# 阅后即焚：多模态大型语言模型是否真正捕捉到图像序列中的事件顺序？

**Yingjin Song, Yupei Du, Denis Paperno and Albert Gatt**
Utrecht University, Utrecht, the Netherlands
{ y.song5, y.du, d.paperno, a.gatt } @uu.nl

## Abstract

本文介绍了 TempVS 基准测试，该测试集中于多模态大语言模型（MLLMs）在图像序列中的时间性基础和推理能力。TempVS 由三个主要测试组成（即事件关系推断、句子排序和图像排序），每个测试都伴随着一个基本的基础测试。TempVS 要求 MLLMs 依赖视觉和语言模态来理解事件的时间顺序。我们评估了 38 种最先进的 MLLMs，结果显示模型在解决 TempVS 时存在困难，与人类能力相比有显著的性能差距。我们还提供了细致的见解，提出了未来研究的有希望的方向。我们的 TempVS 基准测试数据和代码可在 https://github.com/yjsong22/TempVS 获得。

## 1 介绍

多模态大型语言模型（MLLMs）(**???**) 在各种视觉和语言任务中表现出色。同时，对标准化评估框架的需求变得越来越关键。现有的大多数基准测试都集中在涉及单图像的设置上 (**?????**) 。虽然一些基准测试考虑了多图像设置 (**????**) ，但它们主要关注跨图像识别和引用。迄今为止，对更复杂任务的关注相对较少，例如在多个图像中的扎根时间理解和推理。

最近的一些研究已经评估了 MLLMs 在多张图像上的时间理解能力，但仍然存在某些限制。首先，一些任务可以通过依赖单个图像而不是序列来解决 (**??**) 。例如，给定一个图像序列，判断“拉动发条玩具是继续前进还是迅速停止”只需根据最后一张图像即可回答。其次，一些任务严重依赖常识或世界知识 (**??**) ，比如将一组打乱的图像重新排列成正确的烹饪步骤顺序。第三，某些基准测试 (**?**) 使用与图像无关的干扰选项，使模型能够根据对象的存在推断出真实答案。这些因素可能导致基准测试未能真正评估模型对时间序列的理解。此外，现有的基准测试没有为多事件场景设计，因此不足以评估 MLLMs 中复杂的时间序列和关系。

因此，一个问题产生了：现有的多模态大型语言模型（MLLMs）是否真的通过准确排列语言和图像序列中的事件顺序来理解时间？为了解决这个问题，我们提出了 TempVS，一个用于多事件视觉故事图像序列中时间定位和推理的基准。TempVS 包含 2,085 个图像序列（9,803 张图像），覆盖卡通动画、电影和日常生活相册，附有 15,192 道选择题。TempVS 具有三个时间理解和推理任务：事件关系推断、句子排序和图像排序（如 Figure 1 所示）。这些任务伴随着定位任务，以检查模型是否能够匹配与单一文本描述一致的精确图像。我们选择由视觉故事组成的图像序列，每个故事包含几个时间相关但相对独立的事件，因为没有事件可以轻易从前面的事件中预测。这使得在不考虑语言和视觉模式的情况下解决这些任务变得困难。TempVS 挑战模型推理图像序列和文本中的事件顺序（例如，使用“之前”或“之后”描述两个事件的句子，或者一个故事），并整合两者。

我们广泛评估了 38 个 MLLMs，包括从 0.5B 到 78B 的开源模型（例如，LLaVA-OneVision、InternVL2.5、Qwen2-VL、Phi-3.5-vision、DeepSeek-vl2、LLaVA-NeXT-Video）和闭源的 GPT-4o。我们表明，TempVS 对于 SOTA 模型来说是具有高度挑战性的，尤其是在事件关系推理和图像排序任务上。特别是，虽然模型可以准确地将单个事件对接到图像，但它们在需要多模态推理处理序列的主要任务上的表现仍不令人满意。我们进一步分析了语言结构选择、事件间距离以及 Chain-of-Thought 提示的影响。我们的分析揭示了在架构设计、训练目标和/或训练后方法方面可提升时间推理能力的未来改进方向。

**贡献** (1) 我们介绍了 TempVS，这是一个用于评估 MLLMs 在图像序列中的多事件时间定位和推理能力的新基准。(2) 我们对来自不同模型家族和规模的 38 个 MLLMs 进行了广泛评估，突出了与人工标注者相比的性能差距。(3) 我们在评估结果和细粒度分析中的发现表明未来改进的潜在路径。

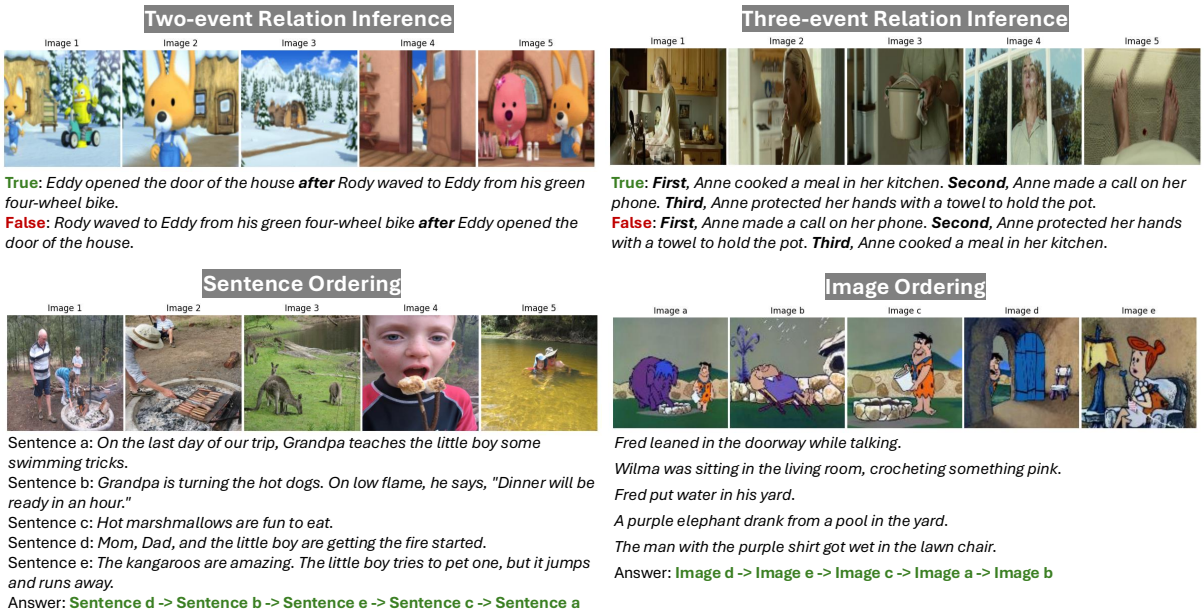



Figure 1: TempVS 基准测试主要测试中的说明性例子。额外的例子在 **??** 中提供。

## 2 相关工作

**多模态大型语言模型** 大型语言模型（LLMs）的进展 (**????**) 推动了多模态 LLMs 的发展，这些模型同时处理视觉和文本信息。最先进的多模态语言模型（MLLMs）(**???????**) 基于 LLMs，集成了一个视觉编码器和一个连接模块。这些模型在许多下游任务中超越了早期一代的多模态模型，这些模型通常基于 BERT 类型的架构 (**?**) 。虽然一些研究集中于训练 MLLMs 以使用交错的图像-文本数据集解释多个图像 (**????**) ，但其理解和推理顺序视觉数据中多事件时间关系的能力仍然很少被探索。

**多图像基准测试** 多图像理解需要多模态大模型（MLLMs）比较、分析和解释多幅图像之间的关系 (**?**) 。诸如 NLVR2 (**?**) 、BLINK (**?**) 、SEED-Bench-2 (**?**) 、Mantis-Eval (**?**) 和 MMT-Bench (**?**) 之类的基准涵盖了一部分多图像任务，重点在于评估模型识别多幅图像间相似性和变化性的能力。DEMON (**?**) 评估 MLLMs 的演示指令跟随能力。Mementos (**?**) 旨在检测为连续图像生成的描述性文本中的对象和行为幻觉。MileBench (**?**) 评估 MLLMs 在长上下文中的性能。MIBench (**?**) 评估多图像指令、多模态知识查询和上下文学习中 MLLM 的能力。MuirBench (**?**) 是一个具有不可回答对照项的综合多图像理解基准，用于测试 MLLMs 的稳健性。MMIU (**?**) 包含大量测试问题，涵盖多种多图像任务和关系。我们提出了 TempVS，这是首个专为图像序列中的多事件时间理解和推理而设计的基准。特别是，它设计旨在避免通过依赖单张图像/帧和常识推理来绕过文本和图像中时间信息的全面整合。

## 3 TempVS 基准测试

TempVS 通过评估事件之间时间关系的文本描述与图像序列的视觉事件顺序之间的一致性，来评估模型理解和推理时间关系的能力。为此，我们创建了三个主要测试（MT）：事件关系推理、句子排序和图像排序。此外，为了研究模型的困难是由于时间理解的挑战还是更基础的落地问题，我们为主要测试创建了相应的落地测试（GT）（§**??** ）。我们在 §3.2 中展示了 TempVS 基准任务的策划过程和统计数据。

TempVS 是从现有的将图像序列与叙述性说明文字配对的数据集中构建的。由于说明文字在细节层次上可能有所不同，我们使用了源数据中的原始说明文字，以及从说明文字中提取主要事件的简化说明文字（我们在 §3.2 中描述了这个过程）。在下文中，我们将一个由 $n$ 张图像组成的图像序列 $S$ 、其相应的说明文字，以及提取出的事件表示为：

$$\mathcal{S} = [(I_1, C_1, E_1), (I_2, C_2, E_2), \ldots, (I_n, C_n, E_n)],$$

，其中 $I_i$ 表示第 $i$ 张图像，$C_i$ 是其关联的原始说明文字，$E_i$ 则是提取出的事件（通过简化说明文字表达）。

### 3.1 主要测试和接地测试

**MT1：事件关系推断** MT1 通过图像序列和文本描述来评估模型对事件时间顺序的理解。文本通过副词性标志（如 after 或 before），或者通过句子的表面顺序明确地描述事件之间的时间

| Statement Template ( $i < j < k$ in image sequence) | | |
|---|---|---|
| Two-event | Pos: $E_j$ after $E_i$ | Neg: $E_i$ after $E_j$ |
| | Pos: $E_i$ before $E_j$ | Neg: $E_j$ before $E_i$ |
| | Pos: $E_j$. Earlier, $E_i$ | Neg: $E_i$. Ealier, $E_j$ |
| | Pos: $E_i$. Then, $E_j$ | Neg: $E_j$. Then, $E_i$ |
| | Pos: $E_i$. $E_j$. | Neg: $E_j$. $E_i$. |
| Three-event | Pos: $E_i$ before $E_j$, and after that, $E_k$<br>Neg: $E_k$ before $E_i$, and after that, $E_j$ | |
| | Pos: $E_i$. Later, $E_j$.Finally, $E_k$.<br>Neg: $E_j$. Later, $E_i$.Finally, $E_k$. | |
| | Pos: First, $E_i$. Second, $E_j$.Third, $E_k$.<br>Neg: First, $E_i$. Second, $E_k$.Third, $E_j$. | |
| | Pos: $E_i$. $E_j$. $E_k$. | Neg: $E_k$. $E_j$. $E_i$. |

Table 1: 用于 MT1 事件关系推理的正面（Pos）和负面（Neg）陈述模板。

关系。从长度为 $n$ 的图像序列中，我们选择（1）两个对 $\{(I_i, E_i), (I_j, E_j)\}$，其中 $i < j \leq n$，确保 $I_i$ 出现在 $I_j$ 之前，且 $E_i$ 发生在 $E_j$ 之前；或者（2）三个对 $\{(I_i, E_i), (I_j, E_j), (I_k, E_k)\}$，其中 $i < j < k \leq n$，并遵循相同的排序约束。这些事件图像对在序列中不必相邻，导致它们之间距离的变化。然后，我们通过应用 Table 1 中的模板来生成正面和负面陈述，以描述这些事件之间的时间关系。[1] 负面陈述保持与正面陈述相同的事件子句和时间连词，同时交换这些子句的位置，使文本表现出不同的时间顺序。[2] 最后，我们推导出包含原始图像序列、正面或负面陈述以及相应答案（True 或 False）的三元组。

**MT2: 句子排序** 这一任务评估模型是否能够根据给定图像序列中事件的时间顺序正确地重新排列一组打乱的句子。因此，MT2 不仅需要理解事件之间的时间关系，还需要考虑文本的连贯性和流畅性。模型需要根据一个有序的图像序列和一组打乱的事件描述，从五个给定选项中选择正确的句子顺序。我们创建了同时包含原始标题（$C$）和提取事件（$E$）的版本。

**MT3: 图像排序** 这个任务与 MT2 相似，但重点在视觉模式，要求模型根据给定的文本描述将一组被打乱的图像重新排列为正确的时间顺序。与 MT2 类似，我们还研究不同的文本风格（即原始字幕或提取的事件描述）是否会影响模型确定正确顺序的能力。

