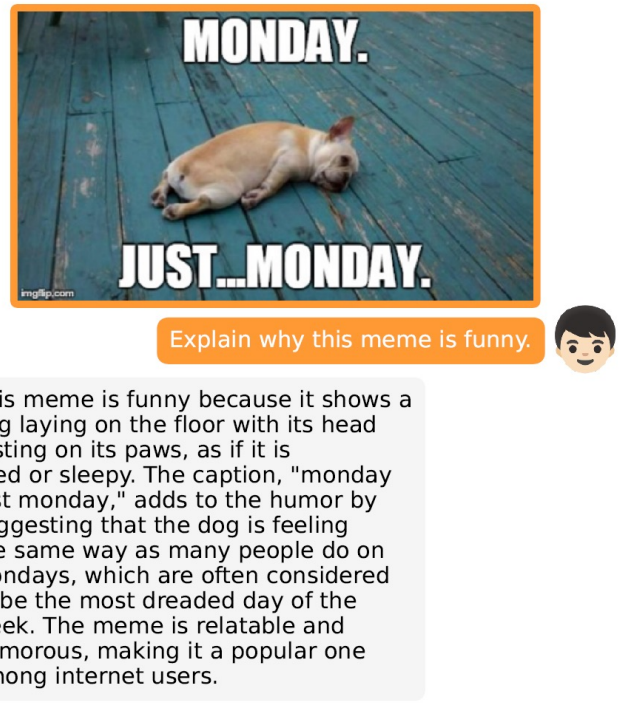
目录

- ChatGPT之前的视觉语言预训练
- 大视觉语言模型的架构和训练
- **大视觉语言模型的评测**
- 大视觉语言模型的能力扩充
- 大语言模型支撑的具身智能（视觉导航）

大视觉语言模型（LVLM）的一体两面

- **展现强大的能力：**光学字符识别（OCR），常识推理（Commonsense Reasoning），表情符理解（meme understanding）
- **难以解决根本的问题：**物体幻觉（Object hallucination）



MiniGPT-4, 2023

Evaluating Object Hallucination in Large Vision-Language Models, 2023

大视觉语言模型间缺乏定量的分析和比较

- 现存的基准主要是**面向任务的：** 特定的输入输出格式
- 大视觉语言模型是**灵活的**，倾向于输出**详细回答:** 自由文本输入输出
- 如何自动地判断模型输出与任务特定的标签是否等价？

Benchmark: VQA v2
Q: Where is skateboarder looking?
GT: down

Benchmark: Visual Entailment
Claim: The player is well-skilled.
GT: entailment

Benchmark: Object Counting
Q: How many persons are there?
GT: 17

LVLMs

Formulation: Open-Ended QA
Prediction: He is looking at down the board.
Judgement: False [EM] / True [Human]

Formulation: Classification
Prediction: The image support the claim.
Judgement: False [EM] / True [Human]

Formulation: Number-Related QA
Prediction: There are more than 17 persons.
Judgement: True [Contain] / False [Human]

[*] indicates the evaluation method. Red and Green represent Wrong and Correct judgement. EM is short for "exactly matched".

