中图法分类号:TP751　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Lin Zehui, Shen Zhongwei. Compositional zero-shot learning with style debiasing and local semantic focusing[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260280.(林泽辉, 沈忠伟. 风格解偏与局部语义聚焦的组合零样本学习[J/OL]. 中国图象图形学报,XXXX:1-15. DOI: 10.11834/jig.260280.)[DOI:10.11834/jig.260280]

风格解偏与局部语义聚焦的组合零样本学习

林泽辉,沈忠伟*

苏州科技大学电子与信息工程学院,苏州　　215009

摘　要:**目的**　组合零样本学习旨在利用训练阶段已见的状态与物体原语识别测试阶段未见的状态—物体组合。现有基于视觉语言预训练模型的方法主要关注组合语义建模和全局跨模态对齐,但仍存在对训练图像风格分布依赖较强、局部判别区域利用不足等问题,导致模型在未见组合场景下泛化能力受限。**方法**　针对上述问题,提出一种风格解偏与局部语义聚焦相结合的组合零样本学习方法。在视觉编码前端,引入风格解偏模块,通过对中间卷积特征的通道均值和标准差进行随机混合,降低模型对颜色、纹理和背景统计模式的依赖;在跨模态交互阶段,设计文本锚引导的局部语义聚焦模块,利用组合、属性和物体分支的文本原型构造语义锚点,对视觉 patch token 进行相似度计算、Top-K 稀疏筛选和重加权,从而突出与当前语义相关的关键局部区域。在此基础上,结合三分支关系建模框架完成属性、物体和组合的联合匹配预测。**结果**　在 MIT-States、UT-Zappos 和 C-GQA 3个基准数据集上进行闭世界和开放世界实验。闭世界设定下,本文方法在 MIT-States、UT-Zappos 和 C-GQA 上的 AUC 分别达到 22.9%、42.8% 和 13.1%;开放世界设定下,AUC 分别达到 7.8%、33.8% 和 3.24%,均优于基线模型。消融实验表明,风格解偏模块和局部语义聚焦模块均能带来稳定性能提升,二者结合后取得最优结果。**结论**　所提方法能够有效缓解视觉风格偏移对组合识别的影响,并增强模型对局部关键语义证据的关注能力,从而提升组合零样本学习中未见状态—物体组合的识别性能和泛化能力。

关键词:组合零样本学习;视觉语言模型;风格解偏;局部语义聚焦;跨模态对齐

Compositional zero-shot learning with style debiasing and local semantic focusing

Lin Zehui, Shen Zhongwei*

School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou; 215009, China

Abstract: **Objective**　Compositional zero-shot learning (CZSL) aims to recognize unseen state-object compositions by transferring knowledge from seen states, objects and their observed combinations. Unlike conventional closed-set recognition, CZSL requires a model not only to identify primitive concepts, but also to recombine them into novel semantic compositions without direct visual supervision. Recent methods based on vision-language pre-trained models have improved CZSL by using large-scale image-text knowledge, prompt learning and cross-modal alignment. However, two issues remain insufficiently explored. First, visual representations may still be biased toward the color, texture and background statistics of training images, which limits their robustness to unseen compositions. Second, state semantics are often determined by

收稿日期:2026-05-17;**修回日期**:2026-08-10

*通信作者:沈忠伟　shenzw@usts.edu.cn

基金项目:国家自然科学基金项目(62306203)

Supported by: National Natural Science Foundation of China(62306203)

local regions, such as the texture of a wet dog, the surface of a rusty car or the cut surface of a sliced tomato. Global feature interaction may introduce irrelevant background information and weaken local discriminative cues. To address these problems, this paper proposes a CZSL method based on style debiasing and local semantic focusing. **Method** The proposed method is built upon a three-branch vision-language framework, including a composition branch, an attribute branch and an object branch. To improve visual robustness, a two-dimensional style debiasing module is inserted into the visual front end. Specifically, after the convolutional patch embedding layer of the CLIP visual encoder, the channel-wise mean and standard deviation of intermediate feature maps are computed. The feature maps are normalized and then reconstructed with mixed style statistics sampled from different images in the same batch. This operation perturbs feature-level style distributions while preserving the main semantic content, thereby reducing the dependence of the model on training-specific visual styles. To enhance local semantic discrimination, a text-anchored local semantic focusing module is further designed. Text prototypes are generated for the composition, attribute and object branches, and their aggregated representations are used as branch-level semantic anchors. The semantic anchor is matched with visual patch tokens through normalized similarity calculation. A Top-K sparse selection strategy is then adopted to retain the most relevant patch tokens, and the selected tokens are re-weighted according to their semantic relevance. The resulting weighted visual token sequence is used for cross-modal relationship modeling. Finally, the three branches jointly predict composition, attribute and object matching scores, and the final state-object prediction is obtained by fusing these scores. **Result** Experiments are conducted on three benchmark datasets, including MIT-States, UT-Zappos and C-GQA, under both closed-world and open-world settings. Four metrics are used for evaluation: Seen accuracy (S), Unseen accuracy (U), Harmonic Mean (H) and Area Under Curve (AUC). In the closed-world setting, the proposed method achieves 50.9% S, 53.7% U, 39.8% H and 22.9% AUC on MIT-States; 67.6% S, 74.4% U, 55.3% H and 42.8% AUC on UT-Zappos; and 41.7% S, 36.4% U, 30.2% H and 13.1% AUC on C-GQA. In the open-world setting, it achieves 49.8% S, 20.3% U, 20.9% H and 7.8% AUC on MIT-States; 67.4% S, 62.1% U, 48.4% H and 33.8% AUC on UT-Zappos; and 41.6% S, 8.3% U, 11.7% H and 3.24% AUC on C-GQA. Compared with the baseline three-branch model, the proposed method consistently improves performance on all three datasets, especially on H and AUC, indicating a better balance between seen and unseen compositions. Ablation experiments on MIT-States show that adding only the style debiasing module improves the closed-world AUC from 22.1% to 22.4%, adding only the local semantic focusing module improves it to 22.6%, and using both modules further improves it to 22.9%. These results verify the complementary effects of the two modules. **Conclusion** This paper presents a compositional zero-shot learning method that combines style debiasing and local semantic focusing. The style debiasing module alleviates the influence of training-specific visual statistics, while the text-anchored local semantic focusing module enhances the selection of discriminative patch-level evidence. By integrating these modules into a three-branch cross-modal framework, the proposed method improves both visual robustness and local semantic discrimination. Experimental results, ablation studies and qualitative analysis demonstrate its effectiveness in recognizing unseen state-object compositions. Future work will explore more fine-grained region-level semantic grounding and stronger external knowledge constraints for difficult compositions with subtle states, weak local cues or complex backgrounds.

Key words: compositional zero-shot learning; vision-language model; style debiasing; local semantic focusing; cross-modal matching

论文引用格式:DOI:10.11834/jig.260280

0 引 言

组合零样本学习(compositional zero-shot learning,CZSL)旨在利用训练阶段已见的状态与物体原语,识别测试阶段未出现过的状态—物体组合,是衡量模型组合泛化能力的重要研究任务(Misra 等,2017;Nagarajan 和 Grauman,2018)。与传统闭集识别不同,组合零样本学习不仅要求模型准确表征状态和物体等基础语义成分,还要求其在缺少目标组合监督样本的情况下,将已学习的原语重新组合并完成联合判别。例如,模型在训练阶段分别学习“湿润”“狗”以及其他相关组合后,应能够识别未直接见过的“湿润的狗”。然而,同一状态在不同物体语境中的视觉表现可能存在明显差异,状态与物体之间

也具有复杂的上下文依赖关系，简单地拼接独立原语表示难以稳定支持未见组合识别。

早期组合零样本学习研究主要围绕组合表示和原语关系建模展开。Misra等人(2017)关注属性在不同物体语境中的上下文依赖问题；Nagarajan和Grauman(2018)将属性建模为作用于物体表征的变换算子；Purushwalkam等人(2019)利用模块化网络完成组合推理；Li等人(2020)从对称性和群结构角度约束状态与物体之间的关系。在此基础上，图嵌入和关系传播机制被用于增强原语及其组合之间的结构表达能力，以缓解单一组合函数表达不足的问题(Naeem等，2021)。这些方法推动了组合表示学习的发展，但其视觉和语义表征通常依赖特定任务数据，面对未见组合及复杂视觉环境时仍存在一定局限。

