

# ROMA's Streaming Dataset

Total: 676,731 samples

27,207

108,986

540,538

## Proactive Data

### Event-Driven Alert

video "Alert me when [a cat] occurs."

video



 Speak Head Label = 1

 Speak Head Label = 0

### Real-Time Narration

video "Describe the video in real time."

video



 Speak Head Label = 1

 Speak Head Label = 0

## Reactive Data

### Backward Tracing

video  ... 

video "What do you see and hear in the video?"

### Forward Prediction

video  ... 

video "What action should I take next in order to put rice into the lunch box?"

### Modality Alignment

video  ... 

video "What is this? Answer the question with a single phrase."