需要一个全面、可靠且易于使用的评价基准

- 评价目标
  - 任务或能力级别的评测

- 核心元素
  - 数据和标签
  - 问题形式化
  - 评价指标

MME：一个系统化的多模态大模型评测基准

- 感知和认知，一共包括14个子任务
- 二元形式化：让模型回答**yes [Y]或no [N]**
- 值得注意的是，所有的指令都是人工设计的

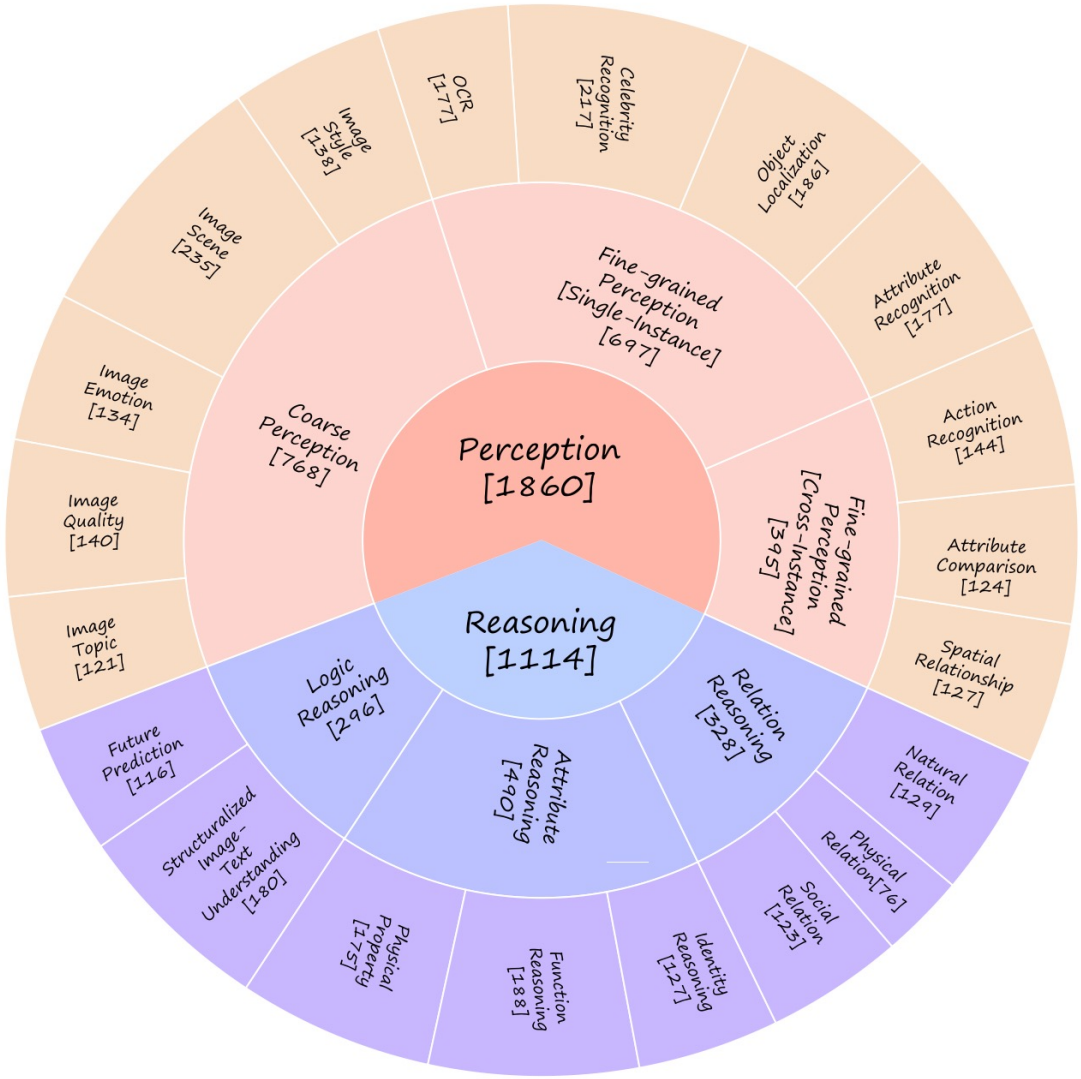
MME的评价策略

- **让模型回答“yes”或“no”**
  - 指令包括两部分，分别是一个简明的问题和一个描述“Please answer yes or no.”。
  - 稳定性测试：对于每张测试图片，人工地设计两条指令，两条指令的问题不同，回答分别是“yes”和“no”。

- **评价指标**
  - “**accuracy**”是根据每个问题计算的。
  - “**accuracy+**”是根据每张图片计算的，其中两个问题都需要被正确回答。
  - **感知分数**　　是所有感知子任务的分数总和。
  - **认知分数**　　以相同的方式计算。

MMBench：一个综合全面的评测基准

- 三个水平的能力维度（L-1到L-3），其中包括20种不同的子能力。
- **L-1：感知**和**推理**
- L-2 感知：1.粗粒度感知，2.细粒度单实例感知，3.细粒度跨实例感知
- L-2认知：1.属性推理，2.关系推理，3.逻辑推理
- **L-3能力**是进一步从L-2能力中划分出来的。