随着视觉语言预训练模型的发展，研究重点逐渐转向利用大规模图文先验知识提高组合泛化能力。Radford等人(2021)提出的CLIP通过大规模图文对比学习获得了较强的开放词汇视觉—语言表征，为组合零样本学习提供了新的技术基础。Nayak等人(2023)提出组合软提示方法CSP，将状态和物体表示为可学习词汇，并通过软提示重组未见组合；Lu等人(2023)提出DFSP，利用可学习软提示和分解式融合模块联合建模视觉与语言信息。近期研究进一步从单一组合提示扩展到多粒度语义建模，通过分别刻画状态、物体及其组合关系，减少模型对已见组合的过拟合。三分支框架分别建模属性、物体和组合语义，实现基础原语与组合关系的协同判别，已成为视觉语言组合零样本学习中的一类重要技术路线(Huang等，2024；Liu等，2026)。

相较于仅在给定候选组合空间内进行识别的闭世界设定，开放世界组合零样本学习需要在更大的组合空间中同时区分合理组合和语义上不成立的组合，更加贴近真实应用环境。现有研究主要通过组合可行性建模，在全部可能的状态—物体排列中筛除不合理组合(Mancini等，2021)。将状态与物体分别建模，并借助外部知识估计组合可行性，能够降低大规模输出空间带来的搜索压力，提高未见组合识别的稳定性(Karthik等，2022)。进一步地，图像—文本交互、组合生成和候选筛选机制也被引入开放世界框架，以增强模型对合理组合的辨别能力(Jayasekara等，2025)。总体而言，现有方法主要关注文本提示构造、原语关系建模、组合语义重组和候选可行性筛选，而对视觉表征自身的鲁棒性及局部关键证据的精细利用仍缺少系统研究。

具体而言，现有视觉语言组合零样本学习方法仍面临视觉风格伪相关和局部语义证据不足两方面问题。一方面，训练数据中的颜色、纹理、背景和成像风格可能与某些已见组合频繁共同出现，使模型将表面视觉统计误认为稳定的状态判别依据。当测试图像中的物体语境或视觉风格发生变化时，这类偶然相关容易削弱模型对未见组合的识别能力。已有域泛化研究表明，视觉风格差异通常反映在中间特征的通道统计量中，对特征统计进行适度扰动有助于降低模型对训练风格的依赖(Zhou等，2021)。然而，现有组合零样本学习方法主要从文本提示、原语解耦和组合关系等角度提高泛化能力，对视觉统计偏置缺少针对性建模。另一方面，状态语义通常集中在少量局部区域。例如，“湿润”主要体现于物体表面的潮湿纹理，“切片”主要体现于切面及边缘结构，“生锈”则主要体现于局部材质变化。现有跨模态交互通常使全部视觉Token参与稠密关系建模，背景和无关区域容易稀释关键局部证据。同时，属性、物体和组合分支所需的视觉信息并不完全一致，使用相同的全局视觉表示服务不同语义分支，难以充分发挥三分支建模的差异性。

针对上述问题，本文提出一种结合风格解偏与局部语义聚焦的三分支组合零样本学习方法。首先，在CLIP视觉编码器卷积Patch Embedding后的二维特征上引入风格解偏模块，通过混合批内不同样本的通道均值和标准差，扰动颜色、纹理和背景等表面统计模式，降低已见组合视觉风格与状态语义之间的过度耦合。该模块仅在训练阶段对浅层视觉特征进行随机扰动，在保留目标轮廓、空间结构和局部相对响应的同时，增强视觉表示的跨风格鲁棒性。其次，设计分支文本锚引导的Top-K局部语义聚焦模块，分别利用属性、物体和组合分支的文本原型构造语义锚点，并在跨模态关系建模前对视觉Patch Token进行相关性评估、稀疏筛选和重新加权。该机制使属性分支更多关注状态变化区域，物体分支侧重目标主体及整体结构，组合分支兼顾状态表现和物体语境。通过“视觉统计解偏—分支局部证据筛选—跨模态关系建模”的递进过程，所提方法同时缓解视觉风格偏置和局部语义稀释问题，提升模型

对未见状态—物体组合的识别能力。

综上所述,本文的主要贡献如下:

1)针对组合零样本学习模型容易依赖训练图像颜色、纹理和背景统计的问题,将二维特征统计解偏机制引入CLIP视觉编码前端,通过训练阶段的受限统计扰动降低模型对特定视觉风格的过度依赖,同时保留与状态和物体识别相关的主要语义信息。

2)提出分支文本锚引导的Top-K局部语义聚焦机制,根据属性、物体和组合分支的文本语义,对视觉Patch Token进行分支级相关性评估、稀疏筛选和重新加权,为不同语义分支提供更具针对性的局部视觉证据。

3)构建融合视觉统计解偏、局部语义聚焦和跨模态关系建模的三分支组合学习框架,并在MIT-States、UT-Zappos和C-GQA三个基准数据集的闭世界与开放世界设定下开展实验。实验结果、消融分析和可视化结果表明,所提方法能够提升未见组合识别性能,并改善已见组合与未见组合之间的平衡能力。

1 方法介绍

本文提出一种基于风格解偏与局部语义聚焦的跨模态组合零样本识别方法,旨在识别由状态与物体构成的未见组合。整体框架如图1所示,在主流的三分支框架上进行两项增强:一是在视觉端引入风格解偏模块,以降低模型对训练图像颜色、纹理和背景统计模式的依赖;二是引入局部语义聚焦模块,以利用文本先验对局部视觉token进行筛选与重加权。整体方法保持属性、物体和组合三分支联合建模范式,同时增强视觉鲁棒性与局部判别性。

1.1 问题定义

组合零样本学习旨在利用训练阶段已见的状态集合 $\boldsymbol{A}=\{a_1,a_2,\cdots,a_{|A|}\}$ 与物体集合 $\boldsymbol{O}=\{o_1,o_2,\cdots,o_{|O|}\}$,识别测试阶段未见的状态—物体组合。式中,$a_i$ 表示第 i 个状态,o_j 表示第 j 个物体,$|\boldsymbol{A}|$ 和 $|\boldsymbol{O}|$ 分别表示状态类别数和物体类别数。记全部状态—物体组合构成的组合空间为 $\boldsymbol{C}=\boldsymbol{A}\times\boldsymbol{O}$,其中,×表示笛卡尔积。训练阶段仅能观测到已见组合集合 $\boldsymbol{C}_{\mathrm{s}}\subset\boldsymbol{C}$,测试阶段需要识别未见组合集合 $\boldsymbol{C}_{\mathrm{u}}\subset\boldsymbol{C}$,且满足 $\boldsymbol{C}_{\mathrm{s}}\cap\boldsymbol{C}_{\mathrm{u}}=\varnothing$。组合零样本学习的核心在于使模型能够由已见组合推广到未见组合。这要求模型既能学习状态和物体各自的独立语义,又能建模二者之间的组合关系,从而适应未见状态—物体组合带来的语义变化与分布差异。

1.2 风格解偏

在组合零样本学习中,模型不仅需要识别未见的状态—物体组合,还需要应对背景纹理、颜色分布和成像风格变化带来的视觉统计偏移。若模型过度依赖训练阶段的表面风格特征,则容易削弱其对未见组合的泛化能力。为缓解这一问题,本文在视觉前端引入风格解偏机制,在中间卷积特征层面对样本风格统计量进行扰动与重组,从而降低模型对单一训练风格分布的依赖。

给定输入图像 $\boldsymbol{X}\in\mathbf{R}^{B\times3\times H\times W}$,首先通过视觉编码器提取中间卷积特征和视觉token序列,记为

$$\boldsymbol{F}=E_v^{\mathrm{f}}(\boldsymbol{X})\in\mathbf{R}^{B\times C\times H'\times W'}\#(1)$$

$$\boldsymbol{P}=E_v^{\mathrm{p}}(\boldsymbol{X})\in\mathbf{R}^{B\times(1+N)\times D}\#(2)$$

