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

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

Github Link

Introduction

This project aims to develop a domain-specific, lightweight Vision-Language Model (VLM) for Visual Question Answering (VQA) focused on bird species identification. We were inspired by the ideas introduced in [1], [2], and [3]. These designs reflect a trade-off between simplicity (linear layers), controllability (cross-attention), and dense fusion (token merging). Our project focuses on the simplest strategy, linear projection for lightweight and interpretable modeling. This project explores training a compact ViT-based image encoder on a bird-only dataset and connecting it with a language model through a simple and interpretable modality integration strategy.

Our motivation is to study whether lightweight VLMs, using minimal connection mechanisms such as a linear projection-based integration, can still achieve strong performance in fine-grained visual reasoning tasks. We focus on the bird domain to take advantage of its rich attribute annotations (e.g., color, shape, part-level details) and to explore VLM interpretability in specialized settings.

Datasets

CUB-200-2011[4]: The Caltech-UCSD Birds 200 (CUB-200-2011) dataset contains 11,788 images spanning 200 bird species, each annotated with species-level class labels. We used this dataset to train and evaluate our vision encoder for fine-grained bird classification. Its well-curated structure and attribute annotations make it a widely used benchmark for visual recognition tasks in ornithology.

Figure 1. Cardinal Images from CUB-200-2011

Bird VQA Dataset (via Hugging Face) [5] : To support multimodal training, we extended CUB-200-2011 using the Bird VQA dataset, which provides 10 natural language descriptions per image in addition to species labels.

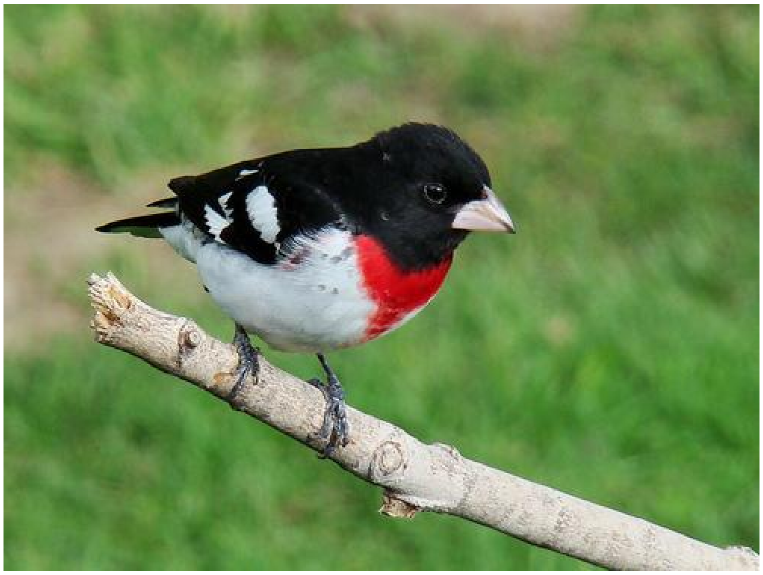

Figure 2. Rose_Breasted_Grosbeak

```
Description: "this bird has a read throat and a white belly. the bird
has a throat that is red and a black crown. a small bird with a black
head, red throat, and white belly. this bird has a white belly and
breast with a red neck and black crown. a small sized bird that has a
white belly and a red chest marking this bird has wings that are
black with a white body this small bird has a red breast with a white
belly and black crown a bird with a black head, red neck and white
body and black and white wings. this bird has a black crown as well
as a tan bill. this bird has wings that are black and has a red
chest"
Label: 57
```

Methodology

1. Preprocessing

To prepare the data for training, we reorganize the CUB-200-2011 dataset into a structured format compatible with standard vision model pipelines, applying consistent resizing and normalization to all images. The dataset is split into training and validation sets based on provided metadata, with images grouped by class for use in fine-grained classification.

