The Wisdom of Crowds: Temporal Progressive Attention for Early Action Prediction – Supplementary Material

Table S1. **Ablation studies across scales $n = \{1, 2, 3, 4\}$ on UCF-101 over different observation ratios** (ρ). Methods are grouped w.r.t. the backbone used. The best overall performance per ρ is in **bold** and the second best results are underlined.

Method	Backbone	dim	Observation ratios (ρ)								
			0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TemPr − (ours)	X3D$_M$	3D	84.8	91.8	92.3	92.6	93.0	93.4	93.5	93.6	93.6
TemPr = (ours)			85.3	92.3	92.8	93.7	93.9	93.9	94.2	94.4	94.3
TemPr ≡ (ours)			87.4	93.3	93.9	94.4	94.0	94.2	94.4	94.9	94.9
TemPr ≣ (ours)			<u>87.9</u>	<u>93.4</u>	<u>94.5</u>	<u>94.8</u>	95.1	**95.2**	**95.6**	<u>96.4</u>	**96.3**
TemPr − (ours)	MoViNet-A4	3D	85.2	92.1	92.5	92.9	93.3	93.7	93.5	93.8	93.7
TemPr = (ours)			85.6	92.9	93.6	94.5	94.4	94.2	94.2	94.6	94.8
TemPr ≡ (ours)			87.3	93.1	**94.9**	94.6	<u>95.2</u>	<u>94.9</u>	94.6	95.1	95.0
TemPr ≣ (ours)			**88.6**	**93.5**	**94.9**	**94.9**	**95.4**	**95.2**	<u>95.3</u>	**96.6**	<u>96.2</u>

Table S2. **Top tower predictors per class and observation ratio for TemPr ≣ .** Towers $\mathcal{T}_1$ ≣ , $\mathcal{T}_2$ ≣ , $\mathcal{T}_3$ ≣ and , $\mathcal{T}_4$ ≣ are highlighted for better readability.

class name	Observation ratios ρ					
	0.1	0.2	0.3	0.5	0.7	0.9
Putting smthng similar to other things ...	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$
Showing smthng behind smthng	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$
Holding smthng	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Poking ... smthng without ... collapsing	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Pretending to sprinkle air onto smthng	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$
Pulling two ends of smthng ... stretched	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Putting smthng into smthng	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_4$
Pretending to turn smthng upside down	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$
Poking a stack of smthng ... collapses	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_e$
Pulling smthng from left to right	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$
Pushing smthng from left to right	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_4$
Pretending to open smthng without ...	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$
Opening smthng	$\mathcal{T}_4$	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$
Showing a photo of smthng ...	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_1$
Stuffing smthng into smthng	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_2$
Putting smthng on the edge of smthng ...	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_1$
Picking smthng up	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$
Closing smthng	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_2$
Putting smthng upright on the table	$\mathcal{T}_4$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_2$
Turning smthng upside down	$\mathcal{T}_3$	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_1$
Pulling two ends of smthng ... two pieces	$\mathcal{T}_3$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_2$	$\mathcal{T}_2$

S1. Cross-scale accuracy and class predictions

Scale configurations. Supplementary to Table 1 in the main text, we consider the two top-performing backbones in Table S1 and ablate over four scale configurations on UCF-101.

For both models, and across observation ratios, Tempr ≣ outperforms all other scale configurations with the most notable improvements on smaller observation ratios. For $\rho = 0.1$ Tempr ≣ demonstrates a +3.1% improvement from Tempr − on X3D$_M$ and +3.6% on MoViNet-A4.

Top tower predictor per class. To better understand the performance of individual towers $\mathcal{T}_i$, we compare their performance across SSsub21 classes. In Table S2, we present the top-performing tower for each class across observation ratios. Overall, we observe that towers trained on larger scales ($\mathcal{T}_3$ ≣ and $\mathcal{T}_4$ ≣) are better suited for classes that also include long-term dependencies. E.g. classes such as *Poking a stack of something without the stack collapsing*, *Pretending to sprinkle air onto something*, *Showing something behind something*, or *Putting something into something*, require a larger part of the action to be observable to become distinguishable. In contrast, towers for smaller scales, are better suited for classes such as *Picking something up*, *Closing something, or Turning something upside down*, which are distinguishable from only a few frames.