式中,$\boldsymbol{X}\in\boldsymbol{R}^{B\times3\times H\times W}$ 表示输入图像批次,B 表示批量大小,3表示输入图像的颜色通道数,H 和 W 分别表示输入图像的高度和宽度;$E_v^{\mathrm{f}}(\cdot)$ 和 $E_v^{\mathrm{p}}(\cdot)$ 分别表示视觉编码器中用于生成二维卷积特征和视觉Token序列的映射;$\boldsymbol{F}\in\boldsymbol{R}^{B\times C\times H'\times W'}$ 表示中间二维特征图,C 表示特征通道数,H' 和 W' 分别表示中间特征图的高度和宽度;$\boldsymbol{P}\in\boldsymbol{R}^{B\times(1+N)\times D}$ 表示包含一个CLS Token和 N 个Patch Token的视觉Token序列,N 表示Patch Token数量,D 表示视觉特征维度。对于第 b 个样本在第 c 个通道上的特征,分别计算其空间均值和标准差:

$$\mu_{b,c}=\frac{1}{H'W'}\sum_{h=1}^{H'}\sum_{w=1}^{W'}\boldsymbol{F}_{b,c,h,w}\#(3)$$

$$\sigma_{b,c}=\sqrt{\frac{1}{H'W'}\sum_{h=1}^{H'}\sum_{w=1}^{W'}\left(\boldsymbol{F}_{b,c,h,w}|\mu_{b,c}\right)^2+\varepsilon}\#(4)$$

式中,ε 为防止数值不稳定的极小常数。然后对特征进行标准化处理:

$$\hat{\boldsymbol{F}}_{b,c,h,w}=\frac{\boldsymbol{F}_{b,c,h,w}-\mu_{b,c}}{\sigma_{b,c}}\#(5)$$

随后,在batch内随机打乱样本顺序,并从Beta分布中采样混合系数 $\lambda_b\sim\mathrm{Beta}(\alpha,\alpha)$。设 $\pi(b)$ 表示与第 b 个样本配对的打乱索引,则混合后的统计量定义为

$$\tilde{\mu}_{b,c}=\lambda_b\mu_{b,c}+(1|\lambda_b)\mu_{\pi(b),c}\#(6)$$

$$\tilde{\sigma}_{b,c}=\lambda_b\sigma_{b,c}+(1|\lambda_b)\sigma_{\pi(b),c}\#(7)$$

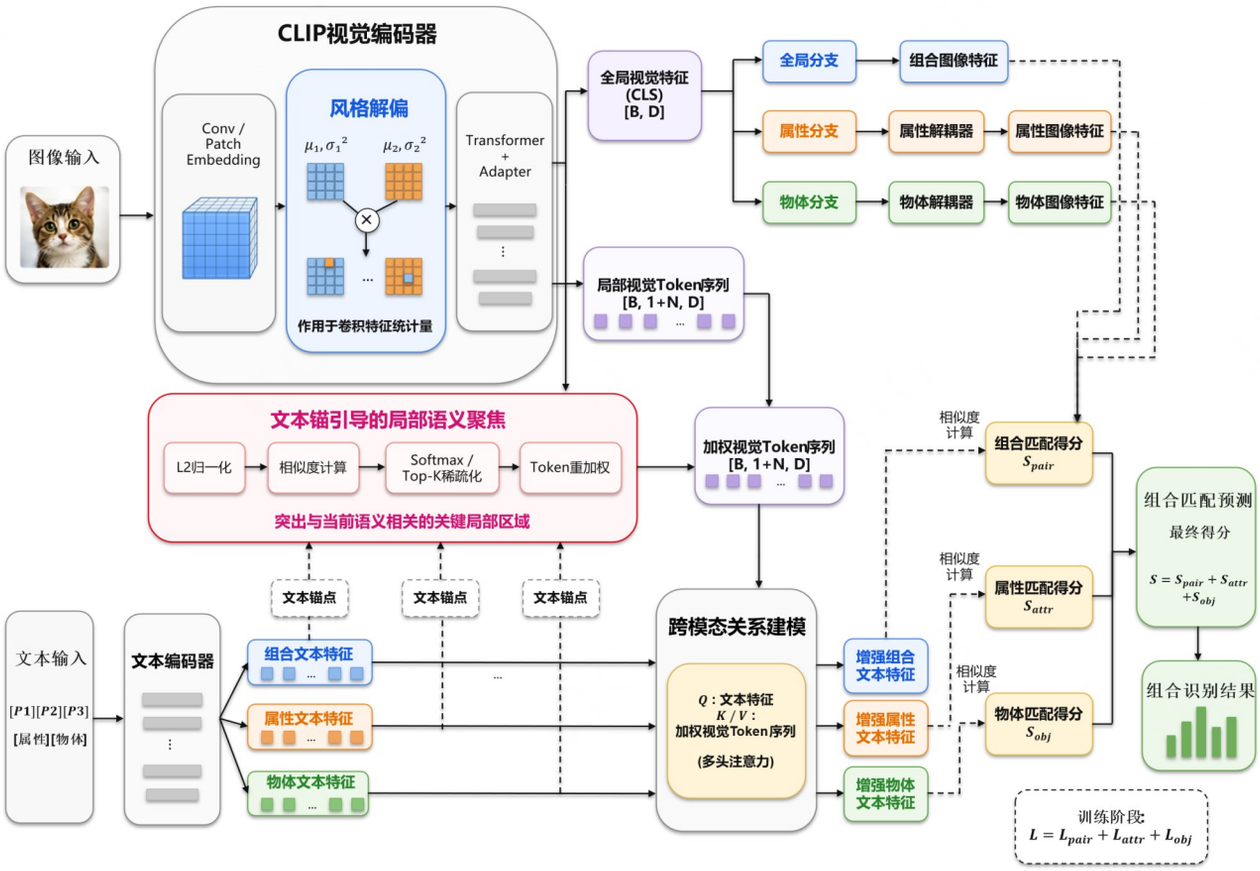

图1　模型整体架构示意图

Fig. 1　Overview of the model architecture

最终得到经风格解偏后的视觉特征：

$$\tilde{F}_{b,c,h,w} = \hat{F}_{b,c,h,w}\tilde{\sigma}_{b,c} + \tilde{\mu}_{b,c} \#(8)$$

式(8)通过重组不同样本的通道统计量，在保持语义内容相对稳定的同时打破训练样本中固有的风格偏置，从而增强模型的跨风格鲁棒性。得到解偏后的中间特征 $\tilde{F}$ 后，继续送入后续视觉分支，生成对应的全局视觉特征和局部视觉 token 表示。这样，后续局部语义聚焦与跨模态关系建模所使用的视觉表示能够减少对训练数据中特定表面统计模式的依赖，同时保留目标结构和局部响应等主要语义信息。

需要说明的是，视觉风格信息与属性判别信息并非完全独立，部分属性本身也可能依赖颜色、纹理和材质等视觉线索。因此，本文的风格解偏并不是去除所有风格信息，而是通过适度扰动训练样本的通道统计量，降低模型对特定颜色分布、纹理强度和背景模式的过度依赖。该操作仅重组特征的通道均值和标准差，原样本中反映目标轮廓、空间结构及局部相对响应的标准化特征仍被保留，因而能够在一定程度上维持属性与物体的主要判别信息。此外，风格解偏仅作用于视觉编码前端的浅层二维特征，并以随机方式在训练阶段启用，验证和测试阶段不执行该操作，从而避免对高层属性语义产生持续干扰。通过这种受限的统计扰动，模型能够在保留有效属性线索的同时，减少对训练数据中特定表面视觉模式的过拟合，提高未见属性-物体组合下的视觉鲁棒性。

1.3　局部语义聚焦

尽管经过风格解偏后的视觉表示具有更强的鲁棒性，但在组合零样本学习中，状态语义往往仅对应图像中的少量局部区域，例如“湿的”“旧的”“碎的”等状态通常集中体现在目标局部纹理或外观细节上。若直接基于全部视觉 token 与文本语义进行交互，则背景区域和无关纹理容易削弱关键判别证据。为此，本文设计文本锚引导的局部语义聚焦机制，通

