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

Ziliang Chen1,3, Xin Huang2, Quanlong Guan1, Liang Lin2, Weiqi Luo1  
1Jinan University 2Sun Yat-sen University 3Pazhou Laboratory

{yanxp3,wangxx35}@mail2.sysu.edu.cn, c.ziliang@yahoo.com, 466783266@qq.com, xdliaang328@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 $\epsilon$-approximate constant modality gap)** Suppose that $f(\cdot)$ and $h_\nu(\cdot)$ denote a visual encoder and a text encoder with prompt template $\nu$, respectively, and a denumerable set of images $\{x_i\}_{i=1}^\infty$ are consistent with the concept set $C$. A constant vector $c$ denotes the modality gap of $\{x_i\}_{i=1}^\infty$ with respect to $f(\cdot)$ and $h_\nu(\cdot)$, if for all arbitrary scalar $\epsilon > 0$ and arbitrary concepts in $C$, there exists a positive integer $N(\epsilon)$, so that given arbitrary positive integers $j$ satisfying $j > N(\epsilon)$, we have a permutation $\{x^{(i)}\}_{i=1}^\infty$ of $\{x_i\}_{i=1}^\infty$ which satisfies $|f(x^{(j)}) - h_\nu(c) - c| < \epsilon$ for arbitrary $c \in C$.

The definition does not enforce each pair of image feature and prompt embedding exactly suits the constant modality gap as [6]. Instead, it employs the $\epsilon$-approximation language of an infinite sequence to define the gap maintained in a range with an 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 $(f(\cdot), h_\nu(\cdot))$ defines the individual $\epsilon$-approximate constant modality gap. Given an 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 $x_i \in X_1$ with the $\epsilon$-approximation to the constant modality gap $\nu$ by $h_\nu(\cdot)$, there exists an image $x_j \in X_2$ with the identical $\epsilon$-approximation to the constant modality gap $\nu$ by $h_\nu(\cdot)$, 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 arbitrary small $\frac{1}{3}\epsilon > 0$, there is a positive integer $N_1(\frac{1}{3}\epsilon)$, so given an arbitrary positive integer $k$ satisfying $k > N_1(\frac{1}{3}\epsilon)$, we have a permutation $\{x^{(i)}\}_{i=1}^\infty$ from $X_1$ which satisfies $|f(x^{(k)}) - h_\nu(c_1) - c| < \frac{1}{3}\epsilon$, given $x^{(k)} \in \{x^{(i)}\}_{i=1}^\infty$. Here we focus on $\{x^{(i)}\}_{i=1}^\infty$, the subchain of $\{x^{(i)}\}_{i=1}^\infty$ 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 arbitrary positive integer $k$ satisfying $k > N_2(\frac{1}{3}\epsilon)$, we may take a permutation $\{x^{(i)}\}_{i=1}^\infty$ from the chain $\{x^{(i)}\}_{i=1}^\infty$ which satisfies $|f(x^{(k)}) - h_\nu(c_2) - c| < \frac{1}{3}\epsilon$, given $x^{(k)} \in \{x^{(i)}\}_{i=1}^\infty$.

Based on the subchain construction, we consider an image $x_i$ drawn from the subchain $\{x^{(i)}\}_{i=1}^\infty$ and another image $x_j$ drawn from $\{x^{(i)}\}_{i=1}^\infty$. Obviously, they satisfy

$$|f(x_i) - h_\nu(c_1) - c| < \frac{1}{3}\epsilon,$$  

$$|f(x_i) - h_\nu(c_2) - c| < \frac{1}{3}\epsilon;$$  

$$|f(x_j) - h_\nu(c_2) - c| < \frac{1}{3}\epsilon.$$

While they also result in

$$|f(x_j) - h_\nu(c_1) - c| = |f(x_i) - h_\nu(c_1) - c + c - f(x_i) + h_\nu(c_2) + f(x_j) - h_\nu(c_2) - c|$$

$$\leq |f(x_i) - h_\nu(c_1) - c| + |f(x_i) - h_\nu(c_2) - c|$$

$$< \frac{1}{3}\epsilon + \frac{1}{3}\epsilon + \frac{1}{3}\epsilon = \epsilon.$$

