A Retrospect to Multi-prompt Learning across Vision and Language (Supplementary Material)

Ziliang Chen[1,3], Xin Huang[2], Quanlong Guan[1], Liang Lin[2] Weiqi Luo[1]
[1]Jinan University [2]Sun Yat-sen University [3]Pazhou Laboratory
{yanxp3,wangxx35}@mail2.sysu.edu.cn, c.ziliang@yahoo.com, 466783266@qq.com, xdliang328@gmail.com, linliang@ieee.org

A.1. Cross-modal Non-identifiability

In order to present the existence proof of cross-modal non-identifiability issue, we provide a rigorous definition to describe the constant modality gap at the individual level:

Definition 1. ***(Individual ϵ-approximate constant modality gap)*** *Suppose that $f(\cdot)$ and $h_{\boldsymbol{v}}(\cdot)$ denote a visual encoder and a text encoder with prompt template $\boldsymbol{v}$, respectively, and a denumerable set of images $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ are consistent with the concept set C. A constant vector $\boldsymbol{c}$ denotes the modality gap of $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ with repect to $f(\cdot)$ and $h_{\boldsymbol{v}}(\cdot)$, if for all arbitary small scalar $\epsilon > 0$ and arbitary concepts in C, there exists a positive integer $N(\epsilon)$, so that given arbitary positive integers j satisfying $j > N(\epsilon)$, we have a permutation $\{\boldsymbol{x}^{(i)}\}_{i=1}^{\infty}$ of $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ which satisfies $|f(\boldsymbol{x}^{(j)})-h_{\boldsymbol{v}}(c)-\boldsymbol{c}|<\epsilon$ for arbitrary $c \in C$.*

The definition does not enforce each pair of image feature and prompt embedding exactly suits the contant modality gap as [6]. Instead, it employs the ϵ-approximation language of an infinite sequence to define the gap maintained in a range with a arbitrary small deviation. It suits the practice with images drawn from a nature distribution.

Proposition 1. ***Individual cross-modal non-identifiability.*** *Suppose a single-prompt learning model $\big(f(\cdot), h_{\boldsymbol{v}}(\cdot)\big)$ defines the individual ϵ-approximate constant modality gap. Given a infinite denumerable set of images X_1 with concepts in C_1 and the other infinite denumerable set of images X_2 with concepts in C_2 ($C_1/C_2 \neq \emptyset$ and $C_1 \cap C_2 \neq \emptyset$), for an image $\boldsymbol{x}_i \in X_1$ with the ϵ-approximation to the constant modality gap $\boldsymbol{v}$ by $h_{\boldsymbol{v}}(c)$, there exists an image $\boldsymbol{x}_j \in X_2$ with the identical ϵ-approximation to the constant modality gap $\boldsymbol{v}$ by $h_{\boldsymbol{v}}(c)$, for arbitrary $c \in C_1/C_2$.*

Proof. Consider two concept sets C_1, C_2, their corresponding image sets X_1, X_2, as well as two concepts $c_1 \in C_1/C_2$, $c_2 \in C_1 \cap C_2$. From the definition above we know that for an arbitary small $\frac{1}{3}\epsilon > 0$, there is a positive integer $N_1(\frac{1}{3}\epsilon)$, so given an arbitary positive integer k satisfying $k > N_1(\frac{1}{3}\epsilon)$, we have a permutation $\{\boldsymbol{x}^{(i_1)}\}_{i_1=1}^{\infty}$ from X_1 which satisfies $|f(\boldsymbol{x}^{(k)})-h_{\boldsymbol{v}}(c_1)-\boldsymbol{c}|<\frac{1}{3}\epsilon$, given $\boldsymbol{x}^{(k)} \in \{\boldsymbol{x}^{(i_1)}\}_{i_1=1}^{\infty}$. Here we focus on $\{\boldsymbol{x}^{(i_1)}\}_{i_1>N_1(\frac{1}{3}\epsilon)}^{\infty}$, the subchain of $\{\boldsymbol{x}^{(i_1)}\}_{i_1=1}^{\infty}$ consistent with the concepts in C_1. Obviously, for $c_2 \in C_1 \cap C_2 \subset C_1$, there also exists an integer $N_2(\frac{1}{3}\epsilon) > 0$, so given an arbitary positive integer k satisfying $k > N_2(\frac{1}{3}\epsilon)$, we may take a repermutation $\{\boldsymbol{x}^{(i_1,i_2)}\}_{i_1>N_1(\frac{1}{3}\epsilon),i_2=1}^{\infty}$ from the chain $\{\boldsymbol{x}^{(i_1)}\}_{i_1>N_1(\frac{1}{3}\epsilon)}^{\infty}$ which satisfies $|f(\boldsymbol{x}^{(k)})-h_{\boldsymbol{v}}(c_2)-\boldsymbol{c}|<\frac{1}{3}\epsilon$, given $\boldsymbol{x}^{(k)} \in \{\boldsymbol{x}^{(i_1,i_2)}\}_{i_1>N_1(\frac{1}{3}\epsilon),i_2=1}^{\infty}$. In the other image set X_2, we may provide a similar deduction. Specifically, given $c_2 \in C_1 \cap C_2 \subset C_2$, for an arbitary small $\frac{1}{3}\epsilon$, it refers to a positive integer $N_3(\frac{1}{3}\epsilon)$, so given an arbitary integer k satisfying $k > N_3(\frac{1}{3}\epsilon)$, we may take a permutation $\{\boldsymbol{x}^{(i_3)}\}_{i_3=1}^{\infty}$ from X_2 which satisfies $|f(\boldsymbol{x}^{(k)})-h_{\boldsymbol{v}}(c_2)-\boldsymbol{c}|<\frac{1}{3}\epsilon$, given $\boldsymbol{x}^{(k)} \in \{\boldsymbol{x}^{(i_3)}\}_{i_3=1}^{\infty}$.

