SIMPLE YET EFFECTIVE SEMI-SUPERVISED KNOWLEDGE DISTILLATION FROM VISION-LANGUAGE MODELS VIA DUAL-HEAD OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised learning (SSL) has emerged as a practical solution for addressing data scarcity challenges by leveraging unlabeled data. Recently, vision-language models (VLMs), pre-trained on massive image-text pairs, have demonstrated remarkable zero-/few-shot performance that often surpasses SSL approaches due to their exceptional generalization capabilities. This gap motivates us to question: how can we effectively *harness the powerful generalization capabilities of VLMs into task-specific models*? Knowledge distillation (KD) offers a natural framework for transferring VLM capabilities, but we identify that it suffers from *gradient conflicts* between supervised and distillation losses. To address this challenge, we propose **D**ual-**H**ead **O**ptimization (**DHO**), which introduces dual prediction heads for each distinct signal. We observe that **DHO** resolves *gradient conflicts*, enabling improved feature learning compared to single-head KD baselines, with practical benefits of minimal computational overhead and test-time hyperparameter tuning without retraining. Extensive experiments across 15 datasets show that **DHO** consistently outperforms KD baselines, often outperforming teacher models with smaller student models. **DHO** also achieves new state-of-the-art performance on both in-distribution ImageNet semi-supervised learning and out-of-distribution generalization across ImageNet variants. We will publicly release our code and model checkpoints to facilitate future research.

1 INTRODUCTION

Vision-language models (VLMs), which learn joint vision-language representations through large-scale pre-training, have shown remarkable zero-shot capabilities across diverse tasks (Radford et al., 2021; Jia et al., 2021). Building upon these strong foundational capabilities, recent work has explored various adaptation strategies, including parameter-efficient approaches such as linear probing (Li et al., 2022; Huang et al., 2024), lightweight adapters (Zhang et al., 2021; Gao et al., 2024), and prompt-based fine-tuning methods (Zhou et al., 2022b;a; Lafon et al., 2025), demonstrating the potential of VLMs for data-limited visual recognition tasks.

In parallel, semi-supervised learning (SSL) has emerged as a practical approach to address data scarcity by leveraging both labeled and unlabeled data (Sohn et al., 2020; Assran et al., 2021; Cai et al., 2022; Zheng et al., 2023). While these methods have shown success to leverage large amounts of unlabeled data, they often struggle to match the impressive zero- and few-shot capabilities of large pre-trained VLMs (Liu et al., 2023b). This discrepancy highlights a fundamental limitation: traditional semi-supervised methods, despite their theoretical appeal, remain suboptimal compared to the rich representations learned by foundation models through massive-scale pre-training.

VLMs excel at zero- and few-shot generalization but may lack the fine-grained discriminative power needed for specific tasks, while models trained on limited labeled data capture task-specific patterns but generalize poorly. This complementary nature motivates us to integrate generalist VLM knowledge with task-specific supervision in semi-supervised learning settings. Therefore, the challenge naturally arises: *how can we effectively transfer the powerful capabilities of large VLMs to task-specific models in semi-supervised settings?*

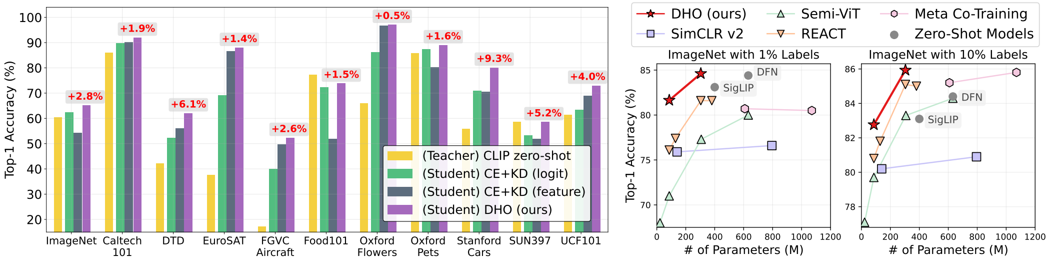

Figure 1: **(Left)**: **DHO** consistently outperforms single-head baselines on 11 datasets under 16-shot semi-supervised setting. The improvements are evaluated in comparison to the second-best one. **(Right)**: **DHO** achieves new SoTA on ImageNet in both 1% and 10% labeled data setting, with fewer parameters.

Knowledge distillation (KD; Hinton, 2015) emerges as a natural solution to the challenge, as it offeres an efficient framework to transfer knowledge from large VLMs to student models while simultaneously leveraging task-specific patterns from limited labeled data. However, existing VLM distillation methods primarily focus on general-purpose training (Yang et al., 2024a; Vasu et al., 2024; Udandarao et al., 2024; Yang et al., 2024c) or employ multi-stage pipelines with unsupervised pre-distillation (Vemulapalli et al., 2024; Wu et al., 2024), which require additional task-specific fine-tuning. Conventional single-stage KD methods, such as logit distillation (Hinton, 2015; Chen et al., 2020), offer a more direct approach, but we find them *suboptimal* in semi-supervised settings.

Through analysis, we identify that this stems from a fundamental problem: *gradient conflicts* between the supervised loss (from limited labeled data) and the distillation loss (from teacher predictions). This misalignment is particularly severe in semi-supervised settings, where the strong and consistent distillation signal from the teacher can overwhelm the weak and potentially noisy signal from scarce labeled data. Such gradient conflicts (■ and ● in Fig. 3) are well-documented to impair effective feature learning (Yu et al., 2020a; Liu et al., 2021; Chen & Er, 2025), preventing the model from finding an optimal balance between task-specific adaptation and general knowledge transfer.

To address the above issue, we propose a *simple yet effective* distillation framework, **DHO** (**D**ual-**H**ead **O**ptimization), which jointly leverages labeled samples and the probabilistic outputs of the teacher model. Specifically, it learns two distinct heads, each optimized with a separate loss: the supervised and the KD loss, respectively. Our analysis reveals that **DHO mitigates gradient conflicts** both in the classification head and the shared feature extractor (● in Fig. 3), arising from the two different training signals and **improves feature representations** compared to baselines (Tab. 5 and Fig. 8). Our framework also enables controlling the relative influence of supervised and teacher predictions through **linear combination** of both head outputs, whose effectiveness is demonstrated in Fig. 6. Additionally, **DHO** enables **post-training hyperparameter adjustment** by tuning the linear combination weights at inference (Fig. 9), eliminating costly training-time hyperparameter search.

We conduct extensive experiments across 15 different datasets including ImageNet (Russakovsky et al., 2015). The experimental results demonstrate the **consistent improvement of DHO** over conventional KD methods across all evaluated datasets (Fig. 1-Left). **DHO** sometimes even **outperforms the zero/few-shot teacher models** with smaller student models with the same labeled data, demonstrating effective task-specific enhancement beyond teacher capabilities (Figs. 4 and 5). Furthermore, **DHO** achieves **new state-of-the-art (SoTA)** performance on ImageNet semi-supervised setting, improving accuracy by 3% and 0.1% with 1% and 10% labeled data, respectively, while using fewer parameters (Fig. 1-Right). Notably, **DHO** can be seamlessly **integrated with existing adaptation techniques** with minimal computational overhead, achieving new **SoTA on Out-of-Distribution (OOD) tasks** across ImageNet distribution-shifted variants (Tab. 2 and §E.11).

Our contributions and findings are summarized as follows:

- We firstly **identify** *gradient conflict* when integrating VLMs' general knowledge with task-specific supervision from limited data. To address this, we propose **D**ual-**H**ead **O**ptimization (**DHO**), which optimizes the supervised and distillation objectives in separate heads.
- **DHO** effectively resolves *gradient conflicts* both in the classification head and the shared feature extractor, leading to improved **feature representations**. Our framework enables flexible **post-training adjustment** of dual head output weights with **minimal computational overhead**.
- In extensive experiments on 15 datasets, **DHO consistently outperforms** baselines and sometimes **surpasses teacher model** performance. It establishes new **SoTA** for **in-distribution ImageNet** (*e.g.*, +3%/+0.1% at 1%/10% labels) with fewer parameters (76M/767M), and also achieves **SoTA OOD generalization** across ImageNet variants when integrated with adaptation methods.

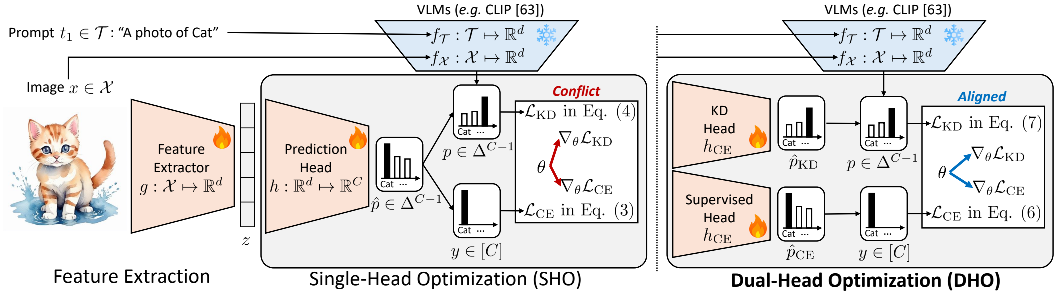
VLMs (e.g. CLIP [63])
Prompt $t_1 \in \mathcal{T}$: "A photo of Cat"
$f_\mathcal{T} : \mathcal{T} \mapsto \mathbb{R}^d$
$f_\mathcal{X} : \mathcal{X} \mapsto \mathbb{R}^d$
Image $x \in \mathcal{X}$
Feature Extractor $g : \mathcal{X} \mapsto \mathbb{R}^d$
z
Prediction Head $h : \mathbb{R}^d \mapsto \mathbb{R}^C$
Cat ...
$\hat{p} \in \Delta^{C-1}$
$p \in \Delta^{C-1}$
$y \in [C]$
Conflict
$\mathcal{L}_{\mathrm{KD}}$ in Eq. (4)
$\nabla_\theta \mathcal{L}_{\mathrm{KD}}$
θ
$\nabla_\theta \mathcal{L}_{\mathrm{CE}}$
$\mathcal{L}_{\mathrm{CE}}$ in Eq. (3)
Feature Extraction
Single-Head Optimization (SHO)
VLMs (e.g. CLIP [63])
KD Head h_{CE}
Supervised Head h_{CE}
$\hat{p}_{\mathrm{KD}}$
$\hat{p}_{\mathrm{CE}}$
$p \in \Delta^{C-1}$
$y \in [C]$
Aligned
$\mathcal{L}_{\mathrm{KD}}$ in Eq. (7)
$\nabla_\theta \mathcal{L}_{\mathrm{KD}}$
θ
$\nabla_\theta \mathcal{L}_{\mathrm{CE}}$
$\mathcal{L}_{\mathrm{CE}}$ in Eq. (6)
Dual-Head Optimization (DHO)

Figure 2: **Conceptual illustration** on KD frameworks, Single-Head Optimization (SHO) and **D**ual-**H**ead **O**ptimization (**DHO**), for semi-supervised settings. As demonstrated in Fig. 3, we observe *gradient conflict* of SHO. In contrast, **DHO mitigates such conflicts** by leveraging dual-head architectures in Fig. 3.

2 Method

2.1 Preliminaries

We begin with preliminaries: a brief background on VLMs, the problem formulation for few-/low-shot learning, and single-head KD baselines. We defer related work—**vision–language pretraining, data-limited adaptation of VLMs, knowledge distillation, and dual-head methods**—to §A.

Background on VLMs. Our work is based on VLMs such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021). These models consist of multimodal encoders: an image encoder $f_\mathcal{X} : \mathcal{X} \to \mathbb{R}^d$ and a text encoder $f_\mathcal{T} : \mathcal{T} \to \mathbb{R}^d$ where $\mathcal{X}$ and $\mathcal{T}$ denote the domains of images and texts, respectively. For zero-shot classification of VLMs across C classes, we use predefined prompt templates, *e.g.*, "a photo of a `[CLASS]`", where `[CLASS]` is the name of class. Given a set of C target classes, *i.e.*, $y \in \{1, \ldots, C\}$, we generate prompted text descriptions $\{t_1, t_2, \ldots, t_C\}$. We obtain the categorical probability vector p using the cosine similarity $\texttt{CosSim}(x, y) = \frac{x^\top y}{||x||_2 \cdot ||y||_2}$ over $\{t_1, t_2, \ldots, t_C\}$, *i.e.*, $p := \sigma([\texttt{CosSim}(f_\mathcal{T}(x), f_\mathcal{T}(t_1))/\zeta, \ldots, \texttt{CosSim}(f_\mathcal{T}(x), f_\mathcal{T}(t_C))/\zeta])$, where σ is the softmax function, $\zeta \in \mathbb{R}_{>0}$ is the temperature scaling (Hinton, 2015), and final classification is determined by $\arg\max_{c \in \{1,\ldots,C\}} p_c$.

Problem Formulation. We focus on transferring knowledge from VLMs to task-specific models under few-shot or low-shot semi-supervised learning scenarios, where both labeled and unlabeled data are utilized. Specifically, given a K-shot and C-class classification problem, we are provided with a labeled dataset $\mathcal{D}^{(l)} = \{(x_n^{(l)}, y_n)\}_{n=1}^N$, where $N = K \times C$ is the total number of labeled examples, and $y_n \in \{1, \ldots, C\}$ denotes the class labels. Additionally, we have access to an unlabeled dataset $\mathcal{D}^{(u)} = \{x_m^{(u)}\}_{m=1}^M$ consisting of M unlabeled images. Low-shot learning represents a more realistic setting than traditional few-shot learning where only a small fraction, *e.g.*, 1% ($\frac{N}{N+M} \approx 0.01$) or 10% ($\frac{N}{N+M} \approx 0.1$) of the total dataset is labeled. Our goal is then to develop a student model by leveraging $\mathcal{D}^{(l)}$ and $\mathcal{D}^{(u)}$, guided by the knowledge of the VLM encoders $f_\mathcal{X}$ and $f_\mathcal{T}$. The student model consists of a **feature extractor** $g(x)$ **parameterzied by** θ and **a linear prediction head** $h(z) = Wz + b \in \mathbb{R}^c$, where $W \in \mathbb{R}^{C \times d}$, $b \in \mathbb{R}^C$, followed by **the softmax function** σ.

Single-head optimization (SHO) of KD. Our method builds on *logit distillation* in semi-supervised settings (Hinton, 2015; Chen et al., 2020) that combines supervised loss $\mathcal{L}_{\mathrm{CE}}$ on the labeled dataset $\mathcal{D}^{(l)}$ with KD loss $\mathcal{L}_{\mathrm{KD}}$ on both labeled and unlabeled datasets $\mathcal{D}^{(l)} \cup \mathcal{D}^{(u)}$, *i.e.*, $\lambda \mathcal{L}_{\mathrm{CE}} + (1 - \lambda)\mathcal{L}_{\mathrm{KD}}$. Specifically, the supervised loss $\mathcal{L}_{\mathrm{CE}}$ and the KD loss $\mathcal{L}_{\mathrm{KD}}$ are defined as follows:

$$\mathcal{L}_{\mathrm{CE}} = \frac{1}{N} \sum_n \ell \left(\sigma(h(z_n^{(l)})), y_n \right), \tag{1}$$

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{N} \sum_n D_{\mathrm{KL}} \left[p_n^{(l)} \| \sigma(h(z_n^{(l)})) \right] + \frac{1}{M} \sum_m D_{\mathrm{KL}} \left[p_m^{(u)} \| \sigma(h(z_m^{(u)})) \right]. \tag{2}$$

where ℓ denotes D_{KL} represent the cross-entropy and Kullback-Leibler divergence, respectively. $z_n^{(l)} = g(x_n^{(l)})$ and $z_m^{(u)} = g(x_m^{(u)}) \in \mathbb{R}^d$ are feature representations obtained by the feature extractor g. $p_n^{(l)}$ and $p_m^{(u)}$ are the categorical probability vectors of labeled $x_n^{(l)}$ and unlabeled data $x_m^{(u)}$, respectively, obtained by teacher VLM encoders $f_\mathcal{X}$ and $f_\mathcal{T}$. Another well-studied single-head KD baseline is *feature distillation*, which leverages mean squared error (MSE) loss to directly align feature representations extracted by the student encoder g and the teacher image encoder $f_\mathcal{X}$. We defer the details of feature distillation for VLMs to CLIP-KD (Yang et al., 2024a).

2.2 DUAL-HEAD OPTIMIZATION (DHO)

Gradient conflicts in SHO. Logit distillation in SHO, as described in §2.1, provides a simple approach for transferring knowledge from VLMs to task-specific models. However, we find that its performance gain is *suboptimal*, which we attribute to **gradient conflicts** between the supervised and KD loss signals which hinder effective feature learning (Yu et al., 2020a; Liu et al., 2021; Chen & Er, 2025). As illustrated in -■- and -●- of Fig. 3, both **the classifier weight W of classifier h and the parameter vector θ of feature extractor g** suffer from gradient conflicts: the cosine similarity between their respective gradients turns negative, *i.e.*, $\texttt{CosSim}(\nabla_\theta \mathcal{L}_{\text{CE}}, \nabla_\theta \mathcal{L}_{\text{KD}}) < 0$ and $\texttt{CosSim}(\nabla_W \mathcal{L}_{\text{CE}}, \nabla_W \mathcal{L}_{\text{KD}}) < 0$, indicating misaligned optimization directions.

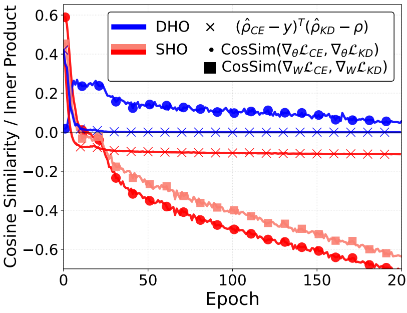

Figure 3: **The average cosine similarity and inner product** over 10 datasets.

