

几何门控视觉语言 模型安全适应

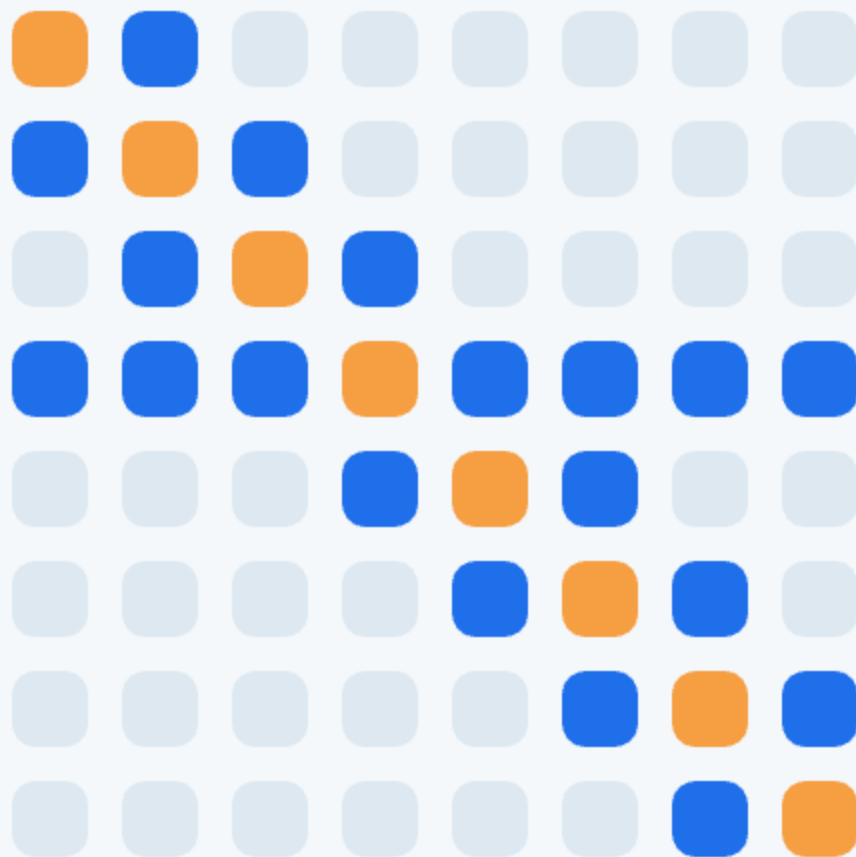
在提示更新进入测试时适应前，先判断
它是否值得信任

G²-TAAP

技术答辩·算法模型创新

参赛团队编号：AIC-2026-63696313

GEOMETRY-GATED
SAFE ADAPTATION



文本原型相似度矩阵

一个问题，一条控制链，三组证据

围绕提示更新的收益与风险展开

- | | | | |
|----|------|----------------------|-------|
| 01 | 风险入口 | 提示改变类别边界，并可能被在线反馈放大 | 03—06 |
| 02 | 方法设计 | TAAP候选、文本几何、安全门、受限融合 | 07—11 |
| 03 | 实验验证 | 匹配基线、负迁移反例、门控取舍 | 12—16 |
| 04 | 落地边界 | 部署、复现、局限与技术来源 | 17—19 |

测试时适应会持续放大初始文本原型的影响

DLAE在无标签测试流中利用预测、缓存与原型状态进行在线适应

共享图文空间

冻结 HQ-CLIP 编码器。图像特征与类别文本原型的余弦相似度决定初始分类。

目标环境变化

设备、光照与拍摄条件改变。目标域缺少可及时使用的人工标注。

在线反馈累积

早期伪标签进入缓存，并继续影响后续原型与预测。

风险入口：提示训练先改变文本边界，在线反馈随后放大其影响

负迁移沿预测、缓存与原型更新形成反馈链

控制点应位于文本原型进入 DLAE 之前



设计对象不是单独的提示准确率，而是提示进入在线适应后的系统行为

Flowers102给出明确反例：静态提升不等于适应后提升

两个独立对照分别暴露直接替换与验证集筛选的风险

直接接入

-0.73 pp

CoOp 基线

94.48%

完整 TAAP

93.75%

静态筛选

95.47%

源域验证准确率

94.23%

接入完整 DLAE

结论：必须在“发布提示权重”这一环节增加风险检查与可回退决策

设计约束与可验证承诺

把技术目标与当前证据分开表达

目标	设计方式	当前证据 / 边界
提示稳定	特征扰动分类与一致性	通过训练目标实现，尚无理论保证
无标签判定	只比较两组文本原型	门控不读取目标图像或标签
风险回退	拒绝时返回 CoOp 权重	张量逐位一致，可直接核验
单路运行	部署仅保留一个权重	不增加第二视觉编码器或缓存

