CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts

CLAP：通过增强提示的对比学习实现内容与风格的分离

Yichao Cai , Yuhang Liu , Zhen Zhang , and Javen Qinfeng Shi
蔡一超，刘宇航，张震，施金峰

Australian Institute for Machine Learning, University of Adelaide, SA 5000, Australia
{yichao.cai,yuhang.liu01,zhen.zhang02,javen.shi}@adelaide.edu.au
澳大利亚阿德莱德大学机器学习研究所，南澳5000，澳大利亚 {yichao.cai,yuhang.liu01,zhen.zhang02,javen.shi}@adelaide.edu.au

Abstract. Contrastive vision-language models, such as CLIP, have garnered considerable attention for various dowmsteam tasks, mainly due to the remarkable ability of the learned features for generalization. However, the features they learned often blend content and style information, which somewhat limits their generalization capabilities under distribution shifts. To address this limitation, we adopt a causal generative perspective for multimodal data and propose contrastive learning with data augmentation to disentangle content features from the original representations. To achieve this, we begin with exploring image augmentation techniques and develop a method to seamlessly integrate them into pre-trained CLIP-like models to extract pure content features. Taking a step further, recognizing the inherent semantic richness and logical structure of text data, we explore the use of text augmentation to isolate latent content from style features. This enables CLIP-like model's encoders to concentrate on latent content information, refining the learned representations by pre-trained CLIP-like models. Our extensive experiments across diverse datasets demonstrate significant improvements in zero-shot and few-shot classification tasks, alongside enhanced robustness to various perturbations. These results underscore the effectiveness of our proposed methods in refining vision-language representations and advancing the state-of-the-art in multimodal learning. [1]
摘要。对比视觉-语言模型，如CLIP，因其学习特征的卓越泛化能力，在多种下游任务中备受关注。然而，它们学习到的特征常常混合了内容和风格信息，这在分布变化下限制了其泛化能力。为解决此限制，我们采用多模态数据的因果生成视角，提出结合数据增强的对比学习方法，以从原始表示中解耦内容特征。为此，我们首先探索图像增强技术，开发了一种方法将其无缝集成到预训练的CLIP类模型中，以提取纯净的内容特征。进一步地，鉴于文本数据固有的语义丰富性和逻辑结构，我们探索文本增强以分离潜在内容与风格特征，使CLIP类模型的编码器专注于潜在内容信息，从而优化预训练模型的表示。我们在多样数据集上的广泛实验表明，在零样本和少样本分类任务中均显著提升，同时增强了对各种扰动的鲁棒性。结果凸显了我们方法在优化视觉-语言表示和推动多模态学习前沿的有效性。[1]

Keywords: Data Augmentation - Latent Variables - Disentanglement
关键词：数据增强 - 潜变量 - 解耦

1　1 Introduction

2　1 引言

Vision-language models, exemplified by CLIP [36], have garnered substantial attention due to their exceptional generalization capabilities, achieved through the learned features, obtained by utilizing a cross-modal contrastive loss [20, 25, 36]. However, despite being pre-trained on extensive datasets, CLIP-like models fall short in disentangling latent content information and latent style information. Consequently, they are not immune to spurious correlations, i.e., style-related information is erroneously utilized to predict task-related labels. These limitations become evident in the presence of distribution shifts or adversarial attacks where spurious correlations often change across different environments. For examples, (1) a notable dependence on specific input text prompts has been reported for zero-shot capabilities [21, 47, 48]; (2) performance decline in few-shot scenarios has been observed in few-shot learning scenarios [13, 36]; and (3) susceptibility to adversarial attacks has been explored [33, 43, 45].
以CLIP [36]为代表的视觉-语言模型因其通过跨模态对比损失[20, 25, 36]学习到的特征具备卓越的泛化能力而备受关注。然而，尽管在大规模数据集上预训练，CLIP类模型在解耦潜在内容信息与潜在风格信息方面仍存在不足。因此，它们无法避免伪相关，即错误地利用与风格相关的信息来预测任务标签。这些局限在分布变化或对抗攻击中尤为明显，因为伪相关往往随环境变化而改变。例如，(1) 零样本能力显著依赖特定输入文本提示[21, 47, 48]；(2) 少样本学习场景中性能下降[13, 36]；(3) 对抗攻击的脆弱性被广泛研究[33, 43, 45]。

[1] Our code is available at https://github.com/YichaoCai1/CLAP
[1] 我们的代码可在 https://github.com/YichaoCai1/CLAP 获取

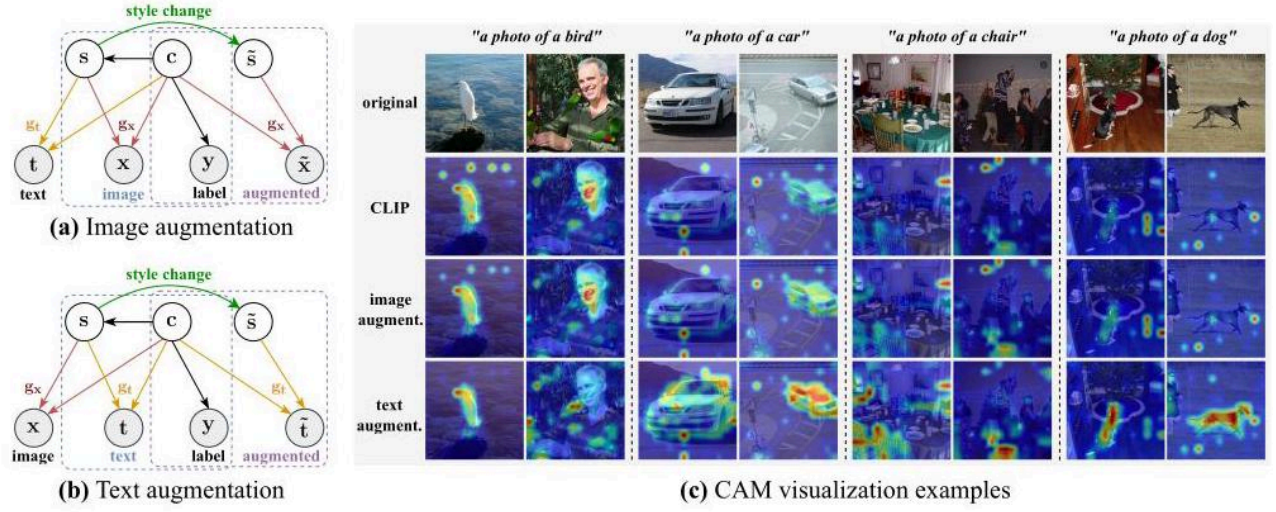


Fig. 1: Causal generative models of vision-language data. Image and text data are generated through distinct underlying deterministic processes, $\mathbf{g_x}$ for images and $\mathbf{g_t}$ for texts,derived from a unified latent space with latent content variables $\mathbf{c}$ and latent style variables $\mathbf{s}$. Latent content $\mathbf{c}$ exclusively determines the sample label $\mathbf{y}$. (a) Soft interventions on latent style variables $\mathbf{s}$ result in $\tilde{\mathbf{s}}$,subsequently generating augmented images $\tilde{\mathbf{x}}$. (b) Due to the same latent space,soft interventions on latent style variables $\mathbf{s}$ can also result in $\tilde{\mathbf{s}}$,producing augmented text $\tilde{\mathbf{t}}$. (c) A qualitative comparison of image features for zero-shot classification using "a photo of a [class]" prompts, visualized using class activation map (CAM) 32, demonstrates that while image augmentation can enhance CLIP features, the features obtained through text augmentation methods predominantly focus on the content.
图1：视觉-语言数据的因果生成模型。图像和文本数据通过不同的确定性过程生成，$\mathbf{g_x}$ 表示图像过程，$\mathbf{g_t}$ 表示文本过程，均源自统一的潜在空间，该空间包含潜在内容变量 $\mathbf{c}$ 和潜在风格变量 $\mathbf{s}$ 。潜在内容 $\mathbf{c}$ 唯一决定样本标签 $\mathbf{y}$ 。(a) 对潜在风格变量 $\mathbf{s}$ 进行软干预产生 $\tilde{\mathbf{s}}$ ，进而生成增强图像 $\tilde{\mathbf{x}}$ 。(b) 由于共享潜在空间，对潜在风格变量 $\mathbf{s}$ 的软干预同样产生 $\tilde{\mathbf{s}}$ ，生成增强文本 $\tilde{\mathbf{t}}$ 。(c) 使用“a photo of a [class]”提示进行零样本分类时，基于类激活图（CAM）32的图像特征定性比较表明，图像增强能提升CLIP特征，而文本增强方法获得的特征则主要聚焦于内容。

Taking a causal perspective, this work begin with a simple yet effective method, image augmentation, to disentangle content and style information within the learned representations of CLIP-like models. This approach is inspired by recent advancements in theoretical development in causal representation learning [41], which demonstrate that augmented image can be interpreted as a result of soft interventions on latent style variables, as depicted in Fig. 1a. Such augmentation results in a natural data pair where content information remains unchanged while style information changes. Consequently, using contrastive learning, it becomes feasible to isolate the invariant content information from the variant style information. Motivated by this theoretical advancement, we propose a practical method to incorporate image augmentation into CLIP-like models to extract content information from the original learned features. Specifically, a disentangled network is designed to fine-tune the pre-trained CLIP model by using a contrastive loss with image augmentation.
从因果视角出发，本工作以一种简单而有效的方法——图像增强，开始解开CLIP类模型学习表示中的内容与风格信息的纠缠。该方法灵感来源于因果表示学习理论发展的最新进展[41]，其表明增强图像可被解释为对潜在风格变量的软干预结果，如图1a所示。此类增强产生了一个自然的数据对，其中内容信息保持不变而风格信息发生变化。因此，利用对比学习，可以将不变的内容信息与变化的风格信息分离开来。受该理论进展的启发，我们提出了一种实用方法，将图像增强融入CLIP类模型中，以从原始学习特征中提取内容信息。具体而言，设计了一个解耦网络，通过使用带有图像增强的对比损失对预训练的CLIP模型进行微调。

Despite the advancements made in disentangling content and style information from the original features learned by CLIP-like models through image augmentation, we recognize an inherent limitation: it is generally challenging to design adequate image augmentations to ensure all style factors change in an image. Theoretically, disentangling content and style information necessitates changes in all style factors [41]. However, inducing sufficient changes in latent style through image augmentation poses challenges due to the high dimensionality and complexity of style information in image data. Achieving significant style variation via artificially designed image augmentation techniques, such as transforming a photograph of a dog into a sketch while preserving content but dramatically altering style, is notably difficult.
尽管通过图像增强在解开CLIP类模型原始特征中的内容与风格信息纠缠方面取得了进展，我们仍认识到一个固有的局限性：设计足够充分的图像增强以确保图像中所有风格因素发生变化通常具有挑战性。从理论上讲，解耦内容与风格信息需要所有风格因素发生变化[41]。然而，由于图像数据中风格信息的高维度和复杂性，通过图像增强诱导潜在风格的充分变化存在困难。通过人为设计的图像增强技术实现显著的风格变化，例如将一张狗的照片转变为素描，同时保持内容不变但风格大幅改变，尤其困难。

Taking a further step, rather than relying on image augmentation, we explore the use of text augmentation to disentangle latent content and style factors. This shift is motivated by two key observations: 1) Vision and language data share the same latent space. Therefore, text augmentation can also be utilized to induce changes in latent style factors instead of image augmentation. 2) Text data inherently possesses high semanticity and logical structure, making it more amenable to property-wise manipulation compared to image data. Consequently, implementing sufficient style changes through text augmentation is more feasible than image augmentation, contributing to isolating content from style information, see Fig. 1c for visual comparison. For instance, transforming text from "a photo of a dog" to "a sketch of a dog" is straightforward in the language modality, whereas achieving a similar transformation in image data is challenging. Inspired by these observations, we posit that introducing style variations through text augmentation, as illustrated Fig. 1b, provides a more effective approach for learning vision-language content features than relying on image augmentation.
更进一步，我们不再依赖图像增强，而是探索使用文本增强来解开潜在的内容与风格因素。这一转变基于两个关键观察：1）视觉与语言数据共享相同的潜在空间，因此文本增强也可用于诱导潜在风格因素的变化，替代图像增强。2）文本数据本质上具有高度的语义性和逻辑结构，相较于图像数据更易于逐属性操作。因此，通过文本增强实现充分的风格变化比图像增强更为可行，有助于内容与风格信息的分离，视觉对比见图1c。例如，将文本从“狗的照片”转为“狗的素描”在语言模态中较为直接，而在图像数据中实现类似转换则较为困难。受此启发，我们认为通过文本增强引入风格变化，如图1b所示，比依赖图像增强更有效地学习视觉-语言内容特征。

In summary, our contributions include: (1) Aimed at disentangling latent content and style factors to refine vision-language features of pre-trained CLIP-like models, we propose constrastive learning with data augmentation to fine tune the original features of pre-trained CLIP-like models from a causal perspective. (2) We present a novel method customized for pre-trained CLIP-like models. This method leverages a disentangled network, which is trained using contrastive learning with image augmentation, to extract latent content features from the learned features provided by image encoder of CLIP-like models. (3) We propose Contrastive Learning with Augmented Prompts (CLAP), to extract latent content features from representations of CLIP-like models. It begins by training a disentangled network using the pre-trained text encoder of CLIP-like models and text augmentation. Subsequently, the trained disentangled network is transferred to the image encoder of CLIP-like models. (4) Experiments conducted on a large real dataset demonstrate the effectiveness of the proposed image augmentation and text augmentation in terms of zero-shot and few-shot performance, as well as robustness against perturbations.
综上所述，我们的贡献包括：（1）针对解耦潜在内容与风格因素以优化预训练CLIP类模型的视觉-语言特征，我们从因果视角提出结合数据增强的对比学习方法，对预训练CLIP类模型的原始特征进行微调。（2）提出一种专为预训练CLIP类模型定制的新方法，利用一个解耦网络，通过带有图像增强的对比学习训练，从CLIP类模型的图像编码器提供的学习特征中提取潜在内容特征。（3）提出对比学习与增强提示（Contrastive Learning with Augmented Prompts，CLAP）方法，从CLIP类模型的表示中提取潜在内容特征。该方法首先利用预训练的CLIP类模型文本编码器和文本增强训练解耦网络，随后将训练好的解耦网络迁移至CLIP类模型的图像编码器。（4）在大规模真实数据集上的实验验证了所提图像增强和文本增强方法在零样本和少样本性能及对扰动的鲁棒性方面的有效性。

3 2 Related Work

4 2 相关工作

Contrastive Vision-Language Models Using a cross-modal contrastive loss, CLIP [36] revolutionarily introduced a scalable contrastive vision-language model by leveraging a large corpus of internet-sourced image-text pairs, demonstrating unprecedented zero-shot learning capabilities and exceptional generalization ability across datasets and supporting numerous downstream tasks [38]. ALIGN 20 expanded the scale of contrastive vision-language modeling, training on up to one billion image-text pairs while integrating the vision transformer's self-attention mechanism [11], which further enhanced performance. Despite their successes, CLIP-like models exhibit sensitivity to input text prompts [21, 48, leading to variable performance across different prompts. Efforts to mitigate this prompt sensitivity through prompt learning and engineering 9.14, 21.47 48 focus on customizing prompts for specific tasks but do not fundamentally enhance CLIP's representations. Furthermore, CLIP-like models are vulnerable to adversarial attacks [4, 12], with current strategies [33, 45] involving adversarial-natural image pairs to improve resilience. Our work diverges from task-specific approaches by aiming to enhance CLIP's representations from a disentanglement perspective, addressing the aforementioned issues inherent in CLIP-like models.
对比视觉-语言模型 CLIP[36]通过跨模态对比损失，利用大规模互联网图文对语料，开创性地引入了可扩展的对比视觉-语言模型，展现了前所未有的零样本学习能力和跨数据集的卓越泛化性能，支持众多下游任务[38]。ALIGN[20]扩大了对比视觉-语言建模的规模，训练了多达十亿的图文对，并融合了视觉变换器的自注意力机制[11]，进一步提升了性能。尽管取得成功，CLIP类模型对输入文本提示敏感[21,48]，导致不同提示下性能波动。为缓解提示敏感性，提示学习与工程方法[9,14,21,47,48]致力于为特定任务定制提示，但未从根本上提升CLIP的表示能力。此外，CLIP类模型易受对抗攻击[4,12]，现有策略[33,45]通过对抗-自然图像对提升其鲁棒性。我们的工作不同于任务特定方法，旨在从解耦视角提升CLIP的表示，解决CLIP类模型固有问题。

Disentangled Representation Learning Aimed at segregating intrinsic latent factors in data into distinct, controllable representations, disentangled representation learning benefits various applications [24, 40, 44]. Specifically, in classification tasks, it's shown that enhancing the model's performance and robustness against data distribution perturbations can be achieved by more effectively disentangling invariant content variables, without needing to identify all intrinsic latent variables completely [22, 26-28]. Within single modalities, studies such as [49] have illustrated that contrastive learning 7.16, 18 can potentially reverse the data generative process, aiding in the separation of representations. Furthermore, 41 suggest that image augmentation can help isolate content variables from the latent space through significant stylistic changes. [19] employs mixture techniques for data augmentation, enabling more abundant cross-modal matches. Diverging from these methods, our approach focuses on employing text augmentation to disentangle latent content variables, introducing a unique approach to learn refined vision-language representations.

解耦表示学习旨在将数据中的内在潜在因素分离为独立且可控的表示，解耦表示学习对多种应用具有益处[24, 40, 44]。具体而言，在分类任务中，研究表明通过更有效地解耦不变内容变量，可以提升模型性能及其对数据分布扰动的鲁棒性，而无需完全识别所有内在潜在变量[22, 26-28]。在单一模态中，诸如[49]的研究表明对比学习（contrastive learning）7.16, 18有可能逆转数据生成过程，有助于表示的分离。此外，文献[41]指出图像增强可以通过显著的风格变化帮助从潜在空间中分离内容变量。[19]采用混合技术进行数据增强，实现更丰富的跨模态匹配。与这些方法不同，我们的方法侧重于利用文本增强来解耦潜在内容变量，提出了一种学习精细视觉-语言表示的独特方法。

5 3 A Causal Generative Model for Multi-Modal Data

6 3 多模态数据的因果生成模型

To understand pretrained CLIP-like models, we investigate the underlying causal generative process for vision-language data. We consider the following causal generative model as depicted in Fig. 1. In the proposed model, the shared latent space ruling vision and language data is divided into two distinct sub-spaces: one corresponding to the latent content variables $\mathbf{c}$ and the other to the latent style variables $\mathbf{s}$. The latent content variables are posited to determine the object label $\mathbf{y}$,a relationship corroborated by prior studies [22,29,31]. Furthermore,to elucidate the correlation between the latent style variable $\mathbf{s}$ and the object variable $\mathbf{y}$,our model incorporates the premise that the latent content variable $\mathbf{c}$ causally influences the latent style variable $\mathbf{s}$,in concordance with the principles of causal representation learning highlighted in recent literature [10, 29, 41]. Additionally, considering the diversity between image data and text data, where information in image data is typically much more details while information in text data tends to be more logically structured nature, we posit distinct causal mechanisms for the generation processes. Our causal generative model is formulated as following structural causal models [2]:
为了理解预训练的类似CLIP模型，我们研究了视觉-语言数据的潜在因果生成过程。我们考虑图1所示的以下因果生成模型。在所提模型中，支配视觉和语言数据的共享潜在空间被划分为两个不同的子空间：一个对应潜在内容变量$\mathbf{c}$，另一个对应潜在风格变量$\mathbf{s}$。潜在内容变量被假设决定对象标签$\mathbf{y}$，这一关系已被先前研究证实[22,29,31]。此外，为了阐明潜在风格变量$\mathbf{s}$与对象变量$\mathbf{y}$之间的相关性，我们的模型引入了潜在内容变量$\mathbf{c}$因果影响潜在风格变量$\mathbf{s}$的前提，这与近期文献中强调的因果表示学习原则相符[10, 29, 41]。另外，考虑到图像数据与文本数据之间的差异，图像数据通常包含更多细节信息，而文本数据则更具逻辑结构性，我们假设生成过程存在不同的因果机制。我们的因果生成模型按照结构因果模型[2]构建如下：

$$\mathbf{s} := \mathbf{g}_{\mathbf{s}}(\mathbf{c}), \mathbf{x} := \mathbf{g}_{\mathbf{x}}(\mathbf{c}, \mathbf{s}), \mathbf{t} := \mathbf{g}_{\mathbf{t}}(\mathbf{c}, \mathbf{s}), \mathbf{y} := \mathbf{g}_{\mathbf{y}}(\mathbf{c}). \tag{1}$$

In Eq. (1),the style variable $\mathbf{s}$ is causally influenced by the content via $\mathbf{g_s}$; $\mathbf{x}$ and $\mathbf{t}$ denote visual and textual data,respectively. Both visual and textual data are causally produced by the shared latent variables $\mathbf{c}$ and $\mathbf{s}$ through distinct, reversible generative processes: $\mathbf{g_x}$ for images and $\mathbf{g_t}$ for text data,respectively. The label $\mathbf{y}$ of a sample is exclusively determined by the content variable $\mathbf{c}$ via $\mathbf{g_y}$. For simplicity,exogenous noises are implicitly assumed but not explicitly represented in the causal generative model's formulation, aligning with the common understanding that each latent variable is influenced by exogenous noise.
在公式(1)中，风格变量$\mathbf{s}$受到内容变量的因果影响，$\mathbf{g_s}$; $\mathbf{x}$和$\mathbf{t}$分别表示视觉和文本数据。视觉和文本数据均由共享潜在变量$\mathbf{c}$和$\mathbf{s}$通过不同且可逆的生成过程因果产生：$\mathbf{g_x}$用于图像，$\mathbf{g_t}$用于文本数据。样本的标签$\mathbf{y}$完全由内容变量$\mathbf{c}$通过$\mathbf{g_y}$决定。为简化起见，外生噪声被隐式假设但未在因果生成模型的表达中显式表示，这与通常理解的每个潜在变量均受外生噪声影响相符。

