SidewalkPilot Steering Model Report

Common challenge-set current-label offline evaluation

Generated 2026-07-18 21:18:42. Every checkpoint is ranked on the same 6,952-image frozen temporal validation subset from the shared 81,237-frame dataset. Series 1/2 images are resized to their required 200x66 input and use their original model-specific preprocessing; Series 3/4 use 320x180. The original 2,224-image Series 1/2 evaluation is retained as a separate historical table, but it is not used for cross-generation ranking. Each Series 4 generation uses paired experiment names: p/r for PC, f/g for CF, and a/c for PCF (final/best respectively).

All-series class-balanced leader on the shared challenge set: 4.0p, with Bal9 34.5%, turn exact 32.1%, and turn +/-1 65.9%. Lowest shared-set MAE: 4.0c at 11.321 deg.

Series 1 uses raw BGR and output scale 86. Series 2 uses output scale 85; v2.0/v2.0b use legacy HSV/CLAHE preprocessing, while v2.1 and newer use raw BGR. Series 1/2 produce one continuous steering value. Series 3.0 and 3.0b produce two continuous values (steering and throttle). Series 3.1 and newer produce 19 raw values: 9 steering-class logits, 9 within-class offsets, and throttle. Series 4 removes throttle and emits one or four 18-value hybrid steering horizons; PC/PCF also consume three previous targets. Every architecture is decoded to current steering degrees before the common metrics are calculated.

Series 1/2 Historical Evaluation Sources

Source	Count	Purpose
D0328_17	315	First dataset relabel, March 28
D0329_15	413	First dataset relabel, March 29
D0425_14	65	Street test
D0426_18	53	Curves and shadows
D0427_18	72	Curved curb
D0429_19	53	Driveway and shadow
D0502_12	154	Shadow fix set
D0502_19	156	Hard turns, curb hugging, smoothness
D0503_17	159	Harsh sidewalk
D0506_20	24	8pm sidewalk failure set
D0510_18	167	v2.3 field-failure run 1: turns, road-right driving, driveways
D0510_19	8	v2.3 field-failure run 2 (short)
D0510_20	585	v2.3 field-failure run 3: turns, road-right driving, driveways

All-Series Shared Challenge Evaluation Sources

Source	Count	Purpose
D0702_16	4230	Series 3 July 2 base set: two manual field runs, 50,684 images
D0707_16	652	Bright-sun HARD/diagonal shadows across the sidewalk (v3.2 shadow-robustness batch)
D0712_16	2070	Series 3 manual field collection, July 12 run 2 (103-frame bad segment removed)

Series 1/2 Historical Class-Balanced Model Ranking

Ranked by macro-average exact recall across the 9 steering buckets, then turn within-one-bucket recall, then MAE. Each metric column has its own red-yellow-green scale: green means better within this table. Signed error is greenest nearest zero. Only the shared-challenge table supports cross-series comparison.

Rank	Model	Checkpoint filename	Prep	Bal9	Turn exact	Turn +/-1	ST exact	MAE	Med	Signed
1	2.3b	SidewalkPilot-v2.3b.pth	raw BGR	79.2%	78.8%	98.4%	62.8%	3.41	2.05	0.41
2	2.4b	SidewalkPilot-v2.4b.pth	raw BGR	78.2%	75.9%	99.5%	55.0%	3.27	2.39	-0.04
3	2.3	SidewalkPilot-v2.3.pth	raw BGR	77.8%	76.6%	98.5%	63.0%	3.30	1.93	0.01
4	2.4	SidewalkPilot-v2.4.pth	raw BGR	77.5%	74.9%	99.5%	55.5%	3.28	2.46	-0.23
5	2.2	SidewalkPilot-v2.2.pth	raw BGR	64.6%	62.4%	92.4%	42.3%	6.68	4.26	-0.22
6	2.2b	SidewalkPilot-v2.2b.pth	raw BGR	64.3%	62.1%	92.1%	42.3%	6.68	4.23	-0.46
7	2.1	SidewalkPilot-v2.1.pth	raw BGR	54.2%	51.0%	79.4%	39.4%	10.80	5.83	-1.15
8	2.0b	SidewalkPilot-v2.0b.pth	HSV/CLAHE -> BGR	54.2%	50.0%	78.8%	38.9%	11.13	5.07	-2.62
9	2.1b	SidewalkPilot-v2.1b.pth	raw BGR	54.1%	50.9%	79.6%	38.2%	10.75	5.80	-1.12
10	2.0	SidewalkPilot-v2.0.pth	HSV/CLAHE -> BGR	53.8%	49.5%	78.9%	38.5%	11.12	5.09	-2.63
11	1.9b	SidewalkPilot-v1.9b.pth	raw BGR	43.2%	41.8%	75.3%	37.3%	12.94	6.60	-2.06
12	1.9	SidewalkPilot-v1.9.pth	raw BGR	43.1%	41.3%	75.1%	36.9%	12.98	6.63	-2.06
13	1.8b	SidewalkPilot-v1.8b.pth	raw BGR	38.7%	35.8%	69.6%	36.0%	15.41	9.36	-0.70
14	1.8	SidewalkPilot-v1.8.pth	raw BGR	38.6%	35.8%	69.5%	36.1%	15.47	9.46	-0.65
15	1.7b	SidewalkPilot-v1.7b.pth	raw BGR	35.4%	35.1%	69.9%	37.5%	15.54	9.53	-0.87
16	1.7	SidewalkPilot-v1.7.pth	raw BGR	35.0%	34.9%	69.2%	37.3%	15.82	9.66	-0.48
17	1.5	SidewalkPilot-v1.5.pth	raw BGR	30.8%	30.7%	63.9%	33.3%	16.82	10.59	-0.89
18	1.5b	SidewalkPilot-v1.5b.pth	raw BGR	30.8%	30.7%	63.9%	33.3%	16.82	10.59	-0.89
19	1.4	SidewalkPilot-v1.4.pth	raw BGR	30.5%	29.4%	63.3%	36.6%	16.81	10.19	-3.99
20	1.6b	SidewalkPilot-v1.6b.pth	raw BGR	30.0%	30.5%	64.9%	30.5%	17.41	11.04	-0.01

Rank	Model	Checkpoint filename	Prep	Bal9	Turn exact	Turn +/-1	ST exact	MAE	Med	Signed
21	1.6	SidewalkPilot-v1.6.pth	raw BGR	29.6%	29.6%	64.7%	29.9%	17.53	11.21	0.21
22	1.4b	SidewalkPilot-v1.4b.pth	raw BGR	29.0%	28.5%	62.2%	40.0%	17.04	10.07	-4.15
23	1.2	SidewalkPilot-v1.2.pth	raw BGR	27.8%	27.4%	61.5%	32.9%	18.64	11.79	-3.55
24	1.2b	SidewalkPilot-v1.2b.pth	raw BGR	27.4%	26.6%	62.4%	31.6%	18.63	11.85	-3.73
25	1.3	SidewalkPilot-v1.3.pth	raw BGR	27.0%	26.2%	62.5%	30.4%	18.55	11.81	-3.43
26	1.3b	SidewalkPilot-v1.3b.pth	raw BGR	26.9%	27.0%	62.5%	32.2%	18.48	11.24	-3.65
27	1.1	SidewalkPilot-v1.1.pth	raw BGR	24.7%	24.1%	56.1%	26.1%	20.85	13.42	-5.16
28	1.1b	SidewalkPilot-v1.1b.pth	raw BGR	24.7%	24.1%	56.1%	26.1%	20.85	13.42	-5.16
29	1.0b	SidewalkPilot-v1.0b.pth	raw BGR	17.6%	17.1%	47.1%	20.6%	25.07	17.45	-7.08
30	1.0	SidewalkPilot-v1.0.pth	raw BGR	15.3%	15.5%	45.1%	18.9%	25.55	18.46	-7.45

All-Series Shared Challenge Class-Balanced Model Ranking

Ranked by macro-average exact recall across the 9 steering buckets, then turn within-one-bucket recall, then MAE. Each metric column has its own red-yellow-green scale: green means better within this table. Signed error is greenest nearest zero. Only the shared-challenge table supports cross-series comparison.

