

**北京航空航天大学计算机学院**

**硕士研究生学位论文**

**文献综述**

**论文题目**：

**专 业**：计算机技术

**研究方向**：智能安全

**研 究 生**：邓彬

**学 号**：ZY2306335

**指导教师**：郭园方

**北京航空航天大学计算机学院**

2024年11月日

摘 要

近年来，基础模型方面的研究取得了显著成效，它们在各种应用中显示出了巨大的前景。随着基础模型在自然语言处理领域取得突破性进展，计算机视觉领域也逐步走上从特定任务视觉模型到视觉基础模型的道路。

**关键词**：对抗样本

Abstract

In recent years, research on foundation models has yielded remarkable results, and they show great promise in a variety of applications. With the breakthrough of foundation models in the field of natural language processing, the field of computer vision has gradually embarked on the road from task-specific vision models to vision foundation models.

**Keywords：**Adversarial Sample;

目 录

1 视觉基础模型研究现状 1

1.1 面向视觉模态任务的视觉基础模型 1

1.2 面向视觉语言模态任务的视觉基础模型 3

2 传统视觉模型对抗样本生成 6

2.1 对抗样本生成方法研究现状 6

2.2 对抗样本迁移性研究现状 8

3 面向视觉基础模型的对抗样本生成 12

3.1 视觉模态对抗样本生成方法 12

3.2 视觉语言模态对抗样本生成方法 15

4 发展趋势和现存的问题 18

5 结论 19

主要参考文献 20

图 目

[**图1 视觉基础模型体系结构** 3](#_Toc152589826)

[**图2 FLAVA架构**[7] 5](#_Toc152589827)

[**图3 FGSM攻击深度视觉模型导致高置信的错误预测**[30] 6](#_Toc152589828)

[**图4 以扰动为中心的攻击框架**[62] 13](#_Toc152589829)

[**图5 AdvCLIP攻击框架**[64] 14](#_Toc152589830)

[**图6 Co-Attack优化扰动原理**[69] 15](#_Toc152589831)

[**图7 块相似性攻击（BSA）**[77] 17](#_Toc152589832)

# 视觉语言模型的研究现状

随着基于大规模数据训练的模型（如BERT[[[1]](#endnote-2)]、DALL-e[[[2]](#endnote-3)]、GPT-3[[[3]](#endnote-4)]）的兴起，人工智能正在经历范式转变。李飞飞等人[[[4]](#endnote-5)]提出基于一类通用模型构建人工智能系统的新兴范式，并将这类通用模型统称为基础模型（Foundation Model）。基础模型定义为在大规模的广泛数据上训练的任何模型（通常使用大规模的自我监督训练），可以适应（例如通过微调）广泛的下游任务。从技术角度来看，基础模型并不新鲜，其基本组成部分，如深度神经网络（Deep Neural Network，DNN）和自监督学习，都已经存在多年。然而，基础模型在自然语言处理（Natural Language Processing，NLP）领域取得了突破性的进展，例如，有1750亿个参数的GPT-3已被有效地用于零样本/少样本学习，在不需要大规模任务相关数据或模型参数更新的情况下实现了令人印象深刻的性能。此外，最近文献中也对不同感知任务的基础模型进行了探索。例如，CLIP[[[5]](#endnote-6)]等预训练的视觉语言模型在不同的下游视觉任务上以零样本方式表现出了绝佳的性能。这意味着计算机视觉（Computer Vision，CV）领域将会效仿NLP领域，走上从特定任务视觉模型到视觉基础模型的道路。

本文将重点关注CV领域中的基础模型，为了与应用于NLP领域的基础模型有所区分，本文将应用于CV领域的基础模型统称为视觉基础模型。按照能够适应的下游任务所涉及的模态，本章将现有的视觉基础模型划分为能面向视觉模态任务的视觉基础模型和面向视觉语言模态任务的视觉基础模型两类，分别能够适应多个视觉模态任务和多个视觉语言模态任务，并以此展开对视觉基础模型的介绍。视觉模态任务涵盖了语义分割、目标检测等多种视觉任务，视觉语言模态任务包括视觉问答、图像文本检索、图像描述等涉及视觉、语言两种模态的下游任务。

## 面向视觉模态任务的视觉基础模型

大规模训练和有效的提示输入是视觉基础模型的重要组成部分。大规模数据是核心，用于预训练的数据集可以分为三大类：图像-文本数据集（如CLIP使用的WebImageText[5]）、部分合成数据集（如SAM使用的SA-1B [[[6]](#endnote-7)]）和组合数据集（如FLAVA使用的PMD [[[7]](#endnote-8)]）。微调技术是基础模型能够适应广泛下游任务的关键，经过大规模训练的模型可以通过微调提高在特定任务上的性能（如开放世界目标检测）和特定功能（如视觉定位），以及通过指令微调（Instruction-tuning）以使模型解决不同的下游视觉任务。

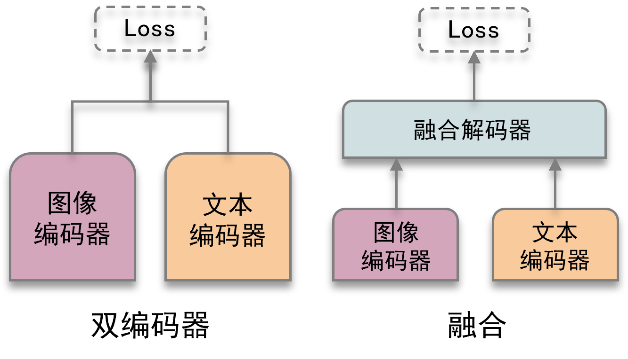
提示工程（Prompt Engineering）最早出现在NLP领域，由于基于提示的方法可以缩小预训练任务和下游任务之间的差距，并有效提高预训练模型性能，甚至有可能接近全参数微调的水平，因此它们在各种NLP任务中表现出显著的优势。最近，受NLP领域提示工程成功的启发，CV领域在利用提示方面也取得了一系列进步。在CV任务中使用提示最早可以追溯到交互式分割，通过用户输入（如点击、边界框、涂鸦等）指导模型准确识别目标边界，进而使分割结果更加精确可靠。随后有不少研究尝试在视觉基础模型上应用提示工程，来引导视觉基础模型生成符合特定要求的输出，使其更加准确、可控和适应不同的任务和应用场景。例如，Lüddecke等人[[[8]](#endnote-9)]提出的CLIPSeg由基于CLIP的图像和文本编码器以及具有跳跃连接的基于Transformer[[[9]](#endnote-10)]的解码器组成。CLIPSeg通过CLIP编码器来获得视觉和文本提示的嵌入，然后将嵌入馈送到解码器，使得CLIPSeg可以根据任意提示生成图像分割。

VPT[[[10]](#endnote-11)]是第一个研究视觉提示的普遍性和可行性的方法，VPT在输入空间引入少量可训练参数作为提示，同时冻结预训练ViT（Vision Transformer）[[[11]](#endnote-12)]模型的参数来进行微调。通过优化这些提示，VPT在推理过程中增强了关于特定视觉模态和任务需求的模型性能。VPT的成功表明了视觉提示的有效性，并为后续工作提供了一个新颖的提示设计视角。受VPT引入可调参数的启发，AdaptFormer[[[12]](#endnote-13)]提出轻量级模块AdaptMLP并将其集成到ViT架构中，进而在动作识别任务上获得优于全参数微调模型的性能。AdaptFormer的设计理念是利用可训练模块，这些模块根据任务特定约束进行定制，使得能够在不更新原始预训练参数的情况下扩展到多种视觉任务。类似的，Jie等人[[[13]](#endnote-14)]提出在ViT中构建卷积旁路（Convpass，Convolutional Bypasses）作为适应模块，来减少与微调相关的计算成本，同时提高预训练模型对特定视觉任务的适应性。与其他参数高效迁移学习（PETL，Parameter-Efficient Transfer Learning）方法不同，Convpass得益于卷积层的硬编码归纳偏置，因此更适合于视觉任务。他们的研究表明，可以为视觉基础模型定制面向视觉的适应模块来提升模型对特定视觉任务的适应性。

Meta AI研究团队旨在创建一个通用的图像分割模型，能够通过提示工程解决新数据上的广泛下游分割任务，进而提出了一种称为Segment Anything Model（SAM）[6]的视觉基础模型。为了增强模型适应提示的灵活性并提高其抗干扰的鲁棒性，SAM分为三个部分：图像编码器、提示编码器和掩膜解码器。这样划分有效地分配了计算成本，从而形成一个适应性强、通用性强的分割模型。Meta AI研究团队收集了11亿个掩膜和1100万个图像，形成大规模数据集SA-1B，并在该数据集上训练SAM。SAM分别使用图像和提示编码器对给定的图像和视觉提示（如框、点或掩膜）进行编码，然后将其组合在预测分割掩膜的轻量级掩膜解码器中。通过提示分割预训练和特定下游任务微调的方法使得SAM利用从提示分割任务中学到的知识来提高在广泛分割问题上的性能。此外，SAM的模块化设计允许对不同提示格式的灵活性和适应性，使其成为应对各种细分挑战的通用解决方案，如医学影像分析（如MedSAM[[[14]](#endnote-15)]、AutoSAM[[[15]](#endnote-16)]）、视频目标跟踪（如SAM-Track[[[16]](#endnote-17)]、TAM（Track Anything Method）[[[17]](#endnote-18)]）、遥感图像实例分割（RSPrompter[[[18]](#endnote-19)]）、图像注释（CAT（Caption AnyThing）[[[19]](#endnote-20)]）等。

