中国科学院大学

University of Chinese Academy of Sciences

硕士学位论文

基于图文对齐的文本生成技术研究

作者姓名：杨哲

指导教师：冯洋 研究员

学位类别：工学硕士

学科专业：计算机科学与技术

培养单位：中国科学院计算技术研究所

2025 年 6 月

Research on Text Generation Technology Based on Image-Text Alignment

A thesis submitted to

University of Chinese Academy of Sciences

in partial fulfillment of the requirement

for the degree of

Master of Engineering

in Computer Science and Technology

By

YANG Zhe

Supervisor: Professor FENG Yang

Institute of Computing Technology, Chinese Academy of Sciences

June, 2025

中国科学院大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。承诺除文中已经注明引用的内容外，本论文不包含任何其他个人或集体享有著作权的研究成果，未在以往任何学位申请中全部或部分提交。对本论文所涉及的研究工作做出贡献的其他个人或集体，均已在文中以明确方式标明或致谢。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日　　期：

中国科学院大学
学位论文授权使用声明

本人完全了解并同意遵守中国科学院大学有关收集、保存和使用学位论文的规定，即中国科学院大学有权按照学术研究公开原则和保护知识产权的原则，保留并向国家指定或中国科学院指定机构送交学位论文的电子版和印刷版文件，且电子版与印刷版内容应完全相同，允许该论文被检索、查阅和借阅，公布本学位论文的全部或部分内容，可以采用扫描、影印、缩印等复制手段以及其他法律许可的方式保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

摘 要

文本生成任务是自然语言处理领域的经典任务之一，其目标是让计算机根据给定的输入自动生成符合要求的自然语言文本。近年来，随着深度学习和大规模预训练模型的发展，文本生成技术在多个领域得到了广泛应用，如机器翻译、视觉问答等。然而，在低资源翻译、视觉知识问答等更具有挑战性的文本生成任务中，由于文本信息的不充足，生成质量往往难以保证。当可用文本信息不足以支持任务需求时，如何保证模型的生成质量仍然是该领域的重要研究课题。

近年来，多模态数据的涌现及多模态技术的快速发展为文本生成任务提供了新的研究思路。当前，互联网上存在大量的图片-文本描述数据，且在如社交媒体等特定场景中，这类数据易于获取，这为利用视觉信息辅助文本生成提供了便利。而图文对齐作为多模态理解的核心技术之一，能够有效利用视觉信息增强文本生成的准确性和丰富性。因此，本文聚焦如何有效利用图片-文本描述数据提升模型的生成能力，递进式地将图文对齐技术引入三种不同的文本信息受限场景：（1）低资源机器翻译：该场景需要在低资源语言平行语料受限的条件下，利用显式的图片-文本描述数据对齐源端低资源语言，并使用有限的平行语料进一步实现源端至目标端的映射;（2）无监督机器翻译：该场景基于无监督方法，需要在不存在任何平行语料的条件下，仅使用图片-文本描述数据辅助建立源端与目标端之间的映射;（3）视觉知识问答：该场景的问答需要外部图片作为额外知识进行辅助，因此需要在没有显式图片-文本描述数据的条件下，使模型自主学习问题和图像的相关性区域，从而提高模型回答准确率。

针对上述三种任务场景，本文基于图文对齐技术，沿着利用图片-文本描述数据建立源端映射，同时使用平行语料进一步建立源端至目标端映射，到仅使用图片-文本描述数据建立源端至目标端映射，最后到自主学习图文相关性区域辅助模型生成这条由易到难的研究路线，依次探索和验证了相应的解决方案。具体取得了如下创新成果：

1、图片辅助的低资源机器翻译

针对低资源条件下神经机器翻译模型因缺乏大规模平行语料而翻译效果较差的问题，本文提出了图片辅助的低资源机器翻译方法。尽管现有的神经机器翻译模型已经取得优越的性能，但其往往依赖大规模平行语料，而这对于低资源语言而言成本过高。本文通过在低资源翻译场景中引入图片，在源端通过图文对齐构建多个低资源语言与特定高资源语言之间的公共语义空间，同时使用高资源语言的平行语料建立源端至目标端的映射，最后实现高资源语言翻译能力到低资源语言的迁移。实验结果表明，结合少量图文对即可有效实现跨模态与跨语言

的语义对齐，在零样本和少样本场景下较纯文本基线模型均取得显著性能提升。

2、基于图文对齐的无监督机器翻译

针对无监督机器翻译中缺乏平行语料实现源端至目标端对齐的问题，本文提出基于图文对齐的无监督机器翻译方法。无监督机器翻译是指在无需任何平行语料的情况下实现源端至目标端对齐，然而，如何在特征表示空间内实现源语言与目标语言的对齐仍是该领域的核心挑战。尽管不同语言的文本表征存在差异，但其视觉描述往往具有共通性。受此启发，本文将图片引入无监督机器翻译，以图片作为源端和目标端的公共语义锚点，借助对比学习方法，同时将源端文本与目标端文本与图片进行对齐，从而实现在不使用平行语料的条件下，仅利用单语图片-文本描述数据建立源端至目标端映射的目标。实验表明，该方法仅需单语图文数据即可实现源语言到目标语言的语义对齐，翻译的效果较现有的纯文本基线模型与多模态无监督机器翻译方法均有显著提升。

3、基于图片后验检索增强的视觉知识问答

针对视觉知识问答场景中，模型检索对应图片数据时检索不准确，生成答案时对图片关注不足的问题，本文提出基于图片后验检索增强的视觉知识问答方法。在视觉知识问答场景中，模型需要检索并引入外部的图片进行辅助并生成答案。现有方法往往直接使用问题文本直接进行检索匹配，由于问题和图像并不存在显式的描述对应关系，因此此类检索方法往往精度较低，且生成效果较差。基于此，本文在训练阶段引入包含答案的后验信息，辅助模型学习到图像中和问题更相关的区域，从而使模型能够自主提取图片的相关性区域，提高生成答案的准确性。本文从检索和生成两个方面进行实验验证，实验结果表明，在检索方面，该方法相较于现有方法能实现更精确的检索；在生成方面，该方法在答案准确率上相比于基线模型有显著提升。

综上所述，本文以提升文本信息受限场景下模型的文本生成能力为核心，从图文对齐的角度出发，依次探索存在图片-文本描述数据与平行语料数据的低资源翻译场景、仅存在图片-文本描述数据的无监督翻译场景以及不存在显式图片-文本描述数据的视觉知识问答场景，对其中存在的多个关键问题设计了相应的解决方案。本文通过逐步减少模型对数据的依赖，提高模型对图片信息的提取能力，增强模型在稀缺标注数据训练条件下的表现。希望本文能够为图文对齐技术在更多现实文本生成场景中的落地给予一定的启发。

关键词：多模态，图文对齐，文本生成，低资源，机器翻译，检索增强

Abstract

Text generation is one of the classic tasks in the field of natural language processing (NLP), with the goal of enabling computers to automatically produce grammatically and semantically correct natural language text based on given inputs. In recent years, with the advancement of deep learning and large-scale pre-trained language models, text generation techniques have been widely applied in various domains, such as machine translation and visual question answering. However, in more challenging text generation tasks like low-resource translation and visual knowledge-based question answering, the quality of generated text often suffers due to insufficient textual information. When the available text is inadequate to meet task requirements, ensuring the generation quality of models remains a critical research topic in this field.

The recent emergence of multimodal data and the rapid development of multimodal technologies have provided new research directions for text generation tasks. Currently, large amounts of image-text paired data are available on the internet, and in specific scenarios such as social media and multimodal encyclopedias, such data is even more accessible, facilitating the use of visual information to assist text generation. As one of the core technologies in multimodal understanding, *image-text alignment* can effectively leverage visual information to enhance the accuracy and richness of generated text. Therefore, this paper focuses on how to utilize image-text paired data to improve model generation capabilities, progressively introducing image-text alignment techniques into three distinct scenarios with limited textual information:

(1) **Low-resource machine translation**: This scenario requires aligning the source low-resource language with explicit image-text descriptions under the constraint of limited parallel corpora, while using scarce parallel data to further establish mappings from the source to the target language.

(2) **Unsupervised machine translation**: Based on unsupervised methods, this scenario necessitates establishing mappings between source and target languages solely using image-text descriptions without any parallel corpora.

(3) **Visual knowledge-based question answering**: In this scenario, answering questions requires external images as supplementary knowledge. Thus, the model should autonomously learn the relevant regions between questions and images to improve answer accuracy in the absence of explicit image-text descriptions.

For these three task scenarios, this paper follows a research trajectory from easy

to challenging—first using image-text descriptions to establish source-side mappings while employing parallel corpora for source-to-target mappings, then relying solely on image-text descriptions for direct source-to-target alignment, and finally enabling autonomous learning of image-text relevance to assist generation. Corresponding solutions are explored and validated, yielding the following innovative contributions:

1. Image-Assisted Low-Resource Machine Translation

To address the poor translation performance of neural machine translation (NMT) models in low-resource settings due to the lack of large-scale parallel corpora, this paper proposes an image-assisted low-resource machine translation method. Although existing NMT models have achieved remarkable performance, they heavily rely on large parallel datasets, which are prohibitively expensive for low-resource languages. By introducing images into low-resource translation, this work constructs a shared semantic space between multiple low-resource languages and a specific high-resource language through image-text alignment. Parallel corpora of the high-resource language are then used to establish source-to-target mappings, ultimately transferring high-resource translation capabilities to low-resource languages. Specifically, a coarse-to-fine contrastive learning approach is designed to align sentence-level and word-level representations across languages. Experimental results show that even with minimal image-text pairs, cross-modal and cross-lingual semantic alignment can be effectively achieved, significantly outperforming text-only baselines in zero-shot and few-shot scenarios.

2. Unsupervised Machine Translation via Image-Text Alignment

To tackle the lack of parallel corpora for source-to-target alignment in unsupervised machine translation (UMT), this paper proposes an image-text alignment-based UMT method. UMT aims to align source and target languages without any parallel data, yet achieving this alignment in feature representation space remains a core challenge. While textual representations vary across languages, their visual descriptions often share commonalities. Inspired by this, images are introduced as shared semantic anchors between source and target languages. Using contrastive learning, both source and target texts are aligned with images, enabling alignment in a shared semantic space without parallel corpora. Experiments demonstrate that this method achieves superior semantic alignment using only monolingual image-text data, outperforming existing text-only baselines and multimodal UMT approaches.

3. Posterior Image Retrieval-Augmented Visual Knowledge Question Answering

To address the issues of inaccurate image retrieval and insufficient visual attention

in visual knowledge-based question answering (VQA), this paper proposes a posterior image retrieval-augmented VQA method. In VQA, models must retrieve and incorporate external images to generate answers. Existing methods typically retrieve images directly using question text, but since questions and images lack explicit descriptive correspondences, retrieval accuracy and generation quality are often poor. To mitigate this, posterior information containing answers is introduced during training, serving as a bridge between images and questions to help the model focus on more relevant visual regions. This allows the model to autonomously extract pertinent image areas, improving answer accuracy. Experiments on retrieval and generation demonstrate that the proposed method achieves higher retrieval precision and significantly improves answer accuracy compared to baseline models.

In summary, this paper centers on enhancing text generation capabilities in scenarios with limited textual information. From the perspective of image-text alignment, it progressively explores low-resource machine translation (with image-text descriptions and parallel data), unsupervised machine translation (with only image-text descriptions), and visual knowledge-based QA (without explicit image-text descriptions), devising solutions for key challenges in each. By gradually reducing model reliance on annotated data while improving visual information extraction, the proposed approaches strengthen model performance under data-scarce conditions. It is hoped that this work will inspire further applications of image-text alignment in real-world text generation scenarios.

Key Words: Multimodal learning, Image-text alignment, Low-resource translation, Machine translation, Retrieval augmentation

目 录

图目录

表目录

第 1 章　绪论

1.1　研究背景与意义

文本生成作为自然语言处理领域的重要分支，其核心目标是赋予计算机系统理解和生成符合人类语言规范文本的能力。这一技术通过建模语言的语法结构、语义关系和语用特征，使机器能够根据结构化数据、非结构化文本或跨模态输入，自动产生符合任务目标，流畅、连贯且符合上下文语境的自然语言输出。近年来，得益于深度学习架构的革新和大规模预训练语言模型（如 GPT、BERT 等）的发展，文本生成技术取得了突破性进展，例如在机器翻译领域，神经机器翻译系统已能实现接近人类水平的翻译质量；而在视觉问答任务中，多模态模型可以结合图像内容生成准确的文本回答。然而，当前技术仍面临诸多挑战性场景的考验。以低资源语言翻译为例，由于平行语料匮乏，模型难以学习到有效的跨语言映射关系；在需要外部知识支持的视觉问答任务中，仅依靠文本输入往往无法获取完整的背景信息。更本质的问题在于，当文本模态提供的信息不足以支撑完整语义理解时，现有系统的生成质量会出现显著下降。

近年来，随着多模态数据的爆炸式增长和跨模态技术的突飞猛进，文本生成领域迎来了全新的发展机遇。在海量互联网数据中，高质量的图片-文本描述数据呈现出规模化、多样化的特点，特别是在视觉社交媒体平台和开放式百科等场景中，这类跨模态数据不仅易于获取，更具有丰富的语义关联性。在这一背景下，图文对齐技术作为多模态理解的关键突破口，展现出其独特的价值。该技术通过深度挖掘视觉特征与语言表征之间的潜在关联，构建起跨模态的语义映射桥梁，从而在多个维度上优化文本生成效果：一方面，通过视觉线索的引入，为模型提供视觉模态的辅助信息，显著提升了生成内容的准确性；另一方面，丰富的视觉信息为文本生成任务提供了额外的信息补充，提高模型在文本信息不充足的条件下的生成效果。

在图文对齐的技术中，由于不同模态的数据在表现形式上有所差异，其背后的语义信息也呈现出多样性和复杂性。除此以外，不同模态的数据之间并没有直接的语义对齐关系，如何在它们之间构建有效的语义桥梁，成为研究的关键。近年来，随着深度学习和表示学习技术的发展，基于嵌入表示的方法[5–8] 逐渐成为主流。这类方法通过将不同模态的数据映射到同一个潜在的语义空间，从而实现跨模态的对齐。在此过程中，对比学习（contrastive learning）[9] 作为一种自监督学习策略，逐渐受到研究者的关注。对比学习通过最大化正样本对之间的相似性和最小化负样本对之间的相似性，使得模型能够更好地捕捉模态内和模态间的语义差异。在多模态数据处理场景中，对比学习不仅能够提升嵌入空间的语义一

致性，还能够有效减少不同模态之间的语义鸿沟。通过将相似样本对（例如，图像与其对应的文本描述）拉近，并将无关的样本对分离开来，对比学习为多模态语义对齐提供了一个有效的优化路径。

基于对比学习的方法，已有的研究工作在各个不同的具体任务中取得了一些进展。一些方法如 CLIP 模型[1] 通过利用卷积神经网络（CNN）[10] 或视觉 Transformer 模型[2] 提取图片的视觉特征，结合基于 Transformer[3] 架构的语言模型，将文本和图片嵌入到一个共享的语义空间中。这类方法的核心思想是通过训练数据的配对，学习不同模态之间的对应关系。与 CLIP 模型不同的是，BLIP2 模型[11] 提出了 Qformer 模型，通过初始化一个可学习的定长序列，将该序列与图片特征进行注意力计算，用该序列提取图像模态中所蕴涵的特征，再将该序列与文本序列拼接后输入模型，以一种间接的方式，将图片和文本对齐至同一空间中 Huang 等[12] 则利用多模态之间的对齐方法，以图片模态为桥梁，实现了跨语言的知识迁移，实现不同语言之间知识的共享。Xu 等[13] 以视频-文本理解为应用场景，以视频中的视觉信息为核心，实现了多种模态之间的对齐。在机器翻译场景中，Nakayama 等[14], Li 等[15] 利用以图片模态为核心的思想，将跨模态技术引入零样本（zero-shot）和少样本（few-shot）翻译中，将多模态数据作为文本数据的补充，实现源端和目标端语言之间的映射。在视觉知识问答场景中，Tan 等[16] 则运用 CLIP 对齐图文模态的特点，将 CLIP 作为多模态特征的提取工具，以 CLIP 提取的多模态特征作为检索的索引，提高多模态大模型的生成效果。

由此可见，图文对齐技术在文本生成任务中具有广泛的应用场景。例如，在图像标注与自动生成描述任务中，研究者可以通过图像对齐文本描述，使得生成的文本与图片内容在语义上保持一致。同样，在低资源翻译场景下，图文对齐可以作为辅助手段建立源端语言到目标端语言的映射。此外，该技术也可应用于其他场景与任务下，在一些智能系统中，比如自主驾驶、智能家居等，多模态数据的融合与对齐也是提升系统智能化水平的重要途径。然而，现有的图文对齐技术仍然面临诸多挑战。多模态数据的异构性、稀疏性和高维性导致了不同模态之间的语义差异显著，传统的对齐方法难以应对这些问题。例如，相同的视觉特征可能对应多种正确的文本语义描述，如何确保语义对齐的精确性和泛化能力仍是研究中的一个重要难题。

综上所述，图文对齐技术为文本生成任务提供了新的解决路径。它通过借助图片的视觉信息为文本生成模型提供额外信息，有望提高低资源机器翻译、无监督机器翻译、视觉知识问答等任务中的性能。然而，这一技术也面临着诸多挑战，包括如何提升语义对齐的精度、如何降低对大量配对数据的依赖、如何处理复杂的语义场景等。因此，深入研究并发展更加高效的基于图文对齐的文本生成技术，不仅具有重要的理论价值，也具有广泛的实际应用前景。

1.2　研究目标

在研究背景与意义这一章节中，已经详细地介绍了基于图文对齐的文本生成技术背景。为了对基于图文对齐的文本生成技术进行深入探索，本文递进式地将图文对齐技术引入三种不同的文本信息受限场景即低资源机器翻译、无监督机器翻译、视觉知识问答场景，针对不同场景中的不同挑战，提出相应的解决方案，提高模型文本生成的能力。同时，总结其中的经验，最终将其应用于更多场景、实现更具应用性的以图文对齐技术为基础的文本生成任务系统。针对三个不同的应用场景，本文分别设立了以下研究目标：

（1）低资源机器翻译场景：目前的通用神经机器翻译模型往往依赖于大规模的高质量平行语料数据进行训练。而对于大部分低资源语种来说，获取大规模高质量平行语料的代价过高，并不具备可行性，这也导致了在低资源场景下，翻译模型效果往往较差的问题。为了解决该问题，当前的方法采用了第三种语言作为中间媒介，获取“源端-中间端”，“中间端-目标端”的平行语料，从而减小语料的获取难度。然而，该方法仍然需要额外的平行语料，并没有在根本上解决低资源场景下数据获取困难的问题。因此，如何在低资源场景下用更小的代价借助额外的信息，尽可能地提高模型的翻译能力，是该场景下的研究目标。

（2）无监督机器翻译场景：无监督机器翻译场景是一种数据需求量更低，应用更广泛的低资源场景，该场景下不存在任何平行语料。无监督机器翻译可以通过迭代式的回译，在完全不需要任何平行语料的条件下实现源端到目标端的翻译。但这种方法需要在进行迭代回译之前，在源端和目标端之间建立一个粗略的初步映射，且最后的翻译效果与映射的效果直接相关。而图像作为一种与语种无关的信息，在辅助文本进行语义对齐方面具有充分潜力，通过图片辅助建立源端语言和目标端语言之间的映射，从而提高无监督机器翻译场景的翻译效果，是无监督机器翻译场景下的研究目标。

（3）视觉知识问答场景：视觉知识问答需要模型根据问题检索相关的额外图片知识，辅助模型生成对应的答案。在该任务中，检索和生成是关键的两个组成部分。而由于在该任务中不存在显式的图片-文本描述数据，且目前的多模态检索方法较为粗略，检索精度不高，检索的准确性对最终大模型的输出质量具有关键性的影响。因此，该任务亟需建立更精确更有效的多模态检索模型。与此同时，有关生成模型的优化方法也较为匮乏，大部分研究者并未对该部分展开相关研究。基于此，建立更强更准确的多模态检索模型，同时提高大模型在视觉知识问答场景下的生成能力，是该场景下的研究目标。

1.3　主要挑战

本节根据具体研究场景的不同，分别分析各个场景中所面临的研究难点。

（1）低资源机器翻译场景：在低资源场景下，所面对的最大问题便是数据不足。神经机器翻译的核心在于建立源端语言到目标端语言的映射，然而当平行语料不充足时，便无法有效建立源端到目标端的映射。而常用的方法如迁移学习仍无法摆脱平行语料的掣肘。与此同时，在一些利用多模态信息的解决方案中，图片往往需要作为输入的一部分参与训练，导致测试时也需要图片的输入，这严重影响了模型的应用性。如何正确利用图片信息，使其辅助文本模型实现语义对齐的同时不影响模型本身的应用性，是该场景下的主要研究难点。

（2）无监督机器翻译场景：在无监督机器翻译场景中，核心挑战在于完全缺乏平行数据，仅依赖单语数据进行跨语言对齐，这导致语义空间对齐困难和模型初始化缺乏监督信号。研究需突破的关键点包括通过单语数据构建语言间的公共语义空间，利用合适的自监督任务（如去噪自编码）强化对齐能力。

（3）视觉知识问答场景：对于视觉知识问答场景，挑战主要源于在该任务场景下，不存在显式的图片-文本描述数据，即对于图片，只有对应的问题文本，而不存在详细的描述性文本。在该前提下，会存在更大的模态鸿沟，跨模态对齐复杂度更高，需捕捉不同模态之间的隐含关联。而高相似性的文本以及包含重复特征元素的图像也对该任务带来了巨大的挑战。因此，在该场景下，需要模型自主学习图片与文本之间的相关性区域，建立更精确的检索方法，提高生成模型的生成能力，从而实现多模态信息的互补与高效利用。

1.4 研究内容

1.4.1 研究框架

针对上述三种任务场景，，依次探索和验证了相应的解决方案。具体取得了如下创新成果：本文的研究逻辑如图1-1所示，本文基于图文对齐技术，沿着利用图片-文本描述数据建立源端映射，同时使用平行语料进一步建立源端至目标端映射，到仅使用图片-文本描述数据建立源端至目标端映射，最后到自主学习图文相关性区域辅助模型生成这条由易到难的研究路线，依次进行 (1) 第3章：图片辅助的低资源翻译研究、(2) 第4章：基于图文对齐的无监督机器翻译研究、(3) 第5章：基于图片后验检索增强的视觉知识问答这三个方向的研究。三个研究工作的核心挑战与解决思路存在逐层递进的关系，对数据资源的依赖逐步减小，模型对图片信息的提取能力逐步增强，最终目的是增强模型在标注数据稀缺条件下的文本生成能力。后文将详细介绍各个部分的具体研究内容。

1.4.2 图片辅助的低资源机器翻译

针对低资源条件下神经机器翻译模型因缺乏大规模平行语料而翻译效果较差的问题，本文提出了图片辅助的低资源机器翻译方法。为了在低资源场景下

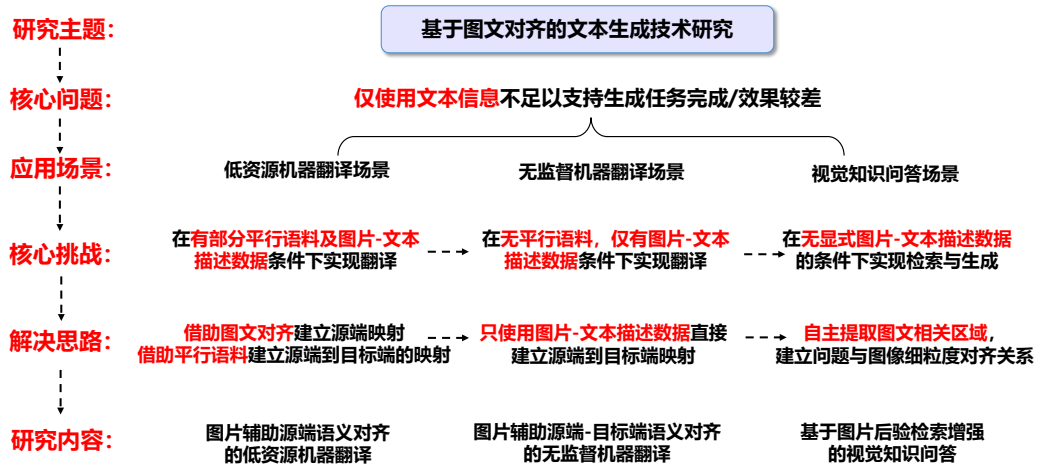


图 1-1　研究脉络图

Figure 1-1　Research Chart

借助图片信息实现从高资源语言到低资源语言的翻译能力迁移，本文提出的解决思路是利用图片-文本描述数据，对齐源端所有语言（包括高资源和低资源语言）的文本特征，并利用跨模态对齐的方法建立不同语种之间的公共语义空间。同时，利用已有的高资源语言平行语料数据建立源端至目标端的映射，从而实现翻译能力迁移的目标。在源端图文对齐的过程中，核心在于利用图片作为中间媒介，使语种不同但语义相同的文本向量表示靠近，而语义不同的向量表示则互相远离。因此，该课题采用对比学习作为跨模态对齐的手段，建立高资源语种和低资源语种之间的公共语义空间。

相对于单一模态的纯文本方法，本研究的核心优势在于充分利用了更易于获得的图片-文本描述数据，引入跨模态对齐方法，大大减小了低资源场景下对数据的需求，显著提高了低资源场景下翻译模型的效果。与此同时，图片在训练的过程中只起到了建立源端语言对齐的作用，并不作为特征输入直接参与模型的训练，因此在测试阶段并不需要图片作为输入，这大大增强了模型的应用性。

1.4.3　基于图文对齐的无监督机器翻译研究

针对无监督机器翻译中缺乏平行语料实现源端至目标端对齐的问题，本文提出基于图文对齐的无监督机器翻译方法。无监督机器翻译的目标是在没有任何平行语料库训练的情况下实现源端语言至目标端语言的映射。一类具有代表性的方法[17–21] 通过单语建模、初步对齐和迭代回译这三个基本组成部分来实现这一目标。单语建模是指在大规模单语语料库上训练模型，以学习不同语言的表征。初步对齐是迭代回译的先导步骤，通过使模型具备粗粒度翻译能力来启动后续流程。之后，利用回译迭代生成伪平行语料，以学习细粒度的源端到目标端对齐映射。正如 Lample 等[18], Huang 等[22] 提到，初步对齐作为迭代回译的起

点，为随后的反向翻译通过迭代完善翻译能力奠定了基础。因此，无监督机器翻译系统的性能在很大程度上取决于高质量的初步对齐。

在初步对齐阶段，当前方法仅仅使用简单的双语对齐词表来进行“词到词”的单一映射，而这远远无法达到高质量映射的要求。且利用双语对齐词表的方法引入了外部的平行语料信息，其本身的获取便存在一定难度，这与无监督机器翻译的场景不符。因此，本研究的具体目标在于如何在仅有单语语料的情况下建立高质量的源端与目标端之间的映射，使模型具备初步的翻译能力。

