Data Descriptions from Large Language Models with Influence Estimation

Chaeri Kim[*]   **Jaeyeon Bae**[*]   **Taehwan Kim**
Ulsan National Institute of Science and Technology(UNIST)
{chaerikim, qowodussla, taehwankim}@unist.ac.kr

Abstract

Deep learning models have been successful in many areas but understanding their behaviors still remains a black-box. Most prior explainable AI (XAI) approaches have focused on interpreting and explaining how models make predictions. In contrast, we would like to understand how data can be explained with deep learning model training and propose a novel approach to understand the data via one of the most common media - language - so that humans can easily understand. Our approach proposes a pipeline to generate textual descriptions that can explain the data with large language models by incorporating external knowledge bases. However, generated data descriptions may still include irrelevant information, so we introduce to exploit influence estimation to choose the most informative textual descriptions, along with the CLIP score. Furthermore, based on the phenomenon of cross-modal transferability, we propose a novel benchmark task named *cross-modal transfer classification* to examine the effectiveness of our textual descriptions. In the experiment of zero-shot setting, we show that our textual descriptions are more effective than other baseline descriptions, and furthermore, we successfully boost the performance of the model trained only on images across all nine image classification datasets. These results are further supported by evaluation using GPT-4o. Through our approach, we may gain insights into the inherent interpretability of the decision-making process of the model.

1 Introduction

Deep learning models have successfully been applied to various fields and achieved high performance (Huang et al., 2017; Vaswani et al., 2017; He et al., 2016; Dosovitskiy et al., 2020). Despite the rapid performance improvement, understanding their behaviors remains a black-box. While most prior explainable AI (XAI) approaches have focused on interpreting and explaining how models make predictions, there are few attempts to explain the data. We would like to understand data since it is one of the most important elements of deep learning performance. Among deep learning models, we try to describe image classes with *human-interpretable* language. For example, in the case of frog class: smooth, moist skin with coloration ranging from green to brown, often featuring various patterns and markings.

Recent studies (Menon and Vondrick, 2023; Maniparambil et al., 2023; Pratt et al., 2023) have utilized large language models (LLMs), pre-trained on vast datasets, to generate textual descriptions that aid vision model performance. However, identifying descriptions that are truly informative and semantically aligned with the class remains a key challenge. To solve this problem, we propose a novel approach to identify the most helpful texts for model training among those generated by the LLM, using influence estimation and CLIP scores. Influence estimation (Koh and Liang, 2017; Pruthi et al., 2020) calculates the positive or negative impact of each training image in predicting a test sample, while CLIP scores measure the similarity between an image and text. Unlike most previous research that applies influence scores only to images, we extend this method to texts by integrating CLIP scores to identify helpful texts for model training. We name the helpful text as *proponent text* and the score for determining the proponent text as *Influence scores For Texts* (IFT).

To obtain the proponent texts, we first write down all the features of each image class with GPT-3.5 (Brown et al., 2020). Instead of merely querying the LLM for class features, we employ a two-stage prompting process using Wikipedia urls to integrate external knowledge bases. This approach not only compensates for potential gaps in detailed knowledge of LLM but also helps mitigate halluci-
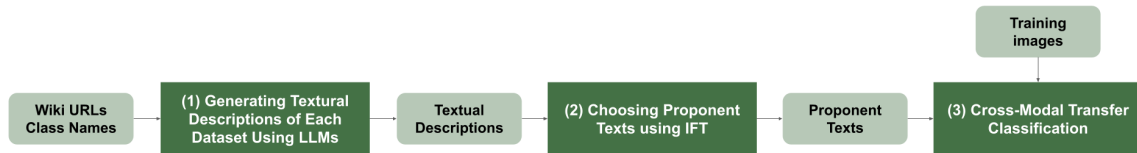

Figure 1: Overview of our framework. (1) 'Generating Textual Descriptions of Each Dataset' process generates the textual descriptions that can explain each class well from given Wikipedia urls, pre-defined prompts, and class names (2) 'Choosing Proponent Texts using IFT' process determines the proponent texts using IFT. (3) 'Cross-Modal Transfer Classification' process first trains model with images, then further trains it with proponent texts.

nations. Among the generated textual descriptions, IFT filters out unhelpful information and retains only proponent texts. Furthermore, leveraging the concept of cross-modal transferability (Zhang et al., 2023), which allows text to serve as input for models originally trained with images, we apply this principle to our approach. We retrain the image-trained model using proponent texts, assigning weights based on the IFT. We call this *'cross-modal transfer classification'*. Compared to image-only training and baselines that generate textual descriptions, our approach achieves superior performance on most of the nine image datasets, demonstrating that our textual descriptions are more helpful in model training than other baselines. Additionally, we evaluate the helpfulness, informativeness, and relevance of the generated descriptions using GPT-4o (Hurst et al., 2024) and find that our method consistently outperforms all baselines across these criteria. Figure 1 shows the overview of our framework.

In sum, our contributions are as follows:

- We propose a novel approach to identify helpful textual descriptions that effectively explain each image class, achieving superior performance in the zero-shot setting compared to baseline descriptions.

- By using the proposed IFT defined as the sum of the influence score and the CLIP score, we can pinpoint the most helpful textual descriptions for model training. Proponent texts selected through IFT contain only relevant information that aids in accurate image classification. Furthermore, we may get an inherent interpretability of where the black-box model focuses on during training.

- We propose *cross-modal transfer classification* as a novel benchmark task based on the cross-modal transferability phenomenon. Extensive experiments with nine datasets show

that the proponent texts are informative and describe each image class well.

2   Related Work

Not only achieving high accuracy, but most research has focused on why the model makes such decisions. Influence function is one of the powerful tools for explaining model decisions. One widely used approach (Koh and Liang, 2017) examines changes in model parameters when input data is perturbed, using second-order optimization techniques to efficiently approximate influence scores. TracIn (Pruthi et al., 2020) calculates the influence of training data on the loss of a test sample by tracing how loss changes during the training process with a fixed test sample.

Efforts to enhance model explainability have also extended to the use of large language models (LLMs). Language Guided Bottlenecks (LaBo) (Yang et al., 2023) is an extended model of Concept Bottleneck Models (Koh et al., 2020) that queries LLMs to collect concepts. Menon and Vondrick (2023) uses descriptions generated through LLMs and proposes an alternative zero-shot classification method named 'classification by description'. However, Menon and Vondrick (2023) highlights some limitations: their generated text descriptions contain non-useful visual cues for the vision-language model and sometimes include repeated text. VDT-Adapter (Maniparambil et al., 2023) and CuPL (Pratt et al., 2023) two approaches that also leverage LLMs to enhance vision-language models. VDT-Adapter (Maniparambil et al., 2023) improves the image classification performance of CLIP by using GPT-4 to generate visually descriptive text prompts, helping the model focus on relevant visual details. Similarly, CuPL (Pratt et al., 2023) (Customized Prompts via Language models) generates category-specific prompts using LLMs, improving zero-shot classification without additional training, and outperforming hand-crafted

