计算机视觉中的推理：分类、模型、任务和方法论

Ayushman Sarkar[1], Mohd Yamani Idna Idris[2], Zhenyu Yu[2*]

[1]Department of Computer Science and Engineering, Birbhum Institute of Engineering and Technology, Suri, 731101, West Bengal, India.
[2*]Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, 50603, Malaysia.

*Corresponding author(s). E-mail(s): yuzhenyuyxl@foxmail.com;
Contributing authors: ayushmansarkar123@gmail.com;
yamani@um.edu.my;

Abstract

Visual reasoning is critical for a wide range of computer vision tasks that go beyond surface-level object detection and classification. Despite notable advances in relational, symbolic, temporal, causal, and commonsense reasoning, existing surveys often address these directions in isolation, lacking a unified analysis and comparison across reasoning types, methodologies, and evaluation protocols. This survey aims to address this gap by categorizing visual reasoning into five major types (relational, symbolic, temporal, causal, and commonsense) and systematically examining their implementation through architectures such as graph-based models, memory networks, attention mechanisms, and neuro-symbolic systems. We review evaluation protocols designed to assess functional correctness, structural consistency, and causal validity, and critically analyze their limitations in terms of generalizability, reproducibility, and explanatory power. Beyond evaluation, we identify key open challenges in visual reasoning, including scalability to complex scenes, deeper integration of symbolic and neural paradigms, the lack of comprehensive benchmark datasets, and reasoning under weak supervision. Finally, we outline a forward-looking research agenda for next-generation vision systems, emphasizing that bridging perception and reasoning is essential for building transparent, trustworthy, and cross-domain adaptive AI systems, particularly in critical domains such as autonomous driving and medical diagnostics.

Keywords: Visual Reasoning, Computer Vision, Graph Neural Networks, neuro-symbolic Models, Causal Inference

1 介绍

计算机视觉在过去十年间取得了巨大进展，从以感知为驱动的应用如图像分类和对象检测转向更复杂的问题，如场景理解、关系建模和视觉问答 [? ? ? ? ?]。然而，如何为视觉系统提供真正的推理能力的问题尚未解决。视觉推理远远超越对象识别，涉及跨对象属性、关系以及时空因果依赖的推断 [? ? ? ? ?]。它作为一个重要的连接纽带，跨越了低级感知和高级认知。虽然取得了显著进展，但在可解释性、跨域通用性和处理复杂任务的能力方面仍存在关键的差距。

这些限制在安全关键应用中更为严重，例如自动驾驶汽车、临床诊断和通过卫星图像进行环境监测，其中透明度和可信性对部署有直接影响 [? ? ? ? ? ?]。尽管深度网络已经展示了非凡的预测能力，但在它们内部发生了不透明的推理 [? ? ?]。现有的可解释人工智能（XAI）方法，如 Grad-CAM [?] 和注意力驱动的显著性图 [?]，大部分是从自然语言处理转移过来的。因此，它们往往忽略了在生物医学图像和遥感图像等应用中更具弹性视觉推理所需的组成、结构和时间信息 [? ? ? ? ? ? ? ? ? ? ? ?]。

大型多模态模型的出现既为视觉推理带来了机遇也带来了挑战。许多近期的方法将语言、提示或大型预训练模型整合到推理流程中 [? ? ? ?]。然而，这种跨模态的整合可能会掩盖特定于视觉的推理机制，并引发关于这些模型是否真正“理解”视觉内容的疑问。现有的调查主要强调推理类型和任务的分类，而相对较少关注包括方法演变、当前评估实践的局限性以及未来研究方向的更广泛视角。

这项调查通过专注于主要基于视觉输入的推理方法，并提出一个统一的分析框架来推进视觉推理系统，从而填补了这些空白。我们将视觉推理分为五种类型（关系型、符号型、时间型、因果型和常识型），并从技术演变、优点和局限性方面对每种类型进行批判性审查。代表性的方法包括图神经网络、模块化注意机制、神经符号流水线和扩展记忆的架构 [? ?]。此外，我们审查了三大评估维度（功能正确性、结构一致性和因果稳健性 [? ? ?]），并讨论了它们在泛化性、可重复性和解释能力方面的局限性。

本研究的主要贡献如下：

- 视觉推理的统一分类法和分析框架：我们将五种类型的推理（关系、符号、时间、因果和常识）统一成一个综合框架，将方法、任务和评估指标结合在一起，揭示推理范式之间的内在关系、区别和互补性。
- 将新兴研究的新方向纳入比较思维的统一推理框架：我们整合了诸如多模态链式思维推理、视觉推理基础模型和更新跨任务基准训练数据集等新兴工作。这些工作在不同推理类型的维度下进行评估，以确定其技术优势、适用范围和弱点。
- 下一代视觉推理系统的前瞻性研究议程：我们指出当前最先进方法在可扩展性、跨域泛化和在弱监督下的可解释性方面的不足。我们概述了包括符号与子符号推理结合、跨域自适应架构设计，以及多类型推理基准测试和测试协议在内的研究方向。

本文的其余部分组织如下。第 2 节介绍了视觉推理的理论基础和分类。第 3 节回顾了核心任务和相关的基准数据集。第 4 节调查了代表性推理方法，包括符号、关系和因果模型。第 5 节讨论了评估协议及其局限性。第 6 节概述了关键挑战和新兴研究方向。最后，图 1 展示了调查的整体结构。

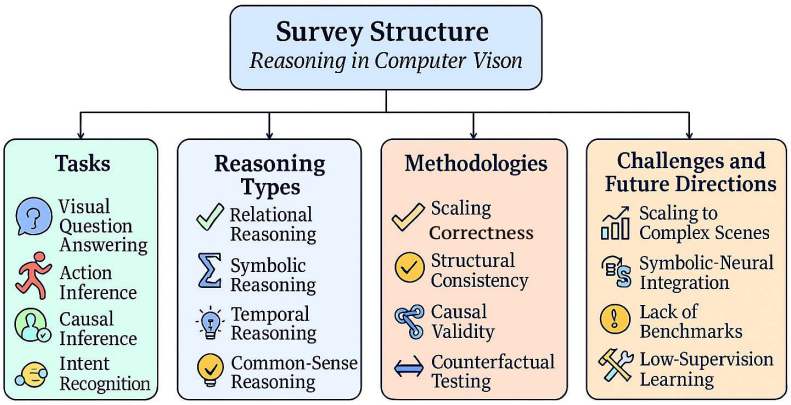

Figure 1 调查结构显示了视觉推理中的推理类型、代表性方法和评估维度。

2 基础

2.1 定义

视觉推理可以系统地分为几种基础类型，每种类型对应一种独特的推理模式。本节概述了文献中广泛研究的五种核心范式：符号推理、关系推理、因果推理、时序推理，以及常识或意图驱动的推理 [? ? ? ? ? ? ? ? ? ?]。

符号推理涉及对源自视觉输入的离散结构化抽象的操控。通常，它依赖于将原始图像转换为中间符号表示，例如场景图、逻辑树或程序，然后进行基于规则的推理。这产生高度可解释的推理流程。例如，一个符号模型可能会推断：”如果对象 A 在对象 B 的左边，并且 B 是红色的，那么 A 就在一个红色对象的左边。”虽然这种明确的可追溯性对于验证和可解释性是有利的，但该方法需要注释过的语义信息，如分割掩码或对象属性，这些信息通常不是直接从像素中提取的 [? ? ? ?]。

相比之下，关系推理强调发现和建模视觉实体之间的成对或更高阶的关系。这些关系可以跨越空间维度（例如“上方”、“后面”）、功能维度（例如“握住”、“支撑”）或语义维度（例如“是... 的一部分”）。关系模型通常通过图神经网络（GNNs）、transformers 或关系网络来实现，而不是使用预定义的规则。通过架构设计，这些模型学习隐式捕捉依赖关系。这种能力对于场景图生成、视觉参考表达理解以及基于类比的图像推理等任务至关重要 [? ? ? ? ? ?]。

因果推理通过建模一个事件影响另一个事件的机制，超越了简单的相关性。在视觉领域，这通常涉及识别方向性依赖关系（例如，“玻璃破碎是因为它掉落了”），而不仅仅是共现。结构因果模型（SCM）、反事实估计器和 do 演算等方法为从视觉数据中推导介入和假设结论提供了基础。这些方法在基于视频的推理中尤为重要，因为事件的时间顺序编码了丰富的因果信号 [? ? ? ? ?]。

时间推理处理的是随时间展开的动态视觉内容。诸如运动预测、动作分割或事件排序等任务需要理解个体对象的时间演变以及实体之间的互动。根据任务复杂度，架构从循环神经网络（如 LSTMs）和时间卷积网络（TCNs）到捕捉长程依赖的基于注意力的时空变换器不等。时间推理在监控、视频质量保证（QA）和自主感知中尤为重要 [? ? ? ? ? ?]。

常识和意图推理涉及对人类行为、目标和上下文线索的隐性理解。它使视觉系统即使在模糊或部分可观测的环境中，也能够做出合理的推断。例如，理解一个人伸手

去拿杯子是意图喝水，这涉及到物理环境、社会规范和先前知识的结合。这类推理在现实世界的应用中发挥着核心作用，包括辅助机器人、驾驶员意图预测和社会场景分析。模型通常结合外部知识库、增强记忆网络或预训练语言模型来支持这种形式的推理 [? ? ? ?]。

2.2 表示

为了支持上述推理类型，视觉系统依赖于严格的数学形式主义和数据表示。这些基础不仅编码视觉语义，还促进结构化和可解释的推理。在下面的部分中，我们总结了视觉推理模型中使用的四种核心表示方法。

场景图。场景图是视觉内容的结构化表示，其中节点表示对象，边表示它们之间的关系。形式上，场景图是一个有向图 $G = (V, E)$，其中 V 表示检测到的实体，$E \subseteq V \times R \times V$ 定义了标记的关系，R 捕捉了诸如“握住”或“旁边”的类型。这些图对于关系推理是基础的，并广泛用于如 VQA、字幕生成和视觉定位等任务 [? ? ? ? ? ?]。

结构因果模型（SCMs）。结构因果模型提供了一个对视觉数据中的因果关系进行建模的原则性框架。它们将每个变量 X_i 表示为其因果父节点 PA_i 和一些外源噪声 U_i 的函数 f_i。因果图 $\mathcal{G}$ 定义了变量之间的依赖关系。通过使用 do 演算，我们可以用诸如 $P(Y|do(X = x))$ 的表达式来估计干预结果。对于涉及干预的视频场景，SCMs 尤其有价值，允许进行“如果物体 A 没有移动会发生什么?” [? ? ? ? ?]这样的反事实分析。

基于逻辑的程序。逻辑驱动的推理在符号视觉抽象上使用形式化的规则集。通过一阶逻辑，这些系统评估如下语句：

$$\forall x, y \, (\mathrm{OnTop}(x, y) \wedge \mathrm{Red}(x) \Rightarrow \mathrm{AboveRedObject}(y)) \tag{1}$$

这些逻辑规则是可解释的，并且可以与神经模块结合，特别是在组合或程序化任务如符号化 VQA 或结构化场景的解释中 [? ? ? ?]。

图和注意力机制。图神经网络（GNNs）和注意力模型是许多现代视觉推理的计算支柱。在 GNNs 中，每个节点 v 基于其邻居 $\mathcal{N}(v)$ 使用以下方法更新其表示：

$$h_v^{(k)} = \sigma\left(\sum_{u \in \mathcal{N}(v)} W^{(k)} h_u^{(k-1)} + b^{(k)}\right) \tag{2}$$

这些模型支持结构化的信息传递，并被用于场景图推理、空间推理和知识感知任务 [? ? ? ? ? ? ? ?]。

同时，基于 Transformer 的注意力通过以下方式计算实体之间的相关性：

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V \tag{3}$$

这些机制在捕捉长距离依赖关系时是不可或缺的，尤其是在时间和多模态上下文中 [? ? ? ? ? ?]。

这些表示工具各自引入了表现力、可扩展性和可解释性之间的权衡。它们的选择深刻影响了视觉系统的推理能力和其输出的透明性。因此，它们构成了解释性视觉推理构建的概念和操作基础。

3 以推理为中心的任务

如表 1 所总结，这些以推理为中心的任务可以分为五种主要类型，它们具有不同的推理能力和代表性基准。视觉推理将计算机视觉从被动感知提升到主动理解，使模型能够解释结构化关系、建模时间动态并揭示因果机制。这些能力对于将低层次视觉特征与高层次认知推断联系起来至关重要。在本节中，我们提出一套核心以推理为中心的任务分类，这些任务需要超越像素级别的识别。对于每个任务类别，我们概述其基本的推理目标、代表性数据集和典型评价设置，为后续章节中的系统比较和方法论分析提供基础。

Table 1 核心视觉推理任务类型、它们的主要推理能力及其代表性数据集的比较。任务描述与第 3 节一致。注意：虽然这些数据集是广泛使用的基准，但许多严重依赖于合成场景或经过整理的领域，限制了它们向不受约束的真实世界环境的直接转移。

Task Type	Reasoning Capability	Representative Datasets
Relational Reasoning	Modeling spatial, semantic, and functional relationships among objects in a scene.	Visual Genome [**?**], VRD [**?**], Open Images V6 [**?**]
Compositional VQA	Performing multi-step symbolic, logical, or compositional reasoning to answer structured visual questions.	CLEVR [**?**], GQA [**?**], VQA v2 [**?**]
Commonsense and Intent Reasoning	Inferring latent goals, motivations, and physical affordances from contextual and background knowledge.	VCR [**?**], VisualCOMET [**?**], Social IQa [**?**]
Causal and Counterfactual Reasoning	Identifying cause–effect dynamics and hypothesizing alternative outcomes under hypothetical interventions.	CausalVQA [**?**], ImageNet-Causal [**?**], SCM Toolkit [**?**]
Temporal Reasoning	Understanding visual event sequences, temporal dependencies, and action transitions over time.	SSv2 [**?**], Charades [**?**], TVQA [**?**]

3.1 关系推理

关系推理评估模型识别和解释视觉实体之间交互的能力，包括空间关系（例如，“上方”，“下方”）、功能依赖（例如，“持有”，“骑”）和语义层级（例如，“部分”，“属于”）。该领域的两个核心任务是视觉关系检测（VRD）和场景图生成（SGG）。

在 VRD 中，目标是提取结构化的关系三元组，如“男人骑自行车”或“狗在桌子下。”SGG 则通过生成一个全面的图表示来扩展这一点，该图编码了所有检测到的物体及它们之间的配对关系。最近的进展利用图神经网络（GNNs）、基于注意力的机制和变压器结构，达到了最先进的性能 [**? ? ?**]。像 Visual Genome [**?**]、VRD [**?**] 和 Open Images V6 [**?**] 这样的基准数据集被广泛采用。重要的是，这些任务的结构化输出支持一系列下游应用，包括视觉问答、图像字幕生成和具身人工智能导航。

尽管如此，目前的关系推理基准测试存在显著的长尾偏差和有限的关系多样性，导致模型在常见的关系类型上表现优异，但难以推广到稀有或依赖于上下文的交互。此外，大多数数据集缺乏对功能或时间依赖性的细粒度注释，限制了在实现健壮的、现实世界关系理解方面的进展。

3.2 组合视觉问答

组合视觉问答扩展了标准的视觉问答任务，要求对图像进行多步推理。模型不再是直接检索答案，而是必须按顺序执行过滤、计数、比较和逻辑推理等操作。CLEVR [?] 和 GQA [?] 等基准数据集专门用于测试这些能力，使用结构化场景和合成问题来降低数据集偏差和捷径学习的影响。

例如，查询“红色立方体是否比蓝色球更多?”需要一连串的推理：检测相关物体，分类它们的颜色和形状，计算每个类别的数量，然后比较这些数量。最先进的方法通常采用神经模块网络、符号推理管道或混合神经符号模型 [? ?] ，产生中间输出——如可执行的程序或视觉轨迹——以增强可解释性和便于调试 [? ?] 。这些特性使得组合式 VQA 在解释性、逐步验证和强大泛化性至关重要的领域中特别相关。

然而，目前基准测试的受控性质限制了其在现实世界中的适用性。场景多样性和语言丰富性仍然有限，合成环境无法捕捉自然图像中的模糊、遮挡和背景噪音。因此，那些在基准测试环境中表现出色的模型在不受约束的场景中往往会经历显著的性能下降，这突显了实验评估与实际应用级别稳健性之间的差距。

3.3 常识和意图推理

一些推理问题需要超越明确的视觉信号进行推断，而是依赖于背景知识、直观物理学或社会约定。此类任务包括预测人类意图、理解合理的未来事件和解释符合现实世界期望的行为。一个经典的例子是推断一个伸手去拿杯子的人很可能是想喝水。

诸如 VisualCOMET [?] 、VCR [?] 和 Social IQa [?] 等数据集通过提示模型假设动机、因果解释或社会反应来测试这些能力。鉴于其定义不明确的性质，此类任务往往需要整合语言模型或外部知识图谱 [?] 。这些类型的推理对于在以人为中心的动态环境中运行的交互式代理、社会意识系统和辅助技术尤为重要。

尽管取得了这些进展，大多数现有的基准仍局限于静态图像和简短的文本提示，这些不足以捕捉意图的流动性、文化规范的微妙之处以及现实世界背景中典型的多主体交互。结果是，即便在精心设计的测试中表现出色的模型，在面对模糊、演变或者具有文化细微差别的情境时，可能会难以泛化，这突显了需要更丰富、纵向的、具有文化多样性的评估环境。

3.4 因果和反事实推理

因果推理侧重于揭示事件发生的原因以及在不同条件下事件可能如何变化。与基于相关性的研究方法不同，因果模型旨在识别和推理干预及其后果。在这一类别中的任务通常包括诸如“是什么导致了玻璃掉落?”或者“如果去除该物体会发生什么?”这类问题。

为支持这一点，像 CausalVQA [?] 、ImageNet-Causal [?] 和反事实生成框架 [?] 的数据集提供了带有因果链的注释视觉场景。像结构因果模型（SCMs）[?] 和反事实逻辑这样的正式框架帮助编码这些关系。在仅结果之外，理解机制至关重要的领域，这种推理尤为重要。包括在自动驾驶、诊断和安全关键决策制定中的应用，评估替代未来或干预措施是可以减轻风险的。

然而，现有的基准常常将因果关系简化为不连续和视觉上显著的事件，忽视了现实环境中常见的潜在混杂因素、远程依赖性以及多因果交互。因此，模型在这样的数据集上可能表现出色，但未能捕捉到动态或安全关键环境中鲁棒泛化所需的细致因果结构。

3.5 时间推理

时间推理涉及到建模视觉事件如何随时间展开。模型需要理解序列：动作的顺序、持续时间和转换，而不是仅仅解释孤立的图像。该领域的核心任务包括动作识别、事件排序、预测以及将自然语言在视频内容中进行时间锚定。

像 Something-Something V2 [**?**] 、Charades [**?**] 和 TVQA [**?**] 这样的数据集，以及最近的基准测试如用于时空推理的 VSTaR [**?**] 和用于交通场景的视频问答的 TUMTraf [**?**] ，挑战着具有依赖于细微运动线索和时序的任务的模型。区分“假装推”和“实际推”等动作取决于微妙的时间动态。为了模拟此类行为，研究人员使用了循环网络（例如 LSTM）、3D 卷积架构，以及最近的时空转换器 [**? ? ? ? ?**] 。

虽然这些进展已大大改善了短期到中期的时间建模，但长时间范围的预测仍然是一个主要瓶颈。遮挡、模糊的过渡以及稀疏的观测常常导致脆弱的表示，无法进行泛化。克服这些限制可能需要具备记忆增强的架构、更丰富的多模态时间线索以及基于结构化常识时间线的推理框架。

4 方法论

以推理为中心的视觉任务需要模型从被动识别转向积极推理、组合整合和结构化知识建模。目前提出的方法通常被归类为四种流行的推理范式：符号推理、神经推理、混合推理和模块化推理。符号推理在通过形式手段实现形式化和可验证性方面最为有效，但需要高度结构化的输入；神经模型在处理噪声和非结构化数据的适应性上表现良好，但基本上无法解释；混合模型尝试兼顾两者优势，但在表示对齐和计算成本上存在问题；模块化系统支持顺序、组合推理，但在开放世界情境中只能在有限程度上扩展。

这些范式不是互斥的，最近的研究越来越关注多范式集成，例如将因果推理嵌入到基于图的模型中，将可微分逻辑引入变压器架构中，并通过大规模视觉-语言模型增强常识基准。这种趋势反映了一种转向，即在统一框架下平衡符号审核性、神经适应性和模块化可解释性的系统，同时解决可扩展性、泛化和跨领域鲁棒性的核心挑战。

表格 2 列出了代表性模型，并讨论了它们的架构范式、推理范围和应用领域，随后对其技术优势和持久瓶颈进行了比较研究——这表明了视觉推理系统未来有前景的方向。符号方法在涉及形式推理痕迹和显式约束检查的应用中很流行，但有高标注和预处理费用。神经方法，特别是图形和变压器风格的网络，具有广泛的表示能力和领域适应性，但在推理的可解释性和对虚假相关性的易受攻击性方面存在困难。混合方法，如神经符号管道和集成知识模型，在结合可解释性和适应性方面显示出前景，但在无缝集成机制和高效训练计划方面仍有不足。模块化推理框架的特点是具有组成化概括和可解释的执行，但往往依赖于精心策划的监督，并在扩展到现实、复杂的世界场景时遇到困难。

最近的进展表明，性能提升越来越多地来自跨范式的融合。例如，将因果推理与图注意力网络结合，或将常识知识库嵌入多模态变压器。这种融合表明，未来的视觉推理系统可能在设计上将是多范式的，优化目标包括可解释性、可扩展性和跨领域的鲁棒性。

Table 2 计算机视觉中具有代表性的视觉推理方法，以推理类型、体系结构范式、骨干模型、推理范围和应用领域进行组织。

Ref.	Method Type	Architecture	Model/Backbone	Scope	Application
[?]	Neuro-Symbolic	Hybrid Pipeline	NS-CL	Global	Visual QA, program execution
[?]	Neuro-Symbolic Reasoning	CNN + Symbolic Executor	CLEVR-based	Global	Scene reasoning, logic tracing
[?]	Symbolic Representations	Hybrid	Custom Symbolic Models	Global	Embodied visual inference
[?]	Modular Reasoning	Controller-Memory	MACNet	Local	Multi-hop VQA, CLEVR
[?]	Neural Module Networks	Modular Neural Network	NMN-VQA	Local	Neural compositional QA
[?]	Modular Pathway Reasoning	Modular Architecture	Custom Modular Nets	Global	Visual scene parsing
[?]	Causal Reasoning	SCM-based	CausalVQA	Local	Counterfactual visual QA
[?]	Causal Disentanglement	SCM-based	Meta-Transfer Objective	Global	Learning robust causal representations
[?]	Relational Reasoning	Graph Networks	GNN-based	Local	Visual relationship modeling
[?]	Scene Graph Reasoning	GNN	Scene Graph GNN	Global	Relationship detection
[?]	Graph-based Explanations	GNN	Custom GNN	Local	Scene graph explanation
[?]	Spatio-Temporal Modeling	3D CNN	I3D, C3D	Local	Video reasoning, action recognition
[?]	Spatio-Temporal Transformers	Transformer	TimeSformer	Global	Space-time reasoning in videos
[?]	Temporal Attention Visualizations	Transformers	Custom Video Models	Local	Attention-based video interpretability
[?]	Commonsense Reasoning	Knowledge-Integrated	VLMs	Global	Social prediction from static images
[?]	Visual Commonsense Reasoning	Multi-modal Networks	VCR model	Global	Social interaction inference
[?]	Knowledge-Enhanced Reasoning	Hybrid	VLM + GNN	Global	Commonsense grounding
[?]	Knowledge Integration	Hybrid	KB-integrated Models	Global	Knowledge-based visual explanation
[?]	Rule-Augmented VLMs	Hybrid	VLM	Global	Explainable rule-guided inference
[?]	Logic Programming Integration	Hybrid	VisualLogic	Global	Formal logic + image analysis
[?]	Attention-based Conditioning	CNN + Language Module	FiLMNet	Local	Feature modulation via language

4.1 符号和神经符号

符号方法依赖于离散的、基于规则的推理，应用于结构化的视觉表示，例如逻辑程序、场景图或文法。它们的主要优势在于透明性，因为每个推理步骤都可以被明确检查和验证。这使得它们特别适合需要严格审计和形式保证的领域。然而，符号方法通常要求干净、结构良好的输入，并且常常无法应对真实世界视觉数据中的模糊性和噪声，从而限制了它们的可扩展性和鲁棒性。

为了解决这些不足，神经符号模型集成了神经感知和符号推理 [? ? ?] 。如图 2 所示，一个基本流程采用目标检测或场景图生成模块将非结构化图像转换为符号对象，并使用程序解析器将自然语言查询转换为形式逻辑程序。然后，程序执行器对这些结构化表示进行符号推理，以生成明确的推理过程并输出答案。这种方法的一个示例是神经符号概念学习器（NS-CL）[?] ，它通过卷积网络将图像编码为符号对象，并

然后应用符号执行来支持高层级的推理。这些混合模型已经在像 CLEVR [?] 这样的结构化数据上表现出色，达到了显著的成分泛化能力和可解释性。

这些方法在逻辑链、概念基础和解释生成的应用中最为有用，其中中间推理过程是至关重要的。然而，它们对精心整理的数据集和手工设计的符号词汇的依赖仍然是开创世界情境中实际应用的障碍。未来的研究越来越多地探讨从非结构化视觉观察中端到端地学习符号抽象，以便神经符号系统能够缩小形式推理与日常图像感知丰富性之间的差距。

最近的进展探索了更灵活的神经符号流程，包括能够进行基于梯度优化的可微逻辑求解器 [?]，以及生成符号程序作为潜在结构的神经符号生成模型 [?]。其他研究利用大型视觉-语言模型（VLMs）从原始图像生成符号场景图或逻辑形式 [? ?]，从而减少了对手工整理的词汇表的依赖。然而，在严重领域转移下的符号落地、对齐从视觉和语言中学习的符号空间以及将推理扩展到无界的开放世界本体时，仍然存在挑战。一个有前景的研究方向是将因果发现与符号抽象结合，使系统不仅能够推理观察到的内容，还能够推理为何发生。

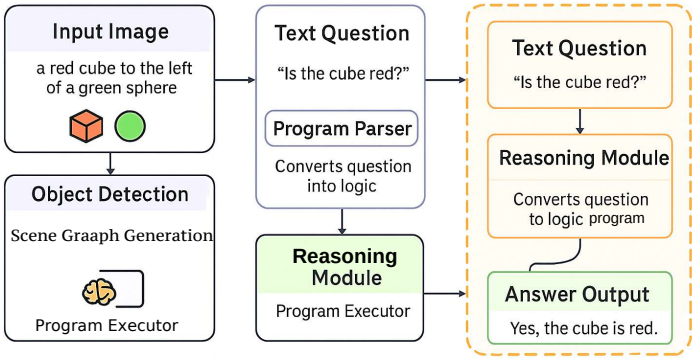

Figure 2 结合视觉解析和符号执行的神经符号推理管道。

4.2 基于关系和图的神经推理

理解对象之间的关系是视觉推理的根本。神经图模型通过物体到节点和交互到边的映射使这些关系显式化。常见的方法有图神经网络（GNNs）、图注意力网络（GATs）和关系网络（RNs），这些方法通过图表示中的上下文信息传播实现关系推理。

如图 3 所示，一个场景被建模为一个图，其中每个检测到的对象（例如，Obj A–D）是一个节点，边（rel1–rel4）表示成对关系，如空间接近、语义共现或物理交互。关系推理通过消息传递展开，其中节点特征通过聚合来自邻居的信息不断被细化。例如，Obj A 可以通过 $A \to B \to D$ 和 $A \to C \to D$ 两种路径影响对 Obj D 的推理，从而实现高阶多跳推断。显式的关系结构有助于场景图补全、对象交互识别和多跳视觉问答等任务，并揭示可解释的中间状态。

尽管这些方法非常成功，但也存在内在的局限性。其性能严重依赖于对象检测和关系标注的精度，一旦在现实世界任务中部署，容易产生级联错误。由于在大图中传播信息的计算复杂性，扩展到人口稀疏的场景或高度动态的场景也存在问题。最近的研究试图通过多模态关系线索的整合、使用预训练的视觉-语言模型作为关系先验，以

及稀疏性意识或分层图表示来提高效率和鲁棒性，从而缓解这些限制。新的研究工作帮助开发了更鲁棒和具有上下文感知的关系推理，以应对开放世界视觉场景。

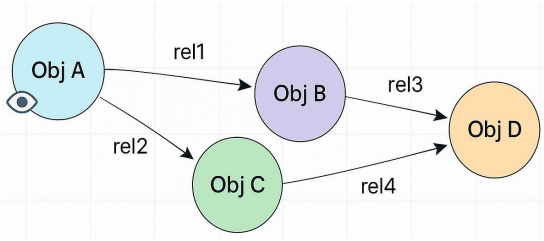

Figure 3 通过对象-对象关系对场景图进行关系推理。

4.3 视觉中的因果推理

虽然整体视觉研究集中于辨别相关性，因果推理则试图辨别为何事情会发生，以及在不同情况下结果会如何不同 [**?** **?**] 。这个范式的核心是结构因果模型（SCM） [**?**] ，它在有向无环图（DAGs）中存储变量的依赖关系，并且能够进行干预和反事实推理。如图 4 所示，节点（因果/效果变量）通过指明因果依赖关系（影响，决定）的有向边相连；闪电符号表示主动干预，叉号圈表示因果删除，放大镜表示反事实查询——这使得关于改变某个原因如何在图中传播的推理成为可能。

在计算机视觉中，诸如 CausalVQA [**?**] 等技术在视觉问答上应用 SCM 原则，将因果数据与观测数据解耦，并推断反事实干预效果。其他方法包括用于特征解耦的因果表示学习，用于防止虚假相关的反事实数据增强，以及在分布转移期间提高鲁棒性的领域泛化方法，如不变风险最小化（IRM）。尽管这些方法前景看好，但它们面临数据集偏差、未观察到的混杂因子和缺乏因果注释的挑战，CLEVR-Causal、VCR-Causal 等标准基准是合成/受限的，并且缺乏在开放世界中因果路径的可解释性。在诸如视频推理等动态任务中，更多的复杂性出现，由于滞后因果效应和多重并发干预，推理变得更加复杂。

我们最近的研究通过将因果发现与自监督表示学习相结合、生成反事实观测值，并从语言输入或传感器观测中注入多模态因果信息，克服了这一缺点，以便为在变化和安全关键的场景中进行机制感知推理做准备。

4.4 模块化和顺序推理架构

模块化架构尝试将大型视觉推理任务划分为更易理解的子问题，每个子问题通过一个专用模块来解决。这些模块的输出随后按顺序或分层装配，以构建整体推理路径。在这种表述中，模块化架构易于组合泛化，并提供逐步解释性 [**?**] ，这使其对于具有明确多步骤推理和中间可由人审核的状态的任务尤其具有吸引力。

一个典型的例子是 MACNet [**?**] ，它使用一个带有显式记忆的递归控制器，连续关注相关的问题组件和图像区域。例如，对于一个像“立方体是否位于球体左侧并且是红色的?”这样的 CLEVR 风格问题，第一步推理会定位立方体，第二步会判断其与球体的空间关系，第三步会验证其颜色属性，从而形成一个透明的推理路径。每一个认知操作对应一个离散的推理步骤，便于对推理过程进行彻底检查。除了 MACNet

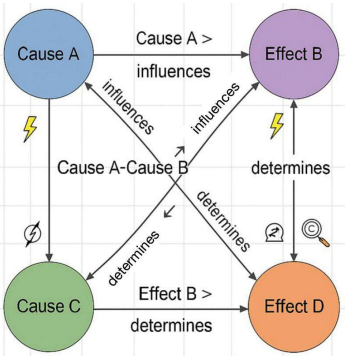

Figure 4 基于结构因果模型（SCM）的因果图用于建模视觉任务中的交互和干预。

之外，还有相关的方法如神经模块网络 [**?**]，它是从学习的模块中即时构建任务特定的网络，以及记忆增强型变压器 [**?**]，可以保留长程依赖并提供反馈驱动的推理。

如表格 3 所示，模块化管道在合成环境（例如，CLEVR）的多跳视觉问答中是最先进的，而在视频序列上进行时间扩展推理时，更倾向于使用内存增强变压器。当需要明确的推理路径时，模块化网络能发挥最佳性能，而诸如内存增强变压器等更为自适应的形式更能适应噪声、不结构化的输入。虽然模块化系统具有可解释性，但其鲁棒性和可扩展性是一个问题。模块化系统通常需要人工策划的监督或程序注释，因此与开放世界数据的兼容性较差。此外，硬模块架构在训练时遇到意外的推理模式时会显得僵化。未来的研究将探索可微分路由和软模块参数化，以使硬模块边界变得更柔软，并使用大型预训练的视觉–语言模型（例如，Flamingo，BLIP-2）作为通用感知骨干网。未来的一个方向是将模块化推理与因果发现或知识基础的先验相结合，实现自适应组合，以更好地推广到开放世界推理和零样本组合同理任务。

从表格 3 中可以看出，在需要可解释性和组成性泛化的任务中，符号和模块化架构占据主导地位，而关系和因果模型更适合动态、语境丰富的环境。时间和常识推理仍然不够成熟，大多数进展来自于将大规模视觉-语言模型作为先验进行整合。

4.5 比较分析与趋势

如表格 2 和 3 所示，符号方法在可解释性方面表现最佳，但需要高度结构化的输入；神经模型提供了灵活性，但代价是可解释性不足；混合框架努力兼顾两者的优点，但却在整合开销上有所折衷；模块化框架提供了组合推理，但在非结构化环境中扩展性较差。最近的研究表明，越来越多的倾向是集成多种范式——例如，将因果发现混入图推理中，将可微分逻辑注入转换器中，或者利用大型视觉-语言模型来落实常识。

在此，我们已确定以下几点：(i) 具有符号审计能力和神经适应性的跨范式融合流水线；(ii) 基于效率扩展意识的架构，以支持高分辨率、时间扩展的输入；以及 (iii) 在多样化、现实的测试套件上进行以泛化为导向的测试，作为未来研究的有前途方向。这种统一方法对于在开放世界、安全关键应用中可运行的下一代视觉推理系统是必需的。

5 评估

测试视觉推理模型远不仅仅是测试输出的准确性；它是在不断变化的条件下测试预测的内容、方式及其成立的原因。与典型的分类问题不同，推理问题包括多步骤推理、

Table 3 通过输入模态、架构、推理类型、核心任务和代表性模型对视觉推理模型进行统一比较。

Input Modality	Architecture	Reasoning Type(s)	Core Task(s)	Representative Models
Image + Text	Neuro-symbolic pipeline	Symbolic, compositional	Visual QA, CLEVR-style logical reasoning	NS-CL [**?**], CLEVR-Hans [**?**]
Image	GNNs, Scene Graphs	Relational, spatial	Scene graph generation, object interaction understanding	Visual Genome GCN [**?**], MoTIFNet [**?**]
Image + Video	SCM-based causal models	Causal, counterfactual	Causal reasoning, VQA under interventions	CausalVQA [**?**], SCM-Vision [**?**]
Video + Text	Spatio-temporal transformers	Temporal, causal	Action recognition, temporal question answering	TimeSformer [**?**], TVQA [**?**]
Image + Question	Modular networks (e.g., MACNet)	Sequential, compositional	Multi-hop VQA in synthetic scenes	MACNet [**?**], NS-VQA [**?**]
Image	Logic-based reasoning	Symbolic, rule-based	Visual program execution, concept grounding	Neural-Symbolic Machines [**?**]
Image	Relation Networks	Pairwise relational reasoning	Comparative QA, Object Relationship Tasks	RN [**?**]
Image + Scene Graph	Graph Attention Networks	Hybrid (symbolic + relational)	Entity linking, semantic relation inference	GAT-ViG [**?**], GIE [**?**]
Video (3D)	Memory-Augmented Transformers	Temporal, commonsense	Future event prediction, video captioning	VINVL [**?**], MMFT [**?**]
Image + KB	Knowledge-integrated models	Commonsense, intent inference	Social reasoning, intention prediction	VisualCOMET [**?**], VCR [**?**]

关系结构，甚至是因果过程——这些都是标量总结分数无法捕捉的方面。当前的评估方法经常过度拟合于简单、狭窄的基准测试，对鲁棒性、可推广性和对所期望推理过程的忠实性等问题仍未解决。

为了解决这些问题，我们采用了一个三管齐下的框架：

- 功能正确性是模型在组合的、多跳的和逻辑丰富的任务上生成正确结果吗？
- 结构一致性是否中间推理路径一致且可解释？
- 因果有效性指的是模型是否捕捉到了因果关系，并对干预作出适当反应。

5.1 功能度量

功能性评估检查模型在需要组合、多跳或逻辑推理的任务中是否产生正确的输出。虽然简单的情况可以用标准准确性来评估，但更具挑战性的情境需要使用与推理复杂性相适应的指标。

基准如 CLEVR [**?**] 和 GQA [**?**] 明确嵌入了多步推理要求，实现了细粒度的错误分析。诊断数据集如 NLVR2 [**?**] 和 ACRE [**?**] 引入了受控的视觉-语言扰动，以测试在分布变化下的鲁棒性。

一个代表性指标是组合推广准确性：

$$\text{Accuracy}_{\text{comp}} = \frac{\text{Correct Answers in Novel Compositions}}{\text{Total Novel Composition Samples}} \tag{4}$$

，它衡量模型推广到熟悉概念的未见组合的能力——这是真正推理的一个标志。除了总体分数之外，功能评估通常将性能分解为子任务（例如，属性比较、集合操作、物体计数），在模块或能力水平上提供可解释的见解。

然而，现有的功能性指标往往在人工或高度控制的环境中起作用，可能会高估在日常世界中的稳健性。大多数任务并不体现对嘈杂的、不完整的或对抗性的视觉观察的推理，而组合性泛化往往仅在有限的概念词汇下进行评估。因此，在基准集合上取得的优异性能可能无法在不受控制的日常情况下产生稳健的推理，这表明需要开发更多样化任务的评估协议，自然数据生态系统，以及跨领域的泛化测试。

除了组成的泛化，功能评估还可以包括诸如推理步骤成功率之类的量，这些量衡量多跳链中正确中间推理步骤的比例，以及事实一致性得分，这些得分估计预测响应与知识基础推理中可验证世界知识之间的一致性。基于强度的功能指标进一步研究模型对控制干扰的鲁棒性，例如对象遮挡、干扰物添加或模态丢失，展示其对视觉-语言噪声的抵抗力。功能指标的选择本身是任务相关的：单跳、封闭域推理可以通过典型的准确性来正确评估，但多跳合成任务需要逐步的正确性和组成泛化准确性作为辅助，而开放域/跨模态任务则需要事实一致性和鲁棒性指标。个别而言，这些改进强调功能正确性并非全有或全无的特性，而是一个与任务相关的能力连续体，每个能力都揭示了视觉推理系统的不同脆弱性/优势。

5.2 结构度量

而功能度量只衡量终端输出，结构评估则关注推理过程，询问模型的内部工作是否由明确定义和可理解的顺序步骤组成。这需要检查中间表示，如注意力图、场景图或模块活动，以验证其内容是否与预期的推理路径一致。例如，对于基于图的模型，可以通过相似性度量来评估预测场景图与真实标注的一致性，如：

$$\text{GraphSim}(G_1, G_2) = \frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} \tag{5}$$

，其中 E_1 和 E_2 分别表示预测图和参考图的边集。更高的分数表示更高的关系推理忠实度。同样，像 MACNet [**?**] 这样的架构产生显式的多步骤推理轨迹，允许评估以检查模块激活是否与预期的逻辑顺序相对应。除了这些之外，结构评估已经扩展到包括注意力的忠实度、模块的相关性以及新兴方法，例如概念图对齐或解耦潜在结构 [**?**]，这些通过使推理过程更加透明来增强可解释性。

最近的发展拓宽了结构评估的范围。多模态对齐度量评估视觉—语言模型中的中间表示是否在不同模态间保持一致。逻辑链一致性测试应用形式推理或符号执行来验证中间步骤的有效性，而基于过程的探测在特定推理阶段引入控制干预以诊断模型行为。其他面向鲁棒性的评估测试中间结构在噪声注入、剪枝或对抗扰动下的稳定性，从而检查推理路径是否具有韧性而非脆弱性。

