

Technical Appendices

A Sparsity of Attention on Image Tokens

We compute the proportion of attention assigned to the top-k image tokens relative to the total attention across all image tokens

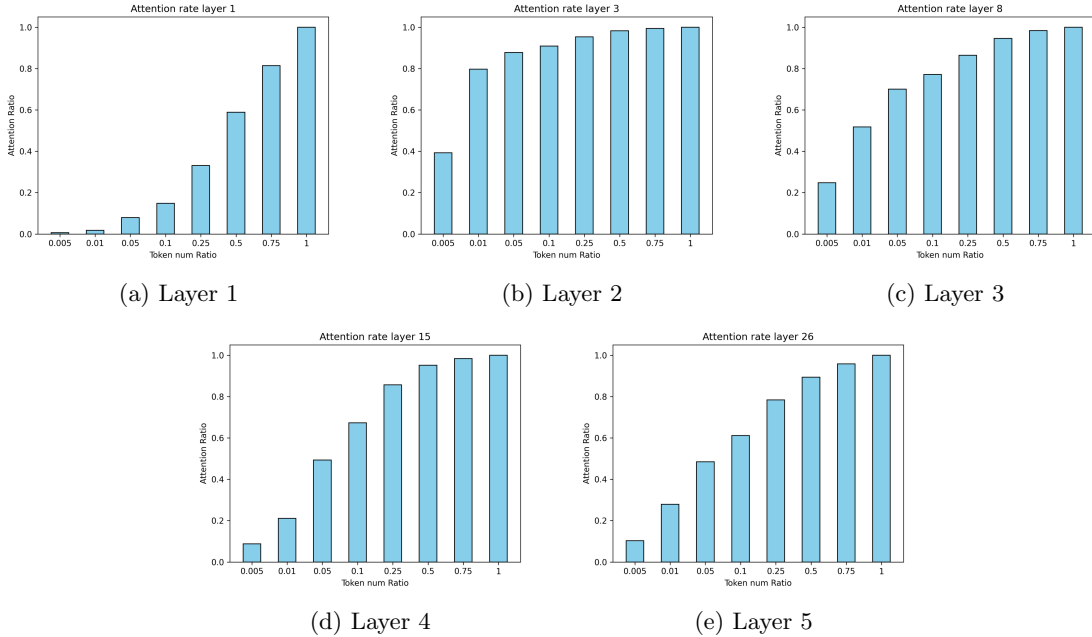


Figure 1: Top token positions selected by attention at different layers.

We observe that in early layers, LVLMs assign nearly uniform weights across all image tokens, making it difficult to distinguish important tokens. However, from mid to late layers, only 25% of the image tokens are sufficient to preserve over 60% of the original attention mass. This further demonstrates the redundancy of image tokens and the effectiveness of our proposed method.

B Effect of Positional Encoding on Attention

We analyze whether positional encodings influence attention patterns by selecting the top 25% attention-weighted tokens per layer and visualizing their spatial positions.