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 \( \epsilon \) is a arbitrary small number, so given each pair of \((f(x_i), h_v(c_1))\) with \( x_i \in X_i \) that satisfy the \( \epsilon \)-approximation with the assumption, we are able to find the other pair \((f(x_i^k), h_v(c_1))\) with \( x_i^k \in X_i \) that satisfy the \( \frac{1}{3}\epsilon \)-approximation and through the reduction above, it results in a images \( x_j \in X_2 \) which also satisfies the \( \epsilon \)-approximation with \( h_v(c_1) \).

From this observation, images \( x_i \) and \( x_j \) distant by the exclusive concept \( c_1 \) are impossible to be identified by the single-prompt learning model \((f(\cdot), h_v(c_1))\).

We can also extend the \( \epsilon \)-approximate constant modality gap from the individual level to the population level:

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

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

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

**Proof.** Similarly with the proof for the individual-level non-identifiability, for an arbitrary small \( \frac{1}{3}\epsilon > 0 \), we use a integer \( N_1(\frac{1}{3}\epsilon) \) to construct the subchain \( \{x^{(i)}\}_{i>N_1(\frac{1}{3}\epsilon)} \subset X_{C_1} \) where \( \sum_{x \in X_i} f(x) - h_v(c_1) - c < \frac{1}{3}\epsilon \) is satisfied for \( c_1 \in C_1 / C_2 \subset C_1 \) with \( \forall X_i \subseteq \{x^{(i)}\}_{i>N_1(\frac{1}{3}\epsilon)} \). Deriving from \( \{x^{(i)}\}_{i>N_1(\frac{1}{3}\epsilon)} \) we are permitted to take its summation to construct \( \{x^{(i_1,i_2)}\}_{i_1>N_1(\frac{1}{3}\epsilon), i_2=1} \). Therefore with an integer \( N_2(\frac{1}{3}\epsilon) > 0 \), we could also extract a subset \( X_2 \) from \( \{x^{(i_1,i_2)}\}_{i_1>N_1(\frac{1}{3}\epsilon), i_2>2N_2(\frac{1}{3}\epsilon)} \) to encourage the similar \( \frac{1}{3}\epsilon \)-approximation \( \sum_{x \in X_2} f(x) - h_v(c_2) - c < \frac{1}{3}\epsilon \), with respect to \( c_2 \in C_1 \cap C_2 \subset C_1 \). Obviously, we can directly extract a subset \( X_3 \) 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 \( \{x^{(i)}\}_{i>N_3(\frac{1}{3}\epsilon)} \subset X_{C_2} \) so that provided with \( \forall X_3 \subseteq \{x^{(i)}\}_{i>N_3(\frac{1}{3}\epsilon)} \) the \( \frac{1}{3}\epsilon \)-approximation \( \sum_{x \in X_3} f(x) - h_v(c_2) - c < \frac{1}{3}\epsilon \) is also achieved.

Based upon the construction above, we now consider the intersection \( X_1 \cap X_2 \). It obviously satisfies

\[
\left| \sum_{x \in X_1 \cap X_2} f(x) - h_v(c_1) - c \right| < \frac{1}{3}\epsilon,
\]

\[
\left| \sum_{x \in X_1 \cap X_2} f(x) - h_v(c_2) - c \right| < \frac{1}{3}\epsilon.
\]

With this regard, we cast an \( \epsilon \)-approximation as follows

\[
\left| \sum_{x \in X_3} f(x) - h_v(c) - c \right| < \frac{1}{3}\epsilon + \frac{1}{3}\epsilon + \frac{1}{3}\epsilon = \epsilon
\]

So given an image group \( X \) consistent with the concept set \( C_1 \) that satisfies the \( \epsilon \)-approximation with the assumption by \( h_v(c_1) \), we may construct the reduction above to obtain \( X' = X_3 \). Although \( X' \) are consistent with the concept set \( C_2 \), it also leads to the \( \epsilon \)-approximation with the assumption by \( h_v(c_1) (c_1 \in C_1 / C_2) \).

