Understanding Retrieval-Augmented Task Adaptation for Vision-Language Models

Yifei Ming [1] **Yixuan Li** [1]

Abstract

Pre-trained contrastive vision-language models have demonstrated remarkable performance across a wide range of tasks. However, they often struggle on fine-trained datasets with categories not adequately represented during pre-training, which makes adaptation necessary. Recent works have shown promising results by utilizing samples from web-scale databases for retrieval-augmented adaptation, especially in low-data regimes. Despite the empirical success, understanding how retrieval impacts the adaptation of vision-language models remains an open research question. In this work, we adopt a reflective perspective by presenting a systematic study to understand the roles of key components in retrieval-augmented adaptation. We unveil new insights on uni-modal and cross-modal retrieval and highlight the critical role of logit ensemble for effective adaptation. We further present theoretical underpinnings that directly support our empirical observations.

1. Introduction

Contrastive vision-language pre-training has emerged as a fundamental cornerstone for a wide array of tasks in natural language processing and computer vision (Radford et al., 2021; Jia et al., 2021; Yang et al., 2022; Li et al., 2022b; Mu et al., 2022; Yu et al., 2022; Sun et al., 2023; Xu et al., 2024). These models excel in capturing the intricate relationships present in both visual and textual data, enabling them to understand context, semantics, and associations holistically. It is now a common practice to employ aligned multi-modal features from web-scale pre-training. However, a challenge arises when these pre-trained models encounter real-world downstream datasets, particularly in low-data (few-shot) scenarios. Such datasets often encompass fine-grained categories that were not adequately represented during the initial pre-training phase, posing a notable hurdle for the models in adapting to these nuanced distinctions.

In the low-data regime, retrieval-augmented adaptation has demonstrated promise, where a wealth of external resources is readily available on the Internet and can be retrieved efficiently to enhance adaptation. Recent works (Udandarao et al., 2023; Zhang et al., 2023) showcase encouraging results by leveraging large-scale text and image databases (Schuhmann et al., 2022). Retrieval-augmented adaptation involves two main steps: first retrieving the most relevant data from an external source, and then adapting to downstream task based on the retrieved samples. While existing works have primarily focused on developing new adaptation algorithms or integrating different knowledge sources, ***there remains a notable gap in understanding how retrieval augmentation impacts adaptation for vision-language models***. Such an understanding is imperative to guide the future development of effective algorithms.

In this work, we adopt a reflective perspective by presenting a systematic study to understand retrieval-augmented adaptation, and establishing new theoretical underpinnings. Our empirical analysis reveals key insights revolving around two aspects: **(1)** the impact of the retrieval method, and **(2)** how retrieved samples help adaptation. First, we show that image-to-image (I2I) retrieval consistently outperforms text-to-image (T2I) retrieval for a wide range of downstream tasks. Under the same retrieval budget, these two retrieval methods differ by the query samples used: I2I employs a few seed images from the target data distribution, whereas T2I employs the textual description of each class label. While both I2I and T2I retrieval introduce distributional shifts *w.r.t.* the target data, we show that I2I achieves strong performance that matches more closely with the oracle when we directly retrieve from the target distribution (*i.e.*, no distributional shifts). Secondly, we show that ensembling the zero-shot prediction together with I2I retrieved samples is the key to improved adaptation performance. For a given test sample, the ensembling is achieved by taking a weighted average between the logit from the retrieved feature cache and the logit of the zero-shot inference. We

[1] Department of Computer Sciences, University of Wisconsin-Madison. Correspondence to: Yifei Ming <ming5@wisc.edu>, Yixuan Li <sharonli@cs.wisc.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

empirically find that without ensembling, the performance of retrieval-augmented adaptation significantly degrades. This new observation complements previous studies that often attribute the success of retrieval to the diversity and quality of samples.

Going beyond empirical analysis, we provide theoretical insights that directly support our empirical observations above. We formalize T2I and I2I retrieval by characterizing the multi-modal feature space with each retrieval scheme. Under realistic assumptions, we analyze how retrieval impacts the modality gap and the shift between the retrieved and target distributions. In particular, we prove that I2I retrieval is superior to T2I retrieval (**Theorem 4.1**) and that logit ensemble is critical for improving CLIP-based adaptation (**Theorem 4.2**) by better leveraging the knowledge encoded in different modalities. Our theoretical results shed light on the key factors in the design of effective retrieval-augmented adaptation algorithms for vision-language models.

Our main contributions are summarized as follows:

- We conduct a timely and systematic investigation into the retrieval-augmented adaptation of vision-language models, where we highlight key components such as the retrieval methods and logit ensemble.
- We provide a finer-grained empirical study with in-depth analysis. We unveil new insights on the critical role of uni-modal retrieval and logit ensemble for effective CLIP-based adaptation in low-data scenarios.
- We develop a novel theoretical framework for retrieval-augmented adaptation and present theoretical results that directly support our empirical observations.
- We further provide a comprehensive ablation study and discuss alternative design choices such as the impact of model architectures, adaptation with a finetuned feature cache, and adaptation with data mixtures.

2. Retrieval-Augmented Task Adaptation

In this section, we first discuss the preliminaries of contrastive vision-language models as well as the external databases employed for retrieval (Section 2.1). Next, we illustrate the two main steps for retrieval-augmented task adaptation: building a feature cache by retrieving relevant samples from the external database (Section 2.2), and performing task adaptation based on retrieved samples (Section 2.3). An illustration of the pipeline is shown in Figure 1.

2.1. Preliminaries

Popular contrastive vision-language models such as CLIP (Radford et al., 2021) adopt a dual-stream architecture with one text encoder $\mathcal{T}: t \to \mathbb{R}^d$ and one image encoder $\mathcal{I}: \mathbf{x} \to \mathbb{R}^d$. The model is pre-trained on a massive web-scale image-caption dataset with a multi-modal contrastive loss, which aligns features from different modalities. This alignment of multi-modal embeddings offers distinct advantages for contemporary large-scale multi-modal vector databases (Schuhmann et al., 2022), enabling efficient retrieval based on semantic similarity.

Zero-shot inference. At inference time, given a test input $\mathbf{x}$, we can obtain the cosine similarity $f_c^{\text{ZOC}}(\mathbf{x}) = \text{sim}(\mathcal{I}(\mathbf{x}), \mathcal{T}(t_c))$ between the visual embedding $\mathcal{I}(\mathbf{x})$ and contextualized representations $\mathcal{T}(t_c)$ for each label $c \in \{1, 2, ..., C\}$. Here the context t_c can be either a generic template such as "`a photo of <CLASS>`" or a textual description of the class. We denote the logit vector of the zero-shot model as $f^{\text{ZOC}}(\mathbf{x}) \in \mathbb{R}^C$, which consists of C cosine similarities. The class prediction can be made based on the maximum cosine similarity among C classes.

External web-scale knowledge base. Pre-trained CLIP models often struggle for downstream datasets with finer-grained categories, which are not well represented in the pre-training dataset. To adapt CLIP models to finer-grained datasets in a low-data scheme, recent works (Liu et al., 2023) demonstrate promising performance by utilizing external resources such as LAION (Schuhmann et al., 2022), a web-scale knowledge base which consists of billions of image-text pairs $\mathcal{S}_L = \{(\mathbf{x}_i, t_i)\}_{i=1}^N$ covering a diverse range of concepts in the real world. Given a fixed budget, we can efficiently build a few-shot cache by retrieving relevant samples from the knowledge base with approximate KNN search (Johnson et al., 2019). We provide details as follows.

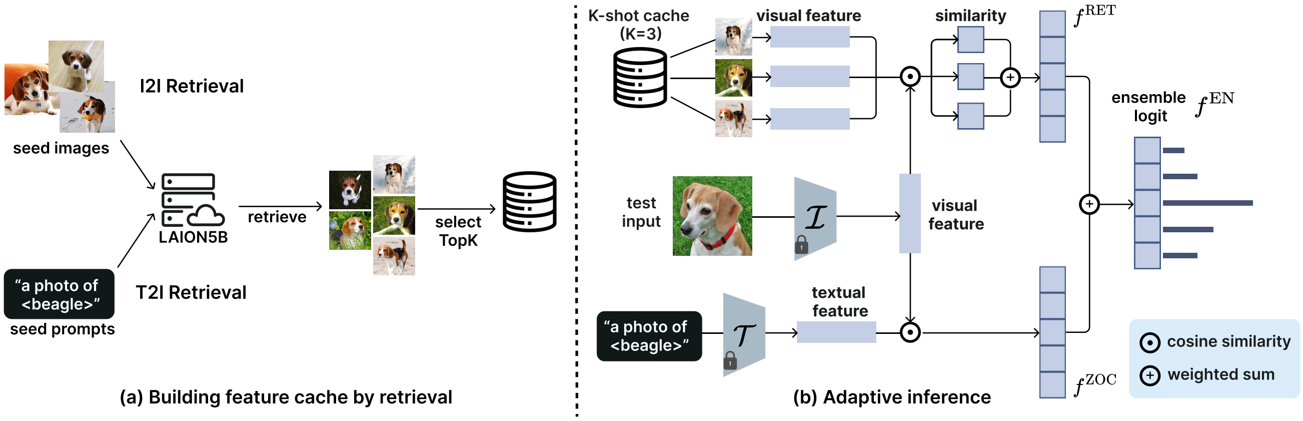
2.2. Building Feature Cache by Retrieval

Given a downstream dataset with C classes: $\mathcal{Y} = \{1, 2, ..., C\}$ and a budget size of KC, we can retrieve K samples per class to build a cache of size KC. For vision-language models, the retrieval methods be categorized as uni-modal and cross-modal retrieval, formalized as follows:

Uni-modal retrieval. We mainly consider image-to-image (I2I) retrieval due to its popularity. For I2I retrieval, we assume access to a small set of query images from the downstream dataset. The query set $\mathcal{Q}_I = \bigcup_{c=1}^C \mathcal{Q}_I^c$, where $\mathcal{Q}_I^c = \{\mathbf{x}_{c,1}, \mathbf{x}_{c,2}, \ldots, \mathbf{x}_{c,n_c}\}$ contains n_c seed images for each class $c \in \mathcal{Y}$. We then retrieve top-K similar images from $\mathcal{S}_L$ per class:

$$\mathcal{R}^{\text{I2I}}(c) = \text{top}_K \left\{ \mathbf{x} \in \mathcal{S}_L : \text{sim}(\mathcal{I}(\mathbf{x}), \mathcal{I}(\mathbf{x}_{c,i})), \mathbf{x}_{c,i} \in \mathcal{Q}_I \right\},$$

where $\text{sim}(\mathcal{I}(\mathbf{x}), \mathcal{I}(\mathbf{x}_{c,i}))$ is the cosine similarity between the image embedding of $\mathbf{x}$ from retrieval database and the query image $\mathbf{x}_{c,i}$, and top_K denotes the operation of selecting the top-K items. We can build a K-shot cache for I2I



(a) Building feature cache by retrieval

(b) Adaptive inference

Figure 1. Illustration of the retrieval-augmented task adaptation framework for CLIP-like models. (a): Given a downstream target dataset, we first retrieve relevant samples from a web-scale database using seed prompts (T2I) or seed images (I2I). We can then build a K-shot cache by selecting the Top-K similar images per class based on CLIP embeddings. (b) At inference time, the final logit f^{EN} of a test input is an ensemble (weighted sum) of logits from the zero-shot model f^{ZOC} and the few-shot cache f^{RET}.

retrieval by taking the union of these sets across all classes:

$$\mathcal{S}_R^{\text{I2I}} = \bigcup_{c \in \mathcal{C}} \left\{ (\mathbf{x}, t) \in \mathcal{S}_L : \mathbf{x} \in \mathcal{R}^{\text{I2I}}(c) \right\}.$$

Cross-modal retrieval. We mainly consider text-to-image (T2I) retrieval. We assume access to class names in the target dataset, also known as "name-only transfer" (Udandarao et al., 2023). The query set $\mathcal{Q}_T = \{t_c\}_{c=1}^{C}$, where t_c is a generic textual description of class c. The retrieved K samples for class c is:

$$\mathcal{R}^{\text{T2I}}(c) = \text{top}_K \left\{ \mathbf{x} \in \mathcal{S}_L : \text{sim}(\mathcal{I}(\mathbf{x}), \mathcal{T}(t_c)), t_c \in \mathcal{Q}_T \right\},$$

where $\text{sim}(\mathcal{I}(\mathbf{x}), \mathcal{T}(t_c))$ is the cosine similarity between the image embedding of $\mathbf{x}$ and the text embedding for class c. The K-shot cache for T2I retrieval is denoted as:

$$\mathcal{S}_R^{\text{T2I}} = \bigcup_{c \in \mathcal{C}} \left\{ (\mathbf{x}, t) \in \mathcal{S}_L : \mathbf{x} \in \mathcal{R}^{\text{T2I}}(c) \right\}.$$

2.3. Task Adaptation with Retrieved Samples

Given a K-shot cache ($\mathcal{S}_R^{\text{I2I}}$ or $\mathcal{S}_R^{\text{T2I}}$) and pre-trained CLIP image and text encoders $\mathcal{I}$ and $\mathcal{T}$, we can perform adaptation *w.r.t.* a fine-grained target dataset. To better understand the effects of retrieved samples, we consider zero-shot adaptation in Section 3, where the cache only consists of retrieved samples. We discuss few-shot adaptation in Section 5, where the cache contains a mixture of samples in the target training set and retrieved samples.

Retrieval-based adaptation. A variety of cache-based adaptation methods have been recently proposed (Zhang et al., 2022a; 2023; Udandarao et al., 2023). At the core, these methods typically obtain a logit ensemble for each test input based on two sources: (1) a logit from the zero-shot CLIP model, and (2) a logit from the cache. Without loss of generality, we consider a representative adaptation framework TipAdaptor (Zhang et al., 2022a). Specifically, given the cache of size CK (consisting of C classes with K retrieved samples per class), we denote the collection of the visual features as $\mathbf{K} = [\mathbf{k}_{1,1}, \mathbf{k}_{1,2}, \cdots, \mathbf{k}_{C,K}] \in \mathbb{R}^{d \times CK}$ where $\mathbf{k}_{c,i} = \mathcal{I}(\mathbf{x}_{c,i})$. For each test input $\mathbf{x}$, we can obtain CK cosine similarities $s_{c,i}(\mathbf{x}) = \text{sim}(\mathcal{I}(\mathbf{x}), \mathbf{k}_{c,i})$. The cosine similarities are then scaled by an exponential function $\tilde{s} : s \mapsto \exp(-\omega + \omega s)$ with a hyperparameter ω that modulates the sharpness. Accordingly, we can obtain an average similarity vector for each class based on visual features, $f_c^{\text{RET}}(\mathbf{x}) = \frac{1}{K} \sum_{i=1}^{K} \tilde{s}_{c,i}(\mathbf{x})$. The final logit of the test sample is an ensemble of logits from the feature cache and zero-shot CLIP prediction:

$$f^{\text{EN}}(\mathbf{x}) = \alpha f^{\text{ZOC}}(\mathbf{x}) + \gamma f^{\text{RET}}(\mathbf{x}),$$

where α, γ weigh the relative importance between two logits. Such a logit ensemble scheme has also been commonly adopted in recent works (Zhang et al., 2023). For completeness, we also discuss learning-based adaptation by setting visual features in $\mathbf{K}$ as learnable parameters. We denote the method as `Ensemble(F)`, where F stands for fine-tuning.