Rank	Model	Checkpoint filename	Prep	Bal9	Turn exact	Turn +/-1	ST exact	MAE	Med	Signed
1	4.0p	SidewalkPilot-v4.0p.onnx	raw BGR 320x180	34.5%	32.1%	65.9%	67.7%	12.40	2.97	0.12
2	4.0a	SidewalkPilot-v4.0a.onnx	raw BGR 320x180	33.5%	30.9%	65.3%	68.1%	12.38	3.11	0.29
3	4.0r	SidewalkPilot-v4.0r.onnx	raw BGR 320x180	32.9%	27.4%	62.6%	77.6%	11.64	1.85	-1.14
4	4.0c	SidewalkPilot-v4.0c.onnx	raw BGR 320x180	32.0%	29.4%	62.9%	75.5%	11.32	1.83	-0.98
5	3.1	SidewalkPilot-v3.1.onnx	raw BGR 320x180	28.8%	27.2%	55.5%	55.7%	22.23	9.44	-3.11
6	3.1b	SidewalkPilot-v3.1b.onnx	raw BGR 320x180	28.2%	25.2%	53.8%	57.1%	20.59	9.33	-4.51
7	3.2	SidewalkPilot-v3.2.onnx	raw BGR 320x180	26.1%	24.3%	55.1%	51.0%	17.46	9.73	-1.45
8	4.0f	SidewalkPilot-v4.0f.onnx	raw BGR 320x180	25.4%	23.5%	56.4%	62.8%	15.62	6.72	1.06
9	4.1a	SidewalkPilot-v4.1a.onnx	raw BGR 320x180	25.3%	22.9%	57.0%	63.2%	14.80	6.35	0.13
10	4.1p	SidewalkPilot-v4.1p.onnx	raw BGR 320x180	25.1%	23.1%	56.5%	65.7%	14.59	5.05	-0.14
11	4.1f	SidewalkPilot-v4.1f.onnx	raw BGR 320x180	25.1%	22.9%	57.8%	61.5%	15.47	7.22	1.18
12	3.4	SidewalkPilot-v3.4.onnx	raw BGR 320x180	24.2%	22.6%	56.2%	64.2%	15.08	6.07	0.42
13	4.1g	SidewalkPilot-v4.1g.onnx	raw BGR 320x180	24.0%	21.2%	56.8%	67.4%	14.65	4.66	0.56
14	3.3	SidewalkPilot-v3.3.onnx	raw BGR 320x180	23.8%	18.5%	47.7%	65.3%	15.17	6.13	-5.41
15	4.1c	SidewalkPilot-v4.1c.onnx	raw BGR 320x180	23.1%	19.5%	52.3%	75.0%	13.57	2.27	-0.98
16	3.4b	SidewalkPilot-v3.4b.onnx	raw BGR 320x180	22.4%	19.1%	51.2%	72.7%	13.99	2.48	-1.57
17	3.2b	SidewalkPilot-v3.2b.onnx	raw BGR 320x180	21.0%	15.4%	43.9%	67.6%	14.64	5.13	-4.54
18	4.0g	SidewalkPilot-v4.0g.onnx	raw BGR 320x180	20.4%	17.1%	46.4%	76.0%	14.12	2.11	-1.86
19	3.3b	SidewalkPilot-v3.3b.onnx	raw BGR 320x180	19.0%	8.6%	36.9%	79.2%	16.08	1.90	-8.19
20	4.1r	SidewalkPilot-v4.1r.onnx	raw BGR 320x180	17.8%	9.2%	37.2%	87.4%	13.38	1.43	-5.77
21	2.3b	SidewalkPilot-v2.3b.pth	raw BGR	17.7%	15.1%	41.0%	30.1%	17.64	11.11	-6.35
22	2.2b	SidewalkPilot-v2.2b.pth	raw BGR	17.7%	14.6%	41.3%	28.9%	17.92	11.66	-6.16
23	2.2	SidewalkPilot-v2.2.pth	raw BGR	17.5%	14.2%	41.5%	29.5%	17.85	11.58	-5.95
24	2.3	SidewalkPilot-v2.3.pth	raw BGR	17.5%	14.8%	40.5%	31.8%	17.44	10.73	-6.73
25	2.0b	SidewalkPilot-v2.0b.pth	HSV/CLAHE -> BGR	17.3%	13.3%	38.9%	46.3%	15.30	7.72	-4.79
26	1.9	SidewalkPilot-v1.9.pth	raw BGR	17.3%	13.2%	38.0%	50.7%	14.77	7.32	-4.96
27	2.0	SidewalkPilot-v2.0.pth	HSV/CLAHE -> BGR	17.2%	13.1%	38.7%	45.9%	15.33	7.74	-4.82

Rank	Model	Checkpoint filename	Prep	Bal9	Turn exact	Turn +/-1	ST exact	MAE	Med	Signed
28	1.9b	SidewalkPilot-v1.9b.pth	raw BGR	17.0%	12.8%	37.8%	50.6%	14.76	7.29	-5.04
29	1.8	SidewalkPilot-v1.8.pth	raw BGR	16.8%	12.9%	38.9%	51.6%	14.57	7.51	-4.44
30	1.8b	SidewalkPilot-v1.8b.pth	raw BGR	16.7%	12.8%	38.5%	51.8%	14.54	7.43	-4.51
31	3.0	SidewalkPilot-v3.0.onnx	raw BGR 320x180	16.7%	16.5%	45.8%	23.5%	18.19	13.12	-2.28
32	2.1	SidewalkPilot-v2.1.pth	raw BGR	16.7%	13.4%	40.2%	36.8%	16.70	10.26	-5.79
33	2.1b	SidewalkPilot-v2.1b.pth	raw BGR	16.2%	13.0%	40.3%	36.8%	16.58	10.15	-5.68
34	3.0b	SidewalkPilot-v3.0b.onnx	raw BGR 320x180	16.2%	15.6%	44.6%	26.3%	17.57	12.63	-2.59
35	2.4	SidewalkPilot-v2.4.pth	raw BGR	15.9%	13.4%	38.5%	33.9%	17.09	10.15	-6.78
36	2.4b	SidewalkPilot-v2.4b.pth	raw BGR	15.9%	13.2%	38.6%	34.0%	17.05	10.10	-6.52
37	1.7b	SidewalkPilot-v1.7b.pth	raw BGR	15.6%	11.4%	35.7%	49.1%	15.61	7.51	-5.65
38	1.1	SidewalkPilot-v1.1.pth	raw BGR	15.3%	12.2%	37.0%	42.0%	17.54	9.32	-6.90
39	1.1b	SidewalkPilot-v1.1b.pth	raw BGR	15.3%	12.2%	37.0%	42.0%	17.54	9.32	-6.90
40	1.3	SidewalkPilot-v1.3.pth	raw BGR	15.1%	10.6%	36.4%	53.4%	15.73	7.25	-6.39
41	1.7	SidewalkPilot-v1.7.pth	raw BGR	15.1%	10.9%	35.9%	49.0%	15.62	7.55	-5.50
42	1.3b	SidewalkPilot-v1.3b.pth	raw BGR	14.9%	10.5%	36.2%	54.2%	15.46	7.03	-6.51
43	1.2	SidewalkPilot-v1.2.pth	raw BGR	14.8%	10.3%	35.4%	52.4%	16.47	7.97	-6.56
44	1.2b	SidewalkPilot-v1.2b.pth	raw BGR	14.6%	10.2%	35.6%	51.1%	16.69	8.02	-6.94
45	1.4b	SidewalkPilot-v1.4b.pth	raw BGR	14.4%	9.9%	33.8%	56.9%	15.98	6.67	-6.17
46	1.4	SidewalkPilot-v1.4.pth	raw BGR	14.3%	9.7%	32.7%	56.5%	15.89	6.86	-6.50
47	1.6b	SidewalkPilot-v1.6b.pth	raw BGR	14.0%	10.2%	34.9%	48.1%	16.54	8.19	-5.39
48	1.0b	SidewalkPilot-v1.0b.pth	raw BGR	13.9%	11.2%	30.8%	17.9%	25.70	19.36	-16.72
49	1.6	SidewalkPilot-v1.6.pth	raw BGR	13.8%	10.1%	34.8%	47.4%	16.54	8.21	-5.39
50	1.5	SidewalkPilot-v1.5.pth	raw BGR	13.1%	8.0%	31.8%	59.3%	16.11	6.35	-5.51
51	1.5b	SidewalkPilot-v1.5b.pth	raw BGR	13.1%	8.0%	31.8%	59.3%	16.11	6.35	-5.51
52	1.0	SidewalkPilot-v1.0.pth	raw BGR	13.1%	10.7%	31.1%	16.7%	25.94	19.65	-16.48