解决其中一个主要测试的前提条件是，我们假设 MLLMs 应该能够将事件描述与多图像序列中的对应图像相匹配。具体而言，给定事件描述 $E_i$ 或标题 $C_i$ 和图像序列 $[I_1, I_2, \ldots, I_n]$，模型需要识别与给定文本描述最匹配的图像的索引（即，$I_i$）。进行基础测试的动机如下：如果模型通过基础测试但未通过相应的主要测试，这表明模型在理解时间顺序方面存在困难，即使它能够准确识别和关联视觉和语言元素。相反，如果模型在主要测试中表现良好但未通过相应的基础测试，这可能表明其成功不是来自真正的时间基础或推理，而是依赖于训练中学习的统计模式、相关性或系统偏差。

### 3.2 基准测试策划与统计

鉴于我们评估模型是否理解跨图像序列和语言的事件时间顺序的目标，我们需要包含多个图像的数据，这些图像形成一个呈现事件的时间序列。我们选择了四个视觉故事数据集：FlintstonesSV (**?**)、PororoSV (**?**)、VIST (**?**) 和 VWP (**?**)。FlintstonesSV 和 PororoSV 旨在用于故事可视化，包含从卡通动画中获取的注释帧。[3] VIST，构建用于视觉故事讲述，源自 Flickr 相册中用户上传的日常生活照片。VWP 具有电影场景序列，并配有对齐的简介。该集合拥有丰富的风格和多样的领域，在评估多模态语言模型的多事件时间基础和推理能力方面发挥着关键作用。

为了确保每个图像序列包含足够数量的以角色为中心的视觉事件，我们使用 Detectron2 识别并保留在至少 60% 图像中可以检测到 PERSON 的图像序列。为避免任意两个事件之间的时间重叠，我们移除任何包含静态动词如“属于”、“爱”和“存在”等的字幕的图像序列。为减少歧义，我们删除以代词开头的字幕，并计算字幕之间的 BERTScore (**?**)，省略具有高度相似字幕的序列。同样地，我们根据它们的 CLIP (**?**) 嵌入之间的余弦相似性去除图像高度相似的图像序列。使用 GPT-4 (**?**) 从原始字幕（$C$）提取简单事件（$E$）；移除任何从中无法提取事件的字幕序列。有关提取过程的详细信息，请参见附录 .1。为了确保 $\{(I_i, E_i), (I_j, E_j)\}$ 中的每个图像-事件对保持独特性，我们使用 CLIP 计算不同图像-文本组合之间的跨模态相似性。基于阈值过滤模糊对，确保对内相似性明显高于对间相似性。一个类似的过程应用于三张图像-事件对的集合 $\{(I_i, E_i), (I_j, E_j), (I_k, E_k)\}$。在 **??** 中，我们提供了静态动词列表、相似性阈值、用于从原始字幕中提取事件的提示以及每个过滤步骤后的数据集统计。

[1] 我们仅使用文本事件来构建 MT1 中的陈述，忽略图片说明。

[2] 对于包含三个事件的否定陈述，我们从五种组合中随机选择一种与肯定陈述不同的组合。

[3] 对于 FlintstonesSV 和 PororoSV 中的主要角色，我们提供了他们外貌的描述，以便与他们的名字相匹配。

| | MT1 (two) | MT1 (three) | MT2 (event) | MT2 (caption) | MT3 |
|---|---|---|---|---|---|
| FlintstonesSV | 2,104 | 916 | 501 | 485 | 565 |
| PororoSV | 864 | 172 | 320 | 326 | 395 |
| VWP | 850 | 208 | 274 | 256 | 316 |
| VIST | 3,742 | 830 | 708 | 551 | 809 |
| TempVS | 7,560 | 2,126 | 1,803 | 1,618 | 2,085 |

Table 2: TempVS 基准统计数据：在 MT1 中，数字表示总语句数；在 MT2 中，数字表示图像序列及对应的顺序打乱的句子集；在 MT3 中，数字表示文本事件或字幕及其关联的顺序打乱的图像集。

**提示和选项生成** 在过滤数据集之后，我们通过将事件与 Table 1 中显示的模板连接起来为 MT1 创建正面和负面陈述。我们依次将每个模板应用于数据集中相应的时间关系组，确保声明类型的均匀分布。此外，我们使用 ChatGPT (?) 为不同任务生成各种提示组件的变体，包括任务说明、答案要求和响应格式（见 Appendix B ），这总共产生了 328 种可能的提示变体。通过在提示中引入足够的多样性，我们降低了结果受到特定提示形式 (?) 影响的风险。所有测试都以多项选择题形式进行。在 MT1 中，选项是 “True” 和 “False”，在样本中交替出现位置（例如，A. True；B. False。和 A. False；B. True。）以防止位置偏差 (?) 。在 MT2 和 MT3 中，一个正确的序列与四个随机打乱的错误序列一起展示，选项标记为 “A” 到 “E”。基础测试使用图像索引作为答案选项。为了确保公平评估，正确答案在整个基准测试中均匀分布在各个选项中。

**质量控制** 为了阻止利用语言偏见的 “盲目” 模型，我们在基准测试中过滤掉仅基于语言模态就可以轻松解决的例子。我们应用了三种单模态 LLMs Phi-3.5-mini-instruct [4B] (?) 、Llama-3.1 [8B] (?) 和 Qwen-2.5-instruct [72B] (?) ，它们是当前 MLLMs 中流行的 LLM 主干。在 MT1 和 MT2 中，如果至少有两个 LLM 能在没有视觉输入的情况下正确回答问题，我们就舍弃该样本。作者进行了人工检查，以排除模糊图像、语法错误和/或语义不合理的陈述，以及图像序列与文本不匹配的情况。

**基准统计** TempVS 由 2,085 个不同的图像序列组成，每个序列都有对应的原始标题和提取的事件。Table 2 展示了 TempVS 中每个数据来源的每项任务的统计数据。基准测试中的大多数图像序列每个包含 5 张图像，除了来自 VWP 的 61 个序列，它们包含 6 到 9 张图像。

# 4 实验

## 4.1 实验设置

我们评估了一系列多样化的最先进模型，具有不同的规模（从 0.5B 到 78B）和不同的视觉和大型语言模型（LLM）骨干。我们选择 DeepSeek-vl2 [3B/16B] (?) ，InternVL2.5 [1B/8B/26B/78B] (?) ，Janus-Pro [1B/7B] (?) ，LLaVA-NeXT-Interleave [0.5B/7B] (?) ，LLaVA-OneVision [0.5B/7B/72B] (?) ，LLaVA-NeXT-Video [7B/34B] (?) ，LongVA [7B] (?) ，Mantis[8B] (?) ，Phi-3-vision [4B]，Phi-3.5-vision [4B] (?) ，和 Qwen2-VL [2B/7B/72B] (?) 。我们还评估了 GPT-4o [2024-11-20]。实现细节见附录 C.1 。[4]

对于多项选择题，我们使用准确率作为衡量指标来评估 MLLMs 的性能。准确率得分被报告用于主要测试（MT）及其对应的基础测试（GT）。在 MT1 中，每个问题与两个或三个事件相关，每个单独的事件都有一个 GT。在 MT2 和 MT3 中，每个问题对应的基础测试数量等于序列中图像的数量。此外，我们引入了一个更严格的指标 GT $_{strict}$ ，它评估模型通过所有对应基础测试的图像序列数量。为了研究主要测试和基础测试性能之间的关系，我们进一步报告 MT|GT $_{strict}$ ，其中，只有当模型通过所有对应基础测试时，其在主要测试上的成功才被视为有效。

对 21 个最先进的 MLLM 的结果如 Table 3 所示。关于所有测试的 38 个 MLLM 的完整量化结果，详见附录 C.2 。我们对实证结果的主要观察如下：

**即使对于最新的多模态大模型来说，TempVS 也是具有挑战性的** InternVL2.5-26B-MPO 在两事件关系推理上取得了最高的性能，而 GPT-4o 在三事件关系推理中领先。InternVL2.5-78B-MPO 在句子排序和图像排序任务中表现优于其他模型。然而，大多数参数小于或等于 7B 的模型表现出随机概率的准确率，大约是 50 %（对于 MT1）和 20 %（对于 MT2 和 MT3）。大多数 MLLM 在两事件和三事件关系推理任务上表现相似，而对于最强的模型（如 InternVL2.5[26B/78B]、LLaVA-OneVision-ov-72B 和 GPT-4o），它们在理解三事件陈述方面的表现甚至略好。此外，句子排序是一个相对简单的任务（最高准确率为 86.3 %），而图像排序则

[4]在评估中，我们将序列中的多张图像水平组合为一张图像。在我们的初步实验中，我们尝试将图像逐一顺序输入模型，并观察到与将它们合并为单个输入相比，性能差异不大（详见 Table 12 的结果）。在发布的 TempVS 基准中，既提供了单独图像输入，也提供了组合的多图像输入。

| | | Two-event Relation (MT1) | | | Three-event Relation (MT1) | | | Sentence Ordering - event (MT2) | | | Sentence Ordering - caption (MT2) | | | Image Ordering - event (MT3) | | Image Ordering - caption (MT3) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GT $_s$ | MT | MT\|GT $_s$ | GT $_s$ | MT | MT\|GT $_s$ | GT $_s$ | MT | MT\|GT $_s$ | GT $_s$ | MT | MT\|GT $_s$ | MT | MT\|GT $_s$ | MT | MT\|GT $_s$ |
| Random | | 4 | 50 | | 0.6 | 50 | | 0.032 | 20 | | 0.032 | 20 | | 20 | | 20 | |
| DeepSeek-vl2 | 3B | 20.8 | 49.7 | 49.7 | 10.5 | 49.8 | 50.6 | 0.8 | 19.5 | 25.0 | 1.5 | 18.2 | 17.9 | 20.4 | 42.9 | 20.7 | 12.5 |
| | 16B | 14.2 | 43.1 | 42.2 | 6.7 | 44.4 | 44.4 | 0.4 | 15.7 | 14.3 | 0.6 | 17.1 | 18.2 | 16.6 | 0.0 | 15.5 | 0.0 |
| InternVL2.5 | 26B | 46.0 | 57.1 | 58.4 | 39.6 | 58.4 | 60.2 | 10.2 | 56.7 | 73.6 | 13.1 | 63.0 | 76.9 | 26.7 | 27.3 | 31.7 | 35.9 |
| | 26B-MPO | 51.3 | 60.3 | 62.1 | 46.0 | 62.1 | 64.7 | 12.6 | 69.9 | 90.8 | 17.0 | 76.9 | 87.3 | 34.4 | 39.7 | 39.5 | 43.7 |
| | 78B | 51.6 | 54.2 | 55.3 | 47.3 | 56.8 | 57.0 | 13.6 | 67.0 | 84.9 | 18.5 | 71.1 | 83.9 | 31.1 | 40.0 | 38.5 | 47.1 |
| | 78B-MPO | 58.8 | 58.5 | 59.9 | 56.5 | 61.4 | 62.6 | 18.4 | 79.8 | 96.6 | 25.9 | 86.3 | 96.4 | 41.0 | 48.8 | 53.8 | 69.7 |
| Janus-Pro | 1B | 2.7 | 48.3 | 48.1 | 0.7 | 46.5 | 42.6 | 0.0 | 18.6 | - | 0.1 | 19.8 | 0.0 | 22.5 | 0.1 | 22.3 | 0.0 |
| | 7B | 4.3 | 35.1 | 34.1 | 0.4 | 32.9 | 39.3 | 0.0 | 17.1 | - | 0.0 | 15.3 | - | 20.9 | - | 21.0 | - |
| LLaVA-NeXT-Interleave | 0.5B | 2.5 | 49.8 | 47.8 | 0.2 | 50.4 | 50.0 | 0.0 | 20.7 | - | 0.0 | 20.7 | - | 20.2 | - | 20.8 | - |
| | 7B | 13.0 | 51.6 | 52.4 | 7.2 | 50.1 | 49.8 | 0.3 | 25.1 | 16.7 | 0.4 | 27.0 | 0.0 | 20.9 | 16.7 | 20.0 | 22.2 |
| LLaVA-OneVision-ov | 0.5B | 8.6 | 45.3 | 45.5 | 2.9 | 48.1 | 47.6 | 0.1 | 18.8 | 0.0 | 0.1 | 18.4 | 0.0 | 19.4 | 100.0 | 19.0 | 0.0 |
| | 7B | 32.8 | 56.0 | 58.0 | 26.0 | 57.5 | 59.8 | 4.5 | 44.2 | 41.2 | 6.8 | 46.9 | 47.6 | 21.3 | 14.3 | 21.6 | 19.8 |
| | 72B | 46.4 | 59.3 | 62.1 | 40.5 | 61.5 | 63.5 | 9.8 | 65.2 | 81.8 | 14.0 | 75.1 | 86.6 | 27.6 | 31.8 | 29.1 | 36.5 |
| LLaVA-NeXT-Video | 7B | 5.8 | 46.0 | 46.0 | 1.4 | 44.9 | 47.0 | 0.0 | 19.0 | - | 0.0 | 18.2 | - | 21.0 | - | 21.3 | - |
| | 34B | 6.5 | 58.5 | 58.4 | 1.6 | 59.5 | 57.6 | 0.1 | 31.8 | 100.0 | 0.0 | 33.4 | - | 19.8 | - | 20.0 | - |
| LongVA | 7B | 8.7 | 54.7 | 56.1 | 2.1 | 56.0 | 61.7 | 0.2 | 34.2 | 66.7 | 0.2 | 35.3 | 50.0 | 19.5 | 0.1 | 19.0 | - |
| Mantis-Idefics | 8B | 12.2 | 51.9 | 53.3 | 4.1 | 52.0 | 51.6 | 0.1 | 22.2 | 0.0 | 0.2 | 20.8 | 0.0 | 18.6 | 0.0 | 19.2 | 50.0 |
| Phi-3.5-vision | 3.4B | 4.0 | 49.0 | 47.7 | 0.8 | 48.8 | 48.3 | 0.0 | 23.1 | - | 0.0 | 25.4 | - | 19.2 | - | 18.3 | - |
| Qwen2-VL-Instruct | 7B | 32.4 | 54.0 | 55.4 | 21.2 | 53.6 | 55.3 | 3.4 | 42.5 | 64.6 | 4.4 | 44.6 | 61.3 | 23.1 | 20.4 | 24.6 | 35.9 |
| | 72B | 31.7 | 54.0 | 56.4 | 20.6 | 55.6 | 60.6 | 3.7 | 46.3 | 64.3 | 5.0 | 55.1 | 70.0 | 26.5 | 47.3 | 28.1 | 47.4 |
| GPT-4o | API | 60.3 | 58.3 | 60.1 | 57.0 | 64.5 | 66.4 | 18.6 | 53.4 | 53.9 | 28.6 | 61.5 | 55.3 | 22.6 | 23.5 | 23.0 | 23.5 |