3. A Finer-Grained Analysis of Retrieval-Augmented Adaptation

Different from recent works on algorithm design and incorporation of new knowledge sources (Zhang et al., 2023; Iscen et al., 2023; Udandarao et al., 2023), the goal of our work is to present a systematic analysis with theoretical insights on how retrieval augmentation impacts adaptation for vision-language models. In this section, we present empirical analysis focusing on the impact of two aspects: retrieval

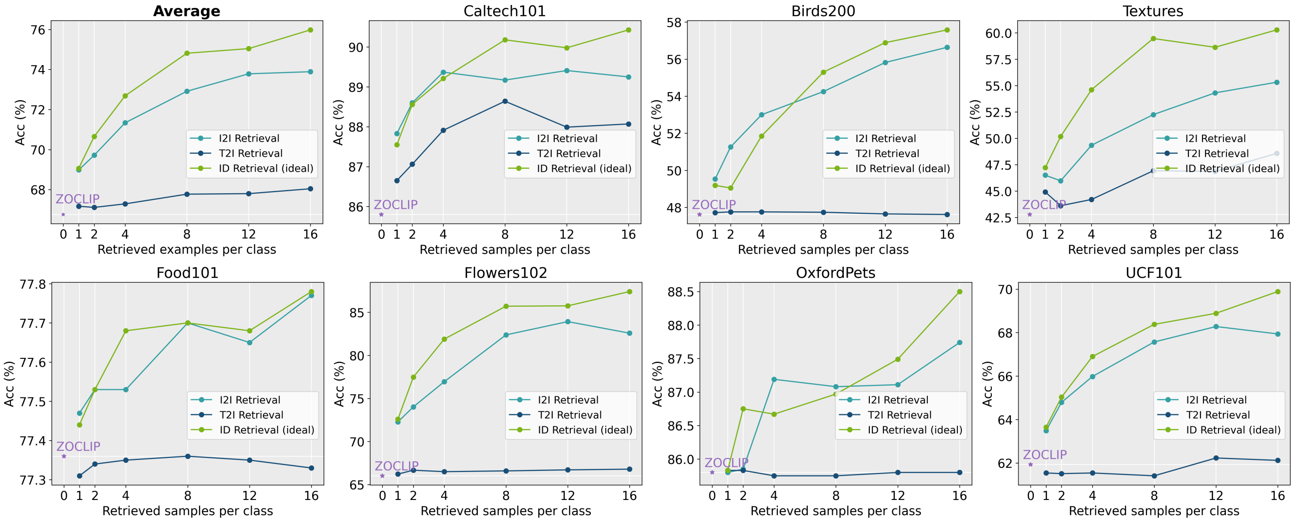


Figure 2. Comparison of adaptation performance (in accuracy) of different retrieval methods. Compared to the zero-shot model (purple star), I2I retrieval significantly improves the performance and consistently outperforms T2I retrieval across shots and datasets.

method (Section 3.2) and logit ensemble with retrieved samples (Section 3.3). We will provide theoretical analysis to support these empirical findings in Section 4. We discuss alternative design choices and ablation studies in Section 5.

3.1. Settings

Datasets. Following prior works (Zhang et al., 2022a), we consider a wide range of real-world datasets that span both common and finer-grained categories: Caltech101 (Fei-Fei et al., 2004), Birds200 (Wah et al., 2011), Food101 (Bossard et al., 2014), OxfordPets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), Textures (Cimpoi et al., 2014), and UCF101 (Soomro et al., 2012).

Implementation details. We use LAION-5B (Schuhmann et al., 2022) as the retrieval database, which consists of 5.85 billion image-text pairs. For T2I retrieval, the default query set contains class descriptions with a prompt template. For I2I retrieval, by default, we use 8 seed images per class as the query set. Based on the query set, we use the clip-retrieval tool[1] for efficient retrieval from LAION-5B. We vary the number of retrieved samples per class $K \in \{1, 2, 4, 8, 16\}$. For adaptation, we use pre-trained CLIP with RN50 backbone as the default. Unless otherwise specified, each reported result is averaged over three independent runs. The ensemble weights of two logits α, γ are tuned on the validation set. Ablation studies on the number of seed images and alternative backbones are in Section 5. Further implementation details can be seen in Appendix A.

3.2. Impact of Retrieval Method

I2I retrieval consistently outperforms T2I retrieval. To better understand the impact of the retrieval method, we compare the adaptation performance (in Accuracy) using I2I and T2I retrieval. The results are shown in Figure 2, where the horizontal axis indicates the number of retrieved samples for each class (shot). As both I2I and T2I retrieval introduce distributional shifts *w.r.t.* the target distribution, we also plot the oracle performance when retrieving samples from the target training set for reference, denoted as ID retrieval (green). Directly retrieving from the target training set can be viewed as performance upper bound.

We observe several salient trends: **(1)** I2I retrieval consistently outperforms T2I retrieval across all shots and datasets. In particular, the gap between I2I and T2I increases when increasing the shot. **(2)** Compared to the zero-shot inference without knowledge augmentation (purple star), I2I retrieval significantly improves the performance. Notably, the gap between I2I retrieval and ID-retrieval (ideal) can be as small as 1% on average (12 shots), highlighting the potential of utilizing retrieved samples in the extremely low-data scheme where one does not have training data in the target dataset. **(3)** While T2I retrieval obtains a diverse collection of samples, the performance gain compared to the zero-shot CLIP for multiple datasets can be marginal. We investigate the reasons by a detailed examination of retrieved samples next and provide theoretical understanding in Section 4 (Theorem 4.1). Similar trends also hold for training-based adaptation, where we finetune the cache features as in Zhang et al. (2022a) (see Figure 10 in Appendix E).

[1] https://github.com/rom1504/clip-retrieval

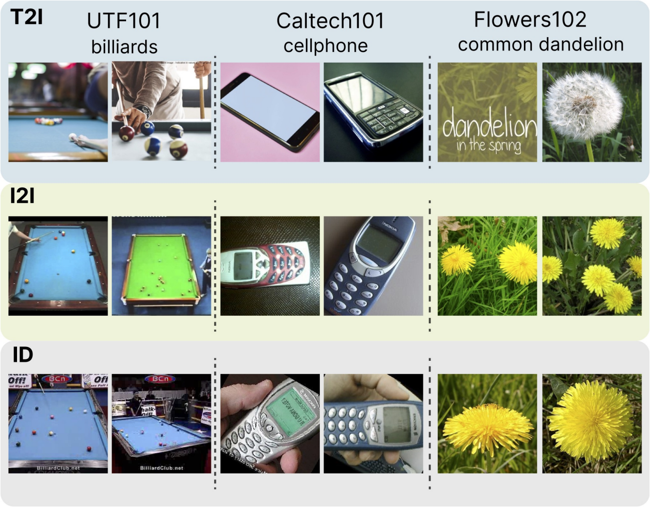


Figure 3. Samples from T2I and I2I retrieval. Top row: the main source of noise for T2I retrieval is semantic ambiguity, as the textual queries (*e.g.*, `a photo of a cellphone`) may not accurately describe the images from target distributions (*e.g.*, cellphones typical in the early 2000s). Middle row: samples retrieved by I2I matches more closely with ID data. Bottom row: images sampled from the target (ID) distribution. More examples can be seen in Appendix C.

A closer look at retrieved samples. To better understand the effects of retrieval, we examine the samples retrieved by T2I and I2I respectively. The results are shown in Figure 3. While T2I retrieval often results in a diverse collection of images corresponding to the class semantics, we find that such diversity may not always be desirable for target task adaptation. For example, when using the query `a photo of a cellphone`, we retrieve images with a broad range of cellphone types. However, the downstream dataset contains cellphones typical in the 2000s with physical keypads. The same phenomenon widely exists in the suite of datasets commonly used in the literature (see Appendix C for more extensive examples) As a result, T2I retrieval can lead to undesirable performance due to semantic ambiguity. In contrast, I2I retrieval mitigates such ambiguity. For example, when using an image of a cellphone with smaller screens and physical keypads, one can retrieve images of older models of cellphones with similar layouts (middle row).

3.3. How Do Retrieved Samples Help Adaptation?

Ensemble with zero-shot prediction is the key. We show that ensembling the zero-shot prediction together with I2I-retrieved samples is the key to improved adaptation performance. The results are shown in Figure 4, where ensemble denotes using $f^{\text{EN}} = \alpha f^{\text{ZOC}} + \gamma f^{\text{RET}}$ with $\alpha, \gamma \in (0, 1)$, RET denotes only using f^{RET} ($\alpha = 0, \gamma = 1$), and ZOCLIP means only using f^{ZOC} ($\alpha = 1, \gamma = 0$). This interesting phenomenon highlights the importance of logit ensembling for adapting vision-language models to downstream tasks. The benefits can also be seen by examining the class-wise performance of RET and Ensemble (see Figure 8 in Appendix B). Similar trends also hold for training-based adaptation, denoted as Ensemble (F), where we finetune the cache features as in Zhang et al. (2022a). Next, we provide further theoretical explanations (Theorem 4.2).

4. Theoretical Understanding

We now provide theory to support our empirical observations and formally understand retrieval-augmented task adaptation. As an overview, **Theorem 4.1** shows why I2I retrieval is superior to T2I retrieval. We further prove that logit ensemble is the key for retrieval-augmented adaptation in **Theorem 4.2**. These two theorems justify our empirical results in Section 3. Full proof is in Appendix D.

4.1. Problem Setup

Given a downstream task with C classes, let $[C] := \{1, 2, \cdots, C\}$. $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_C] \in \mathbb{R}^{d\times C}$ denotes the text embedding matrix for all classes, where $\mathbf{t}_c := \mathcal{T}(t_c) \in \mathbb{R}^d$ and t_c is a generic textual description of class c. Recall that $\mathbf{K} = [\mathbf{k}_{1,1}, \mathbf{k}_{1,2} \ldots, \mathbf{k}_{C,K}] \in \mathbb{R}^{d\times CK}$ denotes the embedding matrix for retrieved images, where $\mathbf{k}_{c,i} := \mathcal{I}(\mathbf{x}_{c,i}) \in \mathbb{R}^d$. For notational simplicity, we assume text and image features are ℓ_2 normalized. Let $\bar{\mathbf{K}} = \frac{\mathbf{K}\mathbf{V}^\top}{K} \in \mathbb{R}^{d\times C}$ contain the average retrieved feature for each class. $\mathbf{V} \in \mathbb{R}^{C\times CK}$ is a sparse matrix containing the one-hot labels for retrieved samples with entries $\mathbf{V}_{i,j} = \mathbb{1}\{i = \tilde{j}\}$ for $i \in [C], j \in [CK]$, where $\tilde{j} := \lceil \frac{j}{K} \rceil$ (Zhang et al., 2022a). For example, when $K = 2, C = 3$, we have:

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

At inference time, let $(\mathbf{x}, y) \sim \mathcal{D}_T$ be a test sample from the target distribution $\mathcal{D}_T$ with label $y \in [C]$ and its visual feature $\mathbf{z} := \mathcal{I}(\mathbf{x})$. The final logit for the test sample can be represented as a weighted sum (ensemble) of logits from the zero-shot CLIP and the feature cache from retrieval:

$$f(\mathbf{x}) = (\alpha\mathbf{T} + \gamma\bar{\mathbf{K}})^\top \mathbf{z},$$

where $0 \le \alpha, \gamma \le 1$.

Given a loss function ℓ (e.g., cross-entropy), the risk on the downstream distribution is $\mathcal{L}(f) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T}[\ell(f(\mathbf{x}), y)]$. To simplify notations, we denote the risk as $\mathcal{R}(\mathbf{Q}) := \mathbb{E}\left[\ell(\mathbf{Q}^\top\mathbf{z}, y)\right]$ for some $\mathbf{Q} \in \mathbb{R}^{d\times C}$. For example, the risk of logit ensemble is $\mathcal{R}(\alpha\mathbf{T} + \gamma\bar{\mathbf{K}})$.

Modality gap and retrieval distribution shift. To understand the impact of retrieval, we characterize the distribu-

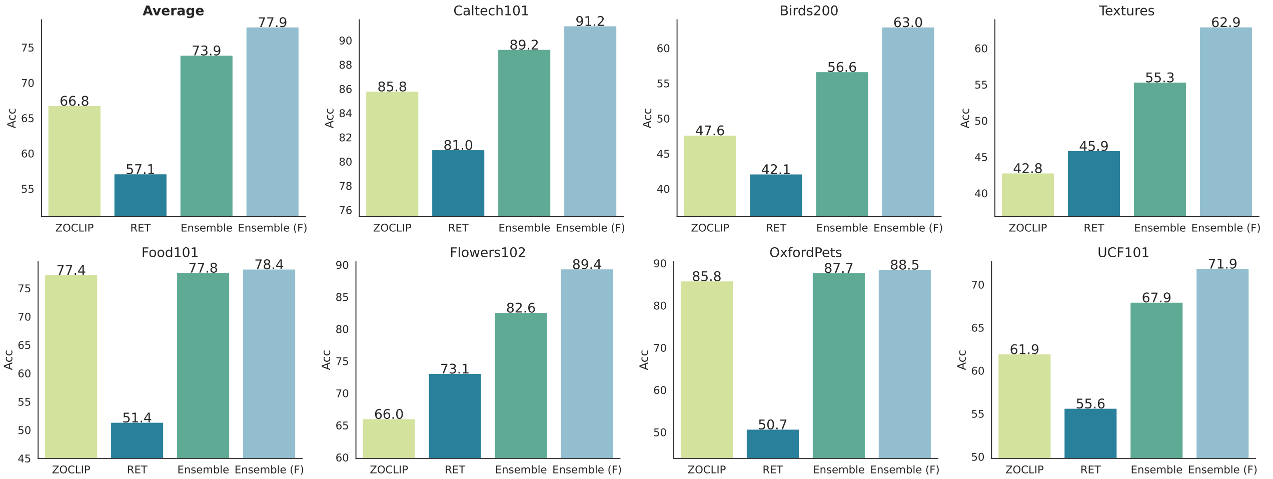


Figure 4. Importance of ensemble for I2I retrieval. Ensemble corresponds to the default logit ensemble: $f^{\text{EN}} = \alpha f^{\text{ZOC}} + \gamma f^{\text{RET}}$ with $\alpha, \gamma \in (0, 1)$. RET denotes only using f^{RET} ($\alpha = 0, \gamma = 1$) and ZOCLIP denotes only using f^{ZOC} ($\alpha = 1, \gamma = 0$). By ensembling the prediction with retrieved samples ($K = 16$), the performance improvement over zero-shot prediction is significant for most datasets.

tional shift between the retrieved data and downstream data in the feature space. We define $\bar{\mathbf{s}}_c := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T}[\mathcal{I}(\mathbf{x})|y = c]$ as the image representation of class $c \in [C]$ based on the downstream distribution. Let $\bar{\mathbf{S}} := [\bar{\mathbf{s}}_1, \ldots, \bar{\mathbf{s}}_C]$. We define the distributional shift between the retrieved data and target data for T2I and I2I retrieval as ξ_c^{T2I} and ξ_c^{I2I} for class c. Let $\xi_{\mathbf{t}} := \max_{c\in[C]} \xi_c^{\text{T2I}}$ and $\xi_{\mathbf{s}} := \max_{c\in[C]} \xi_c^{\text{I2I}}$ (Definition D.4). We can obtain an upper bound for $\xi_{\mathbf{s}}$ and a lower bound for $\xi_{\mathbf{t}}$ by Lemma D.10.

