Attend and Enrich: Enhanced Visual Prompt for Zero-Shot Learning

Man Liu[2,4], Huihui Bai[1,2,4✉], Feng Li[3✉], Chunjie Zhang[2,4],
Yunchao Wei[2,4], Tat-Seng Chua[5], Yao Zhao[2,4]

关注并增强：面向零样本学习的增强型视觉提示

AAAI 2025

1 INTRODUCTION

核心问题

如何在 ZSL中增强模型对“未见类别”的泛化能力，尤其在现有 prompt-based 方法中容易过拟合“已见类别”的主视觉特征，导致在未见类别识别时表现不佳。

1 INTRODUCTION

方法提出1

Prompt 学习过拟合于已见类别，泛化性差

发现了什么问题?

现有的 prompt-based ZSL 方法，虽然可以让模型识别图像，但它们是在 **seen 类别上固定训练的**，导致 prompt 学会的是“如何识别训练图像中的显眼视觉特征”（比如动物的颜色、身体形状），却无法泛化到那些需要理解 **语义细节（如“夜行性”“是否有蹄”）** 的 unseen 类别。

从什么角度解决这个问题?

让Prompt带上语义理解能力

做了什么?

AENet 框架，不直接用视觉 prompt，而是先把 prompt **语义增强**，再融合到视觉特征中。

方法提出2

图像和属性语义之间缺乏一致性，导致融合效果差

发现了什么问题?

视觉图像和属性描述来自不同模态，语义层次和表示粒度不同。直接把图像和属性拼接或平均融合，会因为信息不一致导致模型学不到对齐特征

从什么角度解决这个问题?

先让这两种信息对齐，从而让语义提示真正对图像起作用

做了什么?

概念感知注意力（Concept-Aware Attention, CAA） 模块

- 创造了一个 modal-sharing token，作为**语义锚点**；
- 让图像特征和属性描述都以它为参照进行注意力交互；
- 最终得到两个 “**概念协调（harmonized）” 的表示向量**：图像的、文本的。

方法提出3

prompt 中缺乏对语义细节的建模能力

发现了什么问题?

即使 prompt 和视觉特征对齐了，但 prompt 本身还是靠训练图像优化来的，它容易集中在主视觉区域。

从什么角度解决这个问题?

从属性和图像的融合信息中“预测出一些语义细节残差”，然后加回 prompt

做了什么?

提出**视觉残差预测单元**（VRRU）

- **把 harmonized 图文特征拼接;**
- **用一个轻量线性层（ZLinear）预测出“语义残差信息”z;**
- **然后把这个 z 加回到每一个 prompt token 上，得到 语义增强提示（semantic-enhanced prompt）。**

方法提出4

prompt 增强不一定对 unseen 类别有用，需要语义监督

发现了什么问题?

加入残差z后，怎么确保不是噪声，而是真正“语义相关”的?

从什么角度解决这个问题?

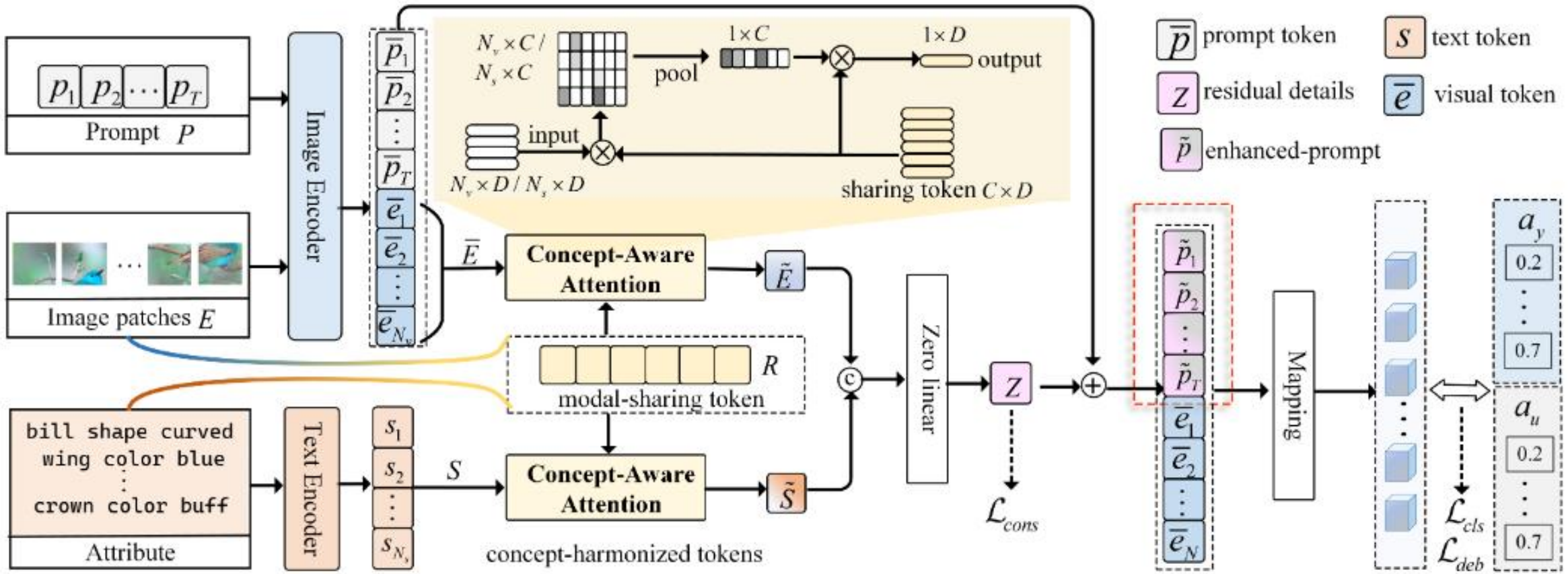
用真实的语义标签去监督z，让它靠近真实属性

做了什么?

