Dynamic Multimodal Evaluation with Flexible Complexity by Vision-Language Bootstrapping

ICLR 2025 Oral

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Introduction

Despite the proliferation of LVLM evaluations, there are **two key challenges** associated with current evaluation benchmarks.

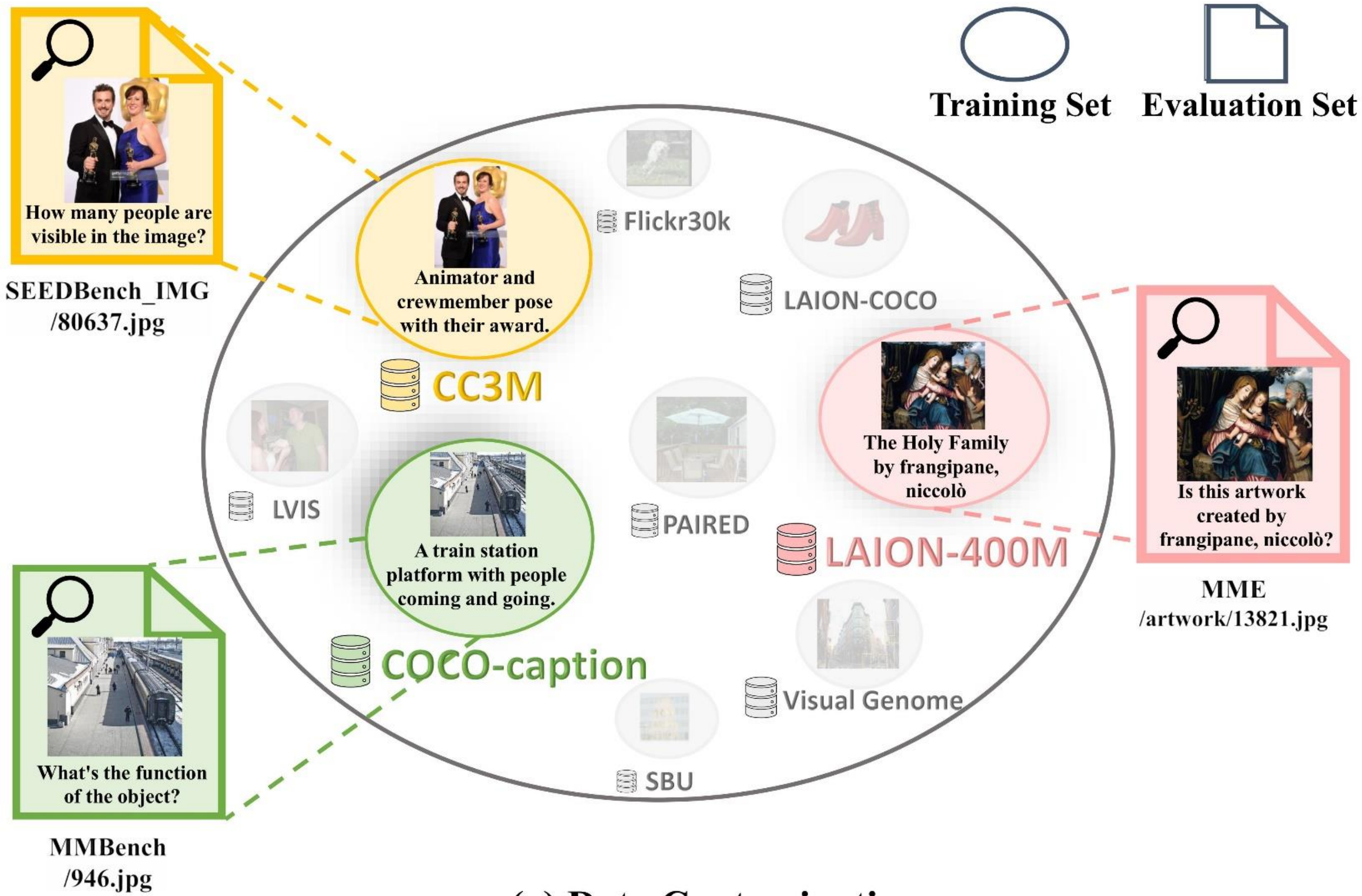
1. **Data contamination**

   Many evaluation benchmarks and training data are constructed from similar sources, leading to a high likelihood of overlap.

