

密 级：机密 ★ 三年

学校代码：10246

学 号：xxxxxxxxxxxx

# 復旦大學

## 硕士 学位 论文

(学术学位)

### 多模态数据合成范式研究

Data Synthesis Paradigm for Multimodal

院 系：计算机科学技术

专 业：人工智能

姓 名：xxx

指 导 教 师：xxx 教授

完 成 日 期：2026 年 1 月 29 日

# 指导小组成员

张五六 工程师

赵 甲 工程师

王三四 讲 师

# 目 录

插图目录	v
表格目录	vii
摘要	ix
Abstract	xi
符号表	xiii
<b>第 1 章 绪论</b>	<b>1</b>
1.1 研究背景 . . . . .	1
1.2 主要贡献 . . . . .	2
1.3 论文组织结构 . . . . .	3
<b>第 2 章 相关工作</b>	<b>5</b>
2.1 视觉模型 . . . . .	5
2.1.1 视觉-语言对比学习与通用对齐 . . . . .	5
2.1.2 从 VLM 到 MLLM：融合大语言模型的多模态生成 . . . . .	5
2.1.3 视觉推理任务与结构化视觉理解 . . . . .	6
2.1.4 生成式视觉模型与可控图像合成 . . . . .	6
2.2 用户图形界面 (GUI) . . . . .	6
2.3 世界知识 . . . . .	6
2.4 多模态数据合成 . . . . .	7
2.4.1 更强 caption 与视觉-语言对齐数据 . . . . .	7
2.4.2 进化式指令构造与代理式数据生成 . . . . .	7
2.4.3 逆向数据工程：从文本到图像的合成闭环 . . . . .	8
<b>第 3 章 面向用户图形界面的视觉语言模型能力提升方案</b>	<b>9</b>
3.1 背景 . . . . .	9
3.2 数据工程：多源融合与高质量构建 . . . . .	9
3.2.1 GUI 数据构造流程 . . . . .	10

3.2.2 开源数据清洗与重构 . . . . .	11
3.2.3 闭源数据构建 . . . . .	12
3.3 模型架构与训练策略 . . . . .	15
3.3.1 视觉编码器与分辨率自适应：AnyRes 机制 . . . . .	15
3.3.2 坐标表示与定位 Token 设计：文本化坐标生成 . . . . .	16
3.3.3 双阶段混合微调（DMT）：GUI 注入与通用对齐 . . . . .	16
3.4 实验与评估 . . . . .	17
3.4.1 评测任务与指标 . . . . .	17
3.4.2 实验结果与分析 . . . . .	18
3.5 本章小结 . . . . .	20
 <b>第 4 章 世界知识能力提升方案</b>	 23
4.1 背景和挑战 . . . . .	23
4.2 知识类别框架 . . . . .	24
4.3 数据工程：世界知识数据构造 Pipeline . . . . .	26
4.3.1 词条生成与图片收集：从知识域到视觉证据 . . . . .	26
4.3.2 图片 Caption 生成：融合视觉描述与知识背景 . . . . .	27
4.3.3 图文相关性过滤：保证视觉-知识对齐质量 . . . . .	29
4.3.4 QA 对生成：从描述到三元组任务 . . . . .	29
4.3.5 QA 对质量过滤：确保训练数据可靠性 . . . . .	31
4.4 训练与实验：从数据类型消融到 CoT 对齐 . . . . .	31
4.4.1 多阶段、多来源数据构建 . . . . .	31
4.4.2 实验一：数据类型消融与冲突分析 . . . . .	32
4.4.3 实验二：引入 Think/CoT 与数据扩充的改进效果 . . . . .	33
4.4.4 最终结果：Chinese SimpleVQA 与 SimpleVQA . . . . .	33
4.5 本章小结 . . . . .	34
 <b>第 5 章 多模态数据构造范式</b>	 37
5.1 背景 . . . . .	37
5.2 构造范式框架 . . . . .	37
5.2.1 构造流程 . . . . .	38
5.2.2 从两类任务实践到统一范式的归纳 . . . . .	42
5.3 模块有效性验证 . . . . .	43
5.4 本章小结 . . . . .	44

<b>第 6 章 结论</b>	<b>47</b>
6.1 研究总结 . . . . .	47
6.2 主要成果 . . . . .	47
6.3 创新点 . . . . .	48
6.4 不足与展望 . . . . .	49
6.4.1 研究不足 . . . . .	49
6.4.2 展望 . . . . .	49
<b>参考文献</b>	<b>51</b>



# 插图目录

4-1 有无世界知识的模型响应对比. 无世界知识：“一个男人和一个女人站在船头.” 有世界知识：“这是 1997 年上映的电影《泰坦尼克号》的经典海报，由詹姆斯·卡梅隆执导，主演是莱昂纳多·迪卡普里奥和凯特·温斯莱特……”	24
4-2 世界知识类别框架（7 个大类与 40 个小类）.	25
4-3 世界知识数据构造 Pipeline.	27
4-4 Caption 数据示例（电影《1917》相关画面）.	28
4-5 三类 QA 与 FinalQA with CoT 示例（格鲁/Gru）.	30
5-1 Auto-Evol 多模态合成数据闭环 Pipeline：数据初始化（含标注）-过滤-原子问题生成-增强与校验-回流重写.	45



# 表格目录

3-1 不同数据配置下的 Grounding 与 Referring 性能对比. . . . .	18
3-2 各类型 UI 控件的 Grounding 性能 (F1 Score). . . . .	19
3-3 Referring 任务各属性预测性能. . . . .	20
4-1 实验一: Part1 数据与不同数据类型配置的消融结果 (Chinese SimpleVQA). Rec 表示 Recognition; Acc 表示 Accuracy; Part1 表示第一批数据; all 表示全部类型数据; e 表示训练轮次; RKF 分别表示 RecQA、KnowQA 与 FinalQA. . . . .	32
4-2 实验二: 引入 Think/CoT 与数据扩充后的改进结果 (Chinese SimpleVQA) . Part12/Part123 表示逐步扩充的数据批次; think 表示 FinalQA 带推理过程的训练设置; Gap 为最优模型相对 qwen2.5-VL-7b 的提升. . . . .	33
4-3 最终结果: Chinese SimpleVQA. Gap 为 Ovis2.5_9B 相对 qwen2.5-VL-7b 的增益. . . . .	34
4-4 最终结果: SimpleVQA. Gap 为 Ovis2.5_9B 相对 qwen2.5-VL-7b 的增益. . . . .	34



# 摘要

多模态大语言模型在通用视觉理解任务上快速发展，但训练数据普遍存在噪声大、对齐弱与推理密度不足等问题，使得“数据质量”逐渐成为能力提升瓶颈。本文首先从两类高难真实需求出发完成方法落地：其一，面向移动端 GUI，构建多源融合与精标语料，并结合高分辨率感知与训练策略提升 Referring 与 Grounding 能力，在 Top 200 应用评测中取得显著提升；其二，面向世界知识，建立知识类别框架与端到端数据 Pipeline，构造 RecQA/KnowQA/FinalQA 并引入 Think/CoT 过程监督，增强“先识别再推理”的推理能力。基于上述实践与实验归因，本文进一步抽象并提出 Auto-Evol 多模态合成数据构造范式，将数据生产从单步“图像 → 文本/问答”生成升级为闭环系统：通过结构化初始化、任务路由与原子问题生成、代理式增强、过滤校验与回流重写，稳定产出高信息密度、高对齐度与高推理密度的训练信号。GUI 与世界知识两条任务线的量化结果为该范式中“结构化锚点/中间表征/过程性监督/评测式质控”等关键设计提供了直接证据，表明该范式与数据工程方法可有效提升模型在真实场景中的定位、理解与知识推理表现，并为后续扩展提供通用框架。

**关键词：**多模态数据合成；Auto-Evol；数据构造范式；GUI 理解；世界知识推理

**中图分类号：**O414.1/65



# Abstract

Multimodal large language models (MLLMs) have advanced rapidly in general vision understanding, yet their progress is increasingly bottlenecked by training data quality, including noise, weak vision–language alignment, and insufficient reasoning density. This thesis first develops and evaluates data and training solutions for two representative real-world challenges. For mobile graphical user interfaces (GUIs), we build multi-source, high-quality annotated data and combine it with high-resolution perception and tailored training strategies to improve core capabilities such as Referring and Grounding, yielding substantial gains on evaluations across top-200 real-world applications. For world knowledge enhancement, we establish a structured knowledge taxonomy and an end-to-end data pipeline, construct a RecQA/KnowQA/FinalQA triplet, and introduce Think/CoT-style process supervision to strengthen the transferable “recognize-then-reason” capability.

Motivated by these practices and evidence, we then abstract and propose Auto-Evol, a paradigm that upgrades conventional one-shot generation (image → text/QA) into a closed-loop multimodal data production system. Auto-Evol integrates structured initialization, task routing with atomic question construction, agent-based augmentation, multi-dimensional filtering and verification, and feedback-driven rewriting, aiming to stably produce training signals with high information density, strong alignment, and high reasoning density. The empirical findings from the GUI and world knowledge tasks provide direct validation for key design choices in Auto-Evol, such as structured anchors, intermediate representations, process supervision, and evaluation-driven quality control. Overall, the proposed paradigm and data engineering methodology effectively improve localization, understanding, and knowledge reasoning in realistic settings, and offer a reusable framework for broader domains.

**Keywords:** Multimodal Data Synthesis; Auto-Evol; Data Construction Paradigm; GUI Understanding; World Knowledge Reasoning

**CLC code:** O414.1/65



# 符号表

$\mathbb{R}$	实数集
$\mathbb{N}$	自然数集
$\mathcal{D}$	数据集
$\mathcal{L}$	损失函数
IoU	交并比 (Intersection over Union)
$B_{\text{pred}}$	预测边界框 (Predicted bounding box)
$B_{\text{gt}}$	标注/真值边界框 (Ground-truth bounding box)
$\tau$	阈值 (如 IoU 判定阈值)
$[l, t, r, b]$	边界框坐标表示 (左、上、右、下)
$R_{\text{ext}}$	外部注释到控件的绑定关系 (Label-to-Control)
$R_{\text{val}}$	当前值到控件的绑定关系 (Value-to-Control)
$x$	输入数据
$y$	输出结果
$f(\cdot)$	映射函数
$\theta$	模型参数
AI	人工智能 (Artificial Intelligence)
ML	机器学习 (Machine Learning)
DL	深度学习 (Deep Learning)
CNN	卷积神经网络 (Convolutional Neural Network)
GAN	生成对抗网络 (Generative Adversarial Network)
GUI	图形用户界面 (Graphical User Interface)
VLM	视觉语言模型 (Vision–Language Model)
MLM	多模态大语言模型 (Multimodal Large Language Model)
OCR	光学字符识别 (Optical Character Recognition)
SFT	监督微调 (Supervised Fine-Tuning)
CoT	链式思维 (Chain-of-Thought)
QA	问答 (Question Answering)
RecQA	识别问答 (Recognition QA)

KnowQA	知识问答 (Knowledge QA)
FinalQA	推理问答 (Reasoning QA)
DMT	双阶段混合微调 (Dual-stage Mixed Tuning)
AnyRes	动态分辨率/任意分辨率机制 (Any Resolution)

# 第 1 章 绪论

## 1.1 研究背景

多模态大语言模型（Multimodal Large Language Model, MLLM）近年来在视觉理解与语言生成方面取得显著进展，并逐步从通用图文理解走向真实应用形态的智能体（Agent）与复杂推理任务。在这一阶段，模型能力的上限越来越受制于训练数据的质量、**对齐与推理密度**，而不仅仅是数据规模：互联网抓取的图文对普遍存在噪声大、语义对齐弱、图像质量参差不齐等问题；人工标注虽然可靠，但成本高且难以覆盖长尾场景、复杂交互形态与多步推理链路。与此同时，模型在真实落地任务中暴露出的误差模式往往具有“结构性”：例如对小目标与文字细节不稳定、对视觉证据与语言描述的绑定松散、对需要先识别再推断的任务易出现短路映射或幻觉。这些现象共同指向一个核心问题：**如何以可控、可迭代的方式生产高质量多模态训练数据**，以稳定提升模型在特定能力维度上的表现<sup>[1-3]</sup>。

从应用需求看，本文关注的两类典型任务进一步凸显了数据问题的复杂性。其一是面向用户图形界面（GUI）的视觉语言能力提升。在 UI Agent 的感知-决策闭环中，模型必须在高分辨率、控件密集且状态易变的界面中完成 **Referring (指代识别)** 与 **Grounding (定位落地)** 等基础能力：既要“看懂”控件语义（按钮含义、输入框字段、图标象形意义），又要以像素级稳定输出目标位置。GUI 场景对训练数据提出了更苛刻的要求：不仅需要精致的几何标注与 OCR 对齐，还需要显式建模控件关系与交互状态，避免“看似可点但不可点”“标签与输入框绑定不稳”等误监督带来的偏差（详见第3章）。这类需求很难仅依赖通用图文对或粗粒度标注来满足，必须通过结构化数据组织、混合标注与严格质控来构建可学习信号。

其二是世界知识能力提升。尽管视觉语言模型能够生成流畅的图像描述，但对于“图像背后隐含的事实、概念与关系”往往缺乏可靠掌握：模型需要将图像中的**具体实例**锚定到知识空间中的**抽象实体与事实**，并在多跳条件约束下完成“先识别再检索/回忆、再筛选推断”的推理链路（详见第4章）。世界知识注入不仅面临知识源分散异构、同名消歧与噪声混入等问题，还面临“视觉证据  $\leftrightarrow$  知识事实”难以可验证对齐的关键挑战：如果训练样本不可验证或答案不唯一，就会引入错误监督并放大幻觉；如果只提供最终答案，模型又容易学习到短路映

射，推理能力提升不显著甚至退化。因此，世界知识任务迫切需要一种能够同时保证对齐可信度与推理过程可学习的数据构造方式。

综合上述两条主线可以看到，多模态数据构造不能仅被视为“生成一些图文或问答对”的单步过程，而应被提升为一种面向能力提升的系统工程：一方面要同时兼顾真实分布覆盖与可控强对齐，另一方面要把复杂能力分解为可监督、可校验的原子单元，并通过过滤、增强与反馈闭环持续迭代数据分布。由此，本文将研究主线聚焦于**多模态数据合成范式**：围绕“数据源组织—结构化初始化—任务路由与生成—多维校验—闭环迭代”的全流程，建立可复用、可落地的数据生产方法学，使合成数据能够在信息密度、对齐度与推理密度三个维度上同时满足训练需求。

在本文中，第3章与第4章分别从 GUI 与世界知识两类高难任务出发，给出面向真实瓶颈的具体数据工程流水线与能力拆解方式：前者强调细粒度定位、OCR 与结构化关系/状态监督，后者强调实体锚定、知识注入与带过程约束的推理监督。这两类实践进一步抽象、统一并上升为一个通用的多模态合成数据构造范式（在后续章节中展开），从而回答本文试图解决的核心问题：**如何以范式化的方法把“零散的一次性合成”升级为“可评测、可迭代、可迁移”的数据生产系统**，并以此稳定提升多模态模型在关键能力维度上的表现。

