Spatio-temporal Fusion Transformer. First, the input is divided into non-overlapping image patches, which are embedded into a high-dimensional space. Each video frame contains a learnable special token used to aggregate information from that frame. To enable information exchange between frames, we introduce the Inter-frame Interaction Transformer, allowing the model to effectively integrate spatiotemporal information in the video. This includes two parts: Temporal Fusion Attention and Self-Propagation Attention. First, Temporal Fusion Attention propagates information across different frames through message tokens, capturing global spatiotemporal dependencies in the video. Then, Self-Propagation Attention refines the frame representation by using message tokens and the frame's embedded patches within each frame.

The Spatio-temporal Fusion Transformer's role is to integrate the representations of all frames to generate a global video representation. First, we add temporal position encoding to the frame representations to retain the order and timing information of the frames. Then, using the Transformer structure, the frame representations are integrated, capturing long-range dependencies between frames. Finally, pooling is applied to the integrated representations to obtain the final video representation $\mathbf{v}$:

$$\mathbf{v} = \text{Pooling}(\text{Transformer}(\mathbf{H} + \mathbf{P}_t))$$

where $\mathbf{H}$ represents the representations of all frames, and $\mathbf{P}_t$ is the temporal position encoding.

The text encoder is responsible for transforming the category label y into a text representation $\mathbf{t}$. However, a fixed text representation is inadequate to reflect the specific content of the video. To address this, we design a video-specific prompt generator to dynamically adapt the text representation to the content of the video. The prompt generator first extracts the global content representation of the video by averaging the representations of all frames, yielding $\mathbf{c}_v$:

$$\mathbf{c}_v = \frac{1}{T}\sum_{i=1}^{T}\mathbf{h}_i$$

where T is the number of frames and $\mathbf{h}_i$ is the representation of the i-th frame. Then, using the Multi-Head Self-Attention (MHSA) mechanism, the text representation $\mathbf{t}$ obtains relevant information from the video content representation $\mathbf{c}_v$ to generate the enhanced prompt representation $\mathbf{p}_t$:

$$\mathbf{p}_t = \text{FFN}(\text{MHSA}(\mathbf{t}, \mathbf{c}_v))$$

Finally, the enhanced prompt representation is fused with the original text representation to produce the final text representation $\mathbf{t}_v$:

$$\mathbf{t}_v = \mathbf{t} + \alpha \cdot \mathbf{p}_t$$

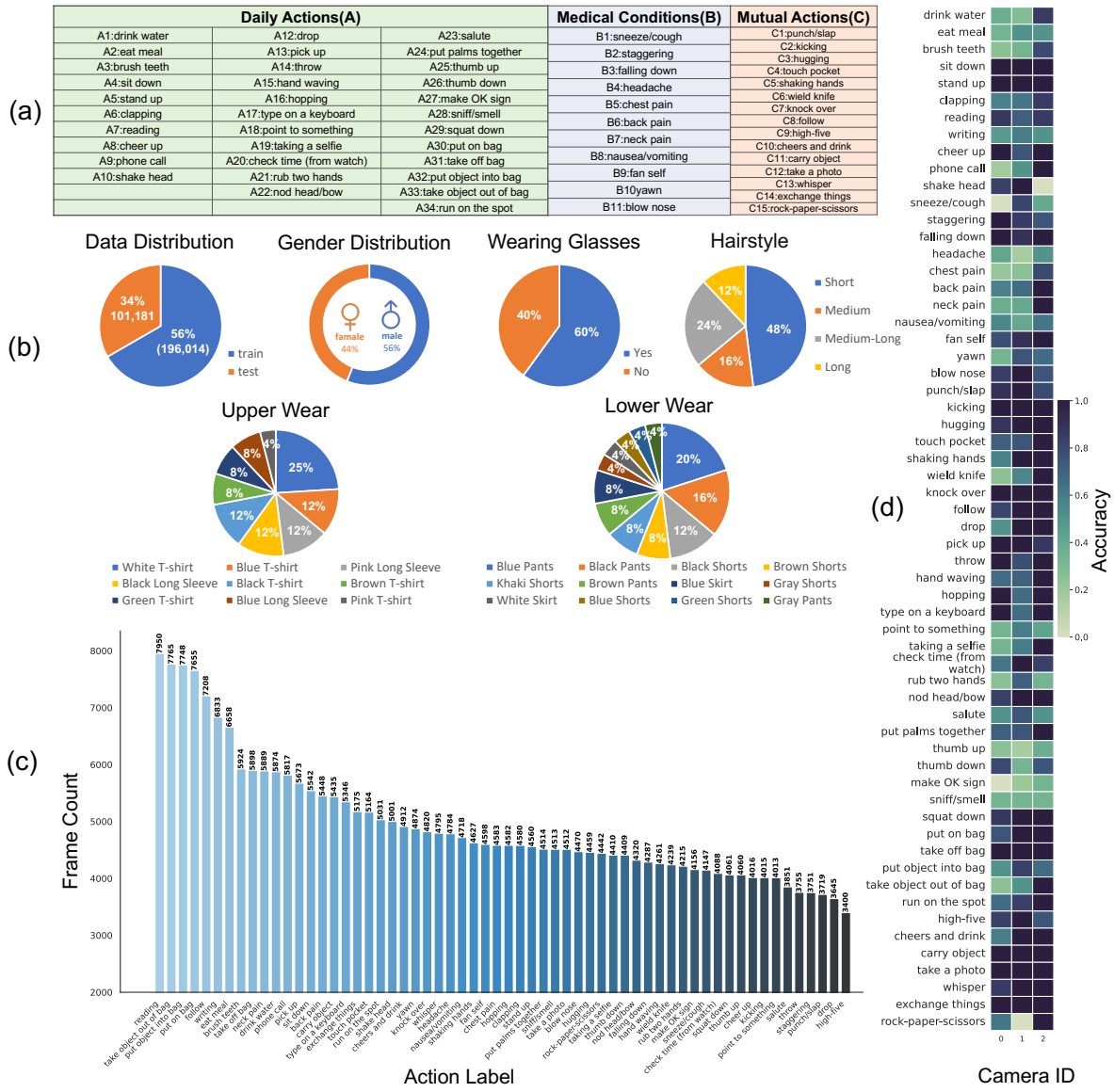

Figure 1: **Details of the Real-Encrypted dataset.** (a) Action Labels: Includes three categories: Daily Actions, Medical Conditions, and Mutual Actions. (b) Privacy Attributes: Data distribution based on attributes such as gender, hairstyle and clothing of the participants, as well as the train/test split ratio. (c) Frame Statistics: The number of video frames corresponding to each Action Label in the Real-Encrypted dataset. (d) Recognition Accuracy: The accuracy of action recognition for each action across three different viewpoints in the Real-Encrypted dataset, evaluated using the PriMo method.

where α is the fusion coefficient, and FFN denotes a feedforward neural network.

During training, we adopt a contrastive learning strategy to align the video and text representations by maximizing the similarity between correctly matched video-text pairs and minimizing the similarity between mismatched pairs. The loss function is expressed as:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(\mathbf{v}, \mathbf{t}_v)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{v}, \mathbf{t}_k)/\tau)}$$

where τ is the temperature parameter that controls the smoothness of the distribution, N is the batch size, $\mathbf{t}_k$ represents the text representation of the k-th sample in the batch, and $\text{sim}(\mathbf{v}, \mathbf{t}_k)$ denotes the cosine similarity between the video representation $\mathbf{v}$ and the text representation $\mathbf{t}_k$.

