Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(1), 9-16
Published Online January 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.151002

Hans汉斯

基于GRPO强化学习算法的视觉语言模型空间推理研究

李亦然

北京邮电大学数学科学学院，北京

收稿日期：2025年11月17日；录用日期：2025年12月19日；发布日期：2025年12月31日

摘　要

视觉语言模型(VLM)在视觉语义任务上表现出色，但在相对深度、三维定位与多实体空间关系等空间几何推理任务中仍存在不足，根源在于高质量空间数据稀缺以及传统监督式训练难以激发模型的空间推理过程。为此，本文提出一套集自动化标注、数据合成与模型训练于一体的端到端框架，以系统提升VLM的空间认知能力。我们构建了无需人工干预的图片空间标注流水线，通过高召回检测、掩码精修与深度相机参数恢复，实现了对照片空间信息的完整标注。在此基础上，我们设计了任务导向的数据合成模块，生成覆盖定性、定量，单跳、多跳的空间推理数据。进一步地，我们基于GRPO算法设计了一套强化学习训练流程，并结合了定制的奖励函数与课程学习训练策略，实现了大模型在空间任务上的稳定训练。实验结果表明，该框架在多个公开与自建基准上超越包括Qwen2.5-VL与InternVL在内的主流模型，验证了语义－几何一致的数据构建与强化学习训练策略优化对提升VLM空间推理能力的有效性。

关键词

空间理解，视觉语言模型，基于规则的强化学习

Spatial Reasoning in Large Vision-Language Models via GRPO-Based Reinforcement Learning

Yiran Li

School of Mathematical Sciences, Peking University of Posts and Telecommunications, Beijing

Received: November 17, 2025; accepted: December 19, 2025; published: December 31, 2025

Abstract

While Visual Language Models (VLMs) excel in visual semantic tasks, they exhibit significant

文章引用: 李亦然. 基于 GRPO 强化学习算法的视觉语言模型空间推理研究[J]. 人工智能与机器人研究, 2026, 15(1): 9-16. DOI: 10.12677/airr.2026.151002

deficiencies in spatial geometric reasoning tasks, such as relative depth estimation, 3D localization, and multi-entity spatial relationships. These limitations stem primarily from the scarcity of high-quality spatial data and the inability of traditional supervised training to effectively elicit spatial reasoning processes within the models. To address these challenges, this paper proposes an end-to-end framework integrating automated annotation, data synthesis, and model training, designed to systematically enhance the spatial cognitive capabilities of VLMs. We construct a fully automated image spatial annotation pipeline that achieves comprehensive annotation of spatial information through high-recall detection, mask refinement, and the recovery of depth and camera parameters, eliminating the need for human intervention. Building upon this, we design a task-oriented data synthesis module to generate spatial reasoning data encompassing qualitative, quantitative, single-hop, and multi-hop scenarios. Furthermore, we develop a reinforcement learning training pipeline based on the Group Relative Policy Optimization (GRPO) algorithm. By incorporating customized reward functions and curriculum learning strategies, we enable the stable training of large models for spatial tasks. Experimental results demonstrate that the proposed framework outperforms mainstream models, including Qwen2.5-VL and InternVL, across multiple public and self-constructed benchmarks. These findings validate the effectiveness of semantic-geometric consistent data construction and optimized reinforcement learning strategies in elevating the spatial reasoning abilities of VLMs.

Keywords

Spatial Understanding, Vision-Language Model, Rule-Based Reinforce Learning

1. 引言

近年来，视觉语言模型(Vision-Language Models, VLMs)在视觉理解与语言推理等广泛任务上取得了突破性进展。然而，与视觉理解和语言推理相比，VLM 在涉及几何结构理解与空间推理的任务上仍表现不足，包括物体间相对深度估计、三维位置推断、多物体空间关系的链式推理等。在这些任务中，模型的输出常包含错误与幻觉，反映出现有训练范式在空间理解领域的系统性缺失[1]。

导致这一瓶颈的根本原因主要体现在两个方面。首先，包含真实三维结构、覆盖多样化空间关系的高质量训练数据极为稀缺，往往依赖密集人工标注，导致数据采集成本高、效率低，难以满足大规模训练需求。其次，主流 VLM 训练目标仍以监督式语言损失为主，缺少模型的主动探索阶段，使模型难以在训练层面形成稳定的空间推理能力。上述两个因素共同限制了 VLM 在复杂空间任务上的泛化能力与可靠性。