To understand this, we first analyze the gradient with respect to (w.r.t) the weight W of a linear head h. The gradients w.r.t the classifier weight W are $\nabla_W \mathcal{L}_{\text{CE}} = (\hat{p} - y)z^\top$ and $\nabla_W \mathcal{L}_{\text{KD}} = (\hat{p} - p)z^\top$, respectively. Their cosine similarity is proportional to $(\hat{p} - y)^\top (\hat{p} - p) \cdot \|z\|^2$:

$$\texttt{CosSim}(\nabla_W \mathcal{L}_{\text{CE}}, \nabla_W \mathcal{L}_{\text{KD}}) \propto (\hat{p} - y)^\top (\hat{p} - p) \cdot \|z\|^2, \tag{3}$$

which misaligns when $(\hat{p} - y)^\top (\hat{p} - p) < 0$ due to prediction mismatch, falling below zero during training (-×- in Fig. 3), leading to gradient conflicts (-■- in Fig. 3).

We observe that **similar gradient conflicts happen on the parameters θ in the feature extractor** g. Let the gradients w.r.t to the feature representation z be $\nabla_z \mathcal{L}_{\text{CE}} = W^\top(\hat{p} - y)$ and $\nabla_z \mathcal{L}_{\text{KD}} = W^\top(\hat{p} - p)$, Then, applying the chain rule, the gradients with respect to the feature extractor parameters θ become $\nabla_\theta \mathcal{L}_{\text{CE}} = \nabla_z \mathcal{L}_{\text{CE}} \cdot \frac{\partial z}{\partial \theta}$ and $\nabla_\theta \mathcal{L}_{\text{KD}} = \nabla_z \mathcal{L}_{\text{KD}} \cdot \frac{\partial z}{\partial \theta}$. Similarly, the cosine similarity for the feature extractor parameters becomes:

$$\texttt{CosSim}(\nabla_\theta \mathcal{L}_{\text{CE}}, \nabla_\theta \mathcal{L}_{\text{KD}}) \propto (\hat{p} - y)^\top W \left(\frac{\partial z}{\partial \theta}\right)^\top \frac{\partial z}{\partial \theta} W^\top (\hat{p} - p) \tag{4}$$

While the complexity of the Jacobian $\frac{\partial z}{\partial \theta}$ makes it difficult to theoretically guarantee gradient conflicts, we empirically observe that **gradient conflicts occur in the parameters θ of the feature extractor** g (-●- in Fig. 3), with $\texttt{CosSim}(\nabla_\theta \mathcal{L}_{\text{CE}}, \nabla_\theta \mathcal{L}_{\text{KD}}) < 0$. Thus, we hypothesize that the gradient conflicts in the feature extractor g are **propagated from the classification head** h.

Dual-head architecture. To mitigate this issue, we propose **D**ual **H**ead **O**ptimization (**DHO**) to decouple $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{KD}}$ via *two independent prediction heads*: $h_{\text{CE}}(z) = W_{\text{CE}} z + b_{\text{CE}}$ and $h_{\text{KD}}(z) = W_{\text{KD}} z + b_{\text{KD}}$, with $W_{\text{CE}}, W_{\text{KD}} \in \mathbb{R}^{C \times d}$ and $b_{\text{CE}}, b_{\text{KD}} \in \mathbb{R}^C$. The corresponding losses are:

$$\mathcal{L}_{\text{CE}} = \frac{1}{N} \sum_n \ell\left(\sigma(h_{\text{CE}}(z_n^{(l)})), y_n\right), \tag{5}$$

$$\mathcal{L}_{\text{KD}} = \frac{1}{N} \sum_n D_{\text{KL}}\left[p_n^{(l)} \| \sigma(h_{\text{KD}}(z_n^{(l)}))\right] + \frac{1}{M} \sum_m D_{\text{KL}}\left[p_m^{(u)} \| \sigma(h_{\text{KD}}(z_m^{(u)}))\right], \tag{6}$$

where the final loss combines both objectives as $\lambda \mathcal{L}_{\text{CE}} + (1 - \lambda)\mathcal{L}_{\text{KD}}$.

Mitigation of gradient conflict in DHO. In **DHO**, gradient conflicts in **the classification head naturally disappear by decoupling the optimization** of W_{CE} and W_{KD}. This separation enables each head to learn distinct signals without interference, reducing prediction mismatch (-×- of Fig. 3). Let $\hat{p}_{\text{CE}} = \sigma(h_{\text{CE}}(z))$ and $\hat{p}_{\text{KD}} = \sigma(h_{\text{KD}}(z))$, then $\nabla_z \mathcal{L}_{\text{CE}} = W_{\text{CE}}^\top (\hat{p}_{\text{CE}} - y)$ and $\nabla_z \mathcal{L}_{\text{KD}} = W_{\text{KD}}^\top (\hat{p}_{\text{KD}} - p)$. The cosine similarity between gradients w.r.t θ in **DHO** is defined as:

$$\texttt{CosSim}(\nabla_\theta \mathcal{L}_{\text{CE}}, \nabla_\theta \mathcal{L}_{\text{KD}}) \propto (\hat{p}_{\text{CE}} - y)^\top W_{\text{CE}} \left(\frac{\partial z}{\partial \theta}\right)^\top \frac{\partial z}{\partial \theta} W_{\text{KD}}^\top (\hat{p}_{\text{KD}} - p), \tag{7}$$

where we empirically find that **gradient conflicts in θ of the feature extractor** g for **DHO** are also resolved, *i.e.*, maintaining positive gradient alignment throughout training (*i.e.*, Eq. 7 > 0; as shown in -●- of Fig. 3). It enables conflict-free representation learning, and leads to **better feature representation** compared to SHO, which empirically validated by linear evaluation in Tab. 5. We defer an algorithm that describes the full training procedure of **DHO** to Alg. 1.

2.3 Dual-head Interpolation

After training, we find that using only one of the two heads at inference is *suboptimal* (see Fig. 6). Motivated by the mixture-of-experts paradigm (Jacobs et al., 1991), we adopt a *simple yet effective* inference rule for **DHO** that linearly interpolates the output probability vectors of two heads:

$$\hat{p}_{\texttt{DHO}} = \alpha\, \sigma\big(h_{\text{CE}}(z)\big) + (1-\alpha)\, \sigma\big(h_{\text{KD}}(z)/\beta\big), \tag{8}$$

where $\alpha \in [0,1]$ balances the supervised and KD heads, $\beta > 0$ is a temperature that softens the KD logits, and the final prediction is $\arg\max_{c\in[C]} \hat{p}_{\texttt{DHO}}$. In practice, we tune α and β on a validation set to reflect dataset-specific supervision quality and teacher accuracy, allowing the model to weight the more reliable source. See Alg. 2 for the full inference procedure.

Effect of α and β. We now demonstrate the effect of tuning α and β using a validation set. Under mild assumptions, **DHO** ε-approximates SHO in ℓ_1 by setting $\alpha = \lambda$ and $\beta = 1$:

Assumption 1 (ε-convergence)**.** *Assume that, after sufficient training, both heads converge to their respective target distributions with ℓ_1-bounded error:*

$$\sup_x \|\sigma(h_{\text{CE}}(z)) - y\|_1 \le \varepsilon, \quad \sup_x \|\sigma(h_{\text{KD}}(z)) - p\|_1 \le \varepsilon, \quad \text{where} \quad \varepsilon \in \mathbb{R}_{>0}. \tag{9}$$

Theorem 1 (Inference equivalence)**.** *Under Assumption 1, by setting $\alpha = \lambda$ and $\beta = 1$, then $\|\hat{p}_{\texttt{DHO}} - \hat{p}_{\text{SHO}}\|_1 \le \varepsilon$, where $\hat{p}_{\text{SHO}}$ is the output of SHO optimally trained with λ.*

Details and proofs are deferred to §B. Theorem 1 implies that $\hat{p}_{\text{SHO}}$ trained with any λ can be approximated by the dual-head interpolation in Eq. 8, by setting $\alpha = \lambda$ and $\beta = 1$. Here, λ is a **training** hyperparameter of SHO, while α and β are **inference** hyperparameters of **DHO**, allowing it to **emulate SHO hyperparameter tuning** without retraining.

Language-aware initialization for VLM students In the case of VLM-to-VLM distillation, we leverage the text encoder $f_{\mathcal{T}}$ of teachers when initializing the dual heads h_{CE} and h_{KD}. Following prior work (Li et al., 2022), we initialize the weights as $W_{\text{CE}}, W_{\text{KD}} \leftarrow [f_{\mathcal{T}}(t_1), \ldots, f_{\mathcal{T}}(t_C)]^\top \in \mathbb{R}^{C\times d}$. We further align the prediction logic of KD head h_{KD}, with the cosine similarity-based approach of the teacher VLMs as follows:

$$h_{\text{KD}} = \frac{1}{\zeta}[\texttt{CosSim}(g(x), w_1), \ldots, \texttt{CosSim}(g(x), w_C)]^\top \in \mathbb{R}^C, \tag{10}$$

where $w_c \in \mathbb{R}^d$ denotes the c-th row of W_{KD}.

3 Experiments

3.1 Experimental Setups

Datasets. For in-distribution evaluation, we use ImageNet (Russakovsky et al., 2015) and 10 fine-grained datasets (Fei-Fei et al., 2004; Parkhi et al., 2012; Krause et al., 2013; Nilsback & Zisserman, 2008; Bossard et al., 2014; Maji et al., 2013; Xiao et al., 2010; Cimpoi et al., 2014; Helber et al., 2019; Soomro, 2012). To assess out-of-distribution (OOD) generalization, we use four ImageNet variants (Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021a;b). See §D for details.

Baselines. We compare **DHO** with conventional single-head KD baselines; **CE**: training only on labeled dataset $\mathcal{D}^{(l)}$ with cross entropy loss (Eq. 1), **KD (logit)**: on unlabeled dataset $\mathcal{D}^{(u)}$ with logit distillation (Eq. 2), and **KD (feature)**: on $\mathcal{D}^{(u)}$ with feature distillation (Yang et al., 2024a). We train on both $\mathcal{D}^{(l)}$ and $\mathcal{D}^{(u)}$ with **CE+KD (logit)** or **CE+KD (feature)**: combining CE with each KD variant using balancing hyperparameter λ. We also consider dual-head KD approaches **SSKD** (He et al., 2021) and **DHKD** (Yang et al., 2024d), though for different purposes as detailed in §A.

For in-distribution evaluation on ImageNet with low-shot settings, we compare against **self and semi-supervised learning** (Chen et al., 2020; Assran et al., 2021; 2022; Cai et al., 2022; Zheng et al., 2023), **CLIP-based-training** (Li et al., 2022; Liu et al., 2023b), **co-training** (Rothenberger & Diochnos, 2023), **KD** (Chen et al., 2020), and **zero-shot VLMs** (Zhai et al., 2023; Fang et al., 2023).

For OOD evaluation, we compare against VLM adaptation methods, including **VPT** (Jia et al., 2022), **CoOp** (Zhou et al., 2022b), **PromptSRC** (Khattak et al., 2023b), and **CasPL** (Wu et al., 2024).

Table 1: Results on **ImageNet** under few-shot semi-supervision using **ResNet-18** and **ResNet-50**. **DHO consistently outperforms all baselines and even the teacher** with ResNet-50 (*e.g.*, +0.7/1.9/2.2/3.4/4.4% with a zero-shot teacher; +1.3/1.5/1.4/1.7/1.4% with a few-shot teacher).

	ResNet-18 trained from scratch						**Self-supervised ResNet-50**					
Method	0-shot	1-shot	2-shot	4-shot	8-shot	16-shot	0-shot	1-shot	2-shot	4-shot	8-shot	16-shot
Single-head KD methods												
KD	51.0	-	-	-	-	-	61.0	-	-	-	-	-
CE	-	0.7	1.1	1.8	3.4	8.2	-	11.4	17.3	26.4	36.7	47.0
CE+KD (feature)	-	17.1	23.5	28.0	32.2	33.8	-	23.0	32.3	41.3	48.2	54.3
CE+KD (logit)	-	50.5	50.6	50.6	51.0	51.2	-	60.4	60.8	61.2	61.6	62.3
Dual-head KD methods												
SSKD (He et al., 2021)	-	42.5	46.2	48.0	50.6	52.0	-	55.2	58.1	60.0	62.3	64.0
DHKD (Yang et al., 2024d)	-	19.7	23.5	23.4	23.7	26.8	-	25.6	34.8	42.7	49.2	55.2
DHO (Ours)	-	51.8	52.4	52.6	53.3	54.5	-	61.0	62.1	62.5	63.7	64.7
DHO−F (Ours)	-	**53.7**	**54.2**	**54.8**	**56.2**	**57.7**	-	**62.3**	**63.1**	**63.9**	**65.5**	**66.8**
Teacher Models (Resnet-50)												
CLIP	60.3	-	-	-	-	-	60.3	-	-	-	-	-
Tip-Adapter-F	-	61.0	61.6	62.5	63.8	65.4	-	61.0	61.6	62.5	63.8	65.4

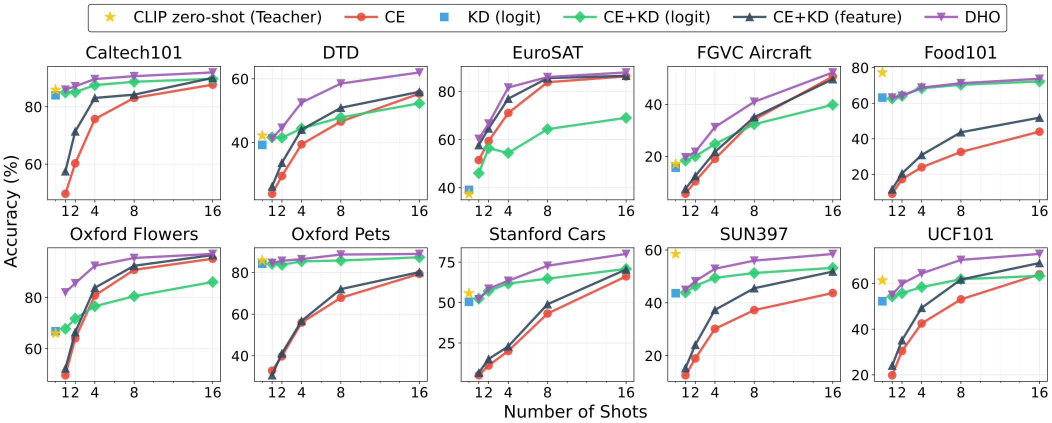

Figure 4: Results on **10 datasets** under few-shot semi-supervision using **ResNet-18** with **zero-shot teacher**.

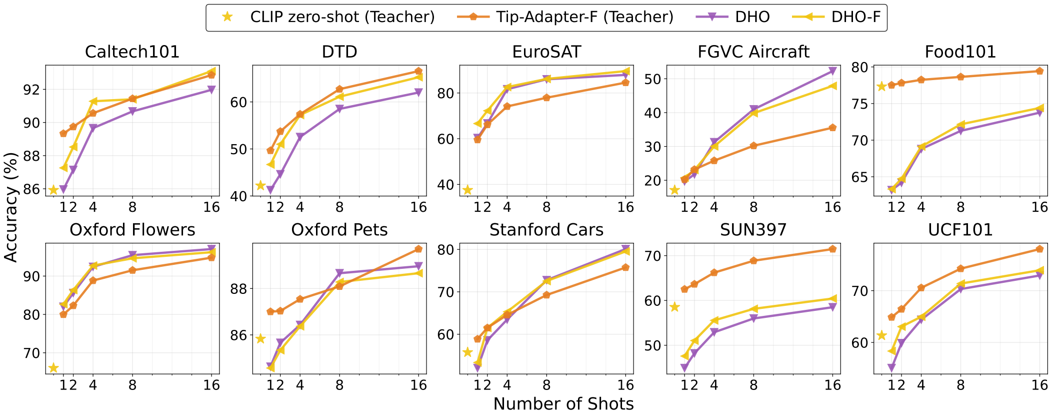

Figure 5: Results on **10 datasets** using **ResNet-18** with either zero- or **few-shot teacher**.

Implementation details. We evaluate across few-shot (1/2/4/8/16-shot) and low-shot (1%/10%) settings, treating remaining data as unlabeled. We adopt Tip-Adapter-F (Zhang et al., 2021) for few-shot teachers, denoting this variant as **DHO−F**. For a fair comparison, we use the same hyperparameters for **DHO** and SHO baselines (except α, β); for other methods, we report the published results. When validation data is unavailable (*e.g.*, ImageNet), we fix $\beta = 0.5$ across all settings. We heuristically set $\alpha = 0.4$ for zero-shot teachers and $\alpha = 0.2$ for few-shot teachers, reflecting the latter's higher reliability. In low-shot settings, we use $\alpha = 0.5$ due to increased label availability. See §C for details.

3.2 Main Results

Effectiveness of DHO compared to conventional KD baselines. Tab. 1 presents results on ImageNet under few-shot semi-supervision using ResNet-18/ResNet-50 student models and ResNet-50 VLM teachers. **DHO consistently outperforms** all baselines across all settings. Other dual-head methods (*i.e.*, SSKD/DHKD), not designed for few-shot semi-supervision, even underperform CE+KD

Table 2: Results on **ImageNet and its distribution-shifted variants**. †: backbone frozen during training. **DHO consistently improves** VLM adaptation methods and **achieves SoTA results** on OOD benchmarks.

Method	Labeled Data	Teacher Model	Unlabeled Data	Val	V2	Sketch	R	A	Avg
CLIP Zero-shot									
CLIP	-	-	-	66.7	60.8	46.2	74.0	47.8	59.1
Linear Evaluation									
Linear Evaluation	1%	-	-	72.8	64.2	47.0	74.6	48.6	61.4
DHO†	1%	✓	-	73.4	65.3	48.3	75.8	49.4	62.4
DHO†	1%	✓	✓	**74.6**	**66.4**	**49.2**	**76.1**	**49.8**	**63.2**
Visual Prompt Tuning									
VPT	1.25%	-	-	73.6	64.6	47.7	75.2	48.7	62.0
VPT+**DHO**	1.25%	✓	-	73.6	65.3	48.7	75.9	49.1	62.5
VPT+**DHO**	1.25%	✓	✓	**75.1**	**66.8**	**50.0**	**76.9**	**50.5**	**63.9**
VLM Text-encoder Prompt Tuning									
CoOp	1.25%	-	-	71.5	64.2	48.0	75.2	49.7	61.7
CoOp+CasPL	1.25%	✓	✓	71.9	64.3	48.3	76.0	-	-
CoOp+**DHO**	1.25%	✓	-	72.8	65.5	49.3	76.4	49.5	62.7
CoOp+**DHO**	1.25%	✓	✓	**73.4**	**66.2**	**49.5**	**77.0**	**50.5**	**63.3**
VLM Multimodal Prompt Tuning									
PromptSRC	1.25%	-	-	71.3	64.4	49.6	77.8	50.9	62.8
PromptSRC+CasPL	1.25%	✓	✓	72.8	65.7	49.7	77.9	-	-
PromptSRC+**DHO**	1.25%	✓	-	73.0	65.3	49.5	77.8	**51.3**	63.4
PromptSRC+**DHO**	1.25%	✓	✓	**73.6**	**66.1**	**49.8**	**78.1**	51.0	**63.7**

on logits. Notably, with ResNet-50, **DHO outperforms the teacher in every few-shot setting** (*e.g.*, +0.7/1.8/2.2/3.4/4.4% or +1.3/1.5/1.4/1.7/1.4% with zero-shot or few-shot teachers).