SSsub21 class accuracies. To further determine the performance of tower predictors in Table S2, we show in Figure S1 the per-class accuracies of all towers for $\rho = 0.3$. Overall, because features are more motion-based compared to UCF-101, coarser scales perform better. Considering the *Putting something on the edge of something so it is not supported and falls down* class, the object will typically fall down only at the end of the action. Therefore, such information is better captured by the coarser scales. Similarly, for *Pretending to sprinkle air onto something*, *pretending* can only be captured over a longer temporal scale. Fine scales perform more favorably for shorter actions such as *Closing something*, *Picking something up*, and *Turning something*

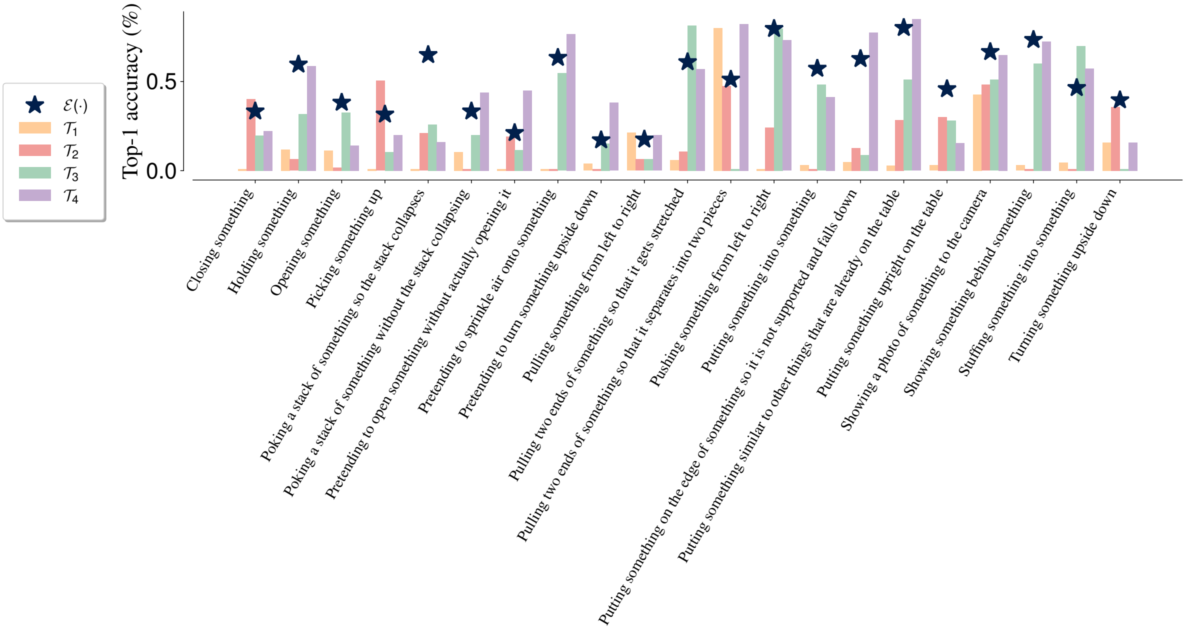
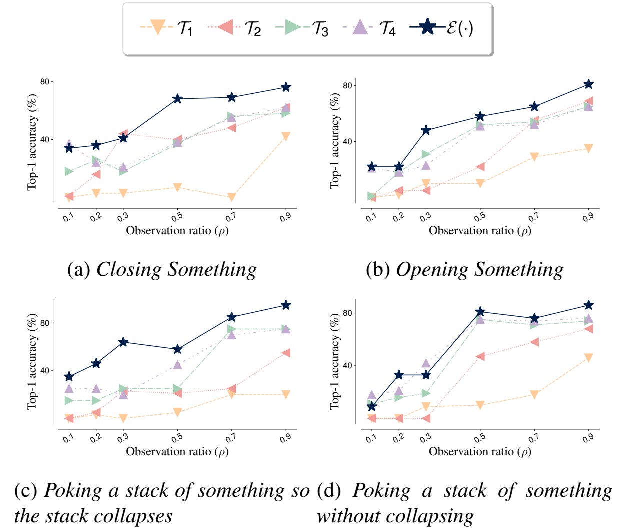


Figure S1. **TemPr SSsub21 class accuracies** over observation ratio $\rho = 0.3$.