## 面向视觉语言模态任务的视觉基础模型

视觉基础模型能够适应于视觉语言模态的不同下游任务得益于提示工程在预训练的视觉模型上的应用。考虑到研究主要集中于CV领域，本文不考虑多模态LLM，即不考虑利用LLM作为核心组件并通过视觉编码器将图像转换为与LLM兼容的格式，来实现LLM与视觉模态对齐的模型。本节所讨论的视觉基础模型主要有两种体系结构，如图1。一种是双编码器结构，使用单独的编码器来处理视觉和语言模态（如CLIP[5]），编码器的输出通过目标函数进行优化；另一种是采用了视觉语言模态融合策略的结构，加入了一个融合解码器，它获取视觉和文本编码器的输出，并学习融合的表示。



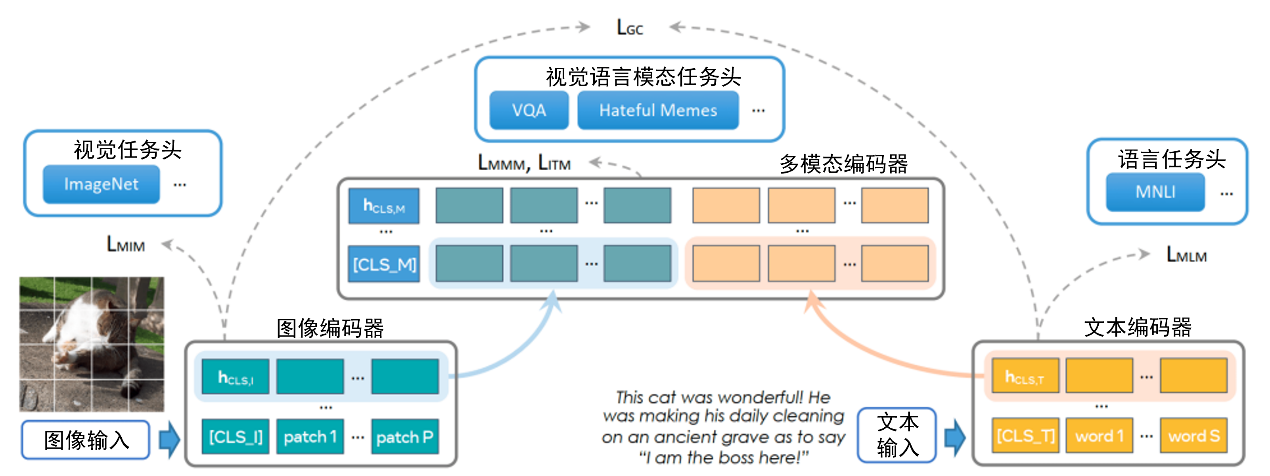
**图1 视觉基础模型体系结构**

传统的深度视觉模型通常使用有监督或弱监督的预训练来预测一组预先确定的类别，这限制了它们的通用性和可用性。Radford等人[5]建议从自然语言中存在的视觉概念中学习感知，并开创性的提出了对比语言图像预训练（CLIP，Contrastive Language-Image Pre-training）模型。为了训练CLIP，他们从互联网收集了共4亿个文本-图像对，形成大规模数据集WebImageText。CLIP采用一个图像编码器（如ViT或大规模的卷积神经网络（Convolutional Neural Network，CNN））来提取图像特征，以及一个文本编码器（如Transformer）来处理文本提示，通过对称交叉熵损失（SCE，Symmetric Cross-Entropy）来最小化正确的图像-文本对之间的余弦相似性和最大化不正确的图像-文本对之间的余弦相似性。这种设计允许CLIP理解视觉和文本信息，并在这两种模态之间建立有意义的关联。CLIP为在CV中应用提示工程提供了新的见解，随后基于对比学习提出的方法大多可以视作CLIP架构的变体，例如，FLIP[[[20]](#endnote-21)]基于自监督对比学习，SLIP[[[21]](#endnote-22)]在图像输入中加入随机掩膜，MaskCLIP[[[22]](#endnote-23)]则在图像和文本输入中均加入随机掩膜。然而，Ghiasi等人[[[23]](#endnote-24)]认为类似CLIP的方法在视觉定位任务中表现不佳，因而提出了OpenSeg。他们先通过分割损失在分割数据集上训练模型，来为每个类别寻找最为匹配的预测分割掩膜，然后借鉴CLIP通过对比学习进行视觉语义对齐。实验表明，基于OpenSeg的模型可以很好的推广到新的数据集并达到优于SOTA（state of the art）模型的性能。

还有一些工作同样没有基于CLIP，而是利用了强大的预训练模型，并采用视觉语言模态融合的策略。Li等人[[[24]](#endnote-25)]提出了基于定位的语言图像预训练模型GLIP（Grounded Language-Image Pre-training），他们开创性的将目标检测任务转换为短语定位任务，即对任一张图片，把标签用句号隔开，拼接成一句话。通过这种方式，所有的目标检测数据集都可以转化为短语定位数据集。他们还引入了深度融合（Deep Fusion）概念：对视觉和语言两个模态的特征向量计算交叉注意力，从而让模型可以在前期阶段就开始进行跨模态的特征学习。受此启发，Liu等人[[[25]](#endnote-26)]将基于Transformer的目标检测器DINO[[[26]](#endnote-27)]与语言预训练相结合，提出了能够适应开集目标检测任务的Grounding-DINO方法。他们将DINO划分为用于特征提取的主干、用于特征增强的颈部和用于区域细化的头部，并在颈部和头部采用对比损失融合语言特征和区域输出。Grounding-DINO在闭集、开集和参考目标检测方面显著优于GLIP。

针对不同模态设计单独编码器的视觉基础模型显示出令人印象深刻的性能，但不太适合需要同时处理视觉和语言两种模态的问题，而应用了视觉语言模态融合策略的视觉基础模型不适用于视觉模态任务。Yu等人[[[27]](#endnote-28)]融合了单编码器、双编码器和编码器-解码器架构，提出了CoCa（Contrastive Captioners）模型。CoCa由一个图像编码器、一个文本编码器以及一个具有交叉注意力层的解耦多模态解码器组成。单模态编码器在类似CLIP的对比损失上进行训练，而解耦多模态解码器采用生成学习方法进行训练，两种方法的结合使得CoCa在单模态和多模态下游任务（如图像分类、视频理解、跨模态检索等）中均表现出色。

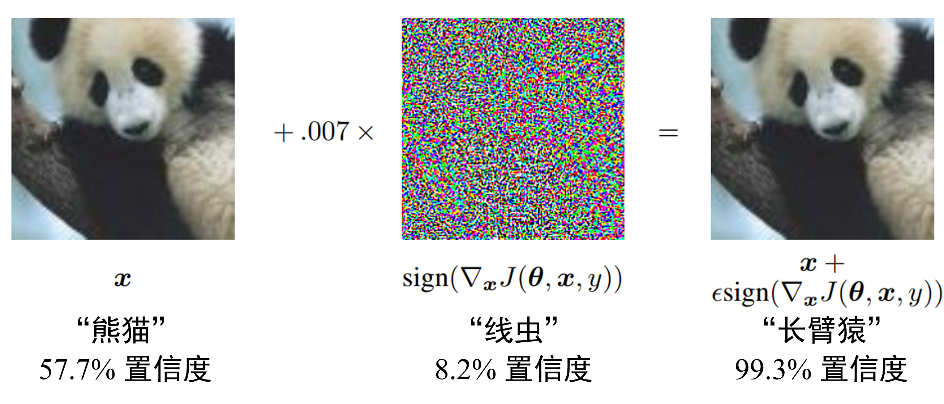
Singh等人[7]认为一个真正的基础模型应该同时能够在视觉模态、语言模态和视觉语言模态任务中表现良好，为此，他们提出了一个名为FLAVA的架构，如图2。FLAVA由图像和文本编码器、多模态编码器、视觉任务头、语言任务头和多模态任务头组成。图像编码器、文本编码器和多模态编码器均采用ViT架构，多模态编码器采用交叉注意力并将前两者的输出进行融合并提供给特定模态任务头。他们融合了多个图像-文本数据集，形成了PMD数据集，并在此数据集上训练FLAVA。为了获得强大的泛化能力，FLAVA在多个单模态和多模态损失上进行训练，包括类似CLIP的对比损失、多模态掩膜建模（MMM，Masked Multimodal Modeling）、图像-文本匹配（Image-text Matching）、图像掩膜建模（MIM，Masked Image Modeling）、语言掩膜建模（MLM，Masked Language Modeling）等。训练包括在监督数据集上对图像和文本编码器进行单模态预训练，然后在图像-文本数据集上进行单模态和多模态联合训练。实验表明，FLAVA在35项视觉、语言和视觉语言任务中展现出绝佳性能。