Based on the subchain construction, we consider an image $\boldsymbol{x}_i$ drawn from the subchain $\{\boldsymbol{x}^{(i_1,i_2)}\}_{i_1>N_1(\frac{1}{3}\epsilon),i_2>N_2(\frac{1}{3}\epsilon)}^{\infty}$ and another image $\boldsymbol{x}_j$ drawn from $\{\boldsymbol{x}^{(i_3)}\}_{i_3>N_3(\frac{1}{3}\epsilon)}^{\infty}$. Obviously, they satisfy

$$\begin{aligned} |f(\boldsymbol{x}_i) - h_{\boldsymbol{v}}(c_1) - \boldsymbol{c}| &< \frac{1}{3}\epsilon; \\ |f(\boldsymbol{x}_i) - h_{\boldsymbol{v}}(c_2) - \boldsymbol{c}| &< \frac{1}{3}\epsilon; \\ |f(\boldsymbol{x}_j) - h_{\boldsymbol{v}}(c_2) - \boldsymbol{c}| &< \frac{1}{3}\epsilon. \end{aligned} \tag{1}$$

While they also result in

$$\begin{aligned} &|f(\boldsymbol{x}_j) - h_{\boldsymbol{v}}(c_1) - \boldsymbol{c}| \\ =\ &|f(\boldsymbol{x}_i) - h_{\boldsymbol{v}}(c_1) - \boldsymbol{c} + \boldsymbol{c} - f(\boldsymbol{x}_i) \\ &\quad + h_{\boldsymbol{v}}(c_2) + f(\boldsymbol{x}_j) - h_{\boldsymbol{v}}(c_2) - \boldsymbol{c}| \\ \leq\ &|f(\boldsymbol{x}_i) - h_{\boldsymbol{v}}(c_1) - \boldsymbol{c}| + |f(\boldsymbol{x}_i) - h_{\boldsymbol{v}}(c_2) \\ &\quad - \boldsymbol{c}| + |f(\boldsymbol{x}_j) - h_{\boldsymbol{v}}(c_2) - \boldsymbol{c}| \\ <\ &\frac{1}{3}\epsilon + \frac{1}{3}\epsilon + \frac{1}{3}\epsilon = \epsilon. \end{aligned} \tag{2}$$

In other words, we also have a denumerable set of images with an arbitrary mutual concept in $C_1 \cap C_2$ whereas can be

inquired with an exclusive concept in C_1/C_2 in terms of the approximation result. Note that, since ϵ is a arbitrary small number, so given each pair of $(f(\boldsymbol{x}_i), h_{\boldsymbol{v}}(c_1))$ with $\boldsymbol{x}_i \in X_1$ that satisfy the ϵ-approximation with the assumption, we are able to find the other pair $(f(\boldsymbol{x}_i^{(k)}), h_{\boldsymbol{v}}(c_1))$ with $\boldsymbol{x}_i^{(k)} \in X_1$ that satisfy the $\frac{1}{3}\epsilon$-approximation and through the reduction above, it results in a image $\boldsymbol{x}_j \in X_2$ which also satisfies the ϵ-approximation with $h_{\boldsymbol{v}}(c_1)$. □

From this observation, images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ distint by the exclusive concept c_1 are impossible to be identified by the single-prompt learning model $\big(f(\cdot), h_{\boldsymbol{v}}(c_1)\big)$.

We can also extend the ϵ-approximate constant modality gap from the individual level to the population level:

Definition 2. ***(Population ϵ-approximate constant modality gap)*** *Suppose that $f(\cdot)$ and $h_{\boldsymbol{v}}(\cdot)$ denote a visual encoder and a text encoder with prompt template $\boldsymbol{v}$, respectively, and a denumerable set of images $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ are consistent with the concept set C. A constant vector $\mathbf{c}$ denotes the modality gap of $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ with repect to $f(\cdot)$ and $h_{\boldsymbol{v}}(\cdot)$, if for all arbitary small scalar $\epsilon > 0$ and arbitary concepts in C, there exists a positive integer $N(\epsilon)$, so that we have a permutation $\{\boldsymbol{x}^{(i)}\}_{i=1}^{\infty}$ of $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ in which the subchain $\{\boldsymbol{x}^{(i)}\}_{i>N(\epsilon)}^{\infty}$ satisfies $|\sum_{\boldsymbol{x}\in\boldsymbol{X}} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c) - \mathbf{c}| < \epsilon$ for an arbitrary concept $c \in C$ and $\boldsymbol{X} \subseteq \{\boldsymbol{x}^{(i)}\}_{i>N(\epsilon)}^{\infty}$.*

From the definition we further derive the formal statment of population-level cross-modal non-identifiability:

Proposition 2. ***Population cross-modal non-identifiability.*** *Suppose a single-prompt learner $\big(f(\cdot), h_{\boldsymbol{v}}(\cdot)\big)$ defines the population ϵ-approximate constant modality gap. Given an image group $\boldsymbol{X}$ consistent with the concept set C_1 that satisfies the ϵ-approximate constant modality gap by $h_{\boldsymbol{v}}(c)$, we are able to construct an image group $\boldsymbol{X}'$ consistent with the concept set C_2 that satisfies the ϵ-approximate constant modality gap by $h_{\boldsymbol{v}}(c)$, in which $c \in C_1/C_2$.*

