A APPENDIX

A DATASET SAMPLES

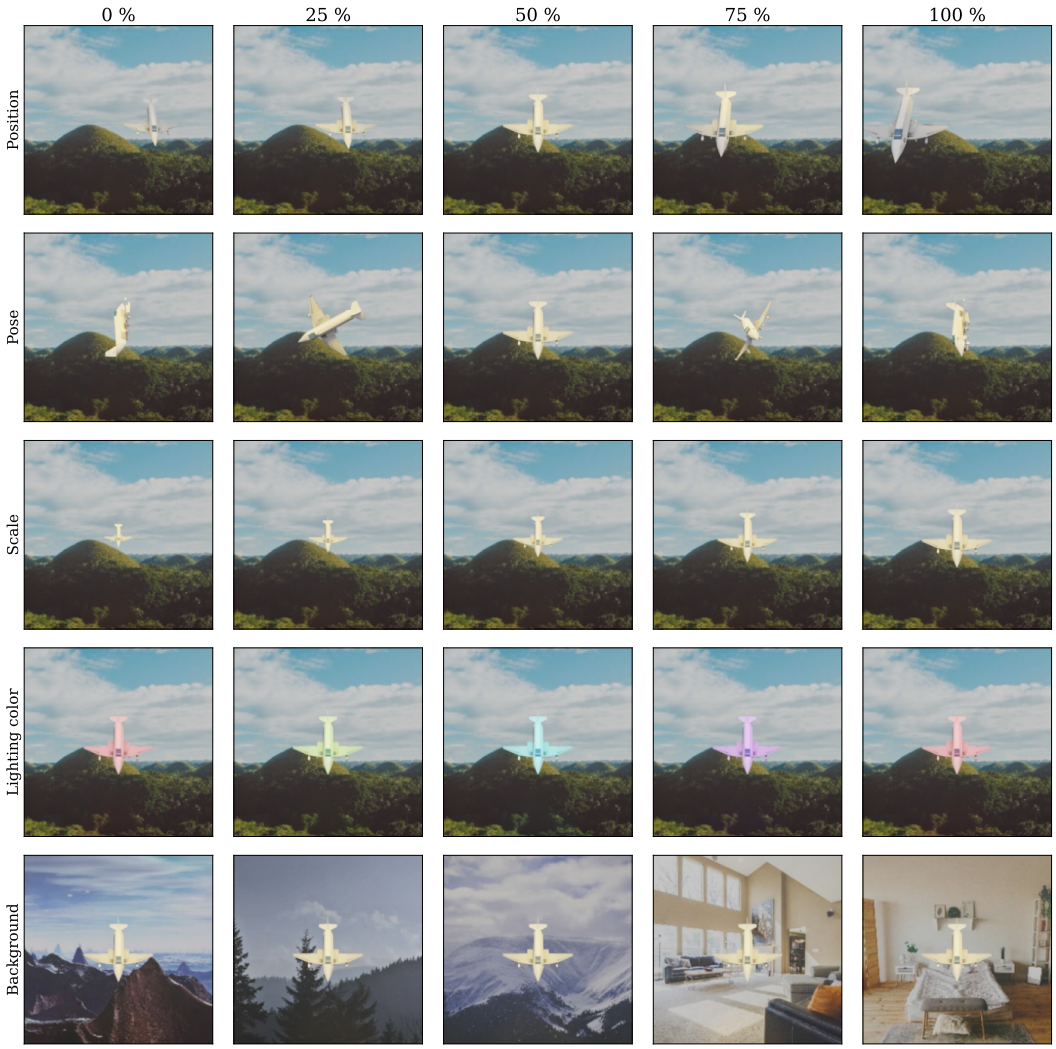


Figure A1: Examples from the dataset illustrating the different factors of variation.

B AGGREGATED PERFORMANCES FOR LINEAR VS FINETUNING AND 21K VS 1K

As can be seen in figure A11, finetuning usually leads to lower drops in performance with high variability rates during training. However linear evaluation is more robust when diversity was not encountered during training. Pretraining on ImageNet21k always improves robustness compared to ImageNet-1k pretraining, whether in finetuning or in linear evaluation. It is worth noting that for translation robustness, all settings exhibit similar performance, and finetuning only benefits ImageNet-1k pretraining.

B.1 CLIP ZERO SHOT CLASSIFICATION

We also evaluate CLIP's robustness using zero-shot classification. We assess both the standard Open AI CLIP model as well as CLIP trained on 2B LAION images. We prompt the model using "a photo of a []". CLIP with LAION-2B accuracies are 31.9% for canonical, 15.9% when pose varies, 18.2% when scale varies, 31.9% when lighting color, and 26.8% background varies. CLIP with trained on

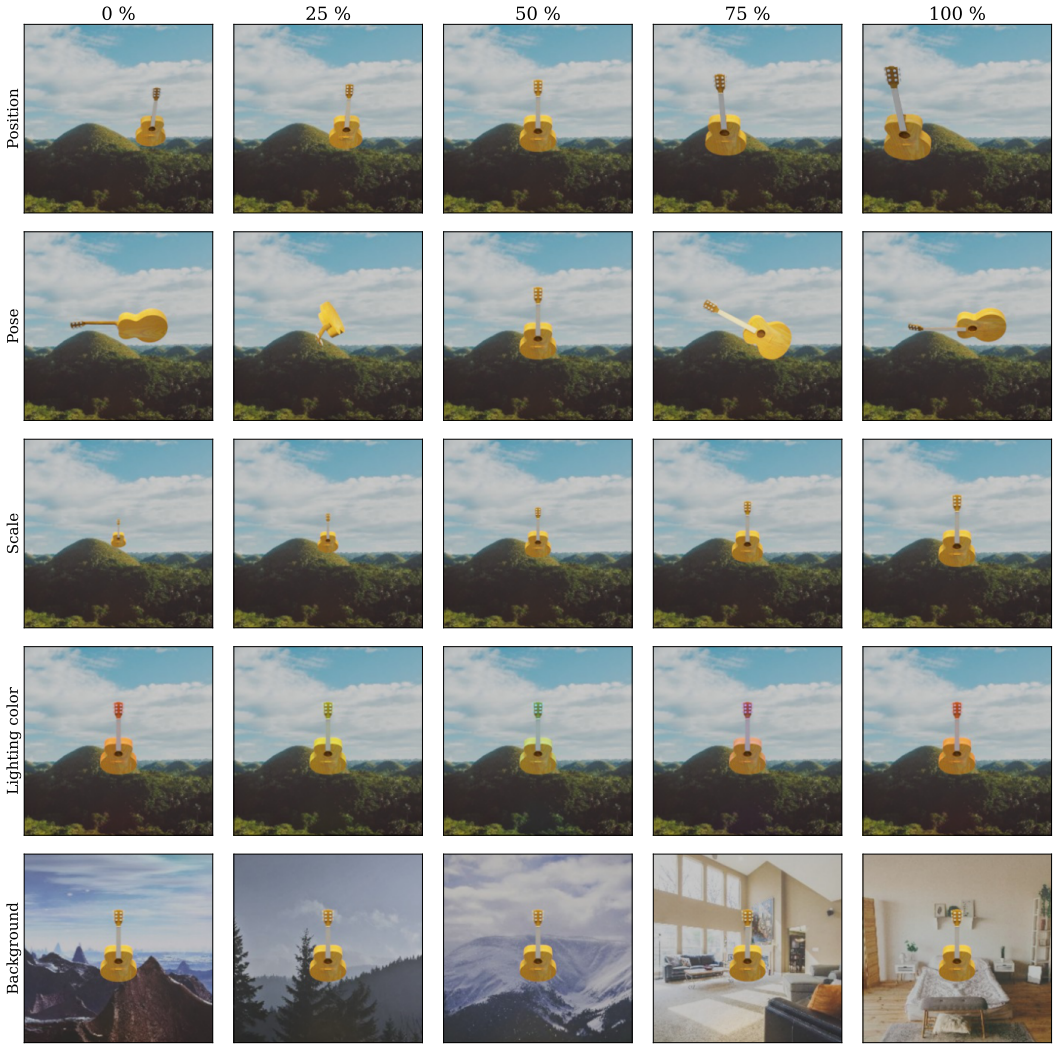


Figure A2: Examples from the dataset illustrating the different factors of variation.

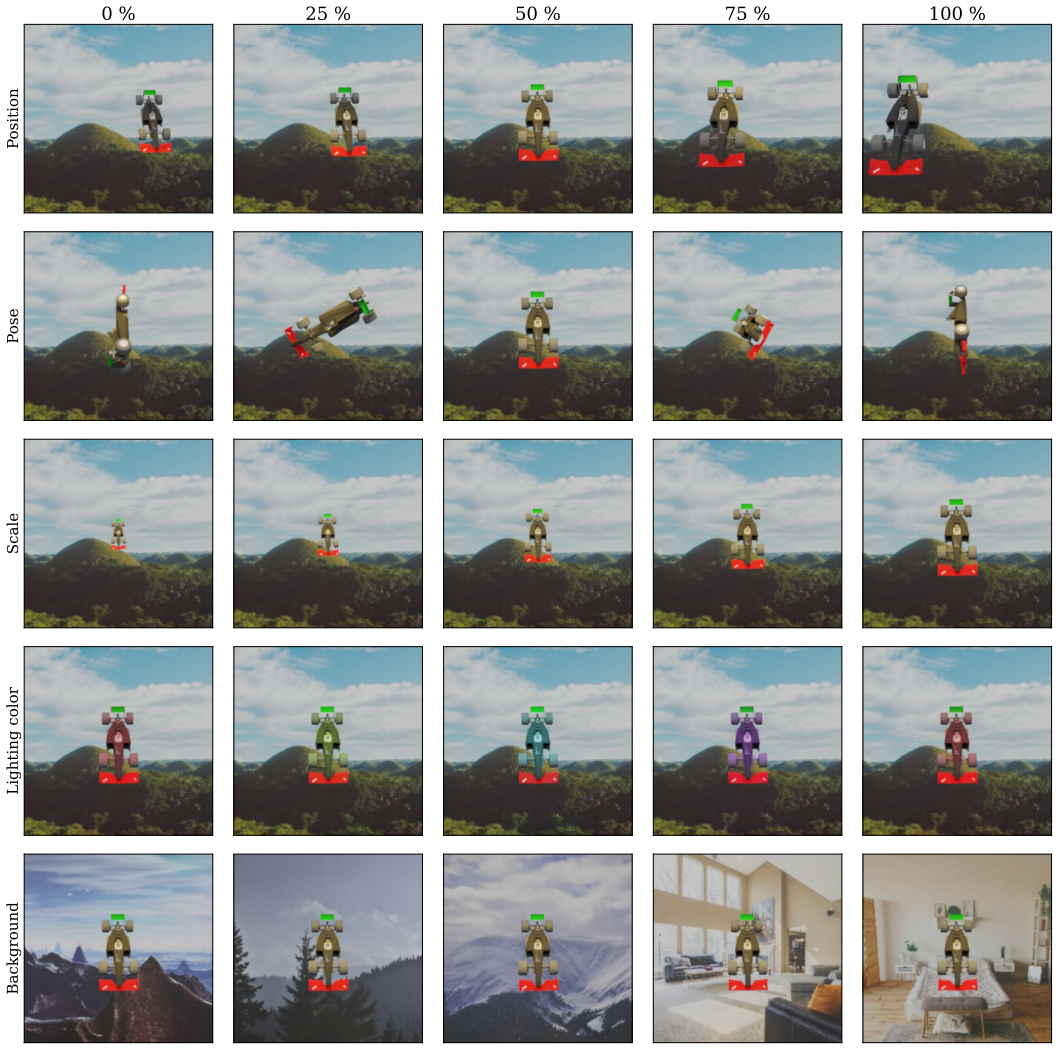


Figure A3: Examples from the dataset illustrating the different factors of variation.