**图2 FLAVA架构**[7]

# 传统视觉模型对抗样本生成

深度模型在计算机视觉领域的应用最早可以追溯到2012年，Krizhevsky等人[[[28]](#endnote-29)]使用CNN在图像分类任务上取得了里程碑式的突破。虽然深度模型在各种各样的计算机视觉任务上能够取得惊人的准确率，有时其表现甚至能够超过人类，但是最近的一些研究表明，基于深度学习的视觉模型在安全方面存在致命缺陷。Szegedy等人[[[29]](#endnote-30)]发现，通过在输入图像上添加非常微小的扰动来构造对抗样本（Adversarial Examples，AE），能够影响深度神经网络的预测结果。这些扰动可以被限制在人类视觉系统不可察觉的范围内，但它们可以完全改变深度视觉模型的输出预测，如图3。



**图3 FGSM攻击深度视觉模型导致高置信的错误预测[****[[30]](#endnote-31)]**

最初，这些能影响预测结果的对抗样本是因图像分类任务而发现的[29]。然而，对于各种主流的计算机视觉任务，它们的存在已经得到了证实。本章主要介绍面向传统视觉模型的对抗样本生成方法。

本章第二节则进一步介绍了基于对抗样本的迁移性的研究。

## 对抗样本生成方法研究现状

图像分类任务是计算机视觉的基础问题之一，早期的对抗样本生成方法更多集中在计算对抗图像的基本算法上，使用图像分类任务作为测试平台。Szegedy等人[29]首次发现了深度视觉模型对对抗扰动的脆弱性，并提出可通过Broyden-Fletcher-Goldfarb-Shanno（L-BFGS）算法将对抗扰动优化问题转化为求解以下问题，以生成人类视觉系统难以察觉的对抗样本：

(2.1)

其中表示自然图像，表示对抗扰动，表示深度视觉模型的输出，为分类损失。然而，在大量图像上求解上述问题需要耗费高昂的计算成本，Goodfellow等人[30]提出了快速梯度符号法（FGSM，Fast Gradient Sign Method）以更高效地计算对抗扰动：

(2.2)

其中表示以为参数的模型的损失函数，表示关于图像的梯度，为符号函数，为控制扰动系数的超参，最终通过计算得到对抗样本。FGSM是现有文献中最具影响力的攻击之一，它的核心是在模型的损失面上执行梯度上升以达到欺骗模型的目的，有许多后续的研究受此启发。例如， Miyato等人[[[31]](#endnote-32)]删除了公式(2.2)中的符号函数，并通过范数对梯度进行归一化来生成对抗样本。Kurakin等人[[[32]](#endnote-33)]同样采用基于范数的归一化，并将FGSM扩展为迭代梯度上升，即I-FGSM（Iterative-FGSM）。在I-FGSM基础上，Dong等人[[[33]](#endnote-34)]建议增加动量项（MI-FGSM，Momentum I-FGSM）来稳定梯度更新方向，避免局部最大值，而Xie等人[[[34]](#endnote-35)]建议以固定概率对输入图像进行变换（如随机调整大小和填充），使I-FGSM在每次迭代中的输入多样化（DI²-FGSM，Diverse Input I-FGSM）。Dong等人[[[35]](#endnote-36)]基于CNN特有的平移不变性进一步提出了一种更高效的方法（TI-FGSM，Translation-Invariant FGSM），并提出通过预定义的核矩阵对未平移图像的梯度进行卷积，来实现加速计算。Lin等人[[[36]](#endnote-37)]从将对抗样本生成看作一个优化过程的角度出发，基于Nesterov加速梯度思想和DNN特有的缩放不变性提出了SI-NI-FGSM（Scale-Invariant Nesterov I-FGSM）方法。此外，投影梯度下降（PGD，Projected Gradient Descent）也是对抗样本生成研究中最具影响力的方法之一，起源于Madry等人[[[37]](#endnote-38)]的开创性工作。PGD采用多次迭代，每次迭代都会将扰动投影到限定范围内，这使得PGD的攻击效果优于FGSM，但也因此需要耗费高昂的计算成本。同时，Madry等人也将I-FGSM视作PGD方法，他们认为I-FGSM所采用的裁剪运算是投影的一种特例。因此，基于I-FGSM的改进方法也可用于PGD。

大多数早期的对抗样本生成方法集中在将扰动的或范数限制在某一范围内来使得扰动不可察觉，而Moosavi-Dezfooli等人[[[38]](#endnote-39)]提出的DeepFool方法没有将扰动范数限制为固定值，而是致力于通过求解以下问题来最小化对抗扰动的范数：

(2.3)

DeepFool方法将数据点周围的决策边界线性化，形成一个凸多面体，以将该点逐渐推到最近的边界上，从而获得最小扰动。此外，Rony等人[[[39]](#endnote-40)]提出了通过将对抗扰动的方向与范数进行解耦，从而在范数较低的情况下成功攻击模型。

在发现深度模型的对抗样本脆弱性开后，针对对抗攻击的防御方法的研究也随之展开。防御蒸馏[[[40]](#endnote-41)]为解决该问题提供了一种有效的方案，它建立在知识蒸馏的基础上。然而，Carlini等人[[[41]](#endnote-42)]设计了一组攻击，通过解决以下优化问题计算范数约束的加性扰动：

(2.4)

其中表示目标深度视觉模型，是满足的函数。他们尝试了的多种解析形式，并在、和伪范数下生成对抗扰动，形成了一系列对抗样本。实验表明，他们所生成的对抗样本在黑盒设置下成功地欺骗了防御蒸馏网络，且对其他防御技术也很有效。C&W攻击通常被认为是一种非常强的攻击，但它需要很高的计算成本。

在图像分类以外的视觉任务中，对抗样本生成方法与用于欺骗分类模型的方法类似，但因不同任务具有各自独特的任务目标，导致了攻击方法更具有针对性。例如，Gu等人[[[42]](#endnote-43)]分析了PGD用于语义分割所存在的问题，提出了特定于分割任务的PGD，即SegPGD。Chow等人[[[43]](#endnote-44)]建议通过欺骗检测器来误导模型输出错误的目标检测结果，并提出了能够同时适应于一阶段和两阶段目标检测网络的TOG（Targeted Objectness Gradient）攻击。

此外，还有一些工作尝试攻击面向视觉语言模态任务的视觉模型，例如，Xu等人[[[44]](#endnote-45)]尝试通过梯度迭代方法优化图像扰动来实现对视觉问答模型的有目标攻击。然而，这些工作提出的攻击方法都是针对基于CNN和递归神经网络（Recurrent Neural Network，RNN）实现的模型，不适用于视觉基础模型。不仅如此，这些工作主要针对特定视觉语言分类任务，而且本质上是基于视觉模态的攻击，这也使得这些方法在视觉语言模态任务上迁移性较差。

## 对抗样本迁移性研究现状

随着对抗样本的存在引起了人们对深度模型的鲁棒性和可靠性的担忧，研究人员发现了一个更有趣的现象——对抗样本具有迁移性[30]。给定输入图像的对抗样本和两个模型和，迁移性指能够欺骗模型的对抗样本也可以欺骗另一个模型，通常和是具有不同架构的模型。从形式上讲，对抗样本的迁移性可表示为：

(2.5)

则生成具有迁移性的对抗扰动的过程可表示为：

(2.6)

其中表示增强迁移性的操作，表示损失函数，则是在优化过程中使用的替代模型。公式(2.6)的问题通常通过PGD[37]来解决。根据现有的研究，本文将增强迁移性的方法分为四类：基于数据增强、基于优化、基于模型组件和基于损失函数。

基于数据增强的方法都是基于应用数据增强策略的基本原理，通过具有随机分量的输入变换来防止对替代模型的过拟合，进而增加对抗样本的迁移性。Xie等人[34]是第一个建议将可微分输入变换应用于干净图像的，即以固定概率对输入图像进行变换使得输入多样化。Dong等人[35]通过平移图像来防止扰动对图像的过拟合，并提出通过在梯度上卷积核矩阵来实现更高效的计算。类似的，Lin等人[36]建议利用缩放不变性来寻找适用于任何经过缩放的输入的扰动。Li等人[[[45]](#endnote-46)]观察到对抗扰动往往具有高度的区域同质性（即相邻像素的梯度往往相似），基于这一发现他们建议将归一化层应用于计算梯度，以促进区域均匀扰动。Wang等人[[[46]](#endnote-47)]提出了一种Mixup方法的变体（Admix），即将不同随机采样图像的一小部分添加到原始图像中（标签保持不变）来增加输入的多样性，从而提高扰动的迁移性。受人类对3D对象上的图像（如印在T-恤上的图像）识别能力的启发，Byun等人[[[47]](#endnote-48)]提出基于对象的多样化输入（ODI，Object-based Diverse Input）方法，通过将图像在一组3D对象上进行渲染来实现数据增强。