为此，本文提出一套端到端的自动化时空数据生成流水线与基于规则的强化学习训练方法。该框架以极低的人力成本构建了大量覆盖全场景、富含空间信息的训练数据，并通过基于规则的强化学习显式增强模型的空间推理能力。整体方法面向真实应用场景，强调高可扩展性、高标注一致性以及对结构化推理目标的适配性。

本文的主要贡献如下：

1) 自动化空间数据标注流水线：提出一条高吞吐、零人工介入的图像空间信息标注流程，结合高召回目标检测(RPN) [2]、掩码精修(SAM) [3]、多模态语义消歧，以及深度/相机内参恢复(Depth Anything V2, MegaSaM) [4] [5]，实现语义与空间几何信息的协同获取，构建可用于空间数据构建的中间表示。

2) 任务导向的空间训练数据自动合成：基于上述标注数据，设计了一个任务导向的数据合成模块，结合模板驱动生成与大模型增强生成两类机制，构建兼具定性与定量属性的空间问答对。同时通过多跳增强策略系统性构建多跳链式空间推理任务，大幅提升训练数据的逻辑复杂度与推理深度。

3) 面向空间推理的基于规则的强化学习框架：提出并实现基于规则的强化学习(GRPO 变体) [6]，构建任务级差异化奖励函数与课程学习机制，从而缓解空间推理中的奖励稀疏与策略不稳定问题，有效提升模型在分布外场景中的泛化性能。

4) 显著优于多项强基线的实验结果：在公开与自构建的空间理解基准(包括相对/绝对深度估计、物体尺度推断、空间关系判断等)上，我们的方法均显著优于多个强基线模型，验证了语义－几何融合数据与规则强化学习对于提升 VLM 空间推理能力的有效性。

2. 相关工作

2.1. 视觉－语言模型与空间推理

近年来，大规模视觉－语言预训练模型在图像描述、视觉问答和跨模态检索等任务中取得了显著进展，但在细粒度空间推理和几何一致性方面仍存在挑战。已有研究尝试通过合成三维渲染数据、引入结构化表示(例如关系图模型)或额外几何监督等方式来提升模型的空间理解能力[1]。然而，这些方法往往依赖于受控的合成场景或昂贵的人工标注，难以有效适应自然图像的丰富多样性。与此不同，我们的方法侧重于自动从互联网级规模的天然图像中恢复几何和语义信息，并直接用于构建面向空间推理的训练任务。这样既兼顾了自然场景的多样性，又提高了几何标注的可用性。

2.2. 自动化数据生成与任务合成

自动化数据合成与增强方法在视觉和语言领域中被广泛采用，包括基于模板的问答生成、基于图形渲染的视觉数据合成，以及最近兴起的利用大型语言模型进行数据扩充(如问题重述、生成困难样本等)。模板方法具有可验证性强的优势，但生成的语言表达多样性不足；基于大模型的方法则能提高语言的自然性，却可能引入语义噪声。为兼顾两者，我们采用了一种模板驱动与大模型增强相结合的混合策略：首先基于模板设计问题以保证题目形式规范且答案可验证，然后使用多模态大模型扩展问题的语言表达和复杂度。最终，通过置信度和几何约束筛选机制过滤出高质量样本，从而同时兼顾数据的规范性和语言的多样性。

2.3. 基于规则的强化学习

在语言生成和对话系统中，使用强化学习(如 PPO 及其在 RLHF 中的应用)来优化生成策略以符合人类偏好或特定任务目标已较为成熟[6]。然而，空间推理任务对奖励设计更加敏感：答案通常允许一定误差，而基于准确匹配的稀疏奖励难以提供稳定的学习信号。为此，有些研究尝试引入连续化奖励、分层奖励或基于规则的判别器来改善训练稳定性。本文借鉴了 GRPO (组内奖励归一化)思路来简化值函数训练的复杂度，并结合任务特定的近似、连续化奖励函数和课程学习机制，旨在针对空间推理任务的容错特性提供更加平滑、稳定的优化信号。

2.4. 本文工作的位置与差异

在综上所述，已有研究在数据合成、几何估计和策略优化等子领域进行了丰富探索，但尚缺乏一条将图像语义标注、像素级空间几何恢复、任务导向数据合成以及基于规则的强化学习有机结合的端到端流水线。本文的主要创新点包括：

