Fighting Spurious Correlations using Language Guided Abstractions

Shika Rao, Nikhil Kommineni, Musonda Sinkala, Manasvin Anand
New York University

December 20, 2024

Abstract

Deep learning models often rely on spurious correlations which are inherently irrelevant to a vision task. But this task is an easy one for humans as we generalize better with few learning examples by learning to identify the underlying reasoning in the visual space through comparison and distinguishing factors. In this work, we propose a novel approach to improve the robustness of small MLLMs (e.g CLIP) to spurious correlations by distilling extracted latent reasoning as auxiliary text information from large MLLMs (e.g LLaVA). This method improves vision-language alignment through a multimodal contrastive loss that explicitly attributes characteristic features of the target object with the object itself in the vision and language modality. On the Waterbirds dataset, our experimental results show that our model achieves a 51.74% increase in worst-group accuracy when compared to pre-trained CLIP with ViT-L/14 backbone. Additionally, we qualitatively demonstrate that our method directs the model's activation maps and clusters the image features towards the actual class, rather than spurious attributes.

Keywords: Spurious Correlations; Knowledge distillation; Vision Language Models

Introduction

Multimodal Large Language Models (MLLMs) have proven effective across various application domains, including robotics [Peng et al., 2024] and self-driving cars [Cui et al., 2023]. However, the size of these models can limit their deployability in such domains. This creates a need for training smaller MLLMs [OpenAI, n.d.], either by compressing larger models [Zhu et al., 2024] or by distilling knowledge into smaller models [Petryk et al., 2022]. Compared to multimodal models like LLaVA[Liu et al., 2023], CLIP[Radford et al., 2021] is a smaller model that is commonly used as a backbone for a wide variety of tasks [Rombach et al., 2022, Lin et al., 2023]. To improve the vision-text alignment and robustness of smaller models like CLIP, knowledge distillation from larger MLLMs is a promising approach, as these larger models contain more latent information that can enhance the performance of smaller models. However, training smaller models in the same way as larger models often results in worse performance due to differences in the size of their function spaces [Cho and Hariharan, 2019].

A significant challenge in vision tasks is that models often learn spurious correlations when trained on biased data. For example, a classifier designed to distinguish between a "landbird" and a "waterbird" might wrongly correlate the background with the bird's category, leading to failures when birds appear outside their usual habitat. As seen in Figure 1, this issue also arises in text-to-image generation models like Stable Diffusion, which uses the CLIP model's text encoder. Improving CLIP's robustness to spurious correlations is critical, as it directly impacts the performance of such generation models. Therefore, there is a need to develop better training frameworks to address the alignment of vision and language in models like CLIP.

In our project, we hypothesize that larger MLLMs like LLaVA perform better than smaller models on spurious correlation tasks because they contain latent visual reasoning information that leads to the identification of important image features. Recent literature on Chain of Thought prompting demonstrates that prompts like "Solve this step-by-step" are effective in reaching correct solutions. Similarly, in our work we improve the representation learning and alignment of smaller MLLMs using prompted Visual Chain of Thought reasoning as auxiliary information. This approach mimics human learning, where understanding the "logic" behind a visual representation enables learning from less data. By prompting large MLLMs to extract "rules," "formulas," or "defining characteristics" through text, we apply this reasoning to better align the text and visual domains of smaller MLLMs like CLIP, thus improving model performance even with a lesser amount of data. As CLIP is commonly used as a backbone, accurately encoding robustness to spurious correlations in the model can lead to benefits in many downstream tasks, such as compositional text-to-image generation (Figure 1).

CLIP may associate an image of an elephant with the text "elephant," without distinguishing between specific features like the trunk, ears, or tusks. This leads to mismatched associations which further leads to spurious correlations. In contrast, our approach leverages reasoning information to create a more explicit, one-to-one mapping between image and text encodings. Specifically, we go beyond merely matching the image of an elephant to the text "elephant." Instead, we match key defining features of the elephant—such as its trunk, big ears, and tusks—directly to corresponding descriptive text and the text class "elephant" itself. This detailed, feature-specific alignment reduces the potential for spurious correlations. As seen in Figure 2, we can see that CLIP has a many-to-one mapping from numerous irrelevant features in the image to text encodings on a sample image from the Waterbirds dataset, leading to spurious correlations. On the other hand, our model displays a more one-to-one mapping between the relevant features of the image and the text class.

