A. Appendix

A.1. Supplementary Experimental

Model	1A	1B	2A	2B	2C	2D	3A	3B	3C	3D	4A
llava-1.5-7b	16.9	16.5	16.4	15.3	16.7	16.0	17.1	15.2	16.7	14.9	17.8
llava-1.6-vicuna-7b	18.0	20.0	18.5	18.3	19.4	20.2	18.9	19.4	17.9	19.7	20.5
internvl3-8b	13.2	7.3	44.5	12.4	36.3	10.9	42.2	13.1	32.7	11.8	35.5
llama-3.2-11b	34.0	24.3	42.0	30.1	39.0	27.7	34.7	22.6	33.8	24.7	57.5
gemma-3-4b	39.9	33.8	40.1	36.4	35.5	33.2	39.4	35.9	36.0	31.7	36.0
qwen-2.5-VL-3b	48.5	45.3	49.4	54.3	43.3	50.3	50.7	55.4	51.6	47.9	58.1
llava-onevision-qwen2-7b	49.7	50.5	48.4	54.6	41.5	52.6	43.9	51.2	46.6	54.1	54.5
ovis-8b	42.9	45.6	40.6	45.1	53.1	46.7	51.1	43.8	58.6	46.8	50.4
qwen-2.5-VL-7b	56.3	59.9	62.8	65.2	59.0	63.5	52.4	56.7	55.3	59.9	63.0
phi4-multimodal	71.4	51.6	71.1	56.2	69.3	55.7	71.0	60.7	68.4	59.1	61.6
Average	39.08	35.48	43.38	38.79	41.31	37.68	42.14	37.40	41.76	37.06	45.49
Model	4B	4C	4D	5A	6A	7A	8A	9A	9B	10A	10B
llava-1.5-7b	18.3	17.6	16.2	20.1	18.8	22.0	16.1	8.2	8.3	7.6	6.4
llava-1.6-vicuna-7b	16.0	18.2	17.8	28.3	26.9	24.5	17.8	14.2	8.4	9.1	5.1
internvl3-8b	14.5	32.3	14.8	65.5	60.7	52.3	45.0	19.1	8.9	17.1	8.9
llama-3.2-11b	22.0	40.9	26.5	28.9	32.8	26.7	30.5	12.2	11.3	13.5	14.0
gemma-3-4b	31.6	32.8	29.5	39.0	44.6	29.0	41.3	30.6	25.0	16.2	19.4
qwen-2.5-VL-3b	51.4	50.3	55.3	77.2	78.9	72.0	83.2	21.2	8.2	14.6	22.1
llava-onevision-qwen2-7b	52.8	50.1	56.8	66.7	73.9	65.5	75.4	42.4	15.0	41.7	23.4
ovis-8b	38.7	60.6	49.0	73.2	82.2	75.0	80.6	65.4	32.7	41.8	33.8
qwen-2.5-VL-7b	64.4	57.5	62.0	73.3	69.4	78.9	80.0	47.0	26.3	46.5	56.4
phi4-multimodal	65.2	64.7	60.8	57.6	49.0	47.7	41.0	21.7	14.6	23.2	18.6
Average	37.49	42.50	38.87	52.98	53.72	49.36	51.09	28.20	15.87	23.13	20.81
Model	11A	11B	12A	12B	13A	13B	14A	14B	15A	15B	**Avg ↑**
llava-1.5-7b	0.5	1.0	8.3	8.3	7.0	7.1	0.0	4.2	0.0	2.7	**11.82**
llava-1.6-vicuna-7b	7.1	3.7	8.3	8.2	4.9	6.1	0.1	7.2	2.1	9.7	**14.20**
internvl3-8b	9.6	6.0	17.9	9.6	16.5	8.5	9.9	4.1	9.4	4.6	**21.72**
llama-3.2-11b	6.6	21.9	13.3	12.3	13.4	16.1	6.6	21.8	5.9	20.5	**24.00**
gemma-3-4b	8.1	9.1	29.5	24.6	16.3	18.6	7.9	6.7	9.2	9.0	**27.37**
qwen-2.5-VL-3b	3.8	0.3	13.1	7.4	16.7	14.1	6.5	0.3	0.4	0.5	**36.01**
llava-onevision-qwen2-7b	41.8	21.2	43.2	17.2	40.3	24.9	40.1	20.1	39.7	22.2	**44.44**
ovis-8b	42.4	25.5	66.4	34.3	40.3	33.2	38.4	26.4	41.9	26.2	**47.90**
qwen-2.5-VL-7b	42.2	14.8	43.3	28.8	38.8	36.8	33.1	16.5	32.7	14.0	**50.52**
phi4-multimodal	6.5	5.2	22.2	17.1	22.6	17.2	7.3	5.1	6.1	5.0	**39.83**
Average	16.86	10.87	26.55	16.78	21.68	18.26	14.99	11.24	14.74	11.44	**31.78**