From this observation, the image groups \( X \) and \( X' \) distinct by the exclusive concept \( c_1 \) are impossible to be identified by the single-prompt learning model \((f(\cdot), h_v(c_1))\).

**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 \( p(X, H; T_i) \) and another EBM-modeled image-prompt joint distribution \( \pi(EBM) (X, H; T_i) \), i.e.,

\[
E_{\phi}(X, H; T_i) = \log \sum_{c<\gamma_i} p_{\phi}(X, H | c)
\]

\[
= \log \sum_{c<\gamma_i} \exp \left( \frac{\text{sim}(f(X), h_v(c); \gamma_i)}{\gamma} \right)
\]

(4)
where the auxiliary energy function of \( q_{(EBM)}(X, H | T_i) \) is defined as

\[
E^{(a)}_{\phi}(X, H; T_i) = -\log \sum_{c \sim V_i} P_{\phi}(X, H)[c] \\
= -\log \sum_{c \sim V_i} \exp \left( \frac{\min(f(X), H(c \mid V_i), U_i)}{\gamma} \right)
\]

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

\[
\min_{\phi} \mathbb{E}_{p(X, H | T_i)} \left[ \sum_{c \sim V_i} -\log P_{\phi}(X, H)[c] \right] - \mathbb{E}_{q_{(EBM)}(X, H | T_i)} \left[ E^{(a)}_{\phi}(X, H; T_i) \right].
\]

Resembling the notations of EMPL objective

\[
\begin{align*}
\min_{\phi} & \mathbb{E}_{p(X, H | T_i)} \left[ \sum_{c \sim V_i} -\log P_{\phi}(X, H)[c] \right] \\
& \text{Generic prompt learning goal} \\
& - \lambda \mathbb{E}_{q_{(EBM)}(X, H | T_i)} \left[ E_{\phi}(X, H; T_i) \right],
\end{align*}
\]

\( \bar{\phi} \) 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.,

\[
\begin{align*}
\min_{\phi} & \mathbb{E}_{p(X, H | T_i)} \left[ \sum_{c \sim V_i} -\log P_{\phi}(X, H)[c] \right] \\
& - \lambda \mathbb{E}_{q_{(EBM)}(X, H | T_i)} \left[ E_{\phi}(X, H; T_i) \right].
\end{align*}
\]

**Lemma 1.** Given two energy-based distributions \( p_{(EBM)} \) and \( q_{(EBM)} \) 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 \( \gamma \).

**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,

\[
\frac{\partial}{\partial \phi} \left( -\mathbb{E}_{q_{(EBM)}(X, H | T_i)} \left[ E^{(a)}_{\phi}(X, H; T_i) \right] \right) = \frac{\partial}{\partial \phi} \left( \mathbb{E}_{q_{(EBM)}(X, H | T_i)} \log \sum_{c \sim V_i} P_{\phi}(X, H)[c] \right)
\]

\[
= \int_{x, h} q_{(EBM)}(x, h | T_i) \frac{\partial}{\partial \phi} \left( \log \sum_{c \sim V_i} P_{\phi}(x, h)[c] \right)
\]

\[
= \int_{x, h} \sum_{c \sim V_i} P_{\phi}(x, h)[c] \frac{1}{Z^{(a)}(\phi)} \frac{\partial}{\partial \phi} \left( \sum_{c \sim V_i} P_{\phi}(x, h)[c] \right)
\]

\[
= - \int_{x, h} \frac{\partial}{\partial \phi} \left( \sum_{c \sim U_i} P_{\phi}(x, h)[c] \right)
\]

\[
= - \int_{x, h} \frac{\partial}{\partial \phi} \left( \frac{Z(\phi)}{Z^{(a)}(\phi)} \int_{x, h} p_{(EBM)}(x, h | T_i) \frac{\partial}{\partial \phi} \log \left( \sum_{c \sim U_i} P_{\phi}(x, h)[c] \right) \right)
\]

\[
= - \gamma \mathbb{E}_{q_{(EBM)}(X, H | T_i)} \frac{\partial}{\partial \phi} \log \left( \sum_{c \sim U_i} P_{\phi}(X, H)[c] \right).
\]