过分支文本原型构造语义锚点，并利用锚点对patch级视觉token进行稀疏筛选和重加权。

首先，在文本端分别构造组合、属性和物体三类提示模板，并将其送入冻结的CLIP文本编码器，得到对应分支的文本原型：

$$\boldsymbol{T}_c \in \mathbf{R}^{K_c \times D}, \boldsymbol{T}_a \in \mathbf{R}^{K_a \times D}, \boldsymbol{T}_o \in \mathbf{R}^{K_o \times D} \#(9)$$

式中，$\boldsymbol{T}_c$、$\boldsymbol{T}_a$、$\boldsymbol{T}_o$分别表示组合、属性和物体分支的文本原型集合；K_c、K_a和K_o分别表示三个分支包含的文本原型数量；D表示文本原型的特征维度，与视觉特征维度保持一致。为了从文本原型中提炼更具代表性的分支语义，引入分支语义锚点。对于分支$b \in \{c, a, o\}$，其锚点定义为

$$\boldsymbol{q}_b = \mathrm{mean}(\boldsymbol{T}_b) = \frac{1}{K_b}\sum_{k=1}^{K_b} \boldsymbol{t}_{b,k} \#(10)$$

式中，$b \in \{c, a, o\}$表示分支索引，其中c、a和o分别对应组合、属性和物体分支；K_b表示分支b中的文本原型数量；$\boldsymbol{t}_{b,k}$表示分支b中的第k个文本原型；mean(·)表示均值聚合操作；$\boldsymbol{q}_b \in \boldsymbol{R}^D$表示由该分支全部文本原型聚合得到的文本语义锚点。

对于视觉端，设风格解偏后得到的视觉token序列为

$$\tilde{P} = \left[\boldsymbol{p}_{\mathrm{cls}}, \boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_N\right] \in \mathbf{R}^{B \times (1+N) \times D} \#(11)$$

式中，$\boldsymbol{p}_{\mathrm{cls}}$为全局CLS token，$\boldsymbol{p}_n$为第$n$个patch token。局部语义聚焦仅作用于patch token，而CLS token保持不变。对分支b而言，首先对patch token和语义锚点进行L_2归一化：

$$\bar{\boldsymbol{p}}_{b,n} = \frac{\boldsymbol{p}_n}{\left\| \boldsymbol{p}_n \right\|_2}, \ \bar{\boldsymbol{q}}_b = \frac{\boldsymbol{q}_b}{\left\| \boldsymbol{q}_b \right\|_2} \#(12)$$

式中，$\bar{p}_{b,n}$和$\bar{q}_b$分别表示经过二范数归一化后的Patch Token和文本语义锚点，$/\!/ \cdot /\!/_2$表示向量的二范数。然后计算第n个patch token与语义锚点之间的相关性分数：

$$s_{b,n} = \frac{\bar{\boldsymbol{p}}_{b,n} \bar{\boldsymbol{q}}_b}{\tau_b} \#(13)$$

式中，$s_{b,n}$表示分支b的第n个Patch Token与文本语义锚点之间的相关性分数，τ_b表示分支b的温度系数，用于调节相关性分数的分布尺度。对所有patch的相关性分数施加softmax，得到初始注意力权重：

$$\alpha_{b,n} = \frac{\exp\left(s_{b,n}\right)}{\sum_{m=1}^{N} \exp\left(s_{b,m}\right)} \#(14)$$

式中，$\alpha_{b,n}$表示分支b中第n个Patch Token对应的初始注意力权重，m表示求和过程中的Patch Token索引，exp(·)表示自然指数函数。为进一步突出关键局部证据，本文采用Top-K稀疏筛选策略。设TopK(·)表示取得分最高的K个位置，则稀疏掩码定义为

$$m_{b,n} = \mathrm{I}\left(n | \mathrm{TopK}\left(\left\{s_{b,m}\right\}_{m=1}^{N}\right)\right) \#(15)$$

式中，$\mathbb{I}(\cdot)$为示性函数。将稀疏掩码作用于初始权重，并重新归一化，得到最终的局部语义权重：

$$\hat{\alpha}_{b,n} = \frac{\alpha_{b,n} m_{b,n}}{\sum_{m=1}^{N} \alpha_{b,m} m_{b,m} + \varepsilon} \#(16)$$

最后利用该权重对patch token进行重加权：

$$\hat{\boldsymbol{p}}_{b,n} = \hat{\alpha}_{b,n} \boldsymbol{p}_n \#(17)$$

式中，$\hat{\boldsymbol{p}}_{b,n}$表示利用最终局部语义权重对第$n$个Patch Token进行重加权后得到的视觉Token。于是，第b个分支对应的加权视觉token序列可写为

$$\hat{\boldsymbol{P}}_b = \left[\boldsymbol{p}_{\mathrm{cls}}, \hat{\boldsymbol{p}}_{b,1}, \hat{\boldsymbol{p}}_{b,2}, \cdots, \hat{\boldsymbol{p}}_{b,N}\right] \#(18)$$

由式(18)可见，该模块输出的并不是单一向量，而是面向不同分支生成的加权视觉token序列。通过这种方式，组合分支、属性分支和物体分支能够分别聚焦于与自身语义更相关的局部视觉区域，从而为后续关系建模和组合匹配提供更具判别性的局部证据。

1.4 关系建模与组合匹配预测

在获得风格解偏后的视觉表示和经文本锚引导聚焦后的分支级局部视觉token后，还需要进一步建立视觉与文本之间的关系，并完成状态、物体和组合的联合匹配预测。为此，本文采用三分支关系建模策略，分别对属性分支、物体分支和组合分支进行跨模态交互，并通过联合监督实现最终组合识别。

对于分支$b \in \{c, a, o\}$，首先对加权视觉token序列$\hat{\boldsymbol{P}}_b$进行池化，得到对应的分支视觉表示。为兼顾局部证据与整体语义，本文采用CLS token与加权patch token平均表示相结合的方式构造分支特征：

$$\boldsymbol{v}_b = \boldsymbol{p}_{\mathrm{cls}} + \frac{1}{N}\sum_{n=1}^{N} \hat{\boldsymbol{p}}_{b,n} \#(19)$$

得到分支视觉表示后，再通过跨模态注意力将视觉特征与对应分支的文本原型进行关系建模。设$\boldsymbol{v}_b$为查询，$\boldsymbol{T}_b$为键和值，则分支级跨模态关系表

示可写为

$$\boldsymbol{r}_b = \mathrm{CrossAttn}\left(\boldsymbol{v}_b | \boldsymbol{T}_b\right) \#(20)$$

式中，CrossAttn(·)表示跨模态注意力操作。进一步采用残差融合得到最终的分支判别表示：

$$\boldsymbol{g}_b = \boldsymbol{v}_b + \boldsymbol{r}_b \#(21)$$

在此基础上，分别计算属性、物体和组合分支与对应文本原型之间的匹配分数：

$$\boldsymbol{s}_a = \frac{\boldsymbol{g}_a \boldsymbol{T}_a^{\top}}{\tau_a}, \boldsymbol{s}_o = \frac{\boldsymbol{g}_o \boldsymbol{T}_o^{\top}}{\tau_o}, \boldsymbol{s}_c = \frac{\boldsymbol{g}_c \boldsymbol{T}_c^{\top}}{\tau_c} \#(22)$$

式中，τ_a、τ_o、τ_c为各分支对应的温度参数。

由于组合零样本学习的最终目标是识别状态—物体对，因此推理阶段需要将三类分支信息进行联合建模。设候选组合$(a_i, o_j) \in \mathcal{C}$，并记其在组合分支中的对应索引为$\phi(i,j)$，则该候选组合的最终匹配得分定义为

$$S\left(a_i | o_j\right) = \lambda_c s_c^{\phi(i,j)} + \lambda_a s_a^i + \lambda_o s_o^j \#(23)$$

式中，$s_c^{\phi(i,j)}$表示组合分支对候选组合$\left(a_i | o_j\right)$的预测分数，s_a^i和s_o^j分别表示属性分支和物体分支对对应状态与物体的预测分数，λ_c、λ_a、λ_o为三分支融合权重。

最终预测结果由最大匹配得分确定：

$$(\hat{a}, \hat{o}) = \arg \max_{(a_i, o_j) \in \mathcal{C}^*} S\left(a_i, o_j\right) \#(24)$$