## 1.2 主要贡献

本文围绕“多模态数据合成范式”这一主线，结合 GUI 与世界知识两类高难任务的实践，主要贡献总结为三点：

- **提出可复用的多模态合成数据构造范式 (Auto-Evol)**：将以往零散的一次性“图 → 文/问答”生成，系统化为面向能力提升的数据生产流程，统一刻画多源数据接入、结构化初始化、任务路由与原子问题生成、增强与多维校验、失败回流重写等关键环节，使高信息密度、高对齐度与高推理密度的数据能够被稳定、迭代地产出与复用（对应范式章节的整体框架与模块设计）。
- **面向 GUI 的高质量数据工程与可执行定位能力提升方案**：针对 GUI “控件密集、文字细小、状态易变”的特征，构建多源融合的数据流水线与结构化监督信号 (bbox/OCR/关系/状态)，并结合高分辨率感知与训练策略提升 Referring 与 Grounding 两项核心能力；在真实应用分布的控件级评测中验证方案有效性（如表3-1、表3-2与表3-3所示）。
- **面向世界知识的可控对齐与推理贯通数据体系**：构建覆盖广、可扩展的世界知识类别框架，并设计“词条 → 图片 → caption → 相关性过滤 → 三种 QA”

的数据 Pipeline；进一步通过引入带推理过程（Think/CoT）的 FinalQA，缓解短路映射与任务冲突，显著提升需要“先识别再推理”的世界知识能力（如表4-2所示）。

## 1.3 论文组织结构

本文的组织结构如下：

第一章（绪论）介绍研究背景、问题动机与本文的主要贡献，并给出全文组织结构。

第二章（相关工作）回顾视觉-语言模型与多模态大模型的发展脉络，以及 GUI 理解、世界知识增强与多模态数据合成相关研究，为本文方法设计提供背景支撑。

第三章（面向 GUI 的能力提升）围绕 Referring 与 Grounding 两项核心能力，给出 GUI 场景的数据工程流水线、结构化监督设计与训练策略（如高分辨率感知与双阶段混合微调），并通过控件级评测验证方法有效性（第3章）。

第四章（世界知识能力提升）构建世界知识类别框架与端到端数据构造 Pipeline，提出 RecQA/KnowQA/FinalQA 的任务分解，并通过引入 Think/CoT 等过程性监督提升“先识别再推理”的世界知识推理能力（第4章）。

第五章（多模态数据构造范式）在前两章任务实践基础上，进一步抽象提出 Auto-Evol 多模态合成数据构造范式，系统阐述闭环 Pipeline 及其关键模块，并结合实验现象对核心设计进行归因式验证。

第六章（结论）总结全文工作与主要结论，分析研究不足，并展望未来研究方向。



# 第 2 章 相关工作

## 2.1 视觉模型

视觉模型的发展经历了从“以视觉编码器为核心的表征学习”到“视觉-语言对齐的基础模型”，再到“面向复杂任务的多模态指令跟随与推理”的演进。本文的研究问题（多模态数据合成范式）与后续两条任务线（GUI、世界知识）都建立在这一演进脉络之上，因此本节重点回顾与本文最相关的视觉-语言模型（VLM）与多模态大模型（MLLM）的关键路线。

### 2.1.1 视觉-语言对比学习与通用对齐

对比学习为视觉与语言建立共享语义空间，是现代视觉-语言预训练的重要基石。CLIP 通过海量图文对进行对比学习，使模型获得了强迁移的视觉表示能力，并在零样本分类等任务上表现突出<sup>[4]</sup>。这类方法的意义在于：它把视觉表征从“固定类别监督”转向“语言可描述的开放世界概念”，为后续多模态生成式建模提供了通用语义对齐能力。

### 2.1.2 从 VLM 到 MLLM：融合大语言模型的多模态生成

随着大语言模型（LLM）能力增强，主流路线逐渐从“检索/匹配式 VLM”走向“以 LLM 为核心的生成式 MLLM”，即利用视觉编码器提取视觉 token，再由 LLM 统一完成理解、生成与推理。Flamingo 采用少样本条件下的视觉-语言融合架构，展示了在多任务多域上的通用能力<sup>[5]</sup>。BLIP-2 进一步提出“冻结图像编码器 + 冻结 LLM”，通过轻量模块完成跨模态桥接，在保持训练效率的同时获得强泛化<sup>[6]</sup>。

在“指令跟随 (instruction following)”范式下，多模态模型通过视觉指令微调（Visual Instruction Tuning）进一步获得更好的对话式交互与任务泛化能力。LLaVA 系列工作系统化探索了视觉指令微调的数据构造与训练策略<sup>[7-8]</sup>。Qwen-VL 则强调更通用的视觉理解与工具化能力，在多类视觉任务与多语言场景下展示了较强表现<sup>[9]</sup>。总体而言，这一阶段的核心变化是：模型能力的瓶颈逐渐从“是否能对齐视觉与语言”转向“是否具备足够高质量、可控结构的监督信号”，这也直接引出本文对多模态数据合成范式的研究动机。

### 2.1.3 视觉推理任务与结构化视觉理解

除通用图文理解外，图表、文档、截图等“结构化视觉输入”对模型的视觉解析与逻辑推理提出了更高要求。ChartQA 提出了包含视觉与逻辑推理的图表问答基准，推动模型从“读图”走向“推断”<sup>[10]</sup>。DePlot 将图表理解转化为“图到表”的结构化转换，从而以可执行中间表征支撑后续推理<sup>[11]</sup>；MatCha 进一步引入数学推理与图表反渲染等机制增强预训练<sup>[12]</sup>。Pix2Struct 提出将“截图解析”作为预训练任务，强调对 UI/网页等截图的结构化理解能力<sup>[13]</sup>。这些工作共同表明：当任务依赖结构化信息（布局、文本、关系）时，仅靠通用 caption 或粗粒度监督往往不足，需要更强的结构化中间表示与针对性数据构造策略。

### 2.1.4 生成式视觉模型与可控图像合成

在图像生成方向，扩散模型推动了高分辨率、高保真生成能力的快速发展。以 SDXL 为代表的高质量扩散模型显著提升了高分辨率图像合成质量<sup>[14]</sup>，为“文本 → 图像”的逆向数据工程提供了可用工具基础。对于本文而言，生成模型的价值不仅在于生成本身，更在于为多模态数据合成提供可控视觉证据，以弥补真实数据的长尾稀缺与弱对齐问题。

## 2.2 用户图形界面 (GUI)

面向 GUI 的视觉理解与交互是多模态模型落地的重要方向之一。与自然图像相比，GUI 截图具有控件密集、文字细小、布局强结构化与状态易变等特征，使得模型需要同时具备（1）细粒度视觉感知能力（小目标定位、OCR 相关细节）；（2）结构化理解能力（控件关系、层级与语义绑定）；（3）可执行输出能力（产生可解析的坐标或动作序列）。这类需求决定了 GUI 领域往往需要更结构化的数据与预训练任务。

相关研究中，Pix2Struct 将截图解析作为预训练任务，强调从截图中恢复结构化表示以提升视觉语言理解能力<sup>[13]</sup>，为“截图类输入”的建模提供了代表性路径。总体来看，GUI 相关工作所体现的共同趋势是：模型要在真实界面中可靠工作，必须依赖高分辨率视觉输入、文本信息对齐、以及对布局/关系的显式建模；因此数据侧不仅需要“屏幕截图 + 描述”，更需要可执行的定位监督与高置信的结构化锚点。这一观察与本文第3章采用的多源融合、结构化标注与质控迭代思路一致。

## 2.3 世界知识

世界知识 (World Knowledge) 能力指模型对事实、概念与关系网络的掌握，以及在视觉证据与知识之间建立可验证链接并完成多步推理的能力。对于视觉

语言模型而言，世界知识增强的难点通常不在于“语言侧是否存储过知识”，而在于“视觉侧是否能稳定锚定实体实例”，以及“推理链路是否可控、可迁移”。

一方面，推理方法与对齐策略对世界知识问答质量至关重要。链式思维提示（Chain-of-Thought, CoT）表明显式推理过程可以显著提升 LLM 的复杂推理能力<sup>[15]</sup>；面向指令跟随的对齐研究（如 InstructGPT）通过人类反馈强化学习提升模型遵循指令与输出质量<sup>[16]</sup>；Orca 等工作进一步探索从强教师模型的“解释轨迹”中进行渐进式学习<sup>[17]</sup>。这些研究共同指出：当任务需要多步推断与事实一致性时，仅监督最终答案可能诱发“短路映射”或幻觉，过程性监督与可验证约束更利于能力迁移。

另一方面，衡量与改进模型事实性也是世界知识增强的重要组成部分。SimpleQA 聚焦短回答场景下的事实性评测，为分析模型在事实正确性上的薄弱点提供了参考<sup>[18]</sup>。在多模态场景中，世界知识能力往往体现为“先识别再调用知识”的链路是否稳定、答案是否唯一可核验。本文第4章即在这一思路下构造 RecQA/KnowQA/FinalQA 分解，并通过引入带 CoT 的 FinalQA 提升可迁移推理能力。

## 2.4 多模态数据合成

多模态大模型进入“数据质量成为瓶颈”的阶段后，如何规模化构造高质量、强对齐且具备推理密度的数据成为关键问题。现有工作大体可归纳为三类：更好的 caption 与对齐、更系统的指令进化与交互扩展、以及利用生成模型进行逆向数据工程。

### 2.4.1 更强 caption 与视觉-语言对齐数据

ShareGPT4V 强调通过更高质量的 caption 提升多模态模型训练效果，表明“更好的描述”可以显著改善视觉-语言对齐与下游能力<sup>[11]</sup>。这类工作揭示：caption 不只是训练目标本身，也可以作为后续过滤、校验与任务构造的中间表征，从而把“图像”转化为可复用、可验证的文本证据。

### 2.4.2 进化式指令构造与代理式数据生成

在纯语言领域，Self-Instruct 提出通过模型自举生成指令数据以实现对齐<sup>[19]</sup>，并启发了多模态领域对“自动化数据生产流程”的探索。AgentInstruct 进一步强调 agentic flow，将数据生成组织为多步骤的代理流程<sup>[20]</sup>。在多模态方向，MMEvol 提出 Evol-Instruct 以增强多模态指令的多样性与能力覆盖<sup>[2]</sup>。在结构化推理任务上，Synthesize Step-by-Step 通过工具、模板与 LLM 组合生成推理型数据，为“可执行/可验证”的数据构造提供了范例<sup>[21]</sup>。这些方法共同体现出“从单轮生成到

流程化生成”的趋势：数据构造需要明确能力目标、提供结构化约束，并引入过滤与迭代机制以控制噪声与幻觉。

### 2.4.3 逆向数据工程：从文本到图像的合成闭环

除“图 → 文”的正向构造外，利用生成模型进行“文 → 图”的逆向数据工程可在源头提升对齐度与可控性。SynthVLM 系统研究了合成图文数据集的构建策略，通过 caption 清洗、扩散生成与自动指标筛选等环节提升图文对齐与图像质量<sup>[3]</sup>；而高质量扩散模型（如 SDXL）为高分辨率图像生成提供了更强基础<sup>[14]</sup>。对于本文所研究的“多模态数据合成范式”而言，正向与逆向两类数据流的结合，使得系统既能覆盖真实分布，又能在长尾与强对齐需求下进行可控补齐。

# 第 3 章 面向用户图形界面的视觉语言 模型能力提升方案

## 3.1 背景

随着多模态大语言模型（Multimodal Large Language Model, MLLM）在通用视觉理解任务上的快速突破，面向真实设备自动化的 **UI Agent**（交互界面代理）逐渐成为落地应用的重要方向。与通用场景不同，移动端与桌面端 GUI 具有控件密集、文字细小、布局结构化、交互状态易变等特征，使得模型在“看懂屏幕”这一前置能力上面临更高门槛。

而在 UI Agent 的感知–决策闭环中，**Referring**（指代识别）与 **Grounding**（定位/落地）是最基础也最关键的两项能力：

- **Referring**: 给定屏幕区域的边界框（Bounding Box, bbox），模型需要输出该区域控件的功能、文本或属性描述，即模型的理解界面能力，例如输入“在 bbox [10,10,50,50] 区域的控件是什么？”，输出“这是搜索按钮”。
- **Grounding**: 给定自然语言指令，模型需要在屏幕截图中定位对应控件并输出其 bbox，即模型定位的能力，例如输入“登陆按钮在哪里？”，输出“bbox [100,200,300,400]”。

尽管通用 MLLM 已具备一定视觉理解能力，但在 GUI 垂直领域仍存在三个主要挑战：(1) **细粒度定位偏差**，小面积控件的坐标难以稳定。(2) **图片信息损失严重**，直接缩放图片后细节与文字信息丢失严重。(3) **视觉与语言难以对齐**，传统检测模型缺乏语义理解能力、视觉语言模型定位效果存在偏差）。为此，本章提出一套以**多源数据工程**为核心、结合**双阶段混合微调（DMT）**与**高分辨率架构优化**的系统方案，目标是在移动端 GUI 最流行前 200 应用的使用场景测试数据中实现 Referring 准确率  $\geq 90\%$ 、单控件 Grounding 准确率  $\geq 90\%$ 。

## 3.2 数据工程：多源融合与高质量构建

数据是 GUI 理解能力提升的基石。本节围绕“**多源融合**”与“**高质量构建**”两条主线，构建开源清洗数据、闭源精标数据与合成数据相互补位的训练语料体系，以保证数据质量可控、可迭代。

### 3.2.1 GUI 数据构造流程

GUI 领域的 Referring 与 Grounding 能力提升，本质依赖于“**理解语义且精准定位**”的数据监督：模型既要学会控件的语义功能（登录/搜索/关闭等），又要学会在高分辨率、密集布局中输出像素级稳定的定位结果。相比自然图像，GUI 场景存在**小目标密集、文字依赖强与交互状态多变**等特点，其中“文字依赖”指的是：移动端控件**种类多样、外观相似但功能多变**，大量关键语义（如“登录/下一步/同意/关闭”以及输入框的字段含义与提示信息）主要通过界面文字表达，必须借助截图中的文本信息（OCR）才能准确判断控件身份与动作目标。使得简单的图文对或粗粒度检测标注难以覆盖真实需求。因此，为了切实提升模型的 GUI 场景下的能力，本节提出了一种可以处理多种开源数据，同时通过重写、人工精标的方法来构建高质量数据的数据构造流程。

整体流程可概括为“**收集多种源数据 → 结构化初始化 → 混合标注与属性补全 → 清洗与结构化聚合 → 任务/QA 数据生成与过滤 → 质量控制与迭代**”，其必要性体现在以下几个方面：