4.2. Main Results

Under realistic assumptions of T2I and I2I retrieval on the pre-trained feature space, we present two key results below. The detailed versions with full proof are in Appendix D.

Theorem 4.1 (Benefit of uni-modal retrieval)**.** *With probability at least $1 - \delta$, the following upper bound of the ensemble risk holds:*

$$\mathcal{R}(\alpha\mathbf{T} + \gamma\bar{\mathbf{K}}) - \mathcal{R}(\bar{\mathbf{S}})$$
$$\le L\left(\alpha\underbrace{\|(\mathbf{T}-\bar{\mathbf{S}})^\top\mathbf{z}\|_2}_{\textit{modality gap}} + \underbrace{\gamma\kappa\sqrt{\frac{8C}{K}\log\frac{C}{\delta}}}_{\textit{retrieval sample complexity}} + \underbrace{\gamma\sqrt{2C\xi}}_{\textit{retrieval shift}}\right),$$

where $L \le \sqrt{\exp(2)+1}$, κ characterizes the inner-class feature concentration (Definition D.1), and ξ is either $\xi_{\mathbf{s}}$ for I2I retrieval or $\xi_{\mathbf{t}}$ for T2I retrieval.

Interpretations: The above upper bound consists of three terms: the gap between the textual and visual modality, the sample complexity of retrieved features which decreases as we increase K, and a term related to the distributional shift induced by the retrieval method. By Lemma D.10, we can further show that I2I provably outperforms T2I retrieval due to a smaller ξ.

Further, to understand the benefit of logit ensemble, we define the following three events:

$$E_1 := \{(\mathbf{x}, y) \sim \mathcal{D}_T | y \neq \operatorname*{argmax}_{c\in[C]} \mathbf{t}_c^\top \mathbf{z}, y \neq \operatorname*{argmax}_{c\in[C]} \bar{\mathbf{k}}_c^\top \mathbf{z}\}$$
$$E_2 := \{(\mathbf{x}, y) \sim \mathcal{D}_T | y = \operatorname*{argmax}_{c\in[C]} \mathbf{t}_c^\top \mathbf{z}, y \neq \operatorname*{argmax}_{c\in[C]} \bar{\mathbf{k}}_c^\top \mathbf{z}\}$$
$$E_3 := \{(\mathbf{x}, y) \sim \mathcal{D}_T | y \neq \operatorname*{argmax}_{c\in[C]} \mathbf{t}_c^\top \mathbf{z}, y = \operatorname*{argmax}_{c\in[C]} \bar{\mathbf{k}}_c^\top \mathbf{z}\}$$

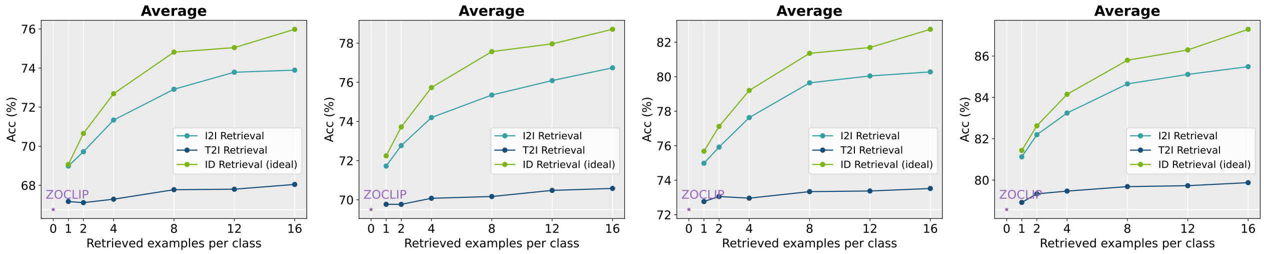
Here E_1 indicates that both f^{ZOC} and f^{RET} incorrectly classify the test sample, while E_2 and E_3 denote the event where only one of them makes a correct prediction. We can see that $\mathcal{R}_{0-1}(f^{\text{ZOC}}) = \Pr(E_1) + \Pr(E_3)$ and $\mathcal{R}_{0-1}(f^{\text{RET}}) = \Pr(E_1) + \Pr(E_2)$.

Theorem 4.2 (Benefit of logit ensemble)**.** *Under realistic assumptions for I2I retrieval, when $\alpha = \gamma = \frac{1}{2}$, we can upper bound the 0-1 risk of logit ensemble:*

$$\mathcal{R}_{0-1}(f) \le \Pr(E_1) + C_1(\Pr(E_2) + \Pr(E_3)) + \rho_c$$

where $C_1 := \rho_d \max\{6\kappa - \nu, 2\kappa + \tau\}$ is a term related to modality gap, inner-class feature concentration, and inter-class separation. ρ_c characterizes the ratio of outliers. See Appendix D for detailed definitions of $\kappa, \tau, \nu, \rho_c$, and ρ_d.

Interpretations: The above theorem characterizes the 0-1 risk upper bound by the modality gap and key properties of retrieved and target distributions. Moreover, logit ensemble utilizes knowledge encoded in different modalities to benefit each other. We can further show that under some conditions (detailed in Appendix D), logit ensemble leads to a lower 0-1 risk (*i.e.,* higher accuracy) than the zero-shot model.



(a) RN50 (b) ViT-B/32 (c) ViT-B/16 (d) ViT-L/14

Figure 5. Impact of architecture. We report the average performance (over all datasets) for I2I retrieval and T2I retrieval under different CLIP backbones and observe consistent trends. Results for individual datasets can be seen in Appendix F.

5. Discussion of Design Choices

In this section, we discuss the impact of other design choices for retrieval-augmented adaptation.

Impact of model architecture. We conduct an ablation study on the impact of model architectures. We consider CLIP with ResNet (RN50) and ViT (Dosovitskiy et al., 2021) backbones (CLIP-B/32, CLIP-B/16, CLIP-L/14), where the vision encoder is based on ViT-B/32 and ViT-L/14, respectively. The results are shown in Figure 5. We observe that a similar trend holds for CLIP with various backbones, where I2I retrieval consistently outperforms T2I retrieval. In particular, larger backbones such as CLIP-L/14 lead to overall superior performance compared to smaller backbones across the number of retrieved samples per class.

Impact of the number of seed images. To investigate the impact of seed images on I2I retrieval, we adjust the number of seed images per class from 2 to 8. The results are shown in Table 1 based on Textures ($K = 16$). We can see that increasing the number of seed images improves the adaptation performance because it is less prone to overfitting to limited retrieved samples. Similar trends also hold for other datasets in the test suite.

Seed #	Method			
	ZOCLIP	RET	Ensemble	Ensemble (F)
2	42.79	38.48	51.77	57.98
4	42.79	44.09	52.96	58.57
8	42.79	45.86	55.32	62.94

Table 1. The impact of the number of seed images (per class) for I2I retrieval. Results are based on RN50 backbone with $K = 16$.

Adaptation with a mixture of ID and retrieved samples. Previously, we have considered only using retrieved samples in the feature cache to better understand the effects of retrieval. When we have access to the few-shot (ID) training set, another practical scenario is to use a mixture of retrieved and ID samples. The results are shown in Figure 6. We report the average performance (over 7 datasets) for I2I retrieval ($K = 16$). EN denotes logit ensemble with only retrieved samples. MIX denotes logit ensemble with a mixture of ID samples and retrieved samples. EN (F) and MIX (F) stand for the finetuned variants. The mixture ratio is 1:1. We observe that mixing ID and retrieved samples further leads to improved performance compared to only using few-shot ID samples. Our observations are consistent with prior works (Udandarao et al., 2023; Zhang et al., 2023) under different logit ensemble schemes, which highlight the potential of retrieval-augmented few-shot adaptation.

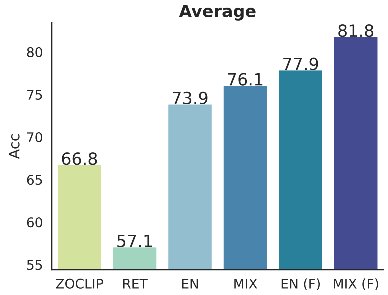


Figure 6. Impact of Mixture of retrieved samples with few-shot ID data. We report the average performance (over all datasets) for I2I retrieval ($K = 16$). EN denotes logit ensemble with only retrieved samples. MIX denotes logit ensemble with a mixture of ID samples and retrieved samples. The mixture ratio is 1:1.

Adaptation with finetuned feature cache. For completeness, we discuss learning-based adaptation by setting the visual features in the cache $\mathbf{K}$ as learnable parameters after initializing from the pre-trained CLIP model. We denote the variant as `Ensemble(F)`, where F stands for fine-tuning. We follow the hyperparameter tuning scheme in Zhang et al. (2022a) and show the results (averaged across all datasets) in Figure 7. We can see that a similar trend holds for training-based adaptation, where I2I retrieval significantly outperforms zero-shot CLIP and T2I retrieval. In the low-shot setting ($K = 1$ or 2), the performance is close to the ideal case (ID retrieval). Full results for individual datasets can be seen in Appendix E.

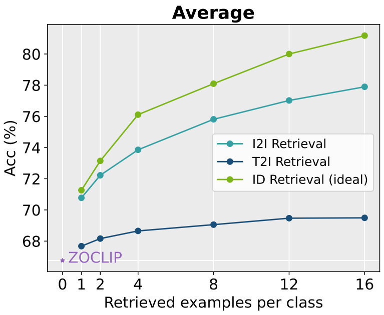


Figure 7. Adaptation with finetuned feature cache. We observe a similar trend as training-free adaptation.

Due to space constraints, we provide additional ablation studies in the Appendix.

6. Related Works

Few-shot task adaptation for vision-language models. Recent years have witnessed the popularity of contrastive language-image pre-training (CLIP) (Radford et al., 2021; Jia et al., 2021; Yang et al., 2022; Li et al., 2022b; Mu et al., 2022; Yu et al., 2022; Zhai et al., 2022; Sun et al., 2023; Zhai et al., 2023; Xu et al., 2024), etc. While CLIP-like models learn aligned multi-modal features, they often struggle on fine-trained datasets with categories not adequately represented during pre-training, which makes adaptation necessary. Recent works propose various promising solutions for adapting the vision-language model in the low-data (few-shot) scheme such as tuning textual prompts (Zhou et al., 2022a;b), visual prompts (Bahng et al., 2022; Chen et al., 2023a), multi-modal prompts (Khattak et al., 2023). Zhang et al. (2022b) use neural architecture search to optimize prompt modules. Lu et al. (2022) optimize prompts by learning prompt distributions. Alternatively, Yu et al. (2023) tune an additional task residual layer. Another line of work utilizes adaptor (Zhang et al., 2022a; Gao et al., 2023; Zhang et al., 2023; Udandarao et al., 2023) by maintaining a memory cache that stores the features of few-shot data. Zhang et al. (2022a) uses an additive logit ensemble with a feature cache from the target training set. In contrast, we focus on the impact of retrieval and build the cache with retrieved samples, rather than the downstream dataset.

Knowledge-augmented adaptation for CLIP. A natural idea for task adaptation is to utilize external knowledge sources by retrieval or synthesis. Sampling from external datasets has shown promising performance in adapting vision models to fine-grained datasets (Liu et al., 2022; Kim et al., 2023). For CLIP-based adaptation, existing methods can be categorized into two regimes, based on the amount of external data utilized. In the high-data regime, Liu et al. (2023) demonstrates promising zero-shot performance by first constructing a large-scale dataset (10M) containing relevant samples retrieved from web-scale databases and then fine-tuning CLIP models on the retrieved dataset. Xie et al. (2023) propose a Retrieval Augmented Module to augment CLIP pre-training on 1.6M retrieved samples. Recently, Iscen et al. (2023) advocated uni-modal search but cross-modal fusion for CLIP adaptation, where the fusion model is trained on 10M samples. Long et al. (2022) demonstrate the promise of retrieval for long-tail visual recognition tasks. In the low-data regime, recent works also enhance the retrieval augmentation pipeline with synthetic samples from pre-trained generative models (Udandarao et al., 2023; Zhang et al., 2023). Beyond augmenting the visual modality, Shen et al. (2022) leverage external text knowledge sources such as WordNet (Fellbaum, 1998) and Wiktionary (Meyer & Gurevych, 2012) to augment captions with class-specific descriptions, while Pratt et al. (2023) perform augmentation by querying large language models. El Banani et al. (2023) use the language guidance to find similar visual nearest neighbors. Li et al. (2022a) establish a benchmark for evaluating the transfer learning performance of language-augmented visual models. In this work, we adopt a reflective perspective and provide a systematic study to understand retrieval-augmented adaptation in the low-data regime and establish new theoretical insights.

Theoretical understanding of multi-modal learning. A few works provide theoretical explanations for multi-modal learning (Zadeh et al., 2020; Huang et al., 2021; Fürst et al., 2022; Chen et al., 2023b). For CLIP models, Liang et al. (2022) demonstrate and provide a systematic analysis of the modality gap between the features of two modalities. Nakada et al. (2023) establish the connection between CLIP and singular value decomposition (SVD) under linear representations. Chen et al. (2023b) develop a theoretical framework to understand the zero-shot transfer mechanism of CLIP. Different from prior works, we focus on the theoretical understanding of retrieval-augmented task adaptation.

7. Conclusion

In this work, we present a timely and systematic investigation for retrieval-augmented adaptation of vision-language models in the low-data regime. Our work offers a finer-grained empirical study, unveiling insights into the impact of cross-modal and uni-modal retrieval. In addition, we highlight the necessity of logit ensemble. We also develop a novel theoretical framework that supports our empirical findings and provides a deeper understanding of retrieval-augmented adaptation. Additionally, our comprehensive ablation study explores various design choices in the retrieval augmentation pipeline. We hope our work will serve as a springboard for future research on algorithm design and theoretical understanding for effective adaptation of vision-language models.