1) 语义–几何协同标注：在互联网图像上实现语义和几何信息的协同标注，并据此生成可验证的空间推理任务；

2) 并行合成策略：设计并行的模板驱动和大模型生成的数据合成策略，以同时兼顾数据的规范性和语言表达的多样性；

3) 空间推理定制优化：提出面向空间推理任务的奖励构造和训练稳定性策略，使强化学习能够在具有一定误差容忍度的空间任务上有效优化模型策略。

以上设计不仅增强了方法的可扩展性，也使得本方法在真实世界多样场景中有效提升了视觉–语言模型的空间推理能力。

3. 方案设计

为了提升大模型的空间推理能力，我们分为两个阶段：空间推理数据的收集和基于规则的强化学习阶段。

3.1. 数据收集

为了提升大模型在空间理解任务中的表现，我们构建了一条全自动、高吞吐并具备强泛化性的空间数据收集与标注流水线。考虑到传统空间数据构建依赖大量人工标注，成本高昂且难以规模化扩展，本方案的目标在于最小化人工介入，通过流水线化的过滤与多模态标注机制，从海量网络图像中持续抽取高质量空间理解数据。总体流程包括图像过滤、物体标注与几何标注三个核心模块。在此基础上，我们进一步构造与空间推理任务紧密结合的强化学习训练数据。

3.1.1. 图像收集与过滤

模型空间能力的泛化性在很大程度上依赖数据的多样性。因此，在初始阶段，我们从多源、多场景的大规模互联网图像集与公开数据集中构建候选池。为降低数据分布有偏而导致模型过拟合，我们对图像进行类别均衡化采样，从而维持各类场景、物体类型和拍摄条件的分布平衡。

针对原始候选数据可能存在的大量低质量样本，我们采用了多级过滤策略。首先，通过亮度、对比度、色彩分布等图像质量统计特征，快速剔除曝光极端、对比度异常或内容退化的样本。其次，为确保保留下来的图像具备充分的空间结构信息从而构建下游空间推理数据，我们利用 CLIP 模型对图像文本对齐特征进行筛选，将包含纯文字内容、缺乏显著空间结构的样本识别为负例并予以移除。通过上述分层过滤策略，我们获得了质量稳定、结构明确、适合空间任务的基础图像集。

3.1.2. 图像物体信息标注

现有物体标注模型在复杂自然场景下仍然容易出现漏检、误检和跨物体描述混淆，难以为后续空间任务提供稳定支持。为此，我们设计了一套多阶段的物体标注机制。首先，为了尽可能地覆盖图片中的所有物体，通过高召回率的候选框检测模型 RPN [2]生成初步目标候选区域；随后，为了获得更可靠的物体轮廓和区域边界，我们基于分割的视觉模型 SAM [2]对候选区域进行掩码级精修。

然而仅依赖局部掩码往往难以对每个物体获得全局一致的语义描述，因此我们结合全局图像与物体掩码，将二者输入多模态大模型，生成准确、唯一且区分度高的物体描述文本。该机制有效抑制了多个物体共享模糊描述的问题，为下游多物体推理与空间关系建模奠定了语义基础。

3.1.3. 图像几何信息标注

为了提取图像中蕴含的空间结构，我们引入了对图像的深度估计与相机参数恢复。具体而言，我们利用了 Depth Anything V2 [4]获取图像稠密深度图，进一步用于估计物体的绝对与相对空间位置。同时，

通过 MegaSaM 模型[5]恢复图像对应的相机内参，使得深度图能够在三维空间中获得统一的尺度与坐标系，从而支持可靠的三维几何结构重建与空间关系解析。

上述语义与几何信息的结合，使得单张图像能够被解析为完备的多模态空间结构表示，包括物体实体、物体位置、像素级深度分布及空间拓扑，为后续强化学习训练数据的构建提供核心支撑。

3.1.4. 基于规则的强化学习数据合成

基于前述语义与几何标注结果，我们设计了面向空间推理任务的训练数据自动合成模块，包含模板驱动与大模型生成两类路径。为了进一步提升强化学习训练数据的覆盖度与难度，二者通过数据增强机制。

在任务构建过程中，我们利用物体语义描述、候选框、掩码边界及深度估计，自动生成空间问答对。总体来说，空间任务可分为定性与定量两大类。定性任务包括物体在图像平面中的绝对方位(如上、下、左、右)、物体之间的相对位置、基于拍摄者视角的相对深度排序等。定量任务则包括物体的边界框尺度估计、带相机内参的绝对深度预测等空间问答。为进一步提高数据的训练价值，我们采用多跳组合策略构造高难度问题，通过将多个基础问题合并，使问题呈现更加复杂的空间推理链路，有助于提升强化学习阶段的模型表现。