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

\( \square \)

Given this observation, if we further prove that optimizing Eq.6 is equivalent with minimizing the KL-divergence between \( p(X, H | T_i) \) and \( q_{(EBM)}(X, H | T_i) \), 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 \( 1 - p_{\phi}(X, H | U_i) = 1 - \sum_{c \sim U_i} P_{\phi}(X, H)[c] \) for matching the observed image-prompt joint distribution so that \( p(X, H | T_i) \) and \( p_{\phi}(X, H | U_i) \) 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 \( r_{(EBM)}(X, H) \) parameterized by \( \phi \) with respect to the energy function \( E_{\phi}(X, H) \), the objective that minimizes the KL-divergence \( D_{KL}(r(X, H) || r_{(EBM)}(X, H)) \) can be opt.
timized via descending the gradient w.r.t. \( \theta \):
\[
\begin{align*}
E_{(x^+, h^+) \sim p(X, H|T_i)} \left[ \frac{\partial E_{\phi}(x^+, h^+)}{\partial \theta} \right] - E_{(x^-, h^-) \sim q_{\pi}^{(EBM)}(X, H|T_i)} \left[ \frac{\partial E_{\phi}(x^-, h^-)}{\partial \theta} \right]
\end{align*}
\]  

(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 \( p(X, H|T_i) \) and \( q_{\pi}^{(EBM)}(X, H|T_i) \) refers to descending the gradient in Eq. 11. With \( \phi \) replaced with a fixed \( \tilde{\phi} \), the gradient can be reformulated to
\[
\begin{align*}
E_{(x^+, h^+) \sim p(X, H|T_i)} & \left[ E_{\tilde{\phi}}^{(a)}(x^+, h^+) \right] - E_{(x^-, h^-) \sim q_{\pi}^{(EBM)}(X, H|T_i)} \left[ E_{\tilde{\phi}}^{(a)}(x^-, h^-) \right].
\end{align*}
\]  

(12)

So we can achieve it by minimizing the KL-divergence between \( p(X, H|T_i) \) and \( q_{\pi}^{(EBM)}(X, H|T_i) \) via the objective
\[
\begin{align*}
E_{(x^+, h^+) \sim p(X, H|T_i)} \left[ E_{\phi}^{(a)}(x^+, h^+) \right] - E_{(x^-, h^-) \sim q_{\pi}^{(EBM)}(X, H|T_i)} \left[ E_{\phi}^{(a)}(x^-, h^-) \right].
\end{align*}
\]

We found that it holds the upper bound as
\[
\begin{align*}
E_{(x^+, h^+) \sim p(X, H|T_i)} & \left[ E_{\phi}^{(a)}(x^+, h^+) \right] - E_{(x^-, h^-) \sim q_{\pi}^{(EBM)}(X, H|T_i)} \left[ E_{\phi}^{(a)}(x^-, h^-) \right] \\
= & \left( x^+, h^+ \right) \sim p(X, H|T_i) \left[ - \log \sum_{c \sim V_i} P_{\phi}(x^+, h^+)[c] \right] + E_{(x^+, h^+) \sim p(X, H|T_i)} \left[ \log \sum_{c \sim V_i} P_{\phi}(x^+, h^+)[c] \right] \\
\leq & \left( x^+, h^+ \right) \sim p(X, H|T_i) \left[ - \log \sum_{c \sim V_i} P_{\phi}(x^+, h^+)[c] \right] + E_{(x^-, h^-) \sim q_{\pi}^{(EBM)}(X, H|T_i)} \left[ \log \sum_{c \sim V_i} P_{\phi}(x^-, h^-)[c] \right] \\
= & E_{p(X, H|T_i)} \left[ \sum_{c \sim V_i} - \log P_{\phi}(X, H)[c] \right] - E_{q_{\pi}^{(EBM)}(X, H|T_i)} \left[ E_{\phi}^{(a)}(X, H; T_i) \right].
\end{align*}
\]