B Real-Encrypted Dataset

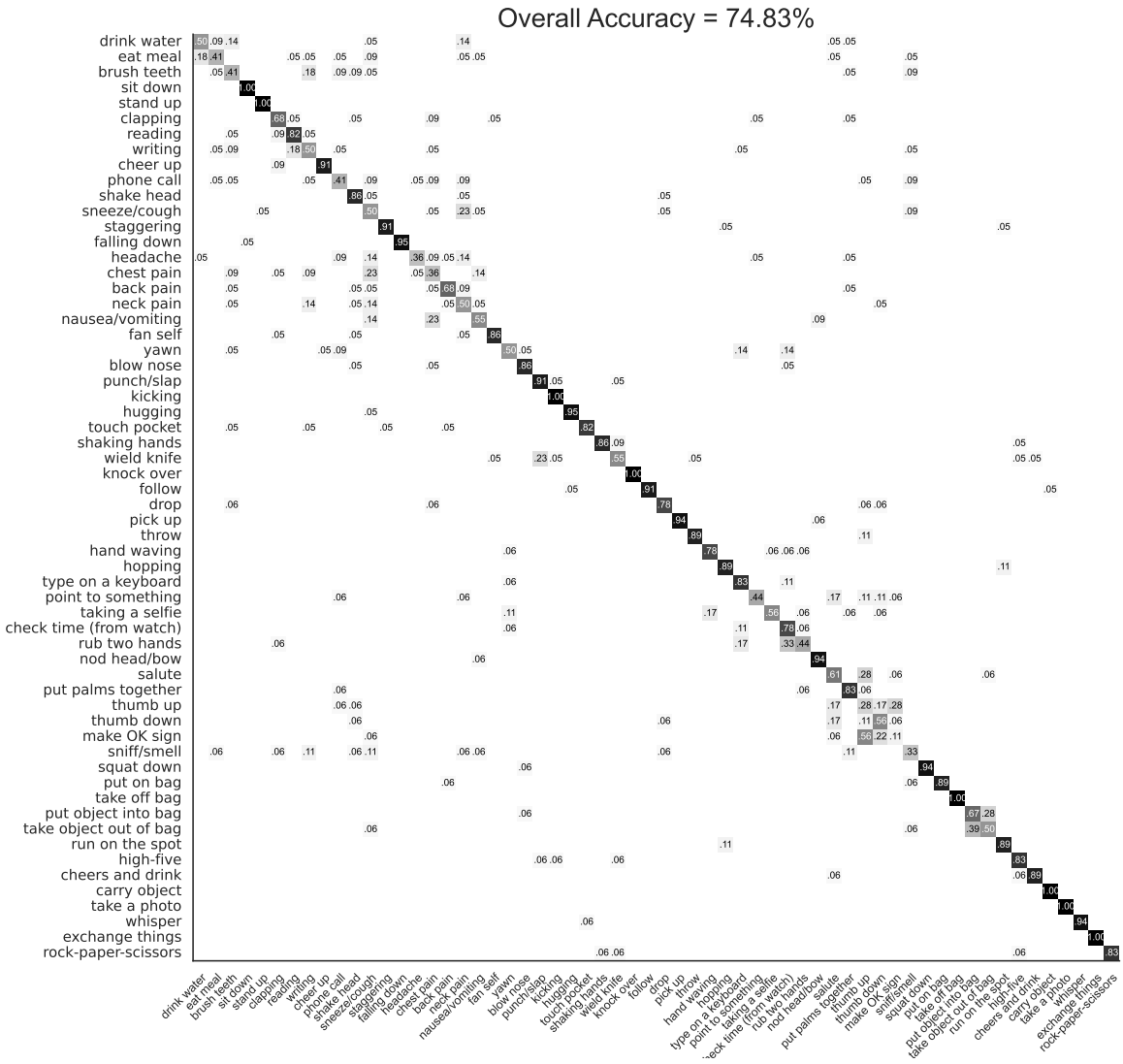

Figure 2: **Confusion matrix on the Real-Encrypted dataset.**

Two methods can be used to generate privacy-protected datasets with Lens Privacy Sealing (LPS). The first method involves capturing existing human action datasets, offering advantages such as large sample size and low acquisition cost, but it cannot accurately replicate real-world lighting conditions. The second method directly applies LPS to existing recording devices to collect human action data in real-world scenarios.

For the first method, we used a 10-layer, 2R, 20C laminating film to physically seal the camera lens and captured the NTU RGB+D 120 dataset to create the privacy-protected dataset, Encrypted-NTU. Like the original NTU RGB+D 120 dataset, Encrypted-NTU contains 114,480 video clips of 120 action categories performed by 106 individuals, with a resolution of 1920x1080.

Since the Encrypted-NTU dataset is solely derived from capturing the NTU dataset and does not represent real-world conditions, it cannot effectively validate the performance of LPS in real-world scenarios. To address this limitation, we introduce Real-Encrypted, a real-world dataset collected using cameras from three different angles, each equipped with 8, 9, and 10 layers of laminating film, respectively, and with a resolution of 1920x1080. Figure 1 provides details of Encrypted-NTU. This dataset includes 60 action categories, encompassing 34 daily activities, 11 health monitoring activities, and 15 public safety activities. The dataset is annotated with six privacy attributes: individual ID, gender, hairstyle, glasses-wearing status, upper-body clothing, and lower-body clothing. Upper-body clothing and lower-body clothing are further categorized into 9 and 12 classes, respectively.

Figure 1(d) illustrates the accuracy of action recognition for each action across three different viewpoints in the Real-Encrypted dataset, evaluated using the PriMo method.The confusion matrix is shown in Figure 2. It can be observed that difficult samples often involve subtle variations, such as "make OK sign." Figure 3 visualizes the Real-Encrypted dataset, presenting results for each sample under different viewpoints and layers of laminating film, annotated with action labels and five privacy attributes.

Figure 3: **Multi-View Visualization of Real-Encrypted dataset.**