Supplementary Material: Seeing is Believing? Evaluating VLMs' Physical Reasoning on Dynamic Objects

Puyin Li
Symbolic Systems
Stanford University
puyinli@stanford.edu

Ella Mao
Graduate School of Business
Stanford University
ellamao@stanford.edu

June 5, 2025

1 VLM Prompt

Below shows the exact prompt we used for VLM.

"Answer the following questions in this exact order"

"Diameter of the ball in pixel",
"Original position of the ball's center at t=0 in pixel",
"Change of pixels of the ball moved between t=0 s and t=0.2 s",
"Change of pixels of the ball moved between t=0.2 and t=0.4 s",
"Velocity at t=0.2 s (pixel/s)",
"Velocity at t=0.4 s (pixel/s)",
"Acceleration of the ball (pixel/s^2)",
"Assuming acceleration is 9.8 m/s^2 in reality, what is the diameter of the ball in the real world (meters)",

"Return a single line of comma-separated pairs like: Metric <k> => <number> where <k> is the index of the question (1-based) and <number> is a decimal-point value. Do not write the metric's real name, units, or any extra words. Pixel units unless the question says meters."

2 Metrics Distribution

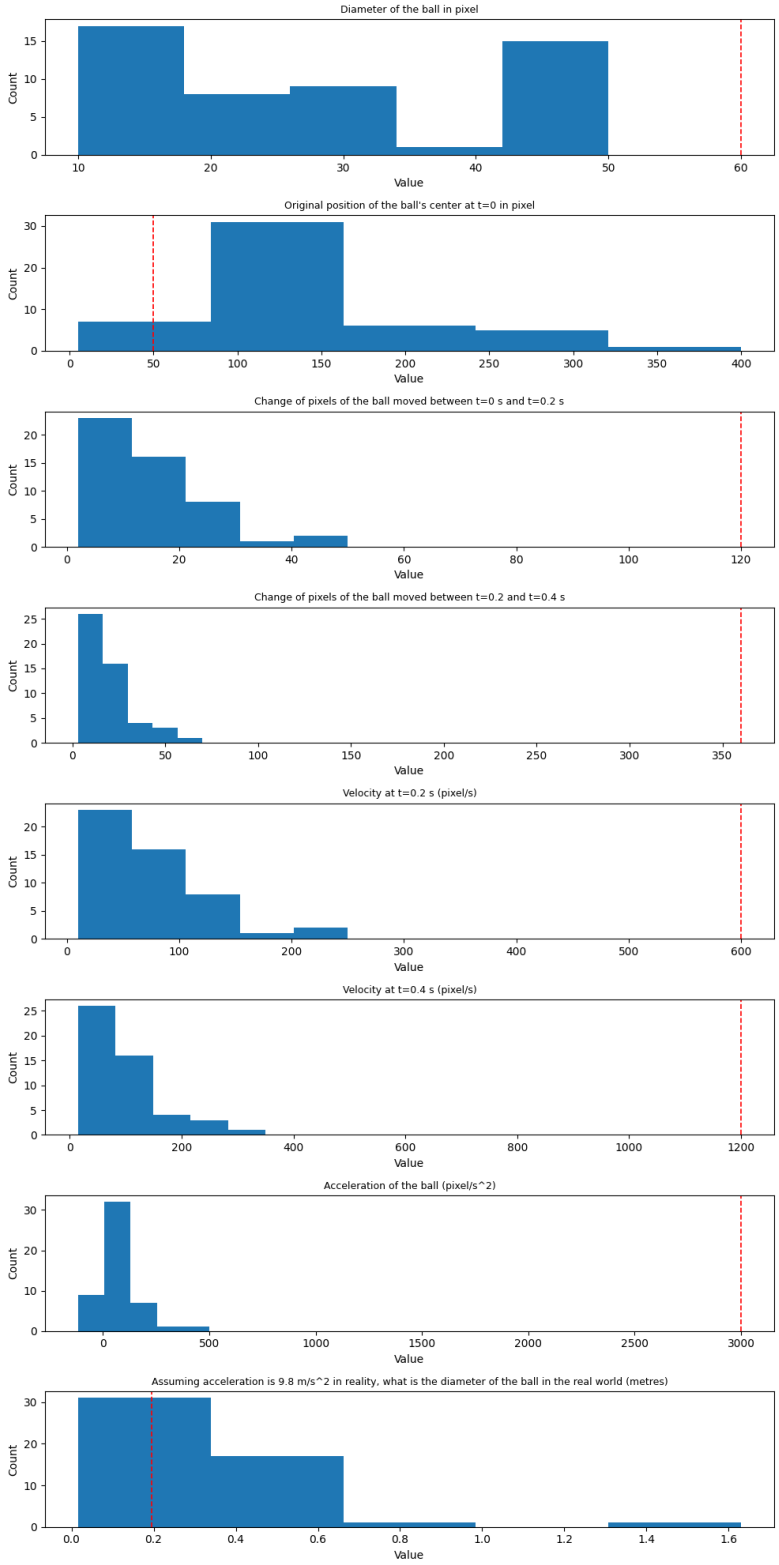

Figure 1: Histogram for S1. Red Ball

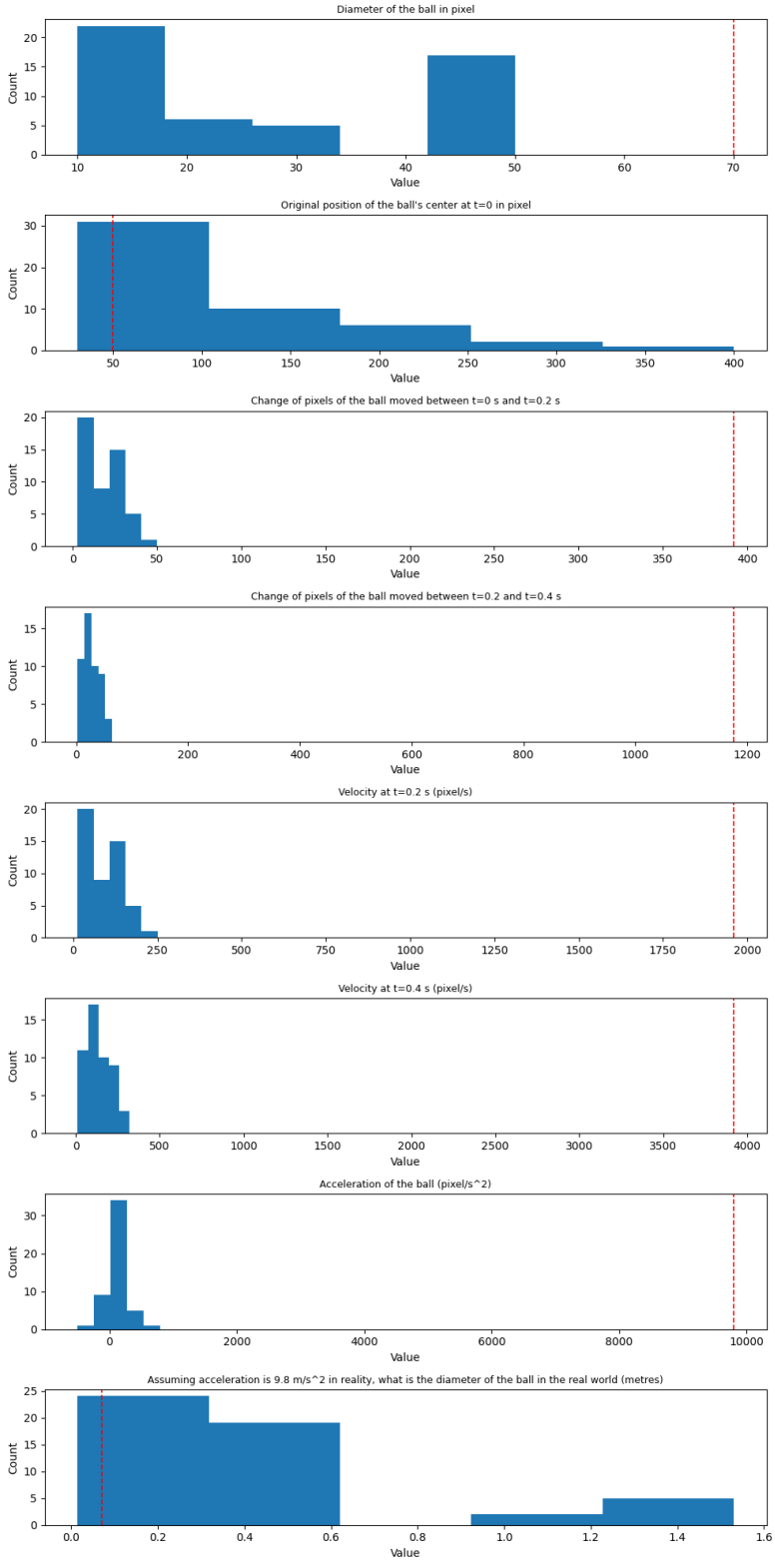

Figure 2: Histogram for S2. Green Ball

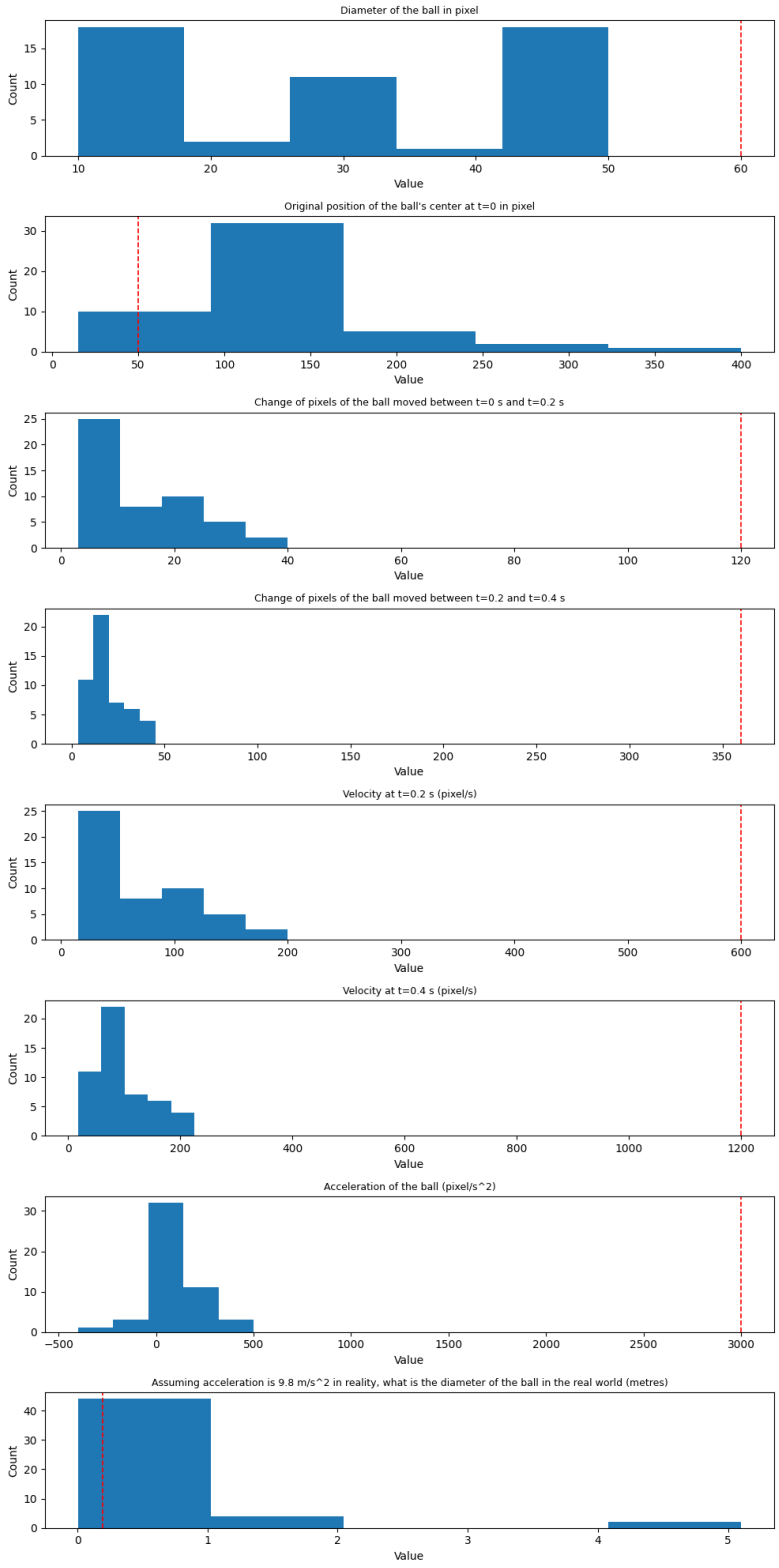

Figure 3: Histogram for S3. Background Color (Orange)