Next, we evaluate **DHO** on 10 additional datasets using the ResNet-18 student and ResNet-50 VLM teachers. In Fig. 4, we observe that **DHO** also **consistently outperforms all baseline methods** across 10 datasets, while relative rankings between baselines vary across datasets. Fig. 5 further demonstrates that the few-shot teacher is more effective than using a zero-shot teacher for **DHO**. Remarkably, **the ResNet-18 student model trained with DHO achieves better performance than the ResNet-50 teacher model** in most cases, demonstrating knowledge transfer capability of **DHO**.

DHO achieves SoTA performance on ImageNet under low-shot settings. We compare **DHO** to previous state-of-the-art (SoTA) methods on ImageNet under 1% and 10% labeled data. All results are taken from published papers, except for **DHO** and CE+KD (logit). As shown in Tab. 3, **DHO** with ViT-L/14 **surpasses the previous SoTA (*e.g.*, +3.0%/+0.1%), while using fewer parameters (*e.g.*, 76M/767M)**, on both 1%/10% labeled data. Notably, CE + KD (logit) outperforms semi-supervised methods with the same parameters, demonstrating that they are **suboptimal compared to methods leveraging the rich representations of VLMs**.

DHO achieves SoTA performance on ImageNet OOD benchmarks. We evaluate **DHO** across various VLM adaptation approaches. As shown in Tab. 2, **DHO** consistently improves different adaptation methods. Compared to CasPL, **DHO** exhibits superior performance across both in-distribution (Val) and out-of-distribution (V2/Sketch/R/A) benchmarks, **establishing new SoTA OOD results** in semi-supervised setting. This suggests that **joint training with labeled supervision and teacher distillation provides a more effective strategy** than the sequential approach of CasPL, aligning with our hypothesis on **DHO** for resolving gradient conflicts. See §E.11 for results with fully trained models.

Table 3: Results on **ImageNet under low-shot settings**. For CT and MCT methods, numbers in parentheses indicate the number of different architectures for co-training.

Method	Architecture	Params (M)	1% (%)	10% (%)
Self and Semi-supervised Learning				
MSN	ViT-B/4	86	75.7	80.2
Semi-ViT	ViT-L/14	307	77.3	83.3
Semi-ViT	ViT-H/14	632	80.0	84.3
CLIP-based Training				
CLIP	ViT-B/16	86	74.3	80.4
REACT	ViT-B/16	86	76.1	80.8
REACT (Gated-Image)	ViT-B/16	129	77.4	81.8
CLIP	ViT-L/14	304	80.5	84.7
REACT	ViT-L/14	304	81.6	85.1
REACT (Gated-Image)	ViT-L/14	380	81.6	85.0
Co-training based Methods				
CT	Multi-arch (2)	608	80.1	85.1
MCT	Multi-arch (2)	608	80.7	85.2
CT	Multi-arch (4)	1071	80.0	84.8
MCT	Multi-arch (4)	1071	80.5	85.8
Knowledge Distillation				
SimCLR v2 distill	ResNet-50 (2×+SK)	140	75.9	80.2
SimCLR v2 self-distill	ResNet-154 (3×+SK)	795	76.6	80.9
CE + KD (logit)	ViT-B/16	86	79.8	80.4
CE + KD (logit)	ViT-L/14	304	83.1	83.6
DHO	ViT-B/16	86	81.6	82.8
DHO	ViT-L/14	304	**84.6**	**85.9**
Zero-shot VLMs				
SigLIP	ViT-SO400M/14	400	83.1	-
DFN	ViT-H/14	632	84.4	-

Addtional results. See §E for additional results of **MobileNetV2**, or comparison to **PCGrad** (Yu et al., 2020b) and **category-aware KD methods** (Zhao et al., 2022; Lv et al., 2024).

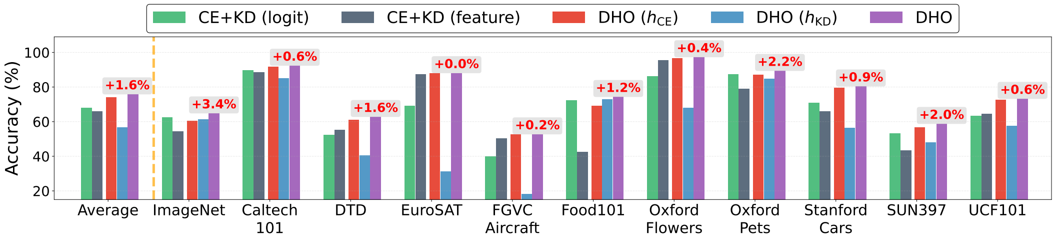

Figure 6: Results of ablation studies on **dual-heads interpolation strategy** in Eq. 8 of **DHO**.

3.3 Analysis

Minimal computational overhead of DHO. As shown in Tab. 4, **DHO** introduces **negligible computational overhead with minimal parameter increase** on the ImageNet with 1000 classes, which can be further reduced down for datasets with fewer classes ($C < 1000$). We further provide inference computational overhead for other models in Tab. 9.

Table 4: **Inference overhead** using RTX 4090.

Model	Params (M)	FLOPs (G)	Throughput (im/s)
ResNet-18	11.69	1.83	3525.7
+ **DHO**	12.20 (+4.4%)	1.83 (+0.0%)	3518.6 (-0.20%)
ResNet-50	25.56	4.14	1018.4
+ **DHO**	27.61 (+8.0%)	4.15 (+0.2%)	1016.4 (-0.20%)

Effectiveness of dual-head interpolation. We evaluate the effectiveness of dual-head interpolation (Eq. 8) by comparing **DHO** with CE+KD (logit), CE+KD (feature), and ablations **DHO** (h_{CE}) and **DHO** (h_{KD}), which predict using only one head (*i.e.*, $\alpha = 1$ or 0). As shown in Fig. 6, **DHO** outperforms **DHO** (h_{CE}) by an average of 1.6% across 11 datasets, with a maximum gain of +3.4% on ImageNet and no degradation on any dataset. Since α and β are inference-time hyperparameters, dual-head interpolation introduces minimal overhead while consistently **improving or maintaining performance**. We also investigate the effectiveness of our adaptive weighting strategy in §E.6. Fig. 7 illustrates three challenging examples: CE head (h_{CE}) is correct in the first, KD head (h_{KD}) in the second, and both fail in the third case, yet the proposed combined prediction is correct—demonstrating the ability of **DHO** to resolve individual head failures. See §F.2 for additional analysis on theses challenging examples.

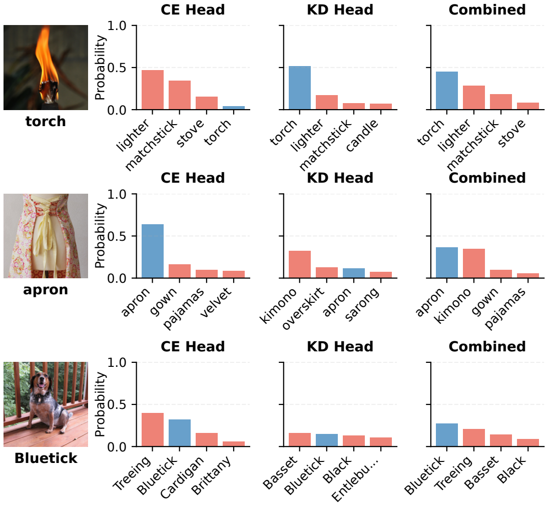

Figure 7: Qualitative results on challenging cases.

Enhanced feature representation of DHO. To validate our claim that mitigating gradient conflicts improves feature representations, we evaluate features using the standard *linear evaluation* protocol (Chen et al., 2020). We train CE+KD (feature), CE+KD (logit), and **DHO** under the 16-shot semi-supervised setting on ImageNet, freeze the feature extractor g, and train a new prediction linear head h_{LE} on the top of g using fully labeled data. As shown in Tab. 5, **DHO** achieves higher Top-1 and Top-5 accuracy than other methods (*e.g.*, +0.9% and 0.5%, respectively). To further assess feature quality, we visualize embeddings z using t-SNE (Van der Maaten & Hinton, 2008) in Fig. 8. Compared to the CE+KD (logit) baseline, **DHO** produces more compact and class-separated feature clusters. These results support our claim that **DHO enhances feature representations** by mitigating gradient conflicts; this improvement leads to better performance of **DHO** compared to SHO, as discussed through Tab. 1, Figs. 4, 5, 10 and 11.

Table 5: Linear evaluation results.

Method	Top-1 (%)	Top-5 (%)
CE+KD (feature) (Yang et al., 2024a)	62.3	85.0
CE+KD (logit) (Chen et al., 2020)	<u>66.2</u>	<u>88.8</u>
DHO	**67.1**	**89.3**

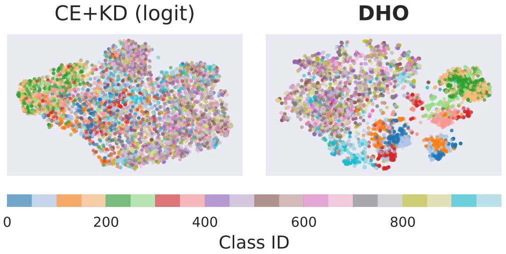

Figure 8: t-SNE visualization.

Effectiveness of Init. and Align in §2.3. To assess the effect of the proposed language-aware initialization and KD-head alignment in §2.3, we run ablations that apply language-aware initialization (**Init.**) to either h_{CE} or h_{KD}, and optionally enable KD-head alignment (**Align.**). We evaluate on ImageNet with 1% labels using a ViT-B/16 student and ViT-L/14 teacher for computational efficiency. As shown in Tab. 6, applying Init. to either head independently improves accuracy (*e.g.*, +0.2/+0.3%), and adding Align. yields additional gains (*e.g.*, +0.4%).

Table 6: Results on **language-aware initialization** and **KD-head alignment** for VLM students on ImageNet with 1% labeled data.

Init (h_{CE}/h_{KD}).	**Align.**	**Accuracy (%)**
✗ / ✗	✗	78.3
✓ / ✗	✗	78.5 (+0.2)
✓ / ✓	✗	78.6 (+0.3)
✓ / ✓	✓	**78.7** (+0.4)

Effect of α and β. In Theorem 1, we show that under mild assumptions (Assumption 1), **DHO** ϵ-approximates the SHO baseline, namely CE+KD (logit), in the ℓ_1 norm by setting α=λ and β=1. This enables **DHO** to emulate SHO hyperparameter tuning at inference time without retraining. To study how α and β affect performance, we visualize a grid search on ImageNet with ResNet-50 under the 16-shot setting. We use ImageNet because it lacks a validation set; thus α and β are heuristically set to 0.4 and 0.5 in §3.1. As shown in Fig. 9, with balanced heads ($\alpha \approx 0.5$), performance remains stable for $\beta \in [0.1, 1]$, and it degrades at extreme α values ($\alpha \approx 0$ or 1) regardless of β. The accuracy peaks at 65.5% with balanced heads ($\alpha \approx 0.5$) and a modest temperature ($\beta \approx 0.3$), close to the performance of our heuristic setting (64.7%). Importantly, **these gains require no additional training**, demonstrating the **efficiency of post-training hyperparameter search** for the proposed dual-head interpolation in §2.3.

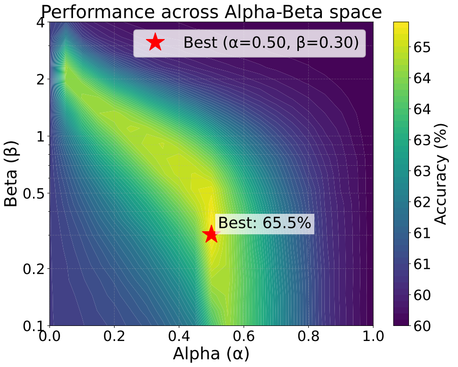

Figure 9: **Grid search results for α and β.**

4 Conclusion, Limitation, and Future Work

We identify the *fundamental challenge* of transferring the zero-/few-shot capabilities of large vision-language models (VLMs) to task-specific models using limited labeled data in semi-supervised settings. Conventional knowledge distillation (KD) methods suffer from *gradient conflicts* where teacher signals and labeled data signals interfere with each other. We propose **DHO** (**D**ual-**H**ead **O**ptimization), a *plug-and-play* framework using two classification heads with separate objectives. It **mitigates gradient conflicts** in both heads and feature extractor, **improving feature representations**, while **enabling flexible post-training hyperparameter adjustment** via linear combination of outputs. Experiments across 15 datasets show **DHO outperforms conventional KD methods**, achieving **state-of-the-art results on both** ImageNet semi-supervised learning with fewer parameters and on out-of-distribution tasks when combined with existing adaptation techniques.

Limitations and future work. While our main focus is to address the core problem of *gradient conflicts* arising from **general knowledge of VLMs and task-specific patterns from labeled data**, we acknowledge several limitations that present opportunities for future research.

We primarily focus on VLMs as general knowledge source, as they provide strong zero-shot and few-shot capabilities in visual recognition tasks. However, we believe the fundamental problem of gradient conflicts between general foundational knowledge and task-specific objectives extends beyond VLMs. This conflict likely emerges in various scenarios where large pre-trained foundation models (such as instruction-tuned language models (Wei et al., 2021; Liu et al., 2023a; Bai et al., 2023; Achiam et al., 2023; Team et al., 2023)) are adapted to specialized downstream tasks. Exploring how **DHO** performs across diverse foundation models and modalities remains an important direction for future work.

Our implementation is limited to visual recognition tasks, as they represent the most fundamental domain in computer vision and provide an ideal testbed for analyzing gradient conflicts in knowledge transfer. Also, VLMs' strong zero-shot and few-shot capabilities in visual recognition tasks make them natural candidates for knowledge distillation. However, extending our approach to more complex visual understanding tasks such as object detection and segmentation would be a promising direction with dedicated architectural adaptations.

Reproducibility Statement

We ensure reproducibility by conducting all experiments on publicly available datasets including ImageNet (Russakovsky et al., 2015), Caltech101 (Fei-Fei et al., 2004), and nine other standard benchmarks detailed in §D. All experimental configurations are fully specified in §3.1, including exact hyperparameters, learning rate and optimizer settings, model architectures (ResNet-18, ResNet-50, MobileNetV2, ViT-B/16, ViT-L/14), and training procedures with dual-head optimization detailed in Algorithm 1. We use publicly available pre-trained models (CLIP ResNet-50, CLIP ViT variants from OpenAI, DINO ResNet-50) with exact checkpoint specifications provided in Tab. 7. The inference procedure with hyperparameters α and β is fully documented in Algorithm 2, with specific values for each setting ($\alpha = 0.4$ for zero-shot, $\alpha = 0.2$ for few-shot teachers, $\beta = 0.5$ when validation unavailable). We commit to releasing our complete codebase, training scripts, pretrained checkpoints for ImageNet, and evaluation protocols upon acceptance. All experiments are conducted using PyTorch with fixed random seeds on NVIDIA RTX 4090 GPUs (4× for ImageNet, 8× for VLM distillation, single GPU for other benchmarks).

Ethics Statement

Our work presents no new ethical concerns as **DHO** is a purely technical contribution for knowledge distillation using existing publicly available datasets (ImageNet, Caltech101, and standard computer vision benchmarks) that contain no personally identifiable information. No additional data collection, human subjects research, or sensitive information processing is involved in this work. We acknowledge that vision-language models may contain biases from their pre-training data, which our distillation framework preserves without amplification. The computational requirements vary by dataset scale (single GPU for most benchmarks, 4× GPUs for ImageNet, 8× GPUs for VLM distillation), which remains modest compared to training large vision-language models from scratch, promoting research accessibility while minimizing environmental impact.

References

Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Compress: Self-supervised learning by compressing representations. *Advances in Neural Information Processing Systems*, 33: 12980–12992, 2020.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Armand Joulin, Nicolas Ballas, and Michael Rabbat. Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8443–8452, 2021.

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, pp. 456–473. Springer, 2022.

Jimmy Lei Ba. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pp. 446–461. Springer, 2014.

Zhaowei Cai, Avinash Ravichandran, Paolo Favaro, Manchen Wang, Davide Modolo, Rahul Bhotika, Zhuowen Tu, and Stefano Soatto. Semi-supervised vision transformers at scale. *Advances in Neural Information Processing Systems*, 35:25697–25710, 2022.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023.

Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3514–3522, 2019.

Jie Chen and Meng Joo Er. Mitigating gradient conflicts via expert squads in multi-task learning. *Neurocomputing*, 614:128832, 2025.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

Yifan Chen, Xiaozhen Qiao, Zhe Sun, and Xuelong Li. Comkd-clip: Comprehensive knowledge distillation for contrastive language-image pre-traning model. *arXiv preprint arXiv:2408.04145*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794–4802, 2019.

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

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pp. 7480–7512. PMLR, 2023.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10995–11005, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Pan Du, Suyun Zhao, Zisen Sheng, Cuiping Li, and Hong Chen. Semi-supervised learning via weight-aware distillation under class distribution mismatch. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16410–16420, 2023.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021a.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1428–1438, 2021b.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021c.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, Rongrong Ji, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *Advances in neural information processing systems*, 35:35959–35970, 2022.

Qingpei Guo, Furong Xu, Hanxiao Zhang, Wang Ren, Ziping Ma, Lin Ju, Jian Wang, Jingdong Chen, and Ming Yang. M^2-encoder: Advancing Bilingual Image-Text Understanding by Large-scale Efficient Pretraining. *arXiv preprint arXiv:2401.15896*, 2024.