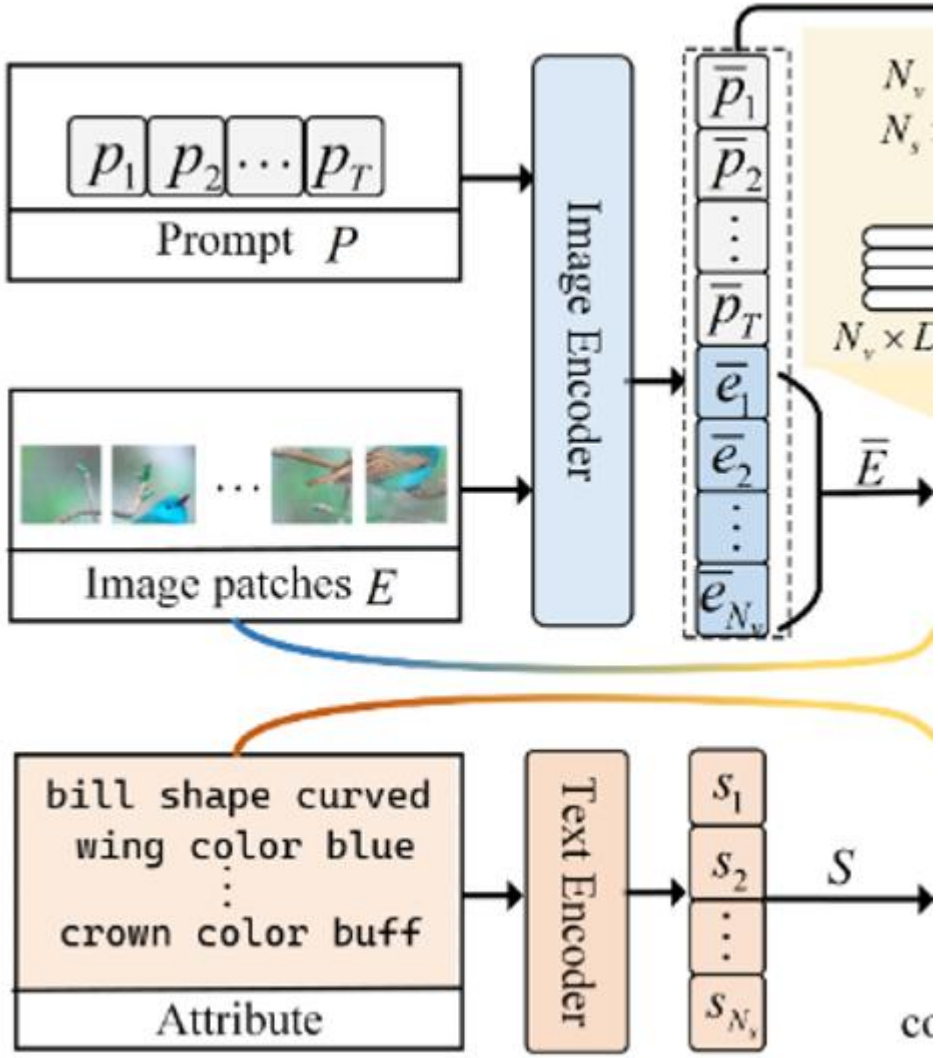
属性一致性损失

- **把 ground-truth 类别属性 ay 通过一个 MLP 映射;**
- **然后用 z 和这个语义投影去做距离最小化;**
- **保证 z 真正代表的是“语义上的残差补充”。**

1 INTRODUCTION



1 INTRODUCTION

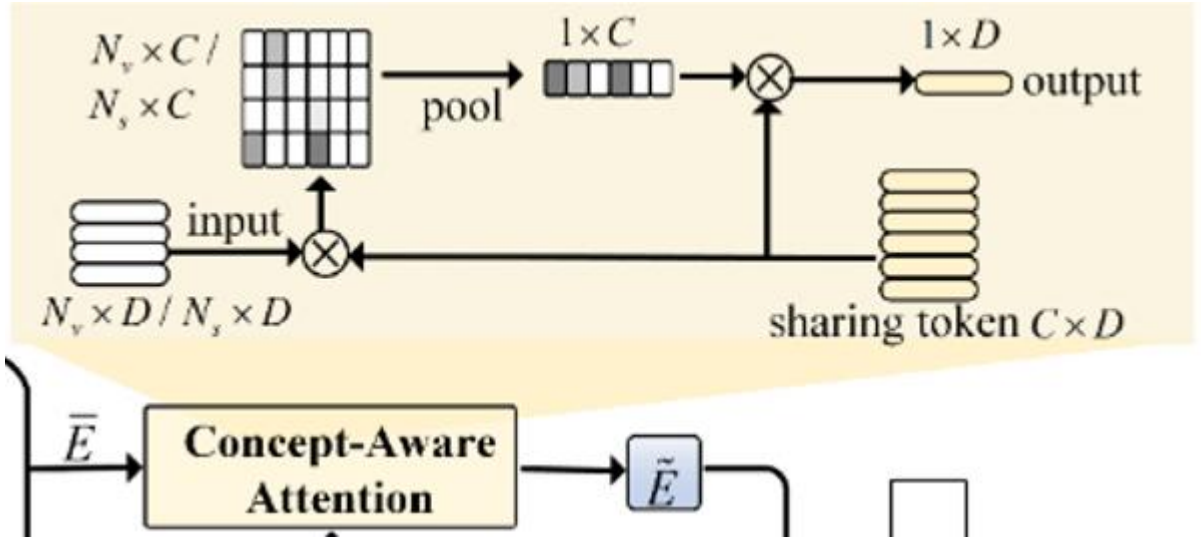


图像+提示输入：E（图像 patch）+初始提示向量 P
输出：视觉 token 表示 Ē ∈ ℝ ^(Nv×D)+嵌入表示 P ∈ ℝ ^(T×D)