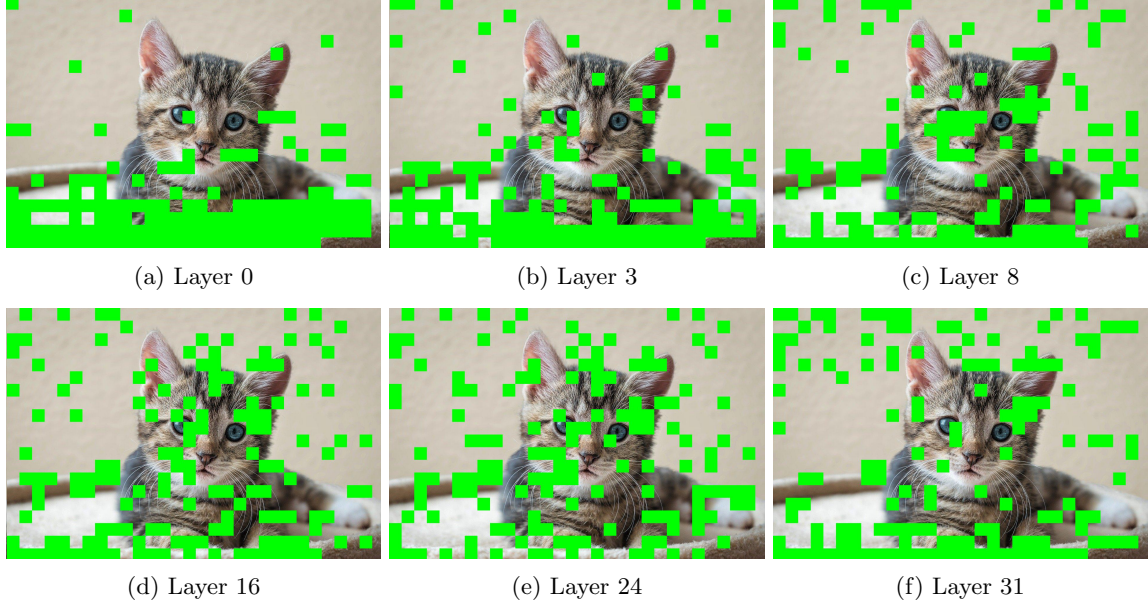


Figure 2: attention map across layers.

It is evident that in early layers, attention is highly influenced by positional encoding. Combined with our second experiment, we hypothesize that the model initially lacks information on the target location and thus distributes attention uniformly. In deeper layers, attention becomes more focused and less dependent on position.

To test this, we flipped the image of a cat vertically and observed the attention pattern:

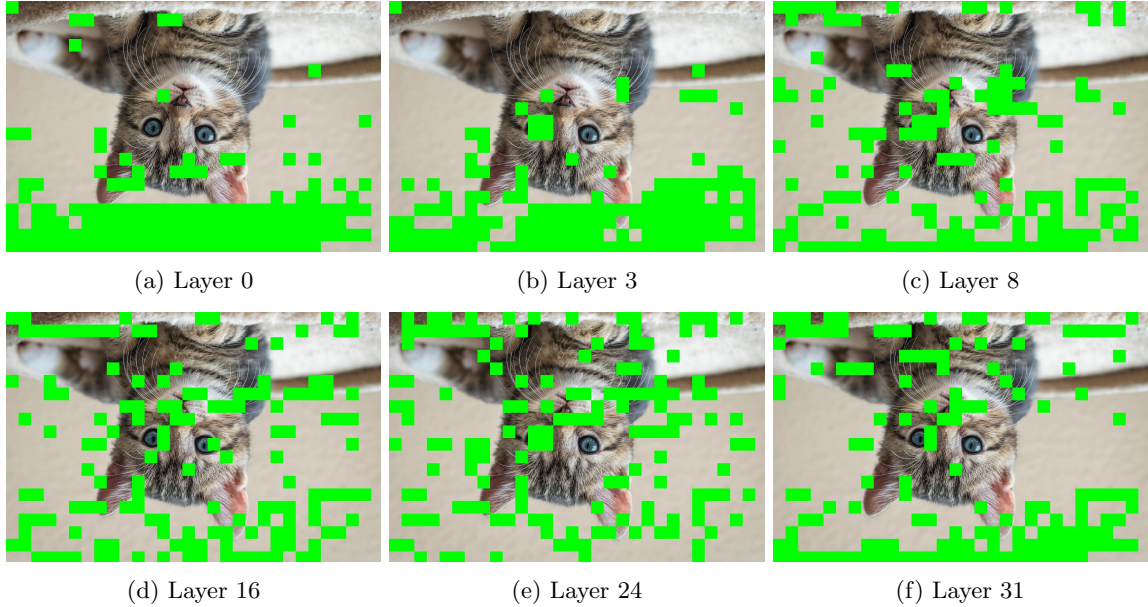


Figure 3: attention map after vertical flipping.

This suggests that positional bias is strong in early layers but diminishes in later ones.

C Average Value Norms of Important Tokens

We compare the average value norm (L2 norm) of different types of tokens (SYS prompt, image tokens, text tokens) across layers for the top 25% important tokens versus all tokens.

Layer	SYS Prompt		Image Tokens		Text Tokens	
	Top 25%	All	Top 25%	All	Top 25%	All
3	1.02	2.80	2.74	2.69	3.00	3.12
8	1.35	3.12	3.61	4.01	3.75	3.74
14	1.66	4.19	4.96	4.78	4.70	5.12
16	1.90	4.16	4.21	4.12	5.15	5.23
18	1.62	4.41	5.65	5.48	5.18	5.17
20	1.12	3.98	5.99	5.86	4.45	4.73
31	3.86	5.09	4.91	4.08	4.80	4.93

Table 1: Comparison of average value norms across token types and layers.

It can be observed that the value norms of important SYS prompt tokens are significantly lower, while the value norms of the other two token types show no clear distinction.