基于优化的方法则是将提升对抗样本迁移性等同于提升模型泛化性。Lin等人[36]强调了生成可迁移对抗样本的过程与训练神经网络的过程具有相似性。具体而言，用于生成对抗样本的模型可以被看作为训练数据，对抗扰动可以被看作为要优化的参数，而目标模型可以被看作测试数据。从这一角度来看，对抗样本的迁移性可以解释为训练模型的泛化能力，因此用于提高神经网络泛化能力的基于优化的方法可以用来提高对抗样本的迁移性。例如，在梯度方向上引入动量是提高神经网络泛化性的最简单和最广泛使用的技术之一。Dong等人[33]提出在I-FGSM方法基础上加入动量项来摆脱较差的局部最大值，从而提高对抗样本的迁移性，而Lin等人[36]则建议将Nesterov加速梯度集成到基于梯度迭代的攻击中，以利用其更新策略更容易、更快地避免陷入局部最优。Wang等人[[[48]](#endnote-49)]受方差缩减技术的启发提出了一种方差调整技术，即在每次迭代时采用前一个数据点邻域中的梯度信息来调整当前数据点的梯度，并将其与上面的方法结合得到方差调整MI-FGSM（VMI-FGSM）和方差调整NI-FGSM（VNI-FGSM）方法。Ma等人[[[49]](#endnote-50)]注意到积分梯度（IG，Integrated Gradient）的一个明显特征是其不受模型结构的影响，仅取决于模型的输入和输出，这种特性可能有助于提高扰动在不同模型架构之间的迁移性，因此提出了将IG纳入MI过程的动量积分梯度（MIG，Momentum Integrated Gradient）方法。

之前的方法中通常采用交叉熵损失作为生成对抗样本的损失函数，Zhang等人[[[50]](#endnote-51)]发现常用的损失算法弱点在于优先考虑速度来欺骗模型，而不是最大化其强度。他们提出了一种新的归一化交叉熵损失来引导logit梯度朝着隐式最大化其与真实类的秩距离的方向更新，以提升top-k强度，并在可迁移黑盒设置下实现了top-k攻击。Zhao等人[[[51]](#endnote-52)]发现现有研究不合理地将扰动优化限制在较少的迭代次数下，这样高估了迁移攻击的难度。他们的研究表明，只要迭代足够的次数，即使是传统的I-FGSM与简单的迁移方法结合，也可以很容易的实现较高的迁移性。他们还证明，利用简单logit损失可以进一步大大提升攻击的迁移性。简单logit损失可表示为：

(2.7)

其中表示在softmax层之前相对于目标类别的logit输出。

还有一些针对对抗样本迁移性的研究侧重于模型本身。例如，Benz等人[[[52]](#endnote-53)]发现批量归一化（BN，Batch Normalization）会导致模型更加依赖非鲁棒特征，并增加对抗攻击的脆弱性。他们证明了在没有BN的情况下训练的替代模型优于采用了BN的替代模型，并且提前停止替代模型的训练也可以增强攻击的迁移性。Wu等人[[[53]](#endnote-54)]发现跳跃连接有助于生成高度可迁移的对抗样本，并提出了跳跃连接梯度方法（Skip Gradient Method）。该方法使用了一个衰减因子来减少来自残差块的梯度，使用更多来自跳跃连接的梯度生成高迁移性的对抗样本。Guo等人[[[54]](#endnote-55)]提出以线性方式执行反向传播的方法（LinBP）增强DNN的线性来提高迁移性。

Liu等人[[[55]](#endnote-56)]首次提出利用具有不同体系结构的多个模型构成的集合来生成可迁移的对抗样本。然而，现有的集成攻击只是统一地融合替代模型的输出，因此不能有效地捕获和放大对抗样本的内在迁移信息。Chen等人[[[56]](#endnote-57)]提出了一种自适应的集成攻击方法（AdaEA，Adaptive Ensemble Attack），通过监测每个模型对对抗样本的贡献差异，自适应地控制每个模型输出的融合。集成攻击的基本思想是利用多个白盒模型的输出来获得平均模型损失，然后应用基于梯度的攻击来生成对抗样本。他们提出在AdaEA上配备AGM和DRF机制，以修改梯度优化过程来增强生成的对抗样本中的迁移信息。这个过程可通过以下公式简单表示：

(2.8)

(2.9)

其中表示个模型的集成梯度，表示可以清除集成梯度中视差部分的滤波器进而同步梯度更新方向，表示逐元素乘法。

此外，Moosavi-Dezfooli等人[[[57]](#endnote-58)]首次针对模型输入研究对抗样本的迁移性。他们提出计算图像不可知的扰动，这些扰动能够在任何图像上以很高的概率欺骗模型。这些扰动因其在不同图像之间的迁移性而被称为通用对抗扰动（UAP，Universal Adversarial Perturbations），其目的是满足以下约束：

(2.10)

其中表示概率，表示干净图像的概率分布，是一个预定义的标量，决定扰动的可接受愚弄率。实验表明，所生成的对抗样本不仅在图像间具有迁移性，而且在不同的模型间也能有良好的迁移性。Moosavi-Dezfooli等人尝试通过DeepFool[38]的思想构建UAP，通过将所有数据点推出它们各自的决策边界，同时将它们反投影到半径为的球上，以此生成的对抗样本能够以50%左右的愚弄率欺骗ImageNet模型。UAP的研究为提升对抗样本的迁移性提供了新的思路。

在对抗样本的迁移性方面的研究仍面临许多挑战。例如，对抗样本迁移到不同架构的模型上时攻击性能仍然会大幅降低，且攻击特定类别比无目标的攻击要困难得多。现有方法大多专注模型间的迁移性，而理解图像间的迁移性与模型间的迁移性之间的相互作用对于提升对抗样本的迁移性也具有重要意义。此外，在传统视觉模型上增强对抗样本迁移性的方法方法主要是针对具有预定义类的视觉分类模型而开发的，而视觉基础模型从不同的角度提取视觉信息，对迁移性提出了新的挑战。

# 面向视觉基础模型的对抗样本生成

与传统深度模型类似，基础模型同样也存在安全性问题。尽管基础模型具有空前的规模，但当前的基础模型在对抗鲁棒性方面收效甚微。已有研究表明，在图像任务上，通过在测试样本上添加微小的、几乎不能被肉眼识别的扰动所构造出的对抗样本（Adversarial Examples），能够使视觉基础模型受到欺骗，输出完全不同的分类结果。例如，Fort[[[58]](#endnote-59)]发现仅通过简单梯度下降优化生成的对抗样本，就可以使得CLIP[5]在零样本分类任务上出现明显错误。这一发现使得基础模型的安全性开始逐步受到重视。

最近，OpenAI等研究团队出于安全性考虑，建议在实际应用场景中应当采用闭源的基础模型。这为基础模型的对抗鲁棒性研究带来了新的挑战，亟待通过对抗样本研究基础模型在强对抗环境中的实际性能，以确保其在复杂场景下的安全性和可靠性。

传统的面向视觉模型的对抗样本生成方法主要针对特定任务设计，而在能够适应多种下游任务的视觉基础模型上不适用。因此，一些研究致力于面向视觉基础模型设计一种具有良好迁移性的对抗样本生成方法。

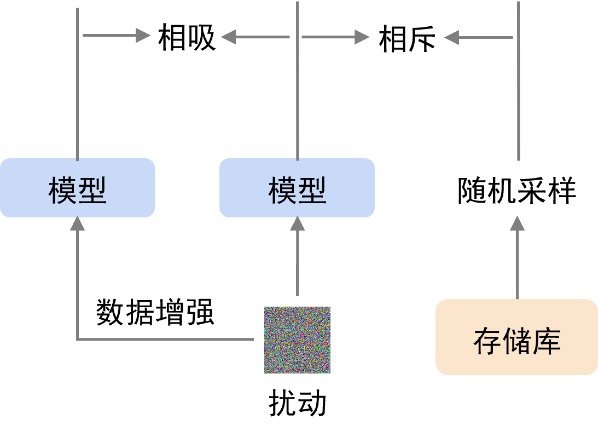
本章将面向视觉基础模型的对抗样本生成方法分为仅扰动图像的视觉模态对抗样本生成和对图像和文本均进行扰动的视觉语言模态对抗样本生成，并从这两个方面分别介绍与面向视觉基础模型的对抗样本生成有关的工作。

