

# Leveraging Cross-Modal Neighbor Representation for Improved CLIP Classification

## Supplementary Material

### I. Analysis of Different Types of Texts

In our proposed CODER, each element of it can be regarded as the image’s matching score with corresponding semantics. Therefore, in the following content, we will explain each text from the perspective of semantics to illustrate the advantages of the new text types we propose. Our Auto Text Generator (ATG) introduces three new types of semantics compared to previous works[1, 3]: Analogous Class-based Semantics, Synonym-based Semantics, and One-to-One Specific Semantics. Table 1 shows examples of how our newly proposed semantics modify the original wrong classification results of images. For ease of demonstration, we only display partial results of the image’s matching scores with the original semantics and the new semantics. Next, we will analyze the functions of these semantics one by one.

#### I.1. The Role of Different Types of Semantics

**Analogous Class-based Semantics.** Introducing analogous classes allows the model to utilize the knowledge of the analogous classes and the relationships between classes in image classification. This mimics the human process of identifying objects by comparing the similarity of unknown classes to known classes. Samples 1 to 5 in Table 1 demonstrate the correction of the images’ original predicted classes by the analogous class-based semantics. Samples 1 and 2 show a side view of a biplane where the propeller is not visible, yet the landing gear is clearly visible. Since the landing gear of most modern airplanes is retracted during flight, the attributes ChatGPT generates for airplanes do not include landing gear. In contrast, for helicopters, the landing gear is generally fixed and conspicuous, so its ChatGPT-generated attributes include landing gear. This leads to the error. After we use the analogous class-based semantics of ground-truth class, we get semantics that more closely resemble the object in the image and then successfully correct the result. Similarly, in Sample 5, motion blur in the image led to a mistaken match with a spinning ceiling fan. By using analogous class-based semantics, we compare the similarity of the object in the image with both “Floor Lamp” and “Industrial Fan”. Since the object in the picture is more similar to “Floor Lamp”, we correct the results accordingly. **Synonym-based Semantics.** Samples 6 to 9 in Table 1 demonstrate the correction of samples by the synonym semantics. These samples show that different synonyms for a class result in completely different matching scores for the same test image. By introducing multiple synonym seman-

tics and selecting the maximum score among these, we effectively address the classification errors caused by CLIP’s varying responses to different synonyms for the same class.

**One-to-One Specific Semantics.** Samples 10 to 12 in Table 1 showcase the correction of samples by the one-to-one specific semantics. Observing these samples, we notice that when the actual class of an image is very similar to the incorrectly predicted class by CLIP, these two classes often share many similar semantics. This leads to difficulties for the original semantics proposed in CLIP and VCD to distinguish between these classes, resulting in classification errors. For instance, in Sample 10, both the Staffordshire Bull Terrier and the Boxer share the semantic of short hair; in Sample 11, Siamese and Birman cats both have blue eyes. Our proposed one-to-one specific semantics leverage ChatGPT’s knowledge to generate key distinguishing features for these specific classes. For example, in Sample 10, the one-to-one specific semantics focus on the difference in the underbite between Staffordshire Bull Terriers and Boxers. For Sample 11, ChatGPT notes the difference in fur length between Siamese and Birman Cats. For Sample 12, the one-to-one specific semantics highlight the flat-faced feature of Persian Cats. These semantics are crucial for differentiating between the two similar classes, thereby helping to correct the image’s original predicted class.

#### I.2. Understanding Failure Cases

We also analyze some common errors of our method.

**Analogous Class-based Semantics.** Regarding the analogous class-based semantics, we have identified five common types of errors:

1. The analogous class-based semantics struggle when the analogous class of the incorrectly predicted class closely resembles the true class. In Sample 1 of Table 2, the analogous class for emu is Cassowary, and for llama is Camel. The emu in the image is misidentified as a camel due to its color resembling that of a camel and being set against a desert-like background.
2. When class names are ambiguous, the analogous class-based semantics may incorrectly associate them with a similar class that doesn’t match the image due to the ambiguous meaning of the word. For example, in Sample 2 of Table 2, “bass” refers to both a fish and a musical instrument. Therefore, ATG generates a wrong analogous class, “upright bass”, that doesn’t match the image. To avoid such errors, it’s essential to clarify class names, like changing “bass” to “bass fish” to remove ambiguity.