(a) *Closing Something*

(b) *Opening Something*

(c) *Poking a stack of something so the stack collapses*

(d) *Poking a stack of something without collapsing*

Figure S2. **TemPr SSsub21 tower accuracies across observation ratios for classes** (a) *Closing Something*, (b) *Opening Something*, (c) *Poking a stack of something so the stack collapses* and (d) *Poking a stack of something without collapsing*.

upside down. For the majority of these classes, informative motions only last a few frames and are thus better addressed by finer scales. Additionally, in Figure S2 we observe that TemPr relies more on coarser scales to capture the differences between visually similar classes. Considering the pairs *Closing something* from Figure S2a and *Opening something* from Figure S2b, as well as *Poking a stack of something so the stack collapses* from Figure S2c and *Poking a stack of something without the stack collapsing* in Figure S2d, there is a stronger reliance to $\mathcal{T}_4$ and $\mathcal{T}_3$, with $\mathcal{T}_2$ only performing better for specific ρ.

Table S3. Tower acc. UCF101.

$\mathcal{T}/\mathcal{E}$	ρ 0.1	0.2	0.3	0.5	0.7	0.9
$\mathcal{T}_4$	78.5	82.3	86.3	84.1	89.3	87.7
$\mathcal{E}(\cdot)$	**84.3**	**90.2**	**90.4**	**91.2**	**92.1**	**92.4**

Table S4. Tower acc. SSsub21.

$\mathcal{T}/\mathcal{E}$	ρ 0.1	0.2	0.3	0.5	0.7	0.9
$\mathcal{T}_4$	26.0	31.6	34.1	36.9	40.6	45.2
$\mathcal{E}(\cdot)$	**28.4**	**34.8**	**37.9**	**41.3**	**45.8**	**48.6**

UCF-101 class accuracies. In Figure S3, we present accuracies for the first 30 classes on UCF-101. Overall, the performance of the aggregation function is equivalent to that of the top-performing tower. For the *BreastStroke* class, the finer scale $\mathcal{T}_1$ outperforms other tower predictors. This is also the case for the *Billiards* class which shows a similar trend with $\mathcal{T}_1$ achieving the best performance. We believe the high accuracy over the fine scales of both *BreastStroke* and *Billiards* classes, is due to their unique appear-

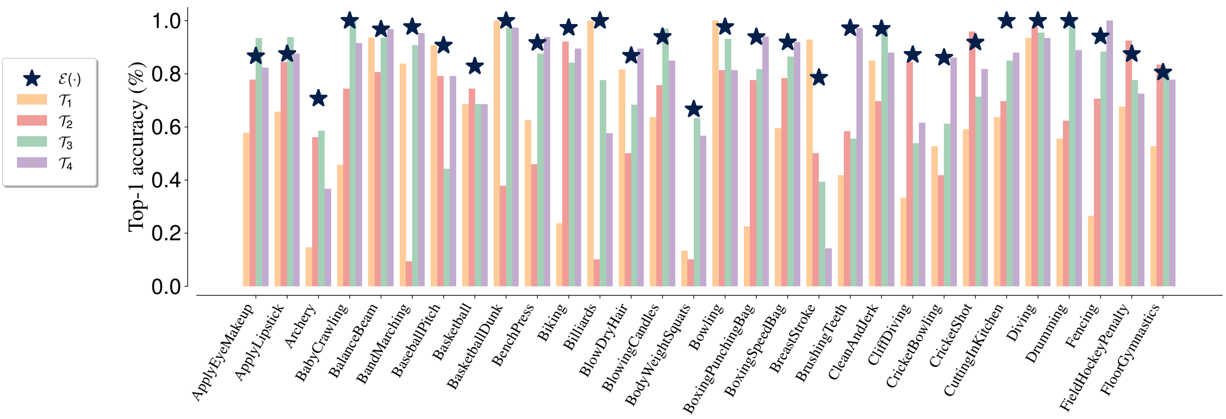


Figure S3. **TemPr UCF-101 class accuracies for the first 30 classes** over observation ratio $\rho = 0.3$.

Table S5. **Tower designs**.

Tower design	ρ 0.2	0.4
MLP ×4	72.4	81.1
MLP ×8	73.1	81.3
(ours)	90.2	90.9

Table S6. **Bottleneck size comparison** based on latent array ($\mathbf{u}$) index dimension (d) used by the cross-attention blocks.

d	Mem. (GB)	Observation ratios (ρ) 0.2	0.4	0.6	0.8
128	1.65	89.1 (-1.1)	89.6 (-1.3)	90.1 (-1.7)	90.7 (-2.3)
256	3.01	90.2	90.9	91.8	92.3
512	5.74	**90.7** (+0.3)	**91.3** (+0.4)	**92.1** (+0.3)	**92.4** (+0.1)

ance and motion features. Thus, for only a small portion of the video, the ongoing action can be correctly predicted.

Tower and aggregation function accuracies. Motivated by class accuracy trends observed in Figure S3 and Figure S1 for UCF-101 and SSsub21, we compare the performance of the final attention tower $\mathcal{T}_4$ to that of the $\mathcal{E}(\cdot)$ aggregator from TemPr . Results for UCF-101 are presented in Table S3 and for SSsub21 in Table S4. Consistent improvements are observed by the predictor ensemble compared to the predictions made from individual towers.

S2. Further ablations

As with the ablation results in Section 4.3 of the main text, we use TemPr with ResNet-18 backbone on UCF-101 for all experiments in this section.

Cross-attention layer replacements. We include tower ablations in Table S5 with ×4/8 MLP layers to assess if the improvements are indeed due to our design. A notable drop is observed with the replacement of the attention towers.

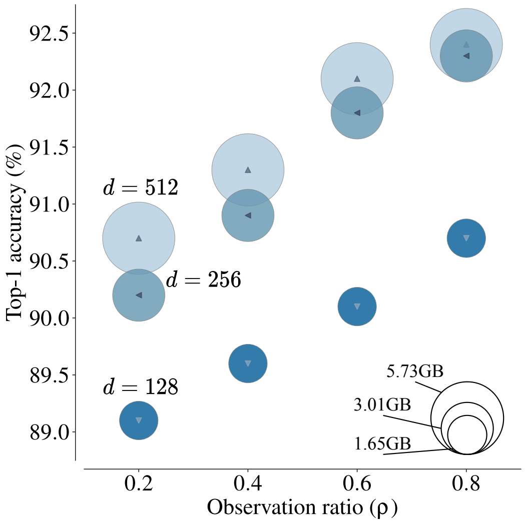


Figure S4. **Bottleneck size** (d) for latent array ($\mathbf{u}$).

Latent array u size: In Figure S4 we present performance results on UCF-101 given different latent array $\mathbf{u}$ sizes d. Size $d = 256$ is shown to be the most cost-effective size as improvements over $d = 128$ range between (1.1-2.3)% while requiring ∼50% less memory than $d = 512$. We additionally detail numerically these individual performances in Table S6. In terms of memory, $d = 128$ requires 1.36GB less than $d = 256$, while $d = 512$ uses 2.73GB more.

Number of self attention blocks. Table S7 demonstrates the impact of the Self MAB number on the accuracy. Increasing the number of self-attention blocks improves accuracy mostly in small observation ratios, while marginally increasing the complexity and memory requirements. We, therefore, adopt $L = 8$ for our model.

Table S7. **Number of self attention blocks** (L)

L	Latency (secs) I (↓)	B (↑)	Pars (M)	FLOPs (G)	Mem. (GB)	ρ 0.2	0.4	0.6	0.8
1	0.31	1.07	20.3	1.29	2.74	70.9	74.8	80.4	86.2
2	0.31	1.09	20.6	1.32	2.78	77.2	76.3	82.8	86.7
4	0.32	1.12	21.5	1.37	2.85	83.4	84.9	85.1	87.4
6	0.32	1.16	22.2	1.42	2.93	88.7	89.5	89.8	90.1
8	0.34	1.27	23.0	1.47	3.01	**90.2**	**90.9**	**91.8**	**92.3**

Table S8. **Ablation on aggregation function.**

(a) SSsub21.

Aggregation	ρ 0.2	0.5
avg	32.3	38.6
softmax	31.4	36.8
ICW	32.4	38.8
adapt. ($\mathcal{E}(\cdot)$)	**34.8**	**41.3**

(b) EK-100.