Impact Statement

The main purpose of this work is to provide a systematic investigation of existing approaches with theoretical understanding. The work can help guide the development of effective and reliable algorithms for retrieval-augmented adaptation of vision-language models. We conducted a thorough manual review to ensure that the retrieved samples do not contain illegal or inappropriate content, and we foresee no immediate negative ethical impact.

Acknowledgement

We would like to thank Zhenmei Shi for in-depth discussions on theoretical analysis. We gratefully acknowledge ICML anonymous reviewers for their helpful feedback and the support from the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation (NSF) Award No. IIS-2237037 & IIS-2331669, Office of Naval Research under grant number N00014-23-1-2643, Philanthropic Fund from SFF, and faculty research awards/gifts from Google and Meta.

References

Bahng, H., Jahanian, A., Sankaranarayanan, S., and Isola, P. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 3: 11–12, 2022.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *The European Conference on Computer Vision (ECCV)*, pp. 446–461, 2014.

Chen, A., Yao, Y., Chen, P.-Y., Zhang, Y., and Liu, S. Understanding and improving visual prompting: A label-mapping perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19133–19143, 2023a.

Chen, Z., Deng, Y., Li, Y., and Gu, Q. Understanding transferable representation learning and zero-shot transfer in clip. *arXiv preprint arXiv:2310.00927*, 2023b.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

El Banani, M., Desai, K., and Johnson, J. Learning visual representations via language-guided sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19208–19220, 2023.

Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop (CVPR-W)*, 2004.

Fellbaum, C. *WordNet: An electronic lexical database*. MIT press, 1998.

Fürst, A., Rumetshofer, E., Lehner, J., Tran, V., Tang, F., Ramsauer, H., Kreil, D., Kopp, M., Klambauer, G., Bitto-Nemling, A., and Hochreiter, S. Cloob: Modern hopfield networks with infoloob outperform clip. In *Advances in neural information processing systems (NeurIPS)*, pp. 20450–20468, 2022.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, pp. 1–15, 2023.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems (NeurIPS)*, 34:10944–10956, 2021.

Iscen, A., Caron, M., Fathi, A., and Schmid, C. Retrieval-enhanced contrastive vision-text models. *arXiv preprint arXiv:2306.07196*, 2023.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023.

Kim, S., Bae, S., and Yun, S.-Y. Coreset sampling from open-set for fine-grained self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7537–7547, 2023.

Kohler, J. M. and Lucchi, A. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning (ICML)*, pp. 1895–1904, 2017.

Li, C., Liu, H., Li, L., Zhang, P., Aneja, J., Yang, J., Jin, P., Hu, H., Liu, Z., Lee, Y. J., et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:9287–9301, 2022a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning (ICML)*, pp. 19730–19742, 2023.

Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations (ICLR)*, 2022b.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multimodal contrastive representation learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 17612–17625, 2022.

Liu, H., Son, K., Yang, J., Liu, C., Gao, J., Lee, Y. J., and Li, C. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15148–15158, 2023.

Liu, Z., Xu, Y., Xu, Y., Qian, Q., Li, H., Ji, X., Chan, A. B., and Jin, R. Improved fine-tuning by better leveraging pre-training data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Long, A., Yin, W., Ajanthan, T., Nguyen, V., Purkait, P., Garg, R., Blair, A., Shen, C., and van den Hengel, A. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6959–6969, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Lu, Y., Liu, J., Zhang, Y., Liu, Y., and Tian, X. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5206–5215, 2022.

Meyer, C. M. and Gurevych, I. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*. na, 2012.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision (ECCV)*, pp. 529–544, 2022.

Nakada, R., Gulluk, H. I., Deng, Z., Ji, W., Zou, J., and Zhang, L. Understanding multimodal contrastive learning and incorporating unpaired data. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4348–4380, 2023.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729, 2008.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

Pratt, S., Covert, I., Liu, R., and Farhadi, A. 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 (ICCV)*, pp. 15691–15701, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2556–2565, 2018.

Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Gan, Z., Wang, L., Yuan, L., Liu, C., et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 15558–15573, 2022.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Udandarao, V., Gupta, A., and Albanie, S. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2725–2736, 2023.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. Caltech-ucsd birds-200-2011. Technical report, 2011.

Xie, C.-W., Sun, S., Xiong, X., Zheng, Y., Zhao, D., and Zhou, J. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19265–19274, 2023.

Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Demystifying clip data. In *International Conference on Learning Representations (ICLR)*, 2024.

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., and Gao, J. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19163–19173, 2022.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Yu, T., Lu, Z., Jin, X., Chen, Z., and Wang, X. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10899–10909, 2023.

Zadeh, A., Liang, P. P., and Morency, L.-P. Foundations of multimodal co-learning. *Information Fusion*, 64:188–193, 2020.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18123–18133, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision (ECCV)*, pp. 493–510, 2022a.

Zhang, R., Hu, X., Li, B., Huang, S., Deng, H., Qiao, Y., Gao, P., and Li, H. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15211–15222, 2023.

Zhang, Y., Zhou, K., and Liu, Z. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022b.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022b.

Appendix

A. Experimental Details

Hardware and software. We run all experiments on NVIDIA GeForce RTX-A6000 GPU. To retrieve samples from the LAION5B database, we build a semantics-based retrieval system with `clip-retrieval` (`https://github.com/rom1504/clip-retrieval`) for fast T2I and I2I retrieval based on textual and visual embeddings of pre-trained CLIP. Our implementation is based on PyTorch 1.12.

Retrieval dataset. We adopt LAION5B as the database for retrieval for three main reasons: (1) Scale: in contrast to prior works that use smaller-scale datasets such as WebVision (Li et al., 2017), Conceptual Captions (Sharma et al., 2018), and ImageNet-21k (Ridnik et al., 2021), LAION5B is a web-scale open-source dataset that contains 5,85 billion CLIP-filtered image-text pairs covering a wide range of concepts in the real world. The diverse concept coverage makes it a reliable source for retrieval (Udandarao et al., 2023). (2) Multi-modal retrieval: one major advantage of LAION is that it computes the textual and visual embeddings of the text-image pairs based on pre-trained CLIP. This provides the foundation for us to conduct a systematic study on both T2I and I2I retrieval. (3) Retrieval efficiency: the development of distributed inference tools such as `clip-retrieval` enable fast index building and efficient retrieval from LAION5B based on approximate KNN search. Such community support for LAION5B makes retrieval more practical compared to alternatives.

Prompts for T2I retrieval. In this work, we use dataset-specific prompts in T2I retrieval to mitigate semantic ambiguity. For example, for Bird200 (Wah et al., 2011), the prompt for T2I retrieval is `A photo of a <CLS>, a type of bird`. The prompts for other datasets can be seen in Table 3. In a recent work (Udandarao et al., 2023), language model-based prompts are used for retrieval. For instance, the prompt for the class `baklava` becomes `baklava is a rich, sweet pastry made with layers of filo dough, nuts, and syrup`. We found that using knowledge-augmented prompts improved the performance of T2I retrieval. However, the performance gain from these prompts was consistently less significant than that observed with I2I retrieval. For example, with an 8-shot setting, the performance (averaged over all datasets) with knowledge-augmented prompts is summarized in Table 2:

Method	ZOC	T2I (original)	T2I (knowledge-augmented)	I2I
AVG ACC	66.76	67.77	68.86	72.91

Table 2. The impact of knowledge-augmented prompts for T2I retrieval.

The results are based on the default setting with RN50 as the vision backbone. The same two key observations hold under alternative prompt strategies: (1) I2I retrieval outperforms T2I retrieval; (2) logit ensemble is essential for superior retrieval performance. By examining the retrieved samples, we identified issues similar to those depicted in Figure 3 when using knowledge-augmented prompts, particularly when the target class contains characteristics not captured by the class names and their descriptions.

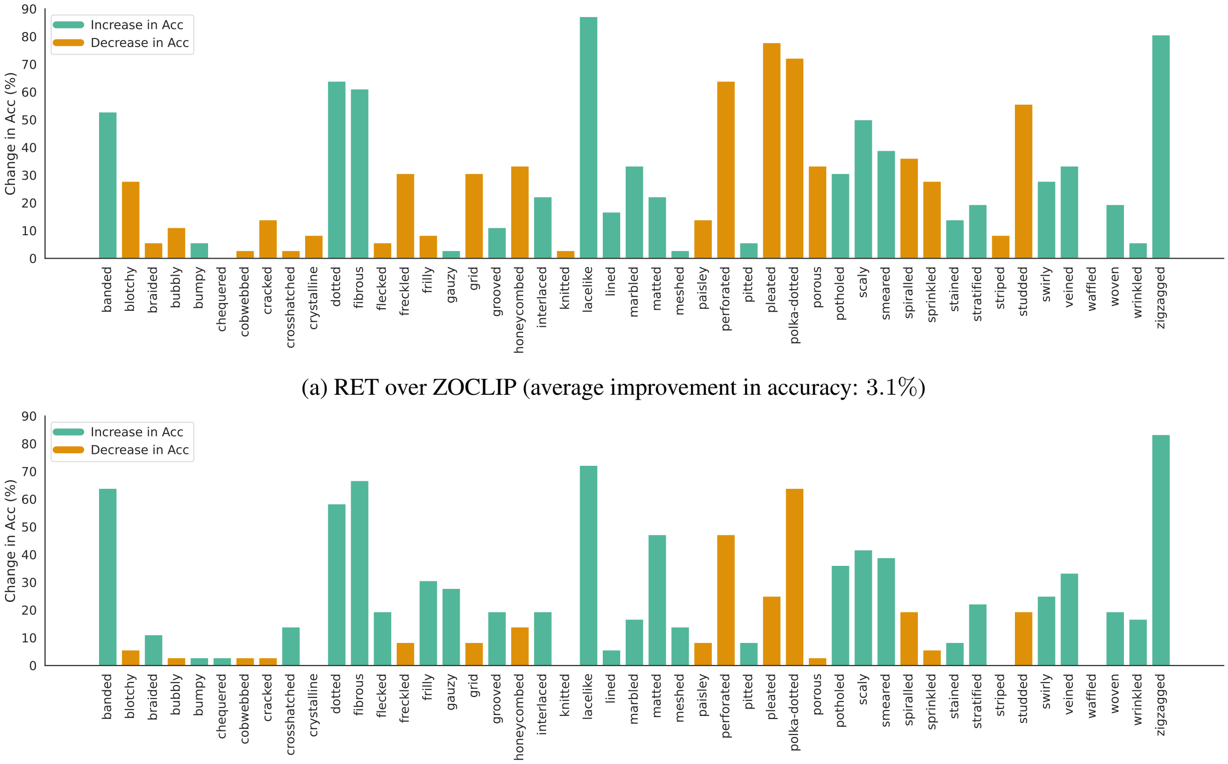
Dataset	Prompt
Caltech101 (Fei-Fei et al., 2004)	`A photo of a <CLS>`
Birds200 (Wah et al., 2011)	`A photo of a <CLS>, a type of bird`
Food101 (Bossard et al., 2014)	`A photo of <CLS>, a type of food`
OxfordPets (Parkhi et al., 2012)	`A photo of a <CLS> pet`
Flowers102 (Nilsback & Zisserman, 2008)	`A photo of a <CLS> flower`
Textures (Cimpoi et al., 2014)	`A photo of <CLS> texture`
UCF101 (Soomro et al., 2012)	`A photo of <CLS> in action`

Table 3. Default prompts for T2I retrieval. In this work, we use dataset-specific prompts to mitigate semantic ambiguity.

Fine-tuning details. As our work focuses on the impact of retrieval, we adopt the fine-tuning scheme in Zhang et al. (2022a) for training-based adaptation, where we set features in the retrieval cache as learnable. For each target dataset, the train, validation, and test split also follow (Zhang et al., 2022a). Specifically, we use AdamW (Loshchilov & Hutter, 2019) as the optimizer with a cosine scheduler. The initial learning rate is set as 0.001 and we finetune for 20 epochs. The hyperparameters such as α, ω, γ are determined based on the validation split of each target dataset.

B. A Closer Look at Logit Ensemble via Classwise Performance

In Section 3.3, we have shown that logit ensemble is essential to CLIP-based adaptive inference with the few-shot cache obtained by retrieval. In this section, we take a finer-grained view by examining the change of accuracy for each class before and after logit ensemble. For better visualization, we use Textures (Cimpoi et al., 2014), a dataset with 47 classes. The results are shown in Figure 8, where green indicates an increase in accuracy while orange denotes a decrease in accuracy. The result for RET vs. ZOCLIP (*i.e.*, before ensemble) is shown in Figure 8a and Ensemble vs. ZOCLIP is shown in Figure 8b. We can clearly observe that (1) before ensemble, RET is inferior to ZOCLIP for multiple classes such as `blotchy` and `freckled`, and `pleated`, as a result of retrieval ambiguity. (2) Logit ensemble significantly mitigates such issue and results in an overall larger proportion of green bars compared to orange bars, as shown in Figure 8b.



(a) RET over ZOCLIP (average improvement in accuracy: 3.1%)

(b) Ensemble over ZOCLIP (average improvement in accuracy: 12.5%)

Figure 8. Change of classwise accuracy before and after logit ensemble. For better visualization, the results are based on Textures (Cimpoi et al., 2014), a dataset with 47 classes. We use I2I retrieval to obtain the few-shot feature cache. We plot the change of accuracy over ZOCLIP for each class before (top row) and after logit ensemble (bottom row). Blue bars indicate an increase in accuracy while orange denotes a decrease in accuracy. (a) Comparison of RET versus ZOCLIP. On average, RET achieves a 3.1% improvement in accuracy compared to ZOCLIP. (b) Comparison of Ensemble versus ZOCLIP. On average, Ensemble achieves a 12.5% improvement in accuracy compared to ZOCLIP. This further highlights the importance of logit ensemble for retrieval-augmented adaptation.

C. Qualitative Analysis of Retrieved Samples

In Section 3.2, we examined the retrieved samples from I2I and T2I retrieval to identify the main sources of errors. Here, we present additional retrieved samples for diverse datasets. The results are depicted in Figure 9, where we contrast samples from T2I retrieval (top row), I2I retrieval (middle row), and the downstream dataset (bottom row). We have two salient observations: (1) As discussed in Section 3.2, T2I retrieval often yields a diverse set of images that match the class semantics. However, this diversity may not always be beneficial for adapting to the target dataset, especially in the few-shot retrieval setting where one is under a limited budget. For example, using the query `a photo of a lobster`, we may not retrieve images of cooked lobsters that often appear in the target dataset. (2) Since T2I retrieval utilizes the class name in the query, it occasionally retrieves images with text on them, rather than images of the actual object. For instance, we retrieve images that feature the text "summer tanager" or "dandelion" (as seen in the 4th and 3rd columns of Figures 9 and 3, respectively). This occurs because the cosine similarity between pairs of `(class name, image of the actual object)` and `(class name, image with the text <class name>)` is similar, based on pre-trained CLIP models. This highlights a prevalent challenge in web-scale cross-modal retrieval systems, such as LAION5B. Conversely, this type of misalignment is rarely encountered in I2I retrieval. Therefore, samples from T2I retrieval can introduce undesirable inductive biases, resulting in limited performance gains over the zero-shot model.

