MicroG-4M Dataset: Statistical Analysis and Evaluation of Fine-Tuning Results

Lei Qi

The following tables summarize the statistics and evaluation results of fine-tuning the models on the MicroG-4M dataset for the human action recognition task.

1 Dataset Partitioning

The dataset was partitioned at the video level to satisfy two criteria:

- **action coverage**—a greedy selection of video clips was performed to ensure the training set contains at least one example of each action class.
- **proportional splitting**—the remaining video clips were randomly shuffled and allocated to the training, validation, and test sets in a 70:10:20 ratio. Once the splits were finalized, all annotation rows were grouped by video, ensuring that annotations for any given video do not span multiple subsets.

Table 1 presents sample-level (row-wise) counts and percentages.
Table 2 presents video-level counts and percentages.

Table 1: Sample-level statistics of the train/val/test splits (total 13,251 samples).

Split	# Samples	Percentage (%)
Train	9,266	69.93
Val	1,329	10.03
Test	2,656	20.04
Total	13,251	100.00

Table 2: Video-level statistics of the train/val/test splits (total 4,759 video clips).

Split	# video clips	Percentage (%)
Train	3,331	69.99
Val	475	9.98
Test	953	20.03
Total	4,759	100.00

2 Evaluation Results

Table 3: Performance comparison of models fine-tuned on MicroG-4M, evaluated on the validation and test sets.

Model				**Validation**				**Test**			
Arch	TC	Backbone	#Params (M)	mAP (%)	F1-score (%)	Recall (%)	AUROC (%)	mAP (%)	F1-score (%)	Recall (%)	AUROC (%)
C2D	8x8	R50	23.61	27.22	12.52	10.34	82.86	29.51	8.09	6.58	83.49
C2D NLN	8x8	R50	30.97	40.42	23.10	20.41	87.11	44.64	28.30	24.86	89.40
I3D	8x8	R50	27.33	40.93	19.78	16.93	86.44	46.41	26.37	22.25	88.79
I3D NLN	8x8	R50	34.68	41.42	24.11	23.00	86.37	47.12	28.07	24.65	88.52
Slow	8x8	R50	31.74	40.32	21.83	19.08	84.55	45.19	26.13	22.77	88.49
Slow	4x16	R50	31.74	42.97	22.73	19.71	85.46	46.37	28.72	25.38	88.30
SlowFast	8x8	R50	33.76	38.76	20.29	17.66	85.91	43.02	22.63	18.98	88.51
SlowFast	4x16	R50	33.76	37.10	17.74	14.90	84.94	42.10	23.69	20.18	87.54
MViTv1	16x4	B-CONV	36.34	17.79	7.89	6.86	72.40	12.86	5.54	4.66	74.63
MViTv2	16x4	S	34.27	17.57	8.31	6.92	72.67	15.14	8.16	7.17	78.61
X3D	13x6	S	2.02	17.59	6.63	5.63	78.27	14.07	5.77	4.52	78.23
X3D	16x5	L	4.37	23.56	8.82	7.38	80.56	18.70	9.15	7.47	78.27

Note: All models has been pretrained on Kinetics400 dataset and continually trained on MicroG-4M. TC denotes the temporal configuration (frame length × sampling rate). #Params indicates the number of parameters (in millions, M).

Table 4: Zero-shot performance on MicroG-4M test set for models pretrained on Kinetics and fine-tuned on AVA.

Model					**Test Result**			
Arch	TC	Backbone	Pretrain	Fine-tune	mAP (%)	F1-score (%)	Recall (%)	AUROC (%)
Slow	8x8	R50	Kinetics 400	AVA v2.2	16.24	2.67	1.99	73.83
SlowFast	32x2	R101	Kinetics 600	AVA v2.2	23.81	6.32	6.62	77.83

Note: All metrics are macro-averaged over action classes. mAP is measured at IoU = 0.5. F1 and AUROC are computed per class and then averaged. TC denotes the temporal configuration (frame length × sampling rate).

3 Per-Class AP after Fine-Tuning on MicroG-4M & AVA

The tables here show the per-class average precision (AP) results for various models, fine-tuned on the MicroG-4M and AVA datasets, when evaluated on the MicroG-4M test set.