Thus, the key contributions of our project are-

1. Developing a novel framework to extract latent reasoning information from large MLLMs and distilling this information to smaller MLLMs to improve their representation learning and alignment.

2. We achieve near State of The Art (SOTA) performance on the Waterbirds dataset with significantly less training data.

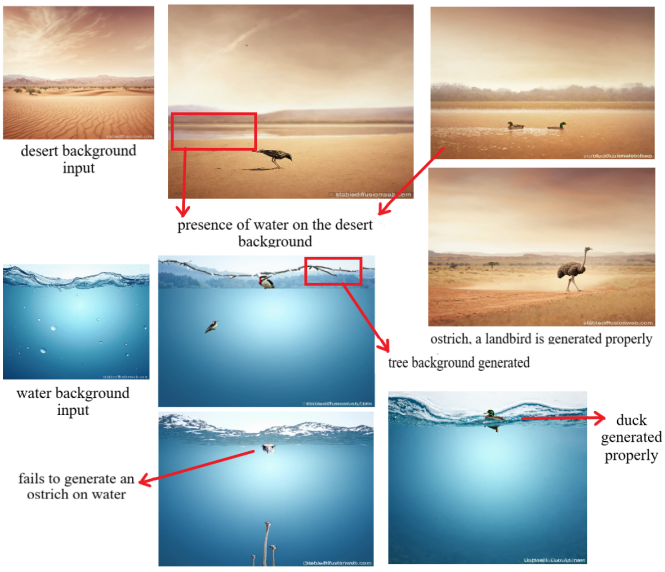
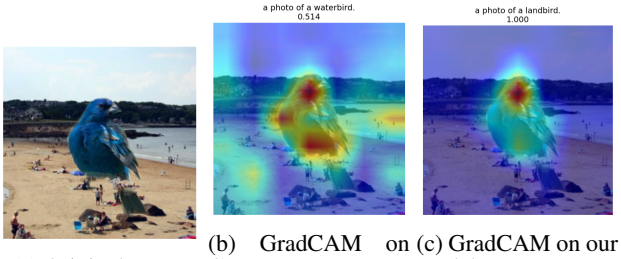


Figure 1: When we input a desert background to Stable Diffusion and prompt it to generate "an image of a bird with webbed feet", we can observe the water background (the spurious correlation) generated along with the bird on the desert background. Similarly, this can be observed in other cases where we give a water or desert background as input and prompt the model to generate "an image of a duck", "an ostrich", and "a woodpecker".



(a) Original Image (b) GradCAM on CLIP. (c) GradCAM on our model.

Figure 2: The landbird in the image is spuriously correlated with the background, which leads the pre-trained CLIP model to predict the class as waterbird. Our approach de-correlates the spurious relationships in the loss function via language which leads to correct class prediction and qualitatively the model identifies only the bird in the image.

Related Work

1. **Language as information for visual tasks:** We draw inspiration from prior works that leverage language in vision learning systems. In [Shtedritski et al., 2023], the authors show that CLIP is better at visual classification when objects are highlighted with a red circle around it. In [Lei et al., 2024], the authors show that the scaffolding technique i.e., overlaying an image with coordinates helps a model better associate relevant texts to the image. These works indicate that directing a model to specific visual features enables the model to extract the information latent in it. In another line of work, authors in [Zheng et al., 2023] came up with a prompting method to extract reasoning information from MLLMs to solve the task of multimodal Chain of Thought reasoning. In our work, we would like to apply the benefits of Visual Chain of Thought reasoning to solve the problem of spurious correlations.

