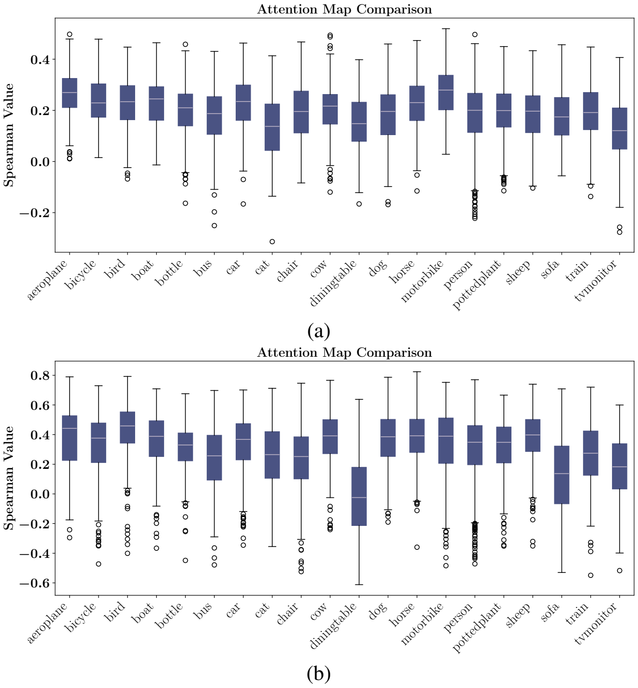
A. Appendix

A.1. Dataset

	Finetuning train Dataset	Finetuning val Dataset	Test Dataset
aeroplane	386	396	217
bicycle	340	300	196
bird	484	510	276
boat	424	403	231
bottle	600	603	356
bus	220	228	236
car	1,038	996	456
cat	481	510	285
chair	1,239	1,268	548
cow	245	242	284
dining table	314	318	167
dog	638	661	298
horse	304	295	204
motorbike	299	298	204
person	4,406	4,517	1,732
potted plant	442	438	321
sheep	392	384	306
sofa	319	313	208
train	254	261	188
tv/monitor	343	351	197
All	13,168	13,292	6,910

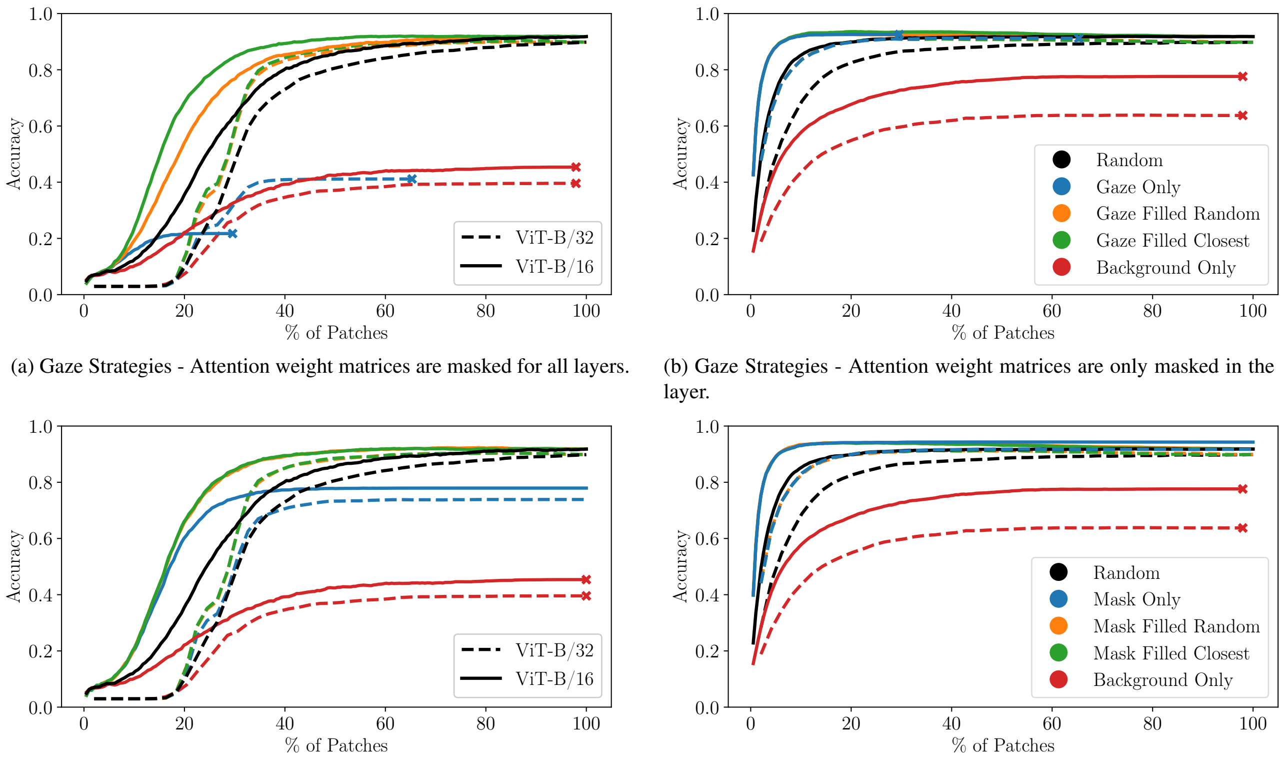
Table A.1. **Dataset overview:** Number of images used for finetuning and gaze token analysis.



(a)

(b)

Figure A.1. **Attention map comparison:** Spearman values for fixation attention maps based on fixation numbers and Transformer attention maps per category for (a) patch size 16 and (b) patch size 32.



(a) Gaze Strategies - Attention weight matrices are masked for all layers.

(b) Gaze Strategies - Attention weight matrices are only masked in the last layer.

(c) Mask Strategies - Attention weight matrices are masked for all layers.

(d) Mask Strategies - Attention weight matrices are only masked in the last layer.

Figure A.2. **Sampling strategy comparison:** Gaze-based sampling strategies compared to mask-based sampling. Sampling strategies perform very similarly, except for Mask Only when all layers are masked. The performance gain over Gaze Only can be explained by the perfect object coverage, leading to more available tokens.