arXiv:2504.12311v5 [cs.CL] 15 Oct 2025

Learning Optimal Prompt Ensemble for Multi-source Visual Prompt Transfer

Enming Zhang[1] Liwen Cao[2] Yanru Wu[1] Zijie Zhao[1] Yang Li[1]*
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University †
[2]Southeast University

Abstract

Prompt tuning has emerged as a lightweight strategy for adapting foundation models to downstream tasks, particularly for resource-constrained systems. As pre-trained prompts become valuable assets, combining multiple source prompts offers a promising approach to enhance generalization for new tasks by leveraging complementary knowledge. However, naive aggregation often overlooks different source prompts have different contribution potential to the target task. To address this, we propose HGPrompt, a dynamic framework that learns optimal ensemble weights. These weights are optimized by jointly maximizing an information-theoretic metric for transferability and minimizing gradient conflicts via a novel regularization strategy. Specifically, we propose a differentiable prompt transferability metric to captures the discriminability of prompt-induced features on the target task. Meanwhile, HGPrompt match the gradient variances with respect to different source prompts based on Hessian and Fisher Information, ensuring stable and coherent knowledge transfer while suppressing gradient conflicts among them. Extensive experiments on the large-scale VTAB benchmark demonstrate the state-of-the-art performance of HGPrompt, validating its effectiveness in learning an optimal ensemble for effective multi-source prompt transfer.

Introduction

With the development of expanding datasets, novel architectures, and improved training algorithms (Chen et al. 2020), a significant number of vision foundation models have been developed (Radford et al. 2021; Dosovitskiy 2020; Liu et al. 2021). Transformer-based pre-trained vision models (PVMs) demonstrate exceptional efficacy across diverse tasks, including image classification and semantic segmentation. While these models exhibit impressive capability, adapting them to downstream applications still presents notable challenges. Full model fine-tuning becomes impractical given the substantial parameter volumes and challenges in low-data scenarios. This paradigm shift has made prompt tuning (Huang, Qian, and Yu 2022; Lester, Al-Rfou, and Constant 2021; Zhou et al. 2022) a key adaptation strategy. By freezing PVMs and adding learnable prompt tokens, it achieves competitive performance with only 0.4% parameter updates, significantly fewer than full fine-tuning.

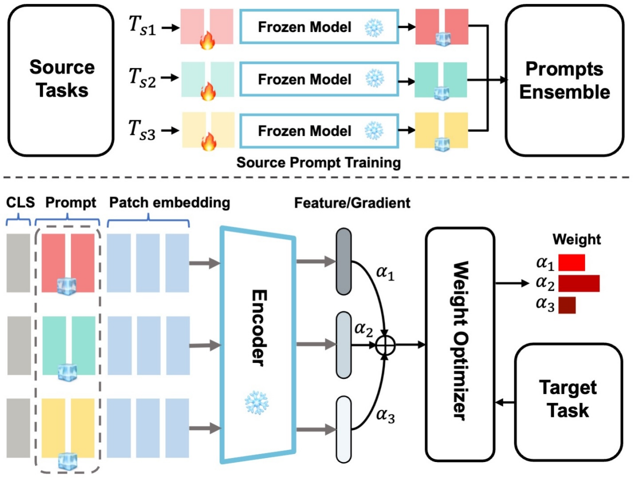


Figure 1: Multi-source prompt transfer framework. Task-specific prompts are tuned via a frozen backbone and statically aggregated for target initialization. Our approach dynamically optimizes source weights through single forward-backward propagation, learning prompt aggregation via an optimization module.

The increasing sophistication of prompt learning has established well-generalized prompts as valuable intellectual assets (Schick and Schütze 2021). This evolution has fostered a practical ecosystem where users can access provider task-specific prompts while maintaining data privacy and model integrity. With the availability of multiple prompts from the prompt pool, these prompts can be utilized in an ensemble way by concurrently assembling them and transferring them to a single pre-trained model, as illustrated in Fig. 1 (Sanh et al. 2021; Wang et al. 2023). However, simply concatenating or averaging source prompts often proves suboptimal, as the knowledge encoded in different prompts may contribute unevenly to the target task and can even lead to representation collapse (Standley et al. 2020).

Effective multi-source transfer necessitates the assignment of adaptive weights to each source prompt. However, conventional methods (Vu et al. 2021; Asai et al. 2022; Su et al. 2022; Zhong et al. 2024) predominantly evaluate the transferability of each prompt in isolation. This approach fails to account for potential interdependencies

[*] Corresponding author. Email: yangli@sz.tsinghua.edu.cn
[†] Shenzhen Key Laboratory of Ubiquitous Data Enabling

when prompts are combined within an ensemble, overlooking complementary effects that can significantly alter overall transferability. Furthermore, existing techniques often rely on heuristic methods, such as computing similarity between prompts parameters (Vu et al. 2021), which typically lack a rigorous theoretical foundation.

To overcome these limitations, we propose a lightweight and theoretically reliable framework that dynamically learns optimal prompt weights. Distinct from prior methods evaluating each prompt in isolation, our key innovation lies in evaluating the transferability of feature ensemble induced by the aggregated prompts. Specifically, we learn the prompt weights by maximizing the H-score, a theoretically grounded, differentiable metric to quantify feature transferability. Unlike conventional approaches relying on heuristics, our method provides an explicit and interpretable measure of each prompt's contribution to the ensemble's transferability, rooted in information-theoretic and statistical principles (Xu et al. 2022).

Moreover, aggregating multiple prompts often introduces detrimental interference between their gradients, which leads to unstable optimization dynamics and suboptimal solution. Building upon the theoretical insight that similar Hessians and Fisher Information reduce inconsistencies in the loss landscape (Parascandolo et al. 2020; Shi et al. 2021; Rame, Dancette, and Cord 2022) , we introduce a simple yet general gradient alignment regularization term in our optimization framework. Specifically, this term match gradient variance from the different source prompts. Minimizing this term encourages consensus during optimization. By resolving these inherent gradient conflicts, our approach develops a prompt ensemble with robust and consistent

Our method achieves state-of-the-art performance through extensive evaluations on the large-scale VTAB benchmark (Zhai et al. 2019), consistently outperforming competitive strategies such as PANDA (Zhong et al. 2024), SPoT, and ATTEMPT. Our approach establishes new benchmarks for future research on multi-source visual prompt transfer. The source code is available in the supplementary material.

