HallE-Control: Controlling Object Hallucination in Large Multimodal Models

Bohan Zhai[1]* **Shijia Yang**[2]* **Chenfeng Xu**[3] **Sheng Shen**[3]

Kurt Keutzer[3] **Chunyuan Li**[1] **Manling Li**[4]

ByteDance Inc.[1], Stanford University[2], UC Berkeley[3], UIUC[4]

Bohan Zhai

UC Berkeley
在 berkeley.edu 的电子邮件经过验证
Machine Learning

Kurt Keutzer

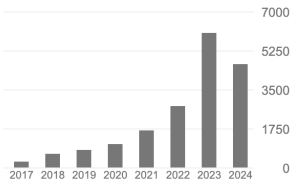
Professor of the Graduate School, EECS, University of California, Berkeley
在 berkeley.edu 的电子邮件经过验证 - 首页
artificial intelligence systems deep learning efficient computation

Chunyuan Li

Microsoft Research, Redmond
在 microsoft.com 的电子邮件经过验证 - 首页
Deep Learning Vision Language

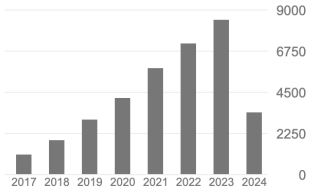
引用次数 查看全部

	总计	2019 年至今
引用	18549	17125
h 指数	62	58
i10 指数	112	107



引用次数 查看全部

	总计	2019 年至今
引用	54702	32046
h 指数	103	69
i10 指数	307	168



Background & Motivation

- LMM在VQA等任务得到很大进展，但是幻觉问题日益凸显，在image caption任务中，幻觉问题尤为严重，具体幻觉为以下三类：
  - Object Existence：出现不存在的对象
  - Object Attribute：不正确的对象属性，如颜色、形状、大小等
  - Object Relationship：不正确的对象交互方式，比如相对位置、交互状态等
- 现有对于幻觉的评估POPE等benchmark是通过VQA的形式来进行评估的，因此对于视觉细节的幻觉评估并不完整。即使是在这种benchmark下表现较好的模型，在实际应用中也存在很严重的幻觉现象。
- Motivation：从LLM decoder的大小、Instruction数据的质量和粒度以及视觉编码器的分辨率等因素探索Object Existence幻觉的根本原因

Hallucination Analysis——Benchmark

- 我们选择 LLaVA、InstructBLIP 和 Shikra：LLaVA 和 Shikra 共享相同的模型结构；Shikra 和 InstructBLIP 使用混合数据集和多任务指令数据；InstructBLIP 仅微调 Q-former，而其他微调投影仪和 LLM。
- 基于VQA的方式：每张图片对应一个yes/no的问题。
- 基于caption的方式：

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}$$

Benchmark	Model	Accuracy↑	Precision↑	Recall↑	F1↑	Yes (%)
POPE - Random	LLaVA_{7B}	73.13	66.95	94.53	78.39	72.78
	LLaVA_{13B}	78.49	73.57	90.93	81.34	63.71
	Shikra_{7B}	86.99	94.77	79.13	86.25	43.04
	InstructBLIP_{7B}	86.60	80.74	96.13	87.77	59.53
POPE - Popular	LLaVA_{7B}	59.87	55.88	93.80	70.03	83.93
	LLaVA_{13B}	70.80	64.73	91.40	75.79	70.60
	Shikra_{7B}	84.35	88.10	79.43	83.54	45.08
	InstructBLIP_{7B}	71.27	64.20	96.13	76.99	74.87
POPE - Adversarial	LLaVA_{7B}	57.06	54.07	93.93	68.63	86.87
	LLaVA_{13B}	63.93	59.03	91.07	71.63	77.13
	Shikra_{7B}	82.88	85.32	79.43	82.27	46.55
	InstructBLIP_{7B}	72.10	65.13	95.13	77.32	73.03
Benchmark	**Model**	**Existence↑**	**Count↑**	**Position↑**	**Color↑**	**Total↑**
MME	LLaVA_{7B}	150.00	48.33	50.00	55.00	303.33
	LLaVA_{13B}	180.00	113.33	55.00	95.00	443.33
	Shikra_{7B}	185.00	118.33	75.00	155.00	533.33
	InstructBLIP_{7B}	185.00	60.00	50.00	125.00	420.00

- 先用MLLM提取caption，然后使用GPT-4提取caption中的object，然后分析所有的object中有多少object是幻觉。
- 左侧：InstructBLIP平均每句0.8对象，LLaVA13B和Shikra每句7.7和7.5个对象。右侧：添加限制输出条件，使得平均句子长度和平均object数量接近。

Table 2: Comparison between CHAIR and our evaluation method, CCEval.

Model	CHAIR				CCEval (Ours)				
	$CHAIR_s$↓	$CHAIR_i$↓	Avg. Length↑	Avg. Object↑	$CHAIR_s$↓	$CHAIR_i$↓	Coverage↑	Avg. Length↑	Avg. Object↑
LLaVA_{7B}	24.1	9.1	42.5	3.7	72.00	19.7	32.74	92.27	9.19
LLaVA_{13B}	60.6	18.4	90.2	7.6	79.00	23.80	33.56	108.02	9.28
Shikra_{7B}	59.1	16.6	91.2	7.5	83.00	24.40	33.29	109.37	9.10
InstructBLIP_{7B}	1.4	1.7	2.3	0.8	72.00	22.30	29.76	108.42	8.04

Hallucination Analysis——LLM

- 扩大MLLM的规模之后，发现POPE指标都得到了一定的改善，但是在CCEval的指标上并没有变化，说明LLM自身不是减少幻觉的主要因素。

Table 3: Performance of LLaVA and InstructBLIP with different sizes of language decoder. LLaVA are trained on CC-595k for stage one and Instruction-150k for stage two.

Benchmark	Model	Accuracy↑	Precision↑	Recall↑	F1↑	Yes (%)
POPE - Random	LLaVA_{7B}	75.77	69.79	93.47	79.91	69.04
	LLaVA_{13B}	78.49	73.57	90.93	81.34	63.71
	LLaVA_{33B}	78.14	73.18	90.93	81.09	64.05
	InstructBLIP_{7B}	86.60	80.74	96.13	87.77	59.53
	$\text{InstructBLIP}_{13B}$	88.73	86.67	92.33	89.41	54.91
POPE - Popular	LLaVA_{7B}	65.07	59.60	93.53	72.81	78.47
	LLaVA_{13B}	70.80	64.73	91.40	75.79	70.60
	LLaVA_{33B}	72.43	66.45	90.60	76.67	68.17
	InstructBLIP_{7B}	71.27	64.20	96.13	76.99	74.87
	$\text{InstructBLIP}_{13B}$	80.53	74.70	92.33	82.59	61.80
POPE - Adversarial	LLaVA_{7B}	57.07	54.07	93.93	68.63	86.87
	LLaVA_{13B}	63.93	59.03	91.07	71.63	77.13
	LLaVA_{33B}	66.30	60.91	91.00	72.98	74.70
	InstructBLIP_{7B}	72.10	65.13	95.13	77.32	73.03
	$\text{InstructBLIP}_{13B}$	73.97	67.53	92.33	78.01	68.37
Benchmark	**Model**	$CHAIR_s$↓	$CHAIR_i$↓	**Coverage↑**	**Avg. Length↑**	**Avg. Object↑**
CCEval (Ours)	LLaVA_{7B}	82.00	25.30	33.58	109.89	9.31
	LLaVA_{13B}	79.00	23.80	33.56	108.02	9.28
	LLaVA_{33B}	82.00	21.80	31.26	106.85	9.07
	InstructBLIP_{7B}	72.00	22.30	29.76	108.42	8.04
	$\text{InstructBLIP}_{13B}$	64.00	16.70	33.60	101.63	8.06