## 视觉模态对抗样本生成方法

前面1.1节介绍了现有的适应于多个视觉任务的视觉基础模型（如SAM[6]等），有一部分工作针对这类模型设计能够在多种视觉任务上适用的对抗样本生成方法。Zhang等人[[[59]](#endnote-60)]首次针对SAM的对抗鲁棒性进行了研究。他们提出了一个称为Attack-SAM的攻击框架，并设计了一个简单有效的ClipMSE损失，在基于提示的掩膜预测任务中能够使SAM被攻击以生成任何所需的掩膜。这项研究表明视觉基础模型在白盒场景中容易受到对抗样本的影响。然而，Attack-SAM在生成对抗样本中需要知道输入的提示位置，而事先了解提示内容是不现实的，并且会使得对抗样本对固定的提示位置过拟合而降低迁移性。

Zheng等人[[[60]](#endnote-61)]在Attack-SAM基础上采用随机化提示点的方法实现对抗样本对提示不可知。此外，他们受Ilyas等人[[[61]](#endnote-62)]启发，提出通过计算对抗样本与其他干净样本的余弦相似度来测量对抗样本特征的相对强度，并将其作为正则项来提升对抗样本特征的强度，进而实现跨模型的黑盒对抗攻击。但是他们提出的方法只是对采用不同图像编码器的SAM具有迁移性。

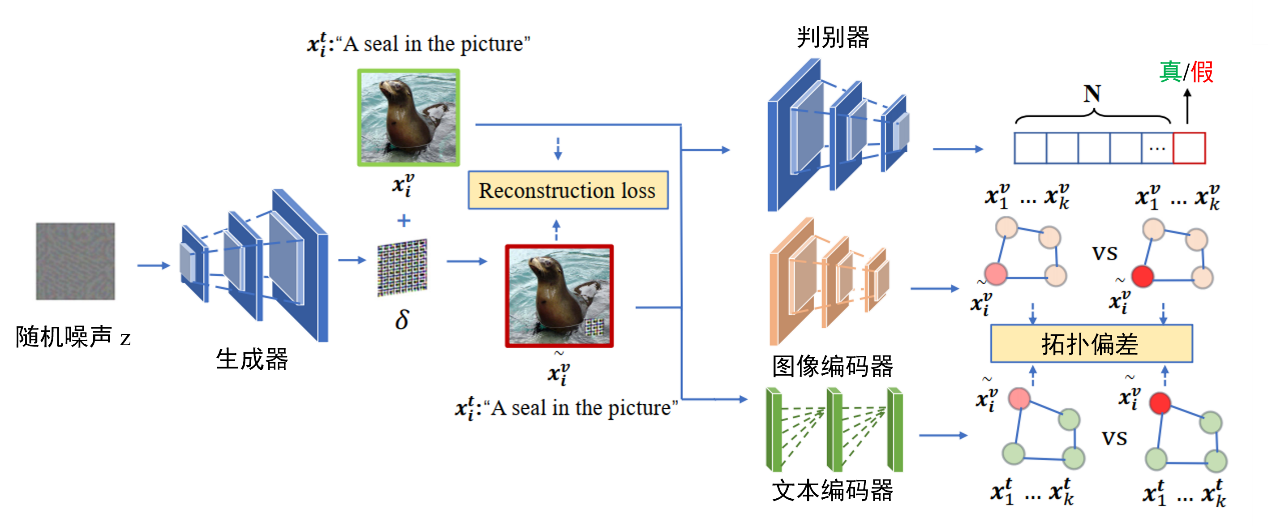
与Zhang等人[59]和Zheng等人[60]不同，Han等人[[[62]](#endnote-63)]致力于研究是否可以使用UAP[57]方法攻击视觉基础模型SAM。他们认为传统的对抗样本生成方法（如PGD[37]、DeepFool[38]等）都是以图像为中心，而这些方法只适用于针对特定图像的攻击，需要在图像发生变化时生成新的扰动。因此，他们提出以扰动为中心的攻击框架，基于自监督的对比学习方法生成UAP：



**图4 以扰动为中心的攻击框架**[62]

具体而言，他们先采用类似PGD的方法基于梯度下降生成对抗扰动，然后将对抗扰动作为对比视图，将对抗扰动增强后作为正样本，选择随机图像作为负样本，采用InfoNCE[[[63]](#endnote-64)]损失优化学习UAP的独立特征。实验表明，最终生成的UAP不仅能够在图像不可知的情形下成功攻击SAM，而且还能够在提示类型不可知的情形下影响SAM输出的掩膜。然而，以上的工作均基于SAM训练生成对抗样本，且仅适用于基于提示的掩膜预测任务，使得这些方法跨任务迁移性差。

此外，还有一部分工作在能够适应于视觉语言模态任务的视觉基础模型上进行，但是仅对图像进行扰动，因此将其归为视觉模态对抗样本。Zhou等人[[[64]](#endnote-65)]首次基于多模态对比学习方法提出了面向视觉基础模型的攻击框架AdvCLIP，能够生成下游不可知的对抗样本。该框架由一个生成器、一个判别器和一个跨模态编码器组成，其中跨模态编码器包含一个图像编码器和一个文本编码器，如图5。生成器通过输入的随机噪声生成对抗扰动，进而与图像组成对抗样本，并输入判别器和跨模态编码器。损失函数由四部分构成：对抗损失，确保对抗样本的特征远离干净图像和文本；拓扑偏差损失，最大化对抗样本与正样本之间的拓扑距离来破坏两者间的拓扑相似性；扰动约束，确保生成的扰动难以被人类视觉系统察觉；GAN损失，使对抗样本与干净图像在判别器上趋于一致，确保对抗样本在视觉上更自然。然而，AdvCLIP要求视觉基础模型的编码器是开源可访问的，即跨模态编码器中的图像编码器和文本编码器来自目标基础模型，同时还要求用户只会微调线性层。这导致AdvCLIP在黑盒场景下不适用。此外，实验结果表明，AdvCLIP仅在图像分类数据集上具有较高的攻击性能，跨任务迁移性差。且由于其仅考虑了对图像的扰动，使得AdvCLIP在图像文本检索任务上的攻击成功率不高。



**图5 AdvCLIP攻击框架**[64]

Zhao等人[[[65]](#endnote-66)]将在视觉基础模型上的黑盒对抗攻击分为基于迁移和基于查询两类，并从这两个方向研究对抗样本生成。在基于迁移的攻击方向，他们提出了两种设计，一种是匹配图像-文本特征（MF-it，Matching image-text features），基于PGD生成能够匹配目标文本的对抗样本；另一种是匹配图像-图像特征（MF-ii，Matching image-image features），他们受Yuksekgonul等人[[[66]](#endnote-67)]的启发，先采用文本生成图像的扩散模型（如Stable Diffusion[[[67]](#endnote-68)]）来生成与目标文本相关的目标图像，进而通过如下方式优化对抗扰动生成对抗样本：

(3.1)

其中表示图像编码器，表示扩散模型，表示目标文本。此外，由于优化对抗样本时不需要扩散模型的梯度信息。因此，对于扩散模型只要求黑盒可访问，即可以使用公开提供的API，这大大降低了生成对抗样本的开销。在基于查询的攻击方向，他们提出了匹配文本-文本特征（MF-tt，Matching text-text features）。他们假设在黑盒场景下可对目标基础模型执行重复查询的操作，即通过提供图像和文本输入来获得文本输出，则可将基于MF-tt的优化过程表示为：

(3.2)

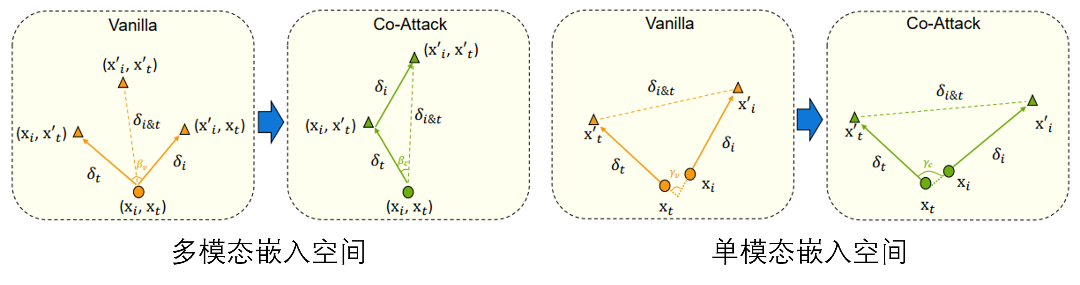
其中表示文本编码器，表示目标基础模型，表示干净文本输入。由于基于黑盒场景的假设，无法在目标基础模型上执行反向传播，Zhao等人采用了随机零阶优化算法（RGF，Random Gradient-Free）[[[68]](#endnote-69)]来解决公式(2.12)的优化问题。此外，他们还发现所提出的基于迁移的攻击方法和基于查询的攻击方法可以组合使用。实验表明，在生成对抗图像过程中加入文本信息能够一定程度上提升其在视觉语言模态任务上的攻击性能，且组合后的方法所生成的对抗样本相比于组合前提升了在不同视觉基础模型上的迁移性。然而，他们只考虑了在适应于同类任务的视觉基础模型上的迁移性，未考虑在不同类任务上的迁移性，且所生成的对抗图像在所测试的大部分视觉基础模型上的攻击成功率仅有50%~60%，这表明在生成视觉模态对抗样本时加入文本信息并不能从根本上改变其在视觉语言模态任务上迁移性差的问题。