Recent seminal work in [41] has demonstrated that the latent content variable $\mathbf{c}$ can be identified up to block identifiability (i.e., $\mathbf{c}$ can be isolated from style variable $\mathbf{s}$),by requiring all latent style variables to change (e.g.,soft interventions on all latent style variables). This change can be achieved through image augmentation,i.e.,the augmented image $\tilde{\mathbf{x}}$ can be interpreted as a generative result of $\tilde{\mathbf{s}}$,which is produced through soft interventions on original latent style variables $\mathbf{s}$. Despite such theoretical advancement,the practical implementation of this theoretical result within CLIP-like models remains unclear. In this study, we propose a practical method to disentangle content and style information within CLIP-like models by employing image augmentation, as detailed in Sec. 4.1. Moreover, we recognize that implementing sufficient changes on all latent style variables $\mathbf{s}$ through text augmentation is more feasible than image augmentation, due to high semanticity and logical structure in text data, we delve into the use of text augmentation to separate content information from style information, as discussed in Sec. 4.2.
近期开创性工作[41]表明，通过要求所有潜在风格变量发生变化（例如对所有潜在风格变量进行软干预），潜在内容变量$\mathbf{c}$可以达到块可识别性（即$\mathbf{c}$可以从风格变量$\mathbf{s}$中分离）。这种变化可以通过图像增强实现，即增强后的图像$\tilde{\mathbf{x}}$可被解释为通过对原始潜在风格变量$\mathbf{s}$进行软干预而生成的$\tilde{\mathbf{s}}$的生成结果。尽管有此理论进展，但该理论结果在类似CLIP模型中的实际实现仍不明确。本研究提出了一种实用方法，通过图像增强在类似CLIP模型中解耦内容和风格信息，详见第4.1节。此外，我们认识到由于文本数据具有高度语义性和逻辑结构，通过文本增强对所有潜在风格变量$\mathbf{s}$实施充分变化比图像增强更为可行，因此我们深入探讨了利用文本增强分离内容信息与风格信息的方法，详见第4.2节。

7 4 Isolating Content from Style with Data Augmentation

8 4 利用数据增强分离内容与风格

In this section, we propose the employment of data augmentation to extract content information from the learned features in pre-trained CLIP-like models. Essentially, data augmentation facilitates the alteration of style factors while preserving content factors. Consequently, leveraging contrastive learning enables the segregation of content information from style information. We delve into two distinct forms of data augmentation, namely image augmentation (Sec. 4.1) and text augmentation (Sec. 4.2).
在本节中，我们提出利用数据增强从预训练的类似CLIP模型中学习的特征中提取内容信息。本质上，数据增强有助于改变风格因素，同时保

持内容因素不变。因此，利用对比学习可以实现内容信息与风格信息的分离。我们探讨了两种不同形式的数据增强，即图像增强（第4.1节）和文本增强（第4.2节）。

8.1 4.1 Isolating Content from Style with Augmented Images

8.2 4.1 利用增强图像分离内容与风格

While recent studies (von et al., 2021) have offered assurance regarding the disentanglement of content and style through contrastive learning with data augmentation, it remains unclear how these theoretical findings can be applied to the realm of vision-language models. We convert the theoretical findings into CLIP-like models in the following. The theoretical findings suggest using In-foNCE loss [34] to extract content information, as outlined below:

尽管近期研究（von et al., 2021）通过带有数据增强的对比学习对内容与风格的解耦提供了理论保证，但这些理论成果如何应用于视觉-语言模型领域仍不明确。我们将在下文中将这些理论成果转化为类似CLIP的模型。理论结果建议使用InfoNCE损失[34]来提取内容信息，具体如下：

$$\mathcal{L}\left(\mathbf{f};\{\mathbf{x}_i,\widetilde{\mathbf{x}}_i\}_{i=1}^b,\tau\right)=-\frac{1}{b}\sum_{i=1}^b\log\frac{\exp\left[\langle\mathbf{f}\left(\mathbf{x}_i\right),\mathbf{f}\left(\widetilde{\mathbf{x}}_i\right)\rangle/\tau\right]}{\sum_{j=1}^b\exp\left[\langle\mathbf{f}\left(\mathbf{x}_i\right),\mathbf{f}\left(\widetilde{\mathbf{x}}_j\right)\rangle/\tau\right]},\tag{2}$$

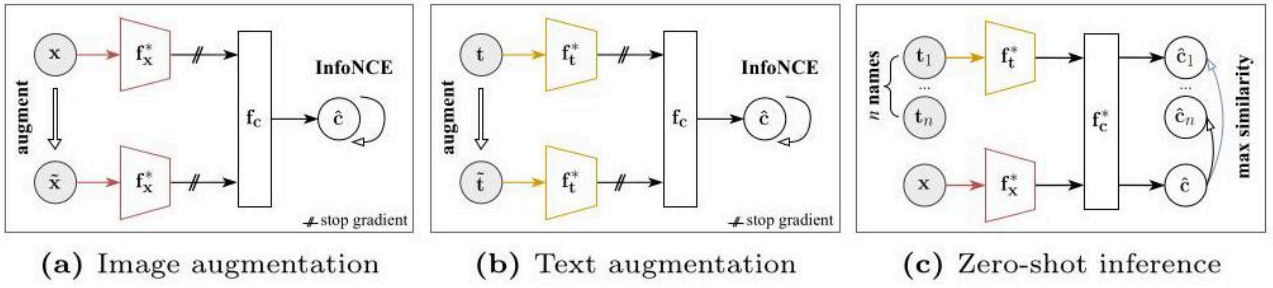


Fig. 2: Refining CLIP through data augmentation. (a) Training involves a disentangled network $\mathbf{f_c}$,utilizing contrastive loss on original and augmented image pairs $\mathbf{x}$ and $\widetilde{\mathbf{x}}$, with CLIP's image encoder $\mathbf{f}_\mathbf{x}^*$ holding frozen gradients. (b) More efficient content feature learning is achieved through contrastive learning with augmented text prompts $\mathbf{t}$ and $\widetilde{\mathbf{t}}$,using the fixed text encoder $\mathbf{f}_\mathbf{t}^*$ of CLIP. (c) Inference stage: The trained disentangled network $\mathbf{f}_\mathbf{c}^*$ integrates with CLIP's text and image encoders, $\mathbf{f}_\mathbf{t}^*$ and $\mathbf{f}_\mathbf{x}^*$,to enable zero-shot inference for an input image $\mathbf{x}$ and class names $\mathbf{t}_1$ to $\mathbf{t}_n$.

图2：通过数据增强优化CLIP。（a）训练阶段涉及一个解耦网络$\mathbf{f_c}$，利用对比损失作用于原始与增强图像对$\mathbf{x}$和$\widetilde{\mathbf{x}}$，CLIP的图像编码器$\mathbf{f}_\mathbf{x}^*$保持梯度冻结。（b）通过对增强文本提示$\mathbf{t}$和$\widetilde{\mathbf{t}}$进行对比学习，使用固定的CLIP文本编码器$\mathbf{f}_\mathbf{t}^*$，实现更高效的内容特征学习。（c）推理阶段：训练好的解耦网络$\mathbf{f}_\mathbf{c}^*$与CLIP的文本和图像编码器$\mathbf{f}_\mathbf{t}^*$和$\mathbf{f}_\mathbf{x}^*$结合，实现对输入图像$\mathbf{x}$及类别名称$\mathbf{t}_1$至$\mathbf{t}_n$的零样本推理。

where $\{\mathbf{x}_i\}_{i=1}^b$ represents a batch of b samples from the training dataset, $\mathbf{f}\left(\mathbf{x}_i\right)$ denotes sample $\mathbf{x}_i$'s features through model $\mathbf{f}$, $\widetilde{\mathbf{x}}_i$ is the augmented counterpart of $\mathbf{x}_i$,and $\langle\mathbf{z}_1,\mathbf{z}_2\rangle$ represents the cosine similarity between two feature vectors, $\mathbf{z}_1$ and $\mathbf{z}_2$,and τ represents the temperature parameter influencing the loss.

其中$\{\mathbf{x}_i\}_{i=1}^b$表示训练数据集中一批b样本，$\mathbf{f}\left(\mathbf{x}_i\right)$表示样本$\mathbf{x}_i$通过模型$\mathbf{f}$，$\widetilde{\mathbf{x}}_i$提取的特征，$\mathbf{f}$，$\widetilde{\mathbf{x}}_i$是$\mathbf{x}_i$的增强对应，$\langle\mathbf{z}_1,\mathbf{z}_2\rangle$表示两个特征向量$\mathbf{z}_1$和$\mathbf{z}_2$之间的余弦相似度，$\tau$表示影响损失的温度参数。

We extend it to refine pre-trained vision-language models, utilizing contrastive learning with augmented images (hereinafter referred to as "Im.Aug"). As illustrated in Fig. 2a, we train a disentangled network on top of CLIP's pre-trained image encoder. To enhance training efficiency and the usability of the proposed method, we freeze the pre-trained image encoder. Based on an InfoNCE loss, the learning objective of Im. Aug is formulated as follows:

我们将其扩展用于优化预训练的视觉-语言模型，利用带增强图像的对比学习（以下简称“Im.Aug”）。如图2a所示，我们在CLIP预训练图像编码器之上训练一个解耦网络。为提高训练效率及方法的实用性，我们冻结了预训练图像编码器。基于InfoNCE损失，Im.Aug的学习目标公式如下：

$$\mathbf{f}_\mathbf{c}^*=\underset{\mathbf{f_c}}{\operatorname{argmin}}\underset{\{\mathbf{x}_i\}_{i=1}^b\in\mathcal{D}_\mathbf{x}}{\mathbb{E}}\mathcal{L}\left(\mathbf{f_c}\circ\mathbf{f}_\mathbf{x}^*;\{\mathbf{x}_i,\widetilde{\mathbf{x}}_i\}_{i=1}^b,\tau\right),\tag{3}$$

where $\mathcal{D}_{\mathbf{x}}$ denotes the training image dataset and b represents the batch size, $\mathbf{f_c}$ is the disentangled network undergoing training. The pre-trained CLIP image encoder is represented by $\mathbf{f}_{\mathbf{x}}^*$,with the asterisk "*" signifying that the model weights remain fixed. The variable $\mathbf{x}_i$ refers to an image sampled from $\mathcal{D}_{\mathbf{x}}$,and $\tilde{\mathbf{x}}_i$ is its augmented view.
其中$\mathcal{D}_{\mathbf{x}}$表示训练图像数据集，b表示批量大小，$\mathbf{f_c}$是正在训练的解耦网络。预训练的CLIP图像编码器用$\mathbf{f}_{\mathbf{x}}^*$表示，星号“*”表示模型权重保持固定。变量$\mathbf{x}_i$表示从$\mathcal{D}_{\mathbf{x}}$中采样的一张图像，$\tilde{\mathbf{x}}_i$是其增强视图。

The composition of the training dataset $\mathcal{D}_{\mathbf{x}}$,the image augmentation techniques used,the structure of the disentangled network $\mathbf{f_c}$,and the utilization of $\mathbf{f}_{\mathbf{c}}^*$ post-training are detailed in the following subsections.
训练数据集$\mathcal{D}_{\mathbf{x}}$的组成、所使用的图像增强技术、解耦网络$\mathbf{f_c}$的结构以及$\mathbf{f}_{\mathbf{c}}^*$后训练的利用方法将在以下小节中详细介绍。

Data Synthesis and Image Augmentation To generate training image data, we combine class names with various image and object attributes to create text prompts for each class. Using a stable diffusion model [39], we produce synthetic images that comprise our training dataset $\mathcal{D}_{\mathbf{x}}$. The creation of template prompts for stable diffusion is based on attributes such as object size, color, image type, and art style. As detailed in Tab. 1, the attributes include 10 colors and 3 sizes for objects, and 8 types and 2 art styles for images. By assembling these attributes into prompts like "a [art style] [image type] of a [object size] [object color] [class]", we generate 480 unique texts for each class, from which one image per prompt is synthesized. Further details on image synthesis and examples are available in Appendix B.1. For the image augmentation procedures, we adopt techniques commonly used in contrastive learning practice [7, 8, 41], specifically random cropping and color distortion.
数据合成与图像增强 为生成训练图像数据，我们将类别名称与各种图像和对象属性结合，创建每个类别的文本提示。利用稳定扩散模型（stable diffusion）[39]，我们生成合成图像，构成训练数据集$\mathcal{D}_{\mathbf{x}}$。稳定扩散模板提示的创建基于对象大小、颜色、图像类型和艺术风格等属性。如表1所示，属性包括10种颜色和3种对象大小，以及8种图像类型和2种艺术风格。通过将这些属性组装成“一个[艺术风格][图像类型]的[对象大小][对象颜色][类别]”的提示，我们为每个类别生成480个独特文本，并从每个提示合成一张图像。图像合成的更多细节和示例见附录B.1。对于图像增强过程，我们采用对比学习实践中常用的技术[7, 8, 41]，具体包括随机裁剪和颜色失真。

Table 1: Template-based prompts. Attributes used to generate text prompts follow the structured format "a [art style] [image type] of a [object size] [object color] [class]", where "[class]" represents the class names.
表1：基于模板的提示。用于生成文本提示的属性遵循结构化格式“一个[艺术风格][图像类型]的[对象大小][对象颜色][类别]”，其中“[类别]”代表类别名称。

Object Color	Object Size	$\mathbf{{ImageType}}$	Art Style
yellow, green, black, blue, multicolored, orange. red, white, brown, purple	large, small, normal sized	painting, cartoon, infograph, sketch, photograph, clipart, mosaic art, sculpture	realistic, impressionistic
物体颜色	物体大小	$\mathbf{{ImageType}}$	艺术风格
黄色，绿色，黑色，蓝色，多色，橙色，红色，白色，棕色，紫色	大号，小号，正常大小	绘画，卡通，信息图，素描，摄影，剪贴画，马赛克艺术，雕塑	写实主义，印象派

Disentangled Network Structure Since the training process is based on CLIP's pre-trained lower-dimensional features, our disentangled network adopts a multi-layer perceptron (MLP) architecture. To fully benefit from the pre-trained CLIP text encoder, we construct a residual MLP featuring a zero-initialized projection, acting as the disentangled network, as depicted in Fig. 3. This design enables learning directly from the pre-trained representation space, avoiding a random starting point, inspired by ControlNet's zero-conv operation [46], which we adapt to a zero-linear operation within our residual MLP.
解耦网络结构 由于训练过程基于CLIP预训练的低维特征，我们的解耦网络采用多层感知机（MLP）架构。为了充分利用预训练的CLIP文本编码器，我们构建了一个带有零初始化投影的残差MLP，作为解耦网络，如图3所示。该设计使得网络能够直接从预训练的表示空间中学习，避免了随机起点，灵感来源于ControlNet的零卷积操作[46]，我们将其改编为残差MLP中的零线性操作。

Within this architecture, the main branch includes a zero-initialized, bias-free linear layer positioned subsequent to the combination of a SiLU activation and a normally initialized linear layer. Conventionally, the dimensions of features before the initial linear layer, situated between the first and second linear layers,and following the second linear layer,are named as the input d_{in} ,latent d_{mid} ,and output d_{out} dimensions,respectively. To rectify any mismatches between the input and output dimensions, the network employs nearest-neighbor downsampling within the shortcut path, thereby ensuring both alignment and the preservation of sharpness for the input features. During the inference stage, a weighting parameter $\alpha > 0$ is introduced to modulate the portion of features emanating from the main branch before their integration with the input features, whereas this parameter remains constant at 1 throughout the training phase.
在该架构中，主分支包括一个零初始化且无偏置的线性层，位于SiLU激活和一个正态初始化线性层的组合之后。通常，初始线性层之前、第一和第二线性层之间以及第二线性层之后的特征维度分别称为输入d_{in}、潜在d_{mid}和输出d_{out}维度。为纠正输入和输出维度之间的任何不匹配，网络在捷径路径中采用最近邻下采样，从而确保对齐并保持输入特征的锐利度。在推理阶段，引入权重参数$\alpha > 0$以调节主分支输出特征与输入特征融合前的比例，而该参数在训练阶段始终保持为1。

Inference After training,the disentangled network $\mathbf{f}_{\mathbf{c}}^{*}$ is utilized following CLIP's image encoder to extract visual content features. Moreover, given that vision-language data generation is rooted in a unified latent space, as depicted in Sec. 3, $\mathbf{f}_{\mathbf{c}}^{*}$ can be seamlessly integrated with CLIP's image and text encoders to enhance zero-shot capabilities. As shown in Fig. 2c,for an image $\mathbf{x}$,the operation is formulated as the composition function $\mathbf{f}_{\mathbf{c}}^{*} \circ \mathbf{f}_{\mathbf{x}}^{*}(\mathbf{x})$,and similarly,for a text $\mathbf{t}$,as $\mathbf{f}_{\mathbf{c}}^{*} \circ \mathbf{f}_{\mathbf{t}}^{*}(\mathbf{t})$. This integration preserves CLIP's zero-shot functionality while achieving refined features through the improved disentanglement of content.
推理 训练完成后，解耦网络$\mathbf{f}_{\mathbf{c}}^{*}$被用于CLIP图像编码器之后以提取视觉内容特征。此外，鉴于视觉-语言数据生成基于统一的潜在空间，如第3节所示，$\mathbf{f}_{\mathbf{c}}^{*}$可以无缝集成于CLIP的图像和文本编码器中以增强零样本能力。如图2c所示，对于图像$\mathbf{x}$，操作被形式化为复合函数$\mathbf{f}_{\mathbf{c}}^{*} \circ \mathbf{f}_{\mathbf{x}}^{*}(\mathbf{x})$，同理，对于文本$\mathbf{t}$，为$\mathbf{f}_{\mathbf{c}}^{*} \circ \mathbf{f}_{\mathbf{t}}^{*}(\mathbf{t})$。该集成保持了CLIP的零样本功能，同时通过改进的内容解耦实现了更精细的特征表达。

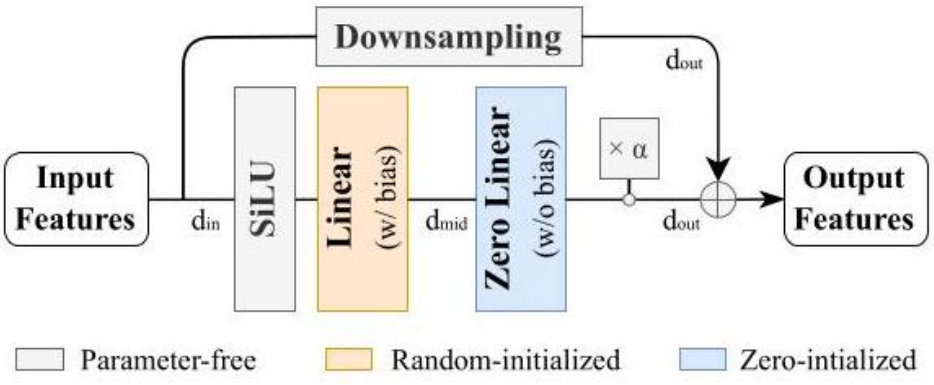


Fig. 3: Structure of the disentangled network. The architecture encompass a residual block featuring a zero-initialized, bias-free linear layer to commence optimization from the input feature space. When the input and output dimension differ, a downsampling operation is utilized to achieve alignment. During inference,a scalar parameter α balance the main branch and input features before combination.
图3：解耦网络结构。该架构包含一个残差块，起始为零初始化且无偏置的线性层，以从输入特征空间开始优化。当输入和输出维度不同时，采用下采样操作以实现对齐。推理时，标量参数α在融合前平衡主分支和输入特征。

8.3 4.2 Isolating Content from Style with Augmented Prompts

8.4 4.2 通过增强提示实现内容与风格的分离

Despite progress in disentangling content and style via image augmentation, adequately altering all style factors in an image remains challenging due to the high dimensionality and complexity of style information in images. Achieving substantial style changes through augmentation, essential for complete disentanglement [41], is difficult with existing image augmentation techniques. On the contrary, text data inherently possesses high semanticity and logical structure, making it more amenable to property-wise manipulation compared to image data. To further exploring the disentanglement of content, we propose Contrastive Learning with Augmented Prompts (CLAP).
尽管通过图像增强在内容与风格解耦方面取得了进展，但由于图像中风格信息的高维度和复杂性，充分改变图像中所有风格因素仍然具有挑战性。通过增强实现显著风格变化是实现完全解耦[41]的关键，但现有图像增强技术难以达到这一目标。相比之下，文本数据本身具有高度语义性和逻辑结构，相较于图像数据更易于逐属性操作。为进一步探索内容解耦，我们提出了基于增强提示的对比学习（Contrastive Learning with Augmented Prompts，CLAP）。