与双语词表相比，单语图片-文本描述数据的获取往往更加容易，在社交媒体中往往存在大量这类数据。而单语图片-文本描述数据与无监督机器翻译的场景恰恰相符，因此，本研究的核心内容在于充分利用额外的图片模态信息，利用跨模态对齐技术，建立更高质量的初步对齐，使模型在该步骤便能够具备较好的初步翻译能力，为后续的迭代回译阶段提供良好的初始条件。

本研究突破了原有无监督机器翻译单一模态的局限性，将多模态图片模态信息与跨模态对齐技术引入无监督场景中，相比于其他方法，其优势在于仅利用了单语图片-文本描述数据建立源端至目标端的初步映射，这也更符合无监督机器翻译的场景需求，使用更低的代价实现了更好的对齐效果。

1.4.4 基于图片后验检索增强的视觉知识问答

针对视觉知识问答场景中，模型检索对应图片数据时检索不准确，生成答案时对图片关注不足的问题，本文提出基于图片后验检索增强的视觉知识问答方法。在视觉知识问答的场景中，即便当前的大语言模型具备强大的生成能力，但是在缺乏辅助视觉信息的前提下，仅依靠文本提供的信息模型往往无法生成准确的输出，甚至可能出现错误的，违背现实的内容。因此，往往需要检索额外的图片信息，来补充模型在某些特定领域知识的不足，从而提高模型生成内容的准确性。当前的方法[16,23,24]往往较为简单，大多直接采用预训练 CLIP 模型中的视觉 Transformer 作为视觉编码器，以 Transformer 编码器作为文本编码器，分别提取视觉特征和文本特征。该方法在处理多模态数据时较为粗糙，没有考虑到图片与问题文本的相关性问题，在检索的过程中包含了大量的语义无关信息，在方法层面仍有较大的改进空间。

基于此，本文分别在检索和生成阶段引入包含答案的后验信息，将该信息作为图像和问题文本的连接点，辅助模型学习到图像中和问题更相关的区域，从而使模型能够自主提取图片的相关性区域，提高生成答案的准确性。具体而言，在检索阶段，本文将带有答案后验信息的图像特征与不包含后验信息的特征进行对齐，加强图像中有关答案后验信息的区域特征；在生成阶段，通过对生成的答案添加相关性权重，使模型自主学习并关注到图像中与问题的相关区域。通过以上方法，本文实现了更为精确的图片检索，并提高了模型的生成准确性。

1.5　论文结构

本文共分为六章，章节的具体内容如下：第一章首先介绍基于图文对齐的文本生成技术的相关研究背景以及其重要研究意义，介绍了本文主要探索的三个文本信息受限的文本生成任务场景及其研究目标，并进一步分析了其中的主要挑战，最后介绍了本文针对所述挑战的解决思路并阐述了具体的研究内容。

第二章主要梳理了基于图文对齐的文本生成技术的应用现状与发展趋势。首先对图文对齐技术进行简单介绍，简述了该技术的典型模型结构。其次，针对该技术的不同应用场景，介绍了目前在这些领域内的主流方法和研究现状，并分析了未来的发展趋势。

第三章介绍了图片辅助的低资源机器翻译。分析了当前低资源机器翻译所存在的系列问题，提出了一种基于句级别和词级别对比学习的语义对齐方法，构建了源端公共语义空间，实现了从高资源语言到低资源语言的翻译能力迁移。随后介绍了实验所用到的数据集、实验细节等，最后展示了实验结果与分析。

第四章介绍了基于图文对齐的无监督机器翻译。分析了在无监督机器翻译场景下所面对的主要挑战，提出了一种仅借助图片-文本描述数据，构建源端与目标端语言的公共语义空间，建立源端到目标端初步映射的方法。其次介绍了本实验所用到的数据集、实验细节以及评价指标等，最后展示了实验结果与分析。

第五章介绍了基于图片后验检索增强的视觉知识问答。分析了在视觉知识问答场景中所存在的主要挑战，并提出了一种借助答案后验信息使模型自主提取图片与文本相关性区域的方法。之后介绍了所使用到的数据集、实验细节等，最后展示了实验的结果与分析。

第六章对上述研究工作进行了总结，指出本文的主要贡献和创新点，最后对基于图文对齐的文本生成技术的未来应用方向与发展进行展望。

第 2 章　研究现状与发展趋势

本章将详细介绍基于图文对齐的文本生成技术的研究现状与发展趋势。2.1 节首先介绍图文对齐技术与主流的图文对齐技术模型。2.2 节将系统介绍目前为止图文对齐技术在低资源机器翻译场景、无监督机器翻译场景、视觉知识问答场景以及其他场景中的应用。2.3 节将总结基于图文对齐的文本生成技术研究现状，并展望未来发展的方向。

2.1　图文对齐技术简介

图文对齐技术，是指通过对比学习等手段，将图片与文本对齐至同一语义空间。这种技术的发展得益于图像处理和自然语言处理的进步，尤其是在多模态数据的深度学习应用中，如何利用图文对齐技术，在不同的应用场景中来辅助任务的进行，已经成为一个重要的研究方向。

在多模态任务中，图片所表达的信息具有高度直观性和语义特征的语种无关性。如下图2-1所示，中文的“一只可爱的猫”和英文的“a cute cat”拥有不同的文本特征表示，但他们共享相同的视觉特征表示。因此，通过图片，文本、音频等其他模态可以被引导映射到共同的语义空间中。这意味着系统能够基于图片的语义表示，推导与其相关的文本描述或音频信息，甚至能够通过跨模态检索来找到其他模态中的相应内容。CLIP 模型作为语义空间对齐的代表性工作，本节以 CLIP 模型为代表介绍该技术常用的模型结构。

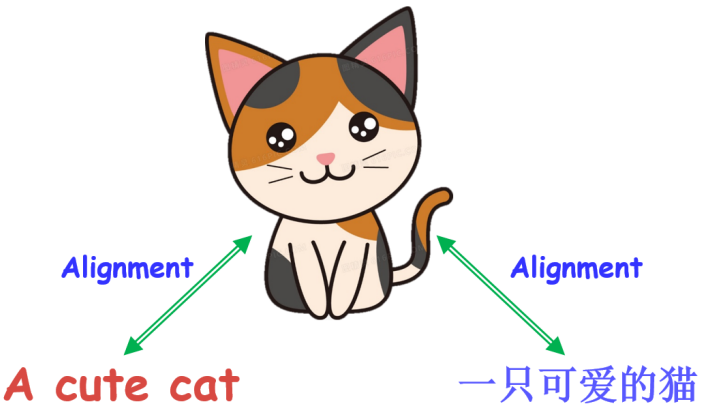


图 2-1　图片的语种无关性

Figure 2-1　Language-independence of images

2.1.1 CLIP 模型简介

CLIP（Contrastive Language-Image Pretraining）[1] 模型是近年来跨模态语义空间对齐技术的代表性模型之一，由 OpenAI 提出并在多模态任务中取得了显著的效果。CLIP 模型的核心目标是将图像和文本映射到同一个共享的语义空间中，借助对比学习（Contrastive Learning）[9] 方法，使得相似的图像与文本对在该空间中距离更近，不相似的对则更远。通过这种方式，CLIP 实现了图像和文本的跨模态对齐，并且具备很强的零样本学习能力。

CLIP 的总体架构基于双塔模型（Dual-tower Model），如图2-2所示，即分别使用两个独立的编码器，一个用于处理图像（视觉模态），另一个用于处理文本（语言模态）。这两个编码器分别将图像和文本编码为固定长度的向量，并在训练过程中通过对比学习的方式，将相似的图像-文本对齐在同一语义空间中。

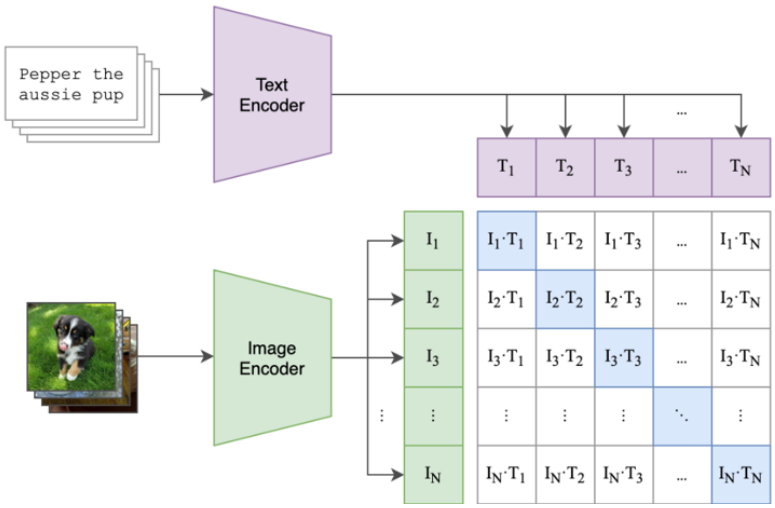


图 2-2 CLIP 模型示意图[1]

Figure 2-2 Overview of the CLIP model[1]

图像编码器 在 CLIP 模型中，图像编码器的作用是将输入的图像转化为一个定长的特征向量。最常见的实现方法是采用深度卷积神经网络模型（CNN）或 Transformer 模型架构：

- 卷积神经网络（CNN）[10]：CLIP 的图像编码器所采用的一种常见模型是 ResNet（Residual Networks）[25]。ResNet 具有较深的层次结构，通过引入残差连接，解决了深层网络中梯度消失的问题，能够高效提取图像中的空间特征。
- 视觉 Transformer（Vision Transformer, ViT）[2]：CLIP 视觉编码器模型的另一种实现是视觉 Transformer 模型，其模型结构如图2-3所示，ViT 利用自注意力机制，能够捕捉图像中的长距离依赖关系。与传统的卷积神经网络不同，ViT 处理图像时不依赖局部卷积操作，而是将图像分成一系列的补丁（patch），再将这

些补丁以一种序列的形式作为输入传入自注意力模块。同时，ViT 还会在补丁序列之前增加一个特殊补丁，用以预测图像整体的类别。该补丁的输出特征往往会作为图像的全局特征被使用。与卷积神经网络相比，ViT 具备更强的全局特征捕捉能力，特别适合处理高分辨率图像。

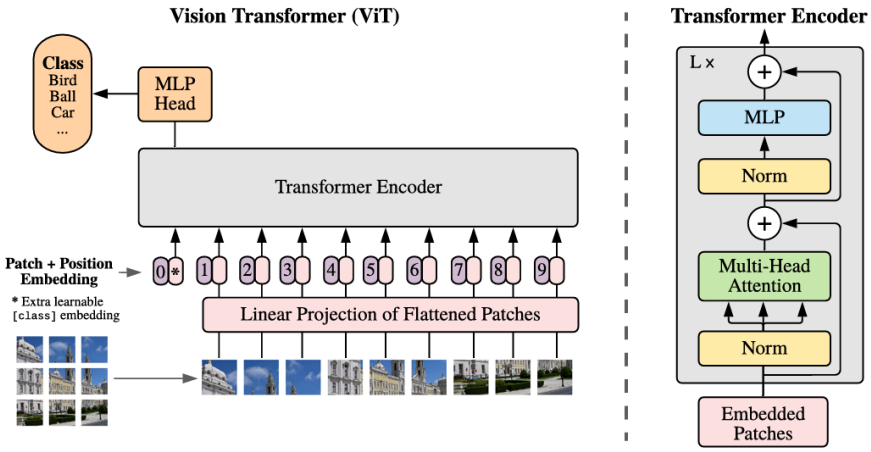


图 2-3　视觉 Transformer 模型示意图[2]

Figure 2-3　Overview of the vision transformer model[2]

当使用 CNN 进行特征提取时，图像编码器的输出将是一个全局特征。当使用 ViT 时，图像编码器的最终输出是一个定长的特征向量，该向量表示了图像的全局（第一个特殊向量）与其余局部语义信息。一般使用全局特征向量用于与文本特征进行对比和对齐。

文本编码器　CLIP 的文本编码器采用 Transformer 模型[3] 的编码器部分。Transformer 是当前自然语言处理任务中最常用的模型结构，其模型结构如图2-4所示，它通过自注意力机制，能够有效处理文本中的长程依赖和复杂语义关系。编码器中的基本块由多头自注意力机制、层归一化、前馈层和残差连接 4 个部分组成。在输入端，设计有词嵌入矩阵，将自然语言句子中的每个词语分别映射为对应的词向量。CLIP 文本编码器的输入是一个自然语言句子，该句子首先被编码成词嵌入表示（Word Embedding），然后通过 Transformer 层进行处理，最后输出一个定长的向量，该向量代表该句子的语义表示。

对比学习　CLIP 模型中最核心的部分是对比学习策略，它通过最大化语义相似的图像和文本之间的相似性，同时最小化不相关图像和文本对之间的相似性来训练模型。具体来说，对比学习的目标是使得在共享语义空间中，相关的图像和文本特征向量相距更近，而不相关的特征向量相距更远。

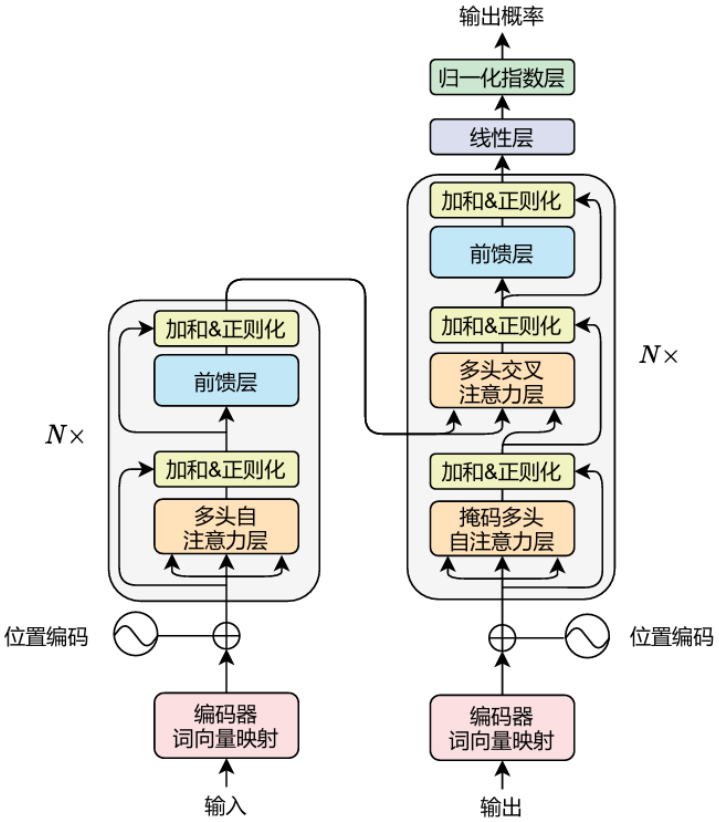


图 2-4　Transformer 模型示意图[3]

Figure 2-4　Overview of the transformer model[3]

给定两个集合 $\mathbf{X}=\{x_i\}_{i=1}^{M}$ 和 $\mathbf{Y}=\{y_i\}_{i=1}^{M}$，对于每个 x_i，正例为 (x_i, y_i)，剩余的 $M-1$ 无关对 $(x_i, y_j)(i \neq j)$ 被视为负例。$\mathbf{X}$ 和 $\mathbf{Y}$ 之间的对比损失定义如下：

$$`122\mathcal{L}_{\text{ctr}}(\mathbf{X},\mathbf{Y}) = -\sum_{i=1}^{M}\log\frac{\exp(s(x_i,y_i)/\tau)}{\sum_{j=1}^{M}\exp(s(x_i,y_j)/\tau)}. \tag{2-1}$$

其中，$s()$ 是余弦相似度函数 $s(a,b)=a^{\top}b/\|a\|\|b\|$ 。τ 是温度超参数，用于控制对硬负样本的惩罚强度[26]。有关温度超参数等的具体作用和更多细节将在第 3 章的方法介绍中进行详细描述。

2.1.2　多模态大语言模型

近年来，在多模态相关的任务中，涌现出越来越多的以多模态大语言模型（Multimodal Large Language Models, MLLMs）为基础的解决方案。多模态大语言模型是大语言模型（Large Language Models, LLMs）在多模态场景下的扩展。大语言模型的起源可追溯至基于深度学习的自然语言处理技术，尤其是 2017 年提出的 Transformer 架构[27]。随着模型规模的扩大（如 GPT-3[28]），大语言模型展现了强大的文本生成、推理和上下文学习能力。然而，纯文本模态的局限性促使

研究者将大语言模型扩展至多模态领域，使其能够处理文本、图像、音频、视频等多模态数据，从而更贴近人类多感官协同的认知模式。

多模态大语言模型的发展遵循两条主要路径：（1）模态对齐与融合：通过跨模态预训练（如 CLIP[29]）对齐文本与图像特征，构建公共语义空间；（2）模型架构扩展：在大语言模型基础上引入视觉编码器（如 ViT[2]）或其他适配器模块，支持多模态输入输出。当前主流的多模态大语言模型可分为以下几类主要范式：（1）基于对齐的预训练模型：例如 Flamingo[30]，通过交叉注意力机制融合文本与视觉特征，支持少样本多模态推理；（2）端到端多模态指令微调模型：如 GPT-4V[31]，LLaVA[32]，InternVL[4]，此类模型将图像编码为与文本兼容的嵌入，直接输入大语言模型生成跨模态响应；（3）检索增强型模型：如 BLIP-2[33]，利用轻量级适配器桥接冻结的视觉编码器与大语言模型，降低训练成本。

这些多模态大语言模型在多模态问答、图像描述生成、具身智能等领域展现了潜力。本文第 5 章所探究的多模态检索增强生成场景就是多模态大模型的常见应用场景之一，其中所采用的多模态大模型为 InternVL 模型，属于端到端多模态指令微调模型。InternVL 模型的工作原理与 LLaVA 模型类似，如图2-5所示，在原有大模型的基础上，增加一个视觉编码器以及线性投影层，将编码得到的视觉片段序列通过线性投影层映射至大模型的输入空间，与文本共同输入模型中，从而获得所需要的答案。

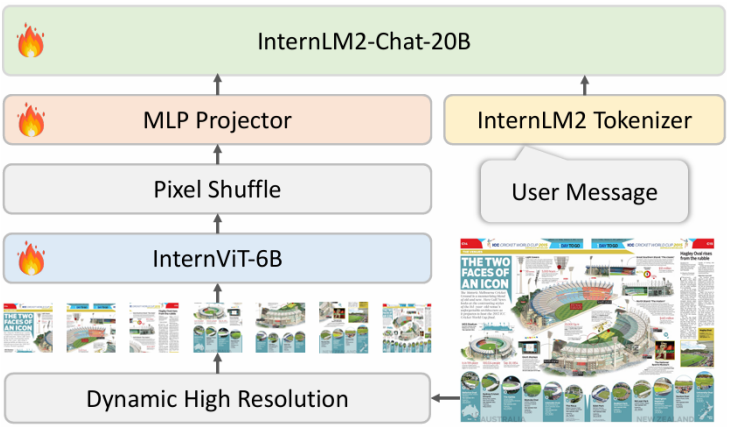


图 2-5　InternVL 模型示意图[4]

Figure 2-5　Overview of the InternVL model[4]

2.2　基于图文对齐的文本生成技术的应用与发展现状

近年来，基于图文对齐的文本生成技术已经取得了一定进展，特别是在深度学习技术的推动下，图文跨模态对齐已成为文本生成任务的一个重要的辅助手段。图片-文本描述数据凭借其丰富的语义关联和可获取性优势，为文本生成任

务提供了有效的跨模态监督信号，推动了视觉引导文本生成、多模态机器翻译等任务的发展。这一技术不仅显著提升了生成文本的准确性和丰富性，还拓展了模型在低资源生成、知识问答等挑战性场景中的应用能力。本节主要从该技术在低资源机器翻译、无监督机器翻译、视觉知识问答三个任务出发，对当前的技术发展现状进行系统介绍。

2.2.1 低资源机器翻译

机器翻译任务是自然语言处理领域的经典任务之一，随着多模态技术与跨模态技术的兴起，机器翻译也向着多模态的场景不断扩展，从而诞生了多模态机器翻译（Multimodal Machine Translation）这一研究领域。多模态机器翻译的产生主要是由于机器翻译任务中存在的文本歧义问题。由于文本歧义的存在，翻译过程中的歧义性错误非常常见。但若能够合理利用额外的视觉模态信息，即可有效缓解这一问题。举例来说，“bank” 同时具有 “银行” 和 “河岸” 两种意思，但是这两种意思所对应的视觉特征表示是截然不同的，这可以帮助模型进行理解和区分。早期的多模态机器翻译方法[34–45] 主要基于递归神经网络（Recurrent Neural Network）结构实现，而近年来的方法[46–54] 则基于 Transformer 模型来构建支持多种模态输入的多模态 Transformer 模型。然而，Caglayan 等[55], Wu 等[56] 通过研究发现发现，在文本信息充足的前提下，机器翻译模型对于视觉信息的利用并不充足，甚至将图片输入改为随机噪声输入也能够实现类似的效果。

而针对上述问题，一部分研究者便将目光转向文本信息不充足的场景，即低资源场景。当前的神经机器翻译模型往往依赖于大规模的翻译语料，这对于一些低资源语言来说是不可行的，因此，将图片信息用于低资源场景作为文本信息的补充则成为了一个新的解决方案。早期的研究[57–60] 尝试使用纯文本方法来解决平行语料缺乏的问题，此类方法使用第三种语言作为中间媒介，即使用源端-中间语言，中间语言-目标端两种翻译数据来实现低资源情境下的零样本翻译，但这类方法并未完全克服需要平行语料的局限性。Nakayama 等[14] 在传统 RNN 结构的基础上，以图片为中心对齐源端和目标端语言，在无需平行语料的情况下实现源端到目标端的映射。Li 等[15] 则利用一种基于通讯的训练方法，以图片作为信息传递的媒介实现少样本翻译。

2.2.2 无监督机器翻译

无监督机器翻译是低资源机器翻译场景的扩展，其旨在完全脱离平行语料，仅使用单语语料来实现翻译的目标。Lample 等[17,18,19] 等人提出了一种具有代表性的无监督机器翻译三阶段范式：单语建模，初步对齐，迭代回译。通过三阶段的训练，建立源端到目标端的简介映射。而后续也有较多的研究，如 Conneau 等[20], Song 等[21] 等均沿用了这一思路，并改进了其中的部分方法。

与机器翻译到多模态场景的拓展类似，无监督多模态机器翻译也是无监督机器翻译方法向多模态场景的延申。在无监督机器翻译场景下，引入额外的视觉模态信息，以增强无监督机器翻译系统的性能。一部分研究参考多模态机器翻译的方法，如 Chen 等[61], Su 等[62] 通过融合视觉和文本信息，构建多模态特征输入模型进行训练，来增强无监督机器翻译模型。此类方法与有监督多模态机器翻译方法对图片的利用方式完全相同，并没有使图片信息真正适配于无监督场景。另一类研究则是以图像为枢纽，实现零样本翻译的目标。例如 Huang 等[22] 在无监督机器翻译的框架下，利用图片对源端语言和目标端语言进行预对齐操作。上述的方法虽然成功地将跨模态对齐技术应用于低资源翻译的场景中，但仍存在诸如未完全脱离对平行数据的依赖，需要引入额外的模型和数据等缺陷，这也是后续研究可以探索的方向。

2.2.3　视觉知识问答

视觉知识问答任务旨在解决需要额外视觉信息的问答任务场景，通过从外部检索额外的相关图片信息，并将检索到的图片以上下文学习[28] 或直接拼接的方式输入到大模型中，从而提升模型生成答案的准确性[63,64]。在检索方面，有几类具有代表性的纯文本检索方法，DPR（Dense Passage Retrieval）[65] 是一种基于稠密向量表示的段落检索方法，它通过双编码器（查询编码器和段落编码器）将查询和段落映射到稠密向量空间，并利用相似度计算实现高效检索。此外，REALM（Retrieval-Augmented Language Model）[66] 和 RAG（Retrieval-Augmented Generation）[63] 等方法通过将外部知识检索与语言模型预训练相结合，构建了检索增强的语言模型，使其能够动态检索相关文档并增强模型的生成与理解能力。这些方法在多类任务中展现了强大的性能提升潜力。

由于纯文本检索方法无法满足视觉知识问答等需要跨模态检索的场景，研究者渐渐将其扩展至多模态领域。Tan 等[16] 利用 CLIP 作为文本和图像特征的编码器，将所提取的特征进行求和平均得到多模态特征，将该特征作为多模态检索知识库的特征索引。Chen 等[23] 利用多模态大模型作为重排模型，将根据文本检索得到的图像特征进行重排，过滤掉与对应图像特征相似的强负例图像，同时在训练的过程中加入强负例图像作为噪声，加强模型的鲁棒性。Ma 等[24] 运用对比学习的方法，在多模态文档检索任务中，将多模态文档和问题进行对齐，提高了多模态文档检索模型的稳定性。而在生成方面，目前的研究并未对该阶段进行深入扩展，Chen 等[23] 仅在该阶段加入负例图片提高模型生成的鲁棒性，因此，在该领域，目前的研究工作还处于初期阶段，仍然有较大的探索空间。

2.2.4 其他多模态场景

不仅仅是以文本模态为主的机器翻译场景，在以语音，视频模态为主的应用场景中，图文对齐技术也得到了较为广泛的应用。这一系列的工作[13,67] 通常利用图像的视觉信息和音频信号的时序特征，通过联合训练的方式，使得图像与音频可以共享一个语义空间。或通过将关键帧或关键图像提取出来，模型可以对视频的内容进行抽象表达，进而与文本、音频等其他模态进行对齐。Huang 等[68] 利用多语言 BART 模型[69] 作为文本特征提取器，利用视频特征提取模型提取视频的特征，并构建公共语义空间，实现了跨模态零样本知识迁移。在实际应用方面，例如在视频配音生成、自动化解说等任务中，系统可以根据图片场景生成对应的语音描述，或者根据语音线索推断相关的图片信息。这一技术未来将在虚拟现实（VR）、增强现实（AR）等多模态融合应用中发挥更大作用。