属性文本输入：S将每个属性词（如“有蹄”、“夜行性”）
用 GloVe 向量表示得到嵌入表示 S ∈ ℝ ^(Ns×D)
Ns： 属性词数量

1 INTRODUCTION

CAA内部结构

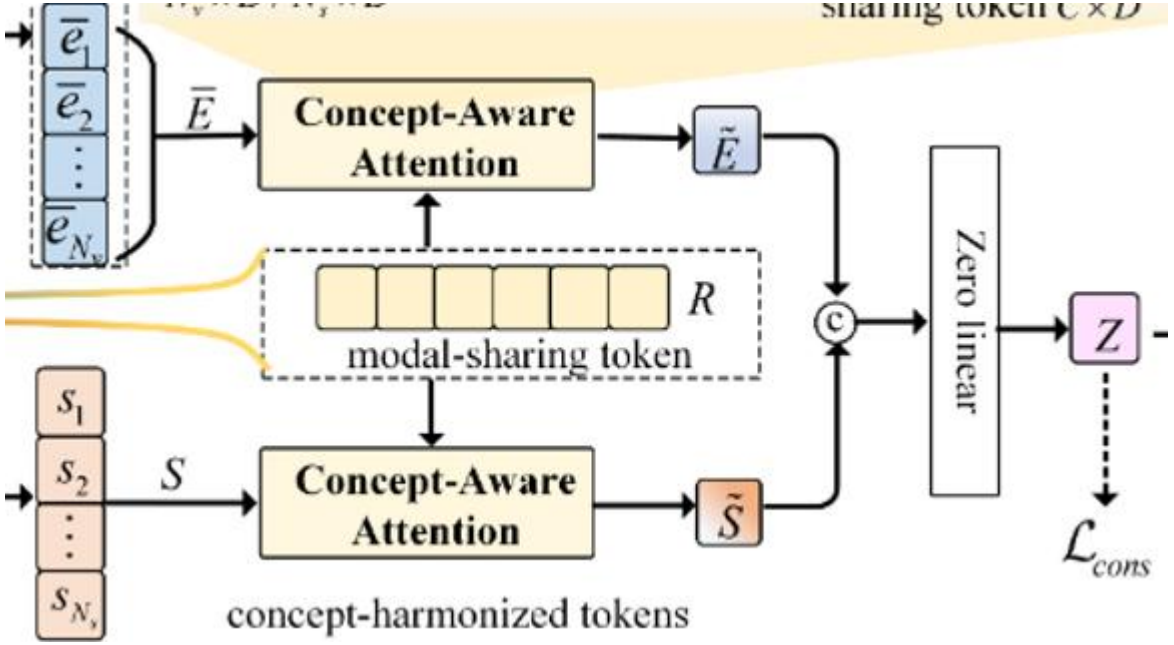


CAA 使用共享 token R作为统一的语义基准，将视觉或文本 token 映射到一个共享的概念空间。

$$\tilde{E} = \mathrm{softmax}(\mathrm{GMP}(Q_{\bar{E}} \cdot K_R^{\mathsf{T}})) \cdot V_R$$

$$\tilde{S} = \mathrm{softmax}(\mathrm{GMP}(Q_S \cdot K_R^{\mathsf{T}})) \cdot V_R$$