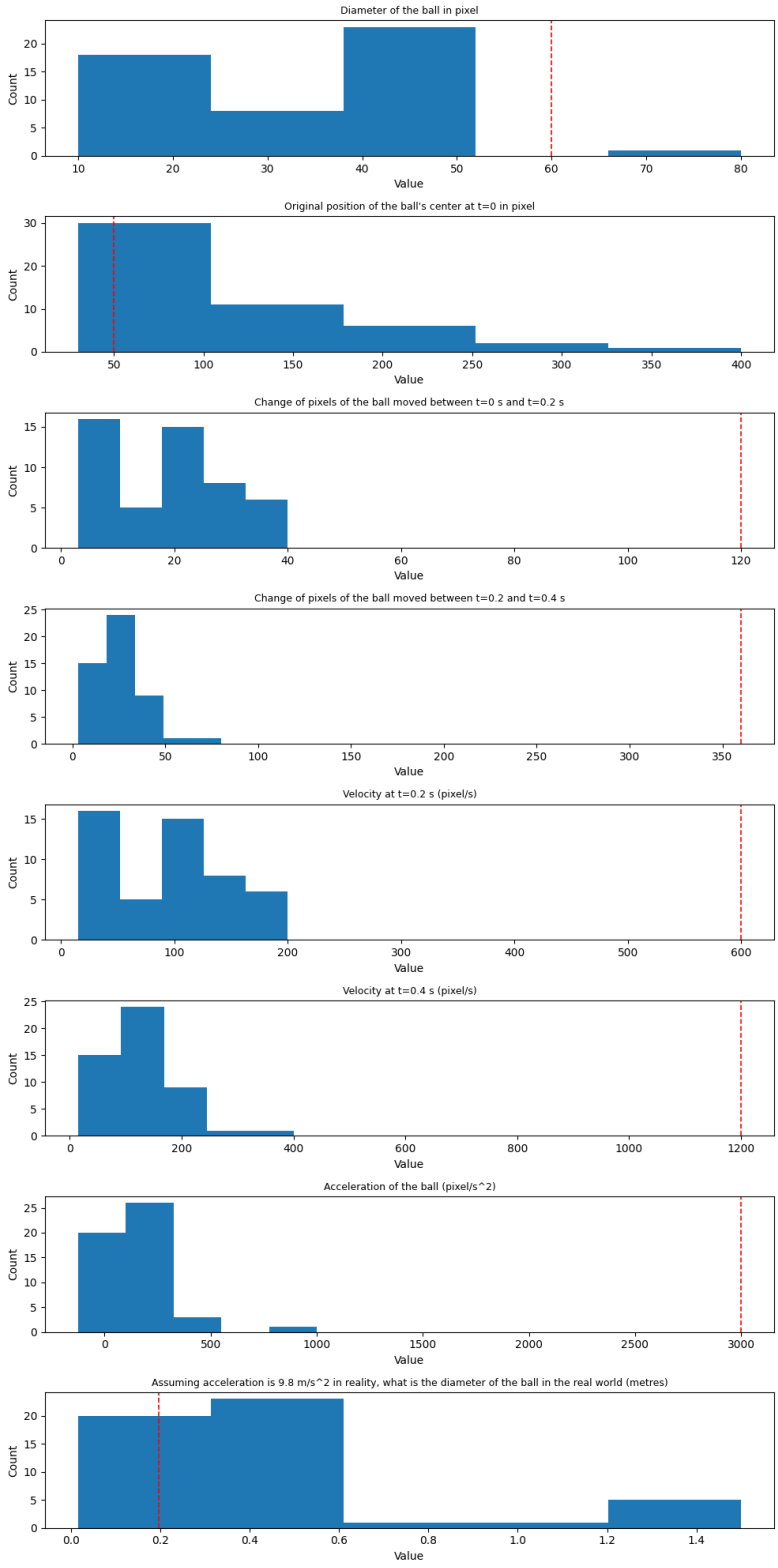

Figure 4: Histogram for S3. Background Color (Red)