MMBench的评价策略

- **循环评价策略**
  - 循环评价将问题提供给VLM多次（使用不同的提示，调换答案的位置），并检查VLM是否在所有尝试中都成功解决了问题。
- **基于ChatGPT的答案抽取**
  - 为了解决VLM自由形式输出的问题，ChatGPT被利用来帮助抽取选择。

The original VL problem:
Q: How many apples are there in the image?
A. 4; B. 3; C. 2; D. 1 GT: A

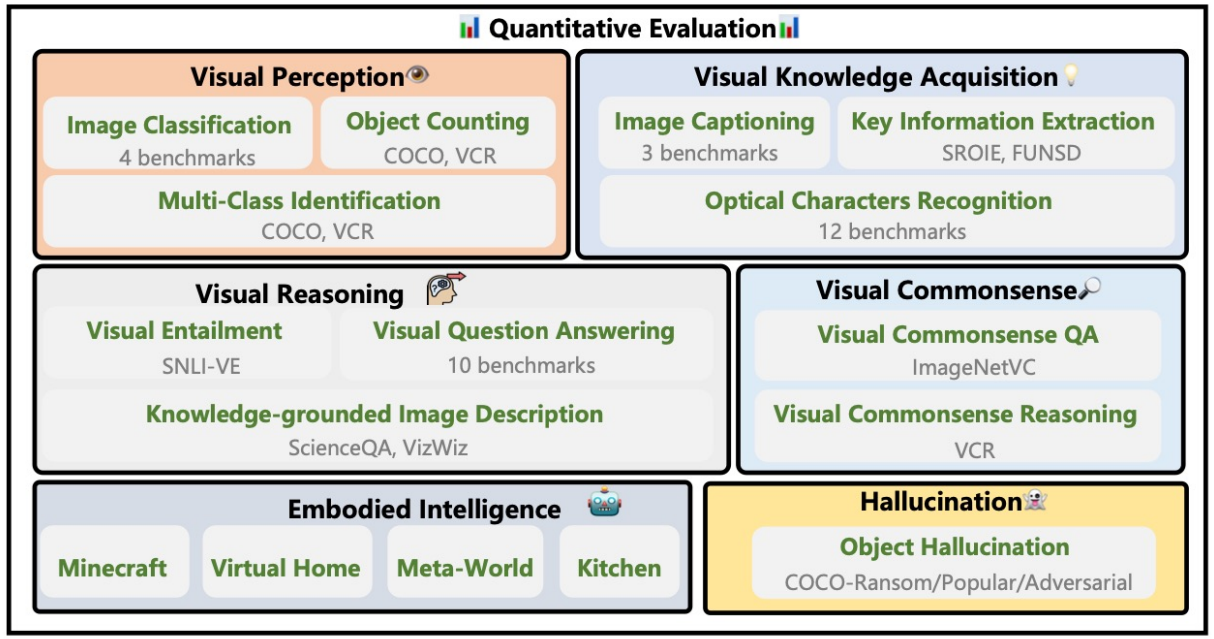
Circular Evaluation

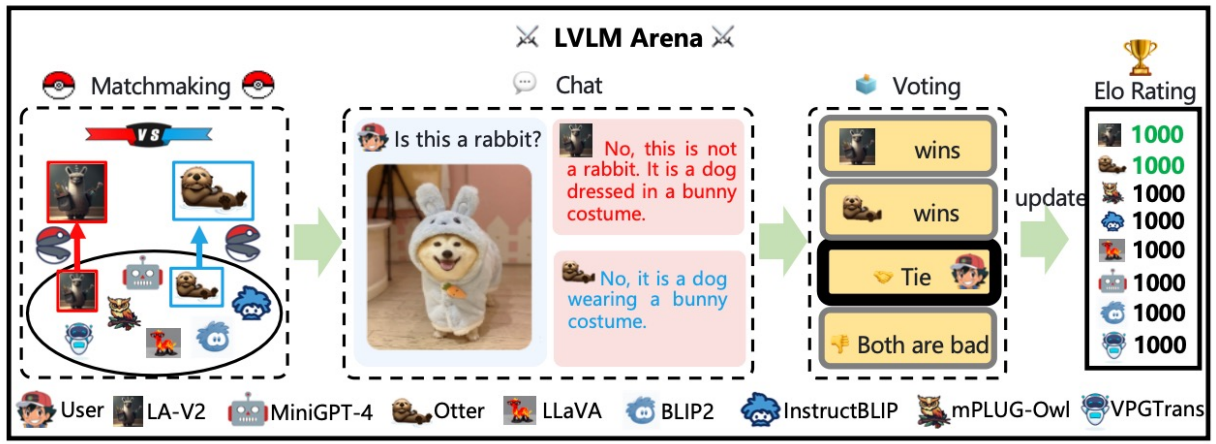
4 Passes in Circular Evaluation (choices with circular shift):
1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓
2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓
3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C ✗
4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.