3. When the object in an image lacks the common semantics of its class, it struggles to match with its analogous classes. As shown in Sample 3 of Table 2, the crab in the image doesn't prominently display typical characteristics, such as two large claws and eight legs, making it difficult to have a high matching score with its analogous classes like King Crab.
4. When the analogous classes of an incorrectly predicted class also appear in the image, it can lead to error. For example, in Sample 4 of Table 2, the analogous class for Sea Horse is Aquarium Decorations and the presence of Aquarium Decorations in the image results in a misclassification.
5. When the label of the image itself is ambiguous, the analogous class-based semantics may also encounter issues. As demonstrated in Sample 5 of Table 2, the image shows an artwork shaped like a dragonfly made from leaves. It's unclear whether to assign the label of a dragonfly or a type of plant to this image, leading to a high matching score with analogous classes like desert plant.

**Synonym-based Semantics.** For synonym-based semantics, when a class name has multiple meanings, its synonyms may correspond to a different meaning that doesn't match the current image, leading to incorrect classification results. For example, in Sample 6 of Table 2, "Bombay" and "Siamese" refer to both cat breeds and place or cultural names, resulting in identified synonyms that may not be cat breeds. This issue can be resolved by clarifying class names to eliminate ambiguity, such as renaming "Bombay" to "Bombay Cat".

**One-to-One Specific Semantics.** Regarding the one-to-one specific semantics, we have identified two types of errors:

1. **The Bottleneck in CLIP's Recognition Ability.** Although the Auto Text Generator (ATG) produces texts highlighting key semantics to differentiate between two classes, CLIP's image-text matching capability may have limitations in certain image-text pairs, preventing the one-to-one specific semantics from producing accurate results. For instance, in Sample 7 of Table 2, ATG identifies that a Bengal Cat has a marbled coat. However, the matching score of the image and the "marbled coat" semantic is low, even though the cat in the image indeed displays a marbled coat. This issue arises from CLIP's inability to precisely calculate the similarity between certain image-text pairs.
2. **The Capacity Limitations of External Experts like ChatGPT.** Although we ask ChatGPT to output key semantics that best distinguish between two classes, ChatGPT may generate incorrect semantics in some cases. For instance, in Sample 8 of Table 2, ChatGPT provides the same semantic for both American Pit Bull Terrier

and Abyssinian, which fails to aid CLIP in correcting the original prediction. This issue arises from LLM's capacity limitations to give the key semantics.

It should be noted that many of the issues mentioned above typically occur only in a minority of hard test cases. In most other test cases, our proposed new semantics have been able to assist CLIP in making correct classifications, as evidenced by the performance improvement of CLIP reflected in Table 1 and Figure 4 of the main text. Moreover, some of these issues can be expected to be effectively resolved in the present or future. For instance, eliminating the ambiguity in class names can address many of the failure cases we just mentioned. With the continuous enhancement of CLIP and Large Language Models, the performance of our method will also improve accordingly.

## II. Analysis of Rerank Stage

In this section, we will further discuss the rerank stage.

### II.1. Reasons of Method Effectiveness

Table 3 illustrates examples of image classification results corrected through the rerank step. Observing the samples in the table, we can see that using the classification scores gap based on one-to-one specific semantics for reranking can correct the original predictions. We believe this is primarily due to three reasons:

1. The semantics that distinctly differentiate a class from various other classes may vary. This leads to the one-to-one specific semantics generated by ATG having a more comprehensive description of the current class, thereby improving classification performance. For example, the most obvious difference between a "cheetah" and a "cougar" is whether there are spots on the body, while the most obvious difference between a "cheetah" and a "snow leopard" is the color of the fur. This means that the one-to-one specific semantics generated for "cheetah" need to focus on both the pattern and color of the fur. And these diverse one-to-one specific semantics can help us describe the class more comprehensively.
2. Our reranking method can be considered as an ensemble of several binary classifiers. These classifiers typically satisfy the "good but different" criterion, ensuring the effectiveness of the ensemble algorithm.
3. Our method effectively utilizes the quantified information of relative advantages between different classes' classification scores. Unlike voting methods that rerank classes based solely on which class receives more votes, our method better captures the model's uncertainty during the process of class prediction. For instance, a small score difference might indicate the model's lack of certainty between two specific class classifications, suggesting that the corresponding classification result may not

be reliable. Therefore, it leads to errors in the answers derived from voting methods.