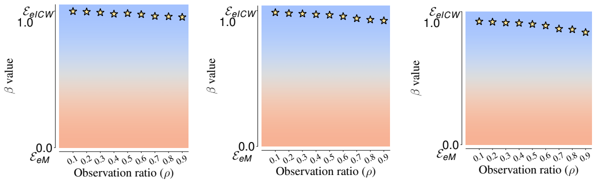
Aggregation	ρ 0.2			0.5		
	V	N	A	V	N	A
avg	21.5	23.9	8.8	51.3	42.2	27.5
softmax	19.4	23.1	8.3	50.7	41.4	24.6
adapt. $\mathcal{E}(\cdot)$	**22.5**	**25.5**	**9.8**	**54.2**	**43.4**	**28.9**

Table S9. **Ablating contributions** with individual and combined replacement.

replacement(s) I. $\mathbf{s}_{1,...,n} \downarrow s_n \times n$	II. $f(\widehat{\mathbf{z}}_i) \downarrow f(\mathbf{z}_i)$	III. $\mathcal{E}(\mathbf{y}_{1,...,n})) \downarrow \overline{f(\widehat{\mathbf{z}})}$	Obs. ratio (ρ) 0.2	0.4	0.6	0.8
Proposed			**90.2**	**90.9**	**91.8**	**92.3**
✗			86.4	88.3	88.8	89.0
	✗		69.4	73.2	78.6	85.5
		✗	89.5	90.1	90.6	91.2
✗	✗		64.3	69.8	75.9	83.4
	✗	✗	67.4	72.8	77.3	84.7
✗		✗	84.2	87.0	87.4	88.3
✗	✗	✗	61.4	67.2	73.5	79.3

SSsub21 and EK-100 aggregation functions. Supplementary to the results in Table 3b for different aggregation functions on UCF-101, we induce additional ablations for SSsub21 and EK-100 in Table S8a and Table S8b respectively. Across both datasets, our proposed adaptive predictor accumulation $\mathcal{E}(\cdot)$performs favorably compared to other aggregation methods. An average improvement of $+5.4\%$ and $+3.8\%$ is observed for UCF-101 and SSsub21.

Combined ablations. Motivated by Table 3 in the main paper, we present combined changes in the model configuration based on our contributions. Setting I. replaces the progressive scales with n copies of the observable video, $\mathbf{s}_{1,...,n} \rightarrow \mathbf{s}_n \times n$. In setting II. the class predictions are made from the extracted CNN features without the utilization of the attention towers $f(\widehat{\mathbf{z}}_i^L) \rightarrow f(\mathbf{z}_i)$. For setting III. the predictor aggregation function is replaced by averaging classifier predictions $\mathcal{E}(f(\widehat{\mathbf{z}}_{1,...,n})) \rightarrow \overline{f(\widehat{\mathbf{z}})}$. On average, a 14.63% accuracy reduction is observed across ratios when predictions are made directly from CNN features. This drop is further amplified when progressive sampling is not used, demonstrating the importance of both the proposed architecture and sub-sampling approach.



(a) TemPr ▬ (b) TemPr ▬ (c) TemPr ▬

Figure S5. **Post-training β values** over obs. ratios on UCF-101.

S3. Predictor aggregation β values

Our proposed adaptive predictor aggregation function relies on a combination of the similarity of predictor probability distributions and their confidences. The trainable parameter of the function defined in Eq. 7 is β which determines the potion of $\underset{eICW}{\mathcal{E}(\cdot)}$ and $\underset{eM}{\mathcal{E}(\cdot)}$ that are used for composing the final aggregated probability distribution.

We visualize the values of the β parameter, for each TemPr configuration that employs multiple scales (▬ , ▬ and ▬) across observation ratios in Figure S5. We use the UCF-101 TemPr models with MoViNet-A4. In general, the β value remains high within 0.98–0.84 for all observation ratios. A small decrease is observed in larger ρ, as independent predictors are exposed to larger portions of the video and can better predict the ongoing action individually.

S4. Additional Qualitative results over tower predictions

We have presented and discussed qualitative results over TemPr ▬ , ▬ , ▬ , ▬ configurations and individual towers $\mathcal{T}_1$ ▬ , $\mathcal{T}_2$ ▬ , $\mathcal{T}_3$ ▬ $\mathcal{T}_4$ ▬ in Section 4.3. Here we provide additional examples in the same format as Figure 4, where predictions differ across TemPr ▬ towers.