2.3 基于图文对齐的文本生成技术的发展趋势

随着多模态学习和深度学习技术的不断发展，基于图文对齐的文本生成技术展现出了巨大的应用潜力和创新空间。其发展趋势可以概括为以下几个方面：

（1）零样本学习与少样本学习的进一步发展：零样本学习（Zero-Shot Learning, ZSL）和少样本学习（Few-Shot Learning, FSL）是跨模态语义对齐技术中的重要应用方向。当前的许多模型，如 CLIP，已经展示了在零样本学习场景下的出色表现，能够在未经过特定任务训练的情况下，根据自然语言描述对图像进行分类和检索。未来的发展趋势是进一步提升模型在零样本和少样本条件下的表现，尤其是跨领域的泛化能力。当前的模型在面对未知领域的数据时，往往会出现性能下降的情况，而随着对齐技术的发展，研究者将致力于提高模型对不同领域、不同数据分布的适应性。具体措施包括自监督学习（Self-Supervised Learning）的广泛应用，自监督学习通过从未标注数据中获取表示学习的能力，有望进一步减少对标注数据的依赖，并提升零样本学习的能力。例如与无监督机器翻译任务相结合，利用单语数据实现零样本翻译。

（2）与更多文本生成任务的深度结合：未来，基于图文对齐的文本生成技术将在更广泛的文本生成任务场景中展现出巨大的应用潜力与发展空间。随着多模态大模型技术的不断突破，该技术有望在以下几个方向实现重要进展：首先，在开放域对话系统中，图文对齐技术将帮助模型更好地理解对话上下文中的视觉语义，实现更具情境感知能力的多轮对话生成；其次，在教育领域的自动题目生成与解答场景，通过结合图文信息可以生成更准确、更具解释性的教学文本；再者，在专业领域的报告生成（如医疗影像报告、工业检测报告）中，该技术能够确保生成文本与视觉证据的高度一致性，显著提升专业文本的可靠性。特别值得注意的是，随着视觉语言预训练模型的演进，图文对齐技术将逐步实现从”

显式对齐” 到” 隐式理解” 的跨越，使模型能够自主挖掘更深层次的跨模态关联，从而在创意写作、个性化内容生成等复杂场景中产生更具创造性的文本输出。这些发展趋势不仅将拓展文本生成技术的应用边界，也将推动多模态人工智能向更高层次的认知智能迈进。

第 3 章　图片辅助的低资源机器翻译

3.1　引言

机器翻译一直是自然语言处理中的核心任务之一。在机器翻译的方法中，神经机器翻译（Neural Machine Translation）展现了其卓越的性能，并成为机器翻译的主流范式。然而，神经机器翻译是一种数据驱动的方法，需要大量的平行数据来进行模型的训练。当数据不足时，训练一个高质量的神经机器翻译模型是不现实的。遗憾的是，世界上有许多语言仍然缺乏足够的训练数据，对于一些语言来说甚至没有任何的平行语料可供参考。因此，有关低资源语言的翻译是神经机器翻译任务中的一个重要挑战。

近年来，研究人员在提高神经机器翻译在低资源语言上的性能方面做了许多尝试。Lample 等[70] 提出了一种无监督方法，利用大量单语数据（>1M）学习语言之间的弱映射关系，但这对于低资源语言来说成本仍然很高。Liu 等[5], Pan 等[7], Gu 等[8], Lin 等[71] 则提出了多语言神经机器翻译模型，通过学习多种语言的共享空间来实现训练集中出现但缺乏对应平行数据的语言之间的翻译。然而，这些方法仍然需要源语言和目标语言以及其他多种语言的辅助平行数据，这对于低资源语言来说仍然不可行。

与此同时，随着多模态任务受到越来越多的关注，诸如图像-文本对这类资源变得更加丰富。受到一些跨模态对齐研究工作[1,72,73] 的启发，本章提出了一种跨模态对比学习方法，以图像为锚点对齐不同源端语言，从而实现低资源语言的零样本和少样本翻译。通过利用一种高资源辅助语言与目标语言之间的平行语料，只需为这些低资源语言获取少量图像-文本对（<0.1M），即可实现从低资源语言到目标语言的翻译。平行句子对用于学习从高资源语言到目标语言的映射，而图像-文本对则通过跨模态对齐为所有源端语言学习一个公共语义空间。通过这种方式，以图像为锚点，学习从低资源语言到目标语言的映射，从而在没有平行语料的情况下实现零样本或少样本翻译。

如图 3-1 所示，高资源语言德语和低资源语言法语通过跨模态对齐结合在一起，从而将德语 → 英语的翻译能力迁移到法语 → 英语。实验结果表明，在零样本和少样本场景下，本章方法在各个测试集均优于基线模型。此外，分析实验表明，本章方法也能够有效实现跨模态和跨语言的对齐。

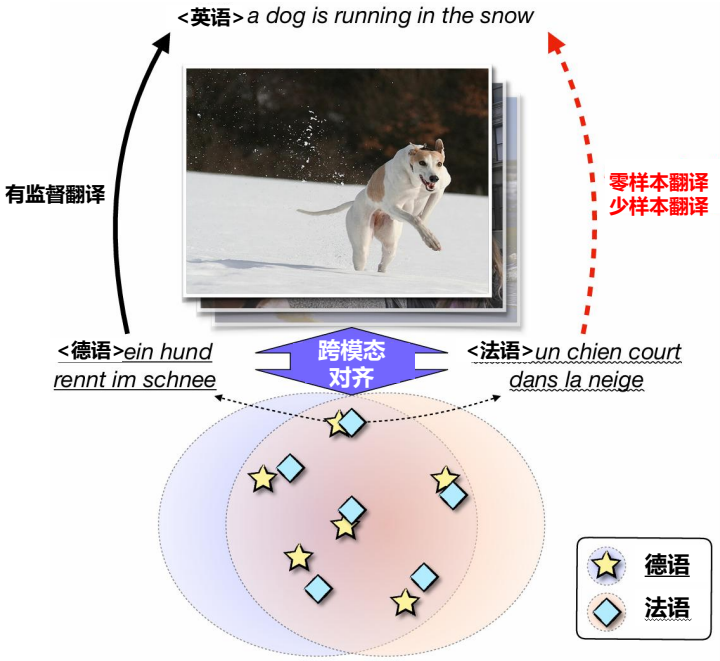


图 3-1 跨模态对齐方法示意图

Figure 3-1 Overview of the cross-modal alignment method

3.2 相关工作

3.2.1 多模态机器翻译

多模态机器翻译（Multimodal Machine Translation）旨在引入视觉模态以增强神经机器翻译，其主要针对机器翻译中出现的歧义问题。由于文本具有多义性，但不同的语义对应不同的视觉表示，将视觉信息与文本信息融合后输入模型进行训练。早期的方法[34–45]主要基于结合注意力机制的 RNN 架构。随着 Transformer 模型的流行，近年来的方法[46–54]便基于 Transformer 进一步提升了多模态翻译系统的性能。然而，Caglayan 等[55], Wu 等[56]发现，在多模态机器翻译任务中，当平行语料充足时，其中的视觉信息部分往往容易被忽略。因此，本章研究的重点之一在于当平行语料不足时，视觉模态所起到的作用。

3.2.2 零样本/少样本机器翻译

由于神经机器翻译系统严重依赖大规模平行语料，研究人员开始探索在平行数据有限的情况下，如何提高翻译系统的性能。目前在这方面已经取得了部分成果，如无监督机器翻译[19,74–78]，这类方法通过大规模的单语数据进行预训练与回译训练学习源端到目标端的间接映射，从而实现了这一目标。多语言机器翻译[5,7,71,79]则通过其他方向的平行语料作为桥梁，实现了这一目标。而另一类研究则关注到了视觉信息的辅助作用，尝试在视觉模态的帮助下实现零样本或少样本翻译[14,15]，但在数据极其有限的情况下未能取得令人满意的性能。基于此，

本章方法扩展了这一研究方向，并在更少数据的情况下实现了更好的性能。

3.2.3　跨模态图文对齐

对比学习是跨模态图文对齐中的代表性方法之一，其在诸多多模态任务中取得了巨大成功，例如跨语言迁移[12]、视频-文本理解[13] 等。其中最具代表性的方法是 CLIP[1]，它通过对比学习实现了图像和文本之间的语义对齐。不仅如此，最近的研究中如 Ye 等[80] 还展示了跨模态对比学习在语音翻译中的强大能力，展现了跨模态图文对齐方法在多模态领域的充分潜力。受到此类研究的启发，本章提出了一种跨模态对比学习方法，以实现零样本和少样本翻译。

3.3　方法介绍

本节内容将详细介绍本章所提出的跨模态对比学习方法，其中包括粗粒度句级别对比学习和细粒度词级别对比学习。

3.3.1　任务定义

该子课题的任务目标是借助特定的高资源语言 $\widehat{L}$，实现从 T 个低资源语言 $L_1, L_2, ..., L_T$ 到目标语言 L_y 的零样本或少样本翻译。对于高资源语言 $\widehat{L}$，其包含三元组数据 $\mathcal{D}_{\widehat{L}} = \{(\mathbf{i}, \mathbf{x}, \mathbf{y})\}$、其中 $\mathbf{i}$ 是图像，$\mathbf{x}$ 和 $\mathbf{y}$ 分别是 $\widehat{L}$ 和 L_y 中的描述。对于每种低资源语言 L_i，只有单语图片-文本描述对数据 $\mathcal{D}_{L_i} = \{(\mathbf{i}, \mathbf{x})\}$ 可用。需要注意的是，不同语言之间不会共享相同的图像。

3.3.2　模型结构

如图3-2所示，本章模型由四个子模块组成：视觉编码器、文本编码器、文本解码器和跨模态对齐模块。

本章方法使用视觉 Transformer 模型（ViT）作为图像编码器来提取视觉特征。视觉编码器首先将图像分割成若干个补丁，然后将嵌入补丁的序列与一个特殊的 [class]）标记输入视觉编码器。最后，图像会被编码为一串向量 $\mathbf{v} = (v_0, v_1, ..., v_m)$，其中 v_0 是 [class] 标记的特征表示，可视为图像的全局表示，而 $\mathbf{v}^p = (v_1, ..., v_m)$ 是图像的局部表示。在接下来的部分中，使用 v_0 进行句子级对比学习，$\mathbf{v}^p$ 则作为词级别对比学习的特征。

文本编码器由 N 层 Transformer 编码器层组成，所有语言（$L_{1...T}$ 和 $\widehat{L}$）共享该编码器层。对于输入句子 $\mathbf{x} = (x_1, ..., x_n)$，文本编码器的输出表示为 $\mathbf{w} = (w_1, ..., w_n)$。文本解码器则由 N 变换解码器层组成。对于句子对 $(\mathbf{x}, \mathbf{y})$，交叉熵损失定义如下：

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{|\mathbf{y}|} \log p(y_i^* | \mathbf{y}_{<i}, \mathbf{x}). \tag{3-1}$$

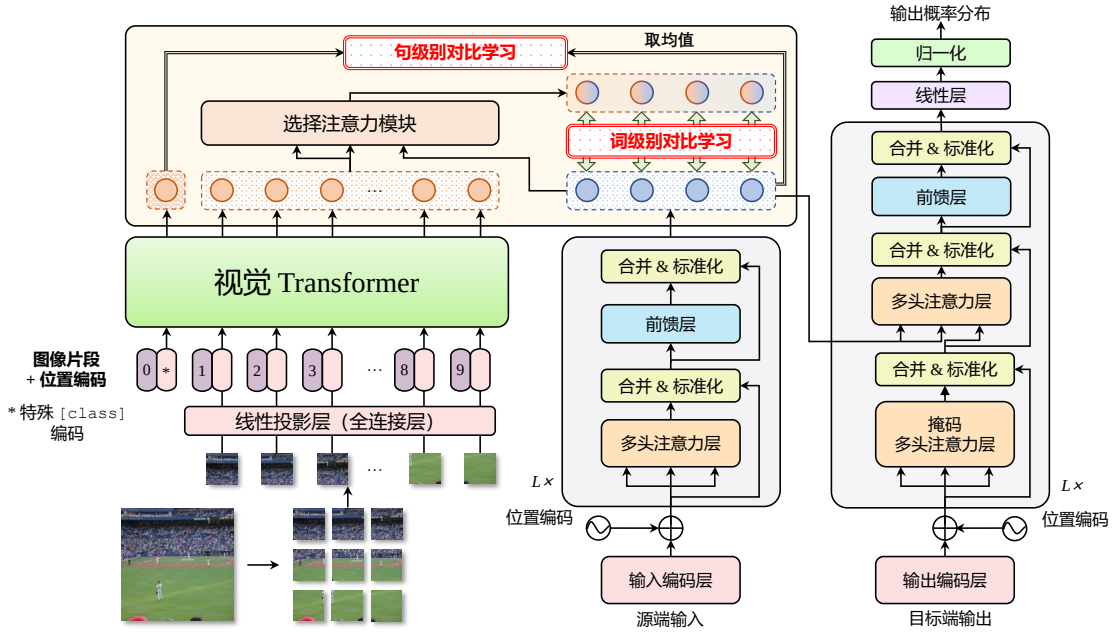


图 3-2　图像辅助的低资源机器翻译模型示意图

Figure 3-2　Overview of the cross-modal zero-shot machine translation model

跨模态对齐模块旨在对齐文本编码器和视觉编码器的输出，其中包含句级别对比学习模块和词级别对比学习模块。

3.3.3　图片辅助的源端语言对齐

对比学习　对比学习（Contrastive Learning）的核心是让相应配对的特征表示更接近，相反，让不相关配对的特征表示远离。给定两个集合 $\mathbf{X} = \{x_i\}_{i=1}^{M}$ 和 $\mathbf{Y} = \{y_i\}_{i=1}^{M}$，对于每个 x_i，正样本为 (x_i, y_i)，而其余的 $M-1$ 个无关对 $(x_i, y_j)(i \neq j)$ 被视为负样本。$\mathbf{X}$ 与 $\mathbf{Y}$ 之间的对比损失定义如下：

$$\mathcal{L}_{\text{ctr}}(\mathbf{X}, \mathbf{Y}) = -\sum_{i=1}^{M} \log \frac{\exp(s(x_i, y_i)/\tau)}{\sum_{j=1}^{M} \exp(s(x_i, y_j)/\tau)}, \tag{3-2}$$

其中，$s()$ 是余弦相似度函数，定义为 $s(a,b) = \frac{a^{\top} b}{|a||b|}$，$\tau$ 是温度超参数，用于控制对难负样本的惩罚力度[26]。

句级别对比学习　句级别对比学习的关键在于对齐不同模态的句子级表征，对于文本和图像的句级别表征，其定义如下：

$$w^s = \frac{1}{n}\sum_{i=1}^{n} w_i, v^s = v_0, \tag{3-3}$$

对于文本特征，采用取全句特征向量平均值的方法获取句级别特征，对于图像特征，则直接采用特殊类别标记的全局特征表示。然后，计算大小为 B 的批次中的对

比学习损失，其文本表示和视觉表示分别为 $\mathbf{W}^s = \{w_1^s, ...w_B^s\}$ 和 $\mathbf{V}^s = \{v_1^s, ...v_B^s\}$。相应的图片和标题配对 (w_i^s, v_i^s) 为正例，其他配对 $(w_i^s, v_j^s)(i \neq j)$ 为负例。最后，句子级对比学习的损失函数定义如下：对于一幅图像 $\mathbf{i}$，其正例为相应的标题 $\mathbf{x}$，其余 $B-1$ 的标题为批次中的负例，其大小为 B，反之亦然。因此，句级别对比学习损失函数可以由以下公式表示：

$$\mathcal{L}_{\text{ctr}}(\mathbf{W}^s, \mathbf{V}^s) = -\sum_{i=1}^{M} \log \frac{\exp(s(\mathbf{W}_i^s, \mathbf{V}_i^s)/\tau)}{\sum_{j=1}^{M} \exp(s(\mathbf{W}_i^s, \mathbf{V}_j^s)/\tau)}. \tag{3-4}$$

在训练过程中，由于在一个批次中存在不同语言的图像-文本对，因此需要先根据语种将一个训练批次分成几个小批次，然后分别计算每种语言的对比学习损失。除此以外，在计算对比学习损失时，还针对目标语言 L_y 计算了基于其图文对数据 $(\mathbf{i}, \mathbf{y})$ 的对比学习损失，该数据来自 $D_{\hat{L}}$，有关目标语言的对比学习作用将在第 3.5.3 节对其影响进行分析。

选择性注意力机制　虽然句级对比学习可以学习到模态之间的粗粒度对齐，但它可能会忽略掉一些局部特征信息，而这些信息对于翻译的结果也是至关重要的。为了更好地实现模态之间的对齐，本章方法在句级别对比学习的基础上提出了词级别对比学习来学习图像和文本之间的细粒度对应关系。

由于图像特征序列和文本序列之间的长度并不相等，无法进行对比学习的训练，为了建立局部图像特征与单词之间的关联模型，本章方法使用选择性注意机制（Selective Attention[54]）模块来提取图像的局部相关特征。对于局部视觉表征 $\mathbf{v}^p = (v_1, ...v_m)$ 和词级别文本表征 $\mathbf{w} = (w_1, ..., w_n)$，选择性注意的查询、关键和值分别是 $\mathbf{w}, \mathbf{v}^p, \mathbf{v}^p$：

$$\mathbf{v}^t = \text{Softmax}\left(\frac{(W_Q \cdot \mathbf{w})(W_K \cdot \mathbf{v}^p)^\top}{\sqrt{d_k}}\right)(W_V \cdot \mathbf{v}^p), \tag{3-5}$$

其中，W_Q、W_K 和 W_V 是可学习的矩阵参数。

词级别对比学习　在经过选择性注意模块提取文本相关特征之后，会得到两个长度相同为 n 的序列 $\mathbf{w} = (w_1, ..., w_n)$ 和 $\mathbf{v}^t = (v_1^t, ..., v_n^t)$。基于此，计算每对序列中的词级别对比损失。具有相同索引 (w_i, v_i^t) 的标记为正例，其他成对的标记 $(w_i, v_j^t)(i \neq j)$ 为负例。对于一个批次中，所有图像-文本对的词级别对比学习损失将相加求和。词级别对比学习损失表示如下：

$$\mathcal{L}_{\text{t-ctr}}(\mathbf{w}, \mathbf{v}^t) = \mathcal{L}_{\text{ctr}}(\mathbf{w}, \mathbf{v}^t) + \mathcal{L}_{\text{ctr}}(\mathbf{v}^t, \mathbf{w}), \tag{3-6}$$

在最终损失计算时，将对所有图文对的词级别对比学习损失进行相加。

3.3.4 训练策略

为了结合句级别和词级别的对比学习方法，本章方法提出了一种两阶段的训练策略，首先通过句级别对比学习粗粒度的跨模态对齐，然后再通过词级别的对比学习进一步学习细粒度的对齐。

粗粒度训练 在训练的第一阶段，模型训练的损失包括高资源语言 $\widehat{L}$ 的翻译任务交叉熵损失以及所有语言（包括目标语言 L_y）的句级别对比学习损失：

$$\begin{aligned}\mathcal{L}_{\text{coarse}} = {} & \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{CE}}(\mathbf{x},\mathbf{y}) + \lambda_s\mathbb{E}_{(\mathbf{i},\mathbf{y})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{s-ctr}}(\mathbf{i},\mathbf{y}) \\ & + \lambda_s\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{s-ctr}}(\mathbf{i},\mathbf{x}) + \lambda_s\sum_{i=1}^{T}\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{L_i}}\mathcal{L}_{\text{s-ctr}}(\mathbf{i},\mathbf{x}),\end{aligned} \tag{3-7}$$

其中，λ_s 是句级别对比学习损失的权重超参数。

细粒度训练 在训练的第二阶段，本章方法在第一阶段句级别对比学习训练，即公式 3-7 的基础上加入了词级别对比损失，其形式如下：

$$\begin{aligned}\mathcal{L}_{\text{fine}} = {} & \mathcal{L}_{\text{coarse}} + \lambda_t\mathbb{E}_{(\mathbf{i},\mathbf{y})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{t-ctr}}(\mathbf{i},\mathbf{y}) \\ & + \lambda_t\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{\widehat{L}}}\mathcal{L}_{\text{t-ctr}}(\mathbf{i},\mathbf{x}) + \lambda_t\sum_{i=1}^{T}\mathbb{E}_{(\mathbf{i},\mathbf{x})\in\mathcal{D}_{L_i}}\mathcal{L}_{\text{t-ctr}}(\mathbf{i},\mathbf{x}),\end{aligned} \tag{3-8}$$

其中，λ_t 是词级别对比学习损失的权重超参数。

少样本训练 经过包含对比学习损失的两阶段训练后，可以直接评估训练完成后的模型在零样本翻译任务上的性能。此外，还可以利用少量低资源语言的额外平行语料数据 $\mathcal{D}_L = (\mathbf{x},\mathbf{y})$ 对模型进行微调，并评估其在少样本翻译任务上的表现。在微调过程中，仅使用交叉熵损失：

$$\mathcal{L}_{\text{finetune}} = \mathbb{E}_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_L}\mathcal{L}_{\text{CE}}(\mathbf{x},\mathbf{y}). \tag{3-9}$$

3.4 实验设置与结果

3.4.1 数据集

在本章实验中，选择德语作为高资源语言，英语作为目标语言，并选择法语和捷克语作为两种低资源语言，测试法语 → 英语和捷克语 → 英语两个语向上模型的零样本和少样本翻译性能。由于德语、法语和捷克语的图片-文本描述数据稀缺，在实验的过程中使用现有的机器翻译模型从两个英语图片-文本描述数据集中翻译生成伪数据。

Multi30K　Multi30K 数据集[81] 包含四种语言的图像标注：英语、德语、法语和捷克语。训练集和验证集分别包含 29,000 和 1,014 个实例。本章实验在 Test2016、Test2017 和 MsCOCO 测试集上对模型进行评估，这些测试集分别包含 1,000、1,000 和 456[1] 个实例。对于捷克语 → 英语任务，仅有 Test2016 测试集可用，因此只在该测试集上进行实验。

MsCOCO　MsCOCO 数据集[82] 是一个英语图像描述数据集，其包含有英语图片-文本描述数据。本章实验中采用其中的 Captioning 2015 数据集进行实验。过滤掉未标注的图像后，共有 121,000 个图像-文本对用以构造伪数据。

VizWiz　VizWiz 数据集[83] 和 MsCOCO 数据集类型，也是一个英语描图像描述数据集，共包含有 30,408 个英语图像-文本对。

伪数据　由于 MsCOCO 和 VizWiz 数据集仅包含图像的英语描述，本章使用预训练的机器翻译模型将英语描述翻译成德语、法语和捷克语。机器翻译模型的详细信息见附录 一。

数据集组成　在生成伪数据后，再将上述三个数据集平均分为三部分，分别用于德语 → 英语、法语 → 英语和捷克语 → 英语三个语向。如表 3-1 所示，每种源语言都包含有 60,136 个图像-文本对，并附有其所对应的源端语言的标注，这些数据将被用于跨模态对比学习。同时，60,136 个德语 → 英语平行句对将被用于翻译任务的训练。所有句子均使用字节对编码（BPE）[84] 分割为子词单元。所有源语言和目标语言共享一个词汇表，词汇表大小为 18K。

表 3-1　数据构成详情

Table 3-1　Detailed dataset statistics

语向	Multi30K	MsCOCO	VizWiz	总和
德语 → 英语	10,000	40,000	10,136	60,136
法语 → 英语	10,000	40,000	10,136	60,136
捷克语 → 英语	9,000	41,000	10,136	60,136

3.4.2　实验设置

本章实验使用经过预训练的 CLIP[1] 模型中的视觉 Transformer 作为 图像编码器。图像块大小为 16×16，分辨率为 224。序列长度为 50，包含一个特殊的 [class]

[1] 由于 MsCOCO 数据集中的 5 个句子出现训练集中，因此被移除。

片段和 49 个局部图像特征片段。源端编码器和目标端解码器基于 Transformer[3] 架构构建。编码器和解码器均有 $N = 6$ 层。注意力头的数量设置为 4。dropout 设置为 0.3，标签平滑值为 0.1。训练时，使用 Adam 优化器[85] 进行优化并进行 2000 次预热更新，学习率设置为 5e-4，每个训练批次最多包含 16K 个子词单元，训练步数最大为 70 步。训练过程中采用两阶段训练策略，第一阶段训练和第二阶段训练各占总步数的一半。

对于句级别的对比学习训练，温度超参数 τ_s 设置为 0.007，权重超参数 λ_s 设置为 5。对于词级别的对比学习训练，τ_t 设置为 0.1，λ_t 设置为 1。

对于少样本翻译，需要从 Multi30K 的训练集中随机抽取 5 组对应数量的平行语料分别进行实验，并报告每次实验结果的均值和标准差。所有实验均在 4 块 TITAN Xp GPU 上完成。本章方法的系统基于 *fairseq*[2] [86] 实现。

3.4.3 评价指标

目前主流的机器翻译质量自动评价指标分为基于统计的评价指标和基于模型的评价指标。本章实验选择了当前广泛使用的基于统计的评价指标：BLEU[87]. BLEU 是最经典的序列级文本评价指标，考虑了从一元组到四元组词语的准确率。为了与基线模型进行更严格的对比，本章实验进一步采用 SacreBLEU[88] 作为最终评价指标。相比于 BLEU，SacreBLEU 提供了统一的分词方法，其评价标准通常比 BLEU 更加严格。

在评估时，需要对最后 5 个检查点取平均作为评估模型，并使用束搜索（beam size 为 5）生成测试文本。实验中使用 sacreBLEU[3] [89] 计算去标记化实例的 BLEU[87] 分数[4]。

3.4.4 基线系统

本章实验的基准系统采用仅使用德语 → 英语平行语料训练的纯文本 Transformer 模型。对于零样本翻译，可以直接对基线模型进行评估。对于少样本翻译，则使用少量低资源语言平行语料对基线模型进行微调后再进行评估。基线模型的所有实验设置均与本章实验所使用的模型相同。

3.4.5 主实验结果

本章主实验在零样本和少样本场景下分别评估了基线模型、仅经过句级别对比学习进行训练的模型（**S-CTR**）以及经过句级别和词级别对比学习训练的模型（**S+T-CTR**）的翻译效果。

[2] https://github.com/pytorch/fairseq

[3] https://github.com/mjpost/sacrebleu

[4] sacreBLEU signature: nrefs:1 | bs:1000 | seed:12345 | case:lc | eff:no | tok:13a | smooth:exp | version:2.0.0

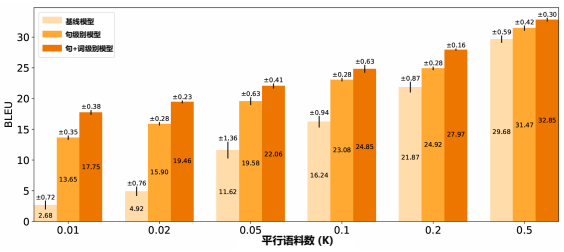
零样本翻译　表 3-2 展示了零样本翻译的结果，可以发现未经过对比学习训练的基线模型不具备任何零样本翻译的能力。而 S-CTR 模型和 S+T-CTR 模型相比基线模型取得了显著提升。与 S-CTR 模型相比，S+T-CTR 模型在各个语向的平均 BLEU 得分上提高了 7.41 分，这证明细粒度的对齐可以显著提升零样本机器翻译的性能。