Related Work

Parameter-efficient Transfer Learning

Parameter-efficient transfer learning is crucial for adapting large pre-trained models. In NLP, methods like adapters (Houlsby et al. 2019a), BitFit (Ben Zaken, Goldberg, and Ravfogel 2022), and LoRA (Hu et al. 2021) tune only 1-5% of parameters. For vision, early work focused on ConvNets (e.g., residual adapters (Rebuffi, Bilen, and Vedaldi 2017)), but vision Transformers (Dosovitskiy 2020) introduced new challenges. While some NLP techniques (e.g., adapters (Chen et al. 2022)) transfer directly, vision-specific approaches like VPT (Jia et al. 2022) (learnable tokens) and VP (Bahng et al. 2022) (pixel-level perturbations) achieve high efficiency with minimal input-space modifications.

Transferability Estimation

Prompt transferability builds on task transferability research (Zamir et al. 2018), as prompts guide frozen models (Feng 2023). Existing metrics for task transferability (Ding et al. 2024; Tran, Nguyen, and Hassner 2019) can inform prompt evaluation. Information-theoretic approaches like H-score (Bao et al. 2019; Ibrahim, Ponomareva, and Mazumder 2022; Wu et al. 2024), LEEP (Nguyen et al. 2020; Agostinelli et al. 2022), and LogME (You et al. 2021) assess feature discriminability and performance prediction. Optimal transport methods like OTCE (Tan, Li, and Huang 2021; Tan et al. 2024) also measure domain-task differences. These foundations support prompt transferability understanding.

Multi-source Prompt Tuning

Prompt-tuning on smaller pre-trained models often underperforms and is highly sensitive to prompt initialization, as evidenced by prior studies (Huang, Qian, and Yu 2022; Lester, Al-Rfou, and Constant 2021). To address these limitations, Prompt Transfer (PoT) methods have been proposed (Vu et al. 2021; Su et al. 2022), which leverage soft prompts learned on source tasks to initialize prompts for target tasks, thereby improving tuning efficiency and performance. SPOT (Vu et al. 2021) explored the use of metrics to predict the best source tasks for prompt transfer, and in parallel, (Su et al. 2022) emphasized how prompt-induced neuron activations play a crucial role in transferability. In addition to single-task transfer, PoT methods have been extended to multi-task settings. For example, ATTEMPT (Asai et al. 2022) proposed mechanisms to aggregate knowledge from multiple source tasks, using attention mechanisms strategies to initialize target prompts. PANDA (Zhong et al. 2024) explicitly addresses the issue of prior knowledge forgetting by distilling task-specific knowledge into the target prompt.

Preliminary

Visual Prompt Tuning

Visual Prompt Tuning (VPT) is a parameter-efficient transfer learning paradigm that adapts pre-trained vision transformers to downstream tasks by learning task-specific prompt embeddings while keeping the original model parameters frozen. This approach introduces a small set of learnable parameters in the form of prompt tokens, which are prepended to the input sequence, enabling efficient adaptation to new tasks without modifying the underlying model architecture. The key advantage of VPT lies in its ability to leverage the rich representations learned by large-scale pre-trained models while requiring significantly fewer trainable parameters compared to full fine-tuning.

Formally, given a pre-trained Transformer with embedding dimension d, we introduce m learnable prompt tokens $P = [p_1, \ldots, p_m] \in \mathbb{R}^{m \times d}$. For an input image X with patch embeddings $E(X) \in \mathbb{R}^{n \times d}$, the combined input sequence becomes $[P; E(X)] \in \mathbb{R}^{(m+n) \times d}$, where m is the prompt length and n is the number of image patches. The model parameters θ remain fixed during training, with gra-