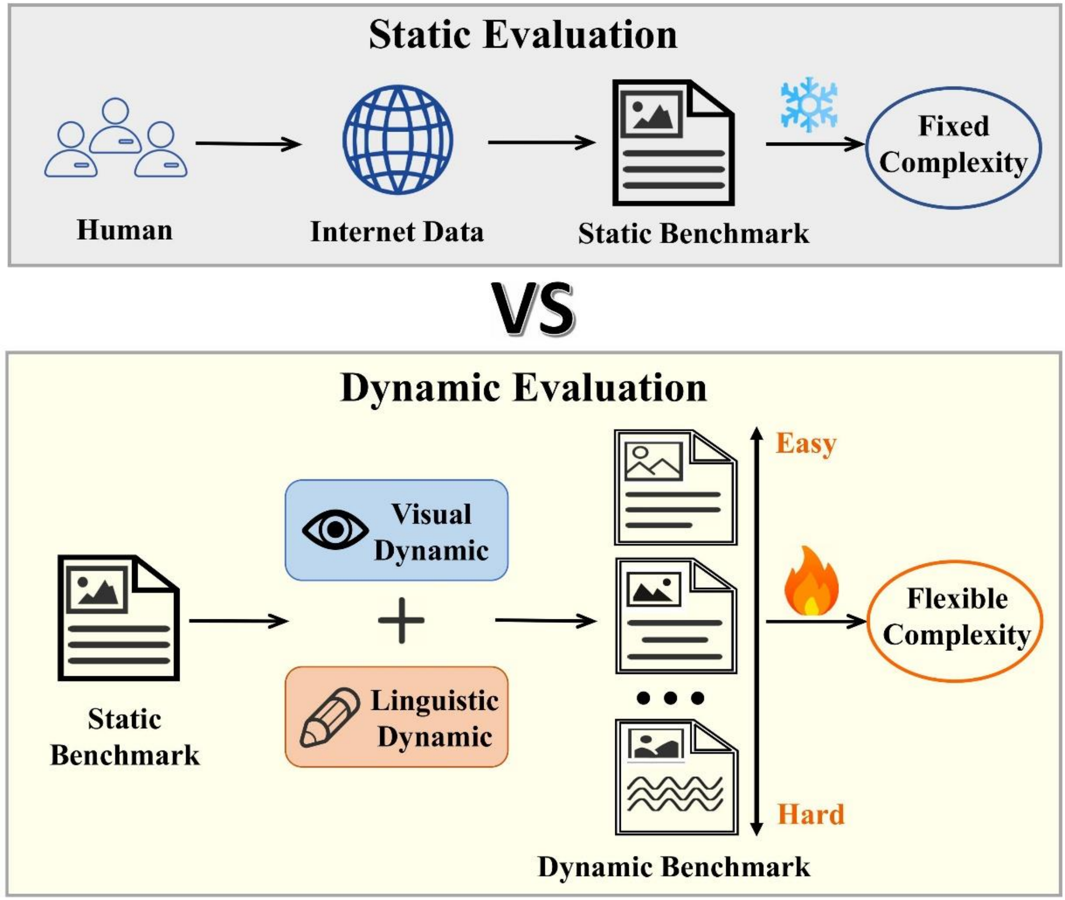
2. **Static dataset with fixed complexity**

   Existing benchmarks for LVLMs are manually collected. Once constructed, they are static with a fixed complexity.

Introduction


Training Set
Evaluation Set
How many people are visible in the image?
SEEDBench_IMG /80637.jpg
Animator and crewmember pose with their award.
CC3M
Flickr30k
LAION-COCO
LVIS
PAIRED
The Holy Family by frangipane, niccolò
LAION-400M
Is this artwork created by frangipane, niccolò?
MME /artwork/13821.jpg
A train station platform with people coming and going.
COCO-caption
Visual Genome
SBU
What's the function of the object?
MMBench /946.jpg

(a) Data Contamination



(b) Dynamic Evaluation vs Static Evaluation

(a) some VQA pairs have overlap with training data.

(b) dynamic evaluation VS previous static evaluation.

Contribution

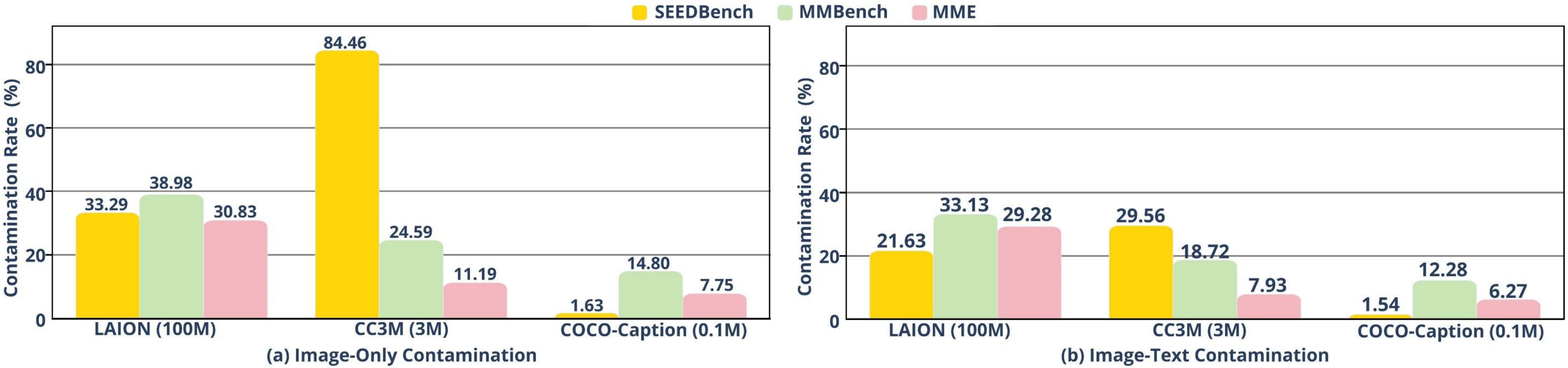
Contributions of Dynamic Multimodal Evaluation (DME) :

1) Delve into the data contamination issue of existing multimodal evaluation benchmarks and find a pronounced overlap between evaluation samples and pre-training data.