式中，$\mathcal{C}^*$在闭世界设定下取测试候选组合集合，在开放世界设定下取全部可行组合集合。

训练阶段，为了同时约束属性、物体和组合三条语义路径，本文采用三分支加权交叉熵联合优化目标：

$$L = \lambda_a \mathcal{L}_a + \lambda_o \mathcal{L}_o + \lambda_c \mathcal{L}_c \#(25)$$

式中，

$$\mathcal{L}_a = -\sum_{i=1}^{K_a} y_a^{(i)} \log p_a^{(i)} \#(26)$$

$$\mathcal{L}_o = -\sum_{j=1}^{K_o} y_o^{(j)} \log p_o^{(j)} \#(27)$$

$$\mathcal{L}_c = -\sum_{k=1}^{K_c} y_c^{(k)} \log p_c^{(k)} \#(28)$$

式中，$y_a^{(i)}$、$y_o^{(j)}$、$y_c^{(k)}$分别表示属性、物体和组合分支的监督标签，$p_a^{(i)}$、$p_o^{(j)}$、$p_c^{(k)}$为各分支经 softmax 归一化后的预测概率。

通过式(25)的联合优化，属性分支、物体分支和组合分支能够在训练过程中形成互补约束：属性分支强化状态判别能力，物体分支增强对象识别稳定性，组合分支则进一步建模状态与物体之间的联合关系。结合前文提出的风格解偏与局部语义聚焦机制，模型能够在保持全局组合建模能力的同时，更有效地抑制风格偏移并突出局部关键证据，从而提升未见状态—物体组合的识别性能。

2 实验结果与分析

2.1 数据集

为验证所提方法在组合零样本学习任务中的有效性，本文在 MIT-States、UT-Zappos 和 C-GQA 3 个基准数据集上开展实验。3 个数据集分别对应通用属性—物体组合、细粒度鞋类属性识别以及大规模复杂场景组合，能够从不同角度评估模型对未见组合的泛化能力。本文遵循现有主流工作的标准划分，并在闭世界与开放世界两种设置下进行统一评测。其中，MIT-States 包含约 53 753 幅图像、115 个状态和 245 个物体，具有组合空间大、背景复杂、风格多样等特点，更适合检验模型在通用场景中的组合泛化能力。UT-Zappos 包含 50 025 幅鞋类图像、16 个状态和 12 个物体，类别空间较小但细粒度差异显著，适合评估模型对局部属性细节的辨别能力。C-GQA 建立在 GQA 场景图基础上，包含 39 298 幅图像、453 个状态和 870 个物体，具有更大的语义空间和更复杂的组合关系，可用于检验模型在大规模组合识别任务中的可扩展性与鲁棒性。

2.2 实验设置

为验证方法有效性，所有实验均在统一环境下进行，采用双 NVIDIA RTX 4090 显卡，并基于 PyTorch 框架实现。图像编码器与文本编码器均采用预训练的 CLIP ViT-L/14，训练过程中冻结主干参数。视觉侧仅对插入 Transformer 各层注意力子层和前馈子层后的 Adapter 进行轻量调优，同时保留属性分支和物体分支对应的 Disentangler 模块。Adapter 的瓶颈维度设置为 64，dropout 设置为 0.1；文本提示采用短语“a photo of”对应的词向量初始化组合、属性和物体三条分支的上下文表示。

风格解偏模块作用于 CLIP 视觉编码器卷积 Patch Embedding 后的二维特征，并仅在训练阶段启用。对于每个训练批次，以概率$p = 0.5$执行风格统计混合，混合系数从对称 Beta 分布 Beta(α, α)中采样，其中α设置为 0.1。局部语义聚焦模块从视觉编

码器输出的256个Patch Token中，选取与分支文本语义锚点相关性最高的$K=64$个Token，即保留25%的局部视觉区域。组合、属性和物体三条分支在计算文本锚点与视觉Token的相似度时，温度系数分别设置为$\tau_c=\tau_a=\tau_o=0.07$。上述参数均依据验证集AUC确定，测试集不参与超参数选择。

训练阶段采用组合、属性和物体三条分支的交叉熵损失进行联合优化，三个分支的损失权重均设置为1.0，即总损失为三项分支损失之和。推理阶段对组合、属性和物体三条分支的预测得分进行等权融合，三个分支的融合权重同样均设置为1.0。模型采用Adam优化器，初始学习率设置为2×10^{-4}，并每隔5轮将学习率衰减为原来的0.5。开放世界设定下，进一步采用与基线方法一致的可行性校准策略，以过滤语义上不合理的属性—物体组合。主实验根据验证集AUC选择最佳训练轮次，并报告多次随机种子运行的平均结果。

2.3 评估指标

为全面评估模型在组合零样本学习任务中的性能，本文遵循现有工作的通用设置，采用Seen(S)、Unseen(U)、Harmonic Mean(H)和Area Under Curve(AUC)4项指标进行评价。其中，S表示模型在已见组合上的识别准确率，U表示模型在未见组合上的识别准确率，分别反映模型对训练分布内样本和新组合样本的识别能力。

由于组合零样本学习中普遍存在已见组合偏置，仅单独报告S或U难以客观反映模型的综合性能，因此进一步采用调和均值H衡量模型在已见组合与未见组合之间的平衡能力，其定义为

$$H=\frac{2SU}{S+U}\#(29)$$

此外，本文采用AUC指标对Seen-Unseen曲线下的面积进行度量，以更全面地反映模型在不同偏置强度下的整体表现。AUC越大，说明模型在不同决策阈值下均能保持较好的已见—未见平衡性能。

本文分别在闭世界和开放世界两种设置下报告上述指标。前者仅在预定义候选组合空间内进行预测，后者则需在更大组合空间中完成识别，因而更具挑战性。通过同时报告两种设置下的S、U、H和AUC，能够较为全面地评估所提方法的组合识别能力与泛化性能。

2.4 实验表现

为全面验证所提方法的有效性，本文在闭世界与开放世界两种评估设定下，分别在MIT-States、UT-Zappos和C-GQA这3个基准数据集上与当前主流方法进行对比，实验结果如表1和表2所示。整体来看，本文方法在3个数据集上的各项指标均取得了优于基线模型和现有主流方法的结果，尤其在H和AUC两项更能反映泛化能力与已见—未见平衡性的指标上表现稳定，说明所提出的风格解偏与局部语义聚焦机制能够有效提升模型对未见状态—物体组合的识别能力。

在闭世界设定下，本文方法在MIT-States、UT-Zappos和C-GQA上的S、U、H和AUC分别达到50.9%、53.7%、39.8%和22.9%，67.6%、74.4%、55.3%和42.8%，41.7%、36.4%、30.2%和13.1%。可以看出，本文方法在通用场景、细粒度属性识别以及大规模复杂组合空间中均保持了较好的优势，验证了所提方法在不同难度任务下的有效性。

在开放世界设定下，各方法性能整体有所下降，但本文方法仍保持领先表现。具体而言，在MIT-States上，本文方法的S、U、H和AUC分别达到49.8%、20.3%、20.9%和7.8%；在UT-Zappos上分别达到67.4%、62.1%、48.4%和33.8%；在C-GQA上分别达到41.6%、8.3%、11.7%和3.24%。上述结果表明，在候选组合空间进一步扩大的情况下，本文方法依然能够保持较强的未见组合识别能力和更好的已见—未见平衡性能。

2.5 消融实验

为验证本文所提出的风格解偏模块和局部语义聚焦模块对模型性能的实际贡献，本文在MIT-States数据集的闭世界设定下进行模块消融实验。以主流三分支框架作为基线模型，在保持训练策略、网络主干和评估协议一致的条件下，分别加入风格解偏模块和局部语义聚焦模块，并与完整模型进行对比。实验结果如表3所示。

从表3可以看出，单独引入风格解偏模块后，模型的S、U、H和AUC分别由基线模型的49.0%、53.0%、39.3%和22.1%提升至49.6%、53.2%、39.4%和22.4%。这说明风格解偏模块能够通过扰动和重组视觉特征统计量，降低模型对训练样本中特定颜色、纹理和背景风格的依赖，从而提升视觉表征的鲁棒性。