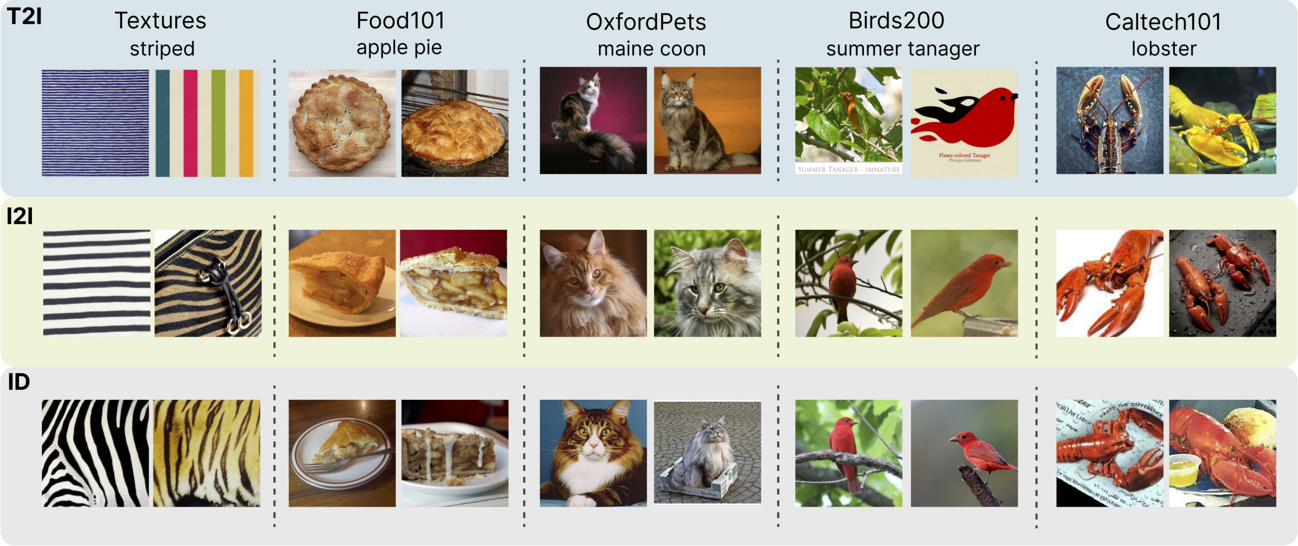


Figure 9. More samples from T2I and I2I retrieval. Top row: the main source of noise for T2I retrieval is semantic ambiguity, as the textual queries (*e.g.*, `striped texture`) may not accurately describe the images from target distributions (bottom row). Middle row: samples retrieved by I2I matches more closely with ID data. Bottom row: images sampled from the target (ID) distribution.

D. Theoretical Understanding

In this section, we provide details on the problem setup, introduce relevant definitions and lemmas, and provide the complete proof for our theoretical results discussed in Section 4. Common notations can be seen in Table 4.

Notation	Description
$[C]$	The set $\{1, 2, \dots, C\}$
$\mathbb{1}$[condition]	Indicator function, equals 1 if the condition is true, 0 otherwise
$\mathcal{T}$	$\mathcal{T} : t \rightarrow \mathbb{R}^d$ is the text encoder of CLIP
$\mathcal{I}$	$\mathcal{I} : \mathbf{x} \rightarrow \mathbb{R}^d$ is the image encoder of CLIP

Table 4. Common notations.

D.1. Problem Setup

We consider a pre-trained CLIP model (Radford et al., 2021) with one text encoder $\mathcal{T} : t \rightarrow \mathbb{R}^d$ and one image encoder $\mathcal{I} : \mathbf{x} \rightarrow \mathbb{R}^d$. We use $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_C] \in \mathbb{R}^{d \times C}$ to denote the text embedding matrix for all classes, where $\mathbf{t}_c := \mathcal{T}(\mathbf{t}_c) \in \mathbb{R}^d$ and t_c is a generic textual description of class c such as "`a photo of <CLASS c>`". For theoretical analysis, we consider training-free adaptation based on retrieved samples. We use the terms "downstream" and "target" dataset interchangeably which refer to the dataset a pre-trained CLIP model is adapted to.

Building feature cache by retrieval. Given a downstream dataset with C classes: $\mathcal{Y} = \{1, 2, ..., C\}$ and a retrieval budget size of KC, we can retrieve K samples per class to build a cache of size KC. Recall that $\mathbf{K} = [\mathbf{k}_{1,1}, \mathbf{k}_{1,2} \dots, \mathbf{k}_{C,K}] \in \mathbb{R}^{d \times CK}$ denotes the embedding matrix for retrieved images, where $\mathbf{k}_{c,i} := \mathcal{I}(\mathbf{x}_{c,i}) \in \mathbb{R}^d$. For notational simplicity, we assume text and image features are ℓ_2 normalized (Radford et al., 2021). In other words, we have $\|\mathbf{z}\|_2 = \|\mathbf{t}_c\|_2 = 1$ for any $\mathbf{z} = \mathcal{I}(\mathbf{x})$ and $\mathbf{t}_c = \mathcal{T}(t_c)$.

Let $\tilde{\mathbf{K}} = \frac{\mathbf{K}\mathbf{V}^\top}{K} = [\tilde{\mathbf{k}}_1, \tilde{\mathbf{k}}_2, \dots, \tilde{\mathbf{k}}_C] \in \mathbb{R}^{d \times C}$ contain the average retrieved feature for each class. $\mathbf{V} \in \mathbb{R}^{C \times CK}$ is a sparse matrix containing the one-hot labels for retrieved samples with entries $\mathbf{V}_{i,j} = \mathbb{1}\{i = \tilde{j}\}$ for $i \in [C], j \in [CK]$, where $\tilde{j} := \lceil \frac{j}{K} \rceil$ (Zhang et al., 2022a). For example, when $K = 2, C = 3$, we have:

$$\mathbf{V} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

We define $\bar{\mathbf{K}} := [\bar{\mathbf{k}}_1, \bar{\mathbf{k}}_2, \dots, \bar{\mathbf{k}}_C]$ as the normalized version where $\bar{\mathbf{k}}_i = \frac{\tilde{\mathbf{k}}_i}{\|\tilde{\mathbf{k}}_i\|_2}$, which will be used in the rest of the section. Note that here the notations are slightly different from Section 4.1 and are more rigorous.

Task adaptation with retrieved cache. At inference time, let $(\mathbf{x}, y) \sim \mathcal{D}_T$ be a test sample from the target distribution $\mathcal{D}_T$ with label $y \in [C]$ and its visual feature $\mathbf{z} := \mathcal{I}(\mathbf{x})$. In some cases, beyond retrieved samples, one also has access to a cache consisting of few-shot training samples from the target distribution. For theoretical analysis, we consider one-shot and denote the feature cache as $\mathbf{S} := [\mathbf{s}_1, \dots, \mathbf{s}_C] \in \mathbb{R}^{d \times C}$. The final logit for the test sample can be represented as a weighted sum (ensemble) of logits from the zero-shot CLIP and the feature cache from retrieved and training samples[2]:

$$f(\mathbf{x}) = (\alpha\mathbf{T} + \beta\mathbf{S} + \gamma\bar{\mathbf{K}})^\top \mathbf{z},$$

where $0 \leq \alpha, \beta, \gamma \leq 1$. Without loss of generality, we assume $\alpha + \beta + \gamma = 1$.

In particular, the zero-shot logit $f^{\text{ZOC}}(\mathbf{x}) := \mathbf{T}^\top \mathbf{z}$ and the retrieval logit $f^{\text{RET}}(\mathbf{x}) := \bar{\mathbf{K}}^\top \mathbf{z}$. In the main paper, we mainly focus on $\beta = 0$ (*i.e.*, one only has access to retrieved samples). We denote the corresponding ensemble logit as $f^{\text{EN}}(\mathbf{x}) = (\alpha\mathbf{T} + \gamma\bar{\mathbf{K}})^\top \mathbf{z}$.

[2] For theoretical analysis, we omit the exponential scaling function to better focus on the effects of ensembling.

Evaluation metric. Given a loss function $\ell(\mathbf{v}, y)$ such as the cross-entropy:

$$\ell(\mathbf{v}, y) = -\log \frac{\exp(\mathbf{v}_y)}{\sum_{i \in [C]} \exp(\mathbf{v}_i)},$$

the population risk on the target distribution is:

$$\mathcal{L}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T}[\ell(f(\mathbf{x}), y)].$$

To simplify notations, we denote the risk as $\mathcal{R}(\mathbf{Q}) := \mathbb{E}\left[\ell(\mathbf{Q}^\top \mathbf{z}, y)\right]$ for some $\mathbf{Q} \in \mathbb{R}^{d\times C}$. For example, the risk of logit ensemble is $\mathcal{R}(\alpha\mathbf{T} + \gamma\bar{\mathbf{K}})$. We also have the error risk $\mathcal{R}_{0-1}$ defined as:

$$\mathcal{R}_{0-1}(f) = 1 - \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T}\left[\mathbb{1}\{\operatorname*{argmax}_{i\in[C]} f(\mathbf{x})_i = y\}\right].$$

D.2. Definitions and Assumptions

Before presenting the main theoretical results, we first introduce the following definitions and assumptions to formalize the retrieval augmented adaptation process based on pre-trained CLIP models.

For class $i \in [C]$, we define $\tilde{\mathbf{s}}_i := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_T}[\mathcal{I}(\mathbf{x})|y = i]$, which is the image representation of class i based on the downstream distribution and $\bar{\mathbf{s}}_i = \frac{\tilde{\mathbf{s}}_i}{\|\tilde{\mathbf{s}}_i\|_2}$ the ℓ_2 normalized version[3]. Let $\bar{\mathbf{S}} := [\bar{\mathbf{s}}_1, \bar{\mathbf{s}}_2, \dots, \bar{\mathbf{s}}_C]$.

Definition D.1 (Inner-class concentration and inter-class separation)**.** We define the inter-class feature separation as $\nu := 1 - \max_{i\neq j} \bar{\mathbf{s}}_i^\top \bar{\mathbf{s}}_j$. We use ρ_c to denote the inner-class feature concentration:

$$\rho_c := \max_{i\in[C]} \Pr\left(\|\mathcal{I}(\mathbf{x}) - \bar{\mathbf{s}}_i\|_2 \geq \kappa | y = i\right)$$

for some positive constant κ.

Definition D.2. Let $\bar{\mathbf{Z}} = [\bar{\mathbf{z}}_1, \dots, \bar{\mathbf{z}}_C] \in \mathbb{R}^{d\times C}$. We define the optimal representations as

$$\bar{\mathbf{Z}}^* = \operatorname*{argmin}_{\bar{\mathbf{Z}}\in\mathbb{R}^{d\times C};\forall i\in[C],\|\bar{\mathbf{z}}_i\|=1} \mathbb{E}[\ell(\bar{\mathbf{Z}}^{*\top}\mathbf{z}, y)].$$

Definition D.3 (Modality gap)**.** We define the modality gap between the pre-trained text distribution and the target distribution (in the visual modality) as $\tau := \max_{i\neq j}(\mathbf{t}_j - \mathbf{t}_i)^\top \bar{\mathbf{s}}_i$, where $i, j \in [C]$.

Definition D.4 (Retrieval distribution shift)**.** We denote the retrieval distribution based on the (text or image) query (denote $\mathbf{t}_c$ or $\mathbf{s}_c$ as $\mathbf{q}_c$) from class c as $\mathcal{D}_{R|\mathbf{q}_c}$. $\tilde{\mathbf{k}}_{\mathbf{q}_c} := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{R|\mathbf{q}_c}}[\mathcal{I}(\mathbf{x})]$ is the average retrieved feature from class c. $\bar{\mathbf{k}}_{\mathbf{q}_c} := \frac{\tilde{\mathbf{k}}_{\mathbf{q}_c}}{\|\tilde{\mathbf{k}}_{\mathbf{q}_c}\|_2}$ denotes the normalized version. We define the distributional shift between target data and T2I and I2I retrieval data for class c as $\xi_c^{\text{T2I}} := 1 - \bar{\mathbf{k}}_{\mathbf{t}_c}^\top \bar{\mathbf{s}}_c$ and $\xi_c^{\text{I2I}} := 1 - \bar{\mathbf{k}}_{\mathbf{s}_c}^\top \bar{\mathbf{s}}_c$. Let, $\xi_{\mathbf{t}} := \max_{c\in[C]} \xi_c^{\text{T2I}}$ and $\xi_{\mathbf{s}} := \max_{c\in[C]} \xi_c^{\text{I2I}}$.

Remarks: Note that $\tilde{\mathbf{k}}_{\mathbf{q}_c}$ is the expected version, while $\tilde{\mathbf{k}}_c$ (defined in Appendix D.1) is the empirical mean of retrieved samples for class $c \in [C]$.

At inference time, for a test sample $(\mathbf{x}, y) \sim \mathcal{D}_T$ with image feature $\mathbf{z} = \mathcal{I}(\mathbf{x})$, one of the following four events can happen:

$$\begin{aligned}
E_1 :=& \{(\mathbf{x}, y) \sim \mathcal{D}_T : y \neq \operatorname*{argmax}_{i\in[C]} \mathbf{t}_i^\top \mathbf{z} \text{ and } y \neq \operatorname*{argmax}_{i\in[C]} \bar{\mathbf{k}}_i^\top \mathbf{z}\} \\
E_2 :=& \{(\mathbf{x}, y) \sim \mathcal{D}_T : y = \operatorname*{argmax}_{i\in[C]} \mathbf{t}_i^\top \mathbf{z} \text{ and } y \neq \operatorname*{argmax}_{i\in[C]} \bar{\mathbf{k}}_i^\top \mathbf{z}\} \\
E_3 :=& \{(\mathbf{x}, y) \sim \mathcal{D}_T : y \neq \operatorname*{argmax}_{i\in[C]} \mathbf{t}_i^\top \mathbf{z} \text{ and } y = \operatorname*{argmax}_{i\in[C]} \bar{\mathbf{k}}_i^\top \mathbf{z}\} \\
E_4 :=& \{(\mathbf{x}, y) \sim \mathcal{D}_T : y = \operatorname*{argmax}_{i\in[C]} \mathbf{t}_i^\top \mathbf{z} \text{ and } y = \operatorname*{argmax}_{i\in[C]} \bar{\mathbf{k}}_i^\top \mathbf{z}\}.
\end{aligned}$$

[3] For any two non-zero vectors $\mathbf{v}_1, \mathbf{v}_2$ with unit norms, we have $\|\mathbf{v}_1 - \mathbf{v}_2\|_2 = \sqrt{2 - 2\mathbf{v}_1^\top \mathbf{v}_2}$.