Proof. Simlarly with the proof for the individual-level non-identifiability, for an arbitary small $\frac{1}{3}\epsilon > 0$, we use a integer $N_1(\frac{1}{3}\epsilon)$ to construct the subchain $\{\boldsymbol{x}^{(i_1)}\}_{i>N_1(\frac{1}{3}\epsilon)}^{\infty} \subset X_{C_1}$ where $|\sum_{\boldsymbol{x}\in\boldsymbol{X}_1} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_1) - \mathbf{c}| < \frac{1}{3}\epsilon$ is satisfied for $c_1 \in C_1/C_2 \subset C_1$ with $\forall\ \boldsymbol{X}_1 \subseteq \{\boldsymbol{x}^{(i_1)}\}_{i_1>N_1(\frac{1}{3}\epsilon)}^{\infty}$. Deriving from $\{\boldsymbol{x}^{(i_1)}\}_{i_1>N_1(\frac{1}{3}\epsilon)}^{\infty}$, we are permitted to take its pumutation to construct $\{\boldsymbol{x}^{(i_1,i_2)}\}_{i_1>N_1(\frac{1}{3}\epsilon),i_2=1}^{\infty}$. Therefore with a integer $N_2(\frac{1}{3}\epsilon) > 0$, we could also extract a subset $\boldsymbol{X}_2$ from $\{\boldsymbol{x}^{(i_1,i_2)}\}_{i_1>N_1(\frac{1}{3}\epsilon),i_2>N_2(\frac{1}{3}\epsilon)}^{\infty}$ to encourage the similar $\frac{1}{3}\epsilon$-approximation $|\sum_{\boldsymbol{x}\in\boldsymbol{X}_2} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_2) - \mathbf{c}| < \frac{1}{3}\epsilon$, with respect to $c_2 \in C_1 \cap C_2 \subset C_1$. Obviously, we can directly extract a subset $\boldsymbol{X}_3$ from the images that are consistent with the concept set C_2: we may take an integer $N_3(\frac{1}{3}\epsilon) > 0$ to motivate the permutation $\{\boldsymbol{x}^{(i_1)}\}_{i_1>N_3(\frac{1}{3}\epsilon)}^{\infty} \subset X_{C_2}$ so that provided with $\forall\ \boldsymbol{X}_3 \subseteq \{\boldsymbol{x}^{(i_1)}\}_{i_1>N_3(\frac{1}{3}\epsilon)}^{\infty}$, the $\frac{1}{3}\epsilon$-approximation $|\sum_{\boldsymbol{x}\in\boldsymbol{X}_3} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_2) - \mathbf{c}| < \frac{1}{3}\epsilon$ is also achieved.

Based upon the construction above, we now consider the intersection $\boldsymbol{X}_1 \cap \boldsymbol{X}_2$. It obviously satsifies

$$\begin{aligned}\Big|\sum_{\boldsymbol{x}\in\boldsymbol{X}_1\cap\boldsymbol{X}_2} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_1) - \mathbf{c}\Big| &< \frac{1}{3}\epsilon,\\ \Big|\sum_{\boldsymbol{x}\in\boldsymbol{X}_1\cap\boldsymbol{X}_2} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_2) - \mathbf{c}\Big| &< \frac{1}{3}\epsilon.\end{aligned} \tag{3}$$

With this regard, we cast an ϵ-approximation as follows

$$\begin{aligned}&\Big|\sum_{\boldsymbol{x}\in\boldsymbol{X}_3} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_1) - \mathbf{c}\Big|\\ =&\Big|\sum_{\boldsymbol{x}\in\boldsymbol{X}_1\cap\boldsymbol{X}_2} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_1) - \mathbf{c} + \mathbf{c} - \sum_{\boldsymbol{x}\in\boldsymbol{X}_1\cap\boldsymbol{X}_2} f(\boldsymbol{x})\\ &\qquad + h_{\boldsymbol{v}}(c_2) + \sum_{\boldsymbol{x}\in\boldsymbol{X}_3} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_2) - \mathbf{c}\Big|\\ \le&\Big|\sum_{\boldsymbol{x}\in\boldsymbol{X}_1\cap\boldsymbol{X}_2} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_1) - \mathbf{c}\Big| + \Big|\sum_{\boldsymbol{x}\in\boldsymbol{X}_1\cap\boldsymbol{X}_2} f(\boldsymbol{x})\\ &\qquad - h_{\boldsymbol{v}}(c_2) - \mathbf{c}\Big| + \Big|\sum_{\boldsymbol{x}\in\boldsymbol{X}_3} f(\boldsymbol{x}) - h_{\boldsymbol{v}}(c_2) - \mathbf{c}\Big|\\ <&\frac{1}{3}\epsilon + \frac{1}{3}\epsilon + \frac{1}{3}\epsilon = \epsilon\end{aligned}$$

So given a image group $\boldsymbol{X}$ consistent with the concept set C_1 that satisfies the ϵ-approximation with the assumption by $h_{\boldsymbol{v}}(c_1)$, we may construct the reduction above to obtain $\boldsymbol{X}' = \boldsymbol{X}_3$. Although $\boldsymbol{X}'$ are consistent with the concept set C_2, it also leads to the ϵ-approximation with the assumption by $h_{\boldsymbol{v}}(c_1)$ ($c_1 \in C_1/C_2$). □

From this observation, the image groups $\boldsymbol{X}$ and $\boldsymbol{X}'$ distint by the exclusive concept c_1 are impossible to be identified by the single-prompt learning model $\big(f(\cdot), h_{\boldsymbol{v}}(c_1)\big)$.

A.2. The Proof of Proposition.3

We present two lemmas to facilitate the proof of the third proposition. The first lemma showing that optimizing each task-specific objective in Eq.7 is equivalent to minimize the KL-divergence between the image-prompt joint distribution $p(X, H|\mathcal{T}_\mathbf{i})$ and another EBM-modeled image-prompt joint distribution $q_{\frac{(\mathrm{EBM})}{\phi}}(X, H|\mathcal{T}_\mathbf{i})$, *i.e.*,

$$\begin{aligned}E_\phi(X, H; \mathcal{T}_i) =&\log \sum_{c\sim U_i} P_{\boldsymbol{\phi}}(X, H)[c]\\ =&\log \sum_{c\sim\mathcal{U}_i} \frac{\exp\big(\frac{\mathrm{sim}(f(X), H(c;\mathcal{V}_i\cup\mathcal{U}_i))}{\gamma}\big)}{\sum_{i=1}^{K}\exp\big(\frac{\mathrm{sim}(f(X), H(c;\mathcal{V}_i\cup\mathcal{U}_i))}{\gamma}\big)},\end{aligned} \tag{4}$$

where the *auxiliary energy function* of $q_{\overline{\phi}}^{(\text{EBM})}(X, H|\mathcal{T}_\mathbf{i})$ is defined as

$$\begin{aligned} E_\phi^{(a)}(X, H; \mathcal{T}_\mathbf{i}) &= -\log \sum_{c\sim\mathcal{V}_i} P_{\boldsymbol{\phi}}(X, H)[c] \\ &= -\log \sum_{c\sim\mathcal{V}_i} \frac{\exp\left(\frac{\text{sim}(f(X), H(c;\mathcal{V}_i\cup\mathcal{U}_i))}{\gamma}\right)}{\sum_{i=1}^{K}\exp\left(\frac{\text{sim}(f(X), H(c;\mathcal{V}_i\cup\mathcal{U}_i))}{\gamma}\right)}, \end{aligned} \tag{5}$$

which is distinct from the energy function $E_\phi(X, H; \mathcal{T}_\mathbf{i})$ applied in EMPL. Given this, we derive the auxiliary objective

$$\begin{aligned} \min_\phi \mathbb{E}_{p(X,H|\mathcal{T}_\mathbf{i})}\Big[\sum_{c\sim V_i} -\log P_{\boldsymbol{\phi}}(X, H)[c]\Big] \\ - \mathbb{E}_{q_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\Big[E_\phi^{(a)}(X, H; \mathcal{T}_i)\Big]. \end{aligned} \tag{6}$$