1 INTRODUCTION



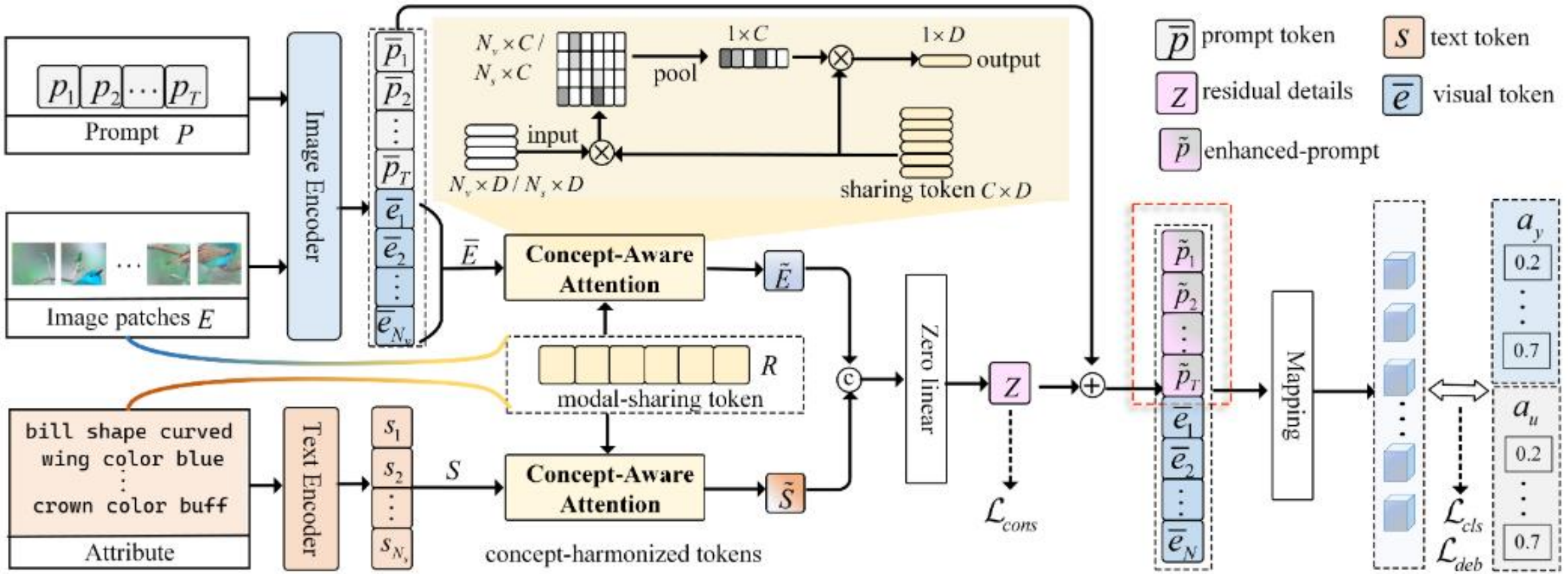
一个专门的线性预测层用来估计残差细节

$$Z = ZLinear\left(\left[\tilde{S}, \tilde{E}\right]\right)$$

确保z真的携带了与真实语义属性相一致的含义
让z尽量靠近该图像所属类别的属性向量ay

$$\mathcal{L}_{cons} = \|Z - \mathrm{MLP}(a_y)\|$$

1 INTRODUCTION



$$\tilde{P} = [\bar{p}_1 + Z, \bar{p}_2 + Z, ..., \bar{p}_T + Z]$$

1 INTRODUCTION

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{deb}\mathcal{L}_{deb}$$

$$\mathcal{L}_{deb} = \|\alpha_s - \alpha_u\|_2^2 + \|\beta_s - \beta_u\|_2^2$$

均值距离　　方差距离

2

EXPERIMENT

Methods	Venue	CUB				SUN				AwA2			
		ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
		acc	*U*	*S*	*H*	*acc*	*U*	*S*	*H*	*acc*	*U*	*S*	*H*
Generative-based Methods													
Composer (Huynh and Elhamifar 2020a)	NeurIPS'20	69.4	56.4	63.8	59.9	62.6	55.1	22.0	31.4	71.5	62.1	77.3	68.8
GCM-CF (Yue et al. 2021)	CVPR'21	–	61.0	59.7	60.3	–	47.9	37.8	42.2	–	60.4	75.1	67.0
SDGZSL (Chen et al. 2021c)	ICCV'21	75.5	59.9	66.4	63.0	–	–	–	–	72.1	64.6	73.6	68.8
CE-GZSL (Han et al. 2021)	CVPR'21	77.5	63.9	66.8	65.3	63.3	48.8	38.6	43.1	70.4	63.1	78.6	70.0
ICCE (Kong et al. 2022)	CVPR'22	78.4	67.3	65.5	66.4	–	–	–	–	72.7	65.3	82.3	72.8
FREE (Chen et al. 2021a)	ICCV'21	–	55.7	59.9	57.7	–	47.4	37.2	41.7	–	60.4	75.4	67.1
HSVA (Chen et al. 2021b)	NeurIPS'21	62.8	52.7	58.3	55.3	63.8	48.6	39.0	43.3	–	59.3	76.6	66.8
LBP (Lu et al. 2021)	TPAMI'21	61.9	42.7	71.6	53.5	63.2	39.2	36.9	38.1	–	–	–	–
f-VAEGAN+DSP (Chen et al. 2023a)	ICML'23	62.8	62.5	73.1	67.4	68.6	57.7	41.3	48.1	71.6	63.7	**88.8**	74.2
SHIP† (Wang et al. 2023)	ICCV'23	–	55.3	58.9	57.1	–	–	–	–	–	–	–	–
Embedding-based Methods													
APN (Xu et al. 2020)	NeurIPS'20	72.0	65.3	69.3	67.2	61.6	41.9	34.0	37.6	68.4	57.1	72.4	63.9
DAZLE (Huynh and Elhamifar 2020b)	CVPR'20	66.0	56.7	59.6	58.1	59.4	52.3	24.3	33.2	67.9	60.3	75.7	67.1
DVBE (Min et al. 2020)	CVPR'20	–	53.2	60.2	56.5	–	45.0	37.2	40.7	–	63.6	70.8	67.0
GEM-ZSL (Liu et al. 2021)	CVPR'21	77.8	64.8	77.1	70.4	62.8	38.1	35.7	36.9	67.3	64.8	77.5	70.6
DPPN (Wang et al. 2021)	NeurIPS'21	77.8	70.2	77.1	73.5	61.5	47.9	35.8	41.0	73.3	63.1	86.8	73.1
GNDAN (Chen et al. 2022c)	TNNLS'22	75.1	69.2	69.6	69.4	65.3	50.0	34.7	41.0	71.0	60.2	80.8	69.0
CLIP (Radford et al. 2021)	ICML'21	–	55.2	54.8	55.0	–	–	–	–	–	–	–	–
CoOP† (Zhou et al. 2022b)	IJCV'22	–	49.2	63.8	55.6	–	–	–	–	–	–	–	–
MSDN (Chen et al. 2022d)	CVPR'22	76.1	68.7	67.5	68.1	65.8	52.2	34.2	41.3	70.1	62.0	74.5	67.7
TransZero (Chen et al. 2022b)	AAAI'22	76.8	69.3	68.3	68.8	65.6	52.6	33.4	40.8	70.1	61.3	82.3	70.2
TransZero++ (Chen et al. 2022a)	TPAMI'22	78.3	67.5	73.6	70.4	67.6	48.6	37.8	42.5	72.6	64.6	82.7	72.5
DUET*† (Chen et al. 2023c)	AAAI'23	72.3	62.9	72.8	67.5	64.4	45.7	45.8	45.8	69.9	63.7	84.7	72.7
I2MVFormer* (Naeem et al. 2023)	CVPR'23	42.1	32.4	63.1	42.8	–	–	–	–	73.6	66.6	82.9	73.8
ZSLViT* (Chen et al. 2024)	CVPR'24	78.9	69.4	**78.2**	73.6	68.3	45.9	**48.3**	47.3	70.2	66.1	84.6	74.2
AENet* (Ours)	–	**80.3**	**73.1**	76.4	**74.7**	**70.4**	**58.6**	45.2	**51.0**	**75.2**	**70.3**	80.1	**74.9**

2 EXPERIMENT

消融实验

Methods	CUB				SUN				AwA2			
	ZSL	GZSL			ZSL	GZSL			ZSL	GZSL		
	acc	*U*	*S*	*H*	*acc*	*U*	*S*	*H*	*acc*	*U*	*S*	*H*
AENet w/o prompt P	70.2	67.0	69.1	68.0	64.5	53.8	25.8	34.8	70.7	61.0	88.2	72.1
AENet w/o residual details Z	75.8	72.3	72.7	72.5	68.0	58.1	38.5	46.2	73.2	65.5	85.1	74.0
AENet w/o concept-aware attention	78.8	75.6	72.1	73.8	68.8	43.6	54.9	48.6	73.7	66.8	81.8	73.5
AENet (w/ all)	80.3	73.1	76.4	74.7	70.4	58.6	45.2	51.0	75.2	70.3	80.1	74.9

Table 2: Ablation study of AENet under the ZSL and GZSL settings on CUB, SUN, and AwA2 datasets.

消融实验

Methods	CUB		AwA2	
	acc	*H*	*acc*	*H*
Linear + Skip	79.8	74.5	74.1	74.5
MLP + Skip	79.5	74.0	74.8	74.3
ZLinear + Gated Skip	79.5	74.4	73.5	74.4
ZLinear + Skip	80.3	74.7	75.2	74.9

Table 3: Ablation study of different implementations for semantic-enhanced prompt distillation.

南京郵電大學
Nanjing University of Posts and Telecommunications

DiPrompT: Disentangled Prompt Tuning for Multiple Latent Domain Generalization in Federated Learning

