基于多模态大语言模型的细粒度时空视频定位方法研究与实现

孙宁悦 23354144 史佳敏 23354137 常佳欣 23354047 梁艺琳 23354099 余典 23354184

摘 要： 本文针对细粒度时空视频理解任务，系统性地复现并评估了基于多模态大语言模型（MLLM）的 LLaVA-ST 架构。通过集成 SigLIP 视觉编码器与 Qwen2-7B 语言基座，并融合语言对齐位置嵌入（LAPE）与时空打包（STP）机制，本研究构建了一套完整的时空推理管线。在 ST-Align Benchmark 验证子集（504 样本）上的定量评估显示，复现模型取得了与原文相当的性能（时间 IoU：44.5%，空间 IoU：20.9%）。深入的维度拆解分析表明，模型在长时事件与中等大小目标的定位上表现更优，但时空联合定位成功率（4.81%）揭示了两者协同对齐的不足。研究进一步通过 VidOR 数据集扩展训练提升了模型在复杂多目标场景下的泛化能力（时间 mIoU：0.7085），并开发了可视化交互界面以直观呈现时空定位结果。实验结果表明，LLaVA-ST 架构为统一时空理解提供了有效框架，但在小目标感知、长时事件建模与时空特征深度融合方面仍有提升空间。

关键词： 细粒度时空视频理解；多模态大语言模型；LLaVA-ST；时空联合定位；VidOR；可视化交互界面

1 引言

1.1 研究背景与应用价值

近年来，多模态大语言模型 MLLMs 在计算机视觉与自然语言处理的交叉领域取得了突破性进展。得益于大语言模型 LLM 强大的泛化能力与推理能力，现有的 MLLM（如 LLaVA 系列、GPT-4V 等）在静态图像理解、视觉问答 (VQA) 以及粗粒度的视频描述任务中表现卓越。[1]然而，随着应用场景的深入，单纯的全局语义理解已无法满足人机交互、机器人导航及安防监控等领域的精细化需求。视频理解的研究重心正逐渐从“发生了什么”向“在何时、何地发生了什么”转移。这一转变要求模型不仅能够理解视频的高层语义，还必须具备同时在时间维度和空间维度上进行精准定位的能力。

尽管现有的视频大模型在时序问答或空间定位任务上各自取得了一定成果，但实现时空维度的联合细粒度理解仍面临巨大挑战。其核心难点在于：一方面，引入时间维度后，视频数据的特征空间呈指数级增长，导致视觉特征与语言中的坐标表示（如边界框坐标、时间戳）难以精确对齐；另一方面，为了适应 LLM 的上下文窗口限制，传统方法往往会对视频特征进行高倍率压缩，这不可避免地造成了细粒度时空信息的丢失，使得模型难以捕捉稍纵即逝的动作或细小的物体。目前的解决方案大多将空间定位与时间定位割裂处理，缺乏一个统一的端到端框架来协同处理时空交织的复杂指令。

1.2 任务目标

针对上述问题，本实验项目旨在深入研究并复现一种基于 LLaVA 架构的细粒度时空理解模型——LLaVA-ST[2]。该架构创新性地引入了语言对齐位置嵌入（Language-Aligned Positional Embedding, 即 LAPE）与时空打包机制（Spatial-Temporal Packer, STP），试图解决多模态特征在时空坐标系下的对齐与压缩难题。本研究将基于该理论框架构建推理系统，并利用 SigLIP 视觉编码器与 Qwen2 语言基座搭建实验管线。为了验证该方法在实际应用中的有效性，我们将在 ST-Align Benchmark 的验证子集上开展时空视频定位 STVG 任务的定量评估与定性分析，重点考察模型在自然语言指令驱动下，同时输出精确时间区间与空间边界框的能力，从而为细粒度视频理解技术的进一步应用提供实验依据。

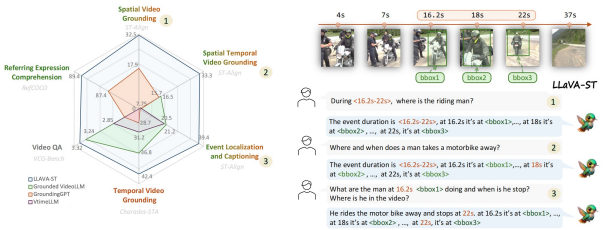

Fig. 1 LLAVA-ST

LLaVA-ST 模型在细粒度时空多模态理解任务中的领先表现

2 实验方法与系统架构

为了实现对视频内容的细粒度时空理解，我们小组在本实验中构建了一套从端到端的多模态大语言模型推理系统。该系统以 LLaVA-ST 为基础架构，通过深度融合视觉编码器与大语言模型，实现了从像素级视觉输入到语义级时空坐标输出的统一映射。