3.1 Quick Table Link

3.1.1 MicroG-4M Model:

- Per-Class AP of MicroG-4M Model: C2D 8x8 R50 5
- Per-Class AP of MicroG-4M Model: C2D NLN 8x8 R50 6
- Per-Class AP of MicroG-4M Model: I3D 8x8 R50 8
- Per-Class AP of MicroG-4M Model: I3D NLN 8x8 R50 16
- Per-Class AP of MicroG-4M Model: SLOWONLY 8x8 R50 10
- Per-Class AP of MicroG-4M Model: SLOWONLY 4x16 R50 7
- Per-Class AP of MicroG-4M Model: SLOWFAST 8x8 R50 14
- Per-Class AP of MicroG-4M Model: SLOWFAST 4x16 R50 11
- Per-Class AP of MicroG-4M Model: MVIT B 16x4 12
- Per-Class AP of MicroG-4M Model: MVITv2 S 16x4 15
- Per-Class AP of MicroG-4M Model: X3D S 13
- Per-Class AP of MicroG-4M Model: X3D L 9

3.1.2 AVA Model:

- Per-Class AP of AVA Model: SLOW 8x8 R50 K400 17
- Per-Class AP of AVA Model: SLOWFAST 32x2 R101 K600 18

3.2 Tables

Table 5: Per-Class AP of MicroG-4M Model: C2D 8x8 R50

ID	AP@50 (%)	Action Name
1	2.23	bend/bow (at the waist)
3	34.03	crouch/kneel
5	50.00	fall down
6	1.10	get up
7	54.68	jump/leap
8	50.00	lie/sleep
9	not detected	martial art
10	27.67	run/jog
11	41.78	sit
12	93.44	stand
14	49.10	walk
17	90.24	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	10.33	close (e.g., a door, a box)
24	0.67	cut
26	49.22	dress/undress clothing
27	3.08	drink
28	59.70	operate spaceship
29	23.51	eat
30	16.60	enter
34	40.57	hit (an object)
36	32.12	lift/pick up
38	1.84	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	28.32	point to (an object)
45	5.71	pull (an object)
46	9.79	push (an object)
47	20.17	put down
48	52.07	read
56	0.24	take a photo
57	0.12	text on/look at a cellphone
58	not detected	throw
59	22.94	touch (an object)
60	1.80	turn (e.g., a screwdriver)
61	32.42	watch (e.g., TV)/any unspecified action
62	31.89	work on a computer
63	not detected	write
64	0.11	fight/hit (a person)
65	22.81	give/serve (an object) to (a person)
66	12.28	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	8.34	hand wave
70	51.39	hug (a person)
72	not detected	kiss (a person)
74	24.62	listen to (a person)
76	not detected	push (another person)
78	0.86	take (an object) from (a person)
79	89.90	talk to (e.g., self, a person, a group)
80	62.35	watch (a person)

Table 6: Per-Class AP of MicroG-4M Model: C2D NLN 8x8 R50

ID	**AP@50 (%)**	**Action Name**
1	12.53	bend/bow (at the waist)
3	40.59	crouch/kneel
5	50.00	fall down
6	5.73	get up
7	38.64	jump/leap
8	100.00	lie/sleep
9	not detected	martial art
10	100.00	run/jog
11	67.66	sit
12	95.65	stand
14	55.82	walk
17	94.70	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	5.74	close (e.g., a door, a box)
24	50.00	cut
26	51.33	dress/undress clothing
27	30.16	drink
28	85.56	operate spaceship
29	41.20	eat
30	31.14	enter
34	43.73	hit (an object)
36	30.06	lift/pick up
38	9.50	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	46.55	point to (an object)
45	4.18	pull (an object)
46	13.38	push (an object)
47	30.91	put down
48	100.00	read
56	0.36	take a photo
57	0.16	text on/look at a cellphone
58	not detected	throw
59	57.67	touch (an object)
60	8.54	turn (e.g., a screwdriver)
61	48.83	watch (e.g., TV)/any unspecified action
62	72.51	work on a computer
63	not detected	write
64	0.16	fight/hit (a person)
65	22.93	give/serve (an object) to (a person)
66	52.80	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	14.93	hand wave
70	100.00	hug (a person)
72	not detected	kiss (a person)
74	52.51	listen to (a person)
76	not detected	push (another person)
78	1.47	take (an object) from (a person)
79	94.58	talk to (e.g., self, a person, a group)
80	68.20	watch (a person)