Chronological Model Growth

Every row uses the same 6,952-anchor shared challenge set, so the metrics are directly comparable across series.

Model	Checkpoint filename	Series	Eval set	Prep	Scale	MAE	Median	Signed	<=2	<=5	<=20	Pred mean
1.0	SidewalkPilot-v1.0.pth	1	Shared	raw BGR	86	25.94	19.65	-16.48	371	965	3528	79.80
1.0b	SidewalkPilot-v1.0b.pth	1	Shared	raw BGR	86	25.70	19.36	-16.72	387	1039	3562	79.55
1.1	SidewalkPilot-v1.1.pth	1	Shared	raw BGR	86	17.54	9.32	-6.90	938	2229	5019	89.37
1.1b	SidewalkPilot-v1.1b.pth	1	Shared	raw BGR	86	17.54	9.32	-6.90	938	2229	5019	89.37
1.2	SidewalkPilot-v1.2.pth	1	Shared	raw BGR	86	16.47	7.97	-6.56	1248	2716	5093	89.72
1.2b	SidewalkPilot-v1.2b.pth	1	Shared	raw BGR	86	16.69	8.02	-6.94	1253	2634	5051	89.34
1.3	SidewalkPilot-v1.3.pth	1	Shared	raw BGR	86	15.73	7.25	-6.39	1407	2785	5197	89.89
1.3b	SidewalkPilot-v1.3b.pth	1	Shared	raw BGR	86	15.46	7.03	-6.51	1340	2814	5264	89.77
1.4	SidewalkPilot-v1.4.pth	1	Shared	raw BGR	86	15.89	6.86	-6.50	1340	2902	5174	89.78
1.4b	SidewalkPilot-v1.4b.pth	1	Shared	raw BGR	86	15.98	6.67	-6.17	1371	2917	5178	90.10
1.5	SidewalkPilot-v1.5.pth	1	Shared	raw BGR	86	16.11	6.35	-5.51	1413	3012	5130	90.77
1.5b	SidewalkPilot-v1.5b.pth	1	Shared	raw BGR	86	16.11	6.35	-5.51	1413	3012	5130	90.77
1.6	SidewalkPilot-v1.6.pth	1	Shared	raw BGR	86	16.54	8.21	-5.39	1126	2447	5130	90.89
1.6b	SidewalkPilot-v1.6b.pth	1	Shared	raw BGR	86	16.54	8.19	-5.39	1147	2482	5152	90.89
1.7	SidewalkPilot-v1.7.pth	1	Shared	raw BGR	86	15.62	7.55	-5.50	1120	2574	5219	90.77
1.7b	SidewalkPilot-v1.7b.pth	1	Shared	raw BGR	86	15.61	7.51	-5.65	1123	2584	5197	90.63
1.8	SidewalkPilot-v1.8.pth	1	Shared	raw BGR	86	14.57	7.51	-4.44	1282	2699	5325	91.84
1.8b	SidewalkPilot-v1.8b.pth	1	Shared	raw BGR	86	14.54	7.43	-4.51	1294	2713	5328	91.76
1.9	SidewalkPilot-v1.9.pth	1	Shared	raw BGR	86	14.77	7.32	-4.96	1247	2678	5362	91.31
1.9b	SidewalkPilot-v1.9b.pth	1	Shared	raw BGR	86	14.76	7.29	-5.04	1247	2686	5367	91.24
2.0	SidewalkPilot-v2.0.pth	2	Shared	HSV/CLAHE -> BGR	85	15.33	7.74	-4.82	1059	2453	5311	91.46
2.0b	SidewalkPilot-v2.0b.pth	2	Shared	HSV/CLAHE -> BGR	85	15.30	7.72	-4.79	1066	2460	5309	91.48
2.1	SidewalkPilot-v2.1.pth	2	Shared	raw BGR	85	16.70	10.26	-5.79	873	1993	5085	90.48
2.1b	SidewalkPilot-v2.1b.pth	2	Shared	raw BGR	85	16.58	10.15	-5.68	879	1993	5109	90.60
2.2	SidewalkPilot-v2.2.pth	2	Shared	raw BGR	85	17.85	11.58	-5.95	655	1659	4868	90.33
2.2b	SidewalkPilot-v2.2b.pth	2	Shared	raw BGR	85	17.92	11.66	-6.16	667	1650	4846	90.12
2.3	SidewalkPilot-v2.3.pth	2	Shared	raw BGR	85	17.44	10.73	-6.73	720	1788	5018	89.54

Model	Checkpoint filename	Series	Eval set	Prep	Scale	MAE	Median	Signed	<=2	<=5	<=20	Pred mean
2.3b	SidewalkPilot-v2.3b.pth	2	Shared	raw BGR	85	17.64	11.11	-6.35	680	1699	4987	89.93
2.4	SidewalkPilot-v2.4.pth	2	Shared	raw BGR	85	17.09	10.15	-6.78	787	1888	5086	89.50
2.4b	SidewalkPilot-v2.4b.pth	2	Shared	raw BGR	85	17.05	10.10	-6.52	785	1886	5085	89.75
3.0	SidewalkPilot-v3.0.onnx	3	Shared	raw BGR 320x180	-	18.19	13.12	-2.28	570	1402	4707	94.00
3.0b	SidewalkPilot-v3.0b.onnx	3	Shared	raw BGR 320x180	-	17.57	12.63	-2.59	642	1544	4827	93.69
3.1	SidewalkPilot-v3.1.onnx	3	Shared	raw BGR 320x180	-	22.23	9.44	-3.11	2571	3000	4439	93.16
3.1b	SidewalkPilot-v3.1b.onnx	3	Shared	raw BGR 320x180	-	20.59	9.33	-4.51	2657	3062	4539	91.77
3.2	SidewalkPilot-v3.2.onnx	3	Shared	raw BGR 320x180	-	17.46	9.73	-1.45	2414	2751	4686	94.82
3.2b	SidewalkPilot-v3.2b.onnx	3	Shared	raw BGR 320x180	-	14.64	5.13	-4.54	3112	3461	5184	91.74
3.3	SidewalkPilot-v3.3.onnx	3	Shared	raw BGR 320x180	-	15.17	6.13	-5.41	2949	3360	4970	90.87
3.3b	SidewalkPilot-v3.3b.onnx	3	Shared	raw BGR 320x180	-	16.08	1.90	-8.19	3542	3899	5073	88.08
3.4	SidewalkPilot-v3.4.onnx	3	Shared	raw BGR 320x180	-	15.08	6.07	0.42	2943	3372	4974	96.69
3.4b	SidewalkPilot-v3.4b.onnx	3	Shared	raw BGR 320x180	-	13.99	2.48	-1.57	3367	3762	5198	94.71
4.0p	SidewalkPilot-v4.0p.onnx	4	Shared	raw BGR 320x180	-	12.40	2.97	0.12	3268	3717	5452	96.40
4.0r	SidewalkPilot-v4.0r.onnx	4	Shared	raw BGR 320x180	-	11.64	1.85	-1.14	3598	4110	5526	95.14
4.0f	SidewalkPilot-v4.0f.onnx	4	Shared	raw BGR 320x180	-	15.62	6.72	1.06	2900	3308	4838	97.33
4.0g	SidewalkPilot-v4.0g.onnx	4	Shared	raw BGR 320x180	-	14.12	2.11	-1.86	3405	3868	5233	94.41
4.0a	SidewalkPilot-v4.0a.onnx	4	Shared	raw BGR 320x180	-	12.38	3.11	0.29	3252	3713	5454	96.57
4.0c	SidewalkPilot-v4.0c.onnx	4	Shared	raw BGR 320x180	-	11.32	1.83	-0.98	3556	4016	5577	95.30
4.1p	SidewalkPilot-v4.1p.onnx	4	Shared	raw BGR 320x180	-	14.59	5.05	-0.14	3008	3470	5063	96.13
4.1r	SidewalkPilot-v4.1r.onnx	4	Shared	raw BGR 320x180	-	13.38	1.43	-5.77	3994	4311	5330	90.51
4.1f	SidewalkPilot-v4.1f.onnx	4	Shared	raw BGR 320x180	-	15.47	7.22	1.18	2861	3245	4869	97.46
4.1g	SidewalkPilot-v4.1g.onnx	4	Shared	raw BGR 320x180	-	14.65	4.66	0.56	3102	3515	5025	96.83
4.1a	SidewalkPilot-v4.1a.onnx	4	Shared	raw BGR 320x180	-	14.80	6.35	0.13	2926	3327	4994	96.41
4.1c	SidewalkPilot-v4.1c.onnx	4	Shared	raw BGR 320x180	-	13.57	2.27	-0.98	3378	3840	5239	95.29