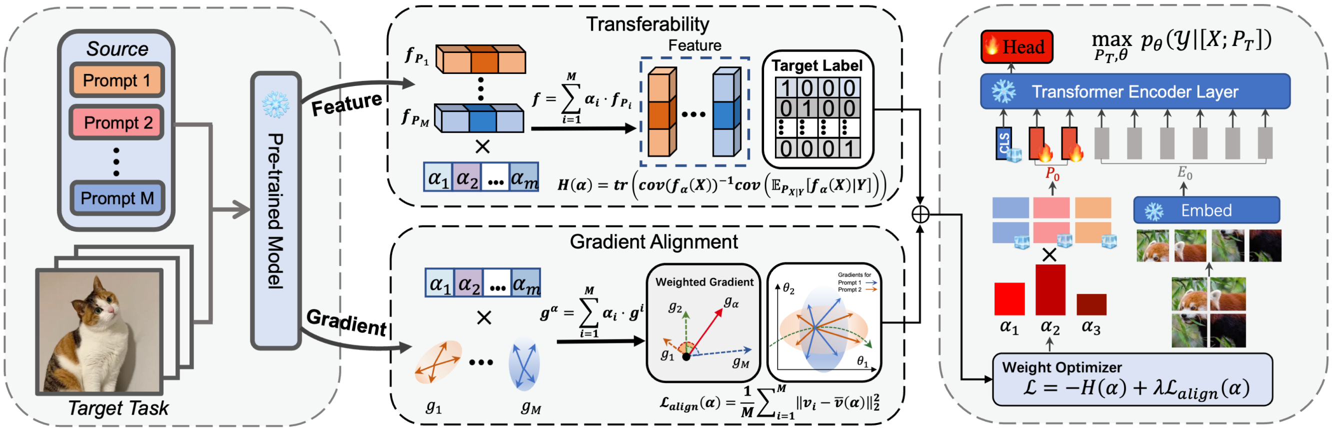

Source
Prompt 1
Prompt 2
Prompt M
Target Task
Pre-trained Model
Feature
Gradient
Transferability
Feature
f_{P_1}
f_{P_M}
$f = \sum_{i=1}^{M} \alpha_i \cdot f_{P_i}$
α_1 α_2 ... α_m
Target Label
1 0 0 0
0 1 0 0
0 0 0 1
$H(\alpha) = tr\left(cov(f_\alpha(X))^{-1} cov\left(\mathbb{E}_{P_{X|Y}}[f_\alpha(X)|Y]\right)\right)$
Gradient Alignment
$g^\alpha = \sum_{i=1}^{M} \alpha_i \cdot g^i$
Weighted Gradient
g_2
g_α
g_1
g_M
θ_2
θ_1
Gradients for
Prompt 1
Prompt 2
$\mathcal{L}_{align}(\alpha) = \frac{1}{M}\sum_{i=1}^{M} \|v_i - \bar{v}(\alpha)\|_2^2$
Head
$\max_{P_T,\theta} p_\theta(\mathcal{Y}|[X;P_T])$
Transformer Encoder Layer
CLS
P_0
E_0
Embed
α_1 α_2 α_3
Weight Optimizer
$\mathcal{L} = -H(\alpha) + \lambda\mathcal{L}_{align}(\alpha)$

Figure 2: Overview of the framework. Given an input image X, the system generates M distinct feature representations $\{f_i\}_{i=1}^{M}$ and corresponding gradients $\{g^i\}_{i=1}^{M}$ through multiple source prompts. These features and gradients are fused using learnable weights $\boldsymbol{\alpha}$ to produce the final combined feature $f_{\boldsymbol{\alpha}}$ and gradient $g^{\boldsymbol{\alpha}}$. The Transferability term evaluates the fused feature distribution against the target class label, and the Gradient Alignment Regularization aligns the prompt gradient variance. The Weight Optimizer jointly optimizes these dual objectives to determine the optimal source weights $\boldsymbol{\alpha}$, which subsequently initialize the target prompt.

dients only propagating through the prompt embeddings P. The prediction probability for class Y is given by:

$$\Pr_\theta(Y|X;P) = \frac{\exp(f_Y([P;E(X)];\theta))}{\sum_{i=1}^{C}\exp(f_i([P;E(X)];\theta))}, \tag{1}$$

where C denotes the number of classes, and $f_i(\cdot)$ represents the pre-trained model's logit output for class i. This formulation allows the model to adapt to new tasks by learning task-specific context through the prompt tokens.

Multi-Source Prompt Transfer

In many real-world scenarios, we often have access to multiple source prompts that can be utilized for the target task. Multi-source prompt transfer aims to harness these related prompts to enhance performance on the target task. Given κ source tasks $\mathcal{S} = \{S_i\}_{i=1}^{\kappa}$ along with their corresponding optimized prompts $\{P_i\}_{i=1}^{\kappa}$, our goal is to construct a target prompt P_T for a new task T by optimally combining the source prompts based on their relevance to the target task.

Let $M \leq \kappa$ denote the number of selected source prompts. We fix the hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ satisfying $\sum_{i=1}^{M}\alpha_i = 1$ and $\alpha_i \geq 0$. Then we simultaneously optimizes both the header parameters θ and the target prompt P_T, where P_T is initialized by a convex combination $P_T = \sum_{i=1}^{M}\alpha_i P_i$ of the frozen source prompts $\{P_i\}_{i=1}^{M}$.

$$\max_{P_T,\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_T}\left[\log P_\theta(y|[x;P_T])\right] \tag{2}$$

This joint optimization learns both task-specific header parameters and target prompt formed by the weighted combination of source prompts. Crucially, the optimization landscape of P_T is highly sensitive to this initialization, making the choice of $\boldsymbol{\alpha}$ a critical factor in final performance. This underscores why learning optimal combination weights is paramount. The weights α_i maintain interpretability by reflecting the relative importance of each source task to the target task. Our method mainly focuses on learning $\boldsymbol{\alpha}$, whose optimized values not only improve transfer performance but also can reveal task relationships and transferability insights.

Methodology

To dynamically learn optimal weights for source prompts, we propose a lightweight framework that jointly maximizes an information-theoretic transferability metric while matching gradient variance through a novel regularization strategy, as illustrated in Fig. 2. First, we present the mathematical formulation of the H-score based transferability metric and establish its theoretical reliability for optimization. Second, we provide a detailed explanation of the gradient alignment regularization, including its theoretical basis and intuition. The framework is designed to be both lightweight and interpretable, capable of serving as a plug-in module for multi-source prompt transfer scenarios.

Measuring Prompt Ensemble Transferability

To overcome the limitations of previous heuristic prompt ensemble strategies that treat prompts independently, we adopt a transferability metric to quantify each prompt's contribution to the combined ensemble. Specifically, we introduce an information theoretic metric for feature transferability based on H-score (Bao et al. 2019; Xu et al. 2022). Unlike conventional assessments that assume transferability correlates with parameter similarity, our proposed metric focuses on the intrinsic informativeness of prompt-induced features, explicitly evaluating the effectiveness of prompt ensembles for the target task. The mathematical formulation of H-score is defined as follows:

Definition 1 *With input data x, label y and feature extractor $f(x)$ (a zero-mean feature function). The one-sided H-score* *of f with regard to the task casting x to y is:*

$$H(\boldsymbol{f}) = \mathrm{tr}\left(\mathrm{cov}(f(X))^{-1}\,\mathrm{cov}\left(\mathbb{E}_{P_{X|Y}}[f(X)|Y]\right)\right). \quad (3)$$

The full derivation is provided in the Appendix. This formulation admits an intuitive interpretation: A high H-score indicates larger inter-class discriminability, characterized by $\mathrm{cov}\left(\mathbb{E}[f(X)|Y]\right)$, and minimized feature redundancy, reflected in a small $\mathrm{tr}(\mathrm{cov}(f(X)))$. Thus, elevated H-scores signify that prompt successfully elicits transfer-effective features from the model.