Table 7: Per-Class AP of MicroG-4M Model: SLOWONLY 4x16 R50

ID	AP@50 (%)	Action Name
1	15.12	bend/bow (at the waist)
3	9.24	crouch/kneel
5	50.00	fall down
6	35.07	get up
7	51.25	jump/leap
8	100.00	lie/sleep
9	not detected	martial art
10	100.00	run/jog
11	72.51	sit
12	95.83	stand
14	55.76	walk
17	94.42	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	17.96	close (e.g., a door, a box)
24	100.00	cut
26	50.21	dress/undress clothing
27	18.62	drink
28	90.29	operate spaceship
29	29.90	eat
30	41.84	enter
34	36.95	hit (an object)
36	35.18	lift/pick up
38	9.47	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	46.82	point to (an object)
45	5.42	pull (an object)
46	43.77	push (an object)
47	25.14	put down
48	87.50	read
56	0.46	take a photo
57	0.10	text on/look at a cellphone
58	not detected	throw
59	54.95	touch (an object)
60	1.70	turn (e.g., a screwdriver)
61	47.80	watch (e.g., TV)/any unspecified action
62	71.99	work on a computer
63	not detected	write
64	0.10	fight/hit (a person)
65	29.33	give/serve (an object) to (a person)
66	44.20	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	17.97	hand wave
70	100.00	hug (a person)
72	not detected	kiss (a person)
74	48.68	listen to (a person)
76	not detected	push (another person)
78	1.10	take (an object) from (a person)
79	94.47	talk to (e.g., self, a person, a group)
80	69.88	watch (a person)

Table 8: Per-Class AP of MicroG-4M Model: I3D 8x8 R50

ID	**AP@50 (%)**	**Action Name**
1	13.77	bend/bow (at the waist)
3	6.08	crouch/kneel
5	100.00	fall down
6	34.31	get up
7	33.93	jump/leap
8	100.00	lie/sleep
9	not detected	martial art
10	100.00	run/jog
11	75.11	sit
12	95.63	stand
14	55.23	walk
17	94.41	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	9.32	close (e.g., a door, a box)
24	20.00	cut
26	55.44	dress/undress clothing
27	44.55	drink
28	94.29	operate spaceship
29	36.31	eat
30	37.51	enter
34	52.01	hit (an object)
36	34.76	lift/pick up
38	10.42	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	48.28	point to (an object)
45	6.08	pull (an object)
46	31.06	push (an object)
47	26.63	put down
48	86.11	read
56	0.63	take a photo
57	0.14	text on/look at a cellphone
58	not detected	throw
59	56.99	touch (an object)
60	6.11	turn (e.g., a screwdriver)
61	46.09	watch (e.g., TV)/any unspecified action
62	71.25	work on a computer
63	not detected	write
64	0.10	fight/hit (a person)
65	26.24	give/serve (an object) to (a person)
66	49.35	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	22.10	hand wave
70	100.00	hug (a person)
72	not detected	kiss (a person)
74	57.34	listen to (a person)
76	not detected	push (another person)
78	1.76	take (an object) from (a person)
79	94.67	talk to (e.g., self, a person, a group)
80	68.77	watch (a person)

Table 9: Per-Class AP of MicroG-4M Model: X3D L

ID	**AP@50 (%)**	**Action Name**
1	5.84	bend/bow (at the waist)
3	1.28	crouch/kneel
5	0.79	fall down
6	0.90	get up
7	1.46	jump/leap
8	0.55	lie/sleep
9	not detected	martial art
10	1.30	run/jog
11	41.39	sit
12	89.19	stand
14	52.47	walk
17	86.27	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	0.66	close (e.g., a door, a box)
24	0.22	cut
26	31.49	dress/undress clothing
27	25.64	drink
28	3.03	operate spaceship
29	5.12	eat
30	24.06	enter
34	1.46	hit (an object)
36	28.18	lift/pick up
38	1.90	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	37.53	point to (an object)
45	0.69	pull (an object)
46	1.40	push (an object)
47	21.52	put down
48	51.27	read
56	0.18	take a photo
57	0.46	text on/look at a cellphone
58	not detected	throw
59	33.42	touch (an object)
60	0.97	turn (e.g., a screwdriver)
61	30.75	watch (e.g., TV)/any unspecified action
62	13.08	work on a computer
63	not detected	write
64	0.12	fight/hit (a person)
65	15.00	give/serve (an object) to (a person)
66	2.33	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	3.20	hand wave
70	0.82	hug (a person)
72	not detected	kiss (a person)
74	11.36	listen to (a person)
76	not detected	push (another person)
78	0.85	take (an object) from (a person)
79	88.59	talk to (e.g., self, a person, a group)
80	50.11	watch (a person)