Hallucination Analysis——data quality

扩大指令微调数据量，范围从 80K 到 2.4M。与在 150K 和 SVIT 上微调的模型相比，在 80K 指令数据上微调的 LLAAVA7B 模型表现出更少的对象存在幻觉。
结果表明，没有质量保证的额外数据可能会增加基于 VQA 和基于字幕的评估的幻觉。

Table 4: Performance of LLaVA$_{7B}$ with different sizes of data. 80K and 158K contains 80K and 158K data respectively, and SVIT contains 2.4M.

Benchmark	Finetune Data	Accuracy↑	Precision↑	Recall↑	F1↑	Yes (%)
POPE - Random	80K	73.13	66.95	94.53	78.39	72.78
	158K	75.77	69.79	93.47	79.91	69.04
	SVIT	52.34	52.00	97.87	67.92	97.01
POPE - Popular	80K	59.87	55.88	93.80	70.03	83.93
	158K	65.07	59.60	93.53	72.81	78.47
	SVIT	50.77	50.43	90.47	64.76	89.70
POPE - Adversarial	80K	57.07	54.07	93.93	68.63	86.87
	158K	58.47	55.00	93.07	69.14	84.6
	SVIT	51.37	50.77	90.33	65.00	88.97
Benchmark	**Finetune Data**	***CHAIR$_i$*↓**	***CHAIR$_s$*↓**	**Coverage↑**	**Avg. Length↑**	**Avg. Object↑**
CCEval (Ours)	80K	72.00	19.70	32.74	92.27	9.19
	158K	82.00	25.30	33.58	109.89	9.31
	SVIT	87.00	23.30	47.46	296.63	18.14

Hallucination Analysis——data quality

Image	Caption	GT Labels
	The image depicts a cozy entertainment room featuring a flat screen TV situated against the wall. A dining table can be seen next to the TV, with some bottles and a cup placed on top of it. Multiple video game controllers, including Wii remotes, are also present on the dining table, suggesting that people are using the TV to play video games. To the far right of the scene, another remote is located close to the wall. A person appears to be sitting or standing behind the dining table, likely either watching the game or waiting for their turn to play. The assortment of bottles on the table suggests that the guests are enjoying drinks during their video game session.	tv: 1 cup: 1 remote: 3 bottle: 4 person: 1 dining table: 1
	In the image, an orange mass transit trolley is making its way through a city. A person is crossing the street in front of the trolley while holding a garbage bag, appearing to be cautious about the approaching vehicle. Another person is standing close to the person crossing the street, and there are two more individuals nearby. In the scene, various vehicles surround the trolley, including cars, buses, and trucks. One of the cars is parked right behind the trolley, while another is situated farther back. Two buses can be seen, with one staying behind the trolley and the other on the right side of it. A truck is also present at the far left side of the scene. A woman nearby is holding a handbag, completing the busy urban setting.	car: 3 bus: 2 truck: 1 train: 1 person: 5 handbag: 1
	The image shows a table in a restaurant, which appears to have been recently used for a meal. The table is set with four place settings, including white dishes. On the table, there are multiple cups, a bowl, and a mix of silverware like forks and spoons laid out. Some of the dishes look dirty, with used napkins and eating utensils scattered around. In the background, two people can be seen, one situated on the left side and another on the right side behind the table. There are also chairs located near the table, with one chair positioned close to the left side and another chair closer to the right side of the frame.	cup: 5 spoon: 3 fork: 2 person: 2 chair: 2 bowl: 1 dining table: 1

Table 12: Quality of training data.

鉴于所有三个数据集都是由 GPT-4 生成的，我们质疑数据的质量。表明训练数据本身可能包含幻觉。附录中给出了一些训练数据示例。有趣的是，我们**发现 GPT-4 根本没有引入额外的对象存在幻觉**——问题在于提供给 GPT-4 的 MSCOCO ground-truth对象，并被**要求严格包含在生成的字幕**中。由于尺寸、分辨率和遮挡等因素，即使是人类观察者也很难接地这些 GT 对象。这导致**视觉编码器**也可能难以有效地定位这些对象。

Hallucination Analysis——Vision Encoder

1. 扩大LLM可以减轻幻觉，但是不明显
2. 扩大数据量实际上会增加幻觉，因为caption中某些对象不会被vision encoder检测
   - 通过增加图像分辨率，通过提高vision encoder的能力，可以显著减少幻觉。

分析解释：

理想情况下object是vision encoder和caption中有一一对应的关系，vision encoder获取object信息作为LLM的上下文知识，但是当vision encoder识别的对象不准确，LLM通过获取vision encoder的信息作为上下文，就会获取与实际信息不相同的参数编码。因此猜测**MLLM的幻觉主要来源于视觉模块未能正确检测object的细节**。

Table 5: Performance of LLaVA with Llama 2_{13B} language decoder and CLIP-Large vision encoder with different input resolutions.

Benchmark	Vision Encoder	$CHAIR_s\downarrow$	$CHAIR_i\downarrow$	Coverage↑	Avg. Length↑	Avg. Object↑
CCEval (Ours)	CLIP-L-112x	79.00	21.70	32.04	110.36	9.12
	CLIP-L-224x	74.00	19.30	32.83	113.03	9.18
	CLIP-L-336x	64.00	16.00	33.37	108.52	8.92

Table 6: Performance of $LLaVA_{7B}$ and with sliding window technique (SW).

Benchmark	Vision Encoder	$CHAIR_s\downarrow$	$CHAIR_i\downarrow$	Coverage↑	Avg. Length↑	Avg. Object↑
CCEval (Ours)	CLIP-L-224x	79.00	21.70	32.04	110.36	9.12
	CLIP-L-336x	79.00	18.90	36.00	111.55	9.19
	CLIP-L-224x (SW)	72.00	18.70	36.89	110.43	8.65

$$\text{CHAIR}_i = \frac{|\{\text{unmatched objects}\}|}{|\{\text{total objects mentioned}\}|}$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with unmatched objects}\}|}{|\{\text{total sentences}\}|}$$

$$\text{Coverage} = \frac{|\{\text{matched objects}\}|}{|\{\text{total ground truth objects}\}|}$$

$$\text{Average Length} = \frac{|\{\text{total words}\}|}{|\{\text{number of examples}\}|}$$