Resembling the notations of EMPL objecitve

$$\begin{aligned} \min_\phi \mathbb{E}_{\mathcal{T}_i}\Bigg[\underbrace{\mathbb{E}_{p(X,H|\mathcal{T}_\mathbf{i})}\Big[\sum_{c\sim V_i} -\log P_{\boldsymbol{\phi}}(X, H)[c]\Big]}_{\text{Generic prompt learning goal}} \\ \underbrace{-\lambda\mathbb{E}_{p_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\Big[E_\phi(X, H; \mathcal{T}_i)\Big]}_{\text{EBM uncertainty modeling}}\Bigg], \end{aligned} \tag{7}$$

$\overline{\phi}$ in Eq.6 denotes the frozen parameters. We will show that optimizing the Eq.6 is equivalent with optimizing the task-specific objective of Eq.7, *i.e.*,

$$\begin{aligned} \min_\phi \mathbb{E}_{p(X,H|\mathcal{T}_\mathbf{i})}\Big[\sum_{c\sim V_i} -\log P_{\boldsymbol{\phi}}(X, H)[c]\Big] \\ - \lambda\mathbb{E}_{p_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\Big[E_\phi(X, H; \mathcal{T}_i)\Big]. \end{aligned} \tag{8}$$

Lemma 1. *Given two energy-based distributions $p_{\overline{\phi}}^{(\text{EBM})}$ and $q_{\overline{\phi}}^{(\text{EBM})}$ with their energy functions defined by Eq.4 and Eq.5, respectively. The optimization of Eq.8 essentially optimizes a combination of one K-way classification objective and Eq.6 with some co-efficient γ.*

Proof. Since the first term across Eq.4 and Eq.5 are identical to solve a K-way classification objective, we are only required to consider the equivalence between their second terms. In specific, we consider the gradient of the second term in Eq.6,

$$\begin{aligned} &\frac{\partial}{\partial\phi}\Big(-\mathbb{E}_{q_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_\mathbf{i})}[E_\phi^{(a)}(X, H; \mathcal{T}_i)]\Big) \\ =&\frac{\partial}{\partial\phi}\Big(\mathbb{E}_{q_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\log\sum_{c\sim\mathcal{V}_i} P_{\boldsymbol{\phi}}(X, H)[c]\Big) \\ =&\int_{\boldsymbol{x},\boldsymbol{h}} q_{\overline{\phi}}^{(\text{EBM})}(\boldsymbol{x}, \boldsymbol{h}|\mathcal{T}_\mathbf{i})\frac{\partial}{\partial\phi}\Big(\log\sum_{c\sim\mathcal{V}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]\Big) \\ =&\int_{\boldsymbol{x},\boldsymbol{h}} \frac{\sum_{c\sim\mathcal{V}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]}{Z^{(a)}(\phi)}\frac{\frac{\partial}{\partial\phi}\big(\sum_{c\sim\mathcal{V}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]\big)}{\sum_{c\sim\mathcal{V}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]} \\ =&\int_{\boldsymbol{x},\boldsymbol{h}} \frac{\frac{\partial}{\partial\phi}\big(\sum_{c\sim\mathcal{V}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]\big)}{Z^{(a)}(\phi)} \\ =&-\int_{\boldsymbol{x},\boldsymbol{h}} \frac{\frac{\partial}{\partial\phi}\big(\sum_{c\sim\mathcal{U}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]\big)}{Z^{(a)}(\phi)} \\ =&-\frac{Z(\phi)}{Z^{(a)}(\phi)}\int_{\boldsymbol{x},\boldsymbol{h}} \frac{\sum_{c\sim\mathcal{U}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]}{Z(\phi)}\frac{\frac{\partial}{\partial\phi}\big(\sum_{c\sim\mathcal{U}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]\big)}{\sum_{c\sim\mathcal{U}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]} \\ =&-\frac{Z(\phi)}{Z^{(a)}(\phi)}\int_{\boldsymbol{x},\boldsymbol{h}} p_{\overline{\phi}}^{(\text{EBM})}(\boldsymbol{x}, \boldsymbol{h}|\mathcal{T}_\mathbf{i})\frac{\partial}{\partial\phi}\log\Big(\sum_{c\sim\mathcal{U}_i} P_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{h})[c]\Big), \end{aligned} \tag{9}$$

in which $Z(\phi)$ and $Z^{(a)}(\phi)$ refer to the partition functions of the energy-based distributions $p_{\overline{\phi}}^{(\text{EBM})}$ and $q_{\overline{\phi}}^{(\text{EBM})}$, respectively. If we take the co-efficient $\gamma = \frac{Z(\phi)}{Z^{(\phi)}(\phi)}$, the gradient above can be restated as:

$$\begin{aligned} &\frac{\partial}{\partial\phi}\Big(-\mathbb{E}_{q_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_\mathbf{i})}[E_\phi^{(a)}(X, H; \mathcal{T}_i)]\Big) \\ =&-\gamma\mathbb{E}_{p_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\frac{\partial}{\partial\phi}\log\Big(\sum_{c\sim\mathcal{U}_i} P_{\boldsymbol{\phi}}(X, H)[c]\Big). \end{aligned} \tag{10}$$