Hence minimizing 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 \( p(X, H|T_i) \) and \( q_{\pi}^{(EBM)}(X, H|T_i) \) 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 \( D_{KL}(p(X, H|T_i)||q_{\pi}^{(EBM)}(X, H|T_i)) \) across different tasks. Given each task-specific objective converged, we obtain \( p(X, H|T_i) \sim q_{\pi}^{(EBM)}(X, H|T_i) \) and \( \sum_{c \sim V_i} P_{\phi}(X, H)[c] = 1 - \sum_{c \sim U_i} P_{\phi}(X, H)[c] = 1 - p_{\phi}(X, H|U_i) \). So the proposition is proved.

\[ \square \]

### 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 executes 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 update 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 \( v \), and further design a memory-efficient method to implement the algorithm. We consider the SGLD sampler running on the prompt embedding space,
\[
h^{t+1} = h^t - \alpha \frac{\partial E_{\phi}(x^{t+1}, h^{t+1})}{\partial h^t} + \sqrt{\alpha} \epsilon_2 \epsilon_2 \sim N(0; I),\]

(13)

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

(14)

where the learnable context and the class in the prompt \( v(c) \) was separated to highlight the optimization process. In this manner, \( h_1(v; c) \) denotes the output to the latent embedding space \( S_{h_1} \) from the orginal prompt while \( h_2(\cdot) \) receives the prompt embeddings in \( S_{h_1} \), which are drawn by the SGLD.
sampler. Given this, we further rewrite the energy function in the iteration $t$:

$$E_\phi(x^{t+1}, h^t(v(c))) = E'_\phi(x^{t+1}, h^t_1(v; c)),$$

(15)

where $x^{t+1}$ denotes the image feature generated by SGLD from the current iteration. In the original formulation (Eq.), $x^{t+1}$ is conditioned on $h^t$ with respect to $v$. In this principle, the image features drawn by SGLD rely on $v$. For the simplicity in our implementation, our algorithm ignores how $x^{t+1}$ changes the learnable context $v$ so that the derivative chain of $E'_\phi(x^{t+1}, h^t_1(v; c))$ would not be branched from $x^{t+1}$. Therefore the SGLD sampling for prompt embedding $h^{t+1}_1$ presents as

$$h^{t+1}_1 = h^t_1 - \frac{\alpha}{2} \frac{\partial E'_\phi(x^{t+1}, h^t_1)}{\partial h^t_1} + \sqrt{\epsilon_2},$$

(16)

in which $h^t_1 = h^t_1(v; c).$ In this regard, we turn to consider the derivative chain rule of $h^{t+1}_1(v; c)$ with respect to the learnable context $v$.

$$\frac{\partial h^{t+1}_1}{\partial v} = \frac{\partial h^t_1}{\partial v} - \frac{\alpha}{2} \frac{\partial E'_\phi(x^{t+1}, h^t_1)}{\partial h^t_1} + \frac{\alpha}{2} \left( I - \frac{\partial^2 E'_\phi(x^{t+1}, h^t_1)}{\partial h^t_1^2} \right)$$

(17)

$$= \frac{\partial h^t_1}{\partial v} \prod_{i=0}^{t-1} \left( I - \frac{\alpha}{2} \frac{\partial^2 E'_\phi(x^{i+1}, h^i)}{\partial h^i_1^2} \right).$$