2. **Robustness to Spurious Correlations in Vision-Language models:** In our work, we identify visual characteristics of the object through text and attribute these features with the image in the classification task to solve the problem of spurious correlations. In a line similar to ours, the authors in [Yang et al., 2023] propose a framework to directly finetune the CLIP model with contrastive losses after detecting spurious the spurious attribute with supervision. However, though the authors leverage text in their work, it is limited to the object class text and the text corresponding to the spurious attribute. We try to leverage auxiliary reasoning descriptions of the object to better align CLIP. In [Zhang et al., 2023], the authors provide theoretical and empirical proofs of the closeness of the text and image representations in CLIP's latent space, driving more insight into the organization and geometry of CLIP's latent space. It further extends this idea to propose that text representations can be effectively used as a proxy for image representations in classification tasks. However, DrML does not attempt to fix the misalignments in CLIP's latent space, but only finetunes the overhead classification layer to handle spurious attributes better. In our work, we achieve performance close to the numbers reported in [Yang et al., 2023, Zhang et al., 2023] with just 25% of the supervised data.

Methodology

In this work, we teach models to focus on relevant parts of an image using text instructions as a guide, to rectify spurious correlations.

We utilize the pretrained world knowledge of Multimodal Large Language Models (MLLMs) to identify and highlight distinguishing features of an image relevant to a task. Our framework consists of the following steps:

1. **Extracting Knowledge from the Teacher Model:** We use a pretrained large MLLM (i.e., LLaVA) as the Teacher model to extract relevant features from images.

Prompts are carefully designed to capture visual reasoning abilities, assuming that large MLLMs are effective at both performing tasks and explaining their reasoning, as suggested by Visual Chain of Thought (VCoT)[Rose et al., 2024] literature. Sample outputs from LLaVa look like,

- *Prompt*: What is the common name of the bird in the image? Give keywords to identify and describe physical traits of the waterbird or landbird class it belongs to. Ignore the background. Be concise.
  Response: Seagull, pointed beak, streamlined body, webbed feet.

2. **Training the Student Model:**
   The extracted image description responses serve as *auxiliary text data* for training the Student model (i.e., CLIP). The Student model aligns the text representations of these responses with the corresponding class label representations in the latent space. Furthermore, the image encoder of the Student model is fine-tuned to align image representations with text representations of these auxiliary image descriptions.

Text-Guided Contrastive Learning The novelty in our approach lies in leveraging supervised contrastive learning and text-guided abstractions to address spurious correlations. Unlike the InfoNCE[Radford et al., 2021] loss function typically used in CLIP, which considers a single positive sample for an anchor, we adopt Supervised Contrastive Loss[Khosla et al., 2021] to align multiple positive samples. The loss function integrates supervised labels and contrastive loss for aligning representations effectively by:

- Pulling together image representations, distinguishing features, and corresponding class labels in the shared embedding space, while supporting multiple positives.
- Pushing apart spurious features and their associated class labels, effectively reducing noise, while avoiding false negatives.

We employ *cosine-similarity* metric as the distance measure for the contrastive loss function. Our loss function can be expressed as,

$$\mathcal{L}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_p)/\tau)}{\sum_{a \in A(i)} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_a)/\tau)}$$

Where,

- $\mathbf{z}_i$ and $\mathbf{z}_p$ are the representations of the anchor and positive samples.
- $P(i)$ is the set of positive samples for anchor i.
- $A(i)$ is the set of all samples in the batch excluding the anchor.
- sim() denotes cosine similarity, and τ is the temperature hyperparameter.

Teacher-Student Alignment The alignment process is two-fold:

1. **Text Representation Alignment:** The text encoder of the Student model is fine-tuned to align auxiliary reasoning text representations (generated by the Teacher model) with the corresponding class label representations. This alignment is achieved using a few augmentations of auxiliary text data for fine-tuning.

2. **Image Representation Alignment:** The image encoder of the Student model is aligned with both the auxiliary reasoning text representations and the class label text representations. This alignment is supervised on a small subset of image samples, ensuring that the image encoder learns to associate relevant visual features with their corresponding labels.

