Appendix

A. Prompts used for recaptioning

For gathering our main set of 1B image captions, we used the following prompt.

> *"The image presented came from a web page called:* `page title` *and had the alt-text:* `alt-text`*. Please describe what is in the image using the alt-text and the page title as a guide to ground your response. For example, if the alt-text contains a specific brand name, use that brand name in your output. Please be descriptive but concise. DO NOT make things up. If you can't tell something with certainty in the image, simply don't say anything about it."*

For gathering our 100M ablation set of very brief captions, we used the following prompt.

> *"The image presented came from a web page called:* `page title` *and had the alt-text:* `alt-text`*. Please very briefly describe what is in the image using the alt-text and the page title as a guide to ground your response. For example, if the alt-text contains a specific brand name, use that brand name in your output. Please describe what is in the image but be extremely concise in your response. I want to emphasize how important it is to be VERY brief. DO NOT make things up. If you can't tell something with certainty in the image, simply don't say anything about it.""*

For gathering our 100M ablation set of captions without conditioning on alt-text or page title, we used the following prompt.

> *"Please describe what is in the image. Please be descriptive but concise. DO NOT make things up. If you can't tell something with certainty in the image, simply don't say anything about it."*

B. Image to text retrieval results

All prior results were given to as text $\rightarrow$ image recall. Following prior work, we also evaluated image $\rightarrow$ text recall.

	COCO	DOCCI-test	DOCCI-full
CLIP [26]	58.4	55.6	41.9
CoCa [35]	63.8	56.7	51.8
BLIP [17]	–	54.7	–
X-VLM [37]	71.6	–	–
VeCLIP [16]	67.8	–	–
Long-CLIP [40]	57.6	38.6	–
MATE [12]	–	62.9	–
TIPS [2]	74.0	57.2	–
Ours	76.2	95.9	91.3

Table 9. **Caption retrieval based on image.** Same as Table 3 on COCO and DOCCI datasets, but on image $\rightarrow$ text retrieval. Note that COCO has multiple text labels for each image, making this task easier than text $\rightarrow$ image retrieval. .

C. Method details for caption statistics.

The plots from Figs. 4 and 5 are based on 1,000 random samples from each dataset (WebLI `alt-text` vs. Gemini Flash 1.5 captions). The plots were generated via Seaborn's displot performing a kernel density estimate, setting cut=0 to avoid putting probability mass on negative caption lengths. The log-likelihood from Fig. 5 was obtained by scoring log-likelihood of `alt-text` vs. Gemini Flash 1.5 captions on a random sample of 1,000 captions each. The model used for scoring was Gemini Pro 1.5 [30], i.e., a larger and higher-quality model than the model used to generate captions. This conforms to the common practice in language modeling of using a large model to score text generation from a smaller model. The log-likelihood of a sequence of tokens $t = (t_1, ..., t_n)$ according to a language model parameterized by θ is calculated as follows:

$$\log p(t|\theta) = \sum_{i=1}^{n} \log p(t_i|t_1, t_2, ..., t_{i-1}; \theta)$$

For length statistics of ablation captions, see Fig. 7.

D. Detailed results on ARO evaluation

We report our performance across the fine-grained splits of ARO in Table 10. We achieve high performance across most subsets of ARO. Notably, our performance on left/right is lower than other spatial relations, in particular much lower than similar relations such as above/below. We suspect there could be an issue with the ground truth labels for this subset. We had a human visualize and mark the correctness of a random sample of 100 left/right labels and determined that the ground truth for 43 out of 100 were either incorrect or ambiguous. Further investigation is warranted.