We can see that $\mathcal{R}_{0-1}(f^{\text{ZOC}}) = \Pr(E_1) + \Pr(E_3)$ and $\mathcal{R}_{0-1}(f^{\text{RET}}) = \Pr(E_1) + \Pr(E_2)$. Next, we formalize the intuitions in Figure 3 as the following definition:

Definition D.5 (Knowledge encoded in different modalities)**.** For a vector $\mathbf{v} \in \mathbb{R}^C$ and a scalar $i \in [C]$, We define $\phi(\mathbf{v}, i, z) := \{j | \mathbf{v}_i - \mathbf{v}_j \leq z\}$. Consider $(\mathbf{x}, y) \sim \mathcal{D}_T$ and $\mathbf{z} = \mathcal{I}(\mathbf{x})$. We define the conditional probability $\rho_d(z)$ as

$$\rho_d(z) = \Pr\left(\phi(\mathbf{T}^\top \mathbf{z}, y, z) \cap \phi(\bar{\mathbf{K}}^\top \mathbf{z}, y, z) \neq \{y\} \middle| E_2 \text{ or } E_3\right).$$

Remarks: $\phi(\mathbf{v}, i, z)$ identifies elements in vector $\mathbf{v}$ that are within a threshold z of the i-th element of $\mathbf{v}$. $\rho_d(z)$ represents the likelihood that, given events E_2 or E_3, the transformed data $\mathbf{z}$ is associated with an incorrect class by both $\mathbf{T}$ and $\bar{\mathbf{K}}$. In practical scenarios, $\rho_d(z)$ is typically small. This is because different modalities usually represent knowledge in distinct ways and, as a result, have different patterns of confusion or error.

Assumption D.6 (Sample representativeness)**.** We assume that the sample for each class is relatively representative, i.e., $\forall i \in [C], \|\mathbf{s}_i - \bar{\mathbf{s}}_i\|_2 \leq \kappa$ for some constant κ.

Assumption D.7 (Retrieved data distribution)**.** We assume that for each class the distribution of retrieved samples is composed of clusters, which exhibit ν separation and κ concentration as defined in Definition D.1. We assume that the retrieval process for a query sample is uniformly sampling from its closest retrieval cluster.

D.3. Main Results and Analysis

Lemma D.8. *We can upper bound the risk $\mathcal{R}(\bar{\mathbf{S}})$ as follows:*

$$\mathcal{R}(\bar{\mathbf{S}}) \leq (1 - \rho_c) \log\left(1 + (C-1)\exp\left(2\kappa - \nu\right)\right) + \rho_c \log\left(1 + (C-1)\exp\left(2\right)\right) \tag{1}$$

where ρ_c, κ, ν defined in Definition D.1 characterize the inner-class concentration and inter-class separation.

Proof. For a test sample $(\mathbf{x}, y) \sim \mathcal{D}_T$ with $\mathbf{z} = \mathcal{I}(\mathbf{x})$. Let $\mathbf{z} = \mathbf{v} + \bar{\mathbf{s}}_y$. By Definition D.1, we have $\Pr\left(\|\mathbf{v}\|_2 \geq \kappa\right) \leq \rho_c$. Thus, we have

$$\begin{aligned}
\mathcal{R}(\bar{\mathbf{S}}) =& \mathbb{E}\left[\ell(\bar{\mathbf{S}}^\top \mathbf{z}, y)\right] && (2)\\
=& \mathbb{E}\left[-\log \frac{\exp\left(\bar{\mathbf{s}}_y^\top \mathbf{z}\right)}{\sum_{i \in [C]} \exp\left(\bar{\mathbf{s}}_i^\top \mathbf{z}\right)}\right] && (3)\\
=& \mathbb{E}\left[\log\left(1 + \sum_{i \neq y} \exp\left(\bar{\mathbf{s}}_i^\top \mathbf{z} - \bar{\mathbf{s}}_y^\top \mathbf{z}\right)\right)\right] && (4)\\
=& \mathbb{E}\left[\log\left(1 + \sum_{i \neq y} \exp\left(\bar{\mathbf{s}}_i^\top \left(\mathbf{v} + \bar{\mathbf{s}}_y\right) - \bar{\mathbf{s}}_y^\top \left(\mathbf{v} + \bar{\mathbf{s}}_y\right)\right)\right)\right] && (5)\\
\leq& (1 - \rho_c)\mathbb{E}\left[\log\left(1 + \sum_{i \neq y} \exp\left(\bar{\mathbf{s}}_i^\top \mathbf{v} + 1 - \nu - \bar{\mathbf{s}}_y^\top \mathbf{v} - 1\right)\right) \middle| \|\mathbf{v}\|_2 \leq \kappa\right] && (6)\\
&+ \rho_c \mathbb{E}\left[\log\left(1 + \sum_{i \neq y} \exp\left(2\right)\right) \middle| \|\mathbf{v}\|_2 \geq \kappa\right] && (7)\\
\leq& (1 - \rho_c)\mathbb{E}\left[\log\left(1 + \sum_{i \neq y} \exp\left(2\|\mathbf{v}\|_2 - \nu\right)\right) \middle| \|\mathbf{v}\|_2 \leq \kappa\right] + \rho_c \log\left(1 + (C-1)\exp\left(2\right)\right) && (8)\\
\leq& (1 - \rho_c) \log\left(1 + (C-1)\exp\left(2\kappa - \nu\right)\right) + \rho_c \log\left(1 + (C-1)\exp\left(2\right)\right). && (9)
\end{aligned}$$

□

Remarks: Lemma D.8 is a tight upper bound. We give a simple toy example here for illustration: consider binary classification on two data points $(\mathbf{x}_1, y_1)$ and $(\mathbf{x}_2, y_2)$. Suppose $\mathbf{z}_1 = \mathcal{I}(\mathbf{x}_1) = -\mathbf{z}_2 = -\mathcal{I}(\mathbf{x}_2)$, we can see that $\mathcal{R}(\bar{\mathbf{S}}) = \mathcal{R}(\bar{\mathbf{Z}}^*) = \log\left(1 + \exp\left(-2\right)\right)$, where $C = 2, \rho_c = \kappa = 0, \nu = 2$.

Lemma D.9. *For a test sample $(\mathbf{x}, y) \sim \mathcal{D}_T$ and its image feature $\mathbf{z} = \mathcal{I}(\mathbf{x})$, with probability at least $1 - \rho_c$, we have*

$$\max_{i \neq y} \mathbf{s}_i^\top \mathbf{z} - \mathbf{s}_y^\top \mathbf{z} \leq 4\kappa - \nu.$$

Proof of Lemma D.9. Let $\mathbf{z} = \mathbf{v} + \bar{\mathbf{s}}_y$. By Definition D.1 and Assumption D.6, we have $\Pr\left(\|\mathbf{v}\|_2 \geq \kappa\right) \leq \rho_c$. Thus, we have with probability at least $1 - \rho_c$ such that

$$\begin{aligned}
\max_{i \neq y} \mathbf{s}_i^\top \mathbf{z} - \mathbf{s}_y^\top \mathbf{z} =& \max_{i \neq y} \left(\mathbf{s}_i - \bar{\mathbf{s}}_i + \bar{\mathbf{s}}_i\right)^\top \left(\mathbf{v} + \bar{\mathbf{s}}_y\right) - \left(\mathbf{s}_y - \bar{\mathbf{s}}_y + \bar{\mathbf{s}}_y\right)^\top \left(\mathbf{v} + \bar{\mathbf{s}}_y\right) && (10)\\
=& \max_{i \neq y} \mathbf{s}_i^\top \mathbf{v} + \left(\mathbf{s}_i - \bar{\mathbf{s}}_i\right)^\top \bar{\mathbf{s}}_y + \bar{\mathbf{s}}_i^\top \bar{\mathbf{s}}_y && (11)\\
& \quad - \mathbf{s}_y^\top \mathbf{v} - \left(\mathbf{s}_y - \bar{\mathbf{s}}_y\right)^\top \bar{\mathbf{s}}_y - \bar{\mathbf{s}}_y^\top \bar{\mathbf{s}}_y && (12)\\
\leq& \max_{i \neq y} \kappa + \kappa + 1 - \nu + \kappa + \kappa - 1 && (13)\\
=& 4\kappa - \nu. && (14)
\end{aligned}$$

□

Remarks: From the above lemma, we can see that if $4\kappa < \nu$, the accuracy of $f^{\mathrm{RET}}(\cdot)$ is at least $1 - \rho_c$.

Lemma D.10 (Retrieval distribution shift bound)**.** *Under Assumption D.6 and Assumption D.7 and suppose that $\mathbf{s}_i$ is in the support of $\mathcal{D}_R$, we have $\xi_{\mathbf{s}} \leq 2\kappa^2$. Furthermore, when the retrieval cluster for $\mathbf{t}_i$ and $\mathbf{s}_i$ are different for any $i \in [C]$, we have $\xi_{\mathbf{t}} \geq \nu - 2\kappa$.*

Proof of Lemma D.10. By Assumption D.6 and Assumption D.7, for any $i \in [C]$, we have

$$\begin{aligned}
\xi_i^{\mathrm{I2I}} =& 1 - \bar{\mathbf{k}}_{\mathbf{s}_i}^\top \bar{\mathbf{s}}_i && (15)\\
=& \frac{1}{2} \left\|\bar{\mathbf{s}}_i - \bar{\mathbf{k}}_{\mathbf{s}_i}\right\|_2^2 && (16)\\
\leq& 2\kappa^2. && (17)
\end{aligned}$$

Furthermore, when the retrieval clusters for $\mathbf{t}_i$ and $\mathbf{s}_i$ are different, by Assumption D.7, we have

$$\begin{aligned}
\xi_i^{\mathrm{T2I}} =& 1 - \bar{\mathbf{k}}_{\mathbf{t}_i}^\top \bar{\mathbf{s}}_i && (18)\\
=& 1 - \left(\bar{\mathbf{k}}_{\mathbf{t}_i}\right)^\top \left(\bar{\mathbf{s}}_i - \bar{\mathbf{k}}_{\mathbf{s}_i} + \bar{\mathbf{k}}_{\mathbf{s}_i}\right) && (19)\\
=& 1 - \left(\bar{\mathbf{k}}_{\mathbf{t}_i}\right)^\top \left(\bar{\mathbf{s}}_i - \bar{\mathbf{k}}_{\mathbf{s}_i}\right) - \bar{\mathbf{k}}_{\mathbf{t}_i}^\top \bar{\mathbf{k}}_{\mathbf{s}_i} && (20)\\
\geq& \nu - \left\|\bar{\mathbf{s}}_i - \bar{\mathbf{k}}_{\mathbf{s}_i}\right\|_2 && (21)\\
=& \nu - \sqrt{2\xi_i^{\mathrm{I2I}}} && (22)\\
\geq& \nu - 2\kappa. && (23)
\end{aligned}$$

□

Theorem D.11 (Benefit of uni-modal retrieval)**.** *Assume the same condition as Lemma D.10, with probability at least $1 - \delta$, the following upper bound of the ensemble risk holds:*

$$\mathcal{R}(\alpha \mathbf{T} + \gamma \bar{\mathbf{K}}) - \mathcal{R}(\bar{\mathbf{S}}) \leq L\left(\alpha \underbrace{\|(\mathbf{T} - \bar{\mathbf{S}})^\top \mathbf{z}\|_2}_{\textit{modality gap}} + \underbrace{\gamma\kappa\sqrt{\frac{8C}{K}\log\frac{C}{\delta}}}_{\textit{retrieval sample complexity}} + \underbrace{\gamma\sqrt{2C\xi}}_{\textit{retrieval shift}}\right), \tag{24}$$

where $L = \sqrt{\exp(2) + 1}$, κ characterizes the inner-class feature concentration (Definition D.1), and ξ is either $\xi_{\mathbf{s}}$ for I2I retrieval or $\xi_{\mathbf{t}}$ for T2I retrieval.

Proof of Theorem D.11. By Lemma D.13 and Lemma D.14, let $L = \sqrt{\exp(2)+1}$, we have:

$$\mathcal{R}(\alpha\mathbf{T} + \gamma\bar{\mathbf{K}}) - \mathcal{R}(\bar{\mathbf{S}}) \le L\left(\alpha\|(\mathbf{T} - \bar{\mathbf{S}})^\top \mathbf{z}\|_2 + \gamma\|(\bar{\mathbf{K}} - \bar{\mathbf{S}})^\top \mathbf{z}\|_2\right). \tag{25}$$

By the vector Bernstein inequality in Lemma D.15 and the union bound, with probability at least $1-\delta$, for any $c \in [C]$:

$$\|\bar{\mathbf{k}}_c - \bar{\mathbf{k}}_{\mathbf{q}_c}\|_2 \le \kappa\sqrt{\frac{8}{K}\log\frac{C}{\delta}}, \tag{26}$$

This bound characterizes the retrieval sample complexity. Moreover, from the definition of the retrieval distributional shift, we have $\|\bar{\mathbf{k}}_{\mathbf{q}_c} - \bar{\mathbf{s}}_c\|_2 = \sqrt{2 - 2\bar{\mathbf{k}}_{\mathbf{q}_c}^\top \bar{\mathbf{s}}_c} = \sqrt{2\xi_c}$, where $\mathbf{q}_c = \mathbf{s}_c$ for I2I retrieval and $\mathbf{q}_c = \mathbf{t}_c$ for T2I retrieval. Therefore, we obtain an upper bound of $\|(\bar{\mathbf{K}} - \bar{\mathbf{S}})^\top \mathbf{z}\|_2$ as:

$$\|(\bar{\mathbf{K}} - \bar{\mathbf{S}})^\top \mathbf{z}\|_2 \le \kappa\sqrt{\frac{8C}{K}\log\frac{C}{\delta}} + \sqrt{2C\xi} \tag{27}$$

We obtain the final bound by putting together Eq. (25) and Eq. (27). □

Remarks: The above upper bound consists of three terms: the gap between the textual and visual modality, the sample complexity of retrieved features which decreases as we increase K, and a term related to the distributional shift induced by the retrieval method. By Lemma D.10, we can see the superiority of I2I over T2I retrieval by comparing $\xi_{\mathbf{s}}$ and $\xi_{\mathbf{t}}$.