接受更新后的性能仍需实验验证；回退等价性不等于跨数据集性能保证

G²-TAAP在部署前决定唯一文本权重

保留 HQ-CLIP、CoOp 与 DLAE，新增候选训练和发布决策



在线计算路径不因门控新增第二分支

TAAP让提示训练显式面向适应风险

仅更新4×512个上下文参数，HQ-CLIP编码器保持冻结

$$\mathcal{L}^{\text{TAAP}} = \mathcal{L}_{\text{ce}}(\mathbf{v}, \mathbf{y}) + 0.5 \mathcal{L}_{\text{ce}}(\tilde{\mathbf{v}}, \mathbf{y}) + 0.2 \mathcal{L}_{\text{cons}} + 0.1 \mathcal{L}_{\text{sep}}$$

源样本分类

维持任务可分性

扰动视图分类

逼近测试视图变化

预测一致性

约束干净/扰动输出

原型分离

减少最近类别拥挤

训练：AdamW · lr 0.002 · 100 epoch · 按源域验证集选择检查点

平均最近类别相似度刻画文本空间拥挤程度

每个类别只保留最相似的其他类别，再对所有类别求平均

$$C(W) = (1/K) \sum_k \max_{j \neq k} w_k^T w_j$$

① 排除自身

去除相似度矩阵对角线，只比较不同类别。

② 找最近竞争者

每个类别取最大非对角相似度。

③ 跨类别平均

得到只依赖文本权重的风险信号。

$C(W)$ 越大表示类别越拥挤；这是风险启发式，不是准确率定理

安全门决定是否更新，固定25%限制更新幅度

三个数据集使用同一规则，门控不读取目标图像或标签

$$g = 1[C(W^T) < C(W^c)]$$

$g = 1$ · 接受候选

候选混淆度下降。保留75% CoOp，吸收25% TAAP。

$g = 0$ · 拒绝候选

候选混淆度未下降。直接使用原 CoOp 文本权重。

$$W^G = \text{Norm}((1 - 0.25g)W^c + 0.25gW^T)$$

Flowers102固定融合仍下降0.20 pp，说明仅限制幅度不足以阻断负迁移

五阶段流程明确训练、发布与测试边界

前四步在部署前完成，第五步处理无标签测试流



测试标签只用于最终离线评估，不参与提示训练、风险判断或在线更新

固定主干和DLAE协议，分层衡量整体收益与门控贡献

DTD同时报告同主干DLAE、CoOp强基线与3次提示训练结果

数据集	类别数	测试样本	识别任务	重复次数
DTD	47	1,692	自然纹理	3 seeds
Caltech101	101	2,465	通用物体	1 run
Flowers102	102	2,463	细粒度花卉	1 run