然而，即使有了所有这些发展，结构化度量仍然存在固有的局限性。许多度量做出了简单化的假设，即选定用于注意力映射的中间表示（例如注意力图、场景图）能够忠实地指示模型的推理，而这一假设在实践中并不一定成立 [**?**]。注意力权重可以在不改变预测的情况下被操控，场景图标签通常是不完整或主观的，而且大多数结构评估都是基于规模很小、精心挑选的数据集，因此生态有效性有限。因此，高结构分数可能代表了对代理表示的一致性，而不是实际的推理能力。为了应对这一问题，我们需要多视图结构验证——交叉检查注意力图、场景图和执行追踪——通过对大规模、现实世界集合的扩展评估，结合因果分析以确保中间追踪反映的是潜在机制而不是假相关。

5.3 因果指标

因果评估将推理分析从相关性扩展到尝试验证模型是否能够在视觉信息中辨别、建模和响应因果关系。与严格的相关方法不同，因果推理直接评估模型对干预的反应及其返回有效反事实的能力。在更高级的情况下，因果评估还评估相关的因果知识是否可以在不同任务或领域之间迁移，从而揭示模型概括因果模式的能力，而不是与某个环境过度拟合。

一个广泛使用的度量是平均因果效应 (ACE)：

$$\text{ACE} = \mathbb{E}[Y \mid do(X = x_1)] - \mathbb{E}[Y \mid do(X = x_0)] \tag{6}$$

，其量化了当将变量 X 从 x_0 设置到 x_1 时，结果 Y 的期望变化。

另一个关键指标是自然直接效应（NDE）：

$$\text{NDE} = \mathbb{E}[Y_{X=1,M=M_{X=0}}] - \mathbb{E}[Y_{X=0}] \tag{7}$$

，其中 M 表示一个中介变量，通过控制间接效应来隔离直接因果影响。

这些指标基于结构性因果模型（SCM）[**?**]，已应用于视觉推理基准测试中，如 CausalVQA [**?**] 和 ImageNet-Causal [**?**]。图 5 展示了一个因果有向无环图（DAG）的例子（Banana → Step → Fall ），支持干预和反事实查询。

同样重要的一个维度是反事实一致性：即当一个关键因果因素被改变时，模型是否适当地调整其预测。例如，在一个“滑倒”视频中，移除香蕉应当改变预测的结果。合成数据集如 CLEVRER [**?**] 提供了受控操作，以系统地测试这一属性。一些较新的基准也在模拟环境中引入动态干预（如物理引擎），以评估模型是否能够在因果依赖关系随着时间变化时调整其推理。

然而，尽管具有诊断价值，这样的数据集往往缺乏现实世界因果关系的复杂性，使得在不受限制环境中的稳健性成为一个未解决的挑战。大多数因果评估程序都是在简单化场景的基础上构建的，这些场景具有明确定义的因果图，仅在极少数情况下能代表真实视觉任务中的模糊性、混杂和反馈。自然场景的真实因果结构通常是不可观察的，因此这些方法只能依赖于合成生成的基准，这可能会夸大模型在非受控环境中的泛化能力。此外，在高维视觉数据中进行干预并非易事——像素级别的变化可能会虚假地引入相关线索——因此测量到的“因果效应”可能反映的是对低级别变化的虚假敏感性，而不是真正的因果推理。这些限制导致了当今因果指标的不连贯性及视觉中野外因果推断的稳健性挑战。

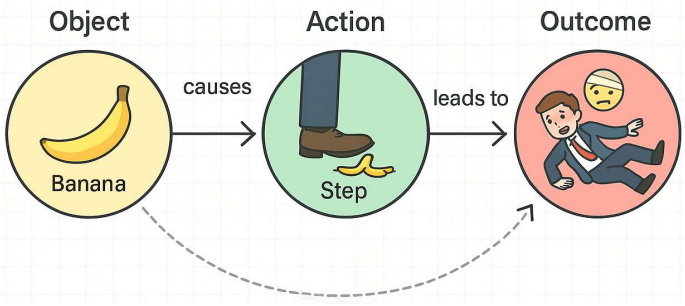

Figure 5 代表视觉依赖性的因果 DAG：例如，香蕉 → 步骤 → 摔倒。

除了常规的准确性之外，评估协议还验证反事实一致性，即当一个主要因果变量被修改后，模型是否能正确修正其预测？如表 4 所总结，没有单一基准能全面捕捉所有推理类型或监督设置。功能性指标衡量结果的正确性；结构性评估探查内部推理过程；因果评估则验证模型对真正推理理解的能力。但现有的指标往往集中于狭窄的推理类型、合成情境或稀疏模式，从而跨范式和实际世界的泛化仍然是一个未解决的问题。因此，全面测试视觉推理需要在现实场景下结合功能性、结构性和因果视角的多方面方案。

Table 4 视觉推理模型的评估维度和指标。

Evaluation Type	Representative Metrics and Criteria
Functional Evaluation	Task Accuracy, Multi-step Reasoning Accuracy, Compositional Generalization Accuracy
Structural Evaluation	GraphSim (Graph Similarity), Attention Map Alignment, Module Trace Faithfulness
Causal Evaluation	Average Causal Effect (ACE), Natural Direct Effect (NDE), Counterfactual Consistency Score

5.4 走向统一推理评价

跨维度整合是至关重要的：高功能评分和低结构保真度可能是利用捷径的标志，而没有功能准确性的稳健因果测量可能表明理论上可行但不可行的推理。未来的评估需要转向多视角推理诊断，其中功能、结构和因果测量共同进行，以揭示失败模式和权衡。

未来具有挑战性的研究方向包括：(i) 构建生态有效的、多模态推理标准，具有因果证明的结构；(ii) 构建自适应评估管道，其任务以对抗性方式适应，以关闭快捷方式漏洞；(iii) 结合基于过程的度量（结构/因果真实性）与训练目标，构建反馈循环，以提高准确性和可解释性。

6 挑战和展望

尽管视觉推理领域的进展迅速，但当前的模型仍面临阻碍其在现实世界中应用的关键限制。与可扩展性、可解释性、泛化能力和监督相关的问题仍然是悬而未决的研究问题。在本节中，我们强调了四个主要挑战，并概述了可能推动该领域构建更强大且认知对齐系统的研究方向。

6.1 可扩展性和泛化

大多数视觉推理模型在面对复杂的、高分辨率的或时间延续的输入时，性能会下降。它们的计算需求使其不适合实时或资源有限的应用，比如机器人或移动设备 [**?** **?**]。此外，在像 CLEVR [**?**] 这样的合成基准上训练的模型常常难以推广到现实场景中，因为在这些场景中视觉和语言的变化更加多样和不可预测。

解决这些问题需要开发既可扩展又模块化的架构。最近的努力包括分层推理管道、像 Longformer [**?**] 这样的稀疏变压器变体，以及随着时间的推移而适应的连续学习框架。同样重要的是提高领域泛化能力，以及训练能够跨异构数据集进行推理而无需