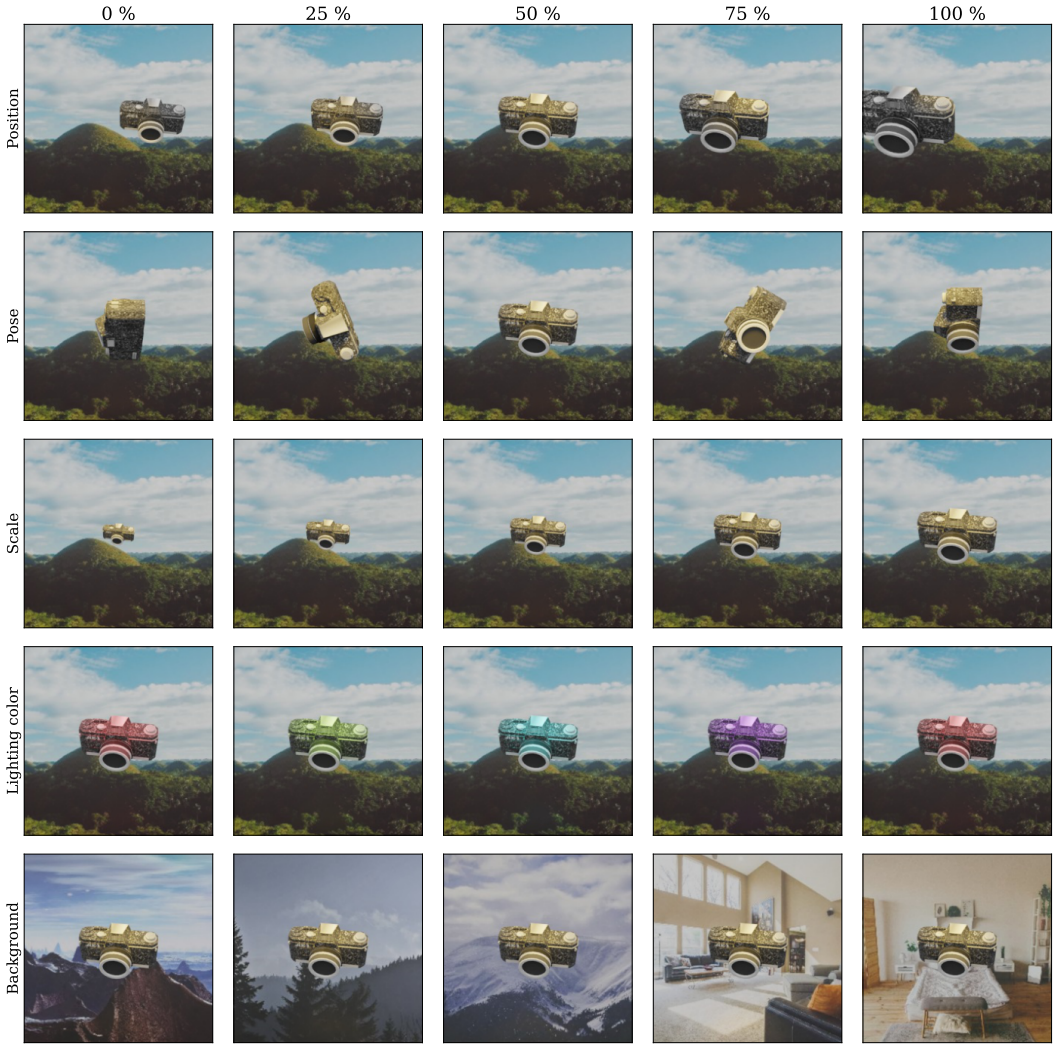


Figure A4: Examples from the dataset illustrating the different factors of variation.

Figure A5: Examples from the dataset illustrating the different factors of variation.

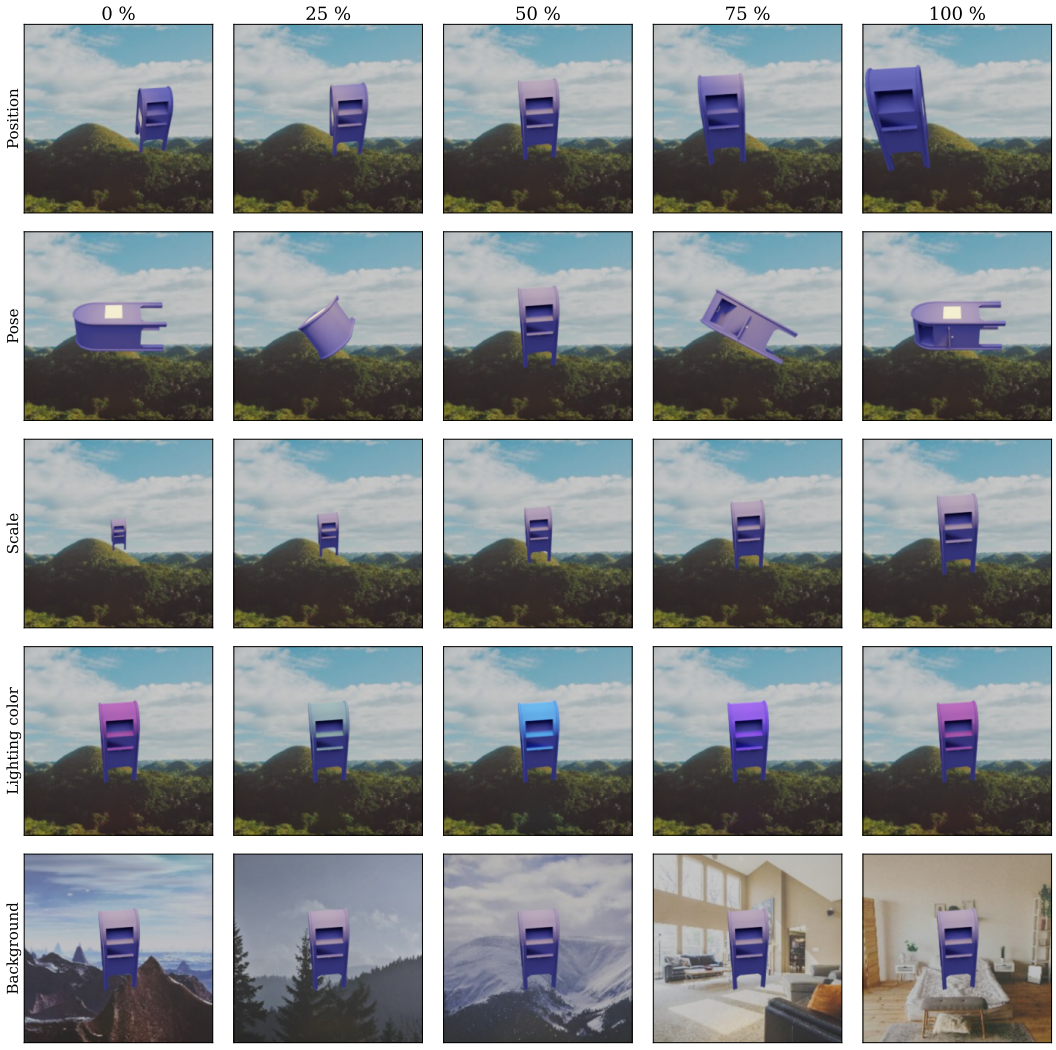


Figure A6: Examples from the dataset illustrating the different factors of variation.

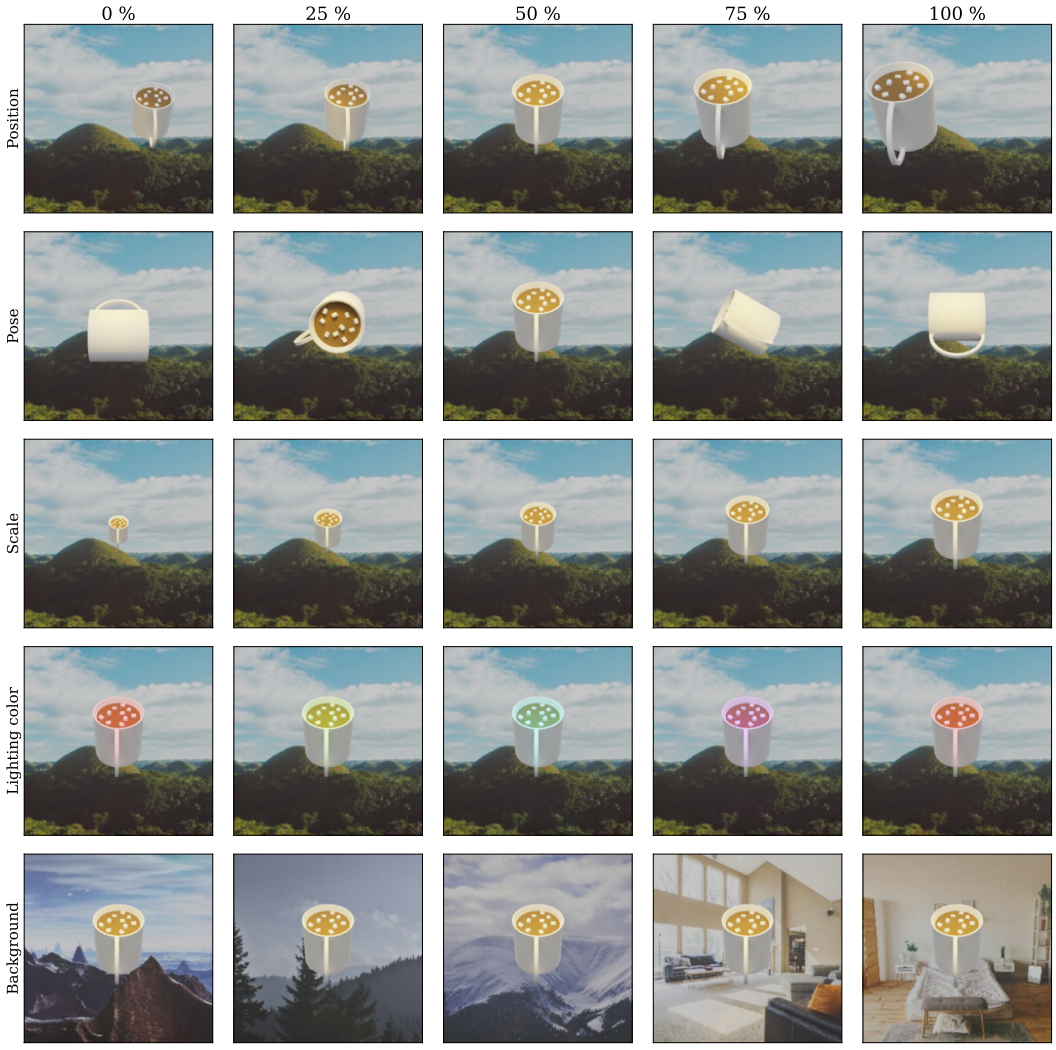


Figure A7: Examples from the dataset illustrating the different factors of variation.