	CLIP	NegCLIP	CLIP-FT	XVLM	BLIP	Flava	Ours
Spatial Relationships	0.56	0.66	0.57	0.74	0.66	0.34	0.83
above	0.48	0.60	0.54	0.80	0.64	0.55	0.84
at	0.59	0.93	0.71	0.72	0.49	0.15	0.93
behind	0.56	0.29	0.34	0.82	0.77	0.28	0.80
below	0.56	0.46	0.48	0.74	0.69	0.44	0.78
beneath	0.80	0.70	0.70	0.80	0.70	0.40	0.80
in	0.63	0.89	0.63	0.73	0.72	0.09	0.99
in front of	0.54	0.75	0.70	0.66	0.55	0.78	0.85
inside	0.50	0.91	0.67	0.69	0.72	0.12	0.93
on	0.52	0.86	0.58	0.86	0.76	0.12	0.98
on top of	0.43	0.75	0.58	0.85	0.79	0.19	0.98
to the left of	0.49	0.50	0.50	0.52	0.51	0.50	0.59
to the right of	0.49	0.50	0.50	0.52	0.49	0.51	0.61
under	0.64	0.43	0.54	0.86	0.73	0.27	0.69
Verbs	0.61	0.86	0.66	0.73	0.56	0.2	0.95
carrying	0.33	0.83	0.75	0.75	0.67	0.08	1.0
covered by	0.47	0.36	0.36	0.61	0.58	0.56	0.97
covered in	0.79	0.50	0.50	0.14	0.29	0.14	0.50
covered with	0.56	0.56	0.50	0.56	0.50	0.19	0.88
covering	0.39	0.58	0.45	0.67	0.55	0.06	0.97
cutting	0.75	0.83	0.83	0.67	0.25	0.00	1.0
eating	0.57	1.00	0.67	0.62	0.52	0.00	1.0
feeding	0.90	0.80	0.80	0.60	0.30	0.20	0.9
grazing on	0.10	0.90	0.30	0.60	0.40	0.50	1.0
hanging on	0.79	1.00	0.93	0.93	0.79	0.00	1.0
holding	0.58	0.97	0.79	0.67	0.44	0.27	1.0
leaning on	0.67	1.00	1.00	0.75	0.58	0.08	1.0
looking at	0.84	1.00	0.68	0.68	0.55	0.26	0.87
lying in	0.47	1.00	0.60	0.87	0.67	0.00	1.0
lying on	0.60	0.88	0.50	0.93	0.75	0.17	1.0
parked on	0.67	0.86	0.38	0.76	0.86	0.00	1.0
reflected in	0.64	0.71	0.57	0.50	0.43	0.43	0.86
resting on	0.38	0.85	0.23	0.92	0.54	0.15	1.0
riding	0.71	0.98	0.78	0.82	0.41	0.02	1.0
sitting at	0.62	1.00	0.88	0.88	0.46	0.00	1.0
sitting in	0.57	0.96	0.78	0.87	0.83	0.30	0.96
sitting on	0.58	0.97	0.78	0.94	0.73	0.14	0.99
sitting on top of	0.50	0.90	0.80	1.00	0.80	0.10	1.0
standing by	0.67	0.92	0.67	0.83	0.67	0.67	0.92
standing in	0.73	0.98	0.69	0.69	0.49	0.05	1.0
standing on	0.60	1.00	0.63	0.83	0.73	0.06	1.0
surrounded by	0.64	0.71	0.64	0.71	0.64	0.79	0.93
using	0.84	1.00	1.00	0.68	0.58	0.00	1.0
walking in	0.70	1.00	0.70	0.60	0.50	0.00	1.0
walking on	0.79	1.00	0.79	0.84	0.42	0.05	1.0
watching	0.45	0.55	0.27	0.59	0.68	0.36	0.82
wearing	0.47	0.99	0.88	0.68	0.48	0.64	1.0
Overall	0.59	0.80	0.64	0.73	0.59	0.24	0.92

Table 10. **Fine-grained results on ARO relations.**. Comparison results from [36].

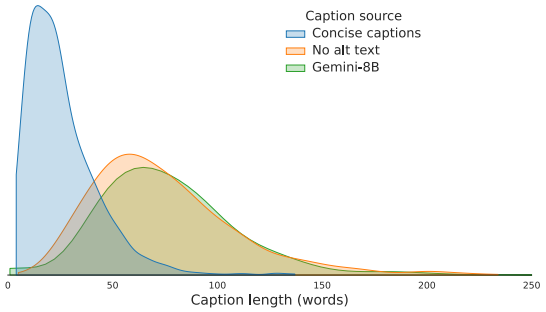

Figure 7. **Caption length comparison for ablation captions.** Concise captions are significantly shorter than our default captions. Using Gemini-8B and removing alt-text conditioning have little impact on the length.