表 3-2　零样本翻译的实验结果

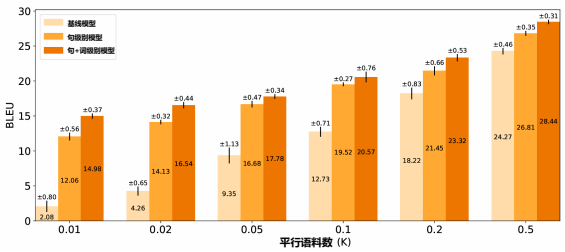
Table 3-2　Results of zero-shot translation

模型	法语 → 英语			捷克语 → 英语	平均值
	Test2016	Test2017	MsCOCO	Test2016	
基线模型	0.30	0.14	0.29	0.09	0.21
句级别对比学习	8.95	7.88	9.32	7.23	8.35
句 + 词级别对比学习	**17.76**	**14.74**	**16.97**	**13.58**	**15.76**

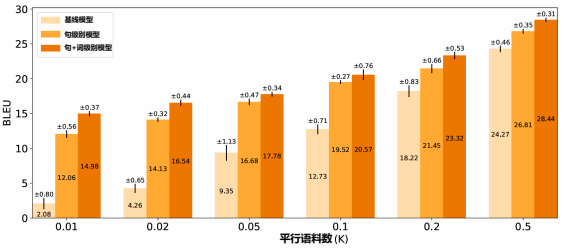
少样本翻译　图 3-3 展示了在四个测试集上少样本翻译的实验结果。在不同数量的平行语料条件下，S+T-CTR 模型均优于基线模型和 S-CTR 模型，这证明了本章方法在少样本场景中的有效性。且多次实验的方差均在合理范围内，这表明平行语料的随机采样对实验结果影响较小，本章方法具有较强的鲁棒性。



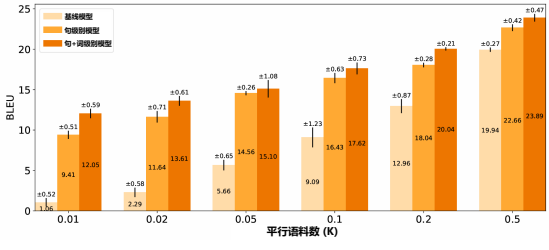
(a) 法语 → 英语 Test2016



(b) 法语 → 英语 Test2017



(c) 法语 → 英语 MsCOCO



(d) 捷克语 → 英语

图 3-3　少样本翻译实验结果

Figure 3-3　Results of few-shot translation

3.5 分析实验

3.5.1 跨模态对齐

本章所提出方法的核心思想是将源端高资源语言和低资源语言的文本和图像在公共语义空间中进行语义对齐。为了验证源端语义对齐是否成功实现，本节进行了下列分析实验，包括文本到图像的检索实验，以及选择性注意力模块的注意力图可视化分析。

文本到图像检索 文本到图像检索是指找到与文本最接近的前 K 张图像。本节分析实验计算了 $K = 1, 5, 10$ 时的 Recall@K 分数。如表 3-3 所示，S-CTR 模型在 R@1/5/10 上相比基线模型分别提升了 34.2/64.2/74.1%，证明了对比学习在跨模态对齐这一任务中的有效性。此外，S+T-CTR 模型相比于 S-CTR 模型在 R@1/5/10 指标上进一步提升了 1.9/1.5/0.7%，这表明细粒度的学习目标能够实现更好的语义对齐。

表 3-3 文本到图像检索在法语到英语 Test2016 数据集上的实验结果

Table 3-3 Text-to-image retrieval on FR→EN Test2016

模型	R@1↑	R@5↑	R@10↑
Baseline	0.2	0.8	1.3
句级别模型	34.4	65.0	75.4
句 + 词级别模型	**36.3**	**66.5**	**76.1**

注意力图 为了进一步验证词级别对比学习在跨模态对齐任务种的效果，本节分析实验提取了选择性注意力模块的注意力图，并进行了可视化。如图 3-4 所示，选择性注意力模块成功地注意到了和文本语义相关的区域。例如，法语单词"gens"（意为"人们"）所对应的图中的三个人区域注意力增强，而单词"maison"（意为"屋顶"）所对应的屋顶区域注意力也得到了增强。

3.5.2 跨语言对齐

第 3.5.1 节分析了本章所提出的对比学习方法在跨模态对齐任务中的有效性。然而，本章方法的最终目标是通过跨模态对齐实现跨语言对齐，即为所有源端语言学习构建一个公共语义空间。

为了对该公共语义空间进行直接的分析，本节实验在零样本场景下比较了基线模型和 S+T-CTR 模型，即没有任何法语 → 英语或捷克语 → 英语的平行数据参与训练。首先需要对 *源编码器*的输出取平均，并使用 T-SNE[90] 工具将其降

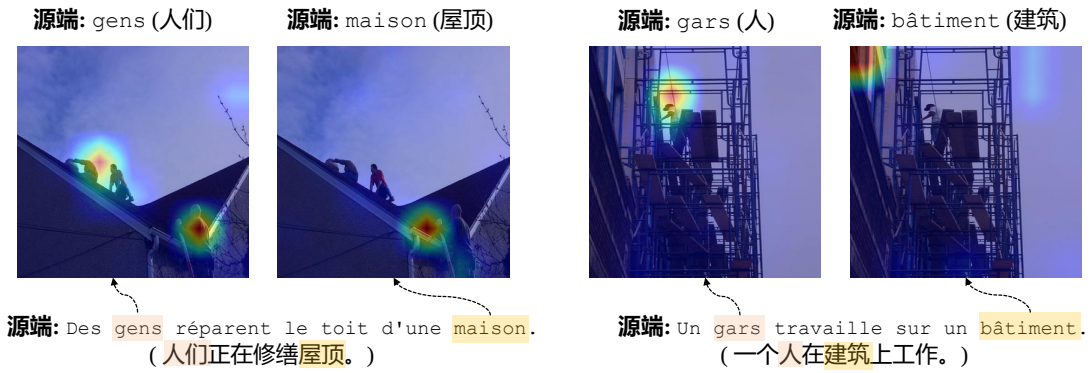
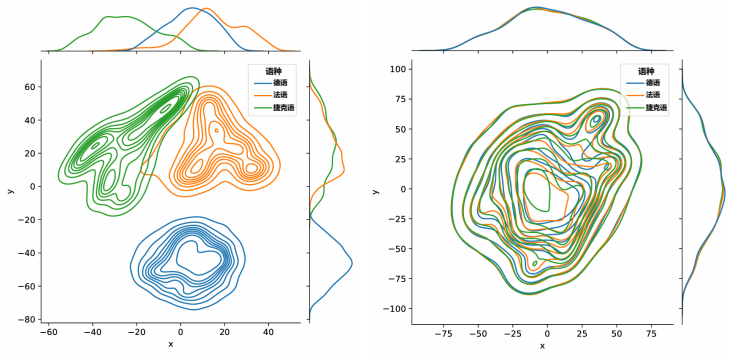


图 3-4　两个实例的选择性注意力可视化图

Figure 3-4　Attention maps of the selective attention module of two cases.

维至二维以进行可视化。如图 3-5 所示，在没有对比学习的情况下，不同源语言的特征表示之间存在明显的区分。相反，在使用对比学习后，三种语言的表示明显重叠，这证明本章方法成功借助跨模态对齐实现了跨语言的对齐。本节实验所使用的数据均来自于 Multi30K Test2016 测试集。



(a) 基线模型　　**(b) S+T-CTR 模型**

图 3-5　零样本场景下德、法、捷克语特征表示可视化

Figure 3-5　Visualization of representations for DE, FR, and CS under the zero-shot scenario

3.5.3　消融实验

目标语言对比学习　在标准的机器翻译中，目标语言通常与源语言相互分离，只在解码器的注意力机制中进行交互。然而，在实验的过程中，发现将目标语言加入对比学习的训练中能够进一步提升模型的效果。如表 3-4 所示，本节对是否加入目标语言展开了消融分析，分别在零样本场景和 100 个平行句对的少样本场景进行了分析实验，发现在没有目标语言对比学习的情况下，模型在零样本和少样本场景下的 BLEU 分数显著下降。因此，可以得出结论，将目标语言加入对比学习中有助于建立源语言和目标语言之间的联系，对最终的翻译效果有益。

表 3-4 目标语言消融实验结果

Table 3-4 Ablation study on contrastive learning of the target language

模型	目标端语言	法语 → 英语		法语 → 英语	
		零样本	少样本	零样本	少样本
基线模型	-	0.24	13.44	0.09	9.10
S-CTR 模型	×	7.81	12.55	6.93	8.67
	✓	8.71	21.49	7.23	16.43
S+T-CTR 模型	×	14.47	13.16	10.97	8.24
	✓	**16.49**	**22.85**	**13.58**	**17.62**

对比损失 vs. L2 损失 在有关跨模态对齐的任务中，对比学习损失并不是拉近模态之间距离的唯一方法，L1 距离、L2 距离损失等均可作为替代。为了分析本章所采用的对比学习损失的优越性，本节实验尝试用 L2 损失替换对比学习损失来对模型进行训练，并在零样本场景和 100 个平行句对的少样本场景下进行评估，具体损失函数如下：

$$\mathcal{L}_{L2} = \sum_{i=1}^{M} \|x_i - y_i\|^2. \tag{3-10}$$

实验结果如表 3-5 所示，在各个条件设置下，对比学习损失的表现均优于 L2 损失，这是因为对比学习损失不仅可以拉近正例之间的距离，还可以拉远负例之间的距离，这对于图像的理解和区分是有利的。而 L2 损失只能直观地拉近正例之间地距离，但不起到区分的作用，模型无法学到精确的图像文本对应关系。

表 3-5 L2 损失和对比学习损失的 BLEU 值结果

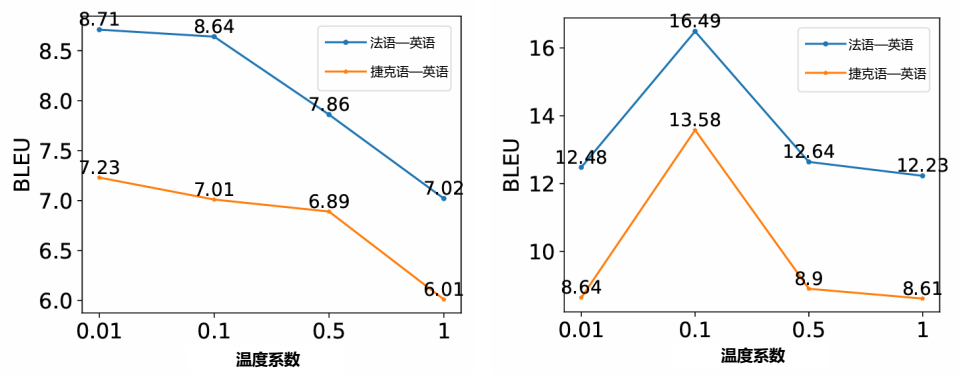
Table 3-5 BLEU scores of models with L2 loss and contrastive loss

模型	损失	法语 → 英语		捷克语 → 英语	
		ZS	FS100	ZS	FS100
基线模型	-	0.24	13.44	0.09	9.10
句级别	L2	8.45	19.60	6.83	15.05
	对比学习	8.71	21.49	7.23	16.43
句 + 词级别	L2	6.02	15.61	5.94	12.76
	对比学习	**16.49**	**22.85**	**13.58**	**17.62**

3.5.4　温度超参数

温度系数 τ 是对比学习损失函数中的一个重要超参数。更小的温度系数可以帮助模型更好地区分正例和负例，但温度系数过小也会使模型过度区分，起到反作用。因此，对于温度系数的选择，本节进行分析实验，分别选择 0.01、0.1、0.5 和 1 四个不同的温度系数值进行实验。图 3-6 展示了在 Multi30K 验证集上不同温度系数条件下的 BLEU 分数。

对于句级别的对比学习，可以观察到较低的温度能够获得更好的结果，因此尝试选择尽可能低的温度系数。然而，实验的过程中发现低于 0.007 的温度系数可能导致梯度爆炸。基于此，最终选择句级别对比学习温度系数 $\tau_s = 0.007$。对于词级别的对比学习，与句级别不同的是，过小的温度系数并不能带来更好的效果，通过实验分析，发现 $\tau_t = 0.1$ 在验证集上取得了最佳结果，主要原因是句子中的词不应被过度区分。



(a) 句级别　　(b) 词级别

图 3-6　不同温度系数的对比学习 BLEU 值

Figure 3-6　BLEU scores of different temperatures for contrastive learning

3.5.5　实例分析

在本节中，主要通过几个示例对本章方法进行定性分析。表 3-6 展示了不同模型的参考翻译结果和实际翻译结果。红色文本 表示语法或词汇错误，(括号中的单词) 表示漏译错误，绿色文本 则表示正确的翻译。

首先，在零样本场景下比较 S-CTR 模型和 S+T-CTR 模型。在案例 1 中，“two trees”未被 S-CTR 模型翻译，而 S+T-CTR 模型正确地翻译了它。案例 2 中同样出现了类似的情况，S-CTR 模型缺失 “a man in a dark blue shirt”，而 S+T-CTR 模型补全了缺失的翻译。这两个案例表明，**细粒度的词元级别对齐可以避免漏译**。

然而值得注意的是，在零样本场景下，S-CTR 模型和 S+T-CTR 模型都可能存在语法问题，这类语法问题可以通过使用少量平行语料进行微调来解决。在案

例 1 中，所得到的翻译短语“playing a game of dirt”显然不合逻辑和语法，而经过额外的 100 条平行语料微调后的模型成功将介词“of”修正为“in”，使其更符合语法。这一现象表明，**仅通过对比学习难以学习到充分的语法知识，但使用少量平行语料进行微调即可弥补这一缺陷**。

表 3-6 Multi30K Test2016 测试集定量分析实例

Table 3-6 Qualitative examples from Multi30K Test2016 set

模型		
实例 1 法语 → 英语		
参考译文	源端	Des enfants sont dehors , jouant dans la terre à côté de deux arbres.
	目标端	Some children are outside playing in the dirt where two trees are.
S-CTR 模型 (零样本)		The young children are playing a game of dirt. *(two trees)*
S+T-CTR 模型 (零样本)		The children are outside playing a game of dirt next to two trees.
S+T-CTR 模型 (少样本)		The children are outside playing a game in the dirt near two trees.
实例 2 捷克语 → 英语		
参考译文	源端	Muž ve žluté košili a muž v tmavém modrém tričku si povídají.
	目标端	A man in a yellow shirt and a man in a dark blue shirt talking.
S-CTR 模型 (零样本)		Man in yellow shirt is crying. *(a man in a dark blue shirt)*
S+T-CTR 模型 (零样本)		Man in yellow shirt *(and)* a man in a blue shirt is smiling.
S+T-CTR 模型 (少样本)		A man in a yellow shirt and a man in a blue shirt is talking.

3.6 本章小结

本章提出了一种基于图文对齐的，包含句级别和词级别对比学习方法的跨模态对齐方法，成功实现了低资源场景下零样本和少样本的翻译目标。实验结果表明，本章所提出的方法在这两种场景下均显著优于基线模型。与此同时，进一步的分析实验表明，本章方法能够成功实现跨模态和跨语言对齐的目标，构造源端语言的公共语义空间，进一步证明了本章方法的有效性。

第 4 章　基于图文对齐的无监督机器翻译

4.1　引言

第 3 章主要介绍了图文对齐方法在低资源翻译场景下的应用，但其中仍存在一定的问题。在进行源端语义对齐时，需要（源语言，目标语言，图片）三元组数据作为高资源数据，同时需要高资源语言的平行语料实现源端至目标端的映射，但事实上，此类数据在低资源场景中并不充足，获取的难度较大。如果只使用（源语言，图片）以及（目标语言，图片）这样的单语图片-文本描述数据实现翻译系统，是本章主要探讨的话题。因此，本章将以图片为锚点的语义对齐方法引入无监督机器翻译的场景中，利用无监督机器翻译不需要任何平行语料即可实现翻译系统的优势，解决上述问题。

无监督机器翻译（Unsupervised Machine Translation）的目标是在没有任何平行语料库用于训练的情况下，实现源端至目标端的映射，将文本从源语言翻译到目标语言。目前一类代表性的方法[17–21] 按照一种三阶段范式实现这一目标：单语建模、初步对齐和迭代回译。单语建模是指在大规模单语语料库上训练模型，以学习不同语言的单语表示。初步对齐作为回译的先验步骤，通过为模型提供粗粒度的翻译能力来启动后续过程。之后，在回译阶段迭代生成伪平行语料库，以学习细粒度的源语言到目标语言的对齐。在三个阶段中，初步对齐具有重要的作用，正如 Lample 等[18] 和 Huang 等[22] 所讨论的，初步对齐作为回译的起点，为回译通过迭代优化的翻译能力奠定了基础。因此，无监督机器翻译系统的性能在很大程度上依赖于初步对齐的质量。

近年来，由于多模态方法的蓬勃发展，越来越多的研究[53–56] 开始利用图像来辅助机器翻译，从而催生了多模态机器翻译（Multi-modal Machine Translation）这一领域。与此同时，一些研究者[14,15,22,62,91] 也将图像的应用扩展到无监督机器翻译，推动了无监督多模态机器翻译（Unsupervised Multi-modal Machine Translation）的发展。例如，Huang 等[22] 利用图像描述模型生成伪平行句子，从而利用数据增强提高翻译的性能。Fei 等[91] 则通过视觉和语言场景图表示输入图像和文本，以学习语义对齐。然而，这些方法与多模态机器翻译的相关方法类似，即在训练过程中融合图像和文本信息或引入额外模型，从而间接地对齐语义空间。实际上，视觉模态作为一种与语言无关的信号，具有将不同语言的相同语义表示直接对齐到共同空间的潜力。此外，单语图像-文本描述数据在社交网络上非常丰富且易于获取。与平行语料不同，此类数据仅需要单语使用者对图像进行描述性标注，而无需双语专家参与进行翻译，这大大降低了语料的获取成本。

基于此，本章提出了一种基于图文对齐的无监督机器翻译方法，以实现更好

的初步对齐。本章方法通过对比学习，利用图像作为枢纽，将源语言和目标语言在语义上对齐到一个共享的潜在语义空间中。具体而言，本章方法通过引入了句子级别的对比学习来学习粗粒度的对齐，以及词级别的对比学习来实现细粒度的对齐。通过句级别和词级别的对比学习，具有相似语义的源端语言和目标端语言在公共语义空间内将拥有相似的表示特征，从而使得模型在第三阶段的回译之前就能够具备更好的初步翻译能力。实验和分析表明，本章方法在翻译性能上显著优于纯文本和多模态基线模型，并有效地实现了更好的源语言到目标语言的对齐，在回译之前为模型提供了良好的初始翻译能力。此外，本章方法在领域外数据集上也表现出了一定的效果，体现了其泛化能力。

4.2 相关工作

在第 3 章的相关工作中，已经对跨模态图文对齐进行了详细的介绍，因此本章主要介绍无监督机器翻译与多模态无监督机器翻译方面的相关工作。

4.2.1 无监督机器翻译

无监督机器翻译（Unsupervised Machine Translation）指的是仅使用单语语料库来实现源端至目标端的对齐，从而完成翻译任务。早期的方法[57–60] 使用第三种语言作为枢纽来实现零样本翻译，但这些方法并未完全克服对平行语料库的依赖。Lample 等[17,18,19] 则提出了一种新颖的无监督方法，该方法通过大规模单语数据初始化模型，并通过回译构建伪平行语料库来训练源语言到目标语言的对齐。在后续的研究中，Conneau 等[20], Song 等[21] 沿用了这一思路，并改进了预训练方法。然而，正如 Lample 等[18] 所提到的那样，源语言到目标语言的对齐具有不确定性。因此，本章方法加入视觉模态作为辅助，以图片为锚点进行对比学习的训练，从而获得更好的语义对齐空间。

4.2.2 多模态无监督机器翻译

无监督多模态机器翻译（Unsupervised Multi-modal Machine Translation）旨在引入视觉模态以增强无监督机器翻译系统的性能。在之前的研究中，一类研究，如 Chen 等[61], Su 等[62] 通过融合视觉和文本信息，构建多模态融合特征来增强无监督机器翻译模型。另一类研究则是以图像为枢纽，连接源端与目标端，从而实现零样本翻译的目标。Nakayama 等[14] 以 RNN 为基础模型，借助图像链接文本特征和图像特征，直接从图像特征中解码目标端语句。同样的，Li 等[15], Huang 等[22] 则以图片为锚点，构建公共语义空间，使相似语义的文本拥有近似的特征表示。然而，这些方法在推理阶段仍然需要图像作为输入，这为方法的实际应用带来了困难。本章方法则进一步扩展了这一研究方向，在推理阶段无需图像作为输入的情况下实现了更好的性能。

4.3　背景介绍

本章方法采用无监督机器翻译的总体技术路线，主要分为以下三个部分：单语建模，初步对齐，迭代回译。其中，本研究在初步对齐环节中引入跨模态技术，利用多模态信息辅助源端与目标端之间建立初步的映射，使模型在该阶段即可具备一定的翻译能力。

4.3.1　任务定义

本章方法的任务目标是仅使用单语图片-文本描述数据，实现 L_x 到 L_y 以及 L_y 到 L_x 的翻译模型。对于语言 L_x 和 L_y，其包含单语图片-文本描述数据 $\mathcal{D}_{L_x} = \{(\mathbf{i}, \mathbf{x})\}$ 和单语图片-文本描述数据 $\mathcal{D}_{L_y} = \{(\mathbf{i}, \mathbf{y})\}$、其中 $\mathbf{i}$ 是图像，$\mathbf{x}$ 和 $\mathbf{y}$ 分别是 L_x 和 L_y 中的描述。需要注意的是，不同语言之间不会共享相同的图像。

4.3.2　无监督机器翻译

单语建模　单语建模的本质是训练一个单语的概率生成模型，这需要使模型学习如何理解和生成该语种的句子。一种常用的训练方法是采用去噪自动编码器（DAE），即训练模型从噪声版本中重建其输入。在 DAE 框架的基础上，研究者们提出了一些改进方案。例如，[17] 采用了随机删除词语和随机排列词语的方式添加噪声，[20] 则使用了跨语言预训练方法，MASS[21] 利用了基于片段的掩码策略来构造噪声，进行序列到序列去噪自编码学习。

而在无监督机器翻译中，并行结构是一种较为常见的单语建模框架，如图4-1（阶段 1）所示，其中源语言和目标语言各自配备一个编码器和解码器，需要同时对训练源端语言和目标端语言的编码器和解码器进行单语建模，即分别对“源端编码器——源端解码器”和“目标端编码器——目标端解码器”两个方向进行去噪自编码器的训练。本章用 $\mathcal{D}_x = \{\mathbf{x}_i\}_{i=1}^{M_x}$ 和 $\mathcal{D}_y = \{\mathbf{y}_i\}_{i=1}^{M_y}$ 分别表示源语言和目标语言的单语数据集。为了构造噪声输入句子，对 $\mathbf{x}$ 和 $\mathbf{y}$ 分别添加噪声 $\delta()$，从而得到 $\delta(\mathbf{x})$ 和 $\delta(\mathbf{y})$。模型的训练目标是最小化如下的交叉熵损失函数：

$$\mathcal{L}_{\mathrm{LM}} = -[\sum_{i=1}^{|\mathbf{x}|} \log P_{S\to S}(x_i|\mathbf{x}_{<i}, \delta(\mathbf{x})) + \sum_{i=1}^{|\mathbf{y}|} \log P_{T\to T}(y_i|\mathbf{y}_{<i}, \delta(\mathbf{y}))]. \tag{4-1}$$

初步对齐　在迭代回译开始之前，初步对齐的目的是为模型提供基础的翻译能力。作为迭代回译过程的起点，初步对齐的质量对模型的最终翻译质量起着举足轻重的作用。[92] 在这个步骤中使用了已有的双语词典，[17,18] 使用了无监督方式构造的双语词典[93]，赋予了模型逐字翻译的能力。但这种翻译能力往往较弱，且忽视了短语和句子本身的流畅性。

本研究中采用对比学习方法，以图像模态为连接点，建立源端目标端公共语

义空间的初步映射。由于相同语义的文本表示在公共空间内相互靠近，源端和目标端的解码器均可从中解码出所需要的语种的译文。

迭代回译 迭代回译是一种为实现源端到目标端对齐而自动生成伪平行句子的方法。如图4-1(阶段 3) 所示，通过重新组合源端编码器和目标端解码器可以得到 $S \rightarrow T$ 上的翻译模型和 $T \rightarrow S$ 语向的翻译模型。这两个模型不断生成伪平行数据，再由更新过后的模型再一次生成伪平行语料进行下一轮迭代，从而提高翻译性能。具体来说，首先将 $\mathbf{x}$ 输入源端编码器，通过目标端解码器的输出产生 $\hat{\mathbf{y}}$。同样，$\mathbf{y}$ 被输入目标端编码器，通过源端解码器得到 $\hat{\mathbf{x}}$。由此，可以获取到伪语料 $\hat{\mathbf{x}}$ 和 $\hat{\mathbf{y}}$。通过这种方式，即可构造伪并行语料对 $(\mathbf{x}, \hat{\mathbf{y}})$ 和 $(\hat{\mathbf{x}}, \mathbf{y})$，通过最小化伪平行语料对之间的交叉熵损失来训练模型：

$$\mathcal{L}_{\mathrm{BT}} = -[\sum_{i=1}^{|\mathbf{x}|} \log P_{T \rightarrow S}(x_i | \mathbf{x}_{<i}, \hat{\mathbf{y}}) + \sum_{i=1}^{|\mathbf{y}|} \log P_{S \rightarrow T}(y_i | \mathbf{y}_{<i}, \hat{\mathbf{x}})]. \tag{4-2}$$