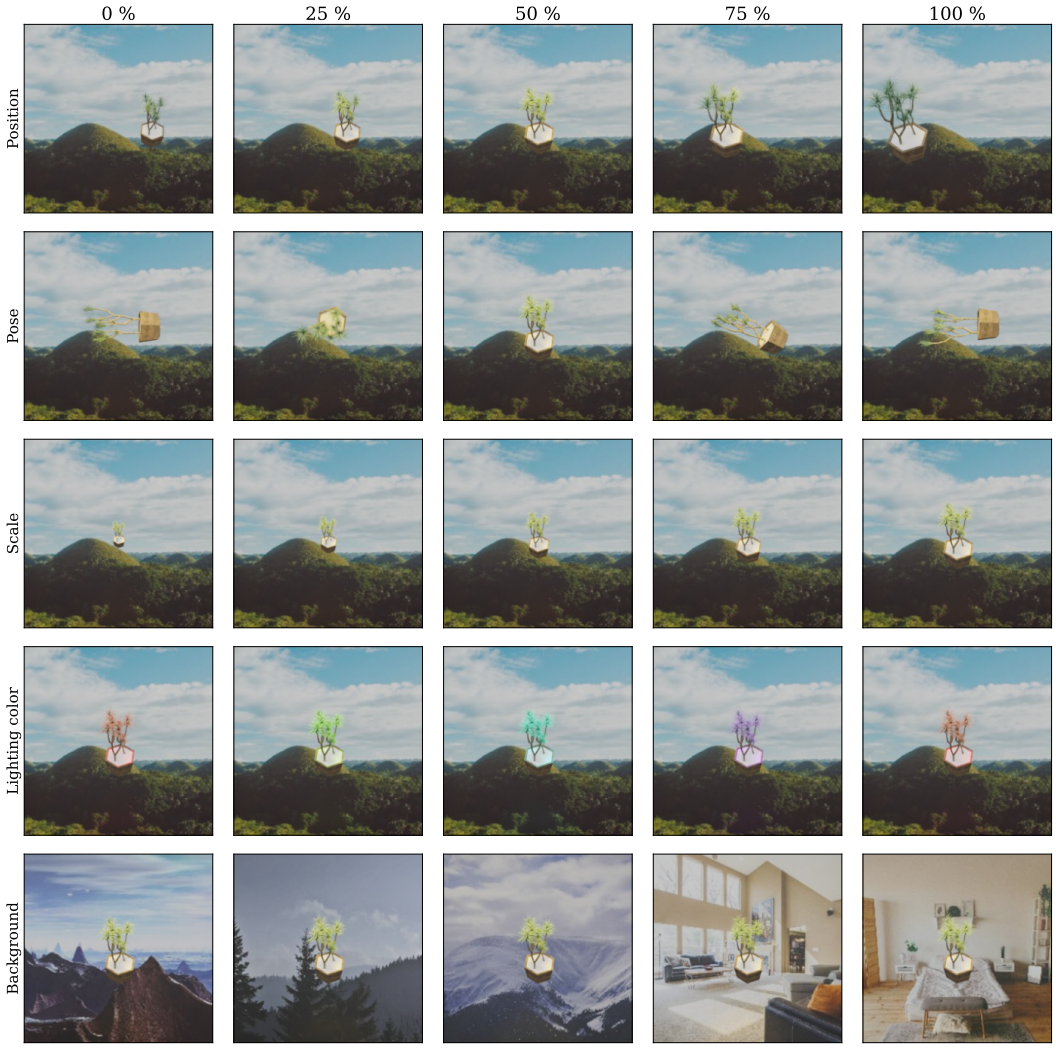


Figure A8: Examples from the dataset illustrating the different factors of variation.

Figure A9: Examples from the dataset illustrating the different factors of variation.

Figure A10: Examples from the dataset illustrating the different factors of variation.

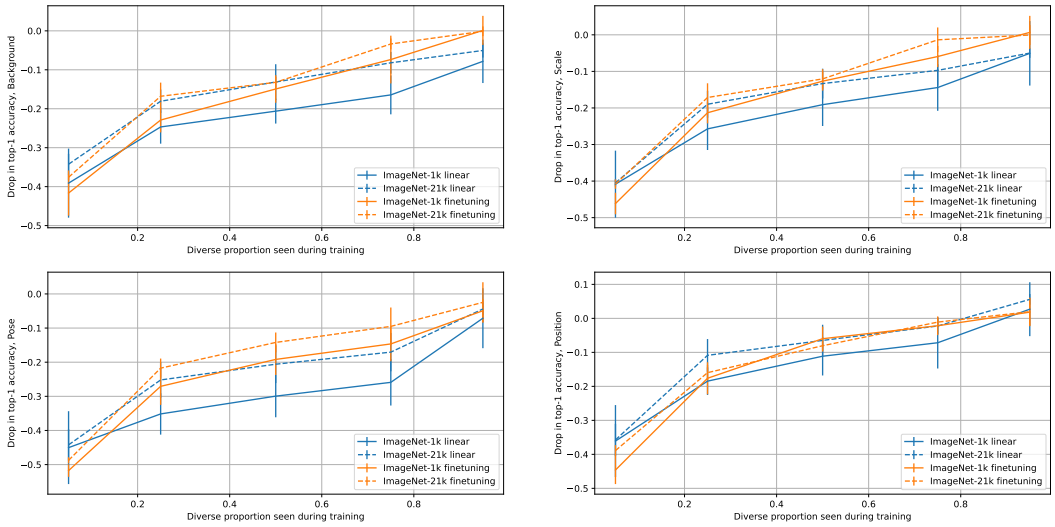


Figure A11: Drops in performance averaged over all methods when varying the proportion of varying examples seen during training.

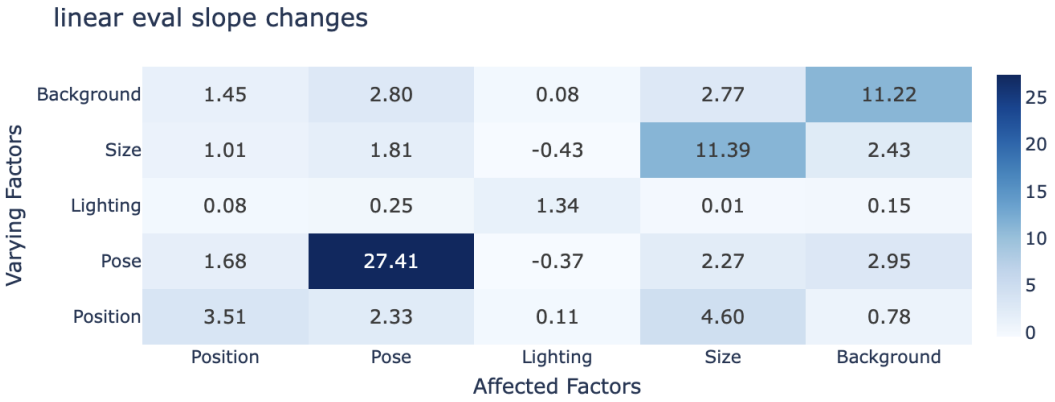


Figure A12: Cross factor changes when the given factor is varying for linear evaluation

400M images has canonical 30.1%, pose 16.0%, scale 18.4%, lighting color 28.3%, and background 23.6% accuracy.

We examine other prompts ("[], an inanimate object", "a photo of a [], an inanimate object", "[], a household item or vehicle") and observe similar classification performance (25-31% accuracy) using these variants.

C Cross Factor Effects When Varying All Instances

The cross factor effects when all instances vary with increasing diversity levels are shown in Figures A12 and A13.

D Varying a Subset of Instances During Training

We show the effect of increasing the number of instances seen varying during training. In Figure A14 we show the effect of each factor. We break down the effect by factor in figures A15 for linear evaluation and finetuning A17. In addition, we show the overall accuracy in tables A1, A5, A2, A3. The val canonical column corresponds to held-out accuracy for canonical and the val diverse corresponds to the accuracy for a changing factor.

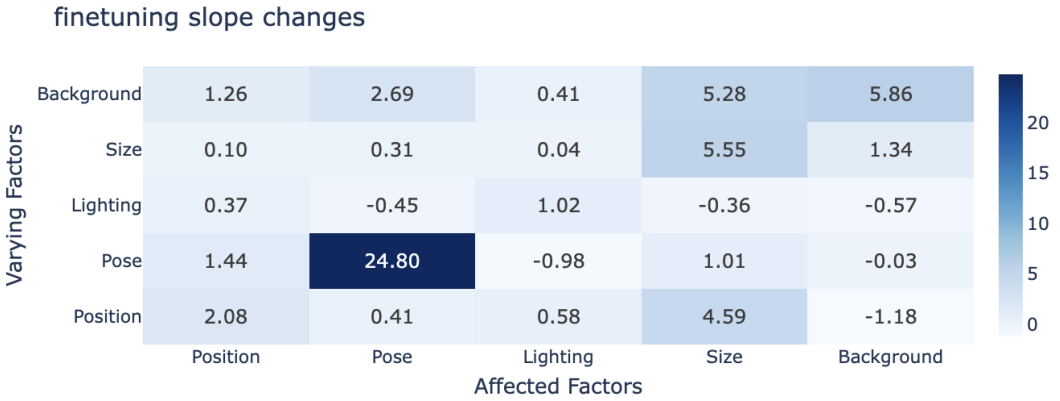


Figure A13: Cross factor changes when the given factor is varying for finetuning

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Translation_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	92.24%	91.37%	92.25%	94.01%	96.33%	77.78%	75.93%	68.52%	68.52%	58.89%	39.82%	57.70%	64.51%	65.45%	66.98%
MAEPretrained	52.54%	55.93%	60.44%	67.91%	83.79%	20.37%	27.78%	33.33%	38.89%	27.78%	9.67%	12.91%	14.26%	15.56%	15.18%
MLPMixerPretrained1k	94.66%	93.98%	94.58%	95.75%	97.25%	77.78%	74.07%	72.22%	66.67%	48.15%	36.86%	53.70%	59.35%	63.05%	63.46%
MLPMixerPretrained21k	94.95%	95.09%	95.19%	96.02%	97.13%	75.93%	75.93%	74.07%	77.78%	70.37%	43.89%	67.36%	71.25%	73.69%	76.18%
ResNet50Pretrained1k	95.00%	94.58%	94.83%	95.79%	97.23%	88.89%	90.74%	87.04%	83.33%	70.37%	44.35%	63.26%	68.99%	70.04%	70.01%
ResNet50Pretrained21k	95.51%	95.30%	95.90%	96.27%	97.39%	77.78%	72.22%	74.07%	74.07%	70.37%	46.61%	68.04%	73.02%	75.90%	77.13%
SimCLRPretrained	96.13%	95.74%	96.22%	96.82%	97.50%	81.48%	79.63%	81.48%	72.22%	70.00%	43.73%	62.96%	69.10%	70.63%	72.02%
ViTPretrained1k	95.79%	96.18%	96.37%	96.80%	97.75%	88.89%	83.33%	77.78%	77.41%	75.93%	49.34%	67.92%	72.74%	75.65%	77.22%
ViTPretrained21k	95.43%	95.01%	95.62%	96.39%	97.50%	83.33%	83.33%	83.33%	77.78%	72.22%	46.59%	67.21%	71.71%	74.02%	75.17%
iBotPretrained1k	96.67%	96.43%	96.49%	97.01%	97.66%	81.48%	81.48%	79.63%	79.63%	72.22%	40.27%	65.23%	73.10%	76.06%	77.33%
iBotPretrained21k	96.84%	96.30%	96.44%	96.92%	97.61%	90.74%	85.19%	87.04%	81.48%	72.22%	47.69%	70.55%	76.57%	78.53%	79.04%

Table A1: Position varying linear eval top-1 accuracy across multiple percentages of varying training instances.