It exactly refers to the gradient of the scecond term in Eq.8. □

Given this observation, if we further prove that optimizing Eq.6 is equivalent with minimizing the KL-divergence between $p(X, H|\mathcal{T}_\mathbf{i})$ and $q_{\overline{\phi}}^{(\text{EBM})}(X, H|\mathcal{T}_\mathbf{i})$, EMPL objective (Eq.8) composed by Eq.8, will also refer to the KL-divergence minimization due to the Lemma.1. It means that Eq.8 encourages $1-p_\phi(X, H|\mathcal{U}_i)=1-\sum_{c\sim\mathcal{U}_i} P_{\boldsymbol{\phi}}(X, H)[c]$ for matching the observed image-prompt joint distribution so that $p(X, H|\mathcal{T}_\mathbf{i})$ and $p_\phi(X, H|\mathcal{U}_i)$ negatively correlate with each other. To facilitate our proof, we recur the second lemma [3]:

Lemma 2. *Given a true image-prompt joint distribution $r(X, H)$, and an EBM-based image-prompt joint distribution $r_{\overline{\theta}}^{(\text{EBM})}(X, H)$ parameterized by ϕ with respect to the energy fuction $E_\theta(X, H)$. The objective that minimizes the KL-divergence $D_{\text{KL}}\Big(r(X, H)||r_{\overline{\theta}}^{(\text{EBM})}(X, H)\Big)$ can be op-*

timized via descending the gradient w.r.t. θ:

$$\mathop{\mathbb{E}}_{(\boldsymbol{x}^+,\boldsymbol{h}^+)\sim r}\left[\frac{\partial E_\theta(\boldsymbol{x}^+,\boldsymbol{h}^+)}{\partial\theta}\right]-\mathop{\mathbb{E}}_{(\boldsymbol{x}^-,\boldsymbol{h}^-)\sim r^{(EBM)}}\left[\frac{\partial E_\theta(\boldsymbol{x}^-,\boldsymbol{h}^-)}{\partial\theta}\right] \tag{11}$$

The proof is derived from the theoretical analysis in [3]. The first term decreases the energy of image-prompt pairs drawn from the true distribution while the second term increases the energy of image-prompt pairs drawn from the energy distribution.

Combined with the previous analysis, we provide the formal proof of Proposition.3:

Proof. In terms of Lemma.2, the KL-divergence minimization between $p(X,H|\mathcal{T}_\mathbf{i})$ and $q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})$ refers to descending the gradient in Eq.11. With ϕ replaced with a fixed $\overline{\phi}$, the gradient can be reformulated to

$$\mathop{\mathbb{E}}_{(\boldsymbol{x}^+,\boldsymbol{h}^+)\sim p(X,H|\mathcal{T}_\mathbf{i})}\left[\frac{\partial E_\phi^{(a)}(\boldsymbol{x}^+,\boldsymbol{h}^+)}{\partial\phi}\right] - \mathop{\mathbb{E}}_{(\boldsymbol{x}^-,\boldsymbol{h}^-)\sim q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\left[\frac{\partial E_\phi^{(a)}(\boldsymbol{x}^-,\boldsymbol{h}^-)}{\partial\phi}\right] \tag{12}$$

So we can achieve it by minimizing the KL-divergence between $p(X,H|\mathcal{T}_\mathbf{i})$ and $q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})$ via the objective

$$\mathop{\mathbb{E}}_{(\boldsymbol{x}^+,\boldsymbol{h}^+)\sim p(X,H|\mathcal{T}_\mathbf{i})}\left[E_\phi^{(a)}(\boldsymbol{x}^+,\boldsymbol{h}^+)\right] - \mathop{\mathbb{E}}_{(\boldsymbol{x}^-,\boldsymbol{h}^-)\sim q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\left[E_\phi^{(a)}(\boldsymbol{x}^-,\boldsymbol{h}^-)\right].$$

We found that it holds the upper bound as

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(\boldsymbol{x}^+,\boldsymbol{h}^+)\sim p(X,H|\mathcal{T}_\mathbf{i})}\left[E_\phi^{(a)}(\boldsymbol{x}^+,\boldsymbol{h}^+)\right] - \mathop{\mathbb{E}}_{(\boldsymbol{x}^-,\boldsymbol{h}^-)\sim q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\left[E_\phi^{(a)}(\boldsymbol{x}^-,\boldsymbol{h}^-)\right] \\
=&\mathop{\mathbb{E}}_{(\boldsymbol{x}^+,\boldsymbol{h}^+)\sim p(X,H|\mathcal{T}_\mathbf{i})}\left[-\log\sum_{c\sim\mathcal{V}_i}P_{\boldsymbol{\phi}}(\boldsymbol{x}^+,\boldsymbol{h}^+)[c]\right] + \mathop{\mathbb{E}}_{(\boldsymbol{x}^-,\boldsymbol{h}^-)\sim q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\left[\log\sum_{c\sim\mathcal{V}_i}P_{\boldsymbol{\phi}}(\boldsymbol{x}^-,\boldsymbol{h}^-)[c]\right] \\
\le&\mathop{\mathbb{E}}_{(\boldsymbol{x}^+,\boldsymbol{h}^+)\sim p(X,H|\mathcal{T}_\mathbf{i})}\left[-\log P_{\boldsymbol{\phi}}(\boldsymbol{x}^+,\boldsymbol{h}^+)[c]\right] + \mathop{\mathbb{E}}_{(\boldsymbol{x}^-,\boldsymbol{h}^-)\sim q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\left[\log\sum_{c\sim\mathcal{V}_i}P_{\boldsymbol{\phi}}(\boldsymbol{x}^-,\boldsymbol{h}^-)[c]\right] \\
=&\mathbb{E}_{p(X,H|\mathcal{T}_\mathbf{i})}\Big[\sum_{c\sim V_i}-\log P_{\boldsymbol{\phi}}(X,H)[c]\Big] - \mathbb{E}_{q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})}\left[E_\phi^{(a)}(X,H;\mathcal{T}_i)\right].
\end{aligned}$$

Hence minmizing the upper bound refers to the objective Eq.6 and in this way, optimizing Eq.6 leads to the gradient descending with Eq.12, which is equivalent to minimize the KL-divergence between the image-prompt distributions $p(X,H|\mathcal{T}_\mathbf{i})$ and $q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_\mathbf{i})$ due to Lemma.2. Besides, according to Lemma.1, we know that optimizing optimizing Eq.6 equals to optimizing Eq.8. Therefore optimizing Eq.8 exactly leads to the KL-divergence minimization $D_{\mathrm{KL}}\Big(p(X,H|\mathcal{T}_i)||q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_i)\Big)$ across different tasks. Given each task-specific objecitve converged, we obtain $p(X,H|\mathcal{T}_i)\to q_{\overline{\phi}}^{(\mathrm{EBM})}(X,H|\mathcal{T}_i)\propto\sum_{c\sim\mathcal{V}_i}P_{\boldsymbol{\phi}}(X,H)[c]$ $=1-\sum_{c\sim\mathcal{U}_i}P_{\boldsymbol{\phi}}(X,H)[c]=1-p_{\boldsymbol{\phi}}(X,H|\mathcal{U}_i)$. So the proposition is proved. □

B. Implementation

In this section, we elaborate the stochastic implementation for EMPL. We first discuss how to derive the gradients to update the prompt template with SGLD, then provide the stochastic version of EMPL and finally provides other details for our implementation.