2) We propose a dynamic multimodal evaluation protocol called vision-language bootstrapping (VLB). VLB can evolve existing benchmarks with visual and linguistic dynamics, obtaining various variants with flexible complexity and reduced data contamination.

3) We perform comprehensive evaluations on a variety of popular LVLMs, indicating that the existing LVLMs still struggle to adapt to different user interactions and intent.

Quantifying Data Contamination Rate in Existing Benchmarks



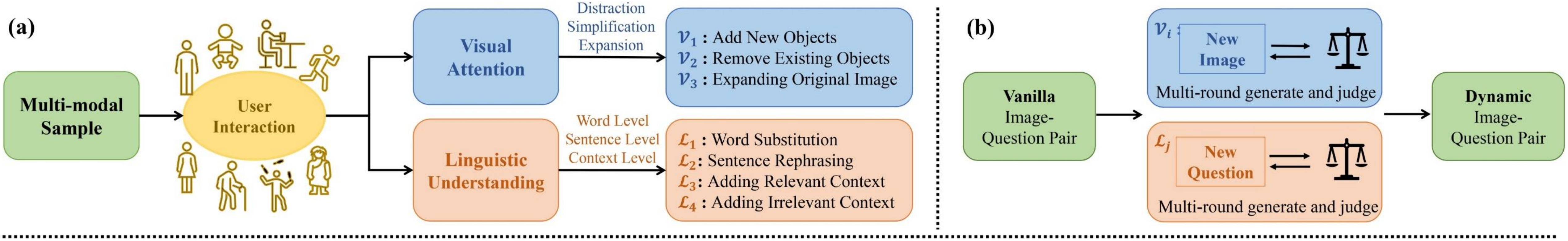
(a) Image-Only Contamination

(b) Image-Text Contamination

We quantify two kinds of data contamination rate between LVLM evaluation samples and pre-training data, both find a pronounced overlap.

Method

Two module of dynamic multimodal evaluation:


(a)
Multi-modal Sample
User Interaction
Visual Attention
Distraction
Simplification
Expansion
$\mathcal{V}_1$: Add New Objects
$\mathcal{V}_2$: Remove Existing Objects
$\mathcal{V}_3$: Expanding Original Image
Linguistic Understanding
Word Level
Sentence Level
Context Level
$\mathcal{L}_1$: Word Substitution
$\mathcal{L}_2$: Sentence Rephrasing
$\mathcal{L}_3$: Adding Relevant Context
$\mathcal{L}_4$: Adding Irrelevant Context
(b)
Vanilla Image-Question Pair
$\mathcal{V}_i$:
New Image
Multi-round generate and judge
$\mathcal{L}_j$
New Question
Multi-round generate and judge
Dynamic Image-Question Pair

(a) Multimodal Bootstrapping Module

By simulating real LVLM‘s user interaction in visual attention and linguistic understanding, we design image (V1, V2, V3) and language (L1, L2, L3, L4) bootstrapping strategies