Table 3: 在 TempVS 基准测试中，针对严格基础测试（GT $_s$）、主要测试（MT）以及所有对应的基础测试通过时的主要测试（MT|GT $_s$），11 个流行的 MLLMs 系列的 21 个变体在零样本平均准确率上的表现。每个指标的最佳模型用粗体标出，第二佳模型用下划线标出。

是一个显著更具挑战性的任务（最高准确率仅为 53.8 %）。在句子和图像排序任务中，GPT-4o 的表现明显逊色于几个表现最佳的开源模型。在语言类型方面，我们发现使用原始标题的句子和图像排序比使用提取的事件更容易。这可能表明模型可能利用了原始标题中的额外上下文细节和时间线索，而这些在简单的提取事件描述中是不可用的。

如 **??** 所示，模型的精度通常随着模型尺寸的增加而提高。然而，随着模型变得更大，边际收益减少。在两事件和三事件关系推理中，这一效应更加明显，较小的模型（7B 或 26B）已经在竞争中获得了更高的精度，有时甚至超过了较大的模型（> 70B）。另外，两种排序任务在较小和较大模型之间表现出更大的性能差距，这可以归因于较大 MLLM 中更强大 LLM 骨架的优秀远程推理能力。令人惊讶的是，DeepSeek-VL2 [3B/16B] 和 Janus-Pro [1B/7B] 是例外，因为其较小的模型在大多数情况下表现优于较大的模型。

我们的结果还强调了高质量后训练的重要性：在相同的模型规模下，InternVL2.5-MPO 在所有评估任务中始终优于 InternVL2.5，尤其是当模型参数超过 7B 时。这些结果表明混合偏好优化（MPO）(**?**) 有效提升了整体多模态时序理解和推理能力。同样，使用直接偏好优化（DPO）微调的模型始终优于仅通过 SFT 的模型。其他比较 LLaVA-NeXT-Video 和 LongVA 系列的结果在附录 C.2 中提供。我们还看到指令微调对复杂推理任务（如图像排序）产生了积极影响（例如，Qwen2-VL-Instruct 在所有任务中均优于 Qwen2-VL）。在 MT1 中，我们观察到 MT 和 MT|GT 之间有轻微差异。这表明，即使模型能够准确识别序列中每一个事件对应的图像，它仍可能缺乏理解文字中事件时间顺序的能力。简而言之，需要不同的能力来对文本描述进行定位和推理其时间关系。在所有定位测试中，GPT-4o 表现最好，但在两个排序任务（MT2 和 MT3）中明显落后于顶尖的开源模型，如 InternVL2.5-MPO[26B/78B]。在 MT2 和 MT3 中，对于大多数大型 MLLMs，MT|GT 准确率在 MT 之上有显著提高，证实了视觉定位在句子或图像排序任务中的基本作用。例如，InternVL2.5-78B-MPO 的 MT|GT 准确率在 MT2 中比 MT 高 16.8（事件）和 10.1（字幕），在 MT3 中高 7.8（事件）和 15.9（字幕）。与此同时，对于像 DeepSeek-vl2[3B] 这样的小模型，MT|GT 准确率有时甚至比 MT 准确率更低。这表明这些模型的推理能力在某种程度上依赖于其定位能力，尽管这不太可能是影响性能的唯一因素。

| | # Image seqs | Accuracy | Fleiss' kappa |
|---|---|---|---|
| Two-event Relation | 60 | 82.5 | 0.728 |
| Three-event Relation | 60 | 81.6 | 0.689 |
| Sentence Ordering (event) | 40 | 81.2 | 0.751 |
| Sentence Ordering (caption) | 40 | 89.3 | 0.764 |
| Image Ordering (event) | 40 | 79.1 | 0.827 |
| Image Ordering (caption) | 40 | 77.9 | 0.742 |

Table 4: 所有主要测试的人类评估结果。

为了评估 TempVS 的质量和估计其难度，我们对 280 个随机选取的图像序列进行了人工评估，这些序列涵盖了 MT1 的所有细粒度声明类型以及 MT2 和 MT3 的两种语言类型（$C$ 和 $E$）。我们在 Prolific 平台上招募了 36 名标注员，并为每个图像序列收集了三份回应，总共收集了 840 份回应。详细信息在 Appendix A 中提供。

在人类平均表现（Table 4）和 SOTA MLLMs 之间存在很大差距。在双事件和三事件关系推理以及图像排序任务中，仍有很大的改进空间。然而，对于句子排序任务，最强的模型 InternVL2.5-78B-MPO 已经接近人类表现。在

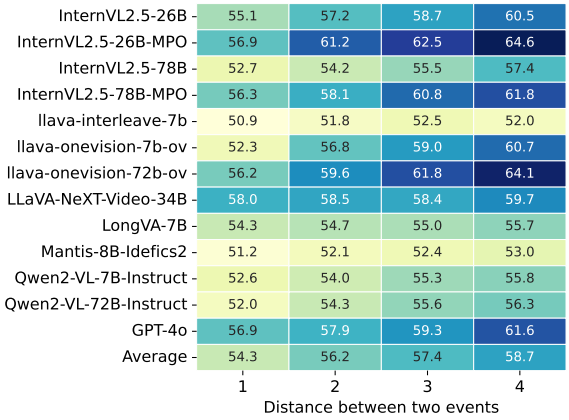



Figure 2: 在序列中两个事件之间有不同距离时，两事件关系推理任务的准确性。

主要任务上，人的一致性相当高，Fleiss' kappa (**?**) 在所有方面均高于 0.68。

### 4.2 进一步分析

**时间表达的影响** 我们进一步分析模型如何理解和推理关于两事件和三事件关系推断任务（MT1）中不同类型的陈述。我们选择这两个任务的前五名模型进行比较（Table 5）。在比较显式和隐式时间事件陈述时（参见 Table 1），我们观察到模型在处理前者时表现一致地更好。这可能是因为句子中时间副词或连接词的存在有助于澄清事件发生的顺序。前五名模型在正面例子上总是取得比负面例子更高的准确率。尽管在我们的基准测试中，“真”和“假”在选项“A”和“B”中均衡分布，但模型表现出更频繁预测“真”的倾向。通过纳入对抗性样本，我们的 TempVS 基准测试有效地揭示了 MLLMs 对于某些答案的偏向，提供了对其组合时间推理能力的稳健评估。

我们还观察到模型在涉及 *在之前*（相应地为 *之后*）的明确标记的时间关系上，以及在 *然后*（相应地为 *早期*）上的更好表现。这些互补的时间副词对之间的关键区别在于，使用 *之前* 和 *然后* 时，文本中事件的顺序与图像序列中的顺序相对应，而使用 *之后* / *更早* 时则不然。因此，我们看到了一些 *图像效应* 的证据：当事件的表面顺序反映其实际顺序（在视觉模式中）时，时间关系对模型来说更容易。这与叙述理解的语篇处理和心理语言学文献中的类似发现相呼应。它还指出了未来在细粒度多模态基准研究中的一个重要途径，即在两个模态的表面特征没有完全对齐的情况下的研究。

**两个事件之间的距离** Figure 2 说明了两个事件关系推理任务的准确性，随着原始序列中事件间的距离从一增加到四，准确性也随之变化。几乎所有模型在距离增加时性能都有所提升。一方面，数据集中较远的视觉事件通常在视觉上更为区别明显。另一方面，尽管在输入的图像之间插入了分隔符，模型仍可能无法有效分开两张距离较近的图像。

**使用思维链提示** Chain-of-Thought（CoT, **?**）是一种广泛使用的方法，通过允许模型在产生最终答案之前生成中间的推理步骤，以增强模型的推理能力。因此，它也可能提升模型在 TempVS 上的时间推理技能。我们使用 InternVL2.5-78B 和 GPT-4o 进行 CoT 实验。详细的提示列在附录 B 中。如 Table 6 所示，CoT 在句子和图像排序上带来了显著的提高，但在事件关系推理（MT1）上的改善有限。这表明逐步推理在排序任务上的潜力。然而，简单的 CoT 对事件关系推理无帮助。我们将把增强模型对这一复杂任务的理解及提升其时间推理能力的方法的研究留待未来工作。

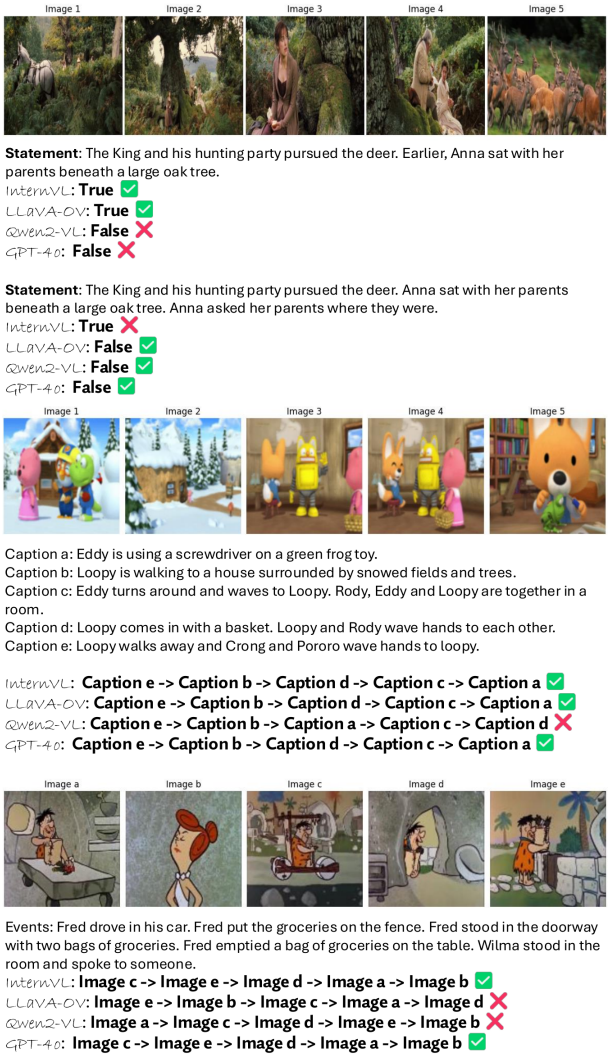



Figure 3: 四个具有代表性的 MLLM 在 TempVS 上的定性案例。

| Tasks | Models | InternVL2.5-26B-MPO | | InternVL2.5-78B-MPO | | llava-onevision-72b-ov | | LLaVA-NeXT-Video-34B | | GPT-4o | | Top-5 Average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Statement Type | Pos | Neg | Pos | Neg | Pos | Neg | Pos | Neg | Pos | Neg | Pos | Neg | Both |
| Two-event Relation | after | 59.9 | 62.9 | 70.7 | 49.1 | 71.4 | 50.6 | 77.9 | 37.1 | 56.2 | 60.4 | 67.2 | 52.0 | 59.6 |
| | before | 66.4 | 66.7 | 71.6 | 57.9 | 70.6 | 59.0 | 77.1 | 41.4 | 53.3 | 73.4 | 67.8 | 59.7 | 63.8 |
| | earlier | 70.8 | 47.0 | 74.3 | 36.8 | 74.1 | 41.5 | 85.9 | 28.8 | 69.1 | 38.5 | 74.8 | 38.5 | 56.7 |
| | then | 65.5 | 59.3 | 73.3 | 44.8 | 71.3 | 50.0 | 86.4 | 35.2 | 61.0 | 63.1 | 71.5 | 50.5 | 61.0 |
| | implict | 65.3 | 38.7 | 74.4 | 31.6 | 74.4 | 30.1 | 86.4 | 28.3 | 64.0 | 43.4 | 72.9 | 34.4 | 53.6 |
| Three-event Relation | before/after | 66.7 | 72.3 | 70.7 | 64.0 | 77.7 | 54.6 | 71.2 | 72.0 | 55.7 | 78.2 | 65.1 | 67.2 | 66.2 |
| | first/second/third | 55.2 | 70.4 | 65.4 | 67.5 | 76.5 | 45.9 | 68.4 | 64.1 | 50.4 | 78.9 | 60.4 | 65.5 | 63.0 |
| | later/finally | 75.2 | 56.3 | 87.5 | 32.5 | 77.4 | 52.8 | 100.0 | 0.1 | 69.4 | 68.3 | 79.8 | 43.1 | 61.5 |
| | implict | 73.3 | 27.1 | 82.6 | 20.3 | 78.7 | 28.4 | 100.0 | 0.3 | 70.3 | 43.7 | 76.3 | 28.4 | 52.4 |