对照：HQ-CLIP+DLAE · CoOp · 完整TAAP · 固定25%融合 · 双层META · G²-TAAP

限制：Caltech101与Flowers102尚缺多随机种子；Flowers102参与问题发现，不作为完全独立的泛化证明

DTD相对同主干DLAE提升11.05个百分点

整体方案收益与几何门控增量分开报告

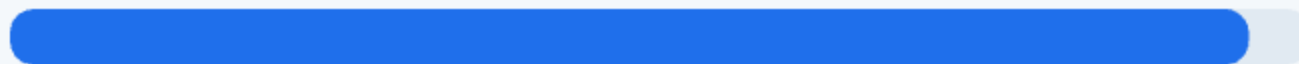
HQ-CLIP + DLAE



65.43%

同主干基础适应

+ CoOp



76.24%

相对DLAE +10.82 pp

+ 完整TAAP



77.36%

收益更高，但跨数据集出现负迁移

+ G²几何门



76.48%

相对DLAE +11.05 pp

+11.05 pp

相对HQ-CLIP+DLAE

65.43 → 76.48

+0.24 pp

相对CoOp+DLAE

76.24 → 76.48

前者体现完整系统收益，后者隔离几何门的严格贡献

门控阻断Flowers102退化，但也拒绝部分有效更新

安全性与收益之间存在可量化取舍

实验	C(CoOp) → C(TAAP)	决策	G ² -TAAP结果
DTD seed 1	0.798171 → 0.798044	接受	76.95% (+0.71)
DTD seed 2 / 3	混淆度未下降	拒绝	均回退76.24%
Caltech101	0.760264 → 0.574165	接受	97.20% (+0.24)
Flowers102	0.660269 → 0.680087	拒绝	94.48% (回退)

DTD seed 1差值仅0.000127：下一步需加入容差阈值、重复计算稳定性与连续门控消融

消融结果区分候选质量、更新幅度与门控作用

“—”表示现有材料未报告该组合

方法	DTD	Caltech101	Flowers102	结论
CoOp基线	76.24	96.96	94.48	匹配基线
完整TAAP	77.36±0.21	97.00	93.75	收益高但发生负迁移
固定25%融合	76.89	97.20	94.28	限制幅度仍不足
验证集类别替换	—	—	94.23	静态验证未保护测试流
双层META	77.01	—	—	可行但未超过TAAP
G ² -TAAP	76.48±0.41	97.20	94.48	门控实现回退

新增训练与几何检查在部署前完成

在线仍为一条模型分支和一套DLAE缓存

2,048

提示参数

4×512上下文token

1

模型分支

不新增视觉编码器

1

缓存系统

沿用DLAE在线路径

$O(K^2d)$

门控复杂度

部署前计算一次

环境	已验证值
Python / PyTorch	3.12.14 / 2.7.1+cu128
open_clip	2.32.0
检查点	13个轻量提示/门控权重，均附SHA-256
核心入口	训练 → 建门控 → 完整测试 → 汇总

墙钟时间来自不同批次，仅作参考，不用于宣称加速

最适合负迁移代价高、目标域无标签的持续识别任务

当前处于公开数据集验证阶段，尚无企业试点或真实用户数据

农业病害识别

季节、设备与光照持续变化。错误更新可能影响后续批次判断。

工业缺陷分类

产线与相机条件变化。系统需要可审计的发布与回退记录。

细粒度物种识别

类别相似度高。文本原型拥挤更容易引发类间混淆。

工程价值

更新前判定 · 单权重部署 · 明确回退 · 检查点可追溯

下一步需要真实业务数据、端侧时延与失败成本验证

当前证据支持“已测范围内的风险控制”

不宣称跨骨干、跨数据集或任意测试流的性能保证

当前边界	下一步验证
仅HQ-CLIP ViT-B/16	增加RN50、ViT-L及不同预训练模型
三个数据集，两个仅单次运行	统一3—5种子并报告置信区间
Flowers102参与问题发现	冻结算法后增加ImageNet-C/A/R/Sketch
门控采用严格小于	验证容差阈值、顺序扰动与连续门控
仅离线公开数据	补充真实业务流、成本和用户反馈

已有技术与本项目改进边界

引用原始论文和官方模型页，避免把第三方贡献表述为原创

已有基础	来源	本项目如何使用
CoOp	Zhou et al., IJCV 2022; arXiv:2109.01134	共享上下文提示参数化
HQ-CLIP	Wei et al., ICCV 2025; arXiv:2507.22431	冻结ViT-B/16图文编码器
DLAE	Wu et al., CVPR 2026	作为在线测试时适应下游
G ² -TAAP	本项目	TAAP训练目标、几何安全门、25%受限融合

官方资源：github.com/KaiyangZhou/CoOp · zxwei.site/hqclip · openaccess.thecvf.com

G²-TAAP

把文本几何风险检查加入提示发布流程

01 候选训练

TAAP面向适应扰动学习提示

02 发布决策

几何改善才允许25%融合

03 失败回退

风险上升时沿用CoOp权重

已验证：DTD相对同主干DLAE +11.05 pp · Flowers102安全回退 · 在线单路运行

感谢聆听 · 敬请指正