As shown in Figure S6, presented over 2 pages, our proposed progressive scales can benefit feature modeling for a variety of action instances e.g. for the *Lunges* instance, the finer scales ($\mathcal{T}_1$ ▬ and $\mathcal{T}_2$ ▬) focus on smaller motions and thus are less influenced by global motion in the video. For *Lunges* and *IceDancing* (form UCF-101), these global motions are similar to those performed for *BodyWeightSquats* and *SalsaSpin*. On the other hand, for the *HighJump* and *SkateBoarding* instances from UCF-101, as well as *hopping* in NTU-RGB and *Pretending to turn something upside down* and *Closing something* in SSsub21, coarse scales are better suited, as motions over larger temporal lengths are more descriptive of the action performed. Failure cases for coarse scales are evident in the chosen examples of *ShavingBeard* from UCF-101, *wipe face* in NTU-RGB, and *turn-off tap* in EPIC-KITCHENS-100, where motions that are descriptive for the class, are performed fast and over shorter temporal durations.

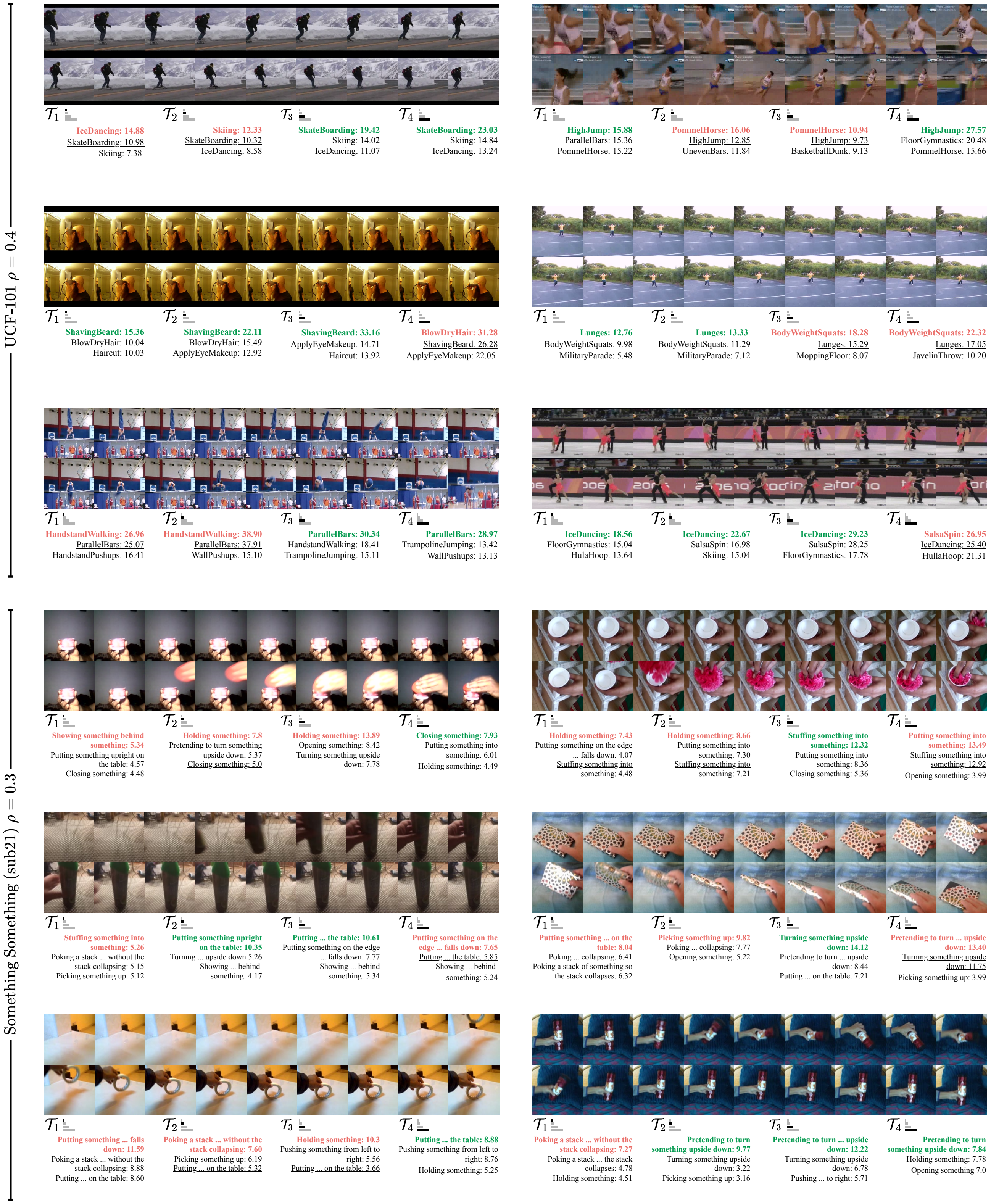


Figure S6. **Instances over UCF-101, SSsub21, NTU-RGB and EK-100**. Top 3 action labels are reported for individual tower predictors $\mathcal{T}_i$ (continues to the next page).