## II.2. The Choice of $K$

In the reranking stage, we rerank the top  $K$  classes from the initial classification results of the first stage. Here  $K$  is a hyperparameter. Since only those images whose true class are within the top  $K$  of the initial prediction results may potentially be corrected by our method. Therefore, in general, the larger the value of  $K$ , the more images that can potentially be corrected, and the better the performance of the method. However, the total number of one-to-one specific test sets involved in the reranking process of the current image is also increasing, leading to greater costs and computational expenses. Therefore, the choice of  $K$  requires a trade-off between performance and expenses. In our paper, we set  $K = 5$ , as we find that for many image classification tasks, CLIP can already largely ensure that the true class of the image is among the top 5 preliminary predicted classes. It should be noted that due to the varying matching capabilities of CLIP for different image-text pairs, as well as the limitations of ChatGPT (as the previous analysis of bad cases of one-to-one specific semantics in L.2), a smaller value of  $K$  might yield better results than a larger  $K$  in some cases.

## II.3. Complexity of Rerank Stage

The one-to-one specific texts are created at the class granularity instead of instance granularity and, once constructed for one class, can be saved and reused for future image classifications. We analyze the computational complexity of the classification process for a single image. Consider there are  $N$  classes,  $K$  general texts per class, and  $K$  one-to-one specific texts per class pair. The feature dimension of CLIP is  $d$ . For simplicity, we focus only on the multiplications for computing similarity scores between image and text features. For the first stage, we need to perform the inner product between the image’s feature vector and the feature vectors of  $N \times K$  texts. This requires  $N \times K \times d^2$  multiplications. For the second stage, we need to perform the inner product between the image’s feature vector and the feature vectors of  $C_5^2$  one-to-one texts. This requires  $C_5^2 \times K \times d^2 = 10 \times K \times d^2$  multiplications. Thus, the total multiplication is  $(N + 10) \times K \times d^2$ , indicating the complexity of the reranking process is small.

## III. More Discussion about CODER

In this section, we provide further discussion about our CODER.

### III.1. Comparison with Related Work

we discuss the differences and innovations of our work compared to some related works.

**Compared with Concept Bottleneck Models.** Some previous works [5, 6] try to train a linear model on the image’s concept scores to complete the image classification task. This model is known as the **Concept Bottleneck Model (CBM)**. Its classification process is based on the weighting sum of various concept scores for the images, which gives it good interpretability. The calculation of these concept scores is based on the computation of cross-modal image-text similarities using CLIP. Our work primarily differs from these methods in terms of key ideas and methods.

For key ideas, previous works about CBM mainly leverage the image-text match scores for model interpretability, while our method uses these scores to construct the image’s neighbor representation for boosting CLIP’s performance. We reveal that this neighbor representation can compensate for the deficiencies in CLIP’s original features while those works about CBM lack such discussion or opinion.

For method, those work [5, 6] also use Large Language Models (LLMs) to get texts with different semantics, but our method differs from them in the implementation detail. Their approach first uses a simple query prompt to get a large candidate text set, then filters it to get a discriminative and diverse subset. However, their filter methods require cumbersome hyperparameter tuning [5] or image-based training [5]. Moreover, their methods may fail to direct LLM to produce diverse texts. In contrast, our method directly guides LLM to produce diverse, high-quality features, bypassing the need for cumbersome filtering, by employing diverse query templates. Our method is simple, without the need for images, filtering process, training process, and hyperparameter tuning process, while still generating diverse texts.

**Compared with External Knowledge-based CLIP Inference Methods.** Previous works [1, 2] have demonstrated the potential of transferring knowledge from external experts like LLMs to enhance CLIP’s performance during the inference stage. However, what knowledge should be transferred from LLMs to better aid models remains an open question with vast research potential. And our ATG has undertaken new explorations for this question. ATG introduces three new text types, including Analogous Class-based Texts, Synonym-based Texts, and One-to-One Specific Texts. Our ATG achieves more beneficial knowledge transfer from LLMs to CLIP, with experiment results showing consistent improvements over previous method [1].

**Comparison with CLIP training-free few-shot image classification methods.** Previous works [4, 7] have also attempted to enhance CLIP’s few-shot performance through a training-free approach. Among these, TIP-Adapter [7] and TIP-X [4] both try to utilize the similarity between test images and few-shot images to correct the original CLIP’s zero-shot classification results based on text classifiers. The

difference is that TIP-Adapter directly uses the image features extracted by CLIP’s image encoder to calculate the similarity between images. While TIP-X computes the similarity using each image’s concept score vectors between the image and the semantics generated by CuPL [2]. Our main difference from these methods is that we are the first to interpret image-text matching scores as images’ neighbor representations. From this perspective, we can conclude that to construct better neighbor representations, we need a diverse and high-quality text set, as suggested by the dense sampling condition of the nearest neighbor algorithm. We demonstrate through experiments that increasing the diversity and quantity of texts can improve the quality of images’ neighbor representations, thereby enhancing the performance of previous clip training-free few-shot image classification methods, such as TIP-Adapter. Previous works did not include our perspective of understanding the image-text matching scores as images’ neighbor representations, nor did they discuss how to construct high-quality image neighbor representations in the CLIP feature space, while we provide the answer — a high-quality and diverse text set is required to meet the dense sampling conditions.