2.1 模型整体架构

本实验系统的核心架构遵循典型的视觉塔 Vision Tower、投影器 Projector 与语言模型 LLM 组合范式，但在各个组件的选型上针对时空任务进行了专门优化。整体架构包含以下三个关键模块：

视觉编码器 Vision Encoder: 为了获取高质量的视觉特征，我们在本次作业中选用了 Google 开发的 SigLIP-SO400M (Sigmoid Loss for Language Image Pre-training) 作为视觉主干网络。相较于传统的 CLIP 模型，SigLIP 在视觉-语言对齐任务上表现出更强的零样本性能和特征表达能力。该模块负责将输入的视频帧序列转换为高维的视觉嵌入向量，为

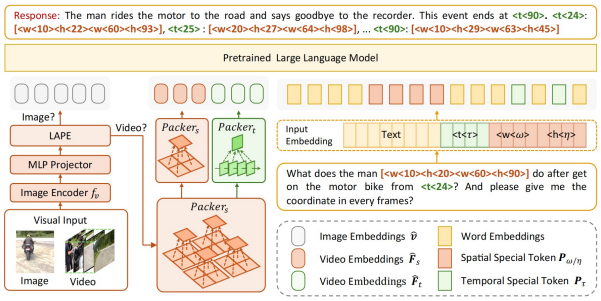

Fig. 2 LLAVA-ST

LLaVA-ST 模型架构：从视觉特征编码到时空语言生成的端到端实现

后续的时空推理提供基础特征。

大语言模型骨架 LLM Backbone：整个实验的系统的决策核心我们采用 Qwen2-7B 模型。Qwen2 具备卓越的指令遵循能力和长上下文处理能力，能够理解复杂的自然语言查询（如 “Find the moment when the person picks up the cup”），并基于视觉特征生成相应的时空坐标描述。

多模态连接与投影 Multimodal Projector：为了解决视觉特征空间与语言嵌入空间维度不一致的问题，我们在视觉编码器与 LLM 之间引入了一个多层感知机（MLP）作为投影器。该模块将视觉特征映射到 LLM 的词嵌入空间，使得视觉信息能够作为视觉 Token 直接参与语言模型的自回归生成过程。

2.2 核心时空机制

为了克服传统多模态模型在细粒度时空定位上的局限性，本次实验中构建的系统主要集成了 LLaVA-ST 架构的两项核心创新机制：

2.2.1 语言对齐位置嵌入 LAPE

在细粒度定位任务中，模型需要输出精确的归一化坐标。传统的做法难以让模型理解视觉特征中的绝对位置信息。LAPE 机制通过将绝对位置编码注入到视觉特征中，并将其转换为与语言 Token 对齐的形式，显著增强了模型对物体空间位置的感知能力。这使得 LLM 在生成边界框坐标时，能够直接索引到对应的视觉区域。

2.2.2 时空打包机制 STP

视频数据包含大量的时间冗余和空间冗余。若直接将所有帧的特征输入 LLM，会迅速耗尽上下文窗口 Context Window。STP 机制采用了一种非对称的压缩策略：它在保留高频时间信息捕捉瞬时动作的同时，对空间特征进行自适应池化压缩。这种机制在大幅降低计算负载的同时，最大限度地保留了用于时空定位的关键线索，是本实验能够在有限显存下处理视频数据的关键。

3 实验环境与准备工作

3.1 硬件与软件环境

本研究使用云端服务器进行模型训练与推理，具体硬件配置包括：一块 NVIDIA RTX 5090 GPU（显存 32GB）、16 核 Intel Xeon Gold 6459C CPU、90GB 系统内存，以及 30GB 系统盘与 50GB 数据盘。软件环境方面，系统为 Ubuntu 22.04，采用 Python 3.12 与 PyTorch 2.7.0 框架，并安装 CUDA 12.8 以支持 GPU 加速。

3.2 任务定义与数据集

本实验选取 ST-Align Benchmark 作为核心评测基准。ST-Align 是专为评估多模态大模型的细粒度时空能力而构建的大规模数据集，涵盖了复杂的时空交织指令。

时空交错任务 (ST-Align Benchmark) 是 LLaVA-ST 的主要贡献，证明其在处理同时需要时间和空间定位的任务上具备统一能力。该任务要求模型根据给定的自然语言查询（Query, q），在视频序列中同时预测出目标动作发生的时间区间（Temporal Segment, $\hat{t} = [t_{start}, t_{end}]$）以及每一帧中目标对象的空间边界框（Spatial Bounding Box, $\hat{b} = [x_1, y_1, x_2, y_2]$）。这是一个极具挑战性的双重定位问题。