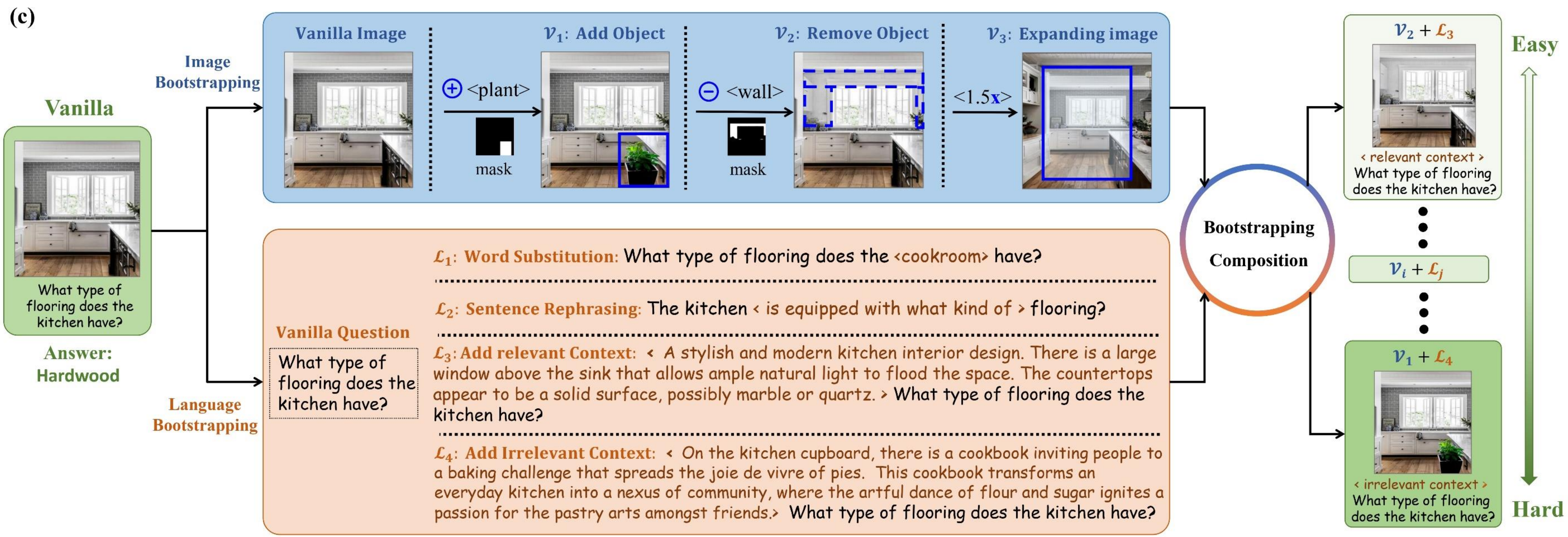
(b) Judge Module

Judge module in ensuring that generated images and questions maintain consistent with the original.

Method

Image Bootstrapping strategy:

V1: Adding new object, **V2:** Remove existing objects, **V3:** expand original images


(c)
Vanilla
Answer:
Hardwood
Image Bootstrapping
Language Bootstrapping
Vanilla Image
$\mathcal{V}_1$: Add Object
⊕ <plant>
mask
$\mathcal{V}_2$: Remove Object
⊖ <wall>
mask
$\mathcal{V}_3$: Expanding image
<1.5x>
Vanilla Question
Bootstrapping Composition
$\mathcal{V}_2 + \mathcal{L}_3$
$\mathcal{V}_i + \mathcal{L}_j$
$\mathcal{V}_1 + \mathcal{L}_4$
Easy
Hard

What type of flooring does the kitchen have?

$\mathcal{L}_1$: **Word Substitution:** What type of flooring does the <cookroom> have?

$\mathcal{L}_2$: **Sentence Rephrasing:** The kitchen < is equipped with what kind of > flooring?

$\mathcal{L}_3$: **Add relevant Context:** < A stylish and modern kitchen interior design. There is a large window above the sink that allows ample natural light to flood the space. The countertops appear to be a solid surface, possibly marble or quartz. > What type of flooring does the kitchen have?

$\mathcal{L}_4$: **Add Irrelevant Context:** < On the kitchen cupboard, there is a cookbook inviting people to a baking challenge that spreads the joie de vivre of pies. This cookbook transforms an everyday kitchen into a nexus of community, where the artful dance of flour and sugar ignites a passion for the pastry arts amongst friends.> What type of flooring does the kitchen have?

< relevant context > What type of flooring does the kitchen have?

< irrelevant context > What type of flooring does the kitchen have?

Language Bootstrapping strategy:

L1: Word substitution, **L2:** Sentence rephrasing, **L3:** Adding relevant context, **L4:** Adding irrelevant context

Method

Compositional Bootstrapping :

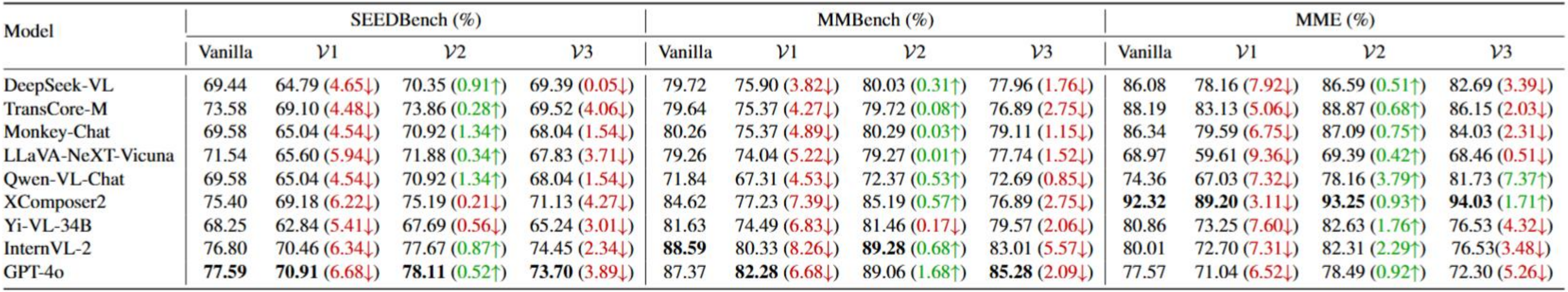
Due to every single dynamic strategy for image and question being atomic, we can investigate two kinds of bootstrapping composition.