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Rotation_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	91.70%	91.43%	92.42%	93.98%	96.28%	81.48%	85.19%	85.19%	79.63%	55.56%	28.64%	41.16%	44.72%	46.17%	48.85%
MAEPretrained	53.90%	57.24%	61.28%	68.44%	83.90%	25.93%	44.44%	38.89%	42.59%	29.63%	6.68%	7.93%	8.58%	8.63%	8.36%
MLPMixerPretrained1k	94.24%	93.76%	94.74%	95.72%	97.28%	74.07%	74.07%	72.22%	70.37%	46.30%	26.84%	39.47%	43.04%	46.51%	48.73%
MLPMixerPretrained21k	94.49%	94.81%	95.03%	95.88%	97.13%	79.63%	77.78%	77.78%	79.63%	72.22%	36.90%	55.46%	61.57%	63.67%	65.50%
ResNet50Pretrained1k	94.72%	94.56%	94.89%	95.74%	97.18%	85.19%	87.04%	85.19%	81.48%	68.52%	34.44%	46.22%	49.90%	51.73%	53.24%
ResNet50Pretrained21k	95.51%	94.96%	95.69%	96.12%	97.30%	77.78%	75.93%	75.93%	72.22%	66.67%	40.29%	57.68%	61.83%	63.91%	65.04%
SimCLRPretrained	95.74%	95.63%	96.16%	96.80%	97.51%	81.48%	83.33%	83.33%	83.33%	62.96%	32.94%	47.35%	52.88%	55.97%	56.91%
ViTPretrained1k	95.71%	95.76%	96.09%	96.79%	97.67%	87.04%	79.63%	81.48%	79.63%	59.26%	39.13%	55.09%	60.14%	64.14%	65.42%
ViTPretrained21k	95.23%	94.89%	95.68%	96.42%	97.45%	85.19%	83.33%	83.33%	83.33%	66.67%	38.25%	54.50%	58.34%	60.94%	62.28%
iBotPretrained1k	96.39%	96.22%	96.41%	96.96%	97.56%	83.33%	81.48%	81.48%	79.63%	72.22%	34.62%	51.88%	58.88%	62.19%	63.11%
iBotPretrained21k	96.44%	96.09%	96.42%	96.92%	97.61%	90.74%	88.89%	90.74%	87.04%	72.22%	41.03%	57.48%	63.69%	65.49%	67.34%

Table A2: Pose linear eval top-1 accuracy across multiple percentages of varying training instances.

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Spot hue_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	92.49%	91.60%	92.46%	94.06%	96.32%	77.78%	75.93%	70.37%	70.37%	59.26%	39.61%	60.78%	66.72%	68.27%	68.64%
MAEPretrained	52.26%	55.43%	60.36%	67.78%	83.62%	22.22%	31.48%	37.04%	37.04%	20.37%	11.51%	15.09%	16.10%	18.91%	15.58%
MLPMixerPretrained1k	95.17%	94.20%	94.98%	95.86%	97.30%	79.63%	77.78%	70.37%	66.67%	57.04%	40.09%	56.90%	64.43%	67.35%	68.93%
MLPMixerPretrained21k	94.89%	95.29%	95.44%	96.13%	97.20%	81.48%	79.63%	72.22%	74.07%	76.30%	44.68%	69.30%	73.32%	76.12%	77.75%
ResNet50Pretrained1k	95.26%	94.81%	95.03%	95.80%	97.23%	87.04%	88.89%	87.04%	83.33%	77.78%	44.08%	62.96%	68.67%	71.81%	72.54%
ResNet50Pretrained21k	95.91%	95.45%	96.00%	96.28%	97.41%	77.78%	75.93%	75.93%	72.22%	71.11%	44.22%	67.22%	70.95%	72.38%	74.39%
SimCLRPretrained	96.30%	95.91%	96.27%	96.84%	97.59%	79.63%	75.93%	74.07%	74.07%	66.67%	43.36%	63.22%	71.32%	73.72%	72.26%
ViTPretrained1k	95.99%	96.36%	96.54%	96.95%	97.80%	90.74%	87.04%	83.33%	81.48%	74.07%	52.53%	69.53%	71.81%	76.01%	77.28%
ViTPretrained21k	95.57%	95.39%	95.84%	96.51%	97.59%	85.19%	81.48%	83.33%	77.78%	75.93%	46.01%	67.50%	71.97%	75.75%	77.06%
iBotPretrained1k	96.50%	96.55%	96.74%	97.12%	97.70%	79.63%	81.48%	81.48%	77.78%	74.07%	42.32%	68.61%	72.75%	76.28%	78.04%
iBotPretrained21k	97.01%	96.42%	96.65%	97.04%	97.64%	88.89%	87.04%	83.33%	83.33%	75.93%	48.83%	73.37%	78.09%	80.39%	81.55%

Table A3: Spot hue linear eval top-1 accuracy across multiple percentages of varying training instances.

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Scale_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	91.81%	91.51%	92.26%	93.90%	96.25%	79.63%	72.22%	74.07%	70.37%	61.11%	36.80%	51.29%	57.05%	56.99%	58.80%
MAEPretrained	51.98%	56.15%	60.87%	68.07%	84.12%	25.93%	35.19%	33.33%	37.04%	35.19%	7.01%	10.74%	11.05%	11.29%	11.44%
MLPMixerPretrained1k	94.27%	93.55%	94.47%	95.59%	97.17%	79.63%	77.78%	70.37%	68.52%	53.70%	32.50%	46.15%	51.66%	55.30%	56.62%
MLPMixerPretrained21k	94.86%	94.98%	95.08%	95.93%	97.09%	79.63%	79.63%	75.93%	75.93%	74.07%	40.55%	60.87%	66.63%	67.52%	70.03%
ResNet50Pretrained1k	94.89%	94.57%	94.84%	95.66%	97.16%	87.04%	90.74%	90.74%	83.33%	72.22%	39.22%	54.23%	59.10%	61.27%	60.89%
ResNet50Pretrained21k	95.51%	95.28%	95.77%	96.08%	97.35%	81.48%	77.78%	77.78%	74.07%	74.07%	41.11%	60.06%	66.41%	69.69%	70.33%
SimCLRPretrained	95.91%	95.53%	96.09%	96.66%	97.48%	83.33%	77.41%	77.78%	72.22%	62.96%	39.79%	54.93%	61.81%	62.93%	62.96%
ViTPretrained1k	95.65%	96.01%	96.18%	96.69%	97.69%	87.04%	83.33%	79.63%	75.93%	72.22%	44.10%	58.13%	63.91%	67.85%	68.82%
ViTPretrained21k	95.06%	94.87%	95.47%	96.30%	97.38%	83.33%	83.33%	83.33%	81.48%	72.22%	41.40%	58.53%	63.25%	65.43%	65.14%
iBotPretrained1k	96.58%	96.37%	96.48%	96.98%	97.60%	84.07%	77.78%	79.63%	77.78%	66.67%	41.34%	58.93%	67.24%	68.66%	69.08%
iBotPretrained21k	96.92%	96.18%	96.39%	96.86%	97.53%	85.19%	77.78%	81.48%	81.48%	75.93%	44.59%	63.11%	68.90%	71.45%	70.82%

Table A4: Scale linear eval top-1 accuracy across multiple percentages of varying training instances

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Background path_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	90.37%	91.88%	92.73%	94.36%	96.70%	85.19%	77.78%	77.78%	77.78%	66.67%	41.30%	54.70%	59.96%	61.75%	63.41%
MAEPretrained	51.51%	56.16%	62.31%	67.08%	84.32%	27.78%	31.48%	37.04%	37.04%	29.63%	9.47%	12.21%	11.84%	13.30%	12.67%
MLPMixerPretrained1k	92.84%	93.82%	94.81%	95.41%	97.38%	75.93%	77.78%	72.22%	74.07%	61.11%	37.39%	49.19%	52.41%	55.75%	56.74%
MLPMixerPretrained21k	95.42%	95.33%	95.62%	96.52%	97.55%	83.33%	81.48%	75.93%	75.93%	77.78%	48.30%	64.37%	66.79%	70.15%	70.84%
ResNet50Pretrained1k	94.35%	95.12%	94.83%	95.93%	97.46%	90.74%	92.59%	88.89%	88.89%	83.33%	44.89%	59.87%	64.15%	66.70%	67.29%
ResNet50Pretrained21k	95.20%	96.40%	95.89%	96.65%	97.67%	81.48%	79.63%	77.78%	77.78%	75.93%	51.21%	65.75%	69.47%	72.01%	73.90%
SimCLRPretrained	95.50%	95.98%	96.14%	96.45%	97.52%	83.33%	83.33%	81.48%	77.78%	72.22%	44.33%	59.82%	65.39%	67.93%	68.93%
ViTPretrained1k	95.72%	96.42%	96.27%	96.57%	97.95%	94.44%	88.89%	88.89%	87.04%	77.78%	50.62%	64.09%	67.14%	72.33%	73.19%
ViTPretrained21k	94.91%	95.22%	96.09%	96.45%	97.71%	83.33%	83.33%	83.33%	83.33%	77.78%	49.36%	64.21%	67.42%	70.77%	72.13%
iBotPretrained1k	96.42%	96.58%	96.60%	97.40%	97.28%	88.89%	83.33%	87.04%	81.48%	79.63%	44.58%	62.59%	68.10%	71.20%	73.36%
iBotPretrained21k	96.16%	96.14%	96.38%	97.37%	97.27%	88.89%	90.74%	90.74%	83.33%	81.48%	51.20%	68.58%	71.57%	74.62%	75.94%

Table A5: Background path linear eval top-1 accuracy across multiple percentages of varying training instances

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Translation_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	97.49%	97.00%	97.25%	97.55%	96.80%	87.04%	90.74%	77.78%	81.48%	75.93%	45.33%	69.74%	74.53%	77.72%	78.19%
MAEPretrained	96.75%	96.63%	97.20%	97.46%	97.91%	83.33%	74.07%	66.67%	64.81%	72.22%	29.17%	51.35%	61.05%	65.16%	70.07%
MLPMixerPretrained1k	96.95%	96.73%	97.17%	97.24%	97.88%	88.89%	77.78%	77.78%	74.07%	64.81%	44.65%	64.56%	70.67%	74.34%	75.95%
MLPMixerPretrained21k	97.71%	97.69%	97.90%	97.98%	98.35%	85.19%	88.89%	85.19%	81.48%	77.78%	46.53%	70.54%	77.27%	79.78%	81.68%
ResNet50Pretrained1k	97.97%	97.92%	97.91%	97.86%	98.27%	87.04%	87.04%	72.22%	81.48%	81.48%	45.05%	64.83%	71.87%	76.56%	79.63%
ResNet50Pretrained21k	97.54%	97.62%	97.74%	97.69%	98.20%	88.89%	83.33%	83.33%	83.33%	77.78%	52.22%	70.96%	75.78%	81.10%	82.45%
SimCLRPretrained	97.40%	97.57%	97.68%	97.81%	98.12%	90.74%	77.78%	87.04%	83.33%	77.78%	45.21%	68.73%	74.59%	76.55%	79.86%
ViTPretrained1k	97.80%	97.88%	98.00%	97.92%	98.28%	90.74%	90.74%	85.19%	81.48%	81.48%	49.30%	71.82%	76.95%	80.39%	82.58%
ViTPretrained21k	97.80%	97.59%	97.89%	97.91%	98.25%	87.04%	88.89%	81.48%	81.48%	87.04%	45.83%	71.80%	75.51%	79.96%	81.86%
iBotPretrained1k	97.77%	97.55%	97.64%	97.77%	98.06%	88.89%	85.19%	79.63%	75.93%	79.63%	45.69%	68.94%	74.76%	76.63%	79.83%
iBotPretrained21k	97.97%	97.83%	97.88%	97.92%	98.13%	88.89%	87.04%	88.89%	81.48%	79.63%	49.84%	70.97%	78.13%	82.53%	84.07%