LVLM-eHub：模型评测擂台赛

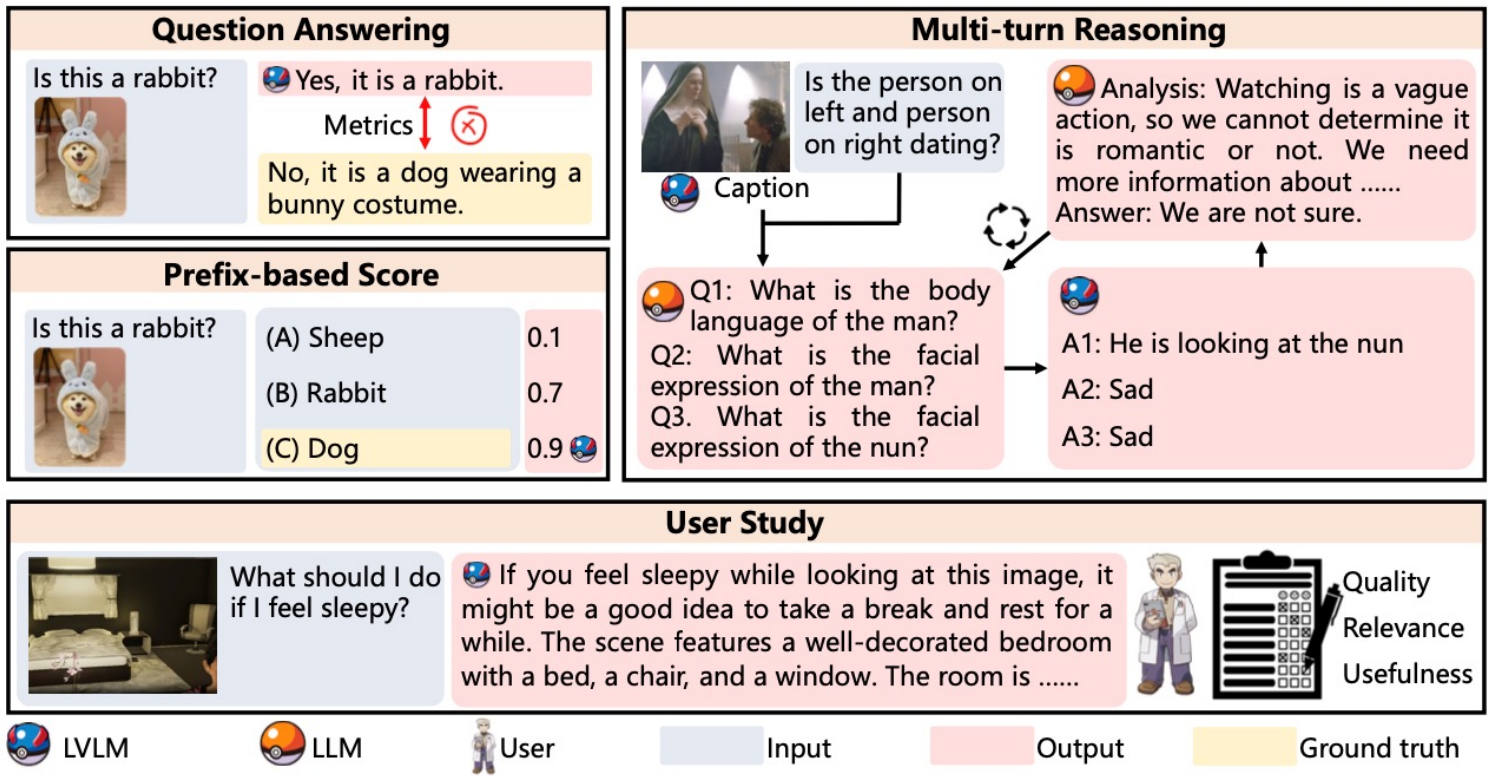
- 模型在定量评测中的**六个关键能力**。
- 针对不同任务/数据集量身定制的评测方法。
- 在一个在线平台LVLM Arena上，用户可以参与在线评价，通过与两个匿名模型聊天并选择他们偏好的模型。