1) **Paired multimodal composition.** We can compose visual bootstrapping **V** and linguistic bootstrapping **L**, obtaining a total of 12 dynamic variants.
2) **Multi-strategy composition.** We can also stack multiple image bootstrapping strategies on a single image or multiple language bootstrapping strategies on the question.

Results of image bootstrapping strategies

Settings:

We selected five popular benchmarks to assess many open-sourced LVLMs and close-source APIs, encompassing MME, MMBench, SEEDBench, MMvet, and LLaVABench.



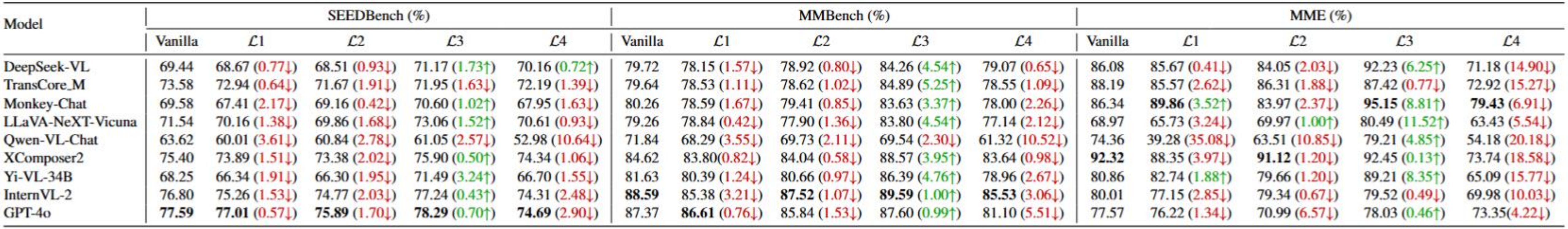
Model	SEEDBench (%)				MMBench (%)				MME (%)			
	Vanilla	$\mathcal{V}1$	$\mathcal{V}2$	$\mathcal{V}3$	Vanilla	$\mathcal{V}1$	$\mathcal{V}2$	$\mathcal{V}3$	Vanilla	$\mathcal{V}1$	$\mathcal{V}2$	$\mathcal{V}3$
DeepSeek-VL	69.44	64.79 (4.65↓)	70.35 (0.91↑)	69.39 (0.05↓)	79.72	75.90 (3.82↓)	80.03 (0.31↑)	77.96 (1.76↓)	86.08	78.16 (7.92↓)	86.59 (0.51↑)	82.69 (3.39↓)
TransCore-M	73.58	69.10 (4.48↓)	73.86 (0.28↑)	69.52 (4.06↓)	79.64	75.37 (4.27↓)	79.72 (0.08↑)	76.89 (2.75↓)	88.19	83.13 (5.06↓)	88.87 (0.68↑)	86.15 (2.03↓)
Monkey-Chat	69.58	65.04 (4.54↓)	70.92 (1.34↑)	68.04 (1.54↓)	80.26	75.37 (4.89↓)	80.29 (0.03↑)	79.11 (1.15↓)	86.34	79.59 (6.75↓)	87.09 (0.75↑)	84.03 (2.31↓)
LLaVA-NeXT-Vicuna	71.54	65.60 (5.94↓)	71.88 (0.34↑)	67.83 (3.71↓)	79.26	74.04 (5.22↓)	79.27 (0.01↑)	77.74 (1.52↓)	68.97	59.61 (9.36↓)	69.39 (0.42↑)	68.46 (0.51↓)
Qwen-VL-Chat	69.58	65.04 (4.54↓)	70.92 (1.34↑)	68.04 (1.54↓)	71.84	67.31 (4.53↓)	72.37 (0.53↑)	72.69 (0.85↓)	74.36	67.03 (7.32↓)	78.16 (3.79↑)	81.73 (7.37↑)
XComposer2	75.40	69.18 (6.22↓)	75.19 (0.21↓)	71.13 (4.27↓)	84.62	77.23 (7.39↓)	85.19 (0.57↑)	76.89 (2.75↓)	**92.32**	**89.20** (3.11↓)	**93.25** (0.93↑)	**94.03** (1.71↑)
Yi-VL-34B	68.25	62.84 (5.41↓)	67.69 (0.56↓)	65.24 (3.01↓)	81.63	74.49 (6.83↓)	81.46 (0.17↓)	79.57 (2.06↓)	80.86	73.25 (7.60↓)	82.63 (1.76↑)	76.53 (4.32↓)
InternVL-2	76.80	70.46 (6.34↓)	77.67 (0.87↑)	74.45 (2.34↓)	**88.59**	80.33 (8.26↓)	**89.28** (0.68↑)	83.01 (5.57↓)	80.01	72.70 (7.31↓)	82.31 (2.29↑)	76.53(3.48↓)
GPT-4o	**77.59**	**70.91** (6.68↓)	**78.11** (0.52↑)	**73.70** (3.89↓)	87.37	**82.28** (6.68↓)	89.06 (1.68↑)	**85.28** (2.09↓)	77.57	71.04 (6.52↓)	78.49 (0.92↑)	72.30 (5.26↓)

Results:

V1 (Adding New Objects) and **V3** (Expanding Original Images) generally result in a decrease in accuracy, while **V2** (Removing Existing Objects) slightly boost the performance.