Series 1/2 Historical Field-Case / Subset MAE

Lower is better. All rows in this table use the same evaluation dataset.

Model	D0328_17	D0329_15	D0425_14	D0426_18	D0427_18	D0429_19	D0502_12	D0502_19	D0503_17	D0506_20	D0510_18	D0510_19	D0510_20
2.3b	2.45	2.57	3.41	3.15	3.42	2.52	3.46	3.23	2.50	4.00	4.47	2.79	4.60
2.4b	3.16	2.92	4.11	4.04	4.08	2.51	4.62	3.58	2.78	2.47	3.13	1.83	3.15
2.3	2.39	2.51	3.30	3.04	3.46	2.54	3.23	3.07	2.38	3.71	4.20	2.65	4.50
2.4	3.14	2.96	4.09	3.97	4.12	2.48	4.66	3.60	2.84	2.48	3.21	1.69	3.16
2.2	3.84	3.62	5.42	4.70	5.66	4.39	5.68	4.74	3.53	2.50	14.71	8.43	10.51
2.2b	3.81	3.56	5.46	4.97	5.61	4.38	5.67	4.72	3.50	2.62	14.65	8.25	10.58
2.1	19.70	20.13	3.69	3.26	3.89	3.32	4.14	3.62	2.47	3.60	10.64	6.88	8.76
2.0b	18.09	20.54	2.69	2.02	2.57	1.70	2.32	2.25	1.66	35.80	11.71	7.49	10.54
2.1b	19.65	19.96	3.65	3.19	3.81	3.33	4.10	3.64	2.42	3.63	10.67	6.86	8.73
2.0	17.97	20.51	2.71	2.03	2.57	1.69	2.35	2.26	1.66	36.00	11.70	7.45	10.59
1.9b	17.10	21.79	1.67	1.49	1.73	1.43	1.36	1.78	28.24	35.89	10.51	5.75	10.87
1.9	17.27	21.89	1.66	1.52	1.74	1.41	1.32	1.79	28.23	35.45	10.48	5.70	10.91
1.8b	17.43	20.92	1.52	1.19	1.42	0.97	1.08	22.57	28.02	37.04	12.89	5.38	14.69
1.8	17.40	20.92	1.51	1.18	1.42	0.96	1.08	22.62	28.08	37.19	13.08	5.39	14.87
1.7b	16.52	20.73	1.39	1.03	1.23	1.01	17.43	22.69	27.59	30.87	10.19	5.57	12.68
1.7	16.55	21.03	1.44	1.12	1.23	0.93	17.51	22.84	27.33	30.78	10.21	5.76	13.50
1.5	16.69	20.32	1.24	0.88	0.99	25.74	18.59	21.31	26.42	26.32	10.75	6.48	15.96
1.5b	16.69	20.32	1.24	0.88	0.99	25.74	18.59	21.31	26.42	26.32	10.75	6.48	15.96
1.4	16.34	20.23	0.88	0.56	31.68	25.83	18.73	24.90	26.43	34.05	10.71	6.82	11.14
1.6b	17.01	20.01	1.24	0.90	1.10	26.22	19.02	22.90	27.36	29.09	10.00	5.66	17.51
1.6	16.84	20.09	1.26	0.84	1.10	26.32	18.90	22.95	27.36	29.10	9.99	5.64	17.99
1.4b	16.94	20.62	1.00	0.72	31.91	26.65	19.21	24.99	26.13	34.16	10.42	7.00	11.30
1.2	16.38	19.88	0.87	39.03	31.76	25.39	20.86	23.31	32.51	36.88	14.69	6.91	11.84
1.2b	16.39	19.86	0.93	38.42	32.11	25.46	21.01	23.36	32.64	36.64	14.57	6.80	11.77
1.3	16.46	20.12	0.87	37.56	34.06	25.27	19.51	23.50	34.04	32.99	13.71	6.20	11.46
1.3b	16.63	20.54	1.02	36.75	35.63	24.12	19.66	23.53	34.63	32.20	13.01	6.25	10.81
1.1	17.12	20.33	1.32	41.55	36.60	33.44	21.42	26.71	33.09	29.72	13.52	5.50	17.34

Model	D0328_17	D0329_15	D0425_14	D0426_18	D0427_18	D0429_19	D0502_12	D0502_19	D0503_17	D0506_20	D0510_18	D0510_19	D0510_20
1.1b	17.12	20.33	1.32	41.55	36.60	33.44	21.42	26.71	33.09	29.72	13.52	5.50	17.34
1.0b	17.36	20.30	49.19	45.40	54.06	37.39	29.15	30.30	40.51	40.61	17.89	15.52	18.26
1.0	18.07	21.08	49.02	46.31	53.69	37.44	29.56	30.14	41.72	40.76	17.95	14.47	18.74

All-Series Shared Challenge Field-Case / Subset MAE

Lower is better. All rows in this table use the same evaluation dataset.