Step 1.
Generating Textual Descriptions of Each Dataset Using LLMs

Step 2.
Choosing Proponent texts using IFT

Step 3.
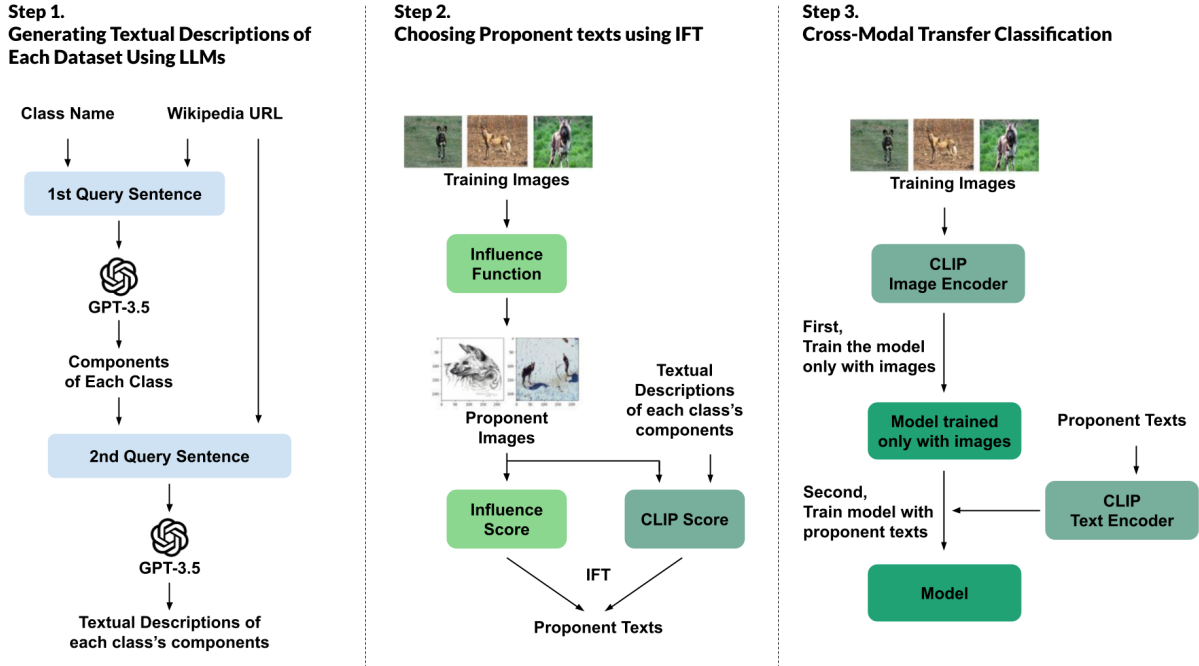**Cross-Modal Transfer Classification**

Figure 2: Details of our approach. (1) Extract class components from class names and obtain textual descriptions using Wikipedia urls. (2) Identify proponent images using influence scores, then combine CLIP scores and influence scores to get proponent texts (3) Train the model with training images, followed by cross-modal transfer training using proponent texts.

prompts across multiple benchmarks. Nevertheless, both approaches have limitations in identifying the most relevant textual descriptions.

To deal with these issues, we define *influence scores for texts* (IFT) composed of influence score and CLIP score. IFT serves as a scoring metric to determine the importance of each text. Through the proposed IFT, our proponent texts include only the information necessary for precise image classification. Moreover, the advent of models trained with vision-language contrastive learning, such as CLIP (Radford et al., 2021), provides cross-modal transferability phenomenon (Zhang et al., 2023). By integrating these insights, we propose a novel benchmark task *cross-modal transfer classification*, which underscores the effectiveness of our textual descriptions in enhancing model performance.

3 Approach

As described in Figure 2, our framework consists of three main steps. In this section, we describe each step in detail.

3.1 Generating Textual Descriptions of Each Dataset Using LLMs

Extracting Components of Each Class From the raw class names of the dataset, we generate ques-

tions by placing the class names into a pre-defined prompt. For datasets with subcategories under a superclass, we query GPT-3.5 for the components of the superclass. For example, in the CUB 200 2011 (Wah et al., 2011) dataset, "bird" is the superclass, and species like "Laysan Albatross" and "Fish Crow" are subcategories. In this case, we ask about the components of "birds" rather than the individual species. The question format is as follows: "*Q : Can you tell me the components of {class name} from the perspective of appearance? A :* ". The extracted components are then used to extract the textual descriptions of each image class. For example, for the 'African hunting dog' class in the Miniimagenet dataset (Vinyals et al., 2016), the components include: body build, coat color, ears, head, and eyes.

Generating Textual Descriptions of Each Class Knowing the components of each class, we can query GPT-3.5 about the appearance of each component. To avoid lacking detailed knowledge and hallucinations, we provide the corresponding Wikipedia url. GPT-3.5 summarizes the relevant information from the Wikipedia url in one line, composed of nouns. If no relevant information is found on Wikipedia, we ask GPT-3.5 to provide a summary based on its existing knowledge. The ques-

$$\text{Influence Score}(i^t, i^v) = \sum_{j=1}^{k} \eta_j \nabla loss(w_{t_j}, i^t) \cdot \nabla loss(w_{t_j}, i^v)) \quad i^t \in I^{train}, i^v \in I^{val} \quad (1)$$

$$\text{IFT}(T_{c'}) = \frac{1}{n_{Ival} \cdot n_{Itrain}} \sum_{i^v \in I^{val}} \sum_{i^t \in I^{train}} (\text{Influence Score}(i^t, i^v) + \text{CLIP Score}(i^t, T_{c'})) \quad (2)$$

tion format is as follows: "*Q : Please summarize the information of appearance about {components of each class} in this {Wikipedia url} in one line composed of nouns. If you couldn't find related information, you must answer general information you know. A : "*

3.2 Choosing Proponent Texts

We use TracIn (Pruthi et al., 2020) to calculate influence scores for images. Based on influence scores, we identify proponent images from training images that aid in predicting each validation image. Then, we calculate the CLIP score, which refers to the cosine similarity between the CLIP embeddings of the proponent images and the corresponding textual descriptions in the validation set. Higher CLIP scores indicate stronger semantic alignment between images and descriptions. With the sum of the influence score and CLIP score, we can calculate *influence scores for texts* (IFT). For notations, let I^{train} and I^{val} denote image samples of the train and validation dataset respectively and T_C denotes the textual descriptions for all classes $\mathcal{C}$. n_{Itrain} is the number of image samples in the training dataset and T_c are the extracted textual descriptions of class c where $c \in \mathcal{C}$.

In equation (1), the loss of model parameterized by w on training image sample i^t can be denoted as $loss(w, i^t)$. In this context, Δloss specifically refers to the cross-entropy loss, as the task involves image classification. Influence score measures the impact of specific training examples on a given validation sample. Since considering only one training sample at a time is impractical, TracIn (Pruthi et al., 2020) introduces practical influence function via k checkpoints $w_{t_1}, w_{t_2}, ..., w_{t_k}$ and minibatches through simple first-order approximation. It is the total reduction in loss on a fixed validation example i^v in the training process. CLIP score(i^t, T_c) computes the correlation between CLIP image embeddings of proponent training image i^t and CLIP text embeddings of textual description for class c T_c.

Equation (2) defines the IFT score. We calculate the average of the sum of influence scores and CLIP scores for the textual descriptions of each image class. We then select ten textual descriptions with the highest IFT scores, naming them *proponent texts*. These proponent texts help us effectively explain and understand each image class.

3.3 Cross-Modal Transfer Classification

Cross-modal transferability (Zhang et al., 2023) states that text inputs can work as good proxies to image inputs trained on a shared image-text embedding space obtained through multi-modal learning. Based on this phenomenon, we can use texts as inputs instead of images for vision models trained with images if images and texts are in a shared embedding space.

We use CLIP (Radford et al., 2021) image encoder and text encoder to align images and proponent texts in the same embedding space. Denote the dimension of CLIP embedding space as D. An input training image i^t is projected into the image embedding space as $e_{it} = CLIP_{ImageEncoder}(i^t)$, where $e_{it} \in \mathcal{R}^D$. Similarly, proponent text T_c for class c is projected into a text embedding $e_{T_c} = CLIP_{TextEncoder}(e_c)$ where $e_{t_c} \in \mathcal{R}^D$. The model, which consists of linear layers, is initially trained on image embeddings and subsequently retrained on proponent text embeddings. While training the model with proponent texts $T_{\mathcal{C}}$, weights are given for each e_{T_c} based on their normalized IFT scores : $\sum_{c \in \mathcal{C}} w_c \cdot CELoss(\hat{y}_{T_c}, y_{T_c})$ while $w_c = \frac{IFT_c}{\sum_{c \in \mathcal{C}} IFT_c}$. Here, $CELoss$ refers to the cross-entropy loss, which measures the discrepancy between the predicted labels $\hat{y}_{T_c}$ and the true labels y_{T_c} of the classes. Because e_{it} and e_{T_c} share the same embedding space $\mathcal{R}^D$, the model can be trained without any additional implementations. For clarity, we include the complete training procedure in Algorithm 1, which first train the classifier only with the training images and then refines it with weighted proponent texts. By combining these two complementary training steps, our approach leverages both modalities in a unified embedding space, thereby enabling effective cross-modal trans-

Algorithm 1 Cross-Modal Transfer Classification Training Algorithm

1: **Input:**
2:    Training images : I^{train}
3:    Textual description for class c : T_c
4:    Frozen CLIP encoders : $\text{CLIP}_{\text{ImageEncoder}}, \text{CLIP}_{\text{TextEncoder}}$
5: **Output:** classifier M
6:    $e_{i_t} \leftarrow \text{CLIP}_{\text{ImageEncoder}}(i_t)$ for each $i_t \in I^{train}$
7:    $e_{T_c} \leftarrow \text{CLIP}_{\text{TextEncoder}}(T_c)$ for each $c \in \mathcal{C}$
8:    $w_c \leftarrow \text{IFT}_c / \sum_{c \in \mathcal{C}} \text{IFT}_c$
9:    **Step 1 (image):** Train M with $\mathcal{L}_{img} = CE(M(e_{i_t}), L_{i_t})$
10:   **Step 2 (text):** Train M with $\mathcal{L}_{txt} = CE(M(w_c \cdot e_{T_c}), L_{T_c})$
11: **return** M

fer classification.