为每个新任务重新训练的模型。这在很大程度上仍未被深入研究，但对于在开放世界环境中的稳健部署至关重要。

尽管有这些进展，大多数“可扩展”或“可泛化”的架构仍然只在受控的基准上进行了验证，未能在不受约束的真实世界流中展示出一致的改进。稀疏注意力和层次化设计经常在推理精度与计算效率之间进行权衡，导致多步骤推理能力的下降。此外，领域泛化策略通常依赖于表面的分布对齐，忽视了因果结构或任务语义的更深层次变化。在不解决这些不匹配的情况下，目前的可扩展性解决方案可能会导致模型在实验室环境中似乎很可靠，但在开放世界部署的组合复杂性下崩溃。

6.2 混合推理范式

神经符号推理结合了神经感知和符号逻辑，为可解释和可泛化的系统带来了巨大前景 [? ? ? ? ?]。这些混合模型尝试将深度学习的适应性与符号表示的结构化表达能力结合起来。然而，这种整合充满了实际挑战：数据表示的不一致、梯度传播的困难以及计算可行性问题阻碍了广泛的采用。

在这个方向的进步需要在多个方面进行创新。一个重点是设计与从神经编码器提取的语义结构对齐的可微分符号算子。另一个是促进神经嵌入和逻辑抽象之间的语义对齐。对噪声的鲁棒性和更好地处理领域转移也很重要。未来的系统可能会通过结合外部知识库受益，从而实现零样本或组合推理能力。

尽管神经符号方法被承诺为“一举两得”的混合体，但绝大多数最先进的方法仅展示了范围狭窄的成功，且通常在精心策划的测试评估中表现出色。在实际应用中，符号成分对噪声或不确定的视觉输入反应脆弱，而神经元素则主导决策过程，使得符号层面降级为事后的过滤器，而不是真正的推理引擎。此外，在动态、实时场景中应用时，其在工程复杂性和计算开销上的集成成本可能超过其解释性改善的好处。如果没有在非结构化、跨领域的情况下进行广泛测试，神经符号推理可能被降级为特殊的研究好奇心，而非可部署的范式。

6.3 基准测试的局限性

许多广泛使用的数据集在范围有限或人工构造的环境中评估推理能力。CLEVR 虽然具有合成性，但缺乏现实世界的复杂性。GQA 和 VCR 提供了更自然的图像，但涵盖的推理技能范围有限。如表 5 所示，没有单一的基准能够充分涵盖所有类型的推理或监督场景。

为了推进这一领域，新数据集应结合现实性和组合性，包含各种类型的推理，并提供具有不同监督水平的多模态输入。丰富的中间注释，例如因果图或推理轨迹，也将使更具信息性的结构和因果评估成为可能 [? ?]。

如今对基准驱动开发的依赖可能反而会生成针对“基准游戏”而优化的模型，而不是优化推理本身。大多数数据不经意间泄露了伪相关性或捷径，导致实现高分而不是真正的多步骤推理。此外，未能在推理分类法和度量定义本身进行正则化意味着跨基准比较无法被信任，从而累积进展受挫。在对抗性构建的基准被社区接受为生态有效且不断变化之前，视觉推理研究可能会被限制在一个评价泡沫中，与开放世界任务的可转化性不足。

Table 5 主要视觉基准的推理类型支持评分。颜色强度表示强度：红色（低），黄色（中），绿色（高）。

Dataset	Relational	Symbolic	Temporal	Causal	Commonsense	Supervision
CLEVR [?]						
CLEVR-X [? ?]						
GQA [? ?]						
VCR [? ?]						
VisualCOMET [? ?]						
CausalVQA [? ?]						
TVQA [? ?]						
ImageNet-Causal [? ?]						
SSv2 [? ?]						
VQA-v2 [? ? ?]						

6.4 弱监督

视觉推理中的一个主要瓶颈是依赖于强标注数据，例如场景图、对象掩码和程序模板。然而，在诸如遥感或医疗等领域，这种标注既昂贵又不可行，因为标注需要专业知识和时间 [? ? ?] 。

最近的方法开始利用弱监督和自监督信号。对比学习 [?] 、掩码图像建模和小样本提示被用于推理任务。像 BLIP 和 Flamingo [?] 这样的视觉语言模型已经表明，预训练的跨模态模型可以用最少的标注数据进行结构化推理。未来的一个关键目标是开发能从间接监督（如奖励、文本反馈或结构先验）中学习的推理代理，并实现跨领域的泛化。有前景的方向包括从原始数据中进行符号自举、自监督逻辑归纳，以及推理模板的跨领域迁移。

虽然弱监督常被视为解决标注瓶颈的方法，但许多现有的“弱监督”设置仍然依赖于隐藏的强监督，例如，隐含有人为挑选的大规模预训练语料库或将标签语义泄露到输入特征的数据集。这导致了一种误导性的泛化感觉，因为模型可能过拟合于预训练数据中的潜在偏见，而不是获得真正的推理能力。此外，在弱监督下的评估通常不一致，缺乏严格的压力测试来验证在非精选领域的鲁棒性。没有将真正的弱监督与隐藏的全面监督分开的标准协议，这个领域可能会高估进展并低估当前方法的脆弱性。

7 结论

这项调查回顾了视觉推理如何将计算机视觉从模式识别转变为认知启发的推断，涵盖了关系、符号、时间、因果和常识五个核心范式，并分析了它们的代表性模型、数据集和评估协议。我们强调进展不仅依赖于准确性，还依赖于评估推理过程的可解释性和因果有效性。尽管在图神经网络、神经符号框架和时空变换器方面取得了进展，但当前系统在可扩展性、对现实世界数据的泛化、神经和符号推理的整合、基准覆盖范围以及对强监督的依赖等方面仍面临关键挑战。从批判性角度看，当今的许多进展仍然受限于基准测试，并偏向于狭窄的、经过精心策划的领域，在不受限制的环境中具有有限的鲁棒性。要克服这些瓶颈，需要更丰富的多模态基准、标准化的因果和结构评估方法，以及自适应的弱监督学习策略，最终实现可以在多样化、开放世界应用中部署的透明和可信赖的推理系统。

8

作者信息

8.1

作者和所属机构 Ayushman Sarkar：计算机科学与工程系，Birbhum 工程技术学院，Suri，731101，西孟加拉邦，印度。
Mohd Yamani Idna Idris：计算机科学与信息技术学院，马来亚大学，吉隆坡，50603，马来西亚。
禹振宇: 马来西亚吉隆坡马来亚大学计算机科学与信息技术学院, 50603, 马来西亚。

8.2

通讯作者 通讯地址: 于振宇 (yuzhenyuyxl@foxmail.com)。

8.3

贡献 Ayushman Sarkar：研究、方法学、撰写——原稿、撰写——审阅 & 编辑。Mohd Yamani Idna Idris：撰写——审阅 & 编辑。Zhenyu Yu：概念化、撰写——审阅 & 编辑。
作者声明没有竞争利益。