Model	D0702_16	D0707_16	D0712_16
4.0p	9.69	11.40	18.24
4.0a	9.68	11.13	18.29
4.0r	9.70	9.86	16.15
4.0c	9.48	10.34	15.39
3.1	15.15	27.26	35.12
3.1b	13.25	26.21	33.82
3.2	15.33	20.39	20.89
4.0f	12.08	17.89	22.15
4.1a	11.48	18.37	20.48
4.1p	11.45	16.95	20.25
4.1f	12.13	17.17	21.76
3.4	11.77	16.06	21.55
4.1g	11.56	15.48	20.71
3.3	11.72	21.05	20.36
4.1c	10.84	16.39	18.26
3.4b	10.86	19.15	18.74
3.2b	11.83	21.18	18.33
4.0g	10.85	16.88	19.92
3.3b	11.08	18.46	25.53
4.1r	10.78	14.99	18.19
2.3b	15.73	18.20	21.38
2.2b	16.15	18.69	21.31

Model	D0702_16	D0707_16	D0712_16
2.2	16.03	18.72	21.30
2.3	15.47	18.11	21.26
2.0b	13.15	17.03	19.15
1.9	12.88	18.52	17.46
2.0	13.17	17.03	19.21
1.9b	12.90	18.50	17.39
1.8	12.97	15.96	17.40
1.8b	12.94	15.89	17.41
3.0	16.85	16.60	21.41
2.1	15.09	18.90	19.29
2.1b	14.98	18.68	19.19
3.0b	16.00	16.67	21.08
2.4	14.34	18.40	22.30
2.4b	14.30	18.49	22.22
1.7b	13.13	16.25	20.46
1.1	14.75	17.38	23.29
1.1b	14.75	17.38	23.29
1.3	13.05	16.04	21.12
1.7	13.19	16.32	20.34
1.3b	13.00	15.46	20.48
1.2	13.46	17.61	22.27
1.2b	13.51	18.12	22.75
1.4b	12.77	17.35	22.11
1.4	12.97	17.25	21.43
1.6b	13.11	18.74	22.86
1.0b	23.77	26.15	29.51
1.6	13.20	18.94	22.62
1.5	13.11	18.54	21.46
1.5b	13.11	18.54	21.46

Model	D0702_16	D0707_16	D0712_16
1.0	23.79	26.67	30.10

Prediction Distribution

Model	Pred min	P05	P25	Median	P75	P95	Pred max	Pred mean	Target mean
4.0p	7.25	69.57	88.71	90.02	92.07	152.14	170.05	96.40	96.28
4.0a	5.87	69.28	88.78	90.06	92.45	152.39	171.42	96.57	96.28
4.0r	7.66	79.29	88.75	89.86	91.44	149.89	169.99	95.14	96.28
4.0c	7.46	79.14	88.92	90.00	91.33	149.83	165.81	95.30	96.28
3.1	1.12	17.09	80.83	89.66	99.73	157.98	179.06	93.16	96.28
3.1b	1.19	18.03	80.79	89.84	99.41	156.15	178.85	91.77	96.28
3.2	1.36	67.28	80.43	90.08	99.95	155.52	177.91	94.82	96.28
4.0f	5.34	68.12	88.66	89.97	99.93	152.24	174.96	97.33	96.28
4.1a	7.18	68.40	88.73	90.09	99.60	150.55	172.77	96.41	96.28
4.1p	1.99	68.71	88.70	90.03	92.05	149.26	175.64	96.13	96.28
4.1f	4.10	68.33	88.80	90.14	100.26	151.28	174.49	97.46	96.28
3.4	2.86	68.98	88.69	90.03	92.41	150.96	177.49	96.69	96.28
4.1g	6.20	68.91	88.72	89.92	91.74	151.03	173.37	96.83	96.28
3.3	2.03	67.61	88.61	90.08	91.64	127.68	173.96	90.87	96.28
4.1c	5.34	69.10	88.95	90.12	91.42	149.83	172.66	95.29	96.28
3.4b	8.12	68.66	88.85	89.91	91.25	148.82	170.13	94.71	96.28
3.2b	1.36	68.32	88.82	90.11	91.72	113.33	178.38	91.74	96.28
4.0g	9.04	79.43	88.54	89.62	90.99	148.98	171.81	94.41	96.28
3.3b	7.24	24.76	88.81	89.85	91.13	111.71	175.42	88.08	96.28
4.1r	21.16	79.58	88.83	89.55	90.50	126.98	159.03	90.51	96.28
2.3b	5.01	64.41	80.17	90.14	99.10	116.09	174.37	89.93	96.28
2.2b	5.01	61.56	80.38	90.52	99.93	117.46	173.02	90.12	96.28
2.2	5.01	61.97	80.73	90.88	100.11	117.30	173.40	90.33	96.28
2.3	5.01	64.58	80.20	89.89	98.38	115.32	174.31	89.54	96.28
2.0b	5.55	71.09	85.29	91.44	97.11	111.05	172.70	91.48	96.28
1.9	8.42	72.04	85.99	91.30	96.72	109.32	175.58	91.31	96.28
2.0	5.54	71.07	85.22	91.46	97.14	110.98	172.75	91.46	96.28
1.9b	7.43	72.01	85.95	91.24	96.58	108.90	175.60	91.24	96.28

Model	Pred min	P05	P25	Median	P75	P95	Pred max	Pred mean	Target mean
1.8	8.36	73.63	86.38	91.20	97.20	110.93	175.97	91.84	96.28
1.8b	8.58	73.72	86.33	91.12	97.03	110.71	175.97	91.76	96.28
3.0	20.05	66.32	82.25	93.94	104.79	123.70	166.77	94.00	96.28
2.1	8.44	66.95	82.29	90.29	98.23	114.59	173.63	90.48	96.28
2.1b	8.51	67.23	82.48	90.47	98.23	114.35	173.53	90.60	96.28
3.0b	25.02	67.29	83.31	93.49	104.30	119.92	159.07	93.69	96.28
2.4	5.04	65.79	81.02	89.85	97.47	111.53	172.68	89.50	96.28
2.4b	5.05	65.98	81.34	90.13	97.67	111.89	172.88	89.75	96.28
1.7b	4.88	71.19	84.83	89.96	96.07	112.04	175.89	90.63	96.28
1.1	4.01	64.77	82.43	88.67	95.56	116.70	175.54	89.37	96.28
1.1b	4.01	64.77	82.43	88.67	95.56	116.70	175.54	89.37	96.28
1.3	4.00	69.41	85.39	89.90	95.09	110.81	175.99	89.89	96.28
1.7	5.23	71.56	84.91	90.14	96.17	112.00	175.84	90.77	96.28
1.3b	4.00	71.02	85.21	89.59	94.61	109.17	176.00	89.77	96.28
1.2	4.04	65.97	84.42	90.41	94.84	113.24	175.98	89.72	96.28
1.2b	4.01	65.63	83.75	90.14	94.64	113.57	175.98	89.34	96.28
1.4b	4.33	71.69	84.81	89.45	93.76	114.46	175.99	90.10	96.28
1.4	4.42	71.57	84.59	89.30	93.65	111.76	175.99	89.78	96.28
1.6b	4.37	68.95	85.73	90.94	96.77	111.71	175.74	90.89	96.28
1.0b	10.66	39.59	64.66	80.17	94.67	116.63	166.57	79.55	96.28
1.6	4.49	68.73	85.56	90.87	96.75	112.12	175.69	90.89	96.28
1.5	4.63	73.49	85.93	89.83	94.25	113.17	175.81	90.77	96.28
1.5b	4.63	73.49	85.93	89.83	94.25	113.17	175.81	90.77	96.28
1.0	10.30	38.08	64.24	80.71	95.79	117.88	166.76	79.80	96.28

Prediction Buckets vs Ground

Bucket columns show prediction counts; exact/off-by-one compare prediction bucket against ground bucket.