As depicted in Fig. 2b, CLAP employs an InfoNCE loss to train a disentangled network atop CLIP's pre-trained text encoder, keeping the encoder's gradients fixed, similar to Im.Aug. Leveraging the simpler structure of text, the template-based prompts previously utilized for synthesizing images now serve as the training text dataset,denoted by $\mathcal{D}_{\mathbf{t}}$. Utilizing the same disentangled network as in Im.Aug, the learning objective of CLAP is outlined as follows:
如图2b所示，CLAP利用InfoNCE损失在CLIP预训练文本编码器之上训练解耦网络，同时保持编码器梯度固定，类似于Im.Aug。借助文本结构的简洁性，先前用于合成图像的模板化提示现作为训练文本数据集，记为$\mathcal{D}_{\mathbf{t}}$。CLAP采用与Im.Aug相同的解耦网络，其学习目标如下所述:

$$\mathbf{f}_{\mathbf{c}}^{*}=\underset{\mathbf{f}_{\mathbf{c}}}{\operatorname{argmin}} \mathbb{E}_{\{\mathbf{t}_i\}_{i=1}^{b} \in \mathcal{D}_{\mathbf{t}}} \mathcal{L}\left(\mathbf{f}_{\mathbf{c}} \circ \mathbf{f}_{\mathbf{t}}^{*}, \left\{\mathbf{t}_i, \tilde{\mathbf{t}}_i\right\}_{i=1}^{b}, \tau\right)+\lambda \mathcal{L}\left(\mathbf{f}_{\mathbf{c}} \circ \mathbf{f}_{\mathbf{t}}^{*}, \left\{\mathbf{t}_i^{c}, \tilde{\mathbf{t}}_i\right\}_{i=1}^{b}, 1\right), \tag{4}$$

where $\mathbf{f}_{\mathbf{t}}^{*}$ denotes the pre-trained CLIP text encoder. The term $\mathbf{t}_i$ references a text prompt from $\mathcal{D}\mathbf{t}$,and $\tilde{\mathbf{t}}_i$ represents its augmented view,produced via prompt augmentation techniques. On the equation's right side, $\mathbf{t}_i^c$ specifies the class name associated with the text prompt $\mathbf{t}_i$. This strategy aims to enhance variations between prompt pairs,especially in cases where the text dataset $\mathcal{D}_{\mathbf{t}}$ has a very limited number of samples. Here, λ serves for adjusting the second term's importance in the total loss function. All other symbols in Eq. 4 maintain their definitions as described earlier.
其中$\mathbf{f}_{\mathbf{t}}^{*}$表示预训练的CLIP文本编码器。术语$\mathbf{t}_i$指的是来自$\mathcal{D}\mathbf{t}$的文本提示，$\tilde{\mathbf{t}}_i$表示通过提示增强技术生成的其增强视图。在方程右侧，$\mathbf{t}_i^c$指定与文本提示$\mathbf{t}_i$相关联的类别名称。该策略旨在增强提示对之间的差异，特别是在文本数据集$\mathcal{D}_{\mathbf{t}}$样本数量非常有限的情况下。这里，λ用于调整总损失函数中第二项的重要性。方程4中的其他符号保持之前的定义不变。

Table 2: Prompt augmentation techniques. Various augmented views are generated from an original text prompt using specific augmentation techniques: OSD (Object Size Deletion), OCD (Object Color Deletion), ITD (Image Type Deletion), ASD (Art Style Deletion), and SPO (Swapping Prompt Order).
表2：提示增强技术。通过特定的增强技术从原始文本提示生成各种增强视图：OSD（对象大小删除）、OCD（对象颜色删除）、ITD（图像类型删除）、ASD（艺术风格删除）和SPO（提示顺序交换）。

Original	OSD	OCD	ITD	ASD	SPO
a realistic painting of a large red car	a realistic painting of a red car	a realistic painting of a large car	a realistic of a large red car	a painting of a large red car	a large red car in a realistic painting
原始	OSD	OCD	ITD	ASD	SPO
一幅大型红色汽车的写实画	一幅红色汽车的写实画	一幅大型汽车的写实画	一幅大型红色汽车的写实画	一幅大型红色汽车的画	一幅写实画中的大型红色汽车

After training, the learned disentangled network is seamlessly integrated with both of CLIP's encoders to extract content representations, as depicted in Fig. 2c.
训练完成后，学习到的解耦网络无缝集成到CLIP的两个编码器中以提取内容表示，如图2c所示。

Prompt Augmentation To ensure text prompts undergo stylistic changes without compromising their content, we have developed specific augmentation techniques for synthetic text prompts. Drawing inspiration from Easy Data Augmentation (EDA) techniques [42], we adapted the Random Deletion (RD) and Random Swap (RS) techniques from EDA, customizing them to suit our prompt structure. To avoid inadvertently altering the content by introducing new object names or changing the core idea of a text prompt, our augmentation methods do not include random word insertions or replacements. Our primary augmentation techniques are Object Size Deletion (OSD), Object Color Deletion (OCD), Image Type Deletion (ITD), Art Style Deletion (ASD), and Swapping Prompt Order (SPO), each applied with a certain probability, as detailed in Tab. 2. Additionally, for down-stream datasets with few categories, to rich the population of training samples, we use an additional augmentation, named IGN (Inserting Gaussian Noise). Following the initializing protocol of prompt learning methods 47, 48, we insert a zero-mean Gaussian noise with 0.02 standard deviation with a noise length equals to 4 , to the tokenized prompts.
提示增强 为确保文本提示在风格上发生变化而不影响其内容，我们为合成文本提示开发了特定的增强技术。借鉴Easy Data Augmentation (EDA) 技术[42]，我们改编了EDA中的随机删除（Random Deletion, RD）和随机交换（Random Swap, RS）技术，针对我们的提示结构进行了定制。为了避免通过引入新的对象名称或改变文本提示的核心思想而无意中改变内容，我们的增强方法不包括随机插入或替换单词。我们的主要增强技术包括对象大小删除（Object Size Deletion, OSD）、对象颜色删除（Object Color Deletion, OCD）、图像类型删除（Image Type Deletion, ITD）、艺术风格删除（Art Style Deletion, ASD）和提示顺序交换（Swapping Prompt Order, SPO），每种方法以一定概率应用，详见表2。此外，对于类别较少的下游数据集，为丰富训练样本数量，我们使用了额外的增强方法，称为插入高斯噪声（Inserting Gaussian Noise, IGN）。按照提示学习方法47、48的初始化协议，我们向分词后的提示中插入均值为0、标准差为0.02、长度为4的高斯噪声。

Intuitively, these prompt augmentation methods parallel random masking techniques used in image augmentation [6, 17]. However, prompt augmentations are more effective and precise than their image counterparts. This effectiveness arises because prompt augmentations can specifically target and eliminate a particular style element without impacting the content, whereas image masking, operating at the pixel or patch level, might inadvertently damage content information or lead to insufficient style changes.
直观来看，这些提示增强方法类似于图像增强中使用的随机遮罩技术[6, 17]。然而，提示增强比图像遮罩更有效且更精确。这种有效性源于提示增强能够针对性地去除特定的风格元素而不影响内容，而图像遮罩在像素或图块级别操作，可能无意中破坏内容信息或导致风格变化不足。

9 5 Experiments

10 5 实验

We conduct three primary experiments to assess our method: (1) zero-shot evaluation with diverse prompts to gauge zero-shot performance and its robustness to prompt perturbations; (2) linear probe tests on few-shot samples to evaluate the efficacy of the learned representations in few-shot settings; and (3) adversarial attack assessments on zero-shot and one-shot classifiers to determine their resistance to adversarial threats. We further conduct an ablative study on hyper-parameters, explore the impact of different prompt augmentation combinations and various sources of training prompts on CLAP's performance, and replicate experiments across different CLIP model sizes.

我们进行了三项主要实验以评估我们的方法：（1）使用多样提示进行零样本评估，以衡量零样本性能及其对提示扰动的鲁棒性；（2）在少样本样本上进行线性探测测试，以评估学习表示在少样本设置中的有效性；（3）对零样本和单样本分类器进行对抗攻击评估，以确定其对对抗威胁的抵抗力。我们还进行了超参数消融研究，探讨了不同提示增强组合和不同训练提示来源对CLAP性能的影响，并在不同规模的CLIP模型上复现实验。

10.1 5.1 Experimental Setup

10.2 5.1 实验设置

Implementation. Im. Aug and CLAP are implemented using the ViT-B/16 CLIP model and executed on an NVIDIA RTX 3090 GPU. To ensure reproducibility, the random seed for all stochastic processes is fixed at 2024. More information on implementation details is provided in Appendix A.1
实现细节。Im. Aug和CLAP均基于ViT-B/16 CLIP模型实现，并在NVIDIA RTX 3090 GPU上运行。为确保可复现性，所有随机过程的随机种子固定为2024。更多实现细节见附录A.1。

Datasets. CLAP is assessed across four multi-domain datasets to examine its performance in varied environments: PACS [23] (4 domains, 7 categories), VLCS 1 (4 domains, 5 categories), OfficeHome [37] (4 domains, 65 categories), and DomainNet [35] (6 domains, 345 categories). For conciseness, we present average results across the domains for each dataset. Detailed experimental outcomes for each domain within these datasets are provided in Appendix A.4.
数据集。CLAP在四个多域数据集上进行评估，以检验其在不同环境下的表现：PACS [23]（4个域，7个类别）、VLCS 1（4个域，5个类别）、OfficeHome [37]（4个域，65个类别）和DomainNet [35]（6个域，345个类别）。为简洁起见，我们展示了每个数据集各域的平均结果。各数据集各域的详细实验结果见附录A.4。

Compute efficiency. CLAP demonstrates faster convergence and shorter training times compared to Im.Aug. For CLAP, training on the PACS and VLCS datasets is completed in roughly 11 minutes, OfficeHome in approximately 14 minutes, and DomainNet in about 47 minutes. In contrast, Im.Aug requires around 16 minutes for PACS and VLCS, 50 minutes for OfficeHome, and 3.3 hours for DomainNet. Both Im.Aug and CLAP maintain CLIP's inference efficiency due to the disentangled network's efficient two-layer MLP structure.
计算效率。与Im.Aug相比，CLAP表现出更快的收敛速度和更短的训练时间。CLAP在PACS和VLCS数据集上的训练时间约为11分钟，OfficeHome约14分钟，DomainNet约47分钟。相比之下，Im.Aug在PACS和VLCS上约需16分钟，OfficeHome约50分钟，DomainNet约3.3小时。由于解耦网络采用高效的两层多层感知机（MLP）结构，Im.Aug和CLAP均保持了CLIP的推理效率。

10.3 5.2 Main Results

10.4 5.2 主要结果

Zero-Shot Performance To assess zero-shot capabilities, CLAP undergoes evaluation using three specific fixed prompts: $\mathbf{ZS(C)}$,utilizing only the class name within "[class]"; ZS(PC), with the format "a photo of a [class]"; and ZS(CP), structured as "a [class] in a photo". To thoroughly examine zero-shot proficiency, a dynamic prompt, $\mathbf{ZS(NC)}$,formatted as "[noise][class]",is also used,where "[noise]" signifies the introduction of Gaussian noise characterized by a mean of 0 and a standard deviation of 0.02 .
零样本性能 为评估零样本能力，CLAP使用三种特定固定提示进行评测：$\mathbf{ZS(C)}$，仅使用“[class]”中的类别名称；ZS(PC)，格式为“a photo of a [class]”；以及ZS(CP)，结构为“a [class] in a photo”。为了全面考察零样本能力，还使用了动态提示$\mathbf{ZS(NC)}$，格式为“[noise][class]”，其中“[noise]”表示引入均值为0、标准差为0.02的高斯噪声。

As presented in Tab. 3, CLAP surpasses both CLIP and Im.Aug across all evaluated prompts for every dataset. Unlike the uniform enhancement in zero-shot performance CLAP achieves over CLIP, Im.Aug displays inconsistent results. A closer examination reveals CLAP's superiority over CLIP is especially significant for the dynamic ZS(NC) prompt. This demonstrates CLAP's effectiveness in significantly improving zero-shot performance compared to the original CLIP representations.
如表3所示，CLAP在所有数据集的所有评估提示中均优于CLIP和Im.Aug。与CLAP对CLIP实现的零样本性能均匀提升不同，Im.Aug的结果表现不稳定。进一步分析显示，CLAP相较于CLIP的优势在动态ZS(NC)提示下尤为显著，证明了CLAP在显著提升零样本性能方面的有效性，相较于原始CLIP表示。

In assessing the model's robustness to prompt perturbations, we examine the variances in zero-shot performance across different prompts by analyzing the range(R)and standard deviation (δ) of results derived from $\mathbf{ZS(C)}, \mathbf{ZS(CP)}$, and $\mathbf{ZS(PC)}$. Additionally,we investigate the decline $\left(\Delta_{(NC)}\right)$ in performance from $\mathbf{ZS(C)}$ to $\mathbf{ZS(NC)}$ as a broad indicator of resilience to noised prompts.
在评估模型对提示扰动的鲁棒性时，我们通过分析由$\mathbf{ZS(C)}, \mathbf{ZS(CP)}$和$\mathbf{ZS(PC)}$得出的结果范围(R)和标准差($\delta$)，考察不同提示下零样本性能的方差。此外，我们还研究了从$\mathbf{ZS(C)}$到$\mathbf{ZS(NC)}$的性能下降$\left(\Delta_{(NC)}\right)$，作为对噪声提示鲁棒性的总体指标。

Table 3: Zero-shot results across three distinct prompts: "C" for "[class]", "CP" for "a [class] in a photo", "PC" for "a photo of a [class]", and a dynamic prompt "NC" for "[noise][class]" showcase that CLAP consistently outperforms CLIP's zero-shot performance across all datasets, whereas image augmentation exhibits mixed outcomes.
表3：三种不同提示下的零样本结果：“C”代表“[class]”，“CP”代表“一张[class]的照片”，“PC”代表“一张[class]的照片”，以及动态提示“NC”代表“[noise][class]”，展示了CLAP在所有数据集上持续优于CLIP的零样本性能，而图像增强则表现出混合结果。

Prompt		Zero-shot performance, avg. 1top-1 acc.(%) (↑)				
		PACS	VLCS	Off.Home	Dom.Net	Overall
ZS(C)	CLIP	95.7	76.4	79.8	57.8	77.4
	Im.Aug	96.5	79.5	77.0	51.5	76.1
	CLAP	97.2	82.6	81.0	58.7	79.9
ZS(CP)	CLIP	95.2	82.0	79.5	57.0	78.4
	Im.Aug	96.3	82.9	75.8	50.7	76.4
	CLAP	97.3	83.4	80.5	58.0	79.8
ZS(PC)	CLIP	96.1	82.4	82.5	57.7	79.7
	Im.Aug	96.5	83.0	78.6	51.6	77.4
	CLAP	97.2	83.4	83.0	59.0	80.6
ZS(NC)	CLIP	90.8	68.3	71.5	51.0	70.4
	Im.Aug	94.8	73.1	67.5	44.0	69.9
	CLAP	97.2	81.0	73.5	52.6	76.1

提示		零样本性能，平均1top-1准确率(%) (↑)				
		PACS	VLCS	办公.家居	家庭.网络	总体
ZS(C)	CLIP	95.7	76.4	79.8	57.8	77.4
	图像增强	96.5	79.5	77.0	51.5	76.1
	CLAP	97.2	82.6	81.0	58.7	79.9
ZS(CP)	CLIP	95.2	82.0	79.5	57.0	78.4
	图像增强	96.3	82.9	75.8	50.7	76.4
	CLAP	97.3	83.4	80.5	58.0	79.8
ZS(PC)	CLIP	96.1	82.4	82.5	57.7	79.7
	图像增强	96.5	83.0	78.6	51.6	77.4
	CLAP	97.2	83.4	83.0	59.0	80.6
ZS(NC)	CLIP	90.8	68.3	71.5	51.0	70.4
	图像增强	94.8	73.1	67.5	44.0	69.9
	CLAP	97.2	81.0	73.5	52.6	76.1

As presented in Tab. 4, CLAP significantly reduces the variance in zero-shot performance across various testing prompts, evidenced by markedly lower values of δ and R ,and a less pronounced decrease in performance with a noised prompt, in contrast to Im.Aug and the baseline representations of CLIP. Although Im. Aug aids in reducing performance variance to some extent, its efficacy is notably inferior to that of CLAP. These findings highlight CLAP's enhanced robustness in maintaining consistent zero-shot performance across a diverse array of prompts.

如表4所示，CLAP显著降低了不同测试提示下零样本性能的方差，表现为δ和R值明显较低，且在带噪提示下性能下降不明显，相较于Im.Aug和CLIP的基线表示。虽然Im.Aug在一定程度上有助于减少性能方差，但其效果明显不及CLAP。这些结果凸显了CLAP在保持多样提示下零样本性能一致性方面的增强鲁棒性。

Few-Shot Performance We conduct evaluations of 1-shot, 4-shot, 8-shot, and 16-shot linear probes across each domain within the four datasets. As illustrated in Fig. 4, CLAP significantly outperforms both CLIP and Im.Aug in few-shot learning scenarios. Notably, in the 1-shot setting CLAP achieves performance improvements over the linear-probe CLIP model by margins of +10%, +3.5% , +2.5% ,and +1.5% on the PACS,VLCS,OfficeHome,and DomainNet datasets, respectively. These improvements are especially significant in comparison to the gains observed with Im.Aug counterparts, underpinning CLAP's efficacy in few-shot scenarios. For detailed quantitative results, please refer to Appendix A.4,

少样本性能 我们在四个数据集的各个领域中进行了1-shot、4-shot、8-shot和16-shot线性探测器的评估。如图4所示，CLAP在少样本学习场景中显著优于CLIP和Im.Aug。尤其在1-shot设置中，CLAP在PACS、VLCS、OfficeHome和DomainNet数据集上分别比线性探测CLIP模型提升了+10%, +3.5%、+2.5%和+1.5%的性能。这些提升相比Im.Aug的增益尤为显著，进一步证明了CLAP在少样本场景中的有效性。详细的定量结果请参见附录A.4。

Adversarial Performance To assess adversarial robustness, zero-shot (ZS(C)) and one-shot classifiers are evaluated against prominent adversarial attack methods, such as FGSM [15, PGD 30, and CW 5, by generating adversarial samples for testing. For FGSM, 1 adversarial iteration is employed, whereas for PGD and CW, 20 iterations are used, all with an epsilon of 0.031 . As indicated in Tab. 5, classifiers utilizing CLAP representations demonstrate superior resilience to these adversarial attacks compared to those based on CLIP representations. Across the four datasets, CLAP's zero-shot and 1-shot classifiers surpass CLIP by margins of +7.6% and +8.5% against FGSM, +1.0% and +11.7% against PGD-20,and +1.1% and +2.3% against CW-20, respectively. These figures notably exceed the performance improvements of +4.4% and +4.6% against FGSM, +0.3% and +6.2% against PGD-20, and 0% and +1.3% against CW-20 achieved by Im.Aug. The result suggests that CLAP efficiently enhances robustness against adversarial attacks in both zero-shot and one-

shot scenarios.

对抗性能 为评估对抗鲁棒性，零样本（ZS(C)）和一样本分类器针对主流对抗攻击方法如FGSM [15]、PGD [30]和CW [5]进行了测试，通过生成对抗样本进行评估。FGSM采用1次对抗迭代，PGD和CW均采用20次迭代，epsilon均为0.031。如表5所示，采用CLAP表示的分类器在抵抗这些对抗攻击方面表现出优于基于CLIP表示的分类器。在四个数据集中，CLAP的零样本和一样本分类器在FGSM攻击下分别比CLIP高出+7.6%和+8.5%，在PGD-20攻击下高出+1.0%和+11.7%，在CW-20攻击下分别提升1.1%和2.3%。这些数值显著超过Im.Aug在FGSM攻击下的+4.4%和+4.6%，PGD-20攻击下的0.3%和6.2%，以及CW-20攻击下的0%和1.3%的性能提升。结果表明，CLAP有效增强了零样本和一样本场景下对抗攻击的鲁棒性。

Table 4: CLAP more effectively reduces zero-shot performance variance across prompts than image augmentation,with R and δ indicating the range and standard deviation for $\mathbf{ZS(C)}, \mathbf{ZS(CP)}$,and $\mathbf{ZS(PC)}$. The decrease $\Delta_{(NC)}$ from $\mathbf{ZS(C)}$ to $\mathbf{ZS(NC)}$ highlights CLAP's enhanced robustness against prompt perturbations.

表4：CLAP比图像增强更有效地降低了不同提示下的零样本性能方差，R和δ分别表示$\mathbf{ZS(C)}, \mathbf{ZS(CP)}$和$\mathbf{ZS(PC)}$的范围和标准差。从$\mathbf{ZS(C)}$到$\mathbf{ZS(NC)}$的下降$\Delta_{(NC)}$凸显了CLAP在应对提示扰动时的增强鲁棒性。

	Metric Method	Performance variance,avg. top-1 acc. (%) (↓)				
		PACS	VLCS	Off.Home	Dom.Net	Overall
R	CLIP	0.9	6.1	3.1	0.8	2.7
	Im.Aug	0.1	3.6	2.8	0.9	1.9
	CLAP	0.1	0.8	2.5	1.0	1.1
δ	CLIP	0.4	2.8	1.4	0.4	1.2
	Im.Aug	0.1	1.7	1.2	0.4	0.8
	CLAP	0.0	0.4	1.1	0.4	0.5
${\Delta }_{\left(NC\right) }$	CLIP	4.9	8.1	8.3	6.8	7.0
	Im.Aug	1.6	6.4	9.5	7.5	6.3
	CLAP	0.0	1.6	7.5	6.1	3.8

	度量方法	性能方差，平均Top-1准确率 (%) (↓)				
		PACS	VLCS	Off.Home	Dom.Net	总体
R	CLIP	0.9	6.1	3.1	0.8	2.7
	图像增强	0.1	3.6	2.8	0.9	1.9
	CLAP	0.1	0.8	2.5	1.0	1.1
δ	CLIP	0.4	2.8	1.4	0.4	1.2
	图像增强	0.1	1.7	1.2	0.4	0.8
	CLAP	0.0	0.4	1.1	0.4	0.5
${\Delta }_{\left(NC\right) }$	CLIP	4.9	8.1	8.3	6.8	7.0
	图像增强	1.6	6.4	9.5	7.5	6.3
	CLAP	0.0	1.6	7.5	6.1	3.8

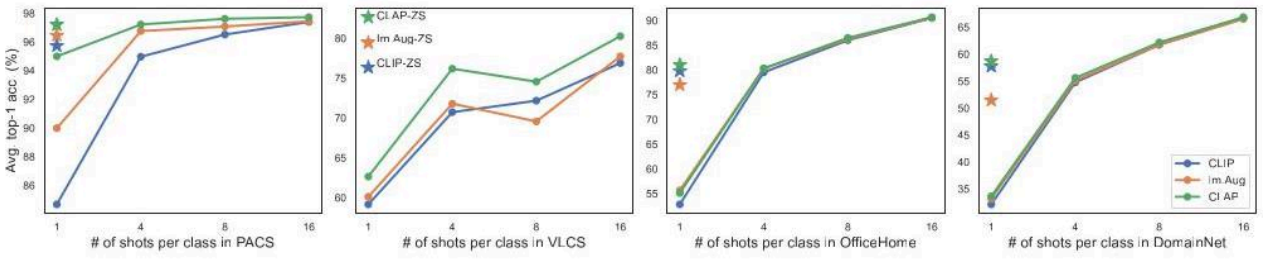


Fig. 4: Few-shot linear probe comparisons of image-encoder features show that CLAP enhances CLIP's few-shot performance more effectively than Im.Aug. In the accompanying figure, "ZS" indicates the zero-shot performance using a "[class]" prompt.