B.1. Energy-based Prompt Gradients with SGLD

Most existing researches executs the SGLD-based gradient update to training examples, in which the model update is alternatively executed with SGLD-based instances. However, EMPL draws training instances from a prompt-image joint distribution where the SGLD-based prompt (embeddings) are intuitively generated from parametric templates. So the parameter udpate is entangled with the SGLD sampling process.

Here we provide more details to reveal the relation between SGLD and the gradients with regards to the learnable contexts $\boldsymbol{v}$, and further design a memory-efficeint method to implement the algorithm. We consider the SGLD sampler running on the prompt embedding space,

$$\boldsymbol{h}^{t+1}=\boldsymbol{h}^t-\frac{\alpha}{2}\frac{\partial E_\phi(\boldsymbol{x}^{t+1},\boldsymbol{h}^t)}{\partial\boldsymbol{h}^t}+\sqrt{\alpha}\epsilon_2,\epsilon_2\sim\mathcal{N}(0;I), \tag{13}$$

Note that SGLD runs on the latent space defined by the text encoder. Suppose that the prompt $\boldsymbol{v}(c)$ goes through the text encoder $\boldsymbol{h}(\cdot)$ to construct the contrastive learning score, we tend to split the text encoder $\boldsymbol{h}(\cdot)$ by two successive neural network modules $h_2(\cdot)$, $h_1(\cdot)$:

$$\boldsymbol{h}(\boldsymbol{v}(c))=h_2\Big(h_1(\boldsymbol{v};c)\Big) \tag{14}$$

where the learnable context and the class in the prompt $\boldsymbol{v}(c)$ was separated to highlight the optimization process. In this manner, $h_1(\boldsymbol{v};c)$ denotes the output to the latent embedding space $\mathcal{S}_{h_1}$ from the orginal prompt while $h_2(\cdot)$ receives the prompt embeddings in $\mathcal{S}_{h_1}$, which are drawn by the SGLD

sampler. Given this, we further rewrite the energy function in the iteration t:

$$E_\phi\big(\boldsymbol{x}^{t+1}, \boldsymbol{h}^t(\boldsymbol{v}(c))\big) = E'_\phi\big(\boldsymbol{x}^{t+1}, h_1^t(\boldsymbol{v}; c)\big), \tag{15}$$

where $\boldsymbol{x}^{t+1}$ denotes the image feature generated by SGLD from the current iteration. In the original formulationn (Eq.), $\boldsymbol{x}^{t+1}$ is conditioned on $\boldsymbol{h}^t$ with respect to $\boldsymbol{v}$. In this principle, the image features drawn by SGLD rely on $\boldsymbol{v}$. For the simplicity in our implementation, our algorithm ignores how $\boldsymbol{x}^{t+1}$ changes the learnable context $\boldsymbol{v}$ so that the derivative chain of $E'_\phi\big(\boldsymbol{x}^{t+1}, h_1^t(\boldsymbol{v}; c)\big)$ would not be branched from $\boldsymbol{x}^{t+1}$. Therefore the SGLD sampling for prompt embedding h_1^{t+1} presents as

$$h_1^{t+1} = h_1^t - \frac{\alpha}{2}\frac{\partial E'_\phi\big(\boldsymbol{x}^{t+1}, h_1^t\big)}{\partial h_1^t} + \sqrt{\alpha}\epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0; I), \tag{16}$$

in which $h_1^t = h_1^t(\boldsymbol{v}; c)$. In this regard, we turn to consider the derivative chain rule of $h_1^{t+1}(\boldsymbol{v}; c)$ with respect to the learnable context $\boldsymbol{v}$,

$$\begin{aligned}\frac{\partial h_1^{t+1}}{\partial \boldsymbol{v}} &= \frac{\partial h_1^t}{\partial \boldsymbol{v}} - \frac{\alpha}{2}\partial\Big(\frac{\partial E'_\phi(\boldsymbol{x}^{t+1}, h_1^t)}{\partial h_1^t}\Big)/\partial \boldsymbol{v}\Big) \\ &= \frac{\partial h_1^t}{\partial \boldsymbol{v}}\left(\mathbf{I} - \frac{\alpha}{2}\frac{\partial^2 E'_\phi(\boldsymbol{x}^{t+1}, h_1^t)}{(\partial h_1^t)^2}\right) \\ &= \frac{\partial h_1^0}{\partial \boldsymbol{v}}\prod_{i=0}^{t}\left(\mathbf{I} - \frac{\alpha}{2}\frac{\partial^2 E'_\phi(\boldsymbol{x}^{i+1}, h_1^i)}{(\partial h_1^i)^2}\right).\end{aligned} \tag{17}$$