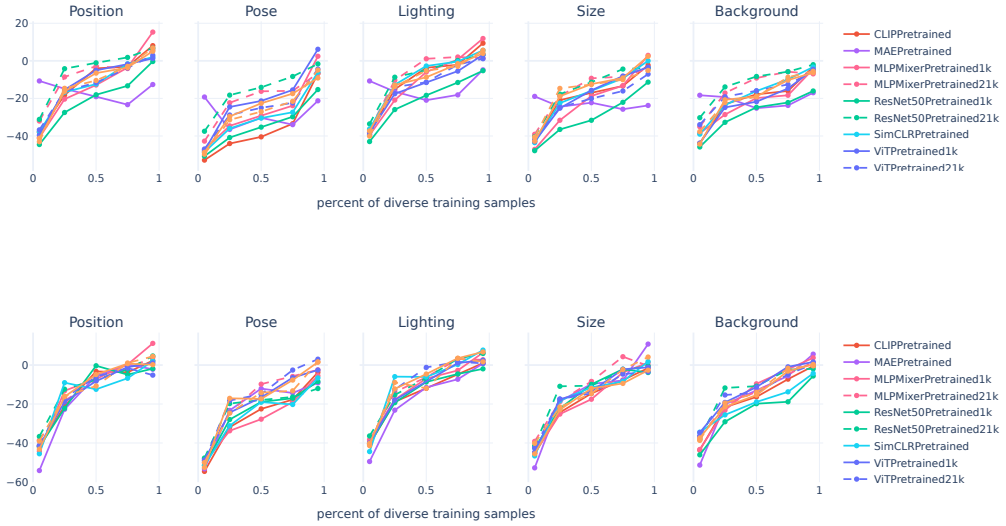
Table A6: Position finetuning top-1 accuracy across multiple percentages of varying training instances

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Rotation_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	97.60%	97.01%	97.30%	97.58%	97.97%	88.89%	88.89%	85.19%	83.33%	72.22%	34.29%	57.01%	62.67%	65.58%	68.45%
MAEPretrained	96.87%	96.75%	97.25%	97.52%	97.95%	75.93%	68.52%	62.96%	68.52%	62.96%	22.64%	45.24%	50.87%	54.18%	53.82%
MLPMixerPretrained1k	96.75%	96.58%	97.08%	97.25%	97.86%	83.33%	85.19%	83.33%	77.78%	64.81%	33.29%	51.43%	55.55%	58.78%	59.35%
MLPMixerPretrained21k	97.68%	97.65%	97.90%	97.90%	98.35%	87.04%	87.04%	77.78%	77.78%	75.93%	38.93%	62.76%	67.97%	71.90%	73.47%
ResNet50Pretrained1k	98.05%	97.81%	97.89%	97.93%	98.24%	84.81%	81.48%	81.48%	81.48%	75.93%	33.29%	53.51%	62.56%	64.45%	67.47%
ResNet50Pretrained21k	97.52%	97.45%	97.65%	97.61%	98.18%	85.19%	79.63%	83.33%	87.04%	85.19%	37.45%	59.85%	65.39%	70.66%	73.13%
SimCLRPretrained	97.37%	97.43%	97.53%	97.74%	98.07%	87.04%	87.04%	83.33%	87.04%	75.93%	35.50%	55.87%	64.34%	66.79%	69.34%
ViTPretrained1k	97.94%	97.87%	97.98%	97.95%	98.31%	88.89%	87.04%	85.19%	77.78%	75.93%	40.13%	62.66%	68.53%	71.32%	73.36%
ViTPretrained21k	97.68%	97.61%	97.90%	97.97%	98.27%	88.89%	79.63%	83.33%	74.07%	70.74%	39.75%	61.42%	68.39%	71.51%	73.76%
iBotPretrained1k	97.49%	97.55%	97.56%	97.75%	98.02%	88.89%	77.78%	83.33%	75.93%	68.52%	36.30%	60.69%	65.98%	68.22%	70.00%
iBotPretrained21k	98.00%	97.79%	97.96%	98.01%	98.19%	88.89%	87.04%	83.33%	85.19%	72.22%	38.86%	62.35%	69.18%	71.96%	73.84%

Table A7: Pose finetuning top-1 accuracy across multiple percentages of varying training instances

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Spot hue_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	97.66%	97.13%	97.40%	97.62%	97.98%	90.74%	88.89%	87.04%	87.04%	81.48%	49.99%	70.06%	75.12%	82.36%	82.48%
MAEPretrained	97.04%	96.67%	97.20%	97.41%	97.95%	79.63%	77.78%	72.22%	74.07%	68.52%	30.10%	54.63%	60.74%	66.78%	69.34%
MLPMixerPretrained1k	97.32%	96.92%	97.20%	97.36%	97.89%	87.04%	83.33%	77.78%	79.63%	72.22%	48.80%	65.60%	70.82%	76.88%	78.59%
MLPMixerPretrained21k	97.80%	97.83%	97.99%	97.99%	98.42%	88.89%	87.04%	81.48%	75.93%	79.63%	49.81%	73.15%	75.23%	79.01%	82.35%
ResNet50Pretrained1k	98.02%	97.99%	98.03%	97.98%	98.28%	88.89%	87.04%	83.33%	81.48%	77.78%	48.48%	67.72%	74.65%	76.90%	75.83%
ResNet50Pretrained21k	97.68%	97.60%	97.87%	97.79%	98.24%	90.74%	87.04%	83.33%	77.78%	75.93%	54.35%	71.69%	76.70%	81.03%	81.86%
SimCLRPretrained	97.60%	97.61%	97.72%	97.87%	98.17%	90.00%	76.30%	85.19%	77.78%	74.07%	45.58%	70.36%	79.09%	78.06%	81.75%
ViTPretrained1k	97.68%	97.93%	NaN	98.00%	98.37%	94.44%	88.89%	NaN	81.48%	81.48%	54.38%	70.94%	NaN	82.53%	83.94%
ViTPretrained21k	97.77%	97.68%	97.94%	98.00%	98.24%	92.59%	85.19%	77.78%	78.15%	81.48%	52.49%	72.55%	76.58%	79.75%	82.39%
iBotPretrained1k	97.91%	97.58%	97.76%	97.88%	98.08%	88.89%	81.48%	79.63%	75.93%	74.07%	48.67%	69.24%	75.01%	79.44%	80.81%
iBotPretrained21k	98.25%	97.87%	97.88%	97.96%	98.13%	90.74%	83.33%	92.59%	79.63%	83.33%	49.39%	74.42%	80.67%	83.05%	85.02%

Table A8: Spot hue finetuning top-1 accuracy across multiple percentages of varying training instances



(a) Training with increasing percentage of variability across all instances using finetuning

Figure A14: Training with increasing percentage of instances seen varying during training using linear evaluation (top) and finetuning (bottom).

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Scale_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	97.68%	97.08%	97.39%	97.63%	97.95%	90.74%	90.74%	87.04%	83.33%	79.63%	45.47%	66.39%	72.46%	75.63%	78.05%
MAEPretrained	96.81%	96.64%	97.19%	97.40%	97.91%	81.48%	70.37%	70.37%	70.37%	53.70%	28.70%	51.73%	60.66%	62.49%	64.47%
MLPMixerPretrained1k	97.23%	96.60%	97.07%	97.24%	97.82%	87.04%	85.19%	83.33%	74.07%	70.37%	41.62%	60.01%	65.78%	70.24%	71.64%
MLPMixerPretrained21k	97.80%	97.80%	97.96%	97.95%	98.37%	85.19%	87.04%	81.48%	72.22%	77.78%	46.02%	68.37%	73.08%	76.51%	77.38%
ResNet50Pretrained1k	97.88%	97.94%	97.95%	97.89%	98.24%	87.04%	85.19%	81.48%	75.93%	79.63%	44.47%	62.86%	71.55%	73.81%	75.80%
ResNet50Pretrained21k	97.40%	97.58%	97.84%	97.72%	98.18%	90.74%	79.63%	81.48%	79.63%	79.63%	49.37%	68.72%	70.89%	76.85%	79.18%
SimCLRPretrained	97.57%	97.54%	97.65%	97.83%	98.10%	88.89%	83.33%	83.33%	83.33%	74.44%	42.25%	65.04%	71.88%	74.89%	76.25%
ViTPretrained1k	97.80%	97.92%	98.05%	97.92%	98.34%	88.89%	83.33%	85.19%	79.63%	79.63%	44.17%	65.86%	71.66%	77.46%	78.46%
ViTPretrained21k	97.77%	97.71%	97.85%	97.99%	98.24%	85.19%	85.19%	85.19%	79.63%	79.63%	42.63%	67.71%	70.61%	74.96%	76.14%
iBotPretrained1k	97.85%	97.61%	97.74%	97.79%	98.04%	90.74%	87.04%	83.33%	83.33%	79.63%	45.14%	64.09%	71.34%	73.90%	76.99%
iBotPretrained21k	97.88%	97.75%	97.86%	97.97%	98.12%	88.89%	87.04%	87.04%	81.48%	75.93%	48.70%	65.52%	72.39%	79.16%	80.04%

Table A9: Scale finetuning top-1 accuracy across multiple percentages of varying training instances

E ALL FACTOR GAPS

We also study the setting where all factors vary during training. In Figures A19 and A20 we show the generalization gaps when all factors vary for linear evaluation and finetuning.