Table 10: Per-Class AP of MicroG-4M Model: SLOWONLY 8x8 R50

ID	AP@50 (%)	Action Name
1	13.78	bend/bow (at the waist)
3	14.04	crouch/kneel
5	33.33	fall down
6	14.66	get up
7	43.14	jump/leap
8	100.00	lie/sleep
9	not detected	martial art
10	100.00	run/jog
11	71.38	sit
12	96.50	stand
14	58.44	walk
17	94.52	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	4.31	close (e.g., a door, a box)
24	100.00	cut
26	56.57	dress/undress clothing
27	36.43	drink
28	65.46	operate spaceship
29	25.09	eat
30	31.95	enter
34	58.30	hit (an object)
36	27.50	lift/pick up
38	23.22	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	46.74	point to (an object)
45	13.78	pull (an object)
46	26.62	push (an object)
47	27.85	put down
48	87.50	read
56	0.69	take a photo
57	0.10	text on/look at a cellphone
58	not detected	throw
59	52.12	touch (an object)
60	4.17	turn (e.g., a screwdriver)
61	41.49	watch (e.g., TV)/any unspecified action
62	70.70	work on a computer
63	not detected	write
64	0.14	fight/hit (a person)
65	32.14	give/serve (an object) to (a person)
66	42.09	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	24.32	hand wave
70	100.00	hug (a person)
72	not detected	kiss (a person)
74	48.34	listen to (a person)
76	not detected	push (another person)
78	3.95	take (an object) from (a person)
79	94.72	talk to (e.g., self, a person, a group)
80	66.71	watch (a person)

Table 11: Per-Class AP of MicroG-4M Model: SLOWFAST 4x16 R50

ID	AP@50 (%)	Action Name
1	7.20	bend/bow (at the waist)
3	14.66	crouch/kneel
5	100.00	fall down
6	4.29	get up
7	59.20	jump/leap
8	100.00	lie/sleep
9	not detected	martial art
10	96.67	run/jog
11	71.34	sit
12	94.73	stand
14	54.77	walk
17	92.44	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	5.19	close (e.g., a door, a box)
24	7.14	cut
26	21.50	dress/undress clothing
27	27.10	drink
28	87.14	operate spaceship
29	31.88	eat
30	51.33	enter
34	45.40	hit (an object)
36	29.19	lift/pick up
38	1.88	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	33.13	point to (an object)
45	25.60	pull (an object)
46	23.91	push (an object)
47	15.95	put down
48	87.50	read
56	0.19	take a photo
57	0.10	text on/look at a cellphone
58	not detected	throw
59	48.29	touch (an object)
60	8.25	turn (e.g., a screwdriver)
61	34.80	watch (e.g., TV)/any unspecified action
62	59.57	work on a computer
63	not detected	write
64	0.16	fight/hit (a person)
65	28.77	give/serve (an object) to (a person)
66	46.11	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	6.26	hand wave
70	100.00	hug (a person)
72	not detected	kiss (a person)
74	44.11	listen to (a person)
76	not detected	push (another person)
78	1.27	take (an object) from (a person)
79	92.61	talk to (e.g., self, a person, a group)
80	66.05	watch (a person)