Proposed Implementation Figure 3 captures our method. For the scope of this project, we propose to use:

- LLaVA[Liu et al., 2023], a large MLLM, as the Teacher model to generate reasoning-based auxiliary text data.
- CLIP[Radford et al., 2021], a smaller MLLM, as the Student model to perform fine-tuning and alignment tasks.

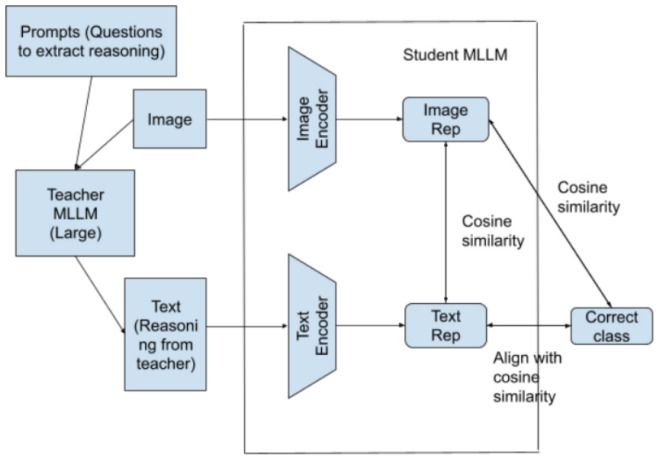


Figure 3: Training Method

Experiments

Setup: CLIP uses two main groups of visual backbones, ResNets(RN) and Vision Transformers(ViT). Typically, ResNet-50 and ViT-L/14@336 are used as prototypes for ResNet and ViT backbones. We use the ViT-B/32 and ViT-L/14 CLIP models as the pretrained models in our experiments for Model 1 and Model 2 respectively. We train these models with a learning rate of 1e-6 for 10 epochs in our experiments.

Dataset	Matching (label and place)	Non-Matching	Total Samples
Train	4555	240	4795
Test	2897	2897	5794
Validation	600	599	1199
Total	8052	3736	11788

Table 1: Summary of matching and non-matching samples across datasets.

Datasets: The primary dataset utilized in our experiments is the Waterbirds Dataset, as described by Sagawa et al., 2019. This dataset was created by extracting bird images and superimposing them onto various backgrounds, allowing us to investigate the spurious correlation between bird species and their respective environments.

The Waterbirds Dataset contains 11,788 images featuring both land and waterbirds across diverse settings. For our experiments, we focused on a subset of 2,000 images to demonstrate the susceptibility of the student model to spurious correlations related to the background-bird type correlation.

During training, the student model is primarily exposed to samples where the bird type labels align with their backgrounds (e.g., landbirds on land and waterbirds in water). We then evaluate the model's performance on a larger set of images where the bird type and background do not match, aiming to establish confidence in its robustness and generalization capabilities. Table **??** summarizes the split of our dataset.

Extracting information from the Teacher: We use the following prompts on LLaVA to extract text descriptions of the object (bird) and the background in the images for fine-tuning.

- **Bird:** Identify the bird in the image and describe distinguishing features of the bird class in the image. Ignore the background, be concise.
- **Background:** Describe the background in the image, ignore the bird.

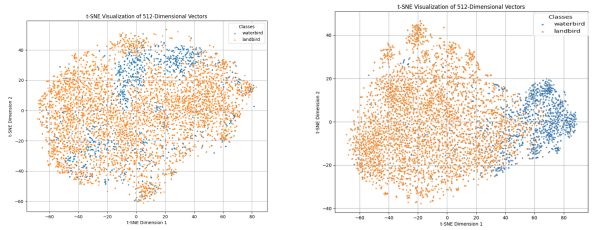
Results

Models 1 and 2 presented in the results table represent the final results after our finetuning. For an thorough overview of the performance metrics across all experimental outcomes, refer to Table 4 in the appendix.