F CLASS GENERALIZATION

In addition to the finetning results, we include here linear evaluation results for class generalization gaps A11.

	train_canonical_top_1_accuracy					val_canonical_top_1_accuracy					val_diverse_Background path_top_1_accuracy				
train_prop_to_vary	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95	0.05	0.25	0.50	0.75	0.95
model															
CLIPPretrained	96.49%	97.05%	97.52%	97.81%	98.01%	94.44%	88.89%	92.59%	87.04%	81.48%	50.79%	67.24%	76.25%	79.79%	80.92%
MAEPretrained	96.20%	96.61%	97.31%	97.63%	97.99%	81.85%	72.22%	77.78%	72.22%	62.96%	30.49%	52.04%	64.77%	66.67%	68.56%
MLPMixerPretrained1k	96.16%	97.03%	97.13%	97.66%	97.76%	90.74%	87.04%	81.48%	75.93%	72.22%	47.41%	63.84%	71.85%	73.76%	75.96%
MLPMixerPretrained21k	98.08%	98.15%	98.09%	98.33%	98.71%	88.89%	92.59%	90.74%	85.19%	79.63%	51.84%	72.00%	77.05%	80.46%	81.10%
ResNet50Pretrained1k	97.71%	97.76%	97.95%	98.04%	98.38%	92.59%	92.59%	90.74%	92.59%	83.33%	46.50%	63.48%	70.90%	73.73%	77.72%
ResNet50Pretrained21k	97.38%	98.09%	97.72%	98.33%	98.34%	88.89%	83.33%	87.04%	81.48%	85.19%	50.87%	71.59%	76.21%	79.13%	83.24%
SimCLRPretrained	97.38%	97.31%	97.77%	97.55%	98.05%	77.78%	87.04%	88.89%	90.74%	83.33%	43.32%	61.47%	69.94%	76.99%	79.07%
ViTPretrained1k	98.12%	97.76%	97.89%	97.73%	98.44%	88.89%	90.74%	87.04%	81.48%	83.33%	54.25%	71.56%	75.73%	80.58%	84.92%
ViTPretrained21k	97.77%	97.49%	97.95%	98.04%	98.37%	91.67%	88.89%	90.74%	87.04%	81.48%	55.17%	73.53%	76.50%	82.28%	81.94%
iBotPretrained1k	97.47%	97.44%	97.61%	98.15%	97.86%	88.89%	92.59%	90.74%	81.48%	79.63%	51.19%	71.17%	75.32%	78.37%	79.93%
iBotPretrained21k	97.69%	97.91%	97.77%	98.17%	97.93%	93.52%	92.59%	90.74%	85.19%	83.33%	55.00%	73.11%	76.62%	83.56%	82.90%

Table A10: Background path finetuning top-1 accuracy across multiple percentages of varying training instances

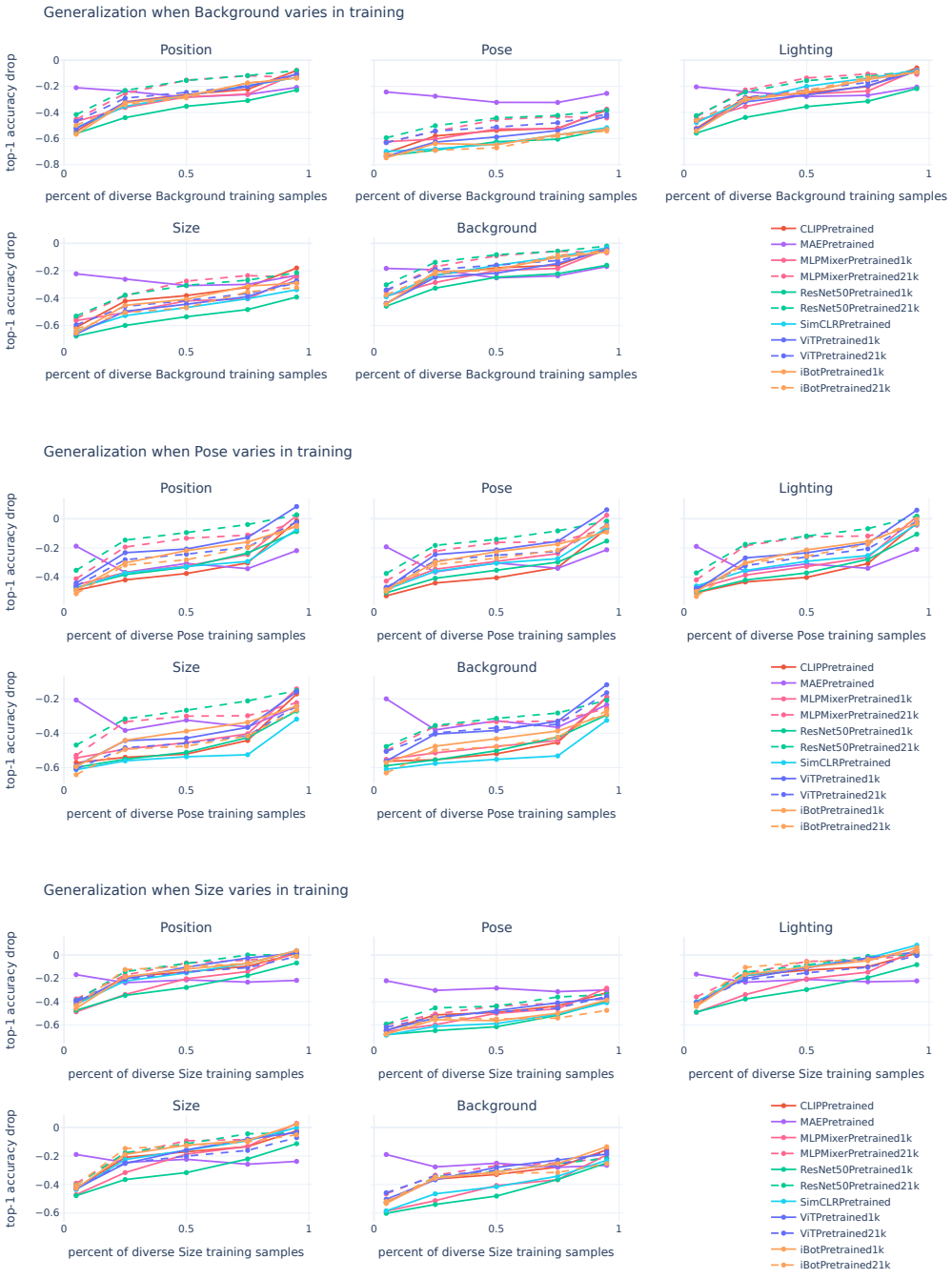


Figure A15: Linear Evaluation Effect of Variability in Training (part 1)

G CROSS FACTOR EFFECTS

We study the effect of varying a factor on the generalization gaps of other factors. In Figures A21 and A22 we show the slopes of the generalization gaps as the number of varying training instances increases during training. We see how varying one factor can also close the robustenss gap of other factors. We also show normalized versions of these plots in A23 and A24.

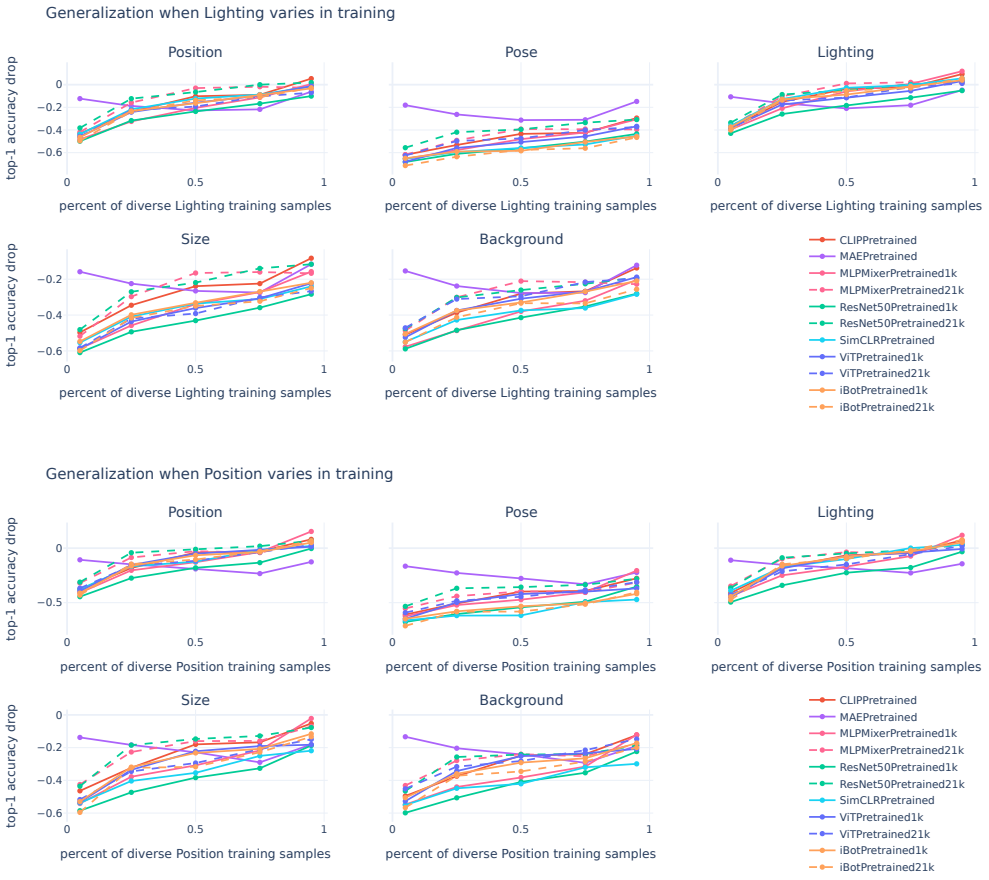


Figure A16: Linear Evaluation Effect of Variability in Training (part 2)

model	Position gap	Pose gap	Lighting color gap	Size gap	Background gap	Average gap
CLIP	-41.69	-46.51	-43.73	-41.51	-48.09	-44.31
MAE	-3.32	-6.64	-7.69	-5.68	-14.61	-7.59
MLPMixer1k	-39.07	-43.70	-43.42	-36.32	-47.79	-42.06
MLPMixer21k	-43.07	-57.34	-42.90	-44.84	-52.09	-48.05
ResNet50-1k	-44.97	-51.06	-43.84	-42.18	-55.67	-47.54
ResNet50-21k	-46.34	-56.89	-44.06	-47.26	-57.81	-50.47
ViT-1k	-47.03	-58.69	-45.24	-48.67	-52.77	-50.48
ViT-21k	-50.41	-56.94	-49.17	-49.76	-55.23	-52.30
iBot-1k	-46.90	-61.71	-46.35	-51.56	-59.88	-53.28
iBot-21k	-53.00	-64.94	-50.99	-56.05	-64.83	-57.96
Average	-41.58	-50.44	-41.74	-42.38	-50.88	-45.40

Table A11: Linear eval class generalization top-1 accuracy gaps: shows validation top-1 accuracy difference between classes (27 randomly selected) seen with diversity and those not.

H Effect of class similarity on models' ability to generalize variation across classes

We study the effect of class similarity by measuring the generalization gaps per class for each factor relative to the class's similarity to the nearest class seen varying during training. If models' are able to generalize variation across classes, we might expect models generalize variation better when the class is similar to one seen varying during training. In Figures A25, A25, A29, and A27.

