



VLM²-Bench: A Closer Look at How Well VLMs Implicitly Link Explicit Matching Visual Cues

进一步了解 VLM 如何隐式链接显式匹配的视觉线索

Jianshu Zhang^{♡*}, Dongyu Yao^{♣*}, Renjie Pi[♡], Paul Pu Liang[♦], Yi R. (May) Fung[♡]
♡HKUST ♣CMU ♦MIT

jianshu.zhang777@gmail.com rainy@cmu.edu rpi@ust.hk
ppling@mit.edu yrfung@ust.hk

Abstract

Visually linking matching cues is a crucial ability in daily life, such as identifying the same person in multiple photos based on their cues, even without knowing who they are. Despite the extensive knowledge that vision-language models (VLMs) possess, it remains largely unexplored whether they are capable of performing this fundamental task. To address this, we introduce **VLM²-Bench**, a benchmark designed to assess whether VLMs can Visually Link Matching cues, with 9 subtasks and over 3,000 test cases.

视觉上匹配线索的能力在日常生活中至关重要，例如根据线索在多张照片中识别出同一个人，即使不知道他们是谁。尽管视觉-语言模型（VLMs）拥有广泛的知识，但它们是否能够执行这一基本任务仍然很大程度上未被探索。为了解决这个问题，我们引入了**VLM²-Bench**，这是一个旨在评估VLMs是否能够Visually Link Matching cues的基准测试，包含9个子任务和超过3,000个测试案例。

Comprehensive evaluation across eight open-source VLMs and GPT-4o, along with further analysis of various language-side and vision-side prompting methods, leads to a total of eight key findings. We identify critical challenges in models' ability to link visual cues, highlighting a significant performance gap where even GPT-4o lags 34.80% behind humans. Based on these insights, we advocate for (i) enhancing core visual capabilities to improve adaptability and reduce reliance on prior knowledge, (ii) establishing clearer principles for integrating language-based reasoning in vision-centric tasks to prevent unnecessary biases, and (iii) shifting vision-text training paradigms toward fostering models' ability to independently structure and

infer relationships among visual cues.¹

对八个开源 VLMs 和 GPT-4o 的综合评估，以及对各种语言侧和视觉侧提示方法的进一步分析，得出了八个关键发现。我们识别了模型在链接视觉线索能力上的关键挑战，突出了一个显著的性能差距，即即使是 GPT-4o 也落后于人类 34.80%。基于这些见解，我们主张 (i) 增强核心视觉能力以提高适应性并减少对先验知识的依赖，(ii) 建立更清晰的原则，将基于语言的推理整合到以视觉为中心的任务中，以防止不必要的偏见，以及 (iii) 将视觉-文本训练范式转向培养模型独立构建和推断视觉线索之间关系的能力。

1 Introduction 引言

Humans constantly link matching visual cues to navigate and understand their environment. For instance, we can determine whether objects, and individuals are the same simply by comparing their distinguishing visual features (1; 27; 35). This ability, often without needing additional background knowledge, is fundamental in our daily interactions with the world around us. However, while current vision-language models (VLMs) (4; 19; 54; 32) have demonstrated extensive knowledge and expanded their capabilities from single-image understanding to handling multiple images and videos, *whether they can effectively link matching visual cues across images or frames—an essential skill for coherent multimodal reasoning—remains an open question.*

人类不断通过关联匹配的视觉线索来导航和理解环境。例如，我们可以通过比较物体和个体的显著视觉特征来确定它们是否相同 (1; 27; 35)。

¹Project page: <https://vlm2-bench.github.io/>.

♦Work was done while student was an intern at HKUST.

项目页面: <https://vlm2-bench.github.io/>。

♣该工作是在学生于香港科技大学实习期间完成的。

*These authors contribute to this work equally.

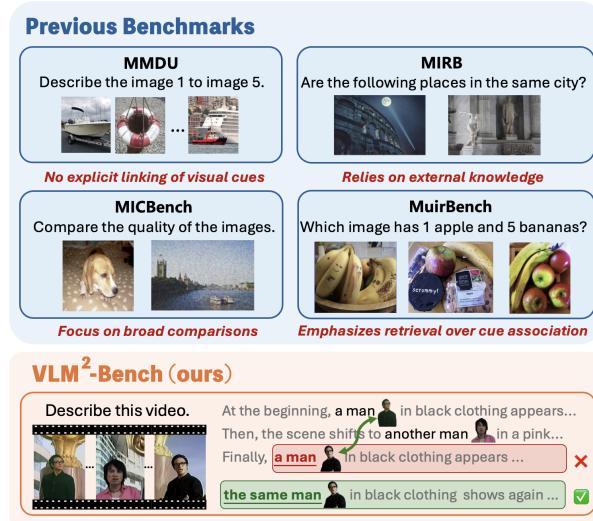


Figure 1: **Previous benchmarks** fail to assess the ability to link matching visual cues, whereas our **VLM²-Bench** explicitly tests this ability, as shown in the example where the model need to identify the reappearance of the same person by linking visual cues, like facial features or clothing, across non-adjacent frames.

现有基准未能评估匹配视觉线索的关联能力，而我们的**VLM²-Bench**明确测试了这一能力，如示例所示，模型需要通过跨非相邻帧的视觉线索（如面部特征或服装）来识别同一人的再次出现。

这种能力通常不需要额外的背景知识，是我们与周围世界日常互动的基础。然而，尽管当前的视觉-语言模型（VLMs）(4; 19; 54; 32)已经展示了广泛的知识，并将能力从单图像理解扩展到处理多图像和视频，它们是否能够有效地跨图像或帧关联匹配的视觉线索——这是连贯多模态推理的关键技能——仍然是一个悬而未决的问题。

As shown in Figure 1, existing benchmarks on multiple images and videos fall short in exploring this fundamental ability as they: (a) do not require explicitly linking visual cues across images or frames (25; 51); (b) rely on external knowledge rather than assessing models’ ability to link explicitly visual cues (56; 23); (c) emphasize broad and abstract visual comparisons rather than specific cue matching (42; 24); and (d) focus on retrieval-based tasks rather than evaluating the direct association of visual cues across different visual contexts (36).

如图 1 所示，现有的多图像和视频基准在探索这一基本能力方面存在不足，因为它们：(a) 不要求显式地跨图像或帧关联视觉线索 (25; 51)；(b) 依赖外部知识，而不是评

估模型显式关联视觉线索的能力 (56; 23)；(c) 强调广泛和抽象的视觉比较，而不是具体的线索匹配 (42; 24)；(d) 专注于基于检索的任务，而不是评估跨不同视觉上下文的视觉线索的直接关联 (36)。

To bridge this gap, we introduce **VLM²-Bench**, a benchmark specifically designed to evaluate how well VLMs visually link matching cues. VLM²-Bench is structured around three types of visual cue connection: *general cue*, *person-centric cue*, and *object-centric cue*, encompassing a total of eight subtasks. To balance scalability and quality, we design a semi-automated pipeline with human verification for further refinement. Additionally, our subtasks cover a variety of QA formats—including T/F, multi-choice, numerical, and open-ended questions—totaling over 3,000 question-answer pairs. To better evaluate model performance, we also design specific metrics tailored to various task.

为了弥补这一差距，我们引入了**VLM²-Bench**，这是一个专门设计用于评估 VLM 在视觉上关联匹配线索能力的基准。VLM²-Bench 围绕三种视觉线索连接类型构建：通用线索、以人为中心的线索和以物体为中心的线索，共包含八个子任务。为了平衡可扩展性和质量，我们设计了一个半自动化的流程，并通过人工验证进行进一步优化。此外，我们的子任务涵盖了多种问答格式——包括 T/F、多选题、数值题和开放式问题——总计超过 3,000 个问答对。为了更好地评估模型性能，我们还设计了针对各种任务的特定指标。

We conduct a comprehensive evaluation of 8 open-source models and GPT-4o on our VLM²-Bench. Despite VLMs generally possessing extensive knowledge, some models perform on par with, or even worse than, the chance-level baseline on our vision-centric tasks. Notably, GPT-4o also underperforms, lagging behind human-level accuracy by 34.80%. This highlights the significant room for improvement in VLMs’ ability to link visual cues. Furthermore, we introduce various language-side and vision-side prompting techniques to explore whether they can enhance the models’ performance on the benchmark. Through experimental results and case studies, we present *eight key observations*, hoping that these insights will guide future improvements in VLMs for vision-centric tasks.

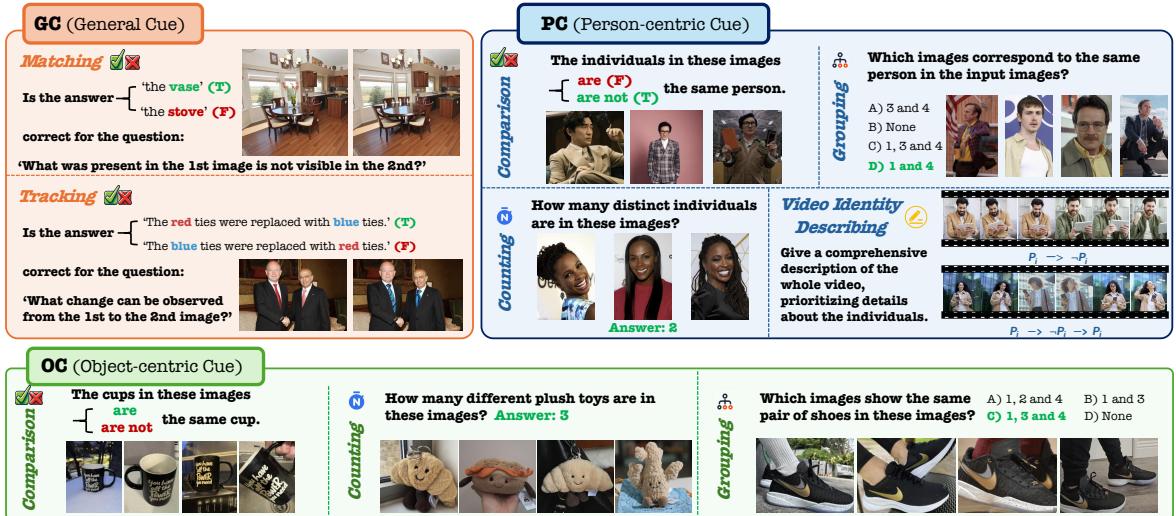


Figure 2: Overview of **VLM²-Bench**. The benchmark is categorized into three subsets based on visual cues: GC (General Cue), OC (Object-centric Cue), and PC (Person-centric Cue), each comprising multiple subtasks. To comprehensively evaluate VLMs’ ability to visually link matching cues, the benchmark includes diverse question formats—T/F , multiple-choice , numerical , and open-ended —ensuring a comprehensive evaluation.

VLM²-Bench 概览。该基准根据视觉线索分为三个子集：GC（通用线索）、OC（以物体为中心的线索）和PC（以人为中心的线索），每个子集包含多个子任务。为了全面评估VLM在视觉上关联匹配线索的能力，基准包括多种问题格式——T/F 、多选题 、数值题 和开放式问题 ——以确保全面评估。

我们在 VLM²-Bench 上对 8 个开源模型和 GPT-4o 进行了全面评估。尽管 VLM 通常拥有广泛的知识，但一些模型在以视觉为中心的任务上表现与随机基线相当，甚至更差。值得注意的是，GPT-4o 也表现不佳，落后于人类水平准确率 34.80%。这突显了 VLM 在关联视觉线索能力方面的显著改进空间。此外，我们引入了多种语言侧和视觉侧的提示技术，以探索它们是否能够提高模型在基准上的表现。通过实验结果和案例研究，我们提出了八个关键观察，希望这些见解能够指导未来在以视觉为中心的任务中对 VLM 的改进。

2 VLM²-Bench

VLM²-Bench 是一个旨在评估模型在处理多张图像或视频时视觉链接匹配线索能力的基准。本节介绍了 VLM²-Bench 的三个主要类别——通用线索 (§2.1)、对象中心线索 (§2.2) 和人物中心线索 (§2.3)——详细说明了它们的相关子任务、数据收集过程以及问答对的构建。

2.1 通用线索 (GC)

GC 旨在评估模型在不同上下文中链接匹配线索的能力，涵盖了广泛的通用线索。给定两张包含匹配和不匹配线索的图像，理想的模型应能准确识别不匹配的线索并关联匹配的线索。