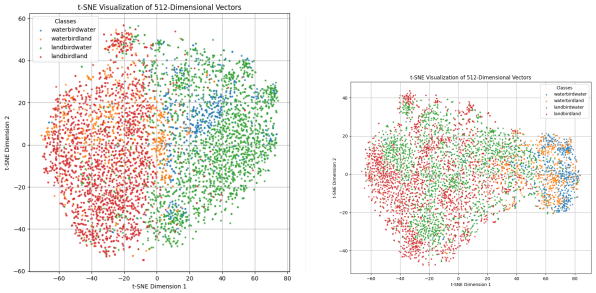
The results listed in Table 2 offer valuable insights into the performance variations among various models applied to the Waterbirds dataset. The fine-tuned CLIP model achieves state-of-the-art benchmarks in comparison to the zero-shot LLAVA and zero-shot CLIP models, which exhibit lower performance metrics. Notably, the fine-tuning CLIP with LLaVA (Model 2) enhances overall accuracy considerably, highlighting the effect of integrating auxiliary data. Furthermore, Model 2 exhibits a significant improvement in worst-group accuracy, demonstrating the model's capacity to effectively mitigate the spurious correlations that persist in the zero-shot LLAVA and zero-shot CLIP models.

The t-SNE plots illustrated in Figure 4 further underscore the effectiveness of fine-tuning in mitigating spurious correlations and improving classification accuracy. In the case of the pre-trained CLIP models, representations are predominantly clustered based on the background type rather than the bird type. Specifically, landbirds and waterbirds depicted in land backgrounds are grouped together, while images featuring water backgrounds also form separate clusters regardless of bird-type. This clustering indicates a reliance on background features for classification rather than the intrinsic characteristics of the birds themselves.

Conversely, the fine-tuned CLIP models exhibit a clear separation of representations according on the bird type, independent of the background type. Clear separation is observed between waterbirds and landbirds, indicating the model's enhanced focus on the bird features after the fine-tuning process.



(a) TSNE Plot for pre-trained CLIP (b) TSNE Plot for the fine-tuned CLIP

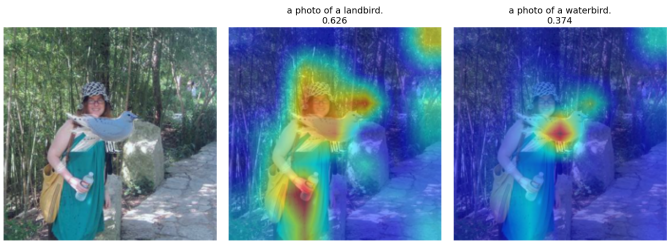


(c) Sub-class TSNE Plot for pre-trained CLIP (d) Sub-class TSNE Plot for the fine-tuned CLIP

Figure 4: 2 and 4 class TSNE plots for the pretrained and fine tuned CLIP. From the 2 class plots, in the pre-trained CLIP there is a lack of a clear separation in representation between waterbirds and landbirds. This hints at the model being unable to accurately classify the birds whereas in the plot for the fine-tuned CLIP there is a clear separation in representation for land and waterbirds hinting at the model being able to classify the bird based on the bird itself. The results for the pre-trained CLIP clearly show that it is spuriously correlating the background with the bird type, as we the images of landbirds on land (RED) and waterbirds on land(ORANGE) clustered together and similar the images with birds on water backgrounds (GREEN and BLUE) are clustered together. In the fine-tuned CLIP, we see much better results, with the clustering split now clearly along the lines of bird type itself and not impacted by the background.

To further gain confidence in the success of methodology in handling spurious correlations, we look at the results of the Gradient-weighted Class Activation Mapping (Grad-CAM, Selvaraju et al., 2019). This approach looks at the gradients

flowing into the final convolutional layer to produce a localization map that visually highlights the regions in our input image that are the most relevant to our model in the classification task. As we can see from Figure 5a, even though the pre-trained CLIP is correct in its classification of the bird as a land bird, the regions of the image it bases its conclusion on is not strictly the bird but also elements from the background such as the trees. Now when we look at the Grad-CAM results from the fine tuned CLIP in 5b, we can clearly see that the area of the image that the model relies on to makes its classification is firmly centered around the bird.



(a) Grad-CAM activation map for pre-trained CLIP

(b) Grad-CAM activation map for fine-tuned CLIP

Figure 5: Grad-CAM activation maps for the pre-trained and fine tuned CLIP models.