表1　MIT-States、UT-Zappos和C-GQA 3个数据集上的闭世界(closed-world)实验结果

Table 1　Closed-world results on MIT-States, UT-Zappos, and C-GQA

方法	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP（Nagarajan 和 Grauman，2018）	14.3	17.4	9.9	1.6	59.8	54.2	40.8	25.9	17.0	5.6	5.9	0.7
LE+（Naeem 等，2021）	15.0	20.1	10.7	2.0	53.0	61.9	41.0	25.7	18.1	5.6	6.1	0.8
TMN（Purushwalkam 等，2019）	20.2	20.1	13.0	2.9	58.7	60.0	45.0	29.3	23.1	6.5	7.5	1.1
SymNet（Li等，2020）	24.2	25.2	16.1	3.0	49.8	57.4	40.4	23.4	26.8	10.3	11.0	2.1
CompCos（Mancini 等，2021a）	25.3	24.6	16.4	4.5	59.8	62.5	43.1	28.1	28.1	11.2	12.4	2.6
CGE（Naeem 等，2021）	28.7	25.3	17.2	5.1	56.8	63.6	41.2	26.4	27.5	11.7	11.9	3.6
Co-CGE（Mancini 等，2024）	31.1	5.8	6.4	1.1	62.0	44.3	40.3	23.1	32.1	2.0	3.4	0.5
SCEN（Li等，2022）	29.9	25.2	18.4	5.3	63.5	63.1	47.8	32.0	28.9	25.4	17.5	5.5
CSP（Nayak 等，2023）	46.6	49.9	36.3	19.4	64.2	66.2	46.6	33.0	28.8	26.8	20.5	6.2
DFSP(i2t)（Lu等，2023）	47.4	52.4	37.2	20.7	64.2	66.4	45.1	32.1	35.6	29.3	24.3	8.7
Troika（Huang 等，2024）	49.0	53.0	39.3	22.1	66.8	73.8	54.6	41.7	41.0	35.7	29.4	12.4
GIPCOL（Xu 等，2024）	48.5	49.6	36.6	19.9	65.0	68.5	48.8	36.2	–	–	–	–
HMSF（Du等，2025）	49.1	52.9	39.1	22.1	65.0	–	–	–	–	–	–	–
本文	**50.9**	**53.7**	**39.8**	**22.9**	**67.6**	**74.4**	**55.3**	**42.8**	**41.7**	**36.4**	**30.2**	**13.1**

注：AoP（Attributes as Operators）；LE+（LabelEmbed+）；TMN（Task-Driven Modular Networks）；CompCos（Compositional Cosine Logits）；CGE（Compositional Graph Embedding）；Co-CGE（Compositional Cosine Graph Embeddings）；SCEN（Siamese Contrastive Embedding Network）；CSP（Compositional Soft Prompting）；DFSP（Decomposed Fusion with Soft Prompt，i2t表示将图像特征融合至文本特征）；GIPCOL（Graph-Injected Soft Prompting for Compositional Learning）；HMSF（Hierarchical Multi-Scale Feature Fusion）。加粗字体表示各列最优结果。“–”表示未在该数据集上进行实验。

当仅加入局部语义聚焦模块时，模型的S、U、H和AUC分别提升至50.1%、53.5%、39.6%和22.6%，整体提升幅度高于单独使用风格解偏模块。该结果表明，利用文本锚点对局部视觉token进行相似度筛选和重加权，能够有效突出与当前状态—物体组合相关的关键局部区域，减少背景和无关纹理对跨模态匹配的干扰，从而增强模型对未见组合的判别能力。

在同时引入两个模块后，本文方法取得最优结果，S、U、H和AUC分别达到50.9%、53.7%、39.8%和22.9%，相比基线模型分别提升1.9、0.7、0.5和0.8个百分点。上述结果说明，风格解偏模块和局部语义聚焦模块具有一定互补性：前者主要增强视觉特征的跨风格稳定性，后者进一步强化局部语义证据的选择性表达。二者协同作用能够有效提升模型在组合零样本学习任务中的识别性能与泛化能力。

2.6　超参数敏感性分析

为分析关键超参数对模型性能的影响，本文在MIT-States数据集的闭世界设置下，分别对局部语义

表2　MIT-States、UT-Zappos和C-GQA 3个数据集上的开放世界(open-world)实验结果

Table 2　Open-world results on MIT-States, UT-Zappos, and C-GQA

方法	MIT-States				UT-Zappos				C-GQA			
	S	U	H	AUC	S	U	H	AUC	S	U	H	AUC
AoP (Nagarajan 和 Grauman, 2018)	16.6	5.7	4.7	0.7	50.9	34.2	29.4	13.7	–	–	–	–
LE+(Naeem等, 2021)	14.2	2.5	2.7	0.3	60.4	36.5	30.5	16.3	19.2	0.7	1.0	0.08
TMN (Purushwalkam等, 2019)	12.6	0.9	1.2	0.1	55.9	18.1	21.7	8.4	–	–	–	–
SymNet(Li等, 2020a)	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
CompCos(Mancini等, 2021a)	25.4	10.0	8.9	1.6	59.3	46.8	36.9	21.3	–	–	–	–
CGE(Naeem等, 2021)	32.4	5.1	6.0	1.0	61.7	47.7	39.0	23.1	32.7	1.8	2.9	0.47
Co-CGE^Closed (Mancini等, 2024)	31.1	5.8	6.4	1.1	62.0	44.3	40.3	23.1	32.1	2.0	3.4	0.53
KG-SP (Karthik 等, 2022)	28.4	7.5	7.4	1.3	61.8	52.1	42.3	26.5	31.5	2.9	4.7	0.78
CSP(Nayak等, 2023)	46.3	15.7	17.4	5.7	64.1	44.1	38.9	22.7	28.7	5.2	6.9	1.20
DFSP (i2t)(Lu 等, 2023)	47.2	18.2	19.1	6.7	64.3	53.8	41.2	26.4	35.6	6.5	9.0	1.95
Troika (Huang 等, 2024)	48.8	18.7	20.1	7.2	66.4	61.2	47.8	33.0	40.8	7.9	10.9	2.70
GIPCOL(Xu等, 2024)	48.5	16.0	17.9	6.3	65.0	45.0	40.1	23.5	–	–	–	–
HMSF(Du等, 2025)	49.2	18.3	19.7	7.1	–	–	–	–	–	–	–	–
本文	**49.8**	**20.3**	**20.9**	**7.8**	**67.4**	**62.1**	**48.4**	**33.8**	**41.6**	**8.3**	**11.7**	**3.24**

注：KG-SP(Knowledge Guided Simple Primitives)；其余方法缩写同表1。加粗字体表示各列最优结果。“–”表示未在该数据集上进行实验。

表3　模块消融实验结果

Table 3　Ablation results

方法	风格解偏	局部语义聚焦	S	U	H	AUC
基线模型	×	×	49.0	53.0	39.3	22.1
基线+风格解偏	√	×	49.6	53.2	39.4	22.4
基线+局部语义聚焦	×	√	50.1	53.5	39.6	22.6
本文方法	√	√	**50.9**	**53.7**	**39.8**	**22.9**

注：“√”表示使用相应模块，“×”表示未使用相应模块。

聚焦模块中的Top-K参数K和风格解偏模块中的Beta分布参数α进行敏感性实验。实验采用控制变量法，每次仅改变一个待分析参数，其余训练配置均与主实验保持一致。考虑到H和AUC能够综合反映模型在已见组合与未见组合之间的平衡能力，本文主要根据这两项指标分析参数变化的影响，同时报告S和U作为参考。所有结果均由3次不同随机种子实验的平均值获得。

2.6.1　Top-K参数分析

局部语义聚焦模块通过Top-K策略保留与分支文本语义锚点相关性较高的视觉Patch Token。为分析局部区域保留数量对模型性能的影响，本文将K

分别设置为32、64、128、192和256。由于CLIP ViT-L/14在224 × 224输入分辨率下共产生256个Patch Token，上述设置分别对应保留12.5%、25%、50%、75%和100%的局部视觉区域。其中，K = 256表示保留全部Patch Token，即不进行稀疏筛选。实验过程中，Beta分布参数α固定为0.1，其余配置保持不变，结果如表4所示。

表4　不同Top-K参数下的实验结果

Table 4　Results with different Top-K values

K	保留比例	S	U	H	AUC
32	12.5	49.9	53.2	39.4	22.5
64	**25.0**	**50.9**	**53.7**	**39.8**	**22.9**
128	50.0	50.6	53.6	39.7	22.8
192	75.0	50.3	53.4	39.6	22.7
256	100.0	50.1	53.3	39.5	22.6