- **收集多种源数据**：为覆盖真实分布与长尾场景，数据来源需同时包含开源 GUI 样本（提供多样布局先验）、闭源真实 App 截图（补齐中文与主流应用生态）、以及必要的合成/重写样本（补足稀缺交互形态与困难指令表达）。这一阶段决定了后续数据分布的上限。
- **结构化初始化**：将不同来源样本统一为可计算、可追溯的结构化记录（截图、分辨率、视图层级、候选控件集合等），并规范坐标表示与字段定义。该步骤的作用是把“图像”转化为可被标注系统与训练系统共同消费的中间表示，为后续的类别映射、关系建模与 OCR 对齐提供稳定接口。
- **混合标注与属性补全**：围绕控件级 bbox 与文字内容进行高质量标注，并补充可交互性与状态等属性。该步骤通常采用“检测 +OCR+ 属性预测”的自动化预标注，并由人工校验与补充兜底。其作用是同时回应 GUI 三大难点：通过紧致 bbox 缓解小控件定位偏差，通过 OCR 字段解决文字依赖，通过状态/可交互标签避免“看似可点但不可点”的误监督。
- **任务/QA 数据生成与过滤**：在结构化标注与控件集合基础上生成面向训练的指令/问答数据。一方面将能力拆解为可监督子任务：**感知类**（定位/识别/OCR）、**理解类**（控件语义解释与指代描述）、**推理类**（结合上下文的动作选择与多步计划）；另一方面可进一步生成描述 UI 功能语义与操作意图的 QA 数据，并通过规则过滤与相似度去重降低噪声与冗余。该步骤的关

键是把“标注信息”组织成模型可学习的交互格式，从而同时对齐语义与布局两条能力链路。

- **质量控制与迭代：**对样本进行一致性与正确性检查（例如 bbox 合法性、OCR 与文本一致性、指令与目标控件匹配、关系图完整性），并将失败样本回收为后续补标/重写/再采集的候选。该步骤是保证数据分布稳定、降低噪声与抑制幻觉的必要条件。

需要强调的是，上述流程虽然并非后续章节提出的多模态数据合成范式，而是从 GUI 任务需求出发的一种可行数据工程路径：它首先明确了 GUI 训练数据必须同时覆盖语义与图形定位，并以“结构化初始化-精细标注-任务化构造-质控迭代”为核心步骤，为后续提出更通用的多模态数据构造框架奠定基础。

### 3.2.2 开源数据清洗与重构

开源数据在规模与多样性上具备优势，但往往存在噪声、布局失真与语义标签不一致的问题。以 RICO 为代表的 GUI 数据集<sup>[22]</sup>包含大量移动端界面及其视图层级（View Hierarchy），然而原始数据中存在不可见节点、错误标注、层级断裂等现象，直接用于训练会导致模型学习到错误的几何与语义先验。本文采用 CLAY（Clean Layout for Android UI）思路对 RICO 进行清洗与增强<sup>[23]</sup>，主要包括：

- **噪声过滤与层级修复：**剔除包含“INVALID”标签、尺寸为 0、透明不可见等节点，并修复视图树结构，确保层级关系可用于后续关系建模与样本构造。
- **语义标签映射：**将原始 Android View 类名映射为更具语义的通用组件类别，以减少类别碎片化带来的学习难度，例如将 android.widget.EditText 及其子类统一映射为 TEXT\_INPUT，将 FloatingActionButton 映射为 BUTTON。
- **任务 Prompt 构造：**在清洗后的结构化数据上生成标准化监督样本，包括：
  - **Grounding：**输入自然语言指令（如“点击右上角的搜索图标”）与页面截图，输出归一化 bbox；
  - **Referring：**输入 bbox 与页面截图，输出组件类别、功能与可见文本（可结合 OCR 字段）。

通过上述处理，开源数据为模型提供了覆盖广的 GUI 布局先验与基础交互语义，但在中文场景、复杂交互与状态理解上仍存在缺口，需要闭源数据与合成数据补齐。

除 RICO 外，本文在开源侧还引入了 **MobileViews**<sup>[24]</sup> 与 **OS-Atlas**<sup>[25]</sup> 两类 GUI 数据，以补齐交互覆盖与定位监督：

- **MobileViews**: MobileViews 是面向移动端智能体与 UI 分析的大规模数据集<sup>[24]</sup>，其 MobileViews-600K 版本包含从 Google Play Store 上两万余应用采集的 60 万级“屏幕截图–视图层次结构（VH）”配对数据。其最新版本基于 DroidBot 并针对大规模采集进行了优化<sup>[26]</sup>，在保持与 DroidBot 输出结构一致的同时，能够捕获更全面的交互细节。本文利用该数据集中控件的类型与组成信息（包括 bbox 与控件内文字）构造训练样本：一方面将控件文字作为 OCR 监督与 Referring 依据，另一方面将 VH 中的层级与控件类型作为语义先验，增强模型对复杂页面结构的泛化能力。
- **OS-Atlas**: 本文主要使用 OS-Atlas 提供的 GUI Grounding 数据集<sup>[25]</sup>，其中目标元素位置存储在 `bbox` 字段，采用  $[left, top, right, bottom]$  的归一化小数表示（每个值均为  $[0, 1]$  范围内相对图像宽高的比例）。在任务构造时，我们将其与其它数据源统一到一致的坐标表示与字段规范（并保持 `bbox` 的几何语义一致），用于补强“指令  $\rightarrow$  目标控件位置”的 Grounding 监督，尤其覆盖长尾指令表述与跨应用布局差异。

### 3.2.3 闭源数据构建

针对开源数据在中文语境、主流 App 生态与复杂交互状态覆盖上的不足，本文构建闭源高质量 GUI 数据，并按迭代版本记为 v1 与 v2：其中 v1 用于早期验证与规范冷启动，重点覆盖文本输入、开关、弹窗等高频交互组件；v2 在 v1 基础上进行二次校验，剔除定义不清的控件并加入负样本（不可点击区域/误导性区域），以提升 Grounding 的判别能力。闭源数据构建强调“可训练”与“可评测”的一致性：既提供控件级几何定位监督，也提供可解释的语义与状态监督，并进一步产出可用于指令学习的 UI 意图问答数据。

**第一阶段：数据标注（混合标注模式）** 该阶段从原始截图开始（可选携带页面 XML/VH 等元数据），采用“自动化预标注 + 人工校验补充 + 纯人工兜底”的混合模式，以兼顾效率与准确性：

- **UI 元素检测**: 使用轻量检测模型（如 YOLO）自动检测图标、按钮、输入框等控件并输出 bbox<sup>[27]</sup>。
- **文字识别（OCR）**: 对候选框区域使用 PaddleOCR 提取控件文字（如“登录”“取消”等），形成文字与位置的对齐监督<sup>[28]</sup>。

- **属性预测**: 利用已训练的多模态模型（如 qwen2vl）对候选控件进行属性补充，例如控件细分类别、可交互性以及选中/开启等状态<sup>[29]</sup>.
- **人工介入与整合**: 标注人员对预标注结果进行校验与修正，并补充模型遗漏的控件；对于复杂样本走纯人工标注分支。最终合并得到统一标注结果，确保每个关键控件的 **BBox**（位置）、类别（类型）与属性（文字/状态/可用性等）准确一致。

**OCR 增强**是闭源数据的重要补强环节。为降低通用 OCR 在艺术字体与复杂背景下的误识别与“幻觉”，本文引入更高精度的文本抽取与校验流程，生成用于训练的高质量中文文本标注，从而让模型在 Referring 任务中能够稳定对齐“控件外观–文字内容–交互语义”三者。

**第二阶段：数据清洗与结构化** 该阶段对标注结果进行质量控制与格式转换，生成训练友好的结构化数据表示：

- **质量校验**: 检查 bbox 合法性、OCR 文本异常（如过长/乱码）、属性冲突（例如“不可交互但标为可点击”）以及控件间的关联关系是否合理。
- **数据提纯与聚合**: 生成包含控件类型、属性、bbox 的纯净样本，并以图片为单位聚合为控件列表（每张截图关联一个包含其所有 UI 元素的集合）。
- **格式转换**: 将结构化控件集合转换为多种下游格式，以支持指代理解 (ref)、定位落地 (grd)、列表/容器理解 (list) 等任务的数据构造与训练。

**第三阶段：QA 数据生成与过滤** 该阶段是将结构化控件信息转化为“可学习的功能语义与操作意图”的核心步骤：将聚合好的控件信息与精心设计的系统提示 (sys prompt) 及少样例 (few-shot examples) 结合，调用强多模态模型（如 GPT-4o/qwen2vl）生成关于界面功能与操作意图的问答对 (QA)<sup>[29–30]</sup>。例如，生成问题：“如何退出登录？”并给出可执行的操作序列描述。为保证质量与覆盖，进一步执行数据过滤与去重：

- **过滤噪音**: 基于规则（如问题过短、答案无意义、指令与目标控件不匹配等）过滤低质量样本；
- **去重降冗余**: 计算图片表征向量 (image embedding)，基于相似度剔除视觉高度相似页面上生成的重复或冗余 QA，避免训练集冗余堆积。

除上述三阶段流程外，针对闭源数据本文还建立了一套标准化的平台屏幕控件标注规范。该规范不仅关注视觉边界 (bbox)，还显式引入控件分类学、控

件关系图谱与状态向量，从而为 Referring 提供可解释的语义监督信号，并为 Grounding 提供“可交互目标”的判别依据。

**细粒度控件分类学 (Taxonomy)** 首先构建了覆盖移动端原子级交互单元与容器级逻辑单元的 23 类核心控件体系，遵循“**原子粒度优先**”原则，以减少布局强相关类别造成的歧义。

- **原子交互类 (Atomic Interaction)**: BUTTON、ICON、TEXT、IMAGE，以及输入与开关类 (TEXT\_INPUT、SWITCH 等)。
- **复合容器类 (Composite Containers)** : NAVIGATION\_BAR、TAB\_BAR、CONTAINER、POPUP、ADVERTISEMENT、NOTIFICATION 等。
- **动态调整类**: SLIDER、PROGRESS\_BAR 及其区域类控件等。

在该规范中，剔除了 CARD\_VIEW、LIST\_ITEM 等布局强相关类别，转而采用“原子粒度优先”的原则，以减少模棱两可的布局标注，并促使模型更关注交互本质。

**语义关系建模 (Directed Relationship Graph)** 传统检测仅预测类别与坐标，缺乏对控件间逻辑的理解。为此，该规范在标注中显式构建控件间的有向关系图 (Directed Relationship Graph)，该信息对 Referring 尤为关键：

- **标签-控件绑定 (Label-to-Control)**: 定义关系  $R_{ext} = (T_{label} \rightarrow C_{target})$ 。例如输入框左侧的“用户名”文本被标记为外部注释，并指向对应的 TEXT\_INPUT，使模型能够更稳定地理解“点击输入用户名的位置”这类指令。
- **值-控件绑定 (Value-to-Control)**：定义关系  $R_{val} = (T_{value} \rightarrow C_{target})$ 。例如输入框中已填写的“张三”、下拉框当前显示的“北京市”被标记为当前值，并指向父控件，辅助模型区分“输入/修改/清空”等意图。
- **功能依赖 (Functional Dependency)**：例如广告弹窗右上角的“关闭 (X)”图标与 ADVERTISEMENT 容器建立依赖关系，使模型学习“如何关闭弹窗/广告”这类意图导航。

**多维属性与状态向量 (Attributes & State Vectors)** 为支持 UI Agent 决策，为每个 bbox 附加多维属性与交互状态信号，核心为交互状态三元组：

- **Activated**: 标注开关开启、Tab 选中、单选框勾选等状态；

- **Interactable**: 区分可交互与置灰不可交互控件;
- **Filled**: 标识输入框/选择器是否已有内容, 辅助判断执行“输入”还是“修改”.

此外, 对于 ICON 类补充图标释义 (结合上下文填写象形含义, 如 “房子 → 主页”), 并对 BUTTON/TEXT\_INPUT 记录内部文本, 用于增强视觉编码器与文本空间的对齐目标.

**人机协同标注与一致性校验 (Human-in-the-Loop)** 为保证标注效率与精度, 采用了半自动化标注工作流: OCR 预处理生成文本 bbox 与内容, 检测模型生成候选控件框, 属性预测模型补全类别与状态; 标注员进行校验与微调, 并引入一致性检查机制 (例如外部注释必须有指向目标、关系图无孤立节点、bbox 紧致且合法), 以降低噪声样本对训练的负面影响.

## 3.3 模型架构与训练策略

在数据工程基础上, 本文进一步从架构与训练两方面优化模型的高分辨率感知与坐标生成能力, 以满足 GUI 小目标定位与中文文本识别的需求.

### 3.3.1 视觉编码器与分辨率自适应: AnyRes 机制

GUI 截图包含密集文字与小图标, 且移动端屏幕常呈现高长宽比 (竖屏). 若直接将原图缩放到视觉编码器的固定正方形输入 (如  $224 \times 224$  或  $336 \times 336$ ), 会产生不可逆的信息损失: 细小图标与字体被压缩后趋于同质化, 导致 Referring 依赖的文本语义对齐变弱、Grounding 需要的像素级几何细节丢失, 进而表现为定位偏差与相邻控件混淆.

为此, 本文采用动态分辨率 (AnyRes) 机制, 其核心思想是同时提供全局布局视野与局部高分辨率细节, 并将二者的视觉特征统一送入语言模型. 具体流程如下:

- **全局缩放 (Global Resize)**: 对原始截图生成一张缩放后的全图, 用于捕捉导航栏、内容区、弹窗遮挡等全局布局关系, 提供稳定的页面结构先验.
- **网格切分 (Grid Cropping)**: 依据原图长宽比自适应选择网格 (如  $1 \times 2$ 、 $2 \times 2$  或  $2 \times 3$ ), 在原始分辨率空间切分得到多个局部 Patch; 每个 Patch 再缩放到视觉编码器的合适输入尺寸, 使编码器能够清晰感知按钮文字、图标边缘与细粒度状态差异.

