连接词汇歧义与视觉：视觉词义消歧迷你综述

Shashini Nilukshi
计算机学院 信息技术学院
科伦坡06区，斯里兰卡
shashini.20221258@iit.ac.lk

Deshan Sumanathilaka
计算机学院 信息技术学院
科伦坡06区，斯里兰卡
deshan.s@iit.ac.lk

arXiv:2602.01193v1 [cs.CL] 1 Feb 2026

摘要

本文对视觉词义消歧（VWSD）进行了简要综述。视觉词义消歧是传统词义消歧（WSD）的多模态扩展，有助于解决视觉语言任务中的词汇歧义。传统词义消歧仅依赖文本和词汇资源，而视觉词义消歧则利用视觉线索，在极少文本输入的情况下确定歧义词的正确含义。本综述考察了从早期多模态融合方法到采用CLIP等对比模型、基于扩散的文本到图像生成以及大语言模型（LLM）支持的新框架的发展。研究梳理了2016至2025年的相关工作，以展示视觉词义消歧通过基于特征的、基于图的以及对比嵌入技术所取得的进展，并重点关注提示工程、微调以及面向多种语言的适配。定量结果表明，基于CLIP的微调模型和经大语言模型增强的视觉词义消歧系统始终优于零样本基线，在平均倒数排名（MRR）上最高可提升6–8%。然而，挑战依然存在，例如上下文方面的局限性、模型偏向常见含义、缺乏多语言数据集，以及需要更好的评估框架。分析指出，CLIP对齐、扩散生成与大语言模型推理的日益融合，是构建强大、具备上下文感知能力且支持多语言的消歧系统的未来路径。

1 引言

自然语言处理（NLP）领域取得了显著进展，尤其是深度学习和基于Transformer的架构的引入。这些模型通过大量数据训练，能够识别语言中的复杂模式，大幅提升了我们应对语言歧义细微之处和复杂性的能力（Yadav等人，2021）。语言歧义仍然是一个持续存在的NLP领域面临的挑战，因为人类语言会根据特定词语或短语周围的上下文表达不同含义（Sumanathilaka 等人，2024年b）。

歧义长期以来一直是自然语言处理领域的一项重大挑战。歧义可能出现在多个层面，主要包括词汇、句法和语义层面，每个层面都会导致语言理解中不同类型的不确定性（Elsharif等人，2025年；Sumanathilaka 等人，2025年）。这种固有的歧义使许多自然语言处理任务从根本上变成了消歧问题。尽管自然语言处理模型不断进步，但由于上下文不足、不同歧义类别相互重叠等因素，克服这些歧义仍然十分困难，因此语言歧义持续成为自然语言处理研究与应用关注的重点（Ortega-Martín等人，2023）。

1.1 词汇歧义与词义消歧

当一个词具有多种潜在含义时，就会产生词汇歧义。例如，在句子“We finally reached the bank”中，bank一词可能指“金融机构”，也可能指“河岸”（Li等人，2024）。短语存在多种解读方式时，也会产生词汇歧义。例如，“I saw the man with a telescope”这句话可能表示那名男子拥有一架望远镜，也可能表示观察者使用望远镜看到了他。这说明，句子中词语的排列方式可能造成多种解读，这与单个词语内部存在的歧义不同（Abeysiriwardanaand Sumanathilaka，2024）。

传统词义消歧（WSD）旨在根据词语的上下文，从多种可能的解释中确定其恰当含义。相关方法通常分为三大类：基于知识的方法、监督方法和无监督方法。基于知识的

方法利用词典和语义网络等词汇资源来推导词义，无需标注数据（Bevilacqua 等人，2021）。监督式技术将词义消歧视为分类问题，利用标注语料库训练依赖上下文特征的模型。相较之下，无监督方法会考察大型文本语料库中的分布模式，根据词义对词语用法进行分类，而无需使用标注信息。每种方法都有各自的优缺点，但传统词义消歧仍然是解决自然语言处理中语言歧义问题的重要组成部分（Sumanathilaka 等人，2024a）。

1.2 视觉词义消歧

视觉词义消歧（VWSD）以 SemEval-2023 任务 1（Raganato 等人，2023a）为例，旨在通过选择与歧义词预期含义相对应的正确图像，解决图像生成中的词汇歧义，而文本上下文往往十分有限。与依赖 WordNet 等文本义项清单的传统词义消歧不同，视觉词义消歧要求将目标词（例如“passenger coach”中的“coach”）与多个选项中的一张正确图像对应起来，这需要精准的多模态理解（Raganato 等人，2023a）。由于上下文有限、干扰图像来自相似领域，且必须实现精准的多模态理解，这项任务颇具挑战性（Grbowiec，2023）。视觉词义消歧利用文本到图像模型（CLIP、DALL-E 2、Stable Diffusion）的进展，将文本语义与视觉语义联系起来。通过在多模态框架中消解词汇歧义，视觉词义消歧提升了生成图像的语义精确度（例如，区分“bat”指的是动物还是器具）。这一进展对图像检索、目标检测以及多语言场景（英语、意大利语、波斯语）等应用至关重要（Katyal 等人，2023）。这项任务依托一个由 Wikidata、Omega Wiki 和 BabelPic 汇编而成的数据集，并使用平均倒数排名（MRR）和首位命中率（HIT@1）等指标进行评估（Raganato 等人，2023a）。

本研究的主要贡献如下：

- 全面比较了视觉词义消歧（VWSD）方法，涵盖从早期特征／图融合到近期对比式 CLIP 以及大语言模型增强框架，并直接给出性能指标（HIT@1、平均倒数排名）。
- 提供了对近期进展（2016–2025年）最早的系统性综述之一，内容涵盖提示工程、大语言模型上下文扩展、微调、多语言适应以及扩散模型，并使用 SemEval-2023 任务 1 等基准测试。
- 探讨视觉词义消歧面临的长期挑战，并提出未来研究的有前景方向。

本文首先对文献进行系统综述，涵盖经典方法和基于大语言模型的现代视觉词义消歧方法，随后讨论面临的挑战和未来方向。本研究涵盖所有视觉词义消歧研究，包括对视觉动词词义消歧（VVSD）的全面考察；作为开创多模态消歧技术的奠基性先驱，VVSD 为后续研究奠定了基础。

2 方法论

本次文献综述采用系统方法，旨在识别并分析有关视觉词义消歧、词义消歧和多模态歧义消解的研究。检索范围涵盖 2016 年至 2025 年发表的同行评审论文，检索来源包括 ACL Anthology、arXiv、IEEE Xplore、SpringerLink、Semantic Scholar 和 Google Scholar 等主要平台，以及 SemEval、WACV 和 CVPR 的会议论文集。检索关键词结合了核心术语和相关术语，例如“视觉词义消歧”“词汇歧义消解”“多模态嵌入”“CLIP 对比式语言-图像预训练”“视觉动词词义消歧”和“基于注意力的融合”。本综述聚焦于更广泛的视觉词义消歧任务，同时深入分析了 2016 至 2022 年的视觉动词词义消歧研究，将其视为一项重要的奠基性先驱工作。

筛选过程采用了前向和后向引文追踪，以确保全面涵盖奠基性研究和新兴研究。纳入标准要求研究提供实证结果、基准测试或模型创新，采用 CLIP、卷积神经网络、基于图的方法或大语言模型增强架构，并以 SemEval-2023 任务1 或类似的多模态数据集作为基准。若研究缺乏足够的实证深度、超出自然语言处理/多模态领域，或属于重复发表，则予以排除。这一严谨的流程确保最终收录的文献具有代表性且紧跟前沿。论文的筛选流程如图 1 所示。

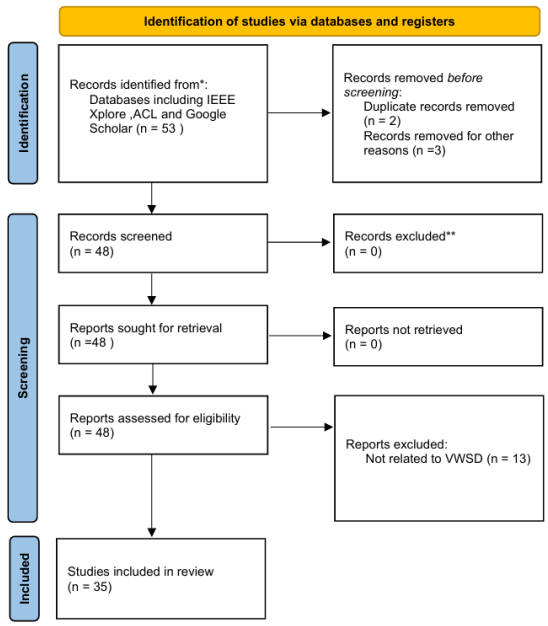

图 1：论文筛选流程的 PRISMA 图

2.1 纳入与排除标准

研究对象/问题（P）：我们纳入了聚焦于视觉词义消歧（VWSD）、视觉动词词义消歧（VVSD）或多模态语境中词汇歧义消解的研究。未专注于词义消歧的研究以及纯理论研究均予以排除。

干预（I）：本综述考察了采用创新模型的实证研究，例如 CLIP、卷积神经网络、基于图的技术或经大语言模型增强的架构，尤其是以 SemEval-2023 任务1 或类似数据集为基准进行评估的研究。我们排除了缺乏实证内容的论文、仅提出理论模型而未进行实际实施的论文，以及非多模态或仅采用基于文本的词义消歧方法的研究。**上下文（C）：**我们纳入了自然语言处理或多模态人工智能领域中，发表于知名学术平台（如 ACL Anthology、CVPR、SemEval）的同行评审研究论文和基准数据集。排除范围包括这些领域之外的研究、重复发表的论文，以及来自未经同行评审或认可度较低的出版平台的论文。

时间范围（T）：如图2所示，本综述纳入了2016年至2025年发表的文献，重点关注多模态融合、对比学习、大语言模型集成和多语言策略方面的进展。2016年以前发表的研究不在本文范围内。

3 现有研究

3.1 早期多模型方法

视觉动词词义消歧（VVSD）是一项独立的专门任务，也是视觉词义消歧（VWSD）的直接前身。它率先利用视觉上下文来厘清动词含义，旨在确定图像所示、与歧义动词相对应的具体动作（Gella 等人，2019）。例如，VVSD 可以区分图像中“play”的不同含义（如演奏乐器、进行体育运动或参与活动），从而提升多模态应用的精确度（Gella 等人，2016）。尽管 VVSD 最初仅关注动词多义现象，但为其开发的成功多模态融合方法促使研究人员将这一方法扩展到所有词类（名词、动词、形容词），进而催生了涵盖范围更广的 VWSD 挑战。早期的 VVSD 研究率先探索了多模态融合，将传统词义消歧与计算机视觉表征相结合，并通常将相关方法归为基于特征或基于图的方法（Gella 等人，2016）。

3.1.1 基于特征的方法

视觉特征：基于卷积神经网络的图像表示

视觉动词词义消歧（VVSD）的一项开创性技术是使用卷积神经网络（卷积神经网络），例如 VGGNet 和 ResNet，来表示图像。这些卷积神经网络提取分层视觉特征，以编码不同层次的抽象。例如，Gella 等人(2016) 使用在 ImageNet 上预训练的模型提取卷积神经网络特征，为与歧义词对应的候选图像生成视觉嵌入。这些视觉向量可作为文本之外的补充信息。

VGG（视觉几何组）网络可从输入图像中提取固定维度的图像特征向量，捕捉空间和语义线索（Simonyan 和 Zisserman，2014年）。ResNet（残差网络）通过残差连接支持更深层架构的训练，从而增强了特征提取能力，生成更强大的图像嵌入（He 等人，2015）。使用这些基于 CNN 的图像特征，使 VVSD 系统能够捕捉细粒度视觉差异；当文本上下文稀缺或存在歧义时，这些差异至关重要。

文本特征：词嵌入和句子表示

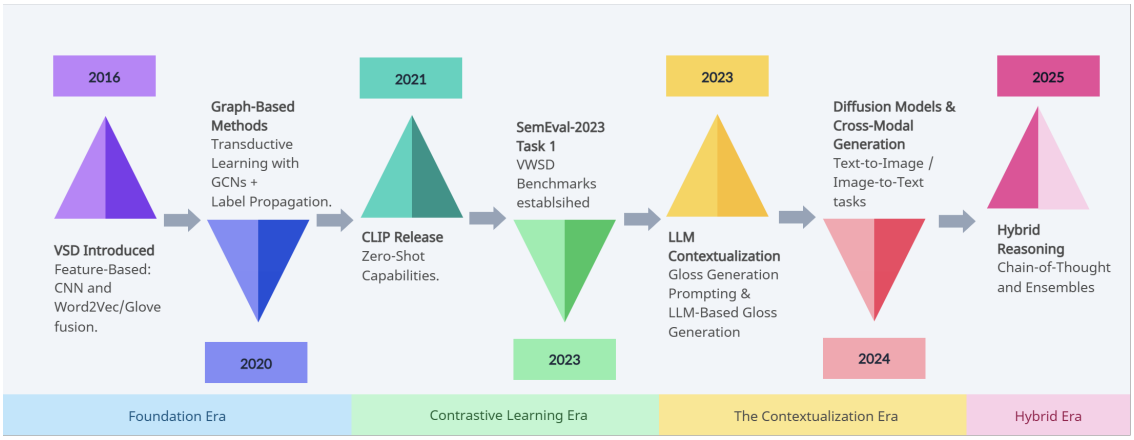

图2：时间线分析

在文本方面，早期的 VVSD 研究使用预训练词嵌入（如 Word2Vec（Mikolov 等人，2013）和 GloVe（Pennington 等人，2014年）来表示歧义词及其最小上下文。这些嵌入基于大型语料库编码语义相似度，为词语提供稠密的向量表示。除简单的词嵌入外，通过对词向量取平均，或使用循环神经网络和长短期记忆网络生成的句子表示，也有助于捕捉上下文。然而，VVSD 中极其有限的上下文（歧义词周围通常只有一两个词）给仅通过文本进行消歧带来了挑战。

多模态表征的整合

• 拼接：最简单的融合技术是在应用分类器或相似度度量之前，将文本与视觉特征向量拼接为联合表示（Gella 等人，2016）。

• 基于注意力的融合：注意力机制根据上下文动态权衡视觉特征与文本特征，使模型能够聚焦于用于消歧的显著特征（Kiela 等人，2019）。

• 典型相关分析（CCA）：CCA被用于学习投影，以最大化视觉模态与文本模态之间的相关性，并将其映射到共享潜在空间，从而改善跨模态对齐（Hardoon等人, 2004; Rasiwasia等人, 2010年）。该方法通过确保融合空间中的相似性，有效地将歧义词与相应图像进行匹配。

3.1.2 基于图的方法

图构建：由多模态特征构建相似度图

基于图的方法将候选图像和词义表示为图中的节点，其中边捕捉由文本特征与视觉特征导出的相似性度量（Vascon等人,2020）。该方法率先针对视觉动词词义消歧进行图构建，通过编码多模态相似性，使词义信息能够经由图的连通性进行传播。该图捕捉如下关系：

1. 基于CNN视觉特征的候选图像之间的相似度。2. 词上下文嵌入之间的语义相似度。3. 结合两种技术的多模态相似度。

这种结构化表示有助于在消歧过程中利用邻域信息和全局图属性（Vascon等人，2020）。

标签传播：传播消歧决策

一旦构建好多模态图，标签传播算法便会通过图结构将已知标签（消歧决策）扩散到未标注节点（没有直接标注的候选词义）。这种半监督学习能够利用大量未标注数据并提升泛化能力。相关技术包括：

• 基于随机游走的传播。
•通过调和函数满足节点标签的平滑性约束。

• 将图卷积网络（GCN）应用于多模态图，以实现端到端学习。（KipfandWelling，2016）。

转导学习方法同时利用有标签和未标注数据，优化在固定测试集上的性能。与归纳式方法不同，这类方法无需在给定数据集之外进行泛化，因此适用于训练数据有限的视觉动词词义消歧任务。通过结合特征提取、图结构和标签传播，转导学习方法在视觉动词词义消歧准确率方面取得了显著提升，尤其是在动词和低频词义上。Vascon等人（2020）使用基于转导学习的图方法，在 VerSe 数据集上取得了最先进的结果。

3.2 基于 CLIP 的方法

CLIP（对比语言-图像预训练）是推动视觉词义消歧发展的变革性模型，它利用海量图文对进行联合训练，创建一个将自然语言与视觉对齐的共享嵌入空间。

3.2.1 零样本方法

CLIP 在共享潜在空间中计算文本与图像嵌入之间的余弦相似度，从而无需针对特定任务进行训练即可直接实现零样本消歧（Radford 等人，2021）。这使系统能够选出在语义上与歧义词上下文最契合的图像（Poth 等人，2023）。通过设计有效的提示来优化输入文本，可以显著提升 CLIP 在视觉词义消歧任务中的表现。例如，“一张用于体育运动的球棒的照片”与“一只倒挂的蝙蝠”能引导 CLIP 关注“bat”的不同词义（Ghahroodi 等人，2023）。近期研究采用提示生成和增强方法，包括加入同义词和定义，以丰富简短的上下文（Li 等人，2023年b）。此外，使用多个提示模板并对结果取平均或采用集成方法，可增强模型对各种上下文表达的鲁棒性（Taghavi 等人，2023）。多样化的模板有助于弥补极简上下文中信息有限的问题。

3.2.2 微调适配

针对视觉词义消歧的 CLIP 联合嵌入空间微调，使用基于任务数据优化的对比损失，提升了跨模态对齐效果，优于零样本基线。SemEval-2023 任务1中的 SLT 在视觉词义消歧数据集上微调了 BLIP，并展现出更优的平均倒数排名（MRR）表现（Molavi和Zeinali，2023）。在 CLIP 的视觉和文本编码器层中插入轻量级适配器，可实现参数高效微调。适配器保留了大部分预训练权重，在降低计算成本的同时提升了任务适应能力（Poth等人，2023）。在相关的自然语言处理和视觉任务上联合训练，可增强视觉词义消歧的泛化能力。例如，同时进行图像描述检索或文本词义消歧任务的训练，可以改善共享表征（Papastavrou等人，2024）。

3.3 大语言模型集成

视觉词义消歧系统正越来越多地利用GPT-3、OPT和InstructGPT等大语言模型，以纳入更多文本上下文并增强多模态消歧。这解决了视觉词义消歧中的主要局限性，即短语级上下文往往无法提供足够的语义细节来实现精确的图像检索。

3.3.1 使用大语言模型生成释义

大语言模型为歧义词生成详细定义，从而为视觉词义消歧模型提供更好的输入。（Li等人，2023年b）表明，使用GPT-3的上下文感知提示在生成词义定义方面远优于早期系统。对于结构化词汇数据库（如WordNet）中缺失的词，这一点尤为明显。该方法称为CADG（上下文感知定义生成），在无需标注数据集或额外模型训练的情况下提高了检索准确率和可扩展性（Kwon等人，2023）。

3.3.2 通过大语言模型扩展上下文

GPT-3 或 OPT 等大语言模型被用于将简短的上下文（通常只是简短的名词短语）扩展为更详细的描述。Li等人（Li等人，2023年b）通过创建额外的文本特征（如同义词、上位词和相关概念），改进了视觉词义消歧的查询。与 CLIP 和 BLIP 等模型的多模态对齐相结合后，这种方法取得了更高的准确率。

3.3.3 基于大语言模型的思维链推理

思维链（CoT）提示正成为一种实用的逐步推理方法。（Kritharoula 等人，2023）及相关研究表明，使用思维链的大语言模型会通过推理各种可能的含义（包括视觉和语言层面）来给出解释。这有助于视觉词义消歧中的消歧选择

在视觉词义消歧中。这种方法能让模型输出更易于理解，预测也更加清晰。

3.4 跨模态生成方法

3.4.1 文本到图像生成

近期进展使视觉词义消歧方法能够生成经模态转换的内容，以辅助消歧。基于扩散的生成模型利用结合上下文的提示，根据歧义词合成图像，从而能够比较生成图像与候选图像，辅助消歧（Yang 等人，2024a）。通过嵌入相似度和人工评估，对生成的特定词义图像与候选图像集进行定量和定性比较，揭示了生成方法在语义对齐方面的潜力（Li等人，2023a）。确保生成图像与文本上下文及候选图像保持一致，有助于减少图像选择中的歧义。

3.4.2 图像到文本生成

生成式图像描述模型为每张候选图像生成详细的文本描述，便于与输入上下文进行语义匹配，通常能提升排序效果和可解释性（Yang 等人，2024a）。将生成的描述与词义释义或扩展上下文进行比较，可在共同的语义空间中整合文本特征和视觉特征。结合原始文本、生成的释义和图像描述中的多种线索，并通过加权融合或注意力机制加以处理，可提升视觉词义消歧的决策效果（Taghavi 等人，2023）。

3.5 多语言与跨语言方法

3.5.1 基于翻译的方法

非英语输入上下文会被翻译为英语，供以英语为主的视觉词义消歧模型处理，从而简化流程，但可能带来翻译引起的歧义（Patil等人，2023）。在多种语言上预训练的模型，包括多语言CLIP变体，可直接编码来自不同语言的上下文，从而在无需翻译的情况下扩大覆盖范围（Laba等人，2024）。利用迁移学习技术，在高资源语言上训练的模型会被微调或适配到低资源语言数据集，以应对数据稀缺问题（Yang 等人，2023a）。

3.5.2 语言无关方法

更多地依赖视觉特征与通用表征，可减少对特定语言上下文的依赖，从而即使在文本极少或含有噪声的情况下也能实现消歧（Yang 等人，2024年b）。所学习的嵌入空间可跨语言和模态对齐语义内容，从而借助语言无关特征促进跨语言视觉词义消歧（Poth等人，2023）。投影或对抗学习等技术将来自不同语言的文本嵌入对齐到共享的图像对齐空间，从而实现无缝的跨语言视觉词义消歧（Papastavrou等人，2024）。

4 研究发现与分析

对视觉词义消歧研究的分析显示，该领域经历了清晰的演进：从纯文本词义消歧逐步转向融合多种模态与语言的方法。传统的词义消歧方法，如基于知识、监督式和无监督式方法，具有较强的可解释性，但在信息有限的上下文中表现不佳（Dixit 等人，2015）。视觉消歧的出现将这一挑战重新定义为多模态检索任务，通过融合嵌入表示来协调文本和视觉含义。早期的多模态系统采用卷积神经网络特征以及 Word2Vec 和 GloVe 嵌入（Gella 等人，2016），此后发展为嵌入对齐框架，例如 CLIP（Radford等人，2021），利用对比学习将图像和文本映射到共享空间中。GPT-3 和 GPT-4 等大语言模型推动了向丰富上下文消歧的转变，能够生成释义、同义词以及与上下文相关的扩展内容（Patil 等人，2023；Papastavrou等人，2024）。

近期研究（Yang 等人，2024a；Molavi和Zeinali，2023）正将基于扩散的生成器与跨模态学习相结合，从而促进不同语言间精准的词义生成和视觉对齐。这一进展表明，研究正趋向于融合 CLIP、生成模型（Stable Diffusion）以及由大语言模型增强的推理能力，进而为构建协调统一的多语言词义消歧系统奠定基础。

各类视觉词义消歧模型的研究结果表明，采用多模态和大语言模型增强策略具有显著且可量化的优势。经过微调的 CLIP 和 BLIP

方法（如 ARPA（Papastavrou等人，2024）和 SLT（Molavi和Zeinali，2023））的表现优于零样本版本，其平均倒数排名（MRR）比基线 CLIP 提高了多达 6–8%。提示驱动生成系统（Li等人，2023年b；Taghavi et al.，2023；Ghahroodi等人，2023）显著提升了 HIT@1 和一致性指标，尤其是在上下文细节有限的情况下。基于大语言模型扩展释义可显著提高词义消歧的准确率；创建定义并加入语义释义（Li等人，2023年b；Patil，2023）有助于消解短语上下文中的歧义。多语言且经过精细微调的模型，例如 FCLL（Yang 等人，2023a）以及各种多语言 CLIP 改编模型（Laba 等，2024），通过使用语言中立嵌入，在英语、意大利语和波斯语测试集上保持了跨语言 MRR，并维持视觉语义连贯性。组合式和混合式方法（例如将集成深度模型与大语言模型结合）取得了 95.77 的平均倒数排名（MRR）和 92.00 的 HIT@1 分数，超越了领先基线模型的表现（Setitra 等人，2025年）。

这篇综述清楚地表明，视觉词义消歧正朝着整合文本推理、视觉落地和释义增强的统一多模态系统发展。当代架构通过结合大语言模型驱动的上下文扩展与对比式视觉文本嵌入方法，实现对上下文敏感的释义判定。

5 视觉词义消歧中的空白与挑战

5.1 上下文与数据约束

当前视觉词义消歧基准测试受到上下文范围狭窄的限制，关注的是非常特定情境中的词义。这一局限导致数据集在多样性和数量上均不足，进而缺少训练和评估有效消歧模型所必需的标注数据。与传统的基于文本的词义消歧不同，后者的模型可以利用丰富的句子或篇章上下文，视觉词义消歧数据集通常只提供极少的上下文，通常仅在歧义目标词旁附上一两个词（Ogezi etal.，2023）。这种对上下文的严重限制造成了不自然的困难消歧情境，可能无法准确反映现实世界中通常能够获得更丰富上下文信息的场景。多项研究表明，这一限制影响重大：使用大语言模型扩充上下文的策略使性能提升了 15-20 %（Li等人，2023年b），凸显了上下文信息对于实现准确消歧至关重要。然而，这些扩充方法也带来了额外挑战，例如大语言模型生成内容可能产生幻觉，以及处理工作流中可能出现错误传播（Kwonet al.，2023）。此外，该领域还面临显著的数据不平衡问题：训练数据集严重偏向英语和单义词示例，而测试集则主要由多义词实例构成，导致分布偏移，阻碍模型泛化（Yang 等人，2023a）。

5.2 模型架构局限性

当前视觉词义消歧研究的主要局限，在于现有视觉语言模型不足以胜任细粒度词义消歧任务。目前领先的先进方法主要采用基于CLIP的架构；尽管这类架构在通用多模态应用中表现出色，但在词义消歧方面仍存在若干显著缺陷。多项研究指出，CLIP倾向于优先选择常见词义，而非较少见或依赖上下文的解释（Dadas，2023），这种偏差严重影响了多义词的消歧效果。此外，无论语义相关性如何，这些模型往往更偏好包含目标词文本表示的图像（Dadas，2023）。CLIP的设计侧重于文本与图像的相似度，而非细致的语义区分，因此与视觉词义消歧任务的需求存在根本性错配。此外，大多数现有系统缺乏端到端可训练架构，往往依赖零样本或静态组件，从而妨碍针对任务进行调优（Li等人，2023a；Taghavi 等人，2023）。

5.3 知识整合的挑战

当前将外部知识源整合到视觉词义消歧系统中的方法面临显著的整合挑战，限制了其有效性。许多研究尝试利用

WordNet、BabelNet 和维基百科等结构化知识库提供定义或语境支持；然而，这些整合工作往往带来的问题多于解决的问题。基于 WordNet 的方法虽然颇受欢迎，却受限于其以英语为中心的设计，且缺乏视觉落地信息（Kwon etal.，2023）。更重要的是，使用大语言模型增强知识仍持续存在幻觉问题。这些模型可能生成看似合理却不正确的定义，从而误导消歧过程（Kwon etal.，2023）。通常用于知识整合的流水线式架构会遭遇错误传播；早期阶段的错误，例如释义检索或义项匹配中的错误，可能会在系统中层层传递，降低最终性能（Kwon et al.，2023）。此外，外部知识访问所需的计算资源以及对 API 的依赖，形成了实际部署障碍，并导致系统性能出现波动（Kwonetal.，2023）。

5.4 多语言与跨语言差距

视觉词义消歧的多语言特性带来了重大挑战，而当前研究才刚刚开始探索这些挑战。该领域明显存在英语偏见。非英语语言受到的关注少得多，性能表现也明显较低（Ogezi 等人，2023）。这一差距源于多种因素：非英语语言的高质量标注多模态数据集数量不足。许多方法依赖翻译，这会引入噪声并扭曲含义。此外，使用英语训练的视觉-语言模型难以实现跨语言迁移（Yang 等人，2024年b）。低资源语言面临的困难更大。例如，乌克兰语在维基百科上的内容有限，也缺乏足够的视觉资源来完成全面的消歧任务（Laba等人，2024）。如何将非英语文本特征与共享视觉表征对齐，仍是一个悬而未决的问题。这一问题源于语言之间的结构差异，以及多语言训练数据的缺乏（Ogezi 等人，2023）。这些多语言挑战严重限制了视觉词义消歧系统在全球范围内的普及，成为其广泛应用的主要障碍。

5.5 评估与泛化问题

视觉词义消歧研究中的评估方法在衡量模型的真实能力及其泛化能力方面存在重要局限。大多数现有研究仅依赖 SemEval-2023 任务1基准测试。这导致评估范围有限，可能无法反映更广泛的消歧问题或现实世界中的表现（Raganato 等人，2023a）。多项研究发现，从开发集转向测试集时，模型性能会明显下降，这表明存在分布偏移问题，削弱了有关模型泛化能力的论断（Berend，2023）。当前基准测试具有人工构造的特性，加剧了这一评估问题；它们可能无法准确体现现实应用中消歧情境的复杂性与多样性。此外，只关注基于准确率的指标，难以深入了解模型在涉及罕见义项、抽象概念或跨领域迁移的情境中的表现。由于缺少用于检验鲁棒性、跨语言公平性或分布外数据表现的标准化评估协议，领域内难以有意义地比较不同方法并追踪真正的进展，形成了显著缺口（Raganato 等人，2023a）。

6 结论

本综述重点介绍了视觉词义消歧领域的重要进展。研究表明，该领域已从传统的基于文本的方法转向有效的多模态系统，将视觉与语言相结合。早期方法采用人工设计的特征和 CNN 嵌入，奠定了坚实的基础。此后，这些方法发展为对比学习模型，例如 CLIP，这类模型能够关联文本和视觉表征。引入大语言模型进行上下文扩展和释义生成，进一步提高了消歧准确率，有助于应对文本上下文有限所带来的关键挑战。

多语言和跨模态生成方法为拓展其在不同语言和领域中的应用提供了有前景的方向。然而，当前系统在处理细粒度语义差异、上下文缺失以及评估挑战方面仍存在困难。将对比对齐、生成式扩散模型和大语言模型推理相结合，为下一代视觉词义消歧系统提供了一个有望取得进展的框架。持续开展的研究应着力解决数据集偏差、拓展多语言资源，并提升可解释性，从而推动其在现实世界中的有效应用。

局限性

本综述存在若干值得注意的局限性。研究筛选范围限定为2016年至2025年发表的研究，因此可能遗漏了多模态或词汇消歧领域更早期的奠基性工作。本综述聚焦于 ACL Anthology 和 arXiv 等特定存储库中的开放获取出版物，可能忽略了需要付费访问的研究或地区性研究。检索时使用了“视觉词义消歧”和“CLIP”等特定关键词，因此可能遗漏采用其他术语或使用未发表数据集的研究。分析重点关注了 SemEval-2023 等英语基准，因此对非英语或低资源语境的了解有限。此外，研究结果主要反映性能指标（HIT@1、平均倒数排名），而非人类可解释性或认知合理性等定性方面。最后，由于排除了正在进行或尚未完成的工作，扩散模型和大语言模型集成方面的最新进展可能未能得到充分体现。

致谢

参考文献

Miuru Abeysiriwardana 和 Deshan Koshala Sumanathilaka。2024。 关于词汇歧义检测和词义消歧的综述。 *arXiv*，abs/2403.16129。

Gábor Berend。2023。 Szegedai 参加 SemEval-2023 任务1：在视觉词义消歧中应用准符号表示。载于 语义评估国际研讨会。

Michele Bevilacqua、Tommaso Pasini、Alessandro Raganato 和 Roberto Navigli。2021。词义消歧的近期趋势：综述。载于 人工智能国际联合会议，第 4330–4338 页。人工智能国际联合会议公司。

Slawomir Dadas.2023. Opi 在 SemEval-2023 任务1中：用于视觉词义消歧的图文嵌入与多模态信息检索。*arXiv*，abs/2304.07127。

Sebastian Diem、Chan Jong Im 和 Thomas Mandl。2023。 希尔德斯海姆大学在 SemEval-2023 任务1中：结合预训练多模态模型与生成模型进行图像消歧。载于第*17*届国际语义评测研讨会论文集（*SemEval-2023*），第130–135页，多伦多，加拿大。计算语言学协会。

Vimal Dixit、Kamlesh Dutta与Pardeep Singh。2015。词义消歧及其方法。CPUH-Research Journal，1(2):54–58。Wala Elsharif、Mahmood Saleh Alzubaidi、James She与Marco Agus。2025年。可视化歧义：分析文生图模型中的语言歧义消解。Comput.，14:19。Spandana Gella、Desmond Elliott与Frank Keller。2019。跨语言视觉动词词义消歧。arXiv，abs/1904.05092。Spandana Gella、Mirella Lapata与Frank Keller。2016。使用多模态嵌入对动词进行无监督视觉词义消歧。arXiv，abs/1603.09188。Omid Ghahroodi、Seyed Arshan Dalili、Sahel Mesforoush与Ehsaneddin Asgari。2023。Sut参加SemEval-2023 任务1：面向视觉词义消歧的提示生成。载于国际语义评估研讨会。Magorzata Grbowiec。2023。Opi pib参加SemEval-2023 任务1：一种基于CLIP并辅以额外词语上下文扩展的解决方案。载于国际语义评估研讨会。David Roi Hardoon、Sándor Szedmák与John Shawe-Taylor。2004。典型相关分析：概述及其在学习方法中的应用。Neural Computation，16:2639–2664。Kaiming He、X. Zhang、Shaoqing Ren与Jian Sun。2015。面向图像识别的深度残差学习。2016 IEEE计算机视觉与模式识别会议（CVPR），第770–778页。Nikita Katyal、Pawan Kumar Rajpoot、Subhanandh Tamilarasu与Joy Mustafi。2023。teampn参加SemEval-2023 任务1：采用零样本多模态方法的视觉词义消歧。载于国际语义评估研讨会。Douwe Kiela、Suvrat Bhooshan、Hamed Firooz与Davide Testuggine。2019。用于图像与文本分类的有监督多模态双向变换器。arXiv，abs/1909.02950。Thomas Kipf与Max Welling。2016。使用图卷积网络的半监督分类。arXiv，abs/1609.02907。Anastasia Kritharoula、Maria Lymperaiou与Giorgos Stamou。2023。面向视觉词义消歧的大语言模型与多模态检索。arXiv预印本arXiv:2310.14025。Sunjae Kwon、Rishabh Garodia、Minhwa Lee、Zhichao Yang与Hong Yu。2023。视觉遇见定义：融入释义信息的无监督视觉词义消歧。arXiv，abs/2305.01788。

Yurii Laba、Yaryna Mohytych、Ivanna Rohulia、Halyna Kyryleyza、Hanna Dydyk-Meush、Oles Dobosevych与Rostyslav O Hryniv。2024。乌克兰视觉词义消歧基准。载于*UNLP*。

Jie Li、Yow-Ting Shiue、Yong-Siang Shih 和 Jonas Geiping。2023a。SemEval-2023 任务1中的增强器：通过提示增强和文本到图像扩散，提升 CLIP 在零样本视觉词义消歧中处理组合性和歧义的能力。*arXiv*，abs/2307.05564。

Keyi Li、Sen Yang、Chenyu Gao 和 Ivan Marsic。2023年b。罗格斯大学多媒体图像处理实验室参加 SemEval-2023 任务1：基于文本增强的视觉词义消歧方法。收录于国际语义评测研讨会。

Margaret Li、Alisa Liu、Zhaofeng Wu 和 Noah A. Smith。2024。面向自然语言处理的歧义类型分类体系。*arXiv*，abs/2403.14072。

Tomas Mikolov、Kai Chen、Gregory S. Corrado 和 Jeffrey Dean。2013。向量空间中词表示的高效估计。收录于国际学习表征会议。

Mohammadreza Molavi 和 Hossein Zeinali。2023。SemEval-2023 任务1：通过使用 BLIP 进行图文检索来增强视觉词义消歧。载于国际语义评测研讨会。

Michael Ogezi、Bradley Hauer、Talgat Omarov、Ning Shi 和 Grzegorz Kondrak。2023年。阿尔伯塔大学参加 SemEval-2023 任务1：用于多语言视觉词义消歧的上下文增强与翻译。载于国际语义评测研讨会。Miguel Ortega-Martín、Óscar García-Sierra、Alfonso Ardoiz、Jorge Álvarez、Juan Carlos Armenteros 和 Adrián Alonso。2023年。ChatGPT 中的语言歧义分析。*arXiv*，abs/2302.06426。Aristi Papastavrou、Maria Lymperaiou 和 G. Stamou。2024年。Arpa：一种利用大型语言模型和 Transformer 推进视觉词义消歧的新型混合模型。*arXiv*，abs/2408.06040。Rahul Patil、Pinal Patel、Charin Patel 和 Mangal Verma。2023年。Rahul Patil 参加 SemEval-2023 任务1：V-WSD：视觉词义消歧。载于国际语义评测研讨会。Jeffrey Pennington、Richard Socher 和 Christopher D. Manning。2014年。GloVe：用于词表征的全局向量。载于自然语言处理经验方法会议。Clifton A. Poth、Martin Hentschel、Tobias Werner、Hannah Sterz 和 Leonard Bongard。2023年。ML MOB 参加 SemEval-2023 任务1：探究 CLIP 在视觉词义消歧中的表现。载于国际语义评测研讨会。

Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。2021。从自然语言监督中学习可迁移视觉模型。载于国际机器学习会议。

Alessandro Raganato、Iacer Calixto、Asahi Ushio、José Camacho-Collados 和 Mohammad Taher Pilehvar。2023a。SemEval-2023 任务1：视觉词义消歧。载于国际语义评测研讨会。

Alessandro Raganato、Iacer Calixto、Asahi Ushio、Jose Camacho-Collados 和 Mohammad Taher Pilehvar。2023年b。SemEval-2023 任务1：视觉词义消歧。载于第*17*届国际语义评测研讨会论文集（*SemEval-2023*），第2227–2234页，多伦多，加拿大。计算语言学协会。

Nikhil Rasiwasia、José Costa Pereira、Emanuele Coviello、Gabriel Doyle、Gert R. G. Lanckriet、R. Levy 和 Nuno Vasconcelos。2010年。一种新的跨模态多媒体检索方法。第 *18* 届 *ACM* 多媒体国际会议论文集。

Insaf Setitra、Praboda Rajapaksha、Aung Kaung Myat 和 Noel Crespi。2025年。利用集成深度模型和大语言模型实现视觉多义性与词义消歧。多媒体工具与应用，84：35727–35759。

Karen Simonyan 和 Andrew Zisserman。2014年。用于大规模图像识别的超深卷积网络。*CoRR*，abs/1409.1556。Deshan Sumanathilaka、Nicholas Micallef 和 Julian Hough。2024a。评估 GPT 在词义消歧方面的潜力：提示工程技术的定量评估。收录于 *2024*年第*15*届*IEEE*控制与系统研究生研究研讨会（*ICSGRC*），第204–209页。Deshan Sumanathilaka、Nicholas Micallef 和 Julian Hough。2025年。GlossGPT：使用少样本思维链提示进行词义消歧的 GPT。《计算机科学程序汇刊》，257:785–792。第16届环境系统、网络与技术国际会议（ANT）/第8届新兴数据与工业4.0国际会议（EDI40）。Deshan Koshala Sumanathilaka、Nicholas Micallef 和 Julian Hough。2024年b。大语言模型能否协助消解歧义？多种大语言模型在词义消歧方面的定量评估。收录于首届网络安全自然语言处理与人工智能国际会议论文集，第97–108页，英国兰开斯特。网络安全自然语言处理与人工智能国际会议。

Zeinab Taghavi、Parsa Haghighi Naeini、Mohammad Ali Sadraei Javaheri、Soroush Rostami Gooran、Ehsaneddin Asgari、Hamid R. Rabiee 和 Hossein Sameti。2023。Ebhaam 在 SemEval-2023 任务1中的方法：一种用于比较视觉词义消歧中跨模态与单模态的方法，基于 CLIP。收录于国际语义评测研讨会。

Sebastiano Vascon、Sinem Aslan、Gianluca Bigaglia、Lorenzo Giudice 和 Marcello Pelillo。2020年。传导式视觉动词词义消歧。*2021 IEEE* 冬季计算机视觉应用会议（*WACV*），第3049–3058页。

A. Yadav、A. Patel 和 M. Shah。2021年。自然语言处理中歧义消解的综合综述。*AIOpen*，2：85–92。

Qihao Yang、Yong Li、Xuelin Wang、Shunhao Li 和 Tianyong Hao。2023a。SemEval-2023 任务1中的南华师范大学 TAM：FCLL：一种用于跨语言视觉词义消歧的细粒度对比语言-图像学习模型。载于国际语义评测研讨会。

Qihao Yang、Yong Li、Xuelin Wang、Shunhao Li 和 Tianyong Hao。2023年b。南华师范大学 TAM 在 SemEval-2023 任务1中的成果：FCLL：一种用于跨语言视觉词义消歧的细粒度对比语言-图像学习模型。载于第 *17* 届国际语义评测研讨会论文集（*SemEval-2023*），第506–511页。

Qihao Yang、Yong Li、Xuelin Wang、Fu Lee Wang 和 Tianyong Hao。2024a。PolCLIP：通过生成多模态互补表示实现统一的图像-文本词义消歧模型。载于 计算语言学协会年会。

Qihao Yang、Xuelin Wang、Yong Li、Lap-Kei Lee、Fu Lee Wang 和 Tianyong Hao。2024年b。MTA：一种用于跨语言视觉词义消歧的轻量级多语言文本对齐模型。*ICASSP 2024——2024* 年 *IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 12166–12170 页。

A 附录

在本附录中，我们对近期视觉词义消歧（VWSD）方法进行了扩展分析。表 1 详细列出了各参考文献中可量化的性能提升和所用技术；表 2则比较了不同类别的 VWSD 方法，并介绍了各自的核心机制和取得的性能指标。

参考文献	Core 技术	具体进展	评估指标	Key 挑战方穿着
Papastavrou 等（2024）（ARPA）	CLIP 微调 微调 + 大语言模型思维链	经过微调的 CLIP with 思维链 大语言模型的推理	平均倒数排名：提升 6–8% ment 超越 CLIP 零样本基准值和准确率超过 15%	缺乏上下文；可解歧义的可解释性 biguation
Molavi 以及 Zeinali（2023）（SLT）	BLIP 微调 微调	优化的 对比视觉词义消歧任务的损失-特定对齐	平均倒数排名：67.91%	跨模态 对齐-细化
Li et al. (2023年 b)	提示 En- 工程 + 大语言模型上下文增强 (CADG)	基于上下文的释义生成，使用 GPT-3 定义，义	HIT@1: 显著提升；一致度 指标: En- 强	严重限制极少的文本语境 text
Taghavi 等人 (2023)	提示驱动的生成	集成提示模板与多模态融合	HIT@1: 显著提升； 一致性指标：有所改善	上下文细节 com- pensation
Ghahroodi 等人（2023）	提示词-驱动生成	精心设计的提示，引导语义义项选择	HIT@1：显著改-进，所需上下文	消歧精度以最少的内 put
Yang 等人 (2023a) (FCLL)	多语言 CLIP 微调- 调整 + 迁移学习	语言无关的 em- 嵌入跨多个语言	英语 HIT@1: 72.56%；平均倒数排名：82.22% （SemEval-2023 1st 名次）	跨语言 MRR 维护（英语、意大利语，FA)
Laba等人 (2024)	多语言 CLIP适配-项	跨语言 迁移 with 视觉 语义连贯性	跨语言 MRR: 在英语中保持一致，IT, FA	多语言 数据集稀缺
Patil 等人 (2023)	基于大语言模型释义扩展-sion	大语言模型生成的 定-义 and 语义释义	准确率：提升词义消歧性能	短语歧义消除
Setitra 等人 (2025年)	混合 En- 似乎 + 大语言模型集成-tion	多个模型输出结合大语言模型的精炼	平均倒数排名：**95.77**；HIT@1: **92.00**	通过模型组合实现最先进的性能
CLIP 基础版-line	零样本 CLIP	直接余弦相似度无需微调	HIT@1：60.5%（英语） 平均倒数排名：73.9%（英语）	基线比较 参考文献

表1：多模态和LLM增强的VWSD方法的可量化性能优势 es

方法类别 gory	核心机制	优势	评估 （准确率, HIT@1、平均倒数排名）
早期 多模态	特征或基于图的 卷积神经网络/Word2Vec 融合	结合视觉（卷积神经网络）以及文本（Word2Vec）特征； 基于图的传播	准确率：最高达 75.6%（无监督）和 83.0%（在 VerSe 数据集 （视觉动词词义消歧）。(Gella 等，2016）
CLIP 零样本	对比嵌入 对齐	No 训练 所需； 利用 预训练视觉-语言对齐	HIT@1： 60.5% （英语）， 37.2% （宏平均）； 平均倒数排名：73.9%（英语），54.4% （宏平均） [SemEval-2023年基线] （Raganato 等，2023年b）
CLIP Fine 微调	任务专属 对齐 通过对比 loss	将 CLIP 适配于视觉词义消歧； 细粒度 对比学习； 参数-高效适配器	HIT@1：65-72%；平均倒数排名：77-82% [TAM： H@1：72.56%， 平均倒数排名：82.22% (Yang 等人, 2023年b)]
大语言模型增强 VWSD	释义/上下文生成-tion 经由 GPT-3/4	充实有限的上下文； 生成定义和 语义扩展	HIT@1：65-76%；平均倒数排名：73-84% [大学 of 希尔德斯海姆：平均倒数排名：73.8% 结合上下文生成] (Diem等人, 2023)
多语言 VWSD	跨语言 CLIP 或 基于翻译的对齐-ment	处理 多种 lan- 种语言（英语， IT， FA）； 语言无关的视觉 表征	英语 Hit@1：84.02%； 意大利语 Hit@1： 84.26%； 波斯语 HIT@1： 64% (Raganato 等，2023年b)

表 2：近期研究中视觉词义消歧方法的性能指标比较