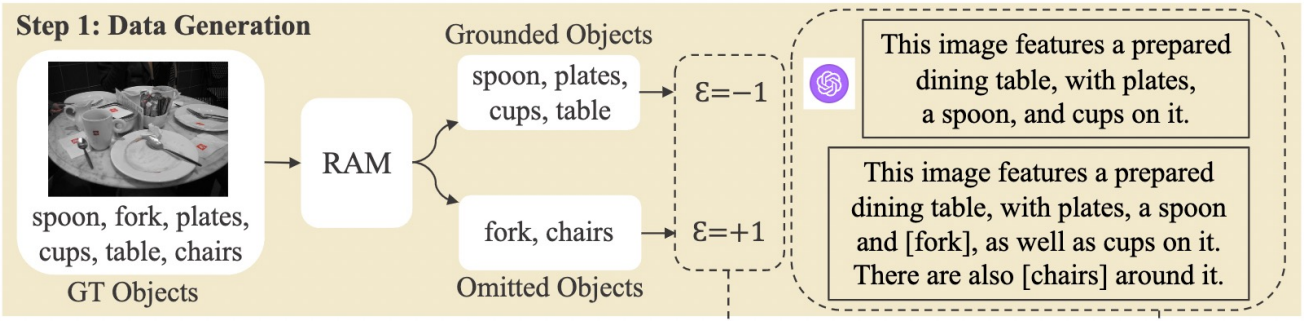
Method

Halle-Control：旨在控制详细字幕中参数知识的程度。为此，我们开发了两个数据集来训练控制器：第一个仅捕获上下文知识，而第二个同时合并上下文和参数知识。此外，我们可以通过在微调数据集中标记未识别的对象来使模型对参数知识发出信号。

1. 对于MSCOCO数据集，使用RAM检测图像里面的object，检测到的作为Grounded Objects没有检测到的作为Omitted Ojbects
2. 上下文数据生成：第一个数据集涉及仅使用来自上下文组的对象生成详细描述。为此，我们将MSCOCO的源标签（包括对象类别、边界框和简短描述）输入到GPT-4。我们遵循LLaVA的描述生成流程，并在附录中提供提示。
3. 参数联合数据生成：第二个数据集包含了上下文和参数知识。从LLaVA的原始详细描述开始，并用特殊标记注释Omitted Object。如果S表示原始图像描述句子，X={x1, ..., xn}代表一组未检测到的对象

$$S_{new} = \text{replace}(S, x_i, [x_i])$$

用括号标记参数对象的目的有两个：一是在推理过程中作为指示器，二是在训练过程中提供提示。



Method

Caption Object Extract Prompt:

User:
I have a description of an image, and I want to get objects from this description and return these objects in a list the object should be a noun, and I don't want duplicated objects. I don't want the scene name to be included, such as some caption describing the image as a scene or depicting a position or a situation or place, this thing is not an object, and doesn't need to be included. Here some objects are inside [] which we want to ignore. Here are some examples:
Example 1:
Input:
caption = "The image features a bathroom sink situated under a large mirror. The sink is accompanied by a soap dispenser, and there are multiple toothbrushes placed around it. A few cups can be seen scattered around the sink area as well. \n \n In addition to the sink, there is a toilet visible to the left side of the bathroom. The overall scene gives an impression of a well-equipped and functional bathroom space. Also a [brush] can been seen."
Answer:
objects = ['sink', 'mirror', 'soap dispenser', 'toothbrush', 'cup', 'toilet']
Here we can see [brush] is ignored because its inside []. bathroom is the place not object, so not included.

Coverage Prompt:

User:
I have two list of objects, list_A and list_B, I want to return a list named uncover which find items in list_B doesn't appear in list_A, sometimes same object can be expressed in different ways in list_A and list_B, we treat different expression but similar meaning objects as matched, not include in mismatch list.
Example 1:
Input:
list_A = ['two cars', 'dark bagpack', 'yellow jacket', 'light', 'brick building', 'wood chair', 'chair', 'green car', 'dining room table', 'bike', 'city street', 'traffic light', 'sedan'] list_B = ['reflection of light', 'view of office building', 'street chair', 'white car', 'red car', 'dark hair']
Answer:
uncover = ['reflection of light', 'dark hair']

In this example
'reflection of light' cannot find matched object in list_A, especially, 'light' is not equal to 'reflection of light'.
'view of office building' in list_B can find matched object 'brick building' although they are not exactly same but they point to similar object.
'street chair' in list_B can find 'chair', 'wood chair' in list_A which is an alternate expression of 'chair'.
'white car' in list_B can find 'two cars' in list_A.
'red car' in list_B can find 'two cars' in list_A.
'dark hair' in list_B cannot find anything similar in list_A

Hallucination Prompt:

User:
I have two lists of objects, list_A, and list_B, I want to return a list hallucination which finds items in list_B don't appear in list_A, sometimes same object can be expressed in different ways in list_A and list_B, we treat different expression but similar meaning objects as matched, not include in mismatch list.

Example 1:
Input:
list_A = ['reflection of light', 'view of office building', 'street chair', 'white car', 'red car', 'dark hair', 'bagpack', 'black shoes', 'dark pants', 'bikes', 'street', 'street light']
list_B = ['two cars', 'dark bagpack', 'yellow jacket', 'light', 'brick building', 'wood chair', 'chair', 'green car', 'dining room table', 'bike', 'city street', 'traffic light', 'sedan']
Answer:
In this example, 'two cars' is just object 'car', we don't care about the number of object. Although 'bikes' and 'bike' is not the same word, but we treat singular nouns and plural nouns as the same thing, so it's not mismatch. Here in list_A's 'street' and list_B's 'city street' are not exactly match but actually, city street can been seen as a kind of street, since city street is still a street, just in city, so they are similar meaning, we don't treat it as a mismatch, even 'city street' seems more specific, but we only still treat it as a match not hallucination. Although there is 'street light' in list_A but 'traffic light' is a different object, 'light' and 'street light' are aiming for providing lights, but 'traffic light"s purpose is providing signal, so they are different object. 'Sedan' is a different kind of car, so 'sedan' match 'car'.

hallucination = ['yellow jacket', 'dining room table', 'traffic light']

Method

Coverage Prompt:

User:
I have two list of objects, list_A and list_B, I want to return a list named uncover which find items in list_B doesn't appear in list_A, sometimes same object can be expressed in different ways in list_A and list_B, we treat different expression but similar meaning objects as matched, not include in mismatch list.
Example 1:
Input:
list_A = ['two cars', 'dark bagpack', 'yellow jacket', 'light', 'brick building', 'wood chair', 'chair', 'green car', 'dining room table', 'bike', 'city street', 'traffic light', 'sedan'] list_B = ['reflection of light', 'view of office building', 'street chair', 'white car', 'red car', 'dark hair']
Answer:
uncover = ['reflection of light', 'dark hair']

In this example
'reflection of light' cannot find matched object in list_A, especially, 'light' is not equal to 'reflection of light'.
'view of office building' in list_B can find matched object 'brick building' although they are not exactly same but they point to similar object.
'street chair' in list_B can find 'chair', 'wood chair' in list_A which is an alternate expression of 'chair'.
'white car' in list_B can find 'two cars' in list_A.
'red car' in list_B can find 'two cars' in list_A.
'dark hair' in list_B cannot find anything similar in list_A