This method is simple yet effective, as it improves performance by only updating the linear layer while the CLIP encoder remains frozen. Furthermore, our approach has a low computational cost. Training with only images on all datasets takes about 2 hours using an NVIDIA 3090 GPU, while cross-modal transfer training with proponent texts requires less than 30 minutes. Additionally, the performance improvement indicates that the proponent texts effectively explain each image class and enhance the training process.

Our novel *cross-modal transfer classification* benchmark aims to improve image classification model performance using effective and helpful textual descriptions. We expect this benchmark to enhance understanding of black-box models by improving performance through human-readable language. Furthermore, it may provide an inherent explanation of the decision-making process of the black-box model.

4 Experiments

In the following sections, we detail our experimental setup, present the results of our approach in comparison. Additionally, we perform ablation studies to further evaluate the contributions of individual components and their impact on overall performance.

4.1 Experimental Setup and Details

Datasets We use nine image datasets for our experiments: CUB 200 2001 (Wah et al., 2011), Miniimagenet (Vinyals et al., 2016), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), OxfordPets (Parkhi et al., 2012), EuroSAT (Helber et al., 2019), Food101 (Bossard et al., 2014), 102flowers (Nilsback and Zisserman, 2008), and Describable Textures Dataset (DTD) (Cimpoi et al.,

2014). We follow the official dataset partitions when available, and use 20% of the training set as the validation set when no official validation set is provided. If no official train/test split exists, we randomly divide the dataset into training, validation, and test sets. Detailed dataset partitions are reported in Section E of Appendix.

Implementation Details We use ViT/32 CLIP (Radford et al., 2021) as the image encoder. We train the linear model using stochastic gradient descent (SGD) with a mini-batch size of 64 with learning rate of 0.1. The vision model, initially trained only with images, is further trained with proponent texts for a total of 30 epochs in all settings. We also use the CosineAnnealingLR learning scheduler with the maximum number of iterations of 200.

To calculate influence scores, we use checkpoints during training. We train a pre-trained ResNet34 with stochastic gradient descent, a mini-batch size of 64, and a learning rate starting at 0.1, divided by 10 every 30 epochs, terminating at 200 epochs. We use checkpoints every 10 epochs.

Baselines We compare our method with other approaches that use large language model to generate text descriptions for image classes (Menon and Vondrick, 2023; Maniparambil et al., 2023; Pratt et al., 2023; Yang et al., 2023). For each baseline, We use the pre-generated descriptions available in the official repositories, adjusting them to fit our experimental setup. For datasets where descriptions are not provided, we reproduce the authors' code to generate the corresponding descriptions.

4.2 Result

Table 1 reports the performance of our approach in zero-shot setting, comparing our textual descriptions with the baseline methods. This setting aims to demonstrate the effectiveness of our textual de-

Datasets	CLIP Zero-shot	Menon	LABO	CuPL	VDT-Adapter	Ours Zero-Shot
CUB 200 2011	38.540%	52.969%	52.917%	**53.349%**	53.162%	<u>53.227%</u>
OxfordPets	81.132%	85.580%	87.196%	**88.814%**	88.140%	<u>88.679%</u>
CIFAR10	88.800%	<u>89.320%</u>	88.709%	88.150%	89.090%	**89.470%**
CIFAR100	61.680%	<u>63.999%</u>	60.460%	63.660%	63.450%	**64.570%**
EuroSat	30.815%	32.630%	28.667%	<u>38.925%</u>	38.457%	**39.148%**
Food101	80.620%	**83.644%**	82.871%	83.339%	83.013%	<u>83.452%</u>
Miniimagenet	81.630%	84.780%	84.890%	84.720%	<u>85.199%</u>	**85.320%**
102flowers	58.730%	66.670%	66.667%	67.643%	<u>68.742%</u>	**69.109%**
DTD	43.085%	37.660%	46.489%	47.553%	<u>48.457%</u>	**48.989%**

Table 1: Accuracy for test images in zero-shot setting. The best performing ones in bold font and underlined represent the second-best performance.

Datasets	Only Images	Menon	LaBo	CuPL	VDT-Adapter	Ours
CUB 200 2011	71.332%	74.905%	73.991%	<u>74.957%</u>	73.699%	**75.130%**
OxfordPets	91.664%	<u>91.711%</u>	90.700%	91.105%	91.431%	**93.396%**
CIFAR10	<u>94.320%</u>	93.470%	93.480%	93.420%	93.520%	**94.820%**
CIFAR100	77.830%	78.380%	78.400%	78.270%	<u>78.430%</u>	**78.650%**
EuroSat	94.296%	93.518%	95.259%	<u>95.293%</u>	94.482%	**96.037%**
Food101	85.848%	<u>86.402%</u>	86.165%	86.145%	86.138%	**86.950%**
Miniimagenet	91.980%	<u>92.030%</u>	91.770%	91.640%	91.740%	**92.480%**
102flowers	96.459%	96.825%	96.948%	96.581%	<u>97.191%</u>	**97.948%**
DTD	72.074%	72.872%	72.713%	72.446%	<u>72.767%</u>	**74.393%**

Table 2: Accuracy for test images when training using only images (Only Images) and cross-modal transfer training with texts (Cross-Modal Transfer Classification). The best performing ones in bold font and underlined represent the second-best performance.

scriptions in a zero-shot context, aligning with the zero-shot setting assumptions in Menon (Menon and Vondrick, 2023). In our approach, we utilize GPT-3.5, and for a fair comparison, we also apply GPT-3.5 to methods that originally used earlier versions, while employing GPT-4 for methods that used GPT-4. Our method outperforms the baselines in most datasets. This result demonstrates that our textual descriptions explain each image class better than the baseline method. Additionally, our 2-stage prompting with Wikipedia urls, proves to be effective, outperforming not only methods that use the same GPT-3.5 but also those that utilize the more advanced GPT-4. To verify the robustness of our method, we conduct additional zero-shot classification experiments using a newer vision-language model, laion/CLIP-ViT-L-14-laion2B-s32B-b82K (Cherti et al., 2023). This model is trained on a larger dataset and reflects recent advances in CLIP training. As shown in Table 7 in the Appendix, our method maintains strong performance across diverse datasets, consistently outperforming baseline methods and demonstrating that our method remains effective across CLIP architectures and continues to outperform most baselines.

Furthermore, as shown in Table 2, most ap-

proaches achieve higher performance with cross-modal transfer training compared to training with images alone. Notably, our approach exhibits the largest performance gains across all datasets when proponent texts are incorporated into cross-modal transfer training. This result further supports that our proponent texts can explain each image class well and the prediction of the black-box model.

As our method consistently outperforms other baselines in both the zero-shot setting and cross-modal transfer classification, this suggests that the improvements stem not from the mere inclusion of textual descriptions but from selecting the most informative ones to enhance model training. If these gains are due to the model's inherent preference for textual supervision (i.e., inductive bias), similar improvements would be observed across all textual description baselines. However, as shown in Table 1 and Table 2, our method consistently achieves greater improvements, indicating that performance gains arise from our influence-guided selection of proponent texts rather than the presence of textual descriptions alone.