For the language model component, we transform bird attribute descriptions into natural language question-answer pairs by using the local llama 3 model with the prompt.

```
You are a helpful assistant for bird visual question answering.
Given the bird description below, generate 3-5 diverse QA pairs.
Ask about its color, size, body parts, behaviors, and overall
appearance.
ONLY respond with a JSON list like this:
[
  {{"question": "...", "answer": "..."}},
  ...
]
Do not use "or", "/", "and", or "&" or " to list multiple answers.
Provide a single best answer or separate answers clearly.
```

These pairs are used both to fine-tune the language model and to evaluate its ability to generate grounded responses. This setup enables assessment of visual-linguistic alignment in a fine-grained, attribute-rich domain.

```
{  "file_name": "American_Redstart_0066_102774.jpg",

  "description": "a black bird with bright orange coverts and upper
coverts... this bird is black with orange and has a very short
beak.",

  "question": "How long is its beak?",

  "answer": "very short"}
```

2. Model Architecture

Our proposed model integrates a vision encoder with a pretrained language model to enable fine-grained visual question answering (VQA) in the bird species domain.

Vision Encoder:
We adopt a ViT-Tiny architecture as the visual backbone, which is trained from scratch on the CUB-200-2011 dataset—a benchmark for fine-grained bird classification. To promote generalization and prevent overfitting on this relatively small dataset, we apply strong regularization techniques during training, including RandAugment for data augmentation, Dropout for stochastic feature masking, and Stochastic Depth to regularize the depth of the network. The encoder outputs a fixed-dimensional embedding for each image, either using the

[CLS] token or a weighted pooling over all patch tokens, depending on the configuration.

Language Model Integration:

To bridge the visual and textual modalities, we project the frozen visual embeddings through a two-layer MLP projection head, which maps them into the text embedding space as input to a pretrained language model. The language model is a generative transformer downloaded from Hugging Face and fine-tuned on synthetic question-answer pairs generated from structured attribute annotations in the CUB dataset. During training, only the two MLP layers are updated, while the vision encoder and language model remain frozen. This approach enables the system to ground visual inputs in natural language without requiring significant computational resources, making it feasible within the constraints of a final project.

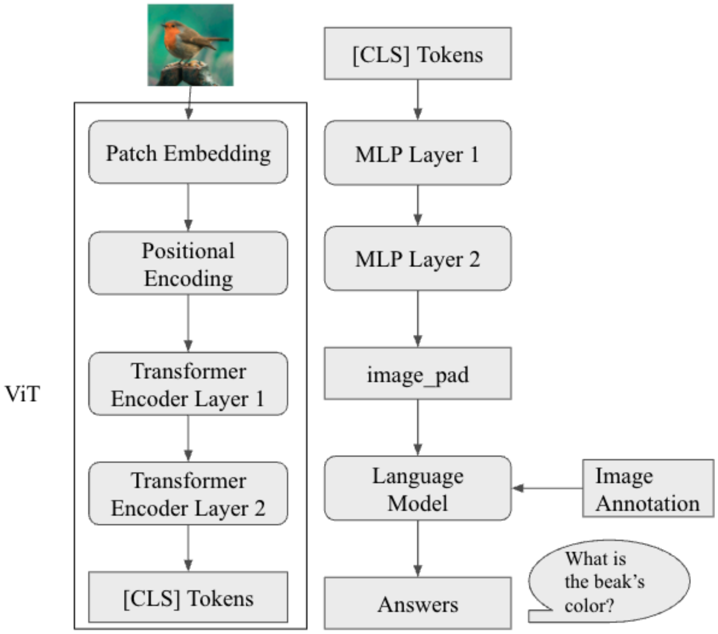

Figure 3. Model Architecture Combining ViT with Pre-Trained Language Model.

Results

We conducted a series of ablation experiments varying key hyperparameters for the ViT-based image encoder. As shown in Table 1, we experimented with different learning rates, batch sizes, pooling strategies, weight decay values, and data augmentation levels. Each configuration was evaluated on bird classification tasks, and the best validation accuracy was reported for each setting. Notably, a learning rate of 1e-3, batch size of 64, and weighted pooling produced the highest individual performance gains, with the weighted pooling strategy achieving 13.8% validation accuracy. This suggests that more nuanced feature aggregation improves the model's ability to distinguish fine-grained categories.