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

Lingxiao He, Wu Liu, Jian Liang, Kecheng Zheng, Xingyu Liao, Peng Cheng, and Tao Mei. Semi-supervised domain generalizable person re-identification. *arXiv preprint arXiv:2108.05045*, 2021.

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

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727, 2022.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23773–23782, 2024.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023a.

Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15190–15200, 2023b.

Gyeongman Kim, Doohyuk Jang, and Eunho Yang. Promptkd: Distilling student-friendly knowledge for generative language models via prompt tuning. *arXiv preprint arXiv:2402.12842*, 2024.

Akisato Kimura, Zoubin Ghahramani, Koh Takeuchi, Tomoharu Iwata, and Naonori Ueda. Few-shot learning of neural networks from scratch by pseudo example optimization. *arXiv preprint arXiv:1802.03039*, 2018.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pp. 264–282. Springer, 2025.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Advances in Neural Information Processing Systems*, 35: 9287–9301, 2022.

Xin-Chun Li, Wen-Shu Fan, Bowen Tao, Le Gan, and De-Chuan Zhan. Exploring dark knowledge under various teacher capacities and addressing capacity mismatch. *arXiv preprint arXiv:2405.13078*, 2024.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in Neural Information Processing Systems*, 34:18878–18890, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15148–15158, 2023b.

He Liu, Yikai Wang, Huaping Liu, Fuchun Sun, and Anbang Yao. Small scale data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6008–6016, 2024.

Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. *arXiv preprint arXiv:1710.07535*, 2017.

Jiaming Lv, Haoyuan Yang, and Peihua Li. Wasserstein distance rivals kullback-leibler divergence for knowledge distillation. *Advances in Neural Information Processing Systems*, 37:65445–65475, 2024.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. *Advances in Neural Information Processing Systems*, 36:60984–61007, 2023.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 5191–5198, 2020.

Marco Mistretta, Alberto Baldrati, Marco Bertini, and Andrew D Bagdanov. Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation. In *European Conference on Computer Vision*, pp. 459–477. Springer, 2025.

K L Navaneet, Soroush Abbasi Koohpayegani, Ajinkya Tejankar, and Hamed Pirsiavash. Simreg: Regression as a simple yet effective tool for self-supervised knowledge distillation. In *British Machine Vision Conference (BMVC)*, 2021.

Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *International Conference on Machine Learning*, pp. 4743–4751. PMLR, 2019.

Dang Nguyen, Sunil Gupta, Kien Do, and Svetha Venkatesh. Black-box few-shot knowledge distillation. In *European Conference on Computer Vision*, pp. 196–211. Springer, 2022.

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

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

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

Gaurav Patel, Konda Reddy Mopuri, and Qiang Qiu. Learning to retain while acquiring: Combating distribution-shift in adversarial data-free knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7786–7794, 2023.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023.

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Jay C Rothenberger and Dimitrios I Diochnos. Meta co-training: Two views are better than one. *arXiv preprint arXiv:2311.18083*, 2023.

Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23681–23690, 2024.

Aditya Singh and Haohan Wang. Simple unsupervised knowledge distillation with space similarity. In *European Conference on Computer Vision*, pp. 147–164. Springer, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

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

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023a.

Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024.

Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, and Xide Xia. Dime-fm: Distilling multimodal and efficient foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15521–15533, 2023b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Minh-Tuan Tran, Trung Le, Xuan-May Le, Jianfei Cai, Mehrtash Harandi, and Dinh Phung. Large-scale data-free knowledge distillation for imagenet via multi-resolution data generation. *arXiv preprint arXiv:2411.17046*, 2024.

Vishaal Udandarao, Nikhil Parthasarathy, Muhammad Ferjad Naeem, Talfan Evans, Samuel Albanie, Federico Tombari, Yongqin Xian, Alessio Tonioni, and Olivier J Hénaff. Active data curation effectively distills large-scale multimodal models. *arXiv preprint arXiv:2411.18674*, 2024.

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

Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15963–15974, 2024.

Raviteja Vemulapalli, Hadi Pouransari, Fartash Faghri, Sachin Mehta, Mehrdad Farajtabar, Mohammad Rastegari, and Oncel Tuzel. Knowledge transfer from vision foundation models for efficient training of small task-specific models. In *International Conference on Machine Learning (ICML)*, 2024.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pp. 10506–10518, 2019.

Kai Wang, Fei Yang, and Joost van de Weijer. Attention distillation: self-supervised vision transformer students need more guidance. *arXiv preprint arXiv:2210.00944*, 2022a.

Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20(4):447–482, 2023.

Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14647–14657, 2022b.

Yuzheng Wang, Dingkang Yang, Zhaoyu Chen, Yang Liu, Siao Liu, Wenqiang Zhang, Lihua Zhang, and Lizhe Qi. De-confounded data-free knowledge distillation for handling distribution shifts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12615–12625, 2024.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Ge Wu, Xin Zhang, Zheng Li, Zhaowei Chen, Jiajun Liang, Jian Yang, and Xiang Li. Cascade prompt learning for vision-language model adaptation. In *European Conference on Computer Vision*, pp. 304–321. Springer, 2024.

Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21970–21980, 2023.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Haohang Xu, Jiemin Fang, Xiaopeng Zhang, Lingxi Xie, Xinggang Wang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Bag of instances aggregation boosts self-supervised distillation. *arXiv preprint arXiv:2107.01691*, 2021.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. Clip-kd: An empirical study of clip model distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15952–15962, 2024a.

Jing Yang, Xiatian Zhu, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Knowledge distillation meets open-set semi-supervised learning. *International Journal of Computer Vision*, pp. 1–20, 2024b.

Kaicheng Yang, Tiancheng Gu, Xiang An, Haiqiang Jiang, Xiangzi Dai, Ziyong Feng, Weidong Cai, and Jiankang Deng. Clip-cid: Efficient clip distillation via cluster-instance discrimination. *arXiv preprint arXiv:2408.09441*, 2024c.

Penghui Yang, Chen-Chen Zong, Sheng-Jun Huang, Lei Feng, and Bo An. Dual-head knowledge distillation: Enhancing logits utilization with an auxiliary head. *arXiv preprint arXiv:2411.08937*, 2024d.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8715–8724, 2020.

Jaemin Yoo, Minyong Cho, Taebum Kim, and U Kang. Knowledge extraction with no observable data. *Advances in Neural Information Processing Systems*, 32, 2019.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

Shikang Yu, Jiachen Chen, Hu Han, and Shuqiang Jiang. Data-free knowledge distillation via feature exchange and activation region constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24266–24275, 2023a.

Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10899–10909, 2023b.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020a.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020b.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12924–12933, 2024a.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Jialu Liu, Michael Bendersky, Marc Najork, and Chao Zhang. Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. *arXiv preprint arXiv:2305.05010*, 2023.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11953–11962, 2022.

Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 2024.

Mingkai Zheng, Shan You, Lang Huang, Chen Luo, Fei Wang, Chen Qian, and Chang Xu. Simmatchv2: Semi-supervised learning with graph consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16432–16442, 2023.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15659–15669, 2023.

Yichen Zhu and Yi Wang. Student customized knowledge distillation: Bridging the gap between student and teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5057–5066, 2021.

Appendix Overview

This appendix provides supplementary material to support the main paper and is organized as follows:

- **Related Work** (§A) discusses previous work relevant to ours, such as vision-language pre-training, data-limited adaptation of VLMs, knowledge Distillation (KD), and dual head approaches.
- **Theoretical Analysis** (§B): provides mathematical foundations and theoretical guarantees for our approach.
- **Algorithms and Implementation** (§C): presents detailed pseudocode (§C.1), implementation specifics (§C.2), and computational overhead analysis (§C.3).
- **Datasets** (§D): describes the datasets used in our experiments, including statistics and preprocessing details.
- **Additional Experiments** (§E): presents MobileNet experiments (§E.3), additional results with KD methods (§E.5), additional results with gradient surgery methods (§E.4), additional results with adaptive weighting (§E.6), and results of out-of-distribution generalization with fully-trained models (§E.11).
- **Additional Analyses** (§F): contains non-linear head design studies (§F.1), and further dual-head investigations (§F.2).

A Related Work

Vision-language pre-training. The emergence of vision-language pre-training has marked a significant breakthrough, enabling the use of extensive image-text pairs collected from the web (Wang et al., 2023; Chen et al., 2023) to train powerful vision encoders transferable to various vision tasks (Gan et al., 2022; Zhang et al., 2024b). Early works such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) leveraged contrastive learning techniques to align images and text into a joint representation space, facilitating zero-shot transfer via language prompts. Building on these foundations, subsequent research has focused on improving vision-language models through enhanced training methodologies (Dong et al., 2023; Gao et al., 2022; Yu et al., 2022; Zhai et al., 2023), as well as scaling models and datasets (Yu et al., 2022; Li et al., 2023; Dehghani et al., 2023; Sun et al., 2023a; Cherti et al., 2023; Fang et al., 2023; Sun et al., 2024; Guo et al., 2024) with their zero-shot transfer capabilities (Jia et al., 2021; Zhai et al., 2022; Pham et al., 2023; Liu et al., 2023b). In contrast, our work focuses specifically on target tasks with compact models, aiming to distill knowledge from these large VLMs effectively.

Data-limited adaptation of VLMs. To preserve pretrained semantic features of VLMs during adaptation with limited data, several approaches have been proposed. Prompt tuning (Lester et al., 2021), initially designed for language models, has been successfully extended to vision tasks. Various methods (Jia et al., 2022; Zhou et al., 2022b;a; Khattak et al., 2023a; Zhu et al., 2023; Khattak et al., 2023b; Menghini et al., 2023; Zhao et al., 2024; Roy & Etemad, 2023; Zhang et al., 2024a; Lafon et al., 2025) have demonstrated the effectiveness of training learnable prompts while keeping the base model frozen. Adapters (Gao et al., 2024; Zhang et al., 2021; Yu et al., 2023b; Silva-Rodriguez et al., 2024) provide an alternative approach by introducing lightweight, trainable modules while maintaining the pre-trained backbone intact. LP++ (Huang et al., 2024) has shown that simple linear layers can effectively adapt CLIP representations in data-limited settings. Note that our work is orthogonal to these approaches: we aim to distill the knowledge of pretrained VLMs into compact models under data-scarce scenarios, making these adaptation methods complementary and applicable to both teacher VLMs in our framework and student models when they are also VLMs.

Knowledge Distillation (**KD**; Hinton, 2015) enables transferring knowledge from large teacher models to compact student architectures, particularly in data-constrained settings. Researchers have explored synthetic data generation (Lopes et al., 2017; Kimura et al., 2018; Nayak et al., 2019; Yoo et al., 2019; Chen et al., 2019; Yin et al., 2020; Fang et al., 2021a; Nguyen et al., 2022; Patel et al., 2023; Yu et al., 2023a; Liu et al., 2024; Tran et al., 2024; Wang et al., 2024), semi-supervised (Chen et al., 2020; He et al., 2021; Du et al., 2023; Yang et al., 2024b), and unsupervised KD using self-supervised teachers (Fang et al., 2021c; Abbasi Koohpayegani et al., 2020; Navaneet et al., 2021; Wang et al., 2022a; Xu et al., 2021; Singh & Wang, 2025). In the VLM domain, recent works (Fang et al., 2021b; Wu et al., 2023; Sun et al., 2023b; Yang et al., 2024a; Vasu et al., 2024; Udandarao et al., 2024; Yang et al., 2024c) distill from large-scale vision-language models into smaller architectures, often using transductive (Kim et al., 2024; Chen et al., 2024) or multi-stage unsupervised strategies (Vemulapalli et al., 2024; Wu et al., 2024; Mistretta et al., 2025). Meanwhile, KD remains challenging due to numerous issues, including model capacity gaps (Cho & Hariharan, 2019; Mirzadeh et al., 2020; Zhu & Wang, 2021; Huang et al., 2022; Li et al., 2024) and inconsistencies between soft and hard targets (Zhang et al., 2023). These challenges are further complicated by misalignment between labeled data and foundational knowledge, especially in few-shot learning scenarios where limited labeled examples may not fully capture the rich semantic understanding of foundation models.

Dual-head approaches. We also consider existing dual-head KD approaches. **SSKD** (He et al., 2021) trains separate heads for labeled and unlabeled data, assuming different data distributions, while **DHKD** (Yang et al., 2024d) introduces a binary KD loss to alleviate neural collapse (Papyan et al., 2020). While these previous KD methods adopt dual-head architectures, they do not target distillation from foundation models or combine predictions at inference. Furthermore, both methods *infer using only the single supervised head* h_{CE} and do not address gradient conflicts arising from zero-/few-shot VLM teachers, whose prediction distributions can significantly differ from the limited labeled data. In contrast, **DHO** explicitly mitigates such gradient conflicts (Fig. 3) by training h_{KD} on both labeled and unlabeled data, and further provides *dual-head interpolation at inference time*, enabling flexible aggregation of supervised and teacher signals. As shown in Theorem 3 and Fig. 9, this interpolation can emulate tuning KD hyperparameters *without retraining*, giving **DHO** both improved performance and minimal hyperparameter tuning cost.

B Theoretical Analysis

In this section, we provide a theoretical analysis of our **D**ual-**H**ead **O**ptimization (**DHO**) framework. We establish that **DHO** effectively addresses single-head logit distillation Hinton (2015); Chen et al. (2020) by decoupling conflicting gradients through specialized heads during training. We prove that post-training, the optimal prediction from our dual-head model—formulated as a weighted combination of the heads' outputs—is mathematically equivalent to the optimal solution of conventional single-head distillation. This equivalence provides theoretical justification for our approach while eliminating gradient conflicts. Furthermore, **DHO** enables efficient adaptation to various datasets through tunable hyperparameters (α and β) without requiring model retraining. Note that in this section we slightly abuse the notation of the main paper for clarity, *e.g.*, we denote p_τ as teacher predictions with temperature scaling τ.

B.1 Single-Head Optimization

We begin by considering two target probability distributions: the ground truth label distribution y and the teacher's softened distribution p_τ for input $x \in \mathcal{X}$, where:

- y represents the ground truth label distribution, typically one-hot encoded vectors where $y_c = 1$ for the true class c and 0 elsewhere
- p_τ denotes the teacher's softened distribution with temperature scaling: $p_\tau = \sigma(z_t/\tau)$, where z_t represents the teacher's logits and σ is the softmax function

Theorem 2 (Optimal Distribution for Single-Head Optimization). *The distribution $\hat{p}^*$ that minimizes the weighted combination of cross-entropy loss with respect to y and Kullback-Leibler divergence with respect to p_τ:*

$$\mathcal{L}(\hat{p}) = \lambda \ell(\hat{p}, y) + (1 - \lambda) D_{\text{KL}}(p_\tau \| \hat{p}) \tag{11}$$

is given by the weighted arithmetic mean:

$$\hat{p}^* = \lambda y + (1 - \lambda) p_\tau \tag{12}$$

where $\lambda \in [0, 1]$ is the weighting hyperparameter.

Proof. We begin by expanding the objective function:

$$\begin{aligned}
\mathcal{L}(\hat{p}) &= \lambda \ell(\hat{p}, y) + (1-\lambda) D_{\mathrm{KL}}(p_\tau \| \hat{p}) && (13)\\
&= -\lambda \sum_{c=1}^{C} y_c \log \hat{p}_c + (1-\lambda) \sum_{c=1}^{C} p_{\tau,c} \log \frac{p_{\tau,c}}{\hat{p}_c} && (14)\\
&= -\lambda \sum_{c=1}^{C} y_c \log \hat{p}_c + (1-\lambda) \sum_{c=1}^{C} p_{\tau,c} \log p_{\tau,c} - (1-\lambda) \sum_{c=1}^{C} p_{\tau,c} \log \hat{p}_c && (15)\\
&= -\sum_{c=1}^{C} [\lambda y_c + (1-\lambda) p_{\tau,c}] \log \hat{p}_c + (1-\lambda) \sum_{c=1}^{C} p_{\tau,c} \log p_{\tau,c} && (16)
\end{aligned}$$

Since the last term is constant with respect to $\hat{p}$, the optimization problem reduces to minimizing:

$$\mathcal{L}'(\hat{p}) = -\sum_{c=1}^{C} [\lambda y_c + (1-\lambda) p_{\tau,c}] \log \hat{p}_c \tag{17}$$

Subject to the probability constraints:

$$\sum_{c=1}^{C} \hat{p}_c = 1, \quad \hat{p}_c \geq 0 \quad \forall c \in \{1, 2, \ldots, C\} \tag{18}$$

Applying the method of Lagrange multipliers with multiplier μ:

$$\mathcal{L}(\hat{p}, \mu) = -\sum_{c=1}^{C} [\lambda y_c + (1-\lambda) p_{\tau,c}] \log \hat{p}_c + \mu \left(\sum_{c=1}^{C} \hat{p}_c - 1 \right) \tag{19}$$

Taking the partial derivative with respect to $\hat{p}_c$ and setting it to zero:

$$-\frac{\lambda y_c + (1-\lambda) p_{\tau,c}}{\hat{p}_c} + \mu = 0 \tag{20}$$

Solving for $\hat{p}_c$:

$$\hat{p}_c = \frac{\lambda y_c + (1-\lambda) p_{\tau,c}}{\mu} \tag{21}$$

Using the constraint $\sum_{c=1}^{C} \hat{p}_c = 1$, and observing that $\sum_{c=1}^{C} y_c = 1$ and $\sum_{c=1}^{C} p_{\tau,c} = 1$ (both being probability distributions):

$$\begin{aligned}
\sum_{c=1}^{C} \hat{p}_c = \sum_{c=1}^{C} \frac{\lambda y_c + (1-\lambda) p_{\tau,c}}{\mu} &= 1 && (22)\\
\frac{1}{\mu} \sum_{c=1}^{C} [\lambda y_c + (1-\lambda) p_{\tau,c}] &= 1 && (23)\\
\frac{1}{\mu} [\lambda \sum_{c=1}^{C} y_c + (1-\lambda) \sum_{c=1}^{C} p_{\tau,c}] &= 1 && (24)\\
\frac{1}{\mu} [\lambda + (1-\lambda)] &= 1 && (25)\\
\mu &= 1 && (26)
\end{aligned}$$

Therefore, the optimal solution is:

$$\hat{p}_c^* = \lambda y_c + (1-\lambda) p_{\tau,c} \tag{27}$$

This weighted arithmetic mean of the two target distributions is the optimal solution that minimizes our objective function. □

B.2 DUAL-HEAD OPTIMIZATION

In our proposed Dual-Head Optimization (**DHO**) framework, we extract shared features $g(x)$ from input x and apply two specialized classification heads:

- $h_{\text{CE}}(z) = W_{\text{CE}}z + b_{\text{CE}}$: optimized exclusively to match ground truth labels using cross-entropy loss $\ell(\sigma(h_{\text{CE}}(z)), y)$
- $h_{\text{KD}}(z) = W_{\text{KD}}z + b_{\text{KD}}$: optimized exclusively to match teacher predictions using KL divergence $D_{\text{KL}}(p_\tau \| \sigma(h_{\text{KD}}(z)/\beta))$

where $z = g(x)$ is the feature representation, and the parameter β controls the temperature during inference, while a fixed temperature of 1 is used during training of the knowledge distillation head.