图4：图像编码器特征的少样本线性探测比较显示，CLAP比图像增强（Im.Aug）更有效地提升了CLIP的少样本性能。图中“ZS”表示使用“[class]”提示的零样本性能。

10.5 5.3 More Analysis

10.6 5.3 更多分析

t-SNE Visualization In our t-SNE visualizations, we examine the representations of CLIP, Im.Aug, and CLAP for all images within the Art Painting domain of the PACS dataset. Fig. 5 shows that CLAP's image representations display a marked inter-class separation and tighter intra-class clustering than those of

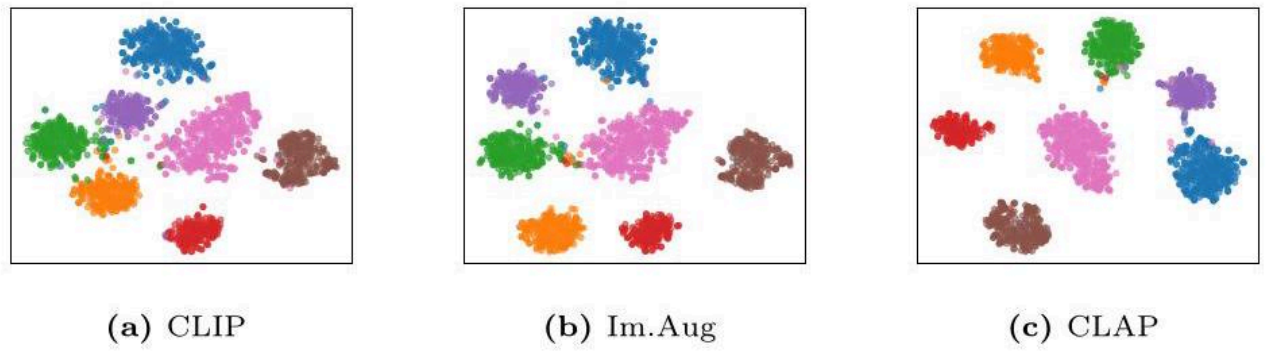
t-SNE可视化 在我们的t-SNE可视化中，我们考察了PACS数据集中艺术绘画领域内CLIP、Im.Aug和CLAP对所有图像的表示。图5显示，CLAP的图像表示相比其他方法表现出更明显的类间分离和更紧密的类内聚类。

Table 5: Image augmentation and CLAP both enhance CLIP's zero-shot with the "[class]" prompt and 1-shot robustness against adversarial attacks, with CLAP showing greater improvements.

表5：图像增强和CLAP均提升了CLIP在使用“[class]”提示的零样本性能及1-shot对抗攻击的鲁棒性，其中CLAP的提升更为显著。

Setting	Method	Avg. top-1 acc. (%) under adversarial attacks(↑)												Avg.
		FGSM				PGD-20				CW-20				
		PACS	VLCS	O.H.	D.N.	PACS	VLCS	O.H.	D.N.	PACS	VLCS	O.H.	D.N.	
ZS(C)	CLIP	86.8	65.6	57.9	22.5	29.1	2.0	10.1	10.7	27.4	1.5	7.4	7.6	29.2
	Im.Aug	88.0	69.6	55.1	37.9	31.3	2.1	10.4	9.0	29.4	1.7	7.0	5.8	31.1
	CLAP	88.7	71.9	58.5	44.2	30.8	3.2	10.6	11.2	29.8	2.3	8.1	8.0	32.7
1-shot	CLIP	66.7	45.2	34.3	22.5	34.8	16.0	5.6	11.3	18.9	0.7	4.5	3.2	23.7
	Im. Aug	79.4	47.1	37.1	23.5	55.2	16.1	8.5	12.5	23.2	0.9	5.1	3.4	28.0
	CLAP	89.6	52.2	37.1	23.9	73.4	21.2	7.4	12.5	27.0	1.1	5.0	3.5	31.9

设置方法		对抗攻击下的平均Top-1准确率(%)（↑）												
		FGSM（快速梯度符号法）				PGD-20（投影梯度下降-20步）				CW-20（Carlini-Wagner攻击-20步）				
		PACS（PACS数据集）	VLCS（VLCS数据集）	O.H.	D.N.	PACS（PACS数据集）	VLCS（VLCS数据集）	O.H.	D.N.	PACS（PACS数据集）	VLCS（VLCS数据集）	O.H.	D.N.	平均值
ZS(C)	CLIP（对比语言-图像预训练）	86.8	65.6	57.9	22.5	29.1	2.0	10.1	10.7	27.4	1.5	7.4	7.6	29.2
	图像增强	88.0	69.6	55.1	37.9	31.3	2.1	10.4	9.0	29.4	1.7	7.0	5.8	31.1
	CLAP	88.7	71.9	58.5	44.2	30.8	3.2	10.6	11.2	29.8	2.3	8.1	8.0	32.7
一次样本	CLIP（对比语言-图像预训练）	66.7	45.2	34.3	22.5	34.8	16.0	5.6	11.3	18.9	0.7	4.5	3.2	23.7
	图像增强	79.4	47.1	37.1	23.5	55.2	16.1	8.5	12.5	23.2	0.9	5.1	3.4	28.0
	CLAP	89.6	52.2	37.1	23.9	73.4	21.2	7.4	12.5	27.0	1.1	5.0	3.5	31.9



(a) CLIP (b) Im.Aug (c) CLAP

Fig. 5: t-SNE visualizations of all images in the Art Painting of PACS dataset show CLAP outperforms the original CLIP and Im.Aug, with clearer inter-class distinctions and tighter intra-class clusters.

图5：PACS数据集中艺术绘画（Art Painting）所有图像的t-SNE可视化显示，CLAP优于原始CLIP和Im.Aug，表现为类间区分更清晰，类内聚类更紧密。

CLIP and Im.Aug. This observation suggests that CLAP's representations are more closely tied to content information and less influenced by style information, in contrast to the other two.
CLIP和Im.Aug。该观察表明，CLAP的表示更紧密地关联内容信息，且较少受风格信息影响，这与另外两者形成对比。

Ablations In Fig. 6, we assess the zero-shot capabilities of our model using two distinct prompts, $\overline{\text{ZS}}\,(\text{C})$ and $\overline{\text{ZS}}\,(\text{PC})$,on the VLCS dataset. This analysis forms part of an ablative study aimed at understanding the influence of various hyper-parameters on model performance. Specifically, we examine: the dimensions of the latent layer within the MLP of the disentangled network, as illustrated in Fig. 6a; the temperature parameter (τ) in the loss function,as depicted in Fig. 6b and the weight coefficient (α) during the inference stage,as shown in Fig. 6c. Our findings indicate that CLAP consistently enhances zero-shot performance across all tested configurations for both prompts, while also significantly reducing the gap between the performances elicited by each prompt. These results underscore the efficacy of CLAP in accommodating a wide range of hyper-parameters.
消融实验 在图6中，我们使用VLCS数据集上的两个不同提示词$\overline{\text{ZS}}\,(\text{C})$和$\overline{\text{ZS}}\,(\text{PC})$评估模型的零样本能力。该分析是消融研究的一部分，旨在理解各种超参数对模型性能的影响。具体而言，我们考察了：图6a所示的解耦网络多层感知机（MLP）中潜在层的维度；图6b所示损失函数中的温度参数(τ)；以及图6c所示推理阶段的权重系数(α)。结果表明，CLAP在所有测试配置和两个提示词下均持续提升零样本性能，同时显著缩小了不同提示词性能间的差距。这些结果凸显了CLAP在适应多种超参数方面的有效性。

Prompt Augmentation Combinations We explore diverse combinations of our tailored prompt augmentation methods and examine Easy Data Augmentation (EDA) techniques [42] on the VLCS dataset. Each tested technique showcases CLAP's enhancements over CLIP, with details available in Appendix A.2.
提示增强组合 我们探索了多种定制提示增强方法的组合，并在VLCS数据集上检验了Easy Data Augmentation（EDA）技术[42]。每种测试技术均展示了CLAP相较于CLIP的提升，具体细节见附录A.2。

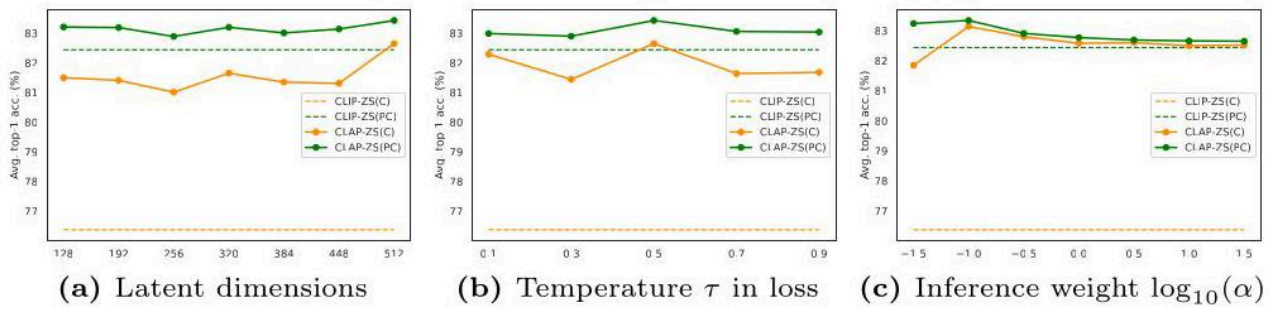


Fig. 6: We conduct ablative study on hyper-parameter choices on the VLCS dataset, including latent dimensions, τ values,and α values during the inference stage. CLAP continuously enhance CLIP's performance throughout the tested values.
图6：我们在VLCS数据集上对超参数选择进行了消融研究，包括潜在维度、推理阶段的τ值和α值。CLAP在所有测试值范围内持续提升CLIP的性能。

Prompt Sources We assess the impact of different training prompt formats, originating from various synthetic sources, on the performance of the VLCS dataset, incorporating EDA techniques. Our evaluation includes our template-based prompts, LLM-generated prompts by ChatGPT-3.5 [3] (with the generation process detailed in Appendix B.2), prompts structured as "a [random] style of [class]," where "[random]" is filled with terms from a random word generator [2] , and prompts produced using the PromptStyler method [9]. The findings indicate that the training prompts with simpler forms tend to yield better performance, with detailed quantitative results presented in Appendix A.3.
提示来源 我们评估了不同训练提示格式对VLCS数据集性能的影响，这些提示来源于多种合成源，并结合了EDA技术。评估内容包括基于模板的提示、由ChatGPT-3.5 [3]生成的大型语言模型（LLM）提示（生成过程详见附录B.2）、结构为“a [random] style of [class]”的提示，其中“[random]”由随机词生成器[2]填充，以及使用PromptStyler方法[9]生成的提示。结果表明，形式较为简单的训练提示往往带来更佳性能，详细定量结果见附录A.3。

Experiments on Different Model Scales In our repeated experiments assessing zero-shot performance on the ViT-L/14 and ResNet50x16 pre-trained with CLIP, we consistently find that CLAP improves zero-shot performance while also reducing performance variances. This consistent observation underscores CLAP's effectiveness in enhancing the quality of CLIP representations. For quantitative details supporting these findings, please see the Appendix C
不同模型规模的实验 在对ViT-L/14和ResNet50x16（均经CLIP预训练）进行的多次零样本性能评估中，我们始终发现CLAP提升了零样本性能并减少了性能波动。这一持续观察强调了CLAP在提升CLIP表示质量方面的有效性。支持这些发现的定量细节请参见附录C。

11 6 Conclusion

12 6 结论

To enhance pre-trained CLIP-like models, this study delves into disentangling latent content variables. Through a causal analysis of the underlying generative processes of vision-language data, we discover that training a disentangled network in one modality can effectively disentangle content across both modalities. Given the high semantic nature of text data, we identify that disentanglement is more achievable within the language modality through text augmentation interventions. Building on these insights, we introduce CLAP (Contrastive Learning with Augmented Prompts) to acquire disentangled vision-language content features. Comprehensive experiments validate CLAP's effectiveness, demonstrating significant improvements in zero-shot and few-shot performance, and enhancing robustness against perturbations. We anticipate that our work will inspire further exploration into disentangling latent variables within vision-language models.
References
为提升预训练的CLIP类模型，本研究深入探讨了潜在内容变量的解耦。通过对视觉-语言数据生成过程的因果分析，我们发现，在单一模态中训练解耦网络能够有效实现跨模态内容的解耦。鉴于文本数据的高度语义特性，我们确定通过文本增强干预在语言模态中更易实现解耦。基于此，我们提出了CLAP（带增强提示的对比学习，Contrastive Learning with Augmented Prompts）以获取解耦的视觉-语言内容特征。全面实验验证了CLAP的有效性，表现为零样本和少样本性能显著提升，并增强了对扰动的鲁棒性。我们期待本工作能激发对视觉-语言模型潜在变量解耦的进一步探索。参考文献

---

2 https://github.com/vaibhavsingh97/random-word
2 https://github.com/vaibhavsingh97/random-word

---

1. Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T.H., Mitliagkas, I.: Generalizing to unseen domains via distribution matching. arXiv preprint arXiv:1911.00804 (2019). https://doi.org/10.48550/arXiv.1911.00804
2. Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T.H., Mitliagkas, I.: 通过分布匹配实现对未见域的泛化。arXiv预印本 arXiv:1911.00804 (2019). https://doi.org/10.48550/arXiv.1911.00804

2. Bollen, K.A.: Structural equations with latent variables, vol. 210. John Wiley & Sons (1989). https://doi.org/10.1002/9781118619179
3. Bollen, K.A.: 含潜变量的结构方程模型，卷210。John Wiley & Sons (1989). https://doi.org/10.1002/9781118619179

3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877-1901 (2020), https://dl.acm.org/doi/abs/10.5555/3495724.3495883
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 等：语言模型是少样本学习者。神经信息处理系统进展33，1877-1901 (2020)，https://dl.acm.org/doi/abs/10.5555/3495724.3495883

4. Carlini, N., Terzis, A.: Poisoning and backdooring contrastive learning. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=iC4UHbQ01Mp
5. Carlini, N., Terzis, A.：对比学习中的投毒与后门攻击。载于：国际学习表征会议（2021），https://openreview.net/forum?id=iC4UHbQ01Mp

5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39-57. IEEE Computer Society (2017). https://doi.org/10.1109/SP.2017.49
6. Carlini, N., Wagner, D.：评估神经网络鲁棒性的探索。载于：2017年IEEE安全与隐私研讨会（SP），第39-57页。IEEE计算机学会（2017）。https://doi.org/10.1109/SP.2017.49

6. Chen, P., Liu, S., Zhao, H., Jia, J.: Gridmask data augmentation. arXiv preprint arXiv:2001.04086 (2020). https://doi.org/10.48550/arXiv.2001.04086
7. Chen, P., Liu, S., Zhao, H., Jia, J.：Gridmask数据增强方法。arXiv预印本 arXiv:2001.04086 (2020)。https://doi.org/10.48550/arXiv.2001.04086

7. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597-1607. PMLR (2020), https://dl.acm.org/doi/abs/10.5555/ 3524938.3525087
8. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.：视觉表征对比学习的简单框架。载于：国际机器学习会议，第1597-1607页。PMLR（2020），https://dl.acm.org/doi/abs/10.5555/3524938.3525087

8. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision

and pattern recognition. pp. 15750-15758 (2021). https://doi.org/10.1109/CVPR46437.2021.01549

9. Chen, X., He, K.：探索简单的孪生网络表征学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集，第15750-15758页（2021）。https://doi.org/10.1109/CVPR46437.2021.01549

9. Cho, J., Nam, G., Kim, S., Yang, H., Kwak, S.: Promptstyler: Prompt-driven style generation for source-free domain generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 15702-15712 (2023). https: //doi.org/10.1109/ICCV51070.2023.01439
10. Cho, J., Nam, G., Kim, S., Yang, H., Kwak, S.：Promptstyler：基于提示的无源域泛化风格生成。载于：IEEE国际计算机视觉会议论文集，第15702-15712页（2023）。https://doi.org/10.1109/ICCV51070.2023.01439

10. Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., Vogt, J.E.: Identifiability results for multimodal contrastive learning. ICLR (2023), https://openreview.net/ forum?id=U_2kuqoTcB
11. Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., Vogt, J.E.：多模态对比学习的可识别性结果。ICLR（2023），https://openreview.net/forum?id=U_2kuqoTcB

11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020), https://openreview.net/forum?id= YicbFdNTTy
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., 等：一张图像胜过16x16个词：大规模图像识别的Transformer（变换器）。载于：国际学习表征会议（2020），https://openreview.net/forum?id=YicbFdNTTy

12. Fort, S.: Adversarial vulnerability of powerful near out-of-distribution detection. arXiv preprint arXiv:2201.07012 (2022). https://doi.org/10.48550/arXiv.2201.07012
13. Fort, S.：强大近似分布外检测的对抗脆弱性。arXiv预印本 arXiv:2201.07012 (2022)。https://doi.org/10.48550/arXiv.2201.07012

13. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132** (2), 581 – 595 (2024) . https://doi.org/10.1007/ s11263-023-01891-x
14. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.：Clip-adapter：通过特征适配器提升视觉-语言模型。国际计算机视觉杂志 **132** (2), 581 – 595 (2024) 。https://doi.org/10.1007/s11263-023-01891-x

14. Ge, C., Huang, R., Xie, M., Lai, Z., Song, S., Li, S., Huang, G.: Domain adaptation via prompt learning. arXiv preprint arXiv:2202.06687 (2022). https://doi.org/ 10.48550/arXiv.2202.06687
15. Ge, C., Huang, R., Xie, M., Lai, Z., Song, S., Li, S., Huang, G.：通过提示学习实现域适应。arXiv预印本 arXiv:2202.06687 (2022)。https://doi.org/10.48550/arXiv.2202.06687

15. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015). https: //doi.org/10.48550/arXiv.1412.6572
16. Goodfellow, I.J., Shlens, J., Szegedy, C.：解释与利用对抗样本。载于：国际学习表征会议（2015）。https://doi.org/10.48550/arXiv.1412.6572

16. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Do-ersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems 33, 21271-21284 (2020), https://dl.acm.org/doi/abs/10.5555/3495724.3497510
17. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., 等：Bootstrap your own latent——一种新的自监督学习方法。神经信息处理系统进展33, 21271-21284 (2020), https://dl.acm.org/doi/abs/10.5555/3495724.3497510

17. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000-16009 (2022). https://doi.org/10.1109/CVPR52688.2022.01553
18. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: 掩码自编码器是可扩展的视觉学习者。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第16000-16009页 (2022). https://doi.org/10.1109/CVPR52688.2022.01553

18. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729-9738 (2020). https://doi.org/10.1109/CVPR42600.2020.00975
19. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: 动量对比用于无监督视觉表示学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第9729-9738页 (2020). https://doi.org/10.1109/CVPR42600.2020.00975

19. Hong, T., Guo, X., Ma, J.: Itmix: Image-text mix augmentation for transferring clip to image classification. In: 2022 16th IEEE International Conference on Signal Processing (ICSP). vol. 1, pp. 129-133. IEEE (2022). https://doi.org/10.1109/ICSP56322.2022.9965292

20. Hong, T., Guo, X., Ma, J.: Itmix：图像-文本混合增强用于将CLIP迁移到图像分类。载于：2022年第16届IEEE国际信号处理会议（ICSP）。第1卷，第129-133页。IEEE (2022). https://doi.org/10.1109/ ICSP56322.2022.9965292

20. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904-4916. PMLR (2021), https://proceedings.mlr.press/v139/jia21b.html
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: 利用噪声文本监督扩展视觉及视觉-语言表示学习规模。载于：国际机器学习会议。第4904-4916页。PMLR (2021), https://proceedings.mlr.press/v139/jia21b.html

21. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113-19122 (2023). https://doi.org/10.1109/CVPR52729.2023.01832
22. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple：多模态提示学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第19113-19122页 (2023). https://doi.org/10.1109/CVPR52729.2023.01832

22. Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., Zhang, K.: Partial disentanglement for domain adaptation. In: International Conference on Machine Learning. pp. 11455-11472. PMLR (2022), https://proceedings.mlr.press/v162/kong22a.html
23. Kong, L., Xie, S., Yao, W., Zheng, Y., Chen, G., Stojanov, P., Akinwande, V., Zhang, K.: 用于领域适应的部分解耦。载于：国际机器学习会议。第11455-11472页。PMLR (2022), https:// proceedings.mlr.press/v162/kong22a.html

23. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: Deeper, broader and artier domain generalization. In: Proceedings of the IEEE international conference on computer vision. pp. 5542-5550 (2017). https://doi.org/10.1109/ICCV.2017.591
24. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.M.: 更深、更广、更艺术化的领域泛化。载于：IEEE国际计算机视觉会议论文集。第5542-5550页 (2017). https://doi.org/10.1109/ICCV.2017.591

24. Li, H., Wang, X., Zhang, Z., Yuan, Z., Li, H., Zhu, W.: Disentangled contrastive learning on graphs. Advances in Neural Information Processing Systems 34, 21872- 21884 (2021), https://dl.acm.org/doi/10.5555/3540261.3541935
25. Li, H., Wang, X., Zhang, Z., Yuan, Z., Li, H., Zhu, W.: 图上的解耦对比学习。神经信息处理系统进展34, 21872-21884 (2021), https://dl.acm.org/doi/10.5555/3540261.3541935

25. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In: International Conference on Learning Representations (2021)
26. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: 监督无处不在：一种数据高效的对比语言-图像预训练范式。载于：国际表征学习会议 (2021)

26. Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., van den Hengel, A., Zhang, K., Shi, J.Q.: Identifiable latent polynomial causal models through the lens of change. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=ia9fKO1Vjq
27. Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., van den Hengel, A., Zhang, K., Shi, J.Q.: 通过变化视角识别潜在多项式因果模型。载于：第十二届国际表征学习会议 (2024), https://openreview.net/forum?id=ia9fKO1Vjq

27. Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., Hengel, A.v.d., Zhang, K., Shi, J.Q.: Identifying weight-variant latent causal models. arXiv preprint arXiv:2208.14153 (2022). https://doi.org/10.48550/arXiv.2208.14153
28. 刘洋, 张志, 龚东, 龚明, 黄斌, A.v.d. Hengel, 张凯, 史建强: 识别权重变异潜在因果模型. arXiv预印本 arXiv:2208.14153 (2022). https://doi.org/10.48550/arXiv.2208.14153

28. Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., Hengel, A.v.d., Zhang, K., Shi, J.Q.: Identifiable latent neural causal models. arXiv preprint arXiv:2403.15711 (2024). https://doi.org/10.48550/arXiv.2403.15711
29. 刘洋, 张志, 龚东, 龚明, 黄斌, A.v.d. Hengel, 张凯, 史建强: 可识别的潜在神经因果模型. arXiv预印本 arXiv:2403.15711 (2024). https://doi.org/10.48550/arXiv.2403.15711

29. Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., Zhang, K., Shi, J.Q.: Identifying latent causal content for multi-source domain adaptation. arXiv preprint arXiv:2208.14161 (2022). https://doi.org/10.48550/arXiv.2208.14161
30. 刘洋, 张志, 龚东, 龚明, 黄斌, 张凯, 史建强: 多源域适应的潜在因果内容识别. arXiv预印本 arXiv:2208.14161 (2022). https://doi.org/10.48550/arXiv.2208.14161

30. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=rJzIBfZAb
31. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: 面向抗对抗攻击的深度学习模型. 载于: 国际学习表征会议 (2018), https://openreview.net/forum?id=rJzIBfZAb

31. Mahajan, D., Tople, S., Sharma, A.: Domain generalization using causal matching. In: International Conference on Machine Learning. pp. 7313-7324. PMLR (2021), https://proceedings.mlr.press/v139/mahajan21b.html

32. Mahajan, D., Tople, S., Sharma, A.: 利用因果匹配进行域泛化. 载于: 国际机器学习会议. 页7313-7324. PMLR (2021), https://proceedings.mlr.press/v139/mahajan21b.html

32. Mamooler, S.: Clip explainability. https://github.com/sMamooler/CLIP_Explainability, accessed: 2024-03-06
33. Mamooler, S.: CLIP可解释性. https://github.com/sMamooler/CLIP_Explainability, 访问时间: 2024-03-06

33. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: Understanding zero-shot adversarial robustness for large-scale models. In: The Eleventh International Conference on Learning Representations (2022), https://openreview.net/forum?id= P4bXCawRi5J
34. Mao, C., Geng, S., Yang, J., Wang, X., Vondrick, C.: 理解大规模模型的零样本对抗鲁棒性. 载于: 第十一届国际学习表征会议 (2022), https://openreview.net/forum?id= P4bXCawRi5J

34. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018). https://doi.org/10.48550/ arXiv.1807.03748
35. Oord, A.v.d., Li, Y., Vinyals, O.: 基于对比预测编码的表示学习. arXiv预印本 arXiv:1807.03748 (2018). https://doi.org/10.48550/ arXiv.1807.03748

35. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE international conference on computer vision. pp. 1406-1415 (2019). https://doi.org/10.1109/ ICCV.2019.00149
36. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: 多源域适应的矩匹配方法. 载于: IEEE国际计算机视觉会议论文集. 页1406-1415 (2019). https://doi.org/10.1109/ ICCV.2019.00149

36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748-8763. PMLR (2021), https://proceedings.mlr.press/v139/radford21a.html
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等: 从自然语言监督中学习可迁移视觉模型. 载于: 国际机器学习会议. 页8748-8763. PMLR (2021), https://proceedings.mlr.press/v139/radford21a.html

37. Rahman, M.M., Fookes, C., Baktashmotlagh, M., Sridharan, S.: Multi-component image translation for deep domain generalization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 579-588. IEEE (2019). https://doi.org/10.1109/WACV.2019.00067
38. Rahman, M.M., Fookes, C., Baktashmotlagh, M., Sridharan, S.: 用于深度域泛化的多组件图像转换. 载于: 2019年IEEE冬季计算机视觉应用会议 (WACV). 页579-588. IEEE (2019). https://doi.org/10.1109/WACV.2019.00067

38. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821-8831. PMLR (2021), https://proceedings.mlr.press/v139/ramesh21a.html
39. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: 零样本文本到图像生成. 载于: 国际机器学习会议. 页8821-8831. PMLR (2021), https://proceedings.mlr.press/v139/ramesh21a.html

39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684-10695 (2022). https://doi.org/10.1109/CVPR52729.2023.01389
40. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 基于潜在扩散模型的高分辨率图像合成. 载于: IEEE/CVF计算机视觉与模式识别会议论文集. 页10684-10695 (2022). https://doi.org/10.1109/CVPR52729.2023.01389

40. Sanchez, E.H., Serrurier, M., Ortner, M.: Learning disentangled representations via mutual information estimation. In: Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXII 16. pp. 205-221. Springer (2020). https://doi.org/10.1007/978-3-030-58542-6_13
41. Sanchez, E.H., Serrurier, M., Ortner, M.: 通过互信息估计学习解耦表示。载于：计算机视觉-ECCV 2020：第16届欧洲会议，英国格拉斯哥，2020年8月23-28日，论文集，第二十二部分16。第205-221页。施普林格（2020）。https://doi.org/10.1007/978-3-030-58542-6_13

41. Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., Locatello, F.: Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems34, 16451-16467 (2021), https://dl.acm.org/doi/10.5555/3540261.3541519
42. Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., Locatello, F.: 通过数据增强的自监督学习可证明地将内容与风格分离。神经信息处理系统进展34，16451-16467（2021），https://dl.acm.org/doi/10.5555/3540261.3541519

42. Wei, J., Zou, K.: Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6382-6388 (2019). https://doi.org/10.18653/v1/D19-1670
43. Wei, J., Zou, K.: EDA：用于提升文本分类任务性能的简单数据增强技术。载于：2019年自然语言处理实证方法会议暨第九届国际联合自然语言处理会议（EMNLP-IJCNLP）论文集。第6382-6388页（2019）。https://doi.org/10.18653/v1/D19-1670

43. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.:

Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959-7971 (2022). https://doi.org/10.1109/ CVPR52688.2022.00780

44. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., 等: 零样本模型的鲁棒微调。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第7959-7971页（2022）。https://doi.org/10.1109/CVPR52688.2022.00780

44. Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: Causalvae: Disentangled representation learning via neural structural causal models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9593-9602 (2021). https://doi.org/10.1109/CVPR46437.2021.00947
45. Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: CausalVAE：通过神经结构因果模型实现解耦表示学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第9593-9602页（2021）。https://doi.org/10.1109/CVPR46437.2021.00947

45. Yang, W., Mirzasoleiman, B.: Robust contrastive language-image pretraining against adversarial attacks. arXiv preprint arXiv:2303.06854 (2023). https://doi.org/10.48550/arXiv.2303.06854
46. Yang, W., Mirzasoleiman, B.: 针对对抗攻击的鲁棒对比语言-图像预训练。arXiv预印本 arXiv:2303.06854（2023）。https://doi.org/10.48550/arXiv.2303.06854

46. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision (ICCV) (2023). https://doi.org/10.1109/ICCV51070.2023.00355
47. Zhang, L., Rao, A., Agrawala, M.: 为文本到图像扩散模型添加条件控制。载于：IEEE国际计算机视觉会议（ICCV）（2023）。https://doi.org/10.1109/ICCV51070.2023.00355

47. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816-16825 (2022). https://doi.org/10.1109/CVPR52688.2022.01631
48. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: 视觉-语言模型的条件提示学习。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第16816-16825页（2022）。https://doi.org/10.1109/CVPR52688.2022.01631

48. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130** (9), 2337 – 2348 (2022) . https://doi.org/10.1007/s11263-022-01653-1
49. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: 学习为视觉-语言模型设计提示。国际计算机视觉杂志 **130** (9), 2337 – 2348 (2022) 。https://doi.org/10.1007/s11263-022-01653-1

49. Zimmermann, R.S., Sharma, Y., Schneider, S., Bethge, M., Brendel, W.: Contrastive learning inverts the data generating process. In: International Conference on Machine Learning. pp. 12979-12990. PMLR (2021), https://proceedings.mlr.press/v139/zimmermann21a.html
50. Zimmermann, R.S., Sharma, Y., Schneider, S., Bethge, M., Brendel, W.: 对比学习逆转数据生成过程。载于：国际机器学习会议。第12979-12990页。PMLR（2021）， https://proceedings.mlr.press/v139/zimmermann21a.html

CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts APPENDIX

CLAP：通过带增强提示的对比学习实现内容与风格的分离 附录

Yichao Cai , Yuhang Liu , Zhen Zhang , and Javen Qinfeng Shi
蔡一超，刘宇航，张震，施建峰

Australian Institute for Machine Learning, University of Adelaide, SA 5000, Australia
{yichao.cai,yuhang.liu01,zhen.zhang02,javen.shi}@adelaide.edu.au
澳大利亚阿德莱德大学机器学习研究所，南澳5000，澳大利亚 {yichao.cai,yuhang.liu01,zhen.zhang02,javen.shi}@adelaide.edu.au

1 Overview of the Appendix:

2 附录概述：

- More details on experiments using the CLIP pre-trained ViT-B/16 model are provided in Appendix A, including implementation details in Appendix A.1, investigations into prompt augmentation combinations in Appendix A.2, analysis of different training prompt sources in Appendix A.3, and detailed experiment results for each dataset in Appendix A.4.
- 关于使用CLIP预训练的ViT-B/16模型的更多实验细节见附录A，包括附录A.1中的实现细节，附录A.2中的提示增强组合研究，附录A.3中不同训练提示来源的分析，以及附录A.4中各数据集的详细实验结果。

- The processes of data synthesis with large models used in our approach are outlined in Appendix B. The image synthesis procedure for

Im. Aug is detailed in Appendix B.1, and the approach for generating "LLM" prompts, used in analyzing prompt sources, is described in Appendix B.2.

- 我们方法中使用的大模型数据合成过程概述见附录B。Im. Aug的图像合成过程详见附录B.1，分析提示来源时使用的“LLM”提示生成方法见附录B.2。

- In Appendix C, we detail our repeated zero-shot experiments conducted with the CLIP pre-trained ViT-L/14 (Appendix C.1) and ResNet50x16 (Appendix C.2 models.
- 在附录C中，我们详细介绍了使用CLIP预训练的ViT-L/14（附录C.1）和ResNet50x16（附录C.2）模型进行的重复零样本实验。

- In section Appendix D, we present discussions covering the underlying rationale for basing CLAP on the CLIP pre-trained models in Appendix D.1, and the impact of image augmentation and text augmentation in Appendix D.2
- 在附录D部分，我们讨论了基于CLIP预训练模型构建CLAP的基本原理（附录D.1）以及图像增强和文本增强的影响（附录D.2）。

3 A More on Experiments with ViT-B/16

4 A 关于ViT-B/16实验的更多内容

4.1 A.1 Implementation Details

4.2 A.1 实现细节

In this section, we detail the implementation of our experiments utilizing the CLIP pre-trained ViT-B/16 model:
本节详细介绍了我们利用CLIP预训练ViT-B/16模型进行实验的实现细节:

Network. The network's output dimension is aligned with the 512-dimensional CLIP features, thereby obviating the need for input feature downsampling. The latent dimensions are tailored to each dataset: 256 for PACS, 448 for OfficeHome, and 512 for VLCS and DomainNet, to accommodate the variety of categories and complexity of datasets. The weight parameter α is adjusted to 0.208 for PACS, 0.056 for VLCS, 0.14 for OfficeHome, and 0.2 for DomainNet, while it is consistently maintained at 1 throughout the training phase.
网络。网络输出维度与512维CLIP特征保持一致，因此无需对输入特征进行下采样。潜在维度根据数据集调整：PACS为256，OfficeHome为448，VLCS和DomainNet为512，以适应类别多样性和数据集复杂性。权重参数α分别调整为PACS的0.208，VLCS的0.056，OfficeHome的0.14，DomainNet的0.2，训练阶段始终保持为1。

Training CLAP. Training parameters are consistent across datasets, employing the Adam optimizer with a learning rate of 0.0001 , limiting training to 8,000 steps with checking the average loss every 480 steps, and instituting early stopping after five checkpoints without a loss decrease of at least 0.01 . Batch sizes are adjusted to 8 for PACS and VLCS, 96 for OfficeHome, and 384 for Domain-Net,with the temperature parameter τ set at 0.5 for PACS and VLCS,and 0.3 for OfficeHome and DomainNet. The loss coefficient λ is set to 1 for PACS and VLCS, and 0.0001 for OfficeHome and DomainNet, due to the first two datasets have less classes. Prompt augmentations, OSD+OCD+SPO, are applied across datasets all with a 0.5 probability. For the PACS and VLCS datasets, Gaussian noise with a zero mean and a standard deviation of 0.02 is randomly inserted at the beginning, middle, or end of the augmented-view prompts to enrich the training samples. In the linear probe evaluations for few-shot analysis, L2 normalization and cross-entropy loss are utilized for training over 1,000 epochs with a batch size of 32 , incorporating early stopping with a patience threshold of 10 epochs and a loss decrease criterion of 0.001 .
CLAP训练。训练参数在各数据集间保持一致，采用Adam优化器，学习率为0.0001，训练限制为8000步，每480步检查一次平均损失，若连续五次检查损失未下降至少0.01则提前停止。批量大小为PACS和VLCS设为8，OfficeHome为96，DomainNet为384，温度参数τ在PACS和VLCS为0.5，OfficeHome和DomainNet为0.3。损失系数λ在PACS和VLCS为1，OfficeHome和DomainNet为0.0001，因前两者类别较少。提示增强采用OSD+OCD+SPO，概率均为0.5。对PACS和VLCS数据集，在增强视图提示的开头、中间或结尾随机插入均值为0、标准差为0.02的高斯噪声以丰富训练样本。少样本线性探针评估中，使用L2归一化和交叉熵损失，训练1000个epoch，批量大小32，采用早停策略，耐心值为10个epoch，损失下降阈值为0.001。

Training Im. Aug. We train a disentangled network using image augmentation, applying the InfoNCE loss with a temperature parameter τ set to 0.5 . This include image augmentation techniques,image cropping (scale $\in [0.64, 1.0]$) and color distortion (brightness $= 0.5$,hue $= 0.3$),each with a probability of 0.5 . Other training and inference configurations for Im. Aug are consistent with those used for CLAP across all datasets.
Im. Aug训练。我们训练了一个使用图像增强的解耦网络，采用InfoNCE损失，温度参数τ设为0.5。包括图像增强技术、图像裁剪（比例$\in [0.64, 1.0]$）和颜色失真（亮度$= 0.5$，色调$= 0.3$），每项概率均为0.5。Im. Aug的其他训练和推理配置与CLAP在所有数据集上的设置一致。

4.3 A.2 Prompt Augmentation Combinations

4.4 A.2 提示增强组合

In Tab. 1, we explore different combinations of our tailored prompt augmentation techniques and EDA (Easy Data Augmentation) [42] techniques on the VLCS dataset. Each combination demonstrates CLAP's effectiveness in enhancing CLIP's performance and reducing performance disparities. The combination of OSD+OCS+SPO+IGN achieves the highest average accuracy and the least variance, outperforming the EDA techniques. Notably, even without incorporating random noise in the augmentations, CLAP significantly surpasses CLIP in handling perturbations on prompts,as evidenced by the largely reduced $\Delta_{(NC)}$.

在表1中，我们探讨了在VLCS数据集上不同定制提示增强技术与EDA（Easy Data Augmentation）[42]技术的组合。每种组合均展示了CLAP提升CLIP性能和减少性能差异的效果。OSD+OCS+SPO+IGN组合实现了最高的平均准确率和最小的方差，优于EDA技术。值得注意的是，即使不加入随机噪声，CLAP在处理提示扰动方面也显著优于CLIP，表现为$\Delta_{(NC)}$大幅降低。

4.5 A.3 Prompt Sources

4.6 A.3 提示来源

In Tab. 2, we examine the effects of various training prompt formats, sourced from different synthetic origins, on the VLCS dataset performance, utilizing
在表2中，我们考察了不同合成来源的训练提示格式对VLCS数据集性能的影响，利用

EDA techniques. The prompt formats are defined as follows: "Template" refers to the template-based prompts fundamental to our primary approach; "LLM" designates prompts created by ChatGPT-3.5 [3], with the generation process elaborated in Appendix B.2: "Random" describes prompts formatted as "a [random] style of [class]," with "[random]" being replaced by terms from a random word generator; and "Prm.Stl." indicates vectorized prompts generated through PromptStyler [9].
EDA技术。提示格式定义如下："Template"指基于模板的提示，是我们主要方法的基础；"LLM"表示由ChatGPT-3.5 [3]生成的提示，生成过程详见附录B.2："Random"描述格式为"a [random] style of [class]"的提示，其中"[random]"由随机词生成器替换；"Prm.Stl."表示通过PromptStyler [9]生成的向量化提示。

Table 1: We evaluate prompt augmentation combinations on the VLCS dataset: OSD (①), OCD (②), ITD (③), ASD (④), SPO (⑤), and IGN (⑥). ZS(Avg.) shows average zero-shot accuracy acoss four distinct inference prompts. CLAP boosts CLIP's accuracy and reduces variances, with ①②⑤⑥ as the optimal combination.
表1：我们在VLCS数据集上评估提示增强组合：OSD（①）、OCD（②）、ITD（③）、ASD（④）、SPO（⑤）和IGN（⑥）。ZS(Avg.)显示四种不同推理提示的平均零样本准确率。CLAP提升了CLIP的准确率并减少了方差，①②⑤⑥为最佳组合。

Metrics	CLIP (base)	EDA	①②③④⑤⑥	①②③④⑤	Avg. top-1 acc. (%) of different augmentations ①②③ ④⑥	①②③④	③④⑤⑥	①②⑤⑥
ZS(Avg.) (↑)	77.3	81.6	82.0	80.1	82.0	79.6	82.1	82.6
$R\left(\downarrow \right)$	6.1	1.9	1.2	2.5	0.9	3.2	1.6	0.8
$\delta \left(\downarrow \right)$	2.8	0.9	0.6	1.2	0.4	1.5	0.7	0.4
${\Delta }_{\left(NC\right) }\left(\downarrow \right)$	8.1	2.3	1.7	3.0	1.8	3.4	2.0	1.6

指标	CLIP（基础版）	EDA	①②③④⑤⑥	①②③④⑤	不同增强方法的平均Top-1准确率（%）①②③ ④⑥	①②③④	③④⑤⑥	①②⑤⑥
零样本（平均）（↑）	77.3	81.6	82.0	80.1	82.0	79.6	82.1	82.6
$R\left(\downarrow \right)$	6.1	1.9	1.2	2.5	0.9	3.2	1.6	0.8
$\delta \left(\downarrow \right)$	2.8	0.9	0.6	1.2	0.4	1.5	0.7	0.4
${\Delta }_{\left(NC\right) }\left(\downarrow \right)$	8.1	2.3	1.7	3.0	1.8	3.4	2.0	1.6

Table 2: We employ EDA augmentation to train CLAP with diverse prompt sources on the VLCS dataset. Each prompt source contributes to improvements in CLIP's zero-shot performance, with "Random" and "Template" prompts, in their simpler forms, yielding better outcomes.
表2：我们采用EDA增强方法，在VLCS数据集上使用多样的提示源训练CLAP。每种提示源都提升了CLIP的零样本性能，其中"随机"和"模板"提示以其较简单的形式表现更佳。

Metrics	CLIP (base)	LLM	Random	Prm.Stl.	Template
ZS(Avg.) (↑)	77.3	78.2	81.6	81.2	81.6
$R\left(\downarrow \right)$	6.1	3.2	0.7	2.7	1.9
$\delta \left(\downarrow \right)$	2.8	1.5	0.3	1.2	0.9
${\Delta }_{\left(NC\right) }\left(\downarrow \right)$	8.1	3.3	2.3	3.0	2.3

指标	CLIP（基础版）	大型语言模型（LLM）	随机	参数风格	模板
零样本（平均）（↑）	77.3	78.2	81.6	81.2	81.6
$R\left(\downarrow \right)$	6.1	3.2	0.7	2.7	1.9
$\delta \left(\downarrow \right)$	2.8	1.5	0.3	1.2	0.9
${\Delta }_{\left(NC\right) }\left(\downarrow \right)$	8.1	3.3	2.3	3.0	2.3

Our experimental results indicate that CLAP, when trained across these varied prompt formats, enhances the performance of CLIP. Notably, despite the complex generation mechanisms of "LLM" and "Prm.Stl." prompts, the simpler, random-styled and template-based prompts demonstrate superior efficacy. However, it is important to highlight that the improvements attributed to these diverse prompt formats, trained with EDA, do not surpass the best performance of the prompt augmentations tailored for template-based prompts.
我们的实验结果表明，CLAP在跨越这些多样化提示格式的训练中，提升了CLIP的性能。值得注意的是，尽管“LLM”和“Prm.Stl.”提示的生成机制较为复杂，但更简单的随机风格和基于模板的提示表现出更优的效果。然而，需要强调的是，这些通过EDA训练的多样化提示格式所带来的改进，并未超过针对基于模板提示的提示增强所达到的最佳性能。