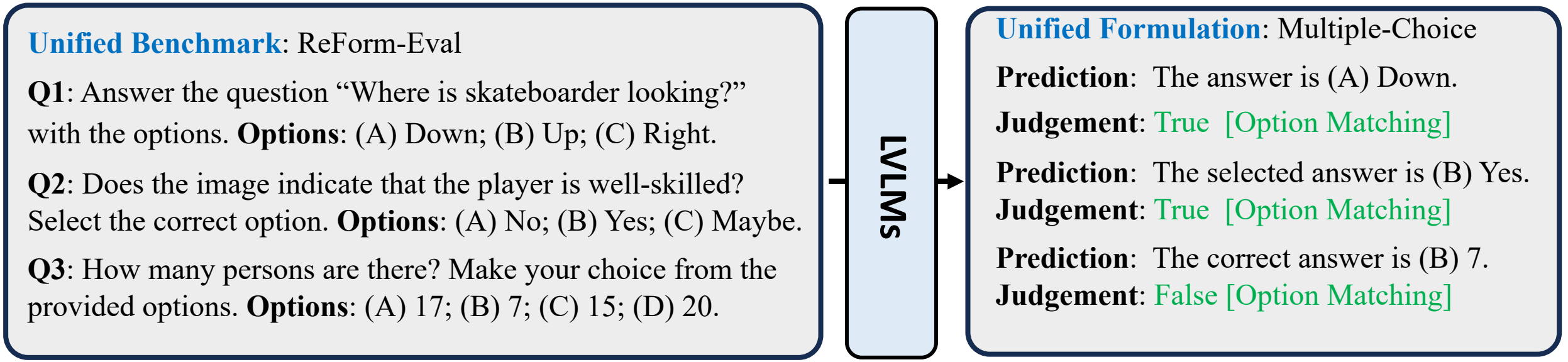
LVLM-eHub的在线评价

- 从模型集合中抽取两个模型。
- 用户与保持匿名的模型交谈。随后，**用户投票选出更好的模型**。
- 包括三个主要组成部分：
  - **配对**
  - **聊天**
  - **投票**



ReForm-Eval：“新瓶装旧酒”的基准构建方法

- 将面向任务的数据样本重新制定为与LVLM兼容的统一格式
  - 特殊的文本生成问题：对于光学字符识别（ORC）和图像描述任务
  - 多选题：对于剩余的其它任务
- 使用统一和兼容的形式实现通用和高效的评估