### III.2. Limitations

Our method has several limitations:

The first limitation is the cost of generating texts using LLMs. When there are a large number of classes, we will also need a relatively large number of texts generated by ATG. This might require more frequent calls to the ChatGPT API for generation, thereby incurring higher API call costs and longer generation times.







The second limitation is the issue of CODER’s dimension. The dimension of CODER is equal to the number of class-related texts generated by ATG, thus it is proportional to the number of classes. When there are few classes, the small number of texts may make it difficult to meet the dense sampling conditions for the construction of images’ CODER, thus affecting the quality of CODER. On the other hand, when there are many classes, the feature dimension of CODER might become excessively high, leading to increased computational resource costs when calculating images’ similarities based on CODER. This problem can be resolved through dimension reduction. However, the dimension reduction process also introduces additional time and computational costs.

Lastly, due to the limitations of CLIP or LLM capabilities, the rerank stage may not be able to correctly modify CLIP’s prediction results for images in some cases.

Regarding the first and third limitations mentioned above, the solution we employed in our experiment is that we only use rerank stage for images with a high probability of error in the initial classification results. We determine whether the initial predicted class of the image is incorrect

by calculating the difference between the largest and the second largest logits values in the initial classification logits predictions for the image. The smaller the difference, the greater the uncertainty of the model’s prediction, indicating a higher likelihood of an incorrect prediction. And for those images with a larger difference in values, we consider CLIP’s classification result to be highly likely correct. We then compare this difference with a preset threshold value, and only the test images with differences smaller than this threshold are subject to re-ranking. For the generation of one-to-one specific texts, we employ a dynamically constructed method: If the one-to-one specific texts needed at the moment have been previously generated, we reuse the previous results. Conversely, we utilize ATG to generate the one-to-one specific texts and then save them.

We will further optimize these limitations in subsequent work.

| Sample ID | Instance   |
|-----------|--|
| 1         |  <p><b>GT Class: Airplanes</b><br/> --- Has jet engines or propellers for propulsion<br/> --- <b>Similar to Biplane</b></p> <p><b>Pred Class: Helicopter</b><br/> --- Has typically two or more sets of landing gear, including skids or wheels<br/> --- <b>Similar to Helicopter Gunship</b></p> |
| 2         |  <p><b>GT Class: Airplanes</b><br/> --- Has jet engines or propellers for propulsion<br/> --- <b>Similar to Biplane</b></p> <p><b>Pred Class: Helicopter</b><br/> --- Has typically two or more sets of landing gear, including skids or wheels<br/> --- <b>Similar to Helicopter Gunship</b></p> |
| 3         |  <p><b>GT Class: Cellphone</b><br/> --- Has rectangular or square shape<br/> --- <b>Similar to Flip phone</b></p> <p><b>Pred Class: Metronome</b><br/> --- Has a small, box-like device with a weighted pendulum or digital display<br/> --- <b>Similar to Musical Instrument</b></p>             |
| 4         |  <p><b>GT Class: Cougar</b><br/> --- A photo of cougar body,<br/> --- <b>Similar to Mountain lion</b></p> <p><b>Pred Class: Wild Cat</b><br/> --- A photo of wild cat<br/> --- <b>Similar to Cheetah</b></p> <p><small>Photo: T. W. Hall</small></p>  |
| 5         |  <p><b>GT Class: Lamp</b><br/> --- Has Light bulb or light source inside the shade<br/> --- <b>Similar to Floor Lamp</b></p> <p><b>Pred Class: Ceiling Fan</b><br/> --- Has possible shadows or blurred motion from spinning blades<br/> --- <b>Similar to Industrial fan</b></p>               |
| 6         |  <p><b>GT Class: Bookstore</b><br/> --- A photo of bookstore<br/> --- <b>A photo of bookshop</b></p> <p><b>Pred Class: Flea Market Indoor</b><br/> --- A photo of Flea Market Indoor</p>  |