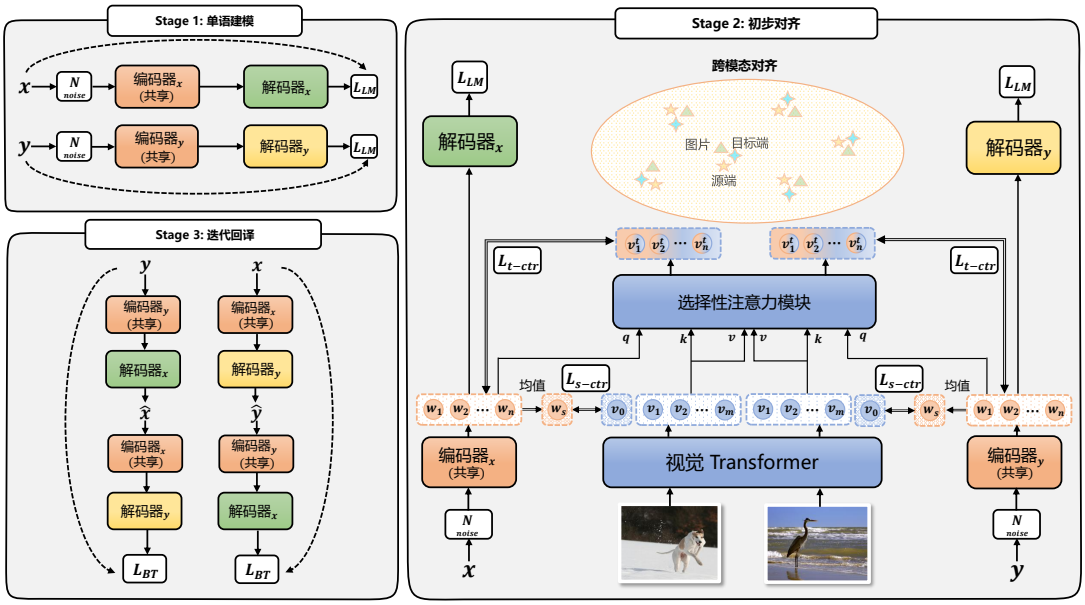


图 4-1 本章方法模型示意图

Figure 4-1 Overview of the model in this chapter

4.4 图片辅助的源端到目标端对齐

在本节中主要介绍所提出的用于建立初步对齐的图文对齐方法。该方法由两部分组成：粗粒度的句子级别对比学习和细粒度的词元级别对比学习[94]。

4.4.1　模型框架

本章方法模型建立在前文介绍的框架之上，该框架由两个编码器和两个解码器组成。对于源语言 x 和目标语言 y，其分别拥有单语的图片-文本描述数据集，其中包含 $(\mathbf{x}_i, \mathbf{i}_i)_{i=1}^{M}$ 和 $(\mathbf{y}_i, \mathbf{i}_i)_{i=1}^{N}$。需要注意的是，不同语言对应的图像并不重复，严格遵循单语图片-文本描述数据的前提。为了实现跨模态图文对齐，本章提出跨模态对比学习模块，该模块包含句级别和词级别的对齐方法。本节将以源语言的对齐过程为例进行说明，目标语言的对齐方式与其相同。

模型层面，源语言编码器和目标语言编码器均由 N 个 Transformer 编码器层组成。为了训练公共语义空间，这两个编码器共享各自的参数。对于输入句子 $\mathbf{x} = (x_1, ..., x_n)$，编码器的输出表示为 $\mathbf{w} = (w_1, ..., w_n)$。而解码器则由 N 个 Transformer 解码器层组成。关于图像编码器，本章方法采用 Vision Transformer (ViT)[2] 来提取视觉特征。ViT 将图像编码为序列 $\mathbf{v} = (v_0, v_1, ..., v_m)$，其中 v_0 是特殊的 [class] 标记，其余部分则表示图像的不同区域。

4.4.2　句级别对比学习

对比学习的核心思想[9] 是在特征表示空间中拉近相互对应的样本（正样本对）之间的距离，同时推远无关样本（负样本对）之间的距离。本章方法首先在句子级别进行粗粒度对齐。具体而言，先对文本编码器的输出取平均，将其作为文本的句子级表示，同时使用 ViT 的特殊 [class] 标记 v_0 所提取出的特征向量作为图像的全局特征：

$$\mathbf{w}^s = \frac{1}{n}\sum_{i=1}^{n} w_i,\ \mathbf{v}^s = v_0. \tag{4-3}$$

通过这种方式，可以将一个大小为 B 的图文批次分别编码为 $\mathbf{W}^s = \mathbf{w}i^s i = 1^B$ 和 $\mathbf{V}^s = \mathbf{v}i^s i = 1^B$。在该批次中，$(\mathbf{w}_i^s, \mathbf{v}_i^s)$ 构成正样本对，而 $(\mathbf{w}_i^s, \mathbf{v}_j^s)(i \neq j)$ 则为负样本对。为了拉近正样本对的距离，同时保持负样本对的区分度，本章采用 InfoNCE 损失[95] 来实现这一目标：

$$\mathcal{L}_{\text{s-ctr}}(\mathbf{x}, \mathbf{i}) = -\sum_{i=1}^{M}[\log \frac{\exp(s(\mathbf{w}_i^s, \mathbf{v}_i^s)/\tau)}{\sum_{j=1}^{M}\exp(s(\mathbf{w}_i^s, \mathbf{v}_j^s)/\tau)} + \log \frac{\exp(s(\mathbf{v}_i^s, \mathbf{w}_i^s)/\tau)}{\sum_{j=1}^{M}\exp(s(\mathbf{v}_i^s, \mathbf{w}_j^s)/\tau)}], \tag{4-4}$$

其中，$s()$ 表示余弦相似度，其计算方式为 $s(a, b) = \frac{a^\top b}{|a||b|}$，$\tau$ 为温度超参数。

4.4.3　词级别对比学习

前文的句子级对比学习已经实现了粗粒度对齐。此外，本章方法进一步利用词级别对比学习来学习细粒度的对齐关系，从而提升模型性能。

在词级别对比学习中，需要关注的对象是每个句子及其对应的图像，并将它们分别编码为序列 $\mathbf{w} = (w_1, ..., w_n)$ 和 $\mathbf{v} = (v_1, ..., v_m)$。由于文本序列与图像序列

的长度不一致，并且图像的全局特征中通常包含冗余信息，本章方法采用选择性注意力机制（selective attention）[54] 来标准化序列长度，并过滤掉无关信息，将 $\mathbf{w}, \mathbf{v}, \mathbf{v}$ 分别作为选择性注意力机制中的查询（query）、键（key）和值（value）。

$$\mathbf{v}^t = \text{Softmax}\left(\frac{(W_Q \cdot \mathbf{w})(W_K \cdot \mathbf{v})^\top}{\sqrt{d_k}}\right)(W_V \cdot \mathbf{v}), \tag{4-5}$$

其中，W_Q、W_K 和 W_V 为可学习的矩阵参数。通过上述模块，可以得到长度一致的文本序列 $\mathbf{w} = (w_1, ..., w_n)$ 和经过选择性注意力变换后的图像序列 $\mathbf{v^t} = (v_1, ..., v_n)$。在此对比学习任务中，$(w_i, v_i^t)$ 构成正样本对，而 $(w_i, v_j^t)(i \neq j)$ 则为负样本对，词级别对比学习的损失函数定义如下：

$$\mathcal{L}_{\text{t-ctr}}(\mathbf{x}, \mathbf{i}) = -\sum_{k=1}^{M}\sum_{i=1}^{|\mathbf{w}|}[\log\frac{\exp(s(w_i, v_i^t)/\tau)}{\sum_{j=1}^{|\mathbf{w}|}\exp(s(w_i, v_j^t)/\tau)} + \log\frac{\exp(s(v_i^t, w_i)/\tau)}{\sum_{j=1}^{|\mathbf{w}|}\exp(s(v_i^t, w_j)/\tau)}]. \tag{4-6}$$

4.5 实验设置与结果

4.5.1 数据集

本章实验使用的数据来自三个数据集：WMT News Crawl、MsCOCO[82] 和 Multi30K[81]。WMT News Crawl 是一个大规模的单语数据集，涵盖多种语言。本章对 2007 至 2017 年的 WMT News Crawl 数据进行随机混合，并选取其中的前 1000 万个句子用于训练。MsCOCO[82] 是一个带有英文标注的图像数据集。具体而言，本章使用的是 Caption 2015 数据集，其中包含 121,000 组图文对。参考 Huang 等[22] 的实验设置，本章将该数据集的一半翻译为德语和法语。Multi30K[81] 是一个多模态机器翻译的基准数据集。其训练集和验证集分别包含 29,000 和 1,014 组德语、法语和英语的句子，并配有相应的图像。在评估阶段，本章使用 Multi30K Test2016、Test2017 和 MsCOCO 测试集对模型进行评测，这些测试集分别包含 1,000、1,000 和 461 个样本。

4.5.2 训练过程

单语建模 参考 Su 等[62] 和 Huang 等[22] 的方法，本章将 WMT 单语语料库的一个 1000 万子集与 14.5K（Multi30K 的一半）数据集的十倍量级结合，得到一个包含 1014 万句子的综合单语数据集。本章采用 MASS[21] 中的单语建模方法，通过掩蔽原始句子中的一个连续片段，并要求解码器重建被掩蔽的部分，更多有关 MASS 方法的细节可参见原文，此处不做赘述。

初步对齐 在这一阶段，本章实验使用的是一个包含每种语言各 75,000 个单语图片-文本描述的数据集，为了确保不同语言中的图像不重叠，该数据集各取

COCO 和 Multi30K 数据集的一半。在此过程中，由于 MASS 方法输出的句子为片段形式，因此本章额外引入了一个词级别重构损失（token mask loss），以使输出的句子更加流畅。具体做法是随机屏蔽输入中的一些词语，并要求解码器输出完整的句子。

迭代回译　最后，在迭代回译阶段，为了与基线系统进行公平比较并保障单语数据的前提，本章使用 Multi30K 单语数据集的一半（14.5K）进行迭代回译训练。值得注意的是，为了增强模型的适用性，不同于大多数 UMMT 系统[22,62]，本章实验在迭代回译的训练过程中未引入任何视觉模态，从而能够得到一个推理时不依赖图像的模型，增强其应用性。

训练数据　在三个训练阶段中，所使用的数据集和数据量均有区别。在单语建模中，本研究采用 WMT News Crawl 数据集中的 10M 单语数据用以单语预训练任务；在初步建模阶段，则使用 MsCOCO 和 Multi30k 数据集中的图文对数据用以训练跨模态图文对齐任务；最后的迭代回译阶段则采用 Multi30k 中的纯文本数据进行训练，具体的训练阶段所使用的数据集和数据量如下表所示：

表 4-1　不同训练阶段所使用的数据集和数据量

Table 4-1　The datasets and data volumes used in different training stages

阶段	数据种类		WMT	Multi30K	MsCOCO
	纯文本	图文对			
单语建模	✓	×	10M	14.5K	-
初步对齐	×	✓	-	14.5K	60.5K
迭代回译	✓	×	-	14.5K	-

4.5.3　实验设置

本章方法的模型基于 Transformer[3] 架构构建，编码器和解码器均有 $N = 6$ 层。注意力头的数量设置为 4，输入嵌入维度为 512，前馈嵌入维度为 1024。dropout 率设置为 0.3，标签平滑设置为 0.1。训练优化方面，本章实验使用 Adam 优化器[85] 并进行 2000 次 warm-up 更新，学习率设置为 5e-4。每个批次最多包含 4096 个词元。在语言建模阶段，共进行 15 步的训练。

本章方法使用 ViT[2] 作为图像编码器，将图像转换为 512 维的嵌入。输出序列的长度为 50，其中包括一个特殊的 [class] 标记和 49 个特征标记。在初步对齐和迭代回译阶段，本章实验保持与单语建模一致的训练参数。此外，在训练过程中引入早停策略，即如果验证集损失在 10 步内未能下降，则终止训练。

在评估阶段，本章实验取最后 5 个检查点的平均参数，并使用束搜索（beam search）进行解码，束宽为 5，并使用 multi-BLEU[87] 评分，通过 multi-bleu.pl 脚本[1]计算模型的 BLEU 分数，同时使用 METEOR 工具[2]计算 METEOR[96] 分数。本章系统的实现基于 fairseq[3][86]。实验在 4 台 NVIDIA 3090 GPU 上进行。

4.5.4 基线系统

本研究中将提出的模型分别与无监督纯文本模型和多模态基线模型进行了比较。为了公平比较，在推理阶段，无监督多模态翻译系统和本章模型保持一致，不接收任何输入图像。纯文本基线包括 MUSE[19]、UNMT[17]、XLM[20] 和 MASS[21]。多模态基线模型包括 UMMT[62], PVP[22], 和 SG[91].

4.5.5 评价指标

本章实验除了采用机器翻译领域常用的 BLEU 指标以外，还采用了 METEOR[96] 指标对翻译的结果进行评估。METEOR 是一个基于单精度的加权调和平均数和单字召回率的度量方法，该指标考虑了整个语料库上的准确率和召回率，最终得出测度，解决了 BLEU 的一些固有缺陷。

4.5.6 主实验结果

本章实验将模型与其他最先进的无监督机器翻译和多模态无监督机器翻译系统进行了比较。如表4-2所示，相较于纯文本模型，本章方法模型在 BLEU 得分上表现出显著的提升。与最先进的纯文本基线 MASS[21] 相比，本章方法在四个语言方向上平均提升了 5.1 个 BLEU 得分，平均提升了 3.2 个 METEOR 得分，这表明跨模态图文对齐在模型中发挥了关键作用。

表4-2的第二部分列出了所有多模态无监督机器翻译系统，为了确保公平比较，这些系统均未提供图像输入进行测试。UMMT 和 PVP 是完全没有图像输入的测试结果。但是由于它们在原始方法中包含图像输入，因此，本章采用由 Fei 等[91] 重新实现，采用了融合伪图像特征方法[53] 加入伪图像特征的结果 UMMT* 和 PVP*。显然，与其他 UMMT 系统相比，本章方法在 BLEU 和 METEOR 指标上都表现出了显著的提升。值得注意的是，本章方法与近期的最先进系统 SG[91] 相比，平均提升了 2.3 个 BLEU 得分和 2.5 个 METEOR 得分，这使得本章方法成为多模态无监督机器翻译领域的新最先进技术。

此外，本章实验还在 Multi30k 中的 Flickr2017 和 MsCOCO2017 测试集上进行了评估，而其他多模态无监督机器翻译系统并未进行此项评测，实验结果如

[1] https://github.com/moses-smt/mosesdecoder/blob/master-/scripts/generic/multi-bleu.perl

[2] https://github.com/cmu-mtlab/meteor

[3] https://github.com/pytorch/fairseq

表4-3所示。由于大多数多模态无监督机器翻译方法没有开放源代码，本章实验将实验结果与纯文本基线模型 MASS[21] 在这两个测试集上的结果进行了比较，本章方法同样取得了显著的改进，这进一步验证了本章方法的有效性。

表 4-2　Multi30k Test2016 数据集实验结果

Table 4-2　Results on Multi30k Test2016

	英语 → 德语		德语 → 英语		英语 → 法语		法语 → 英语		平均	
模型	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
• 纯文本基线模型										
MUSE[93]	15.7	-	5.4	-	8.5	-	16.8	-	11.6	-
UNMT[17]	22.7	-	26.3	-	32.8	-	32.1	-	28.5	-
XLM[20]	28.7	48.7	30.7	31.0	46.3	64.3	42.0	38.1	36.9	45.5
MASS[21]	27.3	48.1	32.3	33.0	47.6	64.5	43.3	38.3	37.6	46.0
• 多模态基线模型										
UMMT[62]	8.4	11.3	7.5	10.8	15.8	12.7	10.2	13.6	10.5	12.1
UMMT*[91]	15.7	17.7	19.3	22.7	30.4	28.4	31.8	30.4	24.3	24.8
PVP[22]	11.1	13.8	14.0	17.2	26.1	23.8	25.7	23.4	19.2	19.6
PVP*[91]	25.4	40.1	27.6	26.0	46.7	58.9	39.0	31.9	34.6	39.0
SG[91]	32.0	52.3	33.6	32.8	**50.6**	64.7	45.5	37.3	40.4	46.7
本研究方法	**36.0**	**55.2**	**38.2**	**36.5**	50.0	**65.3**	**46.6**	**39.7**	**42.7**	**49.2**

表 4-3　Multi30K Flickr2017 和 COCO2017 数据集实验结果

Table 4-3 Results on Multi30K Flickr2017 set and COCO2017 set.

		英语 → 德语		德语 → 英语		英语 → 法语		法语 → 英语		平均	
数据集	模型	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
F17	MASS	22.8	30.3	27.8	43.5	42.5	58.8	38.0	34.8	32.8	41.8
	本研究方法	**28.8**	**34.1**	**31.4**	**49.0**	**44.4**	**60.5**	**41.4**	**37.1**	**36.5**	**45.2**
C17	MASS	24.4	43.5	26.1	30.3	37.5	56.2	36.4	35.0	31.1	41.2
	本研究方法	**27.5**	**46.7**	**27.7**	**32.8**	**39.3**	**57.9**	**40.8**	**37.2**	**33.8**	**43.6**

4.6　分析实验

4.6.1　消融实验

本节对本章方法进行了消融实验分析，以量化每个目标的贡献，实验结果如表4-4所示，表中的字母含义如下：L：单语建模，S：句级别对比学习，T：词级别对比学习，B：迭代回译。根据表中结果，可以分析得到以下结论：(1) 跨模态初始化在模型中起着至关重要的作用。比较第 5 行和第 7 行，可以观察到所有语言方向的 BLEU 得分明显下降了 5.1 分。(2) 单语建模是方法中的另一个重要组成部分，这一步使得模型能够更好地学习单语表示，从而提升后续训练阶段的表现。比较第 2 行和第 4 行，使用单语建模训练的模型提升了 5.3 的 BLEU 得分。

(3) 此外，当比较第 3、4 行与第 6、7 行时，可以发现迭代回译显著提升了翻译性能（大约提升了 6 个 BLEU 分数），这突显了伪平行语料在方法中的重要作用。(4) 另外，比较第 1、3 行与第 2、4 行，基于句级别和词级别对比学习的方法相比仅有句级别对比学习的方法在 BLEU 得分上获得了约 1 分的提升，证明了精细对齐能够带来更好的结果。

表 4-4 消融实验中不同策略的 BLEU 得分

Table 4-4 Ablation studies. BLEU score of different learning strategies

	模型	英语 → 德语	德语 → 英语	英语 → 法语	法语 → 英语	平均
1	S	22.6	25.7	20.3	24.5	23.3
2	S+T	25.1	27.3	20.8	25.6	24.7
3	L+S	26.1	29.4	31.3	30.3	29.3
4	L+S+T	27.5	30.0	31.6	30.8	30.0
5	L+B	27.3	32.3	47.6	43.3	37.6
6	L+S+B	34.6	36.7	49.4	46.1	41.7
7	L+S+T+B (完整)	**36.0**	**38.2**	**50.0**	**46.6**	**42.7**

4.6.2 语义对齐分析

为了分析本章方法所提出的模型能否在潜在空间中实现不同语言之间的语义对齐，本节将从定性和定量两个角度对该问题进行分析。

奇异值差异与有效条件数 为了从定量角度展示本章所提出的方法应用前后公共语义空间的变化，本节对奇异值差异和有效条件数[97] 这两个定量指标展开了分析，以证明方法的有效性。奇异值差异表示了两个特征表示空间之间的信息差异。它测量了通过奇异值分解（SVD）得到的两个特征表示空间的奇异值之间的欧几里得距离。同时，有效条件数衡量了函数输出值在输入发生微小变化时的变化程度。两个指标越小，表明两个语言的表示空间越接近。关于这两个指标的计算细节可见附录二。

如表4-5所示，可以观察到，通过对比学习，英语-德语和英语-法语对的奇异值差异显著减少，有效条件数也相对下降。这证明了，在应用本章方法后，不同语种的语义空间得到了拉近，这进一步强调了本章方法的有效性。

翻译质量 为了进一步分析跨模态图文对齐的有效性，本节对模型在迭代回译之前的翻译性能进行了验证。如表 4-6 所示，其中 MUSE[93] 是一个由双语词典初始化的词对词翻译模型。本章方法模型相较于其他基线方法表现出显著的改

表 4-5　测试集上的奇异值差异以及有效条件数

Table 4-5　Singular value gap (SVG) and effective condition number (ECN) on test sets

	SVG		ECN		
模型	英语-德语	英语-法语	英语	德语	法语
阶段 1	63.3	4.7	24.3	19.4	24.3
阶段 1+2	**2.7**	**0.1**	**16.7**	**17.4**	**17.1**

进，甚至超越了经过回译训练的无监督机器翻译系统[17]。这表明，通过跨模态对比学习，模型成功地学习到了一个共同的语义空间，从而实现了更高质量的对齐。

表 4-6　迭代回译之前的 BLEU 值

Table 4-6　BLEU scores without back-translation

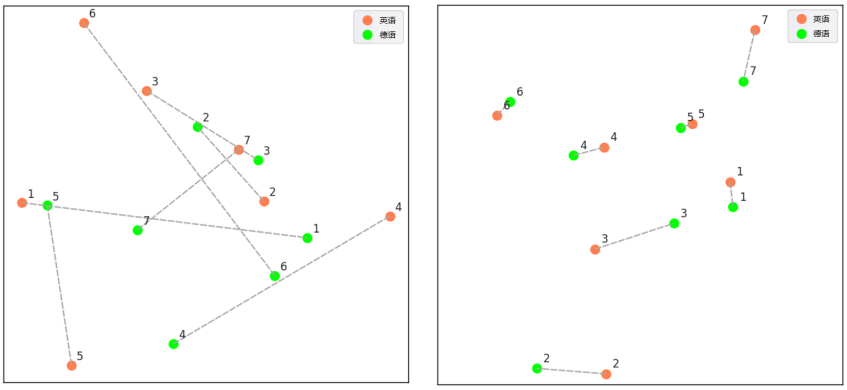
模型	英语 → 德语	德语 → 英语	英语 → 法语	法语 → 英语
MUSE	15.7	5.4	8.5	16.8
MASS	16.7	12.4	16.6	19.7
本研究方法	**27.5**	**29.9**	**31.6**	**30.8**

可视化分析　为了更直观地理解公共语义空间中的源语言和目标语言表示，本节使用主成分分析（PCA）将句子级表示的维度从 512 降至 2，并进行可视化分析。如图 4-2 所示，与基线模型相比，本章方法成功地减少了语义相似的句子表示之间的距离。实验中的测试集采用 Multi30K Test2016 数据集中的德语 → 英语和英语 → 德语部分.

4.6.3　领域外数据集表现

为了进一步验证本章方法的泛化性，本节在常用的 IWSLT 数据集上进行了额外的实验，该数据集常用于纯文本机器翻译，与本章所采用数据集是不同的领域。IWSLT 是一个口语数据集，包含了来自 TED 演讲的各种话题，相比于 Multi30K，更符合现实世界的翻译任务。为了准确评估模型的领域外性能，与之前的研究[91] 不同，本章方法没有引入任何额外的图像或采用文本到图像的检索来寻找匹配的图像。相反，本章方法仅依靠现有的 75K 图文对进行跨模态图文对齐训练，并仅在 IWSLT 数据集上训练迭代回译。

本节在 IWSLT14 EN-DE 和 IWSLT17 EN-FR 数据集上进行了分析实验。EN-DE 方向包含 174K 训练数据和 6.7K 测试数据，而 EN-FR 方向包含 236K 训练数



(a) 阶段 1　　(b) 阶段 1+2

图 4-2　句级别表示可视化分析

Figure 4-2　Visualization of sentence-level representations for DE and EN

据和 8.5K 测试数据。如表 4-7 所示，与强大的纯文本基准模型 MASS[21] 相比，本章方法在所有四个翻译方向上均表现出改进，证明了本章方法在领域外数据集上的有效性。

表 4-7 IWSLT14 英德和 IWSLT17 英法测试集上的 BLEU 值

Table 4-7　BLEU scores on IWSLT14 EN-DE and IWSLT17 EN-FR test sets

模型	英语 → 德语	德语 → 英语	英语 → 法语	法语 → 英语
MASS	22.6	21.9	33.1	31.9
本研究方法	**23.3**	**22.4**	**33.2**	**32.4**

4.6.4　低相似度语言对表现

Marchisio 等[98] 发现，高相似性的语言之间往往更容易实现在语义上对齐，而这对于低相似性的语言之间则更具挑战性。对于英语、德语和法语来说，它们之间有大量共享的词汇，表明这些语言之间的相似度较高。因此，为了探索当应用于低相似性的语言时，本章所提出的语义对齐方法的有效性，本节选择了捷克语进行进一步分析实验。与德语和法语不同，捷克语与英语不属于同一语言家族，英语属于印欧语系，而捷克语属于西斯拉夫语族。实验结果如表 4-8 所示，本章方法在捷克语上的表现仍优于 MASS[21]。这表明本章方法对于低相似性语言仍然具有有效性。

4.6.5　实例分析

本节通过具体的案例对模型的性能进行定性分析。表4-9对比了纯文本模型、未进行回译训练的模型以及完整模型的翻译结果。通过分析两个语言方向的案例，本章所提出的模型相比纯文本模型 MASS 展现出更优的翻译质量，例如案

表 4-8　Multi30K EN-CS Flickr2016 和 Flickr2018 测试集上的实验结果

Table 4-8　Results on Multi30K EN-CS Flickr2016 set and Flickr2018 set

		英语 → 捷克语		捷克语 → 英语		平均值	
数据集	模型	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
F16	MASS	20.1	23.9	27.1	29.3	23.6	26.6
	本研究方法	**24.2**	**26.4**	**30.8**	**32.2**	**27.5**	**29.3**
F18	MASS	16.1	21.2	22.3	27.1	19.2	24.1
	本研究方法	**20.0**	**24.1**	**26.6**	**30.4**	**23.3**	**27.3**

表 4-9　在 Multi30K 数据集上的实例分析

Table 4-9　Qualitative examples on Multi30K test sets

	**模型**		
	实例 1 德语 → 英语		
	参考	原句	Eine frau geht die straße entlang.
		译句	A woman walking down the street.
	无回译训练模型		A woman walks the street.
	MASS 模型		A woman walks down the street at night.
	完整模型		A woman walking down the street.
	实例 2 法语 → 英语		
	Ref.	原句	Un homme en costume tenant une boisson dans un gobelet marchant sur le trottoir, à côté d' un bus.
		译句	A male in a suit holding a beverage in a cup walking down the sidewalk, next to a city bus.
	无回译训练模型		A man in *(a)* costume holding a drink *(in a cup)* in a crosswalk walking on the sidewalk, near a bus.
	MASS 模型		A man in a suit holding a drink in a goggles walking on the sidewalk, at a bus stop.
	完整模型		A man in a suit holding a drink in a mug walking down the sidewalk, next to a bus.