To investigate why our method performs better than the baselines, we visualize the embeddings of each textual description and the embeddings
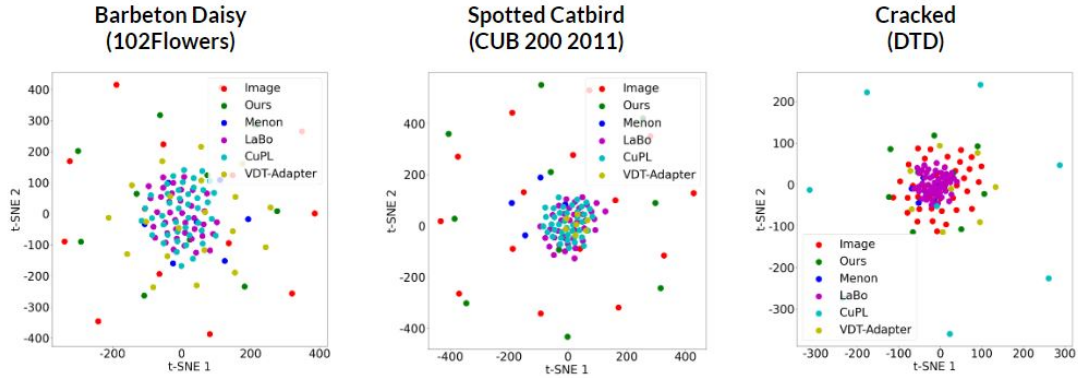
Figure 3: Visualization of the embeddings of textual descriptions of our method and baseline method and the embeddings of the same class images with t-SNE. (a) Barbeton Daisy class of 102flowers dataset (b) Spotted catbird class of CUB 200 2011 dataset (c) cracked class of DTD dataset

of the same class images with t-SNE (Van der Maaten and Hinton, 2008). Figure 3 presents the t-SNE visualizations of selected classes from the 102flowers (Nilsback and Zisserman, 2008), CUB 200 2011 (Wah et al., 2011), and DTD (Cimpoi et al., 2014) datasets, comparing the image embeddings with text embeddings from our method and various baselines. In all three datasets, our method appears to align more closely with the image embeddings compared to the baselines. For instance, in the CUB 200 2011 dataset, the Spotted Catbird class shows that while CuPL (Pratt et al., 2023), LaBo (Yang et al., 2023), and VDT-Adapter (Maniparambil et al., 2023) tend to cluster primarily in the center, our method achieves better alignment with the corresponding image embeddings. These results demonstrate that our approach leads to superior alignment between images and textual descriptions, thereby improving performance in cross-modal transfer tasks. Furthermore, Figure 4 provides examples of proponent images and proponent texts and analysis in Section D of Appendix demonstrates how IFT effectively selects texts that explain each image class by comparing proponent texts with non-proponent texts.

We hypothesize that the performance gains in cross-modal transfer training with proponent texts, compared to image-only training and other description-generating baselines, stem from the model's ability to integrate complementary information from both modalities. Training with images and textual descriptions enables the model to jointly utilize visual features and contextual information. Unlike general textual descriptions from baselines, our proponent texts are filtered using influence estimation and CLIP scores, ensuring greater rele-
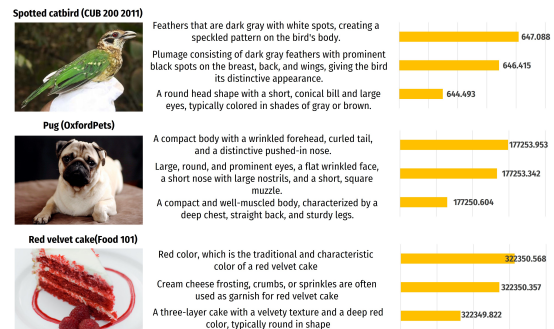
Figure 4: Examples of proponent images, proponent texts and IFT for three classes. (Best viewed at an increased zoom level for clearer details.)

vance and informativeness. Furthermore, by incorporating external knowledge, our descriptions help the model learn richer semantic representations of image classes. The qualitative comparison in Figure 5 demonstrates the differences in textual descriptions generated by various methods for the Blue Jay class of CUB 200 2011 (Wah et al., 2011). Menon (Menon and Vondrick, 2023) provides the most basic descriptions, mentioning only the color of bird and a few physical traits, while LaBO (Yang et al., 2023) and CuPL (Pratt et al., 2023) offer more detailed explanations but with some repetition and a lack of additional visual features. VDT-Adapter (Maniparambil et al., 2023) focuses on specific features like the eyestripe and beak but presents a more mechanical, segmented description. In contrast, our method delivers richer, more detailed descriptions, emphasizing unique facial features such as a 'short, robust beak, a crest on the head, and vibrant blue feathers,' providing a more complete and expressive visual image than the other methods.

Blue jay (CUB 200 2011)

Menon
- a blue bird
- a white chest
- a black neck and head

LaBO
- The Blue Jay is a bold and striking bird with a vibrant blue crest on its head.
- Its feathers are a mix of bright blue, white, and black, giving it a unique and eye-catching appearance.
- Its beak is black and slightly curved, perfect for cracking open nuts and seeds.

CuPL
- The Blue Jay is a brightly colored songbird with a blue body and wings, a white chest, and a black neck and head.
- The Blue Jay is a small to medium sized bird with a distinctive blue color.
- The blue jay is a small to medium sized bird with a strong build.

VDT-Adapter
- Blue jay is a bird with beak that is medium, thick, black
- Blue jay is a bird with head that is blue with a black eyestripe and white cheeks
- Blue jay is a bird with body that is blue upperparts, pale gray underparts

Ours
- Crests consisting of a prominent crown of feathers on the top of their heads, often bright blue in color
- Facial features including a short, robust beak, a crest on the head, and vibrant blue feathers
- A long, blue-tinged tail with white patches on the underside, serving as a visual and behavioral signal in communication

Figure 5: Example of a qualitative sample for evaluating textual descriptions. (Blue Jay class of CUB 200 2011 dataset)

4.3 Ablation Studies

Dataset	Only Images	IF	CLIP	IFT
CUB 200 2011	71.332%	72.834%	72.696%	**74.525%**
OxfordPets	91.664%	92.722%	92.318%	**93.396%**
CIFAR-10	94.320%	94.680%	94.730%	**94.820%**
CIFAR-100	77.830%	77.910%	77.970%	**78.250%**
EuroSAT	94.296%	94.222%	94.407%	**96.037%**
Food101	85.848%	85.974%	85.934%	**86.950%**
Miniimagenet	91.980%	92.040%	92.020%	**92.480%**
102flowers	96.459%	96.460%	96.337%	**97.192%**
DTD	72.074%	73.989%	73.830%	**74.043%**

Table 3: Accuracies of ablation studies on Cross-Modal Transfer Classification with different proponent text determination methods: Influence score(IF), CLIP score(CLIP), and IFT.

To verify the effectiveness of IFT as a scoring metric, we compare cross-modal transfer classification performance when using either the influence score or CLIP score alone instead of IFT in Table 3. For all datasets, compared to when only the influence score or CLIP score is used, the performance is higher when cross-modal transfer training with proponent texts selected using IFT.

We also show examples of proponent texts that change when using only the influence score or CLIP score, and IFT in Figure 6. In this figure, one can see that proponent texts selected using only the CLIP score are just error messages, and rather about the decoration of the bus' appearance and the
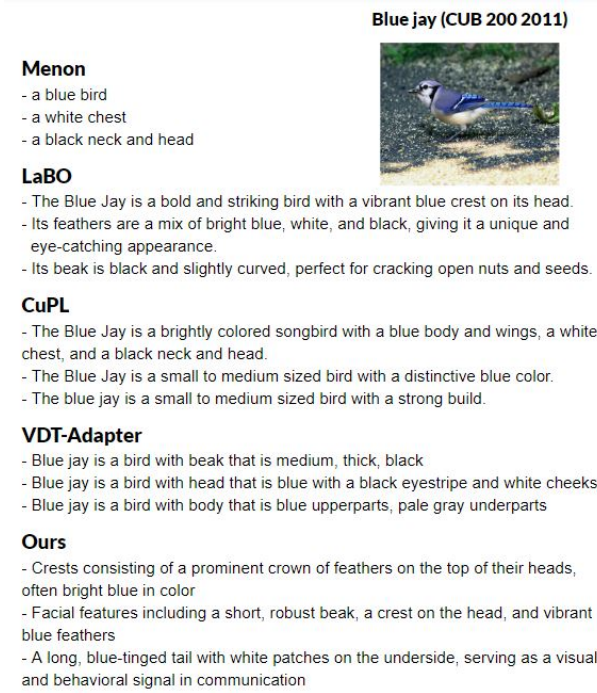
Dataset	Only Images	No Wiki	Wiki
CUB 200 2011	71.332%	72.092%	**74.525%**
OxfordPets	91.664%	92.453%	**93.396%**
CIFAR-10	94.320%	94.490%	**94.820%**
CIFAR-100	77.830%	77.410%	**78.250%**
EuroSAT	94.296%	94.333%	**96.037%**
Food101	85.848%	86.832%	**86.950%**
Miniimagenet	91.980%	91.570%	**92.480%**
102flowers	96.459%	96.703%	**97.192%**
DTD	72.074%	73.830%	**74.043%**