Table 5: 在两事件和三事件关系推理任务中，不同陈述类型的细粒度准确性。“Pos”表示正例，而“Neg”代表负例。在每对正负陈述中，较高的准确度分数用灰色突出显示。

| | InternVL2.5-78B | | | GPT-4o | | |
|---|---|---|---|---|---|---|
| | w/o. CoT | w. CoT | Δ | w/o. CoT | w. CoT | Δ |
| Two-event relation (MT1) | 54.23 | 56.13 | +1.90 | 58.27 | 60.43 | +2.16 |
| Three-event relation (MT1) | 56.79 | 57.20 | +0.41 | 64.52 | 63.38 | -1.14 |
| Sentence Ordering (MT2) | 71.09 | 85.42 | +14.33 | 61.50 | 81.41 | +19.91 |
| Image Ordering (MT3) | 38.49 | 47.54 | +9.05 | 22.97 | 33.77 | +10.80 |

Table 6: 比较有无思维链（CoT）时的模型性能。

**定性案例分析** Figure 3 展示了四个具有代表性的开源多模态大语言模型（MLLMs）之间的特定任务定性结果，即 InternVL2.5-78B-MPO、LLaVA-OneVision-ov-72b、Qwen2-VL-Instruct-72B 和 GPT-4o。进一步的分析表明，当前模型表现不佳的原因在于：(1) 事件定位能力不足，模型无法识别与事件相对应的正确图像；(2) 指令跟随能力差，导致输出不符合所需格式；(3) 对完整序列的理解有限，导致选择的顺序只有部分正确；以及 (4) 在时间推理需要时，难以区分细微不同的图片。

在本文中，我们提出了一项新颖且具挑战性的基准 TempVS，用于评估多模态、多事件中图像序列内 MLLMs 的时间推理能力。在对 38 个先进的 MLLMs 进行评估后，我们发现当前模型通常在推理时间关系以及根据叙述将打乱的图像重新排列到正确顺序方面存在困难。对语言结构、事件距离和连锁思维推理的进一步分析揭示了未来工作的有前景的方向。我们的研究通过揭示多图像场景中多事件时间推理的弱点，为 MLLM 的开发做出了贡献，而 TempVS 为进一步研究提供了宝贵的资源。

**局限性** 尽管我们仔细检查了公开可用的资源库、技术报告和论文，但我们没有找到证据表明所评估的多模态大语言模型（MLLMs）在其训练过程中使用了 TempVS 中包含的数据。然而，对于尚未完全公开或明确说明其训练数据的模型（如 GPT-4o），数据泄漏和污染的可能性仍不明朗。这可能导致对其优势的高估。此外，我们聚焦于选择题，以确保结构化评价和明确的正确性标准，遵循现有的多模态基准 (**?????**) 。然而，其他类型的问题，例如开放性问题，特别是在多图像时间理解和推理中的开放性生成评估，仍值得探索。我们将这些留待未来研究。

在本研究中，我们使用了已发布的数据集和预训练的多模态大型语言模型，它们的使用不存在已知的重大伦理问题。然而，我们承认原始图像-字幕数据中的偏见可能会影响模型及其评估。我们的研究已获得我校伦理委员会的批准，确保符合人类标注过程的伦理指导方针。此外，所有收集的人类标注数据已进行了匿名化处理，以保护参与者的数据隐私和安全。

我们感谢匿名审稿人提供的宝贵意见和建设性反馈。我们也感谢 Letitia Parcalabescu 提出的深刻建议和富有建设性的讨论，这帮助我们提高了这项工作的质量。我们感谢 UU NLP 小组的成员在初步人工评估实验中的协助，以及他们对手稿早期草稿提出的有益建议。

我们过滤掉标题中出现任何形式的静态动词（如在 Table 7 中显示）的样本，以避免描述状态。相比之下，我们只保留含有动态动词的示例来描述动作。

| Base Form | Present Participle | 3rd Person Singular | Past Tense | Past Participle |
|---|---|---|---|---|
| wish | wishing | wishes | wished | wished |
| equal | equaling | equals | equaled | equaled |
| signify | signifying | signifies | signified | signified |
| feel | feeling | feels | felt | felt |
| involve | involving | involves | involved | involved |
| sense | sensing | senses | sensed | sensed |
| sound | sounding | sounds | sounded | sounded |
| detest | detesting | detests | detested | detested |
| want | wanting | wants | wanted | wanted |
| see | seeing | sees | saw | seen |
| forget | forgetting | forgets | forgot | forgot |
| matter | mattering | matters | mattered | mattered |
| contain | containing | contains | contained | contained |
| own | owning | owns | owned | owned |
| taste | tasting | tastes | tasted | tasted |
| dislike | disliking | dislikes | disliked | disliked |
| remember | remembering | remembers | remembered | remembered |
| suppose | supposing | supposes | supposed | supposed |
| resemble | resembling | resembles | resembled | resembled |
| think | thinking | thinks | thought | thought |
| envy | envying | envies | envied | envied |
| depend | depending | depends | depended | depended |
| hate | hating | hates | hated | hated |
| know | knowing | knows | knew | known |
| require | requiring | requires | required | required |
| love | loving | loves | loved | loved |
| appreciate | appreciating | appreciates | appreciated | appreciated |
| need | needing | needs | needed | needed |
| concern | concerning | concerns | concerned | concerned |
| span | spanning | spans | spanned | spanned |
| appear | appearing | appears | appeared | appeared |
| owe | owing | owes | owed | owed |
| weigh | weighing | weighs | weighed | weighed |
| disagree | disagreeing | disagrees | disagreed | disagreed |
| become | becoming | becomes | became | become |
| fear | fearing | fears | feared | feared |
| measure | measuring | measures | measured | measured |
| possess | possessing | possesses | possessed | possessed |
| like | liking | likes | liked | liked |
| look | looking | looks | looked | looked |
| imagine | imagining | imagines | imagined | imagined |
| mind | minding | minds | minded | minded |
| belong | belonging | belongs | belonged | belonged |
| loathe | loathing | loathes | loathed | loathed |
| lack | lacking | lacks | lacked | lacked |
| deserve | deserving | deserves | deserved | deserved |
| mean | meaning | means | meant | meant |
| promise | promising | promises | promised | promised |
| believe | believing | believes | believed | believed |
| prefer | preferring | prefers | preferred | preferred |
| cost | costing | costs | costed | costed |
| hope | hoping | hopes | hoped | hoped |
| recognize | recognizing | recognizes | recognized | recognized |
| include | including | includes | included | included |
| support | supporting | supports | supported | supported |
| understand | understanding | understands | understood | understood |
| comprise | comprising | comprises | comprised | comprised |
| agree | agreeing | agrees | agreed | agreed |
| realize | realizing | realizes | realized | realized |
| value | valuing | values | valued | valued |
| seem | seeming | seems | seemed | seemed |
| hear | hearing | hears | heard | heard |
| doubt | doubting | doubts | doubted | doubted |
| consist | consisting | consists | consisted | consisted |
| smell | smelling | smells | smelled | smelled |

Table 7: 静态动词完整列表。

由于各数据集的域导致图像和文本风格不同，我们通过人工检查和经验调节为每个数据集确定了文本相似度和图像相似度的阈值，如 Table 8 所示。例如，在 FlintstonesSV 中，对于文本相似度，当两个文本的 BERTScore 精确率和召回率都低于 0.98，并且它们的 F1 评分低于 0.96 时，认为它们是不相似的。同样，对于图像相似度，当两个图像的 CLIP 相似度评分低于 0.94 时，认为它们是不相似的。

| | BERTScore | | | CLIP Similarity |
|---|---|---|---|---|
| | precision | recall | f1 | |
| FlintstonesSV | <0.98 | <0.98 | <0.96 | <0.94 |
| PororoSV | <0.96 | <0.96 | <0.95 | <0.90 |
| VWP | <0.98 | <0.98 | <0.97 | <0.95 |
| VIST | <0.92 | <0.92 | <0.90 | <0.88 |

Table 8: 用于数据过滤的相似性阈值。

## .1 用于提取事件的提示

提取的事件与原始字幕之间的主要区别在于细节的层次：前者仅包含基本的事件信息，例如谁做了什么；后者则由人工注解者撰写成完整的故事/叙述。原始字幕通常包含诸如“在……之前”、“然后”或“最后”等时间连接词，这可能使模型能够根据文本提示推断多事件关系并自行排列句子顺序。

我们用来通过 GPT-4 从原始字幕中提取事件的提示如下：

> Given a caption: [the original caption here]. Extract a single, concise and clear event sentence from the provided caption. Ensure the returned sentence satisfies the following criteria: (1) The sentence should contain the event itself without phrases such as "the event is" or "event:". (2) If there are multiple events, extract only a single major event. (3) The sentence must contain only one clause with only one verb. (4) The event should be expressed in simple past tense. (5) If no event is detected, return 'NO_EVENT'.

在 Table 9 中，我们展示了每个数据过滤步骤后剩余图像序列的统计数据。

| | FlintstonesSV | PororoSV | VWP | VIST |
|---|---|---|---|---|
| Original image sequences | 24,433 | 11,444 | 12,627 | 49,700 |
| No stative verbs | 10,378 | 3,114 | 1,995 | 20,294 |
| No starting pronoun | 10,105 | 2,952 | 914 | 12,216 |
| No similar text | 3,092 | 1,952 | 815 | 2,906 |
| No similar image | 636 | 644 | 809 | 2,315 |
| No ambiguous image-text | 633 | 535 | 686 | 2,284 |
| No repetitive image sequences | 633 | 535 | 498 | 1,880 |
| Without No_EVENT | 612 | 535 | 417 | 1,292 |
| Final image sequences after manual check | 565 | 395 | 316 | 809 |
| Final two-event groups | 2104 | 864 | 850 | 3742 |
| Final three-event groups | 916 | 172 | 208 | 830 |

Table 9: 数据过滤过程中每个步骤的图像序列数量。

## A 人类表现调查

我们设计了三个问卷用于人类表现调查，分别对应三个主要任务：事件关系推理、句子排序和图像排序。为了确保参与者充分了解任务，我们在每个问卷的开头提供了任务说明和两个示例问题（如 Figure 4 所示）。此外，Figure 5 展示了参与者需要回答的示例问题。我们从 TempVS 基准中随机选择了 280 个图像序列，并从不同标注者那里收集了每个序列的三个反馈。MT1 的参与者需要完成 30 个问题，而其他两个排序任务的参与者各完成 20 个问题。中位完成时间约为 20 分钟，以确保长时间的集中注意力不会对参与者的判断产生负面影响。我们通过 Prolific 招募了 36 名标注者（18 名女性，18 名男性），每小时薪酬为 č17.1，他们都精通英语并且至少接受过大学水平的教育。