例 1 中的 “at night” 以及案例 2 中的 “googles”、“at a bus stop” 等短语的翻译结果，本章模型可以生成更精准的翻译。在表格中，红色文本 标注翻译错误，绿色文本 标注正确翻译，（括号内词语）标注漏译内容。值得注意的是，回译训练对提升翻译质量（特别是语法准确性）具有关键性作用，未经回译训练的模型常出现语法错误，如案例 1 中 “walks the street” 这类不符合英语表达习惯的句式。这表明回译训练能有效提升生成语句的语法规范性。

4.7 本章小结

在本章中，提出了一种基于图文对齐的无监督机器翻译方法，该方法结合跨模态对比学习方法，使模型能够仅通过单语图片-文本描述数据，学习源语言到目标语言的对齐，从而提高无监督机器翻译中的初步对齐的质量。实验结果表明，与纯文本和多模态基线模型相比，本章方法取得了显著的性能提升，达到了多模态无监督机器翻译领域的先进水平。进一步的分析实验表明，本章方法成功地实现了特征表示空间中不同语言表征的语义对齐，在领域外数据集和低相似度语言上同样有效，具备较强的泛化性。

第 5 章　基于图片后验检索增强的视觉知识问答

5.1　引言

第 3 章和第 4 章主要介绍了图文对齐技术在低资源机器翻译和无监督机器翻译场景下的应用，为了进一步探索在不存在显式图片-描述数据的条件下该方法的有效性，本章将介绍该方法在视觉知识问答场景下的应用。

视觉知识问答是指模型需要通过检索增强的方式，检索与问题相关的额外图片信息，辅助模型生成答案。RAG（Retrieval-Augmented Generation）是检索增强方法的典型代表，通过结合外部知识库中的检索信息，提升了生成任务的准确性和多样性。在 RAG 任务中，模型可以通过检索与问题相关的文本描述或知识，将检索到的知识，以上下文学习[28] 的示例的形式作为模型输入的一部分，帮助生成更符合上下文语义的文本描述或解答。然而，在实际应用中，RAG 任务的数据模态往往不是单一的，还存在大量的文档、图片、表格、视频等多模态数据。在视觉知识问答任务中，模型需要从外部检索与问题内容相关的图片，如背景图片、实物图片等，从而生成更加准确的答案。

在视觉知识问答任务中，不同模态之间的模态鸿沟为检索的准确性带来了较大的挑战。在不存在显式图片-文本描述数据的条件下，图片与文本之间不存在明确的相关性区域。同时，二者之间还具有不同的特征表示空间，因此，在不同的模态之间进行语义对齐具有较大的难度，也是目前亟需解决的问题之一。举例来说，问题文本中可能存在大量的设问信息，仅有少部分描述性词汇；同时，文本模态存在多义性，一个词语往往含有多种含义，需要结合上下文来判断该词语的真正含义，与此同时，图像模态也存在着全局特征中包含大量语义无关西南西等问题。为了实现图文对齐与跨模态检索，Tan 等[16] 利用 CLIP 作为文本和图像特征的编码器，将所提取的特征进行求和平均得到多模态特征，将该特征作为多模态检索知识库的特征索引。Chen 等[23] 利用多模态大模型作为重排模型，将根据文本检索得到的图像特征进行重排，过滤掉与对应图像特征相似的强负例图像，同时在训练的过程中加入负例图像作为噪声，加强模型的鲁棒性。

上述方法较为粗糙，且并未对检索和生成阶段进行深入探索。在训练数据中，答案也包含了一部分重要信息，其也可作为筛选图像相关信息的标准之一。因此本章希望借助答案后验信息，将该信息作为图像和问题文本的连接点，辅助模型学习到图像中和问题更相关的区域，从而使模型能够自主提取图片的相关性区域，提高生成答案的准确性。具体而言，在检索阶段，本文首先以“问题”作为条件和以“问题，答案”为条件生成图像并提取特征，并将带有答案后验信息的图像特征与不包含后验信息的特征进行对齐，加强图像中有关答案后验信

息的区域特征；在生成阶段，通过对生成的答案添加相关性权重，使模型关注到答案后验信息中与图像更相关的关键部分，从而使模型自主学习并提高对图像中与问题的相关区域的关注度，从而实现了更为精确的图片检索，并提高了最终所生成答案的准确率。

5.2 相关工作

本章将介绍有关视觉知识问答任务的相关背景和相关工作。有关图文对齐的相关工作介绍已在第 3 章和第 4 章的相关工作中进行介绍，因此本章主要介绍检索增强以及视觉知识问答任务。

5.2.1 检索增强

在传统的自然语言处理任务中，当面对一些较为复杂或文本信息不充分的任务时，模型往往需要从外部获取与任务相关的外部知识来辅助任务的进行，而这类方法[63,64]也被证明能够有效提升模型的性能。其中，具有代表性的方法如DPR（Dense Passage Retrieval）[65]，一种密集段落检索方法，通过使用双编码器（序列编码器和段落编码器）将查询和段落映射到稠密向量空间，利用相似度计算进行高效检索。相比传统稀疏检索方法，DPR 通过预训练语言模型（如 BERT[99]）生成上下文感知的稠密向量，显著提升了检索精度和效果。REALM（Retrieval-Augmented Language Model）方法[66] 和 RAG（Retrieval-Augmented Generation）方法[63] 则通过结合外部知识检索与语言模型预训练得到一个检索增强的语言模型，使其能够动态检索相关文档并用于增强模型的生成和理解能力。

5.2.2 视觉知识问答任务

在纯文本的自然语言处理任务场景中，检索增强的相关方法已经被证明了其有效性[63–66]。因此，研究者们便尝试将此类方法扩展至多模态场景中。MuRAG[100] 首次提出了多模态检索增强生成框架，该方法构造了一个多模态检索增强生成 Transformer 模型，使其能够访问外部的多模态知识库来增强自身的语言生成。在多模态场景中，视觉知识问答[101–104] 是最具代表性的任务之一。与传统的视觉问答任务不同的是，视觉知识问答往往具备更大的挑战性，其中包含大量的专业知识或者专有名词等，模型需要从外界的知识库中检索相关的文本、图像知识，来辅助问答任务的进行。以 WebQA[103] 任务为例，模型需要根据问题，从外部知识库中检索出最相关的文本或图像信息，在这个过程中，会存在较多的干扰项，如何排除干扰，检索得到最相关的信息，是该任务的核心所在。一类方法[16,100,105,106] 通过借助大语言模型来辅助检索，其中，如 RagVL[23]，该方法通过训练一个重排序大模型，对 CLIP 检索得到的初步检索结果进行重排序，从而提高检索的精度。MI-BART[104] 提出了一种“检索-读取”的两阶段流程，将任

务分为两个不同的子任务。MHyS[107] 则是对查询的内容进行扩展，对图像进行多粒度的文本概括，使信息更加丰富，检索更加精确。上述方法为了实现更高精度的检索，在检索阶段往往依赖于大模型的帮助，如进行数据扩展、重排序等，而本课题希望能够以更小的训练代价来提升检索阶段的效果。

5.3　方法介绍

本章节主要介绍该方法的具体技术路线，主要分为四个部分：任务定义，总体流程，检索模型，生成模型。

5.3.1　任务定义

该子课题的任务场景为视觉知识问答，目标为通过检索到相关的多模态数据文档辅助多模态大模型的生成。首先，构建一个多模态数据文档数据集 $\hat{D}_M = \{(\mathbf{i}, \mathbf{c})\}$，其中的文档格式为（图片，描述）二元组。对于每一个待生成的问题 $\mathcal{D}_M = (\mathbf{q})$，以多模态检索的方式检索得到 K 个最相关的多模态文档数据 $\mathcal{D}_j = (\mathbf{i}_j, \mathbf{c}_j), j = 1, 2, ..., k$。最后，以 K 个多模态文档数据作为补充信息，生成最终所需要的答案 $\mathbf{a}_t$。

5.3.2　总体流程

本章节主要介绍方法的总体流程，主要分为四个部分：构建知识库，初步检索，重排序，生成。流程示意图如图5-1所示。

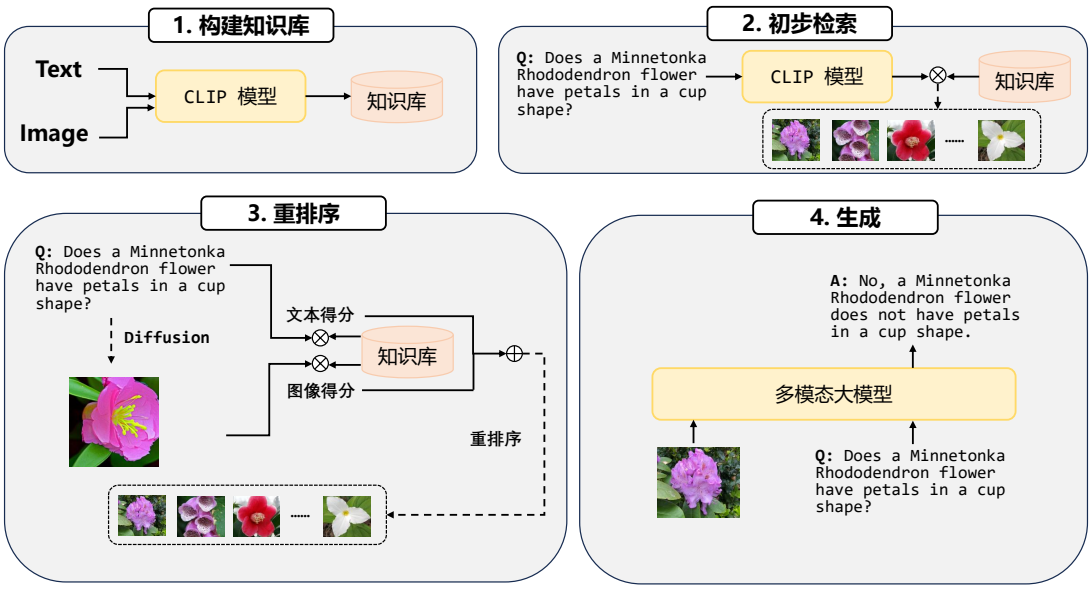


图 5-1　方法总体流程图

Figure 5-1　Overall flowchart of the method

构建知识库 首先需要构建一个多模态知识库，将多模态数据文档数据集 $\hat{D}_M = \{(\mathbf{i}, \mathbf{c})\}$ 编码为特征的形式进行储存，此处使用 CLIP[1] 作为图像和文本的编码器，将图像和对应的标注编码为图像特征和文本特征并进行储存。本文中的多模态数据库由 WebQA 训练集和测试集构成。

$$h_C = \mathbf{CLIP}(\mathbf{c}), h_I = \mathbf{CLIP}(\mathbf{i}), \tag{5-1}$$

其中 h_C,h_I 分别代表标注和图像被 CLIP 提取的特征表示。

初步检索 在得到一个多模态知识库后，首先根据待回答的问题进行初步检索。将问题输入 CLIP 文本编码器中，得到其文本特征 $h_Q = \mathbf{CLIP}(\mathbf{q})$，再将其与知识库中的图像特征计算余弦相似度，以余弦相似度为标准，筛选出与该问题最相关的 K 个图像特征，并读取这 K 个图像特征所对应的原图像。

$$\hat{I} = \underset{i \in \hat{D}_M}{TopK} \ cos(h_Q, h_I), \tag{5-2}$$

此处，K 值取 20，即初步检索得到 20 张与问题最为相关的图像 $\hat{I} = \{\mathbf{i}_1, \mathbf{i}_2, ..., \mathbf{i}_{20}\}$。

重排序 经过初步检索得到的相关图像往往具有较多的相似干扰项，因此需要进行进一步的重排序，以筛选出正确的相关图像信息。首先，根据待回答的问题，经过 Diffusion 模型[108] 生成一张与问题相关的伪图像。将伪图像经过训练得到的检索模型编码，得到其特征 h_{I^*}，有关检索模型的具体内容将在5.3.3中进行介绍。此时，在多个模态维度之间进行相似度计算，问题文本与图像特征之间的相似度得分 s_{TI}，问题文本与知识库中的标注文本特征计算相似度得分 s_{TT}，伪图像则与图像特征计算相似度得分 s_{II}。

$$s_{TI} = cos(h_Q, h_I), s_{TT} = cos(h_Q, h_C), s_{II} = cos(h_I, h_{I^*}), \tag{5-3}$$

将得到的得分与初步检索时的相似度得分进行加权相加，得到最终的得分。

$$s = \lambda_1 \cdot s_{TI} + \lambda_2 \cdot s_{TT} + \lambda_3 \cdot s_{II}, \tag{5-4}$$

此处分别取 $\lambda_1 = 0.5, \lambda_2 = \lambda_3 = 0.25$，最终根据该得分对初步筛选得到的 20 张图像进行重排序，得到其中最相关的 K 张图像。

生成 在经过重排序后，已经完成了检索阶段的所有流程，只需将检索得到的图像与待回答的问题输入多模态大模型中生成所需的答案即可。此时，也可以对多模态大模型进行一定的微调训练，使其能够更适应问题的数据集领域，提升回答的准确率，有关生成模型的相关方法将在5.3.3中进行详细介绍。

5.3.3　自主图文相关性提取

检索阶段　在重排序阶段，需要训练一个重排序模型以提取伪图像的特征，模型示意图如图5-2所示。模型中需要使用扩散（Diffusion）模型来进行图像的生成。Diffusion 模型，与传统的生成模型不同，它通过逐步噪声添加和去噪的过程生成数据，逐步将数据从随机噪声转换为逼近目标数据的样本。在图像生成任务中，Diffusion 模型通过逐步去噪过程，可以从噪声中生成与给定条件相符的图像。以图像生成任务为例，给定标签 c，Diffusion 模型可以在去噪过程中生成符合标签条件的重构图像 i_c。该标签可以是对图像的描述或特定类别信息，从而影响重构图像中特定区域的特征增强。

本研究的核心在于充分利用问题和答案的文本信息，对图像信息进行筛选，以提取与文本信息最相关的图像内容，并将利用问题和答案生成的图像特征与单独利用问题生成的图像特征进行对齐，从而使生成的图像特征更加精确和完整。首先，同时将“问题”与“问题 + 答案”的组合输入 Diffusion 模型生成对应的伪图像，再将伪图像输入 CLIP 视觉编码器中，经过全连接层后得到对应的两个伪图像特征。此时，本章方法希望对两个伪图像特征进行对齐，同时也需要保证图像的特征空间不产生偏移。因此本章采用了两个措施来实现目标，分别为图像标注任务以及真实图片。

图像标注任务，即在训练的过程中不仅仅关注特征之间的对齐，还需要对该特征进行图像标注任务的训练，即要求伪图像特征能够重构出所输入的问题，并在输出端使用 KL 损失对齐由两个伪图像特征生成的概率分布。图像标注任务的主要作用在于保障图像的特征空间不偏离图像本身，保证其语义性不由对齐任务而发生变化。同时由于输出端 KL 损失的存在，也起到了辅助对齐任务的作用。图像标注任务的损失采用交叉熵损失，如下所示：

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i=1}^{|\mathbf{c}|} \log p(c_i^* | \mathbf{c}_{<i}, \mathbf{i}^q) + \log p(c_i^* | \mathbf{c}_{<i}, \mathbf{i}^a), \tag{5-5}$$

其中 c 代表图像对应的标注，$\mathbf{i}^q$,$\mathbf{i}^a$ 分别表示由问题和问题与答案生成的伪图像，而 KL 损失需要拉近两个伪图像特征生成的概率分布，如下所示：

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}[p(c_i^* | \mathbf{c}_{<i}, \mathbf{i}^q) \parallel p(c_i^* | \mathbf{c}_{<i}, \mathbf{i}^a))]. \tag{5-6}$$

真实图像 $\mathbf{i}$ 的加入与图像标注任务的作用类似，主要是保证图像的特征空间不发生错误偏移，导致其过拟合至训练集空间上。分别将两个特征与真实的原有图像特征进行对比学习的训练，以真实图像为锚点，拉近伪图像特征之间的距离。

$$\mathcal{L}_{\mathrm{CTR}} = \mathcal{L}_{\mathrm{ctr}}(\mathbf{i}, \mathbf{i}^q) + \mathcal{L}_{\mathrm{ctr}}(\mathbf{i}, \mathbf{i}^a). \tag{5-7}$$

最终，在检索阶段，模型的训练损失为：

$$\mathcal{L}_R = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{CTR}}. \tag{5-8}$$

如图5-2中的右图所示，由“问题 + 答案”组合输入生成的带有答案后验信息的图像往往包含有更多的准确信息，如关键的“**gold**”颜色信息。以真实图像为锚点，拉近伪图像之间的特征表示距离，从而使由问题生成的伪图像包含有更准确的信息，提高检索的稳定性。

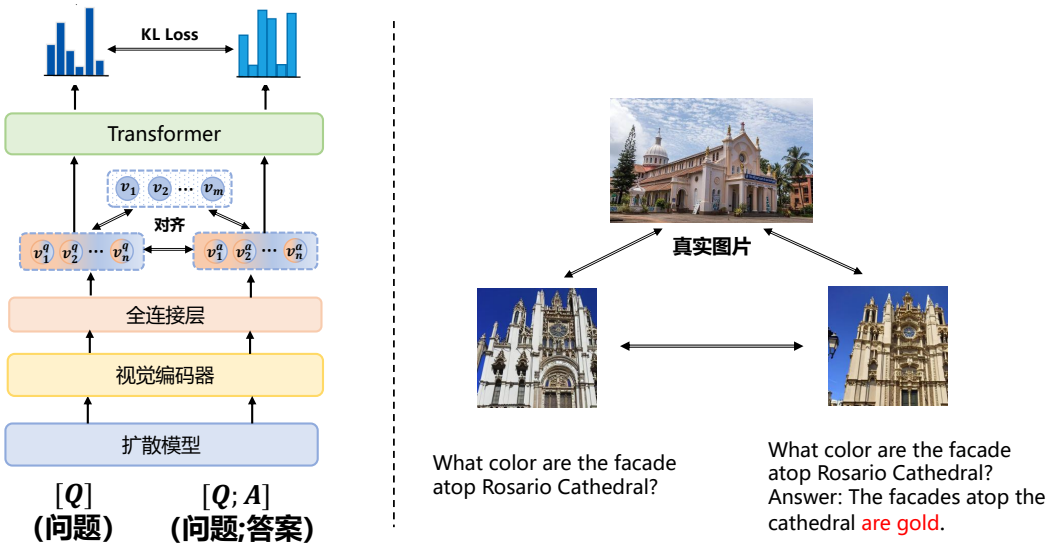


图 5-2　检索模型示意图

Figure 5-2　Overview of the retrieval model

生成阶段　本章方法拟采用 InternVL[4] 模型作为多模态大模型基座，并采用 CLIP 模型中的预训练文本编码器和图像编码器对文本和图像进行编码。并设计图像信息筛选模块，充分利用问题和答案后验信息，对图像的语义信息进行筛选，得到更精确的图像相关性区域表示特征。

InternVL 是一种可支持多种图文格式输入的多模态大模型，具有相对较强的性能和通用性。多模态大模型结合了自然语言处理和计算机视觉的能力，其设计目标是让大语言模型（LLM）不仅能够处理文本输入，还能理解和生成关于图像的语义信息，使得它可以执行如图像描述、视觉问答等复杂的视觉语言任务。其核心在于视觉指令调优 (Visual Instruction Tuning)，类似于在语言模型中进行指令调优。模型通过图像和相应的文本对进行训练，学习如何根据视觉内容进行生成式回答。这个过程使得多模态大模型能够理解复杂的视觉语义信息并生成上下文相关的语言输出。

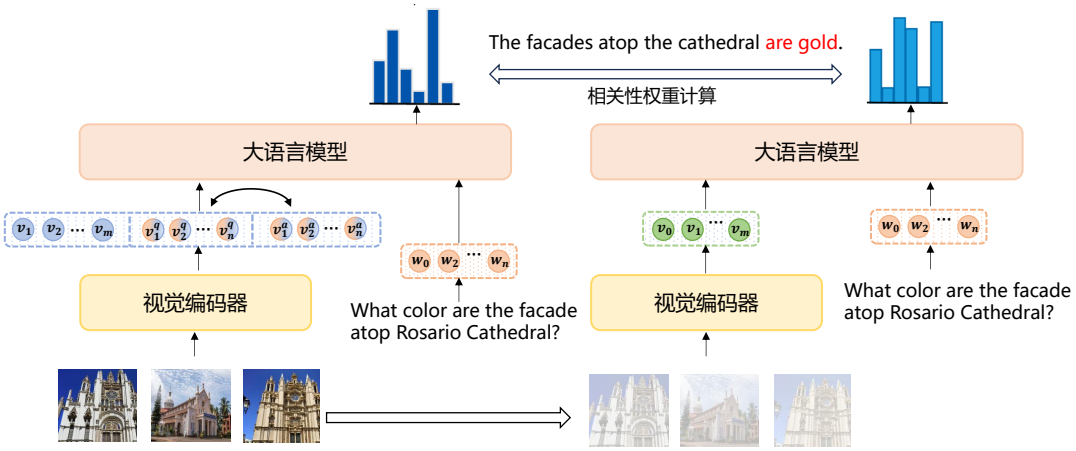


图 5-3　生成模型示意图

Figure 5-3　Overview of the generation model

与传统多模态大模型不同的是，InternVL 模型可支持单轮/多轮对话，单图/多图/纯文本对话，该模型不仅可以拼接多图输入，同时还可通过对输入添加特殊标记符和掩码的形式，使输入的图片可以对应到相关的输入，并区分图像和问题文本的具体位置。举例来说，模型可以支持如下输入："Image-1: <image> Image-2: <image> Describe the two images in detail."。通过这种形式，模型可以输出质量更高，更符合用户需求的输出，并支持多种不同的任务，如图像标注，视觉问答等等，具有更强的通用性。

生成阶段相关性提取　在生成阶段，模型的示意图如图5-3所示。本章方法同样希望模型能够基于答案空间的后验信息提取图像的相关性区域，以生成更加精确的答案，因此，将基于问题生成的伪图像、基于问题和答案生成的伪图像、真实图像均作为图像输入输入至模型中。在视觉编码器的输出端对带有答案后验信息的图像特征与不包含后验信息的图像特征进行对齐。通过这种方式使模型获取来自于答案的后验信息，从而在测试阶段时能够提取可能与答案相关的区域信息，辅助答案的生成。生成阶段对其训练任务的损失函数如下所示：

$$\mathcal{L}_{\mathrm{CTR}} = \mathcal{L}_{\mathrm{ctr}}(\mathbf{i}, \mathbf{i}^{q}) + \mathcal{L}_{\mathrm{ctr}}(\mathbf{i}, \mathbf{i}^{a}). \tag{5-9}$$

由于在对齐任务的基础上，还包含有生成任务本身的训练，因此生成任务在此处可以起到前文所提到的图文标注任务的约束作用，使对齐任务不会脱离真实图像的语义空间。生成任务的损失采用交叉熵损失，如下所示：

$$\mathcal{L}_{\mathrm{G}} = -\sum_{i=1}^{|\mathbf{a}|} \log p(a_i^{*} | \mathbf{a}_{<i}, \mathbf{i}^{q}, \mathbf{i}^{a}, \mathbf{i}), \tag{5-10}$$

其中 $\mathbf{a}$ 表示所生产的答案文本。

除了在生成阶段引入加强后的图像特征以外，参考 Xiao 等[109] 和 Chen 等[23]，通过对生成的答案部分添加权重的方式，倒逼模型提取图像中与文本更为相关的部分。

首先，分别将无噪声版本的图像与加噪版本的图像和问题一起输入大模型中，得到两个 logit 值，计算得到两个 logit 值的差值，将该差值作为第 i 个 token 在 t 时间步的视觉信息的相关性权重 $w_{i,t}$。

$$w_{i,t} = \Delta logits = logit(a^*_{i,t}|\mathbf{a}_{i,<t}, \mathbf{i}, \mathbf{q}) - logit(a^*_{i,t}|\mathbf{a}_{i,<t}, \mathbf{i}^*, \mathbf{q}), \tag{5-11}$$

根据该相关性权重，对大模型的损失函数进行加权，加权的形式如下所示：

$$\mathcal{L}^{i,t}_G = -\frac{w_{i,t}}{\sum_{k=1}^{l} w_{i,k}} \cdot \log p(a^*_{i,t}|\mathbf{a}_{i,<t}, \mathbf{i}, \mathbf{q}). \tag{5-12}$$

视觉信息相关性权重代表加噪后所导致的 logit 值变化，当该变化值较大时，说明该 token 与视觉信息强相关，因此加大其损失函数权重，当该变化值较小时，权重值也较小，说明该 token 不包含重要的视觉相关信息。如图5-4所示，对该图像对应的答案进行相关性权重赋值后，与问题强相关的部分得到了增强，如“round buildings”这一关键信息。除此以外，检索阶段所生成的伪图像也一并作

图 5-4　相关性权重添加示例

Figure 5-4　Example of adding correlation weight

为图像输入进行噪声训练，由于伪图像中包含加强过后的特征，因此，加入伪图像后能够更进一步使模型对视觉信息进行筛选。最终生成阶段的训练损失由对齐任务和添加相关性权重的生成任务损失组成：

$$\mathcal{L}_G = \mathcal{L}_{\mathrm{G}} + \mathcal{L}_{\mathrm{CTR}}. \tag{5-13}$$

5.4　实验设置与结果

5.4.1　数据集

本章方法主要在基于检索的图文问答数据集 WebQA[103] 与 MultimodelQA[110] 数据集上进行实验。WebQA 数据集是一个较为复杂的问答数据集，其要求模型从外部的文本或视觉信息中检索与问题相关的信息，从而辅助问答任务的进行。本章方法采用的是 WebQA 数据集中的视觉问答数据集部分，每一个实例中包含（图像，标注，问题）三个部分，共有训练集 15K, 测试集 2.5K。MultiModalQA 是一个多模态问答数据集，要求模型结合文本、表格和图像等多种模态的信息来回答问题，旨在评估模型在多模态上下文中的推理和整合能力。其数据形式和 WebQA 类似，共包含训练集 2K，测试集 0.2K。

5.4.2　评估指标

本章方法采用常用的评价指标主要分为两个部分，在检索任务中，主要采用召回率（R@K）指标进行评估，即检索得到的前 K 个答案中是否包含正确答案。在生成任务中，针对 WebQA 数据集，本章方法采用生成的答案与真实答案之间的关键实体信息的重叠来作为准确率指标进行评估；针对 MultimodalQA 数据集，则采用精确匹配（Exact Match, EM）指标进行评估。

5.4.3　实验设置

在检索阶段，本章方法中共包含以下模型：Diffusion 模型，CLIP 编码器，图像标注任务 Transformer 模型。Diffusion 模型采用经过预训练的模型[111]，同时，本章方法采用预训练 CLIP[1] 对文本信息进行编码后输入 Diffusion 模型。去噪步骤的数量设置为 50。生成器的种子设置为 0。无分类器引导的规模为 7.5，一次生成图像的批量为 1。CLIP 模型采用的是 CLIP-ViT-B/32 预训练模型。图像标注任务所采用的 Transformer 模型由 6 层编码器和 6 层解码器组成，多头注意力机制头数量为 4，前馈网络内部隐状态维度为 1024，训练过程中最大训练步数值为 80，取最优检查点作为最终使用的模型。

在生成阶段，主要包含大模型 InternVL，微调过程中，均采用 deepspeed 一阶段。对于 1B 模型，每张 GPU 上的批处理规模为 2，对于 2B 模型则为 1。梯度累计值均为 1，学习率均设置为 4e-5, 最大序列长度为 4096。微调均只进行一步，取最后的检查点作为最终模型。

5.4.4　基线模型

本章方法所对比的基线模型有基于 CLIP 的检索方法，即不经过重排序；基于大模型的重排序方法，即先采用 CLIP 筛选出 Top20 个相关图像，再采用多模态大模型对该 20 个图像进行相关性的重排序。