|    |   |  |  |
|----|---|--|--|
| 7  |    | <p><b>GT Class: Youth Hostel</b><br/> --- A photo of youth hostel<br/> --- A photo of hostel<br/> --- A photo of student lodging</p> <p><b>Pred Class: Nursery</b><br/> --- A photo of nursery<br/> --- A photo of baby's room</p>   | <p><b>Score</b><br/> 27.47<br/> 27.68<br/> <b>30.28</b></p> <p>27.34<br/> <b>28.09</b></p> |
| 8  |    | <p><b>GT Class: Bookstore</b><br/> --- A photo of bookstore<br/> --- A photo of bookshop</p> <p><b>Pred Class: General Store Indoor</b><br/> --- A photo of General Store Indoor</p>   | <p><b>Score</b><br/> 26.11<br/> <b>27.47</b></p> <p>26.14</p>                              |
| 9  |    | <p><b>GT Class: Alley</b><br/> --- A photo of alley<br/> --- A photo of alleyway</p> <p><b>Pred Class: Medina</b><br/> --- A photo of medina</p>   | <p><b>Score</b><br/> 24.24<br/> <b>28.41</b></p> <p>27.72</p>                              |
| 10 |   | <p><b>GT Class: Staffordshire bull terrier</b><br/> --- Staffordshire bull terrier which has short, stiff coat<br/> --- More balanced bite (no underbite)</p> <p><b>Pred Class: Boxer</b><br/> --- Boxer which has short-haired dog<br/> --- More prominent underbite</p>        | <p><b>Score</b><br/> 26.36<br/> <b>27.16</b></p> <p>26.70<br/> <b>26.04</b></p>            |
| 11 |  | <p><b>GT Class: Siamese</b><br/> --- Siamese which has blue eyes<br/> --- Shorthaired and sleek coat</p> <p><b>Pred Class: Birman</b><br/> --- Birman which has blue eyes<br/> --- Longer and silkier coat</p>   | <p><b>Score</b><br/> 32.83<br/> <b>32.56</b></p> <p>33.00<br/> <b>31.05</b></p>            |
| 12 |  | <p><b>GT Class: Persian</b><br/> --- Persian which has small ears<br/> --- Flat face with a prominent nose and rounded cheeks</p> <p><b>Pred Class: British Shorthair</b><br/> --- British Shorthair which has small, rounded ears<br/> --- Round face with prominent cheeks</p> | <p><b>Score</b><br/> 25.05<br/> <b>29.55</b></p> <p>27.07<br/> <b>25.64</b></p>            |





Table 1. Examples of using new types of semantics generated by the Auto Text Generator to correct erroneous original predictions in CLIP's zero-shot image classification. The figure demonstrates the different images' matching scores with semantics for each test image's original CLIP-predicted class versus its true class. Here, the black text represents the previous semantics (including CLIP's original class name-based semantics and attribute-based semantics from VCD [1]). Blue text signifies our proposed analogous class-based semantics. Purple text signifies synonym-based semantics, and red text signifies one-to-one specific semantics. The multitude of examples in the figure prove that our newly proposed semantic types can enhance CLIP's classification capabilities.



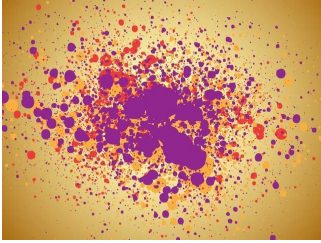

| Sample ID | Instance  |   |
|-----------|---|---|
| 1         |    | <p><b>GT Class: Emu</b><br/> --- A photo of emu<br/> --- <b>Similar to Cassowary</b></p> <p><b>Pred Class: Llama</b><br/> --- A photo of llama<br/> --- <b>Similar to Camel</b></p>   |
| 2         |    | <p><b>GT Class: Bass</b><br/> --- Has Large mouth with sharp teeth<br/> --- <b>Similar to Upright bass</b></p> <p><b>Pred Class: Sea Horse</b><br/> --- Has a body that is covered in bony plates instead of scales<br/> --- <b>Similar to Fish</b></p> |
| 3         |    | <p><b>GT Class: Crab</b><br/> --- A photo of crab<br/> --- <b>Similar to King crab</b></p> <p><b>Pred Class: Tick</b><br/> --- A photo of tick<br/> --- <b>Similar to Bedbug</b></p>  |
| 4         |   | <p><b>GT Class: Crayfish</b><br/> --- A photo of crayfish<br/> --- <b>Similar to Shrimp</b></p> <p><b>Pred Class: Sea Horse</b><br/> --- A photo of sea horse<br/> --- <b>Similar to Aquarium Decorations</b></p>                                       |
| 5         |  | <p><b>GT Class: Dragonfly</b><br/> --- A photo of dragonfly<br/> --- <b>Similar to Damselfly</b></p> <p><b>Pred Class: Joshua Tree</b><br/> --- A photo of joshua tree<br/> --- <b>Similar to Desert Plant</b></p>                                      |
| 6         |  | <p><b>GT Class: Bombay</b><br/> --- A photo of Bombay<br/> --- <b>A photo of Mumbai</b></p> <p><b>Pred Class: Siamese</b><br/> --- A photo of Siamese<br/> --- <b>A photo of Thai</b><br/> --- <b>A photo of Siamese</b></p>                            |