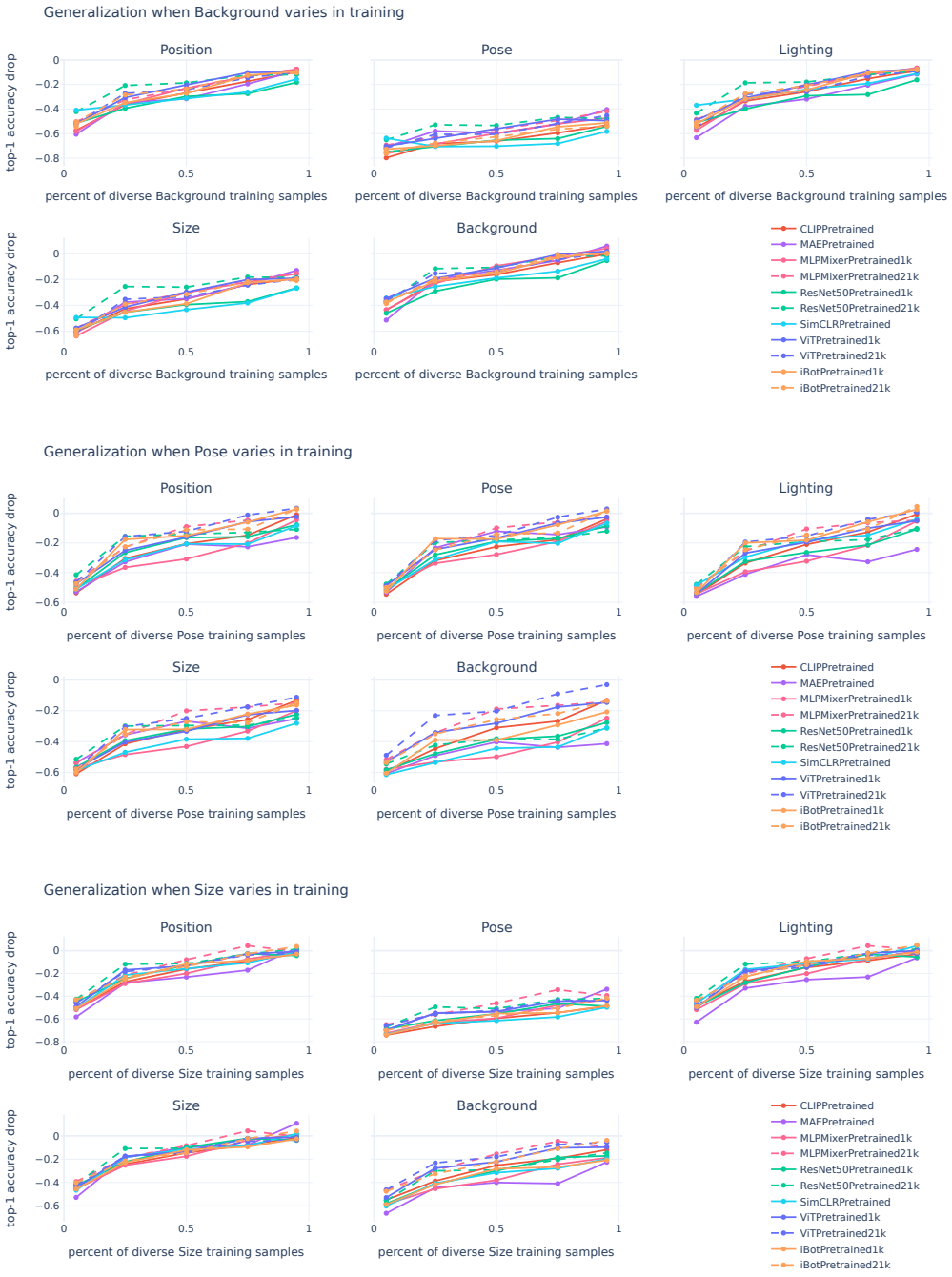


Figure A17: Finetuning Effect of Variability in Training (part 1)

I EXPERIMENTS DETAILS

Tables 2a and 2b show results for the best after 10k steps of training with adam on 6 log scale learning rates (1e-2 to 1e-6) cross validated on canonical top-1 accuracy for validation images.

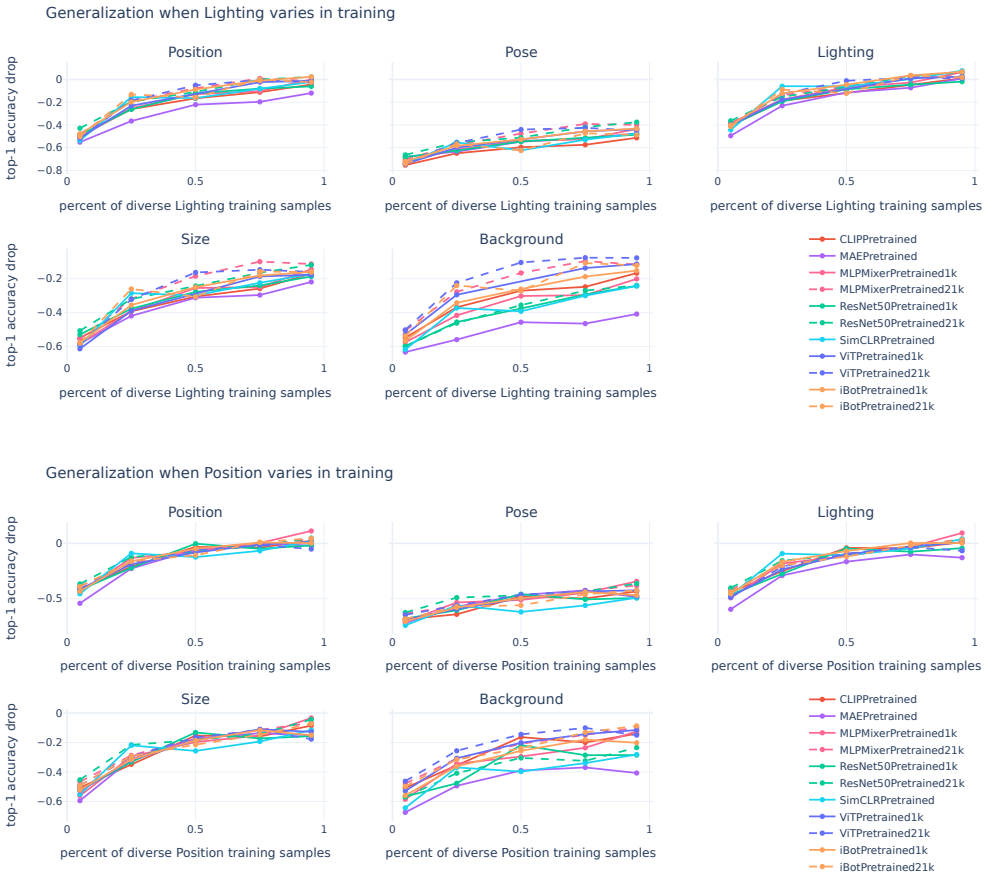


Figure A18: Finetuning Effect of Variability in Training (part 2)

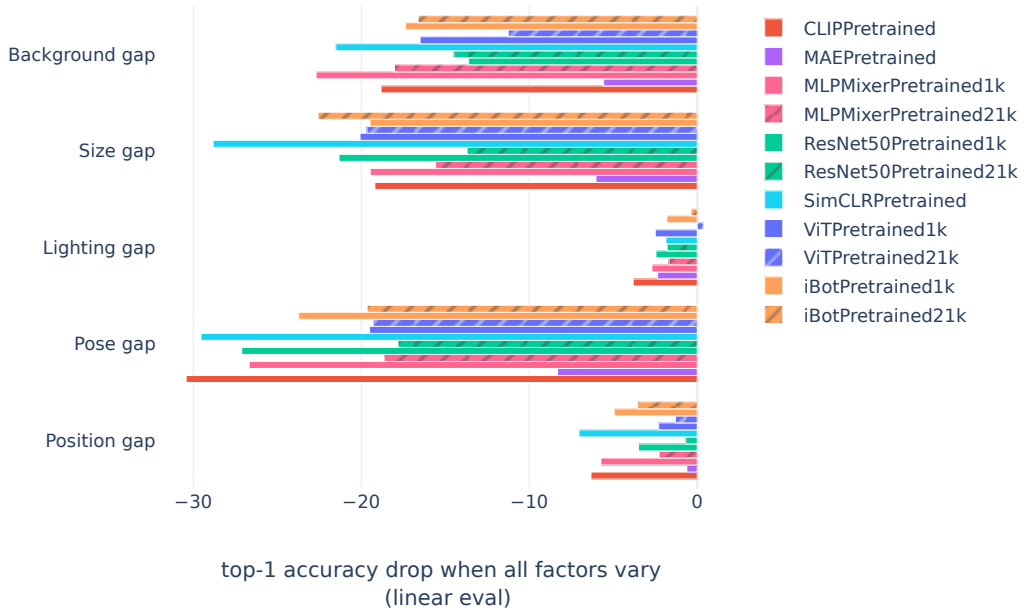


Figure A19: Generalization gaps when all factors vary during training with linear evaluation

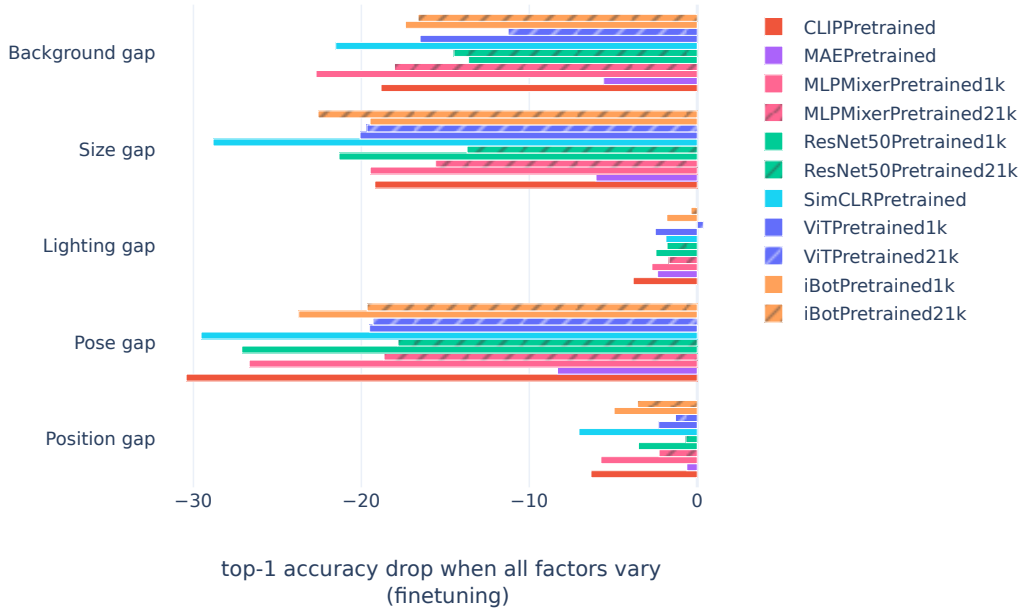


Figure A20: Generalization gaps when all factors vary during training with finetuning

varying factor	Position gap	Pose gap	Lighting gap	Size gap	Background gap
Position	0.385894	0.249188	0.411196	0.350276	0.287304
Pose	0.383425	0.369503	0.400835	0.308550	0.280720
Lighting	0.398418	0.243639	0.385000	0.330450	0.281797
Size	0.385311	0.245047	0.395171	0.352747	0.294040
Background	0.343635	0.209684	0.354958	0.304687	0.300778

Figure A21: Spill over effects: shows the average slope across models when a given factor varies during linear evaluation

varying factor	Position gap	Pose gap	Lighting gap	Size gap	Background gap
Position	0.448903	0.262276	0.475593	0.419783	0.370703
Pose	0.442200	0.454216	0.469521	0.372108	0.353449
Lighting	0.480997	0.277407	0.444130	0.415490	0.384728
Size	0.483358	0.274840	0.487531	0.451954	0.427255
Background	0.425475	0.242605	0.437238	0.406613	0.405921

Figure A22: Spill over effects: shows the average slope across models when a given factor varies for finetuning

varying factor	Position gap	Pose gap	Lighting gap	Size gap	Background gap
Position	1.000000	0.674386	1.068042	0.992995	0.955206
Pose	0.993602	1.000000	1.041130	0.874707	0.933316
Lighting	1.032454	0.659368	1.000000	0.936791	0.936894
Size	0.998488	0.663178	1.026418	1.000000	0.977599
Background	0.890490	0.567474	0.921968	0.863755	1.000000

Figure A23: Normalized Spill over effects: shows the average slope across models when a given factor varies during linear evaluation. Normalization is across rows by dividing the diagonal value to isolate how much more a given spill-over effect than the intended.

varying factor	Position gap	Pose gap	Lighting gap	Size gap	Background gap
Position	1.000000	0.577425	1.070841	0.928819	0.913239
Pose	0.985068	1.000000	1.057171	0.823333	0.870733
Lighting	1.071495	0.610738	1.000000	0.919320	0.947789
Size	1.076755	0.605087	1.097721	1.000000	1.052556
Background	0.947811	0.534118	0.984481	0.899679	1.000000

Figure A24: Normalized Spill over effects: shows the average slope across models when a given factor varies for finetuning. Normalization is across rows by dividing the diagonal value to isolate how much more a given spill-over effect than the intended.