由表4可见，随着K值增加，模型性能整体呈现先提升后下降的变化趋势。当K值较小时，局部语义聚焦模块仅保留少量Patch Token，虽然能够有效排除背景区域，但也可能遗漏与属性或物体语义相关的有效视觉证据。当K值过大时，更多背景区域和非判别性纹理重新参与跨模态关系建模，削弱了局部语义筛选的作用。K设置为64时，模型取得最高的H和AUC，分别达到39.8%和22.9%，表明保留适量的局部视觉区域能够在关键证据保留与无关信息抑制之间取得较好的平衡。相比K = 256的非稀疏设置，K = 64时AUC提高了0.3个百分点，进一步说明Top-K稀疏筛选对局部判别信息提取具有积极作用。因此，本文在其余实验中将K设置为64。

2.6.2　Beta分布参数分析

风格解偏模块通过从对称Beta分布Beta(α, α)中采样混合系数，对不同样本的通道均值和标准差进行组合。参数α决定混合系数的分布形态及风格扰动程度。为分析其对模型性能的影响，本文将α分别设置为0.1、0.3、0.5、0.7和1.0，并将Top-K参数固定为K = 64，实验结果如表5所示。

由表5可见，模型对Beta分布参数具有一定的稳定性，在不同α设置下均能够取得优于基线模型的结果。其中，当α设置为0.1时，模型取得最优的H和AUC。较小的α使采样得到的混合系数更容易接近0或1，从而产生更加多样化的风格统计组合，

表5　不同Beta分布参数下的实验结果

Table 5　Results with different Beta distribution parameters

α	S	U	H	AUC
0.1	**50.9**	**53.7**	**39.8**	**22.9**
0.3	50.7	53.6	39.7	22.8
0.5	50.4	53.4	39.6	22.7
0.7	50.6	53.2	39.6	22.7
1.0	50.2	53.1	39.5	22.6

有助于降低模型对固定训练风格的依赖。随着α增大，混合系数逐渐趋向中间区域，不同样本的通道统计量被更加平均地融合。当混合程度过于集中时，可能会削弱不同视觉风格之间的差异，或者对部分具有判别作用的颜色和纹理信息造成干扰，因而模型性能出现小幅下降。

总体而言，Top-K参数决定局部视觉证据的保留范围，Beta分布参数则控制特征统计扰动的方式。实验结果表明，模型在合理参数范围内具有较好的稳定性，并且K = 64、α = 0.1时能够在局部证据筛选、风格偏置抑制和属性语义保持之间取得较好的平衡。因此，本文在主实验中采用上述参数设置。

2.7　局部语义聚焦可视化

为直观分析局部语义聚焦模块对视觉区域的选择效果，本文选取MIT-States数据集中的部分测试样本，对基线模型与本文方法产生的Patch级语义响应进行可视化。为保证比较的一致性，基线模型和本文方法均采用相同的分支文本语义锚点，与视觉编码器输出的Patch Token计算归一化余弦相似度。基线模型保留所有Patch对应的稠密响应权重，本文方法则展示经过Top-K筛选和重新归一化后实际用于视觉Token重加权的局部语义权重。具体地，将256个Patch权重重新排列为16×16的二维响应图，再通过双线性插值恢复至原始图像尺寸，并叠加于输入图像上。对于同一样本，不同方法和分支采用统一的响应范围和颜色映射，以保证可视化结果具有可比性。

图2展示了基线模型以及本文属性、物体和组合分支的局部响应结果。对于“湿透的狗”，基线模型的响应较为分散，部分注意区域落在背景中，而本文属性分支主要关注狗的毛发表面及其潮湿纹理，物体分支更多覆盖狗的主体轮廓，组合分支则同时

保留毛发状态和物体结构信息。对于“切片番茄”，属性分支的高响应区域主要集中在番茄切面及其边缘结构，物体分支则更关注番茄整体区域。类似地，在“生锈的汽车”和“破损的椅子”等样本中，属性分支分别对锈蚀表面和结构破损区域产生较强响应，而组合分支能够兼顾局部状态线索及其对应的物体语境。

上述结果表明，基线模型在缺少显式局部筛选时容易受到背景区域和非判别性纹理的干扰。本文方法利用分支文本语义对视觉Patch进行相关性评估和稀疏筛选，使属性、物体和组合分支能够形成具有差异性的局部响应。其中，属性分支更倾向于保留状态相关的细粒度视觉证据，物体分支侧重目标主体及整体结构，组合分支则综合建模属性表现与物体语境。该结果从可视化角度验证了局部语义聚焦模块在抑制无关区域和突出关键判别证据方面的有效性。

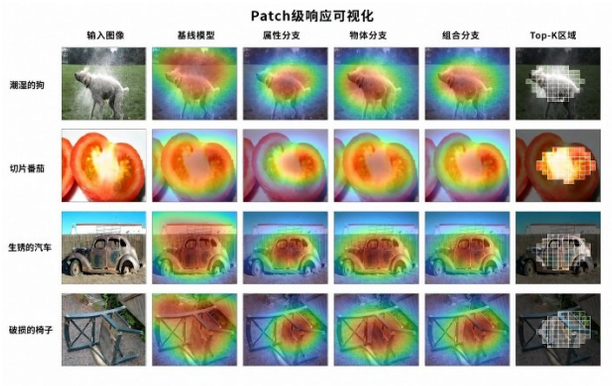

图2　基线模型与本文局部语义聚焦模块的Patch级响应可视化

Fig. 2　Patch-level response visualization of the baseline model and the proposed local semantic focusing module

2.8　模型复杂度与推理效率

为评估所提方法引入的额外计算开销，本文对基线模型及不同模块配置下的参数量、浮点运算量和推理速度进行比较。所有效率实验均在单张NVIDIA RTX 4090显卡上进行，输入图像分辨率统一设置为224×224，batch size设置为1，并采用与主实验一致的推理精度。测试前首先进行100次预热，随后连续推理1000次并取平均值。每次计时前后均执行CUDA同步操作，测试时间不包括图像读取和数据预处理过程。由于不同候选组合的文本特征可以在推理前预先计算并缓存，因此推理速度统计主要包括图像编码、局部语义聚焦、跨模态关系建模和最终得分计算过程。实验结果如表6所示。

表6　不同模型配置的复杂度与推理效率比较

Table 6　Comparison of model complexity and inference efficiency

方法	总参数量/M	可训练参数量/M	FLOPs/G	推理时间/ms	FPS
Troika基线	427.3	12.8	81.4	22.6	44.2
基线+风格解偏	427.3	12.8	81.4	22.6	44.2
基线+局部语义聚焦	427.3	12.8	81.7	23.3	42.9
本文方法	427.3	12.8	81.7	23.4	42.7

注：M表示百万，G表示十亿次浮点运算，ms表示毫秒，FPS表示每秒处理的图像帧数。

由表6可见，风格解偏模块仅对训练阶段的中间特征统计量进行随机混合，不包含额外可学习参数，并在验证和测试阶段关闭。因此，加入该模块后，模型的参数量、FLOPs和推理时间均与基线模型保持一致。该结果表明，风格解偏能够在不增加部署阶段计算负担的情况下，提高视觉表示对训练风格变化的鲁棒性。

局部语义聚焦模块主要包含文本锚点与视觉Patch Token之间的相似度计算、Top-K筛选及Token重加权操作，不引入新的可学习参数。与基线模型相比，加入该模块后FLOPs由81.4 G增加至81.7 G，单幅图像的平均推理时间由22.6 ms增加至23.3 ms，增幅分别约为0.37%和3.10%。完整模型的平均推理时间为23.4 ms，对应推理速度为42.7 FPS。由于整个模型的计算开销主要来自CLIP ViT-L/14视觉编码器和三分支跨模态关系建模，局部Patch筛选所增加的计算量相对有限。