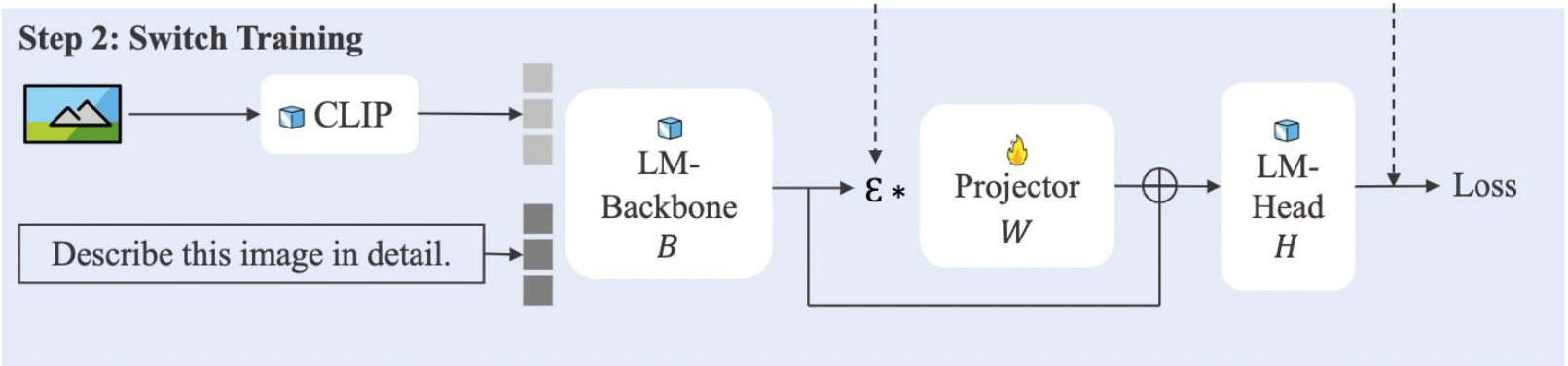
Image:

Caption with $\varepsilon = -1$:
The image depicts a large, clean, and well-organized kitchen with wooden cabinets and white countertops. The kitchen features a center island with various items placed on top of it, such as a knife, a loaf of bread, and some vegetables. There are multiple [bottles] and cups scattered around the kitchen, as well as a vase on the counter. In addition to the island, the kitchen is equipped with a refrigerator, a microwave, and two ovens, ensuring that it is well-equipped for cooking and food preparation. The presence of several vases and potted plants adds a touch of greenery and decoration to the space, making it inviting and pleasant.

Caption with $\varepsilon = 1$:
The scene is a large, well-lit kitchen with a center island. The island is covered in various food items, including [oranges], [apples], [bottles], [cups], [spoons], [knives], [bowls], [bananas], [broccoli], [potted plants], [vases], [bread], [cake], [wine glasses], [potted herbs], [potted flowers], [books], a [clock], a [potted plant] on the counter, a [vase] on the counter, a [bottle] on the counter, a [spoon] on the counter, a [knife] on the counter, a [bowl] on the counter, a [banana] on the counter, a [broccoli] on the counter, a [wine glass] on the counter, a [cup] on the counter, a [potted herb] in the background, a [potted flower] in the background, a [book] in the background, a [clock] in the background, a [potted plant] in the background, a [vase] in the background, a [bottle] in the background, a [spoon] in the background, a [knife] in the background, a [bowl] in the background, a [banana] in the background, a [broccoli] in the background, a [wine glass] in the background, a [cup] in the background. The kitchen is well-equipped with various appliances, including a refrigerator, oven, toaster, and microwave. The refrigerator, oven, and toaster are located along the wall, while the microwave is positioned above the counter. The kitchen also features a [sink] and ample counter space, making it a functional and inviting space for meal preparation and social gatherings.

Method

- 通过增加控制参数ε来解决幻觉问题，例如，+1表示允许想象，-1表示限制想象。让M代表预训练LLM：M(x) = H(ev)，其中H代表LM header，ev = B(x)是来自LM主干的word embedding。修改M′ = M(εW)，从而使word embedding：e′v = ev + εWev，使得派生模型M′(x) = H(B(x) + εW(B(x)))。学习到的Projecter W可以被视为从通用词空间到对象敏感词空间的转换，其中词与词的语义对应优化为对象对应。
- 训练：为了训练这样一个控制参数，使用生成的参数。
  - 对于只有上下文知识的数据，将ε设为-1。
  - 对于同时具有上下文和参数知识的数据，使用ε = 1。只有线性层W在整个训练阶段进行微调。
- 推理：ε可以在区间[-1, 1]内取任何值。ε值为-1表示对参数知识的依赖最小，而ε值为1表示对此类知识的强烈倾向。



Expriment

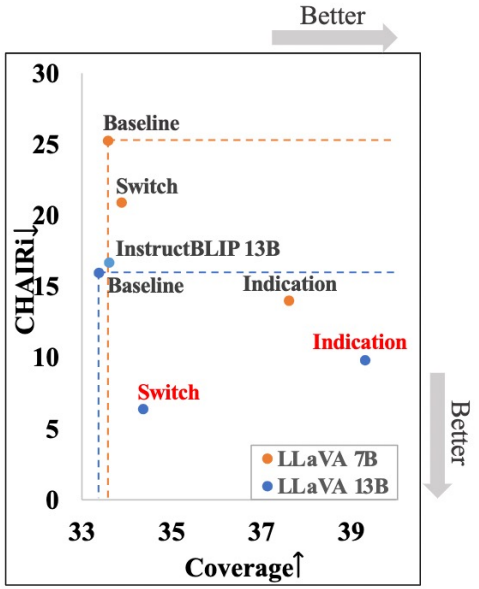
- 联合参数的数据进行微调后的模型试验结果：
  - Only ind：只在括号内的对象进行CCEval评估，较高的幻觉，比baseline还高，说明幻觉主要来自于参数知识。
  - w/o ind：忽略括号内的对象进行CCEval评估，较低幻觉，比baseline低，说明对于参数知识以外的对象，能够很准确的捕捉。
  - w/ ind：包含所有对象的CCEval评估，幻觉显著下降，对象覆盖率也不减少。
- 推理过程的试验，修改ε**的值：1->最大化参数知识。-1->纯粹上下文数据。**
  - **在**ε降低的时候，覆盖率保持基本不变的前提下，幻觉概率得到了明显的改善。

Table 7: Comparison between baselines and the effect of indication on *CCEval*. 'Only ind' means evaluation only on indicated objects; 'w/o ind' means evaluation without indicated objects; 'w/ ind' means evaluation with indicated objects.

Setting	LLM	Resolution	$CHAIR_s$↓	$CHAIR_i$↓	Coverage↑	Avg. Length↑	Avg. Object↑
158K baseline	$LLaVA_{7B}$	224x	82.00	25.30	33.58	109.89	9.31
only ind.	$LLaVA_{7B}$	224x	53.00	63.90	12.01	–	1.66
w/o ind.	$LLaVA_{7B}$	224x	57.00	17.10	37.60	108.94	7.63
w/ ind.	$LLaVA_{7B}$	224x	57.00	14.00	37.62	108.94	9.22
158K baseline	LLaVA Llama 2_{13B}	336x	64.00	16.00	33.37	108.52	8.92
only ind.	LLaVA Llama 2_{13B}	336x	52.00	62.31	19.90	–	1.3
w/o ind.	LLaVA Llama 2_{13B}	336x	52.00	11.62	34.70	106.94	7.23
w/ ind.	LLaVA Llama 2_{13B}	336x	52.00	9.86	39.31	106.94	8.52

Table 8: Performance of *HallE-Control*.