子任务。这里我们介绍了两个子任务：(i) 匹配 (**Mat**) 评估模型在两幅图像中链接相应视觉线索以确定它们是否匹配的能力。模型不仅需要识别差异，还必须关联两幅图像中的相同视觉元素，以识别哪些内容保持不变，哪些内容发生了变化。(ii) 跟踪 (**Trk**) 关注模型跟踪特定视觉线索的能力，该线索仅出现在两幅图像中的一幅中，并确定其如何变化。模型不仅需要检测差异，还必须跨上下文链接线索以理解转换过程。

数据收集。 我们重新利用了两个图像编辑数据集的数据 (40; 16)，其中每个数据样本包括原始图像 I_{ori} 、经过细微修改的编辑图像 I_{edit} 以及描述变化的编辑指令 \mathcal{P} 。我们的数据收集在两个维度上进行。首先，为了确保不匹配线索的多样性，GC 涵盖了各种类型的变化，例如实例级修改（例如添加/删除、交换、属性更改），这些修改专注于特定项目，以及环境级变化。

问答构建。 我们为 **Mat** 和 **Trk** 预定义了一个 T/F 问题模板，其中包含候选答案的占位符（参见附录 ??）。图 3 展示了构建过程，该过程遵循三阶段方法。

手动筛选与精炼： 我们确保 \mathcal{P} 准确反映变化（正确性），唯一对应于修改后的线索（唯一性），并且是明确的（清晰性）。

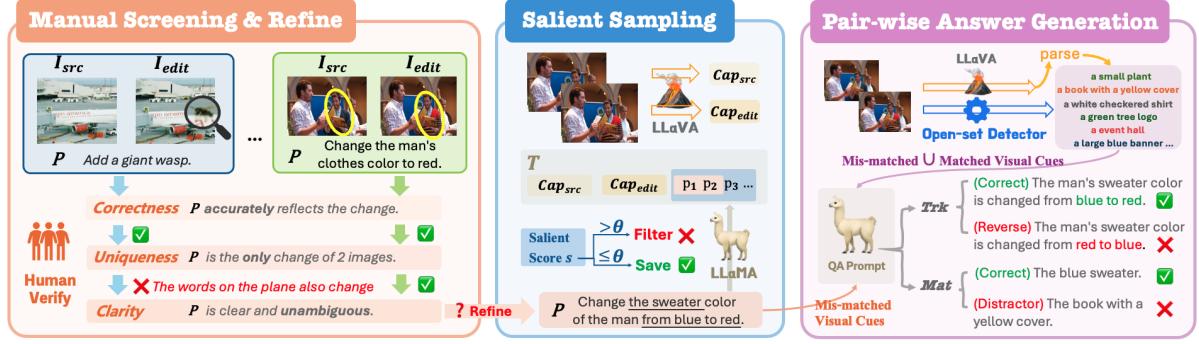


Figure 3: GC 的构建过程: (i) 我们首先基于三个关键标准手动验证编辑后的图像数据。(ii) 然后提示 VLM 为每张图像生成描述, 接着基于显著分数进行过滤, 以保留具有挑战性的案例。(iii) 最后, 从两个来源提取视觉线索, 并将其整合到问答提示中, 指导 LLM 生成正负答案对。

显著采样: 在这里, 我们自动化地移除了过于简单的案例 (例如, 不匹配的线索过于显著)。为此, VLM 首先生成 I_{ori} 和 I_{edit} 的单独描述, 分别记为 Cap_{ori} 和 Cap_{edit} 。然后, 这些描述与 \mathcal{P} 结合成一个段落, 使用预定义的模板 \mathcal{T} (详见表 2)。语言模型 (例如 Llama3-8B (6)) 基于此文本信息分配给 \mathcal{P} 的概率用于计算显著分数, 公式如下:

$$S_{\text{salient}} = \frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \log P_\theta(p_i | C \cup p_{<i}), \quad (1)$$

其中 $\mathcal{P} = \{p_1, p_2, \dots, p_{|\mathcal{P}|}\}$ 表示 \mathcal{P} 的标记化形式, $C = \mathcal{T}(Cap_{ori}, Cap_{edit})$ 表示填充了模板 \mathcal{T} 的上下文。分数低于 θ (此处为 -2.0) 的样本被保留, 确保基准包含更多需要细致视觉线索关联的挑战性示例。

成对答案生成: 最后, 我们使用双级方法提取视觉线索。首先, 从 VLM 生成的描述中解析的线索弥补了开放集检测器在处理分布外场景时的局限性。同时, 开放集检测器 (43) 提取了 VLM 可能忽略的细粒度线索。利用这些提取的线索, 我们提示 LLM 为 Mat 和 Trk 生成一对答案, 每个答案由一个正例和一个负例组成。

2.2 对象中心线索 (OC)

OC 旨在评估模型使用对象中心线索链接与日常对象相关的匹配线索的能力。即使首次遇到某个对象, 一个良好对齐的模型也应能够利用其独特的视觉线索建立关联, 使其能够在不同场景中识别和跟踪该对象。这种能力对于现实世界部署中的连贯感知和交互至关重要。

子任务。 根据链接线索解决问题的复杂性, 我们在 OC 中定义了三个子任务。**(i) 比较 (Cpr)** 要求模型确定出现在不同图像中的对象是否相同。此任务主要评估模型感知视觉一致性或

变化的能力。值得注意的是, 我们观察到模型在进行二元决策时表现出显著的模型特定偏差 (10; 48; 31; 18), 导致结果与其实际能力之间存在差异。为了缓解这一问题, 我们引入了一致性对验证, 其中对于每个陈述 (例如, “X 是 Y”, 答案为 T), 我们生成相应的否定 (例如, “X 不是 Y”, 答案为 F)。只有当模型正确回答两个陈述时, 才被认为正确, 确保其决策的一致性。**(ii) 计数 (Cnt)** 涉及识别唯一对象的数量, 要求模型不仅识别变化或一致性, 还要跟踪不同的线索以避免重复计数同一对象。**(iii) 分组 (Grp)** 是最具挑战性的任务, 要求模型识别同一对象的所有实例, 基于跨多幅图像的精确线索匹配。

数据收集。 我们手动收集了各种类别的日常对象 (例如宠物、杯子)。对于每个类别, 我们定义了多个子类别, 并收集了一组图像 \mathcal{I}_{O_i} ——四幅描绘同一对象在不同场景中的图像。此外, 我们还收集了一组 \mathcal{I}_{-O_i} , 其中包含四幅不同对象的图像, 每幅图像都包含与 \mathcal{I}_{O_i} 匹配的一些视觉线索, 这些图像用作干扰项。

问答构建。 对于每个子任务, 我们定义了一个问题模板, 其中包含 \mathcal{I}_{O_i} 的占位符, 使我们能够根据不同对象定制问题 (参见附录 ??)。对于答案生成, 我们首先根据预定义规则整理多图像序列。对于每个特定序列, 我们生成与 Cpr 、 Cnt 和 Grp 相关的问题的正确答案。

2.3 人物中心线索 (PC)

PC 旨在评估模型链接人物中心线索的能力。虽然模型无法记住每个个体, 但它应具备通过利用独特的视觉线索 (如面部特征、服装或身体姿势) 在不同图像或帧中关联同一个人的能力。这种能力对于确保对人类动作的连贯感知至关重要, 也是现实世界 VLM 应用的基本要求。

子任务。与 OC 的子任务类似（参见 §2.2），PC 包括 (i) 比较 (**Cpr**)、(ii) 计数 (**Cnt**) 和 (iii) 分组 (**Grp**)。然而，与对象不同，个体可以通过视频中的动作进行观察。因此，我们引入了 (iv) 视频身份描述 (**VID**)。此子任务评估模型是否能够通过分析包含该人物的视频描述正确链接同一个人。

数据收集。 我们手动选择了若干个体，每个个体记为 \mathcal{P}_i 。对于每个个体，我们收集了 $\mathcal{I}_{\mathcal{P}_i}$ ——四幅描绘同一个人的图像。对于每幅图像 $I_i \in \mathcal{I}_{\mathcal{P}_i}$ ，我们选择具有最高 CLIP 相似度的干扰图像 $I_{-i} \notin \mathcal{I}_{\mathcal{P}_i}$ (11)。这使我们能够获得大多数线索匹配的不同个体的图像。对于 **VID** 子任务，我们收集了不同个体的视频，记为 $V_{\mathcal{P}_i}$ ，并将每个视频与另一个具有高度相似线索（例如动作、场景、服装）的不同个体的视频 $V_{-\mathcal{P}_i}$ 配对。然后我们构建了两个视频序列：(i) $\mathcal{P}_i \rightarrow \neg\mathcal{P}_i$ ，评估模型区分个体的能力。(ii) $\mathcal{P}_i \rightarrow \neg\mathcal{P}_i \rightarrow \mathcal{P}_i$ ，评估模型是否检测到变化并将 \mathcal{P}_i 的最后一次出现与其首次出现链接起来。the two individuals. (ii) $\mathcal{P}_i \rightarrow \neg\mathcal{P}_i \rightarrow \mathcal{P}_i$, which examines whether the model can detect changes between \mathcal{P}_i and $\neg\mathcal{P}_i$, and link the final occurrence of \mathcal{P}_i to its first appearance.

两个个体。(ii) $\mathcal{P}_i \rightarrow \neg\mathcal{P}_i \rightarrow \mathcal{P}_i$ ，用于检验模型是否能够检测到 \mathcal{P}_i 和 $\neg\mathcal{P}_i$ 之间的变化，并将最后一次出现的 \mathcal{P}_i 与其第一次出现联系起来。

QA Construction. The construction for the overall QA in PC’s *Cpr*, *Cnt*, and *Grp* sub-tasks follows a similar approach to OC. For the **VID** task, we emphasize the model’s ability to describe individuals when designing open-ended questions, aiming to better test the model’s capacity to link individuals appearing in different scenes.

QA 构建。 PC 的 *Cpr*、*Cnt* 和 *Grp* 子任务的整体 QA 构建遵循与 OC 类似的方法。对于 **VID** 任务，我们在设计开放式问题时强调模型描述个体的能力，旨在更好地测试模型在不同场景中链接个体的能力。

2.4 Benchmark Statistics

Our benchmark is organized into three main categories, comprising a total of 9 subtasks. After careful verification, it contains 3,060 question-answer pairs, with varying formats including T/F, multi-choice (MC), numerical (Nu), and open-ended (Oe). To ensure the quality of the annotations, we perform

an inter-annotator agreement (IAA) evaluation (33) involving three annotators, resulting in a high Fleiss’ Kappa score (7) of 0.983. Figure 4 presents the distribution of these subtasks across the three categories, along with the breakdown of different question formats. For additional details, refer to Appendix ??.

我们的基准分为三个主要类别，共包含 9 个子任务。经过仔细验证，它包含了 3,060 个问答对，格式多样，包括 T/F (对/错)、多选 (MC)、数值 (Nu) 和开放式 (Oe)。为了确保注释的质量，我们进行了三个注释者之间的互评一致性 (IAA) 评估 (33)，得到了较高的 Fleiss’ Kappa 分数 (7) 0.983。图 4 展示了这些子任务在三个类别中的分布，以及不同问题格式的细分。更多详细信息，请参阅附录 ??。

3 Evaluation

3.1 Metric Design

T/F (Matching, Tracking, Comparison) : Accuracy is computed based on paired evaluation, where a response is correct only if it answers *T* (ground-truth True) and *F* (ground-truth False) correctly. The overall accuracy across N test pairs is:

T/F (匹配, 跟踪, 比较): 准确率基于配对评估计算，只有当回答正确回答了 *T* (真实值为真) 和 *F* (真实值为假) 时，才被认为是正确的。在 N 个测试对上的总体准确率为：

$$Acc_{pair} = \frac{\sum_{i=1}^N (T_i^+ \cap F_i^-)}{N}, \quad (2)$$

where T^+ and F^- denote correct predictions for *T* and *F*, respectively.

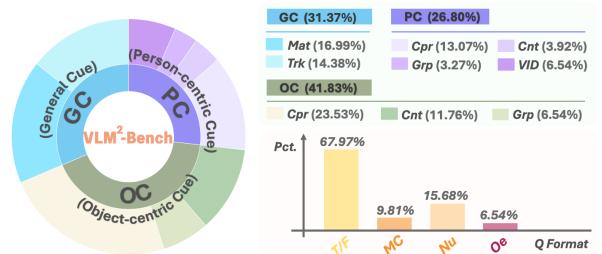


Figure 4: Statistical overview of **VLM²-Bench**. The pie chart shows the distribution of 9 subtasks across the 3 main categories of visual cues. The bar plot illustrates the percentage breakdown by question format.