综合性能与效率可以看出，本文方法在不增加可学习参数的情况下，仅引入少量Patch相似度计算与排序开销，便能够提升未见组合识别性能及已见

—未见组合之间的平衡能力。因此，所提风格解偏和局部语义聚焦机制在模型性能与推理效率之间取得了较好的折中，具有一定的实际应用价值。

2.9 成功与失败案例分析

为进一步直观分析本文方法在未见组合识别中的预测效果，本文选取 MIT-States 数据集中的部分成功预测与失败预测样例进行定性可视化，结果如图3所示。其中，上排为成功预测样例，下排为失败预测样例。每个样例下方分别给出真实标签和模型预测标签。

从成功案例可以看出，当图像中的状态线索较为明显且目标区域较完整时，本文方法能够较准确地识别状态—物体组合。例如，对于“湿透的狗”，模型能够捕获狗毛表面的潮湿纹理；对于“切片番茄”，模型能够关注切面区域的颜色和形态特征；对于“生锈的汽车”和“破损的椅子”，模型也能够根据局部表面纹理和结构变化给出正确预测。这说明本文提出的文本锚引导局部语义聚焦机制能够帮助模型突出与当前组合语义相关的关键视觉区域，从而提升状态与物体组合的判别能力。

同时，失败案例也反映出本文方法仍存在一定局限。当不同状态之间具有相似视觉表现、局部状态线索不够明显或背景干扰较强时，模型仍可能产生混淆。例如，“老旧的椅子”被误判为“破损的椅子”，“干燥的狗”被误判为“湿透的狗”等，说明对于语义抽象或视觉线索不明显的状态，模型仍难以稳定判断。

综上，定性结果表明，本文方法在状态线索清晰、局部判别区域明显的样本上具有较好的识别能力，但在细粒度状态相近、物体外观相似或状态语义较抽象的场景下仍存在一定误判。

3 结 论

本文针对组合零样本学习中视觉风格偏置与局部判别证据不足的问题，提出一种融合风格解偏与局部语义聚焦的三分支视觉—语言方法。该方法通过风格解偏增强视觉表征鲁棒性，并利用文本锚点对局部视觉 token 进行筛选与重加权，从而提升模型对未见状态—物体组合的识别能力。在 MIT-States、UT-Zappos 和 C-GQA 3 个基准数据集上的实验结果表明，本文方法在闭世界和开放世界设定下

our method on the state-object compositional recognition task

图3 本方法在状态—物体组合识别任务中的部分成功预测与失败预测示例

Fig. 3 Representative success and failure predictions of

均取得了稳定提升，尤其在 H 和 AUC 等反映已见—未见平衡能力的指标上表现较好。消融实验进一步验证了风格解偏模块和局部语义聚焦模块的有效性，说明前者能够缓解视觉统计偏移带来的干扰，后者能够增强模型对局部关键语义证据的关注能力。定性可视化结果也表明，本文方法在状态线索清晰、目标区域较完整的样本上具有较好的组合识别效果。

尽管如此，本文方法仍存在一定局限。当状态语义较为抽象、不同状态之间视觉差异细微，或关键区域被遮挡时，模型仍可能发生误判。此外，当前局部语义聚焦主要依赖文本锚点与 patch token 的相似性，在多目标复杂场景中的精确定位能力仍有提升空间。未来将进一步研究更细粒度的区域级语义对齐机制，并结合外部知识增强组合可行性建模，以提升模型在复杂开放场景下的泛化能力与应用价值。

参考文献(References)

Du W L, Bao X L, Zhao W, Xu X F, Lu X Y and Zhang R H. 2025. A novel compositional zero-shot learning approach based on hierarchical multi-scale feature fusion. Engineering Applications of Artificial Intelligence, 159: 111633 [DOI: 10.1016/j.engappai.2025.111633]

Huang S T, Gong B, Feng Y T, Zhang M, Lv Y L and Wang D L. 2024. Troika: multi-path cross-modal traction for compositional zero-shot learning//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 24005-24014 [DOI: 10.1109/CVPR52733.2024.02266]

Liu J, Tao C B, Shen Z W, Luo X Z and Cao F. 2026. Cross-modal com-

positional zero-shot recognition via disentanglement and soft prompt. Journal of Image and Graphics, 31(3): 0180-0191 (刘杰, 陶重犇, 沈忠伟, 罗喜召, 曹峰. 2026. 显式语义解耦与软提示驱动的组合零样本识别. 中国图象图形学报, 31(3): 0180-0191) [DOI: 10.11834/jig.250265]

Wang S C, Huang Q, Zhang Y F, Li X, Nie Y Q and Luo G C. 2022. Review of action recognition based on multimodal data. Journal of Image and Graphics, 27(11): 3139-3159 (王帅琛, 黄倩, 张云飞, 李兴, 聂云清, 雒国萃. 2022. 多模态数据的行为识别综述. 中国图象图形学报, 27(11): 3139-3159) [DOI: 10.11834/jig.210786]

Hudson D A and Manning C D. 2019. GQA: a new dataset for real-world visual reasoning and compositional question answering//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 6700-6709 [DOI: 10.1109/CVPR.2019.00686]

Isola P, Lim J J and Adelson E H. 2015. Discovering states and transformations in image collections//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 1383-1391 [DOI: 10.1109/CVPR.2015.7298744]

Jayasekara H, Pham K, Saini N and Shrivastava A. 2025. Unified framework for open-world compositional zero-shot learning//Proceedings of 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE [DOI: 10.1109/WACV61041.2025.00761]

Karthik S, Mancini M and Akata Z. 2022. KG-SP: knowledge guided simple primitives for open world compositional zero-shot learning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 9336-9345 [DOI: 10.1109/CVPR52688.2022.00912]

Li X Y, Yang X, Wei K, Deng C and Yang M L. 2022. Siamese contrastive embedding network for compositional zero-shot learning//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE: 9326-9335 [DOI: 10.1109/CVPR52688.2022.00911]

Li Y L, Xu Y, Mao X H and Lu C W. 2020. Symmetry and group in attribute-object compositions//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 11316-11325 [DOI: 10.1109/CVPR42600.2020.01133]

Lu X C, Guo S, Liu Z M and Guo J C. 2023. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 23560-23569 [DOI: 10.1109/CVPR52729.2023.02256]

Mancini M, Naeem M F, Xian Y Q and Akata Z. 2021. Open world compositional zero-shot learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 5222-5230 [DOI: 10.1109/CVPR46437.2021.00518]

Mancini M, Naeem M F, Xian Y Q and Akata Z. 2024. Learning graph embeddings for open world compositional zero-shot learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(3): 1545-1560 [DOI: 10.1109/TPAMI.2022.3163667]

Misra I, Gupta A and Hebert M. 2017. From red wine to red tomato: composition with context//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 1160-1169 [DOI: 10.1109/CVPR.2017.129]

Naeem M F, Xian Y Q, Tombari F and Akata Z. 2021. Learning graph embeddings for compositional zero-shot learning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 953-962 [DOI: 10.1109/CVPR46437.2021.00101]

Nagarajan T and Grauman K. 2018. Attributes as operators: factorizing unseen attribute-object compositions//Proceedings of the 15th European Conference on Computer Vision (ECCV 2018). Munich, Germany: Springer: 172-190 [DOI: 10.1007/978-3-030-01246-5_11]

Nayak N V, Yu P and Bach S H. 2023. Learning to compose soft prompts for compositional zero-shot learning//Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net

Purushwalkam S, Nickel M, Gupta A and Ranzato M. 2019. Task-driven modular networks for zero-shot compositional learning//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE: 3592-3601 [DOI: 10.1109/ICCV.2019.00369]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Xu G Y, Chai J Y and Kordjamshidi P. 2024. GIPCOL: graph-injected soft prompting for compositional zero-shot learning//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE: 5762-5771 [DOI: 10.1109/WACV57701.2024.00567]

Yu A and Grauman K. 2014. Fine-grained visual comparisons with local learning//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus, USA: IEEE: 192-199 [DOI: 10.1109/CVPR.2014.32]

Zhou K Y, Yang Y, Qiao Y and Xiang T. 2021. Domain generalization with MixStyle//Proceedings of the 9th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net

作者简介

林泽辉, 男, 硕士研究生, 主要研究方向为组合零样本学习。
E-mail: 2411041008@post.usts.edu.cn

沈忠伟，通信作者，男，讲师，主要研究方向为视频分析、动作识别和深度学习。E-mail：shenzw@usts.edu.cn