Control value	LLM	Resolution	$CHAIR_s$↓	$CHAIR_i$↓	Coverage↑	Avg. Length↑	Avg. Object↑
1	$LLaVA_{7B}$	224x	89.00	26.60	32.80	108.54	9.72
0.5	$LLaVA_{7B}$	224x	85.00	27.92	34.02	109.33	8.81
-0.5	$LLaVA_{7B}$	224x	81.00	24.88	35.87	118.08	8.04
-1	$LLaVA_{7B}$	224x	76.00	20.90	33.88	133.79	8.02
1	LLaVA Llama 2_{13B}	336x	65.00	14.58	36.14	102.18	8.37
0.5	LLaVA Llama 2_{13B}	336x	65.00	14.44	32.32	103.51	8.45
-0.5	LLaVA Llama 2_{13B}	336x	66.00	13.79	33.07	105.57	8.41
-1	LLaVA Llama 2_{13B}	336x	43.00	6.37	34.37	136.28	8.79



总结

- MLLM的幻觉与LLM自身的幻觉关系较小
- 与指令微调数据集质量关系较大
  - 数据集自身带有幻觉，出现没有出现的物体
  - 数据集中标注了过于难以观察的object，但是模型并没有观察到object，因此迫使模型训练幻觉
- Vision Encoder的能力不够强大，图像分辨率不够清晰，以至于MLLM并不能很好地获取视觉信息，vision embedding作为LLM的上下文的时候本身就带有模糊表示，以致于LLM理解为幻觉。

InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks

Zhe Chen[2,1†], Jiannan Wu[3,1†], Wenhai Wang[1,4], Weijie Su[6,1†], Guo Chen[2,1†], Sen Xing[5], Muyan Zhong[5], Qinglong Zhang[1], Xizhou Zhu[5,7,1], Lewei Lu[7,1], Bin Li[6], Ping Luo[3], Tong Lu[2], Yu Qiao[1], Jifeng Dai[5,1✉]
[1]OpenGVLab, Shanghai AI Laboratory [2]Nanjing University
[3]The University of Hong Kong [4]The Chinese University of Hong Kong [5]Tsinghua University
[6]University of Science and Technology of China [7]SenseTime Research

https://github.com/OpenGVLab/InternVL

Watch 35 | Fork 221 | Star 3k

Code

109 Commits

2 months ago

2 months ago

About

[CVPR 2024 Oral] InternVL Family: A Pioneering Open-Source Alternative to GPT-4V. 接近GPT-4V表现的可商用开源多模态对话模型

internvl.github.io/

- `2024/05/29` : 🚀 We release the Mini-InternVL-Chat series, which includes two models: Mini-InternVL-Chat-2B-V1-5 and Mini-InternVL-Chat-4B-V1-5. Our small models achieve impressive performance with minimal size: the 2B model delivers 80% of the performance with only 8% of the model size, and the 4B model achieves 90% of the performance with just 16% of the model size. For more details, please check our blog.
- `2024/05/28` : Thanks to the lmdeploy team for providing AWQ quantization support. The 4-bit model is available at OpenGVLab/InternVL-Chat-V1-5-AWQ.
- `2024/05/13` : 🔥 InternVL can now be used as the text encoder for diffusion models to support multilingual generation natively in over 110 languages worldwide. See MuLan for more details.
- `2024/04/28` : We release the INT8 version of InternVL-Chat-V1-5, see HF link.
- `2024/04/28` : We achieve the SOTA performance (75.74) on the Infographics VQA benchmark, see here.
- `2024/04/18` : InternVL-Chat-V1-5 has been released at HF link, approaching the performance of GPT-4V and Gemini Pro on various benchmarks like MMMU, DocVQA, ChartQA, MathVista, etc.
- `2024/02/27` : InternVL is accepted by CVPR 2024! 🎉
- `2024/02/24` : InternVL-Chat models have been included in the VLMEvalKit.
- `2024/02/21` : InternVL-Chat-V1-2-Plus achieves SOTA performance on MathVista (59.9), MMBench (83.8), and MMVP (58.7). See our blog for more details.
- `2024/02/12` : InternVL-Chat-V1-2 has been released. It achieves 51.6 on MMMU val and 82.3 on MMBench test. For more details, please refer to our blog, SFT data or try our demo. The model is now available on HuggingFace, and both training/evaluation data and scripts are open-sourced.
- `2024/02/04` : InternVL-Chat-V1-1 achieves 44.67% on MMVP, higher than GPT-4V!
- `2024/01/27` : We release 448 resolution model, achieving 76.6 on MMBench dev, see here.
- `2024/01/24` : InternVL-Chat-V1-1 is released, it supports Chinese and has stronger OCR capability, see here.
- `2024/01/16` : We release our customized mmcv/mmsegmentation/mmdetection code, integrated with DeepSpeed, which can be used for training large-scale object detection and semantic segmentation models.

Author list

Zhe Chen (陈喆)

PhD candidate, Nanjing University
在 smail.nju.edu.cn 的电子邮件经过验证 - 首页
Computer Vision Foundation Model

引用次数

	总计	2019 年至今
引用	1244	1244
h 指数	12	12
i10 指数	13	13

How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites — 7 — 2024
Z Chen, W Wang, H Tian, S Ye, Z Gao, E Cui, W Tong, K Hu, J Luo, Z Ma, ...
arXiv preprint arXiv:2404.16821

InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks — 20 — 2024
Z Chen, J Wu, W Wang, W Su, G Chen, S Xing, Z Muyan, Q Zhang, X Zhu, ...
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Oral

Towards Ultra-Resolution Neural Style Transfer via Thumbnail Instance Normalization — 20 — 2022
Z Chen, W Wang, E Xie, T Lu, P Luo
Proceedings of the AAAI Conference on Artificial Intelligence 36 (1), 393-400

Vision Transformer Adapter for Dense Predictions — 374 — 2022
Z Chen, Y Duan, W Wang, J He, T Lu, J Dai, Y Qiao
International Conference on Learning Representation (ICLR)

Tong Lu

Nanjing University
没有经过验证的电子邮件地址 - 首页
Computer Vision Foundation Models

引用次数 查看全部

	总计	2019 年至今
引用	12160	11258
h 指数	39	34
i10 指数	122	96

Jifeng Dai

Associate Professor of EE, Tsinghua University; Adjuct Researcher of Shanghai AI Laboratory
在 tsinghua.edu.cn 的电子邮件经过验证 - 首页
computer vision deep learning

引用次数 查看全部

	总计	2019 年至今
引用	40801	37319
h 指数	49	47
i10 指数	75	75

Background

- Vision and Vision-Language foundation Model的发展以及参数量远远不及LLM。
  - Disparity in parameter scales：最大的LLM已经达到1000B，但是广泛使用的Vision Encoder依旧处于1B左右的大小，这可能会导致约束LLM的能力
  - Inconsistent representation：纯视觉数据训练的或与BERT/自训练Language Encoder对齐训练出的模型与LLM表征不一致。
  - Inefficient connector：当前主流的Connector为MLP和Q-former，轻量或随机初始化，可能无法捕捉跨模态的理解和生成的模态间关系。
- 训练数据质量参差不齐
- 对齐视觉和语言模态的训练方式也颇具争议