From this observation, we can compute the gradient of the energy function with respect to the learnable context $\boldsymbol{v}$:

$$\begin{aligned}&\frac{\partial E'_\phi\big(\boldsymbol{x}^{t+1}, h_1^t(\boldsymbol{v}(c))\big)}{\partial \boldsymbol{v}} = \frac{\partial E'_\phi\big(\boldsymbol{x}^{t+1}, h_1^t(\boldsymbol{v}(c))\big)}{\partial h_1^t}\frac{\partial h_1^t}{\partial \boldsymbol{v}} \\ =&\frac{\partial E'_\phi\big(\boldsymbol{x}^{t+1}, h_1^t\big)}{\partial h_1^t}\frac{\partial h_1^{t-1}}{\partial \boldsymbol{v}}\left(\mathbf{I} - \frac{\alpha}{2}\frac{\partial E'_\phi(\boldsymbol{x}^t, h_1^{t-1})}{(\partial h_1^{t-1})^2}\right) \\ =&\frac{\partial E'_\phi\big(\boldsymbol{x}^{t+1}, h_1^t\big)}{\partial h_1^t}\frac{\partial h_1^0}{\partial \boldsymbol{v}}\underbrace{\prod_{i=0}^{t-1}\left(\mathbf{I} - \frac{\alpha}{2}\frac{\partial^2 E'_\phi(\boldsymbol{x}^{i+1}, h_1^i)}{(\partial h_1^i)^2}\right)}_{\Delta_{t-1}(\boldsymbol{x}, \boldsymbol{v}; c)}\end{aligned} \tag{18}$$

Note that $\frac{\partial h_1^0}{\partial \boldsymbol{v}}$ is independent with the SGLD sampling process; $\frac{\partial E'_\phi\left(\boldsymbol{x}^{t+1}, h_1^t\right)}{\partial h_1^t}$ can be computed through SGLD. Δ_{t-1} refers to a series of matrix multiplication across the second-order derivatives of the energy function with respect to different image feature and prompt embedding pairs drawn by SGLD. Obviously, they follow the multiplication chain as

$$\Delta_t(\boldsymbol{x}, \boldsymbol{v}; c) = \Delta_{t-1}(\boldsymbol{x}, \boldsymbol{v}; c)\left(\mathbf{I} - \frac{\alpha}{2}\frac{\partial^2 E'_\phi(\boldsymbol{x}^t, h_1^{t-1})}{(\partial h_1^{t-1})^2}\right). \tag{19}$$

Algorithm 1. Energy-based Prompt Gradient per Training Batch

Input:
Images $\{\boldsymbol{x}_j\}_{j=1}^M$, Unseen class names $\mathcal{U}_i$. Learnable context $\boldsymbol{v}$.
Output: The EBM-based prompt gradient $\nabla_{\boldsymbol{v}} E'_\phi(\{\boldsymbol{x}_j\}_{j=1}^M, \mathcal{U}_i)$.

Initiate $\nabla_{\boldsymbol{v}} E'_\phi = \mathbf{0}$;
while $c \sim \mathcal{U}_i$ **do** (Parallel execution)
  Initiate $h_1(\boldsymbol{v}; c)$ and $\frac{\partial h_1^0}{\partial \boldsymbol{v}} = \frac{\partial h_1(\boldsymbol{v}; c)}{\partial \boldsymbol{v}}$;
  for j=1:M **do** (Parallel execution)
    Initiate $\Delta(\boldsymbol{x}_j, \boldsymbol{v}; c) = \mathbf{I}$.
    for t=0:T **do**
      Update $\boldsymbol{x}_j^{t+1}, h_1^t$ by a SGLD sampler;
      $\nabla_{\boldsymbol{v}} E'_\phi \leftarrow \nabla_{\boldsymbol{v}} E'_\phi + \frac{\partial E'_\phi\left(\boldsymbol{x}^{t+1}, h_1^t\right)}{\partial h_1^t}\frac{\partial h_1^0}{\partial \boldsymbol{v}}\Delta(\boldsymbol{x}_j, \boldsymbol{v}; c)$;
      $\Delta(\boldsymbol{x}_j, \boldsymbol{v}; c) \leftarrow \Delta(\boldsymbol{x}_j, \boldsymbol{v}; c)\Big(\mathbf{I} - \frac{\alpha}{2}\frac{\partial^2 E'_\phi(\boldsymbol{x}^t, h_1^{t-1})}{(\partial h_1^{t-1})^2}\Big)$.
    end for
  end for
end while
return $\nabla_{\boldsymbol{v}} E'_\phi$.

We observe that $\{\Delta_i(\boldsymbol{x}, \boldsymbol{v}; c)\}_{i=1}^{t-1}$ can be disposed after we obtained $\Delta_t(\boldsymbol{x}, \boldsymbol{v}; c)$ with the second-order derivative of the energy function with respect to the updated image feature and prompt embeddings. Given this, we only need to store a screenshot of $\Delta(\boldsymbol{x}, \boldsymbol{v}; c) \leftarrow \Delta_t(\boldsymbol{x}, \boldsymbol{v}; c)$ for each iteration, which is enough to formulate the gradient $\frac{\partial E'_\phi\left(\boldsymbol{x}^{t+1}, h_1^t(\boldsymbol{v}(c))\right)}{\partial \boldsymbol{v}}$ without redundant computation.

The pipeline of computing the energi-based prompt gradient for each training batch can be summarized into Algorithm.1.

B.2. Stochastic Algorithm of EMPL

Algorithm 2. Energy-based Multi-prompt Learning (Stochastic)

Input: An open vocabulary V consists of observed class names $\mathcal{V}$ and unseen class names $\mathcal{U}$; Training images $X = \{\boldsymbol{x}_j, c_j\}_{j=1}^N$ with labels as class names in $\mathcal{V}$; the prompt template parameter ϕ .
Output: the optimal prompt template parameter ϕ^{*a}.

Initiate $f(\cdot), h_{\boldsymbol{v}}(\cdot)$ with CLIP;
while not converged **do**
  Draw a training batch $\mathcal{X}$ from X with K_1 observed classes in $\mathcal{V}$, then draw $K-K_1$ unseen classes $\mathcal{U}_i$ in $\mathcal{U}$;
  Obtain the gradient $\nabla_{\boldsymbol{v}} P(\mathcal{X})$ of the first term in Eq.8 by the base prompt learning algorithm with $\mathcal{X}$;
  Obtain $\nabla_{\boldsymbol{v}} E'_\phi(\mathcal{X}, \mathcal{U}_i)$ by Algorithm.1;
  Update $\boldsymbol{v} \leftarrow \boldsymbol{v} + \mathrm{Opt}(\nabla_{\boldsymbol{v}} P(\mathcal{X}) - \lambda \nabla_{\boldsymbol{v}} E'_\phi(\mathcal{X}, \mathcal{U}_i))^b$
end while
return $\phi^* \leftarrow \boldsymbol{v}^*$.