Experiment

Results of language bootstrapping strategies



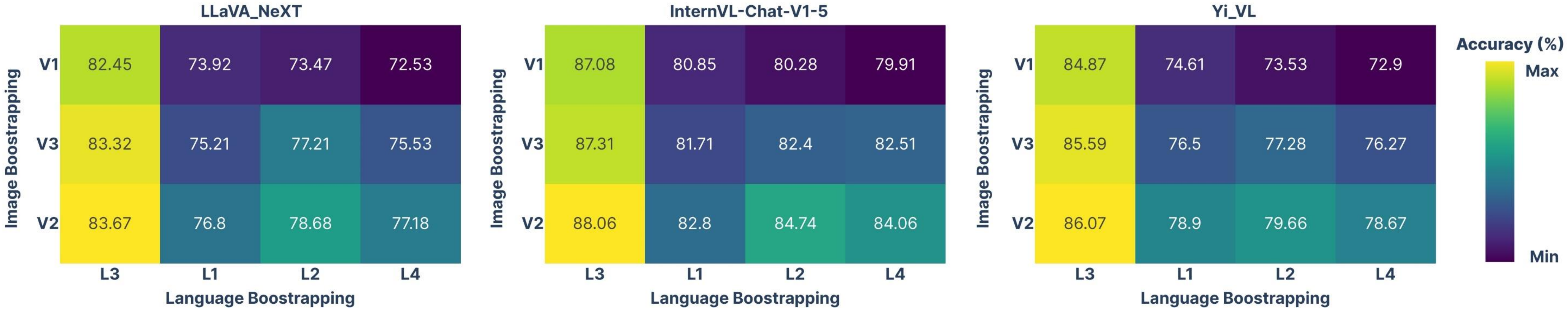
Model	SEEDBench (%)					MMBench (%)					MME (%)				
	Vanilla	$\mathcal{L}1$	$\mathcal{L}2$	$\mathcal{L}3$	$\mathcal{L}4$	Vanilla	$\mathcal{L}1$	$\mathcal{L}2$	$\mathcal{L}3$	$\mathcal{L}4$	Vanilla	$\mathcal{L}1$	$\mathcal{L}2$	$\mathcal{L}3$	$\mathcal{L}4$
DeepSeek-VL	69.44	68.67 (0.77↓)	68.51 (0.93↓)	71.17 (1.73↑)	70.16 (0.72↑)	79.72	78.15 (1.57↓)	78.92 (0.80↓)	84.26 (4.54↑)	79.07 (0.65↓)	86.08	85.67 (0.41↓)	84.05 (2.03↓)	92.23 (6.25↑)	71.18 (14.90↓)
TransCore_M	73.58	72.94 (0.64↓)	71.67 (1.91↓)	71.95 (1.63↓)	72.19 (1.39↓)	79.64	78.53 (1.11↓)	78.62 (1.02↓)	84.89 (5.25↑)	78.55 (1.09↓)	88.19	85.57 (2.62↓)	86.31 (1.88↓)	87.42 (0.77↓)	72.92 (15.27↓)
Monkey-Chat	69.58	67.41 (2.17↓)	69.16 (0.42↓)	70.60 (1.02↑)	67.95 (1.63↓)	80.26	78.59 (1.67↓)	79.41 (0.85↓)	83.63 (3.37↑)	78.00 (2.26↓)	86.34	**89.86** (3.52↑)	83.97 (2.37↓)	**95.15** (8.81↑)	**79.43** (6.91↓)
LLaVA-NeXT-Vicuna	71.54	70.16 (1.38↓)	69.86 (1.68↓)	73.06 (1.52↑)	70.61 (0.93↓)	79.26	78.84 (0.42↓)	77.90 (1.36↓)	83.80 (4.54↑)	77.14 (2.12↓)	68.97	65.73 (3.24↓)	69.97 (1.00↑)	80.49 (11.52↑)	63.43 (5.54↓)
Qwen-VL-Chat	63.62	60.01 (3.61↓)	60.84 (2.78↓)	61.05 (2.57↓)	52.98 (10.64↓)	71.84	68.29 (3.55↓)	69.73 (2.11↓)	69.54 (2.30↓)	61.32 (10.52↓)	74.36	39.28 (35.08↓)	63.51 (10.85↓)	79.21 (4.85↑)	54.18 (20.18↓)
XComposer2	75.40	73.89 (1.51↓)	73.38 (2.02↓)	75.90 (0.50↑)	74.34 (1.06↓)	84.62	83.80(0.82↓)	84.04 (0.58↓)	88.57 (3.95↑)	83.64 (0.98↓)	**92.32**	88.35 (3.97↓)	**91.12** (1.20↓)	92.45 (0.13↑)	73.74 (18.58↓)
Yi-VL-34B	68.25	66.34 (1.91↓)	66.30 (1.95↓)	71.49 (3.24↑)	66.70 (1.55↓)	81.63	80.39 (1.24↓)	80.66 (0.97↓)	86.39 (4.76↑)	78.96 (2.67↓)	80.86	82.74 (1.88↑)	79.66 (1.20↓)	89.21 (8.35↑)	65.09 (15.77↓)
InternVL-2	76.80	75.26 (1.53↓)	74.77 (2.03↓)	77.24 (0.43↑)	74.31 (2.48↓)	**88.59**	85.38 (3.21↓)	**87.52** (1.07↓)	**89.59** (1.00↑)	**85.53** (3.06↓)	80.01	77.15 (2.85↓)	79.34 (0.67↓)	79.52 (0.49↓)	69.98 (10.03↓)
GPT-4o	**77.59**	**77.01** (0.57↓)	**75.89** (1.70↓)	**78.29** (0.70↑)	**74.69** (2.90↓)	87.37	**86.61** (0.76↓)	85.84 (1.53↓)	87.60 (0.99↑)	81.10 (5.51↓)	77.57	76.22 (1.34↓)	70.99 (6.57↓)	78.03 (0.46↑)	73.35(4.22↓)