Assumption 2 (ε-Convergence)**.** We assume that after sufficient training, both heads have converged to their respective target distributions with bounded error:

$$\sup_x \|\sigma(h_{\text{CE}}(z)) - y\|_1 \le \varepsilon, \quad \sup_x \|\sigma(h_{\text{KD}}(z)/\beta) - p_\tau\|_1 \le \varepsilon \tag{28}$$

where $\|\cdot\|_1$ denotes the ℓ_1 norm and $\varepsilon > 0$ is a small constant.

Theorem 3 (Inference Equivalence Under ε-Convergence)**.** *Under Assumption 2, by combining the outputs of both heads as:*

$$\hat{p}_{\texttt{DHO}} = \alpha \cdot \sigma(h_{CE}(z)) + (1-\alpha) \cdot \sigma(h_{KD}(z)/\beta), \quad \textit{where } \alpha = \lambda \tag{29}$$

we obtain a prediction that approximates the optimal single-head solution with bounded error:

$$\|\hat{p}_{\texttt{DHO}} - \hat{p}^*\|_1 \le \varepsilon \tag{30}$$

Proof. We analyze the ℓ_1 distance between the **DHO** prediction and the optimal solution:

$$\begin{aligned}
\|\hat{p}_{\texttt{DHO}} - \hat{p}^*\|_1 &= \|\alpha \cdot \sigma(h_{\text{CE}}(z)) + (1-\alpha) \cdot \sigma(h_{\text{KD}}(z)/\beta) - \lambda y - (1-\lambda)p_\tau\|_1 && (31)\\
&= \|\lambda(\sigma(h_{\text{CE}}(z)) - y) + (1-\lambda)(\sigma(h_{\text{KD}}(z)/\beta) - p_\tau)\|_1 && (32)\\
&\le \lambda\|\sigma(h_{\text{CE}}(z)) - y\|_1 + (1-\lambda)\|\sigma(h_{\text{KD}}(z)/\beta) - p_\tau\|_1 && (33)\\
&\le \lambda\varepsilon + (1-\lambda)\varepsilon = \varepsilon && (34)
\end{aligned}$$

where we applied the triangle inequality for the ℓ_1 norm and used Assumption 2.

Therefore, we have established that:

$$\hat{p}_{\texttt{DHO}} \approx_\varepsilon \hat{p}^* \tag{35}$$

where $\approx_\varepsilon$ denotes approximation with ℓ_1 error bound ε. □

Lemma 1 (Temperature Matching via KL Divergence)**.** *Assume the knowledge distillation head is trained to minimize KL divergence with respect to the teacher's predictions at temperature 1, such that:*

$$D_{\text{KL}}(p_1 \| \sigma(h_{KD}(z))) \le \delta \tag{36}$$

Then, setting the temperature parameter $\beta = \tau$ at inference time allows the KD head to approximate the teacher's prediction at temperature τ with error bound:

$$\|\sigma(h_{KD}(z)/\beta) - p_\tau\|_1 \le \sqrt{2\delta} \tag{37}$$

Proof. When logits are properly scaled and under appropriate conditions of the softmax function, we can reasonably approximate:

$$D_{\text{KL}}(p_\tau \| \sigma(h_{\text{KD}}(z)/\tau)) \approx D_{\text{KL}}(p_1 \| \sigma(h_{\text{KD}}(z))) \le \delta \tag{38}$$

Applying Pinsker's inequality, which establishes a relationship between KL divergence and the L1 norm difference between probability distributions:

$$\|p_\tau - \sigma(h_{\text{KD}}(z)/\tau)\|_1 \le \sqrt{2D_{\text{KL}}(p_\tau \| \sigma(h_{\text{KD}}(z)/\tau))} \le \sqrt{2\delta} \tag{39}$$

To ensure ε-convergence between the KD head at temperature τ and the teacher's prediction at temperature τ, it is sufficient to guarantee:

$$\sqrt{2\delta} \leq \varepsilon \Rightarrow \delta \leq \frac{\varepsilon^2}{2} \tag{40}$$

□

Corollary 1 (Optimal **DHO** Configuration)**.** *With proper training ensuring ε-convergence of both heads, dual-head optimization with temperature parameter $\beta = \tau$ and mixing parameter $\alpha = \lambda$ approximates the optimal single-head objective with error bounded by ε:*

$$\hat{p}_{\textbf{DHO}} \approx_\varepsilon \hat{p}^* = \lambda y + (1 - \lambda) p_\tau \tag{41}$$

This demonstrates that our ***DHO*** *approach achieves the same theoretical optimality as SHO.*

C Algorithms and Implementation

C.1 Pseudocode

We present the pseudocode for **DHO** in Algs. 1 and 2 for training and inference, respectively.

Algorithm 1 **DHO** Training with zero-shot CLIP (Radford et al., 2021) teacher

```
Input: labeled set $\mathcal{D}^{(l)} = \{(x_i^{(l)}, y_i)\}_{i=1}^N$, unlabeled set $\mathcal{D}^{(u)} = \{x_j^{(u)}\}_{j=1}^M$,
       student feature extractor $g$, prediction heads $h_{\text{CE}}, h_{\text{KD}}$, teacher encoders $f_\mathcal{X}, f_\mathcal{T}$,
       prompt template "A photo of [CLASS]", temperature scaling factors $\zeta, \eta$,
       balancing hyperparameter $\lambda$,
       supervised mini-batch size $B$, and unsupervised mini-batch size $B'$.
while not converged do
    Sample mini-batch $\mathcal{B}^{(l)} = \{(x_b^{(l)}, y_b)\}_{b=1}^B$ from $\mathcal{D}^{(l)}$, $\mathcal{B}^{(u)} = \{x_{b'}^{(u)}\}_{b'=1}^{B'}$ from $\mathcal{D}^{(l)} \cup \mathcal{D}^{(u)}$.
    // Process labeled data
    for each $(x_b^{(l)}, y_b) \in \mathcal{B}^{(l)}$ do
        $z_b^{(l)} \leftarrow g(x_b^{(l)})$
        $\hat{p}_{\text{CE},b}^{(l)} \leftarrow \sigma(h_{\text{CE}}(z_b^{(l)}))$
        $\hat{p}_{\text{KD},b}^{(l)} \leftarrow \sigma(\frac{1}{\eta} h_{\text{KD}}(z_b^{(l)}))$
        $p_b^{(l)} \leftarrow \sigma\left(\frac{1}{\zeta \cdot \eta}[\texttt{CosSim}(f_\mathcal{X}(x_b^{(l)}), f_\mathcal{T}(t_1)), \ldots, \texttt{CosSim}(f_\mathcal{X}(x_b^{(l)}), f_\mathcal{T}(t_C))]^\top\right)$
    end for
    // Process unlabeled data
    for each $x_{b'}^{(u)} \in \mathcal{B}^{(u)}$ do
        $z_{b'}^{(u)} \leftarrow g(x_{b'}^{(u)})$
        $\hat{p}_{\text{KD},b'}^{(u)} \leftarrow \sigma(\frac{1}{\eta} h_{\text{KD}}(z_{b'}^{(u)}))$
        $p_{b'}^{(u)} \leftarrow \sigma\left(\frac{1}{\zeta \cdot \eta}[\texttt{CosSim}(f_\mathcal{X}(x_{b'}^{(u)}), f_\mathcal{T}(t_1)), \ldots, \texttt{CosSim}(f_\mathcal{X}(x_{b'}^{(u)}), f_\mathcal{T}(t_C))]^\top\right)$
    end for
    // Compute losses and update
    $\mathcal{L}_{\text{CE}} \leftarrow \frac{1}{B} \sum_{b=1}^B \ell(\hat{p}_{\text{CE},b}^{(l)}, y_b)$
    $\mathcal{L}_{\text{KD}} \leftarrow \frac{1}{B} \sum_{b=1}^B D_{\text{KL}}(\hat{p}_{\text{KD},b}^{(l)} || p_b^{(l)}) + \frac{1}{B'} \sum_{b'=1}^{B'} D_{\text{KL}}(\hat{p}_{\text{KD},b'}^{(u)} || p_{b'}^{(u)})$
    $\mathcal{L} \leftarrow \lambda \mathcal{L}_{\text{CE}} + (1 - \lambda) \mathcal{L}_{\text{KD}}$
    Update parameters of $g$, $h_{\text{CE}}$, $h_{\text{KD}}$ using $\nabla \mathcal{L}$
end while
```

Algorithm 2 Dual-Head Optimization Inference

```
Input: an image $x$, feature extractor $g$, prediction heads $h_{\text{CE}}$, $h_{\text{KD}}$, linear coefficient $\alpha$, temperature scaling $\beta$
$z \leftarrow g(x)$
$\hat{p}_{\text{CE}} \leftarrow \sigma(h_{\text{CE}}(z))$
$\hat{p}_{\text{KD}} \leftarrow \sigma(h_{\text{KD}}(z)/\beta)$
$\hat{p} \leftarrow \alpha \cdot \hat{p}_{\text{CE}} + (1 - \alpha) \cdot \hat{p}_{\text{KD}}$
$\hat{y} \leftarrow \arg\max_c(\hat{p}_c)$
Return: $\hat{y}$
```

C.2 Implementation Details

Architecture choices for teacher and student models. For zero-shot teachers, we use CLIP ResNet-50 (Radford et al., 2021) in few-shot settings and VIT-H/14 from DFN (Fang et al., 2023) in low-shot settings. For few-shot teachers, we adopt Tip-Adapter-F (Zhang et al., 2021), a learnable adapter model, and denote the corresponding variant as **DHO-F**. On ImageNet, to avoid label leakage, we either train ResNet-18 from scratch or use a self-supervised ResNet-50 from DINO (Caron et al., 2021). For other datasets, we use ResNet-18 and MobileNetV2 (Sandler et al., 2018) without such concerns. In low-shot settings, we use CLIP ViT-B/16 and ViT-L/14 (Radford et al., 2021).

Few-/low-shot settings. Tab. 7 provides a comprehensive overview of the implementation details for our experiments on 1) few-shot semi-supervised settings on ImageNet and 10 datasets, 2) low-shot semi-supervised settings on ImageNet, and 3) VLM-based adaptation methods.

OOD settings. Tab. 8 provides the implementation details for our out-of-distribution (OOD) generalization experiments on 1) full training model evaluation and 2) adaptation methods including linear evaluation, visual prompt tuning, and VLM-based methods.

Table 7: Implementation details for our experiments across different settings.

Few-shot Semi-supervised Settings on ImageNet	
Model Configuration	**Student Training Details**
• **Student:** ResNet18 (He et al., 2016) from scratch or ResNet50 from DINO (Caron et al., 2021) • **Input size:** 224×224 • **Zero-shot Teacher:** ResNet50 from CLIP (Radford et al., 2021) • **Few-shot Teacher:** ResNet50 from Tip-Adapter-F (Zhang et al., 2021) • **Teacher input size:** 224×224 • **labeled data:** $K \in \{1, 2, 4, 8, 16\}$ shots • **ζ, η, and λ:** $0.01, 2, 0.5$ • **α and β:** $\alpha = 0.4$, $\beta = 0.5$ (zero-shot); $\alpha = 0.2$, $\beta = 0.5$ (few-shot)	• **Epochs:** 20 • **Optimizer:** AdamW (β_1=0.9, β_2=0.999) • **Learning rate:** 1×10^{-3}, weight decay: 1×10^{-2} • **Batch size:** 512 (labeled: 256, unlabeled: 256) • **Scheduler:** Cosine decay without warmup • **Augmentation:** Random crops (x0.5-1.0), horizontal flips
Few-shot Semi-supervised Settings on 10 Fine-Grained Datasets	
Model Configuration	**Student Training Details**
• **Student:** ResNet18 (He et al., 2016) or MobileNet (Sandler et al., 2018) pre-trained on ImageNet under supervision • **Input size:** 224×224 • **Zero-shot Teacher:** ResNet50 from CLIP (Radford et al., 2021) • **Few-shot Teacher:** ResNet50 from Tip-Adapter-F (Zhang et al., 2021) • **Teacher input size:** 224×224 • **labeled data:** $K \in \{1, 2, 4, 8, 16\}$ shots • **ζ, η, and λ:** $0.01, 2, 0.5$ • **α and β:** determined by validation	• **Epochs:** 200 • **Optimizer:** AdamW (β_1=0.9, β_2=0.999) • **Learning rate:** 1×10^{-3}, weight decay: 1×10^{-2} • **Batch size:** 128 (labeled: 64, unlabeled: 64) • **Scheduler:** Cosine decay without warmup • **Augmentation:** Random crops (x0.5-1.0), horizontal flips
Low-shot Semi-supervised Settings on ImageNet	
Model Configuration	**Student Training Details**
• **Student:** CLIP ViT-B/16 or ViT-L/14 (Radford et al., 2021) • **Input size:** 224×224 (ViT-B/16) or 336×336 (ViT-L/14) • **Zero-shot Teacher:** CLIP ViT-L/14 or ViT-H/14 (Fang et al., 2023) • **Teacher input size:** 336×336 (ViT-L/14) or 378×378 (ViT-H/14) • **Few-shot Teacher:** N/A • **labeled data:** 1% ($\frac{N}{N+M} \approx$ 0.01) or 10% ($\frac{N}{N+M} \approx 0.1$) of training data • **$\zeta$, η, and λ:** $0.01, 2, 0.5$ • **α and β:** $\alpha = 0.5$, $\beta = 0.5$	• **Epochs:** 32 • **Optimizer:** AdamW (β_1=0.9, β_2=0.999) • **Learning rate:** 5×10^{-5}, weight decay: 5×10^{-2} • **Batch size:** 512 (labeled: 256, unlabeled: 256) • **Scheduler:** Cosine warmup decay (5000 steps) • **Augmentation:** Random crops (x0.5-1.0), horizontal flips

Table 8: Implementation details for our out-of-distribution generalization experiments.

Full Training	
Model Configuration	**Training Details**
• **Student:** CLIP ViT-B/16 (Radford et al., 2021) • **Student input size:** 224×224 • **Zero-shot Teacher:** CLIP ViT-L/14 (Radford et al., 2021) • **Teacher input size:** 336×336 • **Labeled data:** 1% and 10% ImageNet • ζ**,** η**, and** λ: $0.01, 2, 0.5$ • α **and** β**:** 0.5 and 1	• **Epochs:** 32 • **Optimizer:** AdamW (β_1=0.9, β_2=0.999) • **Learning rate:** 5×10^{-5}, weight decay: 5×10^{-2} • **Batch size:** 512 (labeled: 256, unlabeled: 256) • **Scheduler:** Cosine warmup decay (5000 steps) • **Augmentation:** Random crops (x0.5-1.0), horizontal flips
Adaptation Methods (Linear Evaluation & Visual Prompt Tuning)	
Method Configuration	**Training Details**
• **Linear evaluation** (Caron et al., 2021) • **Visual prompt tuning** (Jia et al., 2022) • **Frozen backbone:** CLIP ViT-B/16 (Radford et al., 2021) • **Input size:** 224×224 • **Zero-shot Teacher:** CLIP ViT-L/14 (Radford et al., 2021) • **Teacher input size:** 336×336 • **Labeled data:** 1% and 10% ImageNet • ζ**,** η**, and** λ: $0.01, 2, 0.5$ • α **and** β**:** 0.5 and 1	• **Epochs:** 20 • **Optimizer:** AdamW (β_1=0.9, β_2=0.999) • **Learning rate:** 5×10^{-5}, weight decay: 5×10^{-2} • **Batch size:** 512 (labeled: 256, unlabeled: 256) • **Scheduler:** Cosine warmup decay (5000 steps) • **Augmentation:** Random crops (x0.5-1.0), horizontal flips
Adaptation Methods (Prompt Tuning)	
Method Configuration	**Training Details**
• **Prompt tuning:** CoOp (Zhou et al., 2022b), PromptSRC (Khattak et al., 2023b) • **Frozen backbone:** CLIP ViT-B/16 (Radford et al., 2021) • **Input size:** 224×224 • **Zero-shot Teacher:** CLIP ViT-L/14 (Radford et al., 2021) • **Teacher input size:** 336×336 • **Labeled data:** 1% and 10% ImageNet • ζ**,** η**, and** λ: $0.01, 2, 0.5$ • α **and** β**:** 0.5 and 1	• **Prompt tuning:** Following PromptSRC (Khattak et al., 2023b) configurations • **Comparison:** CasPL (Wu et al., 2024) with domain-specific unlabeled data

C.3 Computational Costs

Inference overhead of DHO over SHO. Tab. 9 presents computational overheads at inference time introduced by DHO over SHO for all the architectures in this paper, such as MobileNetV2 (Sandler et al., 2018), ResNet-18 (He et al., 2016), ResNet-50 (He et al., 2016), ViT-B/16 (Dosovitskiy, 2020), and ViT-L/16 (Dosovitskiy, 2020).

Table 9: Inference overhead using RTX 4090 across different architecture.