Table 12: Per-Class AP of MicroG-4M Model: MVIT B 16x4

ID	**AP@50 (%)**	**Action Name**
1	5.67	bend/bow (at the waist)
3	1.67	crouch/kneel
5	0.87	fall down
6	0.85	get up
7	2.72	jump/leap
8	0.10	lie/sleep
9	not detected	martial art
10	0.85	run/jog
11	42.79	sit
12	88.58	stand
14	23.21	walk
17	82.37	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	0.83	close (e.g., a door, a box)
24	0.19	cut
26	2.34	dress/undress clothing
27	2.94	drink
28	2.56	operate spaceship
29	2.98	eat
30	4.61	enter
34	3.15	hit (an object)
36	8.56	lift/pick up
38	0.91	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	17.09	point to (an object)
45	0.74	pull (an object)
46	0.98	push (an object)
47	5.22	put down
48	2.38	read
56	0.11	take a photo
57	0.55	text on/look at a cellphone
58	not detected	throw
59	13.17	touch (an object)
60	11.45	turn (e.g., a screwdriver)
61	20.95	watch (e.g., TV)/any unspecified action
62	12.65	work on a computer
63	not detected	write
64	0.14	fight/hit (a person)
65	4.46	give/serve (an object) to (a person)
66	2.34	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	3.47	hand wave
70	1.54	hug (a person)
72	not detected	kiss (a person)
74	13.07	listen to (a person)
76	not detected	push (another person)
78	1.90	take (an object) from (a person)
79	87.52	talk to (e.g., self, a person, a group)
80	48.65	watch (a person)

Table 13: Per-Class AP of MicroG-4M Model: X3D S

ID	AP@50 (%)	Action Name
1	1.32	bend/bow (at the waist)
3	1.69	crouch/kneel
5	0.76	fall down
6	0.72	get up
7	1.28	jump/leap
8	0.41	lie/sleep
9	not detected	martial art
10	2.05	run/jog
11	49.45	sit
12	89.49	stand
14	48.25	walk
17	84.34	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	0.46	close (e.g., a door, a box)
24	0.48	cut
26	2.49	dress/undress clothing
27	2.70	drink
28	2.89	operate spaceship
29	2.66	eat
30	12.82	enter
34	1.48	hit (an object)
36	13.43	lift/pick up
38	1.15	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	22.38	point to (an object)
45	0.90	pull (an object)
46	1.32	push (an object)
47	5.21	put down
48	2.94	read
56	0.18	take a photo
57	0.64	text on/look at a cellphone
58	not detected	throw
59	18.09	touch (an object)
60	4.14	turn (e.g., a screwdriver)
61	30.55	watch (e.g., TV)/any unspecified action
62	12.45	work on a computer
63	not detected	write
64	0.12	fight/hit (a person)
65	3.26	give/serve (an object) to (a person)
66	2.29	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	3.98	hand wave
70	1.44	hug (a person)
72	not detected	kiss (a person)
74	8.76	listen to (a person)
76	not detected	push (another person)
78	0.92	take (an object) from (a person)
79	89.15	talk to (e.g., self, a person, a group)
80	47.75	watch (a person)

Table 14: Per-Class AP of MicroG-4M Model: SLOWFAST 8x8 R50

ID	AP@50 (%)	Action Name
1	9.49	bend/bow (at the waist)
3	13.40	crouch/kneel
5	33.33	fall down
6	4.71	get up
7	42.14	jump/leap
8	100.00	lie/sleep
9	not detected	martial art
10	100.00	run/jog
11	69.19	sit
12	95.12	stand
14	59.34	walk
17	93.20	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	50.90	close (e.g., a door, a box)
24	12.50	cut
26	60.18	dress/undress clothing
27	19.56	drink
28	88.50	operate spaceship
29	31.44	eat
30	33.16	enter
34	36.63	hit (an object)
36	31.17	lift/pick up
38	9.72	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	43.55	point to (an object)
45	2.27	pull (an object)
46	41.22	push (an object)
47	22.75	put down
48	80.56	read
56	0.58	take a photo
57	0.22	text on/look at a cellphone
58	not detected	throw
59	54.52	touch (an object)
60	5.09	turn (e.g., a screwdriver)
61	43.00	watch (e.g., TV)/any unspecified action
62	72.29	work on a computer
63	not detected	write
64	0.11	fight/hit (a person)
65	28.65	give/serve (an object) to (a person)
66	43.13	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	18.60	hand wave
70	100.00	hug (a person)
72	not detected	kiss (a person)
74	48.19	listen to (a person)
76	not detected	push (another person)
78	6.62	take (an object) from (a person)
79	93.98	talk to (e.g., self, a person, a group)
80	64.77	watch (a person)