DiPrompT：用于联邦学习中多潜在域泛化的解耦提示调优

CVPR
2024

南京郵電大學
Nanjing University of Posts and Telecommunications

目录 | CONTENTS

01 研究背景 Introduction

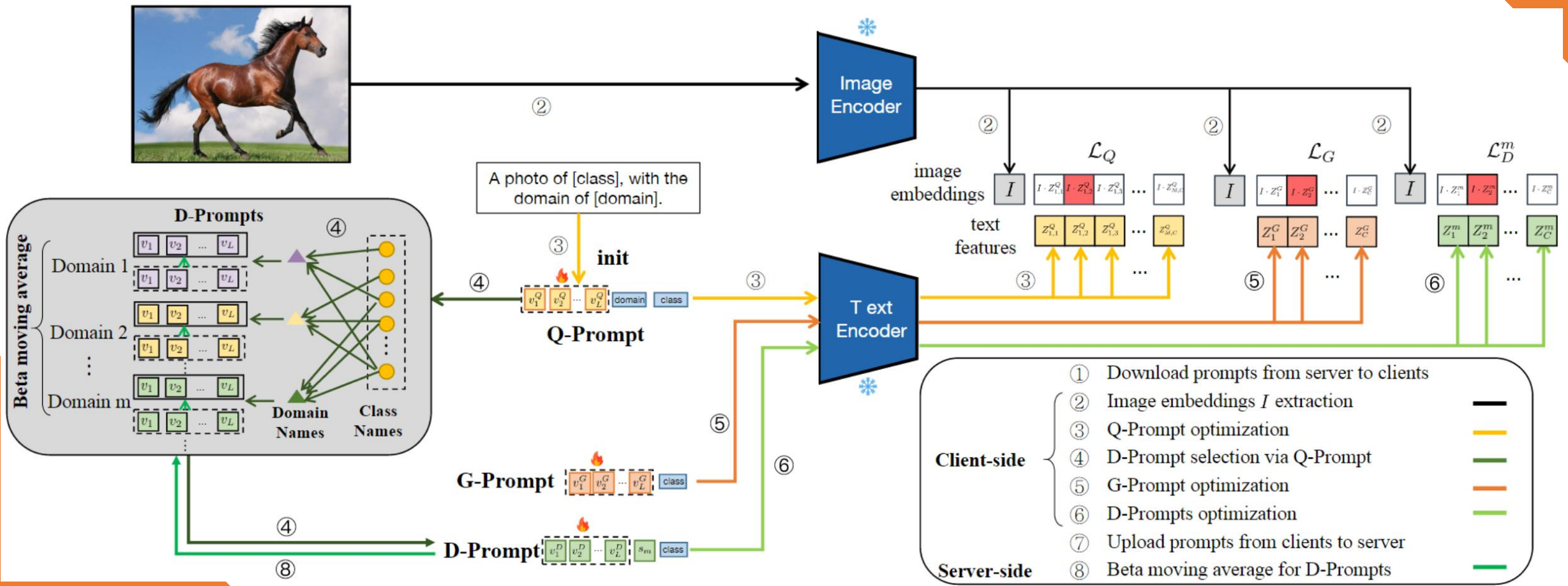
核心问题：

- **客户端数量与源域数量不匹配**：在现实场景中，客户端数量显著超过源域数量 ($K \gg M$)，导致数据分布复杂，单一模型难以同时提取全局特征和域特定知识。
- **域标签未知**：在许多实际场景中，域标签 (domain labels) 不可用，模型需要动态推断样本的潜在域归属

创新点：

- **解耦提示调优** (Disentangled Prompt Tuning)：提出全局提示 (G-Prompt) 和域提示 (D-Prompt) 的解耦机制，以分别提取全局知识和域特定知识
- **动态查询机制** (Q-Prompt)：设计了 Q-Prompt，用于在训练和推理过程中动态分配样本的潜在域标签
- **协作集成机制**：推理阶段结合全局提示和域提示的协作集成，动态权重化地利用不同提示的知识