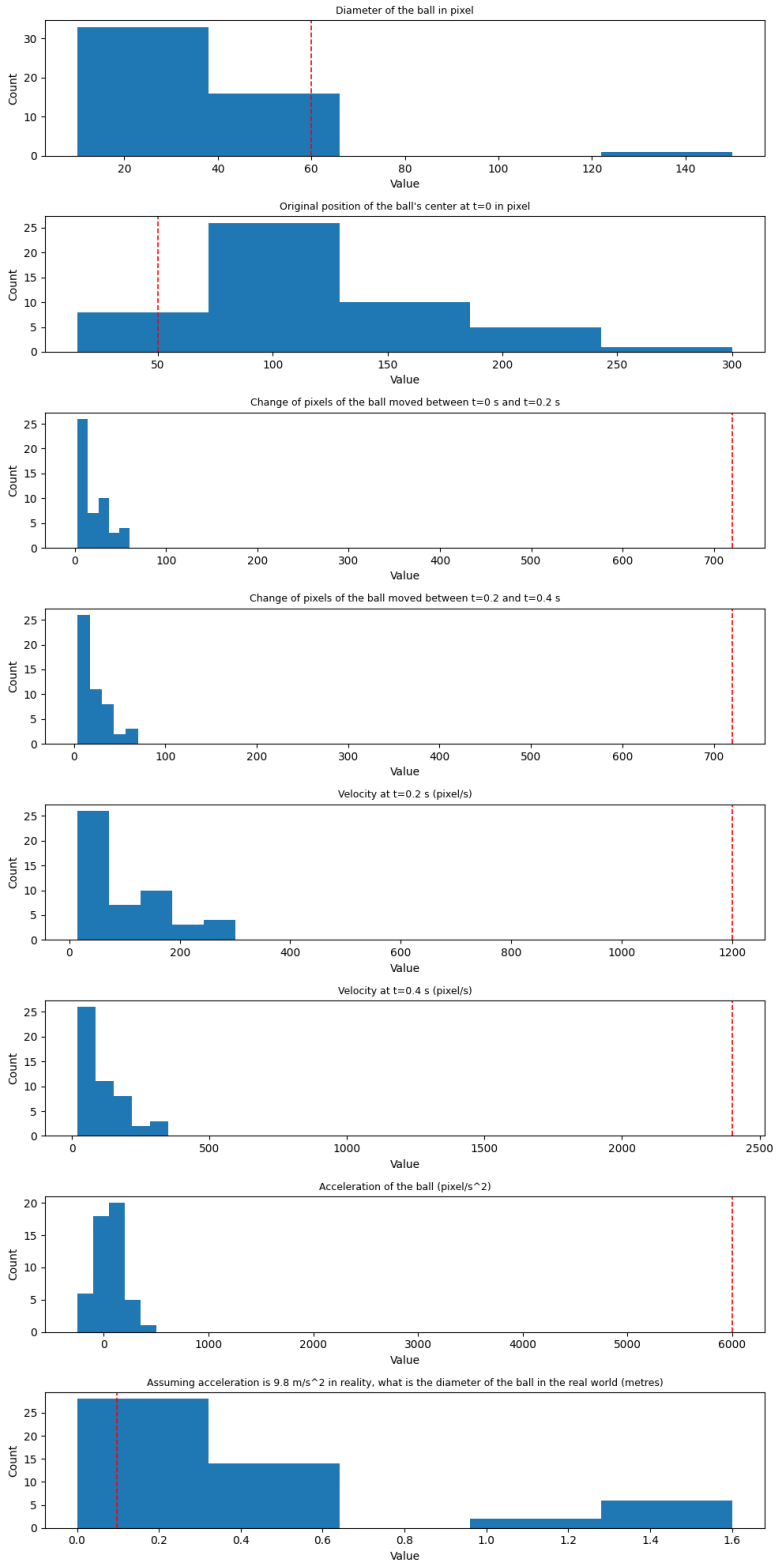

Figure 5: Histogram for S4. Double Gravity

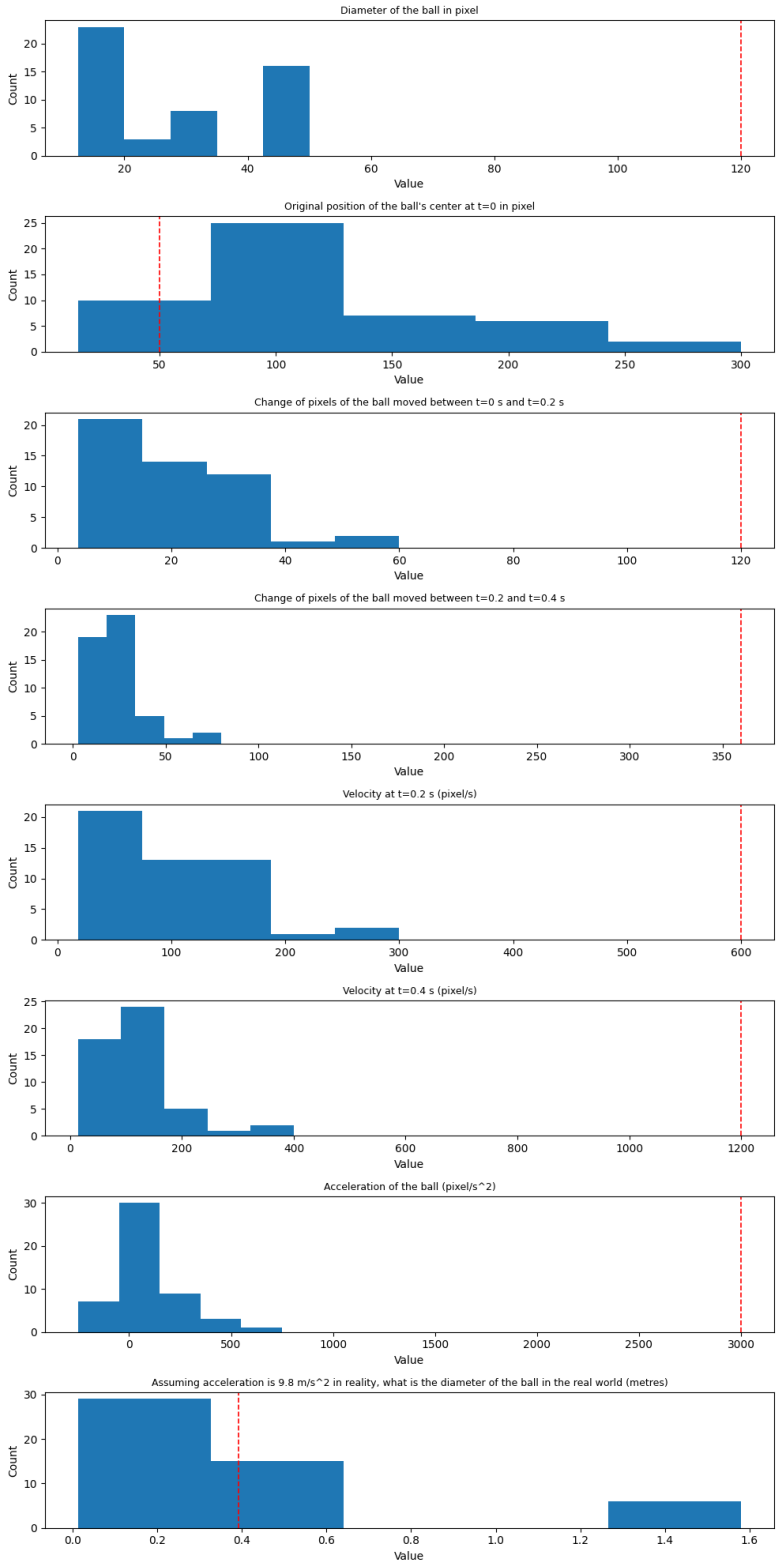

Figure 6: Histogram for S5. Double Size

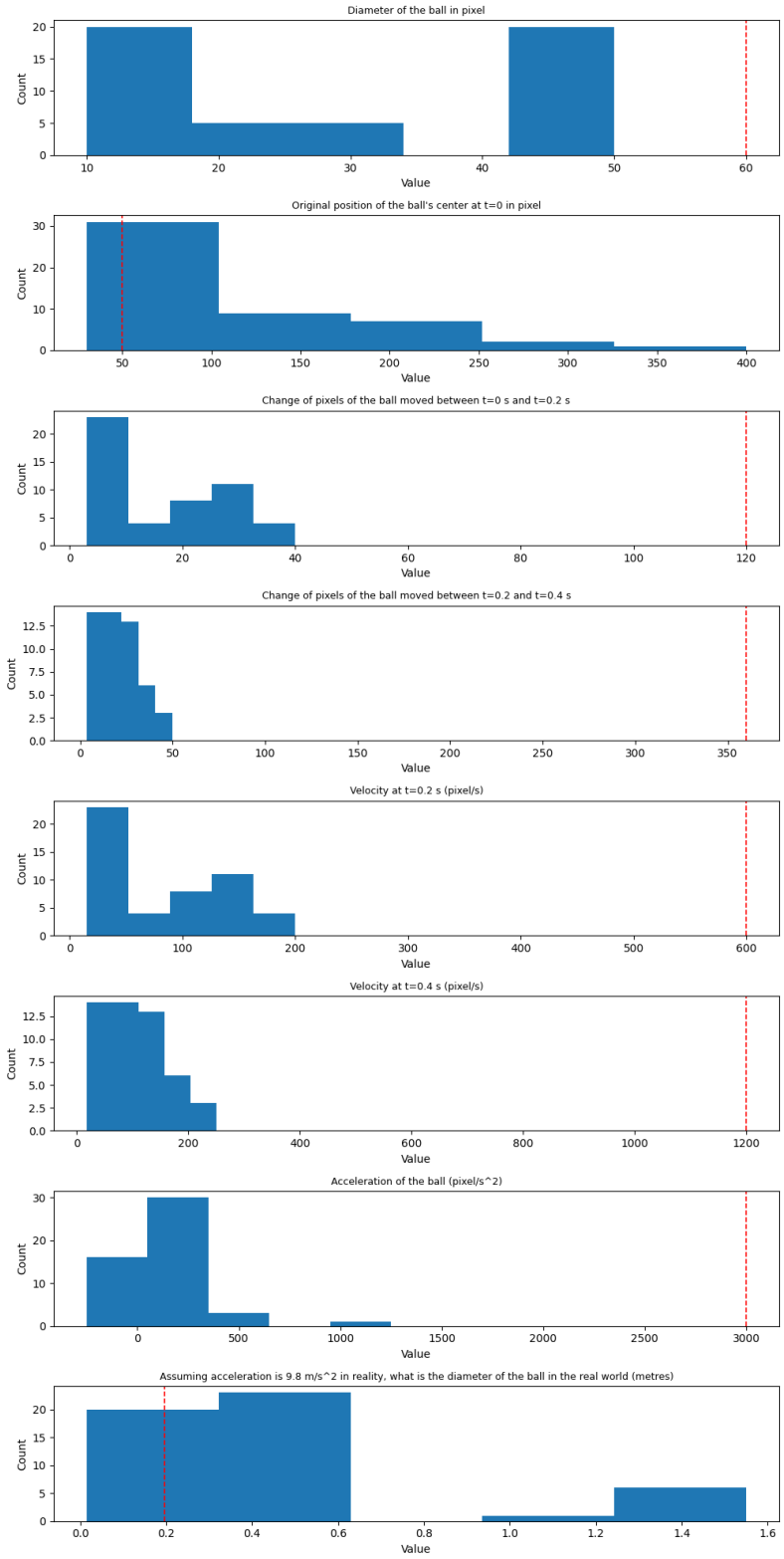

Figure 7: Histogram for S6. Tennis Ball

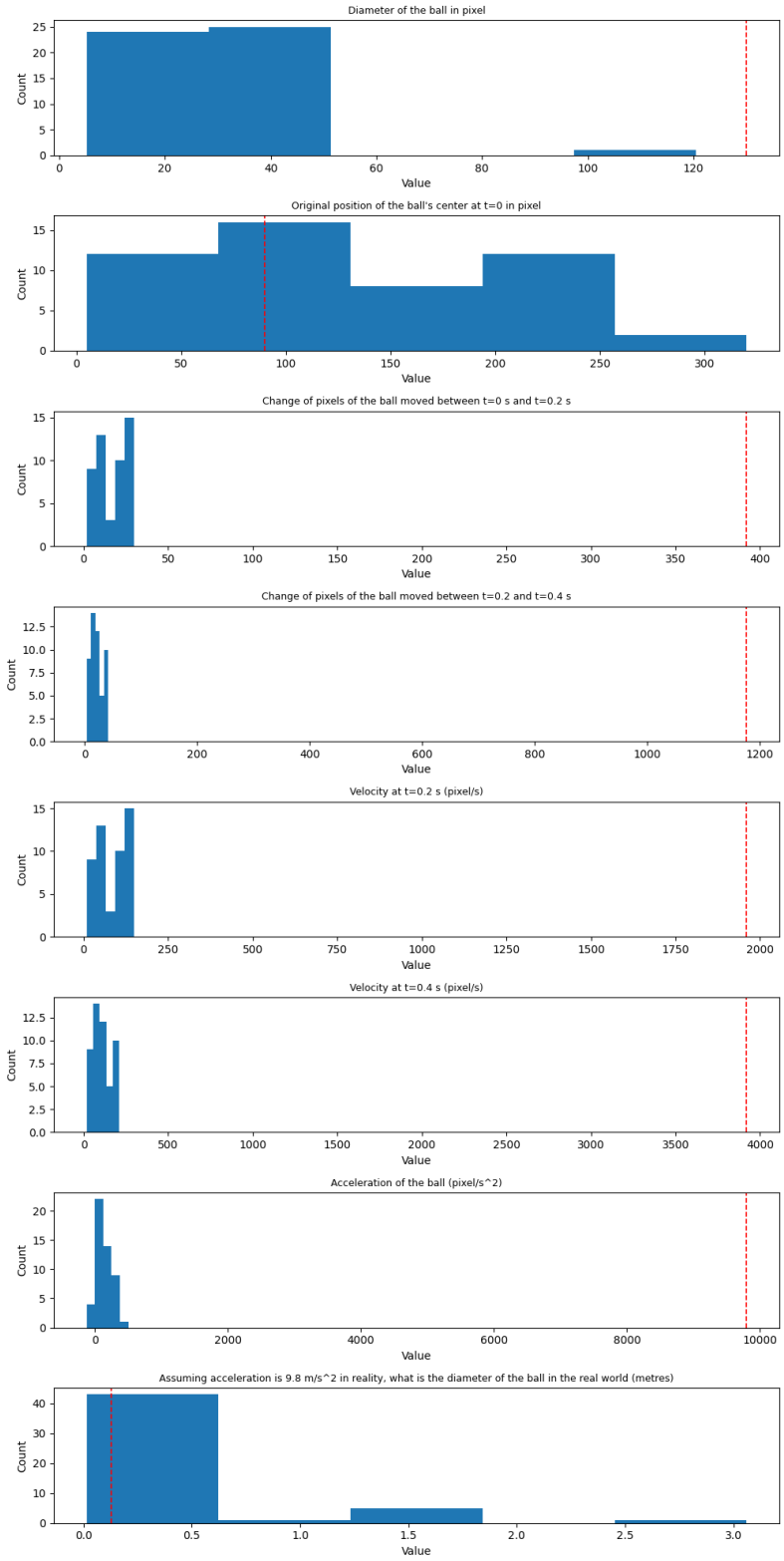

Figure 8: Histogram for S7. Weenie Toy

3 Values of Ground Truth and Results from VLM

Table 1: S1. Red Ball

Metric	Ground Truth	Mean	Std. Dev.
1	60.0	29.8	14.8
2	50.0	139.7	76.1
3	120.0	16.2	11.2
4	360.0	19.6	13.5