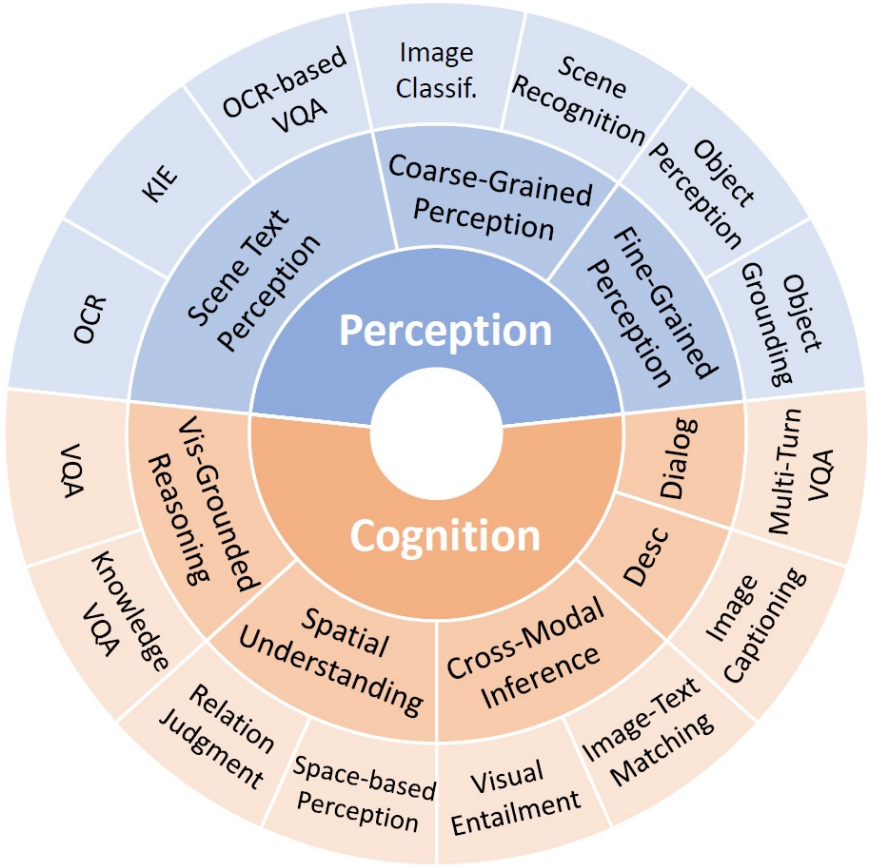


[*] indicates the evaluation method. Red and Green represent Wrong and Correct judgement. EM is short for “exactly matched”.

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

ReForm-Eval构建

- 61个现有基准数据集，来自2个主要类别、8个子类别和15个任务
- 专门的文本生成：
  - 视觉描述任务
  - OCR相关任务
- 多选题：
  - 标签 → 正选项
  - 难负选项:
    - 分类：类别之间的语义关系
    - 开放式QA: ChatGPT生成
    - 其它：任务特定的策略

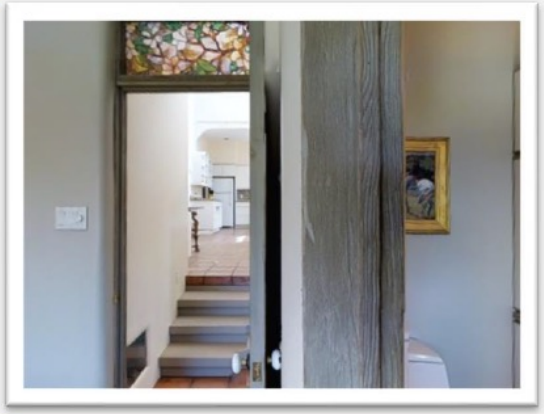


ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

ReForm-Eval构建：空间理解

- 从Matterport3D中构建MP3D-spatial，用于在真实世界的VLN评估

Question: Describe the spatial connection between vase and mantel within the image.
Options:
(A) The vase is inside the mantel;
(B) The vase is right of the mantel;
(C) The vase is next to the mantel;
(D) The vase is on the top of the mantel.
Answer: (D)



Question: In the image, point out the object that has the greatest distance from you.
Options:
(A) picture;
(B) refrigerator;
(C) stairs;
(D) unknown.
Answer: (B)

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

ReForm-Eval评价：统一的形式

- 文本生成问题的评估取决于场景：
  - 视觉描述：
    - 为了简洁的输出，限制了最大的输出长度
    - 指标：CIDEr（参照BLIP-2）
  - OCR相关：
    - 指标：token-level精度，出现在输出中的目标token的比例
- 多选题
  - 输出中的选项匹配：检测输出中的 “(A)” 等选项标记
  - 指标：准确性
  - 挑战：现存的LVLM可能并不会遵循多选指令
    - E.g., 生成 “Blue” 而不是“(A) Blue” 或 “(A)”

解耦LVLM的指令遵循能力

- 黑盒方法：上下文学习（In-Context Learning）
  - 指导LVLM通过ICL以所需格式生成：

$X_{\text{system-message}}$
Human: Can you see the image? Options: (A) Yes; (B) No; (C) Not Sure; (D) Maybe.
Assistant: The answer is (A) Yes.
Human: X_{question} Options: X_{options}
Assistant: The answer is

  - 其中上下文样本仅有文本内容并且不提供图片的信息。
- 白盒方法：似然率的（likelihood）评价
  - 计算每个选项的likelihood，并选择可能性最高的选项：

$$\hat{c} = \arg\max_{c^i \in C} P_\theta(c^i|v,q) = \arg\max_{c^i \in C} \sum_{t=1}^{t_c} P_\theta(c_t^i|v,q,c_{<t}^i)$$