Table 5. Complete accuracies over all 32 test cases, which case coding rule can be observed in the Sec. 3. Each code's evaluation or score is being represented by 1000 samples evenly distributed across the respective count range.

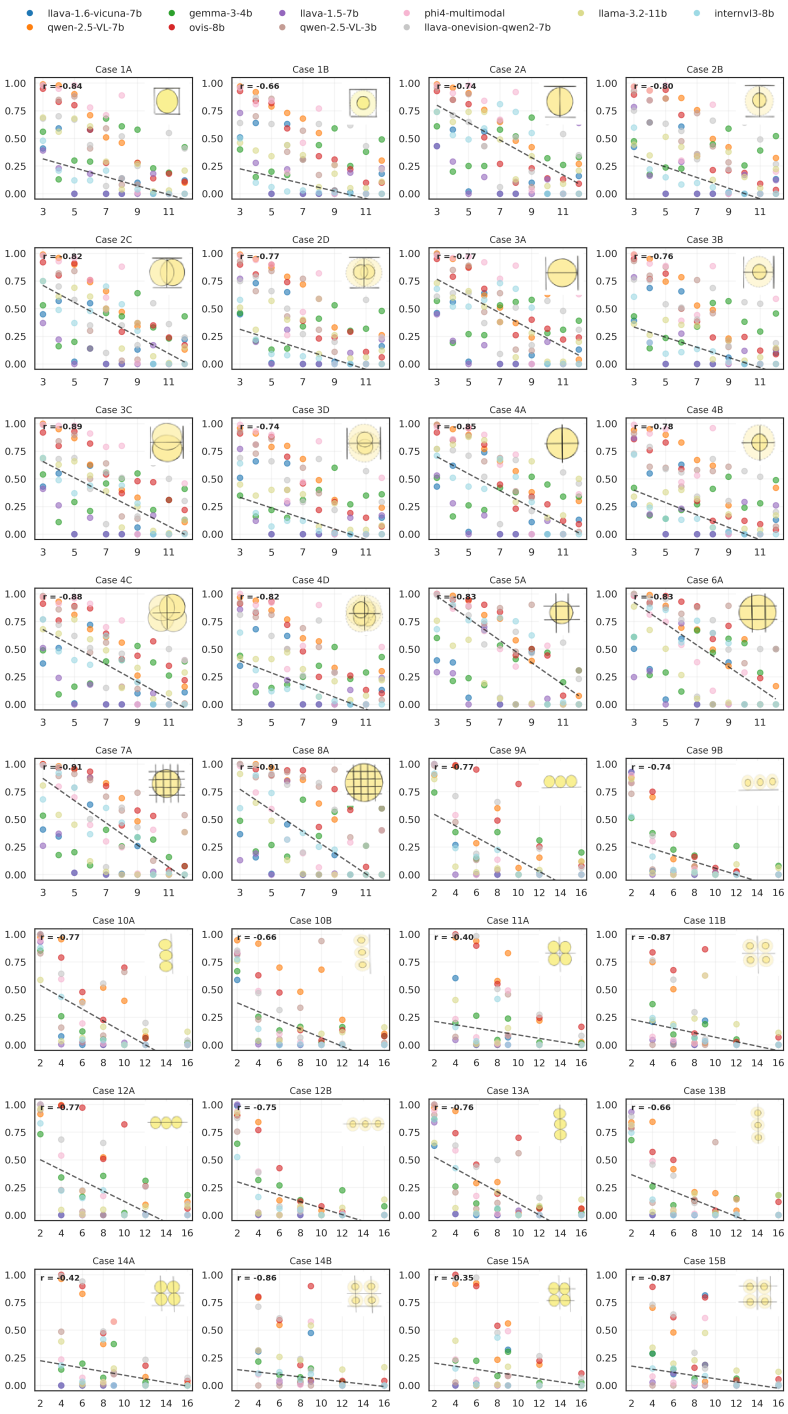

Figure 7. **Accuracy vs. Ground Truth Count.** *Left:* Scatter plots show a consistent negative correlation ($r \approx -0.78$) between count magnitude and accuracy across diverse geometric