Given a frozen visual encoder f_θ and M source prompts $\{P_i\}_{i=1}^{M}$ pre-trained on distinct tasks, for an input image $X \in \mathcal{X}$, the i-th source prompt feature extraction is defined as:

$$f_{P_i}(X) = f_\theta\left([x_0; P_i; E(X)]\right) \in \mathbb{R}^h \quad (4)$$

where $E(X) \in \mathbb{R}^{n\times d}$ denotes image patch embeddings, $x_0 \in \mathbb{R}^d$ the [CLS] token, and h the feature dimension. The optimal combination weights α are determined by maximizing the H-score of the weighted feature sum, which yields the most transferable prompted feature representation.

Definition 2 *Given source-specific features $\{f_{P_j}\}_{j=1}^{M}$, the optimal feature weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^\top \in \mathbb{R}^M$ are determined by:*

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} H\left(\sum\nolimits_{j=1}^{M} \alpha_j \cdot f_{P_j}\right) \quad \textit{s.t.} \quad \sum\nolimits_{j=1}^{M} \alpha_j = 1 \quad (5)$$

We then verify the benign property of the proposed optimization problem by proving that the optimal objective H-score is a convex function of $\boldsymbol{\alpha}$.

Theorem 1 *Given input data X, labels Y, and $\{f_{P_i}\}_{i=1}^{n}$ with $\sum_{i=1}^{n} \alpha_i = 1$, the H-score of the weighted feature is a convex quadratic form:*

$$H(f) = H\left(\sum\nolimits_{i=1}^{n} \alpha_i f_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j \cdot \mathrm{tr}\left(\mathbb{E}_{P_Y}\left[\mathbb{E}_{P_{X|Y}}[f_i(X)|Y] \cdot \mathbb{E}_{P_{X|Y}}[f_j(X)|Y]^\top\right]\right) \quad (6)$$

With above theoretical guarantee, the optimization problem can be reliably solved using gradient descent based methods, as detailed in Algorithm 1. We provide the complete proof of Theorem 1 in the Appendix.

Algorithm 1: HGPrompt: Training Process

```
Input: Target data $\mathcal{D}_T = \{(x_i, y_i)\}_{i=1}^{N}$, source prompts $\{P_j\}_{j=1}^{M}$, learning rate $\eta$, hyperparameter $\lambda$
Output: Optimal weights $\boldsymbol{\alpha}^*$
Initialize $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$ with $\sum_{j=1}^{M} \alpha_j = 1$
for epoch = 1 to $K$ do
    $\mu_y(\boldsymbol{\alpha}) = \mathbb{E}_{X|Y}[f_{\boldsymbol{\alpha}}(X)|Y = y]$
    $H(\boldsymbol{\alpha}) = \mathrm{tr}(\mathrm{cov}(f_{\boldsymbol{\alpha}})^{-1}\mathrm{cov}(\{\mu_y\}))$
    Compute gradient variance: $\{v_i\}_{i=1}^{M}$ via Eq.(8)
    Evaluate gradient alignment regularization: $\mathcal{L}_{\text{align}}(\boldsymbol{\alpha})$ via Eq.(9)
    Compute total loss: $\mathcal{L}(\boldsymbol{\alpha}) = -H(\boldsymbol{\alpha}) + \lambda\mathcal{L}_{\text{align}}(\boldsymbol{\alpha})$
    Update weights: $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta\nabla_{\boldsymbol{\alpha}}\mathcal{L}$
end for
```

Gradient Alignment Regularization

Each prompt encodes task-specific knowledge. However, directly aggregating these prompts often leads to cross-interference between prompts, where independent evaluation fails to account for their synergistic or conflicting interactions. To address these issues, we propose aligning the gradient directions of all prompts, ensuring they collectively guide the model toward a unified optimization trajectory.

Building upon the gradient agreement principles from multi-task learning (Yu et al. 2020; Shi et al. 2021; Liu et al. 2023; Rame, Dancette, and Cord 2022), we propose a novel gradient variance matching objective for multi-source prompt transfer. Given M source tasks with optimized prompts $\{P_i\}_{i=1}^{M}$, we compute the gradient of the loss with respect to each source prompt P_i as:

$$g^i = \nabla_{P_i}\mathcal{L}(f_\theta([x_0; P_i; E(X)]), y). \quad (7)$$

For each source prompt P_i, we compute its gradient variance:

$$v_i = \mathrm{Var}(G) = \frac{1}{N-1}\sum_{j=1}^{N}(g_j^i - g_j^\alpha)^2, \quad (8)$$

where $g^\alpha = \frac{1}{M}\sum_{i=1}^{M}\alpha_i g^i$ is the weighted mean of gradients, and $\mathbf{G} = [g^i]_{i=1}^{M}$ is the $N \times |P|$ prompt gradient matrix and N is the batch size of samples used for gradient computation. We adapt the regularization to promote gradient variance alignment among source prompts:

$$\mathcal{L}_{\text{align}}(\boldsymbol{\alpha}) = \frac{1}{M}\sum_{i=1}^{M}\|v_i - \bar{v}(\boldsymbol{\alpha})\|_2^2, \quad (9)$$

where the mean gradient variance is defined as $\bar{v}(\boldsymbol{\alpha}) = \frac{1}{M}\sum_{i=1}^{M} v_i(\boldsymbol{\alpha})$. Balanced with a hyperparameter coefficient $\lambda > 0$, this regularization penalty complements the original H-score objective,

$$\mathcal{L}(\boldsymbol{\alpha}) = -H(\boldsymbol{\alpha}) + \lambda\mathcal{L}_{\text{align}}(\boldsymbol{\alpha}). \quad (10)$$

Theoretical Analysis Our gradient alignment regularization $\mathcal{L}_{\text{align}}$ builds on established theoretical foundations in domain-invariant learning (Parascandolo et al. 2020), which seeks to identify invariant mechanisms in data by finding model parameters that exhibit consistent behavior across different domains. To quantify the consistency of the loss landscape around the optimal parameter θ^* across domains, the inconsistency score is defined as follows:

Definition 3 *Given a model parameter θ^*, the inconsistency score $\mathcal{I}^\epsilon(\theta^*)$ is defined as:*

$$\mathcal{I}^\epsilon(\theta^*) = \max_{A,B}\ \max_{\theta\in N^\epsilon_{A,\theta^*}} \left|\mathcal{R}_B(\theta) - \mathcal{R}_A(\theta^*)\right|, \quad (11)$$

where $\theta \in N^\epsilon_{A,\theta^*}$ if there exists a continuous path in parameter space between θ and θ^* along which the risk $\mathcal{R}_A$ remains within ϵ of $\mathcal{R}_A(\theta^*)$, for $\epsilon > 0$.