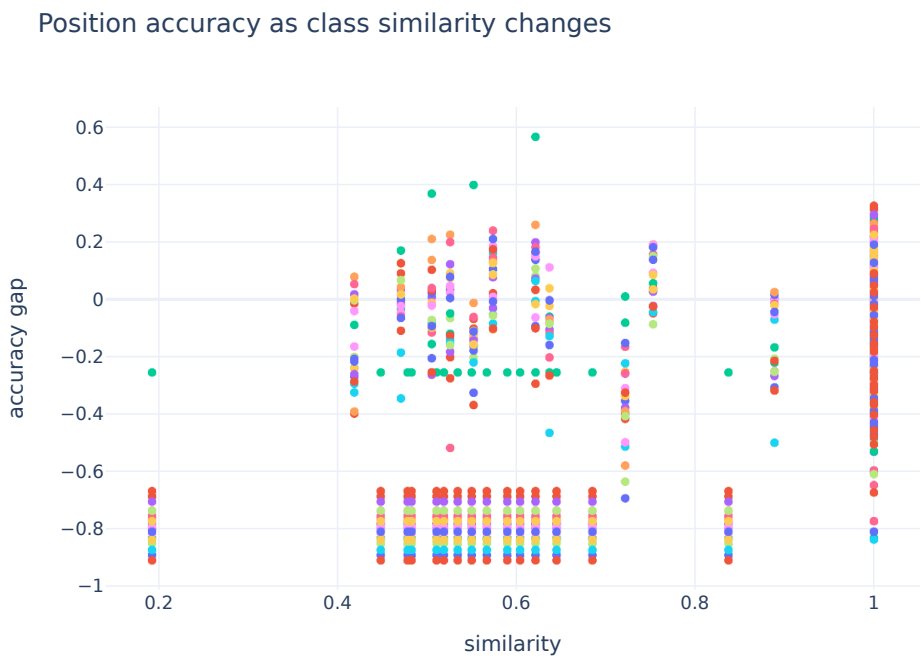


Figure A25: Position gap as class similarity to nearest neighbor increases to classes seen varying during training.

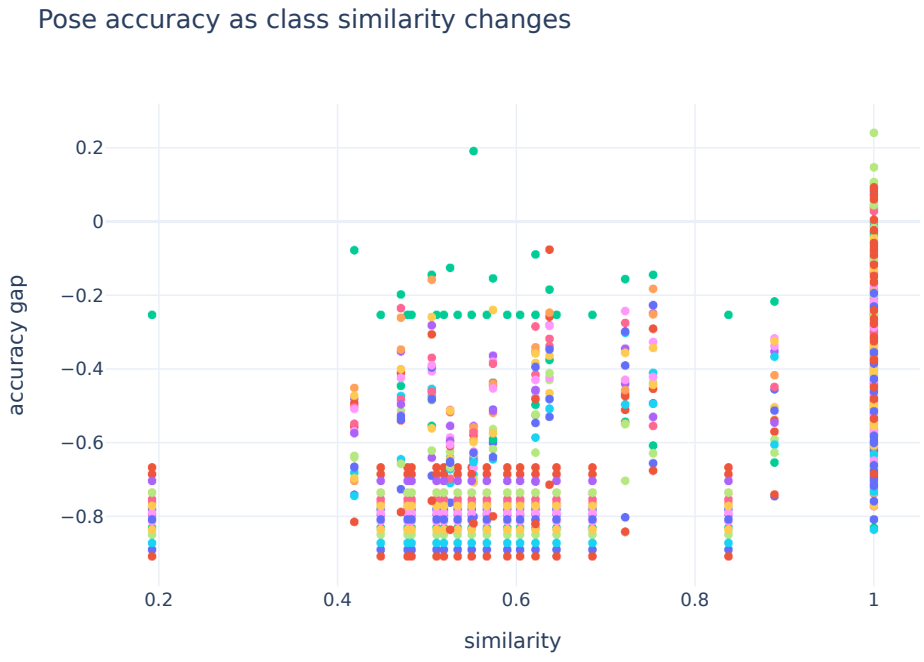


Figure A26: Pose gap as class similarity to nearest neighbor increases to classes seen varying during training.

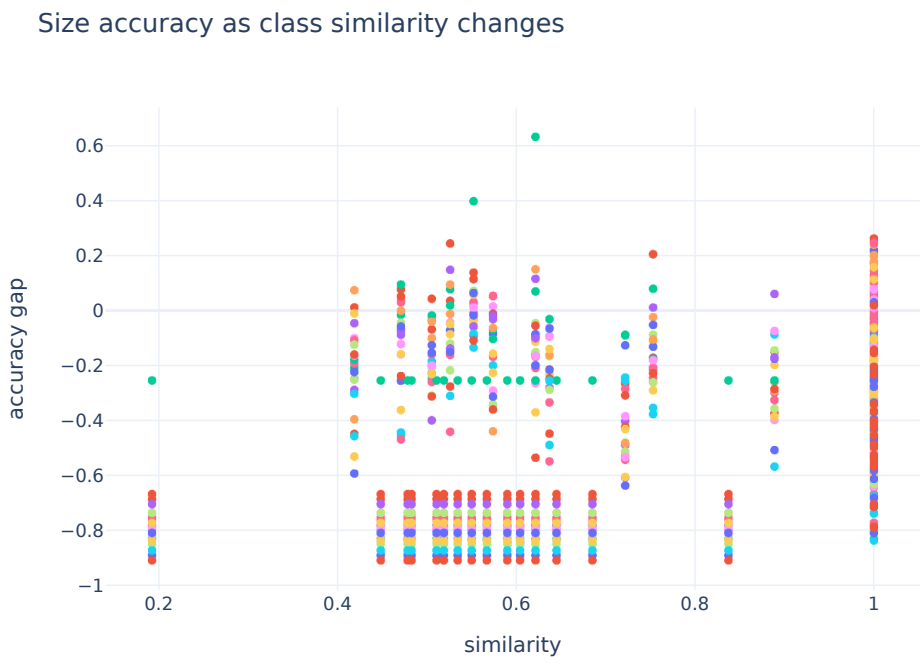


Figure A27: Scale gap as class similarity to nearest neighbor increases to classes seen varying during training.

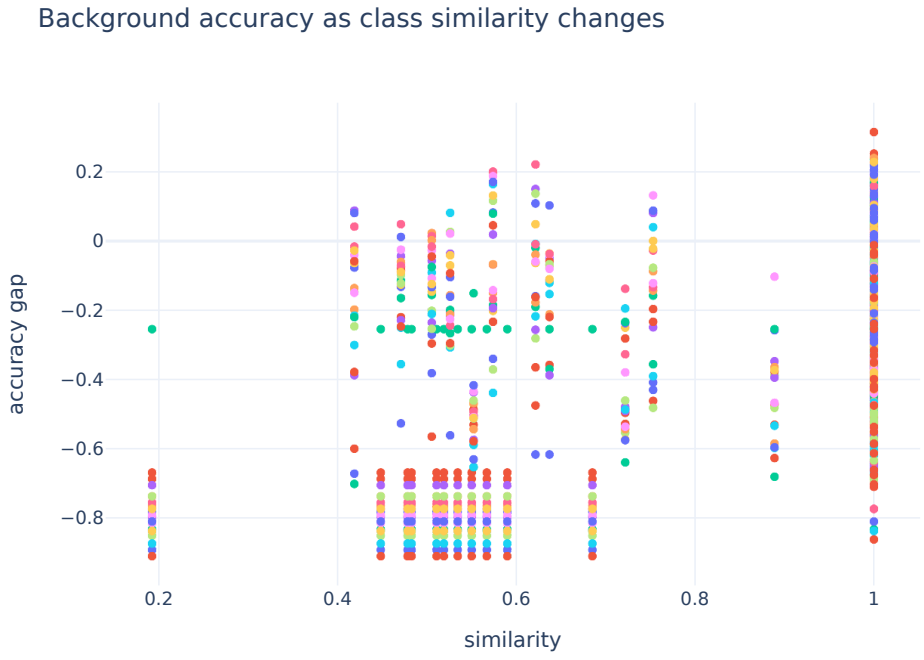


Figure A28: Background gap as class similarity to nearest neighbor increases to classes seen varying during training.

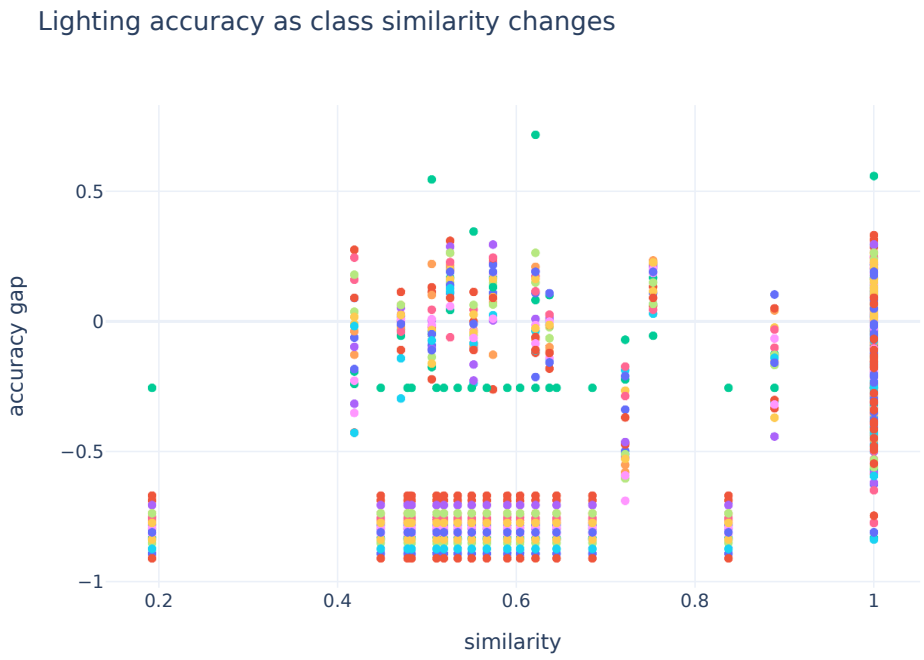


Figure A29: Lighting color gap as class similarity to nearest neighbor increases to classes seen varying during training.