- **位置对齐与编码 (Patch Position Encoding)**: 为避免“看清局部但不知道其在整屏哪里”，为每个 Patch 附加其网格索引/相对位置编码，使语言模型能够在推理时对齐“局部细节  $\leftrightarrow$  全局位置”.
- **特征融合 (Feature Fusion)**: 分别提取全局图与各局部 Patch 的视觉特征序列，并在 Embedding 层进行拼接（或按顺序组织后拼接）作为统一的视觉 token 输入语言模型。这样模型既能利用全局图理解页面结构，又能从局部 Patch 读取细小文字与图标细节，从而同时提升语义理解与精确定位。

AnyRes 带来的主要代价是视觉 token 数量增加，从而提升计算开销。本文在网格数量选择上遵循“**足够看清关键细节**”与“**控制推理成本**”之间的折中：在典型竖屏场景下优先采用少量 Patch 覆盖主要区域，以在可接受的计算预算内显著提升 Grounding 稳定性与中文 OCR 相关的 Referring 准确率。

### 3.3.2 坐标表示与定位 Token 设计：文本化坐标生成

为充分利用自回归语言模型的生成能力，本文采用文本化坐标表示，将 bbox 输出为  $[0, 1000]$  范围的归一化整数序列（如 `<box>250, 500, ...</box>`）。同时引入 `<|box_start|>`、`<|box_end|>`、`<|ref_start|>` 等特殊 Token，区分普通文本生成与定位任务生成，降低坐标串与自然语言串的混淆。

其中，**坐标归一化**的目的在于消除不同设备分辨率与长宽比带来的尺度差异：将像素坐标映射到固定范围后，模型学习到的是“相对位置与相对尺寸”的几何概念，而非对某一固定分辨率的硬编码记忆。这使得同一指令在不同屏幕尺寸（例如  $1080 \times 1920$  与  $1440 \times 2560$ ）下仍能输出一致的定位结果，并便于与数据工程中来自不同来源的数据（开源/闭源/合成）进行统一标注与混合训练。同时，使用整数文本序列输出坐标可避免额外回归头设计，并与语言模型的离散生成范式天然兼容。

**特殊 Token** 的引入则承担“结构化约束与可解析性”两项作用：一方面，`<|box_start|>`、`<|box_end|>` 等标记显式划分坐标片段的边界，减少模型将坐标数字与自然语言混写导致的格式错误；另一方面，这类标记使下游系统能够以确定性规则解析模型输出，将 bbox 作为可执行的动作目标（例如点击/框选），从而降低“输出不可用”带来的失败率。对于训练而言，特殊 Token 也相当于为定位任务提供了更清晰的任务提示（task delimiter），有助于缓解定位生成与对话生成之间的干扰，提升坐标生成的稳定性与一致性。

### 3.3.3 双阶段混合微调 (DMT)：GUI 注入与通用对齐

为兼顾 GUI 能力与通用对话推理能力，本文采用双阶段混合微调（Dual-stage Mixed Tuning, DMT）策略。该策略的出发点是：当模型同时学习“专用能

力”（GUI 定位、OCR 密集对齐）与“通用能力”（多模态对话、指令遵循与推理）时，简单的多任务混合训练往往引发两类问题：其一，在数据量较大或任务差异较大时容易出现**能力冲突**（不同任务目标相互干扰，导致某些能力不升反降）；其二，若采用顺序微调，模型又容易发生**灾难性遗忘**（后续通用对齐覆盖掉前期学到的 GUI 定位能力）。DMT 通过“先专用、再混合回放”的方式，在两者之间取得平衡。

- **阶段一：GUI 领域知识注入 (Domain Adaptation)**. 训练数据 100% 为 GUI 相关数据（开源清洗数据 + 闭源精标数据 +OCR 密集任务），以 bbox 预测与文本对齐为主；训练中冻结视觉编码器大部分参数，重点微调语言模型部分，使模型快速获得稳定的屏幕元素感知与坐标生成能力。
- **阶段二：通用能力恢复与对齐 (General Alignment)**. 在保持 GUI 能力的同时引入通用多模态对话数据（约 80–90%），并回放阶段一的高质量 GUI 样本（约 10–20%）。通过调节混合比例  $k$  平衡灾难性遗忘与能力冲突，实践表明保留约 15–20% 的 GUI 回放数据即可稳定维持定位精度。

从数据混合角度看，已有研究表明：在**低资源**条件下，多源数据混合可能带来协同效应；而在**高资源**条件下，简单扩大混合数据总量更容易触发能力冲突，且性能退化往往更多来自“数据总量与任务差异”的叠加，而非比例本身。因此，阶段二采用“通用为主、GUI 少量回放”的设计：一方面继续提升通用对话与推理能力，另一方面用比例  $k$  的 GUI 回放样本提供持续的定位与 OCR 对齐约束，显式缓解遗忘。同时，相比直接多任务学习，DMT 更能保留通用能力；相比纯顺序微调，DMT 又能通过回放显著降低专用能力遗忘，从而实现 GUI 专用能力与通用能力的更优折中。

## 3.4 实验与评估

### 3.4.1 评测任务与指标

围绕 Referring 与 Grounding 两项核心能力，本文采用控件级评测作为主要指标：

- **Referring 准确率**：给定目标 bbox，模型输出的类别/功能/文本描述与标注答案一致的比例；对于含 OCR 文本的样本，额外关注关键实体与短语的匹配正确率。
- **Grounding 准确率**：给定自然语言指令，模型预测 bbox 与标注 bbox 达到预设重叠阈值（如  $\text{IoU} \geq \tau$ ）且命中目标控件的比例；同时统计在密集小控件场景下的误差分布（偏移量、误命中率）。

其中，IoU（Intersection over Union）用于度量预测框与标注框的空间重叠程度，定义为两者交集面积与并集面积之比：

$$\text{IoU} = \frac{\text{area}(B_{\text{pred}} \cap B_{\text{gt}})}{\text{area}(B_{\text{pred}} \cup B_{\text{gt}})} \in [0, 1].$$

在 GUI 场景中，许多控件（如图标、页面指示器、滑块把手）尺寸较小且相邻密集，轻微的像素级偏移就可能导致 IoU 显著下降。因此本文同时报告不同阈值下的结果：较宽松的  $\text{IoU} \geq 0.1$  更关注“是否命中正确控件区域”，而更严格的  $\text{IoU} \geq 0.5$  则更强调框的几何精确度与边界贴合度。评测样本覆盖 Top 200 应用的高频页面与交互组件，包含中文文本、弹窗浮层与多状态控件等困难模式，以检验模型在真实分辨率与真实布局下的泛化能力。

### 3.4.2 实验结果与分析

为验证多源数据工程的有效性，实验设计主要从**数据维度**进行对比：开源清洗数据、闭源精标数据及其组合（并配合负样本、关系图谱与状态向量等监督信息）。

表3-1从数据维度汇总了不同数据配置下的 Grounding 与 Referring 性能。可以看到，仅使用基线模型时两项指标均较低；引入开源清洗数据后，Grounding F1 从 0.324 提升至 0.557，Referring F1 从 0.523 提升至 0.612；进一步融合闭源精标数据后，Grounding F1 达到 0.846，Referring F1 达到 0.905，验证了多源数据融合策略的有效性。

数据配置	Grounding F1	Referring F1
Baseline	0.324	0.523
Baseline + 开源数据	0.557	0.612
Baseline + 开源 + 闭源	<b>0.846</b>	<b>0.905</b>

表 3-1 不同数据配置下的 Grounding 与 Referring 性能对比。

表3-2展示了模型在各类型 UI 控件上的 Grounding 性能。可以看到，在 IoU 阈值为 0.1 时，整体 F1 达到 0.923；在更严格的  $\text{IoU} \geq 0.5$  阈值下，整体 F1 仍保持 0.846。其中，按钮、输入框、弹窗、开关等高频交互控件的定位精度较高 ( $F1 > 0.94$ )，而页面指示器、滑块等小尺寸或形态多变的控件定位难度较大。

表3-3展示了模型在 Referring 任务上的属性预测性能。整体加权平均 F1 达到 0.905，其中“类型”属性预测最为准确 ( $F1=0.944$ )，“内部注释”（控件内显示文本）的识别也保持较高水平 ( $F1=0.915$ )。相比之下，“当前值”属性的召回

控件类型	数量	F1 (IoU≥0.1)	F1 (IoU≥0.5)
文本	3768	0.945	0.862
图标	2298	0.855	0.737
按钮	1579	0.972	0.947
图片	347	0.850	0.836
导航栏	298	0.983	0.930
选项区	183	0.945	0.863
输入框	167	0.976	0.946
选项框	132	0.871	0.750
弹窗	64	0.969	0.969
开关	38	1.000	0.947
页面指示器	29	0.724	0.483
滚动选择器	26	1.000	0.962
广告	22	0.909	0.864
地图	15	0.933	0.867
滑块	14	0.714	0.500
日期选择器	13	1.000	1.000
多重滚动选择器	12	1.000	1.000
滑块容器区	7	1.000	0.857
通知	6	1.000	0.500
<b>ALL</b>	<b>9018</b>	<b>0.923</b>	<b>0.846</b>

表 3-2 各类型 UI 控件的 Grounding 性能 (F1 Score).

率较低 (0.551)，主要因为该属性仅在滑块、进度条等少数控件上有意义，样本稀疏且形态多样。

属性	实际样本数	预测总数	精确率	召回率	F1
内部注释	5773	5619	0.927	0.903	0.915
图标含义	1624	1485	0.898	0.821	0.858
外部注释	1927	1814	0.831	0.783	0.806
当前值	198	146	0.747	0.551	0.634
是否不可交互	585	529	0.817	0.739	0.776
是否已激活	545	487	0.867	0.774	0.818
类型	10072	9853	0.954	0.934	0.944
<b>加权平均</b>	<b>20724</b>	<b>19933</b>	<b>0.923</b>	<b>0.889</b>	<b>0.905</b>

表 3-3 Referring 任务各属性预测性能.

综合表3-2与表3-3的细粒度统计，可进一步归纳两类主要误差来源：(1) 极密集区域的近邻混淆（图标间距过小导致误点相邻控件）、(2) 文本与控件关系未显式建模时的语义歧义（如“用户名/手机号”标签与输入框的绑定不稳定）。对应地，本文在数据工程侧提供了针对性缓解：通过关系图谱与状态向量为控件间关系与页面状态提供显式监督以降低语义歧义；引入负样本与工具校验减少“看起来像可点但不可点”的误命中；同时以多源数据覆盖更多页面布局与控件形态，提升模型在真实分辨率与真实布局下的泛化能力。

## 3.5 本章小结

本章面向移动端 GUI 这一“控件密集、文字细小、状态易变”的垂直场景，提出了一套提升视觉语言模型屏幕理解能力的系统方案，围绕 Referring 与 Grounding 两项核心能力给出从数据到训练的完整落地路径。本文的主要工作与贡献可概括为：

- **多源数据工程流水线**: 从任务需求出发构建“收集 → 结构化初始化 → 混合标注与属性补全 → 任务/QA 生成 → 质控迭代”的数据流程，融合开源清洗数据 (RICO/CLAY、MobileViews、OS-Atlas)<sup>[22-25]</sup>、闭源精标数据 (v1/v2) 与必要的合成/重写样本，兼顾分布覆盖与质量可控，为 GUI 能力持续迭代提供数据基础。

- **面向可解释语义与可执行定位的标注规范:** 提出细粒度控件分类学（23类核心控件）、有向关系图谱（如标签-控件绑定、值-控件绑定与功能依赖）以及多维属性与状态向量（Activated/Interactable/Filled、图标释义、内部文本等），并通过人机协同与一致性校验降低噪声，显式缓解 GUI 中的语义歧义与“看似可点但不可点”的误监督问题.
- **高分辨率感知与坐标生成训练设计:** 采用 AnyRes 分辨率自适应机制保留全局布局与局部细节，并以文本化坐标与专用 Token 约束 bbox 生成的可解析性；结合 DMT 双阶段混合微调在 GUI 专用能力注入与通用对齐之间取得平衡，缓解能力冲突与灾难性遗忘.
- **系统评测与量化验证:** 在 Top 200 应用的真实页面分布上进行控件级评测，给出不同数据配置与细粒度控件/属性维度的结果统计. 实验表明，引入开源数据与闭源精标数据可显著提升整体性能（Grounding F1 由 0.324 提升至 0.846，Referring F1 由 0.523 提升至 0.905），并揭示了密集小控件近邻混淆与文本绑定歧义等关键误差模式，为后续 UI Agent 的任务规划与执行奠定基础.



# 第 4 章 世界知识能力提升方案

## 4.1 背景和挑战

近年来，大型视觉语言模型（Vision-Language Model, VLM）通过在海量图文对上的预训练，习得了强大的视觉感知与语言生成能力。然而，这类能力往往停留在”**视觉层面的描述**”，即对颜色、形状、人物姿态等可见信息进行总结，难以进一步回答”**图像背后隐含的事实、概念与关系**”。

本文将**世界知识**（World Knowledge）定义为对物理世界、社会文化、历史事件、科学技术等领域的事实、概念、关系与常识的综合掌握，其本质是一个庞大且交织的知识网络。图4-1展示了一个直观的对比案例：对于经典电影《泰坦尼克号》的标志性场景，缺乏世界知识的模型仅能给出浅层描述——”一个男人和一个女人站在船头”；而具备世界知识的模型则能识别出这是 1997 年上映的电影《泰坦尼克号》，由詹姆斯·卡梅隆执导，主演是莱昂纳多·迪卡普里奥和凯特·温斯莱特，并进一步理解这一场景描绘了主角杰克和露丝在船头”飞翔”的浪漫时刻，象征着自由和爱情，以及这部电影在全球取得的巨大商业成功和文化影响力。这种从”看见”到”理解”的跃迁，正是世界知识赋予模型的核心能力。

然而，提升 VLM 的世界知识能力并非简单”增加数据量”即可解决，世界知识注入在实践中往往同时受到**三大困境**的制约：

- **知识整合与收集 (Integration)**：目前的世界知识通常存在于庞大的文本语料库或结构化的知识图谱中，知识源虽然丰富，但要从中抽取并组织成**可被模型学习**的训练样本不容易。其难点体现在：知识高度分散且表达形式异构（不同来源对同一实体的称呼、粒度与侧重点不一致），同时存在噪声与时效性差异；此外，直接“搬运知识”难以保证与视觉内容可验证对应，导致构造出的样本在训练时容易引入错误监督或不可学习信号。
- **视觉与语言对齐 (Alignment)**：大语言模型往往已经具备一定的世界知识能力，但视觉侧未必能将图像中的**具体实例 (Specific Instance)**与文本知识中的**抽象描述 (Abstract Description)**精确关联。例如，图像中出现一只特定的鸟，文本知识可能包含该鸟类的学名、习性与分布范围，但模型难以将这些抽象属性稳健地“落到”当前图像实例上，从而出现“知道但



图 4-1 有无世界知识的模型响应对比. 无世界知识：“一个男人和一个女人站在船头.”有世界知识：“这是 1997 年上映的电影《泰坦尼克号》的经典海报，由詹姆斯·卡梅隆执导，主演是莱昂纳多·迪卡普里奥和凯特·温斯莱特……”

看不出来”或“看出来但说不对”的错配. 更进一步，同一实体在不同视角、光照、遮挡或风格化图像中呈现差异显著，也会放大对齐难度并影响泛化能力.

- **多跳推理能力 (Reasoning)**: 许多复杂世界知识问题需要模型进行多步推理，将多个知识点链接后再得出结论. 例如，回答“图中这位导演拍摄的、获得奥斯卡最佳影片的电影是哪部？”，模型需要先识别图中人物为某位导演，再检索/回忆其作品列表，最后按“获奥斯卡最佳影片”这一条件进行筛选得到唯一答案. 该过程要求模型同时具备稳定的实体识别、可控的知识调用以及可执行的条件推断链路，对当前 VLM 仍是显著挑战.

## 4.2 知识类别框架

世界知识增强的首要前提，是建立一个可覆盖、可扩展、可操作的知识类别框架. 若缺乏统一的类别坐标系，世界知识将以“零散事实”的形式分布在不同来源中，不仅难以进行系统性采样与补全，也难以在训练与评测中对模型能力进行可解释的归因分析. 因此，本文构建分层分类体系，包含 7 个大类与 40 个小类（如图4-2所示），用于约束数据构造过程并为后续框架化的数据流水线提供统一接口.