以上提到的视觉模态对抗样本生成方法都是将视觉基础模型作为替代模型进行训练，而相比于传统深度视觉模型，视觉基础模型参数规模更大、网络更深，需要足够的内存空间存储模型参数等信息来支持基础模型的运行，且在迭代优化生成对抗样本时，由于网络更深，导致训练的收敛速度更慢，需要更多的时间和计算资源，这使得生成对抗样本的开销增加。

## 视觉语言模态对抗样本生成方法

前面1.1节介绍了现有的适应于多种视觉语言任务的视觉基础模型（如CLIP[5]等），而视觉模态对抗样本在视觉语言模态任务上的迁移性较差。虽然Zhou等人[64]和Zhao等人[65]的工作证明了在视觉模态对抗样本生成过程中加入语言模态的信息，能够一定程度上提升其在视觉语言模态任务上的攻击性能，但是提升仍然有限。从实验数据上来看，现有的视觉模态对抗样本生成方法在视觉语言模态任务上的迁移性仍较差。因此，一些工作尝试在图像和文本上均进行扰动，研究设计面向视觉基础模型的视觉语言模态对抗样本生成方法。

Zhang等人[[[69]](#endnote-70)]首次在视觉基础模型上针对视觉语言模态任务提出了一种多模态交互的攻击方法，称作协同多模态对抗攻击（Co-Attack，Collaborative Multimodal Adversarial Attack）。对于采用融合策略的视觉基础模型，Co-Attack通过使扰动后的多模态嵌入远离原始多模态嵌入来实现协同扰动文本和图像（如图6左），而对于采用对比学习来使视觉与语言模态对齐的视觉基础模型，Co-Attack通过使扰动后的图像嵌入远离扰动后的文本嵌入来实现对单模态嵌入的攻击（如图6右）。具体而言，他们提出先通过BERT-Attack[[[70]](#endnote-71)]生成文本扰动，然后将文本扰动作为文本输入，采用类似PGD的方式生成图像扰动，采用Zhang等人[[[71]](#endnote-72)]所提出的KL（Kullback-Leibler）发散损失来最大化嵌入表示。他们认为Co-Attack相对于不考虑多模态交互的方法能够产生更大的扰动，如图5中。然而Co-Attack以单一图像-文本为输入，对抗样本会对训练的图像-文本对过拟合。



**图6 Co-Attack优化扰动原理**[69]

Co-Attack虽然考虑了视觉模态和语言模态的交互，但是它只考虑了白盒场景。为此，Lu等人[[[72]](#endnote-73)]提出了一种集合级引导攻击（SGA，Set-level Guidance Attack），并首次探索了适应于视觉语言模态任务的视觉基础模型在黑盒场景下的对抗鲁棒性。他们在Co-Attack基础上将单一的图像-文本对扩展为图像集-文本集，并使用来自不同模态的配对数据作为监督信号来引导对抗样本的优化方向。具体而言，SGA通过数据增强将输入的图像扩展为图像集，接着为图像集中的每一张图像匹配相近的多个文本描述构成文本集，然后为文本集中每个文本描述生成对应的对抗文本，形成对抗文本集，再通过对抗文本集优化生成对抗图像。此外，在迭代优化对抗图像和对抗文本的过程中，SGA会逐步拉远图像和文本在特征空间中的距离，从而破坏跨模态交互，以提升对抗样本的迁移性。SGA虽然实现了出色的跨模型迁移性，但是其生成对抗样本的开销非常大，从原本的单个图像-文本对扩展为图像集-文本集，计算量会成倍增加。此外，虽然SGA通过增强方法将单一的图像-文本对扩展为图像集-文本集，但是仍然是以单一的图像-文本作为输入，且采用的数据增强方法并不是生成新信息，而是混合现有信息，这会使得图像集与输入图像更加相关，进而导致生成的对抗样本对输入图像过拟合。

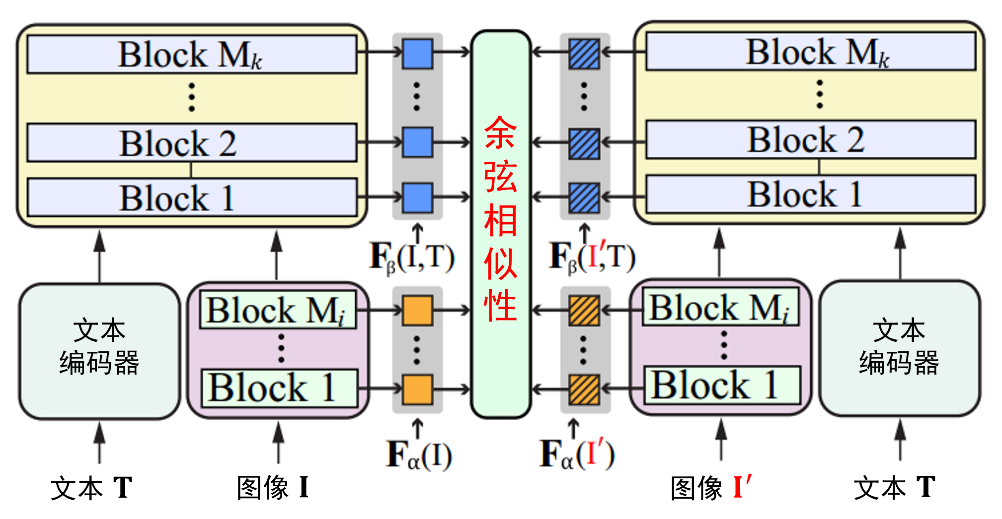
Wang等人[[[73]](#endnote-74)]从增强迁移性的角度同样基于对比学习提出了一种在视觉基础模型上生成视觉语言模态对抗样本的方法。受Guo等人[[[74]](#endnote-75)]启发，他们使用Gumble-softmax分布[[[75]](#endnote-76)]将离散分布的文本转化为一个可微的采样过程，使得能够通过梯度下降优化生成文本扰动。在生成视觉语言模态对抗样本过程中，他们结合了MI-FGSM[33]和DI-FGSM[34]方法来优化图像扰动，不同于Co-Attack方法，他们基于梯度同时生成对抗文本和对抗图像。为了提高视觉语言模态对抗样本的迁移性，他们采用对比学习，包括模态内对比学习和图像-文本对比学习，在不同模态使对抗样本特征远离原始样本。具体而言，他们采用了如下的对比损失函数：

(3.3)

(3.4)

其中为InfoNCE[63]损失，和分别表示对抗文本和对抗图像。在图像-文本对比学习中，以文本为对比视图时，采用干净图像作为正样本，随机文本作为负样本；在以图像为对比视图时，采用干净文本作为正样本，干净图像以及对其增强后的图像作为负样本。在模态内对比学习中，采用随机图像作为正样本，干净图像及对其增强后的图像作为负样本。此外，他们还采用了负对数似然损失来提高对抗文本的流畅性，采用BERT分数[[[76]](#endnote-77)]来约束对抗文本中的语义变化，以避免与原始文本产生较大的语义差异。他们将上述损失合并作为生成多模态对抗样本的最终损失函数，并通过超参分别控制不同损失的约束强度。实验表明，该方法优于Co-Attack，能够在视觉模态和视觉语言模态任务中都具有更高的攻击性能，且在不同视觉基础模型具有良好的迁移性。然而，由于基于单个图像-文本对生成对抗样本，会出现对训练的图像-文本对过拟合的问题。

Yin等人[[[77]](#endnote-78)]提出了一种任务不可知且能够在黑盒场景下攻击视觉基础模型的对抗样本生成方法VLAttack，能够适应于视觉模态和视觉语言模态的多种任务，如图像分类、视觉问答、图像描述、视觉蕴涵等。不同于Co-Attack，VLAttack先对图像进行攻击，他们提出了一种新的块相似性攻击（BSA，Block-wise Similarity Attack）策略。他们通过研究发现，如果预训练模型从干净输入和扰动输入中学习的特征表示显著不同，则这种扰动有可能可以迁移到微调的模型上，以改变其对下游任务的预测。如图7，他们通过最大化预训练模型的图像编码器和Transformer编码器的基于块的中间表示之间的余弦距离来优化图像扰动。此外，他们认为BERT-Attack[70]方法生成的语言模态对抗样本已足够强大，因此直接采用BERT-Attack优化文本扰动。为了充分学习不同模态扰动之间的相互联系，他们提出了一种新的迭代交叉搜索攻击（ICSA，Iterative Cross-Search Attack）策略。具体而言，VLAttack通过多次迭代来优化对抗图像，在第一次对抗图像生成中使用干净文本作为BSA的文本输入，在之后的每次迭代中使用多个对抗文本作为BSA的文本输入。然而，VLAttack过分关注输入图像的特征，并通过迭代加强了对图像整体特征的学习，这会导致对抗样本过拟合到训练的图像。当输入图像改变时，需要重新训练来生成新的对抗样本。