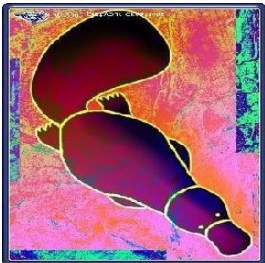
|   |   |   |
|---|---|---|
| 7 |  | <p><b>GT Class: Bengal</b><br/> --- A photo of Bengal <span style="float: right;"><b>Score</b><br/>33.09</span></p> <p>--- <b>Spotted or marbled coat with a variety of colors, such as brown, silver, or snow</b> <span style="float: right;"><b>28.89</b></span></p> <p><b>Pred Class: Abyssinian</b><br/> --- A photo of Abyssinian <span style="float: right;">30.46</span></p> <p>--- <b>Short, ticked coat with a warm reddish-brown color</b> <span style="float: right;"><b>30.25</b></span></p>                          |
| 8 |  | <p><b>GT Class: American Pit Bull Terrier</b> <span style="float: right;"><b>Score</b></span><br/> --- A photo of american pit bull terrier <span style="float: right;">32.61</span></p> <p>--- <b>Short coat and various color patterns</b> <span style="float: right;"><b>31.41</b></span></p> <p><b>Pred Class: American Bulldog</b><br/> --- A photo of american bulldog <span style="float: right;">32.80</span></p> <p>--- <b>Short coat and various color patterns</b> <span style="float: right;"><b>31.71</b></span></p> |

Table 2. **Bad case examples in zero-shot image classification using new semantics generated by the Auto Text Generator.** Here, black text represents the previous semantics (including CLIP’s original class name-based semantics and attribute-based semantics from VCD [1]). **Blue text** signifies our proposed analogous class-based semantics. **Purple text** signifies synonym-based semantics, and **red text** signifies one-to-one specific semantics.



| Sample ID | Instance   |
|-----------|--|
| 1         |  <p><b>GT Class: Chihuahua</b> Ori Value: 26.35 <b>Gap</b></p> <p>--- vs Staffordshire Bull Terrier----- 0.39</p> <p>--- vs Miniature Pinscher ----- 0.42</p> <p>--- vs Boxer ----- 0.19</p> <p>--- vs American Pit Bull Terrier ----- 0.61</p> <p>--- <b>Mean Gap</b> ----- <b>0.04</b></p> <p><b>Pred Class: Staffordshire Bull Terrier</b> Ori Value: 30.23</p> <p>--- vs Miniature Pinscher----- -0.10</p> <p>--- vs Chihuahua ----- -0.39</p> <p>--- vs Boxer ----- -0.21</p> <p>--- vs American Pit Bull Terrier ----- 0.02</p> <p>--- <b>Mean Gap</b> ----- <b>-0.17</b></p> |
| 2         |  <p><b>GT Class: Beagle</b> Ori Value: 30.73 <b>Gap</b></p> <p>--- vs Basset Hound ----- -0.12</p> <p>--- vs American Bulldog ----- 3.20</p> <p>--- vs Boxer ----- 3.49</p> <p>--- vs Saint Bernard ----- 5.76</p> <p>--- <b>Mean Gap</b> ----- <b>3.08</b></p> <p><b>Pred Class: Basset Hound</b> Ori Value: 31.12</p> <p>--- vs Beagle ----- 0.12</p> <p>--- vs American Bulldog ----- 1.90</p> <p>--- vs Boxer ----- 2.14</p> <p>--- vs Saint Bernard ----- 4.64</p> <p>--- <b>Mean Gap</b> ----- <b>2.20</b></p>  |
| 3         |  <p><b>GT Class: Bengal</b> Ori Value: 21.61 <b>Gap</b></p> <p>--- vs Egyptian Mau----- 0.0</p> <p>--- vs British Shorthair ----- 0.51</p> <p>--- vs Abyssinian ----- 1.05</p> <p>--- vs Siamese ----- 1.49</p> <p>--- <b>Mean Gap</b> ----- <b>0.76</b></p> <p><b>Pred Class: Egyptian Mau</b> Ori Value: 31.61</p> <p>--- vs British Shorthair ----- 0.61</p> <p>--- vs Abyssinian ----- 0.81</p> <p>--- vs Bengal ----- 0.0</p> <p>--- vs Siamese ----- 0.95</p> <p>--- <b>Mean Gap</b> ----- <b>0.59</b></p>  |
| 4         |  <p><b>GT Class: Great pyrenees</b> Ori Value: 26.33 <b>Gap</b></p> <p>--- vs Samoyed----- 1.05</p> <p>--- vs Keeshond ----- 3.39</p> <p>--- vs Saint Bernard ----- 3.27</p> <p>--- vs American Bulldog----- 2.83</p> <p>--- <b>Mean Gap</b> ----- <b>2.63</b></p> <p><b>Pred Class: Samoyed</b> Ori Value: 32.93</p> <p>--- vs Great Pyrenees----- -1.05</p> <p>--- vs Keeshond ----- 0.81</p> <p>--- vs Saint Bernard ----- 3.42</p> <p>--- vs American Bulldog ----- 2.21</p> <p>--- <b>Mean Gap</b> ----- <b>2.32</b></p>   |