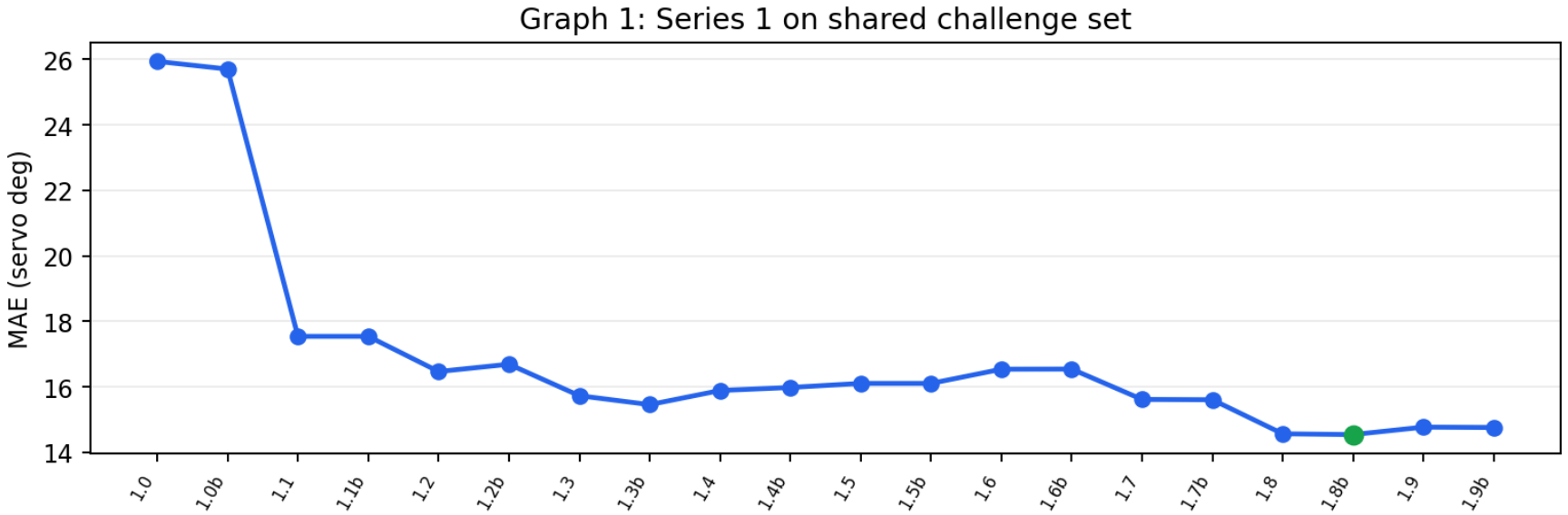
Model	HL 0-45	L 45-75	SL 75-85	ST 85-95	SR 95-105	R 105-135	HR 135-180	Exact	Off by <=1
4.0p	36	329	1018	3873	178	897	621	58.9%	81.4%

Model	HL 0-45	L 45-75	SL 75-85	ST 85-95	SR 95-105	R 105-135	HR 135-180	Exact	Off by <=1
4.0a	28	365	931	3906	251	831	640	59.0%	81.3%
4.0r	75	255	624	4509	77	966	446	64.6%	81.7%
4.0c	38	299	848	4404	110	684	569	63.4%	82.8%
3.1	637	364	762	3291	257	729	912	48.2%	68.3%
3.1b	588	380	793	3398	214	860	719	48.9%	69.1%
3.2	178	760	877	3003	758	617	759	44.3%	71.8%
4.0f	79	460	807	3742	188	843	833	52.4%	74.1%
4.1a	71	440	890	3761	224	843	723	52.7%	75.9%
4.1p	77	359	925	3945	147	749	750	54.3%	76.5%
4.1f	66	459	862	3686	177	900	802	51.6%	74.0%
3.4	79	384	923	3858	168	782	758	53.1%	75.5%
4.1g	69	349	822	4076	110	804	722	55.0%	76.1%
3.3	174	694	588	4069	548	536	343	52.2%	73.0%
4.1c	62	328	693	4596	76	568	629	59.1%	78.0%
3.4b	59	376	764	4526	34	668	525	57.6%	77.1%
3.2b	171	270	852	4200	585	580	294	52.7%	77.1%
4.0g	77	166	842	4767	105	347	648	58.4%	77.8%
3.3b	465	205	170	5182	416	311	203	57.8%	73.7%
4.1r	164	156	335	5809	87	102	299	63.0%	76.3%
2.3b	67	1025	1395	1978	1573	833	81	26.7%	66.6%
2.2b	68	1033	1391	1884	1494	981	101	26.0%	65.6%
2.2	69	1001	1351	1912	1529	995	95	26.3%	66.2%
2.3	69	1011	1418	2075	1533	773	73	27.6%	67.3%
2.0b	38	510	1137	2943	1718	503	103	36.6%	75.2%
1.9	36	420	1054	3254	1647	476	65	39.6%	75.8%
2.0	37	514	1148	2915	1734	501	103	36.3%	75.1%
1.9b	38	412	1072	3245	1651	471	63	39.4%	75.9%
1.8	16	401	990	3317	1558	624	46	40.5%	75.1%
1.8b	17	393	1001	3339	1541	615	46	40.5%	75.0%

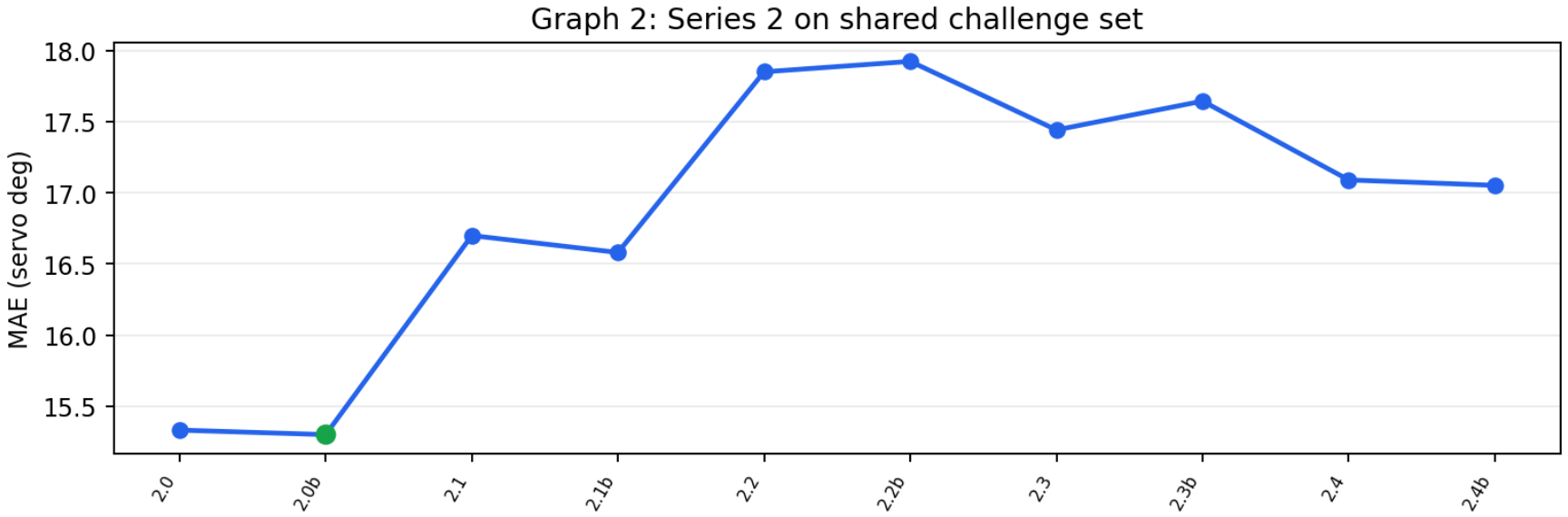
Model	HL 0-45	L 45-75	SL 75-85	ST 85-95	SR 95-105	R 105-135	HR 135-180	Exact	Off by <=1
3.0	30	848	1234	1554	1569	1621	96	23.8%	63.1%
2.1	61	748	1390	2403	1452	813	85	31.0%	69.6%
2.1b	55	733	1383	2409	1479	812	81	30.9%	69.9%
3.0b	16	807	1173	1753	1564	1570	69	25.4%	64.4%
2.4	56	846	1502	2244	1600	621	83	28.6%	69.6%
2.4b	57	827	1452	2250	1635	647	84	28.6%	69.7%
1.7b	51	451	1271	3202	1379	514	84	37.9%	74.1%
1.1	118	659	1521	2822	1055	641	136	33.5%	69.2%
1.1b	118	659	1521	2822	1055	641	136	33.5%	69.2%
1.3	100	446	1108	3543	1148	549	58	40.7%	73.3%
1.7	55	445	1257	3199	1380	534	82	37.7%	74.0%
1.3b	77	415	1181	3608	1138	484	49	41.0%	74.1%
1.2	103	586	1134	3439	1089	497	104	39.9%	71.2%
1.2b	115	607	1222	3359	1042	508	99	39.0%	70.8%
1.4b	69	414	1305	3737	863	437	127	42.5%	72.6%
1.4	65	415	1362	3720	863	428	99	42.2%	72.6%
1.6b	117	431	1029	3184	1562	521	108	36.8%	72.1%
1.0b	524	2302	1232	1202	893	709	90	17.3%	46.7%
1.6	116	429	1076	3152	1541	528	110	36.2%	72.0%
1.5	87	309	1082	3922	995	437	120	43.5%	72.4%
1.5b	87	309	1082	3922	995	437	120	43.5%	72.4%
1.0	583	2206	1199	1136	929	812	87	16.5%	46.4%