4.7 A.4 Detailed Results on ViT-B/16

4.8 A.4 ViT-B/16的详细结果

Details on Zero-Shot Evaluations We present the domain-level zero-shot performance with various prompts across each dataset in Tab. 3. CLAP consistently enhances CLIP's zero-shot performance across these different prompts. Given that CLAP exclusively utilizes text data for training, it does not compromise CLIP's inherent ability to generalize across domains, which is acquired from its extensive training dataset. Rather, by achieving a more effective disentanglement of content, it unequivocally enhances CLIP's zero-shot performance across all dataset domains.
关于零样本评估的详细信息 我们在表3中展示了各数据集上使用不同提示的领域级零样本性能。CLAP在这些不同提示下持续提升了CLIP的零样本性能。鉴于CLAP仅利用文本数据进行训练，它并未削弱CLIP从其大规模训练数据中获得的跨领域泛化能力。相反，通过实现更有效的内容解耦，CLAP明确提升了CLIP在所有数据集领域的零样本性能。

Table 3: Domain-level zero-shot results of the ViT-B/16 model on the test datasets.
表3：ViT-B/16模型在测试数据集上的领域级零样本结果。

		Domain-level avg. top-1 acc. (%) of zero-shot performance usig ViT-B/16 (↑)											
Dataset Domains		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	Im. Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	96.4	96.9	97.5	93.4	97.0	97.6	97.4	97.6	97.6	87.8	93.5	97.1
	C.	98.9	99.0	98.9	99.0	99.2	99.0	99.1	99.0	98.9	95.4	97.6	98.8
	P.	99.9	99.9	99.9	99.3	99.6	99.9	99.9	99.9	99.9	93.1	99.0	99.9
	S.	87.7	90.1	92.5	89.2	89.6	92.5	88.1	89.4	92.3	87.1	89.3	93.1
VLCS	C.	99.7	99.8	99.9	99.9	99.9	99.9	99.9	99.9	99.9	87.0	96.0	99.9
	L.	61.8	66.2	67.7	69.9	70.4	70.4	70.2	70.2	70.7	55.9	59.9	65.9
	S.	70.1	74.8	78.0	73.3	76.0	77.2	73.6	76.4	76.9	61.4	66.2	75.3
	V.	73.9	77.1	84.9	84.8	85.4	86.0	86.1	85.6	86.2	68.9	70.3	82.9
OfficeHome	A.	80.5	79.0	81.8	80.1	76.0	81.6	83.2	78.7	83.2	73.0	69.2	73.6
	C.	64.6	59.6	66.4	63.7	58.9	65.4	68.1	61.9	69.0	57.0	52.0	60.4
	P.	86.3	83.6	87.5	86.6	83.4	87.2	89.1	86.6	89.7	77.2	72.3	78.9
	R.	88.0	85.9	88.5	87.6	84.8	87.7	89.8	87.2	90.0	79.0	76.5	81.1
DomainNet	C.	71.0	64.3	71.9	70.5	62.1	72.0	71.3	63.4	72.8	63.2	53.9	64.6
	I.	48.6	40.5	50.6	47.7	40.7	49.5	47.8	40.0	50.5	42.9	35.0	45.1
	P.	66.6	59.1	67.7	66.0	59.0	67.3	66.5	59.8	68.4	57.2	50.4	59.4
	Q.	14.9	12.4	15.2	13.3	11.5	13.8	14.1	11.8	14.3	12.0	9.2	13.1
	R.	82.6	76.6	83.1	82.2	75.8	82.2	83.4	78.2	83.7	75.2	67.9	75.6
	S.	63.1	56.1	63.7	62.2	55.0	63.1	63.4	56.4	64.4	55.7	47.5	57.6

数据集	领域	使用ViT-B/16的零样本性能领域级平均Top-1准确率（%）（↑）											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP
PACS	A.	96.4	96.9	97.5	93.4	97.0	97.6	97.4	97.6	97.6	87.8	93.5	97.1
	C.	98.9	99.0	98.9	99.0	99.2	99.0	99.1	99.0	98.9	95.4	97.6	98.8
	P.	99.9	99.9	99.9	99.3	99.6	99.9	99.9	99.9	99.9	93.1	99.0	99.9
	S.	87.7	90.1	92.5	89.2	89.6	92.5	88.1	89.4	92.3	87.1	89.3	93.1
VLCS	C.	99.7	99.8	99.9	99.9	99.9	99.9	99.9	99.9	99.9	87.0	96.0	99.9
	L.	61.8	66.2	67.7	69.9	70.4	70.4	70.2	70.2	70.7	55.9	59.9	65.9
	S.	70.1	74.8	78.0	73.3	76.0	77.2	73.6	76.4	76.9	61.4	66.2	75.3
	V.	73.9	77.1	84.9	84.8	85.4	86.0	86.1	85.6	86.2	68.9	70.3	82.9
OfficeHome	A.	80.5	79.0	81.8	80.1	76.0	81.6	83.2	78.7	83.2	73.0	69.2	73.6
	C.	64.6	59.6	66.4	63.7	58.9	65.4	68.1	61.9	69.0	57.0	52.0	60.4
	P.	86.3	83.6	87.5	86.6	83.4	87.2	89.1	86.6	89.7	77.2	72.3	78.9
	R.	88.0	85.9	88.5	87.6	84.8	87.7	89.8	87.2	90.0	79.0	76.5	81.1
DomainNet	C.	71.0	64.3	71.9	70.5	62.1	72.0	71.3	63.4	72.8	63.2	53.9	64.6
	I.	48.6	40.5	50.6	47.7	40.7	49.5	47.8	40.0	50.5	42.9	35.0	45.1
	P.	66.6	59.1	67.7	66.0	59.0	67.3	66.5	59.8	68.4	57.2	50.4	59.4
	Q.	14.9	12.4	15.2	13.3	11.5	13.8	14.1	11.8	14.3	12.0	9.2	13.1
	R.	82.6	76.6	83.1	82.2	75.8	82.2	83.4	78.2	83.7	75.2	67.9	75.6
	S.	63.1	56.1	63.7	62.2	55.0	63.1	63.4	56.4	64.4	55.7	47.5	57.6

Details on Few-Shot Evaluations We display the quantitative results of few-shot performance in Tab. 4. CLAP consistently enhances the few-shot capabilities, showcasing improvements across test datasets at a closer domain level.
关于少样本评估的详细信息 我们在表4中展示了少样本性能的量化结果。CLAP持续提升了少样本能力，在更接近领域层面的测试数据集上均表现出改进。

Details on Adversarial Evaluations In Tab. 5, we detail our adversarial performance evaluations for PACS, VLCS, OfficeHome, and DomainNet, respectively. CLAP enhances both zero-shot and one-shot performance across all domains of the tested datasets. While Im.Aug boosts one-shot robustness against adversarial tasks, its impact on zero-shot adversarial robustness is inconsistent.
关于对抗性评估的详细信息 在表5中，我们分别详细介绍了PACS、VLCS、OfficeHome和DomainNet的数据集上的对抗性性能评估。CLAP提升了所有测试数据集领域的零样本和一样本性能。虽然Im.Aug增强了一样本对抗任务的鲁棒性，但其对零样本对抗鲁棒性的影响不稳定。

Details on Ablative Analysis In Tab. 6, we provide detailed results from our analysis on zero-shot performance using various combinations of prompt augmentations. Additionally, in Tab. 7, we present the outcomes of our ablative studies focusing on the hyperparameters τ ,latent dimension,and α ,respectively, each evaluated domain-wise. The results indicate that CLAP is effective across a wide range of hyperparameters.
关于消融分析的详细信息 在表6中，我们提供了使用不同提示增强组合进行零样本性能分析的详细结果。此外，在表7中，我们展示了针对超参数τ、潜在维度和α的消融研究结果，均按领域分别评估。结果表明CLAP在广泛的超参数范围内均表现有效。

5 B Data Synthesis

6 B 数据合成

6.1 B.1 Synthetic Image Generation

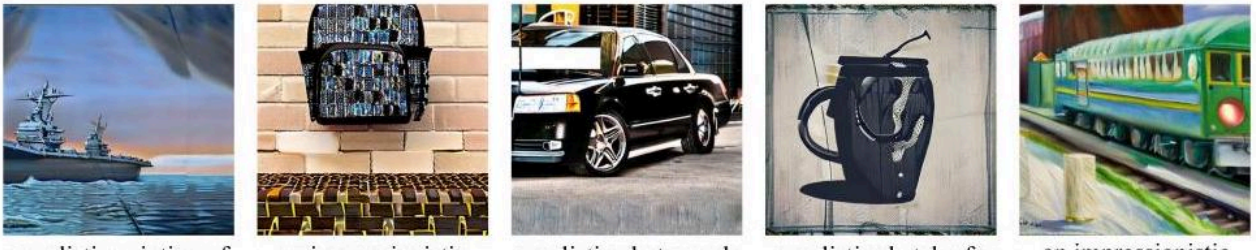
6.2 B.1 合成图像生成

We employ the stable diffusion [39] v2.1 model for generating synthetic images used in our comparing experiments, specifically utilizing the Stable Diffusion
我们采用稳定扩散（Stable Diffusion）[39] v2.1模型生成用于对比实验的合成图像，具体使用了Stable Diffusion

Table 4: Domain-level few-shot results of the ViT-B/16 model using the test datasets.
表4：ViT-B/16模型在测试数据集上的领域级少样本结果。

Dataset	Domains	Domain-level avg. top-1 acc. (%) of few-shot performance of ViT-B/16 (†)															
		1-shot			4-shot			8-shot			16-shot			32-shot		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	79.5	84.1	94.5	92.4	96.4	97.2	95.1	97.2	98.4	97.9	98.1	98.4	98.8	99.1	98.9
	C.	86.7	96.1	98.3	96.8	98.6	99.2	98.8	98.9	99.3	99.5	99.2	99.5	99.6	99.6	99.6
	P.	97.4	99.8	99.9	99.6	99.8	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9	99.9
	S.	75.1	80.0	87.3	91.1	92.3	92.5	92.3	92.3	92.9	92.4	92.6	93.1	93.9	94.2	94.1
VLCS	C.	99.2	99.7	99.8	99.9	99.8	99.9	99.8	99.7	99.9	99.7	99.9	99.9	99.9	100.0	99.9
	L.	41.3	41.3	41.1	56.7	57.0	59.8	46.2	36.8	48.3	59.4	60.4	62.6	60.4	60.7	61.9
	S.	45.3	46.1	50.8	61.9	63.7	69.0	67.4	67.7	71.3	75.9	76.8	80.9	77.4	78.6	81.0
	V.	50.9	53.4	59.0	64.5	66.7	76.1	75.4	74.1	78.7	72.6	73.9	77.7	85.7	86.1	87.9
OfficeHome	A.	42.6	45.1	43.9	76.8	77.6	77.7	84.8	86.0	85.5	91.8	92.1	92.1	97.4	97.5	97.5
	C.	40.1	45.0	43.8	69.9	70.2	70.5	75.8	75.9	76.6	81.6	81.6	81.6	89.0	89.0	89.2
	P.	70.2	73.3	73.4	89.7	90.3	90.2	93.8	93.7	93.9	95.7	95.7	95.8	97.7	97.6	97.6
	R.	58.4	59.3	59.4	81.7	83.1	82.9	89.7	89.5	89.9	92.9	92.7	93.2	95.8	95.8	95.8
DomainNet	C.	42.1	43.6	43.8	66.8	67.5	67.8	74.2	74.3	74.6	78.5	78.6	78.8	82.8	82.8	82.7
	L	19.5	20.8	21.0	38.5	39.3	39.7	46.7	47.0	47.3	53.2	53.2	53.6	60.0	59.9	60.1
	P.	32.1	33.5	34.2	60.5	60.9	61.5	68.0	68.0	68.7	72.5	72.6	73.0	76.7	76.6	76.8
	Q.	15.2	15.3	15.3	30.0	29.6	29.9	37.1	36.4	36.8	43.8	43.4	43.5	49.4	49.1	49.0
	R.	50.8	52.1	52.7	76.7	77.0	77.6	81.7	81.9	82.2	84.0	83.9	84.3	85.9	85.9	86.0
	S.	33.1	33.9	34.8	56.2	56.6	57.2	62.9	62.9	63.7	67.8	67.7	68.1	72.5	72.3	72.6

数据集	领域	ViT-B/16（†）少样本性能的领域级平均top-1准确率（%）															
		1次样本			4次样本			8次样本			16次样本			32次样本		
		CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP
PACS	A.	79.5	84.1	94.5	92.4	96.4	97.2	95.1	97.2	98.4	97.9	98.1	98.4	98.8	99.1	98.9
	C.	86.7	96.1	98.3	96.8	98.6	99.2	98.8	98.9	99.3	99.5	99.2	99.5	99.6	99.6	99.6
	P.	97.4	99.8	99.9	99.6	99.8	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9	99.9
	S.	75.1	80.0	87.3	91.1	92.3	92.5	92.3	92.3	92.9	92.4	92.6	93.1	93.9	94.2	94.1
VLCS	C.	99.2	99.7	99.8	99.9	99.8	99.9	99.8	99.7	99.9	99.7	99.9	99.9	99.9	100.0	99.9
	L.	41.3	41.3	41.1	56.7	57.0	59.8	46.2	36.8	48.3	59.4	60.4	62.6	60.4	60.7	61.9
	S.	45.3	46.1	50.8	61.9	63.7	69.0	67.4	67.7	71.3	75.9	76.8	80.9	77.4	78.6	81.0
	V.	50.9	53.4	59.0	64.5	66.7	76.1	75.4	74.1	78.7	72.6	73.9	77.7	85.7	86.1	87.9
OfficeHome	A.	42.6	45.1	43.9	76.8	77.6	77.7	84.8	86.0	85.5	91.8	92.1	92.1	97.4	97.5	97.5
	C.	40.1	45.0	43.8	69.9	70.2	70.5	75.8	75.9	76.6	81.6	81.6	81.6	89.0	89.0	89.2
	P.	70.2	73.3	73.4	89.7	90.3	90.2	93.8	93.7	93.9	95.7	95.7	95.8	97.7	97.6	97.6
	R.	58.4	59.3	59.4	81.7	83.1	82.9	89.7	89.5	89.9	92.9	92.7	93.2	95.8	95.8	95.8
DomainNet	C.	42.1	43.6	43.8	66.8	67.5	67.8	74.2	74.3	74.6	78.5	78.6	78.8	82.8	82.8	82.7
	L	19.5	20.8	21.0	38.5	39.3	39.7	46.7	47.0	47.3	53.2	53.2	53.6	60.0	59.9	60.1
	P.	32.1	33.5	34.2	60.5	60.9	61.5	68.0	68.0	68.7	72.5	72.6	73.0	76.7	76.6	76.8
	Q.	15.2	15.3	15.3	30.0	29.6	29.9	37.1	36.4	36.8	43.8	43.4	43.5	49.4	49.1	49.0
	R.	50.8	52.1	52.7	76.7	77.0	77.6	81.7	81.9	82.2	84.0	83.9	84.3	85.9	85.9	86.0
	S.	33.1	33.9	34.8	56.2	56.6	57.2	62.9	62.9	63.7	67.8	67.7	68.1	72.5	72.3	72.6



a realistic painting of a large blue aircraft carrier

an impressionistic mosaic art of a small black backpack

a realistic photograph of a large black car

a realistic sketch of a large black mug

an impressionistic painting of a normal sized green train

Fig. 1: Examples of synthetic images created with SDv2.1 and associated prompts.
图1：使用SDv2.1及相关提示词创建的合成图像示例。

v2-1 Model Card available on Hugging Face [1] . For each class across the four datasets, we produce 480 images using our synthetic template prompts as input for the stable diffusion model. All generated images are of 512×512 resolution. Examples of these synthetic images alongside their corresponding text prompts are displayed in Fig. 1.
v2-1模型卡可在Hugging Face [1] 获取。对于四个数据集中的每个类别，我们使用合成模板提示词作为稳定扩散模型（stable diffusion model）的输入，生成480张图像。所有生成的图像分辨率均为512×512。这些合成图像及其对应的文本提示词示例展示于图1中。

6.3 B.2 LLM Prompts Generation

6.4 B.2 大型语言模型（LLM）提示词生成

We utilize ChatGPT-3.5 [3] to create the LLM prompts employed in our comparative analysis of different prompt sources. Fig. 2 illustrates the process of prompting ChatGPT-3.5 to generate text prompts for specific class names. For each class, we produce 120 samples, and below are a few examples from the generated prompts:
我们利用ChatGPT-3.5 [3] 创建用于不同提示词来源比较分析的大型语言模型提示词。图2展示了提示ChatGPT-3.5生成特定类别名称文本提示词的过程。对于每个类别，我们生成120个样本，以下是部分生成提示词示例：

- Bird:
- 鸟类：

---

https://huggingface.co/stabilityai/stable-diffusion-2-1

Table 5: Domain-level results under adversarial attacks of ViT-B/16 on the datasets.
表5：ViT-B/16模型在各数据集上的领域级对抗攻击结果。

Dataset	Domains	Domain-level avg. top-1 acc. (%) under adversarial attackings using ViT-B/16 (†)																	
		FGSM						PGD-20						CW-20					
		ZS-C			1-shot			ZS-C			1-shot			ZS-C			1-shot		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	76.3	79.3	79.3	61.2	78.0	87.3	1.7	2.2	1.8	16.0	42.1	63.1	1.5	2.0	2.3	0.5	1.1	1.7
	C.	94.9	95.0	94.0	66.5	84.2	95.1	33.3	37.7	35.6	33.3	57.2	86.1	28.8	34.0	33.2	11.9	23.6	31.8
	P.	91.6	90.3	91.7	67.4	80.8	92.1	5.7	7.0	6.7	27.1	55.0	69.8	4.7	4.9	5.8	0.7	2.7	4.1
	S.	84.5	87.5	89.8	71.6	74.6	83.8	75.8	78.4	79.2	63.0	66.3	74.6	74.5	76.8	77.9	62.7	65.4	70.3
VLCS	C.	55.3	53.8	55.5	25.8	28.8	25.3	4.4	5.1	4.7	2.0	5.2	2.5	2.9	3.1	3.5	0.7	1.2	1.0
	L.	49.4	45.5	50.6	27.0	32.6	30.4	15.2	14.9	16.0	6.4	8.9	8.0	12.4	11.2	13.0	6.1	8.3	7.7
	S.	61.7	58.1	62.5	48.0	46.9	51.6	13.2	13.9	14.0	8.6	10.7	10.0	9.2	8.8	10.2	8.3	7.9	8.4
	V.	65.3	63.2	65.6	36.5	40.1	41.0	7.5	7.9	7.9	5.3	9.4	8.9	5.2	4.8	5.6	2.9	2.8	2.9
OfficeHome	A.	55.3	53.8	55.5	25.8	28.8	25.3	4.4	5.1	4.7	2.0	5.2	2.5	2.9	3.1	3.5	0.7	1.2	1.0
	C.	49.4	45.5	50.6	27.0	32.6	30.4	15.2	14.9	16.0	6.4	8.9	8.0	12.4	11.2	13.0	6.1	8.3	7.7
	P.	61.7	58.1	62.5	48.0	46.9	51.6	13.2	13.9	14.0	8.6	10.7	10.0	9.2	8.8	10.2	8.3	7.9	8.4
	R.	65.3	63.2	65.6	36.5	40.1	41.0	7.5	7.9	7.9	5.3	9.4	8.9	5.2	4.8	5.6	2.9	2.8	2.9
DomainNet	C.	57.8	50.9	58.8	33.3	34.3	35.0	21.6	18.7	22.8	18.4	19.6	20.0	15.8	12.5	16.6	7.0	7.5	7.8
	I.	35.8	28.0	37.0	12.2	13.3	13.2	6.1	3.7	6.7	4.6	5.3	5.1	3.3	1.9	3.7	0.9	0.9	0.9
	P.	43.9	39.0	44.3	18.4	20.6	20.3	3.1	2.8	3.3	8.6	10.4	9.9	1.8	1.3	1.9	0.3	0.3	0.3
	Q.	12.9	10.3	13.2	10.9	10.8	11.1	8.4	6.8	8.6	5.4	5.4	5.6	7.1	5.4	7.4	4.9	4.8	5.1
	R.	62.1	55.9	62.4	34.5	35.9	36.5	7.1	6.5	7.5	17.6	19.7	19.6	4.5	3.4	4.7	1.2	1.4	1.4
	S.	49.1	43.3	49.7	25.7	26.0	27.5	17.8	15.5	18.6	13.6	14.4	15.1	13.4	10.2	13.9	5.0	5.2	5.6