|   |   |  |
|---|---|--|
| 5 |    | <p><b>GT Class: British Shorthair</b> Ori Value: 33.43 <b>Gap</b></p> <p>--- vs Russian Blue ----- -0.22</p> <p>--- vs Siamese ----- 1.81</p> <p>--- vs Persian ----- -0.81</p> <p>--- vs Abyssinian ----- 1.88</p> <p>--- <b>Mean Gap</b> ----- <b>1.18</b></p> <p><b>Pred Class: Russian Blue</b> Ori Value: 34.04</p> <p>--- vs British Shorthair ----- -0.22</p> <p>--- vs Siamese ----- 1.66</p> <p>--- vs Persian ----- -0.29</p> <p>--- vs Abyssinian ----- 2.44</p> <p>--- <b>Mean Gap</b> ----- <b>0.90</b></p> |
| 6 |    | <p><b>GT Class: Stratified</b> Ori Value: 29.32 <b>Gap</b></p> <p>--- vs Marbled ----- 0.88</p> <p>--- vs Veined ----- -0.15</p> <p>--- vs Fibrous ----- 1.10</p> <p>--- vs Wrinkled ----- 1.32</p> <p>--- <b>Mean Gap</b> ----- <b>0.78</b></p> <p><b>Pred Class: Marbled</b> Ori Value: 31.57</p> <p>--- vs Stratified----- -0.88</p> <p>--- vs Veined ----- 0.32</p> <p>--- vs Fibrous ----- 0.39</p> <p>--- vs Wrinkled ----- 0.88</p> <p>--- <b>Mean Gap</b> ----- <b>0.18</b></p>                                  |
| 7 |   | <p><b>GT Class: Sprinkled</b> Ori Value: 22.31 <b>Gap</b></p> <p>--- vs Dotted ----- 1.95</p> <p>--- vs Smearred ----- 1.29</p> <p>--- vs Swirly ----- 2.99</p> <p>--- vs Porous ----- 1.76</p> <p>--- <b>Mean Gap</b> ----- <b>2.00</b></p> <p><b>Pred Class: Dotted</b> Ori Value: 23.92</p> <p>--- vs Sprinkled ----- -1.95</p> <p>--- vs Smearred ----- 1.15</p> <p>--- vs Swirly ----- 1.00</p> <p>--- vs Porous ----- 0.98</p> <p>--- <b>Mean Gap</b> ----- <b>0.29</b></p>  |
| 8 |  | <p><b>GT Class: Fibrous</b> Ori Value: 26.70 <b>Gap</b></p> <p>--- vs Porous ----- 0.81</p> <p>--- vs Matted ----- 0.90</p> <p>--- vs Marbled ----- 2.12</p> <p>--- vs Wrinkled ----- 0.81</p> <p>--- <b>Mean Gap</b> ----- <b>1.16</b></p> <p><b>Pred Class: Porous</b> Ori Value: 27.67</p> <p>--- vs Fibrous ----- -0.81</p> <p>--- vs Matted ----- -0.12</p> <p>--- vs Marbled ----- 1.34</p> <p>--- vs Wrinkled ----- 0.73</p> <p>--- <b>Mean Gap</b> ----- <b>0.29</b></p>   |