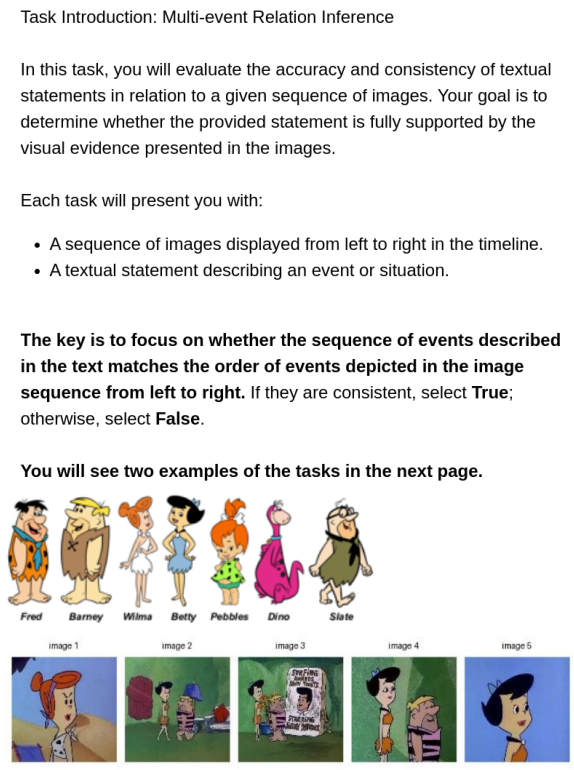



Figure 4: 人类问卷开始时显示的任务说明和示例问题的示例

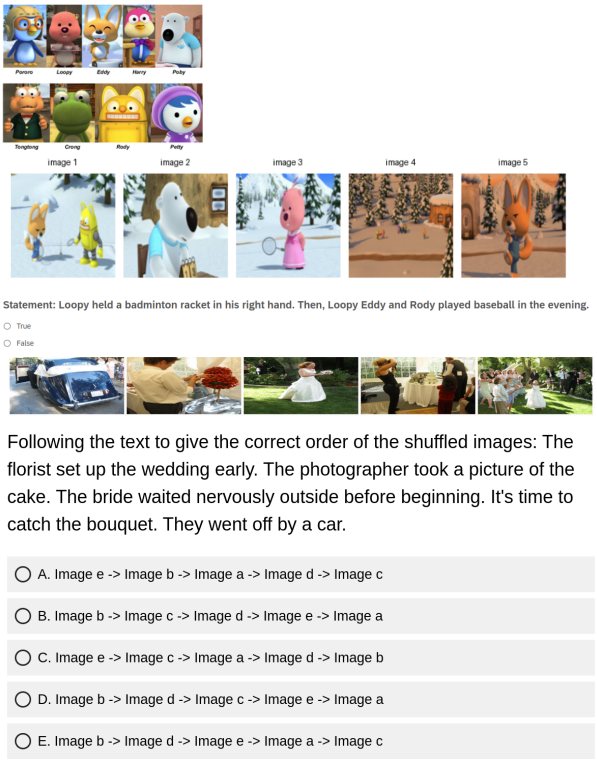



Figure 5: 人工标注者使用的示例问题界面

## B 提示

**提示模板** 我们用于每个任务的提示示例如下所示。我们的提示包括角色描述（仅针对来自 FlintstonesVS 和 PororoVS 的图像序列）、任务说明、问题文本、响应格式和选项（仅用于主要测试），以及一个指示答案开始的前缀。

GT: Grounding Test
User:
[Character Description] Description of character appearance in the images: Fred is chubby, has black hair, a large nose and wears an orange and black spotted short-sleeved loincloth with a blue scarf.
[Task Instruction] Pick the image from the image sequence that accurately represents the event. When making the choice, focus on the evidence presented in the sequence of images from left to right.
[Question Text] The event is: Fred was sitting in a room on a stool.
[Response Format] Submit the right number of the image in the sequence as your answer only without additional reasoning or repetition of the instructions.
The answer is:

MT1: Event Relation Inference
[Character Description] Description of character appearance in the images: Pororo is a gentoo penguin with an orange-yellow beak wearing a tan aviator's helmet and orange goggles.
[Task Instruction] Is the statement completely accurate and consistent with the content in the sequence of images? When making the choice, focus on the evidence presented in the sequence of images from left to right.
[Question Text] The statement is: A snowball dropped down a small snow hill after Pororo threw a snowball.
[Response Format] Submit only the right option letter as your answer, e.g., Option [Letter].
[Options] Options are: A. True; B. False.
The answer is:

MT2: Sentence Ordering
[Task Instruction] You are given an ordered image sequence and several sentences in a random order. Your task is to analyze the content of the sequence of images from left to right and rearrange the sentences into the correct chronological or logical order. Read the image sequence from left to right. Use the images' content to guide your sentence ordering. Avoid assumptions not supported by the image sequence or sentences.
[Question Text] The shuffled sentences are:
Sentence a: John prepares to shoot again and fires.
Sentence b: The shadowy figure is hit with the rifle blast.
Sentence c: John takes aim at a shadowy figure up above.
Sentence d: Ken stops the horses and prepares to leave the stagecoach.
Sentence e: Ken is approaching a house out on the prairie on his stagecoach with his two horses.
[Response Format] Do not provide explanations or repeat the prompt. Select from the following options and your answer should only be in the format: Option [Letter].
[Options]
A. Sentence c -> Sentence b -> Sentence a -> Sentence e -> Sentence d;
B. Sentence b -> Sentence e -> Sentence d -> Sentence a -> Sentence c;
C. Sentence e -> Sentence b -> Sentence c -> Sentence d -> Sentence a;
D. Sentence c -> Sentence a -> Sentence e -> Sentence d -> Sentence b;
E. Sentence d -> Sentence c -> Sentence e -> Sentence b -> Sentence a.
The answer is:

MT3: Image Ordering
[Task Instruction] The text narrates a story or event sequence. Use your vision-language reasoning to reorder the images to reflect the narrative structure. Carefully read the provided text. Focus on the events, actions, and details described to reorder the images logically.The images are labeled in order as Image a, Image b, Image c, Image d, and Image e, and so on if there are more photos.
[Question Text] The events are: Some people came for the family gathering today. The girls enjoyed some fruit. We played on the swings. The boys lounged in the chair. Grandpa put his grandson on his knee.
[Response Format] Select from the following options and your answer should be in the format: 'Option [Letter]'. Respond with the correct option only, avoiding any explanations or repetition.
[Options]
A. Image b -> Image c -> Image d -> Image a -> Image e;
B. Image e -> Image d -> Image c -> Image b -> Image a;
C. Image b -> Image e -> Image d -> Image a -> Image c;
D. Image a -> Image c -> Image e -> Image b -> Image d;
E. Image c -> Image b -> Image a -> Image e -> Image d.
The answer is:

对于提示中的每个组件，我们生成多个变体（如 Table 10 中所列）。通过组合不同的组件，我们最终生成了 328 个不同的提示。

**思维链提示**

| Task | CoT Prompt |
|---|---|
| MT1 | Analyze the provided image sequence to determine whether the following statement is True or False. When making the choice, carefully examine the evidence presented in the sequence of images from left to right. First, describe the key details and changes observed in each image. Then, explain how these details support or contradict the given statement. Finally, based on your step-by-step reasoning, conclude whether the statement is True or False. Ensure your response follows this format: the reasoning process should be enclosed within <think> and </think> tags, and the final answer should be enclosed within <answer> and </answer> tags. |
| MT2 | You are presented with an ordered image sequence and several sentences in random order. Your task is to determine the correct order of the sentences based on the context, events, or details observable in the images. Multiple-choice options are provided, with each option representing a possible sequence of the sentences. Think step by step, using the content of the images to guide your reasoning. Avoid assumptions not supported by the image sequence or the sentences.<br>Format your response as follows: Enclose your step-by-step reasoning process within <think> and </think> tags. Enclose the selected option (e.g., Option A, Option B) within <answer> and </answer> tags. |
| MT3 | Order the following images in the correct sequence based on the content of the story. Compare each image with the text description, carefully analyzing the sequence of events to determine the proper order of the unordered images.<br>Response should be in this format:<br><think> First, examine the text description to identify key events and their chronological order. Next, analyze each image to match it with the corresponding event described in the text. Consider visual cues, actions, and details in the images that indicate the progression of the story. Arrange the images accordingly to reflect the correct sequence of events. </think><br><answer> Option [Your Choice] </answer> |

Table 11: 不同任务的思维链提示。

# C 实验

## C.1 实现细节

我们评价了 38 个现有的多模态大模型（MLLM），用于多事件时间定位和推理，包括专有和开源模型。所有开源模型使用其在 HuggingFace 上提供的官方预训练版本进行评估。所评估开源模型的详细配置列在 Table 13 中。为最小化随机性，我们将温度设置为 0。开源模型的实验是在单个 NVIDIA H100 GPU 上进行的。对于超过 70B 参数的模型，我们采用混合精度（FP16），否则采用全精度。

根据 **?** 、**?** 、**?** 和 **?** ，我们将连续的图像合并为一个输入，用一条细白带将它们分开，并用一个索引标记每个图像（例如，当图像按顺序排列时标记为“图像 1”，或者当图像被打乱时标记为“图像 a”）。在初步实验中，我们测试了逐个顺序输入图像，但观察到与如 Table 12 所示的合并输入相比，性能差异很小。组合图像的使用也受到当前 MLLMs 的实际限制的驱动，其中许多只支持单一图像输入。此外，由于某些模型的上下文长度有限，同时输入多张图像可能导致令牌截断和信息丢失 (**?**) 。在发布的 TempVS 基准和评估工具包中，我们提供了单个图像输入和合并的多图像输入，以便人们可以根据自己的需要使用我们的基准来评估模型。

| | |
|---|---|
| Task Instruction - definition | |
| MT1 | - Is the statement completely accurate and consistent with the content in the sequence of images?<br>- Analyze the provided image sequence to determine whether the following statement is True or False.<br>- Review the sequential images and decide whether the provided statement is True or False.<br>- Is the statement entirely consistent and supported by the images in the sequence?<br>- Examine the visual evidence in the provided sequence of images to determine whether the statement is True or False.<br>- Does the statement fully match the information presented in the ordered images? Taking the content in the image seqeunce into account, can you decide whether the statement is True or False?<br>- Given the multiple images provided in order, can you select the correct answer from True or False considering statement? |
| MT2 | - You are given an ordered image sequence and several sentences in a random order. Your task is to analyze the content of the sequence of images from left to right and rearrange the sentences into the correct chronological or logical order.<br>- Here is an left-to-right image seqeunce and some sentences in random order. Your goal is to determine the correct order of the sentences based on the context, events, or details observable in the images. Consider the visual elements in the sequential images to infer the logical or temporal sequence.<br>- The images are presented in a correct order, but the sentences are not. Your task is to reorder the sentences to match the sequence of events or details in the images from left to right.<br>- You are provided with a sequence of images and several randomly ordered sentences. Your task is to: 1. Understand the context of the image sequence; 2. Identify how each sentence relates to the image; 3. Rearrange the sentences to form a coherent order.<br>- Using your multimodal understanding and reasoning of the ordered image sequence and the randomly shuffled sentences , arrange the sentences in the correct sequence to match the flow of events in the image seqeunce.<br>- Presented is an image sequence in the order from left to right along with several unordered sentences. Your task is to determine the correct sequence of the sentences by analyzing the context, events, or details depicted in the images. Use the visual elements in the images to deduce the logical or chronological order. |
| MT3 | - Rearrange the following images in the correct order based on content in the story.<br>- The provided paragraph describes a sequence of events. Arrange the images in the correct chronological order to match the story.<br>- The images are out of order compared to the text. Identify and reorder them to match the described sequence.<br>- The text narrates a story or event sequence. Use your vision-language reasoning to reorder the images to reflect the narrative structure.<br>- Using your understanding of the text and image content, arrange the images in the correct sequence to match the flow of events in the text.<br>- The paragraph describes events in a specific timeline. Use multimodal reasoning to reorder the images in the correct sequence. |
| GT | - Based on the event described, select the image that best matches it from the following options.<br>- Pick the image from the image sequence that accurately represents the event.<br>- Which is the most fitting image for the described event? Pick from the choices below.<br>- If you were to illustrate the event, which picture would you use?<br>- Find the image from the sequence of images that represents the event most precisely.<br>- Identify the picture that best represents the given event. |
| Task Instruction - requirement | |
| MT1 | - When making the choice, focus on the evidence presented in the sequence of images from left to right.<br>- Choose the correct option based on the content in the sequential images from left to right.<br>- Only use the left-to-right content of the image sequence to inform the decision.<br>- Evaluate the statement strictly based on the information shown in the sequence of images from left to right. |
| MT2 | - Read the image sequence from left to right. Use the images' content to guide your sentence ordering. Avoid assumptions not supported by the image seqeunce or sentences.<br>- Understand images from left to right in the order. Focus on the visual content of the images to determine the correct order of the sentences. Make sure the reordered sentences form a clear and coherent narrative or description.<br>- Follow the sequence of images from left to right and use their content to determine the correct sentence order. Do not rely on assumptions that are not supported by the images or sentences.<br>- Interpret the images in their left-to-right order, focusing on their visual details to arrange the sentences correctly. Ensure that the reordered sentences create a logical and coherent narrative or description. |
| MT3 | - Carefully read the provided text. Focus on the events, actions, and details described to reorder the images logically.<br>- Focus on matching the actions and events shown in the images with the details described in the text.<br>- Identify key events and details from the text and use them to determine the proper order of the images.<br>- Compare the images with the text description, focusing on the sequence of events to arrange unordered images correctly. |
| GT | - Make your choice by considering the evidence shown in the sequence of images from left to right.<br>- Select the correct option using only the content presented in the sequential images from left to right.<br>- Rely on the left-to-right order of the image sequence to guide your decision.<br>- Assess the statement exclusively based on the information depicted in the sequence of images from left to right. |
| Response Format | |
| MT1 | - No need to give reasoning process. Submit only the right option letter as your answer, e.g., Option [Letter].<br>- Do not tell the reasons of your decision. Provide the most suitable choice letter in the format of 'Option [Letter]' as your response only.<br>- Do not repeat the prompt or include reasons. Only return the correct option letter in the form of 'Option [Letter]' as your response.<br>- Please exclude explanations in the response. Offer the most proper choice letter in the format of 'Option [Letter]' as your answer only. |
| MT2 | - Do not provide explanations or repeat the prompt. Select from the following options and your answer should only be in the format: Option [Letter].<br>- Provide your answer from the following choices only in the format: 'Option [Letter]' without explanations or repeating the instructions.<br>- Choose the correct option from the choices provided below and output your answer only as 'Option [Letter]', avoiding any explanations or repetition. |
| MT3 | - Select from the following options and your answer should be in the format: 'Option [Letter]'. Respond with the correct option only, avoiding any explanations or repetition.<br>- Do not provide explanations or restate the question. Provide your answer from the following choices only in the format: Option [Letter].<br>- Choose the correct option from the choices provided below and submit your answer only as 'Option [Letter]', without any justifications or repetition of the prompt. |
| GT | - Submit the right number of the image in the sequence as your answer only without additional reasoning or repetition of the instructions.<br>- Provide only the most suitable image number as your response, avoiding any explanations or repetition.<br>- Only return the correct image number in the provided sequence without additional reasoning details or repetition of the instructions. |