VLM²-Bench 的统计概览。饼图显示了 9 个子任务在 3 个主要视觉线索类别中的分布。条形图展示了不同问题格式的百分比分布。

Baselines or Models	GC		OC			PC				Overall*	
	Mat	Trk	Cpr	Cnt	Grp	Cpr	Cnt	Grp	VID	Avg	Δ_{human}
Chance-Level	25.00	25.00	50.00	34.88	25.00	50.00	34.87	25.00	-	33.72	-61.44
Human-Level	95.06	98.11	96.02	94.23	91.92	97.08	92.87	91.17	100.00	95.16	0.00
LLaVA-OneVision-7B	16.60	13.70	47.22	56.17	27.50	62.00	46.67	37.00	47.25	39.35	-55.81
LLaVA-Video-7B	18.53	12.79	54.72	62.47	28.50	62.00	66.91	25.00	59.00	43.32	-51.84
LongVA-7B	14.29	19.18	26.67	42.53	18.50	21.50	38.90	18.00	3.75	22.59	-72.57
mPLUG-Owl3-7B	17.37	18.26	49.17	62.97	31.00	63.50	58.86	26.00	13.50	37.85	-57.31
Qwen2-VL-7B	27.80	19.18	68.06	45.99	35.00	61.50	58.59	49.00	16.25	42.37	-52.79
Qwen2.5-VL-7B	35.91	43.38	71.39	41.72	47.50	80.00	57.98	69.00	46.50	54.82	-40.34
InternVL2.5-8B	21.24	26.03	53.33	55.23	46.50	51.50	60.00	52.00	5.25	41.23	-53.93
InternVL2.5-26B	30.50	30.59	43.33	51.48	52.50	59.50	59.70	61.00	21.75	45.59	-49.57
GPT-4o	37.45	39.27	74.17	80.62	57.50	50.00	90.50	47.00	66.75	60.36	-34.80

Table 1: Evaluation results on **VLM²-Bench**, covering *Mat* (Matching), *Trk* (Tracking), *Cpr* (Comparison), *Cnt* (Counting), *Grp* (Grouping), and *VID* (Video Identity Describing). The highest, second, and third highest scores are highlighted. *: Overall excludes the *VID* due to the lack of a chance-level baseline for open-ended tasks.

VLM²-Bench 上的评估结果，涵盖 *Mat* (匹配)、*Trk* (跟踪)、*Cpr* (比较)、*Cnt* (计数)、*Grp* (分组) 和 *VID* (视频身份描述)。最高、第二高 和 第三高 的分数被突出显示。*: 总体排除了 *VID*，因为开放式任务缺乏随机水平基线。

其中 T^+ 和 F^- 分别表示对 T 和 F 的正确预测。

Numerical (Counting): Absolute matching alone does not effectively reflect the severity of errors in numerical responses. To measure the extent of the error between the predicted count \hat{N}_i and ground truth N_i , we introduce Acc_{num} . The first step is to calculate the normalized error:

Numerical (计数): 绝对匹配无法有效反映数值响应中的错误严重程度。为了衡量预测值 \hat{N}_i 与真实值 N_i 之间的误差程度，我们引入了 Acc_{num} 。首先计算归一化误差：

$$\epsilon_i = \frac{|\hat{N}_i - N_i|}{\max(N_i - 1, N_i^{img} - N_i)}, \quad (3)$$

where N_i^{img} is the number of input images. We define $w_i = \max(\{N_i^{img}\}_{i=1}^n)/N_i^{img}$ to penalize errors in cases with fewer images and introduce α as an error amplification factor. The final accuracy over n cases is
其中 N_i^{img} 是输入图像的数量。我们定义 $w_i = \max(\{N_i^{img}\}_{i=1}^n)/N_i^{img}$ 以惩罚图像数量较少的情况下错误，并引入 α 作为误差放大因子。最终在 n 个案例上的准确率为：

$$Acc_{num} = 1 - \frac{1}{n} \sum_{i=1}^n w_i \cdot \epsilon_i^{-\alpha}. \quad (4)$$

Multi-choice (Grouping): Accuracy is the proportion of correctly predicted choices.

Multi-choice (分组): 准确率是正确预测选项的比例。

Open-ended (Video Identity Describing): We use GPT-4o to score model's descriptions, in combination with rule-based scoring prompts. The final accuracy Acc_{oe} is obtained by averaging the scores of all open-ended responses and rescaling them to the range of [0,1]. Additionally, we perform manual verification of GPT-4o's scoring. For each model, we randomly sample 20 scored responses for review, and find only 2 instances with discrepancies, resulting in an accuracy rate of 98.89% (178/180). Refer to Appendix A for more details.

Open-ended (视频身份描述): 我们使用 GPT-4o 对模型的描述进行评分，并结合基于规则的评分提示。最终准确率 Acc_{oe} 通过对所有开放式回答的评分进行平均并将其缩放到 [0,1] 范围内得到。此外，我们对 GPT-4o 的评分进行了人工验证。对于每个模型，我们随机抽取 20 个评分回答进行审查，发现只有 2 个实例存在差异，准确率为 98.89% (178/180)。更多细节请参见附录 A。

3.2 Evaluation Setup

Evaluated Models. 我们评估了八个支持多图像或视频输入的开源 VLM:

LLaVA-OneVision (19)、LLaVA-Video (54)、LongVA (53)、mPLUG-Owl3 (47)、Qwen2-VL (37)、Qwen2.5-VL (32) 和 InternVL2.5 (4)。此外，我们还纳入了商业模型 GPT-4o (12) 进行比较。

Evaluated Models. We evaluate eight open-source VLMs that support multiple-image or video input: LLaVA-OneVision (19), LLaVA-Video (54), LongVA (53), mPLUG-Owl3 (47), Qwen2-VL (37), Qwen2.5-VL (32), and InternVL2.5 (4)。Additionally, we include the commercial model GPT-4o (12) for comparison.

Baselines. We introduce chance-level and human-level baselines (details are in Appendix ??).

Baselines. 我们引入了机会水平和人类水平的基线（详细信息见附录 ??）。

3.3 Results and Findings

Results. Table 1 presents the comprehensive performance of various models across the three categories – General Cue (GC), Object-centric Cue (OC), and Person-centric Cue (PC) – of our VLM²-Bench, covering a total of nine subtasks. **Results.** 表 1 展示了各种模型在我们 VLM²-Bench 的三个类别——通用线索 (GC)、对象中心线索 (OC) 和人物中心线索 (PC) ——中的综合表现，涵盖了总共九个子任务。

Finding I: Simple tasks for humans pose significant challenges for VLMs. We observe that humans achieve near-perfect accuracy across most tasks in our VLM²-Bench. In contrast, even GPT-4o, a state-of-the-art model, performs significantly lower than humans, with an overall performance gap of 34.80%。For open-source models, many show performance comparable to the chance-level baseline or only slightly outperform it. Specifically, for the VID, humans can easily achieve 100% accuracy in distinguishing and linking individuals in a video. However, even the best-performing model, GPT-4o, reaches only 66.75%。Errors mainly arise from failing to recognize individuals after changes or misidentifying reappearing persons as new。

Finding I: 对人类来说简单的任务对 VLM 来说却具有显著挑战性。我们观察到，人类在我们 VLM²-Bench 的大多数任务中几乎达到了完美的准确率。相比之下，即使是目前最先进

的模型 GPT-4o，其表现也显著低于人类，总体表现差距为 34.80%。对于开源模型，许多模型的表现仅与机会水平基线相当或略高于基线。具体来说，对于 VID，人类可以轻松达到 100% 的准确率来区分和链接视频中的个体。然而，即使是表现最好的模型 GPT-4o，也仅达到 66.75%。错误主要源于无法识别变化后的个体或将重新出现的人误认为是新人。

Finding II: Relatively consistent error patterns in Mat and Trk of GC. Table ?? shows that models struggle with mismatched cues due to swap in Mat, which requires linking two completely different cues. To identify what has changed, models must first link and match all the other cues in the context before they can determine that the swapped cue has been transformed. This task requires a deeper understanding of how cues relate to each other across different instances. In contrast, Trk challenges models with mismatched cues due to add/remove, which focuses on tracking how a specific cue changes. This suggests that when there is a cue that appears only once, the model struggles to link the non-appearing cue with the appearing cue to track the transformation process effectively. This limitation reveals models' difficulty in handling cases where certain cues are missing but still need to be linked to understand the dynamic changes.

Finding II: 在 GC 的 Mat 和 Trk 中，模型的错误模式相对一致。表 ?? 显示，模型在由于交换导致的线索不匹配时表现不佳，这需要链接两个完全不同的线索。为了识别发生了什么变化，模型必须首先链接并匹配上下文中的所有其他线索，然后才能确定交换的线索已被转换。这项任务需要更深入地理解不同实例之间线索的关联方式。相比之下，Trk 通过增加/删除导致的线索不匹配来挑战模型，这侧重于跟踪特定线索的变化。这表明，当某个线索只出现一次时，模型难以将未出现的线索与出现的线索链接起来，以有效跟踪转换过程。这一限制揭示了模型在处理某些线索缺失但仍需链接以理解动态变化时的困难。

Finding III: 模型在链接人物中心线索时比对象中心线索表现更好。我们选择了排名前三的开源模型 (Qwen2.5-VL-8B、InternVL2.5-8B、InternVL2.5-26B)，并比较了它们在 OC 和 PC 中三个共享任务 (Cpr, Cnt, Grp) 上的表现。结果显示，模型在人物中心线索上的表现优于对象中心线索。

Finding III: Models perform better in linking person-centric cues than object-centric cues. We selected the top three open-source models (Qwen2.5-VL-8B, InternVL2.5-8B, InternVL2.5-26B) and compared their performance on the three shared tasks (*Cpr*, *Cnt*, *Grp*) in both OC and PC. Results show that models perform better in linking person-centric cues than object-centric cues. That, on average, the performance on PC is higher than on OC by 7.65%, 9.75%, and 11.83% for the tasks of *Cpr*, *Cnt*, *Grp*, respectively.

平均而言，在*Cpr*, *Cnt*, *Grp*任务中，PC上的性能分别比OC高7.65%，9.75%和11.83%。

This could be due to the fact that, during training on person-related data, models are likely provided with explicit person names as anchors to person-centric cues, which helps the models better distinguish different individuals.

这可能是由于在与人相关的数据训练过程中，模型可能会被提供明确的人名作为以人为核心的线索的锚点，这有助于模型更好地区分不同的个体。

In contrast, objects are typically trained using general category names, which may not provide such clear distinctions.

相比之下，对象通常使用通用类别名称进行训练，这可能无法提供如此清晰的区别。

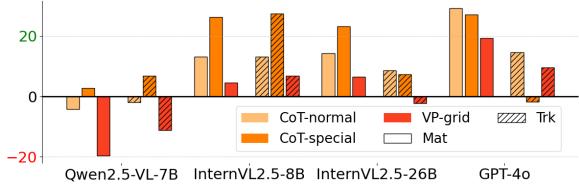
Additionally, these models might have been specifically trained on large datasets that emphasize differentiating and linking individuals (28; 5), thereby enhancing their ability to link person-centric cues.