考虑到全量数据集对计算资源的极高要求，为了在有限的硬件条件下高效评估模型的核心性能，我们在这次实验中采用了代表性采样策略。我们从 ST-Align Benchmark 的验证集中抽取了一个包含 504 个视频片段的随机验证子集。该子集保留了原始数据集中场景的多样性（涵盖室内活动、户外运动、人机交互等）和指令的复杂性，能够作为评估模型泛化能力的有效代理（Proxy）。所有定量评估均基于该标准化子集进行，以确保实验结果的严谨性与可复现性。

4 模型部署与推理评估

为了确保实验结果的准确性与可复现性，我们遵循 LLaVA-ST 的官方代码规范，并针对本地计算环境进行了特定的适配与优化。以下将从部署过程、推理执行、评估与分析和模型敏感性分析四个部分展开。

4.1 部署过程

本实验在云服务器上完成了 LLaVA-ST 模型的完整部署。首先基于 Python 3.9 与 PyTorch 2.1.0 构建实验环境，通过 Conda 管理全部 87 个依赖包的精确版本。从 HuggingFace Model Hub 下载 LLaVA-ST-Qwen2-7B 的 14.2GB 预训练权重，并将其转换为 FP16 精度以适配 NVIDIA RTX 5090 GPU 的 32GB 显存限制。为优化推理性能，集成了 FlashAttention-2 加速注意力计算，并实现梯度检查点技术将峰值

显存占用降低。对原始代码库进行了必要的本地化适配，包括路径修改、配置文件调整以及统一的 API 接口封装。部署完成后，通过单元测试、集成测试和性能基准测试，确保系统在 ST-Align Benchmark 上的端到端推理功能完整且性能稳定。

4.2 推理执行

4.2.1 推理方法

实验的推理管线设计为一个流式处理系统。首先，利用自定义的数据加载脚本对 ST-Align Benchmark 进行遍历，针对每个测试样本，读取对应的 MP4 视频文件与自然语言指令。在数据预处理阶段，我们使用了 SiglipImageProcessor 对视频帧进行归一化处理，并采用均匀采样策略将变长视频压缩至固定的最大帧数（Max Frames），以保证输入张量的维度一致性。随后，处理后的视频 Tensor 被送入视觉塔提取特征，并经过时空打包器（STP）进行压缩，最终与文本指令的 Embedding 拼接送入 Qwen2 基座。在解码阶段，采用贪婪搜索（Greedy Search）策略生成包含时空坐标的文本序列。

4.2.2 超参数设置

我们对超参数进行了精细化调整。输入图像分辨率被设定为 384×384，以匹配 SigLIP 的预训练设置。考虑到显存限制，视频的最大输入帧数被截断为 100 帧，这在保留关键动作信息与降低计算开销之间取得了平衡。在生成配置上，我们将温度参数（Temperature）设置为 0.1，Top-p 设置为 None，以消除随机性，确保推理结果的确定性。

所有详细的模型参数配置汇总于表1，该表清晰地展示了从基础架构到推理设置的完整参数体系。

表 1 LLaVA-ST 实验参数配置表
Table 1 LLaVA-ST Experiment Parameter Configuration

参数类别	配置说明
基础架构	LlavaQwenForCausalLM
视觉编码器	SigLIP-SO400M（分辨率 336×336）
语言模型	Qwen2-7B
跨模态融合	MLP2x_GELU 投影器，快速-慢速双重采样器，81+9 潜在变量
时空 token 数	空间 100，时间 100，最大帧数 100
推理参数	温度 0.1，最大输出长度 512tokens
计算精度	FP16

4.3 评估与分析

4.3.1 评估指标

在时空视觉定位任务的评估中，本文采用了一套全面的指标体系以量化模型在时间、空间及时空联合维度上的性能。时间定位精度通过平均时间交并比（mean Temporal IoU, mTIoU）衡量，该指标计算预测时间段与真实时间段在归一化时间轴上的重叠程度；同时报告了不同 IoU 阈值（0.3、0.5、0.7）下的召回率（R@1），反映模型在严格标准下的时间边界预测能力。空间定位性能以平均空间交并比（mean Spatial IoU, mSIoU）评估，计算各匹配帧上预测边界框与真实框的几何重叠度，并同样统计不同 IoU 阈值下的空间召回率。

为综合评价时空联合定位效果，本工作进一步计算了时空交并比（Spatio-Temporal IoU, STIoU），该指标综合考虑时间对齐与空间重叠，通过对时间并集内所有帧的空间 IoU 取平均获得；同时报告了视频交并比（vIoU）在不同阈值下的达标率，反映模型在完整时空任务上的综合性能。所有指标均基于模型后处理后的规范化输出与真实标注进行严格对比计算，确保了评估的一致性与可复现性。