02 核心算法 Methods



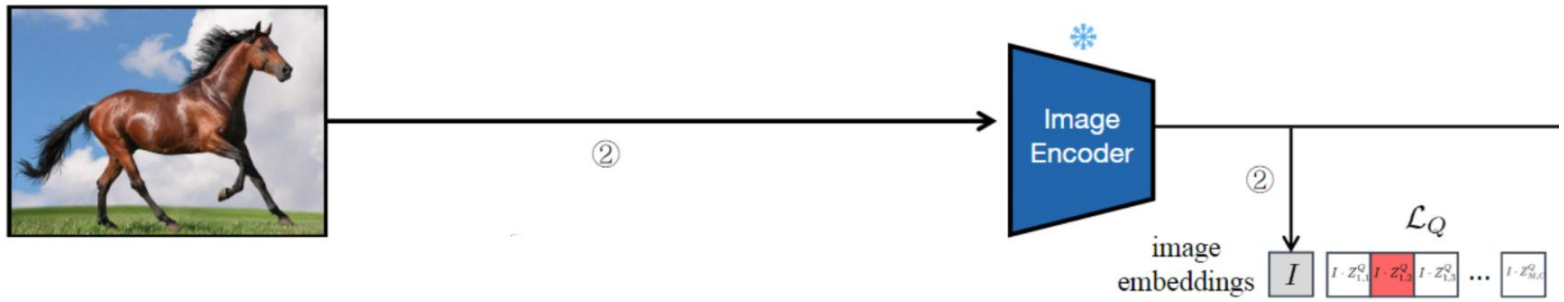
02 算法流程 Methods

下载提示：训练之前，模型的提示（包括初始全局提示V^Q和初始域提示V_m^Q）从服务器下载到每个客户端

V^Q：全局提示　　V_m^Q：域提示，针对第 m 域学习到的文本提示

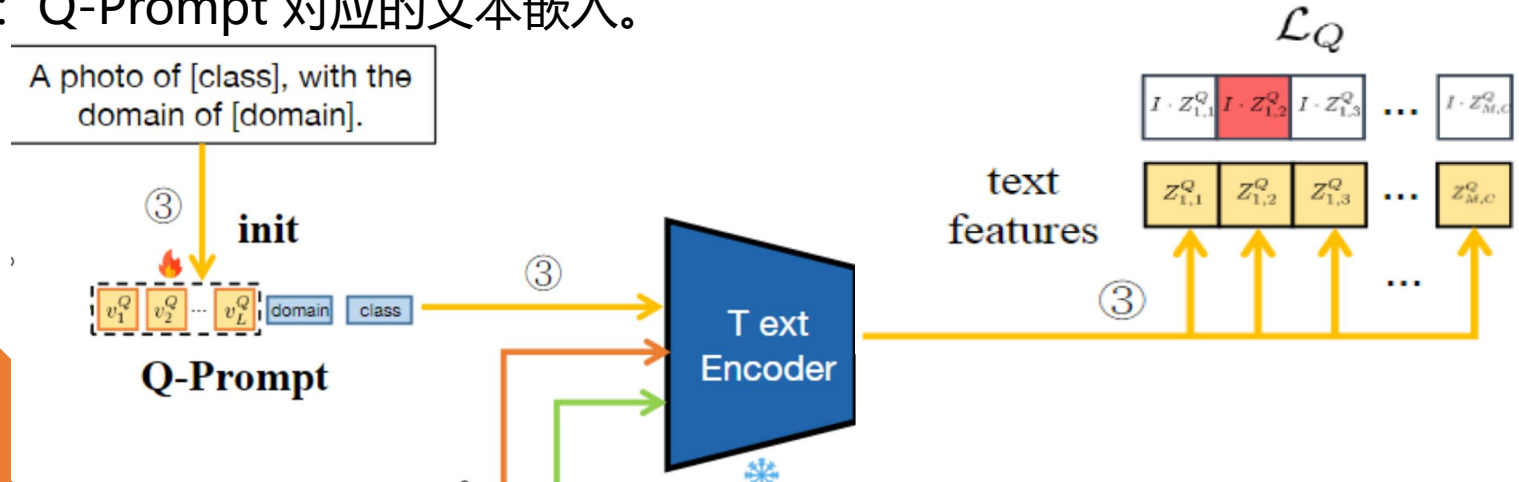
②图像嵌入提取：为每个客户端提取输入图像的特征。

I：图像嵌入向量



③Q-Prompt 优化：调整文本提示,为每个输入图像选择最合适的域标签

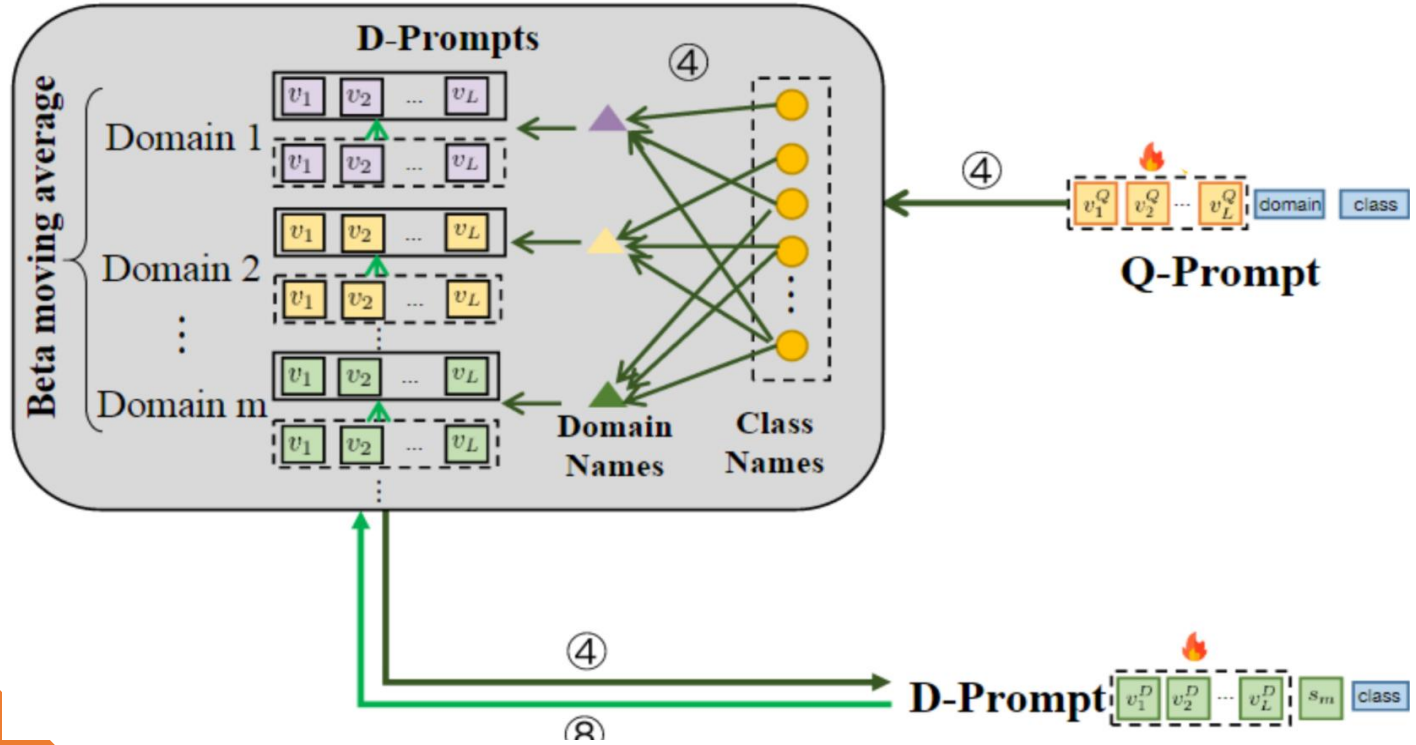
Z^Q：Q-Prompt 对应的文本嵌入。



02 算法流程 Methods

④D-Prompt 选择：根据Q-Prompt选择合适的域提示

Z_m^D：域提示对应的文本嵌入

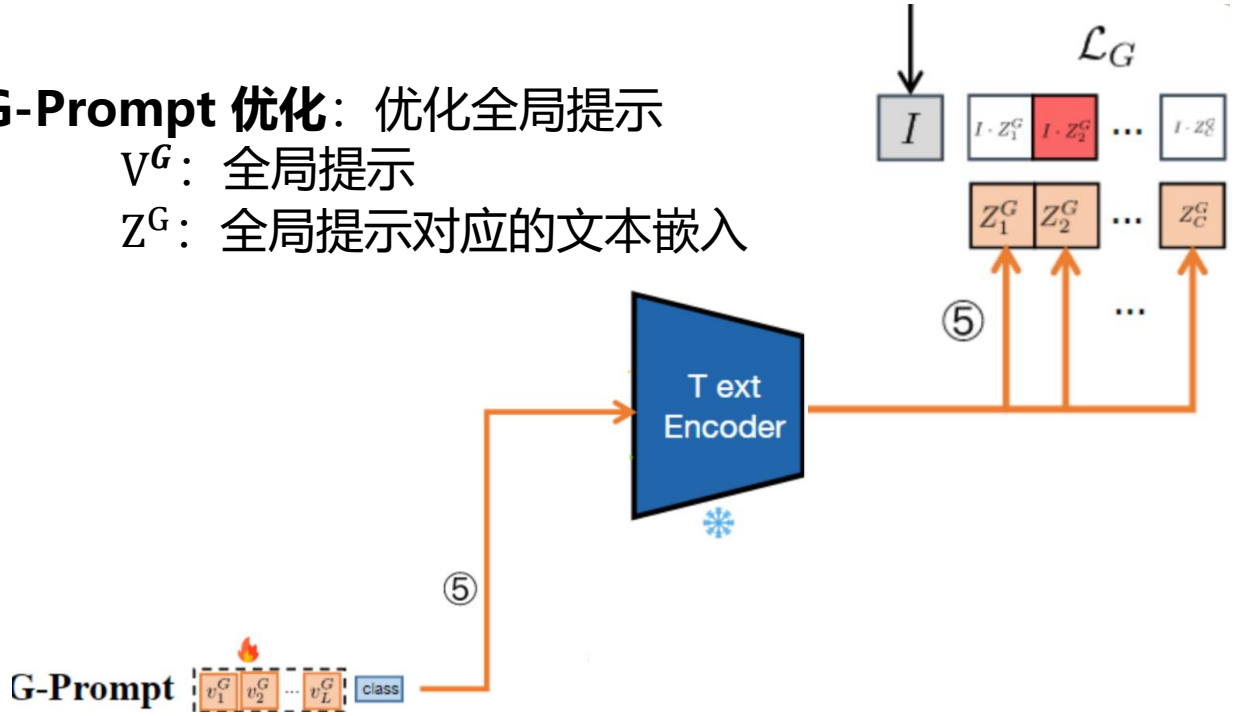


02 算法流程 Methods

⑤G-Prompt 优化：优化全局提示

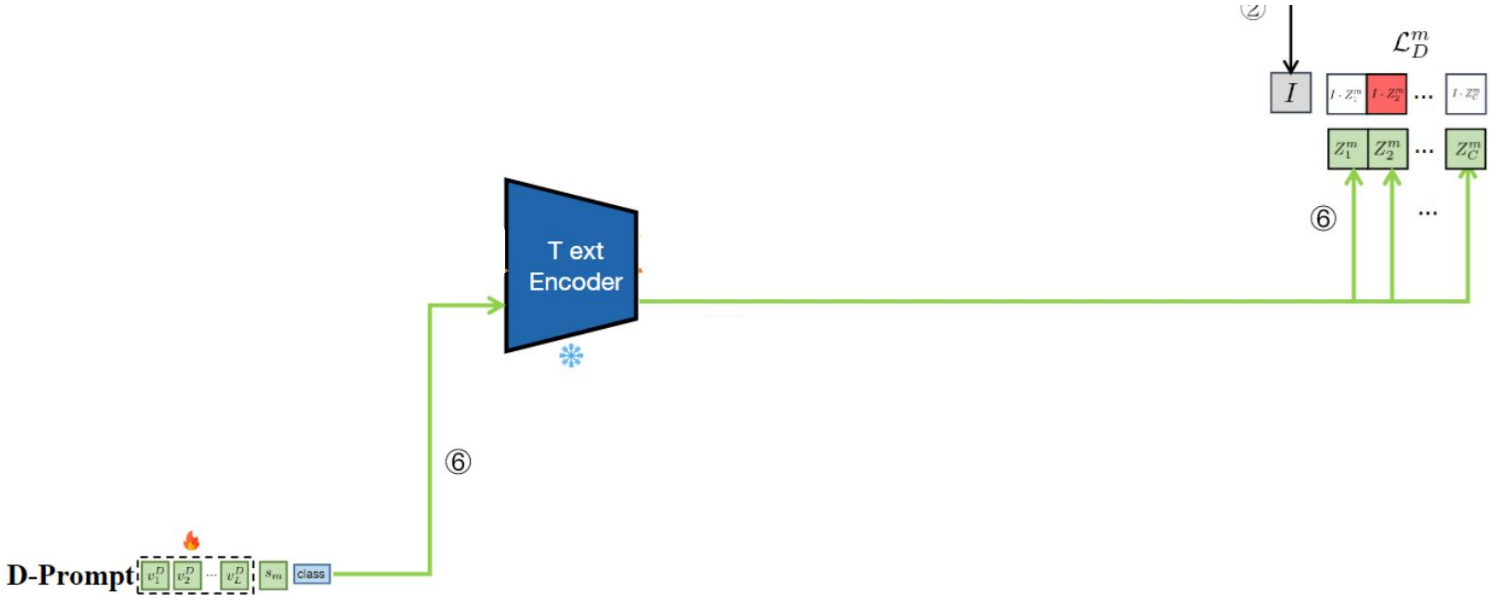
V^G：全局提示

Z^G：全局提示对应的文本嵌入



⑥D-Prompt 优化：通过优化每个域的提示,专注于每个域的特定知识

Z_m^D：域提示对应的文本嵌入



02 核心算法 Methods

全局提示的损失函数：

$$\mathcal{L}_G(x, y) = \mathbb{E}_{x,j}\mathcal{L}_{ce}(y, P_g(y = j|x)). \qquad (2)$$

全局提示的预测概率：

$$P_g(y = j|x) = \frac{\exp(\mathrm{sim}(I, Z_j)/\tau)}{\sum_{i=1}^{C} \exp(\mathrm{sim}(I, Z_i)/\tau)}, \qquad (1)$$