3.2. 基于规则的强化学习

3.2.1. 基于规则的强化学习框架概述

为了进一步提升模型对空间场景的推理与决策能力，我们采用基于规则的强化学习算法。具体而言，我们采用了 GRPO 算法，该算法避免了传统强化学习方法中训练独立价值函数模型的复杂性，转而通过组内奖励归一化直接推导优势函数。这种设计不仅简化了训练流程，还通过增加模型与环境的交互学习，使模型在空间任务目标下逐步提升生成结果的准确性。相比单纯的监督式训练，此类方法能够在数据分布外的空间场景中更好地保持推理一致性，达到更好的泛化效果。

3.2.2. 奖励函数设计

奖励函数是基于规则强化学习中最关键的元素，其设计直接决定模型能够学习的方向和训练的稳定性。空间推理任务具有明显不同于传统语言任务的奖励需求，而通用奖励机制通常无法适应这些特定要求。传统的数学物理任务往往答案确定，因此奖励函数往往采用正则匹配的形式，而空间推理的任务允许有一定的误差，匹配式的奖励函数会导致稀疏奖励问题。为了解决这个问题，我们针对不同空间任务类别分别构建奖励函数。对于边界框预测任务，我们采用了 IOU 误差来得到奖励。对于绝对深度估计任务，我们设计了近似奖励函数，即使和真实值存在误差，也会根据误差大小给予奖励。对于相对深度估计等规则较难判断的任务，我们利用大模型来判断推理结果的奖励。在此框架下，模型能够在训练过程中获得明确、可量化且与空间语义一致的优化信号。

3.2.3. 强化学习细节优化

利用不加改造的 GRPO 算法做强化学习会有大量的训练不稳定性情况发生，近期的研究指出[7]，熵坍缩是其中的主要原因。为此我们采用了课程学习与训练超参调整两种方式保证训练中的稳定性。

空间推理任务的难度较高，模型在初始阶段容易出现梯度不稳定与奖励稀疏的问题。为缓解模型初期学习困难与后期学习学不到知识，我们引入课程学习策略，以人类学习过程为启发，通过从简单任务逐步过渡到复杂任务的方式，实现数据难度与模型能力的动态匹配。具体而言，我们通过离线过滤与在线过滤相结合的方法，根据模型在训练过程中的通过率实时筛除过难或过易样本，使训练数据的有效性与挑战度保持在合理区间内，从而提升模型收敛速度与泛化能力。

对于超参方面，我们采用了大 batch size、高温采样、clip-higher、定时更换参考模型基准的技术，保

障了训练的稳定与能力的持续增强[7]。

4. 实验设计与结果分析

4.1. 实验配置细节

本实验基于 Qwen2.5-VL-7B 多模态大模型[8]。在基于规则的强化学习阶段，为了增大模型探索的多样性，全局 batch size 设置为 256，每个样本推理条数设置为 16。大模型和视觉编码器的学习率统一设置为 10^{-6}。我们用了 AdamW 优化器，betas 设为 0.9，权重衰减系数设为 0.01。KL 系数设为 0.01。训练长度设置为 8192。为了给模型更大的探索空间，我们调高了 clip 策略的阈值上限到 0.25，下限依旧保持 0.2。为了保持采样多样性，同时避免输出低概率 token 影响训练，我们将采样温度设置为 1，top p 设置为 0.7。

4.2. 模型效果

为了验证本文数据集和模型训练的效果。我们选取了包括 CV-bench、VSI-bench [9]在内的多个大模型领域空间理解的测评榜，将榜单中关于物体空间关系、相对深度估计、绝对深度估计、距离估计等类别的小类题目归类，并基于多个评测榜的分数平均值打分，结果如表 1。

Table 1. Comparison of VLMs on spatial reasoning benchmark
表 1. 在空间能力上的多模态模型能力对比

模型	空间关系	相对深度	绝对深度	距离估计
Qwen2.5-VL-7B [8]	34.5	40.8	48.2	32.5
LLaVA-OneVision-7B [10]	32.7	37.3	37.6	39.4
SpaceR-7B [11]	61.9	38.6	40.5	31.4
InternVL-8B [12]	63.5	34.4	43.6	38.3
SAT-7B [13]	64.2	41.1	45.1	40.6
本文方案	67.8	58.3	53.7	55.8

由表 1 的实验结果可以看出，我们的方案在各个空间能力维度上都取得了优异的效果。相比于训练前的基线模型 Qwen2.5-VL-7B，在空间能力上有着较大的提升。

4.3. 消融实验

为了验证基于空间标注而合成的不同类型数据的有效性，我们对每个合成类别做了消融实验，来验证每种数据的有效性。结果如表 2。

Table 2. Dataset ablation study
表 2. 数据的消融实验结果

模型	空间关系	相对深度	绝对深度	距离估计
基础模型	34.5	40.8	48.2	32.5
+物体定位	38.7	42.6	49.3	38.7
+空间关系	56.7	45.8	50.2	38.9
+相对深度	57.8	53.6	51.1	41.6