此外，这些模型可能已经在强调区分和链接个体的大型数据集上进行了专门训练(28; 5)，从而增强了它们链接以人为核心的线索的能力。

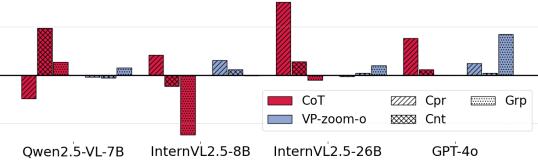
4 How Prompting Methods affect VLMs

在本节中²，我们研究了各种提示方法（语言侧和视觉侧）以评估它们对VLM²-Bench

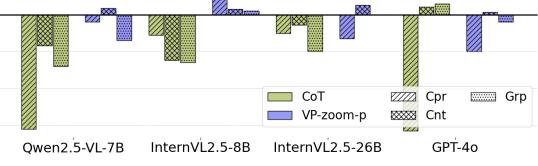
²由于篇幅限制，我们在本节中引用了附录中的大多数案例研究、图表和细节。



(a) CoT-normal, CoT-special 和 VP-grid 在 GC 上的结果。



(b) CoT 和 VP-zoom-o 在 OC 上的结果。



(c) CoT 和 VP-zoom-p 在 PC 上的结果。

Figure 5: 在 VLM²-Bench 上应用不同提示方法时的性能提升或下降 (%)。

性能的影响。我们选择了表现最好的三个开源模型 (Qwen2.5-VL-8B、InternVL2.5-8B、InternVL2.5-26B)，以及 GPT-4o，并探索了不同的 CoT (15; 41) 和视觉提示 (VP) (17; 45) 方法 (详见附录 A)。目标是研究这些技术是否可以提高基准测试的性能，并找出导致其成功或失败的根本原因。

4.1 Probing for General Cue (GC)

方法. (i) **CoT-normal** (表 16) 鼓励模型逐步解决问题，使其能够通过推理解决问题。(ii) **CoT-special** (表 17) 引导模型使用更接近人类通常的思维方式来解决问题。(iii) **VP-grid** (图 9) 改编自之前的工作 (17)，用于我们的任务，通过在图像上叠加点阵作为视觉锚点，提供位置参考并增强模型的性能。

发现 IV: 语言推理有助于模型在逻辑上连接视觉线索。 从图 5a 可以看出，CoT-normal 和 CoT-special 这两种语言推理方法在大多数情况下对模型性能有积极影响。如图 ?? 所示，CoT-special 通过首先让模型明确写出每张图像中的线索，然后使用语言进行推理，从而提高了性能。这一过程通过结构化任务并提供更清晰的逻辑指导，帮助降低了模型的错误率。这表明，当模型在连接一般视觉线索时，使用语言来帮助构建逻辑流程可能是有益的。

发现 V: 视觉提示的有效性取决于模型解释提示线索和视觉内容的能力。 如图 5a 所示, VP-grid 对 QwenVL2.5 的 GC 性能产生了负面影响, 与普通方法相比, 性能显著下降。图 ?? 揭示了这种下降源于模型难以理解提示中的视觉坐标, 导致对线索的误解, 并使其在普通设置下原本正确的案例失败。然而, 如图 A.2 所示, GPT-4o 通过有效利用视觉提示引入的线索, 并结合其强大的视觉感知能力, 成功解决了一个之前错误的案例。

4.2 Probing for Object-centric Cue (OC)

方法. (i) CoT(表 16)。(ii) VP-zoom-o(图 10) 使用开放集检测器(30)获取边界框, 然后裁剪以将模型的注意力集中在对象中心的线索上。通过消除不相关的非对象线索并强调对象中心的线索, 它增强了模型更好地关注最相关视觉信息的能力。

发现 VI: 语言的开放性可能会阻碍对象分组。

与连接实例级线索的 GC 不同, OC 需要基于细粒度的视觉细节对相似对象进行分组。如图 5b 所示, 使用 CoT 的 InternVL2.5 在此任务上表现不佳, 因为语言的开放性导致对细微视觉线索的覆盖有限(见图 ??), 并且对相同线索的表示不一致, 引入了歧义, 使得模型更难可靠地对齐和分组匹配对象。

发现 VII: 放大对象线索对更强的模型有益, 而对其他模型影响较小。 从图 5b 可以看出, 对于像 GPT-4o 这样具有强大视觉能力的模型, 我们的 VP-zoom-o 方法进一步提升了性能。对于其他模型, 该方法至少确保了性能与原始方法持平, 而不会导致任何性能下降 zoom-o method further enhances performance. For other models, this method at least ensures that the performance remains on par with the vanilla approach, without causing any degradation.

4.3 Probing for Person-centric Cue (PC)

Methods. (i) CoT (Table 16). (ii) VP-zoom-p (Figure 11) utilizes a face detector (9) to obtain bounding boxes of faces—the most distinguishing feature of different individuals. It then crops the image to focus only on the face, thereby minimizing the interference from distractor cues such as clothing and other background elements.

Methods. (i) CoT (Table 16). (ii) VP-zoom-p (Figure 11) 利用人脸检测器(9) 获取人脸的边界框——这是不同个体最具区分性的特征。然

后裁剪图像以仅聚焦于人脸, 从而最小化来自干扰线索(如服装和其他背景元素)的干扰。

Finding VIII: CoT and visual prompting fail to improve linking on highly abstract person-centric cues, leading to a performance drop. From Figure 5c, we observe that for almost all models, neither CoT (language-based) nor VP-zoom-p (vision-based) lead to improved performance. This is because facial features are highly abstract, and CoT methods struggle to effectively describe them in words. Additionally, VP-zoom-p fails because current models' visual capabilities are insufficient to accurately perceive facial features.

Finding VIII: CoT 和视觉提示在高度抽象的人为中心线索上未能改善链接, 导致性能下降。 从图 5c 中我们观察到, 对于几乎所有模型, 无论是基于语言的 CoT 还是基于视觉的 VP-zoom-p 都没有带来性能的提升。这是因为面部特征高度抽象, CoT 方法难以用语言有效描述它们。此外, VP-zoom-p 失败的原因是当前模型的视觉能力不足以准确感知面部特征。

Finding IX: Models in comparison tasks prioritize their knowledge over direct matching. In Figure 5c, we observe that CoT leads to a significant drop in GPT-4o's performance in Cpr. We find that when the model compares whether two images depict the same person, its reasoning chain tends to focus on identifying who each person is first, leading to poor performance for individuals it has not encountered before. However, for Grp, the model tends to describe the appearance features of each person. For models with strong visual capabilities like GPT-4o, this approach naturally leads to relatively higher performance improvements.

Finding IX: 模型在比较任务中优先考虑其知识而非直接匹配。 在图 5c 中, 我们观察到 CoT 导致 GPT-4o 在 Cpr 中的性能显著下降。我们发现, 当模型比较两幅图像是否描绘同一个人时, 其推理链倾向于首先识别每个人是谁, 这导致对于其未遇到过的个体表现较差。然而, 对于 Grp, 模型倾向于描述每个人的外观特征。对于像 GPT-4o 这样具有强大视觉能力的模型, 这种方法自然会导致相对较高的性能提升。

5 Related Work

Advancements in vision-language models (12; 32; 53; 19; 47; 4; 22) have significantly broadened their capabilities. Previously restricted to processing single-image inputs, many VLMs can now handle multi-image and even video inputs, allowing them to capture richer and more dynamic visual contexts. Additionally, with access to a growing volume of high-quality visual-textual paired training data (29; 8; 2; 55; 38), these models have shown substantial improvements in perceiving subtle visual cues and their relationships, enabling them to engage in more nuanced reasoning about visual content. Furthermore, VLMs are increasingly applied in real-world scenarios, including navigation (39), planning (46), and autonomous driving (13), solidifying their role in bridging vision and language for practical applications. However, to truly integrate into everyday life, VLMs still have significant room for improvement when it comes to more fundamental but common visual tasks, such as those assessed in our benchmark.

视觉-语言模型的进展 (12; 32; 53; 19; 47; 4; 22) 显著扩展了其能力。以前仅限于处理单张图像输入，许多视觉-语言模型现在能够处理多张图像甚至视频输入，使它们能够捕捉更丰富和动态的视觉上下文。此外，随着访问越来越多的高质量视觉-文本配对训练数据 (29; 8; 2; 55; 38)，这些模型在感知细微视觉线索及其关系方面表现出显著改进，使它们能够对视觉内容进行更细致的推理。此外，视觉-语言模型越来越多地应用于现实场景中，包括导航 (39)、规划 (46) 和自动驾驶 (13)，巩固了它们在连接视觉和语言以用于实际应用中的作用。然而，要真正融入日常生活，视觉-语言模型在更基础但常见的视觉任务方面仍有很大的改进空间，例如我们基准测试中评估的任务。

Benchmarking vision-language models plays a critical role in guiding their future development (21; 49; 3). These benchmarks typically focus on assessing the models' fine-grained perception (18; 34), reasoning abilities (26; 50; 44), commonsense knowledge (52). In addition, evaluations targeting multi-image and video inputs are designed to measure the new competencies that

VLMs require as their visual context extends. These tasks include captioning (52; 51), retrieval (36; 20), comparison (42; 14), and temporal reasoning (24). However, existing benchmarks focus on evaluating VLMs' ability to interpret visual cues based on their knowledge. In contrast, humans typically solve such tasks by explicitly matching visual cues without relying on extensive background knowledge. To better assess whether they can replicate this human-like ability, we propose VLM²-Bench, which focuses on linking and matching explicit visual cues.

视觉-语言模型的基准测试在指导其未来发展方面起着关键作用 (21; 49; 3)。这些基准测试通常侧重于评估模型的细粒度感知 (18; 34)、推理能力 (26; 50; 44) 和常识知识 (52)。此外，针对多图像和视频输入的评估旨在衡量视觉-语言模型在视觉上下文扩展时所需的新能力。这些任务包括图像描述 (52; 51)、检索 (36; 20)、比较 (42; 14) 和时间推理 (24)。然而，现有的基准测试侧重于评估视觉-语言模型基于其知识解释视觉线索的能力。相比之下，人类通常通过显式匹配视觉线索来解决此类任务，而不依赖于广泛的背景知识。为了更好地评估它们是否能够复制这种类似人类的能力，我们提出了 VLM²-Bench，该基准测试侧重于连接和匹配显式视觉线索。

6 Takeaways

Based on our findings, we highlight three key areas for future improvements:

基于我们的发现，我们强调了未来改进的三个关键领域：

- **Strengthening Fundamental Visual Capabilities.** Improving core visual abilities not only enhances overall performance but also increases adaptability. A stronger visual foundation maximizes the effectiveness of visual prompting and reduces reliance on prior knowledge, enabling models to operate more independently in vision-centric tasks.

加强基础视觉能力。 提升核心视觉能力不仅能够增强整体性能，还能提高适应性。更强的视觉基础能够最大化视觉提示的效果，并减少对先验知识的依赖，使模型在视觉中心任务中能够更独立地运行。

- **Balancing Language-Based Reasoning in Vision-Centric Tasks.** Integrating

language into vision-centric tasks requires careful calibration. Future research should establish clearer principles on when language-based reasoning aids visual understanding and when it introduces unnecessary biases, ensuring models leverage language appropriately.

在视觉中心任务中平衡基于语言的推理。
将语言整合到视觉中心任务中需要仔细的校准。未来的研究应建立更清晰的原则，明确何时基于语言的推理有助于视觉理解，何时会引入不必要的偏差，确保模型能够适当地利用语言。

- **Evolving Vision-Text Training Paradigms.** Current training paradigms focus heavily on emphasizing vision-language associations. However, as models expand their visual context window, their ability to reason purely within the visual domain becomes increasingly crucial. We should prioritize developing models that can structure, organize, and infer relationships among visual cues.

发展视觉-文本训练范式。当前的训练范式主要强调视觉-语言关联。然而，随着模型扩展其视觉上下文窗口，它们在纯视觉领域内的推理能力变得越来越重要。我们应优先开发能够构建、组织和推断视觉线索之间关系的模型。

7 Conclusion

In summary, we introduce VLM²-Bench, a novel benchmark designed to probe the capability of vision-language models (VLMs) in visually linking matching cues, an essential yet underexplored skill for models in everyday visual reasoning.

总之，我们介绍了VLM²-Bench，这是一个新颖的基准测试，旨在探索视觉-语言模型（VLMs）在视觉上连接匹配线索的能力，这是模型在日常视觉推理中至关重要但尚未充分探索的技能。

Through extensive evaluations and further analysis of prompting techniques applied on our benchmark, we identify 8 key findings. Notably, even GPT-4o falls 34.80% behind human performance. Based on these insights, we advocate for advancements in fundamental visual capabilities, better integration of language-based reasoning, and the evolution of vision-text training paradigms to improve

VLMs' performance in vision-centric tasks.

通过广泛的评估和对应用于我们基准测试的提示技术的进一步分析，我们确定了8个关键发现。值得注意的是，即使是GPT-4o也落后于人类表现34.80%。基于这些见解，我们主张在基础视觉能力、更好地整合基于语言的推理以及视觉-文本训练范式的演进方面取得进展，以提高VLMs在以视觉为中心的任务中的表现。

Limitations

VLM²-Bench focuses on evaluating visual cue linking but does not cover all possible scenarios. Additionally, while it provides valuable insights, its scale is limited, and model performance may not fully generalize to all real-world settings. Automated evaluation constraints limit the inclusion of open-ended questions in our benchmark, impacting the assessment of models' vision-centric reasoning abilities. Expanding task diversity and refining evaluation methods remain important directions for future work.