Shared Challenge-Set Graphs

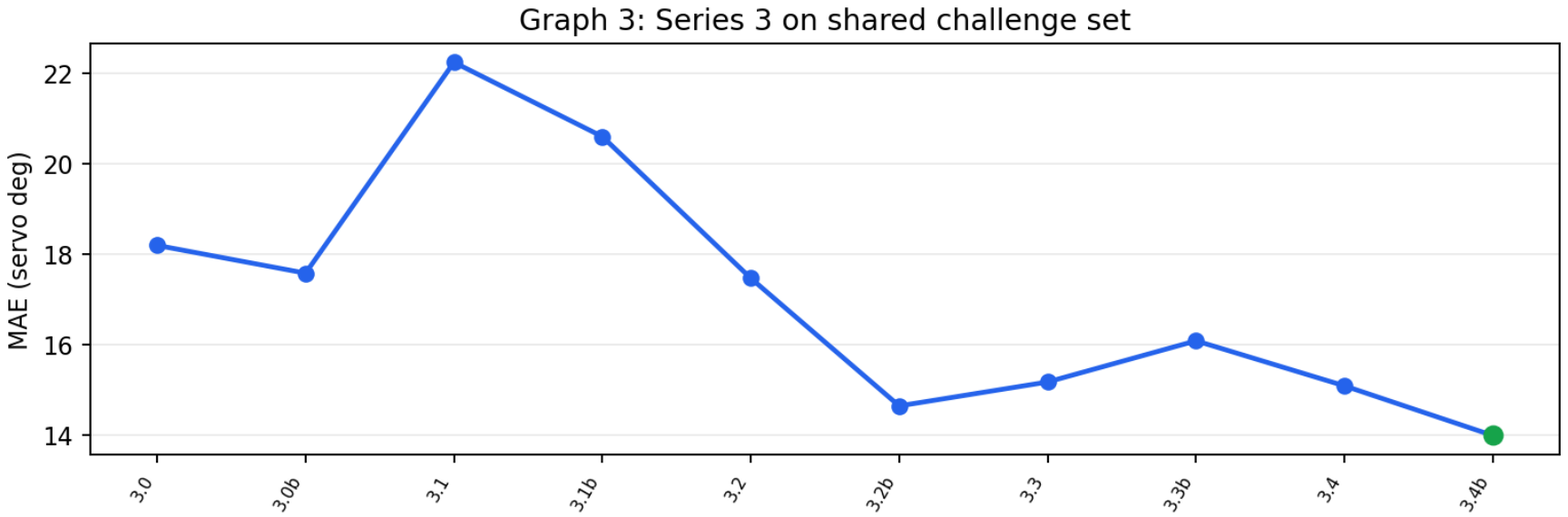
Graph 1: Series 1 on shared challenge set



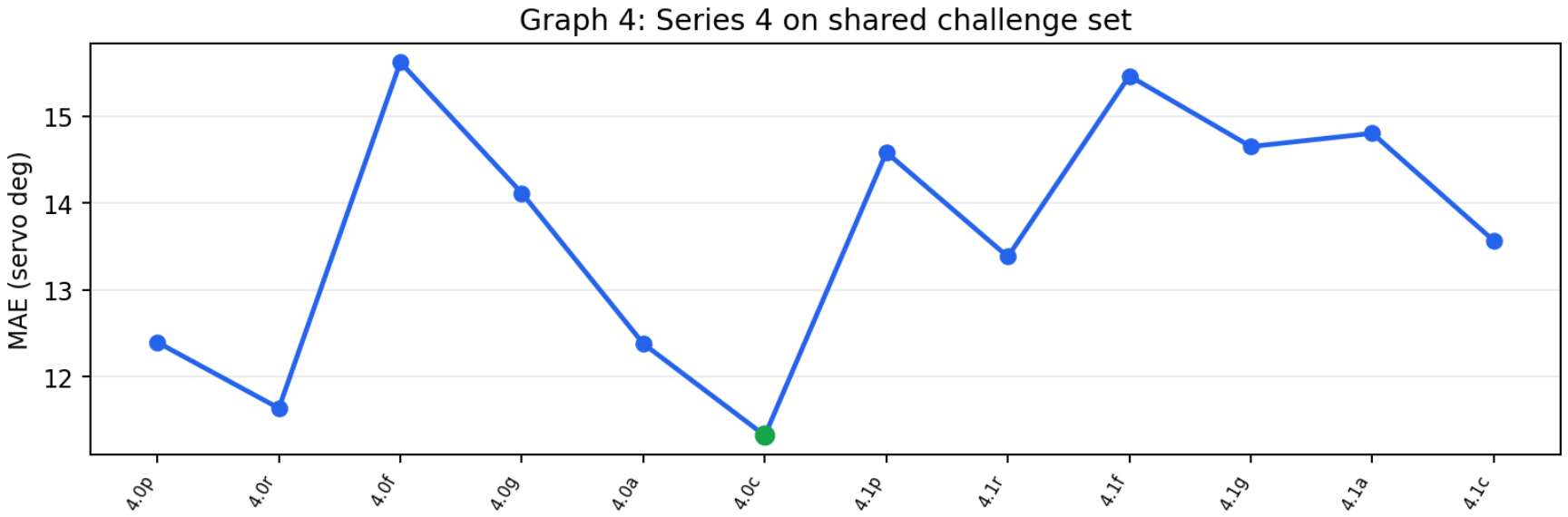
Graph 2: Series 2 on shared challenge set



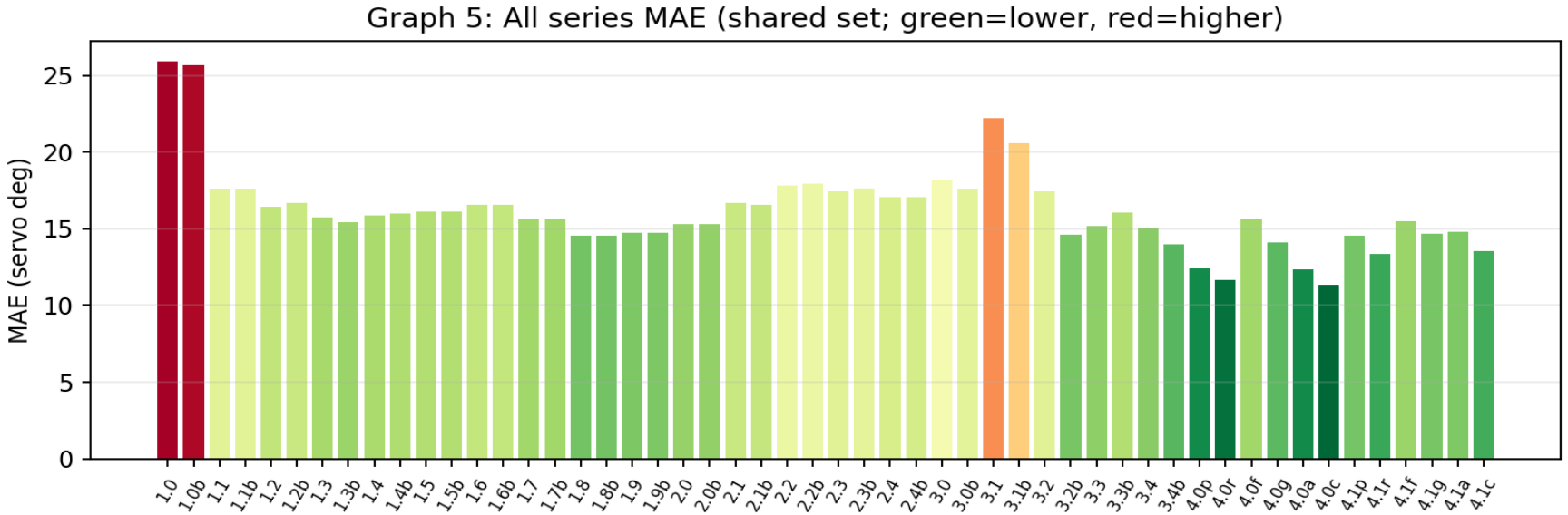
Graph 3: Series 3 on shared challenge set