数据集领域		使用ViT-B/16 (†)在对抗攻击下的领域级平均Top-1准确率（%）																	
		FGSM						PGD-20						CW-20					
		零样本分类			一次性学习			零样本分类			一次性学习			零样本分类			一次性学习		
			CLIP图像增强CLAP		CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP
PACS	A.	76.3	79.3	79.3	61.2	78.0	87.3	1.7	2.2	1.8	16.0	42.1	63.1	1.5	2.0	2.3	0.5	1.1	1.7
	C.	94.9	95.0	94.0	66.5	84.2	95.1	33.3	37.7	35.6	33.3	57.2	86.1	28.8	34.0	33.2	11.9	23.6	31.8
	P.	91.6	90.3	91.7	67.4	80.8	92.1	5.7	7.0	6.7	27.1	55.0	69.8	4.7	4.9	5.8	0.7	2.7	4.1
	S.	84.5	87.5	89.8	71.6	74.6	83.8	75.8	78.4	79.2	63.0	66.3	74.6	74.5	76.8	77.9	62.7	65.4	70.3
VLCS	C.	55.3	53.8	55.5	25.8	28.8	25.3	4.4	5.1	4.7	2.0	5.2	2.5	2.9	3.1	3.5	0.7	1.2	1.0
	L.	49.4	45.5	50.6	27.0	32.6	30.4	15.2	14.9	16.0	6.4	8.9	8.0	12.4	11.2	13.0	6.1	8.3	7.7
	S.	61.7	58.1	62.5	48.0	46.9	51.6	13.2	13.9	14.0	8.6	10.7	10.0	9.2	8.8	10.2	8.3	7.9	8.4
	V.	65.3	63.2	65.6	36.5	40.1	41.0	7.5	7.9	7.9	5.3	9.4	8.9	5.2	4.8	5.6	2.9	2.8	2.9
OfficeHome	A.	55.3	53.8	55.5	25.8	28.8	25.3	4.4	5.1	4.7	2.0	5.2	2.5	2.9	3.1	3.5	0.7	1.2	1.0
	C.	49.4	45.5	50.6	27.0	32.6	30.4	15.2	14.9	16.0	6.4	8.9	8.0	12.4	11.2	13.0	6.1	8.3	7.7
	P.	61.7	58.1	62.5	48.0	46.9	51.6	13.2	13.9	14.0	8.6	10.7	10.0	9.2	8.8	10.2	8.3	7.9	8.4
	R.	65.3	63.2	65.6	36.5	40.1	41.0	7.5	7.9	7.9	5.3	9.4	8.9	5.2	4.8	5.6	2.9	2.8	2.9
DomainNet	C.	57.8	50.9	58.8	33.3	34.3	35.0	21.6	18.7	22.8	18.4	19.6	20.0	15.8	12.5	16.6	7.0	7.5	7.8
	I.	35.8	28.0	37.0	12.2	13.3	13.2	6.1	3.7	6.7	4.6	5.3	5.1	3.3	1.9	3.7	0.9	0.9	0.9
	P.	43.9	39.0	44.3	18.4	20.6	20.3	3.1	2.8	3.3	8.6	10.4	9.9	1.8	1.3	1.9	0.3	0.3	0.3
	Q.	12.9	10.3	13.2	10.9	10.8	11.1	8.4	6.8	8.6	5.4	5.4	5.6	7.1	5.4	7.4	4.9	4.8	5.1
	R.	62.1	55.9	62.4	34.5	35.9	36.5	7.1	6.5	7.5	17.6	19.7	19.6	4.5	3.4	4.7	1.2	1.4	1.4
	S.	49.1	43.3	49.7	25.7	26.0	27.5	17.8	15.5	18.6	13.6	14.4	15.1	13.4	10.2	13.9	5.0	5.2	5.6

- A pair of vibrant macaws converse in a lush, tropical rainforest, depicted in a lively, exotic wildlife painting.
- 一对色彩鲜艳的金刚鹦鹉在郁郁葱葱的热带雨林中交谈，描绘于一幅生动、异域风情的野生动物画作中。

- A solitary eagle watches over a vast, rugged canyon at sunrise, portrayed in a majestic, wilderness landscape photograph.
- 一只孤独的鹰在日出时分俯瞰广袤崎岖的峡谷，呈现于一幅雄伟的荒野风景摄影作品中。

- Dog:
- 狗：

- A sleek Whippet races in a competitive dog track, illustrated in a fast-paced, dynamic sports style.
- 一只矫健的惠比特犬在竞速赛道上飞奔，描绘于一幅节奏明快、充满动感的体育风格画作中。

- A sturdy and reliable English Bulldog watching over a small shop, its solid presence reassuring to the owner.
- 一只结实可靠的英国斗牛犬守护着一家小店，其稳重的存在令店主安心。

- Car:
- 车：

- A quirky art car parades through the streets in a colorful festival, captured in a fun, expressive style illustration.
- 一辆古怪的艺术车在街头节日游行中穿行，捕捉于一幅有趣且富有表现力的插画风格中。

- A high-tech, autonomous car maneuvers through a smart city environment, portrayed in a futuristic, sci-fi digital art piece.
- 一辆高科技的自动驾驶汽车在智能城市环境中穿梭，呈现于一幅未来感十足的科幻数字艺术作品中。

- Chair:
- 椅子：

- A folding chair at an outdoor wedding, elegantly decorated and part of a beautiful ceremony.
- 一把户外婚礼用的折叠椅，优雅装饰，是美丽仪式的一部分。

- A high-end executive chair in a law firm, projecting authority and professionalism.
- 一把高端律师事务所的行政椅，彰显权威与专业。

- Person:
- 人：

- An energetic coach motivates a team on a sports field, illustrated in an inspiring, leadership-focused painting.
- 一位充满活力的教练在运动场上激励团队，描绘于一幅鼓舞人心、聚焦领导力的画作中。

- A graceful figure skater glides across an ice rink, captured in a delicate, winter-themed pastel drawing.
- 一位优雅的花样滑冰选手在冰场上滑行，捕捉于一幅细腻的冬季主题粉彩画中。

Table 6: Zero-Shot Performance on VLCS Dataset Across Varied Augmentation Combinations and Prompt Sources: ① Random Object Size Deletion, ② Random Object Color Deletion, ③ Random Image Type Deletion, ④ Random Art Style Deletion, ⑤ Random Swapping Order, ⑥ Addition of Gaussian Noise.
表6：VLCS数据集在多种增强组合和提示源下的零样本性能：①随机物体大小删除，②随机物体颜色删除，③随机图像类型删除，④随机艺术风格删除，⑤随机交换顺序，⑥添加高斯噪声。

Method Domains		Avg. top-1 acc. (%) (↑) of different augmentations and prompts on VLCS										
		CLIP (base)	①②③	④⑤⑥	①②③	④⑤	①②③	④⑥	①②③	④	③④⑤	⑥
ZS(C)	C.	99.7	99.9	99.8	99.9	99.8	99.9	99.9	97.9	99.7	99.9	99.9
	L.	61.8	66.6	62.3	67.0	62.2	66.2	67.7	66.2	69.0	67.3	66.5
	S.	70.1	78.1	75.5	78.0	74.3	78.5	78.0	73.2	76.9	73.5	76.9
	V.	73.9	82.8	80.6	83.2	79.3	82.7	84.9	72.6	81.8	81.8	81.9
ZS(CP)	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9
	L.	69.9	69.3	67.9	69.6	68.4	70.0	70.4	69.3	70.4	71.2	69.7
	S.	73.3	77.6	76.4	76.7	75.9	78.8	77.2	76.2	75.2	75.1	78.0
	V.	84.8	85.3	84.0	85.3	84.2	85.1	86.0	77.0	84.2	86.0	84.6
ZS(PC)	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
	L.	70.2	70.0	68.0	70.1	68.5	70.0	70.7	67.5	70.6	71.8	70.0
	S.	73.6	76.6	75.6	76.0	74.8	77.8	76.9	76.9	75.1	74.9	78.2
	V.	86.1	85.7	84.7	85.7	84.5	85.5	86.2	78.2	84.6	86.8	84.8
ZS(NC)	C.	87.0	99.8	99.6	99.8	99.4	99.7	99.9	95.3	98.6	99.6	99.8
	L.	55.9	65.2	61.3	65.6	60.5	65.4	65.9	63.0	66.7	64.0	64.7
	S.	61.4	75.6	70.3	75.2	68.3	74.9	75.3	68.9	73.3	69.8	73.0
	V.	68.9	80.1	75.2	80.4	73.8	79.4	82.9	69.3	79.6	77.2	78.6

方法领域		VLCS上不同增强和提示的平均Top-1准确率(%)（↑）										
		CLIP（基础版）	①②③	④⑤⑥	①②③	④⑤	①②③	④⑥	①②③	④	③④⑤	⑥
零样本（C）	C.	99.7	99.9	99.8	99.9	99.8	99.9	99.9	97.9	99.7	99.9	99.9
	L.	61.8	66.6	62.3	67.0	62.2	66.2	67.7	66.2	69.0	67.3	66.5
	S.	70.1	78.1	75.5	78.0	74.3	78.5	78.0	73.2	76.9	73.5	76.9
	V.	73.9	82.8	80.6	83.2	79.3	82.7	84.9	72.6	81.8	81.8	81.9
零样本（CP）	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9
	L.	69.9	69.3	67.9	69.6	68.4	70.0	70.4	69.3	70.4	71.2	69.7
	S.	73.3	77.6	76.4	76.7	75.9	78.8	77.2	76.2	75.2	75.1	78.0
	V.	84.8	85.3	84.0	85.3	84.2	85.1	86.0	77.0	84.2	86.0	84.6
零样本（PC）	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
	L.	70.2	70.0	68.0	70.1	68.5	70.0	70.7	67.5	70.6	71.8	70.0
	S.	73.6	76.6	75.6	76.0	74.8	77.8	76.9	76.9	75.1	74.9	78.2
	V.	86.1	85.7	84.7	85.7	84.5	85.5	86.2	78.2	84.6	86.8	84.8
零样本（NC）	C.	87.0	99.8	99.6	99.8	99.4	99.7	99.9	95.3	98.6	99.6	99.8
	L.	55.9	65.2	61.3	65.6	60.5	65.4	65.9	63.0	66.7	64.0	64.7
	S.	61.4	75.6	70.3	75.2	68.3	74.9	75.3	68.9	73.3	69.8	73.0
	V.	68.9	80.1	75.2	80.4	73.8	79.4	82.9	69.3	79.6	77.2	78.6

Method Domains		Avg. top-1 acc. (%) (↑) of different augmentations and prompts on VLCS										
		③④⑤	⑥	①②⑤	⑥	EDA	LLM	Rand.	Pr.	St.	Temp.	
ZS(C)	C.	99.9	97.9	99.7	99.9	99.9	99.9	99.9	99.9	99.9	99.9	
	L.	66.2	67.7	66.2	69.0	67.3	66.5					
	S.	78.5	78.0	73.2	76.9	73.5	76.9					
	V.	82.7	84.9	72.6	81.8	81.8	81.9					
ZS(CP)	C.	99.9	99.9	99.8	99.9	99.9	99.9	99.9	99.9	99.9	99.9	
	L.	70.0	70.4	69.3	70.4	71.2	69.7					
	S.	78.8	77.2	76.2	75.2	75.1	78.0					
	V.	85.1	86.0	77.0	84.2	86.0	84.6					
ZS(PC)	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	
	L.	70.0	70.7	67.5	70.6	71.8	70.0					
	S.	77.8	76.9	76.9	75.1	74.9	78.2					
	V.	85.5	86.2	78.2	84.6	86.8	84.8					
ZS(NC)	C.	99.7	99.9	95.3	98.6	99.6	99.8					
	L.	65.4	65.9	63.0	66.7	64.0	64.7					
	S.	74.9	75.3	68.9	73.3	69.8	73.0					
	V.	79.4	82.9	69.3	79.6	77.2	78.6					

方法领域		VLCS上不同增强和提示的平均Top-1准确率(%)（↑）										
		③④⑤	⑥	①②⑤	⑥	EDA	LLM	随机	概率	温度		
零样本（C）	C.	99.9	97.9	99.7	99.9	99.9	99.9					
	L.	66.2	67.7	66.2	69.0	67.3	66.5					
	S.	78.5	78.0	73.2	76.9	73.5	76.9					
	V.	82.7	84.9	72.6	81.8	81.8	81.9					
零样本（CP）	C.	99.9	99.9	99.8	99.9	99.9	99.9					
	L.	70.0	70.4	69.3	70.4	71.2	69.7					
	S.	78.8	77.2	76.2	75.2	75.1	78.0					
	V.	85.1	86.0	77.0	84.2	86.0	84.6					
零样本（PC）	C.	99.9	99.9	99.9	99.9	99.9	99.9					
	L.	70.0	70.7	67.5	70.6	71.8	70.0					
	S.	77.8	76.9	76.9	75.1	74.9	78.2					
	V.	85.5	86.2	78.2	84.6	86.8	84.8					
零样本（NC）	C.	99.7	99.9	95.3	98.6	99.6	99.8					
	L.	65.4	65.9	63.0	66.7	64.0	64.7					
	S.	74.9	75.3	68.9	73.3	69.8	73.0					
	V.	79.4	82.9	69.3	79.6	77.2	78.6					

7 C Experiments on Other CLIP Model Scales

8 C 其他CLIP模型规模的实验

8.1 C.1 Experiments on ViT-L/14

8.2 C.1 ViT-L/14上的实验

We refined the output dimension to align with the input dimension of 768 . The chosen latent dimensions were 448 and 640 for PACS and VLCS, respectively, and 768 for both OfficeHome and DomainNet. The inference weighting α was set to 0.1 for PACS, 0.03 for VLCS, 0.14 for OfficeHome, and 0.2 for Domain-Net. All other training configurations remained consistent with the ViT-B/16 experiments across each dataset. The training configuration for Im. Aug was set the same as CLAP for each dataset,with the inference weighting α being 0.1 for PACS and 0.03 for the other three datasets.
我们将输出维度调整为与输入维度768一致。选择的潜在维度分别为PACS的448和VLCS的640，OfficeHome和DomainNet均为768。推理权重α设置为PACS的0.1，VLCS的0.03，OfficeHome的0.14，DomainNet的0.2。其他训练配置与各数据集上的ViT-B/16实验保持一致。Im. Aug的训练配置与每个数据集上的CLAP相同，推理权重α为PACS的0.1，其他三个数据集为0.03。

Table 8 showcases the zero-shot results for the ViT-L/14 model using four distinct prompts, following the protocol established for the ViT-B/16 experiments. These results demonstrate that CLAP is more efficient than Im. Aug in enhancing zero-shot performance. Moreover, Tab. 9 illustrates that CLAP significantly reduces variations in zero-shot performance across different prompts, thereby confirming CLAP's performance improvements over CLIP across a range of model sizes. Detailed domain-level results are presented in Tab. 10, offering an in-depth analysis.
表8展示了ViT-L/14模型在四种不同提示词下的零样本（zero-shot）结果，遵循ViT-B/16实验的协议。这些结果表明，CLAP在提升零样本性能方面比Im. Aug更高效。此外，表9显示CLAP显著减少了不同提示词间零样本性能的波动，进一步验证了CLAP在不同模型规模上相较于CLIP的性能提升。详细的领域级结果见表10，提供了深入分析。

8.3 C.2 Experiments on ResNet50x16

8.4 C.2 ResNet50x16上的实验

To validate our approach on different model structures, we repeated zero-shot experiments on the ResNet50x16 model pre-trained with CLIP. Since the output dimension of CLIP is the same as ViT-B/16, we used the same training configuration as ViT-B/16 for training Im. Aug and CLAP. For inference, we refined the weighting coefficient α to0.1,1,0.03,and 0.1 for Im.Aug,and0.03,0.2,0.06, and 0.1 for CLAP, for PACS, VLCS, OfficeHome, and DomainNet respectively.
为验证我们方法在不同模型结构上的适用性，我们在预训练的CLIP ResNet50x16模型上重复了零样本实验。由于CLIP的输出维度与ViT-B/16相同，我们采用了与ViT-B/16相同的训练配置来训练Im. Aug和CLAP。推理时，我们将权重系数α调整为Im. Aug在PACS、VLCS、OfficeHome和DomainNet上的分别为0.1、1、0.03和0.1，CLAP分别为0.03、0.2、0.06和0.1。

Table 7: Ablative study of hyperparameters on VLCS dataset using ViT-B/16 model.
表7：使用ViT-B/16模型在VLCS数据集上的超参数消融研究。

Hyper- parameters	Value	Avg. top-1 acc. (%) (↑) using ViT-B/16 on VLCS dataset											
		ZS (C)				ZS (CP)				ZS (PC)			
		C.	L.	S.	V.	C.	L.	S.	V.	C.	L.	S.	V.
	0.1	99.9	67.6	77.5	84.2	99.9	70.9	74.9	85.9	99.9	71.2	74.6	86.3
	0.3	99.9	66.3	77.2	82.4	99.9	69.9	76.7	85.2	99.9	69.9	76.4	85.4
τ	0.5	99.9	67.7	78.0	84.9	99.9	70.4	77.2	86.0	99.9	70.7	76.9	86.2
	0.7	99.9	65.9	77.7	83.1	99.9	68.9	77.9	84.9	99.9	69.6	77.7	85.0
	0.9	99.9	66.0	77.6	83.3	99.9	69.0	77.9	85.0	99.9	69.7	77.5	85.0
	128.0	99.9	66.0	77.6	82.6	99.9	70.0	77.4	85.4	99.9	70.1	77.1	85.7
	192.0	99.9	64.9	77.9	83.0	99.9	68.9	78.0	85.6	99.9	69.0	77.8	86.0
	256.0	99.9	63.8	77.6	82.7	99.9	67.6	78.7	84.8	99.9	67.8	78.6	85.2
Lantent dim.	320.0	99.9	66.0	77.8	82.9	99.9	69.2	78.1	85.3	99.9	69.7	77.7	85.5
	384.0	99.9	65.8	76.9	82.8	99.9	69.4	77.5	85.3	99.9	69.6	77.0	85.5
	448.0	99.9	65.8	77.4	82.1	99.9	69.7	77.6	84.9	99.9	69.9	77.1	85.6
	512.0	99.9	67.7	78.0	84.9	99.9	70.4	77.2	86.0	99.9	70.7	76.9	86.2
	${10}^{-{1.5}}$	99.9	66.5	77.9	83.1	99.9	70.4	77.1	86.0	99.9	70.3	76.6	86.1
	${10}^{-1}$	99.9	69.5	77.5	85.7	99.9	70.4	77.1	86.2	99.9	70.9	76.5	86.1
	${10}^{-{0.5}}$	99.9	70.6	75.2	85.5	99.9	70.7	75.7	85.9	99.9	71.0	75.1	85.7
α	${10}^{0}$	99.8	71.5	73.5	83.5	99.9	71.7	74.4	85.8	99.8	72.3	73.5	85.5
	${10}^{0.5}$	99.8	72.0	73.1	85.5	99.8	72.2	73.7	85.7	99.8	72.5	72.9	85.6
	${10}^{1}$	99.8	72.1	72.8	85.4	99.8	72.3	73.4	85.7	99.8	72.5	72.9	85.5
	${10}^{1.5}$	99.8	72.1	72.8	85.4	99.8	72.2	73.3	85.7	99.8	72.6	72.7	85.5

超参数	数值	使用ViT-B/16在VLCS数据集上的平均Top-1准确率(%)（越高越好）											
		ZS (C)				ZS (CP)				ZS (PC)			
		C.	L.	S.	V.	C.	L.	S.	V.	C.	L.	S.	V.
	0.1	99.9	67.6	77.5	84.2	99.9	70.9	74.9	85.9	99.9	71.2	74.6	86.3
	0.3	99.9	66.3	77.2	82.4	99.9	69.9	76.7	85.2	99.9	69.9	76.4	85.4
τ	0.5	99.9	67.7	78.0	84.9	99.9	70.4	77.2	86.0	99.9	70.7	76.9	86.2
	0.7	99.9	65.9	77.7	83.1	99.9	68.9	77.9	84.9	99.9	69.6	77.7	85.0
	0.9	99.9	66.0	77.6	83.3	99.9	69.0	77.9	85.0	99.9	69.7	77.5	85.0
	128.0	99.9	66.0	77.6	82.6	99.9	70.0	77.4	85.4	99.9	70.1	77.1	85.7
	192.0	99.9	64.9	77.9	83.0	99.9	68.9	78.0	85.6	99.9	69.0	77.8	86.0
	256.0	99.9	63.8	77.6	82.7	99.9	67.6	78.7	84.8	99.9	67.8	78.6	85.2
潜在维度	320.0	99.9	66.0	77.8	82.9	99.9	69.2	78.1	85.3	99.9	69.7	77.7	85.5
	384.0	99.9	65.8	76.9	82.8	99.9	69.4	77.5	85.3	99.9	69.6	77.0	85.5
	448.0	99.9	65.8	77.4	82.1	99.9	69.7	77.6	84.9	99.9	69.9	77.1	85.6
	512.0	99.9	67.7	78.0	84.9	99.9	70.4	77.2	86.0	99.9	70.7	76.9	86.2
	${10}^{-{1.5}}$	99.9	66.5	77.9	83.1	99.9	70.4	77.1	86.0	99.9	70.3	76.6	86.1
	${10}^{-1}$	99.9	69.5	77.5	85.7	99.9	70.4	77.1	86.2	99.9	70.9	76.5	86.1
	${10}^{-{0.5}}$	99.9	70.6	75.2	85.5	99.9	70.7	75.7	85.9	99.9	71.0	75.1	85.7
α	${10}^{0}$	99.8	71.5	73.5	83.5	99.9	71.7	74.4	85.8	99.8	72.3	73.5	85.5
	${10}^{0.5}$	99.8	72.0	73.1	85.5	99.8	72.2	73.7	85.7	99.8	72.5	72.9	85.6
	${10}^{1}$	99.8	72.1	72.8	85.4	99.8	72.3	73.4	85.7	99.8	72.5	72.9	85.5
	${10}^{1.5}$	99.8	72.1	72.8	85.4	99.8	72.2	73.3	85.7	99.8	72.6	72.7	85.5

Table 11 showcases the zero-shot results for ResNet50x16 model across different prompts, substantiating that CLAP is more effective than Im.Aug in refining CLIP features. Moreover, Tab. 12 illustrates that both Im.Aug and CLAP reduce variations in zero-shot performance across different prompts, with the improvement of CLAP being more significant. The results validate our approach across different model scales, including both ViT-based and CNN-based structures. Domain-level results are detailed in Tab. 13.