Table 15: Per-Class AP of MicroG-4M Model: MVITv2 S 16x4

ID	AP@50 (%)	Action Name
1	2.68	bend/bow (at the waist)
3	1.16	crouch/kneel
5	0.51	fall down
6	1.32	get up
7	1.83	jump/leap
8	0.65	lie/sleep
9	not detected	martial art
10	3.57	run/jog
11	59.83	sit
12	93.34	stand
14	59.48	walk
17	89.20	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	0.90	close (e.g., a door, a box)
24	0.15	cut
26	2.75	dress/undress clothing
27	1.95	drink
28	2.57	operate spaceship
29	2.37	eat
30	13.26	enter
34	1.75	hit (an object)
36	10.77	lift/pick up
38	1.39	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	19.54	point to (an object)
45	0.85	pull (an object)
46	2.06	push (an object)
47	5.70	put down
48	2.35	read
56	0.17	take a photo
57	0.36	text on/look at a cellphone
58	not detected	throw
59	23.41	touch (an object)
60	0.85	turn (e.g., a screwdriver)
61	25.04	watch (e.g., TV)/any unspecified action
62	14.24	work on a computer
63	not detected	write
64	0.12	fight/hit (a person)
65	3.56	give/serve (an object) to (a person)
66	2.01	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	3.54	hand wave
70	1.61	hug (a person)
72	not detected	kiss (a person)
74	10.97	listen to (a person)
76	not detected	push (another person)
78	0.99	take (an object) from (a person)
79	92.49	talk to (e.g., self, a person, a group)
80	59.60	watch (a person)

Table 16: Per-Class AP of MicroG-4M Model: I3D NLN 8x8 R50

ID	AP@50 (%)	Action Name
1	23.33	bend/bow (at the waist)
3	18.58	crouch/kneel
5	100.00	fall down
6	13.34	get up
7	19.45	jump/leap
8	100.00	lie/sleep
9	not detected	martial art
10	100.00	run/jog
11	76.00	sit
12	95.28	stand
14	62.78	walk
17	94.13	carry/hold (an object)
20	not detected	climb (e.g., a mountain)
22	2.87	close (e.g., a door, a box)
24	50.00	cut
26	64.48	dress/undress clothing
27	12.82	drink
28	94.29	operate spaceship
29	36.82	eat
30	41.77	enter
34	40.77	hit (an object)
36	29.54	lift/pick up
38	24.28	open (e.g., a window, a car door)
41	not detected	play musical instrument
43	40.30	point to (an object)
45	27.61	pull (an object)
46	34.89	push (an object)
47	25.34	put down
48	90.00	read
56	0.36	take a photo
57	0.09	text on/look at a cellphone
58	not detected	throw
59	50.16	touch (an object)
60	13.57	turn (e.g., a screwdriver)
61	46.89	watch (e.g., TV)/any unspecified action
62	69.93	work on a computer
63	not detected	write
64	0.15	fight/hit (a person)
65	28.20	give/serve (an object) to (a person)
66	53.25	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	33.38	hand wave
70	100.00	hug (a person)
72	not detected	kiss (a person)
74	52.89	listen to (a person)
76	not detected	push (another person)
78	1.62	take (an object) from (a person)
79	94.82	talk to (e.g., self, a person, a group)
80	68.00	watch (a person)

Table 17: Per-Class AP of AVA Model: SLOW 8x8 R50 K400

ID	AP@50 (%)	Action Name
1	2.93	bend/bow (at the waist)
2	not detected	crawl
3	6.12	crouch/kneel
4	not detected	dance
5	50.00	fall down
6	34.05	get up
7	69.16	jump/leap
8	0.65	lie/sleep
9	not detected	martial art
10	9.76	run/jog
11	8.47	sit
12	67.97	stand
13	not detected	swim
14	24.66	walk
15	not detected	answer phone
16	not detected	brush teeth
17	81.26	carry/hold (an object)
18	not detected	catch (an object)
19	not detected	chop
20	not detected	climb (e.g., a mountain)
21	not detected	clink glass
22	2.46	close (e.g., a door, a box)
23	not detected	cook
24	0.99	cut
25	not detected	dig
26	5.76	dress/put on clothing
27	36.90	drink
28	16.36	drive (e.g., a car, a truck)
29	7.12	eat
30	3.85	enter
31	not detected	exit
32	not detected	extract
33	not detected	fishing
34	0.59	hit (an object)
35	not detected	kick (an object)
36	6.29	lift/pick up
37	not detected	listen (e.g., to music)
38	2.67	open (e.g., a window, a car door)
39	not detected	paint
40	not detected	play board game
41	not detected	play musical instrument
42	not detected	play with pets
43	10.77	point to (an object)
44	not detected	press
45	1.14	pull (an object)
46	11.02	push (an object)
47	4.95	put down
48	1.09	read
49	not detected	ride (e.g., a bike, a car, a horse)
50	not detected	row boat
51	not detected	sail boat
52	not detected	shoot
53	not detected	shovel
54	not detected	smoke
55	not detected	stir