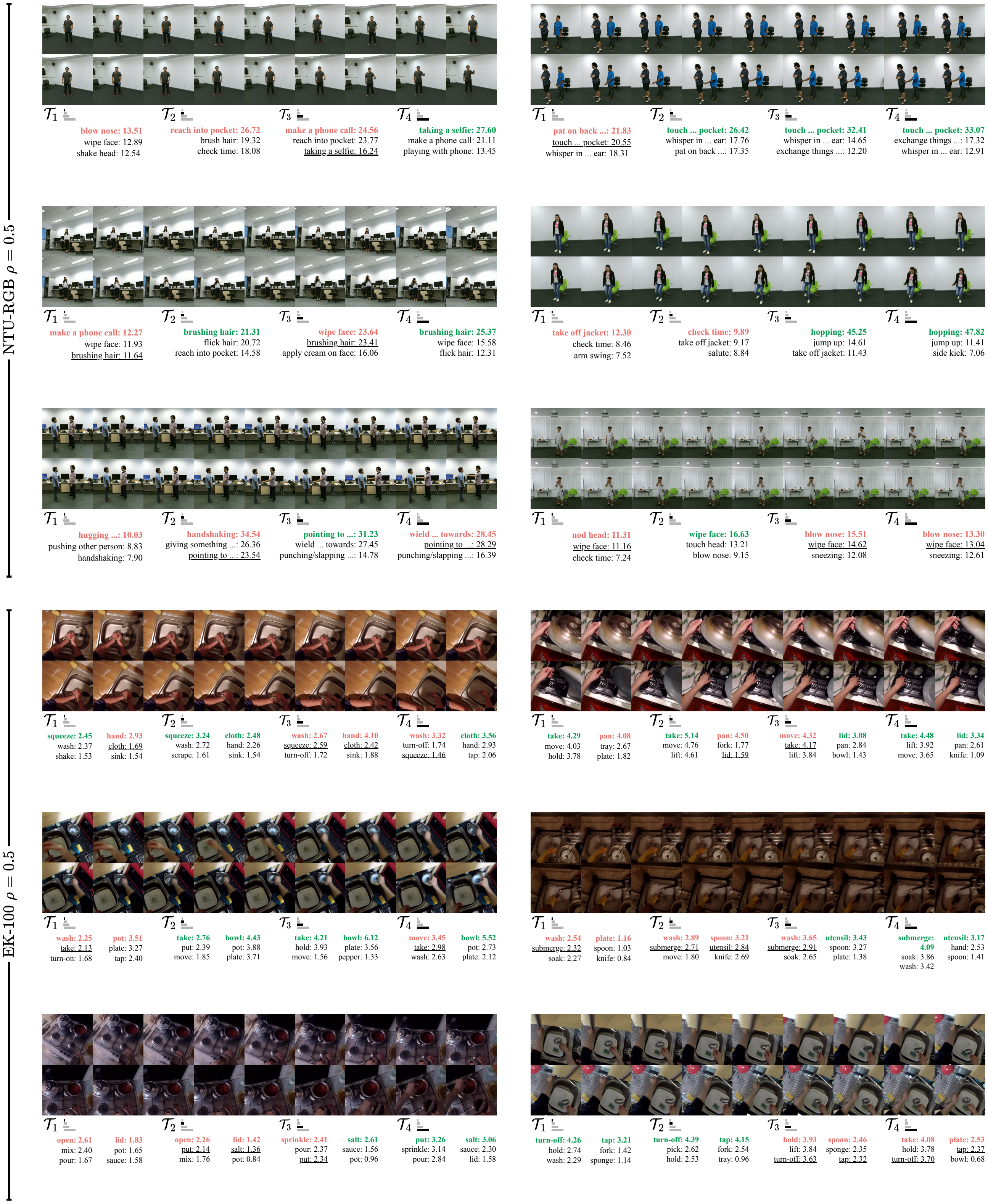


Figure S6. **Instances over UCF-101, SSsub21, NTU-RGB and EK-100**. Top 3 action labels are reported for individual tower predictors ($\mathcal{T}_i$).