Table 4: Accuracies of ablation studies on Cross-Modal Transfer Classification with(Wiki) and without Wikipedia url provided to GPT-3.5(No Wiki).

material of the bus' wheel. It is hard to say that this information is helpful as visual cues to classify the bus images. The proponent texts selected using the influence score contain information about the appearance of not only the bus but also the structure inside the bus. On the other hand, all of the proponent texts determined using IFT provide clues to classify the bus images, and one can see what factors help the model to make a decision. Through IFT, we can resolve issues mentioned in baseline methods, such as visual cues that do not helpful for models when classifying images, or the same text being repeated. Through this, we can see that IFT is appropriate as a score metric for proponent texts.

To evaluate the effectiveness of our 2-stage prompting with Wikipedia URLs, we compare cross-modal transfer performance when GPT-3.5 is provided with Wikipedia URLs versus when it is not. Table 4 presents the results, showing that across all datasets, providing Wikipedia URLs improves performance. Figure 7 presents examples of proponent texts generated with Wikipedia URLs. Without Wikipedia URLs, the selected proponent texts provide only general information about birds rather than class-specific details, such as "Feathered appendages used for flight and balance" or "Feathers in various colors and patterns, providing insulation, protection, and aiding in flight." In contrast, when Wikipedia URLs are used, the descriptions become more class-specific, such as "Blackish gray feathers with a white head, dark eyes, and a yellow bill" or "Feathers consisting of black and white coloring, with long wings and a hooked bill." Additionally, utilizing external knowledge bases helps mitigate the hallucination effect of GPT-3.5. Figure 7 shows that descriptions generated with external knowledge correctly depict the beak as long, whereas those generated without it inaccurately describe the beak as short. These results demonstrate that our 2-stage prompting with

BUS (CIFAR-100)

Influence Score

A long and rectangular shape with a flat front, a spacious interior, and large windows for visibility.

Large vehicles designed to carry passengers, typically characterized by a long and rectangular body with multiple rows of seats inside.

Large glass panels providing views, natural light, and ventilation on the sides of the bus.

CLIP Score

The information could not be found.

Different colors and finishes of paint used to decorate and protect the exterior of buses.

Wheels made of rubber or other durable materials, connected to the axles of the bus to facilitate movement.

IFT

A long and rectangular shape with a flat front, a spacious interior, and large windows for visibility.

Large entrance points on the side of the bus, typically made of metal and equipped with handles, locks, and windows.

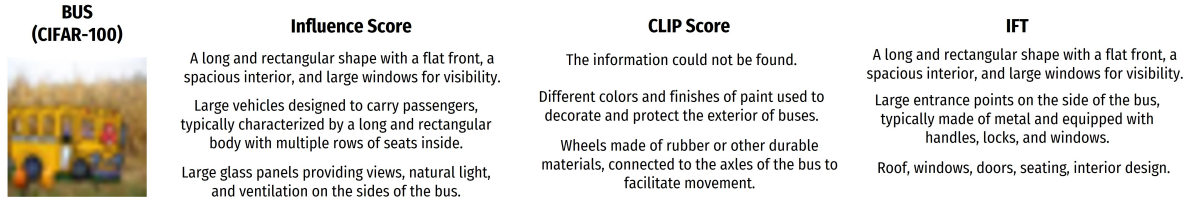Roof, windows, doors, seating, interior design.

Figure 6: Examples of selected proponent texts when using influence score (Influence Score), CLIP score (CLIP Score), and IFT (Bus class of CIFAR-100 dataset).

Black Footed Albatross (CUB 200 2011)

Proponent Texts(No Wiki)

Feathers in various colors and patterns, providing insulation, protection, and aiding in flight.

Feathered appendages used for flight and balance, typically composed of a framework of bones and covered in feathers.

A short, pointed structure used for eating and manipulating food, typically found on birds.

Proponent Texts(Wiki)

A long, strong, and hooked beak that is adapted for catching and consuming prey.

Blackish gray feathers with a white head, dark eyes, and a yellow bill.

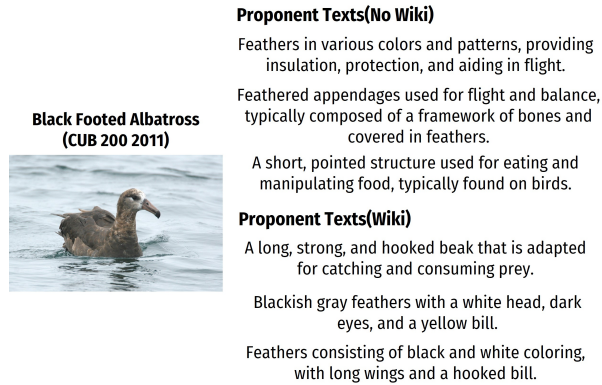Feathers consisting of black and white coloring, with long wings and a hooked bill.

Figure 7: Examples of the selected proponent texts when we provide Wikipedia urls to GPT-3.5 or not. (Black Footed Albatross class of CUB 200 2011 dataset)

Wikipedia URLs effectively provides class-specific details while reducing hallucinations.

	Menon	LaBo	CuPL	VDT	Ours
Helpful	9.87%	8.88%	20.72%	29.28%	**31.25%**
Informative	0.66%	15.13%	32.57%	21.71%	**33.55%**
Relevant	7.89%	9.54%	27.63%	**28.29%**	**28.29%**

Table 5: Top-1 Ratings (%) from GPT-4o for evaluation of textual descriptions across three criteria: Helpful, Informative, and Relevant.

	Menon	LaBo	CuPL	VDT	Ours
Helpful	2.31	1.99	2.11	2.02	**1.6**
Informative	2.61	2.09	1.90	1.95	**1.57**
Relevant	2.83	2.62	2.31	2.28	**1.69**

Table 6: Average ranks (lower is better) from GPT-4o for evaluation of textual descriptions across three criteria: Helpful, Informative, and Relevant.

4.4 Qualitative Evaluation with GPT-4o

We additionally conduct evaluations using GPT-4o to compare the quality of textual descriptions generated by our baselines. For this experiment, we randomly select 100 classes from the nine benchmark datasets. From each class, we sampled three descriptions per method (CuPL, LaBo, Menon, VDT, and Ours), along with two original images as visual references. These two images and the five descriptions (one for each method) are then provided to GPT-4o for evaluation across three criteria: Helpful, Informative, and Relevant. This setup produces three evaluation instances per class, yielding 300 evaluation cases in total. We adopt two metrics: 1) **Top-1 Rating** (Lin et al., 2023): The proportion of times each method is rated as the best for a given criterion, 2) **Ranking Average** (Xu et al., 2022): The mean rank of each method, where ties are permitted when descriptions are of similar quality. Table 5 reports the Top-1 Ratings, and Table 6 presents the Ranking Averages.

Our method consistently ranked highest across all metrics and criteria, demonstrating strong performance in generating helpful, informative, and relevant descriptions. The detailed evaluation process and prompt templates for GPT-4o evaluations are provided in Appendix C.

5 Conclusion

We propose a simple yet effective approach that generates textual descriptions that can explain the image data well using 2-stage prompting that utilizes external knowledge. By leveraging the generated textual description with the proposed IFT, defined as the sum of influence score and CLIP score, we can determine *proponent texts*, which are informative texts for explaining each image class. Furthermore, we propose a novel benchmark task named *cross-modal transfer classification*. By training the vision model with these proponent texts, we achieve improved performance compared to training with images alone. Our approach enables a language-based understanding of data, potentially enhancing the interpretability of model predictions. In quantitative evaluations, including zero-shot classification and cross-modal transfer classification, our method consistently outperforms baselines. Ablation studies, qualitative analyses, and GPT-4o evaluation collectively validate that our generated descriptions are more helpful in model training than those of the baselines.