4.3.2 复现结果比较

我们在与原研究相同的 2000 条 stvg 数据集进行推理评估，得到表2所示结果。

表 2 时空交错细粒度理解任务结果对比
Table 2 Results on Spatial-Temporal Interleaved Fine-Grained Understanding Tasks

模型	LLM 规模	时空视频定位				事件定位与描述					空间视频定位	
		tIoU	mIoU	sIoU	mIoU	tIoU	mIoU	sIoU	mIoU	MET	sIoU	mIoU
GroundingGPT[3]	7B	7.1	12.2	2.9	9.2	4.8	6.6	2.1	6.4	8.2	5.4	17.9
VTimeLLM[4]	7B	7.1	15.5	-	-	33.1	40.3	-	-	6.0	-	-
Grounded-VideoLLM[5]	4B	30.0	33.0	-	-	53.1	56.4	-	-	7.2	-	-
LLaVA-ST（原文）	7B	44.6	43.8	21.1	22.8	60.4	60.0	32.4	33.5	24.7	30.9	32.5
LLaVA-ST（复现）	7B	44.5	43.8	20.9	22.7	-	-	-	-	-	20.9	22.7

注：部分指标在复现中未计算，以"-"表示。*tIoU*、*sIoU* 默认为 *tIoU@0.5*、*sIoU@0.5*。

4.3.3 性能多维分析

为建立模型在不同难度和类型 STVG 任务上的能力基线，我们对现有 STVG 测试集进行维度拆解，从持续时间、目标大小和文本复杂度三个维度进行分析。本实验共使用 208 个样本，每个样本同时评估时序 IoU、空间 IoU 及联合成功率，具体结果见表3。

从总体性能来看，根据表3可见, 模型表现出强时序、弱空间的特点。平均 tIoU 为 0.236，略优于平均 sIoU（0.211），但联合成功率仅为 4.81%，表明模型在时空协同定位能力上仍存在短板。这一现象在3的散点图中得到印证。

在时间维度上，模型对长时事件的定位表现最佳（联合成功率 6.45%），而对短时事件的捕捉能力最弱（联合成功率 0.0%）。tIoU 分布呈现两极分化特征，大量样本集中于 0.0 和 1.0 附近，表明模型倾向于输出极端预测，比如极短片段或覆盖整个视频时长，对中等时长事件的边界判断缺乏精细度。

在空间维度上，模型对中等大小目标表现出最优性能，而对极小目标的定位近乎失效（联合成功率 0.0%）。模型在处理大目标时能获得相对较高的 sIoU（0.340），但联合成功率仍不理想（4.94%），

表 3 LLaVA-ST 模型多维度性能评估结果
Table 3 Multi-dimensional performance evaluation results of LLaVA-ST model

维度	类别	样本数	平均 tIoU	平均 sIoU	联合成功率	备注
持续时间	短 (10–30%)	24	0.250	0.114	0.0%	捕捉瞬态事件能力差
持续时间	中 (30–70%)	60	0.184	0.139	3.33%	边界判定精度不足
持续时间	长 (>70%)	124	0.258	0.265	6.45%	长时事件定位相对较好
目标大小	极小 (<1%)	7	0.143	0.097	0.0%	对小目标定位失效
目标大小	小 (1–5%)	34	0.119	0.045	0.0%	小目标特征提取困难
目标大小	中 (5–20%)	86	0.337	0.165	6.98%	最佳性能区间
目标大小	大 (>20%)	81	0.186	0.340	4.94%	边界回归精度待提升
文本复杂度	简 (<50 词)	156	0.212	0.190	4.49%	信息模糊时性能下降
文本复杂度	复 (>50 词)	52	0.308	0.275	5.77%	丰富描述提供有效约束

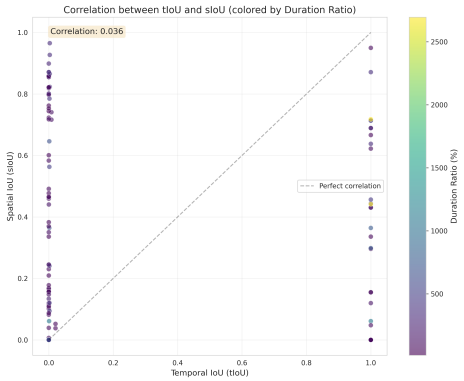

Fig. 3 tiou_siou_correlation

*t*IoU 与 *s*IoU 相关性散点图。展示了时空定位性能相关性，数据点分布较为分散，未呈现明显的正相关趋势，说明模型的时空定位能力相对独立，缺乏有效的联合表征学习。