Table 10: 我们用来生成提示的不同组件的所有变体。

| | Two-event Relation Inference | | Three-event Relation Inference | | Sentence Ordering- event | | Image Ordering - event | |
|---|---|---|---|---|---|---|---|---|
| | MT | MT\|GT *strict* | MT | MT\|GT *strict* | MT | MT\|GT *strict* | MT | MT\|GT *strict* |
| InternVL2_5-26B-MPO | 0.5962 | 0.6053 | 0.6089 | 0.633 | 0.6557 | 0.8867 | 0.3203 | 0.392 |
| InternVL2_5-78B-MPO | 0.5972 | 0.5811 | 0.5948 | 0.6013 | 0.7378 | 0.9543 | 0.3697 | 0.4527 |
| llava-interleave-qwen-7b | 0.5132 | 0.5325 | 0.5149 | 0.521 | 0.2807 | 0.132 | 0.2083 | 0.1459 |
| Mantis-8B-Idefics2 | 0.5052 | 0.5154 | 0.5111 | 0.5377 | 0.2217 | 0 | 0.1868 | 0 |
| LLaVA-NeXT-Video-34B-DPO | 0.5412 | 0.5303 | 0.5166 | 0.5524 | 0.2953 | 0 | 0.1975 | 0.25 |
| LongVA-7B-DPO | 0.5479 | 0.5437 | 0.5257 | 0.5508 | 0.3371 | 0.75 | 0.1912 | 0 |

Table 12: 当多个图像分别输入到模型中时，??? 个 MLLMs 的结果。

| HuggingFace Model ID | # Params | Vision Backbone | Base LLM |
|---|---|---|---|
| deepseek-ai/deepseek-vl2-tiny | 3.4B | SigLIP-SO400M-384 | DeepSeekMoE |
| deepseek-ai/deepseek-vl2-small | 16.1B | SigLIP-SO400M-384 | DeepSeekMoE |
| OpenGVLab/InternVL2_5-1B | 0.9B | InternViT-300M-448px-V2_5 | Qwen2.5-0.5B-Instruct |
| OpenGVLab/InternVL2_5-1B-MPO | 0.9B | InternViT-300M-448px-V2_5 | Qwen2.5-0.5B-Instruct |
| OpenGVLab/InternVL2_5-8B | 8.1B | InternViT-300M-448px-V2_5 | internlm2_5-7b-chat |
| OpenGVLab/InternVL2_5-8B-MPO | 8.1B | InternViT-300M-448px-V2_5 | internlm2_5-7b-chat |
| OpenGVLab/InternVL2_5-26B | 25.5B | InternViT-6B-448px-V2_5 | internlm2_5-20b-chat |
| OpenGVLab/InternVL2_5-26B-MPO | 25.5B | InternViT-6B-448px-V2_5 | internlm2_5-20b-chat |
| OpenGVLab/InternVL2_5-78B | 78.4B | InternViT-6B-448px-V2_5 | Qwen2.5-72B-Instruct |
| OpenGVLab/InternVL2_5-78B-MPO | 78.4B | InternViT-6B-448px-V2_5 | Qwen2.5-72B-Instruct |
| deepseek-ai/Janus-Pro-1B | 1.0B | ViT-L-16-SigLIP-384 | DeepSeek-LLM-1.5b-base |
| deepseek-ai/Janus-Pro-7B | 7.0B | ViT-L-16-SigLIP-384 | DeepSeek-LLM-7b-base |
| llava-hf/llava-interleave-qwen-0.5b-hf | 0.9B | SigLIP-400M | Qwen1.5-0.5B-Chat |
| llava-hf/llava-interleave-qwen-7b-hf | 8.1B | SigLIP-400M | Qwen1.5-7B-Chat |
| llava-hf/llava-interleave-qwen-7b-dpo-hf | 8.1B | SigLIP-400M | Qwen1.5-7B-Chat |
| lmms-lab/llava-onevision-qwen2-0.5b-ov | 0.9B | siglip-so400m-patch14-384 | Qwen2-0.5B |
| lmms-lab/llava-onevision-qwen2-0.5b-si | 0.9B | siglip-so400m-patch14-384 | Qwen2-0.5B |
| lmms-lab/llava-onevision-qwen2-7b-ov | 8.0B | siglip-so400m-patch14-384 | Qwen2-7B |
| lmms-lab/llava-onevision-qwen2-7b-si | 8.0B | siglip-so400m-patch14-384 | Qwen2-7B |
| lmms-lab/llava-onevision-qwen2-72b-ov-sft | 73.2B | siglip-so400m-patch14-384 | Qwen2-72B |
| lmms-lab/llava-onevision-qwen2-72b-si | 73.2B | siglip-so400m-patch14-384 | Qwen2-72B |
| llava-hf/LLaVA-NeXT-Video-7B-hf | 7.1B | SigLIP-400M | vicuna-7b-v1.5 |
| llava-hf/LLaVA-NeXT-Video-7B-DPO-hf | 7.1B | SigLIP-400M | vicuna-7b-v1.5 |
| llava-hf/LLaVA-NeXT-Video-34B-hf | 34.8B | SigLIP-400M | vicuna-33b-v1.3 |
| llava-hf/LLaVA-NeXT-Video-34B-DPO-hf | 34.8B | SigLIP-400M | vicuna-33b-v1.3 |
| lmms-lab/LongVA-7B | 7.9B | SigLIP-400M | Qwen2-7B-Instruct |
| lmms-lab/LongVA-7B-DPO | 7.9B | SigLIP-400M | Qwen2-7B-Instruct |
| TIGER-Lab/Mantis-8B-Idefics2 | 8.4B | idefics2-8b | Mistral-7B-v0.1 |
| TIGER-Lab/Mantis-8B-siglip-llama3 | 8.5B | SigLIP | LLaMA-3-8B |
| microsoft/Phi-3-vision-128k-instruct | 3.8B | CLIP ViT-L/14 | Phi-3-mini-128k-instruct |
| microsoft/Phi-3.5-vision-instruct | 4.2B | CLIP ViT-L/14 | Phi-3.5-mini-instruct |
| Qwen/Qwen2-VL-2B | 2.2B | CLIP ViT-L/14 | Qwen2-1.5B |
| Qwen/Qwen2-VL-2B-Instruct | 2.2B | CLIP ViT-L/14 | Qwen2-1.5B |
| Qwen/Qwen2-VL-7B | 7.6B | CLIP ViT-L/14 | Qwen2-7B |
| Qwen/Qwen2-VL-7B-Instruct | 7.6B | CLIP ViT-L/14 | Qwen2-7B |
| Qwen/Qwen2-VL-72B | 72.7B | CLIP ViT-L/14 | Qwen2-72B |
| Qwen/Qwen2-VL-72B-Instruct | 72.7B | CLIP ViT-L/14 | Qwen2-72B |

Table 13: 所有评估过的开源 MLLMs 的详细信息：HuggingFace 模型 ID，参数数量，视觉骨干网，基础 LLM。

## C.2 38 个 MLLM 的评价结果

在本节中，我们展示了 38 个最新的多模态大语言模型（MLLMs）在 TempVS 基准测试上的完整定量结果。MT1 事件关系推理任务的结果显示在 Table 14 中，而 Table 15 展示了 MT2 句子排序任务的结果，Table 16 报告了 MT3 图像排序任务的结果。需要注意的是，GT 表示整体匹配评估，它测量了在整个基准中正确匹配到对应图像的事件数量。GT $_{strict}$ 表示严格匹配评估，它计算了序列中每个事件与其对应图像正确匹配的图像序列数量。

| | Two-event Relation Inference | | | | Three-event Relation Inference | | | |
|---|---|---|---|---|---|---|---|---|
| | GT | GT $_{strict}$ | MT | MT\|GT $_{strict}$ | GT | GT $_{strict}$ | MT | MT\|GT $_{strict}$ |
| deepseek-vl2-tiny | 0.4159 | 0.2082 | 0.4971 | 0.4965 | 0.4292 | 0.1052 | 0.4976 | 0.5064 |
| deepseek-vl2-small | 0.4116 | 0.1416 | 0.4311 | 0.4223 | 0.4306 | 0.0669 | 0.4436 | 0.4435 |
| InternVL2_5-1B | 0.3636 | 0.1702 | 0.4102 | 0.4073 | 0.3896 | 0.085 | 0.3895 | 0.4270 |
| InternVL2_5-1B-MPO | 0.3879 | 0.1881 | 0.3727 | 0.3709 | 0.4002 | 0.0953 | 0.3736 | 0.3569 |
| InternVL2_5-8B | 0.5655 | 0.3976 | 0.5426 | 0.5540 | 0.6246 | 0.3278 | 0.5438 | 0.5556 |
| InternVL2_5-8B-MPO | 0.6230 | 0.4697 | 0.5619 | 0.5736 | 0.6871 | 0.4118 | 0.558 | 0.5714 |
| InternVL2_5-26B | 0.6181 | 0.4604 | 0.5705 | 0.5843 | 0.6824 | 0.3958 | 0.5835 | 0.6022 |
| InternVL2_5-26B-MPO | 0.6521 | 0.5132 | 0.6032 | 0.6212 | 0.7096 | 0.4595 | 0.6212 | 0.6474 |
| InternVL2_5-78B | 0.6555 | 0.5157 | 0.5423 | 0.5532 | 0.7154 | 0.4733 | 0.5679 | 0.5698 |
| InternVL2_5-78B-MPO | 0.7046 | 0.5881 | 0.5847 | 0.5992 | 0.7771 | 0.5653 | 0.6139 | 0.6258 |
| Janus-Pro-1B | 0.2359 | 0.0273 | 0.4827 | 0.4814 | 0.2558 | 0.0073 | 0.4645 | 0.4259 |
| Janus-Pro-7B | 0.2399 | 0.0429 | 0.3509 | 0.3406 | 0.2555 | 0.0038 | 0.3287 | 0.3929 |
| llava-interleave-qwen-0.5b-hf | 0.2172 | 0.0245 | 0.4976 | 0.4779 | 0.2182 | 0.0019 | 0.5036 | 0.5012 |
| llava-interleave-qwen-7b-hf | 0.3521 | 0.1300 | 0.5161 | 0.5239 | 0.4112 | 0.0723 | 0.5007 | 0.4981 |
| llava-interleave-qwen-7b-dpo-hf | 0.3531 | 0.1358 | 0.5198 | 0.5357 | 0.4102 | 0.0845 | 0.5173 | 0.5319 |
| llava-onevision-qwen2-0.5b-ov | 0.304 | 0.0859 | 0.4528 | 0.4545 | 0.3343 | 0.0286 | 0.4807 | 0.4764 |
| llava-onevision-qwen2-0.5b-si | 0.2153 | 0.0235 | 0.4496 | 0.4429 | 0.2156 | 0.0013 | 0.3587 | 0.3731 |
| llava-onevision-qwen2-7b-ov | 0.5176 | 0.3278 | 0.5602 | 0.5804 | 0.578 | 0.2604 | 0.5753 | 0.5979 |
| llava-onevision-qwen2-7b-si | 0.4575 | 0.2505 | 0.5292 | 0.5419 | 0.4895 | 0.1403 | 0.5546 | 0.5654 |
| llava-onevision-qwen2-72b-ov-sft | 0.6215 | 0.4641 | 0.5928 | 0.6213 | 0.6844 | 0.405 | 0.6154 | 0.6349 |
| llava-onevision-qwen2-72b-si | 0.5418 | 0.3593 | 0.5296 | 0.5393 | 0.6015 | 0.2804 | 0.5248 | 0.5298 |
| LLaVA-NeXT-Video-7B-hf | 0.2731 | 0.0575 | 0.4603 | 0.4603 | 0.2972 | 0.0135 | 0.4486 | 0.4723 |
| LLaVA-NeXT-Video-7B-DPO-hf | 0.2718 | 0.0596 | 0.4672 | 0.466 | 0.2981 | 0.014 | 0.4520 | 0.4615 |
| LLaVA-NeXT-Video-34B-hf | 0.2742 | 0.0652 | 0.5847 | 0.5839 | 0.3053 | 0.0159 | 0.5947 | 0.5763 |
| LLaVA-NeXT-Video-34B-DPO-hf | 0.3028 | 0.0825 | 0.5330 | 0.5386 | 0.3289 | 0.0205 | 0.5248 | 0.5263 |
| LongVA-7B | 0.3015 | 0.0874 | 0.5468 | 0.5613 | 0.3138 | 0.0208 | 0.5596 | 0.6169 |
| LongVA-7B-DPO | 0.3241 | 0.1132 | 0.5319 | 0.5582 | 0.3449 | 0.0386 | 0.5233 | 0.5350 |
| Mantis-8B-Idefics2 | 0.3452 | 0.1218 | 0.5193 | 0.533 | 0.3586 | 0.041 | 0.5197 | 0.5164 |
| Mantis-8B-siglip-llama3 | 0.2444 | 0.0443 | 0.5238 | 0.5368 | 0.2392 | 0.0065 | 0.5251 | 0.5833 |
| Phi-3-vision-128k-instruct | 0.2226 | 0.0379 | 0.5196 | 0.5219 | 0.2332 | 0.0065 | 0.5132 | 0.4583 |
| Phi-3.5-vision-instruct | 0.2235 | 0.0400 | 0.4904 | 0.4774 | 0.2316 | 0.0078 | 0.4877 | 0.4828 |
| Qwen2-VL-2B | 0.3053 | 0.088 | 0.4935 | 0.4842 | 0.3188 | 0.0313 | 0.4717 | 0.4923 |
| Qwen2-VL-2B-Instruct | 0.3815 | 0.1738 | 0.5271 | 0.5314 | 0.4003 | 0.0777 | 0.5051 | 0.5052 |
| Qwen2-VL-7B | 0.4032 | 0.1664 | 0.5122 | 0.4982 | 0.4233 | 0.0899 | 0.4991 | 0.4603 |
| Qwen2-VL-7B-Instruct | 0.5144 | 0.3239 | 0.5397 | 0.5542 | 0.5415 | 0.2124 | 0.5358 | 0.5534 |
| Qwen2-VL-72B | 0.4906 | 0.2997 | 0.5461 | 0.5738 | 0.5167 | 0.1935 | 0.5429 | 0.5900 |
| Qwen2-VL-72B-Instruct | 0.5078 | 0.3167 | 0.5395 | 0.5643 | 0.5427 | 0.2059 | 0.5563 | 0.6062 |
| GPT-4o | 0.7043 | 0.6034 | 0.5827 | 0.6005 | 0.7807 | 0.5704 | 0.6452 | 0.6644 |