Model	**Params (M)**	**FLOPs (G)**	**Throughput (im/s)**
MobileNetV2	3.50	0.33	2978.4
+ **DHO**	4.79 (+36.5%)	0.34 (+3.0%)	2971.2 (-0.24%)
ResNet-18	11.69	1.83	3525.7
+ **DHO**	12.20 (+4.4%)	1.83 (+0.0%)	3518.6 (-0.20%)
ResNet-50	25.56	4.14	1018.4
+ **DHO**	27.61 (+8.0%)	4.15 (+0.2%)	1016.4 (-0.19%)
ViT-B/16	86.57	16.87	290.2
+ **DHO**	87.34 (+0.9%)	16.87 (+0.0%)	290.1 (-0.02%)
ViT-L/16	304.33	59.70	255.1
+ **DHO**	305.35 (+0.3%)	59.70 (+0.0%)	255.6 (+0.18%)

Training time and hardware requirements. Tab. 10 presents the training time required for our experiments. For VLM distillation experiments, which represent the most resource-intensive component of our work, we used 8× NVIDIA RTX 4090 GPUs. The ViT-H/14 to ViT-L/14 distillation required approximately 80 hours, while the ViT-H/14 to ViT-B/16 and ViT-L/14 to ViT-B/16 distillations required approximately 40 and 28 hours, respectively. For the ViT-H/14 to ViT-L/14 distillation, we implemented gradient accumulation with 4 steps and mixed precision training (Micikevicius et al., 2017) to optimize computational efficiency. For ImageNet experiments, we used 4× NVIDIA RTX 4090 GPUs, with ResNet-18 and ResNet-50 models requiring approximately 6 and 8 hours of training time, respectively. We provide these details to facilitate reproduction of our results and to give researchers a clear understanding of the computational resources needed to implement our approach at scale.

Table 10: Training time and hardware.

Student	**Teacher**	**Training Time**	**Hardware**
ResNet-18	ResNet-50	≈ 6 hours	4× RTX 4090
ResNet-50	ResNet-50	≈ 8 hours	4× RTX 4090
ViT-B/16	ViT-L/14	≈ 28 hours	8× RTX 4090
ViT-B/16	ViT-H/14	≈ 40 hours	8× RTX 4090
ViT-L/14	ViT-H/14	≈ 80 hours	8× RTX 4090

Inference overhead improvements with ToMe. To further improve the computational efficiency of our approach, we explored integrating Token Merging (ToMe) (Bolya et al., 2022) with DHO. ToMe is a technique that reduces the number of tokens in ViTs by merging similar tokens to improve the efficiency of ViTs. Tab. 11 shows that combining DHO with ToMe significantly reduces computational costs with minimal impact on performance.

Table 11: Performance and inference overhead of DHO with Token Merging (ToMe) on ImageNet under low-shot semi-supervised settings using RTX 4090.

Method	**Labeled**	**Accuracy (%)**	**Params (M)**	**FLOPs (G)**	**Throughput (im/s)**
DHO	1%	81.6	87.22	17.58	243.35
DHO + ToMe	1%	81.4 (–0.2)	87.22	13.12 (–25.4%)	323.39 (+32.9%)
DHO	10%	82.8	87.22	17.58	238.11
DHO + ToMe	10%	82.5 (–0.3)	87.22	13.12 (–25.4%)	308.49 (+29.6%)

D DATASETS

Table 12: Overview of datasets used in our experiments, organized into three categories: **(top)** standard classification datasets, **(middle)** ImageNet, and **(bottom)** ImageNet variants for out-of-distribution (OOD) evaluation. For few-shot semi-supervised learning experiments, we report both the absolute number of labeled samples and their percentage relative to the full training set.

Dataset	# Classes	# Train	# Val	# Test	# Labeled (1-shot)	# Labeled (16-shot)
Fine-grained 10 Datasets						
Caltech101 (Fei-Fei et al., 2004)	100	4,128	1,649	2,465	100 (2.42%)	1,600 (38.76%)
OxfordPets (Parkhi et al., 2012)	37	2,944	736	3,669	37 (1.26%)	592 (20.11%)
StanfordCars (Krause et al., 2013)	196	6,509	1,635	8,041	196 (3.01%)	3,136 (48.18%)
Flowers102 (Nilsback & Zisserman, 2008)	102	4,093	1,633	2,463	102 (2.49%)	1,632 (39.87%)
Food101 (Bossard et al., 2014)	101	50,500	20,200	30,300	101 (0.20%)	1,616 (3.20%)
FGVCAircraft (Maji et al., 2013)	100	3,334	3,333	3,333	100 (3.00%)	1,600 (48.00%)
SUN397 (Xiao et al., 2010)	397	15,880	3,970	19,850	397 (2.50%)	6,352 (40.00%)
DTD (Cimpoi et al., 2014)	47	2,820	1,128	1,692	47 (1.67%)	752 (26.67%)
EuroSAT (Helber et al., 2019)	10	13,500	5,400	8,100	10 (0.07%)	160 (1.19%)
UCF101 (Soomro, 2012)	101	7,639	1,898	3,783	101 (1.32%)	1,616 (21.15%)
Coarse-grained Dataset						
ImageNet (Russakovsky et al., 2015)	1,000	1.28M	-	50,000	1,000 (0.08%)	16,000 (1.25%)
ImageNet OOD Variants						
ImageNet-V2 (Recht et al., 2019)	1,000	-	-	10,000	-	-
ImageNet-Sketch (Wang et al., 2019)	1,000	-	-	50,889	-	-
ImageNet-A (Hendrycks et al., 2021b)	200	-	-	7,500	-	-
ImageNet-R (Hendrycks et al., 2021a)	200	-	-	30,000	-	-

We evaluated our approach on 11 diverse datasets, with ImageNet (Russakovsky et al., 2015) serving as our primary benchmark. The datasets span general object recognition (Russakovsky et al., 2015; Fei-Fei et al., 2004), fine-grained classification tasks (vehicles (Krause et al., 2013; Maji et al., 2013), natural entities (Nilsback & Zisserman, 2008; Parkhi et al., 2012; Bossard et al., 2014)), and specialized domains (scenes (Xiao et al., 2010), textures (Cimpoi et al., 2014), remote sensing (Helber et al., 2019), and human actions (Soomro, 2012)). Additionally, we conduct experiments on four out-of-distribution test sets to further validate the model's generalization capabilities. To assess our model's robustness to distribution shifts, we evaluate it on several challenging variants of ImageNet: ImageNet-v2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a).

We summarize the overview of datasets used in Tab. 12, these datasets exhibit diversity in their characteristics, with varying numbers of classes and samples per dataset. This diversity enabled us to thoroughly validate our method across different few-shot semi-supervised learning scenarios by systematically varying the ratios between labeled and unlabeled samples.

E ADDITIONAL EXPERIMENTS

E.1 EXPERIMENTS ON MOTIVATION

We observe that existing KD methods, such as logit distillation (Hinton, 2015), face gradient conflict in **few-/low-shot semi-supervised settings**. To investigate whether gradient conflicts are severe or only present in few-/low-shot semi-supervised settings, we conduct additional experiments by using full labels on ImageNet and 8 datasets using ResNet-18/ResNet-50 as student/teacher. First of all, we observe **no gradient conflicts at all in any fully supervised settings**, *i.e.*, $\texttt{CosSim}(\nabla_\theta \mathcal{L}_{\text{CE}}, \nabla_\theta \mathcal{L}_{\text{KD}}) > 0$ throughout training. This is because the distillation signal from a fully supervised teacher aligns well with the labeled data when the teacher is trained on the same dataset as the student. However, in few-/low-shot semi-supervised settings, this alignment often breaks down, as shown in Fig. 3. Tab. 13 also reports results on ImageNet using a ResNet-18 student and a ResNet-50 teacher. With fully labeled data, the gap is significantly reduced (0.4), while **DHO** outperforms CE+KD (logit) in few-shot settings, validating that **DHO** is specifically tailored to few-/low-shot semi-supervised settings.

Table 13: Results on ImageNet using a ResNet-18 student and a ResNet-50 teacher. 100% denotes using the fully labeled dataset with a fully supervised teacher.

Method	1-shot	2-shot	4-shot	8-shot	16-shot	100%
CE+KD (logit)	50.5	50.6	50.6	51.0	51.2	66.3
DHO	51.8	52.4	52.6	53.3	54.5	66.7

E.2 Experiments on Feature Quality

In this section, we provide more insights into the impact of gradient conflict on **feature learning**. To measure feature quality, we introduce the following definition:

Definition 1 (Feature quality)**.** *Let $Z \in \mathbb{R}^{N\times d}$ be the feature matrix of X (i.e., $Z = [g(x_1), \ldots, g(x_N)]^\top$) and $Y \in \mathbb{R}^{N\times C}$ the one-hot encoded target matrix. For $\lambda \geq 0$, we define the **feature quality** of Z as the expected risk (↓) of ridge regression:*

$$\mathcal{L} = \mathbb{E}_{z,y}\left[\|y - z^\top \hat{W}_\lambda\|_2^2\right], \quad \text{where } \hat{W}_\lambda = (Z^\top Z + \lambda N I_d)^{-1} Z^\top Y.$$

We adopt the expected risk of ridge regression (↓) as the measure of feature quality for simplicity (while linear probing is widely used and we also use it in Tab. 5).

Let $g_1 := \nabla_\theta \mathcal{L}_{\text{CE}}(\theta)$, $g_2 := \nabla_\theta \mathcal{L}_{\text{KD}}(\theta)$, $g := \nabla_\theta \mathcal{L}(\theta)$, and $G = \frac{1}{2}g_1 + \frac{1}{2}g_2$. We assume:

Assumption 3 (Gradient cone)**.**

$$g = \beta_1 g_1 + \beta_2 g_2 + r_\perp, \qquad \beta_1, \beta_2 \geq 0, \qquad r_\perp^\top g_1 = r_\perp^\top g_2 = 0.$$

This means that the probe gradient lies within the non-negative cone spanned by g_1 and g_2, up to an orthogonal residual $r_\perp$—that is, both g_1 and g_2 **contribute positively to improving feature quality**.

Taking a vanilla SGD step ($\theta - \eta G$, $\eta > 0$), with first-order Taylor expansion:

$$\mathcal{L}(\theta - \eta G) = \mathcal{L}(\theta) - \eta\, g^\top G + \mathcal{O}(\eta^2) \approx \mathcal{L}(\theta) - \eta g^\top G. \tag{42}$$

We neglect the $\mathcal{O}(\eta^2)$ term (*e.g.*, $\eta < 10^{-3}$ makes it negligible). Under the Assumption 3:

$$\begin{aligned} g^\top G &= \tfrac{1}{2}\, g^\top (g_1 + g_2) \\ &= \tfrac{1}{2}\,(\beta_1 g_1 + \beta_2 g_2 + r_\perp)^\top (g_1 + g_2) \\ &= \tfrac{1}{2}\big(\beta_1 \|g_1\|_2^2 + \beta_2 \|g_2\|_2^2\big) + \tfrac{1}{2}(\beta_1 + \beta_2)\, g_1^\top g_2. \end{aligned} \tag{43}$$

If gradients are aligned ($g_1^\top g_2 > 0$), then Eq. 43 is positive, accelerating the decrease in Eq. 42, and thus **the expected probe risk decreases more**.

To empirically validate this, we report $\hat{\mathcal{L}} := \frac{1}{M}\sum_{i=1}^{M} \|y_i - z_i^\top \hat{W}_\lambda\|_2^2$ on the test set over 200 epochs with the same setup as Fig. 4:

Table 14: **Feature quality** throughout training, computed using Definition 1.

10 Datasets (Average)	1	11	21	31	41	51	61	71	81	91	101	111	121	131	141	151	161	171	181	191
CE+KD (logit)	4.974	4.664	4.638	4.608	4.601	4.598	4.595	4.595	4.594	4.595	4.595	4.595	4.596	4.596	4.596	4.597	4.598	4.598	4.599	4.599
DHO	4.672	4.435	4.413	4.391	4.390	4.390	4.390	4.393	4.393	4.394	4.395	4.396	4.396	4.398	4.399	4.400	4.400	4.400	4.401	4.401

Tab. 14 shows that across all epochs, $\hat{\mathcal{L}}$ in Definition 1 of **DHO** is lower than that of CE+KD (logit), **confirming our analysis on feature quality**.

E.3 Experiments on MobileNet

To demonstrate the versatility of our **DHO** approach beyond ResNet (He et al., 2016) and ViT (Dosovitskiy, 2020) architectures, we extended our experiments to the MobileNetV2 (Sandler et al., 2018), which is specifically designed for real-world applications with compact models. We maintained identical experimental settings as described in §C.2, using MobileNetV2 as the student model while distilling from CLIP ResNet50.

As illustrated in Fig. 10, our **DHO** consistently outperforms all single-head baseline methods, demonstrating its effectiveness on lightweight model architectures along with ResNet18. Furthermore,

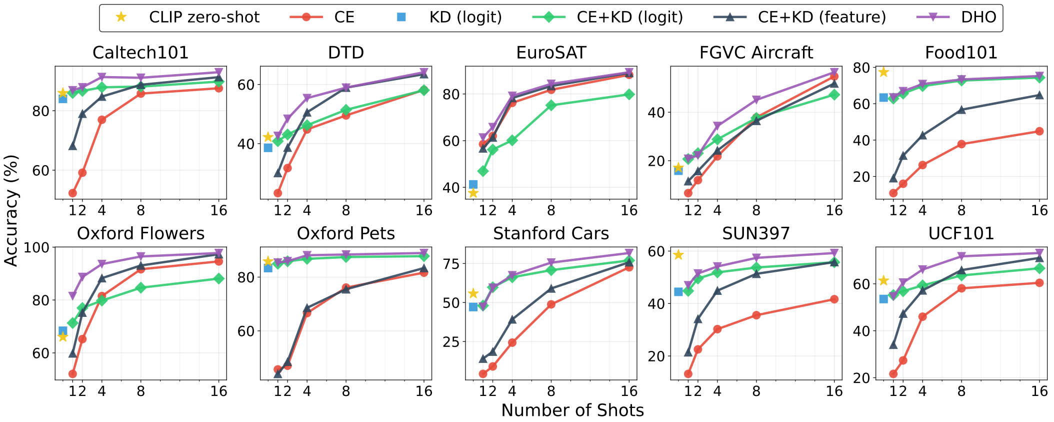

Figure 10: Results on **10 datasets** under few-shot semi-supervision using **MobileNetV2** with **zero-shot teacher** (Radford et al., 2021).

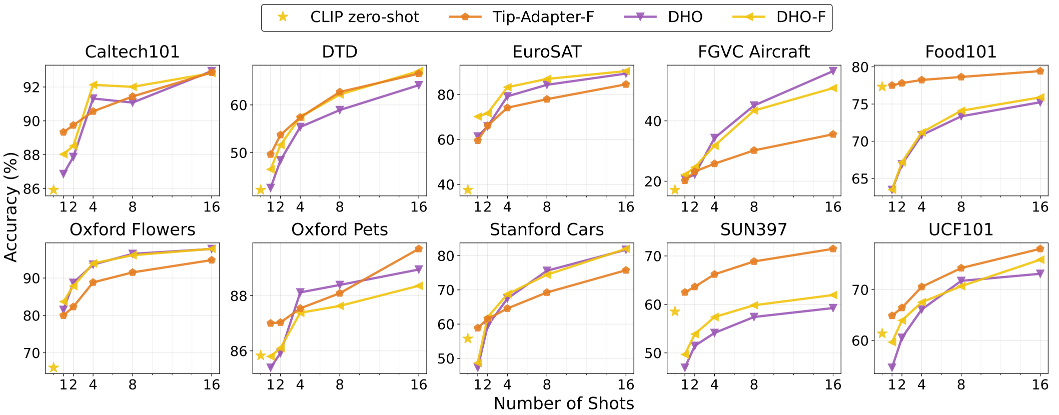

Figure 11: Results on **10 datasets** using **MobileNetV2** with either zero- or **few-shot teacher** (Zhang et al., 2021).

Fig. 11 reveals patterns similar to our ResNet18 experiments regarding few-shot integration. Our method successfully incorporates few-shot teacher knowledge, although we observe that the few-shot teacher does not consistently yield improvements over the zero-shot teacher. Notably, our distilled MobileNetV2 model sometimes achieves superior performance to the zero- and few-shot teachers (ResNet-50) despite having significantly fewer parameters. This pattern of outperforming both zero-shot and few-shot teachers mirrors the observations from our main experiments, further validating the effectiveness of our approach across different architectural families.

E.4 Additional Results with Gradient Surgery Methods

In this section, we present experimental results comparing **DHO** with PCGrad (Yu et al., 2020b), a well-known gradient surgery method from the multi-task learning literature. PCGrad addresses gradient conflicts by projecting conflicting gradients to resolve conflicts post hoc.

We compare **DHO** with PCGrad using 8 of the 10 datasets (without Food101, Sun397) with ResNet-18 as the student model. The results are shown in Tab. 15.

We observe that PCGrad improves over CE+KD (logit), but still underperforms compared to **DHO**. Moreover, PCGrad incurs additional memory and computational costs ($\mathcal{O}(|\theta|)$) due to gradient projection and storage, whereas **DHO** remains lightweight ($\mathcal{O}(d \times C)$) and simple to implement.

Table 15: Performance comparison of PCGrad with **DHO** on 8 datasets (average) using ResNet-18.

Method	1-shot	2-shot	4-shot	8-shot	16-shot
CE+KD (logit)	56.2	58.9	61.7	65.8	69.8
PCGrad	56.8	60.3	62.1	67.5	71.6
DHO	**60.1**	**63.7**	**70.2**	**75.4**	**79.1**

The superior performance of **DHO** can be attributed to its approach of avoiding gradient conflicts at the source by isolating the learning dynamics of each objective via dual heads, rather than resolving conflicts after they arise. Additionally, **DHO** provides dynamic interpolation capability at inference time between supervised and distillation signals, which standard gradient-based methods do not offer.

E.5 Additional Results with KD Methods

In this section, we present experimental results combining our **DHO** method with additional distillation approaches: Decoupled Knowledge Distillation (DKD) (Zhao et al., 2022) and Wasserstein Knowledge Distillation (WKD) (Lv et al., 2024). Both DKD and WKD are orthogonal to **DHO** since their losses can be applied directly to the outputs of h_{CE}. DKD decouples the target class from non-target classes, and WKD computes a kernel matrix within each class. Both approaches require ground-truth labels, restricting their use to the small labeled dataset $\mathcal{D}^{(l)}$.