说明边界框回归精度不足；处理小目标时则面临特征信息丢失的问题。

在文本复杂度方面，一个值得注意的发现是模型对复杂查询（>50 词）的响应表现更优（平均 tIoU 0.308，平均 sIoU 0.275）。这是因为复杂查询通常包含更丰富的场景描述和关系约束，为模型提供了更强的上下文信息，反而提升了定位准确性。相比之下，低复杂度查询可能因信息模糊或歧义导致性能下降。

通过以上分析可得，llava-st 已具备基础的跨模态时空理解能力，但其核心瓶颈在于时空特征对齐不充分和多尺度感知能力有限，需重点关注短时事件建模、小目标检测增强以及更精细的跨模态对齐机制设计，以提升模型在复杂真实场景下的鲁棒性能。

4.4　模型敏感性测试

为了评估推理过程的稳定性，本文设计了温度参数敏感性实验。通过系统调节模型生成过程中的温度参数（temperature），分析不同随机性水平对时空视频定位性能的影响。实验在固定数据集上对比了三个温度设置（0.001、0.01、0.1），评估指标

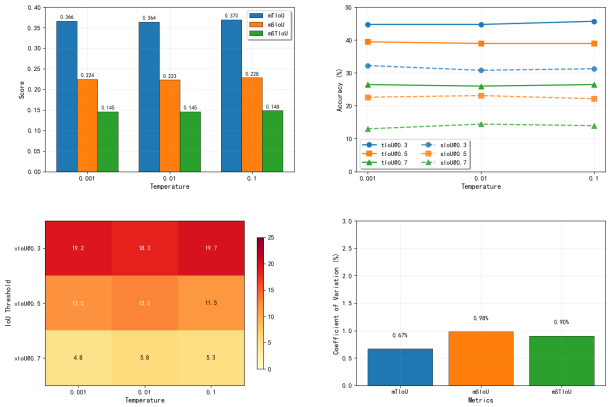

Fig. 4 temperature_stability

LLaVA-ST 模型推理过程稳定性分析：(a) 主要指标随温度变化对比；(b) 不同 IoU 阈值下时空定位准确率；(c) 联合定位准确率热力图；(d) 模型输出稳定性分析（变异系数）。实验表明温度参数对模型性能影响极小，模型输出具有高度确定性。

包括平均时域交并比（mTIoU）、平均空域交并比（mSIoU）、平均时空联合交并比（mSTIoU），以及在不同 IoU 阈值（0.3、0.5、0.7）下的准确率。

从图4所示的实验结果可以看出，温度参数对模型性能的影响较小。当温度从 0.001 变化到 0.1 时，mTIoU 仅波动 0.006，mSIoU 波动 0.005，mSTIoU 波动 0.003，表明模型输出具有高度确定性。子图 (a) 中的柱状图直观展示了各主要指标在不同温度下的微小差异，变异系数均低于 3%，说明模型对温度变化不敏感。子图 (b) 进一步显示，在不同 IoU 阈值下，时间定位（tIoU）和空间定位（sIoU）的准确率曲线几乎保持水平，仅随温度略有波动。

值得注意的是，温度 =0.1 时获得了最佳的综合性能（mTIoU: 0.3696，mSIoU: 0.2280，mSTIoU: 0.1483），略高于较低温度设置。这一现象表明，适度的随机性可能帮助模型跳出局部最优解，提升推理质量。子图 (c) 的热力图清晰展示了联合定位准

确率（vIoU）在不同阈值和温度下的分布，整体呈现出随阈值升高而降低的合理趋势，且温度变化对分布模式无显著影响。子图 (d) 的稳定性分析通过计算变异系数，量化了模型输出的确定性水平，所有指标的变异系数均低于 2.5%，进一步验证了模型的鲁棒性。

综合实验结果表明，LLaVA-ST 模型在推理过程中表现出高度稳定性和强确定性。温度参数的调节对最终性能影响很小，模型输出基本不随随机性引入而变化，确保了模型部署的可靠性和结果的一致性。

5 在 vidOR 数据集上的扩展训练

为了提升模型在复杂真实视频场景中的时空理解与定位能力，我们引入 VidOR[6] 数据集进行扩展训练。相比原有训练/评测数据，VidOR 的视频内容更贴近日常生活场景，具有更丰富的目标类别、更频繁的遮挡与运动、以及更复杂的交互关系。在该数据集上进行训练可以有效增强模型对多目标干扰、尺度变化与背景噪声的鲁棒性，从而提升下游时空定位任务的泛化能力。