表11展示了ResNet50x16模型在不同提示词下的零样本结果，证明CLAP在优化CLIP特征方面比Im.Aug更有效。此外，表12显示Im.Aug和CLAP均减少了不同提示词下零样本性能的波动，其中CLAP的提升更为显著。结果验证了我们的方法在不同模型规模上的适用性，包括基于ViT和基于CNN的结构。领域级结果详见表13。

9 D Discussion

10 D 讨论

10.1 D.1 Rationale behind CLAP's Foundation on CLIP

10.2 D.1 CLAP基于CLIP的理论依据

The primary challenge in cross-modal transferability lies in the significant domain gap between text and image data, which typically hinders the direct application of models trained in one modality to another. For a causal explaination, despite the consistency of the content variable that dictates the object label across modalities, the generative processes from latent variables to observations inherent to each modality differ markedly. The CLIP model, trained on a comprehensive dataset of image-text pairs with a symmetric InfoNCE loss, significantly ameliorates this issue. By aligning the features of text and images into similar patterns, it facilitates leveraging a network trained atop the CLIP encoder of one modality as a viable proxy for the other. Consequently, this allows for the direct application of the disentangled network trained in the text modality atop CLIP's image encoder to refine representations.
跨模态迁移的主要挑战在于文本与图像数据之间存在显著的领域差异，这通常阻碍了在一种模态上训练的模型直接应用于另一种模态。因果解释上，尽管决定对象标签的内容变量在模态间保持一致，但从潜变量到观测的生成过程在各模态中存在显著差异。CLIP模型通过在大规模图文对数据集上采用对称InfoNCE损失训练，显著缓解了这一问题。它将文本和图像的特征对齐到相似的模式，使得基于CLIP编码器训练的网络能够作为另一模态的有效代理。因此，可以直接将文本模态上训练的解耦网络应用于CLIP的图像编码器，以优化表示。

You
你

```
I have some class names, such as "dog", "car", "person", "chair". I want you to generate various
我有一些类别名称，比如"狗"、"汽车"、"人"、"椅子"。我希望你生成这些对象在不同艺术风格图片中的各种风格。
```

style of these objects in different art-style pictures. Providing me with a list of different sentences
请给我一份不同句子的列表，描述包含该对象的场景，带有不同的对象属性（如对象大小、状态或年龄（如适用）、颜色）、对象可能出现的不同背景；不同的图像类型（如照片、绘画等）。

that decribe a scene contaning the object, with different object attributes (such as object size,

condition or age(if applicable), colors), different backgrounds the object may apear in; different

image types (such as photo, painting, etc).

ChatGPT
ChatGPT

```
Absolutely, "Il provide you with a list of descriptive sentences for each of the class names "dog",
当然，我会为每个类别名称"狗"、"汽车"、"人"、"椅子"提供一份描述性句子列表，涵盖各种对象属性、背景和图像类型。
```

"car", "person", "chair," incorporating various object attributes, backgrounds, and image types.
每个句子将描述一个独特的场景：

Each sentence will describe a unique scene:

Fig. 2: The prompting method we use for generating text prompts with ChatGPT-3.5.
图2：我们使用ChatGPT-3.5生成文本提示的方法。

Table 8: Zero-shot performance across four prompts ("C", "PC", "CP") and 1 noised prompts ("NC") with CLIP pre-trained ViT-L/14 model. CLAP demonstrates consistent gains in zero-shot performance across all datasets, validating its effectiveness.
表8：CLIP预训练ViT-L/14模型在四种提示词（“C”、“PC”、“CP”）和一种噪声提示词（“NC”）下的零样本性能。CLAP在所有数据集上均表现出持续的零样本性能提升，验证了其有效性。

	Prompt Method.	Zero-shot performance,avg. top-1 acc. (%) (↑)				
		PACS	VLCS	OfficeHome	DomainNet	Overall
ZS(C)	CLIP	97.6	77.1	85.9	63.2	80.9
	Im.Aug	98.3	78.5	86.0	63.4	81.6
	CLAP	98.5	80.7	87.5	64.2	82.7
ZS(CP)	CLIP	97.3	80.6	86.0	62.0	81.5
	Im.Aug	98.3	81.1	86.1	62.4	82.0
	CLAP	98.5	81.4	87.9	63.7	82.9
ZS(PC)	CLIP	98.4	81.7	86.5	63.5	82.5
	Im.Aug	98.6	81.9	86.6	63.7	82.7
	CLAP	98.6	82.2	88.0	64.5	83.3
ZS(NC)	CLIP	91.0	65.5	77.1	55.4	72.3
	Im.Aug	95.6	69.3	77.1	55.7	74.4
	CLAP	98.5	73.1	81.3	58.3	77.8

	提示方法。	零样本性能，平均Top-1准确率（%） （↑）				
		PACS	VLCS	OfficeHome	DomainNet	总体
ZS(C)	CLIP	97.6	77.1	85.9	63.2	80.9
	图像增强	98.3	78.5	86.0	63.4	81.6
	CLAP	98.5	80.7	87.5	64.2	82.7
ZS(CP)	CLIP	97.3	80.6	86.0	62.0	81.5
	图像增强	98.3	81.1	86.1	62.4	82.0
	CLAP	98.5	81.4	87.9	63.7	82.9
ZS(PC)	CLIP	98.4	81.7	86.5	63.5	82.5
	图像增强	98.6	81.9	86.6	63.7	82.7
	CLAP	98.6	82.2	88.0	64.5	83.3
ZS(NC)	CLIP	91.0	65.5	77.1	55.4	72.3
	图像增强	95.6	69.3	77.1	55.7	74.4
	CLAP	98.5	73.1	81.3	58.3	77.8

10.3 D.2 Impact of Image and Text Augmentations

10.4 D.2 图像和文本增强的影响

Identifying pure content factors poses a significant challenge. This difficulty primarily arises from the need for finding effective augmentations of observational data to alter style factors significantly while preserving content integrity.
识别纯内容因素是一项重大挑战。这一难点主要源于需要找到有效的观测数据增强方法，以显著改变风格因素，同时保持内容的完整性。

Through the cross-modal alignment provided by CLIP, we discovered that disentangling in one modality can seamlessly improve representations in both modalities. The impact of image augmentations has been well-explored and found effective at preserving content, but traditional methods do not impose sufficient changes to remove all style information. Our exploration of text augmentations reveals that the logical structure of text and the relative ease of implementing style changes can have a significant impact on achieving disentanglement. However, more efficient methods are worthy of exploration.
通过CLIP（Contrastive Language-Image Pre-training）提供的跨模态对齐，我们发现单一模态中的解耦能够无缝提升两种模态的表示效果。图像增强的影响已被充分研究，且被证明在保持内容方面有效，但传统方法未能施加足够的变化以完全去除所有风格信息。我们对文本增强的探索表明，文本的逻辑结构及实现风格变化的相对简易性对实现解耦具有显著影响。然而，更高效的方法仍值得进一步探索。

Table 9: CLAP reduces the variance in zero-shot performance across different prompts with CLIP pre-trained ViT-L/14 model.
表9：CLAP在使用CLIP预训练的ViT-L/14模型时，减少了不同提示词下零样本性能的方差。

Metric	Method		Zero-shot variance,avg. top-1 acc. (%) (↓)				
			PACS	VLCS		DomainNet	Overall
R		CLIP	1.0	4.6	0.6	1.5	1.9
		Im.Aug	0.3	3.4	0.6	1.3	1.4
		CLAP	0.1	1.5	0.4	0.7	0.7
δ		CLIP	0.4	2.0	0.3	0.6	0.8
		Im.Aug	0.1	1.5	0.3	0.5	0.6
		CLAP	0.0	0.6	0.2	0.3	0.3
${\Delta }_{\left(NC\right) }$		CLIP	6.6	11.5	8.8	7.8	8.7
		Im.Aug	2.7	9.2	8.9	7.7	7.1
		CLAP	0.1	7.7	6.3	5.9	5.0

度量	方法		零样本方差，平均Top-1准确率（%） （↓）				
			PACS	VLCS		DomainNet	总体
R		CLIP	1.0	4.6	0.6	1.5	1.9
		图像增强	0.3	3.4	0.6	1.3	1.4
		CLAP	0.1	1.5	0.4	0.7	0.7
δ		CLIP	0.4	2.0	0.3	0.6	0.8
		图像增强	0.1	1.5	0.3	0.5	0.6
		CLAP	0.0	0.6	0.2	0.3	0.3
${\Delta }_{\left(NC\right) }$		CLIP	6.6	11.5	8.8	7.8	8.7
		图像增强	2.7	9.2	8.9	7.7	7.1
		CLAP	0.1	7.7	6.3	5.9	5.0

Table 10: Domain-level zero-shot results of the ViT-L/14 model on the test datasets.
表10：ViT-L/14模型在测试数据集上的领域级零样本结果。

Datasets	Domains	Domain-level avg. top-1 acc. (%) of zero-shot performance using ViT-L/14 (↑)											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	97.2	98.0	98.8	96.8	98.0	98.5	98.7	98.8	98.9	85.6	91.6	98.6
	C.	99.5	99.6	99.8	98.3	99.6	99.7	99.5	99.6	99.7	95.9	98.1	99.6
	P.	99.9	100.0	100.0	99.4	99.5	100.0	99.9	100.0	99.9	91.1	97.5	99.9
	S.	93.8	95.7	95.5	94.8	96.0	95.7	95.4	95.9	95.8	91.5	95.2	95.8
VLCS	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	87.5	87.9	94.4
	L.	57.4	60.1	64.3	71.3	71.6	72.6	71.7	72.0	72.6	53.8	59.7	60.7
	S.	71.0	72.4	74.4	66.2	67.4	66.8	69.9	70.4	69.9	55.9	60.5	62.9
	V.	80.0	81.6	84.3	85.2	85.7	86.2	85.1	85.3	86.4	65.0	69.3	74.3
OfficeHome	A.	86.2	86.3	87.7	85.7	86.2	88.1	87.0	87.0	87.8	78.1	77.1	80.7
	C.	73.3	73.4	75.7	73.8	73.4	76.0	73.1	73.5	76.0	65.9	66.3	70.6
	P.	92.0	91.8	93.6	92.3	92.4	94.3	92.9	92.8	94.1	80.7	81.0	86.8
	R.	92.2	92.7	93.0	92.2	92.4	93.4	93.1	93.3	93.9	83.8	84.0	86.9
DomainNet	C.	78.4	78.5	79.1	77.5	77.7	78.8	79.4	79.4	79.7	70.0	70.4	72.8
	I.	52.9	53.0	54.6	50.4	50.7	53.6	51.7	52.0	53.9	45.3	45.2	48.8
	P.	70.4	70.8	72.4	68.9	69.9	72.1	69.9	70.6	72.7	59.9	60.3	64.8
	Q.	21.5	21.6	22.5	20.6	20.9	21.7	22.6	22.8	22.9	17.9	18.4	20.2
	R.	85.8	85.9	85.9	85.3	85.5	85.7	86.3	86.4	86.2	77.5	77.5	78.7
	S.	70.2	70.4	70.7	69.4	69.8	70.6	71.0	71.3	71.5	62.0	62.2	64.6

数据集	领域	使用ViT-L/14进行零样本性能的领域级平均Top-1准确率（%） （↑）											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP
PACS	A.	97.2	98.0	98.8	96.8	98.0	98.5	98.7	98.8	98.9	85.6	91.6	98.6
	C.	99.5	99.6	99.8	98.3	99.6	99.7	99.5	99.6	99.7	95.9	98.1	99.6
	P.	99.9	100.0	100.0	99.4	99.5	100.0	99.9	100.0	99.9	91.1	97.5	99.9
	S.	93.8	95.7	95.5	94.8	96.0	95.7	95.4	95.9	95.8	91.5	95.2	95.8
VLCS	C.	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	87.5	87.9	94.4
	L.	57.4	60.1	64.3	71.3	71.6	72.6	71.7	72.0	72.6	53.8	59.7	60.7
	S.	71.0	72.4	74.4	66.2	67.4	66.8	69.9	70.4	69.9	55.9	60.5	62.9
	V.	80.0	81.6	84.3	85.2	85.7	86.2	85.1	85.3	86.4	65.0	69.3	74.3
OfficeHome	A.	86.2	86.3	87.7	85.7	86.2	88.1	87.0	87.0	87.8	78.1	77.1	80.7
	C.	73.3	73.4	75.7	73.8	73.4	76.0	73.1	73.5	76.0	65.9	66.3	70.6
	P.	92.0	91.8	93.6	92.3	92.4	94.3	92.9	92.8	94.1	80.7	81.0	86.8
	R.	92.2	92.7	93.0	92.2	92.4	93.4	93.1	93.3	93.9	83.8	84.0	86.9
DomainNet	C.	78.4	78.5	79.1	77.5	77.7	78.8	79.4	79.4	79.7	70.0	70.4	72.8
	I.	52.9	53.0	54.6	50.4	50.7	53.6	51.7	52.0	53.9	45.3	45.2	48.8
	P.	70.4	70.8	72.4	68.9	69.9	72.1	69.9	70.6	72.7	59.9	60.3	64.8
	Q.	21.5	21.6	22.5	20.6	20.9	21.7	22.6	22.8	22.9	17.9	18.4	20.2
	R.	85.8	85.9	85.9	85.3	85.5	85.7	86.3	86.4	86.2	77.5	77.5	78.7
	S.	70.2	70.4	70.7	69.4	69.8	70.6	71.0	71.3	71.5	62.0	62.2	64.6

A promising direction for future research is to explore efficient combinations of both modalities to enhance disentangled semantics. As each modality has its unique advantages-Text data recapitulates properties well since it is preprocessed by human intelligence, while image data is more precise in depicting the exact same objects or events due to its more detailed nature - the impact of combining augmentations of both modalities could be substantial.

未来研究的一个有前景的方向是探索两种模态的高效组合以增强解耦语义。由于每种模态各有其独特优势——文本数据因经过人工智能预处理而能很好地概括属性，而图像数据因其更为细致的特性在描绘相同对象或事件时更为精确——结合两种模态的增强方法可能带来显著影响。

Table 11: Zero-shot performance with CLIP pre-trained ResNet50x16 model. CLAP demonstrates consistent enhancement across all datasets, validating its effectiveness.

表11：使用CLIP预训练的ResNet50x16模型进行零样本性能测试。CLAP在所有数据集上均表现出持续提升，验证了其有效性。

		Zero-shot performance,avg. top-1 acc. (%) (↑)				
		PACS	VLCS	OfficeHome	DomainNet	Overall
ZS(C)	CLIP	96.1	70.4	80.4	57.1	76.0
	Im.Aug	96.4	74.7	80.4	57.1	77.2
	CLAP	97.0	79.9	81.6	58.0	79.1
ZS(CP)	CLIP	95.0	73.5	79.0	56.1	75.9
	Im.Aug	95.7	75.8	79.3	56.5	76.8
	CLAP	96.7	80.3	79.9	57.4	78.6
ZS(PC)	CLIP	96.5	78.4	81.7	57.1	78.4
	Im.Aug	97.0	79.8	81.8	57.4	79.0
	CLAP	96.8	80.1	82.5	58.2	79.4
ZS(NC)	CLIP	86.4	61.2	69.3	48.2	66.3
	Im.Aug	88.3	71.3	69.5	48.7	69.4
	CLAP	94.9	80.1	71.9	50.6	74.4

		零样本性能，平均Top-1准确率（%）（↑）				
		PACS	VLCS	OfficeHome	DomainNet	综合
ZS(C)	CLIP	96.1	70.4	80.4	57.1	76.0
	图像增强	96.4	74.7	80.4	57.1	77.2
	CLAP	97.0	79.9	81.6	58.0	79.1
ZS(CP)	CLIP	95.0	73.5	79.0	56.1	75.9
	图像增强	95.7	75.8	79.3	56.5	76.8
	CLAP	96.7	80.3	79.9	57.4	78.6
ZS(PC)	CLIP	96.5	78.4	81.7	57.1	78.4
	图像增强	97.0	79.8	81.8	57.4	79.0
	CLAP	96.8	80.1	82.5	58.2	79.4
ZS(NC)	CLIP	86.4	61.2	69.3	48.2	66.3
	图像增强	88.3	71.3	69.5	48.7	69.4
	CLAP	94.9	80.1	71.9	50.6	74.4

Table 12: CLAP consistently reduces variances in zero-shot performance across different prompts with CLIP pre-trained ResNet50x16 model, validating its effectiveness.

表12：CLAP在使用CLIP预训练的ResNet50x16模型时，持续降低了不同提示下零样本性能的方差，验证了其有效性。

Metric	Method	Zero-shot variance,avg. top-1 acc. (%) (↓)				
		PACS	VLCS			Overall
R	CLIP	1.5	8.0	2.7	1.1	3.3
	Im. Aug	1.3	5.1	2.5	0.9	2.4
	CLAP	0.3	0.4	2.6	0.8	1.0
δ	CLIP	0.6	3.3	1.1	0.5	1.4
	Im.Aug	0.5	2.2	1.0	0.4	1.0
	CLAP	0.1	0.2	1.1	0.3	0.4
${\Delta }_{\left(NC\right) }$	CLIP	9.7	9.3	11.1	8.9	9.7
	Im.Aug	8.1	3.5	10.9	8.5	7.7
	CLAP	2.1	-0.1	9.7	7.5	4.8

度量	方法	零样本方差，平均Top-1准确率（%）（↓）				
		PACS	VLCS			总体
R	CLIP	1.5	8.0	2.7	1.1	3.3
	图像增强	1.3	5.1	2.5	0.9	2.4
	CLAP	0.3	0.4	2.6	0.8	1.0
δ	CLIP	0.6	3.3	1.1	0.5	1.4
	图像增强	0.5	2.2	1.0	0.4	1.0
	CLAP	0.1	0.2	1.1	0.3	0.4
${\Delta }_{\left(NC\right) }$	CLIP	9.7	9.3	11.1	8.9	9.7
	图像增强	8.1	3.5	10.9	8.5	7.7
	CLAP	2.1	-0.1	9.7	7.5	4.8

Table 13: Domain-level zero-shot results using RestNet50x16 on the test datasets.

表13：在测试数据集上使用RestNet50x16进行领域级零样本测试的结果。

Datasets	Domains	Domain-level avg. top-1 acc. (%) of zero-shot performance using RN50x16 (†)											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP	CLIP	Im.Aug	CLAP
PACS	A.	95.7	95.8	97.2	93.7	95.5	96.5	95.7	96.7	96.8	81.2	84.0	94.5
	C.	98.3	98.2	99.0	98.1	98.8	99.0	98.6	98.7	98.9	92.3	93.2	98.0
	P.	98.9	98.6	99.9	98.4	97.8	99.8	99.8	99.9	99.9	85.3	87.3	95.2
	S.	91.5	93.1	91.9	89.7	90.8	91.5	91.8	92.9	91.6	86.9	88.6	92.1
VLCS	C.	96.8	97.1	99.3	99.7	99.4	99.3	99.7	99.6	99.4	75.6	89.3	99.4
	L.	53.4	60.8	65.9	51.6	58.9	67.3	59.5	68.1	66.8	54.1	60.6	67.0
	S.	63.2	70.9	69.5	68.0	72.9	69.5	72.9	73.7	69.0	52.0	66.7	69.5
	V.	68.4	70.1	85.2	74.5	72.1	85.3	81.7	78.0	85.2	63.1	68.5	84.5
OfficeHome	A.	82.2	82.5	83.5	79.7	79.9	80.6	82.0	82.4	83.4	67.7	68.9	72.1
	C.	63.0	62.9	64.7	61.7	62.2	62.8	65.4	65.3	66.1	54.6	55.0	56.8
	P.	88.2	87.9	89.0	87.4	87.5	88.5	90.0	89.9	90.6	75.4	75.3	78.2
	R.	88.1	88.2	89.1	87.3	87.5	87.6	89.2	89.5	89.7	79.5	78.9	80.3
DomainNet	C.	69.0	68.9	69.6	68.6	68.6	69.4	69.9	70.0	70.4	59.5	60.1	61.4
	I.	51.0	51.1	52.7	48.2	49.0	50.6	48.2	48.9	50.7	41.2	41.6	44.3
	P.	65.2	65.6	66.5	63.7	64.4	65.6	65.4	65.9	67.0	53.5	54.3	56.8
	Q.	11.8	11.9	12.7	12.3	12.6	13.1	11.8	12.2	12.7	9.3	9.7	11.0
	R.	82.1	82.2	83.1	81.6	81.8	82.6	83.3	83.4	83.8	72.9	73.0	74.7
	S.	63.2	63.1	63.6	62.0	62.4	63.4	63.9	63.8	64.6	53.1	53.4	55.3

数据集	领域	使用RN50x16 (†)进行零样本性能测试的领域级平均Top-1准确率（%）											
		ZS(C)			ZS(CP)			ZS(PC)			ZS(NC)		
		CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP	CLIP	图像增强	CLAP
PACS	A.	95.7	95.8	97.2	93.7	95.5	96.5	95.7	96.7	96.8	81.2	84.0	94.5
	C.	98.3	98.2	99.0	98.1	98.8	99.0	98.6	98.7	98.9	92.3	93.2	98.0
	P.	98.9	98.6	99.9	98.4	97.8	99.8	99.8	99.9	99.9	85.3	87.3	95.2
	S.	91.5	93.1	91.9	89.7	90.8	91.5	91.8	92.9	91.6	86.9	88.6	92.1
VLCS	C.	96.8	97.1	99.3	99.7	99.4	99.3	99.7	99.6	99.4	75.6	89.3	99.4
	L.	53.4	60.8	65.9	51.6	58.9	67.3	59.5	68.1	66.8	54.1	60.6	67.0
	S.	63.2	70.9	69.5	68.0	72.9	69.5	72.9	73.7	69.0	52.0	66.7	69.5
	V.	68.4	70.1	85.2	74.5	72.1	85.3	81.7	78.0	85.2	63.1	68.5	84.5
OfficeHome	A.	82.2	82.5	83.5	79.7	79.9	80.6	82.0	82.4	83.4	67.7	68.9	72.1
	C.	63.0	62.9	64.7	61.7	62.2	62.8	65.4	65.3	66.1	54.6	55.0	56.8
	P.	88.2	87.9	89.0	87.4	87.5	88.5	90.0	89.9	90.6	75.4	75.3	78.2
	R.	88.1	88.2	89.1	87.3	87.5	87.6	89.2	89.5	89.7	79.5	78.9	80.3
DomainNet	C.	69.0	68.9	69.6	68.6	68.6	69.4	69.9	70.0	70.4	59.5	60.1	61.4
	I.	51.0	51.1	52.7	48.2	49.0	50.6	48.2	48.9	50.7	41.2	41.6	44.3
	P.	65.2	65.6	66.5	63.7	64.4	65.6	65.4	65.9	67.0	53.5	54.3	56.8
	Q.	11.8	11.9	12.7	12.3	12.6	13.1	11.8	12.2	12.7	9.3	9.7	11.0
	R.	82.1	82.2	83.1	81.6	81.8	82.6	83.3	83.4	83.8	72.9	73.0	74.7
	S.	63.2	63.1	63.6	62.0	62.4	63.4	63.9	63.8	64.6	53.1	53.4	55.3