**图7 块相似性攻击（BSA）**[77]

综上所述，现有的视觉语言模态对抗样本生成方法均存在对训练数据过拟合的问题，当输入数据发生改变时，需要重新计算扰动来生成与之对应的对抗样本。此外，视觉语言模态对抗样本生成方法大都基于视觉基础模型训练生成对抗样本，同样存在训练开销大的问题。

总体而言，虽然已有一些工作在视觉基础模型上针对对抗样本的迁移性进行了研究，但面向视觉基础模型的对抗样本生成的研究还处于早期阶段，仍有许多问题尚未得到探索。

# 发展趋势和现存的问题

对于面向视觉基础模型的对抗样本生成方法，目前已有的工作可以归纳为两个方面：

1. 仅扰动图像的视觉模态对抗样本生成；
2. 对图像和文本均进行扰动的视觉语言模态对抗样本生成。

对于第一个方面，现有的视觉模态对抗样本生成方法主要以提升在多个任务上的迁移性和在黑盒场景下的迁移性为目标。然而，现有的方法在视觉基础模型上生成的视觉模态对抗样本仅对一类视觉任务有效，在不同类视觉任务上攻击性能较差，其跨任务迁移性受限。此外，视觉模态对抗样本在视觉语言模态任务上的迁移性也较差。未来的工作主要面向提升对抗样本的跨任务迁移性。

对于第二个方面，现有的视觉语言模态对抗样本生成方法主要以加强不同模态间的交互性和提升在适应不同类型任务的视觉基础模型上的迁移性为目标，但是现有的采用多模态对比学习的方法大多以图像为中心，这会使得对抗样本对训练的图像或图像-文本对过拟合，影响攻击性能。因此，未来的工作将会关注对通用扰动表征的学习，以及解决对抗样本对训练数据过拟合的问题。

此外，现有的方法均以基础模型作为替代模型，而在基础模型上训练的开销过大，计算成本高，且基础模型的运行对硬件有一定要求。因此，未来的工作将寻求训练成本更低且不降低对抗样本攻击性能的替代模型，来减小生成对抗样本的计算开销。

# 结论

随着基础模型范式的普及，越来越多的视觉基础模型涌现，然而其安全性也逐渐成为关注的焦点。因此，一些研究人员展开对视觉基础模型的对抗鲁棒性的研究，面向视觉基础模型的对抗样本生成是研究视觉基础模型在强对抗环境下的实际性能的主要方法之一。本篇综述首先概述了视觉基础模型的发展现状，然后详细介绍了当前主流的面向传统视觉模型的对抗样本生成方法，并总结了在视觉基础模型上的对抗样本生成相关的工作，最后分析了当前该领域的不足，提出了展望。希望本综述的见解能够帮助研究者探索在视觉基础模型上生成更有效的对抗样本的方法。