Future work and Conclusion

While our work demonstrated substantial performance improvement on the Waterbirds dataset, further experimentation on more complex datasets, such as WILDS-FMoW [Koh et al., 2021], is needed to test our method. Evaluating the impact of our fine-tuned model on the generalization performance of CLIP in text-to-image generation settings is another important experiment that could yield valuable insights. As mentioned in the Introduction, models like Stable Diffusion, which use CLIP's text encoder, suffer from spurious correlations that lead to the generation of incorrect images. We hypothesize that integrating our fine-tuned CLIP text encoder into the Stable Diffusion pipeline could help better handle text prompts that contain spurious correlations.

Another line of exploration is enhancing prompt quality. This can be achieved by applying segmentation masks on input images for LLaVA, allowing it to focus on the relevant regions of the image and refine the generated prompts, improving the quality of responses. Otherwise, we can experiment with stronger teacher models too to further improve prompt quality.

Regarding our method, we currently focus only on the distinguishing attributes of the object in the classification task. However, encoding relative attributes is also crucial for addressing spurious correlations. More subtle differences, such as "smaller size" or "lighter color," are not captured when we focus solely on the attributes. Incorporating another form of text prompting that considers the relative differences between classes and aligns this with the corresponding text representation may help mitigate spurious correlations.

Additionally, testing our framework with various student models other than CLIP, and different teacher models beyond LLaVA, would demonstrate the extensibility and generalizability of our approach in improving any small multimodal model.

Table 2: Results on the Waterbirds dataset.
Model 1: ViT-B/32 Clip, Model 2: ViT-L/14 Clip
Details on these models can be found in the reference section of this report

	Zero shot ViT-L/14 Clip	Zero shot ViT-B/32 Clip	Zero Shot LLAVA	Binary Clip ViT-L/14	Binary Clip ViT-B/32	Fine-tuned Clip*[4]	Model 1	Model 2	DrML
Overall Accuracy	83.69%	86.70%	78.97%	89.35%	80.29%	**97.20%**	90.61%	96.10%	-
Waterbird Accuracy	48.44%	67.30%	70.98%	76.79%	65.11%	-	82.32%	**87.62%**	-
Landbird Accuracy	93.73%	61.80%	68.71%	92.92%	84.61%	-	92.97%	**98.51%**	-
Waterbird on water	63.55%	95.79%	94.24%	92.52%	89.56%	-	**93.46%**	90.50%	97.00
Waterbird on land	33.02%	77.57%	63.71%	61.06%	40.65%	**89.70%**	68.07%	84.74%	55.64
Landbird on water	88.69%	39.33%	42.66%	85.99%	69.98%	-	75.48%	**97.83%**	82.71
Landbird on land	98.94%	84.35%	94.77%	**99.87%**	99.25%	-	97.52%	99.20%	98.93

References

Cho, J. H., & Hariharan, B. (2019). On the efficacy of knowledge distillation.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., Gao, T., Li, E., Tang, K., Cao, Z., Zhou, T., Liu, A., Yan, X., Mei, S., Cao, J., … Zheng, C. (2023). A survey on multimodal large language models for autonomous driving.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., & Krishnan, D. (2021). Supervised contrastive learning.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., … Liang, P. (2021). Wilds: A benchmark of in-the-wild distribution shifts.

Lei, X., Yang, Z., Chen, X., Li, P., & Liu, Y. (2024). Scaffolding coordinates to promote vision-language coordination in large multi-modal models.

Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., & Lin, T.-Y. (2023). Magic3d: High-resolution text-to-3d content creation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 300–309.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning.

OpenAI. (n.d.). Gpt-4o mini: Advancing cost-efficient intelligence [Accessed: 2024-12-19].

Peng, A., Bobu, A., Li, B. Z., Sumers, T. R., Sucholutsky, I., Kumar, N., Griffiths, T. L., & Shah, J. A. (2024). Preference-conditioned language-guided abstraction.