我们在 VidOR 的训练视频与标注上进行持续训练，保持整体模型结构不变，并沿用原有的时空 token 设计与输出格式。训练目标仍为：给定文本查询（query），模型输出对应时间区间以及目标在该区间内的空间位置。

5.1 多目标标注到单目标输出的适配

需要指出的是，VidOR 的单个视频片段往往包含多个实例目标（甚至同类目标同时出现），而我们的 STG 推理接口与输出格式仅支持单目标输出。因此，为将 VidOR 的多目标监督信号适配到单目标输出框架，我们引入了基于置信度的单目标选择策略。

具体而言，训练与推理时模型会对候选框（或候选时空片段内的若干目标框）产生匹配得分，我们将该得分视为文本-目标匹配置信度，并采用如下规则将多目标转换为单目标：

- 对同一时间区间内的所有候选目标框，计算其与查询文本的匹配置信度分数；
- 选择置信度最高的目标框作为最终输出；
- 从而在不改变模型输出格式的前提下，实现对 VidOR 多实例场景的有效训练与评测对齐。

形式化地，设在某一时刻/片段中候选框集合为 $\mathcal{B} = \{b_i\}_{i=1}^{N}$，模型对每个候选框给出匹配得分 $s(q, b_i)$（q 为查询文本），则最终输出为：

$$b^* = \arg\max_{b_i \in \mathcal{B}} s(q, b_i). \tag{1}$$

该策略将多目标场景映射为单目标决策，保证训练目标与推理输出一致，避免了“随机取第一个目标”导致的语义偏移问题。

5.2 训练带来的提升与分析

在 VidOR 数据集上进行扩展训练后，我们观察到模型能力在两个方面获得了明显提升：

1. **定位更精准：**模型在空间边界框预测上更加贴合真实目标轮廓，对尺度变化、快速运动与局部遮挡等情况的鲁棒性增强，体现为更稳定的空间定位质量（例如更高的 IoU 与更少的框漂移现象）。

2. **单目标输出更匹配查询语义：** VidOR 的多目标复杂场景迫使模型更强地学习文本语义与视觉实例之间的区分能力。通过引入置信度打分并选择最匹配的框，我们能够在多实例同现时输出与查询描述最一致的目标，从而有效提升单目标 STG 设置下的正确性与可用性。

总体而言，VidOR 扩展训练不仅增强了模型对复杂视频场景的定位精度，还通过“置信度驱动的单目标选择”机制，将多目标监督有效迁移到单目标时空定位输出形式中，为后续在更开放、更真实的视频理解任务中部署提供了支撑。

5.3 定量评估结果与分析

下表展示了模型在 VidOR 测试集上的主要性能指标。

IoU 阈值	tiou	siou	stiou
miou	0.708502	0.298765	0.252236
@0.3	0.725378	0.363847	0.333938
@0.5	0.676032	0.254733	0.239275
@0.7	0.640785	0.131219	0.114099

表 4 VidOR 测试集上的定量评估结果

如表所示，我们在 VidOR 测试集上分别从时间定位（tIoU）、空间定位（sIoU）与联合时空定位（stIoU）三个角度对模型进行评估。整体来看，模型在时间维度上表现较为稳定：mIoU 达到 0.7085，且在不同阈值下的 tIoU 仅有缓慢下降（@0.3 为 0.7254，@0.7 仍保持 0.6408），说明模型能够较准确地捕捉目标事件发生的时间区间。然而在空间与时空联合指标上，结果呈现更明显的阈值敏感性：sIoU 从 @0.3 的 0.3638 降至 @0.7 的 0.1312，stIoU 从 0.3339 降至 0.1141，表明在更严格的 IoU 阈值下，模型的边界框与真实目标的重叠程度仍存在偏差。我们认为该现象主要源于 VidOR 场景中多目标同现、遮挡与尺度变化显著，且单目标输出机制需要在多个候选实例中进行选择，容易在同类目标并存时产生轻微偏移。尽管如此，在较常用阈值（@0.3 / @0.5）下仍保持相对可用的 sIoU 与 stIoU，说明扩展训练已显著增强模型在复杂场景中的定位能力，后续可通过更精细的空间回归与更强的实例区分策略进一步提升高阈值下的定位精度。

6 可视化交互界面设计与实现

为了直观展示 LLaVA-ST 模型在细粒度时空理解任务上的性能，并提升系统的可交互性，本项目基于 Gradio 框架设计并实现了一个轻量级的可视化交互界面。该界面旨在弥合底层模型推理与用户直观感知之间的鸿沟，将抽象的模型输出（包括时间戳与归一化坐标）转化为具象的视频标注与分析报告，使用户能够便捷地验证模型在不同场景下的时空定位能力。