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,CLS and weighted, weighted pooling	Weighted	13.8%
Weight Decay	0.1, 0.001	0.1	7.85%
Augmentation	minimal,standard	standard	6.35%

Table 1. ViT Performance with Different Parameters

In Table 2, we compare the performance of different ViT configurations, including a ViT model we built and trained from scratch, the same ViT augmented with weighted pooling, and a fully pretrained ViT. While the scratch-trained ViT models showed limited performance, the pretrained ViT reached a significantly higher accuracy of 73.2%. This indicates that pretraining on large-scale vision datasets can yield representations that transfer well to fine-grained classification tasks in specialized domains like bird classification.

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

Table 2. Different ViT Models' Performance

In our evaluation of the lightweight Vision-Language Model for fine-grained bird species reasoning, we apply a combination of keyword match accuracy, full sentence exact match accuracy, and partial word overlap to assess the performance of our model. These metrics capture the VLM's ability to complete visual question answering.

Keyword Match Accuracy: This metric assesses whether the model's output includes any of the key attributes related to the bird species in the question-answer pairs. A high keyword match indicates that the model is able to correctly identify important attributes such as color, shape, or patterns in the bird images. For example, a query like "What is the color of the bird's tail?" should have a correct prediction if the tail color is mentioned in the answer, such as "red." Our results showed that the model achieved a keyword match accuracy of 43%, indicating that it was able to extract relevant descriptive cues in nearly half of the evaluated examples.

Full Sentence Exact Match Accuracy: This metric measures how closely the model's prediction matches the entire ground truth sentence. Given the complexity of fine-grained visual reasoning, a perfect match is difficult, but this measure helps gauge how accurately the model can replicate the specific attributes mentioned in the ground truth answer. In our case, the full sentence exact match accuracy was 0%, suggesting that the model often diverged from the precise phrasing or details of the expected responses.

Partial Word Overlap Average: This metric looks at the overlap between the predicted text and the ground truth, quantifying how much of the predicted answer aligns with the correct answer in terms of shared words. This evaluation allows us to consider cases where the model may not generate the exact wording but still captures significant portions of the expected answer. For example, if the model predicted "tail" instead of the full answer "It has a curved tail," it would score 0.2 for a partial overlap. The model achieved a partial word overlap average of 5.61%, highlighting that while exact matches were rare, the model occasionally generated answers that partially aligned with the ground truth content.

Model	Vision-Conditioned LLM (ViT + LM)
Keyword Match Accuracy	43%
Average Partial Overlap:	5.61%
Full Sentence Exact Match Accuracy:	0

The keyword match accuracy yields 43% after various rounds of training and prompt engineering. This indicates that the VLM is capable of extracting some key features in fine-grained tasks like bird species classification. In contrast, full sentence match accuracy and partial word overlap is very low, suggesting that the model is not generating text that captures the specificity of ground truth answers.

Challenges

Throughout the project, we encountered substantial technical and research design challenges across both the vision and language components. We made extensive efforts to address these issues through systematic experimentation, architectural tuning, and careful evaluation design.

On the vision side, we initially trained a ViT-Tiny model from scratch on the CUB-200-2011 dataset. Despite strong regularization techniques such as RandAugment, Dropout, and Stochastic Depth, the model severely overfit: training accuracy reached 100%, while validation accuracy stagnated around 10%. We explored switching to a larger dataset with 48,000 images [6], but it unexpectedly underperformed on validation, likely due to domain mismatch or lower data quality. We ultimately chose to continue with CUB-200-2011 and focused on tuning our architecture. Strategies like adjusting learning rates, batch sizes, and weight decay, as well as replacing the standard CLS token with a weighted pooling strategy, helped improve validation accuracy from 8.8% to 13.5%.

In the language component, we developed a lightweight vision-language model (VLM) by projecting ViT-based visual embeddings into the input space of a pre-trained language model, Phi-3 Mini. Training these projection layers directly on our bird QA dataset led to poor generalization, likely due to the limited size of the data. To address this, we experimented with pretraining the VLM on the broader CC3M dataset. This introduced additional engineering challenges—handling large-scale data, resolving out-of-memory (OOM) issues, and ensuring stable optimization. We mitigated these with mixed-precision training, smaller batch sizes, adaptive learning rates, and sampling from smaller CC3M subsets. We also explored increasing the number of <|image_pad|> tokens to better align ViT patch outputs with the language model input.