Despite these limitations, we conducted experiments on ImageNet (Tab. 16) and on 8 of the 10 datasets (without Food101, Sun397) as shown in Tab. 17. The results demonstrate that **DHO** substantially improves the performance of these state-of-the-art KD methods. Notably, WKD+**DHO** consistently outperforms **DHO** alone in most settings, demonstrating the extensibility of **DHO** due to its simplicity.

Table 16: Performance comparison of DKD and WKD with and without **DHO** on ImageNet. Numbers in parentheses show improvement over the base method.

Method	1-shot	2-shot	4-shot	8-shot	16-shot
DHO	51.8	52.4	52.6	53.3	54.5
DKD	8.9	15.0	20.6	28.2	34.9
DKD+**DHO**	47.1 (+38.2)	44.3 (+29.3)	40.9 (+20.3)	40.2 (+12.0)	42.4 (+7.5)
WKD	11.0	17.0	17.0	27.7	34.6
WKD+**DHO**	**53.2** (+42.2)	**53.3** (+36.3)	**53.3** (+36.3)	**54.0** (+26.3)	**54.8** (+20.2)

Table 17: Performance comparison of DKD and WKD with and without **DHO** on 8 datasets (average). Numbers in parentheses show improvement over the base method.

Method	1-shot	2-shot	4-shot	8-shot	16-shot
DHO	**60.1**	63.7	70.2	75.4	79.1
DKD	28.7	41.3	55.4	66.3	73.1
DKD+**DHO**	46.2 (+17.5)	51.2 (+9.9)	61.8 (+6.4)	69.8 (+3.5)	74.9 (+1.8)
WKD	30.0	38.2	53.3	66.1	73.6
WKD+**DHO**	59.6 (+29.6)	**64.5** (+26.3)	**71.2** (+17.9)	**75.7** (+9.6)	**79.6** (+6.0)

E.6 Additional Results with Adaptive Weighting

To further explore the potential of **DHO**, we implement an entropy-based adaptive weighting mechanism. Let the entropy of a probability vector $p \in \Delta^{C-1}$ be $H(p) = -\sum_{c=1}^{C} p_c \log p_c$. We compute the adaptive weight α as:

$$\alpha = \frac{\exp(-H(\hat{p}_{\text{CE}}))}{\exp(-H(\hat{p}_{\text{CE}})) + \exp(-H(\hat{p}_{\text{KD}}))} \tag{44}$$

where $\hat{p}_{\text{CE}}$ and $\hat{p}_{\text{KD}}$ are the output probability vectors from h_{CE} and h_{KD}, respectively. The final prediction is then computed as $\hat{p} = \alpha \cdot \hat{p}_{\text{CE}} + (1 - \alpha) \cdot \hat{p}_{\text{KD}}$.

The intuition behind this approach is that lower entropy (higher confidence) predictions should receive higher weights in the final ensemble. When one head produces more confident predictions than the other, the adaptive weighting mechanism automatically emphasizes the more certain prediction. The results show that the entropy-based adaptive weighting method is not proved to be effective in these experiments. While the adaptive weighting does not consistently outperform the fixed interpolation approach, this is likely due to modern neural networks producing overconfident predictions, making entropy an unreliable proxy for uncertainty without proper calibration. However, we believe adaptive weighting could be beneficial with well-calibrated models or alternative uncertainty measures.

E.7 Sensitivity of Single-head KD to Hyperparameters

To investigate the sensitivity of single-head KD baseline, *i.e.*, CE + KD (logit), we conduct additional experiments by varying the loss balancing hyperparameter $\lambda \in \{0.0, 0.1, \ldots, 0.9, 1.0\}$ and the temperature hyperparameter $\tau \in \{0.1, 0.3, 0.5, 1, 2, 4\}$, resulting in total 55 training runs. Fig. 12 shows the interpolated results from these 55 experiments on ImageNet using a ResNet-50 student and a CLIP ResNet-50 teacher, following the same experimental setup in Fig. 9. We observe that the single-head KD baseline is also sensitive to the hyperparameters (λ, τ), requiring intensive training to obtain the best combination. Crucially, the best performance of CE + KD (logit) is 63.0%, which is less than the performance of **DHO** (64.7%) with heuristically selected (α, β).

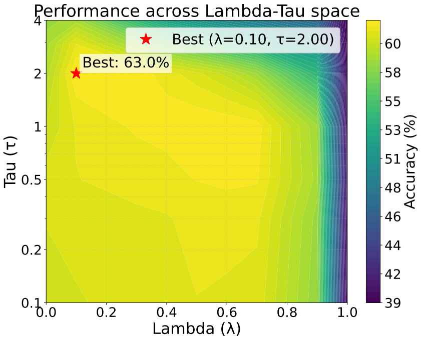

Figure 12: Grid search results for λ and τ for **CE + KD (logit)**.

E.8 Comparison under Few-shot Teacher

Table 18: Results on 10 datasets (average) using ResNet-18 with either zero- or **few-shot teachers**.

Method	0-shot	1-shot	2-shot	4-shot	8-shot	16-shot
CE + KD (logit)	–	55.6	58.2	61.1	64.8	68.4
CE + KD (logit)-F	–	59.8	63.9	68.9	72.6	76.2
DHO	–	58.9	62.2	68.4	73.1	76.5
DHO-F	–	61.1	64.7	69.5	73.6	76.9
Zero-shot teacher	58.8	–	–	–	–	–
Few-shot teacher	–	65.0	67.2	70.4	73.3	76.9

In Tab. 1 and Fig. 5, we observe the improvement of **DHO** under few-shot teachers, *i.e.*, **DHO-F**. In this section, we also compare **DHO-F** against single-head KD baselines under few-shot teachers, *i.e.*, CE + KD (logit)-F. Specifically, we train and evaluate CE+KD (logit)-F under the same experimental setup as Fig. 5, and summarize the results in Tab. 18. We observe that the baseline improves with the few-shot teacher; however, **DHO-F** still outperforms the baseline even under the same few-shot teacher setting.

E.9 Comparison to Simple 2-Stage Baseline

Table 19: Comparison with **simple two-stage baselines** with/without **FixMatch** (Sohn et al., 2020) on 10 datasets (average) using ResNet-18 with a zero-shot teacher.

Method	FixMatch	0-shot	1-shot	2-shot	4-shot	8-shot	16-shot
KD	-	53.9	-	-	-	-	-
CE + KD (logit)	-	-	55.6	58.2	61.1	64.8	68.4
2-stage	-	-	53.0	53.6	58.5	63.7	68.6
2-stage	✓	-	53.5	54.6	59.8	66.7	70.5
DHO	-	-	58.9	62.2	68.4	73.1	76.5
Zero-shot teacher	-	58.8	-	-	-	-	-

One possible way to mitigate gradient conflict in few-/low-shot settings with VLM teachers is to train a single-head classifier in two distinct stages. Specifically, we can (1) train a single-head classifier using only the KD loss $\mathcal{L}_{\text{KD}}$ in Eq. 2 on both the labeled dataset $\mathcal{D}^{(l)}$ and the unlabeled dataset $\mathcal{D}^{(u)}$, and then (2) fine-tune the classifier using only the CE loss $\mathcal{L}_{\text{CE}}$ in Eq. 1 on the labeled dataset $\mathcal{D}^{(l)}$. To investigate the effectiveness of **DHO** over this simple 2-stage baseline, we conduct additional experiments using the exact setup employed in Fig. 4 and Tab. 7 on the same 10 datasets, with the only exception that we train each stage with 100 epochs to make up a total 200 epochs for fair comparison. We also consider using a representative semi-supervised learning method, *i.e.*, **FixMatch** (Sohn et al., 2020), in the second stage to prevent overfitting. For FixMatch, we follow the original hyperparameters $\tau = 0.95$ and $\lambda_u = 1$, using RandAugment and Cutout as strong augmentations. The only exception is $\mu = 7$, which we set to 1 (*i.e.*, a 1:1 ratio of labeled to unlabeled samples per mini-batch), matching the setup of DHO.

As shown in Tab. 19, the two-stage baseline performs similarly to CE + KD (logit) when a comparatively larger number of labeled samples is available (*e.g.*, 8-/16-shot), but it significantly degrades when labeled data is scarce (*e.g.*, 1-/2-/4-shot). This result is intuitive: during the second stage of fine-tuning on the labeled data only, as the amount of labeled data decreases, the classifier becomes more prone to overfitting. As expected, applying FixMatch in the second stage improves performance by preventing overfitting. However, DHO consistently achieves the best performance, even compared to this two-stage baseline. This indicates that DHO is more efficient in practice, since it avoids splitting training into multiple stages and does not require additional fine-tuning cycles, thereby reducing engineering complexity (*e.g.*, tuning τ, λ_u, μ, and strong augmentations of FixMatch).

E.10 DHO IS ORTHOGORNAL TO SSL METHODS

Due to the simplicity of DHO, we observe that it can be seamlessly integrated with, and improves upon, recent category-aware KD methods (Zhao et al., 2022; Lv et al., 2024), as shown in §E.5. In this section, we further demonstrate that DHO is also orthogonal to semi-supervised learning (SSL) methods such as DebiasedPL (Wang et al., 2022b), which leverages **debiased pseudo labels** to train a classifier. Specifically, we train h_{CE} using DebiasedPL, while h_{KD} is trained with the KD loss of DHO, *i.e.*, $\mathcal{L}_{\text{KD}}$ in Eq. 6.

We conduct experiments on ImageNet with 1% labels on ImageNet using ResNet-50 with a zero-shot VLM teacher. We use a batch size of 384, an EMA teacher model, multi-view augmentation, and 50 training epochs. Because the Google Drive link in the official codebase for the 1% ImageNet split has expired, we randomly sample a new 1% split, which may introduce minor discrepancies relative to the numbers reported in the original paper. As shown in Tab. 20, DHO improves DebiasedPL by **+1.4**, demonstrating that DHO is both simple and complementary to strong SSL approaches.

Table 20: Comparison with DebiasedPL (Wang et al., 2022b).

Method	ImageNet 1%
DebiasedPL	63.9
DebiasedPL + DHO	65.3 **(+1.4)**
Zero-shot VLM teacher	60.3

Table 21: Performance of ViT-B/16 and ViT-L/14 distilled from ViT-H/14 with entropy adaptive weighting under different percentages of labeled data.

Method	ViT-B/16		ViT-L/14	
	1%	10%	1%	10%
DHO	**81.66**	**82.78**	84.59	**85.94**
DHO+Ent	**81.66**	82.65	**84.60**	85.92

E.11 OUT-OF-DISTRIBUTION EVALUATION UPON FULLY TRAINED MODEL

We provide the evaluation on out-of-distribution datasets with fully trained model in Tab. 22. DHO significantly outperformed zero-shot baselines on similar-distribution variants (ImageNet-V2, ImageNet-Sketch) across both ViT-B/16 and ViT-L/14 architectures, but showed performance degradation on out-of-distribution datasets (ImageNet-R, ImageNet-A), suggesting increased distribution overfitting from full model training. Interestingly, ViT-B/16 models distilled from ViT-L/14 handled shifted distributions better than those taught by the larger ViT-H/14 DFN (Fang et al., 2023), despite the latter's superior performance on shifted distributions such as ImageNet-R and ImageNet-A. We attribute this to the shared training background between ViT-B/16 and ViT-L/14 in the CLIP framework (Radford et al., 2021), which appears to better preserve generalization capabilities during the adaptation. This points to an important insight: **our method works best on out-of-distributions when the teacher and student models share similar training distributions**, suggesting that successful knowledge distillation also depends on the alignment between teacher and student than just the teacher's raw capabilities.

Table 22: Accuracy(%) of **DHO** with full training model on the ImageNet distribution-shifted variants.

Student Model	**Params (M)**	**Labeled Data**	**Teacher Model**	**Val**	**V2**	**Sketch**	**R**	**A**
ViT-B/16 Student								
ViT-B/16 (Radford et al., 2021)	86M	zero-shot	-	66.7	60.8	46.2	**74.0**	**47.0**
ViT-B/16	86M	1%	ViT-L/14	78.7	70.1	48.0	70.9	41.1
ViT-B/16	86M	10%	ViT-L/14	80.8	71.3	47.4	71.7	41.4
ViT-B/16	86M	1%	ViT-H/14	81.6	72.6	50.6	65.5	35.6
ViT-B/16	86M	10%	ViT-H/14	**82.8**	**73.6**	**50.7**	67.7	37.8
ViT-L/14 Student								
ViT-L/14 (Radford et al., 2021)	304M	zero-shot	-	75.3	68.3	59.2	**86.5**	**74.6**
ViT-L/14	304M	1%	ViT-H/14	84.6	77.0	61.5	79.9	60.8
ViT-L/14	304M	10%	ViT-H/14	**85.9**	**77.8**	**61.7**	82.8	64.4
Zero-shot VLM								
ViT-H (Fang et al., 2023)	632M	zero-shot	-	83.6	77.2	71.7	92.3	77.4

F Additional Analysis

F.1 Additional Analysis on Non-Linear Head Design

To further investigate the architectural advantages of dual head optimization, we conducted experiments with non-linear head designs, replacing the linear heads used in our main experiments. We design a non-linear classifier with a sequence of layers: an initial linear projection layer, followed by layer normalization (Ba, 2016), GELU activation (Hendrycks & Gimpel, 2016), dropout (Srivastava et al., 2014), and a final linear classification layer. We compared **DHO** with three non-linear configurations; **DHO+NL-Head-CE**: non-linear CE head, **DHO+NL-Head-KD**: non-linear KD head, and **DHO+NL-Head-CE+KD**: non-linear both CE and KD heads. All experiments followed the few-shot semi-supervised setting detailed in §C.2.

Table 23: Results of different **non-linear head configurations** on **11 datasets** including **ImageNet** under few-shot semi-supervision using **ResNet-18** with **zero-shot teacher** (Radford et al., 2021). We report averaged accuracy for 10 visual recognition datasets except the ImageNet.

	ImageNet					**Average of 10 tasks**				
Configuration	1-shot	2-shot	4-shot	8-shot	16-shot	1-shot	2-shot	4-shot	8-shot	16-shot
DHO (base)	61.7	62.2	62.6	63.8	65.1	58.9	**62.2**	68.4	**73.1**	**76.5**
DHO+NL-Head-CE	61.7	61.9	62.2	63.1	64.8	**59.3**	62.1	68.0	72.7	76.2
DHO+NL-Head-KD	**62.1**	**62.6**	**62.9**	**64.0**	**65.9**	58.3	62.1	67.8	72.4	**76.5**
DHO+NL-Head-CE+KD	62.0	62.3	62.6	63.8	65.4	59.1	**62.2**	**68.6**	72.4	**76.5**

Performance Analysis. Our experiments revealed key findings regarding head architecture (Tab. 23, Fig. 13). On ImageNet, non-linear KD heads consistently outperformed linear ones, suggesting complex architectures better capture teacher predictions. Conversely, non-linear CE heads degraded performance, likely due to overfitting on limited labeled data. While dual non-linear heads outperformed fully linear configurations, they were less effective than non-linearity in the KD head alone.

Table 24: Results on **dual-heads interpolation strategy** of different **non-linear head configurations** on **ImageNet** under 16-shots semi-supervised setting.

Configuration	**CE Head**	**KD Head**	**Combined**
DHO (base)	60.64	61.55	65.37
DHO+NL-Head-CE	60.18	61.39	64.91
DHO+NL-Head-KD	60.95	61.76	**65.97**
DHO+NL-Head-CE+KD	**61.66**	**61.81**	65.59

On the other 10 datasets, optimal configurations varied considerably with no consistently superior approach. This highlights that non-linear transformation effectiveness depends strongly on dataset characteristics and head functionality. Given comparable performance but superior computational efficiency, we adopted linear head architectures for all subsequent experiments.

Head Decomposition Analysis. Analysis under the 16-shot semi-supervised setting revealed complex relationships between architectural choices and head-wise performance as shown in Tab. 24. Non-linear CE branches decreased CE head performance from 60.64% to 60.18% despite increased parameters. Conversely, non-linear KD heads improved both heads: CE accuracy increased to 60.95%

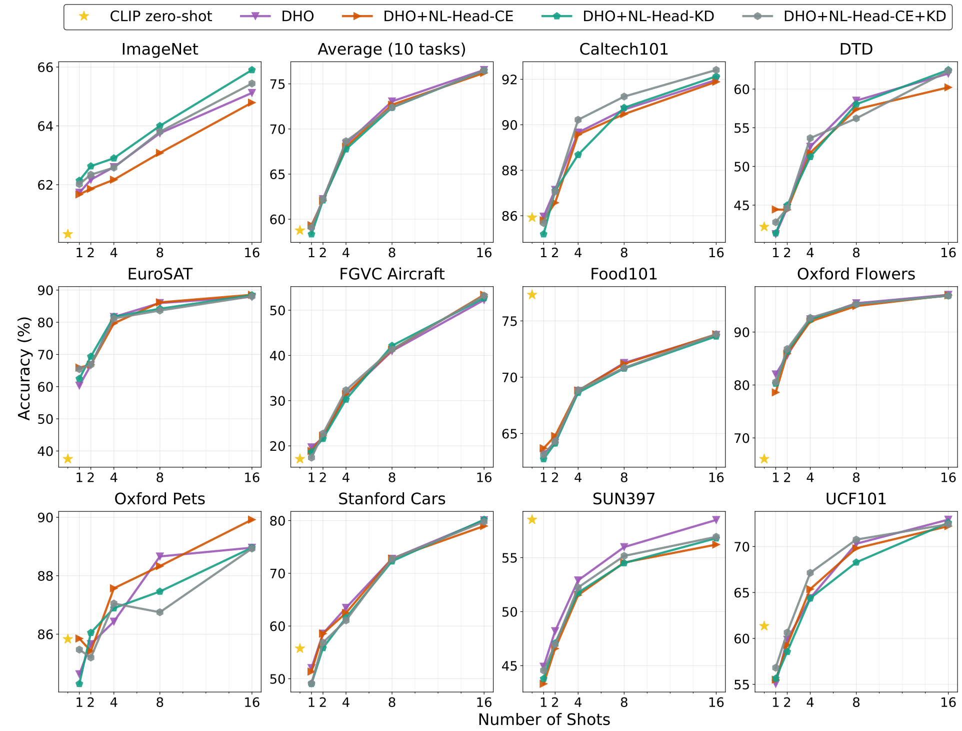

Figure 13: Results of different **non-linear head configurations** on **11 datasets** including **ImageNet** under few-shot semi-supervision using **ResNet-18** with **zero-shot teacher** (Radford et al., 2021).

and KD prediction to 61.76%. However, dual non-linear heads reduced combined performance from 65.97% to 65.59%, suggesting head specialization may compromise joint feature representation. These findings highlight the complex interplay between architectural decisions and multi-head learning dynamics.