图像嵌入和类别 j 对应的文本嵌入的相似度

全部图像嵌入和类别 i 对应的文本嵌入的相似度求和

域提示的损失函数：

$$\mathcal{L}_D^m(x,y) = \mathcal{L}_{ce}^m + \mathcal{L}_{cont}^m$$

$$= \mathcal{L}_{ce}^m - \log \frac{\exp(\text{sim}(V_m^D \cdot \tilde{V}_m))}{\sum_{i=1}^{M} \exp(\text{sim}(V_m^D \cdot V_i^D))}, \quad (3)$$

$$\mathcal{L}_{ce}^m = \mathbb{E}_{x,j}\mathcal{L}_{ce}(y, P_m(y = j|x))$$

$$= \frac{\exp(\text{sim}(I, Z_j^m))/\tau)}{\sum_{i=1}^{C} \exp(\text{sim}(I, Z_i^m)/\tau)}$$

域提示的加权聚合：

$$V_m^{D,r+1} = V_m^{D,r} + \frac{\sum_{i=1}^{K}(|\mathcal{D}_i| * \mathcal{I}_{m,i}) \cdot \triangle V_{m,i}^{D,r+1}}{\sum_{i=1}^{K}(|\mathcal{D}_i| * \mathcal{I}_{m,i})}, \quad (4)$$

$$= V_{m,i}^{D,r+1} - V_m^{D,r}$$

动态查询的概率：

$$P(y=j,d=m|x)=\frac{\exp(\mathrm{sim}(I,Z_{j,m}^{Q})/\tau)}{\sum_{p=1}^{C}\sum_{q=1}^{M}\exp(\mathrm{sim}(I,Z_{p,q}^{Q})/\tau)}, \quad (6)$$

