

# SARM

## Key Take-aways

### 1. A Dual Reward Model Architecture

$$\hat{S}_{1:N} = \arg \max_{i \in \{1, \dots, k\}} \Pi_{1:N|i}, \quad \hat{S}_t \in \{1, \dots, k\},$$

1.  $\hat{S}_{1:N}$  → int  
→  $\hat{\alpha}_k$  subtask segment  
 $\hat{S}_{1:N}$   
 $\hat{\alpha}_k$  is annotated before  
 $P_k = \sum_{j=1}^k \bar{\alpha}_j$

MLP Fusion Net

Transformer Sequential Aggregator



Stage Estimator

Classification Head

Subtask Estimator

Regression Head

$\hat{\tau}_{1:N}$

$$\hat{y}_{1:N} = \hat{P}_{k-1,1:N} + \bar{\alpha}_{k,1:N} \hat{\tau}_{1:N}, \quad \hat{y}_{1:N} \in [0, 1].$$

$\hat{c}_{1:N} \in [0, 1]$

Pos Embedding

Linear Projector

Linear Projector

State Normalizer

Image Frames

Joint State

Figure 2: Overview of **SARM**, stage-aware reward modeling. **Left:** SARM overview, which includes both a stage estimator and subtask estimator. First the task stage is predicted from the observations. This prediction is additionally passed into the subtask estimator which predicts a scale value of the progress within the stage. **Right:** An overview of the estimator architecture which is replicated for both the stage estimator and the subtask estimator.

### \* Note,

1.  $N$  a sequence of  $N$  images
2. give  $\bar{\alpha}_k$  from labeling (i.e., annotation)

Labeling by subtask priors: Let trajectory  $i$  have total length  $T_i$  and be segmented into  $K$  subtasks with lengths  $\{L_{i,k}\}_{k=1}^K$ . We estimate a dataset-level prior proportion for each subtask

$$\bar{\alpha}_k = \frac{1}{M} \sum_{i=1}^M \frac{L_{i,k}}{T_i}, \quad \bar{\alpha}_k \geq 0, \quad \sum_{k=1}^K \bar{\alpha}_k = 1,$$

where  $M$  is the number of trajectories.

-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-
-	-	-	-