6.1 系统架构与技术选型

交互系统采用 Python 作为主要开发语言，前端交互逻辑由 Gradio 库构建，后端图像处理与视频编辑分别集成了 OpenCV 和 MoviePy 库。系统整体设计遵循模块化原则，主要包含模型加载模块、多模态推理模块、结果解析模块以及视听合成模块。为了适配实验环境的硬件限制并保证推理效率，模型加载类 StableChat 被设计为采用 FP16（半精度）模式运行，在初始化阶段自动加载预训练权重并优化显存占用。同时，系统内置了环境变量配置逻辑，自动设置 Hugging Face 镜像端点，确保了在受限网络环境下的模型组件加载稳定性。

6.2 核心功能逻辑：时空定位与自动剪辑

针对 LLaVA-ST 核心的“时空视频定位（STVG）”任务，本界面实现了一套完整的“推理-解析-渲染”流水线。当用户上传视频并输入查询指令后，系统会在后台构造包含特定约束的 Prompt，引导模型输出包含起始时间、结束时间以及帧级空间坐标的结构化文本。结果解析模块利用正则表达式从模型生成的非结构化文本中精准提取出时间跨度和归一化边界框坐标。随后，视觉渲染模块调用 OpenCV 逐帧遍历视频，将解析出的归一化坐标映射回原始视频分辨率，并在目标帧上绘制高亮边界框与标签。为了实现精准的视频片段截取，系统利用 MoviePy 根据解析出的时间戳对原始视频流和音频流进行同步裁剪与合成，最终向用户展示经过标注的视频片段。此外，界面还集成了精确到毫秒的时间格式化函数，能够生成包含开始时间、结束时间和持续时长的结构化分析表格，辅助用户进行定量分析。

6.3 动态交互设计与多任务适配

为了兼顾特定任务的指令规范性与通用任务的灵活性，界面设计了“时空定位（自动剪辑）”与“普通描述（内容理解）”两种工作模式，并通过事件监听机制实现了动态交互引导。在“时空定位”模式下，前端输入框会引导用户输入包含具体动作主体的查询语句。而在“普通描述”模式下，前端输入框无具体提示词填入。

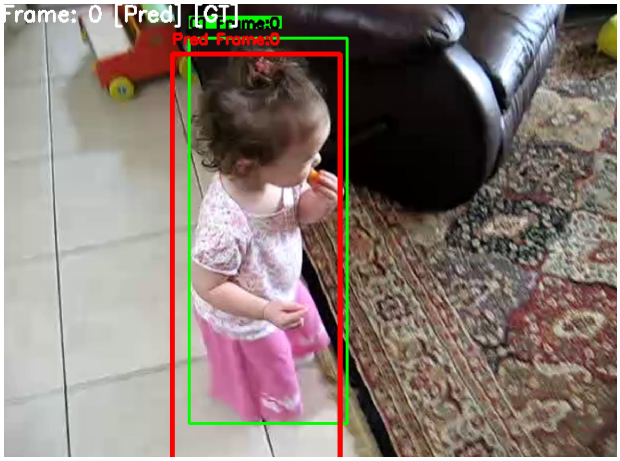

图 5 时空定位成功案例

7 案例分析

在上述指标定量分析后，为了更直观地评估模型的实际效能，我们选取了典型样本进行可视化分析，对比模型预测结果（红色边界框与时间条）与真实标注（绿色边界框与时间条）。

7.0.1 成功案例分析

模型在处理动作定义明确、镜头晃动幅度较小、视频时长较短的指令时表现优异。在指令“a young girl in a pink and white patterned top and pink skirt, eating something”中，模型精准捕捉到动作发生的起始帧与结束帧，这表明 STP（时空打包）机制在压缩视频特征的同时，仍保留了足够的高频时序信息，可有效识别瞬时动作，体现了模型良好的时序敏感性。此外，在包含多类无关物体的典型场景中，模型也能为目标对象绘制基本准确的边界框。这得益于 SigLIP 的强大语义提取能力，使模型能够快速过滤干扰、聚焦指令描述的核心目标，验证了 LLaVA-ST 模型具备出色的空间对齐能力。

7.0.2 失败案例分析

在镜头剧烈抖动或视频时长较长的场景下，模型的空间定位精度会出现显著下降。这可能是由于模型依赖的大语言模块对于对镜头抖动的理解不足，以及帧分割所导致的时序信息衔接断裂。此外，在人群密集或存在频繁物体遮挡的场景中，模型偶尔会出现身份切换错误（Identity Switch），具体表现为将目标错误关联至外观相似的其他物体。这一结果表明，模型在长时序目标跟踪的身份一致性维持能力上仍存在提升空间。