Q-Prompt 的优化目标：
同时最小化了当前提示和动量平均提示之间的损失（通过 MSE 损失）以及它们之间的 KL 散度

$$\mathcal{L}_Q=\mathcal{L}_{mse}+\mathcal{L}_{KL}=\mathbb{E}_{x,y}\sum_{m=1}^{M}[(Z_{y,m}^{Q})-(\hat{Z}_{y,m}^{Q})]^2 + \mathcal{D}_{KL}(\mathrm{sim}(I,Z_{y,m}^{Q}),\mathrm{sim}(I,\hat{Z}_{y,m}^{Q}))], \quad (7)$$

$$=g(\{V^{Q},c_y,s_m\})$$

$$=g(\{\hat{V}^{Q},c_y,s_m\})$$

域描述

类标签

Q-Prompt向量集合

当前提示

动量提示

泛化能力验证：

03 实验结果 Experiment

个性化能力验证：

①Pathological Non-IID setting

Table 3: Accuracy comparison (%) on the Pathological Non-IID setting over 10 clients.

Methods	OxfordPets	Flowers102	DTD	Caltech101	Food101
CoOp (Zhou et al., 2022b)	83.21±1.30	70.14±0.76	44.23±0.63	87.37±0.44	70.43±2.42
PromptFL (Zhou et al., 2022b)	90.79±0.61	72.80±1.14	54.11±0.22	89.70±1.99	77.31±1.64
PromptFL+FT (Cheng et al., 2021)	91.23±0.50	72.31±0.91	53.74±1.36	89.70±0.25	77.16±1.56
Prompt+PER (Arivazhagan et al., 2019)	89.50±1.62	72.11±1.35	50.23±0.82	86.72±1.45	71.29±1.87
Prompt+Prox (Li et al., 2020)	89.24±0.41	66.40±0.29	44.26±1.11	89.41±0.55	76.24±1.94
Prompt+AMP (Huang et al., 2021)	80.21±0.44	69.10±0.13	47.16±0.92	87.31±1.60	74.48±1.71
pFedPrompt (Guo et al., 2023a)	91.84±0.41	86.46±0.15	77.14±0.09	96.54±1.31	92.26±1.34
FedPGP	**98.96±0.42**	**99.29±0.03**	**91.52±0.41**	**98.90±0.19**	**95.52±0.15**

②Dirichlet Non-IID setting

Table 5: Accuracy comparison (%) on the Dirichlet Non-IID setting in CIFAR-10 and CIFAR-100 over 100 clients.

Methods	CIFAR-10	CIFAR-100
CLIP (Radford et al., 2021)	87.52±0.56	64.83±0.49
CoOp (Zhou et al., 2022b)	93.13±0.34	74.78±0.41
PromptFL (Zhou et al., 2022b)	92.32±0.79	73.72±0.61
Prompt+Prox (Li et al., 2020)	91.79±0.46	71.08±0.89
FedPGP	**94.82±0.37**	**77.44±0.15**

消融实验：

①正样本对和低秩适应对泛化能力的影响

Table 6: Accuracy (%) of ablation study on adaption and additional loss for clients' local classes and Base-to-novel generalization.

Methods	Local	Base	Novel	HM
FedPGP w/o Positive	94.63	84.68	77.75	85.13
FedPGP w/ Full-rank Adaption	**98.57**	48.00	63.40	64.17
FedPGP	95.67	**85.69**	**81.75**	**87.33**

②负样本对对个性化能力的影响

Table 7: Accuary (%) of ablation study on additional loss for personalization.

Methods	OxfordPets	Flowers102	DTD	Caltech101	Food101
FedPGP w/o Negative	97.65±0.20	98.63±0.11	90.78±0.31	98.48±0.17	94.72±0.18
FedPGP	**98.96±0.42**	**99.29±0.03**	**91.52±0.41**	**98.90±0.19**	**95.52±0.15**

03 实验结果 Experiment

消融实验：

μ的不同设置：

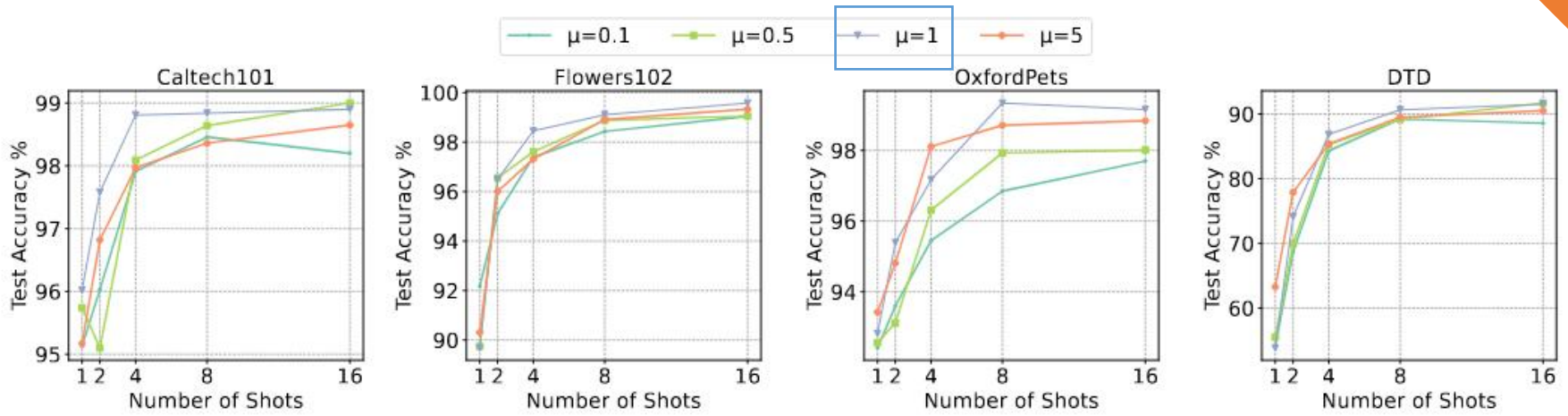


Figure 2: Quantitative comparisons on four datasets across varying shot numbers and parameter μ of contrastive loss in FedPGP over 10 clients.


南京郵電大學
1942
Nanjing University of Posts and Telecommunications
南京郵電大學
Nanjing University of Posts and Telecommunications

THANKS!