Theorem D.12 (Benefit of logit ensemble)**.** *Assume the same condition as Lemma D.10. For I2I retrieval with $\alpha = \gamma = \frac{1}{2}, \beta = 0$, we have*

$$\mathcal{R}_{0-1}(f) \le \Pr(E_1) + (\Pr(E_2) + \Pr(E_3))\rho_d(\max\{6\kappa - \nu, 2\kappa + \tau\}) + \rho_c \tag{28}$$

Proof of Theorem D.12. We define the events $E_c = \{\|\mathcal{I}(\mathbf{x}) - \bar{\mathbf{s}}_i\|_2 \ge \kappa \text{ and } y = i, \forall i \in [C]\}$. We also define events $E_d(z) = \left\{\phi(\mathbf{T}^\top \mathbf{z}, y, z) \cap \phi(\bar{\mathbf{K}}^\top \mathbf{z}, y, z) = \{y\}\right\}$. Note that we have $\Pr(E_d(z)|E_2 \text{ or } E_3) = 1 - \rho_d(z)$. By Definition D.5, we have

$$\begin{aligned}
\max_{(\mathbf{x},y)\in E_c, y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} =& \max_{(\mathbf{x},y)\in E_c, y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top (\mathbf{z} - \bar{\mathbf{s}}_i + \bar{\mathbf{s}}_i) && (29)\\
=& \max_{(\mathbf{x},y)\in E_c, y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top (\mathbf{z} - \bar{\mathbf{s}}_i) + (\mathbf{t}_j - \mathbf{t}_i)^\top \bar{\mathbf{s}}_i && (30)\\
\le& 2\kappa + \tau. && (31)
\end{aligned}$$

By Lemma D.9 and Assumption D.6 and Assumption D.7, conditional on E_c, we have the logits gap $\max_{i\neq y} \bar{\mathbf{k}}_i^\top \mathbf{z} - \bar{\mathbf{k}}_y^\top \mathbf{z} \le 6\kappa - \nu$. Let $\mathrm{ACC}(f) = 1 - \mathcal{R}_{0-1}(f)$. Then, we get

$$\begin{aligned}
\mathrm{ACC}(f) =& \Pr\left(y = \operatorname*{argmax}_i \frac{1}{2}\mathbf{t}_i^\top \mathbf{z} + \frac{1}{2}\bar{\mathbf{k}}_i^\top \mathbf{z}\right) && (32)\\
\ge& \Pr(E_4) + \Pr(E_c \cap E_2)\Pr\left(y = \operatorname*{argmax}_i \mathbf{t}_i^\top \mathbf{z} + \bar{\mathbf{k}}_i^\top \mathbf{z} \middle| E_c \cap E_2\right) && (33)\\
&+ \Pr(E_c \cap E_3)\Pr\left(y = \operatorname*{argmax}_i \mathbf{t}_i^\top \mathbf{z} + \bar{\mathbf{k}}_i^\top \mathbf{z} \middle| E_c \cap E_3\right) && (34)\\
=& \Pr(E_4) + \Pr(E_c \cap E_2)\Pr\left(\max_{y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < 0 \middle| E_c \cap E_2\right) && (35)\\
&+ \Pr(E_c \cap E_3)\Pr\left(\max_{y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < 0 \middle| E_c \cap E_3\right). && (36)
\end{aligned}$$

Now, we prove that $E_d(6\kappa - \nu) \cap E_c \cap E_2 \subseteq \{\max_{y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < 0\} \cap E_c \cap E_2$.

For any $(\mathbf{x}, y) \in E_d(6\kappa - \nu) \cap E_c \cap E_2$ and $y = i \neq j$,

- if $j \in \phi(\mathbf{T}^\top \mathbf{z}, y, z)$ and $j \notin \phi(\bar{\mathbf{K}}^\top \mathbf{z}, y, z)$, we have $(\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < 0 - (6\kappa - \nu) \leq 0$;
- if $j \notin \phi(\mathbf{T}^\top \mathbf{z}, y, z)$ and $j \in \phi(\bar{\mathbf{K}}^\top \mathbf{z}, y, z)$, by Lemma D.9, $(\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < -(6\kappa - \nu) + 6\kappa - \nu = 0$;
- if $j \notin \phi(\mathbf{T}^\top \mathbf{z}, y, z)$ and $j \notin \phi(\bar{\mathbf{K}}^\top \mathbf{z}, y, z)$, we have $(\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < -(6\kappa - \nu) - 6\kappa - \nu < 0$.

Thus, we have $\max_{y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < 0$. Therefore, $E_d(6\kappa - \nu) \cap E_c \cap E_2 \subseteq \{\max_{y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < 0\} \cap E_c \cap E_2$.

Similarly, by $\max_{(\mathbf{x},y)\in E_c, y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} \leq 2\kappa + \tau$, we have $E_d(2\kappa + \tau) \cap E_c \cap E_3 \subseteq \{\max_{y=i\neq j} (\mathbf{t}_j - \mathbf{t}_i)^\top \mathbf{z} + (\bar{\mathbf{k}}_j - \bar{\mathbf{k}}_i)^\top \mathbf{z} < 0\} \cap E_c \cap E_3$.

Thus, as E_2 and E_3 are disjoint and union bound, we have

$$\begin{aligned} \mathrm{ACC}(f) \geq & \Pr(E_4) + \Pr(E_c \cap E_2) \Pr(E_d(6\kappa - \nu) | E_c \cap E_2) \qquad (37) \\ & + \Pr(E_c \cap E_3) \Pr(E_d(2\kappa + \tau) | E_c \cap E_3) \qquad (38) \\ \geq & \Pr(E_4) + (\Pr(E_2) + \Pr(E_3))(1 - \rho_d(\max\{6\kappa - \nu, 2\kappa + \tau\})) - \rho_c. \qquad (39) \end{aligned}$$

We finish the proof by following $\mathrm{ACC}(f) = 1 - \mathcal{R}_{0-1}(f)$ and $\Pr(E_1) + \Pr(E_2) + \Pr(E_3) + \Pr(E_4) = 1$. □

Remarks: The above theorem characterizes the 0-1 risk upper bound by the modality gap and key properties of retrieved and target distributions. Moreover, logit ensemble utilizes knowledge encoded in different modalities to benefit each other. When $(\Pr(E_2) + \Pr(E_3))(1 - \rho_d(\max\{6\kappa - \nu, 2\kappa + \tau\})) - \rho_c \geq \max\{\Pr(E_2), \Pr(E_3)\}$, we can see that logit ensemble leads to a lower 0-1 risk (*i.e.,* higher accuracy) compared to the zero-shot model. This happens when the modality gap τ is small and the test data exhibits good clustering properties.

D.4. Auxiliary Lemmas

Lemma D.13 (Lipschitz continuity of cross-entropy loss)**.** *When $y \in [C]$, the cross-entropy loss $\ell(\mathbf{v}, y)$ is L-Lipschitz on the hyper-cube, i.e., $\mathbf{v} \in [-1, 1]^C$, where $L = \sqrt{\exp(2) + 1}$.*

Proof of Lemma D.13. Note that since $\ell(\cdot, y) : \mathbb{R}^C \to \mathbb{R}$ is differentiable, it is sufficient to find L such that $\|\nabla \ell(\cdot, y)\|_2 \leq L$. Let $s = \sum_{i\in[C]} \exp(\mathbf{v}_i)$. Applying calculus rules we have that

$$\frac{\partial \ell}{\partial \mathbf{v}_y} = \frac{\exp(\mathbf{v}_y) - s}{s} \quad \text{and} \quad \frac{\partial \ell}{\partial \mathbf{v}_i} = \frac{\exp(\mathbf{v}_y + \mathbf{v}_i)}{s} \;\; \forall i \neq y. \tag{40}$$

Thus,

$$\begin{aligned} \|\nabla \ell(\cdot, y)\|_2^2 = & \frac{\left(\sum_{i\neq y} \exp(\mathbf{v}_i)\right)^2 + \exp(2\mathbf{v}_y) \left(\sum_{i\neq y} \exp(2\mathbf{v}_i)\right)}{s^2} \qquad (41) \\ \leq & \frac{s^2 + \exp(2\mathbf{v}_y) s^2}{s^2} \qquad (42) \\ \leq & \exp(2) + 1. \qquad (43) \end{aligned}$$

Thus, we have $L = \sqrt{\exp(2) + 1}$. □

Lemma D.14 (Bounded logits)**.** *For an input with visual feature $\mathbf{z} \in \mathbb{R}^d$, if $\mathbf{Q}$ is a convex combination among $\{\mathbf{T}, \mathbf{S}, \bar{\mathbf{S}}, \bar{\mathbf{K}}\}$, we have $\mathbf{Q}^\top \mathbf{z} \in [-1, 1]^C$.*

Proof of Lemma D.14. From the definitions of matrices $\mathbf{T}, \mathbf{S}, \bar{\mathbf{S}}, \bar{\mathbf{K}} \in \mathbb{R}^{d\times C}$ defined in Appendix D.1 and Appendix D.2, we have that the Euclidean norm of each column in $\mathbf{T}, \mathbf{S}, \bar{\mathbf{S}}, \bar{\mathbf{K}}$ and $\mathbf{z}$ is smaller or equal to 1. Thus, their convex combination $\mathbf{Q}$ multiplied by $\mathbf{z}$ satisfies $\mathbf{Q}^\top \mathbf{z} \in [-1, 1]^C$. □

Lemma D.15 (Vector Bernstein inequality. Lemma 18 in Kohler & Lucchi (2017))**.** *Let $\mathbf{v}_1, ..., \mathbf{v}_n \in \mathbb{R}^d$ be independent vector-valued random variables and assume that each one is centered, uniformly bounded with variance bounded above:*

$$\mathbb{E}[\mathbf{v}_i] = 0 \text{ and } \|\mathbf{v}_i\|_2 \leq B_2 \text{ as well as } \mathbb{E}[\|\mathbf{v}_i\|_2^2] \leq \sigma^2. \tag{44}$$

Let $\widehat{\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_i$. Then we have for $0 < \epsilon < \sigma^2/B_2$,

$$\Pr(\|\widehat{\mathbf{v}}\|_2 \geq \epsilon) \leq \exp\left(-n \cdot \frac{\epsilon^2}{8\sigma^2} + \frac{1}{4}\right). \tag{45}$$

E. Training-based Adaptation

In Section 5, we have shown the average performance of training-based adaptation, where the feature cache is finetuned (based on the RN50 backbone). In this section, we report the performance for each dataset. The results are shown in Figure 10. The result for each dataset is consistent where I2I retrieval outperforms T2I retrieval and zero-shot CLIP when varying the shot number from 2 to 16.

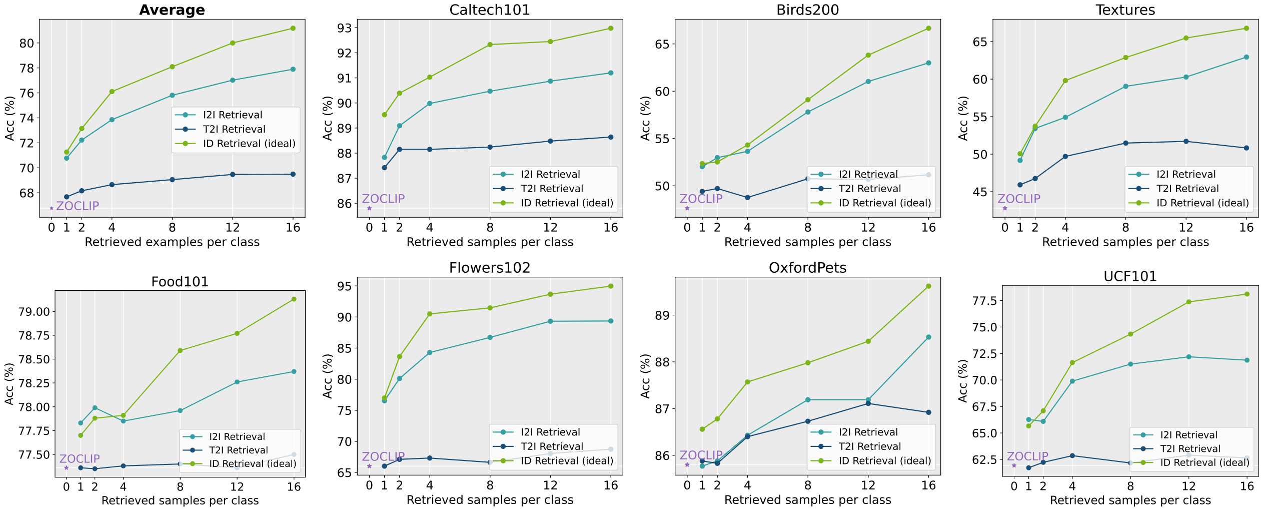


Figure 10. Comparison of retrieval method on adaptation with finetuned feature. Results are based on RN50. We observe a trend similar to training-free adaptation, where I2I retrieval consistently outperforms T2I retrieval and zero-shot CLIP.

F. Impact of Architecture

In Section 5, we show the average performance over all datasets for I2I retrieval and T2I retrieval under different CLIP backbones and observe consistent trends. The results for individual datasets can be seen in Figure 11 (training-free adaptation based on ViT-B/32), Figure 12 (training-based adaptation based on ViT-B/32), Figure 13 (training-free adaptation based on ViT-B/16), Figure 14 (training-based adaptation based on ViT-B/16), Figure 15 (training-free adaptation based on ViT-L/14), and Figure 16 (training-based adaptation based on ViT-L/14).

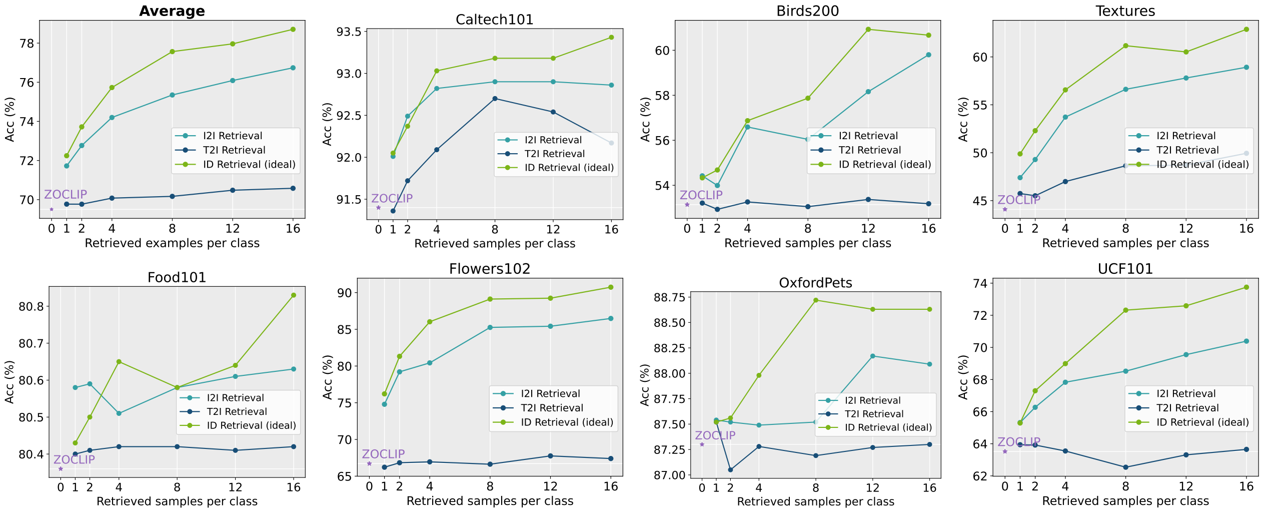


Figure 11. Impact of model architecture. Results are based on ViT-B/32 (training-free).