Table 14: MT1: 事件关系推理的完整结果

在我们的基准测试中，仅语言模型在过滤步骤之前和之后的结果显示在 Table 17 。对于每个仅语言模型，第一行显示问题移除之前的结果，第二行显示移除之后的结果。我们观察到，过滤掉那些仅语言模型容易回答的问题后，它们的性能急剧下降，接近或低于随机猜测的水平。

**过滤文本可回答问题前后的结果比较** 我们在 Table 18 中提供了 5 个有代表性的 MLLM 在这个过滤步骤前后的性能结果。我们观察到所有 MLLM 的性能都下降，表明在去除那些仅通过文本就能回答的问题后，我们的基准测试变得更加具有挑战性。这进一步强调了 MLLM 对视觉内容真正理解的必要性。

| | Ordering Sentences (events) | | | | Ordering Sentences (captions) | | | |
|---|---|---|---|---|---|---|---|---|
| | GT | GT $_{strict}$ | MT | MT\| GT $_{strict}$ | GT | GT $_{strict}$ | MT | MT\| GT $_{strict}$ |
| deepseek-vl2-tiny | 0.3701 | 0.0084 | 0.1953 | 0.25 | 0.4415 | 0.0154 | 0.1823 | 0.1786 |
| deepseek-vl2-small | 0.3021 | 0.0037 | 0.1574 | 0.1429 | 0.3735 | 0.0061 | 0.1707 | 0.1818 |
| InternVL2_5-1B | 0.3462 | 0.0089 | 0.2253 | 0.1176 | 0.3589 | 0.0105 | 0.2037 | 0.2105 |
| InternVL2_5-1B-MPO | 0.3636 | 0.0084 | 0.2168 | 0.4375 | 0.3833 | 0.0099 | 0.2032 | 0.3889 |
| InternVL2_5-8B | 0.5294 | 0.0711 | 0.4589 | 0.5778 | 0.5574 | 0.0765 | 0.5424 | 0.6259 |
| InternVL2_5-8B-MPO | 0.5747 | 0.0963 | 0.5663 | 0.7377 | 0.6183 | 0.1415 | 0.6289 | 0.7704 |
| InternVL2_5-26B | 0.5765 | 0.1016 | 0.5674 | 0.7358 | 0.6137 | 0.1311 | 0.63 | 0.7689 |
| InternVL2_5-26B-MPO | 0.6089 | 0.1258 | 0.6989 | 0.9079 | 0.6491 | 0.1696 | 0.7693 | 0.8734 |
| InternVL2_5-78B | 0.6121 | 0.1363 | 0.6695 | 0.8494 | 0.6536 | 0.1845 | 0.7109 | 0.8388 |
| InternVL2_5-78B-MPO | 0.6604 | 0.1842 | 0.7984 | 0.9657 | 0.7064 | 0.2594 | 0.8634 | 0.9639 |
| Janus-Pro-1B | 0.2344 | 0 | 0.1858 | - | 0.2497 | 0.0006 | 0.1982 | 0 |
| Janus-Pro-7B | 0.2409 | 0 | 0.1705 | - | 0.2623 | 0 | 0.1525 | - |
| llava-interleave-qwen-0.5b-hf | 0.2161 | 0 | 0.2068 | - | 0.2269 | 0 | 0.2065 | - |
| llava-interleave-qwen-7b-hf | 0.3363 | 0.0032 | 0.2505 | 0.1667 | 0.3477 | 0.0039 | 0.2704 | 0 |
| llava-interleave-qwen-7b-dpo-hf | 0.3332 | 0.0047 | 0.2679 | 0.3333 | 0.3409 | 0.0099 | 0.2996 | 0.3333 |
| llava-onevision-qwen2-0.5b-ov | 0.2931 | 0.0005 | 0.1884 | 0 | 0.3118 | 0.0006 | 0.1839 | 0 |
| llava-onevision-qwen2-0.5b-si | 0.2106 | 0 | 0.1895 | - | 0.2199 | 0 | 0.1944 | - |
| llava-onevision-qwen2-7b-ov | 0.4839 | 0.0447 | 0.4421 | 0.4118 | 0.5118 | 0.0683 | 0.4692 | 0.4758 |
| llava-onevision-qwen2-7b-si | 0.426 | 0.0168 | 0.4284 | 0.6875 | 0.4119 | 0.0116 | 0.4537 | 0.7619 |
| llava-onevision-qwen2-72b-ov-sft | 0.5838 | 0.0984 | 0.6516 | 0.8182 | 0.6179 | 0.1399 | 0.7506 | 0.8661 |
| llava-onevision-qwen2-72b-si | 0.5115 | 0.0411 | 0.61 | 0.8077 | 0.5464 | 0.0743 | 0.6691 | 0.7852 |
| LLaVA-NeXT-Video-7B-hf | 0.2645 | 0 | 0.1895 | - | 0.2742 | 0 | 0.1823 | - |
| LLaVA-NeXT-Video-7B-DPO-hf | 0.2666 | 0 | 0.1963 | - | 0.2736 | 0 | 0.1938 | - |
| LLaVA-NeXT-Video-34B-hf | 0.2646 | 0.0005 | 0.3184 | 1 | 0.2831 | 0 | 0.3343 | - |
| LLaVA-NeXT-Video-34B-DPO-hf | 0.2882 | 0.0016 | 0.3095 | 0 | 0.2984 | 0.0006 | 0.3133 | 0 |
| LongVA-7B | 0.286 | 0.0016 | 0.3421 | 0.6667 | 0.2909 | 0.0022 | 0.353 | 0.5 |
| LongVA-7B-DPO | 0.3089 | 0.0026 | 0.3547 | 0.8 | 0.2997 | 0.0028 | 0.3618 | 0.6 |
| Mantis-8B-Idefics2 | 0.3326 | 0.0011 | 0.2216 | 0 | 0.2868 | 0.0022 | 0.2081 | 0 |
| Mantis-8B-siglip-llama3 | 0.2406 | 0.0005 | 0.2142 | 0 | 0.2461 | 0 | 0.2153 | - |
| Phi-3-vision-128k-instruct | 0.2243 | 0 | 0.2316 | - | 0.2214 | 0 | 0.2329 | - |
| Phi-3.5-vision-instruct | 0.2262 | 0 | 0.2311 | - | 0.2331 | 0 | 0.2539 | - |
| Qwen2-VL-2B | 0.2873 | 0.0004 | 0.1957 | 0 | 0.3274 | 0.0011 | 0.1995 | 0 |
| Qwen2-VL-2B-Instruct | 0.3638 | 0.0058 | 0.1658 | 0.0909 | 0.333 | 0.0015 | 0.2014 | 0.5 |
| Qwen2-VL-7B | 0.4592 | 0.0028 | 0.3097 | 0.25 | 0.4879 | 0.0076 | 0.3276 | 0.5455 |
| Qwen2-VL-7B-Instruct | 0.4813 | 0.0342 | 0.4253 | 0.6462 | 0.5098 | 0.0441 | 0.446 | 0.6125 |
| Qwen2-VL-72B | 0.4539 | 0.0258 | 0.4363 | 0.7143 | 0.4914 | 0.0435 | 0.5319 | 0.7215 |
| Qwen2-VL-72B-Instruct | 0.4756 | 0.0368 | 0.4632 | 0.6429 | 0.5085 | 0.0496 | 0.5507 | 0.7 |
| GPT-4o | 0.6708 | 0.1863 | 0.534 | 0.5389 | 0.7231 | 0.2357 | 0.615 | 0.5532 |

Table 15: MT2 的完整结果：包含事件/描述的句子排序。

| | Ordering Images (events) | | Ordering Images (captions) | |
|---|---|---|---|---|
| | MT | MT\| GT $_{strict}$ | MT | MT\| GT $_{strict}$ |
| deepseek-vl2-tiny | 0.2042 | 0.4286 | 0.2072 | 0.125 |
| deepseek-vl2-small | 0.1663 | 0 | 0.1553 | 0 |
| InternVL2_5-1B | 0.2107 | 0.1429 | 0.1997 | 0.0625 |
| InternVL2_5-1B-MPO | 0.2052 | 0.1538 | 0.2002 | 0.2667 |
| InternVL2_5-8B | 0.2766 | 0.2783 | 0.2651 | 0.4113 |
| InternVL2_5-8B-MPO | 0.2986 | 0.3841 | 0.318 | 0.3816 |
| InternVL2_5-26B | 0.2666 | 0.2727 | 0.3165 | 0.3589 |
| InternVL2_5-26B-MPO | 0.344 | 0.3971 | 0.3949 | 0.4369 |
| InternVL2_5-78B | 0.311 | 0.4 | 0.3849 | 0.4709 |
| InternVL2_5-78B-MPO | 0.4099 | 0.4883 | 0.5382 | 0.6966 |
| Janus-Pro-1B | 0.2247 | - | 0.2227 | 0 |
| Janus-Pro-7B | 0.2092 | - | 0.2097 | - |
| llava-interleave-qwen-0.5b-hf | 0.2017 | - | 0.2077 | - |
| llava-interleave-qwen-7b-hf | 0.2087 | 0.1667 | 0.1997 | 0.2222 |
| llava-interleave-qwen-7b-dpo-hf | 0.2167 | 0.3846 | 0.2127 | 0.1905 |
| llava-onevision-qwen2-0.5b-ov | 0.1942 | 1 | 0.2082 | 0 |
| llava-onevision-qwen2-0.5b-si | 0.2032 | - | 0.1902 | - |
| llava-onevision-qwen2-7b-ov | 0.2132 | 0.1429 | 0.2157 | 0.1983 |
| llava-onevision-qwen2-7b-si | 0.2077 | 0.35 | 0.2047 | 0.4375 |
| llava-onevision-qwen2-72b-ov-sft | 0.2756 | 0.3182 | 0.2906 | 0.3647 |
| llava-onevision-qwen2-72b-si | 0.2546 | 0.3553 | 0.2461 | 0.2677 |
| LLaVA-NeXT-Video-7B-hf | 0.2102 | - | 0.2132 | - |
| LLaVA-NeXT-Video-7B-DPO-hf | 0.2067 | - | 0.2082 | - |
| LLaVA-NeXT-Video-34B-hf | 0.1977 | 0 | 0.1997 | - |
| LLaVA-NeXT-Video-34B-DPO-hf | 0.1887 | 0.3333 | 0.1932 | 0 |
| LongVA-7B | 0.1952 | - | 0.1902 | - |
| LongVA-7B-DPO | 0.1962 | 1 | 0.1937 | 0 |
| Mantis-8B-Idefics2 | 0.1862 | 0 | 0.1922 | 0.5 |
| Mantis-8B-siglip-llama3 | 0.2002 | 0 | 0.1987 | - |
| Phi-3-vision-128k-instruct | 0.1857 | - | 0.1897 | - |
| Phi-3.5-vision-instruct | 0.1922 | - | 0.1832 | - |
| Qwen2-VL-2B | 0.1917 | - | 0.1859 | 0 |
| Qwen2-VL-2B-Instruct | 0.1438 | 0.125 | 0.1483 | 1 |
| Qwen2-VL-7B | 0.1917 | 0 | 0.2067 | 0.1818 |
| Qwen2-VL-7B-Instruct | 0.2312 | 0.2041 | 0.2456 | 0.3585 |
| Qwen2-VL-72B | 0.2416 | 0.3548 | 0.2406 | 0.4561 |
| Qwen2-VL-72B-Instruct | 0.2651 | 0.4727 | 0.2811 | 0.4737 |
| GPT-4o | 0.2257 | 0.2353 | 0.2297 | 0.2349 |