6 Limitation

This work has several limitations. First, the extracted textual descriptions can vary depending on the pre-defined prompts and the large language model(LLM) used, affecting performance. Second, as the dataset size increases, so does the computational load required to calculate influence scores, resulting in higher computational costs. In our future work, we plan to experiment with various prompts and large language models and also explore methods to efficiently compute influence scores even as dataset sizes increase. Third, since our method relies on LLMs such as GPT-3.5/4, it may inherit the inherent biases present in these models. While we leverage external knowledge sources (e.g., Wikipedia URLs) to improve factual consistency and mitigate hallucinations, these strategies may not fully resolve deeper representational or social biases embedded in the LLMs' pretraining data.

References

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Jiayu Lin, Rong Ye, Meng Han, Qi Zhang, Ruofei Lai, Xinyu Zhang, Zhao Cao, Xuanjing Huang, and Zhongyu Wei. 2023. Argue with me tersely: Towards sentence-level counter-argument generation. *arXiv preprint arXiv:2312.13608*.

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. 2023. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 262–271.

Sachit Menon and Carl Vondrick. 2023. Visual classification via description from large language models. In *The Eleventh International Conference on Learning Representations*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Shusheng Xu, Xingxing Zhang, Yi Wu, and Furu Wei. 2022. Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11556–11565.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197.

Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. 2023. Diagnosing and rectifying vision models using language. In *International Conference on Learning Representations (ICLR)*.

A 2-stage Prompting with Wikipedia url

Since large language models generate different quality of answers depending on prompting, we provide more detailed implementation details of our 2-stage prompting with the Wikipedia url. We refer to the setting of baseline method (Menon and Vondrick, 2023). We use the prompting structure of Q: A: and provide GPT-3.5 desired output examples.

First, when we are extracting components of each class, we remind the form of questions:

Q : Can you tell me the components of {class name} from the perspective of appearance?
A :

Prior to the above question, we provide additional examples like this:

Q : Can you tell me the components of American Bulldog from the perspective of appearance?
A : 1. Coat Type and Texture 2. Coat Color 3. Body Build 4. Size 5. Head 6. Muzzle and Nose 7. Eyes 8. Ears 9. Tail 10. Legs and Paws 11. Coat Patterns 12. Facial Features 13. Unique Breed Traits

Since we provide these examples to GPT-3.5, when we ask a question, the answers in the order of 1, 2, 3 are given. So we can get the desired answers by simply removing the numbers.

Second, when we are extracting textual descriptions of each class using the components of each class and the Wikipedia url, we remind the form of questions:

Dataset	Zero-Shot	Menon	LaBo	CuPL	VDT-Adapter	Ours
CUB 200 2011	62.703%	63.807%	63.928%	65.084%	64.273%	**66.396%**
OxfordPets	89.488%	88.679%	88.814%	**93.127%**	92.453%	92.562%
CIFAR10	93.660%	92.640%	93.570%	93.120%	93.099%	**93.819%**
CIFAR100	75.690%	75.150%	74.880%	74.510%	75.250%	**75.730%**
EuroSAT	40.778%	38.630%	41.815%	**45.444%**	42.926%	43.740%
Miniimagenet	85.910%	85.590%	85.530%	86.220%	86.280%	**86.435%**
102flowers	65.690%	68.987%	70.085%	70.452%	71.062%	**72.161%**
DTD	51.968%	51.277%	54.628%	58.404%	58.404%	**59.989%**

Table 7: Zero-shot image classification results with laion/CLIP-ViT-L-14-laion2B (Cherti et al., 2023) model.

Q : Please summarize the information of appearance about {components of each class} in this {Wikipedia url} in one line composed of nouns. If you couldn't find related information, you must answer general information you know. A :

Prior to the above question, we also provide additional examples like this:

Q : Please summarize the information of appearance about nose in this url https : //en.wikipedia.org/wiki/AmericanBulldog in one line composed of nouns. If you couldn't find related information, you must answer general information you know like the above questions.
A : A short to medium-length muzzle with a nose that can be black, brown, or pigmented, often matching the coat color, and it is a distinctive feature on the breed's square-shaped head.

Since we ask for a "composed of nouns" answer, we can get a detailed and informative answer without refining it.

B Proponent Texts and IFT

In this section, we provide additional examples of proponent texts and IFT in Figure 9, 10, 11, 12, 13.

C GPT-4o Evaluation Details

To objectively assess the quality of textual descriptions generated by each method, we conduct a third-party evaluation using GPT-4o, a vision-language model with strong reasoning capabilities.

Evaluation Setup. We randomly select 100 classes across all nine benchmark datasets. For each class, we sample three descriptions from each of the five methods (Menon et al., LaBo, CuPL, VDT-Adapter, and Ours), resulting in 15 descriptions per class. We also sample two reference images per class from the original dataset. The evaluation is conducted in the zero-shot setting and produces 300 evaluation instances in total (3 per class).

Evaluation Criteria. For each evaluation instance, GPT-4o is presented with two reference images from the same class and five textual descriptions (one per method) in randomized order. The model is then asked to evaluate each description according to the following three criteria: **Helpfulness:** how useful the description is in understanding the visual content. **Informativeness:** how much specific and relevant detail the description provides. **Relevance:** how well the description aligns with the given images.

Prompt Template for Top-1 Rating. The following prompt is used to collect Top-1 rating judgments from GPT-4o. The model is instructed to return only the highest-ranked description for each evaluation criterion, along with a brief reasoning.
You are a vision-language model evaluator.
Given two images and five textual descriptions, your task is to rank the descriptions for each of the following three criteria and output only the top-1 ranking description (or group of descriptions if equally best) along with a short rationale. The criteria are:
*1. **Helpful**: Does the description help distinguish or understand the two images effectively?*
*2. **Informative**: Does the description provide detailed and meaningful content?*
*3. **Relevant**: Does the description accurately reflect the visual content of the two images?*
You are allowed to assign the same rank to multiple descriptions if you believe they are equally strong for a given criterion, but please output only the top-ranked group for each criterion.

Newfoundland (OxfordPets)

Proponent Texts

1. A massive, broad head with a muzzle that is wide and deep, and a nose that is large and may be black or brown.
2. Drop ears that are set well back and level with the eyes, and they are covered in long, thick, and waterproof fur.
3. Black, brown, gray, or Landseer (black and white) coat colors, typically thick, dense, and double-layered.

Textual Descriptions Not Chosen For Proponent Texts

1. Medium-sized, round and set apart eyes with a soft, kind expression in the Newfoundland breed.
2. A large breed with a massive and powerful build, and males typically weighing between 130-150 pounds.
3. Groomed facial features including a wide, strong head, well-developed jowls, and a gentle expression.

Osteospermum (102flowers)

Proponent Texts

1. Ray florets are typically purple or white, radiating out from a central disk, forming the petals of the daisy-like calyx.
2. Large daisy-like flowers with vibrant colors and dark centers, giving the plants a decorative and eye-catching style.
3. Daisies with daisy-like flowerheads composed of petals that are often white, purple, or pink in color.

Textual Descriptions Not Chosen For Proponent Texts

1. Fragrance may refer to the pleasant smell emitted by flowers and plants, which varies depending on the species.
2. Small to medium-size flowers that can range in diameter from 2 to 4 inches.
3. A central reproductive organ in the flower of Osteospermum plants, consisting of the stigma, style, and ovary.

Figure 8: Examples of proponent texts and non-proponent texts for the Newfoundland class of the OxfordPets dataset and the Osteospermum class of the 102flowers dataset.

Images:
Image A: <Insert Image A>
Image B: <Insert Image B>
Descriptions:
1. "{description1}"
2. "{description2}"
3. "{description3}"
4. "{description4}"
5. "{description5}"
For each criterion, please output only the top-1 ranking description(s) along with a short rationale for why these descriptions are the best.
Output format:
Helpful Ranking:
Top-1*: Descriptions 1*
Reason: "These descriptions clearly highlight key differences between the two images, such as beak shape and feather patterns."
—
Informative Ranking:
Top-1*: Descriptions 2*
Reason: "They include specific visual details such as color, size, and structural features."
—
Relevant Ranking:
Top-1*: Description 5*
Reason: "Highly aligned with actual visible features in both images."
Please ensure your ranking is thoughtful and grounded in what is visible in the two images.