F.2 Additional Dual-Head Analysis

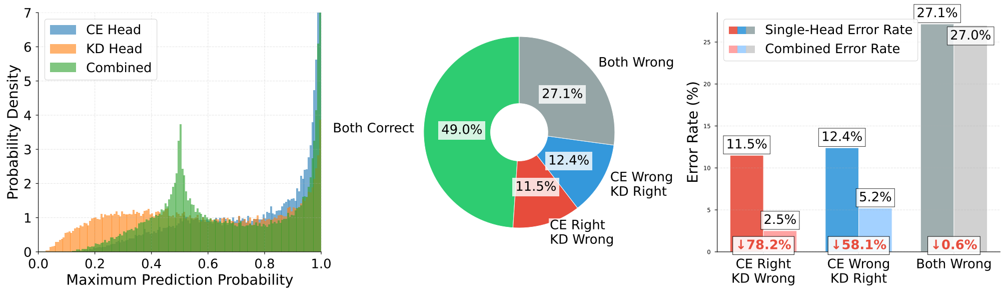

Figure 14: Analysis of **DHO** on the ImageNet under 16-shot semi-supervised setting. **(Left)** Maximum probability distributions for predictions from CE head, KD head, and their combined output. **(Middle)** Prediction agreement diagram analysis, categorizing cases where both heads are correct, only one head is correct, and both heads are incorrect. **(Right)** Error reduction analysis comparing single-head failure cases against improvements achieved through combined predictions.

In this section, we further analyze the prediction behavior of **DHO**. As shown in Fig. 14 (left), despite sharing feature representations, the CE head (h_{CE}), trained on labeled data, produces sharper predictions, whereas the KD head (h_{KD}), guided by teacher distillation, generates smoother distributions. Prediction agreement analysis (Fig. 14, middle) shows that the two heads agree in 76.2% of cases while complementing each other: the CE head correctly classifies 11.5% of cases where the KD head fails, and vice versa for 12.4%. Error reduction analysis (Fig. 14, right) further demonstrates that our combined approach reduces failure rates from 11.5% to 2.5% for the KD head and from 12.4% to 5.2% for the CE head, confirming the effectiveness of **DHO**.

We also present additional qualitative results of **DHO**, both on ImageNet (see Figs. 15 and 16) and on other 10 datasets (see Figs. 17 and 18).

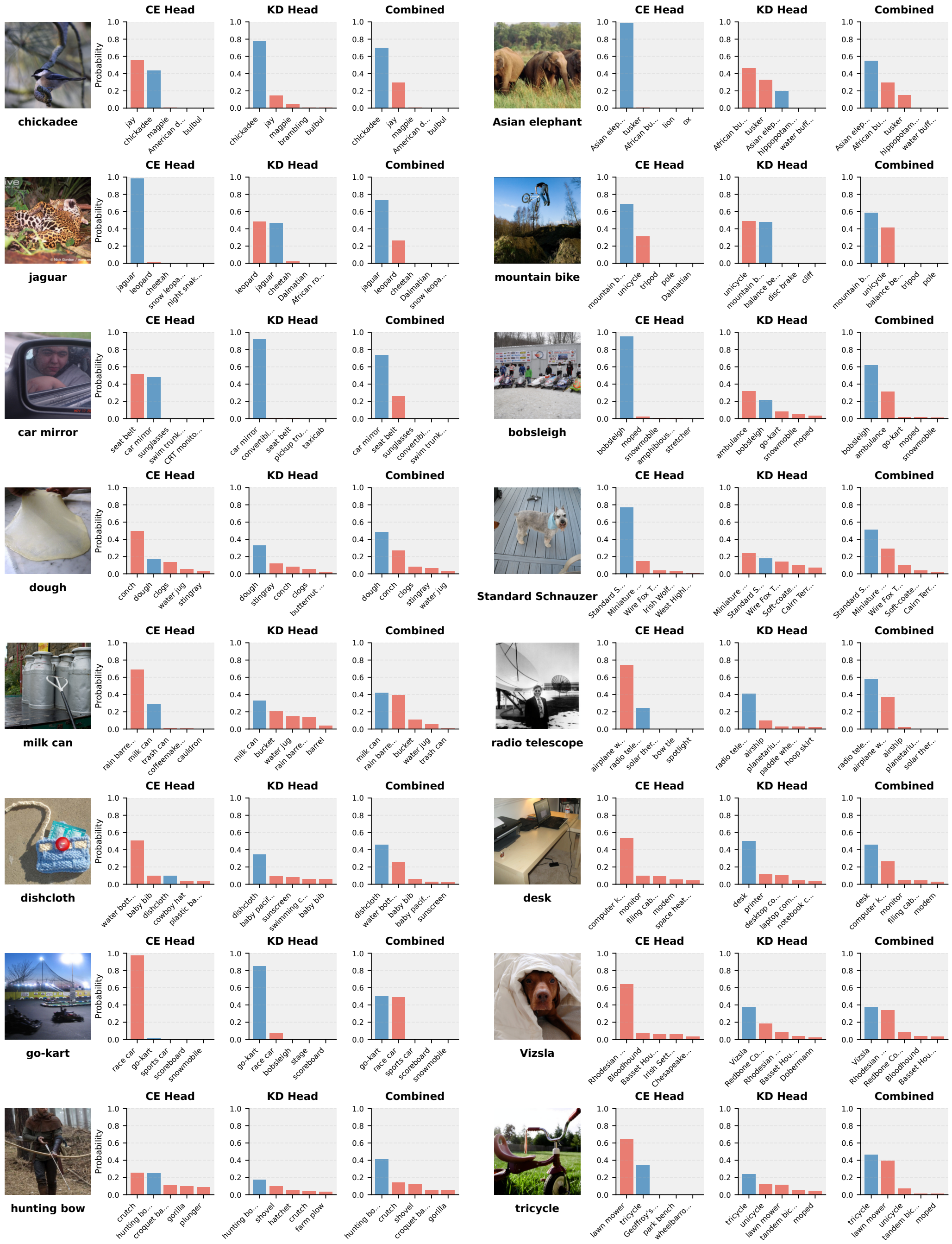

Figure 15: Additional qualitative results on ImageNet under 16-shot semi-supervised setting.

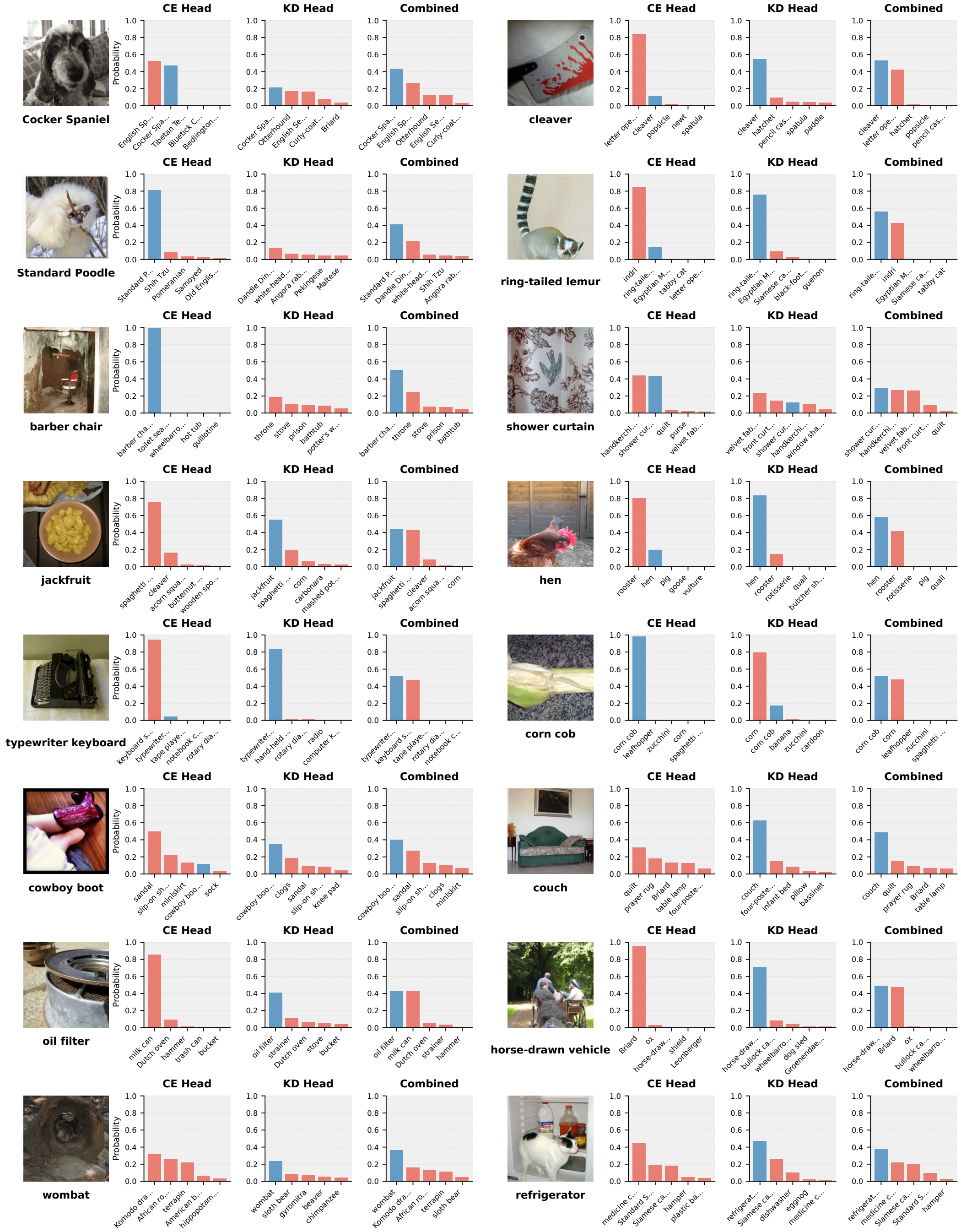

Figure 16: Additional qualitative results on ImageNet under 16-shot semi-supervised setting.

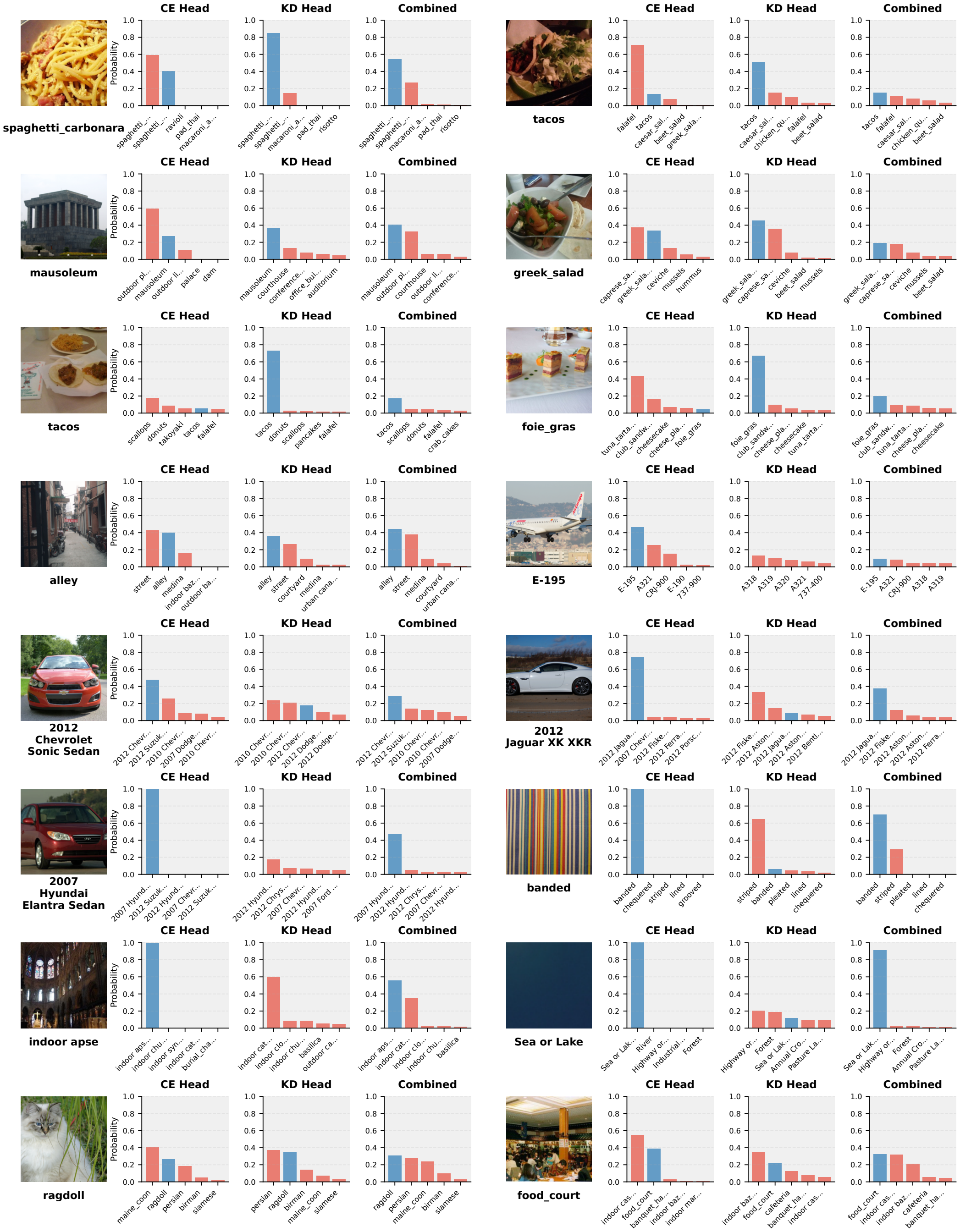

Figure 17: Additional qualitative results on other 10 datasets for models trained under 16-shot semi-supervised setting.

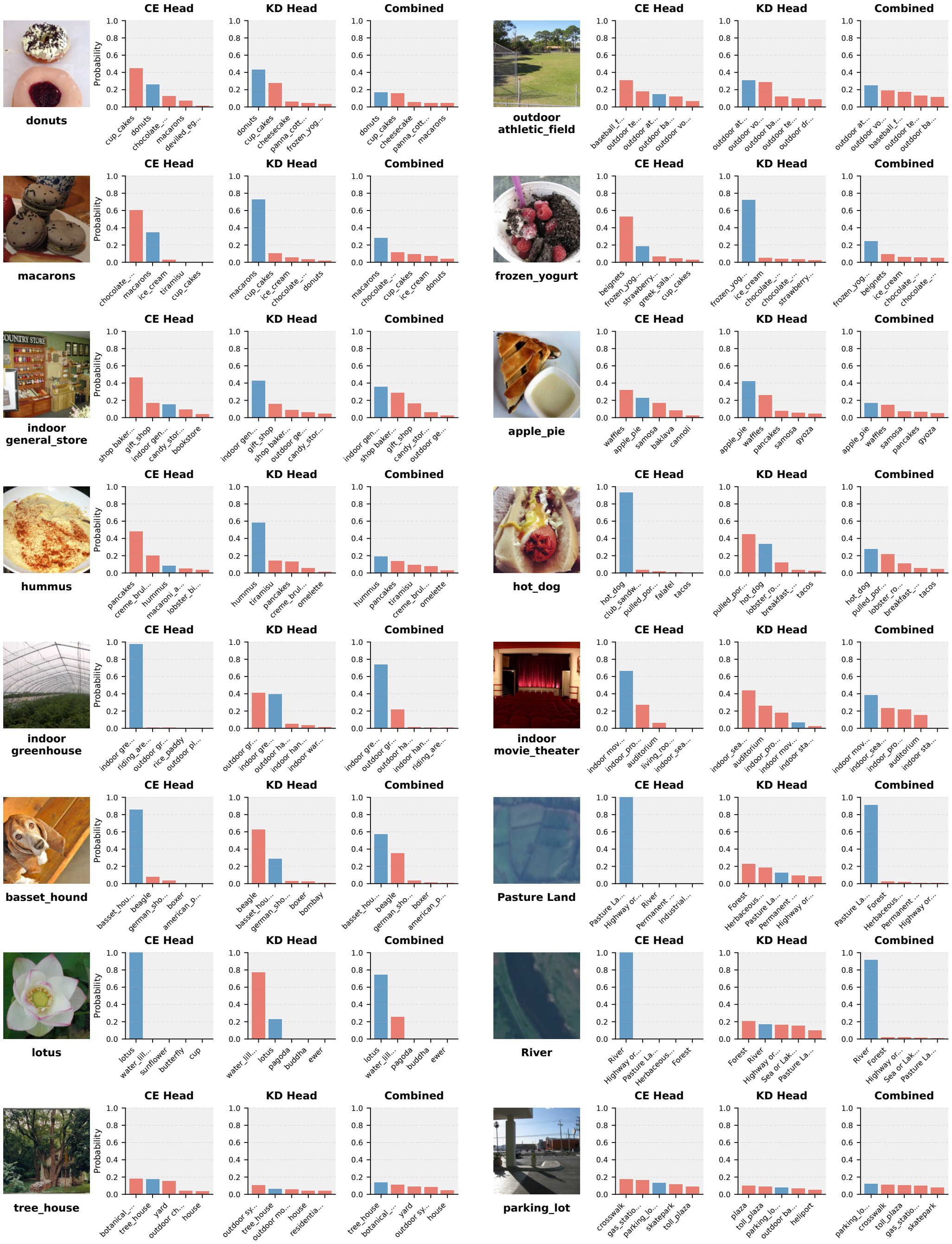

Figure 18: Additional qualitative results on other 10 datasets for models trained under 16-shot semi-supervised setting.

G The Use of LLMs

We used LLMs solely for light editing such as correcting grammatical errors and polishing some words. They did not contribute to research ideation, experiments, analysis, or substantive writing.