表 5-1 检索实验结果

Table 5-1 Results of retrieval R@K

	WebQA				MultimodalQA			
模型	R@1	R@2	R@5	R@10	R@1	R@2	R@5	R@10
• ***CLIP-TopK* 检索方法**								
CLIP	42.94	55.20	68.75	77.63	73.91	82.61	93.04	95.22
• **基于大模型的重排序方法**								
LLaVA-1.5-13B	34.99	45.35	65.87	80.56	66.72	72.61	90.87	95.22
Qwen-VL-chat	36.55	47.64	67.22	80.42	61.20	67.83	87.39	93.91
mPLUG-Owl2	32.14	43.26	63.80	79.38	62.41	68.26	89.57	92.61
本研究方法	**52.19**	**63.51**	**75.20**	**80.72**	**87.39**	**95.22**	**98.26**	**98.26**

5.4.5 主实验结果

检索实验 如表5-1所示，实验结果表明，本研究提出的检索方法在 WebQA 和 MultimodalQA 两个多模态数据集上均展现出全面优势。在 WebQA 数据集中，本方法的 R@1 准确率达到 52.19%，较 CLIP-TopK 基线方法提升 21.5%，同时显著超越 LLaVA-1.5-13B、Qwen-VL-chat 等基于大模型的重排序方法 15–20 个百分点。这一差距在更高阶指标（R@5、R@10）中虽有所缩小，但本方法仍以 75.20% 和 80.72% 的准确率保持领先，特别是在 R@10 指标上突破了其他模型长期徘徊在 80% 以下的瓶颈，显示出在跨模态检索场景下优越的性能。与现有方法对比，相较于依赖庞大参数量的 LLaVA-13B 等模型，本方法在 WebQA 的 R@1 上以更轻量的架构实现 52.19% 的准确率，表明其通过精细的语义对齐方法设计而非单纯扩大模型规模来提升性能。这些技术特性使得该方法在保持较高计算效率的同时，仍能应对复杂多模态场景中的语义鸿沟问题。未来研究可进一步探索该方法在更大规模或噪声更强的跨模态数据集中的泛化能力，以及不同模态对齐策略对高精度检索的具体影响。

另外值得注意的是，模型在不同数据集上的表现差异揭示了任务特性的影响。这种差异可能源于 WebQA 任务中更复杂的跨模态关联或更高的噪声干扰，而本方法在更具挑战性的 WebQA 环境中仍能保持 R@1 超过 50%，验证了其鲁棒性。另一方面，基于大模型的重排序方法效果反而弱于不经过重排序的 CLIP 检索方法，这可能是因为未经过微调的多模态大模型在相关性评估领域弱于使用大规模数据进行对比学习训练的 CLIP 模型。

生成实验 表5-2中是生成实验的结果，评估的指标为回答准确率。

在无检索增强的情况下，纯文本的基线模型在检索增强问答数据集上性能

表 5-2　生成实验结果

Table 5-2　Results of generation accuracy

		WebQA			MultimodalQA
模型	重排序	单图	多图	平均	EM
• 无检索增强					
Qwen2-0.5B-Instruct	-	17.29	19.33	18.20	10.43
internlm2-chat-18b	-	23.25	32.58	27.40	10.43
gpt-3.5-turbo-0125	-	40.80	54.49	46.88	25.22
InternVL2-1B	-	26.10	43.57	33.86	19.57
InternVL2-2B	-	30.37	48.20	38.29	25.22
• 有检索增强，无训练微调					
InternVL2-1B	×	59.69	58.21	58.95	55.22
InternVL2-1B	✓	60.41	58.21	59.31	61.30
InternVL2-2B	×	53.01	57.34	55.17	56.09
InternVL2-2B	✓	54.89	56.68	55.79	60.87
• 有检索增强，有训练微调					
InternVL2-1B	×	67.65	77.55	72.60	58.47
InternVL2-1B	✓	68.85	77.84	73.35	61.10
InternVL2-2B	×	68.32	78.52	73.42	62.26
InternVL2-2B	✓	69.91	78.25	74.08	67.59

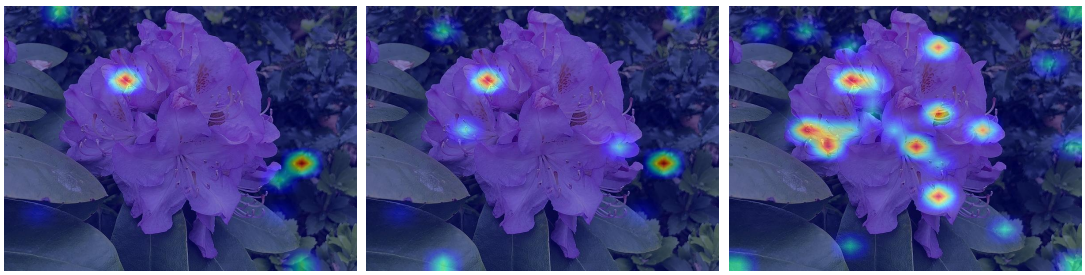
普遍较低。例如，Qwen2-0.5B-Instruct 在 WebQA 数据集上的单图和多图准确率分别为 17.29 和 19.33，平均仅为 18.20；引入检索增强后，模型的表现有显著提升。例如，InternVL2-1B 模型在 WebQA 上的平均准确率从 33.86 提升到 58.95，MultimodalQA 的 EM 得分从 19.57 提升到 55.22。与此同时，重排序方法可以有效提高生成准确率。例如，InternVL2-1B 在启用重排序后，WebQA 的平均准确率仅从 58.95 提升到 59.31，MultimodalQA 的 EM 得分从 55.22 提升到 61.30。

在引入训练微调后，模型性能进一步提升。例如，InternVL2-1B 在 WebQA 上的平均准确率从 58.95（无微调）提升到 72.60（有微调）。而重排序在训练微调后的模型中仍然表现出一定的积极作用。例如，InternVL2-1B 在启用重排序后，WebQA 的平均准确率提升 2 分，MultimodalQA 的 EM 得分提升 2.5 分。由此，进一步验证了重排序方法的有效性。

5.5 分析实验

5.5.1 可视化分析

为了直观地验证模型自主提取图片与问题文本相关性区域的能力，本节对模型进行可视化分析。首先提取大模型中每一层的注意力矩阵，提取其中与问题中关键词语的相关部分，之后将该部分进行可视化，等比例与原图片进行重合，观察模型根据问题关键词关注到图片中的区域变化。如下图5-5所示，从左至右的三张图分别代表未经过微调训练的模型、只经过一般微调训练的模型、经过相关性提取微调训练的模型。可以明显地观察到，经过相关性提取训练的模型对图片中与问题相关的关键区域给予了更多的关注。



(a) Does a Minnetonka Rhododendron flower have petals in a cup shape?

(b) Are any other round buildings visible in the skyline near the Tokyo Sky Tree tower?

图 5-5 注意力可视化分析

Figure 5-5 Visualization of attention maps

5.5.2 消融分析

本节对本章方法在 WebQA 数据集上进行消融分析，以 InternVL2-1B 模型为例，实验结果如表5-3所示：

当模型既无检索增强也无训练微调时，性能显著下降。单图准确率从 69.90

降至 37.10，多图准确率从 76.79 降至 38.40，平均准确率从 74.85 降至 37.75。这表明检索增强和训练微调这两个部分对模型性能的提升至关重要，尤其是在多图任务中，性能下降尤为明显。

当模型仅无训练微调时，性能也有所下降。平均准确率从 74.85 降至 59.31。这说明训练微调对模型性能的提升有显著作用。当模型未进行噪声处理时，平均准确率从 74.85 降至 72.25。这表明噪声处理对模型性能的提升同样具有积极作用作用。当模型无重排序时，模型的表现同样有小幅下降。平均准确率从 74.85 降至 72.70。这说明重排序和噪声处理都是本章方法的有效组成部分。

表 5-3　消融实验结果

Table 5-3　Results of ablation study

	WebQA		
模型	单图	多图	平均
InternVL2-1B	68.85	77.84	73.35
无检索，无训练	37.10	38.40	37.75
无训练	60.41	58.21	59.31
无噪声	68.10	76.39	72.25
无重排序	67.65	77.55	72.60

5.5.3　检索评估

为了对检索的结果进行进一步评估，本节将分析检索实验的精确率、召回率、F1 分数三个指标。由于 MultimodalQA 为单图检索，因此三个指标均相等。如表5-4所示，本研究提出的方法在三个指标上 WebQA 和 MultimodalQA 两个多模态检索任务中均显著优于 CLIP-TopK 基线。在 WebQA 数据集上，本方法的精确率、召回率和 F1 分数分别达到 43.27%、58.85% 和 49.87%，较基线模型（37.12%、50.83%、42.91%）分别提升 6.15、8.02 和 6.96 个百分点。其中，精确率的提升（43.27% vs. 37.12%）表明模型返回的 Top-K 结果中误检比例显著降低，这对需要高置信度检索的应用场景尤为重要。而 F1 值的提升（42.91% → 49.87%）则反映了该方法在平衡查准率与查全率方面的有效性——F1 作为调和平均数，其提升说明模型在高精确率和高召回率之间达成了更优权衡。

5.5.4　实例分析

为了更直观地表现本章方法中加入噪声对模型输出的影响，本节对检索过程中所遇到的实例进行分析。如下表5-5所示，绿色字体代表正确回答，红色字体代表错误回答，仅用 CLIP 排序得到的错误检索所检索到的图像与真实图像十分相似，从而导致了生成答案的误判，影响了答案的准确性。运用本章方法进行

表 5-4　检索指标评估

Table 5-4　Results of retrieval

	WebQA			MultimodalQA		
模型	精确率	召回率	F1	精确率	召回率	F1
CLIP-TopK	37.12	42.94	39.82	73.91	73.91	73.91
本研究方法	**43.27**	**52.19**	**47.32**	**87.39**	**87.39**	**87.39**

重排序之后即可找到真实图像，并给出与参考答案非常接近的精确回答。因此，这进一步说明本章方法可以实现细粒度的精确检索。

表 5-5　在 WebQA 数据集上的实例分析

Table 5-5　Qualitative examples on WebQA test sets

	**成功检索**	
	问题	Where is the clock on the front of the Fenchurch Street station with respect to the rest of the building?
	生成回答	The clock on the front of the Fenchurch Street station is located at the top center of the building, just above the entrance.
	参考回答	The clock on the front of the Fenchurch Street station is on the top and middle of the building.
	**失败检索**	
	问题	Where is the clock on the front of the Fenchurch Street station with respect to the rest of the building?
	生成回答	The clock on the front of the Fenchurch Street station is located on the right side of the building, near the entrance.
	参考回答	The clock on the front of the Fenchurch Street station is on the top and middle of the building.

5.6　本章小结

本研究针对视觉知识问答中模型检索对应图片数据时检索不准确，生成答案时对图片关注不足的难题，提出一种基于基于图片后验检索增强生成框架。该方法分为检索与生成两阶段：检索阶段通过生成问题对应的图像特征与“问题-答案”对应的带有答案后验信息的图像特征，并将二者进行对齐，利用优化后的特征辅助重排序以提升检索精度；生成阶段将伪图像与原始图像共同输入多模态大模型，通过引入相关性权重机制，根据输出的词级 logit 差异动态调整损失权重（差异大则强化图像相关词学习），使模型自主学习到图片与问题文本的相关性区域。实验表明，该方法显著提升检索召回率与生成准确率，尤其在微调后生

成准确率增幅明显，验证了后验特征与相关性权重机制的有效性。

第 6 章　总结和展望

6.1　研究工作总结

随着深度学习技术的发展，依托于预训练大语言生成模型，文本生成技术在多个领域得到了广泛的应用。当文本模态能够提供充足的信息时，模型能够根据所给定的输入输出符合任务目标且符合语法规范的输出，然而，当面对文本模态信息受限的挑战性场景时，由于文本信息不足以支撑任务目标的完成，模型往往会产生幻觉，输出错误的答案。而随着多模态领域的兴起，图片-文本描述数据这一具有高度语义关联性的数据进入了研究者的视野，此类数据数量大且易于获取。因此，利用该数据，将图文对齐技术引入文本生成中，成为了该领域的一个新兴研究课题。如何利用额外的图片信息，对受限的文本信息进行补充，从而提高模型的生成质量，是目前的研究重点。本文从该角度出发，依次由易到难地探索三种文本受限的文本生成任务：（1）低资源机器翻译：利用图片-文本描述数据对齐源端语言，并使用平行语料对齐源端和目标端，从而实现翻译能力从源端高资源语言到低资源语言的迁移、（2）无监督机器翻译：在没有任何平行语料的前提下，仅利用图片-文本描述数据对齐源端与目标端，使模型具备一定的初步翻译能力、（3）视觉知识问答：在没有显式图片-文本描述数据的条件下，通过引入答案后验信息，使模型自主学习图片与问题文本的相关性区域，并提高对该区域的关注，从而提高生成答案的准确性。

本文根据上述研究路线，具体展开以下三点研究工作：

1、图片辅助的低资源机器翻译

针对低资源场景下，神经机器翻译模型缺乏足够平行语料的问题，本文提出了一种图片辅助的低资源机器翻译方法。本文的出发点在于利用图片-文本描述数据与有限的平行语料，将高资源语言的翻译能力迁移至低资源语言上，基于此，首先使用平行语料建立源端到目标端的映射，并在现有平行语料建立的映射基础上，引入额外的图片信息，并将这些额外的图片信息作为源端各个语言之间的语义锚点，对齐源端各个语言之间的表示，从而实现翻译能力迁移的目标。在语义对齐的方法上，本文提出一种包含粗粒度句级别和细粒度词级别对比学习方法的跨模态对齐方法。先将句级别特征与图片的全局特征进行粗粒度的对齐，而后再通过选择性注意力机制，在每一句中进行词级别的细粒度对齐，从而构建源端公共语义空间。实验结果表明，本文方法能够使模型具备一定的零样本翻译效果，并在少样本翻译任务中显著超过纯文本基线模型。进一步的定量以及定性分析实验表明，本文能够成功实现跨模态和跨语言对齐的目标，构造源端语言的公共语义空间，进一步证明了所提出方法的有效性。

2、基于图文对齐的无监督机器翻译

针对无监督机器翻译场景下源端与目标端之间对齐困难的问题，本文提出了一种基于图文对齐的无监督机器翻译方法。在无监督机器翻译的相关研究中，有一种具有代表性的三阶段方法，该方法先通过单语建模为模型赋予单语理解能力，再通过初步对齐赋予模型粗略的翻译能力，最后借助迭代回译建立源端与目标端之间的简介映射。而本文的核心在于结合了图文对齐方法，仅使用图片-文本描述数据在初步对齐阶段直接学习到源端和目标端之间的语义对齐，提高初步对齐的质量，使模型在该阶段获得的较好的翻译能力，提高后续的迭代回译结果质量。语义对齐的方法选择上，本文同样选择效果较好的对比学习方法进行训练。实验结果表明，与纯文本无监督机器翻译模型和先进的多模态无监督机器翻译模型相比，本文所提出的方法在各个指标和语向上均取得了显著的优势。进一步的分析实验表明，本文所提出的方法在源端和目标端之间进行了直接的语义对齐，成功训练得到了一个公共语义空间。本文还在训练集领域外的纯文本翻译数据集上进行了评估，结果表明方法仍然有效，体现了方法的泛化性。

3、基于图片后验检索增强的视觉知识问答

针对视觉知识问答场景中，模型检索对应图片数据时检索不准确，生成答案时对图片关注不足的问题，本文提出基于图片后验检索增强的视觉知识问答方法。在视觉知识问答场景中，模型需要检索并引入外部的图片对模型进行检索增强。现有方法往往直接使用问题文本直接进行检索匹配，但不同于图文匹配任务，由于视觉知识问答场景中问题和图像并不存在显式的描述对应关系，因此此类检索方法往往会检索到错误的图片，对模型的生成起到误导作用。基于此，本文引入包含答案的后验信息，将该信息作为图像和问题文本的连接点，辅助模型学习到图像中和问题更相关的区域，从而使模型能够自主提取图片的相关性区域，提高检索的准确性与生成答案的质量。本文分别针对检索和生成设置了评估实验，实验结果表明，在检索方面，该方法相较于现有方法能实现更精确的检索；与此同时，生成阶段经过微调训练后，生成答案的准确率也明显提高，证明了本文方法的有效性。

6.2 未来工作展望

本文从三个文本信息首先的文本生成任务场景出发，分别针对各个场景下所存在的问题，以图文对齐技术作为基点，提出了相应的解决方案。但实际的应用场景远远不只有本文所提出的三个场景，随着智能感知技术的快速发展，越来越多的多模态场景也随之涌现。与此同时，多模态大模型的高速发展也为诸多小模型所不能解决的问题提供了可能性。本节将对一些未来的研究方向，提出以下几点展望：

动态场景　在本文所涉及到的多模态数据中，仍然以图像信息为主，未出现视频等包含更丰富更复杂信息的多模态数据。而现有静态图像对齐机制难以捕捉视频中的时序语义演变，如动作连贯性和 3D 空间的空间结构特性，如物体遮挡关系，这导致动态多模态对齐面临关键帧提取低效与特征融合粗糙等问题。针对以上问题，可以有以下解决方案：构建时空的混合编码架构，举例来说，时空维度可以采用双通道模型进行编码，时间维度采用时序模型对原始视频帧进行建模，空间维度设计层次化图卷积网络，从局部几何特征到全局形状特征逐层提取空间语义信息，并通过动态门控机制融合时空信息。除此以外，还可以采取渐进性的对齐策略，先通过粗粒度对齐，通过跨模态注意力定位视频关键片段，再进行细粒度对齐，在关键片段区域进行时空特征匹配，实现动词短语与运动模式的细粒度关联。目前针对视频模态的研究也已经成为了一个热点研究方向，而这也是多模态研究中一个值得长期探索的关键方向。

小样本场景　本文所研究的应用场景中包含了低资源场景与无监督场景，但由于数据集方面的限制并没有对真实的低资源数据进行实验。在真实低资源场景下，还会面对更多的问题和挑战。例如，单语数据不充足，无法训练得到高质量单语模型；低资源语言往往具有较多特殊的词汇、语法等，这对于模型对文本特征的理解带来了较大的困难。在未来的研究中，需要针对此类场景进一步提升模型在极端低资源数据条件下的表现，尤其是跨领域的泛化能力，进一步减少模型对标注数据的依赖。而在图像方面，也同样存在该方面的问题，例如小物体识别、低清晰度图像识别等。而对于以上问题，也可以进行联合训练，将二者结合构建跨语言-跨模态原型库，引入组合语义训练，强制模型学习跨模态对应关系。综上所述，有关小样本场景的扩展仍具有较大的研究空间。

多模态大模型场景　多模态大模型已经逐渐成为多模态领域研究中的主流，而多模态大模型中的语义对齐技术正面临从” 粗粒度关联” 向” 细粒度认知” 的转变。随着视觉-语言预训练模型的规模化发展，未来研究需突破以下关键方向：其一，跨模态因果推理的构建，需建立可解释的对齐机制以区分语义关联中的因果性与相关性，而不是仅仅依赖统计。其二，多模态场景和多语言场景的结合，目前多模态大模型的文本方面仍聚焦于英语，缺少有关多语言场景的拓展工作，利用语义对齐技术将英语能力迁移至其他语言也是一个值得探索的方向；其三，更多模态的拓展，目前已经有越来越多的多模态大模型工作聚焦于多种模态的联合输入，而不仅局限于图文双模态。如何在多种模态特征进行语义对齐，实现更合理、精确的语义融合也是一个亟待解决的议题。

参考文献

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C/OL]//Meila M, Zhang T. Proceedings of Machine Learning Research: volume 139 Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. PMLR, 2021: 8748-8763. http://proceedings.mlr.press/v139/radford21a.html.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [C/OL]//9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. https://openreview.net/forum?id=YicbFdNTTy.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C/OL]//Guyon I, von Luxburg U, Bengio S, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017: 5998-6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[4] Chen Z, Wu J, Wang W, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks [EB/OL]. 2024. https://arxiv.org/abs/2312.14238.

[5] Liu Y, Gu J, Goyal N, et al. Multilingual denoising pre-training for neural machine translation [J/OL]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742. https://doi.org/10.1162/tacl_a_00343.

[6] Lin Z, Pan X, Wang M, et al. Pre-training multilingual neural machine translation by leveraging alignment information [C/OL]//Webber B, Cohn T, He Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 2649-2663. https://doi.org/10.18653/v1/2020.emnlp-main.210.

[7] Pan X, Wang M, Wu L, et al. Contrastive learning for many-to-many multilingual neural machine translation [C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 244-258. https://aclanthology.org/2021.acl-long.21. DOI: 10.18653/v1/2021.acl-long.21.

[8] Gu S, Feng Y. Improving zero-shot multilingual translation with universal representations and cross-mapping [C/OL]//Goldberg Y, Kozareva Z, Zhang Y. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 6492-6504. https://aclanthology.org/2022.findings-emnlp.485/. DOI: 10.18653/v1/2022.findings-emnlp.485.

[9] Sohn K. Improved deep metric learning with multi-class n-pair loss objective [C/OL]// Lee D, Sugiyama M, Luxburg U, et al. Advances in Neural Information Processing Sys-

tems: volume 29. Curran Associates, Inc., 2016. https://proceedings.neurips.cc/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

[10] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition [J/OL]. Proc. IEEE, 1998, 86(11): 2278-2324. https://doi.org/10.1109/5.726791.

[11] Li J, Li D, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models [C/OL]//Krause A, Brunskill E, Cho K, et al. Proceedings of Machine Learning Research: volume 202 International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. PMLR, 2023: 19730-19742. https://proceedings.mlr.press/v202/li23q.html.

[12] Huang P, Patrick M, Hu J, et al. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models [C/OL]//Toutanova K, Rumshisky A, Zettlemoyer L, et al. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. Association for Computational Linguistics, 2021: 2443-2459. https://doi.org/10.18653/v1/2021.naacl-main.195.

[13] Xu H, Ghosh G, Huang P, et al. Videoclip: Contrastive pre-training for zero-shot video-text understanding [C/OL]//Moens M, Huang X, Specia L, et al. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021. Association for Computational Linguistics, 2021: 6787-6800. https://doi.org/10.18653/v1/2021.emnlp-main.544.

[14] Nakayama H, Nishida N. Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot [J/OL]. Mach. Transl., 2017, 31(1-2): 49-64. https://doi.org/10.1007/s10590-017-9197-z.

[15] Li Y, Ponti E M, Vulic I, et al. Emergent communication pretraining for few-shot machine translation [C/OL]//Scott D, Bel N, Zong C. Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020. International Committee on Computational Linguistics, 2020: 4716-4731. https://doi.org/10.18653/v1/2020.coling-main.416.

[16] Tan C, Wei J, Sun L, et al. Retrieval meets reasoning: Even high-school textbook knowledge benefits multimodal reasoning [J/OL]. CoRR, 2024, abs/2405.20834. https://doi.org/10.48550/arXiv.2405.20834. DOI: 10.48550/ARXIV.2405.20834.

[17] Lample G, Conneau A, Denoyer L, et al. Unsupervised machine translation using monolingual corpora only [C]//International Conference on Learning Representations (ICLR). 2018.

[18] Lample G, Ott M, Conneau A, et al. Phrase-based & neural unsupervised machine translation [C/OL]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 5039-5049. https://www.aclweb.org/anthology/D18-1549. DOI: 10.18653/v1/D18-1549.

[19] Lample G, Conneau A, Ranzato M, et al. Word translation without parallel data [C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. https://openreview.net/forum?id=H196sainb.

[20] Conneau A, Lample G. Cross-lingual language model pretraining [C/OL]//Wallach H M, Larochelle H, Beygelzimer A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. 2019: 7057-7067. http://papers.nips.cc/paper/8928-cross-lingual-language-model-pretraining.

[21] Song K, Tan X, Qin T, et al. MASS: masked sequence to sequence pre-training for language generation [C/OL]//Chaudhuri K, Salakhutdinov R. Proceedings of Machine Learning Research: volume 97 Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA. PMLR, 2019: 5926-5936. http://proceedings.mlr.press/v97/song19d.html.

[22] Huang P Y, Hu J, Chang X, et al. Unsupervised multimodal neural machine translation with pseudo visual pivoting [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 8226-8237. https://aclanthology.org/2020.acl-main.731. DOI: 10.18653/v1/2020.acl-main.731.

[23] Chen Z, Xu C, Qi Y, et al. MLLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training [J/OL]. CoRR, 2024, abs/2407.21439. https://doi.org/10.48550/arXiv.2407.21439. DOI: 10.48550/ARXIV.2407.21439.

[24] Ma X, Lin S C, Li M, et al. Unifying multimodal retrieval via document screenshot embedding [EB/OL]. 2024. https://arxiv.org/abs/2406.11251.

[25] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition [C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016: 770-778. https://doi.org/10.1109/CVPR.2016.90.

[26] Wang F, Liu H. Understanding the behaviour of contrastive loss [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 2495-2504.

[27] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[28] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners [J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[29] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//International Conference on Machine Learning. 2021: 8748-8763.

[30] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning [J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.

[31] OpenAI. Gpt-4 technical report [J]. arXiv preprint arXiv:2303.08774, 2023.

[32] Liu H, Li C, Li Y, et al. Visual instruction tuning with large language models [J]. arXiv preprint arXiv:2304.08485, 2023.

[33] Li J, Li D, Xiong C, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models [J]. arXiv preprint arXiv:2301.12597, 2023.

[34] Caglayan O, Barrault L, Bougares F. Multimodal attention for neural machine translation [J/OL]. CoRR, 2016, abs/1609.03976. http://arxiv.org/abs/1609.03976.

[35] Huang P Y, Liu F, Shiang S R, et al. Attention-based multimodal neural machine translation [C/OL]//Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Berlin, Germany: Association for Computational Linguistics, 2016: 639-645. https://www.aclweb.org/anthology/W16-2360. DOI: 10.18653/v1/W16-2360.

[36] Calixto I, Elliott D, Frank S. DCU-UvA multimodal MT system report [C/OL]//Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. Berlin, Germany: Association for Computational Linguistics, 2016: 634-638. https://www.aclweb.org/anthology/W16-2359. DOI: 10.18653/v1/W16-2359.

[37] Delbrouck J B, Dupont S. An empirical study on the effectiveness of images in multimodal neural machine translation [C/OL]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 910-919. https://www.aclweb.org/anthology/D17-1095. DOI: 10.18653/v1/D17-1095.

[38] Caglayan O, Aransa W, Bardet A, et al. LIUM-CVC submissions for WMT17 multimodal translation task [C/OL]//Proceedings of the Second Conference on Machine Translation. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 432-439. https://www.aclweb.org/anthology/W17-4746. DOI: 10.18653/v1/W17-4746.