VLM²-Bench专注于评估视觉线索的链接，但并未涵盖所有可能的场景。此外，尽管它提供了有价值的见解，但其规模有限，模型性能可能无法完全推广到所有现实世界中的设置。自动化评估的限制使得我们的基准测试中无法包含开放式问题，这影响了模型视觉中心推理能力的评估。扩展任务多样性和改进评估方法仍然是未来工作的重要方向。

References

- [1] Vicki Bruce and Andrew W Young. 1986. **Understanding face recognition.** *British journal of psychology*, 77 (Pt 3):305–27.
- [2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. **Sharegpt4v: Improving large multimodal models with better captions.** *Preprint*, arXiv:2311.12793.
- [3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.
- [4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.
- [5] Dawei Dai, Xu Long, Li Yutang, Zhang Yuan-hui, and Shuyin Xia. 2024. **Humanvlm: Foundation for human-scene vision-language model.** *Preprint*, arXiv:2411.03034.
- [6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- [7] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- [8] Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. **Imageinwords: Unlocking hyper-detailed image descriptions.** *Preprint*, arXiv:2405.02793.
- [9] Adam Geitgey. 2016. Machine learning is fun! part 4: Modern face recognition with deep learning. *Medium. Medium Corporation*, 24:2016.
- [10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.
- [11] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.
- [12] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- [13] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2024. **Senna: Bridging large vision-language models and end-to-end autonomous driving.** *Preprint*, arXiv:2410.22313.
- [14] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. 2024. Img-diff: Contrastive data synthesis for multimodal large language models. *arXiv preprint arXiv:2408.04594*.
- [15] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- [16] Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. 2023. Imagenhub: Standardizing the evaluation of conditional image generation models. *arXiv preprint arXiv:2310.01596*.
- [17] Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. 2024. **Scaffolding coordinates to promote vision-language coordination in large multi-modal models.** *Preprint*, arXiv:2402.12058.
- [18] Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. Natural-bench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*.
- [19] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- [20] You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, et al. 2025. Magician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *arXiv preprint arXiv:2501.05767*.
- [21] Paul Pu Liang, Akshay Goindani, Talha Chafeekar, Leena Mathur, Haofei Yu, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2024. Hemm: Holistic evaluation of multimodal foundation models. *arXiv preprint arXiv:2407.03418*.

- [22] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42.
- [23] Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. 2024. Mibench: Evaluating multimodal large language models over multiple images. *arXiv preprint arXiv:2407.15272*.
- [24] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.
- [25] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. 2024. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*.
- [26] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.
- [27] Romina Palermo and Gillian Rhodes. 2007. *Are you always on my mind? a review of how face perception and attention interact*. *Neuropsychologia*, 45:75–92.
- [28] Renjie Pi, Jianshu Zhang, Tianyang Han, Jipeng Zhang, Rui Pan, and Tong Zhang. 2024. Personalized visual instruction tuning. *arXiv preprint arXiv:2410.07113*.
- [29] Renjie Pi, Jianshu Zhang, Jipeng Zhang, Rui Pan, Zhekai Chen, and Tong Zhang. 2024. Image textualization: An automatic framework for creating accurate and detailed image descriptions. *arXiv preprint arXiv:2406.07502*.
- [30] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. *Grounded sam: Assembling open-world models for diverse visual tasks*. *Preprint*, arXiv:2401.14159.
- [31] Jongyoong Song, Sangwon Yu, and Sungroh Yoon. 2024. Large language models are skeptics: False negative problem of input-conflicting hallucination. *arXiv preprint arXiv:2406.13929*.
- [32] Qwen Team. 2025. *Qwen2.5-vl*.
- [33] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.
- [34] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.
- [35] Anne Treisman and Garry A. Gelade. 1980. *A feature-integration theory of attention*. *Cognitive Psychology*, 12:97–136.
- [36] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. 2024. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.
- [37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- [38] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *Preprint*, arXiv:2307.06942.
- [39] Kasun Weerakoon, Mohamed Elnoor, Germesh Seneviratne, Vignesh Rajagopal, Senthil Hariharan Arul, Jing Liang, Mohamed Khalid M Jaffar, and Dinesh Manocha. 2024. Behav: Behavioral rule guided autonomy using vlms for robot navigation in outdoor scenes. *Preprint*, arXiv:2409.16484.
- [40] Cong Wei, Zheyang Xiong, Weiming Ren, Xinrun Du, Ge Zhang, and Wenhui Chen. 2024. Omnidit: Building image editing generalist models through specialist supervision. *arXiv preprint arXiv:2411.07199*.
- [41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. *Chain-of-thought prompting elicits reasoning in large language models*. *Preprint*, arXiv:2201.11903.
- [42] Haoning Wu, Hanwei Zhu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Annan Wang, Wenxiu Sun, Qiong Yan, et al. 2025. Towards open-ended visual quality comparison. In *European Conference on Computer Vision*, pages 360–377. Springer.

- [43] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Li-juan Wang. 2022. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*.
- [44] Shujin Wu, Yi Fung, Sha Li, Yixin Wan, Kai-Wei Chang, and Heng Ji. 2024. MACAROON: Training vision-language models to be your engaged partners. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7715–7731, Miami, Florida, USA. Association for Computational Linguistics.
- [45] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2023. Fine-grained visual prompting. *Preprint*, arXiv:2306.04356.
- [46] Zhutian Yang, Caelan Garrett, Dieter Fox, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. 2024. Guiding long-horizon task and motion planning with vision language models. *Preprint*, arXiv:2410.02193.
- [47] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *Preprint*, arXiv:2408.04840.
- [48] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.
- [49] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- [50] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- [51] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.
- [52] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- [53] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.
- [54] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.
- [55] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.
- [56] Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

.1 每个子任务和问题类型的详细信息

通用提示 (GC).

匹配 (Mat). 我们收集了 260 个真/假 (T/F) 对，重点关注视觉实例与文本描述之间的一致性（例如，对象的存在、基本属性）。每个 T/F 对形成两个不同的查询（一个为真，一个为假），总共生成 520 个查询。

跟踪 (Trk). 我们设计了 220 个 T/F 对，测试跨帧的对象或实体连续性的理解。例如，一个问题可能会问同一个对象是否在后续帧中重新出现。每个 T/F 对同样生成两个查询，总计 440 个。

以对象为中心的提示 (OC). 所有视觉查询案例都基于我们构建的 360 个图像序列。有关图像序列的详细信息可以在第 .4 节中找到。

比较 (Cpr). 该子任务测试模型在不同帧之间比较对象属性（例如，大小、颜色、数量）的能力。我们生成了 360 个 T/F 对，每个对生成两个查询（总计 720 个）。在这 360 对中，我们保持了真与假的地真值答案的 1:2 比例（即 120 个真，240 个假）。

计数 (Cnt). 我们提供了 360 个数值问题，每个问题要求计算给定场景或序列中的对象数量。可能的数值答案通常是小整数（例如，1, 2, 3），反映了相关对象的数量。

分组 (Grp). 我们生成了 200 个多项选择题 (MC)，要求根据某些标准对对象进行分组（例如，AAB, ABC, AAAB, AABC, ABCD）。每个问题提供多个分组配置选项以及一个“无”选项，该选项可以作为正确答案或干扰项。对于长度为 4 的图像序列，选项包括各种可能的分组（例如，两个相同，三个相同等），以及至少一个额外的干扰分组，以确保足够的挑战性。

以人为中心的提示 (PC). 与 OC 类似，PC 的 260 个图像序列和 200 个视频片段的构建细节在第 .5 节中有详细说明。

比较 (Cpr). 我们创建了 200 个 T/F 对（总计 400 个查询），重点比较与一个或多个人类个体相关的属性或动作。地真值在 100 个真和 100 个假之间保持平衡。

计数 (Cnt). 该子任务涉及 120 个数值问题，要求计算序列中的人数或某些动作的频率。典型的数值答案范围从 1 到 4，考虑到每个视觉序列的范围。

分组 (Grp). 我们提供了 100 个基于至少包含三个图像的序列的 MC 问题，其中至少两个图像包含相同的主要“元人类”。目标是根据外观、角色或动作识别正确的分组。与 OC-Grp 类似，每个问题包括一个“无”选项，作为正确答案或干扰项。

开放式 (VID). 我们引入了 200 个开放式查询，重点关注各种以人为中心的方面，例如识别角色或描述活动。这些问题允许模型回答更具灵活性，并评估生成与上下文相关答案的能力。

.2 注释质量和一致性

正如正文中所提到的，三位注释者审查了所有 3,060 个问题-答案对。注释者间一致性研究显示，共识率高达 98.74%，确保了数据的准确性和一致性。

.3 总结

我们的构建方法确保了对象中心和以人为中心的推理的平衡覆盖，以及基本通用提示（如元素匹配和跟踪）的覆盖。包含多种问题类型 (T/F、MC、数值和开放式) 进一步促进了视觉语言模型的全面评估。正文中的图 4 展示了这些子任务及其问题格式的分布。我们相信，VLM²-Bench 的丰富性和多样性使其成为推进多模态研究的强大平台。我们为手动筛选图像编辑数据构建的图形用户界面 (GUI)。

显著采样。 图 7 和表 2 中的伪代码展示了第 2.1 节中提到的显著采样分数的计算过程。

成对答案生成的提示。 表 3 和 4 提供了用于生成成对答案的完整提示，这些提示用于我们的评估任务。提示的设计旨在指导语言模型为每个任务生成两个不同的答案——一个正例 (T) 答案和一个负例 (F) 答案。双答案格式旨在捕捉预期的响应及其直接对立面，从而提供更平衡的模型理解洞察。

.4 OC (对象中心提示)

数据收集。 为了构建数据集，我们遵循结构化方法收集对象中心图像，如图 8 所示。我们总共手动收集了 320 张对象图像。

主要元对象选择。 我们预定义了 8 种常见对象类型，每种类型包含 5 个元对象。对于每个元对象，我们收集了四张从不同角度和场景条件下拍摄的同一对象的图像。

干扰元对象选择。 为了构建有意义的对象图像序列，我们为每个主要元对象引入了视觉干扰元素，称为“干扰元对象”。具体来说，对于每个主要元对象，我们收集了四张属于同一对象类别中不同但视觉相似的元对象的图像。这些图像是根据预定义的视觉提示混淆原则选择的，确保它们为视觉语言模型提供有意义的挑战。我们确保每个干扰图像属于不同的干扰元对象，从根本上保证最终构建的序列中不同元对象的数量严格遵循我们的设计。选择干扰元对象的原则如图 8 的外环所示。

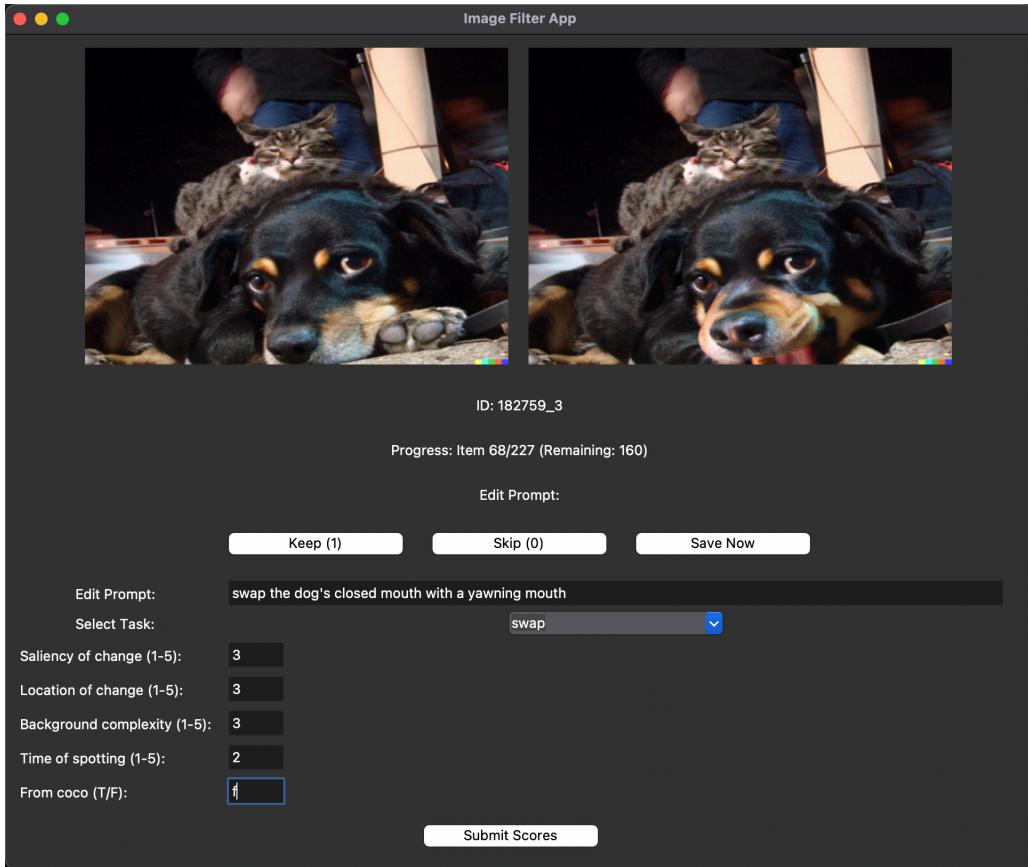


Figure 6: 用于手动筛选图像编辑数据并优化编辑提示的 GUI，应用于 General Cue (GC) 中。

假设你正在查看两张图像：

图像 1: <Cap_src>

图像 2: <Cap_edit>

从图像 1 到图像 2，变化可以总结为：<P>

Table 2: 显著分数计算的模板，每个样本包含三个占位符。

任务描述

给定第一张图像和第二张图像之间的变化，你需要为问题“在第二张图像中可以观察到哪些新元素，而这些元素在第一张图像中不存在？”生成四个选项（此问题根据不匹配的提示类型而变化，此处展示的是“添加/删除”类别中的“添加”问题）。记住，选项的长度应相似。此外，你的回答应以“选项：”开头。