Method

- Large-scale vision-language foundation model
  - Parameter-balanced vision and language component：6B的vision encoder和8B的LLM中间件

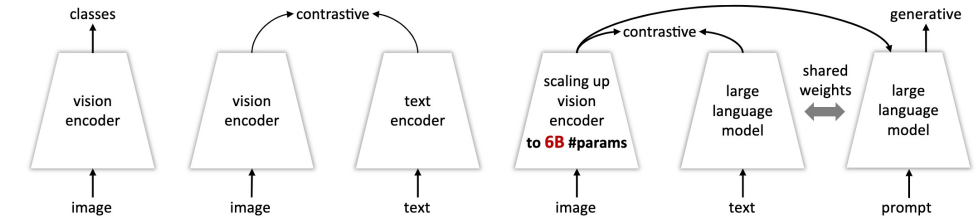


Figure 1. **Comparisons of different vision and vision-language foundation models.** (a) indicates the traditional vision foundation model, *e.g.* ResNet [57] pre-trained on classification tasks. (b) represents the vision-language foundation models, *e.g.* CLIP [117] pre-trained on image-text pairs. (c) is our InternVL, which presents a workable way to align the large-scale vision foundation model (*i.e.*, InternViT-6B) with the large language model and is versatile for both contrastive and generative tasks.

  - 使用预训练多语言LLaMA来初始化LLM中间件实现表征一致性
  - Progressive image-text alignment：先大规模噪声文本-图像对对比学习，细粒度数据的生成学习

Method

InternViT-6B：使用vanilla ViT，扩展到6B，扩展权衡因素（精度、速度、稳定性），使用超参数搜索：model depth{32, 48, 64, 80}，head dimension{64, 128}，MLP ratio{4, 8}

name	width	depth	MLP	#heads	#param (M)
ViT-G [173]	1664	48	8192	16	1843
ViT-e [23]	1792	56	15360	16	3926
EVA-02-ViT-E [130]	1792	64	15360	16	4400
ViT-6.5B [128]	4096	32	16384	32	6440
ViT-22B [37]	6144	48	24576	48	21743
InternViT-6B (ours)	3200	48	12800	25	5903

QLLaMA：使用预训练多语言LLaMA（7B）初始化，随机初始化96个可学习Query和cross attention layer（共1B）。

优势：

1. LLM预训练参数初始化，是的InternViT的表征能够与LLM对齐。
2. 即使冻结LLM也可以取得很好的性能。
3. 可以应用于对比学习，因为视觉表征自身已经很好地和大语言模型对齐。

Method

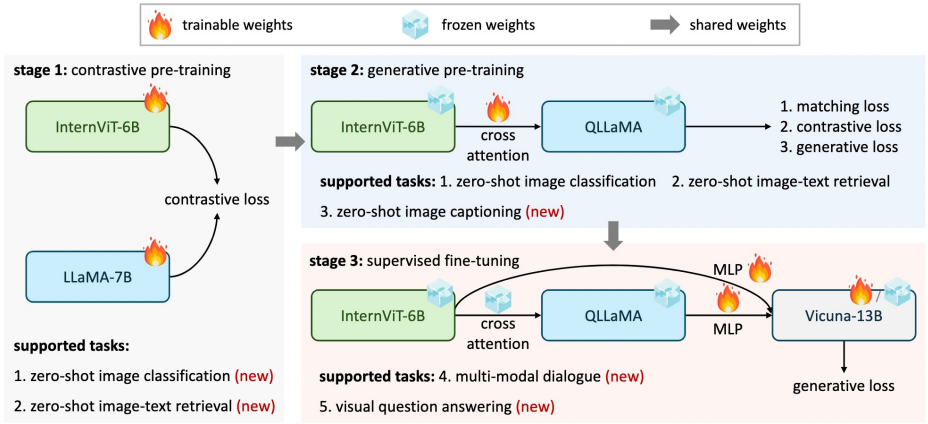


Figure 3. **The training strategy of the proposed InternVL model.** It consists of three progressive stages, including vision-language contrastive training, vision-language generative training, and supervised fine-tuning. These stages effectively leverage public data from diverse sources, ranging from noisy image-text pairs on the web to high-quality caption, VQA, and multi-modal dialogue datasets.

dataset	characteristics		stage 1		stage 2	
	language	original	cleaned	remain	cleaned	remain
LAION-en [120]	English	2.3B	1.94B	84.3%	91M	4.0%
LAION-COCO [121]		663M	550M	83.0%	550M	83.0%
COYO [14]		747M	535M	71.6%	200M	26.8%
CC12M [20]		12.4M	11.1M	89.5%	11.1M	89.5%
CC3M [124]		3.0M	2.6M	86.7%	2.6M	86.7%
SBU [112]		1.0M	1.0M	100%	1.0M	100%
Wukong [55]	Chinese	100M	69.4M	69.4%	69.4M	69.4%
LAION-multi [120]	Multi	2.2B	1.87B	85.0%	100M	4.5%
Total	Multi	6.03B	4.98B	82.6%	1.03B	17.0%

Table 2. **Details of the training data for InternVL in stage 1 and stage 2.** Among them, LAION-en [120], LAION-multi [120], COYO [14], and Wukong [55] are web-scale image-text pairs data. LAION-COCO [121] is a synthetic dataset with high-quality captions from LAION-en. CC12M [20], CC3M [124], SBU [112] are academic caption datasets. "Multi" means multilingual.

task	#samples	dataset
Captioning	588K	COCO Caption [22], TextCaps [126]
VQA	1.1M	VQAv2 [54], OKVQA [104], A-OKVQA [122], IconQA [99], AI2D [71], GQA [64]
OCR	294K	OCR-VQA [107], ChartQA [105], DocVQA [29], ST-VQA [12], EST-VQA [150], InfoVQA [106], LLaVAR [182]
Grounding	323K	RefCOCO/+/g [103, 170], Toloka [140]
Grounded Cap.	284K	RefCOCO/+/g [103, 170]
Conversation	1.4M	LLaVA-150K [92], SVIT [183], VisDial [36], LRV-Instruction [90], LLaVA-Mix-665K [91]

Table 3. **Details of the training data for InternVL in stage 3.** We collect a wide range of high-quality instruction data, totaling approximately 4 million samples. For a fair comparison, we only use the training split of these datasets.