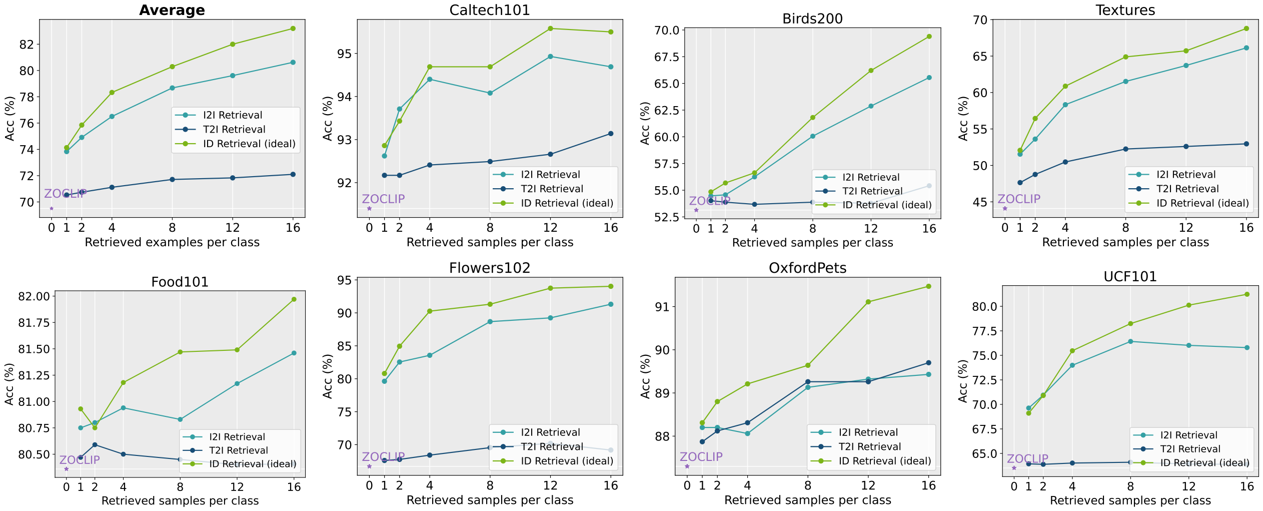


Figure 12. Impact of model architecture. Results are based on ViT-B/32 (feature cache finetuned).

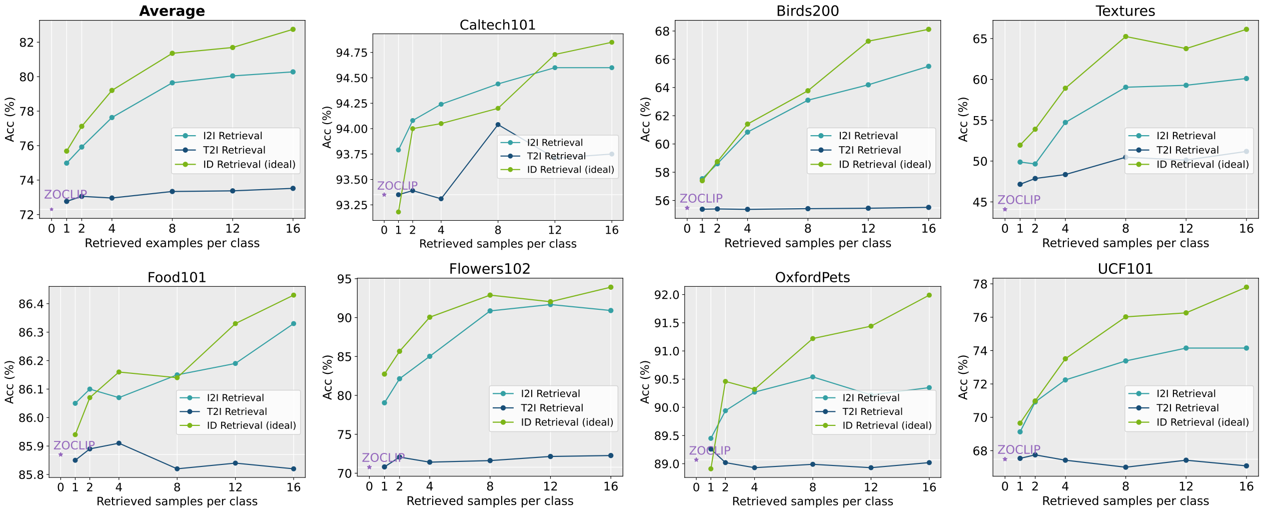


Figure 13. Impact of model architecture. Results are based on ViT-B/16 (training-free).

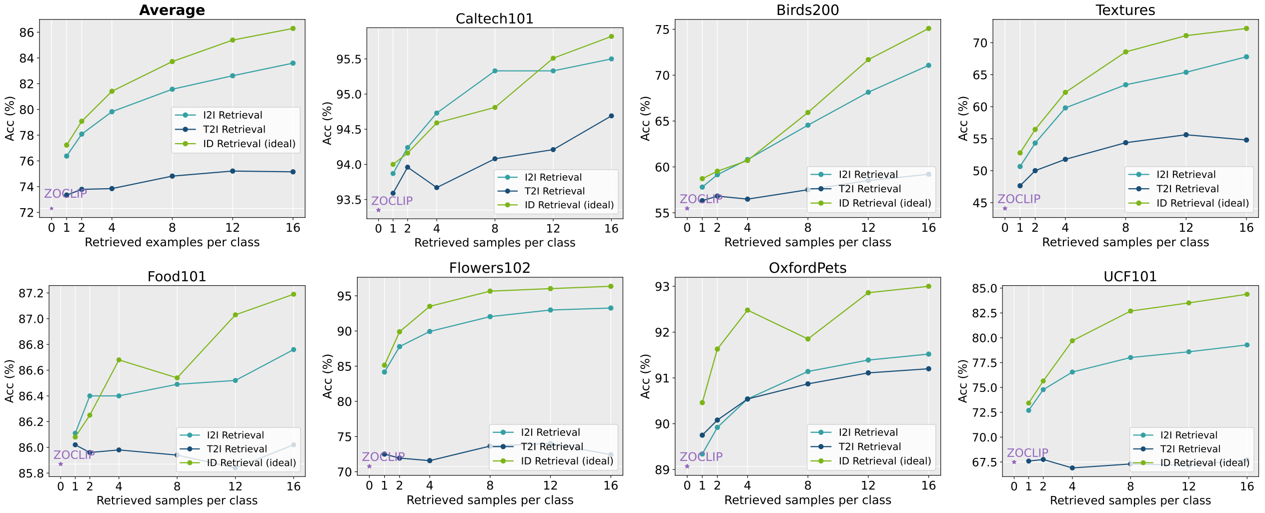


Figure 14. Impact of model architecture. Results are based on ViT-B/16 (feature cache finetuned).

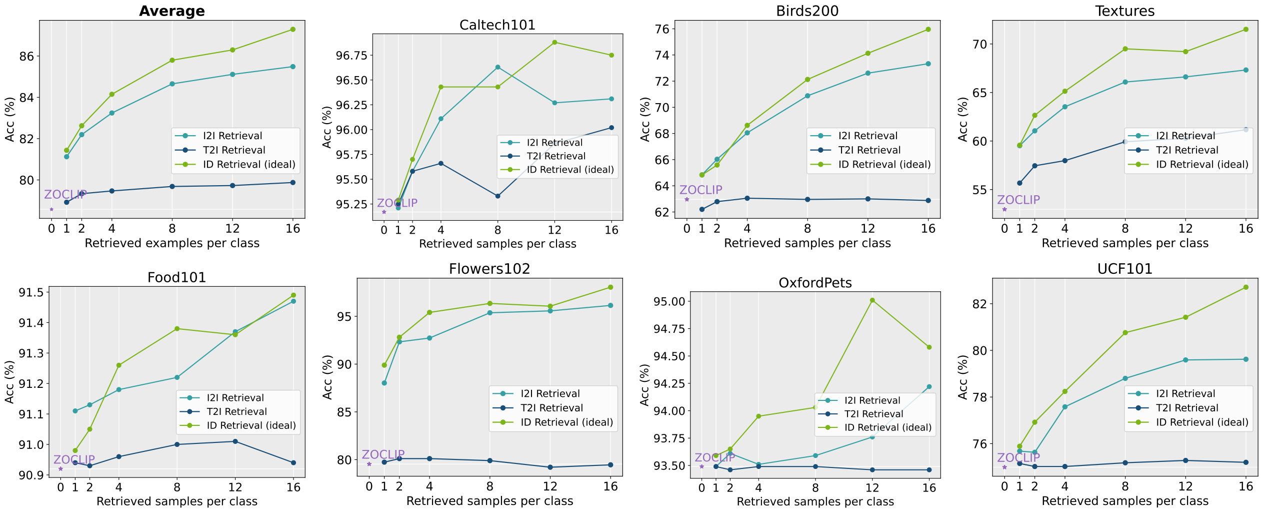


Figure 15. Impact of model architecture. Results are based on ViT-L/14 (training-free).

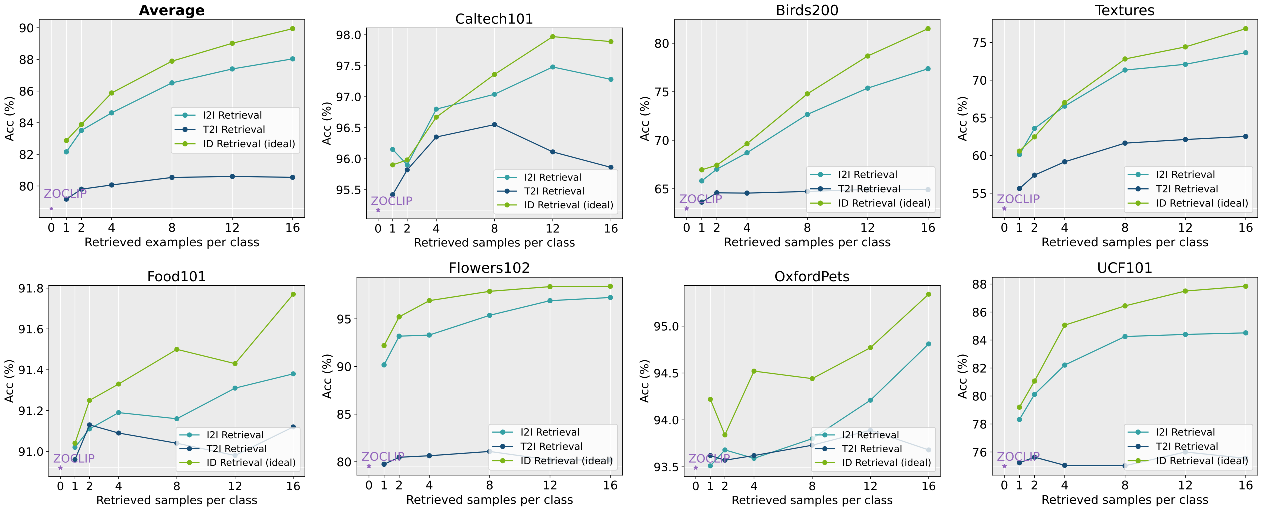


Figure 16. Impact of model architecture. Results are based on ViT-L/14 (feature cache finetuned).

G. Ablation on the ensemble weight scale $\gamma : \alpha$

In the main paper, we set the ensemble weights as tunable hyperparameters. In this section, we conduct an additional ablation study on the ratio of $\gamma : \alpha \in \{0.1, 0.5, 1, 2, 5, 7.5, 10, 15, 20, 50\}$ across different $\omega \in \{0.1, 0.5, 1, 2, 5, 10, 20, 50\}$. We observe that a moderate $\gamma : \alpha$ ratio yields superior performance and the optimal ratio is dataset-dependent. As a concrete example, Table 5 displays the accuracy on each dataset for various $\gamma : \alpha$. The results are based on the RN50 backbone, 8 shot, and $\omega = 2$ with I2I retrieval. For most datasets, a relatively larger ratio (e.g., 5) yields better performance. For Food101, a smaller $\gamma : \alpha$ ratio (e.g., 0.5) suffices.

$\gamma : \alpha$	**Dataset**						
	Caltech 101	Birds200	Textures	Food101	Flowers102	OxfordPets	UCF101
0.1	86.09	47.95	43.32	77.44	66.42	85.77	62.41
0.5	87.42	49.57	44.98	**77.65**	67.52	86.37	64.23
1.0	88.44	50.74	46.75	77.54	69.02	87.14	65.61
2.0	89.01	52.78	48.58	77.17	71.54	**87.35**	66.90
5.0	**89.21**	**53.94**	**50.77**	74.52	78.08	85.04	**66.98**
7.5	88.03	53.64	50.00	71.85	**78.28**	82.20	66.19
10	86.77	52.23	49.17	69.33	76.17	78.6	65.27
15	85.31	50.17	47.87	65.08	74.18	73.15	63.79
20	84.38	48.55	46.87	62.11	75.96	69.04	62.46
50	82.11	43.53	44.62	54.17	76.98	56.75	58.68

Table 5. Ablation on the ensemble weight scale $\gamma : \alpha$.

H. Extension beyond CLIP-like models

In the main paper, we mainly consider pre-trained CLIP-like models due to their wide applicability. To explore whether our findings can be generalized to other vision-language models, in this section, we conduct experiments based on BLIP-2 (Li et al., 2023). Our experiments are based on the feature extraction pipeline from `https://github.com/salesforce/LAVIS`. Table 6 displays the performance (accuracy) when only using the logit from the zero-shot model (ZOC), only using the logit from the retrieval cache (RET), and using an ensemble of logits (Ensemble) for T2I and I2I retrieval, respectively. The same observations also hold for BLIP-2: (1) I2I retrieval consistently outperforms T2I retrieval; (2) Ensemble with the zero-shot prediction is essential. The results are based on 8 shot, and we observe that similar trends hold consistently across other shots from 2 to 16.

Dataset	**Method**				
	ZOCLIP	RET (T2I)	RET (I2I)	Ensemble (T2I)	Ensemble (I2I)
Caltech101	88.19	86.61	91.44	90.14	91.76
Textures	46.16	50.95	58.10	53.31	61.41
Food101	73.39	75.81	71.94	79.66	80.56
Flowers102	41.41	59.44	83.56	62.53	85.87
UCF101	67.57	68.49	73.78	70.63	73.17

Table 6. Extension of findings beyond CLIP-like models. We evaluate the performance of pre-trained BLIP-2 on diverse datasets. The two key observations still hold: (1) I2I retrieval consistently outperforms T2I retrieval; (2) Ensemble with the zero-shot prediction is essential.