成对设计

在这两个选项中，你需要仅包含对象名称，但要具体：

1. 正确答案（你需要从编辑信息中推断出 * 唯一 * 的对象）
2. 干扰项（你需要从描述中随机选择一个 * 仅 * 在描述中的对象，但与正确答案对象不同）

上下文示例

编辑信息：

在人物的左手中添加一把向下倾斜的武士刀。

描述：

图像描绘了一个穿着传统日本盔甲的人，站在一个雾气缭绕的雪景中。盔甲细节丰富，似乎由金属制成，带有各种带扣和绑带。人物戴着覆盖整个脸部的黑色面具，增添了神秘和隐秘的外观。背景中有石灯笼和其他传统日本建筑，部分被雾气遮挡。整体氛围宁静而略带诡异，雾气增添了神秘和孤立感。场景暗示了一个历史或奇幻背景，可能是一个武士或忍者在雪雾环境中的场景。

选项：

正确答案：武士刀

干扰项：黑色面具

任务

编辑信息：

< 编辑提示 >

描述：

< 描述 >

Table 3: 用于在 General Cue (GC) 的匹配 (Mat) 子任务中生成成对答案的提示。

任务描述

给定第一张图像和第二张图像之间的变化，你需要为问题“从第一张图像到第二张图像可以观察到哪些关键的视觉差异？”生成四个选项。记住，选项的长度应相似。此外，你的回答应以“选项：”开头，并且必须包含正确答案和直接反向答案。

成对设计

在这两个选项中，你需要包含：

1. 正确答案（你需要从编辑信息中推断）
2. 直接反向答案（你需要从编辑信息中推断并将其改为相反的内容）

上下文示例

编辑信息：

将黑色忍者手套替换为适合服务的干净白色手套。

描述：

图像描绘了一个穿着正式服装的人，站在门口。该人穿着黑色燕尾服，白色衬衫和黑色领结。他们拿着一个托盘，托盘上有几个物品。托盘上有一个小玻璃容器、一个瓶子和一个小白色物体，可能是盐瓶或类似物品。该人还戴着黑色手套，这是服务或正式用餐场景中的典型装备。背景显示了一个带有黄铜铰链的木门和浅色墙壁。场景似乎是室内，可能是在家中或正式场所。

选项：

正确答案：黑色忍者手套被替换为干净的白色手套。

直接反向答案：干净的白色手套被替换为黑色忍者手套。

任务

编辑信息：

< 编辑提示 >

描述：

< 描述 >

Table 4: 用于在 General Cue (GC) 的跟踪 (Trk) 子任务中生成成对答案的提示。

Algorithm 1 显著分数计算

```
1 # cap_src: 源图像的描述
2 # cap_edit: 编辑后图像的描述
3 # T: 构建段落的模板
4 # P: 编辑提示
5 input_text = concat(cap_src, cap_edit, T)
6 in_tokens = tokenizer.encode(input_text)
7 out_tokens = tokenizer.encode(P)
8 log_sum = 0
9 tokens = in_tokens
10
11 # 模型前向传播
12 for i in range(1, len(out_tokens)):
13     outputs = model(tokens)
14     logits = outputs.logits
15
16     # 提取下一个 token 的对数概率
17     probs = log_softmax(logits[0, -1, :])
18     prob = probs[out_tokens[i]]
19     log_sum += prob
20
21     # 更新输入序列
22     tokens = concat(tokens, out_tokens[i])
23
24 # 将对数概率总和归一化为显著分数
25 salient_score = log_sum / len(out_tokens)
26
27 # 返回: 显著分数
```

Figure 7: GC 构建过程中显著采样阶段的显著分数计算伪代码。

图像来源。 根据对象的性质，我们从各种来源收集图像：

- **毛绒对象：** 毛绒玩具的图像完全来自 [Jellycat 网站](#) 及其评论部分，其中多样化的用户上传图像提供了各种对象角度和场景。
- **宠物对象：** 对于宠物类别的元对象，我们从社交媒体上受欢迎的宠物摄影师的账户中获取图像。我们还包含了一只由作者之一拥有的布偶猫的图像。因此，这种方法确保了数据集中的每个宠物元对象都属于同一只猫或狗，最小化了与视觉提示混淆无关的变异性。
- **其他对象：** 大多数图像来自 [亚马逊](#) 产品列表和包含用户上传照片的评论部分。数据集的一小部分是通过 Google Lens 图像搜索策划的，其中使用特定的视觉干扰提示来检索并手动选择图像。指导此选择过程的详细视觉提示原则可以在图 8 中找到。

图像序列构建。 OC 中的图像序列构建（共 360 个序列）遵循表 5 中的结构。更多具体细节如下：

两图像序列 (`image_seq_len = 2`)

1. **仅主元对象 (AA)**: 从同一个主元对象中随机抽取两张图像。构建 40 个序列（每个主元对象一个）。

2. **主元对象 + 干扰元对象 (AB)**: 从主元对象中随机选择一张图像，并从对应的干扰元对象中选择一张图像。构建 40 个序列。

三图像序列 (`image_seq_len = 3`)

1. **仅主元对象 (AAA)**: 从同一个主元对象中随机抽取三张图像。构建 40 个序列。
2. **主元对象 + 干扰元对象 (AAB)**: 从主元对象中选择两张图像，并从干扰元对象中选择一张图像。图像顺序被打乱。构建 40 个序列。
3. **主元对象 + 干扰元对象 (ABC)**: 从主元对象中选择一张图像，并从不同的干扰元对象中选择两张图像。构建 40 个序列。

四图像序列 (`image_seq_len = 4`)

1. **仅主元对象 (AAAA)**: 从同一个主元对象中随机抽取四张图像并打乱顺序。构建 40 个序列。
2. **主元对象 + 干扰元对象 (AAAB)**: 从同一个主元对象中随机抽取三张图像，并从干扰元对象中选择一张图像。构建 40 个序列。
3. **主元对象 + 干扰元对象 (AABC)**: 从主元对象中选择两张图像，并从不同的干扰元对象中选择两张图像。构建 40 个序列。
4. **主元对象 + 干扰元对象 (ABCD)**: 从主元对象中选择一张图像，并从不同的干扰元对象中选择三张图像。构建 40 个序列。

问题模板。 表 6、7 和 8 列出了对象中心线索任务的详细标准问题模板（包含格式说明），包括 3 个子任务：比较 (cpr)、计数 (Cnt) 和分组 (Grp)。

5 PC (人物中心线索)

数据收集。 我们主要从 <https://www.imdb.com/> 收集元人类的图像，部分来自演员或女演员的社交媒体。

主元人类选择。 Asian, Black, and White) and genders (Male and Female). For every race-gender combination, we select five main meta-humans, each contributing four images, yielding a total of 120 images.

我们的数据集在不同种族群体中均匀分布（亚洲人、黑人和白人）和性别（男性和女性）。对于每个种族-性别组合，我们选择五个主要



Figure 8: 对象中心提示 (OC) 图像的结构化设计概述。**中心层 (主要元对象)**: 最内层代表预定义的 8 个对象类别，这些类别是我们数据集的基础。这些类别包括宠物、毛绒玩具、包、书、杯子、衬衫、鞋子和玩具。每个类别包含 4 个主要元对象。

中间层 (每个类别中的示例元对象): 中心周围的每个部分展示了其类别中的代表性主元对象。这些元对象作为数据收集的核心实例。例如，宠物类别包括猫和狗，而包类别包括背包、书包和时尚包。

外层 (干扰元对象 & 视觉干扰原则): 最外层的环展示了 4 个干扰元对象中的一个，这些干扰元对象专门选择用于创建具有挑战性的图像序列。每个干扰元对象与其对应的主元对象共享一个或多个干扰性视觉线索。

Num	Src	图像序列构建过程	Cpr	cnt	Grp
2	AA	从同一个对象 O_i 中随机抽取 2 张图像 $\mathcal{I}_{O_i} = \{I_i, I_j\}$, 并打乱顺序。	T	2	-
2	AB	从 \mathcal{I}_{O_i} 中选择 1 张图像 I_i , 并从干扰集 $\mathcal{I}_{\neg O_i}$ 中选择 1 张图像 $I_{\neg i}$, 随机打乱顺序。	F	1	-
3	AAA	从同一个对象 O_i 中随机抽取 3 张图像 $\mathcal{I}_{O_i} = \{I_i, I_j, I_k\}$, 并打乱顺序。	T	3	-
3	AAB	从同一个对象 O_i 中随机抽取 2 张图像 $\mathcal{I}_{O_i} = \{I_i, I_j\}$, 并从干扰集 $\mathcal{I}_{\neg O_i}$ 中选择 1 张图像 $I_{\neg i}$, 随机打乱顺序。	F	2	$[I_i, I_j]$
3	ABC	从同一个对象 O_i 中随机抽取 1 张图像 $\mathcal{I}_{O_i} = \{I_i\}$, 并从干扰集 $\mathcal{I}_{\neg O_i}$ 中选择 2 张图像 $\{I_{\neg i}, I_{\neg j}\}$, 随机打乱顺序。	F	3	[]
4	AAAA	从同一个对象 O_i 中随机抽取 4 张图像 $\mathcal{I}_{O_i} = \{I_i, I_j, I_k, I_p\}$, 并打乱顺序。	T	4	-
4	AAAB	从同一个对象 O_i 中随机抽取 3 张图像 $\mathcal{I}_{O_i} = \{I_i, I_j, I_k\}$, 并从干扰集 $\mathcal{I}_{\neg O_i}$ 中选择 1 张图像 $I_{\neg i}$, 随机打乱顺序。	F	2	$[I_i, I_j, I_k]$
4	AABC	从同一个对象 O_i 中随机抽取 2 张图像 $\mathcal{I}_{O_i} = \{I_i, I_j\}$, 并从干扰集 $\mathcal{I}_{\neg O_i}$ 中选择 2 张图像 $\{I_{\neg i}, I_{\neg j}\}$, 随机打乱顺序。	F	3	$[I_i, I_j]$
4	ABCD	从同一个对象 O_i 中随机抽取 1 张图像 I_i , 并从干扰集 $\mathcal{I}_{\neg O_i}$ 中选择 3 张图像 $\{I_{\neg i}, I_{\neg j}, I_{\neg k}\}$, 随机打乱顺序。	F	3	[]

Table 5: 对象中心线索 (OC) 任务的多图像序列构建总结。

OC-Cpr 正面问题:

根据图像判断以下陈述：‘这些图像中的 $\{obj\}$ 是同一个 $\{obj\}$ 。’仅提供一个正确答案：‘T’ (正确) 或 ‘F’ (错误)。回答只能是 ‘T’ 或 ‘F’。

GT 答案: **T**

OC-Cpr 负面问题:

根据图像判断以下陈述：‘这些图像中的 $\{obj\}$ 不是 同一个 $\{obj\}$ 。’仅提供一个正确答案：‘T’ (正确) 或 ‘F’ (错误)。回答只能是 ‘T’ 或 ‘F’。

GT 答案: **F**

OC-Cnt 问题:

根据以下规则回答问题：你只需要提供 *一个* 正确的数字答案。例如，如果你认为答案是 ‘1’，你的回答应该只是 ‘1’。问题是：输入图像中有多少个不同的 $\{obj\}$?

GT 答案: **3** (示例答案)

Table 7: 用于对象中心线索 (OC) 计数 (Cnt) 子任务的问题模板。

的元人类，每个元人类贡献四张图像，总共生成 120 张图像。

To ensure consistency, all selected individuals are within a similar age range, preventing significant age-related facial changes that could interfere with identity recognition. Additionally, each actor's appearance remains relatively consistent in terms of makeup and overall styling, ensuring that different images of the same meta-human retain distinct yet comparable visual cues (e.g. face shape, eye

Table 6: 用于对象中心线索 (OC) 比较 (Cpr) 子任务的一致性对评估的问题模板。

OC-Grp 问题：

根据以下规则回答问题：你只需要提供 *一个* 正确答案，从以下选项中选择。例如，如果你认为正确答案是 ‘B) 1 和 2’，你的回答应该是 ‘B) 1 和 2’。

问题是：输入图像中哪些图像显示了相同的 $\{obj\}$ ？选项：A) 1 和 3; B) 无; C) 2 和 3; D) 1 和 2。

GT 答案：A) 1 和 3 (示例答案)

Table 8: 用于对象中心线索 (OC) 分组 (Grp) 子任务的问题模板。

spacing, nose structure, and lip contours). By preserving these features, we avoid manipulating a single individual’s visual cues that could potentially mislead VLMs. Rather, we ensure that the evaluation genuinely tests whether the model can visually link matching cues to recognize the same or different individuals without prior identity knowledge.