Results:

Except for **L3** (Adding relevant context), the other strategies **L1(**Word substitution), **L2 (**Sentence rephrasing), **L4 (**Adding irrelevant context) all result in a degradation of LVLM performance.

This highlights the challenge LVLMs face in addressing questions posed by different individuals from diverse backgrounds. Conversely, incorporating image relevant captions helps the model better understand questions.

Results of composing image and language bootstrapping strategies



LLaVA_NeXT

	L3	L1	L2	L4
V1	82.45	73.92	73.47	72.53
V3	83.32	75.21	77.21	75.53
V2	83.67	76.8	78.68	77.18

InternVL-Chat-V1-5

	L3	L1	L2	L4
V1	87.08	80.85	80.28	79.91
V3	87.31	81.71	82.4	82.51
V2	88.06	82.8	84.74	84.06

Yi_VL

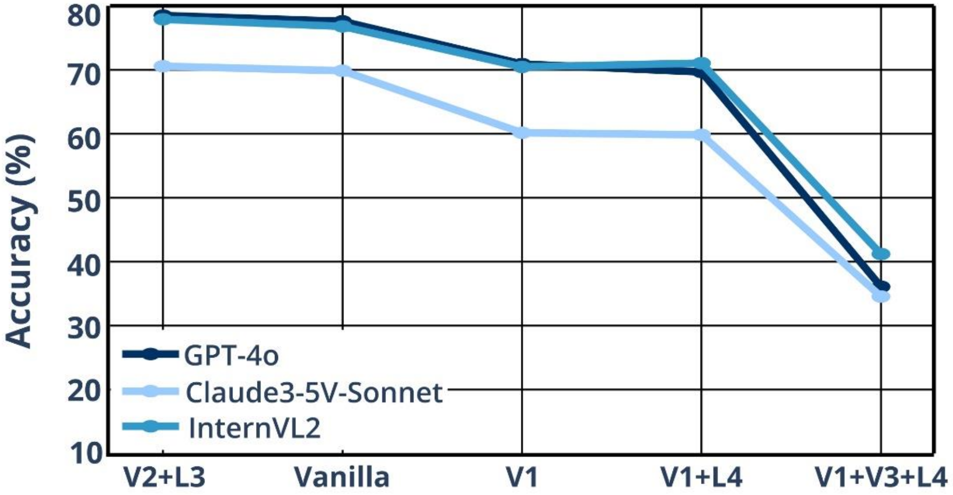
	L3	L1	L2	L4
V1	84.87	74.61	73.53	72.9
V3	85.59	76.5	77.28	76.27
V2	86.07	78.9	79.66	78.67

Paired multimodal composition:

We combine our strategies on Vision and Language and obtained 12 dynamic variants.

We observe that these LVLMs exhibit varying performance when faced with different variants, which means they possess different complexity.