[a] ϕ mostly refers to $\boldsymbol{v}$ while is limited to the single learnable context
[b] Opt denotes an optimizer.

Provided with the gradient update based on the energy-based function in Algorithm.1, we present the stochastic optimization derived from the EMPL objective (Algorithm.2). Briefly speaking, the stochastic EMPL constructs a task $\mathcal{T}_i$ by drawing a training batch per iteration, in which the images with labels construct a supervised learning objective (*i.e.*, the first term in Eq.8) derived from the base prompt learning counterpart (*e.g.*, CoOp or ProDA); then it tasks the training images with the unseen class names to compute the energy-based prompt gradient by Algorithm.1; finally, it combines the gradients according to the EMPL objective to update the learnable parameter with an optimizer.

B.3. Other implementation details

The implementation of EMPL depends on the baselines CoOp and ProDA. We exactly follow their implementations to facilitate the first term of Eq.8 and take all their learnable prompt contexts to construct the energy-based gradient produced by Algorithm.2. Without loss of generality, we set $\lambda = 0.1$, $K{-}K_1{=}4$ (K_1 is exactly the same with the base code of CoOp and ProDA). The SGLD sampler is implemented at the second last transformer layer to generate prompt embeddings. And we use the SGLD sampler with a constant step size of 2 and a standard deviation of 1e-3. The number of updates in each SGLD round is set to 25. As for SGLD-based prompting, we execut alternatively the SGLD to generate 8 prompt embeddings for each image per class, then the prediction is achieved either by the original multi-prompt paradigm (ProDA) or by the average of voting from all prompt-based contrastive learning predictions (CoOp).

C. Complementary Experiments

In this section, we provide some empirical studies complementary to the experiments shown in our paper. We first convey some empirical evidences to verify whether multi-prompt learning can help to relieve the non-identifiability issue; then we visualize the multi-prompt embeddings generated by EMPL.

C.1. Multi-label Classification

In order to justify the non-identifiability issue, we evaluate single-prompt and multi-prompt learners on the dataset with diverse visual concepts contained within each image. We evaluate the CoOp and ProDA in the multi-label classification setup. Specifically, we initialize different prompt learning models with CLIP then trained the models with their algorithms on NUS-WIDE [1]. The implementation is derived from `https://github.com/sunanhe/MKT` [2] where the prompt-tuning objective across different baselines are replaced by the ranking loss to achieve the multi-label classification task. We evaluate the baseline models with their mAP and F1 (K=3) score across zero-shot learning (ZSL) and generalized zero-shot learning (GZSL) setups and besides, we also consider their prediction consistency. In particular, for each test image, we consider another corresponding test image (counterpart) with some shared classes whereas we inquire the first test image with the exclusive classes that only the second test image belongs to [1]. Given this, we observe whether we can simultaneously inquire the first test image with the class it belongs to and with the class it does not belongs to, respectively. It futher refers to the consistency rate (CR, *i.e.*, how many images could be simultaneously inquired with the exclusive classes that they belong to and they do not belong to, respectively) to measure the severity of non-identifiability. Obviously, the higher indicates more serious.

Table 1. The performance based on F1 score, mAP and CR across the zero-shot learning and generalized zero-shot learning setups.

Prompt learners	ZSL			GZSL		
	F1	mAP	CR	F1	mAP	CR
CoOp	29.5	32.5	0.32	21.1	16.8	0.25
ProDA	30.9	33.6	0.16	21.1	17.1	0.11
ProDA(×4)	31.2	34.0	0.12	21.5	17.2	0.08

As reported in Table.1, multi-prompt baselines outperform single-prompt learning models with clear margins in the ZSL setup. Besides, we also observe that the CR significantly drops as the number of prompts increases. It implies that the non-identifiability has been relieved and probably contributes to the performance. In the GZSL setup, the performance difference across baselines are inconspicuous while the CR is still significantly decreased as the number of prompts increases.

C.2. Visualization

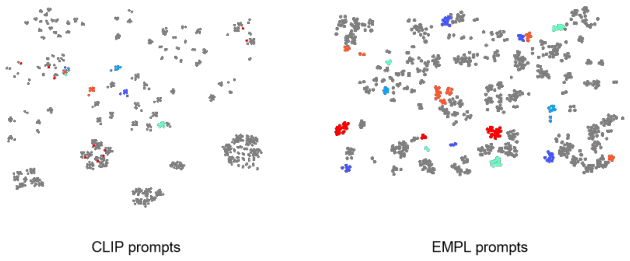

Figure 1. The t-SNE visualization [5] for the input embedding spaces between CLIP and EMPL. Different colors indicate different prompts with different categories (best viewed in color).

Here we provide the t-SNE visualization [5] to the input embeddings of the text encoder, which are extracted from CLIP and EMPL (+CoOp). For CLIP, the hand-crafed descriptive texts with 50 categories are fed into the embed-

[1]For the images containing multiple exclusive classes, we prefer a pair of class names with the closest distance between their word embeddings [4]. The selection for the counterpart image depends on the alphabetical order.

ding layer to generate the t-SNE visualization. Since EMPL does not generate prompts in the input embedding space, we backpropagate SGLD-based prompt embeddings from the latent embedding layer to the input embedding layer. It generates the virtual input embeddings for the t-SNE visualization, which also contains the consistent 50 categories with the CLIP's visualization.

As illustrated in Fig.1, we found that CLIP's prompt inputs are scattered in the space more sparsely and only a few of narrow ranges refer to the textual descriptions with a specific category. Instead, prompt embeddings generated by EMPL correspond to more condense virtual input embeddings. Besides, the SGLD sampler encourages to generate multiple prompts for each category description so that their virtual input embeddings cover more areas in the input embedding space and for each base prompt embedding input, the SGLD sampler creates a set of derived prompt embeddings, which are reflected by a series of virtual input embeddings surrounding the base prompt input embedding. It implies the capability of multi-prompting.

References

[1] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 6

[2] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. *arXiv preprint arXiv:2207.01887*, 2022. 6

[3] Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*, 2016. 3, 4

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 6

[5] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[6] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *The Eleventh International Conference on Learning Representations*. 1