Stage1对比学习：清洗数据用大规模的噪声文本-图像对用CLIP的方式进行训练，ViT -> Vision Embedding， LLaMA[EOS] -> Language Embedding。
目标：能够实现zero-shot的分类和检索任务。
Stage2生成任务预训练：使用stage1的参数，冻结ViT和LLM，训练query和cross attention，过滤出较高质量的Caption数据及部分文本-图像对，损失函数参考BLIP2: ITC+ITM+ITG损失。
目标：提取强大的视觉表示，进一步实现特征空间对齐。
Stage3Fine-tuning：使用MLP与先有的常用LLM连接，使用高质量指令数据集进行微调，由于之前已经对齐到了LLM的特征，因此即使LLM时冻结的也可以实现比较好的效果。

Method

config	stage 1	stage 2
image enc. weight init.	random init. [7]	from stage 1
text enc. weight init.	from [32]	from stage 1
image enc. peak learning rate	1e-3	frozen
text enc. peak learning rate	1e-4	frozen
cross attn peak learning rate	–	5e-5
learning rate schedule	cosine decay	cosine decay
optimizer	AdamW [98]	AdamW [98]
optimizer hyper-parameters	$\beta_1, \beta_2 = 0.9, 0.95$	$\beta_1, \beta_2 = 0.9, 0.98$
weight decay	0.1	0.05
input resolution	$196^2 \rightarrow 224^2$	224^2
patch size	14	14
total batch size	164K	20K
warm-up iterations	5K	2K
total iterations	175K	80K
samples seen	28.7B	1.6B
drop path rate [63]	uniform (0.2)	0.0
data augmentation	random resized crop	random resized crop
numerical precision	DeepSpeed bf16 [118]	DeepSpeed bf16 [118]
trainable / total parameters	13B / 13B	1B / 14B
GPUs for training	640×A100 (80G)	160×A100 (80G)

Table 20. **Training settings of InternVL's stage 1 and stage 2.** "$196^2 \rightarrow 224^2$" means we initially train at a 196×196 resolution, and later switch to 224×224 resolution for the final 0.5 billion samples, for higher training efficiency.

config	retrieval fine-tuning
image-text data	Flickr30K [116] / Flickr30K-CN [77]
peak learning rate	1e-6
layer-wise lr decay rate	InternViT-6B (0.9), QLLaMA (0.9)
learning rate schedule	cosine decay
optimizer	AdamW [98]
optimizer hyper-parameters	$\beta_1, \beta_2 = 0.9, 0.999$
weight decay	0.05
input resolution	364^2
patch size	14
total batch size	1024
warm-up iterations	100
training epochs	10
drop path rate [63]	0.3
data augmentation	random resized crop & flip
numerical precision	DeepSpeed bf16 [118]
trainable / total parameters	14B / 14B
GPUs for training	32×A100 (80G)

Table 21. **Training settings of retrieval fine-tuning.** We fine-tune InternVL on Flickr30K and Flickr30K-CN separately.

config	ImageNet linear probing
peak learning rate	0.2
learning rate schedule	cosine decay
optimizer	SGD
optimizer momentum	0.9
weight decay	0.0
input resolution	224^2
patch size	14
total batch size	1024
warm-up epochs	1
training epochs	10
data augmentation	random resized crop & flip
GPUs for training	8×A100 (80G)

Table 22. **Training settings of ImageNet linear probing.**

config	linear probing / head tuning / full tuning
peak learning rate	4e-5
layer-wise lr decay rate	– / – / 0.95
learning rate schedule	polynomial decay
optimizer	AdamW [98]
optimizer hyper-parameters	$\beta_1, \beta_2 = 0.9, 0.999$
weight decay	0.0 / 0.05 / 0.05
input resolution	504^2
patch size	14
total batch size	16
warm-up iterations	1.5K
total iterations	80K
drop path rate [63]	0.0 / 0.0 / 0.4
data augmentation	default augmentation in MMSeg [31]
numerical precision	DeepSpeed bf16 [118]
GPUs for training	8×A100 (80G)

Table 23. **Training settings of ADE20K semantic segmentation.** We list the hyperparameters for three different configurations, including linear probing, head tuning, and full-parameter tuning.

Experiment

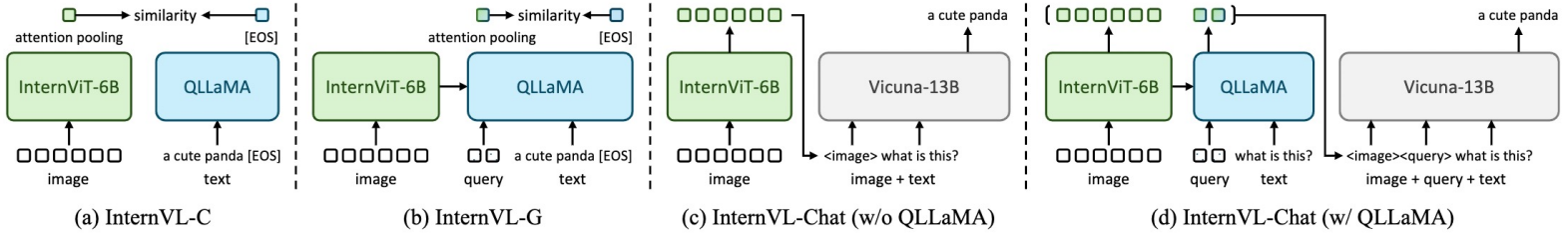


Figure 4. **Different ways to use InternVL.** By flexibly combining the vision encoder and the language middleware, InternVL can support various vision-language tasks, including contrastive tasks, generative tasks, and multi-modal dialogue.

视觉感知任务：使用ViT的特征进行全连接或池化进行分类

对比任务：使用(a)/(b)，将ViT的token和[EOS]的特征进行余弦相似度比较。

生成任务：使用ViT/QLLaMA的输出作为LLM的前缀token，生成后续token。

多模态对话：InternVL作为Vision component，既可以直接使用ViT，也可以使用QLLaMA。

method	IN-1K	IN-A	IN-R	IN-V2	IN-Sketch	ObjectNet	Δ↓	avg.
OpenCLIP-H [67]	78.0	59.3	89.3	70.9	66.6	69.7	5.7	72.3
OpenCLIP-g [67]	78.5	60.8	90.2	71.7	67.5	69.2	5.5	73.0
OpenAI CLIP-L+ [117]	76.6	77.5	89.0	70.9	61.0	72.0	2.1	74.5
EVA-01-CLIP-g [130]	78.5	73.6	92.5	71.5	67.3	72.3	2.5	76.0
OpenCLIP-G [67]	80.1	69.3	92.1	73.6	68.9	73.0	3.9	76.2
EVA-01-CLIP-g+ [130]	79.3	74.1	92.5	72.1	68.1	75.3	2.4	76.9
MAWS-ViT-2B [128]	81.9	–	–	–	–	–	–	–
EVA-02-CLIP-E+ [130]	82.0	82.1	94.5	75.7	71.6	79.6	1.1	80.9
CoCa* [169]	86.3	90.2	96.5	80.7	77.6	82.7	0.6	85.7
LiT-22B* [37, 174]	85.9	90.1	96.0	80.9	–	87.6	–	–
InternVL-C (ours)	**83.2**	**83.8**	**95.5**	**77.3**	**73.9**	**80.6**	**0.8**	**82.4**

(a) ImageNet variants [38, 60, 61, 119, 141] and ObjectNet [8].

method	EN	ZH	JP	AR	IT	avg.
M-CLIP [16]	–	–	–	–	20.2	–
CLIP-Italian [11]	–	–	–	–	22.1	–
Japanese-CLIP-ViT-B [102]	–	–	54.6	–	–	–
Taiyi-CLIP-ViT-H [176]	–	54.4	–	–	–	–
WuKong-ViT-L-G [55]	–	57.5	–	–	–	–
CN-CLIP-ViT-H [162]	–	59.6	–	–	–	–
AltCLIP-ViT-L [26]	74.5	59.6	–	–	–	–
EVA-02-CLIP-E+ [130]	82.0	3.6	5.0	0.2	41.2	–
OpenCLIP-XLM-R-B [67]	62.3	42.7	37.9	26.5	43.7	42.6
OpenCLIP-XLM-R-H [67]	77.0	55.7	53.1	37.0	56.8	55.9
InternVL-C (ours)	**83.2**	**64.5**	**61.5**	**44.9**	**65.7**	**64.0**

(b) Multilingual ImageNet-1K [38, 76].