  - 其中$C = \{c^i\}_{i=1}^N$ 是选项，v 是图像，q 是问题，P_θ 通过LVLM建模。

大视觉语言模型的输出稳定性评价

- 大模型对提示敏感：
  - 每个样本使用不同但等效的提示进行多次测试
  - 不同的指令模板、打乱选项、随机选项标记
  - 最终性能是多次测试的平均值
- 不稳定性测量：
  - 预测分布的熵（仅适用于多选题）
  - $e = -\sum_{i=1}^{N} p_i \log(p_i)$ where $p_i = \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}(\hat{c_j} = c_i)$
  - 其中 M 是多次测试的数量，$\hat{c_j}$是第j次测试的结果。

ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks，2023

Reform-Eval的总体实验

- 总体性能：13种方法的16个模型，具有不同的基座

Model	Model Architecture					
	Vis Encoder	LLM	Connection Module	#oP	#oTP	#oVT
BLIP-2	ViT-G/14	FlanT5-XL	Q-Former	3.94B	106.7M	32
InstructBLIP$_F$	ViT-G/14	FlanT5-XL	Q-Former	4.02B	187.2M	32
InstructBLIP$_V$	ViT-G/14	Vicuna-7B	Q-Former	7.92B	188.8M	32
LLaVA$_V$	ViT-L/14	Vicuna-7B	Linear	7.05B	6.74B	256
LLaVA$_{L_2}$	ViT-L/14	LLaMA2-7B	Linear	7.05B	6.74B	256
MiniGPT4	ViT-G/14	Vicuna-7B	Q-Former+Linear	7.83B	3.1M	32
mPLUG-Owl	ViT-L/14	LLaMA-7B	Perceiver	7.12B	384.6M	65
PandaGPT	ImageBind	Vicuna-7B+LoRA	Linear	7.98B	37.8M	1
IB-LLM	ImageBind	LLaMA-7B+LoRA+BT	BindNet+Gate	8.61B	649.7M	1
LA-V2	ViT-L/14	LLaMA-7B+BT	Linear+Adapter+Gate	7.14B	63.1M	10
mmGPT	ViT-L/14	LLaMA-7B+LoRA	Perceiver+Gate	8.37B	23.5M	64
Shikra	ViT-L/14	Vicuna-7B	Linear	6.74B	6.44B	256
Lynx	ViT-G/14	Vicuna-7B+Adapter	Perceiver	8.41B	688.4M	64
Cheetor$_V$	ViT-G/14	Vicuna-7B	Query+Linear+Q-Former	7.84B	6.3M	32
Cheetor$_{L_2}$	ViT-G/14	LLaMA2-Chat	Query+Linear+Q-Former	7.84B	6.3M	32
BLIVA	ViT-G/14	Vicuna-7B	Q-Former+Linear	7.92B	194.6M	32

PS: Underlined represents a trainable component. "BT" represents bias-tuning . "BindNet" represents bind network.

Table 7: Model architecture of different LVLMs. "#oP", "#oTP", and "#oVT" are number of total parameters, number of trainable parameters, and number of visual tokens, respectively.

ReForm-Eval的优点

- 数据丰富的全面评价：
  - 评估维度包含感知到推理
  - 重新制定了61个基准数据集，丰富的数据收集（~3000样本每维度（MMBench/MME大小的10倍））
  - 无需人工标注
- 高效的评价：
  - 基于统一形式的通用评价方法
  - 无需任务特定的评价方法（在LVLM-ehub中的）
  - 无需ChatGPT或人工的帮助（在LAMM和MMBench中的）

ReForm-Eval的优点

- 可靠的评价：
  - 黑盒和白盒两种辅助LVLM多项选择题的评估方法
- 不稳定性感知的评价：
  - 使用不同但等效的提示对同一样本进行多次测试
  - 对多选题的直接的不稳定性测量

ReForm-Eval开源了！

- https://github.com/FudanDISC/ReForm-Eval/

小结

- 大模型展示了强大的综合能力，对于它的评价也变得复杂

- （1）任务/能力的多样性
- （2）评价方法的高效性
- （3）输出结果的稳定性
- （4）测试样本的不可见性