5	600.0	81.0	56.1
6	1200.0	97.9	67.3
7	3000.0	88.7	116.9
8	0.2	0.3	0.3

Table 2: S2. Green Ball

Metric	Ground Truth	Mean	Std. Dev.
1	60.0	31.2	16.2
2	50.0	124.5	78.8
3	120.0	17.7	11.2
4	360.0	21.7	11.9
5	600.0	88.7	56.2
6	1200.0	108.2	59.5
7	3000.0	119.6	221.4
8	0.2	0.5	0.4

Table 3: S3_1. Background Color (Orange)

Metric	Ground Truth	Mean	Std. Dev.
1	60.0	32.3	20.6
2	50.0	127.8	75.4
3	120.0	17.2	9.6
4	360.0	21.4	11.4
5	600.0	86.1	48.0
6	1200.0	107.0	57.0
7	3000.0	105.3	114.8
8	0.2	0.5	0.6

Table 4: S3_2. Background Color (Red)

Metric	Ground Truth	Mean	Std. Dev.
1	60.0	34.1	17.4
2	50.0	126.8	73.4
3	120.0	19.4	11.1
4	360.0	24.4	14.7
5	600.0	97.1	55.6
6	1200.0	122.1	73.5
7	3000.0	128.2	185.3
8	0.2	0.5	0.4

Table 5: S4. Double Gravity

Metric	Ground Truth	Mean	Std. Dev.
1	60.0	32.4	22.7
2	50.0	114.3	54.2
3	720.0	20.1	14.9
4	720.0	21.9	15.6
5	1200.0	100.7	74.6
6	2400.0	109.5	77.8
7	6000.0	46.7	134.4
8	0.1	0.5	0.5

Table 6: S5. Double Size

Metric	Ground Truth	Mean	Std. Dev.
1	120.0	29.2	15.4
2	50.0	116.7	63.0
3	120.0	18.8	13.1
4	360.0	22.6	15.8
5	600.0	95.4	65.3
6	1200.0	114.0	78.3
7	3000.0	108.8	182.8
8	0.4	0.4	0.4

Table 7: S6. Tennis Ball

Metric	Ground Truth	Mean	Std. Dev.
1	70.0	29.0	16.0
2	50.0	124.0	73.0
3	392.0	20.6	12.0
4	1176.0	26.0	15.4
5	1960.0	102.9	60.1
6	3920.0	129.7	77.5
7	9800.0	150.9	197.2
8	0.1	0.5	0.4

Table 8: S7. Weenie Toy

Metric	Ground Truth	Mean	Std. Dev.
1	130.0	32.3	20.6
2	90.0	127.8	75.4
3	392.0	17.2	9.6
4	1176.0	21.4	11.4
5	1960.0	86.1	48.0
6	3920.0	107.0	57.0
7	9800.0	105.3	114.8
8	0.1	0.5	0.6