类别制定的作用与必要性主要体现在三方面：

1. **指导词条生成与采样**: 在每个子类的定义约束下，词条生成能够被限定在明确语义边界内，避免“概念混淆”（如将作品名误作人物名）与“粒度漂



图 4-2 世界知识类别框架 (7 个大类与 40 个小类).

- 移”（如将系列与单部作品混用），从而提升词条集合的可控性与可复用性.
2. 支持长尾覆盖与迭代补全：按子类统计实体与样本分布，可显式暴露稀缺类别与薄弱子类，便于后续定向补齐与均衡采样.
  3. 作为训练/评测的统计维度：类别标签为分析不同知识域的对齐难度、误差模式与泛化瓶颈提供了可解释的分组依据.

为保证后续“词条生成 → 图片收集 → 标注与 QA 构造”的一致性，本文为每个大类与子类均给出可操作的边界定义，定义内容包括：**覆盖范围**（该类包含哪些实体/概念）、**排除规则**（易混淆但不属于该类的情况）、**命名规范**（中英文/别名/译名的处理方式）以及**典型示例**（用于提示生成与校验）。例如，**建筑景观/地标建筑**强调“可定位的实体建筑与场所”，并排除“抽象建筑风格”；**文化娱乐/电影**以“独立作品”为基本粒度，区分系列名与单部片名；**品牌商标/商业 Logo**强调“可识别的品牌标记”，并区分公司主体与产品线名称。

在该框架下，代表性大类如下：

- **自然景观**: 自然形成的地理实体与景观要素, 强调可定位与可辨识性(如河流、山脉、森林、沙漠等).
- **生物**: 具有明确物种/类群归属的生命体实体, 强调可用学名/别名统一归并(如知名动物、植物、真菌等).
- **建筑景观**: 人造建成环境中的可识别实体, 强调“具体建筑/场所”而非抽象风格(如地标建筑、宗教场所等).
- **人物**: 具有可追溯身份的真实人物实体, 强调同名消歧与多语别名对齐(如政治人物、演艺明星、历史名人等).
- **品牌商标**: 可被视觉识别的品牌标记与商业符号, 强调品牌与产品的边界区分(如商业Logo、产品包装等).
- **文化娱乐**: 文化作品及其衍生实体, 强调作品粒度与系列关系的规范化(如电影、游戏、动漫、书籍等).
- **科学技术**: 科技相关的可识别对象或现象, 强调术语标准化与概念边界(如特定设备、科学现象等).

该框架为后续章节中的词条生成、图片收集与 QA 构造提供了稳定、可复用的“坐标系”, 使世界知识数据构造从零散经验转为可控流程, 并为后续框架的建立与扩展提供结构化支撑.

## 4.3 数据工程: 世界知识数据构造 Pipeline

世界知识增强的关键在于“把分散知识变成可学习的多模态监督”. 本文设计了如图4-3所示的世界知识数据构造 Pipeline, 该流程包含词条生成与图片收集、图片 Caption 生成、图文相关性过滤、QA 对生成与 QA 对质量过滤等核心环节, 形成从知识源到可训练数据的完整转化路径.

### 4.3.1 词条生成与图片收集: 从知识域到视觉证据

对应图4-3中的“词条生成和图片收集”环节, 本文首先在前述知识类别框架的约束下, 为每个子类生成可控的词条集合, 并围绕词条自动化收集图片, 以形成“实体 $\leftrightarrow$ 视觉证据”的锚点, 为后续图文对齐与知识学习打下基础:

- **类别制定**: 基于前述分类体系的定义边界, 确定每个子类的覆盖范围与排除规则, 从源头保证后续词条生成与样本采样的语义一致性与可追溯性.

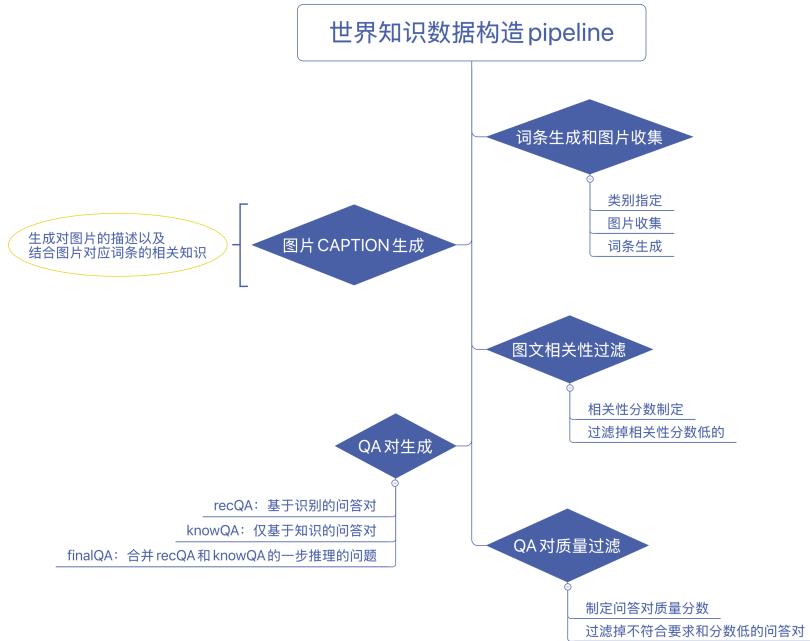


图 4-3 世界知识数据构造 Pipeline.

- **词条生成**: 词条生成采用两种互补方式. 其一, 使用大模型依据类别/子类定义生成该子类中“最热门、最具公众关注度”的核心词条（通常每类约 150~200 个），以覆盖主流高频知识；其二，人工补充部分新近出现或相对冷门但具有研究价值的词条，以提升长尾覆盖并缓解仅学习头部概念带来的偏置.
- **图片收集**: 针对每个词条，本文实现了一个自动化的图片收集 agent 脚本，可调用 Google 图片搜索并自动抓取该词条返回结果中的前 100~500 张图片. 该步骤将抽象词条落到具体视觉实例上，为后续“图像  $\leftrightarrow$  词条  $\leftrightarrow$  知识”对齐提供充分样本基础；同时保留来源、分辨率、时间等元信息，以支持去重与质量追踪.

在这一阶段，数据工程的目标是”实体可控”与”图像高质量”：实体可控保证知识监督可追溯，图像高质量保证视觉侧对齐不被噪声主导.

### 4.3.2 图片 Caption 生成：融合视觉描述与知识背景

对应图4-3中的”图片 CAPTION 生成”环节，本文为每张自动抓取的图片生成结构化 caption，其作用不仅是“生成更丰富的描述”，更是为后续多个关键环节提供可依赖的中间表征：(1) **图文相关性过滤**需要 caption 帮助模型判断图片是否真的与词条匹配；(2) **知识要素提取**可从 caption 中抽取实体属性、别名与背景事实；(3) **QA 对生成**可将 caption 作为问题构造与答案验证的上下文依据.



图4-4 Caption 数据示例（电影《1917》相关画面）.

尤其在相关性过滤场景下，由于图片是自动抓取的，实际会混入与词条无关的图片（例如同名歧义、检索噪声、相似外观实体等）。仅依赖视觉模型直接判别相关性容易受限于大模型自身的识别能力与先验偏差，因此本文将 caption 与词条知识一并提供给过滤模型作为判据，从而提升过滤的稳健性与可解释性。

图4-4给出了caption数据的一个示例。可以看到，caption不仅描述了画面中的关键视觉元素（两名士兵、英军制服、草地与田野等），还补充了与该视觉内容强相关的背景知识（电影名称、上映年份、导演、拍摄手法与剧情要点）。这类“视觉描述 + 背景知识”的融合表达，一方面为后续相关性过滤提供了更可验证的判断依据，另一方面也为知识要素提取与QA构造提供了可复用的文本上下文。

**PROMPT:** 请详细观察这张图片，描述其中的主要内容，并提供相关的背景知识。

**RESPONSE:** 这张图片来自电影《1917》(2019)。在图片中可以看到两名士兵穿着英军制服，躺在草地上。背景是一片开阔的田野，远处有一些模糊的植被和空旷的天空。

《1917》是一部关于第一次世界大战的战争片，由萨姆·门德斯执导。这部电

影以其独特的拍摄手法闻名，采用了一种类似单镜头的方式，使观众感觉仿佛是在实时跟随士兵的旅程。影片的情节围绕着两名年轻的英国士兵，他们接到一项看似不可能完成的任务：穿越敌方领土，传递一条至关重要的信息，以阻止一场可能导致 1600 人阵亡的伏击。

影片的视觉效果和沉浸式体验获得了极高的评价，尤其是在构建紧张感和真实感方面。电影中展现了战争的残酷和士兵们在战场上的勇气与牺牲，获得了很多观众和评论家的好评。

### 4.3.3 图文相关性过滤：保证视觉-知识对齐质量

对应图4-3中的”图文相关性过滤”环节，本文通过以下方式保证数据质量：

- **相关性分数制定**：采用大模型进行判别，并设置 1~5 的离散相关性分数以刻画“图片  $\leftrightarrow$  词条”的匹配强度：5 分表示完全相关（图片直接包含标签内容），4 分表示高度相关（作品/组成部分等紧密关系），3 分表示中等相关（可联想但不直接对应），2 分表示低度相关（需较强推理才能建立联系），1 分表示完全不相关。模型在输出分数时同时给出简要理由，并采用固定格式（如“相关性分数:[分数]”）便于程序解析与自动化处理。
- **阈值化过滤与尽可能保留**：考虑到大模型判别仍存在误差，本文并非“一刀切”地仅保留最高分样本，而是通过设定分数阈值与抽样复核策略，尽可能保留潜在相关图片（如 4 分、部分 3 分样本），以在覆盖率与噪声控制之间取得平衡。过滤后的样本将进入后续 QA 构造，从源头降低“错图错标”导致的对齐噪声。

### 4.3.4 QA 对生成：从描述到三元组任务

对应图4-3中的”QA 对生成”环节，为了解耦感知与推理，本文构造三类互补问答对：

- **RecQA (Recognition QA)**：对应“模型看见图片”的能力，问题围绕“这是谁/这是什么/来自哪里”等识别任务，答案为实体或关键属性。其目标是将视觉特征稳定锚定到词条实体，解决“看到但叫不出名字”的问题。
- **KnowQA (Knowledge QA)**：对应“模型知道关于该实体的世界知识”的能力，问题强调外部知识（历史、文化、科学属性等），形式上可不依赖视觉信息，但与图片实体强相关。其目标是向模型注入可泛化的背景事实，解决“认识了但不了解”的问题。
- **FinalQA (Reasoning QA)**：对应“模型理解图片并能调用知识推理”的能力，问题需要“先识别再用知识推理”，可是一跳或多跳链路（如识别实体



图 4-5 三类 QA 与 FinalQA with CoT 示例 (格鲁/Gru).

→ 回忆知识 → 条件筛选). 其目标是打通 RecQA 与 KnowQA 形成的知识链路，推动模型从“描述/识别”走向“解释/推断”.

三类 QA 分别对应不同能力层次：RecQA 偏感知/识别，KnowQA 偏理解/知识，FinalQA 偏推理/筛选. 这种拆分避免模型直接学习“图像 → 最终答案”的捷径，从而提升泛化与可解释性.

图4-5展示了三类 QA 在同一实体上的组织方式. RecQA 首先要求模型从图像中识别出“戴着条纹围巾的光头男性角色”为格鲁 (Gru)；KnowQA 进一步询问与该实体相关、但必须依赖外部知识才能回答的问题（例如《神偷奶爸》系列的制作公司）；FinalQA 则将二者合并为一步推理问题，使模型必须先完成识别，再调用知识得到最终答案. 此外，为了促进可解释推理与避免“直接背答案”，本文还构造带有简要推理过程的 FinalQA (FinalQA with CoT)，要求模型先给出识别结论，再给出问题答案.

#### RecQA (基于识别的回答对)

**PROMPT:** 图片中戴着条纹围巾的光头男性角色是谁?

**RESPONSE:** 格鲁 (Gru).

#### KnowQA (仅基于知识的回答对)

**PROMPT:** 格鲁 (Gru) 所在的《神偷奶爸》系列电影是由哪家公司制作的?

**RESPONSE:** 照明娱乐 (Illumination).

#### FinalQA (合并 RecQA 与 KnowQA 的一步推理问题)

**PROMPT:** 图片中戴着条纹围巾的光头男性角色所在的系列电影是由哪家公司制作的?

**RESPONSE:** 照明娱乐 (Illumination).

**FinalQA with CoT (带简要推理过程)**

**PROMPT:** 图片中戴着条纹围巾的光头男性角色是照明娱乐 (Illumination) 制作的哪部动画电影系列的主角?

**RESPONSE:** 图片中戴着条纹围巾的光头男性角色是格鲁 (Gru)，他是动画电影系列《神偷奶爸》的主角.

### 4.3.5 QA 对质量过滤：确保训练数据可靠性

对应图4-3中的”QA 对质量过滤”环节，本文采用多级质控保证知识可靠与对齐有效：

- **规则约束与质量分数制定：**本文使用大模型对“问题-答案”对进行规则校验，重点约束：问题必须同时依赖图像信息与外部知识（既不能纯靠知识回答，也不能纯靠图像即可回答）；问题应当客观、明确且无歧义；答案必须唯一且不随时间变化；避免一个问题中同时询问两个子问题等。基于满足程度给出 1~3 的质量分数，其中 3 分表示完全合格，2 分表示勉强合格但必须满足“唯一答案/时效稳定/单问题”这些关键约束，1 分表示不合格。模型输出时给出简要理由，并采用固定格式（如“合格分数:[分数]”）便于解析。
- **低质量过滤与回收：**对低分样本进行过滤或回收重写，重点剔除多答案、表述模糊、主观化或时效性强的问题，从而保证训练信号的可学习性与评测的一致性。

## 4.4 训练与实验：从数据类型消融到 CoT 对齐

本节将训练策略与实验分析合并阐述：首先通过消融实验验证不同数据类型对世界知识能力的影响；随后针对推理指标“增益不显著甚至下降”的现象，引入带推理过程的 Think/CoT 数据并扩充数据分布，给出改进后的结果；最后报告在 **Chinese SimpleVQA** 与 **SimpleVQA** 上的最终性能，并分析与强基线之间的差距。

### 4.4.1 多阶段、多来源数据构建

训练过程中，本文采用多批次策略逐步扩充数据分布：

- **Batch 1 / Part1：**使用最强模型构建第一批数据 (Part1)，确保高精度与高一致性；

- **Batch 2 及后续 / Part12, Part123:** 引入不同模型与更多未见词条扩充数据分布，降低单一模型偏置，并提升长尾覆盖.