[39] Calixto I, Liu Q. Incorporating global visual features into attention-based neural machine translation. [C/OL]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 992-1003. https://aclanthology.org/D17-1105. DOI: 10.18653/v1/D17-1105.

[40] Delbrouck J, Dupont S. Multimodal compact bilinear pooling for multimodal neural machine translation [J/OL]. CoRR, 2017, abs/1703.08084. http://arxiv.org/abs/1703.08084.

[41] Calixto I, Liu Q, Campbell N. Doubly-attentive decoder for multi-modal neural machine translation [C/OL]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 1913-1924. https://www.aclweb.org/anthology/P17-1175. DOI: 10.18653/v1/P17-1175.

[42] Libovický J, Helcl J. Attention strategies for multi-source sequence-to-sequence learning [C/OL]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 196-202. https://www.aclweb.org/anthology/P17-2031. DOI: 10.18653/v1/P17-2031.

[43] Caglayan O, Bardet A, Bougares F, et al. LIUM-CVC submissions for WMT18 multimodal translation task [C/OL]//Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Belgium, Brussels: Association for Computational Linguistics, 2018: 597-602. https://aclanthology.org/W18-6438. DOI: 10.18653/v1/W18-6438.

[44] Zhou M, Cheng R, Lee Y J, et al. A visual attention grounding neural model for multimodal machine translation [C/OL]//Proceedings of the 2018 Conference on Empirical

Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, 2018: 3643-3653. https://www.aclweb.org/anthology/D18-1400. DOI: 10.18653/v1/D18-1400.

[45] Helcl J, Libovický J, Variš D. CUNI system for the WMT18 multimodal translation task [C/OL]//Proceedings of the Third Conference on Machine Translation: Shared Task Papers. Belgium, Brussels: Association for Computational Linguistics, 2018: 616-623. https://aclanthology.org/W18-6441. DOI: 10.18653/v1/W18-6441.

[46] Ive J, Madhyastha P, Specia L. Distilling translations with visual awareness [C/OL]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 6525-6538. https://www.aclweb.org/anthology/P19-1653. DOI: 10.18653/v1/P19-1653.

[47] Yao S, Wan X. Multimodal transformer for multimodal machine translation [C/OL]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 4346-4350. https://www.aclweb.org/anthology/2020.acl-main.400. DOI: 10.18653/v1/2020.acl-main.400.

[48] Yin Y, Meng F, Su J, et al. A novel graph-based multi-modal fusion encoder for neural machine translation [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 3025-3035. https://www.aclweb.org/anthology/2020.acl-main.273. DOI: 10.18653/v1/2020.acl-main.273.

[49] Liu P, Cao H, Zhao T. Gumbel-attention for multi-modal machine translation [J]. arXiv preprint arXiv:2103.08862, 2021.

[50] Lin H, Meng F, Su J, et al. Dynamic context-guided capsule network for multimodal machine translation [C/OL]//MM '20: Proceedings of the 28th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2020: 1320–1329. https://doi.org/10.1145/3394171.3413715.

[51] Caglayan O, Kuyu M, Amac M S, et al. Cross-lingual visual pre-training for multimodal machine translation [C/OL]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, 2021: 1317-1324. https://www.aclweb.org/anthology/2021.eacl-main.112.

[52] Zhang Z, Chen K, Wang R, et al. Neural machine translation with universal visual representation [C/OL]//International Conference on Learning Representations. 2020. https://openreview.net/forum?id=Byl8hhNYPS.

[53] Fang Q, Feng Y. Neural machine translation with phrase-level universal visual representations [C/OL]//Muresan S, Nakov P, Villavicencio A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 5687-5698. https://aclanthology.org/2022.acl-long.390/. DOI: 10.18653/v1/2022.acl-long.390.

[54] Li B, Lv C, Zhou Z, et al. On vision features in multimodal machine translation [C/OL]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics

(Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 6327-6337. https://aclanthology.org/2022.acl-long.438. DOI: 10.18653/v1/2022.acl-long.438.

[55] Caglayan O, Madhyastha P, Specia L, et al. Probing the need for visual context in multimodal machine translation [C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4159-4170. https://www.aclweb.org/anthology/N19-1422. DOI: 10.18653/v1/N19-1422.

[56] Wu Z, Kong L, Bi W, et al. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation [C/OL]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online: Association for Computational Linguistics, 2021: 6153-6166. https://aclanthology.org/2021.acl-long.480. DOI: 10.18653/v1/2021.acl-long.480.

[57] Firat O, Sankaran B, Al-onaizan Y, et al. Zero-resource translation with multi-lingual neural machine translation [C/OL]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016: 268-277. https://www.aclweb.org/anthology/D16-1026. DOI: 10.18653/v1/D16-1026.

[58] Chen Y, Liu Y, Cheng Y, et al. A teacher-student framework for zero-resource neural machine translation [C/OL]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, 2017: 1925-1935. https://www.aclweb.org/anthology/P17-1176. DOI: 10.18653/v1/P17-1176.

[59] Cheng Y, Yang Q, Liu Y, et al. Joint training for pivot-based neural machine translation [C/OL]//Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017. 2017: 3974-3980. https://doi.org/10.24963/ijcai.2017/555.

[60] Johnson M, Schuster M, Le Q V, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation [J/OL]. Transactions of the Association for Computational Linguistics, 2017, 5: 339-351. https://www.aclweb.org/anthology/Q17-1024. DOI: 10.1162/tacl_a_00065.

[61] Chen Y, Liu Y, Li V O. Zero-resource neural machine translation with multi-agent communication game [C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

[62] Su Y, Fan K, Bach N, et al. Unsupervised multi-modal neural machine translation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 10482-10491.

[63] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks [C]//NeurIPS. 2020: 9459-9474.

[64] Borgeaud S, Mensch A, Hoffmann J, et al. Improving language models by retrieving from

trillions of tokens [C/OL]//Chaudhuri K, Jegelka S, Song L, et al. Proceedings of Machine Learning Research: volume 162 International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. PMLR, 2022: 2206-2240. https://proceedings.mlr.press/v162/borgeaud22a.html.

[65] Karpukhin V, Oguz B, Min S, et al. Dense passage retrieval for open-domain question answering [C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6769-6781. https://aclanthology.org/2020.emnlp-main.550. DOI: 10.18653/v1/2020.emnlp-main.550.

[66] Guu K, Lee K, Tung Z, et al. Realm: Retrieval-augmented language model pre-training [C/OL]//Proceedings of the 37th International Conference on Machine Learning (ICML). 2020: 9459-9474. https://proceedings.mlr.press/v119/guu20a.html.

[67] Li B, Lv C, Zhou Z, et al. On vision features in multimodal machine translation [C/OL]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 6327-6337. https://aclanthology.org/2022.acl-long.438. DOI: 10.18653/v1/2022.acl-long.438.

[68] Huang P, Patrick M, Hu J, et al. Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models [C/OL]//Toutanova K, Rumshisky A, Zettlemoyer L, et al. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021. Association for Computational Linguistics, 2021: 2443-2459. https://doi.org/10.18653/v1/2021.naacl-main.195. DOI: 10.18653/V1/2021.NAACL-MAIN.195.

[69] Lewis M, Liu Y, Goyal N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension [C/OL]//Jurafsky D, Chai J, Schluter N, et al. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. Association for Computational Linguistics, 2020: 7871-7880. https://doi.org/10.18653/v1/2020.acl-main.703. DOI: 10.18653/V1/2020.ACL-MAIN.703.

[70] Lample G, Conneau A, Denoyer L, et al. Unsupervised machine translation using monolingual corpora only [C]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. 2018.

[71] Lin Z, Pan X, Wang M, et al. Pre-training multilingual neural machine translation by leveraging alignment information [C/OL]//Webber B, Cohn T, He Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020. Association for Computational Linguistics, 2020: 2649-2663. https://doi.org/10.18653/v1/2020.emnlp-main.210.

[72] Li W, Gao C, Niu G, et al. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning [C/OL]//Zong C, Xia F, Li W, et al. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long

Papers), Virtual Event, August 1-6, 2021. Association for Computational Linguistics, 2021: 2592-2607. https://doi.org/10.18653/v1/2021.acl-long.202.

[73] Fang Q, Ye R, Li L, et al. STEMM: Self-learning with speech-text manifold mixup for speech translation [C/OL]//Muresan S, Nakov P, Villavicencio A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: Association for Computational Linguistics, 2022: 7050-7062. https://aclanthology.org/2022.acl-long.486/. DOI: 10.18653/v1/2022.acl-long.486.

[74] Lample G, Ott M, Conneau A, et al. Phrase-based & neural unsupervised machine translation [C/OL]//Riloff E, Chiang D, Hockenmaier J, et al. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. Association for Computational Linguistics, 2018: 5039-5049. https://aclanthology.org/D18-1549/.

[75] Lample G, Conneau A, Denoyer L, et al. Unsupervised machine translation using monolingual corpora only [C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. https://openreview.net/forum?id=rkYTTf-AZ.

[76] Ren S, Zhang Z, Liu S, et al. Unsupervised neural machine translation with SMT as posterior regularization [C/OL]//The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 2019: 241-248. https://doi.org/10.1609/aaai.v33i01.3301241.

[77] Sennrich R, Zhang B. Revisiting low-resource neural machine translation: A case study [C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 211-221. https://aclanthology.org/P19-1021. DOI: 10.18653/v1/P19-1021.

[78] Ruiter D, España-Bonet C, van Genabith J. Self-supervised neural machine translation [C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, 2019: 1828-1834. https://aclanthology.org/P19-1178. DOI: 10.18653/v1/P19-1178.

[79] Aharoni R, Johnson M, Firat O. Massively multilingual neural machine translation [C/OL]// Burstein J, Doran C, Solorio T. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics, 2019: 3874-3884. https://doi.org/10.18653/v1/n19-1388.

[80] Ye R, Wang M, Li L. Cross-modal contrastive learning for speech translation [C/OL]// Carpuat M, de Marneffe M C, Meza Ruiz I V. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics,

2022: 5099-5113. https://aclanthology.org/2022.naacl-main.376/. DOI: 10.18653/v1/2022.naacl-main.376.

[81] Elliott D, Frank S, Sima'an K, et al. Multi30k: Multilingual english-german image descriptions [C/OL]//Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany. The Association for Computer Linguistics, 2016. https://doi.org/10.18653/v1/w16-3210.

[82] Lin T, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context [C/OL]// Fleet D J, Pajdla T, Schiele B, et al. Lecture Notes in Computer Science: volume 8693 Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V. Springer, 2014: 740-755. https://doi.org/10.1007/978-3-319-10602-1_48.

[83] Gurari D, Zhao Y, Zhang M, et al. Captioning images taken by people who are blind [C/OL]//Vedaldi A, Bischof H, Brox T, et al. Lecture Notes in Computer Science: volume 12362 Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVII. Springer, 2020: 417-434. https://doi.org/10.1007/978-3-030-58520-4_25.

[84] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units [C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 1715-1725. https://aclanthology.org/P16-1162. DOI: 10.18653/v1/P16-1162.

[85] Kingma D P, Ba J. Adam: A method for stochastic optimization [C/OL]//ICLR (Poster). 2015. http://arxiv.org/abs/1412.6980.

[86] Ott M, Edunov S, Baevski A, et al. fairseq: A fast, extensible toolkit for sequence modeling [C/OL]//Ammar W, Louis A, Mostafazadeh N. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations. Association for Computational Linguistics, 2019: 48-53. https://doi.org/10.18653/v1/n19-4009.

[87] Papineni K, Roukos S, Ward T, et al. Bleu: a method for automatic evaluation of machine translation [C/OL]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA. ACL, 2002: 311-318. https://aclanthology.org/P02-1040/. DOI: 10.3115/1073083.1073135.

[88] Post M. A call for clarity in reporting BLEU scores [C/OL]//Bojar O, Chatterjee R, Federmann C, et al. Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018. Association for Computational Linguistics, 2018: 186-191. https://doi.org/10.18653/v1/w18-6319.

[89] Post M. A call for clarity in reporting BLEU scores [C/OL]//Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics, 2018: 186-191. https://www.aclweb.org/anthology/W18-6319. DOI: 10.18653/v1/W18-6319.

[90] Laurens V D M, Hinton G. Visualizing data using t-sne [J]. Journal of Machine Learning Research, 2008, 9(2605): 2579-2605.

[91] Fei H, Liu Q, Zhang M, et al. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination [C/OL]//Rogers A, Boyd-Graber J, Okazaki N. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 5980-5994. https://aclanthology.org/2023.acl-long.329/. DOI: 10.18653/v1/2023.acl-long.329.

[92] Klementiev A, Irvine A, Callison-Burch C, et al. Toward statistical machine translation without parallel corpora [C/OL]//Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics, 2012: 130-140. https://aclanthology.org/E12-1014.

[93] Conneau A, Lample G, Ranzato M, et al. Word translation without parallel data [J/OL]. CoRR, 2017, abs/1710.04087. http://arxiv.org/abs/1710.04087.

[94] Yang Z, Fang Q, Feng Y. Low-resource neural machine translation with cross-modal alignment [C/OL]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 10134-10146. https://aclanthology.org/2022.emnlp-main.689.

[95] van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding [EB/OL]. 2019. https://arxiv.org/abs/1807.03748.

[96] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C/OL]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: Association for Computational Linguistics, 2005: 65-72. https://aclanthology.org/W05-0909.

[97] Dubossarsky H, Vulić I, Reichart R, et al. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures [C/OL]//Webber B, Cohn T, He Y, et al. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020: 2377-2390. https://aclanthology.org/2020.emnlp-main.186. DOI: 10.18653/v1/2020.emnlp-main.186.

[98] Marchisio K, Duh K, Koehn P. When does unsupervised machine translation work? [C/OL]// Barrault L, Bojar O, Bougares F, et al. Proceedings of the Fifth Conference on Machine Translation. Online: Association for Computational Linguistics, 2020: 571-583. https://aclanthology.org/2020.wmt-1.68.

[99] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J/OL]. arXiv preprint arXiv:1810.04805, 2019. https://arxiv.org/abs/1810.04805.

[100] Chen W, Hu H, Chen X, et al. Murag: Multimodal retrieval-augmented generator for open question answering over images and text [C/OL]//Goldberg Y, Kozareva Z, Zhang Y. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. Associ-

ation for Computational Linguistics, 2022: 5558-5570. https://doi.org/10.18653/v1/2022.emnlp-main.375. DOI: 10.18653/V1/2022.EMNLP-MAIN.375.

[101] Marino K, Rastegari M, Farhadi A, et al. OK-VQA: A visual question answering benchmark requiring external knowledge [C/OL]//IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, 2019: 3195-3204. http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html. DOI: 10.1109/CVPR.2019.00331.

[102] Schwenk D, Khandelwal A, Clark C, et al. A-OKVQA: A benchmark for visual question answering using world knowledge [C/OL]//Avidan S, Brostow G J, Cissé M, et al. Lecture Notes in Computer Science: volume 13668 Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII. Springer, 2022: 146-162. https://doi.org/10.1007/978-3-031-20074-8_9.

[103] Chang Y, Cao G, Narang M, et al. Webqa: Multihop and multimodal QA [C/OL]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022. IEEE, 2022: 16474-16483. https://doi.org/10.1109/CVPR52688.2022.01600.

[104] Penamakuri A S, Gupta M, Gupta M D, et al. Answer mining from a pool of images: Towards retrieval-based visual question answering [C/OL]//Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China. ijcai.org, 2023: 1312-1321. https://doi.org/10.24963/ijcai.2023/146. DOI: 10.24963/IJCAI.2023/146.

[105] Yu B, Fu C, Yu H, et al. Unified language representation for question answering over text, tables, and images [C/OL]//Rogers A, Boyd-Graber J, Okazaki N. Findings of the Association for Computational Linguistics: ACL 2023. Toronto, Canada: Association for Computational Linguistics, 2023: 4756-4765. https://aclanthology.org/2023.findings-acl.292/. DOI: 10.18653/v1/2023.findings-acl.292.

[106] Gu N, Fu P, Liu X, et al. Light-PEFT: Lightening parameter-efficient fine-tuning via early pruning [C/OL]//Ku L W, Martins A, Srikumar V. Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics, 2024: 7528-7541. https://aclanthology.org/2024.findings-acl.447/. DOI: 10.18653/v1/2024.findings-acl.447.

[107] Li P, Si Q, Fu P, et al. Multimodal hypothetical summary for retrieval-based multi-image question answering [J/OL]. CoRR, 2024, abs/2412.14880. https://doi.org/10.48550/arXiv.2412.14880. DOI: 10.48550/ARXIV.2412.14880.

[108] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models [J/OL]. arXiv preprint arXiv:2006.11239, 2020. https://arxiv.org/abs/2006.11239.

[109] Xiao X, Wu B, Wang J, et al. Seeing the image: Prioritizing visual correlation by contrastive alignment [C/OL]//Globersons A, Mackey L, Belgrave D, et al. Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada,

December 10 - 15, 2024. 2024. http://papers.nips.cc/paper_files/paper/2024/hash/37294f033582ac0064bf90fa557c2573-Abstract-Conference.html.

[110] Talmor A, Yoran O, Catav A, et al. Multimodalqa: complex question answering over text, tables and images [C/OL]//9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021. https://openreview.net/forum?id=ee6W5UgQLa.

[111] von Platen P, Patil S, Lozhkov A, et al. Diffusers: State-of-the-art diffusion models [EB/OL]. 2022. https://github.com/huggingface/diffusers.

[112] Ott M, Edunov S, Grangier D, et al. Scaling neural machine translation [C/OL]//Bojar O, Chatterjee R, Federmann C, et al. Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018. Association for Computational Linguistics, 2018: 1-9. https://doi.org/10.18653/v1/w18-6301.

[113] Luong M T, Manning C D. Achieving open vocabulary neural machine translation with hybrid word-character models [C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016: 1054-1063. https://aclanthology.org/P16-1100. DOI: 10.18653/v1/P16-1100.

[114] Roy O, Vetterli M. The effective rank: A measure of effective dimensionality [C/OL]//15th European Signal Processing Conference, EUSIPCO 2007, Poznan, Poland, September 3-7, 2007. IEEE, 2007: 606-610. https://ieeexplore.ieee.org/document/7098875/.

[115] N.J. Higham P G P M F S, M.R. Dennis, Tanner J. The princeton companion to applied mathematics [M]. 2016.

附录 一　用于伪数据的翻译模型

本节附录中主要介绍第 3 章方法中用于构建伪数据的翻译模型的详细信息。对于英语 → 德语和英语 → 法语语向，实验中使用 Ott 等[112][1] 提供的预训练模型，该模型包含 6 层编码器和解码器。注意力头的数量设置为 16。英语 → 德语的 dropout 设置为 0.3，英语 → 法语的 dropout 设置为 0.1。标签平滑设置为 0.1。

对于英语 → 捷克语语向，我们在 WMT2015 英语 → 捷克语训练集上训练了一个 Transformer-base 模型，该训练集包含约 15M 平行数据。模型包含 6 层编码器和解码器。注意力头的数量设置为 8。dropout 和标签平滑均设置为 0.1。

实验在 WMT 测试集 newstest2014 上评估英语 → 德语和英语 → 法语模型，并在 newstest2015 上评估英语 → 捷克语模型。评估结果如表 附录一-1 所示，实验中所采用的预训练翻译模型均具备可靠的性能。

表 附录一-1　构造伪数据所用的翻译模型 BLEU 值

Table 附录一-1　BLEU scores of translation models for constructing the pseudo data.

语向	模型	BLEU
英语 → 德语	Vaswani 等[3]	28.4
	Ott 等[112]	29.3
英语 → 法语	Vaswani 等[3]	41.0
	Ott 等[112]	43.2
英语 → 捷克语	Luong 等[113]	20.7
	Vaswani 等[3]	25.2

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/translation

附录 二 奇异值差异和有效条件数

在本节附录中，将详细解释奇异值间隙和有效条件数的计算方法。给定一个单语数据集 $\mathbf{x} = (x_1, ..., x_n)$，它可以通过编码器转换为一个 (n, d) 维的嵌入矩阵 $\mathbf{X}$。利用奇异值分解（SVD），可以得到一个对角矩阵 Σ，其中主对角线由 d 个奇异值组成，$\sigma_1, \sigma_2, ..., \sigma_d$，按降序排列。本节附录将使用从该分解中获得的奇异值对表示空间进行定量分析。

奇异值差异 奇异值差异（SVG）衡量的是从两个矩阵 $\mathbf{X}_1$ 和 $\mathbf{X}_2$ 的分解中获得的奇异值之间的差异值。该度量使用平方欧几里得距离来量化在对数变换后的对应奇异值之间的差异。

$$SVG(\mathbf{X}_1, \mathbf{X}_2) = \sum_{i=1}^{d} (\log \sigma_i^1 - \log \sigma_i^2)^2 \tag{附 1-1}$$

其中，σ_i^1 和 σ_i^2 分别是对应于两个嵌入矩阵 $\mathbf{X}_1$ 和 $\mathbf{X}_2$ 的排序后的奇异值。当两个嵌入变得更加接近时，奇异值之间的差异减小；反之，当它们发生偏离时，SVG 值增加。

有效条件数 有效条件数衡量输入 $\mathbf{X}$ 中的微小扰动在输出中被放大的程度。首先，需要引入有效排序的概念。由于较小的奇异值与噪声相关联，我们应当找到最后一个有效奇异值。Roy 等[114] 提出了一种在计算输入矩阵 $\mathbf{X}$ 的所谓有效秩之前考虑奇异值全频谱的方法：

$$erank(\mathbf{X}) = \lfloor e^{H(\Sigma)} \rfloor \tag{附 1-2}$$

其中，$H(\Sigma)$ 是矩阵 $\mathbf{X}$ 的归一化奇异值分布的熵，$\bar{\sigma} i = \frac{\sigma_i}{\sum_{i=1}^{d} \sigma_i}$，其计算公式为 $H(\Sigma) = -\sum_{i=1}^{d} \bar{\sigma}_i \log \bar{\sigma}_i$。在获得有效秩后，参考 N.J. Higham 等[115]，有效条件数定义为其第一个（最大）奇异值与最后一个有效奇异值（有效秩）的比值。

$$\kappa_{ecn}(\mathbf{X}) = \frac{\sigma_1}{\sigma_{erank(\mathbf{X})}} \tag{附 1-3}$$

有效条件数越小，表示在给定输入扰动下输出中扰动的放大程度越小，这表明特征表示的稳定性越高。

致 谢

时光荏苒，转眼间我的研究生生涯即将画上句号。回首这段求学之路，充满了挑战与收获，也离不开许多人的帮助与支持。在此，我谨向所有关心、帮助过我的人致以最诚挚的谢意。

首先，我要衷心感谢我的导师冯洋老师。冯老师渊博的学识、严谨的治学态度、敏锐的学术洞察力以及平易近人的为人处世风格，都深深影响着我。在整个研究生生涯的科研过程中，冯老师都给予了我悉心的指导和无私的帮助。更重要的是，冯老师不仅教会了我如何做研究，更锻炼了我除了学术与科研之外的能力，如口头汇报、计划安排等，同时也教了我许多如何为人处事的态度，而这也将使我受益终身。还记得在科研和论文写作陷入瓶颈时，是冯老师耐心地与我讨论，帮我理清思路，指明方向；在实验遇到困难时，是冯老师鼓励我不要轻言放弃，并给予我宝贵的建议。冯老师的谆谆教诲和殷切期望，我将永远铭记于心。

其次，我要感谢实验室的秘书程一老师和裴晓雪老师。程老师和裴老师工作认真负责，待人热情周到，为实验室的日常运转和同学们的科研学习提供了极大的便利。无论是科研相关事务，还是日常事务的办理，程老师和裴老师总是尽心尽力，为我们创造了良好的学习和科研环境。

感谢实验室的每一位帮助过我的同学们，感谢李秀星、谷舒豪、邵晨泽、李泽康、李绩成、郭登级、张绍磊、张倬诚、伍烜 甫、田畅、房庆凯、马铮睿、黄浪林、刘龙祥、桂尚彤师兄和刘舒曼、欧蛟、郭雯钰、赵彤钰师姐，你们是我初入实验室的榜样，也是我这一路尚前行的指路明灯。是房庆凯师兄帮助我在初入实验室时快速入门多模态机器翻译，并在遇到问题时不厌其烦地与我进行讨论，在科研受挫时积极思考解决方案，最后帮助我发表了研究生生涯的第一篇学术论文。刘龙祥师兄也在工作方面给予了非常多的建议，让我在找工作的阶段能够有更明确的目标，避开一些弯路和陷阱。感谢与我同级的郭守涛、张珂豪、鄢子文，是你们陪伴我度过了这段难忘的时光。我们一起讨论学术问题，一起分享生活趣事，互相鼓励，共同进步。你们的陪伴和支持，是我前进的动力。感谢卜梦煜、周奘 、雨田、温卓凡、乔康裕、高博飞、崔璐毅、洪运、石响师弟，你们是实验室未来的顶梁柱，也祝你们未来在各自的科研领域能够获得更多的成就。

我还要感谢我的父母。尽管北京和温州相隔千里，但你们天气预报的界面上永远有一页属于北京，我总是能在天气转凉时接到你们“多穿衣服”、“多喝热水”的叮嘱。你们无私的爱和默默的支持，是我最坚强的后盾。无论我遇到什么困难，你们总是鼓励我勇敢面对，并给予我最大的理解和支持。你们的爱，是我一路上前进的最大动力。

最后，特别感谢方泽铭同学，6 年的陪伴，你是我迷茫时、难过时、孤独时的依靠，感谢你陪伴我的每一分每一秒，感谢这一路上你对我的付出。

路漫漫其修远兮，吾将上下而求索。研究生生涯的结束并不是人生的终点，未来的路还很长，还有更多更大的挑战在等待着我，我将带着你们的期望和祝福，继续努力，不断前行！

2025 年 6 月

作者简历及攻读学位期间发表的学术论文与其他相关学术成果

作者简历:

2018 年 09 月——2022 年 06 月，在浙江大学电气工程学院院获得学士学位。

2022 年 09 月——2025 年 06 月，在中国科学院大学计算技术研究所攻读硕士学位。

已发表（或正式接受）的学术论文:

(1) **Zhe Yang**, Qingkai Fang, Yang Feng. Low-resource Neural Machine Translation with Cross-modal Alignment. The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022), December 7-11, 2022, 10134–10146, online & offline.

(2) Shaolei Zhang, Qingkai Fang, **Zhe Yang**, Yang Feng. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. The 13th International Conference on Learning Representations (ICLR 2025), April 24-28, 2025, Singapore.

参加的研究项目及获奖情况:

(1) 国家自然科学基金面上项目，非自回归神经机器翻译关键技术研究，项目批准号 62376260，2024/01-2027/12

(2) 国家重点研发计划科技创新 2030- “新一代人工智能” 重大项目课题，人机行为与情境常识的大规模知识处理与推理，课题号 2018AAA0102502，2019/12-2023/12

(3) 2023-2024 中国科学院大学三好学生