Table 16: MT3 的完整结果：含有事件/说明的图像排序。

| | Two-event Relation | Three-event Relation | Sentence Ordering - Event | Sentence Ordering - Caption |
|---|---|---|---|---|
| Phi-3.5-mini-instruct | 0.5031 | 0.4999 | 0.3257 | 0.3368 |
| | 0.1798 | 0.2352 | 0.1048 | 0.1242 |
| Llama-3.1-8B | 0.5023 | 0.5059 | 0.2091 | 0.1995 |
| | 0.2767 | 0.2094 | 0.0952 | 0.1095 |
| Qwen2.5-72B-Instruct | 0.5361 | 0.5349 | 0.5260 | 0.6108 |
| | 0.2952 | 0.3010 | 0.2722 | 0.3203 |

Table 17: 文本生成模型在过滤掉可用文字回答的问题前后的表现。对于每个文本生成模型，第一行包含在移除掉至少两个文本生成模型可以正确回答的问题之前的结果，第二行包含移除这些问题之后的结果。

| | Two-event Relation | Three-event Relation | Sentence Ordering - Event | Sentence Ordering - Caption |
|---|---|---|---|---|
| InternVL2_5-26B-MPO | 0.6643 | 0.6743 | 0.7410 | 0.7940 |
| | 0.6032 | 0.6212 | 0.6989 | 0.7693 |
| InternVL2_5-78B-MPO | 0.6755 | 0.6948 | 0.8334 | 0.9037 |
| | 0.5847 | 0.6139 | 0.7984 | 0.8634 |
| llava-onevision-qwen2-72b-ov-sft | 0.6712 | 0.6841 | 0.7211 | 0.8093 |
| | 0.5928 | 0.6154 | 0.6516 | 0.7506 |
| LLaVA-NeXT-Video-34B-hf | 0.7158 | 0.6828 | 0.4014 | 0.3993 |
| | 0.5847 | 0.5947 | 0.3184 | 0.3343 |
| Qwen2-VL-72B-Instruct | 0.6505 | 0.6533 | 0.5698 | 0.6606 |
| | 0.5395 | 0.5563 | 0.4632 | 0.5507 |

Table 18: 五个 MLLM 在过滤掉文本可回答的问题前后的表现。对于每个 MLLM，第一行包含在移除至少能被两个仅语言 LLM 正确回答的问题之前的结果，第二行包含在移除这些问题之后的结果。

**仅使用图像进行图像排序任务的表现** 我们在没有文本描述的情况下执行图像排序任务。如 Table 19 所示，即使是表现最好的多模态大模型，仅基于图像进行排序仍然极具挑战性，因为它们的表现接近随机猜测。这进一步证实了我们的基准要求模型需要同时理解来自文本和图像的时间信息。模型不能仅仅依赖先验知识来推断图像的正确顺序。这正是我们选择视觉故事作为基准数据来源的原因：仅凭单一模态无法实现准确的时间推理。

| | with event | with caption | image only |
|---|---|---|---|
| InternVL2_5-26B-MPO | 0.3440 | 0.3949 | 0.2430 |
| InternVL2_5-78B-MPO | 0.4099 | 0.5382 | 0.2709 |
| llava-onevision-qwen2-72b | 0.2756 | 0.2906 | 0.2023 |

Table 19: 三种 MLLM 在使用事件文本、原始标题和仅图片的图像排序任务中的性能。

我们在 TempVS 基准中提供了更多示例。

Image a

Image b

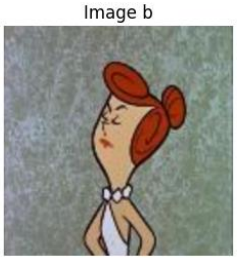


Image c

Image d

Image e

Captions: Fred is in his car driving somewhere. Fred is holding groceries at a fence outside. He puts the groceries on the fence and then jumps over the gate. Fred is standing in the doorway with two bags of groceries and yelling someone in the house. Fred is in the kitchen. He is emptying a bag of groceries on the table. He is speaking to someone off-screen. Wilma is standing in the room. She is speaking while having her elbows bent.

Events: Fred drove in his car. Fred put the groceries on the fence. Fred stood in the doorway with two bags of groceries. Fred emptied a bag of groceries on the table. Wilma stood in the room and spoke to someone.

Correct image sequence: **Image c -> Image e -> Image d -> Image a -> Image b**

Image a

Image b

Image c

Image d

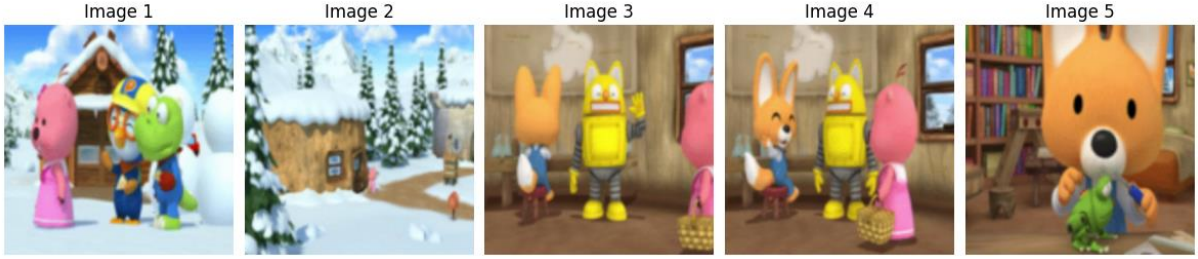
Image e

Captions: We had fun riding the black roller coaster at the fair. The lights lit up the night and the rides made us all dizzy. The dragon coaster was mom's favorite. The arcade games had the funniest stuffed monkeys as prizes. We threw a million darts trying to win one!

Events: We took the black roller coaster. The lights lit up the night. Mom favored the dragon coaster. The arcade games had stuffed monkeys as prizes. We threw the darts.

Correct image sequence: **Image c -> Image e -> Image d -> Image a -> Image b**

Image 1 Image 2 Image 3 Image 4 Image 5



Event a: Eddy used a screwdriver on a green frog toy.
Event b: Loopy walked to a house.
Event c: Eddy turned around and waved to Loopy.
Event d: Loopy came in with a basket.
Event e: Loopy walked away from Crong and Pororo.

Caption a: Eddy is using a screwdriver on a green frog toy.
Caption b: Loopy is walking to a house surrounded by snowed fields and trees.
Caption c: Eddy turns around and waves to Loopy. Rody, Eddy and Loopy are together in a room.
Caption d: Loopy comes in with a basket. Loopy and Rody wave hands to each other.
Caption e: Loopy walks away and Crong and Pororo wave hands to loopy.

Correct sentence sequence: **Sentence e -> Sentence b -> Sentence d -> Sentence c -> Sentence a**

Image 1 Image 2 Image 3 Image 4 Image 5

Event a: Robert found many different shops and services inside the building.
Event b: Jon sat in his room planning his next move.
Event c: Amy and Robert came in their car to meet Jon.
Event d: Robert loaded his gun in the lift.
Event e: Robert entered the building.

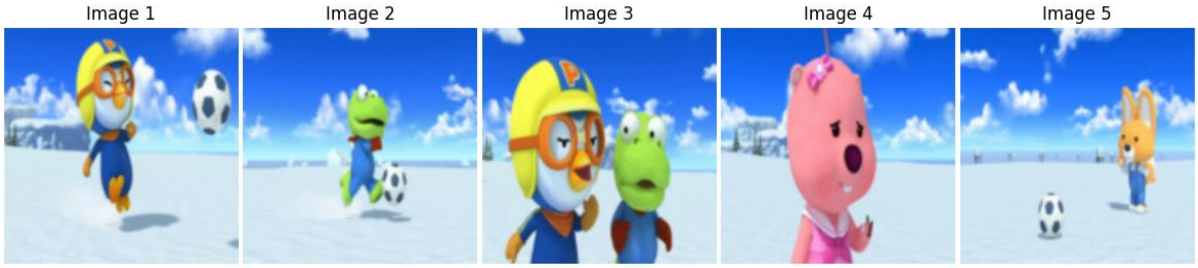
Caption a: Robert finds many different shops and services inside building. He is confused.
Caption b: Jon is sitting in his room, planning his next move.
Caption c: Amy and Robert are coming in their car to meet Jon.
Caption d: Robert gets inside lift and loads his gun. He intends to kill Jon as soon as find him.
Caption e: Robert goes inside the building and begins to search for Jon.

Correct sentence sequence: **Sentence b -> Sentence c -> Sentence e -> Sentence a -> Sentence d**


**True**: The King and his hunting party pursued the deer after the parents told Anna they were in the middle of the King's forest.
**False**: The parents told Anna they were in the middle of the King's forest after the King and his hunting party pursued the deer.

**True**: The King and his hunting party pursued the deer. Earlier, Anna sat with her parents beneath a large oak tree.
**False**: Anna sat with her parents beneath a large oak tree. Earlier, the King and his hunting party pursued the deer.

**True**: Anna sat with her parents beneath a large oak tree. Anna asked her parents where they were. The King and his hunting party pursued the deer.
**False**: The King and his hunting party pursued the deer. Anna sat with her parents beneath a large oak tree. Anna asked her parents where they were.




**True**: Pororo kicked the ball high. Crong received a ball from Pororo.
**False**: Crong received a ball from Pororo. Pororo kicked the ball high.

**True**: Pororo kicked the ball high. Then, Loopy failed to kick the ball.
**False**: Loopy failed to kick the ball. Then, Pororo kicked the ball high.

**True**: Pororo and Crong shouted to Loopy before Loopy failed to kick the ball, and after that, Eddy kicked the ball.
**False**: Eddy kicked the ball before Pororo and Crong shouted to Loopy, and after that, Loopy failed to kick the ball.

Image 1 Image 2 Image 3 Image 4 Image 5


**True**: Granpa made some last-minute repairs before the kids had a tickle fight in the family room.
**False**: The kids had a tickle fight in the family room before Granpa made some last-minute repairs.

**True**: The mother, father, and son listened to a story at dinner. Earlier, Granpa made some last-minute repairs.
**False**: Granpa made some last-minute repairs. Earlier, the mother, father, and son listened to a story at dinner.

**True**: First, Granpa made some last-minute repairs. Second, the kids had a tickle fight in the family room. Third, Nina took out a chocolate cake.
**False**: First, Nina took out a chocolate cake. Second, Granpa made some last-minute repairs. Third, The kids had a tickle fight in the family room.

Image 1 Image 2 Image 3 Image 4 Image 5

**True**: Wilma and betty talked in the living room. Earlier, Fred sat at the kitchen table with a boy.
**False**: Fred sat at the kitchen table with a boy. Earlier, Wilma and Betty talked in the living room.

**True**: Fred and Barney talked to each other while driving in the car after Fred choked the men in the blue dress with a cap in the dining room.
**False**: Fred choked the men in the blue dress with a cap in the dining room after Fred and Barney talked to each other while driving in the car.

**True**: Fred choked the men in the blue dress with a cap in the dining room. Fred and Barney talked to each other while driving in the car. Wilma and Betty talked in the living room.
**False**: Fred and Barney talked to each other while driving in the car. Fred choked the men in the blue dress with a cap in the dining room. Wilma and Betty talked in the living room.