# Lightweight Vision-Language Modeling for Fine-Grained Bird Species Reasoning

by Treehouse: Shufan He (she68), Yitong Liu (yliu336), Yunjia Zhao (yzhao291)

## Introduction

We developed a lightweight Vision-Language Model (VLM) for fine-grained Visual Question Answering (VQA) on bird species. Our model combines a compact ViT-based vision encoder—trained from scratch on a bird-specific dataset—with a pretrained language model using a simple linear projection.

Figure 1. Bird Images from CUB-200-2011

## Data

**CUB-200-2011:** Contains 11,788 images across 200 bird species with species labels. Used to train the vision encoder for classification.

**Bird VQA Dataset (via Hugging Face):** Extends CUB with 10 natural language descriptions per image and species labels. Supports multimodal training for captioning and VQA, enabling alignment of visual features with rich textual data.

## Methodology

**Vision Encoder:** A ViT-Tiny architecture trained from scratch on the CUB-200-2011 dataset with strong regularization to learn domain-specific features for bird classification.

**Language Model Integration:** Image embeddings from the frozen ViT encoder are projected into the space of a pretrained language model using a two-layer MLP. The language model is then fine-tuned on VQA-style question-answer pairs generated from bird attribute annotations to enable answering questions about bird classification and feature descriptions.

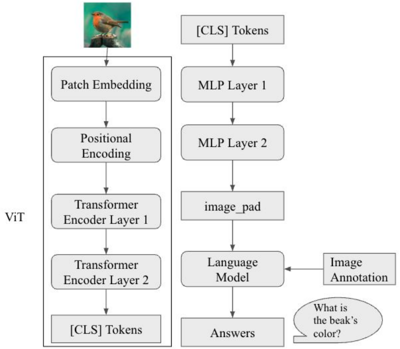


Figure 2. Model Architecture

## Results

| Parameter | Configuration Options | Best Value | Val accuracy |
|---|---|---|---|
| Learning Rate | 1e-3,1e-4,1e-5 | 1e-3 | 11.32% |
| Batch Size | 16,32,64 | 64 | 8.13% |
| pooling | cls,weighted pooling | weighted pooling | 13.5% |
| Weight Decay | 0.1, 0.001 | 0.1 | 7.85% |
| Augmentation | minimal,standard | standard | 6.35% |

| Model | Accuracy in Classification | Accuracy in Question Answering (keyword matching) |
|---|---|---|
| ViT | 8.8% | NA |
| ViT + weighted pooling | 13.5% | NA |
| Pre-trained ViT | 73.2% | NA |

## Discussion

The primary challenge we faced was **severe overfitting** when training the ViT-Tiny model from scratch. Despite achieving perfect training accuracy (100%) after 100 epochs, the validation accuracy remained extremely low (~10%), indicating the model was failing to generalize. Early attempts at regularization and model tuning only modestly improved validation accuracy by ~3%. However, when we switched to a pretrained ViT, validation accuracy dramatically increased to ~70%, highlighting the importance of leveraging pretrained representations in low-data regimes.

Despite the challenges, our project revealed several useful lessons. Our modular design enabled us to easily swap between a vision encoder trained from scratch and one initialized with pretrained weights, which proved critical. We also find pretraining to be essential: the pretrained ViT backbone significantly outperformed the scratch-trained version, highlighting the value of strong visual representations in low-data regimes.