#### 4.4.2 实验一：数据类型消融与冲突分析

本实验的目的在于确定不同数据类型对世界知识能力是否具有正向影响，并定位可能的任务冲突。评测指标包括识别侧（Correct、Accuracy、F1）与推理侧（Correct、Accuracy、F1），分别对应 Recognition 与 Final 两类任务。

实验结果如表4-1所示，可以得到以下结论：

- **与强基线仍有差距：**尽管数据增强带来提升，但与 qwen2.5-VL-7b 相比<sup>[29]</sup>，各项指标仍存在明显差距，说明世界知识能力提升仍受限于模型容量、数据覆盖与对齐质量等因素。
- **Image Caption 对识别显著有益：**加入 caption 类数据能为识别任务提供更稳定的视觉-文本锚点，从而对 recognition 相关指标产生正向提升。
- **FinalQA 对推理提升不显著甚至下降：**直接加入 FinalQA（尤其是仅监督最终答案的 Direct Answer 形式）并未稳定提升 final 指标，部分设置下甚至下降。
- **FinalQA 与 RecQA 存在冲突：**当 FinalQA 强调“直达最终答案”时，模型容易学习“图像特征 → 答案”的短路映射，反而削弱 RecQA 对实体识别与稳健锚定的学习；同时识别不稳又会反过来限制 FinalQA 的推理上限，导致二者在训练中互相掣肘。

Model	Correct_Rec	Acc_Rec	F1_Rec	Correct_Final	Acc_Final	F1_Final
qwen2.5-VL-7b	41.3	44.6	42.9	39.5	42.1	40.8
Ovis2.5_9B_base	27.9	28.3	28.1	36.8	<u>37.2</u>	37.0
Ovis2.5_9B_part1_all_e1	<u>32.9</u>	<u>33.3</u>	<u>33.1</u>	36.3	36.3	36.3
Ovis2.5_9B_part1_all_e2	32.2	32.8	32.5	36.2	36.2	36.2
Ovis2.5_9B_part1_RKFQA_e1	31.2	31.6	31.4	<u>37.1</u>	37.1	<u>37.1</u>
Ovis2.5_9B_part1_RKQA_e1	31.6	32.0	31.8	35.2	35.3	35.2

表 4-1 实验一：Part1 数据与不同数据类型配置的消融结果（Chinese SimpleVQA）。Rec 表示 Recognition；Acc 表示 Accuracy；Part1 表示第一批数据；all 表示全部类型数据；e 表示训练轮次；RKF 分别表示 RecQA、KnowQA 与 FinalQA。

### 4.4.3 实验二：引入 Think/CoT 与数据扩充的改进效果

由于实验一中推理指标不升反降，我们进一步验证推理过程监督的作用。核心观察是：当 FinalQA 仅监督最终答案时，模型更容易“记答案”；而当 FinalQA 要求输出简要推理过程（Think/CoT）<sup>[15]</sup>，模型被迫先完成识别再调用知识，从而更有效地提升 final 相关指标。

基于这一发现，我们采取如下改进思路并在表4-2中给出结果：

- **保留所有类型数据**：同时保留 caption、RecQA、KnowQA、FinalQA 等多种数据，以维持“视觉锚定 + 知识注入 + 推理贯通”的能力闭环；
- **扩充未见词条**：继续扩充模型未见过的新词条与长尾实体，提升泛化能力与覆盖面；
- **FinalQA 增加 CoT 过程**：将 FinalQA 从 Direct Answer 升级为带简要推理过程的形式，使模型学习“先识别再推理”的过程性能力；
- **增加 KnowQA 与 FinalQA 密度**：同一张图片可生成多个 KnowQA 与 FinalQA 样本，提高知识链路覆盖与训练信号密度。

Model	Correct_Rec	Acc_Rec	F1_Rec	Correct_Final	Acc_Final	F1_Final
qwen2.5-VL-7b	41.3	44.6	42.9	39.5	42.1	40.8
Ovis2.5_9B_part12	33.9	34.3	34.1	36.6	36.7	36.7
Ovis2.5_9B_think_part12	34.4	37.6	35.9	38.6	41.3	39.9
Ovis2.5_9B_think_part123	<b>57.5</b>	<b>60.7</b>	<b>59.0</b>	<b>51.3</b>	<b>54.2</b>	<b>52.7</b>
<b>Gap</b>	+16.2	+16.1	+16.1	+11.8	+12.1	+11.9

表 4-2 实验二：引入 Think/CoT 与数据扩充后的改进结果（Chinese SimpleVQA）。Part12/Part123 表示逐步扩充的数据批次；think 表示 FinalQA 带推理过程的训练设置；Gap 为最优模型相对 qwen2.5-VL-7b 的提升。

从表4-2可见，Think/CoT 对 final 相关指标提升显著，且在进一步扩充数据分布（Part123）后，recognition 与 final 均获得同步大幅提升。这说明：世界知识能力的提升不仅依赖数据规模与覆盖，更依赖训练信号的结构——显式的推理过程能够促使模型学习可迁移的推理策略，而非在训练分布上“背答案”。

### 4.4.4 最终结果：Chinese SimpleVQA 与 SimpleVQA

在完成上述训练与数据构造策略后，一共产出了 80w 条数据，在 Chinese SimpleVQA 与 SimpleVQA 上的最终结果，分别见表4-3与表4-4。总体上，Ovis2.5\_9B

在中文基准上对 recognition 指标达到并超过部分基线，但在 final 推理指标上与更大模型（如 Qwen2.5-VL-72B）仍存在差距，说明后续仍可从模型容量、检索增强或更高质量推理数据等方向继续提升。

Model	Correct_Rec	Acc_Rec	F1_Rec	Correct_Final	Acc_Final	F1_Final
qwen2.5-VL-7b	41.3	44.6	42.9	39.5	42.1	40.8
Qwen2.5-VL-72B	45.7	48.5	47.1	<b>49.0</b>	<b>53.4</b>	<b>51.1</b>
Ovis2.5_9B (Ours)	<b>47.4</b>	<b>53.6</b>	<b>50.3</b>	48.1	52.1	50.0
Gap	+6.1	+9.0	+7.4	+8.6	+10.0	+9.2

表 4-3 最终结果：Chinese SimpleVQA. Gap 为 Ovis2.5\_9B 相对 qwen2.5-VL-7b 的增益。

Model	Correct	Accuracy	F1
qwen2.5-VL-7b	43.2	45.6	44.3
Qwen2.5-VL-72B	<b>49.4</b>	52.2	<b>50.8</b>
Ovis2.5_9B (Ours)	45.8	<b>52.8</b>	49.0
Gap	+2.6	+7.2	+4.7

表 4-4 最终结果：SimpleVQA. Gap 为 Ovis2.5\_9B 相对 qwen2.5-VL-7b 的增益。

## 4.5 本章小结

本章围绕视觉语言模型在世界知识理解上的不足，系统构建了面向世界知识增强的数据体系与训练范式，为后续统一框架的提出提供了可复用的工程化基础。本文的主要贡献与创新点可概括为：

- **提出可操作的世界知识类别框架：**构建 7 个大类、40 个子类的分层分类体系，并强调每个类别/子类的覆盖范围、排除规则与命名规范，从而将“世界知识”从零散事实组织为可采样、可统计、可扩展的结构化坐标系，直接服务于后续词条生成与长尾覆盖。
- **给出端到端的世界知识数据构造 Pipeline：**围绕“词条生成 → 自动化图片收集 → caption 生成 → 相关性过滤 → QA 对生成 → 质量过滤”建立稳定流水线。尤其通过 caption 作为中间表征支撑图文相关性判别，并引入 1~5 分相关性评分机制，在覆盖率与噪声控制之间取得更稳健的平衡。
- **设计 RecQA/KnowQA/FinalQA 的三元组任务分解并引入 CoT 监督：**从“看见 → 知道 → 理解并推理”的能力递进出发，将感知、知识与推理解耦

为三类互补监督信号，避免模型学习“图像 → 答案”的捷径；进一步将 FinalQA 升级为带简要推理过程的 FinalQA with CoT，使模型学习可迁移的推理链路而非记忆答案。

- **通过消融实验揭示冲突并验证改进有效性：**实验表明 caption 对识别指标具有稳定增益，而仅提供 Direct Answer 形式的 FinalQA 对推理指标提升不显著甚至下降，反映出 FinalQA 与 RecQA 在训练中存在短路映射带来的冲突；引入 Think/CoT 并扩充未见词条与多问一图数据后，final 相关指标得到显著改善，并在 Chinese SimpleVQA 与 SimpleVQA 上取得具有竞争力的最终结果。



# 第5章 多模态数据构造范式

## 5.1 背景

当前多模态大语言模型（Multimodal Large Language Model, MLLM）的训练正面临“**数据质量而非数据数量**”成为主要瓶颈的阶段性转折。一方面，网络抓取的图文对存在噪声大、对齐弱（Misalignment）、图像质量参差不齐（水印、模糊）等问题；另一方面，人工标注成本高且难以覆盖复杂推理与多样化交互格式。已有研究（如 ShareGPT4V、MMEvol、SynthVLM）均表明：**高信息密度、高对齐度与高推理密度的数据**，往往比盲目扩大规模更能有效提升模型能力<sup>[1-3]</sup>。

为系统性描述并复用高质量数据生产经验，本文提出 **Auto-Evol** 多模态合成数据构造范式：将“**合成数据生成**”从单步的“图 → 文/问答”重构为一个包含**正反双向链路、任务导向的原子问题生成、代理式生成与进化增强、多维校验与闭环反馈**的工程系统。该范式可统一承载不同任务域的数据生产需求：例如第3章面向 GUI 的 Referring/Grounding 能力提升与第4章面向世界知识的 RecQA/KnowQA/FinalQA 体系，均可视为 Auto-Evol 在不同任务路由下的具体实例。更进一步，本章并不将范式作为“抽象概念”孤立呈现，而是以第3章与第4章的数据构造方法为起点，结合 ShareGPT4V、MMEvol、SynthVLM、Synthesize Step-by-Step 等工作的可复用组件<sup>[1-3,21]</sup>，总结得到可落地、可迭代、可验证的多模态数据构造方法学。

## 5.2 构造范式框架

图5-1给出了 Auto-Evol 范式的闭环框架。该框架从数据源侧同时接入**原生图片与生成图片**两类输入，通过数据初始化统一格式并沉淀生成所需的源信息（可包含自动/人工标注产物），随后进行质量过滤；再将样本路由到“**任务导向的原子问题生成**”，并按感知、理解、推理三类代理生成多任务指令；最后通过数据增强与数据校验提升多样性、难度与正确性，若校验不过则回流重写/重生成，最终产出可用于监督微调（SFT）与后训练（Post-training）的高质量合成数据。

Auto-Evol 的构造流程可逐模块映射到图5-1的闭环节点。下文结合第3章（GUI）与第4章（世界知识）的实践，依次讲解各模块的作用。

### 5.2.1 构造流程

**(1) 数据源：原生图片与生成图片的双向数据流** Pipeline 顶部同时接入两类数据源：

- **原生图片（真实收集）**：来自真实应用截图、公开数据集与搜索收集的高清图片。GUI 章使用真实设备分辨率截图以保留小控件细节；世界知识章以“词条 → 图片”的方式确保实体可控。
- **生成图片（Text-to-Image）**：当长尾场景稀缺或需要严格对齐时，可采用“文本 → 图像”的逆向数据工程。SynthVLM 表明，先清洗 caption 再用扩散模型生成高分辨率图像，并用 CLIPScore+SSIM 筛选，可从源头提高图文对齐度与图像清晰度，从而以更少数据达到更强效果<sup>[3,31–32]</sup>。

双向数据流的价值在于：**正向流（图 → 文）**擅长覆盖真实分布，**反向流（文 → 图）**擅长制造“强对齐、可控分布”，两者互补可显著提升数据的覆盖与可信度。

**(2) 数据初始化：统一格式与任务路由的前置条件** “数据初始化”的核心目标是将不同来源的原始材料整理为后续**指令/QA 生成阶段**可直接提取或参考的**源信息池**：包括可追溯的元信息（来源、分辨率、类别/词条等）、必要的路由标签，以及与任务相关、可被复用的结构化线索。这里的“源信息”不要求跨数据源完全同构，而是强调**为生成提供可依赖的参照物**。在实践中，初始化阶段通常会沉淀两类关键产物：

- **自动生成的结构化描述/要素（用于蒸馏的中间表征）**：以世界知识数据为例，初始化阶段往往会先为图片生成结构化 caption，并显式区分“画面可见信息”（主体、属性、关系、时空线索等）与“背景知识要点”（年份、导演、题材、所属系列等可核验事实）。该 caption 随后既可用于图文相关性过滤，也可作为知识要素抽取与 RecQA/KnowQA/FinalQA 构造时的主要上下文依据，从而把“图像与知识源”转化为可复用、可校验的中间表征。与此同时，为提升大模型蒸馏的可靠性，caption/要素抽取常配套引入：固定输出格式与字段约束（例如先给出实体识别结论，再给出若干可核验事实）、自一致性采样/多次生成投票、以及“生成–判别”式复核（judge 过滤不一致或缺乏证据支撑的内容），将不确定与幻觉样本在数据侧前置淘汰。以词条“1917（电影）”为例，caption 会被要求同时给出图中“士兵/军装/战场环境”等可见证据与“2019、萨姆·门德斯、战争片”等知识要点；随后系统依据 caption 与词条的一致性保留高相关样本，并据此生成识别锚定的 RecQA 与需要知识调用/筛选的 KnowQA、FinalQA。

- **人工标注与上下文采集（高置信锚点）**：人工标注适用于需要像素级精确度、存在强歧义/同名消歧、或需要遵循严格规则与一致性约束的场景，其作用是为后续生成提供稳定锚点与可复核依据。以 GUI 任务为例，可沉淀候选控件集合、紧致 bbox、OCR 文本、控件可交互性/状态（如 Activated/Interactable/Filled）以及标签-控件绑定等关系信息，从而支撑 Grounding/Referring 等任务的可执行监督；以世界知识任务为例，可对词条集合进行边界校正与别名对齐、处理同名实体歧义，并对关键样本进行抽检复核，确保“图片  $\leftrightarrow$  词条  $\leftrightarrow$  知识”的对应关系可追溯、可验证。

**(3) 过滤：相关性与准确性的第一道闸门** “过滤相关性、准确性”是合成数据可用性的核心保障，其目标是尽早剔除“错图错标/弱对齐/不可验证”的样本，避免噪声在后续指令生成与训练中被放大。以第4章为例，图片来自自动检索，天然会混入同名歧义、相似实体、无关插图等噪声；因此在进入 QA 构造前，系统会基于“图片  $\leftrightarrow$  词条”的匹配强度进行**相关性评分**（1–5 分），并要求输出简要理由与固定格式，便于程序解析与自动化处理。在阈值化过滤的同时，为兼顾覆盖率与长尾，流程通常会保留高分样本并对部分中等分样本进行抽样复核，从而在“尽可能保留潜在相关证据”与“抑制错配监督”之间取得平衡。

与之相比，SynthVLM 更偏向使用**自动指标**（如 CLIPScore、SSIM 等）对图文对齐度与图像质量进行强筛选<sup>[3,31–32]</sup>，其优势在于可规模化、可重复；而世界知识任务中的相关性过滤更强调**语义一致性与可解释性**（尤其是同名消歧与实体锚定）。两者共同指向同一结论：过滤不仅是去噪，更是**控制训练信号可信度与信息密度的关键环节**。

**(4) 任务导向的原子问题生成（三基础能力/三生成链路）** 在 Auto-Evol 中，“**把任意多模态任务统一分解为感知、理解、推理三类基础能力，并为每一类能力设计对应的数据生成链路与监督信号**”。因此，三代理生成不仅是工程上的分工，更是多模态模型能力结构的抽象——其他任务数据（如 GUI 定位与交互、世界知识识别与问答、多轮对话、工具调用等）都可以视为三类能力的组合与不同权重的路由结果。

具体而言，Pipeline 中的三条生成链路可概括为：

- **感知链路（Perception）**：以图像为主要证据，生成可直接对齐到视觉区域/实体的监督信号（如 bbox、OCR 文本、实体识别锚定）。例如在 GUI 中，感知链路产出紧致 bbox 与文本对齐，使模型学会在密集小控件场景下稳定“指哪打哪”；在世界知识中，感知链路对应 RecQA：先把图片中的具体实例可靠地锚定到词条实体，解决“看见但叫不出名字/叫错名字”的问

题.