续表

+绝对深度	58.1	56.7	53.6	49.3
+相机内参	62.2	58.2	53.6	54.1
+多跳推理	67.8	58.3	53.7	55.8

由表 2 的实验结果可以看出，合成的不同类型数据都较好地达到了预期的目标。能让模型在同分布测试集上获得显著的提升，同时在相近任务上也有一定的正向促进。

5. 结论与展望

本文提出了一套面向多模态大模型空间推理能力提升的自动化数据构建流水线和一套基于规则的强化学习训练框架。针对现有 VLM 在空间几何推理能力不足、输出不稳定以及高质量空间数据稀缺等核心问题，我们从数据与训练两端进行了系统性设计。结合了高召回目标检测、掩码精修、语义去歧义、深度估计与相机参数恢复等模块，我们建立了一个可在大规模互联网图像上运行的语义–几何协同标注流水线，从而以最低人工成本生成高覆盖度、高一致性的空间理解数据。其次，我们提出了任务驱动的数据自动合成机制与基于规则的强化学习训练框架，通过定制化奖励函数、组内归一化优势估计以及课程学习策略，使模型能够在复杂多样的空间任务中获得稳定、细粒度且符合空间推理特征的优化信号。最终，我们在多个空间推理基准上显著超过当前强基线模型，证明了自动化空间数据构建与规则强化学习在提升 VLM 空间推理能力方面的有效性。

尽管本文在大规模空间数据生成与策略优化方面取得了初步成效，但仍有若干值得进一步探索的方向。从单静态图像扩展到真实三维或多视角场景：目前的几何标注依赖单目深度估计与相机内参恢复，几何一致性仍可能受限于模型噪声。未来可探索结合多视角数据、3DGS 等隐式场景重建模型，以获得更高保真度的三维结构，为复杂空间推理任务奠定更坚实的几何基础。面向动态视频与时序推理的数据构建：随着空间任务从静态关系扩展到动态因果关系、运动轨迹预测与跨帧空间链式推理，单帧几何表示已难以覆盖完整语义。未来可将本文框架可扩展到视频级标注与时序一致性建模，使其能够构建时空混合、多层级的推理数据。

参考文献

[1] Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X. and Liu, S. (2024) SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models. In: *Advances in Neural Information Processing Systems*, Curran Associates Inc, 135062-135093.

[2] Ren, S., He, K., Girshick, R. and Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/tpami.2016.2577031

[3] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., *et al*. (2023). Segment Anything. 2023 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Paris, 1-6 October 2023, 4015-4026. https://doi.org/10.1109/iccv51070.2023.00371

[4] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J. and Zhao, H. (2024) Depth Anything V2. In: *Advances in Neural Information Processing Systems* (*NeurIPS*), Curran Associates Inc, 21875-21911.

[5] Li, Z., Wang, Q., Zhang, F. and Tan, P. (2025) MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Monocular Videos of Dynamic Scenes. *Proceedings of the IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 10486-10496.

[6] Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Xiao, M., Li, Y.K., Wu, Y. and Guo, D. (2024) DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. https://arxiv.org/abs/2402.03300

[7] He, J., Liu, J., Liu, C.Y., Yan, R., Wang, C., *et al*. (2025) Skywork Open Reasoner 1 Technical Report.

https://arxiv.org/abs/2502.06657

[8] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., *et al*. (2025) Qwen2.5-VL Technical Report. https://arxiv.org/abs/2502.13923

[9] Yang, J., Yang, S., Gupta, A., Han, R., *et al*. (2024) Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. https://arxiv.org/abs/2406.18385

[10] Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z. and Li, C. (2024) LLaVA-OneVision: Easy Visual Task Transfer. https://arxiv.org/abs/2408.03326

[11] Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F. and Sun, X. (2025) Spacer: Reinforcing MLLMs in Video Spatial Reasoning. https://arxiv.org/abs/2501.01805

[12] Deng, N., Gu, L., Ye, S., He, Y., Chen, Z., Li, S., Wang, H., Wei, X., Yang, T., Dou, M., *et al*. (2025) InternSpatial: A Comprehensive Dataset for Spatial Reasoning in Vision-Language Models. https://arxiv.org/abs/2502.14028

[13] Ray, A., Duan, J., Brown, E., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., *et al*. (2024) SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models. https://arxiv.org/abs/2412.07755