# 主要参考文献

1. [] Devlin J, Chang M, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C] // North American Chapter of the Association for Computational Linguistics (2019). NAACL, 2019. [↑](#endnote-ref-2)
2. [] Ramesh A, Pavlov M, Goh G, et al. Zero-Shot Text-to-Image Generation[C] // Proceedings of the 38th International Conference on Machine Learning, PMLR 139:8821-8831, 2021. [↑](#endnote-ref-3)
3. [] Brown T, Mann B, et al. Language Models are Few-Shot Learners[C] // 34th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2020, pp. 1877-1901. [↑](#endnote-ref-4)
4. [] Bommasani R, Hudson D, Adeli E, Altman R, et al. On the opportunities and risks of foundation models[R/OL]. https://crfm.stanford.edu/report.html. 2021. [↑](#endnote-ref-5)
5. [] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners[J]. OpenAI blog, 1(8):9, 2019. [↑](#endnote-ref-6)
6. [] Kirillov A, Mintun E, Ravi N, et al. Segment anything[EB/OL]. arXiv:2304.02643, 2023. [↑](#endnote-ref-7)
7. [] Singh A, Hu R, Goswami V, Couairon G, et al. FLAVA: A Foundational Language And Vision Alignment Model[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 15638-15650. [↑](#endnote-ref-8)
8. [] Lüddecke T, Ecker Alexander S. Image Segmentation Using Text and Image Prompts [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 7086-7096 [↑](#endnote-ref-9)
9. [] Vaswani A, Shazeer N, Paramar N, et al. Attention is All You Need[C] // 31th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2017:6000-6010. [↑](#endnote-ref-10)
10. [] Jia M, Tang L, Chen B, et al. Visual prompt tuning[C] // European Conference on Computer Vision (ECCV). Springer, 2022:709-727. [↑](#endnote-ref-11)
11. [] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C] // 9th International Conference on Learning Representations (ICLR). ICLR, 2021. [↑](#endnote-ref-12)
12. [] Chen S, Ge C, Tong Z, et al. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition[C] // 36th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2022. [↑](#endnote-ref-13)
13. [] Jie S, Deng Z. Convolutional Bypasses Are Better Vision Transformer Adapters [EB/OL]. arXiv: 2207.07039, 2022. [↑](#endnote-ref-14)
14. [] Ma J, He Y, Li F, Han L, et al. Segment Anything in Medical Images[EB/OL]. arXiv:2304.12306, 2023. [↑](#endnote-ref-15)
15. [] Shaharabany T, Dahan A, Giryes R, Wolf L. AutoSAM: Adapting SAM to Medical Images by Overloading the Prompt Encoder[EB/OL]. arXiv:2306.06370, 2023. [↑](#endnote-ref-16)
16. [] Cheng Y, Li L, Xu Y, et al. Segment and Track Anything[EB/OL]. arXiv: 2305.06558, 2023. [↑](#endnote-ref-17)
17. [] Yang J, Gao M, Li Z, et al. Track Anything: Segment Anything Meets Videos [EB/OL]. arXiv:2304.11968, 2023. [↑](#endnote-ref-18)
18. [] Chen K, Liu C, Chen H, et al. RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation based on Visual Foundation Model[EB/OL]. arXiv:2306.16269, 2023. [↑](#endnote-ref-19)
19. [] Wang T, Zhang J, Fei J, et al. Caption Anything: Interactive Image Description with Diverse Multimodal Controls[EB/OL]. arXiv:2305.02677, 2023. [↑](#endnote-ref-20)
20. [] Li Y, Fan H, Hu R, et al. Scaling Language-Image Pre-training via Masking[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023, pp. 23390-23400. [↑](#endnote-ref-21)
21. [] Mu N, Kirillov A, Wagner D, et al. SLIP: Self-supervision meets Language-Image Pre-training[C] // European Conference on Computer Vision (ECCV). Springer, 2022:529-544. [↑](#endnote-ref-22)
22. [] Zhou C, Loy C, Dai B. Extract Free Dense Labels from CLIP[C] // European Conference on Computer Vision (ECCV). Springer, 2022:696-712. [↑](#endnote-ref-23)
23. [] Ghiasi G, Gu X, Cui Y, Lin T. Scaling Open-Vocabulary Image Segmentation with Image-Level Labels[C] // European Conference on Computer Vision (ECCV). Springer, 2022:540-557. [↑](#endnote-ref-24)
24. [] Li L, Zhang P, Zhang H, Yang J, et al. Grounded Language-Image Pre-training[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 10965-10975. [↑](#endnote-ref-25)
25. [] Liu S, Zeng Z, Ren T, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection[EB/OL]. arXiv:2303.05499, 2023. [↑](#endnote-ref-26)
26. [] Caron M, Touvron H, Misra I, et al. Emerging Properties in Self-Supervised Vision Transformers[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021, pp. 9650-9660. [↑](#endnote-ref-27)
27. [] Yu J, Wang Z, Vasudevan V, et al. CoCa: Contrastive Captioners are Image-Text Foundation Models[EB/OL]. arXiv:2205.01917, 2022. [↑](#endnote-ref-28)
28. [] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Conv-olutional Neural Networks[C] // 26th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2012. [↑](#endnote-ref-29)
29. [] Szegedy C, Zaremba C, Sutskever I, Bruna J, et al. Intriguing Properties of Neural Networks[EB/OL]. arXiv:1312.6199, 2013. [↑](#endnote-ref-30)
30. [] Goodfellow I, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples [EB/OL]. arXiv:1412.6572, 2014. [↑](#endnote-ref-31)
31. [] Miyato T, Maeda S, Koyama M, Ishii S. Virtual Adversarial Training: A Regula-rization Method for Supervised and Semi-Supervised Learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.41, no.8, pp. 1979-1993, 1 Aug. 2019. [↑](#endnote-ref-32)
32. [] Kurakin A, Goodfellow I, Bengio S. Adversarial Examples in the Physical World[C] // 5th International Conference on Learning Representations (ICLR). ICLR, 2017. [↑](#endnote-ref-33)
33. [] Dong Y, Liao F, Pang T, et al. Boosting Adversarial Attacks With Momentum[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018, pp. 9185-9193. [↑](#endnote-ref-34)
34. [] Xie C, Zhang Z, Zhou Y, et al. Improving Transferability of Adversarial Examples With Input Diversity[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019, pp. 2730-2739. [↑](#endnote-ref-35)
35. [] Dong Y, Pang T, Su H, Zhu J. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019, pp. 4312-4321. [↑](#endnote-ref-36)
36. [] Lin J, Song C, He K, et al. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks[C] // 8th International Conference on Learning Representations (ICLR). ICLR, 2020. [↑](#endnote-ref-37)
37. [] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[C] // 6th International Conference on Learning Representations (ICLR). ICLR, 2018. [↑](#endnote-ref-38)
38. [] Moosavi-Dezfooli, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016, pp. 2574-2582. [↑](#endnote-ref-39)
39. [] Rony J, Hafemann L, Oliveira L, et al. Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019, pp. 4322-4330. [↑](#endnote-ref-40)
40. [] Papernot N, McDaniel P, Wu X, et al. Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks[C] // 37th IEEE Symposium on Security and Privacy (SP). IEEE, 2016, pp. 582-597. [↑](#endnote-ref-41)
41. [] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C] // 38th IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 39-57. [↑](#endnote-ref-42)
42. [] Gu J, Zhao H, et al. SegPGD: An Effective and Efficient Adversarial Attack for Evaluating and Boosting Segmentation Robustness[C] // European Conference on Computer Vision (ECCV). Springer, 2022:308-325. [↑](#endnote-ref-43)
43. [] Chow K, Liu L, et al. TOG: Targeted Adversarial Objectness Gradient Attacks on Real-time Object Detection Systems[EB/OL]. arXiv:2004.04320, 2020. [↑](#endnote-ref-44)
44. [] Xu X, Chen X, Liu C, Rohrbach A, et al. Fooling Vision and Language Models Despite Localization and Attention Mechanism[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018, pp. 4951-4961. [↑](#endnote-ref-45)
45. [] Li Y, Bai S, Xie C, et al. Regional Homogeneity: Towards Learning Transferable Universal Adversarial Perturbations Against Defenses[C] // European Conference on Computer Vision (ECCV). Springer, 2020:795-813. [↑](#endnote-ref-46)
46. [] Wang X, He X, Wang J, He K. Admix: Enhancing the Transferability of Adver-sarial Attacks[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021, pp. 16158-16167. [↑](#endnote-ref-47)
47. [] Byun J, Cho S, Kwon M, Kim H, Kim C. Improving the Transferability of Targeted Adversarial Examples Through Object-Based Diverse Input[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 15244-15253. [↑](#endnote-ref-48)
48. [] Wang X, He K. Enhancing the Transferability of Adversarial Attacks Through Variance Tuning[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021, pp. 1924-1933. [↑](#endnote-ref-49)
49. [] Ma W, Li Y, Jia X, Xu W. Transferable Adversarial Attack for Both Vision Transformers and Convolutional Networks via Momentum Integrated Gradients[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023, pp. 4630-4639. [↑](#endnote-ref-50)
50. [] Zhang C, Benz P, Karjauv A, et al. Investigating Top-k White-Box and Trans-ferable Black-Box Attack[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 15085-15094. [↑](#endnote-ref-51)
51. [] Zhao Z, Liu Z, Larson M. On Success and Simplicity: A Second Look at Trans-ferable Targeted Attacks[C] // 35th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2021, pp. 6115-6128. [↑](#endnote-ref-52)
52. [] Benz P, Zhang C, Kweon I. Batch Normalization Increases Adversarial Vulner-ability and Decreases Adversarial Transferability: A Non-Robust Feature Perspective[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021, pp. 7818-7827. [↑](#endnote-ref-53)
53. [] Wu D, Wang Y, Xia S, et al. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets[EB/OL]. arXiv:2022.05990, 2020. [↑](#endnote-ref-54)
54. [] Guo Y, Li Q, Chen H. Backpropagating Linearly Improves Transferability of Adv-ersarial Examples[C] // 34th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2020, pp. 85-95. [↑](#endnote-ref-55)
55. [] Liu Y, Chen X, Liu C, Song D. Delving into Transferable Adversarial Examples and Black-box Attacks[C] // 5th International Conference on Learning Representations (ICLR). ICLR, 2017. [↑](#endnote-ref-56)
56. [] Chen B, Yin J, Chen S, et al. An Adaptive Model Ensemble Adversarial Attack for Boosting Adversarial Transferability[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023, pp. 4489-4498. [↑](#endnote-ref-57)
57. [] Moosavi-Dezfooli S, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017, pp. 1765-1773. [↑](#endnote-ref-58)
58. [] Fort S. Adversarial examples for the OpenAI CLIP in its zero-shot classification regime and their semantic generalization[DB/OL]. https://stanislavfort.com/blo-g/OpenAI\_CLIP\_adversarial\_examples. 2021. [↑](#endnote-ref-59)
59. [] Zhang C, Zhang C, Kang T, Kim D. Attack-SAM: towards attacking segment anything model with adversarial examples[EB/OL]. arXiv:2305.00866, 2023. [↑](#endnote-ref-60)
60. [] Zheng S, Zhang C. Black-box targeted adversarial attack on segment anything (SAM)[EB/OL]. arXiv:2310.10010, 2023. [↑](#endnote-ref-61)
61. [] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features[C]//33th Conference on Neural Information Processing Systems (NeurIPS). NIPS, 2019:125-136. [↑](#endnote-ref-62)
62. [] Han D, Zheng S, Zhang C. Segment Anything Meets Universal Adversarial Perturbation[EB/OL]. arXiv:2310.12431, 2023. [↑](#endnote-ref-63)
63. [] Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding[EB/OL]. arXiv:1807.03748, 2019. [↑](#endnote-ref-64)
64. [] Zhou Z, Hu S, Li M, et al. AdvCLIP: Downstream-agnostic Adversarial Examples in Multimodal Contrastive Learning[C] // Proceedings of the 31th ACM International Conference on Multimedia (MM). ACM MM, 2023, pp. 6311-6320. [↑](#endnote-ref-65)
65. [] Zhao Y, Pang T, Du C, et al. On Evaluating Adversarial Robustness of Large Vision-Language Models[EB/OL]. arXiv:2305.16934, 2023. [↑](#endnote-ref-66)
66. [] Yuksekgonul M, Bianchi, et al. When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?[C] // 11th International Conference on Learning Representations (ICLR). ICLR, 2023. [↑](#endnote-ref-67)
67. [] Rombach R, Blattmann A, Lorenz D, et al. High-Resolution Image Synthesis With Latent Diffusion Models[C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022, pp. 10684-10695. [↑](#endnote-ref-68)
68. [] Nesterov Y, Spokoiny V. Random Gradient-Free Minimization of Convex Func-tions[J]. Foundations of Computational Mathematics, 2017, pp. 527-566. [↑](#endnote-ref-69)
69. [] Zhang J, Yi Q, Sang J. Towards Adversarial Attack on Vision-Language Pre-training Models[C] // Proceedings of the 30th ACM International Conference on Multimedia (MM). ACM MM, 2022, pp. 5005-5013. [↑](#endnote-ref-70)
70. [] Li L, Ma R, Guo Q, et al. BERT-ATTACK: Adversarial Attack Against BERT Using BERT[C] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). EMNLP, 2020, pp. 6193-6202. [↑](#endnote-ref-71)
71. [] Zhang H, Yu Y, Jiao J, et al. Theoretically Principled Trade-off between Robustness and Accuracy[C] // Proceedings of the 36th International Conference on Machine Learning, PMLR 97:7472-7482, 2019. [↑](#endnote-ref-72)
72. [] Lu D, Wang Z, Wang T, et al. Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Language Pre-training Models[C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023, pp. 102-111. [↑](#endnote-ref-73)
73. [] Wang Y, Hu W, Dong Y, et al. Exploring Transferability of Multimodal Adversarial Samples for Vision-Language Pre-training Models with Contrastive Learning[EB/OL]. arXiv:2308.12636, 2023. [↑](#endnote-ref-74)
74. [] Guo C, Sablayrolles A, et al. Gradient-based Adversarial Attacks against Text Transformers[EB/OL]. arXiv:2104.13733, 2021. [↑](#endnote-ref-75)
75. [] Jang E, Gu S, Poole B. Categorical Reparameterization with Gumbel-Softmax [EB/OL]. arXiv:1611.01144, 2016. [↑](#endnote-ref-76)
76. [] Zhang T, Kishore V, Wu F, et al. BERTScore: Evaluating Text Generation with BERT[EB/OL]. arXiv:1904.09675, 2020. [↑](#endnote-ref-77)
77. [] Yin Z, Ye M, Zhang T, et al. VLAttack: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models[EB/OL]. arXiv:2310.04655, 2023. [↑](#endnote-ref-78)