Prompt Template for Mean Ranking. The following prompt is used to collect mean ranking scores from GPT-4o for each evaluation instance. The model is instructed to rank five descriptions per criterion, with ties allowed.

You are a vision-language model evaluator.
Given two images and five textual descriptions, your task is to rank the descriptions for each of the following three criteria:
*1. **Helpful**: Does the description help distinguish or understand the two images effectively?*
*2. **Informative**: Does the description provide detailed and meaningful content?*
*3. **Relevant**: Does the description accurately reflect the visual content of the two images?*
You are allowed to assign the same rank to multiple descriptions if you believe they are equally strong for a given criterion.
Images:
Image A: <Insert Image A>
Image B: <Insert Image B>
Descriptions:
1. "{description1}"
2. "{description2}"
3. "{description3}"
4. "{description4}"
5. "{description5}"
For each criterion, please list the descriptions grouped by rank, with a short rationale for each group.
Output format:
Helpful Ranking:
Rank 1*: Descriptions 2, 5*
Reason: "These descriptions clearly highlight key differences between the two images, such as beak shape and feather patterns."
Rank 2*: Description 1*
Reason: "Provides some useful context but lacks comparative elements."
Rank 3*: Descriptions 3, 4*

Reason: "These are vague or unrelated to distinguishing the images."

—

Informative Ranking:
Rank 1: Descriptions 1, 5
Reason: "They include specific visual details such as color, size, and structural features."

...

Relevant Ranking:
Rank 1: Description 5
Reason: "Highly aligned with actual visible features in both images."

...

Please make sure your ranking is thoughtful and grounded in what is visible in the two images.

D Analysis

In this section, we demonstrate that IFT can select texts that describe each image class well by showing examples of proponent texts and non-proponent texts.

Figure 8 shows examples of texts selected as proponent texts and those that are not. For the Newfoundland class of the OxfordPets dataset, non-proponent texts contain general information such as expression or weight, which do not aid in classifying Newfoundland images. For the Osteospermum class of the 102flowers dataset, non-proponent texts include information about fragrance, growth, and reproductive organs, which are not helpful for image classification.

These examples illustrate that IFT ensures proponent texts contain information that effectively helps the model classify images.

E Dataset Details

Dataset	Classes	Train Size	Validation Size	Test Size
CUB 200 2011	200	4800	1194	5794
Miniimagenet	100	40000	10000	10000
CIFAR-10	10	40000	10000	10000
CIFAR-100	100	40000	40000	10000
OxfordPets	37	5910	697	742
Food101	101	75750	10100	15150
EuroSAT	10	18900	5400	2700
102flowers	102	6552	818	819
DTD	47	1880	1880	1880

Table 8: Dataset partitions of the train, validation , and test set of total nine datasets we use.

We use a total of nine datasets. Additionally, we report the size of the train set, validation and test set of the datasets we use in Table 8.

CUB 200 2001 (Wah et al., 2011) This dataset contains a total of 200 bird species, with each species having around 30 images.

Miniimagenet (Vinyals et al., 2016) This dataset consists of 100 classes, each containing 600 images. The classes are drawn from a larger dataset called ImageNet (Deng et al., 2009), which contains over a million images across thousands of classes. It is introduced in Vinyals et al. (2016) as a benchmark dataset for few-shot learning, but we use 500 images for each class as a train set and 100 images as a test set.

CIFAR-10 (Krizhevsky et al., 2009) This is a widely-used dataset in machine learning research, consisting of 60,000 32x32 color images in 10 classes.

CIFAR-100 (Krizhevsky et al., 2009) This is an extension of CIFAR-10 dataset, consisting of 60,000 32x32 color images in 100 classes.

OxfordPets (Parkhi et al., 2012) The dataset is a collection of approximately 7,349 images containing 37 different breeds of cats and dogs. It is commonly used for tasks such as image classification and object detection in computer vision research.

EuroSAT (Helber et al., 2019) This dataset is used for classifying geographical land cover types based on satellite imagery. It comprises 10 distinct geographical landscape classes, including categories like forests, cropland, roads, buildings, and rivers, among others.

Food101 (Bossard et al., 2014) The dataset is a widely used collection of food images that is primarily employed for food recognition and image classification tasks. It consists of approximately 101,000 images, each depicting a specific food item from one of the 101 distinct food categories.

102flowers (Nilsback and Zisserman, 2008) The dataset is a collection of flower images, consisting of 102 different categories, each representing a distinct species or type of flower.

Describable Textures Dataset (DTD) (Cimpoi et al., 2014) This dataset is a collection of diverse texture images, containing 47 different categories with distinct visual features and structures.

Red winged blackbird

Black feathers with distinctive red and yellow shoulder patches, a long and slender body, and a pointed beak — 639.12

Bright yellow eyes with a black iris, providing a striking contrast against the black feathers of the bird — 637.46

Blackbird with a distinctive red-and-yellow shoulder patch, a pointed bill, and a long, deeply forked tail — 637.322

Yellow breasted Chat

Plumage patterns including streaks, spots, and bars, with a yellow breast as a distinguishing feature — 605.977

A short, slightly curved beak that is typically dark in color — 605.396

Short-taloned feet with strong claws, allowing for perching and hopping movements — 603.665

Scissor tailed flycatcher

Long, scissor-shaped tail feathers that are vibrant and distinctive in color, such as shades of blue or red — 336.916

A long, slender tail with elongated outermost feathers, giving it a distinctive scissor-like appearance — 336.664

White, light gray, and black plumage with a reddish-pink abdomen and elongated tail feathers — 335.244

American goldfinch

Plumage patterns that include a bright yellow body, black wings with white markings, and a black cap on the head — 1291.121

Small feet with sharp claws to help it cling to branches and perch on plants — 1289.333

Golden eyes, black feathers, and yellow plumage characterize the appearance of the American goldfinch — 1288.53

Figure 9: Examples of proponent images, proponent texts and IFT for four classes of CUB 200 2011 dataset.

Alpine sea holly

Purple-blue petals surrounding a central disc of stamens, each consisting of a filament and an anther — 123428.563

A dense, spiky cluster of small flowers arranged in a spherical or cylindrical shape — 123428.096

Blue or purple flowers that are surrounded by spiky bracts in shades of silvery-gray or green — 123425.124

Balloon flower

Smooth, waxy-textured leaves and vibrant balloon-shaped flowers — 348167.983

Purple, blue, pink, or white flowers with a balloon-like shape — 348166.071

A slender stem-like structure that connects the balloon-shaped flower to the main stem or branch of the plant — 348165.669

Camellia

Globular form with layered petals and a prominent central boss — 583367.245

Large, showy, and colorful petals arranged in layers around a central cluster of stamens and pistils — 583366.28

Wide range of colors including white, pink, red, and variegated, with some varieties having multiple colors on a single flower — 583363.499

Red ginger

Bright, showy, petal-like structures that are typically red in color and surround the central pistil — 113226.93

Petals that are vibrant red, elongated and curved, forming a spiral pattern around the central cone-like structure — 113224.89

Sepals are green, leaf-like structures that protect and support the flower bud — 113224.529

Figure 10: Examples of proponent images, proponent texts and IFT for four classes of 102flowers dataset.

Sphynx

Hairless cats with skin that can be various colors and patterns, including solid, mottled, or spotted — 453286.323

A bald or hairless tail that is thin and whip-like in appearance — 453286.302

Deep wrinkles are a distinguishing feature of the Sphynx breed, appearing on the forehead, around the eyes, and on the neck — 453285.455

Pug

A compact body with a wrinkled forehead, curled tail, and a distinctive pushed-in nose — 177253.953

Large, round, and prominent eyes, a flat wrinkled face, a short nose with large nostrils, and a short, square muzzle — 177253.342

A compact and well-muscled body, characterized by a deep chest, straight back, and sturdy leg — 177250.604

Chihuahua

Large, round eyes, a prominent forehead, a small nose, and characteristic butterfly ears — 340005.86

Short-haired chihuahuas have a wide range of coat colors, including but not limited to black, brown, white, fawn, and various patterns and markings — 340004.881

A tiny toy breed with a compact body and a small size — 340003.9

Keeshond

Long, thick double coat with a dense undercoat and a harsh, straight outer coat, usually in gray, black, and cream color combinations, featuring a distinctive "spectacles" pattern around the eyes — 142555.723

A wedge-shaped head with a broad skull, erect ears, and a well-defined stop — 142553.918

Keeshonds have a fox-like face with a broad, wedge-shaped head, well-set ears, and dark, almond-shaped eyes — 142553.909

Figure 11: Examples of proponent images, proponent texts and IFT for four classes of OxfordPets dataset.