为了确保一致性，所有选定的个体都在相似的年龄范围内，以防止与年龄相关的显著面部变化干扰身份识别。此外，每个演员的外貌在化妆和整体造型方面保持相对一致，确保同一元人类的不同图像保留独特但可比较的视觉线索（例如脸型、眼距、鼻子结构和唇形）。通过保留这些特征，我们避免了对单个个体的视觉线索进行操纵，以免误导视觉语言模型 (VLMs)。相反，我们确保评估真正测试模型是否能够在没有先验身份知识的情况下，通过视觉链接匹配线索来识别相同或不同的个体。

Distractor Meta-human Selection. To introduce challenging distractors in our sequences, we compute the CLIP embedding for every image and store these embeddings in a reference base. When a distractor image is needed, we perform an image-to-image similarity search within this base to identify the most visually similar image that originates from a different meta-human. This fine-grained matching ensures that the distractor image closely resembles the main meta-human’s image, leading to more challenging image sequences.

干扰元人类选择。 为了在我们的序列中引入具有挑战性的干扰项，我们计算每张图像的 CLIP 嵌入，并将这些嵌入存储在参考库中。当

需要干扰图像时，我们在此库中进行图像到图像的相似性搜索，以识别来自不同元人类的最视觉相似的图像。这种细粒度的匹配确保干扰图像与主要元人类的图像非常相似，从而生成更具挑战性的图像序列。

Discussion on Why Objects Require Dedicated Distractors, While Humans Do Not. In object-centric tasks, objects are categorized into eight distinct types, with substantial differences among different types (e.g. pets and bags). Therefore, each main meta-object requires dedicated distractors from the same object type to ensure meaningful comparisons. In contrast, humans belong to a single category, meaning that any meta-human can serve as a distractor for another. Given that we compute CLIP embeddings to select visually similar distractors, the constructed image sequences already present a significant challenge without the need for type-specific distractors. We also ensure diversity by selecting five main meta-humans for each race-gender pair, providing a sufficiently large pool from which to choose suitable distractors. Corresponding to our hypothesis, in the final curated sequences, most distractor meta-humans chosen were of the same race or gender as the main meta-human. Additionally, as shown in Table 1, these curated image sequences along with our designed questions effectively challenge tested models, revealing their limited performances in visually linking matching cues on person-centric data.

关于为什么对象需要专用干扰项而人类不需要的讨论。 在以对象为中心的任务中，对象被分为八种不同的类型，不同类型之间存在显著差异（例如宠物和包）。因此，每个主要的元对象都需要来自同一对象类型的专用干扰项，以确保有意义的比较。相比之下，人类属于单一类别，这意味着任何元人类都可以作为另一个元人类的干扰项。鉴于我们通过计算 CLIP 嵌入来选择视觉上相似的干扰项，构建的图像序列已经提出了重大挑战，而无需特定类型的干扰项。我们还通过为每个种族-性别对选择五个主要元人类来确保多样性，从而提供了一个足够大的池来选择合适的干扰项。与我们的假设一致，在最终策划的序列中，大多数选择的干扰元人类与主要元人类属于相同的种族或性

别。此外，如表1所示，这些策划的图像序列以及我们设计的问题有效地挑战了测试模型，揭示了它们在视觉链接以人为核心的数据中的匹配线索方面的有限性能。

Images Sequence Construction. The construction of image sequences in PC (a total of 260 sequences) follows the structure in Table ???. More specific details are listed below:

图像序列构建。 PC 中的图像序列构建（总共 260 个序列）遵循表 ?? 中的结构。更多具体细节如下：

Two-Image Sequences (`image_seq_len = 2`)

1. **Main Meta-Human Only (PP):** Two images are randomly selected from the same main meta-human, resulting in 50 sequences.
2. **Main Meta-Human + Distractor Meta-Human (PQ):** One image is randomly selected from the main meta-human, and the other from a distractor meta-human. The order of the images is shuffled. This results in 50 sequences.

两图像序列 (`image_seq_len = 2`)

1. **仅主要元人类 (PP):** 从同一主要元人类中随机选择两张图像，生成 50 个序列。
2. **主要元人类 + 干扰元人类 (PQ):** 从主要元人类中随机选择一张图像，从干扰元人类中选择另一张图像。图像顺序被打乱。生成 50 个序列。

Three-Image Sequences (`image_seq_len = 3`)

1. **Main Meta-Human Only (PPP):** Three images are randomly sampled from the same main meta-human. 20 sequences are constructed.
2. **Main Meta-Human + Distractor Meta-Human (PPQ):** Two images are selected from the main meta-human, and one from a single distractor meta-human. The order of images is shuffled. 30 sequences are constructed.
3. **Main Meta-Human + Distractor Meta-Humans (PQR):** One image is selected

from the main meta-human, while the other two come from distinct distractor meta-humans. The order is shuffled. 10 sequences are constructed.

三图像序列 (`image_seq_len = 3`)

1. **仅主要元人类 (PPP):** 从同一主要元人类中随机抽取三张图像。构建 20 个序列。
2. **主要元人类 + 干扰元人类 (PPQ):** 从主要元人类中选择两张图像，从单个干扰元人类中选择一张图像。图像顺序被打乱。构建 30 个序列。
3. **主要元人类 + 多个干扰元人类 (PQR):** 从主要元人类中选择一张图像，另外两张来自不同的干扰元人类。顺序被打乱。构建 10 个序列。

Four-Image Sequences (`image_seq_len = 4`)

1. **Main Meta-Human Only (PPPP):** All four images are sampled from the same main meta-human. The order is shuffled. 30 sequences are constructed.
2. **Main Meta-Human + Distractor Meta-Human (PPPQ):** Three images are sampled from the main meta-human, while one is selected from a single distractor meta-human. The order is shuffled. 20 sequences are constructed.
3. **Main Meta-Human + Distractor Meta-Humans (PPQR):** Two images are selected from the main meta-human, while two are selected from distinct distractor meta-humans. The order is shuffled. 20 sequences are constructed.
4. **Main Meta-Human + Distractor Meta-Humans (PQRS):** One image is selected from the main meta-human, while three are selected from distinct distractor meta-humans. The order is shuffled. 30 sequences are constructed.

四图像序列 (`image_seq_len = 4`)

1. **仅主要元人类 (PPPP):** 所有四张图像均来自同一主要元人类。顺序被打乱。构建 30 个序列。

2. **主要元人类 + 干扰元人类 (PPPQ):** 从主要元人类中抽取三张图像，从单个干扰元人类中选择一张图像。顺序被打乱。构建 20 个序列。
3. **主要元人类 + 多个干扰元人类 (PPQR):** 从主要元人类中选择两张图像，另外两张来自不同的干扰元人类。顺序被打乱。构建 20 个序列。
4. **主要元人类 + 多个干扰元人类 (PQRS):** 从主要元人类中选择一张图像，另外三张来自不同的干扰元人类。顺序被打乱。构建 30 个序列。

, 无论是 $P \rightarrow \neg P$ 还是 $P \rightarrow \neg P \rightarrow P$ 格式，都确保了每个视频片段在采样过程中都有帧被包含：

- **均匀采样 (8/16 帧):** 每个片段根据视频总长度贡献相应数量的帧。由于在一个拼接视频中，所有采样的片段长度相同，该方法保证每个片段至少有 2 帧可以作为模型输入帧被采样。
- **帧率采样 (1fps):** 由于帧以固定速率采样， $P \rightarrow \neg P$ 和 $P \rightarrow \neg P \rightarrow P$ 的结构确保了每个片段在序列中的位置无论如何，都有足够的时间被捕捉到多帧。

模型名称	均匀采样 (8/16)	帧率采样 (1fps)
LLaVA-OneVision-7B	✓	
LLaVA-Video-7B	✓	
LongVA-7B	✓	
mPLUG-Owl3-7B	✓	
Qwen2-VL-7B	✗	
Qwen2.5-VL-7B	✗	
InternVL2.5-8B	✓	
InternVL2.5-26B	✓	
GPT-4o	✓	

Table 9: 不同视频采样方法的视觉语言模型比较。

因此，通过保持每个片段时间结构的完整性， $P \rightarrow \neg P$ 和 $P \rightarrow \neg P \rightarrow P$ 格式有效地确保了每个片段都能为所有模型的最终采样帧输入贡献帧。

问题模板。 表 10、表 11、表 12 和表 13 展示了以人为中心的线索任务 (Person-centric Cue task) 的详细标准问题模板，涵盖了四个子任务：比较 (PC-Cpr)、计数 (PC-Cnt)、分组 (PC-Grp) 和视频身份描述 (PC-VID)。

PC-Cpr 正面问题:

根据图像判断以下陈述：‘这些图像中的个体是同一个人。’请仅提供一个正确答案：‘T’ (正确) 或 ‘F’ (错误)。回答 ‘T’ 或 ‘F’。

正确答案：**T**

PC-Cpr 负面问题:

根据图像判断以下陈述：‘这些图像中的个体不是同一个人。’请仅提供一个正确答案：‘T’ (正确) 或 ‘F’ (错误)。回答 ‘T’ 或 ‘F’。

正确答案：**F**

Table 10: 用于比较 (Cpr) 子任务的一致性对评估的问题模板。

PC-Cnt 问题:

“根据以下规则回答问题：你只需提供 * 一个 * 正确的数字答案。例如，如果你认为答案是 ‘1’，你的回答应仅为 ‘1’。问题是：输入图像中有多少个不同的个体？”

正确答案：**2 (示例答案)**

Table 11: 用于计数 (Cnt) 子任务的问题模板。

PC-Grp 问题:

根据以下规则回答问题：你只需提供 * 一个 * 正确答案，从以下选项中选择。例如，如果你认为正确答案是 ‘B) 2 和 3’，你的回答应仅为 ‘B) 2 和 3’。问题是：输入图像中哪些图像对应同一个人？选项：A) 无；B) 2 和 3；C) 1 和 3；D) 1 和 2。”

正确答案：**D) 1 和 2 (示例答案)**

Table 12: 用于分组 (Grp) 子任务的问题模板。

PC-VID 问题:

“对整个视频进行全面描述，优先描述视频中的个体细节。”

Table 13: 用于视频身份描述 (VID) 子任务的问题模板。

A 关于提示方法的更多细节

A.1 LLM 作为评估者的提示

当模型回答我们的自由形式的 PC-VID 问题时，它们的回答由 GPT-4o 使用表 14 和 15 中详述的评分提示进行评估。具体来说，对于遵循 $\mathcal{P} \rightarrow \neg\mathcal{P}$ 序列的视频，GPT-4o 评估模型是否明确区分第一个个体 (\mathcal{P}) 和第二个个体 ($\neg\mathcal{P}$) 是不同的。在这种情况下，如果模型成功做出这种区分，则得分为 1；否则，得分为 0。

对于呈现 $\mathcal{P} \rightarrow \neg\mathcal{P} \rightarrow \mathcal{P}$ (PQP) 模式的视频，评估更为细致。评估模型 (GPT-4o) 检查两个方面：(1) 模型是否正确识别出有两个不同的个体 (即 \mathcal{P} 和 $\neg\mathcal{P}$)，以及 (2) 模型是否明确识别出最终出现的是与第一个个体 (\mathcal{P}) 相同的个体。完美识别这两个方面得分为 2，而正确区分个体但未明确将最终出现与第一个个体联系起来得分为 1。如果模型未能区分个体，则得分为 0。

A.2 在 VLM²-Bench 上探索的提示方法

CoT (普通 CoT)。 普通版本的链式思维提示如表 16 所示。我们简单地要求模型“逐步思考”，以确保自我反思和自我纠正，以及透明的思考过程。

CoT-特殊版本 (用于 GC)。 表 17 展示了链式思维提示的特殊版本。根据任务特点，我们仔细分析了人类如何通过视觉链接匹配线索来处理 GC 中的问题，然后精心设计了此提示，以模仿人类的视觉链接过程。

VP-grid (用于 GC)。 图 9 展示了带有网格辅助的视觉提示 (VP-grid) 的完整版本。在这里，我们遵循 (17)，在输入图像上打印一组点阵，并将图像顺序维度与笛卡尔坐标连接为 (图像顺序索引, 列索引, 行索引)。在详细的文本提示设计中，我们还整合了对网格的引用和解释，使视觉语言模型 (VLMs) 能够利用这种视觉辅助作为空间和视觉匹配的参考。

VP-zoom-o 用于 OC。 在图 10 中，我们展示了用于 OC 的视觉提示过程。我们利用 Grounded-SAM (30) 模型根据对象类型检测边界框，然后裁剪“放大”的对象作为图像输入，以进一步生成视觉问答对。

VP-zoom-p 用于 PC。 视觉提示过程与 OC 类似 (图 11)。我们使用人脸检测模型 (9) 来“放大”个体的脸部，并遮挡其他不相关信息。

A.3 Case for CoT prompting in Object-centric Cue Task

A.4 对象中心提示任务中的 CoT 提示案例

The task design for Object-centric cue (OC) and person-centric cue (PC) requires multiple images (more than 2) as sequence input. We observe that, unlike General Cue (GC) tasks where models are required to link instance-level cues, OC tasks demand that models group similar objects based on fine-grained visual features. As illustrated in Figure 5b, models using the CoT approach sometimes struggle to provide a comprehensive overview of vision-based cues across a sequence of images.