cases. *Right:* Detailed breakdown for LLaVA-1.5 and Qwen2.5-VL reveals specific "blind spots" (circled in red) where models achieve 0% accuracy for specific numbers (e.g., 7, 11), evidencing strong linguistic priors.

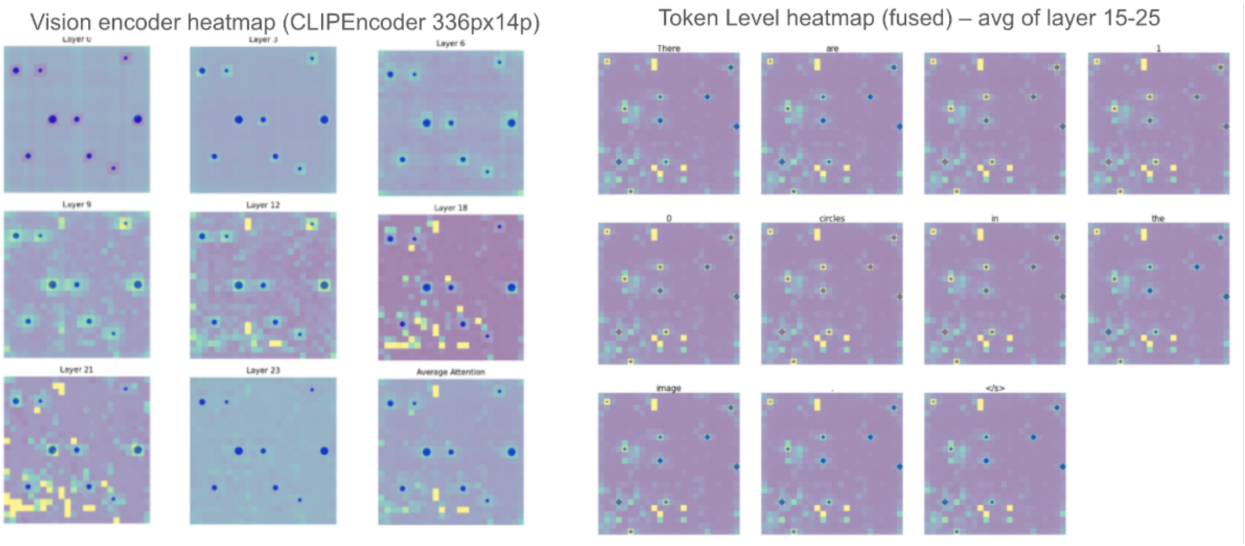

Figure 8. **Fused Token Heatmaps.** Averaged over deep LLM layers (15–25), the attention becomes diffuse and misaligned. The signal "washes out," failing to retain the instance-level separation required for accurate counting.