Bibimbap

Thinly sliced vegetables, such as carrots, cucumbers, and mushrooms, as well as protein toppings like beef or tofu, and garnishes such as sesame seeds, scallions, and kimchi. — 332865.976

A mixture of vegetables, meat, and rice, often topped with a fried egg, served in a hot stone bowl. — 332865.652

Gochujang, a spicy red chili paste, is a common sauce used in bibimbap. — 332863.74

Chocolate cake

Glossy, smooth, and decorative layer made from melted chocolate, icing, or other sweet toppings, applied to the surface of a cake. — 626368.972

A variety of colors including shades of brown, from light to dark, as well as the possibility of additional decorative elements such as frosting or sprinkles. — 626368.571

Moist and decadent cake layers, typically round in shape, with a height ranging from one to three layers. — 626367.901

Churros

A cylindrical shape with a ridged surface. — 412924.442

Crispy, golden outer coating. — 412923.552

Sugar, cinnamon, chocolate, or other toppings and dips to add flavor and visual appeal to the churros — 412921.305

Red velvet cake

Red color, which is the traditional and characteristic color of a red velvet cake — 322350.568

Cream cheese frosting, crumbs, or sprinkles are often used as garnish for red velvet cake — 322350.357

A three-layer cake with a velvety texture and a deep red color, typically round in shape — 322349.822

Figure 12: Examples of proponent images, proponent texts and IFT for four classes of Food101 dataset.

33865

Arctic fox

Seasonal variations in the Arctic fox include changes in fur color to blend with the surrounding environment, such as white in the winter and brown or gray in the summer. — 664.998

Facial features such as a pointed snout, small ears, and bushy tail in the arctic fox. — 664.806

Short, triangular ears with a rounded tip and thick fur that helps insulate against the cold. — 662.398

Boxer

Short coat, square head, brachycephalic skull shape, strong jaws, muscular body, cropped ears, docked tail, powerful bite, energetic personality, friendly disposition, hunting instinct. — 764.084

A docked tail that is set high and carried erect, being a defining characteristic of the breed. — 762.493

Brindle, fawn, or white coat colors. — 761.774

Cliff

Brown, gray, red, white, and various shades of these colors, depending on the composition of the cliff's rocks. — 587.454

Sheer vertical face or steep slope formed by erosion, often found near bodies of water. — 587.143

The appearance of the surrounding environment of a cliff includes features such as rocks, vegetation, water, and wildlife. — 584.551

Frying pan

The bottom design of a frying pan can vary, including options such as flat, ridged, or grooved surfaces. — 1124.249

A frying pan is typically round or oval-shaped and can vary in size, with different diameters available for different cooking needs. — 1123.447

Handle, base, pan, iron, stainless steel, non-stick surface, cooking, lid — 1120.786
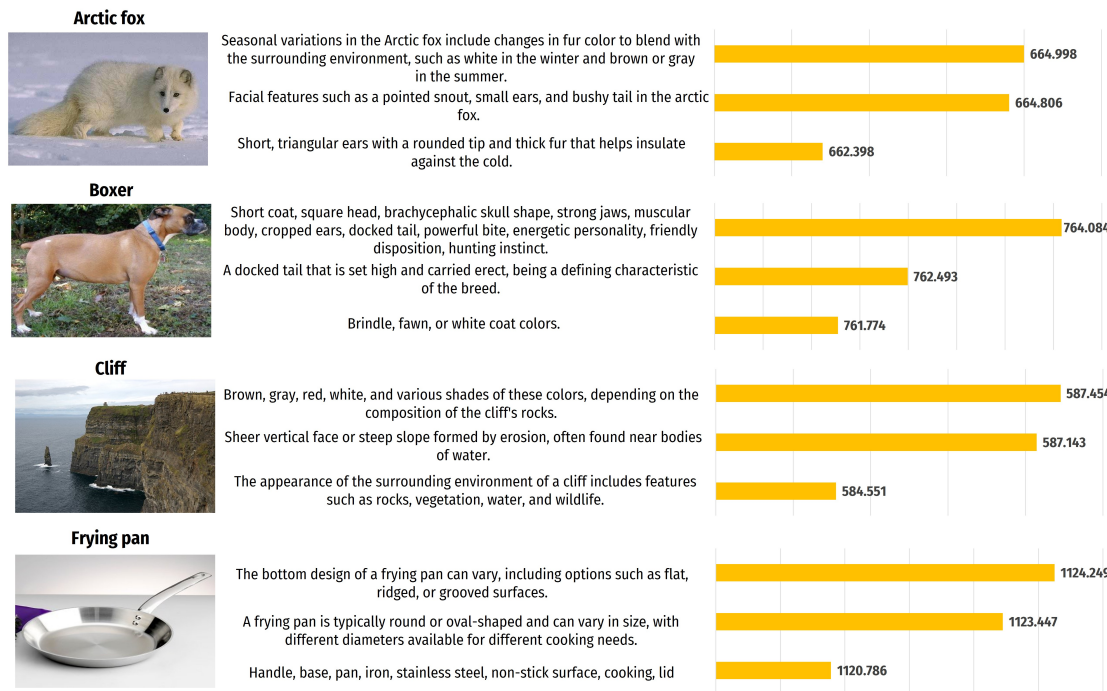
Figure 13: Examples of proponent images, proponent texts and IFT for several classes of Miniimagenet dataset.

Menon
- delicate, intricate pattern of thin, sticky threads
- radiating from a central point
- often with a spiral pattern

LaBO
- The spider web glistens in the sunlight, reflecting a rainbow of colors.
- It is intricate and delicate in design, with thin strands crisscrossing in different directions.
- From afar, it almost looks like a work of art, with its symmetrical patterns and geometric lines.

CuPL
- A spider web is a thin, sticky, arachnid-produced silk fiber that, in most cases, is used to catch prey.
- A spider web is a network of silken threads that a spider spins to catch prey.
- A spider web is a circular web that is often found in the homes of spiders.

VDT-Adapter
- The predominant color of a spider web is usually white or translucent.
- The surface quality of a spider web is typically smooth and silky.
- A spider web is generally small to medium in size.

Ours
- A complex and intricate geometric arrangement of silk threads forming concentric circles and radial spokes, often with a central hub and various supporting strands, designed to capture prey and provide structural stability
- Fine, strong and elastic fibers that are produced by spiders and used to create intricate webs for trapping prey
- Silk threads arranged in a circular or geometric pattern, typically found in outdoor areas such as gardens, forests, and bushes

Spider Web (Miniimagenet)

Figure 14: Example of a qualitative sample for evaluating textual descriptions. (Spider Web class of Miniimagenet dataset)

Menon
- small size
- short coat
- black, brown, or red coloration

LaBO
- The miniature Pinscher is a small breed of dog, standing at an average height of 10-12 inches.
- They have a compact and muscular body, giving them a sturdy and agile appearance.
- Their head is slightly elongated with a moderately long muzzle and small, alert ears that stand erect.

CuPL
- A miniature pinscher is a small, short-haired dog with smooth, glossy fur.
- The miniature pinscher is a small, short-haired dog with a stout body.
- A miniature pinscher is a small, slim dog with a short coat.

VDT-Adapter
- Miniature Pinscher is small sized dog
- Miniature Pinscher is a dog with fur length that is short
- Miniature Pinscher is a dog with fur texture that is smooth

Ours
- A small and compact head with a well-defined stop, high-set ears, and a distinctively shaped skull
- Clear, self-colored, or two-toned with specific markings such as rust or tan
- Ears that are set high on the head and are usually cropped to stand erect, along with a well-defined stop and a strong, blunt muzzle

Figure 15: Example of a qualitative sample for evaluating textual descriptions. (Miniature Pinscher class of Oxfordpets dataset)