- **理解链路 (Understanding)**: 在感知锚点基础上，生成高密度语义解释与可复用上下文（如结构化 caption、关系解释、背景知识片段），提升语义覆盖与可解释性。以世界知识为例，初始化阶段沉淀的结构化 caption 把“画面可见信息”与“背景知识要点”组织成稳定上下文，理解链路据此构造 KnowQA，将可泛化的事实知识与实体绑定；以 GUI 为例，理解链路对应 Referring 类数据：给定 bbox 输出控件功能/文本/属性及其与周边元素的关系，从而让模型学会“看懂”而不仅是“定位”。
- **推理链路 (Reasoning)**: 在“感知锚定 + 语义/知识上下文”之上，生成需要多步条件推断的监督信号，并尽可能加入过程性约束以避免短路学习。以世界知识为例，FinalQA 要求模型先识别实体再调用知识完成筛选/推断，并在加入 Think/CoT 后显著提升推理指标，说明推理链路必须提供“先识别再推理”的过程监督；以 GUI 为例，推理链路对应包含多步计划与工具调用约束的任务数据，使模型学习在页面状态、控件关系与目标约束下生成可执行的动作序列。

通过上述三链路路由，Auto-Evol 把“难任务”拆成“可控生成、可控校验”的组合单元：感知链路提供可对齐的锚点，理解链路提供可复用的语义/知识上下文，推理链路提供过程性约束与可验证推断，从而系统性降低错配与幻觉传播，并提升数据的可迁移性与覆盖度。

**(5) 数据增强：细粒度、难度、指令多样性的系统注入** Pipeline 中的“数据增强”可理解为对训练信号的三条系统性“加密”方向，它们既对应数据侧要解决的三类瓶颈，也对应模型希望被强化能力，即关注细节、攻克难题、与人交互的能力：

- **细粒度增强 (对齐粒度 → 感知能力)**: 针对多模态数据常见的“对齐粒度不够/忽略长尾细节”问题，围绕图像中的次要物体、背景细节与空间关系补充监督，使模型学习到更细的视觉证据与文本描述之间的对应关系。MMEvol 的 *Fine-grained Perception Evolution* 提供了一个可操作范式<sup>[2]</sup>：利用检测框等视觉约束，将问题生成从“主目标描述”扩展到“被忽略区域/长尾物体”的定向提问，从而减少“看见但不说/说错”的幻觉风险。对应到本文任务中，GUI 场景的小控件与小字体天然属于细粒度信息，补充紧致 bbox 与 OCR 对齐监督，本质上就是对感知链路的信息密度加密。

- **难度增强（推理密度 → 推理能力）**: 针对“指令复杂性不足/缺乏深度推理步骤”的问题，通过多步推理、条件约束与可执行步骤显式提升推理密度。例如可以将推理抽象为“视觉操作链”（如定位、OCR、存在性判断、计数/计算等原子操作），并要求在输出答案前生成可追踪的推理步骤，从而把“难题”转化为可校验的过程监督。本文在世界知识任务中对 FinalQA 加入 Think/CoT（要求先识别实体，再调用知识完成筛选/推断），并在实验中观察到推理指标显著提升，正体现了“过程性推理监督”对推理能力增强的关键作用。
- **指令多样性增强（交互覆盖 → 理解与指令遵循能力）**: 针对“指令形式单一/交互覆盖不足”的问题，将同一语义目标改写为多样化的交互形式（如结构化 JSON、代码片段、角色设定、多轮对话等），以提升模型对不同用户表达与输出约束的鲁棒性。MMEvol 的 *Interaction Evolution* 强调在保持视觉约束的前提下扩展输出格式与交互方式<sup>[2]</sup>，避免仅做“文本复杂化”而脱离图像证据；在 Auto-Evol 中，该方向对应理解链路的格式多样化注入，使模型既学会“说对”，也学会“按要求说”。

**(6) 数据校验：LLM-as-a-Judge + 工具/执行校验的多维验证** 高质量合成数据必须“可验证”。更准确地说，**数据校验就是对数据做评测**：围绕“我们希望模型最终具备什么能力”，反向定义数据应满足的作用与特性，并在生成后用可操作的评测准则筛选与修正数据分布。总之想要什么样的数据，就要用对应方法校验才能得到，例如这三类比较普遍的校验：

- **难度校验（过滤“过于简单”的样本）**: 若目标是让模型解决难题，则需要避免训练集中被大量“简单题”稀释。实践中可用**能力较弱的模型或规则基线先行作答**：若弱模型在不依赖关键视觉/知识证据的情况下即可轻易答对，则该样本更可能属于低难度或存在捷径，可降权、剔除或重写以提升推理密度。
- **正确性与可验证性校验（确保监督信号“对”）**: 若目标是让模型切实学会某项能力，则监督必须可靠。可用**更强的模型进行一致性复核**（例如答案唯一性、推理链条自洽、与图像证据相符），并对推理类样本引入**工具/执行校验**（如代码执行、规则验证、可解析约束检查）来验证最终答案与中间步骤的一致性，从源头抑制错误监督与幻觉传播。
- **覆盖与多样性校验（保证可泛化的交互学习）**: 若目标是提升交互与指令遵循鲁棒性，则数据需要在任务类型、表达方式与输出格式上足够多样。可

通过分布统计与约束检查（任务路由占比、模板去重、格式合法性、长尾覆盖等）评估并调整数据，使模型既学会“答对”，也学会“按要求答”。

**(7) 回流重写/重生成：闭环提升与分布自适应** 当某类任务通过率持续偏低时，系统触发右侧“重写”回路：要么对指令/答案进行重写（保持图像不变），要么回流到“生成图片”分支重新合成更契合任务的图像，再进入流程。该闭环使数据系统具备“自我修复”的能力，避免一次性生成导致的质量不可控的同时也不要浪费以得倒的信息。

### 5.2.2 从两类任务实践到统一范式的归纳

需要强调的是，Auto-Evol 并非“从概念出发”的抽象框图，也不只是对两章工作的事后总结，而是建立在**系统性相关工作调研与本文多模态数据/训练实践之上的方法学抽象**：一方面，ShareGPT4V、SynthVLM、MMEvol、Synthesize Step-by-Step 等工作分别从高密度对齐、逆向数据工程、进化式增强与过程可执行推理等角度提供了可复用的模块化思想<sup>[1-3,21]</sup>；另一方面，本文在多模态任务落地过程中形成了关于多源数据组织、结构化标注、过程性监督与质量控制的一系列工程经验。下面以 GUI 与世界知识两条主线作为代表性实例说明该范式的归纳来源：

- **GUI 主线（第3章）：**以**结构化初始化与细粒度标注**为核心，将“截图”转化为可计算的控件集合（bbox、OCR、状态向量、关系图谱），再围绕 Referring/Grounding 生成任务化监督，并通过多源融合与质控迭代提升稳定性。该主线强调“**可执行定位**”与“**可解释语义**”的双对齐，并体现了细粒度对齐与分层能力路由在复杂交互场景中的必要性。
- **世界知识主线（第4章）：**以**实体锚定与知识注入**为核心，通过“词条 → 图片 → caption → 相关性过滤 → RecQA/KnowQA/FinalQA”建立从知识源到可学习多模态监督的转化链路，并通过**带推理过程（Think/CoT）的 FinalQA**强化“先识别再推理”的可迁移推理策略。该主线强调“**可验证对齐**”与“**推理链路贯通**”，对应了过程监督与可校验链路对世界知识推理能力提升的关键作用。
- **跨任务复用的工程实践（本文其他多模态工作）：**除上述两类任务外，本文在多模态数据构造与训练中还沉淀了可跨任务复用的通用做法，例如：以统一的“源信息池”承接不同数据源与标注形态、以三基础能力链路统一组织监督信号、以及以“难度/正确性/覆盖多样性”为目标对数据进行评测

式校验。这些实践与调研结论共同促成了 Auto-Evol 作为统一范式的可迁移性与可扩展性。

二者看似任务不同，但其共同结构恰好对应 Auto-Evol 的闭环：统一表示承接多源输入，原子化任务拆解把复杂能力拆成可监督信号，分层代理生成提升信息密度与多样性，多维过滤/校验与回流控制噪声与幻觉并推动持续迭代。因此，Auto-Evol 可被视为“面向不同任务路由的数据生产系统”，既能覆盖 GUI 这类强几何/强交互任务，也能覆盖世界知识这类强对齐/强推理任务。

### 5.3 模块有效性验证

Auto-Evol 的关键模块并非凭经验堆叠，其必要性可在第 4 章与第 5 章的实验结果中得到直接验证。下面从“模块 → 证据”的角度进行归因式总结。

**(A) 多源融合与结构化监督的必要性（对应初始化（含标注）/过滤）** 在 GUI 任务中，仅靠基线模型性能较低；引入开源清洗数据后指标显著提升，而进一步融合闭源精标数据后提升更为明显：如表3-1所示，Grounding F1 由 0.324 提升至 0.846，Referring F1 由 0.523 提升至 0.905。该结果表明，**多源覆盖 + 结构化标注 + 质量控制**能够显著提高定位与理解的稳定性，是 Auto-Evol “数据初始化–协作标注–过滤”的直接实证。

**(B) 中间表征与相关性过滤的必要性（对应 caption 与对齐校验）** 在世界知识任务中，caption 不仅提供“更丰富的描述”，更作为可依赖的中间表征支撑相关性过滤与后续 QA 构造。实验一的结论指出：**Image Caption 对识别显著有益**（表4-1），说明“中间表征 + 对齐筛选”能够为实体锚定提供更稳健的训练信号，对应 Auto-Evol 中“过滤相关性、准确性”与“相关性校验”的设计动机。

**(C) 显式推理过程数据对推理能力的显著增益（对应原子化/推理代理/执行校验）** 第4章进一步揭示：仅监督最终答案的 FinalQA 可能诱发“短路映射”，导致推理指标提升不显著甚至下降；而当 FinalQA 加入简要推理过程（Think/CoT）并扩充数据分布后，final 相关指标**出现显著提升**。如表4-2所示，在 Chinese SimpleVQA 上，Ovis2.5\_9B\_think\_part123 相对 qwen2.5-VL-7b 在 F1\_Final 上提升 11.9 个点（同时 F1\_Rec 提升 16.1 个点），验证了“**显式推理链路监督**”对可迁移推理能力的关键作用。这一现象与 Auto-Evol 强调的“原子化任务拆解 + 推理代理生成 + 多维校验（含可执行/一致性校验）”相一致：通过过程性监督与可验证机制，才能在提升推理强度的同时抑制幻觉与短路学习。

## 5.4 本章小结

本章围绕“数据质量成为瓶颈”的现实约束，提出并系统化阐释了**Auto-Evol**多模态数据构造范式。其核心意义在于：将以往零散、一次性的合成数据生成，升级为一个**可路由、可评测、可闭环迭代**的数据生产系统，从而稳定地产出高信息密度、高对齐度与高推理密度的训练信号。

具体贡献体现在三点：**(1) 统一能力视角**——以感知/理解/推理三基础能力为坐标系，把不同任务的数据需求统一到三条生成链路与可复用监督信号上，避免“为每个任务重新设计一套数据体系”；**(2) 评测式质控**——将数据校验显式定义为对数据的评测，围绕难度、正确性/可验证性与覆盖多样性进行筛选与分布修正，降低错配监督与幻觉传播风险；**(3) 闭环可持续迭代**——通过过滤、增强、校验与回流重写把失败样本转化为下一轮优化目标，使数据分布能够随任务瓶颈自适应演化。

范式的有效性在后续章节得到量化支撑：在 GUI 任务中，多源融合与结构化监督显著提升 Grounding/Referring 性能（表3-1）；在世界知识任务中，引入中间表征与过程性推理监督（Think/CoT）带来推理指标的显著增益（表4-2）。因此，Auto-Evol 既为第3章与第4章提供统一的数据工程底座，也为后续扩展到更多多模态场景提供了可复用的方法学框架。

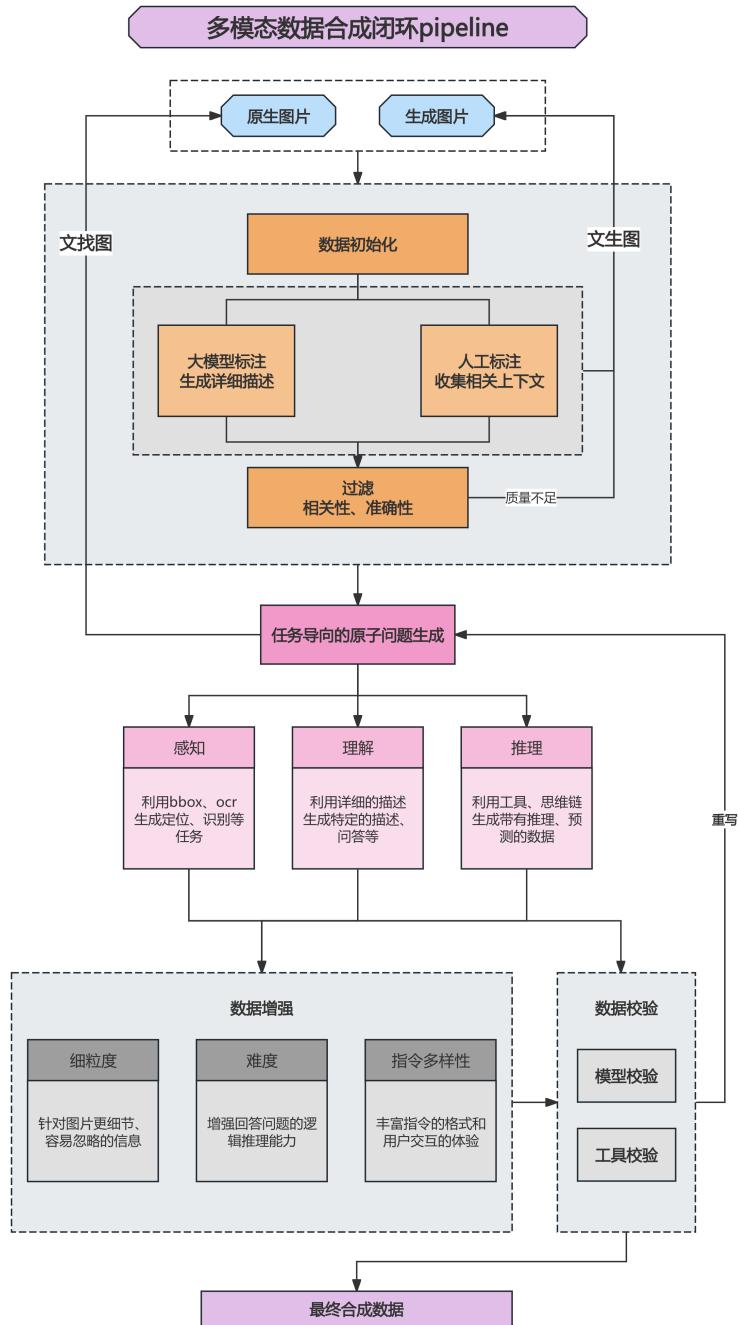


图 5-1 Auto-Evol 多模态合成数据闭环 Pipeline: 数据初始化（含标注）-过滤-原子问题生成-增强与校验-回流重写.