A critical discovery was that the Tiny ViT weights we initially used had been pre-trained on CUB-200-2011 itself. Even though the encoder was frozen, it leaked task-specific information, artificially inflating early-stage performance. We corrected this by reinitializing the model with ImageNet-pretrained weights to ensure a fair and unbiased evaluation. Despite retraining and careful monitoring of loss dynamics, the VLM's performance remained modest—reflecting the inherent difficulty of fine-grained visual question answering (VQA). Unlike generic object recognition, fine-grained tasks require subtle visual distinctions and precise attribute grounding, which are hard to achieve without domain-specific supervision.

In retrospect, we realized that our architecture further limited performance - both the ViT and the language model were frozen during training, with only the projection layers being updated. This constrained the model's ability to learn new visual-linguistic mappings. In addition, the evaluation dataset consisted of only 100 samples, making the results more sensitive to noise and harder to generalize. Low performances in metrics like exact match and partial overlap revealed that while the model can produce plausible responses, it often failed to align precisely with the ground truth.

Lastly, we faced important research design challenges in prompt engineering and evaluation. Defining effective prompts for QA tasks required balancing specificity with generality across attribute types like color, size, and shape. Even minor changes in phrasing had notable impacts on model output. On the evaluation side, conventional metrics like exact match proved too rigid, while lexical overlap metrics missed semantic nuances. We adopted a combination of keyword matching, partial word overlap, and manual inspection to better reflect the model's ability to ground visual features in language.

Reflection

Working on this project helped us better understand the challenges in building vision-language systems, especially when dealing with detailed tasks in areas with limited data. One of the most important lessons we learned was that strong generalization requires strong foundations, especially in the form of robust, pretrained models. Training a ViT from scratch on a narrow dataset like CUB-200-2011 proved to be ineffective and limiting even after applying extensive regularization. In contrast, transfer learning from pretrained vision models is a more promising path. This suggests that general-purpose representations are strong and useful when domain-specific data is limited.

We also realized that vision-language alignment is delicate and highly dependent on data quality and architectural design. Simply projecting visual features into a language model's input space is insufficient unless the features are well-structured and semantically rich. This challenge was further compounded by the difficulty of language generation conditioned on high-dimensional, often noisy visual inputs. Our experience showed how easily misalignment can occur in multimodal setups, especially when the training data lacks diversity or specificity.

Despite modest results, this process taught us the importance of iterative experimentation, principled debugging, and thoughtful design in tackling ambiguous, open-ended tasks. Ultimately, this project deepened our understanding of multimodal modeling and highlighted clear opportunities for future improvement, such as using more diverse and domain-relevant datasets, incorporating architectures with explicit cross-modal attention mechanisms, and adopting training regimes that balance contrastive objectives with generative alignment.

For future work, we could explore more advanced vision encoders and multimodal fusion techniques that go beyond simple feature projection, such as cross-attention or contrastive alignment methods. We're also interested in leveraging larger, more diverse datasets to improve generalization in low-resource domains. Additionally, experimenting with more flexible language decoders and prompt-based learning could help generate more accurate and context-aware outputs.

References

[1] Haotian Liu, Chunyuan Li, Qingyang Wu, et al. "Visual Instruction Tuning." *arXiv preprint arXiv:2304.08485*, 2023. https://arxiv.org/abs/2304.08485

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. "Flamingo: Visual language models with frozen vision and language models." *arXiv preprint arXiv:2204.14198*, 2022. https://arxiv.org/abs/2204.14198

[3] Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, et al. "Fuyu-8B: Open-access multimodal models for understanding and generating text and images." *arXiv preprint arXiv:2310.08491*, 2023. https://arxiv.org/abs/2310.08491

[4] Wah, Catherine, et al. "The Caltech-UCSD Birds-200-2011 Dataset." *California Institute of Technology*, 2011. https://www.vision.caltech.edu/datasets/cub_200_2011/

[5] Alkzar90. "CC6204-Hackaton-Cub-Dataset." *HuggingFace Datasets*, 2023. https://huggingface.co/datasets/alkzar90/CC6204-Hackaton-Cub-Dataset

[6]126.7k images https://images.cv/dataset/bird-image-classification-dataset