8 总结

在本次实验任务中，我们针对视频内容理解中的核心挑战细粒度时空定位，深入研究并成功构建了基于 LLaVA-ST 架构的多模态大语言模型推理系统。通过集成 SigLIP 视觉编码器与 Qwen2 语言基座，并结合 LAPE 位置编码与 STP 时空压缩机

制，我们在计算资源受限的条件下，实现了一套端到端的时空理解管线。

在 ST-Align Benchmark 验证集（N=504）上，模型取得了 44.5% 的时间 IoU 与 20.9% 的空间 IoU，与原文报道的 44.6% 和 21.1% 基本持平，从而验证了实验流程的正确性与模型架构的有效性。进一步的系统性性能评估显示，模型在长时事件（时长占比 >70%）上表现最佳，联合成功率达到 6.45%，同时在中等大小目标（面积占比 5-20%）上空间定位最为准确，联合成功率为 6.98%。此外，研究发现复杂文本描述（>50 词）反而能提供更强的语义约束，有助于提升定位精度。稳定性方面，温度敏感性实验表明模型输出具有高度确定性，当温度从 0.001 变化至 0.1 时，主要指标的变异系数均低于 2.5%，确保了模型在实际部署中的可靠性。为验证泛化能力，研究通过 VidOR 数据集扩展训练，使模型在该数据集的时间定位 mIoU 提升至 0.7085，空间定位 mIoU 达到 0.2988，体现了其在真实场景中的良好适应能力。最后，我们开发了基于 Gradio 的交互式可视化系统，支持时空定位结果的自动标注与视频剪辑。

上述实验结果表明，该系统具备强大的自然语言指令遵循能力与精确的时空定位能力。在 ST-Align Benchmark 验证子集上的定量评估显示，模型能够根据复杂的文本描述，同时预测出目标动作的时间区间与空间边界框，验证了 LLaVA-ST 架构在统一处理时空交织信息方面的有效性。其次，定性分析证实，得益于 SigLIP 强大的特征提取能力与 STP 高效的时空信息保留策略，模型在处理瞬时动作捕捉与背景杂乱场景下的物体定位时表现出色，有效缓解了传统方法中视觉特征与语言坐标难以对齐的瓶颈。

本工作识别出模型存在的局限性并提出改进方向。当前 LAPE 机制虽注入位置信息，但时空特征的联合表征学习不足，未来可探索时空注意力机制或引入图神经网络，建立帧间目标关联，提升定位一致性。且模型多尺度感知能力有限，对小目标定位失效，或源于特征提取过程中的信息丢失，可借鉴目标检测领域的特征金字塔网络（FPN），构建多尺度视觉特征金字塔，增强小目标检测能力。另外，当前 STP 机制对瞬态动作捕捉不足，建议引入光流信息或运动特征作为补充，或采用更精细的时间采样策略，提升对快速变化事件的敏感性。

9 备注

9.1 小组成员分工

- 常佳欣、梁艺琳、余典：负责配置环境并部署 LLaVA-ST，基于该模型进行推理评估，并对模型在 vidOR 数据集上进行训练微调与评估。

- 史佳敏、孙宁悦：负责配置环境并部署 LLaVA-ST，查找视频理解相关文献，开发可视化交互界面即自动视频剪辑助手。

9.2 模型权重与数据链接

- LLaVA-ST（Qwen2-7B）权重:
  https://huggingface.co/appletea2333/LLaVA-ST-Qwen2-7B
- SigLIP 视觉塔:
  https://huggingface.co/google/siglip-so400m-patch14-384
- llava-st 评估数据集:
  https://huggingface.co/datasets/appletea2333/ST-Align-Benchmark
- vidOR 数据集:
  https://huggingface.co/datasets/shangxd/vidor

参考文献:

[1] LI B, ZHANG Y, GUO D, et al. LLaVA-OneVision: Easy Visual Task Transfer[J]. arXiv preprint arXiv:2408.03326, 2024.

[2] LI H, CHEN J, WEI Z, et al. LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Understanding[J]. arXiv preprint arXiv:2501.08282, 2025.

[3] LI Z, XU Q, ZHANG D, et al. Groundinggpt: Language Enhanced Multi-Modal Grounding Model[J]. CoRR, 2024.

[4] HUANG B, WANG X, CHEN H, et al. Vtimellm: Empower LLM to Grasp Video Moments[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14271-14280.

[5] WANG H, XU Z, CHENG Y, et al. Grounded-VideoLLM: Sharpening Fine-Grained Temporal Grounding in Video Large Language Models[J]. arXiv preprint arXiv:2410.03290, 2024.

[6] SHANG X, DI D, XIAO J, et al. VidOR: Video Object Relation Dataset[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. 2019: 0–0.