# 第 6 章 结论

## 6.1 研究总结

本文围绕“**多模态数据合成范式**”这一主线展开研究，针对多模态大语言模型（MLLM）训练中“**数据质量成为瓶颈**”的现实问题，探索如何以可控、可迭代、可验证的方式构造高质量训练数据，从而稳定提升模型在真实场景中的关键能力。相比仅依赖互联网抓取图文对或高成本人工标注，本文强调以数据工程视角把合成数据生产升级为系统化流程：在覆盖真实分布的同时提供强对齐与高推理密度监督，并通过多维质控抑制错配与幻觉传播。

为此，本文提出了**Auto-Evol** 多模态合成数据构造范式（见第 3 章），将以往单步的“图 → 文/问答”生成重构为包含**正反双向数据流、任务导向原子问题生成、代理式生成与进化增强、多维校验与闭环回流**的闭环 Pipeline。该范式以“感知/理解/推理”三基础能力为统一坐标系，把不同任务的数据需求抽象为可路由、可复用的生成链路，并将“**数据校验**”明确为对数据进行评测式验证，从而在数据侧前置控制训练信号的可信度与信息密度。

在范式落地层面，本文以两类高难任务作为代表性实例验证其有效性。其一是面向用户图形界面（GUI）的屏幕理解与可执行定位能力提升（第3章）：通过多源数据融合、结构化初始化、混合标注与属性补全、任务化构造与质控迭代，构建同时覆盖语义理解与几何定位的高质量监督信号，并结合高分辨率感知与训练策略提升 Referring 与 Grounding 性能。其二是面向世界知识的识别–知识注入–推理贯通能力提升（第4章）：构建可操作的世界知识类别框架与端到端数据 Pipeline，利用 caption 作为中间表征支撑相关性过滤与 QA 构造，并通过引入带推理过程（Think/CoT）的 FinalQA 缓解短路映射，显著增强“先识别再推理”的可迁移推理能力。

综上，本文从范式方法学、数据工程落地与实验验证三方面给出了一条可复用的路径：以闭环数据生产系统支撑不同任务域的合成数据构造，并以可验证的监督信号驱动多模态模型能力提升。

## 6.2 主要成果

本文的主要成果可概括为以下几个方面：

- 提出并系统化阐释**Auto-Evol** 多模态数据构造范式：给出从数据源接入、结

构化初始化、过滤、原子问题生成、增强与校验到回流重写/重生成的闭环 Pipeline，并以“感知/理解/推理”三基础能力统一组织多模态监督信号，使范式具备可路由、可评测与可迭代特性（第3章）。

- **构建面向 GUI 的多源高质量数据体系并提升 Referring/Grounding 能力：**针对控件密集与文字细小等难点，构建开源清洗数据、闭源精标数据与必要合成数据的互补语料，并引入 bbox/OCR/关系图谱/状态向量等结构化监督；实验表明多源融合与结构化监督显著提升定位与理解稳定性（第3章）。
- **构建面向世界知识的类别框架与三元组任务体系并提升推理能力：**建立 7 大类 40 子类的知识类别框架，提出“词条 → 图片 → caption → 相关性过滤 → RecQA/KnowQA/FinalQA”的 Pipeline，并通过 Think/CoT 形式的 FinalQA 提供过程性推理监督，在评测基准上获得显著提升（第4章）。

## 6.3 创新点

本文的创新点主要体现在以下三方面：

1. **范式层：将合成数据生成升级为闭环数据生产系统。**不同于将合成数据视为一次性“生成若干图文/QA 对”的做法，Auto-Evol 从工程系统视角构建闭环 Pipeline，并显式引入过滤、增强、校验与回流机制，使数据分布能够随任务瓶颈自适应演化，从而稳定产出高对齐与高推理密度的数据（第3章）。
2. **方法层：以三基础能力（感知/理解/推理）统一组织多任务监督与路由。**本文将多模态能力结构抽象为三条生成链路：感知链路提供可对齐锚点（如 bbox/OCR/实体识别），理解链路提供高密度语义与可复用上下文（如结构化 caption/关系解释），推理链路提供过程性约束与可验证推断（如带 CoT 的 FinalQA）。该统一视角使不同任务域的数据构造可复用同一套方法学组件，并便于进行消融与归因分析（第3章、第3章与第4章）。
3. **验证层：以“评测式质控”反向定义并验证数据有效性。**本文将数据校验视为对数据的评测，围绕难度、正确性/可验证性与覆盖多样性定义可操作的校验与过滤策略，降低错配监督与幻觉传播风险；并在 GUI 与世界知识两类任务中通过定量实验证关键模块（多源融合、caption 中间表征、过程性推理监督等）的必要性与贡献（第3章与第4章）。

## 6.4 不足与展望

尽管本文围绕多模态数据合成范式进行了系统探索并取得一定效果，但仍存在若干局限，有待后续进一步研究与完善。

### 6.4.1 研究不足

- 对自动化判别与强模型生成的依赖仍较强。** 在相关性过滤、质量打分、重写与推理链路生成等环节，流程往往需要依赖能力更强的模型作为生成器或判别器。尽管这种“LLM-as-a-Generator/Judge”的范式显著提升了自动化程度，但其成本、稳定性与偏置问题仍会影响数据分布与最终训练效果。
- 可验证性与工具化校验仍有提升空间。** 本文强调数据应可验证，但在部分任务（尤其是开放域世界知识）中，完全自动、可执行的事实核验仍较困难；当知识源不一致或存在时效差异时，如何构造严格唯一、可核验的监督信号仍需要更强的检索与证据链机制。
- 任务覆盖与范式泛化尚未完全展开。** 本文以 GUI 与世界知识两条任务线验证范式有效性，但 Auto-Evol 作为通用范式仍需要在更多模态与场景中验证，例如视频理解、时序交互、具身智能、多模态工具调用等，以进一步检验其可迁移性与可扩展性。
- 评测体系仍存在不足。** 当前评测主要围绕特定基准与任务指标进行，尚缺乏与数据生产闭环严格对齐的、覆盖“对齐度/推理密度/可执行性/安全性”的综合评价体系，导致部分数据改动的收益难以被快速、稳定地观测与归因。

### 6.4.2 展望

- 更强的证据链与工具校验：从“LLM 判别”走向“可执行验证”。** 未来可将检索增强（RAG）、结构化知识库与可执行工具（如规则验证、代码执行、可解析约束检查）更深度地嵌入数据校验环节，构建“证据 → 结论”的可追溯链路，以系统性降低幻觉与错误监督。
- 面向多模态 Agent 的数据闭环：从静态监督到交互式自改进。** GUI 场景天然具备可交互与可执行特征，未来可引入环境回放与任务执行反馈，把“模型输出是否可执行/是否达成目标”作为校验信号，形成更接近真实 Agent 训练的闭环数据生产与评测体系。

- **范式扩展到更多模态与任务域.** 将 Auto-Evol 扩展到视频、音频、三维与具身场景，探索时序一致性、跨帧对齐、长程规划等问题下的原子化任务拆解与质控策略，以验证范式在更复杂场景下的通用性.
- **效率与规模化：降低生成成本并提升覆盖.** 未来可研究更高效的生成与筛选策略（如分层采样、主动学习式采样、低成本模型预筛 + 高成本模型复核的级联机制），在保证质量的同时扩大长尾覆盖并降低整体数据生产成本.
- **数据治理与安全约束.** 随着合成数据规模扩大，数据的版权合规、隐私保护与安全对齐将更加关键. 未来可在数据初始化与过滤阶段引入更系统的治理策略与安全评测，确保数据生产链路在可控与可审计的前提下运行.

# 参考文献

- [1] CHEN L, LI J, DONG X, et al. Sharegpt4v: Improving large multi-modal models with better captions[A/OL]. 2023. <https://arxiv.org/abs/2311.12793>.
- [2] XU R, WEI J, LIN X, et al. Mmevol: Empowering multimodal large language models with evol-instruct[A/OL]. 2024. <https://arxiv.org/abs/2409.05840>.
- [3] LIU Z, LIANG H, LI B, et al. Synthvlm: Towards high-quality and efficient synthesis of image-caption datasets for vision-language models[C/OL]//Proceedings of the 33rd ACM International Conference on Multimedia. ACM, 2025. DOI: [10.1145/3746027.3758222](https://doi.org/10.1145/3746027.3758222).
- [4] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.
- [5] ALAYRAC J B, DONAHUE J, LUC P, et al. Flamingo: a visual language model for few-shot learning[C/OL]//Advances in Neural Information Processing Systems (NeurIPS): Vol. 35. 2022: 23716-23736. [https://proceedings.neurips.cc/paper\\_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d438-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d438-Abstract-Conference.html).
- [6] LI J, LI D, SAVARESE S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International Conference on Machine Learning. PMLR, 2023: 19730-19742.
- [7] LIU H, LI C, WU Q, et al. Visual instruction tuning[C]//Advances in Neural Information Processing Systems: Vol. 36. 2024.
- [8] LIU H, LI C, LI Y, et al. Improved baselines with visual instruction tuning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2024: 26296-26306.
- [9] BAI J, BAI S, YANG S, et al. Qwen-vl: A frontier large vision-language model with versatile abilities[A]. 2023.

- [10] MASRY A, LONG D X, TAN J Q, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning[C]//Findings of the Association for Computational Linguistics: ACL 2022. 2022: 2263-2279.
- [11] LIU F, EISENSCHLOS J M, PICCINNO F, et al. Deplot: One-shot visual language reasoning by plot-to-table translation[C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 10724-10732.
- [12] LIU F, PICCINNO F, KRICHENE S, et al. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 12756-12770.
- [13] LEE K, JOSHI M, TURC I, et al. Pix2struct: Screenshot parsing as pretraining for visual language understanding[C]//International Conference on Machine Learning. PMLR, 2023: 18893-18912.
- [14] PODELL D, ENGLISH Z, LACEY K, et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis[A]. 2023.
- [15] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Advances in Neural Information Processing Systems: Vol. 35. 2022: 24824-24837.
- [16] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Advances in Neural Information Processing Systems: Vol. 35. 2022: 27730-27744.
- [17] MUKHERJEE S, MITRA A, JAWAHAR G, et al. Orca: Progressive learning from complex explanation traces of gpt-4[A]. 2023.
- [18] WEI J, WAINAKH Y, TU R, et al. Simpleqa: Measuring short-form factuality in large language models[A/OL]. 2024. <https://arxiv.org/abs/2411.04368>.
- [19] WANG Y, KORDI Y, MISHRA S, et al. Self-instruct: Aligning language models with self-generated instructions[C/OL]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023: 13484-13508. <https://aclanthology.org/2023.acl-long.754>.
- [20] MITRA A, DEL CORRO L, ZHENG G, et al. Agentinstruct: Toward generative teaching with agentic flows[A/OL]. 2024. <https://arxiv.org/abs/2407.03502>.

- [21] LI Z, JASANI B, TANG P, et al. Synthesize step-by-step: Tools, templates and llms as data generators for reasoning-based chart vqa[A/OL]. 2024. <https://arxiv.org/abs/2403.16385>.
- [22] DEKA B, HUANG Z, FRANZEN C, et al. Rico: A mobile app dataset for building data-driven design applications[C/OL]//Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology (UIST). 2017. DOI: [10.1145/3126594.3126651](https://doi.org/10.1145/3126594.3126651).
- [23] LI G, BAECHLER G, TRAGUT M, et al. Learning to denoise raw mobile ui layouts for improving datasets at scale[C/OL]//Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI). 2022. <https://doi.org/10.1145/3491102.3502042>.
- [24] GAO L, ZHANG L, WANG S, et al. Mobileviews: A million-scale and diverse mobile gui dataset[A/OL]. 2024. <https://arxiv.org/abs/2409.14337>.
- [25] WU Z, WU Z, XU F, et al. Os-atlas: A foundation action model for generalist gui agents[A/OL]. 2024. <https://arxiv.org/abs/2410.23218>.
- [26] HoneyNet Project. Droidbot: A lightweight ui-guided test input generator for android[EB/OL]. <https://honeynet.github.io/droidbot/>.
- [27] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 779-788. <https://arxiv.org/abs/1506.02640>.
- [28] DU Y, LI C, GUO R, et al. Pp-ocr: A practical ultra lightweight ocr system[A/OL]. 2020. <https://arxiv.org/abs/2009.09941>.
- [29] BAI S, CHEN K, LIU X, et al. Qwen2.5-vl technical report[A/OL]. 2025. <https://arxiv.org/abs/2502.13923>.
- [30] OPENAI. Gpt-4o system card[A/OL]. 2024. <https://arxiv.org/abs/2410.21276>.
- [31] HESSEL J, HOLTZMAN A, FORBES M, et al. Clipscore: A reference-free evaluation metric for image captioning[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 7514-7528. <https://aclanthology.org/2021.emnlp-main.595/>. DOI: [10.18653/v1/2021.emnlp-main.595](https://doi.org/10.18653/v1/2021.emnlp-main.595).

- [32] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity[J/OL]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612. DOI: [10.1109/TIP.2003.819861](https://doi.org/10.1109/TIP.2003.819861).
- [33] WANG Y, MISHRA S, ALIPOORMOLABASHI P, et al. Supernaturalinstructions: Generalization via declarative instructions on 1600+ tasks[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 5085-5109.

## **复旦大学**

## **学位论文独创性声明**

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。论文中除特别标注的内容外，不包含任何其他个人或机构已经发表或撰写过的研究成果。对本研究做出重要贡献的个人和集体，均已在论文中作了明确的声明并表示了谢意。本声明的法律结果由本人承担。

作者签名：\_\_\_\_\_ 日期：\_\_\_\_\_

## **复旦大学**

## **学位论文使用授权声明**

本人完全了解复旦大学有关收藏和利用博士、硕士学位论文的规定，即：学校有权收藏、使用并向国家有关部门或机构送交论文的印刷本和电子版本；允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。涉密学位论文在解密后遵守此规定。

作者签名：\_\_\_\_\_ 导师签名：\_\_\_\_\_ 日期：\_\_\_\_\_