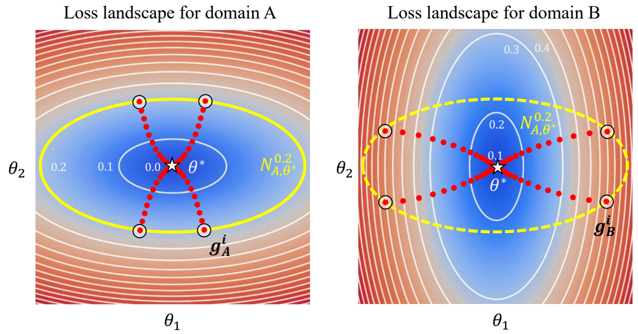


Figure 3: Loss landscapes for a two-parameter model, showing conflicting gradient variances $\{g_i^{(A)}\}_{i=1}^{n_A}$ and $\{g_i^{(B)}\}_{i=1}^{n_B}$ around θ^*. This shows the case where a nearby solution $\theta \in N^{\epsilon}_{A,\theta^*}$ maintains equivalent risk $\mathcal{R}_A(\theta) \approx \mathcal{R}_A(\theta^*)$ in domain A but exhibits higher risk in domain B.

This concept is illustrated in Figure 3, which demonstrates that minima with low consistency fail to generalize to new environments (Deutsch 2011). The inconsistency score $\mathcal{I}$ increases when the loss landscapes around θ^* present conflicting geometric structures across different domains.

Theorem 2 *Let θ^* be a simultaneous local minimum across domains with positive definite Hessians. Under the quadratic bowl assumption and for sufficiently small $\epsilon > 0$:*

$$\mathcal{I}(\theta^*) = \max_{A,B}\left(|R_B(\theta^*) - R_A(\theta^*)| + \max_{\frac{1}{2}\theta^\top H_A \theta \le \epsilon} \frac{1}{2}|\theta^\top H_B \theta|\right). \tag{12}$$

We provide the complete proof of Theorem 2 in the Appendix. The first term captures the loss landscape mismatch through domain-level risk differences. We will prove and show that $\mathcal{L}_{\text{align}}$ forces this term to be small in Appendix. For the second term, we employ a diagonal approximation of the Hessians for analysis. In that case, $H_e = \text{diag}(\lambda_1^e, \dots, \lambda_h^e)$ with $\forall i \in \{1, \dots, h\}, \lambda_i^e > 0$, the curvature term can be expressed as:

$$\begin{aligned}\max_{\frac{1}{2}\theta^\top H_A \theta \le \epsilon} \frac{1}{2}\theta^\top H_B \theta &= \max_{\|\tilde{\theta}\|_2^2 \le 2\epsilon} \sum_i \tilde{\theta}_i^2 \lambda_i^B / \lambda_i^A \\ &= \epsilon \cdot \max_i \lambda_i^B / \lambda_i^A.\end{aligned} \tag{13}$$

This result demonstrates that the second term diminishes when H_A and H_B have similar eigenvalues. Consequently, enforcing $H_A = H_B$ reduces inconsistencies in the loss landscape, thereby enhancing generalization performance. As we elaborate in the Appendix, our proposed $\mathcal{L}_{\text{align}}$ effectively aligns domain-level Hessians through gradient variance matching, by leveraging the fundamental connections between gradient variance, Fisher Information, and the Hessian.

Experiments

We evaluate the proposed approach for a wide range of downstream recognition tasks with pre-trained Transformer backbones. We first describe our experimental setup, including the pre-trained backbone and downstream tasks and a brief introduction to other transfer learning methods.

Setup

Datasets. We experiment on a collection of 13 datasets from V-tab-1k (Zhai et al. 2019). VTAB is a collection of dieverse visual classification tasks, which encompasses three distinct categories of tasks: Natural, featuring images taken with conventional cameras; Specialized, containing data acquired through specialized devices, such as satellite sensors; and Structured, which demands spatial reasoning, like counting objects. We provide more detailed descriptions of the datasets in the Appendix.

Implementation Details. We implement all experiments on NVIDIA A800-80GB GPUs. For a fair comparison, all methods use a ViT-B/16 backbone pre-trained on ImageNet-21k, and the number of prompt tokens is set to 50. We follow the original configurations, eg. number of image patches divided, existence of `[CLS]`, etc. We train the prompt on all the source tasks for 10 epochs for source prompt training. We use 2000 samples from each source task for each target task to compute the transferability loss and gradient alignment loss.

Baselines. We compare our approach to eleven recent methods, categorizing them as follows: (1) Methods that retrain the classification head: **PARTIAL-k** (Zhang, Isola, and Efros 2016) fine-tunes only the last k layers of the backbone while freezing others; **MLP-k** utilizes a multilayer perceptron with k layers as the classification head instead of a linear layer. (2) Methods that update a subset of backbone parameters or add new trainable modules: **Adapter** (Houlsby et al. 2019b) inserts new MLP modules with residual connections into transformer layers; **SIDETUNE** (Zhang et al. 2020) trains a side network and linearly interpolates between pre-trained features and side-tuned features before feeding them into the head; **BIAS** (Ben Zaken, Goldberg, and Ravfogel 2022) fine-tunes only the bias terms of the pre-trained backbone. (3) Prompt transfer methods: **Average** directly uses the mean of source prompt embeddings; **Single-Best** selects the source prompt with optimal transfer performance; **Visual Prompt Tuning (VPT)** (Jia et al. 2022) initializes target prompt embeddings randomly; **SPoT** (Vu et al. 2022) calculates similarity between source and target prompt embeddings; **ATTEMPT** (Asai et al. 2022) mixes pre-trained source and target prompts via an attention mechanism; **PANDA** (Zhong et al. 2024) measures cosine similarity between task embeddings as a transferability proxy.

Main Results

Our experimental evaluation across 13 diverse vision tasks, as detailed in Tab. 1, demonstrates HGPrompt's superiority over 13 baselines using a ViT-B/16 backbone pre-trained on ImageNet-21k. The proposed method achieves state-of-the-art performance with an average accuracy of 60.3%, surpassing prior multi-source prompt transfer approaches. HG-

Method	Cifar100	DTD	Flowers102	Pets	SVHN	EuroSAT	DMLab	sNORB-Azim	sNORB-Ele	dSpr-Loc	dSpr-Ori	Clevr-Count	Clevr-Dist	**Average**
Linear	61.7	58.6	96.6	83.9	32.7	83.9	30.6	12.2	20.3	12.6	18.2	32.1	28.6	44.0
PARTIAL-1	64.4	60.3	97.5	86.0	36.3	87.8	32.5	16.5	21.8	31.3	39.2	41.3	32.1	49.8
MLP-2	39.3	43.0	88.5	76.3	28.0	80.4	29.7	12.5	20.3	24.5	30.8	31.5	29.5	41.1
MLP-3	41.9	46.2	90.5	78.4	30.3	83.9	30.7	14.1	21.5	25.9	33.1	33.8	30.2	43.1
MLP-5	38.1	44.1	90.8	79.1	28.8	81.2	30.5	13.9	20.4	22.5	33.2	33.0	29.1	41.9
MLP-9	38.6	46.1	92.1	81.2	28.0	84.2	31.0	14.7	22.9	19.7	33.2	39.0	28.3	43.0
Adapter	73.8	61.7	97.5	86.6	32.7	85.3	29.4	11.9	19.5	22.4	20.8	40.1	35.1	47.4
SIDETUNE	53.5	58.7	93.4	77.2	17.6	37.2	26.7	10.6	15.1	13.2	13.6	20.3	19.4	35.1
BIAS	70.8	57.5	97.2	85.1	45.3	89.7	31.2	13.5	23.2	63.3	39.7	49.1	54.5	56.2
VPT	56.0	57.4	97.3	82.5	61.4	88.9	36.7	15.3	14.1	42.8	35.5	34.8	51.0	51.8
Single-Best	63.9	60.2	97.0	83.4	63.2	89.5	36.1	18.3	18.9	57.1	36.4	38.3	51.9	54.9
Average	64.8	61.8	96.1	84.2	64.4	90.6	36.3	17.2	21.1	59.5	34.1	37.5	50.8	55.2
SPoT	75.6	63.7	97.7	86.3	70.4	92.1	37.3	19.4	23.3	65.0	36.0	41.5	52.8	58.5
ATTEMPT	67.8	62.1	96.1	85.1	69.0	91.0	36.2	17.9	23.5	61.2	35.0	43.5	51.2	56.9
PANDA	74.1	61.3	96.5	86.2	**71.2**	90.8	37.8	19.4	24.0	67.7	37.3	42.8	**53.9**	58.7
HGPrompt	**75.9**	**64.2**	**98.1**	**87.4**	71.0	**92.6**	**38.1**	**20.3**	**24.9**	**68.1**	**40.4**	**49.3**	53.5	**60.3**

Table 1: Performance comparison across diverse vision tasks using a Vision Transformer (ViT-B/16) backbone pre-trained on ImageNet-21k. The second-best results are underlined, while the best results are highlighted in bold. All reported values represent the average accuracy obtained from three independent runs, with the highest average accuracy achieved by our method.

Table 2: Ablation Study on Framework Components

$\mathbf{H}(\alpha)$	$\mathcal{L}_{\text{align}}$	Cifar	DTD	Pets	Euro	**Avg**
×	×	60.4	57.8	82.7	89.1	72.5
✓	×	74.6	62.3	85.9	91.2	78.5
×	✓	74.1	61.9	85.5	90.8	78.1
✓	✓	**75.9**	**64.2**	**87.4**	**92.6**	**80.0**

Prompt excels in fine-grained recognition tasks, achieving top results on Flowers102 and Oxford Pets. It also outperforms all baselines in texture analysis on DTD and maintains competitive performance on CIFAR100. Notably, the method establishes new state-of-the-art results in geometric reasoning tasks, including sNORB-Azimuth and dSprite-Orientation, with significant improvements in complex visual reasoning tasks like Clevr-Count. While PANDA retains an advantage in SVHN, HGPrompt exhibits a more balanced and robust performance across all task categories, highlighting its effectiveness.

Ablation Study

The ablation study in Table 2 systematically evaluates each component's contribution in our framework. Additional dataset results are provided in the Appendix. The baseline method, which directly optimizes weights by minimizing cross-entropy loss on the target task, achieves 72.5% average accuracy. Using the H-score objective alone improves performance to 78.5%, validating its effectiveness for feature discriminability evaluation. Similarly, employing $\mathcal{L}_{\text{align}}$ as the sole optimization objective yields 78.1% accuracy. Most notably, combining both components produces the best performance 80.2%, demonstrating their complementary roles in achieving optimal transfer learning results.

Analysis and Discussion

Evaluation on Prompt Weights To demonstrate the effectiveness of our learnt Prompt Weights, we pretrained a set of source prompts and evaluated their zero-shot transfer accuracy, as shown in Fig. 4. We plotted the weights calculated by SPoT, PANDA, and our proposed HGPrompt method. Our approach more accurately reflects semantic task affinities, indicating that our proposed metric can better distinguish different task relationships: similar tasks exhibit larger prompt transferability. For example, tasks involving natural scenes—such as Flowers, Pets, and DTD—demonstrate higher inter-task transferability, a pattern largely captured by HGPrompt. To systematically validate this finding, we conducted a quantitative analysis presented in Table 3. We computed Spearman's rank correlation between the predicted weights and the actual zero-shot transfer accuracy, confirming that our metric achieves superior correlation compared to existing approaches. In contrast, the results reveal that SPoT and PANDA struggle to accurately evaluate task-relevant semantic information, exhibiting significant fluctuations. Complete results are provided in the appendix.

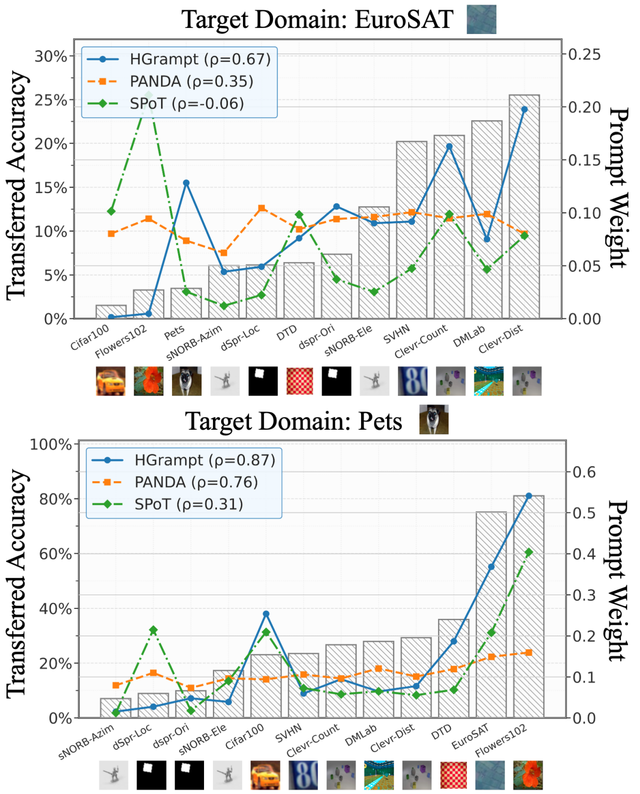


Figure 4: Prompt Weights analysis for 12 source prompts. Bar plots represent single-source transfer accuracy (left axis), while line plots indicate prompt weights (right axis).

Table 3: Spearman's ρ correlation scores.

	Cifar	C-di	d-Lo	DML	SVHN	**Avg**
SPoT	0.552	0.175	-0.168	0.112	-0.147	0.105
PANDA	0.916	0.441	0.552	0.713	0.224	0.569
HGPrompt	**0.944**	**0.664**	**0.853**	**0.727**	**0.853**	**0.808**

Performance Scaling with Source Prompts Number As shown in Fig.5, our method demonstrates progressively stronger performance advantages over PANDA and SPOT when using DTD as the target domain, particularly as the number of source prompts increases from 3 to 11. Results for other domains can be found in Appendix. This scaling behavior highlights our approach's superior capability in effectively utilizing larger prompt collections. While absolute accuracy shows consistent improvement with additional source prompts, the system eventually approaches an inherent efficiency ceiling.

Representation Space Visualization To analyze the effect of prompt ensemble on the learned representation of the target test data, we present t-SNE visualizations of ViT feature embeddings in four different prompt transfer methpds in Fig.6. As shown in Fig.6.d, our method shows better class discriminability than other baselines. Instead of scattered clusters, objects from the same category form tightly grouped regions with clear separation boundaries. The visualization underscores the effectiveness of our method in constructing a coherent and well-structured feature space for transfer learning.

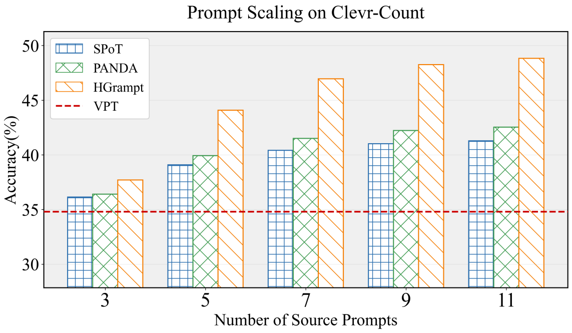
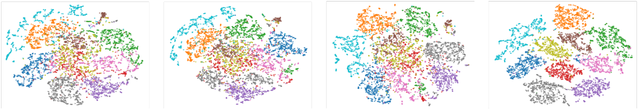


Figure 5: Performance scaling with increasing source prompts on Clever-Count target domain.



(a) SPoT (b) ATTEMPT (c) PANDA (d) HGPrompt

Figure 6: t-SNE Visualization of representations on EuroSAT (10 Classes). Each color corresponds to a distinct class.

Parameter Analysis We analyze how the regularization coefficient λ influences our results. Performance trends across benchmarks reveal optimal outcomes when both objectives contribute comparably, indicating mutual reinforcement between feature transferability and gradient alignment. The full results are provided in the appendix.

Discussion While our current work has demonstrated the effectiveness of visual prompting within transformer architectures, we acknowledge its limitations in terms of architectural specificity and modality constraints. Future work in these directions may require novel approaches to prompt design and adaptation, potentially drawing inspiration from recent advances in multimodal learning and architecture-agnostic representation techniques. The ultimate goal would be to establish prompting as a truly universal interface for model adaptation and control, transcending specific architectural choices or modality limitations.

Conclusion

In this work, we introduce HGPrompt, a novel framework for multi-source prompt transfer that explicitly optimizes the prompt ensemble. Our methodology determines optimal source weights by maximizing the H-score while matching gradient variance, thereby effectively quantifying the transferability of the source prompt ensemble. By dynamically balancing feature discriminability with generalization, HGPrompt leverages complementary information across prompts while simultaneously suppressing interference. Our contributions establish a solid foundation for advancing multi-source prompt transfer, offering both theoretical and practical insights to enhance foundation model adaptability.

References

Agostinelli, A.; Uijlings, J.; Mensink, T.; and Ferrari, V. 2022. Transferability metrics for selecting source model ensembles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7936–7946.

Asai, A.; Salehi, M.; Peters, M.; and Hajishirzi, H. 2022. ATTEMPT: Parameter-Efficient Multi-task Tuning via Attentional Mixtures of Soft Prompts. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6655–6672. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*.

Bao, Y.; Li, Y.; Huang, S.-L.; Zhang, L.; Zheng, L.; Zamir, A.; and Guibas, L. 2019. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE international conference on image processing (ICIP)*, 2309–2313. IEEE.

Ben Zaken, E.; Goldberg, Y.; and Ravfogel, S. 2022. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 1–9. Dublin, Ireland: Association for Computational Linguistics.

Chen, S.; Ge, C.; Tong, Z.; Wang, J.; Song, Y.; Wang, J.; and Luo, P. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35: 16664–16678.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Deutsch, D. 2011. *The beginning of infinity: Explanations that transform the world.* penguin uK.

Ding, Y.; Jiang, B.; Yu, A.; Zheng, A.; and Liang, J. 2024. Which Model to Transfer? A Survey on Transferability Estimation. *arXiv preprint arXiv:2402.15231*.

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

Feng, L. 2023. Learning to Predict Task Transferability via Soft Prompt. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8829–8844. Singapore: Association for Computational Linguistics.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019a. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*, 2790–2799. PMLR.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019b. Parameter-Efficient Transfer Learning for NLP. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2790–2799. PMLR.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Y.; Qian, K.; and Yu, Z. 2022. Learning a better initialization for soft prompts via meta-learning. *arXiv preprint arXiv:2205.12471*.

Ibrahim, S.; Ponomareva, N.; and Mazumder, R. 2022. Newer is not always better: Rethinking transferability metrics, their peculiarities, stability and performance. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 693–709. Springer.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727. Springer.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Liu, X.; Zhong, Y.; Zhang, Y.; Qin, L.; and Deng, W. 2023. Enhancing generalization of universal adversarial perturbation through gradient aggregation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4435–4444.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Nguyen, C.; Hassner, T.; Seeger, M.; and Archambeau, C. 2020. Leep: A new measure to evaluate transferability of learned representations. In *International Conference on Machine Learning*, 7294–7305. PMLR.

Parascandolo, G.; Neitz, A.; Orvieto, A.; Gresele, L.; and Schölkopf, B. 2020. Learning explanations that are hard to vary. *arXiv preprint arXiv:2009.00329*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rame, A.; Dancette, C.; and Cord, M. 2022. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*, 18347–18377. PMLR.

Rebuffi, S.-A.; Bilen, H.; and Vedaldi, A. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Schick, T.; and Schütze, H. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 255–269. Online: Association for Computational Linguistics.

Shi, Y.; Seely, J.; Torr, P. H.; Siddharth, N.; Hannun, A.; Usunier, N.; and Synnaeve, G. 2021. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*.

Standley, T.; Zamir, A.; Chen, D.; Guibas, L.; Malik, J.; and Savarese, S. 2020. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, 9120–9132. PMLR.

Su, Y.; Wang, X.; Qin, Y.; Chan, C.-M.; Lin, Y.; Wang, H.; Wen, K.; Liu, Z.; Li, P.; Li, J.; Hou, L.; Sun, M.; and Zhou, J. 2022. On Transferability of Prompt Tuning for Natural Language Processing. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3949–3969. Seattle, United States: Association for Computational Linguistics.

Tan, Y.; Li, Y.; and Huang, S.-L. 2021. Otce: A transferability metric for cross-domain cross-task representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15779–15788.

Tan, Y.; Zhang, E.; Li, Y.; Huang, S.-L.; and Zhang, X.-P. 2024. Transferability-guided cross-domain cross-task transfer learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Tran, A. T.; Nguyen, C. V.; and Hassner, T. 2019. Transferability and hardness of supervised classification tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1395–1405.

Vu, T.; Lester, B.; Constant, N.; Al-Rfou, R.; and Cer, D. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*.

Vu, T.; Lester, B.; Constant, N.; Al-Rfou', R.; and Cer, D. 2022. SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5039–5059. Dublin, Ireland: Association for Computational Linguistics.

Wang, Z.; Panda, R.; Karlinsky, L.; Feris, R.; Sun, H.; and Kim, Y. 2023. Multitask prompt tuning enables parameter-efficient transfer learning. *arXiv preprint arXiv:2303.02861*.

Wu, Y.; Wang, J.; Wang, W.; and Li, Y. 2024. H-ensemble: An Information Theoretic Approach to Reliable Few-Shot Multi-Source-Free Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15970–15978.

Xu, X.; Huang, S.-L.; Zheng, L.; and Wornell, G. W. 2022. An information theoretic interpretation to deep neural networks. *Entropy*, 24(1): 135.

You, K.; Liu, Y.; Wang, J.; and Long, M. 2021. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*, 12133–12143. PMLR.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3712–3722.

Zhai, X.; Puigcerver, J.; Kolesnikov, A.; Ruyssen, P.; Riquelme, C.; Lucic, M.; Djolonga, J.; Pinto, A. S.; Neumann, M.; Dosovitskiy, A.; et al. 2019. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*.

Zhang, J. O.; Sax, A.; Zamir, A.; Guibas, L.; and Malik, J. 2020. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 698–714. Springer.

Zhang, R.; Isola, P.; and Efros, A. 2016. Colorful image colorization. In Matas, J.; Sebe, N.; Welling, M.; and Leibe, B., eds., *Computer Vision - 14th European Conference, ECCV 2016, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 649–666. Germany: Springer. ISBN 9783319464862.

Zhong, Q.; Ding, L.; Liu, J.; Du, B.; and Tao, D. 2024. PanDa: Prompt Transfer Meets Knowledge Distillation for Efficient Model Adaptation. *IEEE Transactions on Knowledge and Data Engineering*, 36(9): 4835–4848.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.