|    |   |   |
|----|---|---|
| 9  |    | <p><b>GT Class: Fibrous</b> Ori Value: 22.84 <b>Gap</b></p> <p>--- vs Woven ----- 0.66<br/> --- vs Matted ----- 0.13<br/> --- vs Cobwebbed ----- 0.96<br/> --- vs Braided ----- 0.11<br/> --- <b>Mean Gap</b> ----- <b>0.47</b></p> <p><b>Pred Class: Woven</b> Ori Value: 24.85</p> <p>--- vs Fibrous ----- -0.66<br/> --- vs Matted ----- 0.32<br/> --- vs Cobwebbed ----- 0.17<br/> --- vs Braided ----- 0.49<br/> --- <b>Mean Gap</b> ----- <b>0.08</b></p>                 |
| 10 |    | <p><b>GT Class: Chair</b> Ori Value: 23.29 <b>Gap</b></p> <p>--- vs Windsor Chair----- 0.67<br/> --- vs Wheelchair----- 4.21<br/> --- vs Beaver ----- 2.05<br/> --- vs Barrel ----- 1.15<br/> --- <b>Mean Gap</b> ----- <b>2.02</b></p> <p><b>Pred Class: Windsor Chair</b> Ori Value: 23.92</p> <p>--- vs Chair ----- -0.67<br/> --- vs Wheelchair ----- 2.92<br/> --- vs Beaver ----- 0.65<br/> --- vs Barrel ----- 1.56<br/> --- <b>Mean Gap</b> ----- <b>1.11</b></p>       |
| 11 |  | <p><b>GT Class: Crayfish</b> Ori Value: 18.96 <b>Gap</b></p> <p>--- vs Lobster ----- 0.20<br/> --- vs Scorpion ----- 0.09<br/> --- vs Sea Horse ----- 6.43<br/> --- vs Crab ----- 3.89<br/> --- <b>Mean Gap</b> ----- <b>2.86</b></p> <p><b>Pred Class: Lobster</b> Ori Value: 26.34</p> <p>--- vs Crayfish ----- -0.20<br/> --- vs Scorpion ----- 1.09<br/> --- vs Sea Horse ----- 4.19<br/> --- vs Crab ----- 4.62<br/> --- <b>Mean Gap</b> ----- <b>2.42</b></p>             |
| 12 |  | <p><b>GT Class: Platypus</b> Ori Value: 16.79 <b>Gap</b></p> <p>--- vs Beaver ----- 0.73<br/> --- vs Water Lilly ----- 6.28<br/> --- vs Sea Horse ----- 3.65<br/> --- vs Yin Yang ----- 0.34<br/> --- <b>Mean Gap</b> ----- <b>2.75</b></p> <p><b>Pred Class: Beaver</b> Ori Value: 16.68</p> <p>--- vs Platypus ----- -0.73<br/> --- vs Water Lilly ----- 6.04<br/> --- vs Sea Horse ----- 3.57<br/> --- vs Yin Yang ----- 4.71<br/> --- <b>Mean Gap</b> ----- <b>3.39</b></p> |


|    |   |   |             |
|----|---|---|-------------|
| 13 |  | <b>GT Class: Lotus</b> Ori Value: 17.89         | <b>Gap</b>  |
|    |   | --- vs Water Lilly -----                        | -0.51       |
|    |   | --- vs Yin Yang -----                           | 1.37        |
|    |   | --- vs Buddha -----                             | 5.22        |
|    |   | --- vs Crocodile Head -----                     | 3.94        |
|    |   | --- <b>Mean Gap</b> -----                       | <b>2.50</b> |
|    |   | <b>Pred Class: Water Lilly</b> Ori Value: 22.23 |             |
|    |   | --- vs Lotus -----                              | 0.51        |
|    |   | --- vs Yin Yang -----                           | 0.27        |
|    |   | --- vs Buddha -----                             | 3.16        |
|    |   | --- vs Crocodile Head -----                     | 3.16        |
|    |   | --- <b>Mean Gap</b> -----                       | <b>1.70</b> |

Table 3. **Examples of correcting original classification results through reranking based on One-to-One Specific CODER.** By introducing one-to-one semantics to accentuate features with the maximum differences between classes, and leveraging the rich information contained in the relative strengths of predictive scores across classes, we successfully rectify the original wrong predicted results.

## References

- [1] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023. 1, 3, 6, 8
- [2] Sarah M. Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. *CoRR*, abs/2209.03320, 2022. 3, 4
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1
- [4] Vishaal Udandara, Ankush Gupta, and Samuel Albanie. Susx: Training-free name-only transfer of vision-language models. *CoRR*, abs/2211.16198, 2022. 3
- [5] An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, and Julian J. McAuley. Learning concise and descriptive attributes for visual recognition. In *ICCV*, pages 3067–3077. IEEE, 2023. 3
- [6] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pages 19187–19197, 2023. 3
- [7] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapt: Training-free adaption of CLIP for few-shot classification. In *ECCV*, pages 493–510, 2022. 3