Results of composing image and language bootstrapping strategies

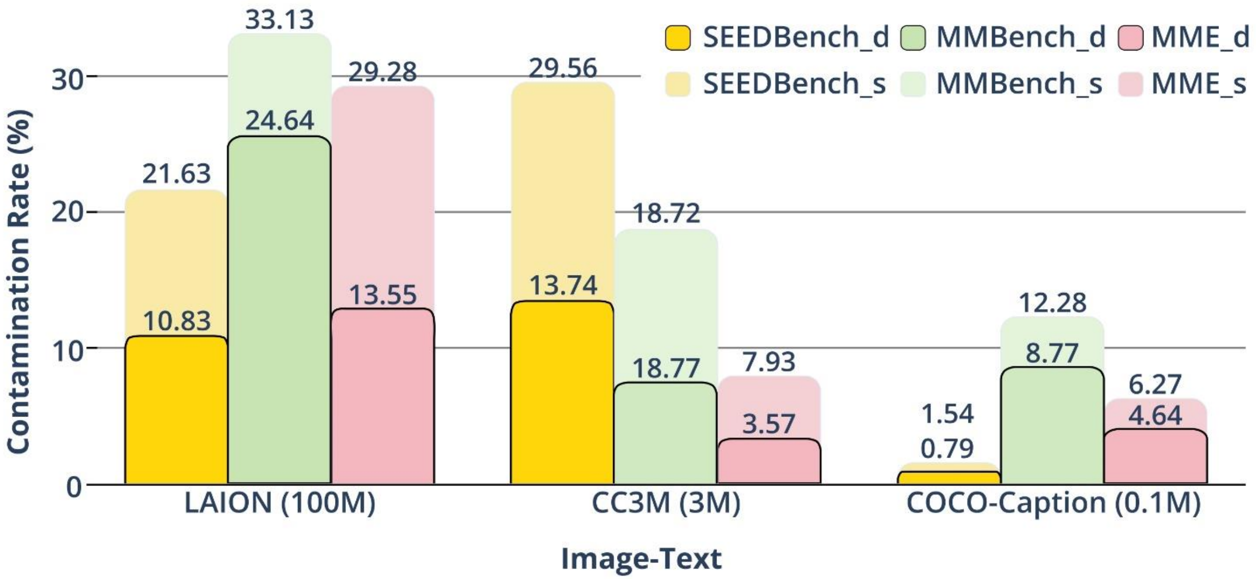


Multi-strategy composition:

As the number of hard strategies we use increases from 0-3, the difficulty of the benchmark also increases, leading to a gradual decline in the accuracy of LVLM.

Experiment

Can our VLB reduce data contamination?

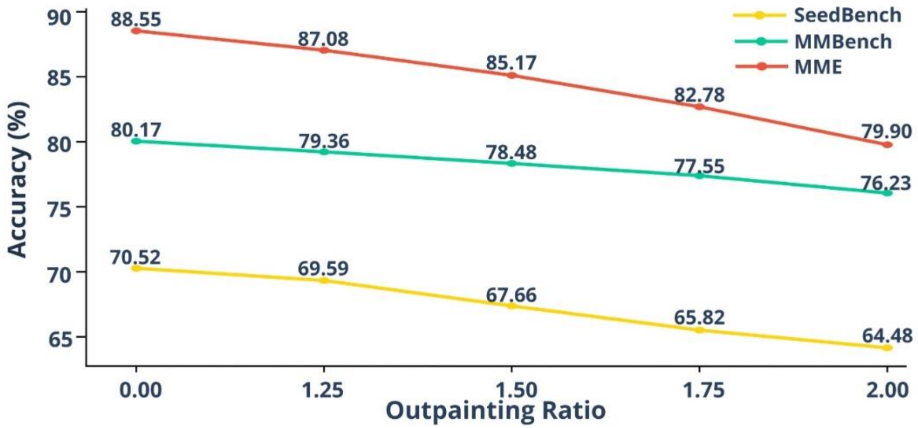


Yes, we applied the same methods as the vanilla to redetect the hardest variant, and found a significant reduction in data contamination rate among training sets.

Experiment

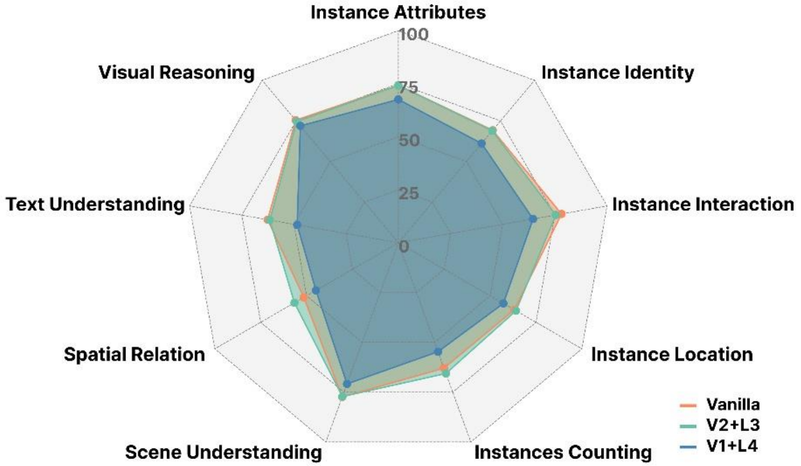
The effect of image expanding ratio

As the expanding ratios increase, the accuracy of the models in correctly answering questions decreases, and the rate of this decrease becomes steeper.



The effect of VLB on different tasks

Performance changes more significantly in tasks like `Instance Interaction', `Text Understanding' and `Spatial Relation’.



上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

THANKS FOR WATCHING

SJTU