ID	AP@50 (%)	Action Name
56	0.15	take a photo
57	0.10	text on/look at a cellphone
58	not detected	throw
59	16.83	touch (an object)
60	2.22	turn (e.g., a screwdriver)
61	5.10	watch (e.g., TV)
62	12.81	work on a computer
63	not detected	write
64	0.31	fight/hit (a person)
65	3.01	give/serve (an object) to (a person)
66	3.23	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	4.42	hand wave
70	4.40	hug (a person)
71	not detected	kick (a person)
72	not detected	kiss (a person)
73	not detected	lift (a person)
74	15.32	listen to (a person)
75	not detected	play with kids
76	not detected	push (another person)
77	not detected	sing to (e.g., self, a person, a group)
78	0.41	take (an object) from (a person)
79	82.53	talk to (e.g., self, a person, a group)
80	48.18	watch (a person)

Table 18: Per-Class AP of AVA Model: SLOWFAST 32x2 R101 K600

ID	AP@50 (%)	Action Name
1	8.01	bend/bow (at the waist)
2	not detected	crawl
3	1.60	crouch/kneel
4	not detected	dance
5	100.00	fall down
6	12.91	get up
7	44.37	jump/leap
8	0.25	lie/sleep
9	not detected	martial art
10	67.56	run/jog
11	14.09	sit
12	74.96	stand
13	not detected	swim
14	25.34	walk
15	not detected	answer phone
16	not detected	brush teeth
17	87.85	carry/hold (an object)
18	not detected	catch (an object)
19	not detected	chop
20	not detected	climb (e.g., a mountain)
21	not detected	clink glass
22	13.35	close (e.g., a door, a box)
23	not detected	cook
24	100.00	cut

ID	AP@50 (%)	Action Name
25	not detected	dig
26	13.47	dress/put on clothing
27	15.17	drink
28	2.26	drive (e.g., a car, a truck)
29	21.29	eat
30	5.37	enter
31	not detected	exit
32	not detected	extract
33	not detected	fishing
34	0.69	hit (an object)
35	not detected	kick (an object)
36	14.22	lift/pick up
37	not detected	listen (e.g., to music)
38	1.91	open (e.g., a window, a car door)
39	not detected	paint
40	not detected	play board game
41	not detected	play musical instrument
42	not detected	play with pets
43	15.23	point to (an object)
44	not detected	press
45	1.23	pull (an object)
46	5.94	push (an object)
47	5.78	put down
48	3.17	read
49	not detected	ride (e.g., a bike, a car, a horse)
50	not detected	row boat
51	not detected	sail boat
52	not detected	shoot
53	not detected	shovel
54	not detected	smoke
55	not detected	stir
56	0.29	take a photo
57	0.17	text on/look at a cellphone
58	not detected	throw
59	23.75	touch (an object)
60	0.86	turn (e.g., a screwdriver)
61	8.49	watch (e.g., TV)
62	44.95	work on a computer
63	not detected	write
64	0.35	fight/hit (a person)
65	27.04	give/serve (an object) to (a person)
66	9.26	grab (a person)
67	not detected	hand clap
68	not detected	hand shake
69	11.24	hand wave
70	25.22	hug (a person)
71	not detected	kick (a person)
72	not detected	kiss (a person)
73	not detected	lift (a person)
74	26.66	listen to (a person)
75	not detected	play with kids
76	not detected	push (another person)
77	not detected	sing to (e.g., self, a person, a group)
78	2.00	take (an object) from (a person)
79	91.51	talk to (e.g., self, a person, a group)
80	48.29	watch (a person)