对象中心提示 (OC) 和人物中心提示 (PC) 的任务设计需要多个图像 (超过 2 个) 作为序列输入。我们观察到，与通用提示 (GC) 任务不同，OC 任务要求模型基于细粒度的视觉特征对相似对象进行分组。如图 5b 所示，使用 CoT 方法的模型有时难以提供跨图像序列的基于视觉提示的全面概述。

A detailed case in Figure ?? is provided by InternVL2.5-26B’s response. The ground truth and Vanilla responses correctly identify that there is no grouping for the same meta-object in the sequence, with the answer ‘D) None’. In the CoT response, the model states: “The second and third images both have dinosaurs wearing sunglasses”. Although the description here is true, its ambiguity and lack of detailed coverage lead the model to incorrectly select option C) 2 and 3, rather than the correct option D) None. Because if we take a closer look at the design on the backpack in image 3, the dinosaur with sunglasses is actually holding a keyboard instead of a skateboard in image 2. This is a distractive visual matching cue we intend to capture during the distractor meta-object selection. This major difference should have prevented models from grouping image 2 and image 3 together.

图 ?? 中提供了一个详细的案例，展示了 InternVL2.5-26B 的响应。真实情况和 Vanilla 响应正确地识别出序列中没有对相同元对象进行分组，答案为 ‘D) None’。在 CoT 响应中，模型表示：“第二和第三张图像都有戴着太阳镜的恐龙”。尽管这里的描述是正确的，但其模糊性和缺乏详细覆盖导致模型错误地选择了选项 C) 2 和 3，而不是正确的选项 D) None。因为

任务

你正在评估模型准确区分视频中顺序出现的两个不同个体 P 和 Q 的能力（首先是 P，然后是 Q）。给定一个描述，你的任务是确定模型是否明确识别出第一个个体（P）和第二个个体（Q）是不同的个体。

返回格式

你只需在"Score:" 后返回一个数字。如果你认为模型正确识别出两个出现属于不同的个体，返回"Score: 1"。如果你认为模型未能明确说明有两个不同的个体，返回"Score: 0"。

描述

<模型的描述>

Table 14: 用于 VID 的评分提示（当视频属于 $P \rightarrow \neg P$ 类别时）。

任务

你正在评估模型准确区分视频中顺序出现的两个不同个体 P 和 Q 的能力，视频遵循 PQP 模式（首先是 P，然后是 Q，最后是 P）。给定一个描述，你的任务是确定模型是否明确识别出：(1) P 和 Q 是不同的个体，以及 (2) 最终场景中的个体与第一个个体（P）相同。

返回格式

你只需在"Score:" 后返回一个数字。

(1) 如果模型正确描述视频遵循 PQP 序列，明确识别出第一个和最后一个出现属于同一个个体（P），而中间出现的是不同的个体（Q），返回"Score: 2"。

(2) 如果模型正确识别出视频中有两个不同的个体（P 和 Q），但未明确提到最后一个场景返回到 P，返回"Score: 1"。

(3) 如果模型未能识别出有两个不同的个体出现（例如，将所有出现视为同一个个体或未区分 P 和 Q），返回"Score: 0"。

描述

<模型的描述>

Table 15: 用于 VID 的评分提示（当视频属于 $P \rightarrow \neg P \rightarrow P$ 类别时）。

<Question>

Let's think 'step by step' to answer this question, you need to output the thinking process of how you get the answer.

Table 16: 用于 GC（这里我们表示为 CoT-normal，以区别于表 17 中专门为 GC 设计的 CoT-special）、OC 和 PC 的 CoT 提示。

<Question>

Use the following 4 steps to answer the question:

Step 1. Understand the Question

- Identify the question's purpose.
- Check for any format requirements.

Step 2. Perceive (List Elements)

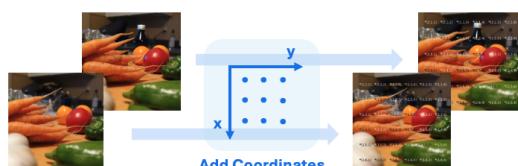
- List every details in each image respectively.
- Note positions and attributes of elements.

Step 3. Connect (Compare & Reason)

- Compare corresponding elements in each image.
- List all the unchanged elements and the changed element.

Step 4. Conclude (Answer the Question)

Table 17: 专门为 GC 设计的 CoT-special 提示。



<Question>

Here's the instruction you need to strictly follow to approach this question:

Two images are provided, each overlaid with a grid of dots arranged in a matrix with dimensions h by w . Each dot on this grid is assigned a unique set of three-dimensional coordinates labeled as (t, x, y) . The first coordinate, " t " distinguishes the two images—“1” for the first image, “2” for the second. The remaining coordinates, “ x ” and “ y ,” specify each dot’s location, where within any column x increases from top to bottom, and within any row y increases from left to right.

This labeling system is intended to help you identify, reference, connect, and compare objects across both images. Now, use the following 4 steps to answer the question.

Step 1. Understand the Question - Identify the question’s purpose.
- Check for any format requirements.

Step 2. Perceive (List Elements and coordinates) - For all the objects in the ‘Options’ of the question, identify them in each image separately, double check their existence. If the object exists then output its nearest coordinates. - Output format like ‘Image1: apple at coordinates (1, 2, 3)... Image2: banana at coordinates (2, 4, 5)’

Step 3. Connect (Compare & Reason) - Use the grid coordinates to connect objects across the two images, observing any similarities or differences at the same (x, y) positions.

Step 4. Conclude (Answer the Question) - If a specific output format is required (e.g., “MY_ANSWER: ...”), follow it exactly. Include the transparent thinking process in your answer, and make sure you output the final *ONE* answer after ‘MY_ANSWER:’, just like ‘MY_ANSWER: D’

Figure 9: VP-grid 在 GC 中如何工作的示意图。

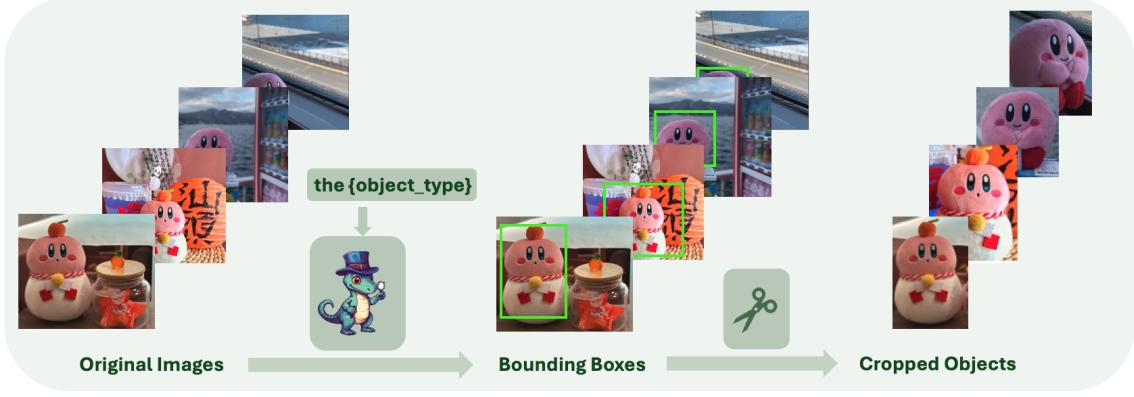


Figure 10: VP-zoom-o 在 OC 中如何工作的示意图。

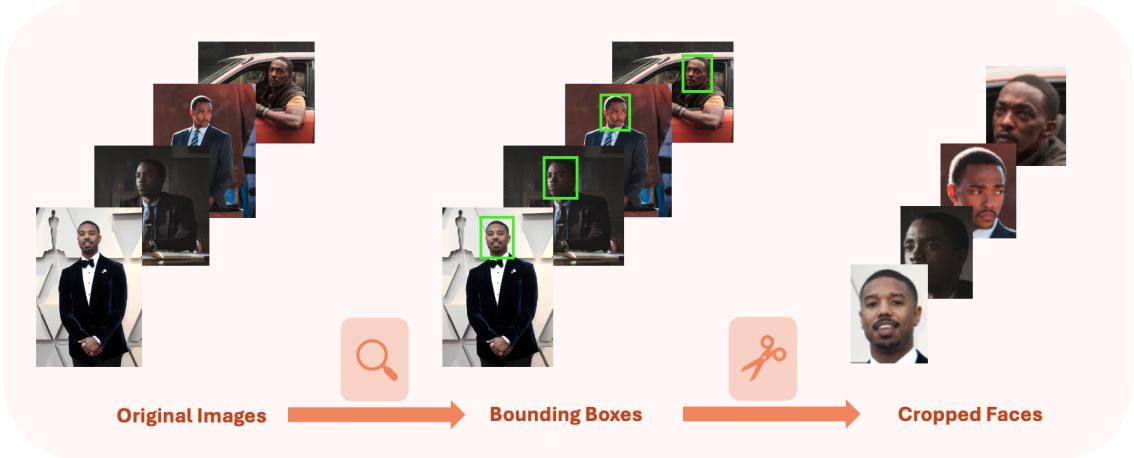


Figure 11: VP-zoom-p 在 PC 中如何工作的示意图。

如果我们仔细观察图像 3 中背包上的设计，戴着太阳镜的恐龙实际上拿着的是键盘，而不是图像 2 中的滑板。这是我们希望在干扰元对象选择期间捕捉到的分散视觉匹配提示。这一重大差异本应阻止模型将图像 2 和图像 3 分组。

According to our findings, this misgrouping occurs for two main reasons:

1. Insufficient Overview of Visual Cues:

The CoT prompt does not force the model to systematically verify all critical details across multiple images. As a result, the model overlooks nuanced differences, such as the design discrepancy on the backpack in image 3, where the dinosaur holds a keyboard rather than a skateboard.

2. Variability in Descriptive Language:

The open-ended language generated by the CoT approach can lead to inconsistent descriptions. In this case, the model generalized the visual cue of a "dinosaur

design" without capturing the specific attribute (i.e., the object the dinosaur is holding), which is crucial for correct grouping.

根据我们的发现，这种错误分组的发生主要有两个原因：

1. 视觉提示的概述不足：CoT 提示并未强制模型系统地验证多个图像中的所有关键细节。因此，模型忽略了细微的差异，例如图像 3 中背包上的设计差异，恐龙拿着的是键盘而不是滑板。

2. 描述语言的变异性：CoT 方法生成的开放式语言可能导致不一致的描述。在这种情况下，模型泛化了“恐龙设计”的视觉提示，而没有捕捉到特定属性（即恐龙拿着的物体），这对于正确的分组至关重要。

Thus, the lack of structured guidance in the CoT prompt leads to the dropping or misinterpretation of critical cues, resulting in incor-

rect grouping decisions for multi-image sequences in OC tasks. This analysis underscores the importance of more detailed structured intermediate reasoning strategies, such as those provided by a tailored CoT-special prompt, to ensure that all relevant visual details are captured and compared accurately.

因此，CoT 提示中缺乏结构化指导会导致关键提示的丢失或误解，从而导致 OC 任务中多图像序列的错误分组决策。这一分析强调了更详细的结构化中间推理策略的重要性，例如由定制的 CoT 特殊提示提供的策略，以确保所有相关的视觉细节都被准确捕捉和比较。