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

$$\frac{\partial E'_\phi(x^{t+1}, h^t(v; c))}{\partial v} = \frac{\partial E'_\phi(x^{t+1}, h^t_1(v; c))}{\partial h^t_1} \frac{\partial h^t_1}{\partial v}$$

$$= \frac{\partial E'_\phi(x^{t+1}, h^t_1)}{\partial h^t_1} \frac{\partial h^t_1}{\partial v} \prod_{i=0}^{t-1} \left( I - \frac{\alpha}{2} \frac{\partial^2 E'_\phi(x^{i+1}, h^i)}{\partial h^i_1^2} \right)$$

$$= \frac{\partial E'_\phi(x^{t+1}, h^t_1)}{\partial h^t_1} \prod_{i=0}^{t} \left( I - \frac{\alpha}{2} \frac{\partial^2 E'_\phi(x^{i+1}, h^i)}{\partial h^i_1^2} \right) \Delta_{t-1}(x, v; c)$$

(18)

Note that $\frac{\partial^0 h}{\partial v}$ is independent with the SGLD sampling process; $\frac{\partial E'_\phi(x^{t+1}, h^t_1)}{\partial h^t_1}$ can be computed through SGLD. $\Delta_{t-1}$ 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

$$\Delta_t(x, v; c) = \Delta_{t-1}(x, v; c) \left( I - \frac{\alpha}{2} \frac{\partial^2 E'_\phi(x^{i+1}, h^i)}{\partial h^i_1^2} \right).$$

(19)

We observe that $\{\Delta_t(x, v; c)\}_{t=1}^{t-1}$ can be disposed after we obtained $\Delta_t(x, v; c)$ 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 $\Delta(x, v; c) \leftarrow \Delta_t(x, v; c)$ for each iteration, which is enough to formulate the gradient $\frac{\partial E'_\phi(x^{t+1}, h^t_1(v; c))}{\partial v}$ without redundant computation.

The pipeline of computing the energy-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 $X = \{x_j, c_j\}_{j=1}^{N}$ with labels as class names in $\mathcal{V}$; the prompt template parameter $\phi^\ast$.

**Output:** the optimal prompt template parameter $\phi^\ast$.

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

Notes: $\phi^\ast$ mostly refers to $v$ while is limited to the single learnable context $v$.

$^\ast$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 \( T_i \) 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.

### B3. 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 execute alternatively the SGLD to generate 8 prompt embeddings for each image per class, then the prediction is achieved either by the original multiprompt 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 multiprompt 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. 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 belong to, respectively. It further 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.

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).](Image)

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

<table>
<thead>
<tr>
<th>Prompt learners</th>
<th>ZSL</th>
<th>GZSL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>mAP</td>
</tr>
<tr>
<td>CoOp</td>
<td>29.5</td>
<td>32.5</td>
</tr>
<tr>
<td>ProDA</td>
<td>30.9</td>
<td>33.6</td>
</tr>
<tr>
<td>ProDA(×4)</td>
<td>31.2</td>
<td>34.0</td>
</tr>
</tbody>
</table>

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 visu-
alization, which also contains the consistent 50 categories
with the CLIP’s visualization.

As illustrated in Fig.1, we found that CLIP’s prompt in-
puts are scattered in the space more sparsely and only a few
of narrow ranges refer to the textual descriptions with a spe-
cific category. Instead, prompt embeddings generated by
EMPL correspond to more condense virtual input embed-
dings. 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 em-
bedding space and for each base prompt embedding input,
the SGLD sampler creates a set of derived prompt embed-
dings, which are reflected by a series of virtual input em-
beddings surrounding the base prompt input embedding. It
implies the capability of multi-prompting.

References

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

[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.

tive models with energy-based probability estimation. arXiv

and Jeff Dean. Distributed representations of words and
phrases and their compositionality. Advances in neural in-
formation processing systems, 26, 2013.

data using t-sne. Journal of machine learning research,
9(11), 2008.

[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 In-
ternational Conference on Learning Representations.