Graph 4: Series 4 on shared challenge set



Graph 5: All series MAE (shared set; green=lower, red=higher)



Steering-Bucket Confusion (9-class) - latest Series 3/4 models

Rows = the TRUE steering bucket, columns = the model's PREDICTED bucket. Green diagonal = correct; red off-diagonal = confusion. Shading = share of that true bucket (each row sums to ~100%). Buckets: HL 0-45, L 45-60, L+ 60-75, SL 75-85, ST 85-95, SR 95-105, R 105-120, R+ 120-135, HR 135-180.

v4.1c · exact bucket 57% · within one bucket 75%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	8	1	18	4	15	2	·	·	2
L	3	·	33	22	20	·	·	·	10
L+	8	·	77	66	72	·	·	·	·
SL	11	·	68	123	144	·	3	2	6
ST	26	·	127	464	3557	29	170	77	291
SR	1	·	1	3	163	4	22	9	36
R	·	·	·	4	186	9	34	21	61
R+	·	·	2	·	161	23	69	35	72
HR	5	·	1	7	278	9	73	53	151

v4.1a · exact bucket 50% · within one bucket 72%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	12	3	13	6	11	·	·	·	5
L	5	6	24	29	16	·	·	·	8
L+	2	9	76	79	53	1	·	1	2
SL	5	7	83	137	115	1	3	3	3
ST	43	22	193	612	2998	117	329	98	329
SR	1	·	·	7	120	16	34	23	38
R	·	·	1	5	132	27	48	34	68
R+	·	·	3	2	106	42	90	30	89
HR	3	·	·	13	210	20	107	43	181

v4.1g · exact bucket 53% · within one bucket 73%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	13	4	14	6	11	·	·	·	2
L	2	1	32	27	19	1	·	·	6
L+	3	6	68	77	67	·	·	1	1
SL	5	2	67	134	137	·	8	·	4
ST	39	4	146	551	3197	50	254	152	348
SR	·	·	1	5	141	4	30	20	38
R	·	·	2	6	155	15	23	51	63
R+	·	·	1	4	121	33	67	51	85
HR	7	·	1	12	228	7	75	72	175

v4.1f · exact bucket 49% · within one bucket 70%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	14	4	11	6	9	·	·	·	6
L	3	4	36	23	16	·	·	·	6
L+	4	8	75	76	57	·	·	·	3
SL	4	4	90	136	112	2	5	1	3
ST	35	10	210	598	2916	80	371	128	393
SR	·	·	2	4	130	9	36	16	42
R	·	·	2	4	138	27	39	37	68
R+	·	·	2	2	110	39	80	39	90
HR	6	·	1	13	198	20	105	43	191

v4.1r · exact bucket 63% · within one bucket 75%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	7	3	8	2	29	·	·	·	1
L	9	4	14	13	40	1	·	1	6
L+	15	3	41	43	115	·	·	·	6
SL	21	6	15	68	240	1	·	·	6
ST	101	6	49	207	4146	51	·	55	126
SR	1	·	1	·	212	5	·	5	15
R	1	·	·	·	265	6	·	15	28
R+	3	2	·	1	302	4	1	8	41
HR	6	2	2	1	460	19	·	17	70

v4.1p · exact bucket 52% · within one bucket 73%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	12	2	19	4	11	·	·	·	2
L	6	1	28	30	18	·	·	·	5
L+	5	2	75	76	62	1	·	·	2
SL	5	1	70	155	115	·	5	3	3
ST	43	8	145	633	3113	63	290	87	359
SR	1	1	2	5	145	9	26	16	34
R	·	·	1	8	139	21	37	35	74
R+	1	·	2	4	114	26	88	39	88
HR	4	·	2	10	228	27	87	36	183

v4.0c · exact bucket 61% · within one bucket 80%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	18	4	6	6	15	1	·	·	·
L	6	3	36	20	23	·	·	·	·
L+	1	·	85	66	70	·	1	·	·
SL	2	·	57	185	112	·	·	·	1
ST	11	6	102	563	3579	51	159	61	209
SR	·	·	·	4	124	15	54	16	26
R	·	·	·	2	133	15	75	39	51
R+	·	·	·	2	105	11	104	63	77
HR	·	·	·	·	243	17	64	48	205

v4.0a · exact bucket 56% · within one bucket 78%

T\P	HL	L	L+	SL	ST	SR	R	R+	HR
HL	17	7	5	10	10	·	·	·	1
L	1	12	32	25	18	·	·	·	·
L+	·	5	92	67	59	·	·	·	·
SL	1	2	61	194	96	3	·	·	·
ST	8	16	132	627	3230	139	228	62	299
SR	·	·	1	4	105	27	53	14	35
R	·	·	·	1	109	31	78	43	53
R+	1	·	·	·	84	28	105	78	66
HR	·	·	·	3	195	23	82	88	186

Notes

- All 52 checkpoints in this report were evaluated on the same 6,952 sequence-valid frozen validation anchors, so the top-level cross-series metrics are directly comparable.

- The original 2,224-image Series 1/2 evaluation is retained under historical_evaluation in the JSON and in explicitly historical PDF tables.

- Series 2 v2.0/v2.0b used their required HSV/CLAHE preprocessing; all other Series 1/2 models used raw BGR.

- MAE and within-degree counts can reward straight collapse. Use the class-balanced and turn-recall columns for selection, with MAE as supporting evidence.

- Offline MAE does not prove real-world reliability. The car can still fail on lighting, turns, driveways, road-edge ambiguity, speed, and sensor conditions.

- The shared-set comparison measures each complete checkpoint and training pipeline. It does not isolate architecture from differences in training data or augmentation.

- Historical checkpoint source mixes are not inferred from trainer defaults. A model should be called CARLA-assisted only when its saved command, run configuration, or source-count log proves that CARLA roots were used.

- The Series 3/4 report set is held out by the frozen 100-frame window split and requires three valid previous and three valid future frames without split crossings or timestamp gaps.

- The shared source dataset contains 81,237 curated real field frames across five manual-driving runs from July 2 through July 12, 2026; 6,952 anchors satisfy the common report contract.

- The hold-last baseline repeats the most recent previous target. It is a persistence reference for Series 4 history models, not a deployed controller behavior.

- Series 3 throttle labels are near-constant at full throttle, so this report evaluates steering only; throttle control remains disabled pending varied-throttle data.

- The July 13 field comparison selected v3.4 from the cases presented. A later supervised comparison found v4.0f viable but not clearly better than v3.4; v4.0p/v4.0r/v4.0a/v4.0c repeatedly held prior steering predictions and were not drivable. The v4.1 checkpoints are offline-only pending runtime integration and field testing.