Petryk, S., Dunlap, L., Nasseri, K., Gonzalez, J., Darrell, T., & Rohrbach, A. (2022). On guiding visual attention with language specification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18092–18102.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models.

Rose, D., Himakunthala, V., Ouyang, A., He, R., Mei, A., Lu, Y., Saxon, M., Sonar, C., Mirza, D., & Wang, W. Y. (2024). Visual chain of thought: Bridging logical gaps with multimodal infillings.

Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359.

Shtedritski, A., Rupprecht, C., & Vedaldi, A. (2023). What does clip know about a red circle? visual prompt engineering for vlms.

Yang, Y., Nushi, B., Palangi, H., & Mirzasoleiman, B. (2023). Mitigating spurious correlations in multi-modal models during fine-tuning.

Zhang, Y., HaoChen, J. Z., Huang, S.-C., Wang, K.-C., Zou, J., & Yeung, S. (2023). Diagnosing and rectifying vision models using language.

Zheng, G., Yang, B., Tang, J., Zhou, H.-Y., & Yang, S. (2023). Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models.

Zhu, X., Li, J., Liu, Y., Ma, C., & Wang, W. (2024). A survey on model compression for large language models.

Appendix

Training Configurations

We also experiment with adding an extra linear and mlp layers as classification layers on top of the vision and text encoder. In this configuration, we optionally freeze CLIP's vision and text encoders to test out all possible training configurations. Table 4 and 5 capture these training configurations and their performance results. Furthermore, Network Type *modified_clip* in Table 5 indicates the presence of additional projection layers on top of CLIP's image and text encoders, with Layer Type *linear* indicating the use a linear layer as the projection layer and Layer Type *mlp* indicating the use of a multi layer perceptron as the projection layer. The presence of *probe* indicates that we froze CLIP's image and text encoder and only trained the projection head during finetuning.

Subject	**LLaVA Prompt**	**Sample Output**
Bird	Identify the bird in the image and describe distinguishing features of the bird in the image. Ignore the background, be concise.	The bird in the image is a brown and white bird with a black head. It is flying over a swimming pool, possibly landing on the edge of the pool.
Bird	What is the common name of the bird in the image? Give keywords to identify and describe physical traits of the waterbird or landbird class it belongs to. Ignore the background. Be concise.	Seagull, pointed beak, streamlined body, webbed feet.
Background	Describe the background in the image, ignore the bird. Be concise.	The background in the image is a body of water, which appears to be a lake or a pond.
Background	Give keywords describing the background. Ignore the bird. Be concise.	Mountains, water, city, trees.

Table 3: Prompts and corresponding sample outputs for various subjects.

Table 4: The full set of results from experiments conducted on the Waterbird dataset is presented. Models 1 and 2 achieved the highest performance and were therefore selected as the final models, with their results reported. The remaining results correspond to variations of Models 1 and 2, the details of which have been omitted from this report for brevity.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Overall Accuracy	90.61%	96.10%	85.12%	85.23%	84.28%	85.90%	82.84%
Waterbird Accuracy	82.32%	87.62%	81.31%	80.76%	73.21%	73.05%	65.97%
Landbird Accuracy	92.97%	98.51%	86.21%	86.50%	87.43%	89.56%	87.65%
Waterbird on water	93.46%	90.50%	93.30%	93.46%	91.74%	91.74%	91.12%
Waterbird on land	68.07%	84.74%	69.31%	68.07%	54.36%	54.36%	40.81%
Landbird on water	75.48%	97.83%	75.17%	75.48%	80.40%	80.40%	75.79%
Landbird on land	97.52%	99.20%	97.25%	97.52%	98.71%	98.71%	99.51%

Table 5: Model details

Model	Mode	Network Type, Layer Type for Modified Clip	Background Consider	Include Class Text in Image Training
model 4	text image concat no bg	clip, NA	FALSE	Automatically set to True
model 5	text image concat no bg	modified clip, linear	FALSE	Automatically set to True
model 6	text image concat no bg	modified clip, linear probe	FALSE	Automatically set to True
model 7	text image concat no bg	modified clip, mlp	FALSE	Automatically set to True