Without any human intervention, evolving multiple CBM algorithms with Escher yields state of the art performance.

ViT-L-14	LM4CV	LM4CV+ESCHER
CIFAR-100	84.48	**89.63**
CUB-200-2011	63.26	**83.17**
Food101	94.77	**94.90**
NABirds	76.58	**78.21**
Oxford Flowers	94.80	**96.86**
Oxford IIIT Pets	92.50	**92.86**
Stanford Cars	86.84	**93.76**

Table 1. Performance of LM4CV [23] and LM4CV evolved with ESCHER on multiple fine-grained classification problems. ESCHER improves upon LM4CV's performance in all datasets while utilizing no extra human annotations.

Dataset	CbD	CbD+ESCHER
CIFAR-100	76.20	**77.80**
CUB-200-2011	62.00	**63.33**
Food101	93.11	**93.58**
NABirds	53.61	**54.30**
Oxford Flowers	79.41	**81.37**
Stanford Cars	75.65	**77.14**

Table 3. Performance for Classify by Descriptions (CbD) [15] and CbD evolved with ESCHER on multiple fine-grained classification datasets in a zero-shot learning setting. CbD+ESCHER improves upon CbD's performance in all datasets.

Naively sampling labels from an LLM does not help; feedback from visual critic is key.

ViT-L-14	LM4CV	LM4CV+ Many Concepts	LM4CV +ESCHER
CUB-200-2011	63.26	66.09	**83.17**
Food101	94.77	94.77	**94.90**
Stanford Cars	86.84	86.84	**93.76**

Table 4. Results on an ablation of ESCHER's library learning component. For LM4CV, we replace the concepts learned with library learning with an equal number of concepts sampled from an LLM. We find that concepts evolved with ESCHER still outperform naively sampling more concepts – suggesting that feedback from a VLM critic is essential for LM4CV+ESCHER's performance.

Qualitative Results

This is a **Male Ring-necked pheasant.**

With no iterations, the model confuses this for a **Female Ring-necked pheasant** because:

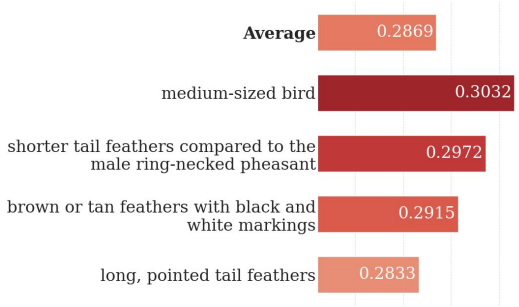
While the **true class** has lower aggregate activation because:

After 5 iteration with Escher, the model predicts this as an **Male Ring-necked pheasant** because:

Qualitative Results

This is a **Male Ring-necked pheasant.**

With no iterations, the model confuses this for a **Female Ring-necked pheasant** because:



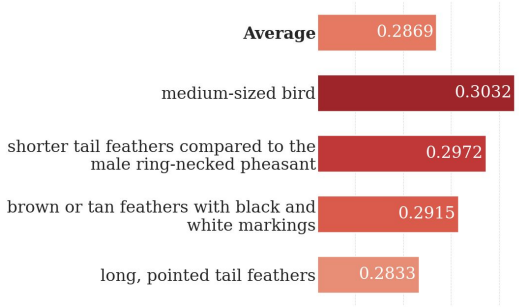
While the **true class** has lower aggregate activation because:

After 5 iteration with Escher, the model predicts this as an **Male Ring-necked pheasant** because:

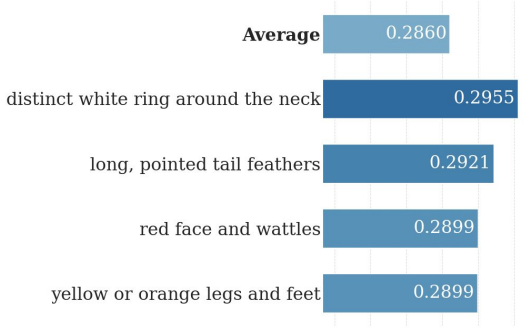
Qualitative Results

This is a **Male Ring-necked pheasant.**

With no iterations, the model confuses this for a **Female Ring-necked pheasant** because:



While the **true class** has lower aggregate activation because:

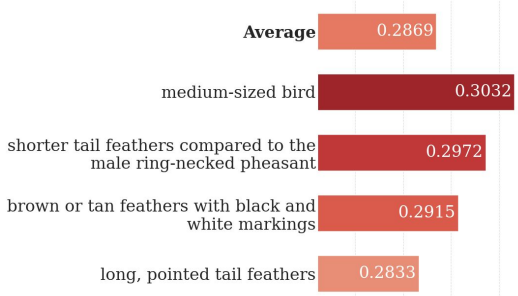


After 5 iteration with Escher, the model predicts this as an **Male Ring-necked pheasant** because:

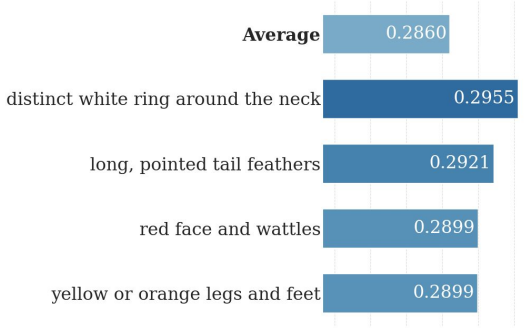
Qualitative Results

This is a **Male Ring-necked pheasant.**

With no iterations, the model confuses this for a **Female Ring-necked pheasant** because:



While the **true class** has lower aggregate activation because:



After 5 iteration with Escher, the model predicts this as an **Male Ring-necked pheasant** because: