Appendix

Anonymous submission

Appendix. I - Proofs and Derivations

Proof of Definition 1

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and output space respectively. Denote the transferred feature representation by $f : \mathcal{X} \to \mathbb{R}^k$. For a classification task, let $h_f : \mathcal{X} \times \mathcal{Y} \to [0,1]^{|\mathcal{Y}|}$ be a predictor function with the log-loss function $L(f(x), y)$ for a given (x, y) sample. The traditional machine learning approach uses stochastic gradient descent to minimize $L(h) = \mathbb{E}_{X,Y}[L(f(x), y)]$. We will show that the optimal log loss when f is given can be characterized analytically using concepts in information theory and statistics.

Definition 1. *The Divergence Transition Matrix (DTM) of discrete random variables X and Y is a $|\mathcal{Y}| \times |\mathcal{X}|$ matrix $\hat{B}$ with entries:*

$$\hat{B}_{y,x} = \frac{P_{XY}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}} - \sqrt{P_Y(y)}\sqrt{P_X(x)}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Given m training examples $\{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$, the loss $L(f, \theta)$ is:

$$L(f,\theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{|\mathcal{Y}|} \nVdash\{y^{(i)} = k\} \log\left(\frac{e^{-\theta_k^T f(x^{(i)})}}{\sum_{j=1}^{|\mathcal{Y}|} e^{-\theta_j^T f(x^{(i)})}}\right)$$

Using concepts in Euclidean information geometry, it is shown in (Huang et al. 2019) that under a local assumption, for a given feature dimension k:

$$\text{argmin}_{f,\theta} = \text{argmin}_{\Psi \in \mathbb{R}^{|\mathcal{X}| \times k}, \Phi \in \mathbb{R}^{|\mathcal{Y}| \times k}} \frac{1}{2}\|\tilde{B} - \Psi\Phi^{\mathrm{T}}\|_F^2 + o(\epsilon^2) \tag{1}$$

By defining $\phi(x) = \sqrt{P_X(x)}f(x)$, the optimal Ψ^* is:

$$\Psi^* = \tilde{B}\Phi(\Phi^{\mathrm{T}}\Phi)^{-1} \tag{2}$$

Substituting (2) into (1), the log loss has a closed-form solution:

$$\|\tilde{B}\|_F^2 - \|\tilde{B}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\|_F^2 \tag{3}$$

Let X, x, $\mathcal{X}$ and P_X represent a random variable, a value, the alphabet and the probability distribution respectively. $\sqrt{P_X}$ denotes the vector with entries $\sqrt{P_X(x)}$ and $[\sqrt{P_X}] \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ denotes the diagonal matrix of $\sqrt{P_X}$. For joint distribution P_{YX}, $\mathbf{P}_{YX} \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$ represents the probability matrix. Given k feature functions $f_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \ldots, k$, let $f(x) = [f_1(x), \ldots, f_k(x)] \in \mathbb{R}^k$ be the feature vector of x, and $F = [f(x_1)^\top, \ldots, f(x_{|\mathcal{X}|})^\top]^\top \in \mathbb{R}^{|\mathcal{X}| \times k}$ be the feature matrix over all elements in $\mathcal{X}$. We can further rewrite $||\tilde{B}\Phi(\Phi^T\Phi)^{-\frac{1}{2}}||_F^2$ as follows.

$$\begin{aligned}\|\tilde{B}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\|_{\mathrm{F}}^2 &= \mathrm{tr}\left((\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\Phi^{\mathrm{T}}\tilde{B}^{\mathrm{T}}\tilde{B}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\right)\\ &= \mathrm{tr}\left((\Phi^{\mathrm{T}}\Phi)^{-1}\Phi^{\mathrm{T}}\tilde{B}^{\mathrm{T}}\tilde{B}\Phi\right)\end{aligned} \tag{4}$$

Since any feature function can be centered by subtracting the mean, without the loss of generality, we assume $\mathbb{E}[f(X)] = 0$. Using the one-to-one correspondence between Φ and F, i.e. $\Phi = [\sqrt{\mathrm{P}_X}]F \in \mathbb{R}^{|\mathcal{X}| \times k}$, we have:

$$\begin{aligned}\Phi^{\mathrm{T}}\Phi &= \left([\sqrt{\mathrm{P}_X}]F\right)^{\mathrm{T}}\left([\sqrt{\mathrm{P}_X}]F\right)\\ &= \mathbb{E}[f(X)^{\mathrm{T}}f(X)]\\ &= \mathrm{cov}(f(X))\end{aligned} \tag{5}$$

The DTM matrix $\tilde{B}$ introduced in Definition 1 can be written in matrix notation:$\tilde{B} = \left[\sqrt{\mathrm{P}_Y}\right]^{-1}\mathrm{P}_{YX}\left[\sqrt{\mathrm{P}_X}\right]^{-1} - \sqrt{\mathrm{P}_Y}\sqrt{\mathrm{P}_X}^{\mathrm{T}}$ Then we have:

$$\begin{aligned}\tilde{B}\Phi = &\left(\left[\sqrt{\mathrm{P}_Y}\right]^{-1}\mathrm{P}_{YX}\left[\sqrt{\mathrm{P}_X}\right]^{-1} - \sqrt{\mathrm{P}_Y}\sqrt{\mathrm{P}_X}^{\mathrm{T}}\right)\\ &\left[\sqrt{\mathrm{P}_X}\right]F\\ = &\left[\sqrt{\mathrm{P}_Y}\right]\left([\mathrm{P}_Y]^{-1}\mathrm{P}_{YX}F - 1 \cdot \mathbb{E}[f(X)]^{\mathrm{T}}\right),\end{aligned} \tag{6}$$

where 1 is a column vector with all entries 1 and length $|\mathcal{Y}|$. It follows that:

$$\begin{aligned}\Phi^{\mathrm{T}}\tilde{B}^{\mathrm{T}}\tilde{B}\Phi &= \left([\mathrm{P}_Y]^{-1}\mathrm{P}_{YX}F - 1 \cdot \mathbb{E}[f(X)]^{\mathrm{T}}\right)^{\mathrm{T}}\\ &\quad [\mathrm{P}_Y]\left([\mathrm{P}_Y]^{-1}\mathrm{P}_{YX}F - 1 \cdot \mathbb{E}[f(X)]^{\mathrm{T}}\right)\\ &= \mathbb{E}_{P_Y}\Big[\left(\mathbb{E}[f(X)|Y] - 1 \cdot \mathbb{E}[f(X)]^{\mathrm{T}}\right)\\ &\quad \cdot \left(\mathbb{E}[f(X)|Y] - 1 \cdot \mathbb{E}[f(X)]^{\mathrm{T}}\right)\Big]\\ &= \mathrm{cov}\left(\mathbb{E}[f(X)|Y]\right)\end{aligned} \tag{7}$$

By substituting (7) and (5) into (4), we have:

$$\|\tilde{B}\Phi(\Phi^{\mathrm{T}}\Phi)^{-\frac{1}{2}}\|_F^2 = \mathrm{tr}\left(\mathrm{cov}(f(X))^{-1}\mathrm{cov}(\mathbb{E}[f(X)|Y])\right) \tag{8}$$

Proof of Theorem 1

$$\begin{aligned} H(\boldsymbol{f}) &= H\left(\sum_{i=1}^{n} \alpha_i \cdot \boldsymbol{f}_i\right)\\ &= \mathrm{tr}\left(\mathrm{cov}\left(\mathbb{E}_{P_{X|Y}}\left[\sum_{i=1}^{n}\alpha_i\cdot \boldsymbol{f}_i(X)\Big|Y\right]\right)\right)\\ &= \mathrm{tr}\Bigg(\mathbb{E}_{P_Y}\Bigg[\sum_{i=1,j=1}^{n,n}\alpha_i\alpha_j\cdot \\ &\qquad \mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_i(X)|Y]\cdot \mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_j(X)|Y]^{\mathrm{T}}\Bigg]\Bigg) \qquad (9)\\ &= \sum_{i=1,j=1}^{n,n}\alpha_i\alpha_j\cdot \mathrm{tr}\Big(\mathbb{E}_{P_Y}\big[\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_i(X)|Y] \\ &\qquad \cdot \mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_j(X)|Y]^{\mathrm{T}}\big]\Big) \\ &= \sum_{i=1,j=1}^{n,n}\alpha_i\alpha_j\cdot\sum_{k=1}^{d}\Big(\mathbb{E}_{P_Y}\big[\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_i^{(k)}(X)|Y] \\ &\qquad\cdot \mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_j^{(k)}(X)|Y]\big]\Big)\end{aligned}$$

We then denote this quadratic form as $\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{\alpha}$ and prove that it is positive semi-definite. Firstly, we have $\boldsymbol{F} = \sum_{k=1}^{d}\boldsymbol{F}_k$, where k denotes feature dimension and $\{\boldsymbol{F}_k\}_{i,j} = \mathbb{E}_{P_Y}[\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_i^{(k)}(X)|Y]\cdot\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_j^{(k)}(X)|Y]]$.

For all k, consider $\forall \boldsymbol{b}\in\mathbb{R}^n$, we have:

$$\begin{aligned} \boldsymbol{b}^{\mathrm{T}}\boldsymbol{F}_k\boldsymbol{b} &= \sum_{i=1,j=1}^{n,n} b_i\mathbb{E}_{P_Y}\Big[\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_i^{(k)}(X)|Y] \\ &\qquad \cdot\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_j^{(k)}(X)|Y]\Big]b_j \\ &= \mathbb{E}_{P_Y}\left[\left(\sum_{j=1}^{n}\left(\sum_{i=1}^{n} b_i\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_i^{(k)}(X)|Y]\right)\right.\right. \\ &\qquad \left.\left.\cdot\,\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_j^{(k)}(X)|Y]b_j\right)\right] \\ &= \mathbb{E}_{P_Y}\left[\left(\sum_{i=1}^{n} b_i\cdot\mathbb{E}_{P_{X|Y}}[\boldsymbol{f}_i^{(k)}(X)|Y]\right)^2\right]\geq 0 \end{aligned} \tag{10}$$

Hence, $\boldsymbol{F}_k$ is positive semi-definite, and $\boldsymbol{F} = \sum_{k=1}^{d}\boldsymbol{F}_k$ is a sum of positive semi-definite matrix is also positive semi-definite. Therefore, $H(\boldsymbol{f}) = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{\alpha}$ is a convex quadratic form.

Proof of Theorem 2

Assumption 1. *We make the quadratic bowl assumption around the local minima θ^* on all domains: $\forall e\in\mathcal{E}$,*

$$\mathcal{R}_e(\theta) = \mathcal{R}_e(\theta^*) + \frac{1}{2}(\theta-\theta^*)^{\top}H_e(\theta-\theta^*), \tag{11}$$

where H_e is positive definite of eigenvalues $\lambda_1^e \geq \cdots \geq \lambda_h^e > 0$.

Remark 1. *Assumption 1 is milder on $N_{e,\theta^*}^{\epsilon}$ for low ϵ. Indeed, when $\epsilon\to 0$, then $\max_{\theta\in N_{e,\theta^*}^{\epsilon}}\|\theta-\theta^*\|_2^2 \to 0$ and the quadratic approximation coincides with the second-order Taylor expansion around θ^*. Moreover, this approximation is common in optimization (Schaul, Zhang, and LeCun 2013; Jastrzebski et al. 2017).*

Proposition 1. *Let $\epsilon>0$, weights θ^*. $\forall(A,B)\in\mathcal{E}^2$, with $N_{A,\theta^*}^{\epsilon}$ the largest path-connected region of weights space where the risk $\mathcal{R}_A$ remains in an ϵ interval around $\mathcal{R}_A(\theta^*)$, we note:*

$$\begin{aligned} \mathcal{I}^{\epsilon}(A,B) &= \max_{\theta\in N_{A,\theta^*}^{\epsilon}}|\mathcal{R}_B(\theta)-\mathcal{R}_A(\theta^*)|, \\ R(A,B) &= \mathcal{R}_B(\theta^*)-\mathcal{R}_A(\theta^*), \\ H^{\epsilon}(A,B) &= \max_{\frac{1}{2}(\theta-\theta^*)^{\top}H_A(\theta-\theta^*)\leq\epsilon}\frac{1}{2}(\theta-\theta^*)^{\top}H_B(\theta-\theta^*). \end{aligned} \tag{12}$$

If $\forall(A,B)\in\mathcal{E}^2$ such as $R(A,B)<0$, we have:

$$\epsilon \leq -R(A,B)\times\frac{\lambda_h^A}{\lambda_1^B}, \tag{13}$$

then under previous Assumption 1,

$$\max_{(A,B)\in\mathcal{E}^2}\mathcal{I}^{\epsilon}(A,B) = \max_{(A,B)\in\mathcal{E}^2}(R(A,B)+H^{\epsilon}(A,B)) \tag{14}$$

Proof. We first prove that, under quadratic Assumption 1, $\forall A\in\mathcal{E}, N_{A,\theta^*}^{\epsilon} = \{\theta | \, |\mathcal{R}_A(\theta)-\mathcal{R}_A(\theta^*)|\leq\epsilon\}$. Indeed, the former is always included in the latter by definition. Reciprocally, be given θ in the latter, $\{\lambda\theta^* + (1-\lambda)\theta \mid \lambda\in[0,1]\}$ linearly connects θ^* to θ in parameter space with the risk $\mathcal{R}_A$ remaining in an ϵ interval around $\mathcal{R}_A(\theta^*)$ because $\forall\mu\in[0,1]$ we have $|\mathcal{R}_A(\mu\theta^*+(1-\mu)\theta)-\mathcal{R}_A(\theta^*)| = (1-\mu)^2|\mathcal{R}_A(\theta)-\mathcal{R}_A(\theta^*)|\leq(1-\mu)^2\epsilon\leq\epsilon$.

Therefore $\forall(A,B)\in\mathcal{E}^2$:

$$\begin{aligned} \mathcal{I}^{\epsilon}(A,B) &= \max_{|\mathcal{R}_A(\theta)-\mathcal{R}_A(\theta^*)|\leq\epsilon}|\mathcal{R}_B(\theta)-\mathcal{R}_A(\theta^*)| \\ &= \max_{\substack{\frac{1}{2}(\theta-\theta^*)^{\top}H_A \\ (\theta-\theta^*)\leq\epsilon}}\left|R(A,B)+\frac{1}{2}(\theta-\theta^*)^{\top}H_B(\theta-\theta^*)\right| \end{aligned} \tag{15}$$

As the Hessians are positive, $H^{\epsilon}(A,B)>0$. We now need to split the analysis based on the sign of $R(A,B)$.

Case $R(A,B)\geq 0$

Both $R(A,B)$ and $H^{\epsilon}(A,B)$ are non-negative. Removing the absolute value from the RHS of Eq. 15 gives:

$$\mathcal{I}^{\epsilon}(A,B) = R(A,B)+H^{\epsilon}(A,B). \tag{16}$$

Taking the maximum over $(A,B)\in\mathcal{E}^2$ where $R(A,B)\geq 0$ gives:

$$\max_{\substack{(A,B)\in\mathcal{E}^2\\ R(A,B)\geq 0}}\mathcal{I}^{\epsilon}(A,B) = \max_{\substack{(A,B)\in\mathcal{E}^2\\ R(A,B)\geq 0}}\big(R(A,B)+H^{\epsilon}(A,B)\big). \tag{17}$$

Case $R(A,B) < 0$

Leveraging λ_1^B the largest eigenvalue from H_B and λ_h^A the lowest eigenvalue from H_A, we upper bound:

$$H^\epsilon(A,B) \leq \max_{\frac{\lambda_h^A}{2}\|\theta-\theta^*\|_2^2 \leq \epsilon} \frac{\lambda_1^B}{2}\|\theta-\theta^*\|_2^2 = \epsilon \times \frac{\lambda_1^B}{\lambda_h^A}. \quad (18)$$

Then Eq. 13 gives $H^\epsilon(A,B) < -R(A,B)$. Thus the number inside the absolute value from the RHS of Eq. 15 is negative. This leads to: $\mathcal{I}^\epsilon(A,B) = -R(A,B) - H^\epsilon(A,B)$ $< -R(A,B) = R(B,A) < \mathcal{I}^\epsilon(B,A)$. Thus the max over $\mathcal{E}^2$ of function $(A,B) \rightarrow \mathcal{I}^\epsilon(A,B)$ cannot be achieved for (A,B) with $R(A,B) < 0$. We obtain:

$$\max_{(A,B)\in\mathcal{E}^2} \mathcal{I}^\epsilon(A,B) = \max_{(A,B)\in\mathcal{E}^2|R(A,B)\geq 0} \mathcal{I}^\epsilon(A,B). \quad (19)$$

Similarly, $R(A,B) + H^\epsilon(A,B) \leq 0 < R(B,A) + H^\epsilon(B,A)$. Thus the max over $\mathcal{E}^2$ of function $(A,B) \rightarrow (R(A,B) + H^\epsilon(A,B))$ cannot be achieved for (A,B) with $R(A,B) < 0$. We obtain:

$$\begin{aligned} &\max_{(A,B)\in\mathcal{E}^2}(R(A,B) + H^\epsilon(A,B)) \\ &= \max_{\substack{(A,B)\in\mathcal{E}^2 \\ R(A,B)\geq 0}}(R(A,B) + H^\epsilon(A,B)) \end{aligned} \quad (20)$$

Combining Eq. 17, Eq. 19 and Eq. 20, we conclude the proof.

□

$\mathcal{L}_{\textbf{align}}$ matches the domain-level Hessians

The Hessian matrix $H = \sum_{i=1}^n \nabla_\theta^2 \ell\left(f_\theta\left(\boldsymbol{x}^i\right), \boldsymbol{y}^i\right)$ is of key importance in deep learning. Yet, H cannot be computed efficiently in general. In contrast, we use the fact that the diagonal of H is approximated by the gradient variance $\nabla\text{Var}(\boldsymbol{G})$.

The Hessian and the 'true' Fisher Information Matrix (FIM). The 'true' FIM $F = \sum_{i=1}^n \mathbb{E}_{\hat{\boldsymbol{y}}\sim P_\theta(\cdot|\boldsymbol{x}^i)}\left[\nabla_\theta \log p_\theta(\hat{\boldsymbol{y}} \mid \boldsymbol{x}^i)\nabla_\theta \log p_\theta(\hat{\boldsymbol{y}} \mid \boldsymbol{x}^i)^\top\right]$ approximates the Hessian H.

The 'true' FIM and the 'empirical' FIM. Yet, $\boldsymbol{F}$ remains costly as it demands one backpropagation per class. That's why most empirical works (e.g., (Dangel, Tatzel, and Hennig 2021)) approximate the 'true' FIM $\boldsymbol{F}$ with the 'empirical' FIM $\tilde{\boldsymbol{F}} = \boldsymbol{G}_e^\top \boldsymbol{G}_e = \sum_{i=1}^n \nabla_\theta \log p_\theta(\boldsymbol{y}^i|\boldsymbol{x}^i)\nabla_\theta \log p_\theta(\boldsymbol{y}^i|\boldsymbol{x}^i)^\top$ (Martens 2020) where $p_\theta(\cdot|\boldsymbol{x})$ is the density predicted by f_θ on input $\boldsymbol{x}$. While $\boldsymbol{F}$ uses the model distribution $P_\theta(\cdot|X)$, $\tilde{\boldsymbol{F}}$ uses the data distribution $P(Y|X)$. Despite this key difference, $\tilde{\boldsymbol{F}}$ and $\boldsymbol{F}$ were shown to share the same structure and to be similar up to a scalar factor (Thomas et al. 2020).

The 'empirical' FIM and the gradient covariance. Critically, $\tilde{\boldsymbol{F}}$ is nothing else than the unnormalized uncentered covariance matrix when ℓ is the negative log-likelihood. Thus, the gradient covariance matrix $\boldsymbol{C} = \frac{1}{n-1}\left(\boldsymbol{G}^\top\boldsymbol{G} - \frac{1}{n}\left(\mathbf{1}^\top\boldsymbol{G}\right)^\top\left(\mathbf{1}^\top\boldsymbol{G}\right)\right)$ of size $|\theta| \times |\theta|$ and $\tilde{\boldsymbol{F}}$ are equivalent (up to the multiplicative constant n) at any first-order stationary point: $\boldsymbol{C} \propto \tilde{\boldsymbol{F}}$. Overall, this suggests that $\boldsymbol{C}$ and $\boldsymbol{H}$ are closely related (Jastrzebski et al. 2017).

Critically, our regularization operates on the gradient variance $\text{Var}(\mathbf{G})$, which corresponds to the diagonal elements of the gradient covariance matrix $\mathbf{C}$. Let $\mathbf{a} = (\theta, p)^\top$ denote the compound parameter vector, where θ represents fixed model parameters and p corresponds to trainable prompt embeddings. The gradient covariance matrix is defined as:

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{\theta\theta} & \boldsymbol{C}_{\theta p} \\ \boldsymbol{C}_{p\theta} & \boldsymbol{C}_{pp} \end{bmatrix} = \begin{bmatrix} \text{Var}\left(\nabla_\theta f\right) & \text{Cov}\left(\nabla_\theta f, \nabla_p f\right) \\ \text{Cov}\left(\nabla_p f, \nabla_\theta f\right) & \text{Var}\left(\nabla_p f\right) \end{bmatrix}. \quad (21)$$

Under the constraint that θ is fixed, this matrix simplifies to:

$$\boldsymbol{C} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_{pp} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \text{Var}\left(\nabla_p f\right) \end{bmatrix}, \quad (22)$$

since $\text{Var}\left(\nabla_\theta f\right) = \mathbf{0}$ and all covariance terms vanish due to the constancy of θ-gradients. The trace of this covariance matrix therefore reduces to:

$$\text{tr}\left(\boldsymbol{C}\right) = \text{tr}\left(\text{Var}\left(\nabla_p f\right)\right). \quad (23)$$

This simplification demonstrates that gradient alignment in the original composite parameter space (θ, p) reduces precisely to considering gradient variance in the lower-dimensional prompt embedding space p, significantly decreasing optimization complexity. By explicitly aligning these diagonal components across tasks, we implicitly enforce consistency in the Hessian diagonal of the optimization landscape near local optima.

Appendix. II - Experimental Results

Ablation Study

The extended ablation results across other datasets are provided in the Tab. 1. Both components individually improve upon the baseline in every task domain. The combined approach maintains its superiority in all cases.

Evaluation on Prompt Weights

Supplementary results across other task domains confirm the patterns observed in Fig. 1. Our weight calculation method maintains superior Spearman correlation compared to PANDA and SPoT in most tested domains. These comprehensive results validate our method's robustness in capturing genuine task affinities across diverse domains.

Performance Scaling with Source Prompts Number

Results of most target domains demonstrate the scaling pattern in Fig. 2, with our method maintaining performance advantages over PANDA and SPOT. The efficiency ceiling varies by domain, and notably, more complex tasks exhibit significantly greater performance gains as the number of prompts increases – indicating enhanced knowledge transfer from pretrained source prompts. Crucially, our approach consistently achieves saturation at considerably higher accuracy levels than baseline methods across most tested scenarios.

Table 1: Ablation Study on Framework Components.

$\mathbf{H}(\alpha)$	$\mathcal{L}_{\text{align}}$	Flo	SVHN	DML	sNO-A	sNO-E	dS-L	dS-O	Cle-C	Cle-D	**Avg**
×	×	97.0	66.5	34.3	16.1	20.8	63.4	35.7	38.2	48.5	46.7
✓	×	97.9	68.5	35.7	18.6	23.1	66.3	36.2	42.8	50.9	48.9
×	✓	97.5	67.1	36.9	17.2	22.8	65.9	35.8	42.1	50.3	48.4
✓	✓	**98.1**	**71.0**	**38.1**	**20.3**	**24.9**	**68.1**	**40.4**	**49.3**	**53.5**	**51.5**

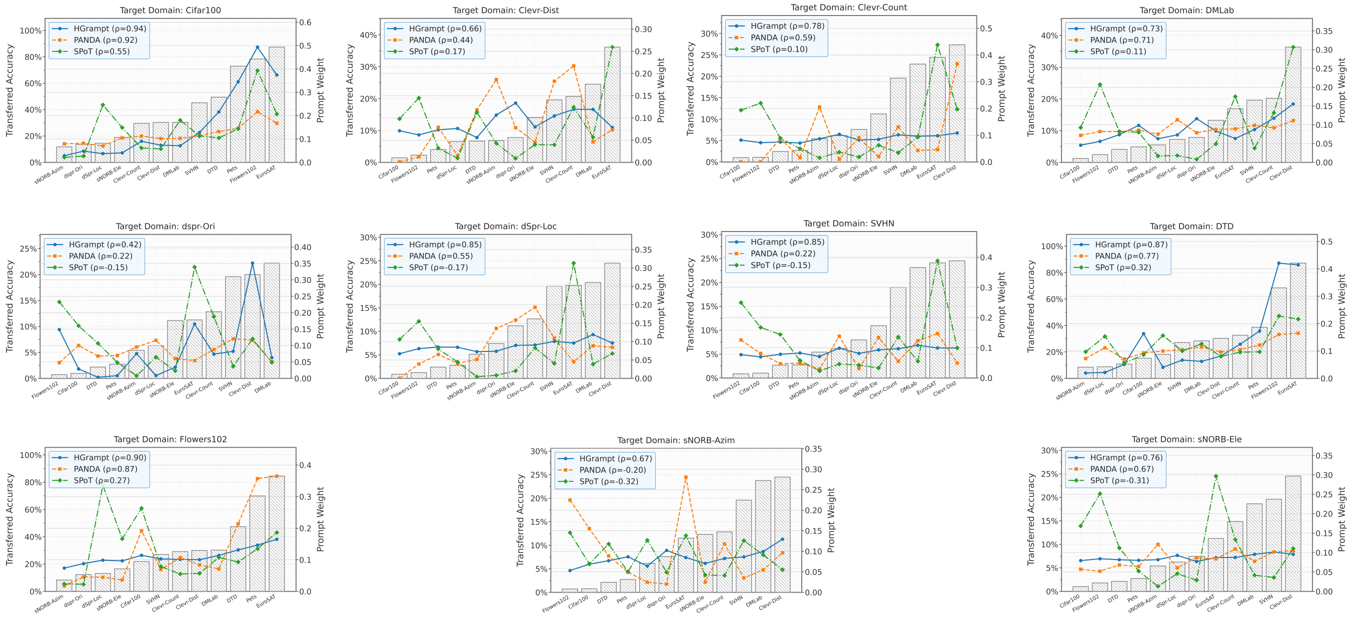

Figure 1: Prompt Weights analysis for 12 source prompts. Bar plots represent single-source transfer accuracy (left axis), while line plots indicate prompt weights (right axis).

Parameter Analysis

The full sensitivity analysis is provided in the Tab. 2. We analyze how the objective balancing influences results, with implementation noting that instead of the formulation in Eq. 10, we maintain proportional equivalence between objectives due to their differing scales. Performance trends reveal optimal outcomes when $\lambda = 1$, indicating mutual reinforcement between feature transferability and gradient alignment.

Table 2: Regularization coefficient performance.

Dataset	**Regularization coefficient** λ						
	0.1	0.2	0.5	1.0	2.0	5.0	10.0
Cifar100	72.1	74.0	75.8	**75.9**	75.8	75.5	74.2
DTD	63.0	64.0	63.5	**64.2**	63.1	63.0	62.9
Flowers102	94.5	96.0	97.3	**98.1**	97.6	96.4	95.5
Pets	87.3	86.9	86.3	87.4	**87.7**	87.1	84.7
SVHN	67.5	70.5	70.2	**71.0**	70.9	70.8	69.3
EuroSAT	89.5	91.9	91.4	**92.6**	92.3	91.0	88.5
DMLab	34.6	36.5	37.8	**38.1**	37.6	36.7	37.4
sNORB-Azim	19.2	20.0	19.6	**20.3**	19.7	19.7	17.2
sNORB-Ele	22.5	24.2	24.5	**24.9**	24.0	23.1	22.3
dSpr-Loc	66.6	67.4	67.3	**68.1**	67.4	66.9	65.8
dSpr-Ori	39.1	38.8	39.3	**40.4**	40.1	40.1	36.7
Clevr-Count	48.4	48.0	48.4	**49.3**	48.3	49.0	49.0
Clevr-Dist	**53.8**	51.4	53.1	53.5	52.8	53.2	51.5

Appendix. III - Visualization of Datasets

CIFAR-100 The CIFAR-100 dataset shown in Fig. 3 contains 100 classes of common objects organized into 20 superclasses, featuring diverse categories such as aquatic mammals (beaver, dolphin, otter, seal, whale), fish species (aquarium fish, flatfish, ray, shark, trout), flowers (orchids, poppies, roses, sunflowers, tulips), food containers (bottles, bowls, cans, cups, plates), and various vehicles (bicycle, bus, motorcycle, train).

DTD The Describable Textures Dataset shown in Fig. 4 comprises 5,640 texture images categorized into 47 human-perceived texture attributes. This collection captures a wide

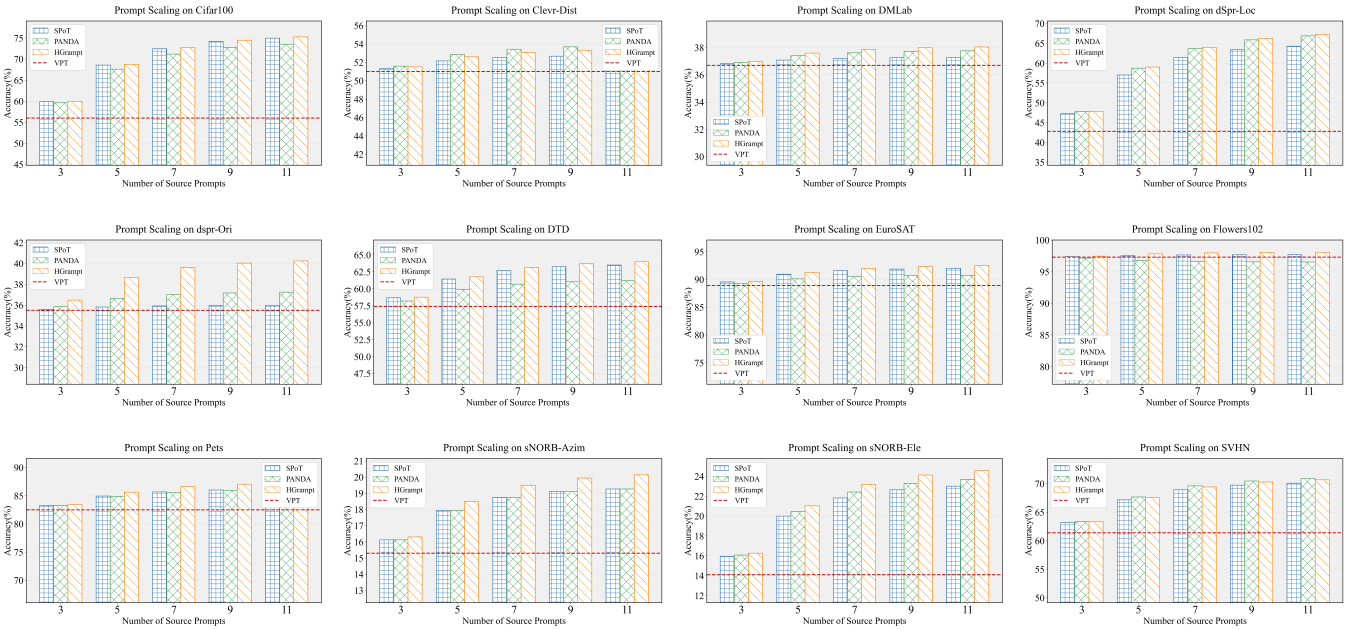

Figure 2: Performance scaling with increasing source prompts.

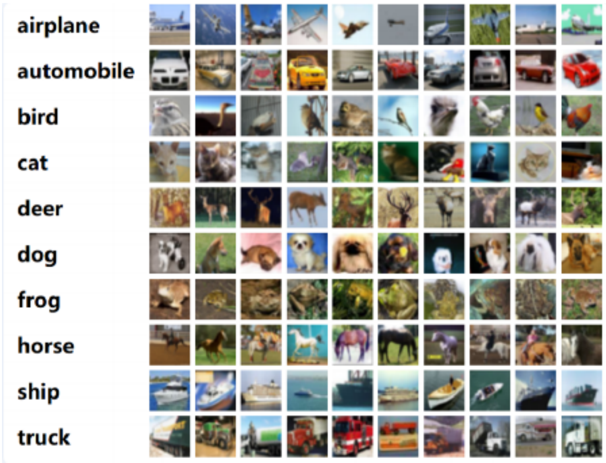

Figure 3: Representative images from CIFAR-100 showing diverse object categories.

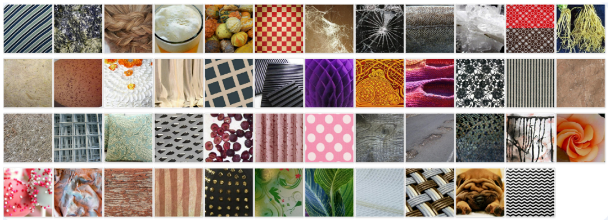

Figure 4: Representative samples from DTD.

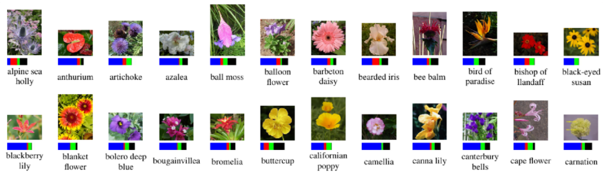

Figure 5: Varieties of floral specimens in the Flowers102 dataset.

spectrum of visual textures in natural environments, providing a challenging benchmark for material recognition tasks.

Flowers102 Developed by the University of Oxford, the Flowers102 dataset shown in Fig. 5 contains 8,189 images across 102 flower species, providing a comprehensive collection for fine-grained visual categorization of botanical specimens under varying photographic conditions.

Pets The Pets dataset shown in Fig. 6 features 7,349 images across 37 distinct cat and dog breeds, offering a challenging benchmark for fine-grained breed recognition with significant intra-class variation in animal appearance, pose, and imaging conditions.

SVHN Collected from Google Street View imagery, the SVHN dataset shown in Fig. 7 contains over 600,000 digit images for number recognition tasks, available in both full-digit and cropped 32×32 formats, capturing digits in diverse real-world contexts with complex backgrounds.

EuroSAT Based on Sentinel-2 satellite imagery, EuroSAT shown in Fig. 8 provides 27,000 high-resolution (10m) images across 10 land cover categories including agricultural areas, forests, grasslands, and urban zones, covering diverse geographical regions throughout Europe.

Figure 6: Representative cat and dog breeds from the Oxford-IIIT Pets dataset.

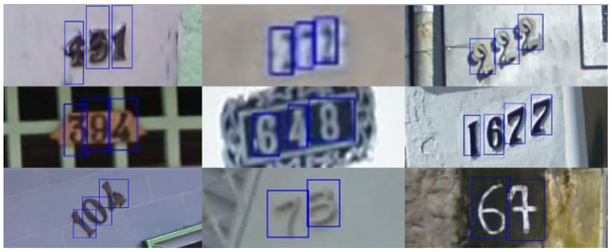

Figure 7: Naturalistic digit images from the SVHN dataset.

CLEVR CLEVR shown in Fig. 9 is a synthetic visual reasoning dataset featuring procedurally generated scenes with geometric objects and associated compositional questions, with specialized VTAB tasks including object counting (CLEVR Count: 3-10 objects) and depth estimation (CLEVR Distance: 6 distance intervals).

DMLab The DeepMind Lab environment shown in Fig. 10 provides 3D navigation frames annotated with distance relationships to objects, featuring six interaction categories based on proximity to fruits (apples, melons, lemons) at three distance levels: nearby, far, and very far.

dSprites dSprites shown in Fig. 11 is a 2D shape dataset with 737,280 binary images generated from six disentangled factors (color, shape, scale, rotation, position), with VTAB tasks focusing on location prediction (16 horizontal position bins) and orientation estimation (16 rotation bins).

SmallNORB SmallNORB shown in Fig. 12 contains 48,600 stereo images of 50 toys across 5 categories, captured under controlled variations including 6 lighting conditions, 9 elevation angles (30-70°), and 18 azimuth angles (0-340°), with VTAB tasks for azimuth prediction (18 categories) and elevation estimation (9 categories).

Figure 8: Land cover classification examples from EuroSAT.

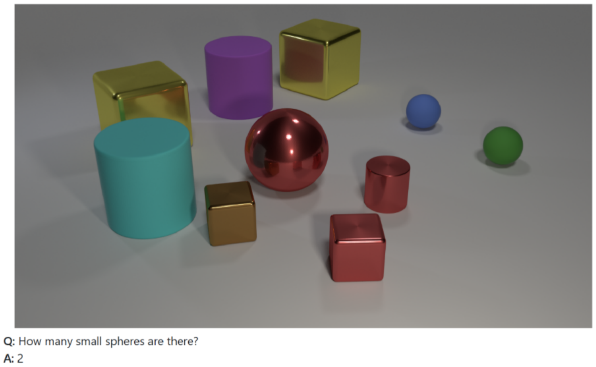

Figure 9: Compositional scenes and reasoning tasks in CLEVR.

References

Dangel, F.; Tatzel, L.; and Hennig, P. 2021. ViViT: Curvature access through the generalized Gauss-Newton's low-rank structure. *arXiv preprint arXiv:2106.02624*.

Huang, S.-L.; Makur, A.; Wornell, G. W.; and Zheng, L. 2019. On universal features for high-dimensional learning and inference. *arXiv preprint arXiv:1911.09105*.

Jastrzebski, S.; Kenton, Z.; Arpit, D.; Ballas, N.; Fischer, A.; Bengio, Y.; and Storkey, A. 2017. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*.

Martens, J. 2020. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146): 1–76.

Schaul, T.; Zhang, S.; and LeCun, Y. 2013. No more pesky learning rates. In *International conference on machine learning*, 343–351. PMLR.

Thomas, V.; Pedregosa, F.; Merriënboer, B.; Manzagol, P.-A.; Bengio, Y.; and Le Roux, N. 2020. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, 3503–3513. PMLR.

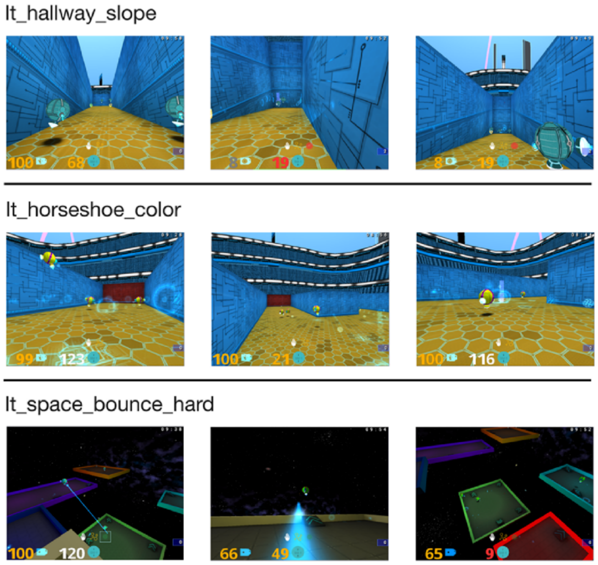

Figure 10: First-person perspective in DMLab's 3D environment.

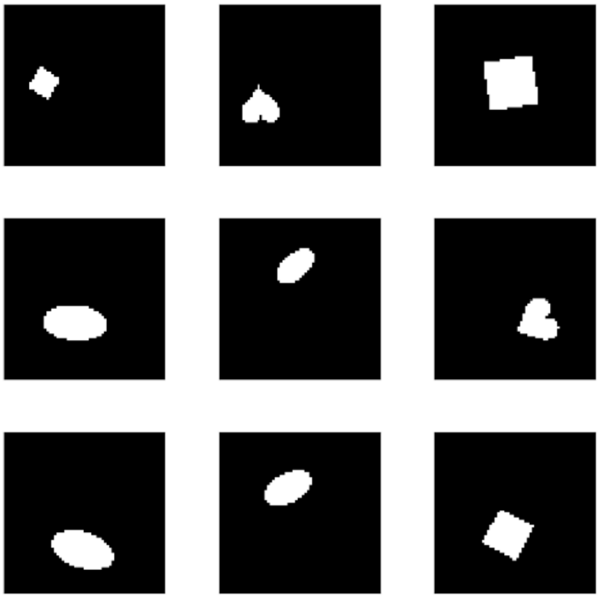

Figure 11: Shape variations in dSprites demonstrating disentangled factors.

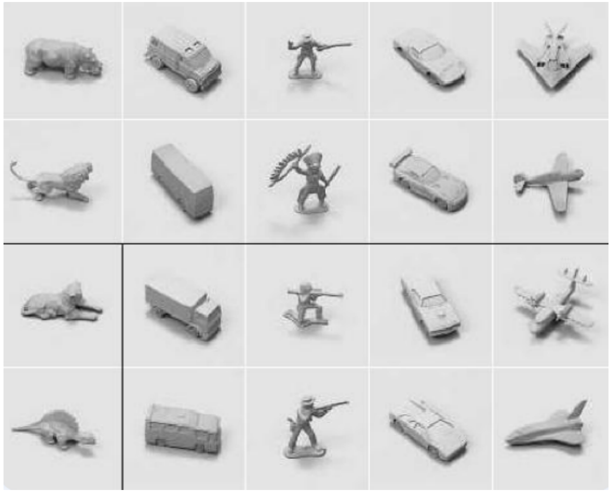

Figure 12: Object variations under different poses in Small-NORB.