Table 6. **Comparison of zero-shot image classification performance.** "Δ↓": The gap between the averaged top-1 accuracy and the IN-1K top-1 accuracy. *CoCa [169] and LiT-22B [37] use the private JFT-3B dataset [173] during training. Multilingual evaluation involves 5 languages, including English (EN), Chinese (ZH), Japanese (JP), Arabic (AR), and Italian (IT).

method	#F	K400 [17]		K600 [18]		K700 [19]	
		top-1	avg.	top-1	avg.	top-1	avg.
OpenCLIP-g [67]	1	–	63.9	–	64.1	–	56.9
OpenCLIP-G [67]	1	–	65.9	–	66.1	–	59.2
EVA-01-CLIP-g+ [130]	1	–	66.7	–	67.0	–	60.9
EVA-02-CLIP-E+ [130]	1	–	69.8	–	69.3	–	63.4
InternVL-C (ours)	1	**65.9**	**76.1**	**65.5**	**75.5**	**56.8**	**67.5**
ViCLIP [152]	8	64.8	75.7	62.2	73.5	54.3	66.4
InternVL-C (ours)	8	**69.1**	**79.4**	**68.9**	**78.8**	**60.6**	**71.5**

Table 8. **Comparison of zero-shot video classification results on Kinetics 400/600/700.** We report the top-1 accuracy and the mean of top-1 and top-5 accuracy. "#F" denotes the number of frames.

Experiment

method	multi-lingual	Flickr30K (English, 1K test set) [116]						COCO (English, 5K test set) [22]						avg.
		Image → Text			Text → Image			Image → Text			Text → Image			
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
Florence [171]	×	90.9	99.1	–	76.7	93.6	–	64.7	85.9	–	47.2	71.4	–	–
ONE-PEACE [143]	×	90.9	98.8	99.8	77.2	93.5	96.2	64.7	86.0	91.9	48.0	71.5	79.6	83.2
OpenCLIP-H [67]	×	90.8	99.3	99.7	77.8	94.1	96.6	66.0	86.1	91.9	49.5	73.4	81.5	83.9
OpenCLIP-g [67]	×	91.4	99.2	99.6	77.7	94.1	96.9	66.4	86.0	91.8	48.8	73.3	81.5	83.9
OpenCLIP-XLM-R-H [67]	✓	91.8	99.4	99.8	77.8	94.1	96.5	65.9	86.2	92.2	49.3	73.2	81.5	84.0
EVA-01-CLIP-g+ [130]	×	91.6	99.3	99.8	78.9	94.5	96.9	68.2	87.5	92.5	50.3	74.0	82.1	84.6
CoCa [169]	×	92.5	99.5	99.9	80.4	95.7	97.7	66.3	86.2	91.8	51.2	74.2	82.0	84.8
OpenCLIP-G [67]	×	92.9	99.3	99.8	79.5	95.0	97.1	67.3	86.9	92.6	51.4	74.9	83.0	85.0
EVA-02-CLIP-E+ [130]	×	93.9	99.4	99.8	78.8	94.2	96.8	68.8	87.8	92.8	51.1	75.0	82.7	85.1
BLIP-2† [81]	×	97.6	100.0	100.0	89.7	98.1	98.9	–	–	–	–	–	–	–
InternVL-C (ours)	✓	94.7	99.6	99.9	81.7	96.0	98.2	70.6	89.0	93.5	54.1	77.3	84.6	86.6
InternVL-G (ours)	✓	**95.7**	**99.7**	**99.9**	**85.0**	**97.0**	**98.6**	**74.9**	**91.3**	**95.2**	**58.6**	**81.3**	**88.0**	**88.8**

method		Flickr30K-CN (Chinese, 1K test set) [77]						COCO-CN (Chinese, 1K test set) [84]						avg.
WuKong-ViT-L [55]	×	76.1	94.8	97.5	51.7	78.9	86.3	55.2	81.0	90.6	53.4	80.2	90.1	78.0
R2D2-ViT-L [159]	×	77.6	96.7	98.9	60.9	86.8	92.7	63.3	89.3	95.7	56.4	85.0	93.1	83.0
Taiyi-CLIP-ViT-H [176]	×	–	–	–	–	–	–	–	–	–	60.0	84.0	93.3	–
AltCLIP-ViT-H [26]	✓	88.9	98.5	99.5	74.5	92.0	95.5	–	–	–	–	–	–	–
CN-CLIP-ViT-H [162]	×	81.6	97.5	98.8	71.2	91.4	95.5	63.0	86.6	92.9	69.2	89.9	96.1	86.1
OpenCLIP-XLM-R-H [67]	✓	86.1	97.5	99.2	71.0	90.5	94.9	70.0	91.5	97.0	66.1	90.8	96.0	87.6
InternVL-C (ours)	✓	90.3	98.8	99.7	75.1	92.9	96.4	68.8	92.0	96.7	68.9	91.9	96.5	89.0
InternVL-G (ours)	✓	**92.9**	**99.4**	**99.8**	**77.7**	**94.8**	**97.3**	**71.4**	**93.9**	**97.7**	**73.8**	**94.4**	**98.1**	**90.9**

Table 7. **Comparison of zero-shot image-text retrieval performance.** We evaluate the retrieval capability in English using the Flickr30K [116] and COCO [22], as well as in Chinese using Flickr30K-CN [77] and COCO-CN [84]. †BLIP-2 [81] is finetuned on COCO and zero-shot transferred to Flickr30K, contributing to the enhanced zero-shot performance on Flickr30K.

method	glue layer	LLM decoder	COCO	Flickr30K	NoCaps
Flamingo-9B [3]	Cross-Attn	Chinchilla-7B	79.4	61.5	–
Flamingo-80B [3]	Cross-Attn	Chinchilla-70B	84.3	67.2	–
KOSMOS-2 [115]	Linear	KOSMOS-1	–	66.7	–
PaLI-X-55B [24]	Linear	UL2-32B	–	–	**126.3**
BLIP-2 [81]	QFormer	Vicuna-13B	–	71.6	103.9
InstructBLIP [34]	QFormer	Vicuna-13B	–	82.8	121.9
Shikra-13B [21]	Linear	Vicuna-13B	–	73.9	–
ASM [149]	QFormer	Husky-7B	–	**87.7**	117.2
Qwen-VL [5]	VL-Adapter	Qwen-7B	–	85.8	121.4
Qwen-VL-Chat [5]	VL-Adapter	Qwen-7B	–	81.0	120.2
Emu [131]	QFormer	LLaMA-13B	112.4	–	–
Emu-I [131]	QFormer	LLaMA-13B	117.7	–	–
DreamLLM [41]	Linear	Vicuna-7B	115.4	–	–
InternVL-G (ours)	Cross-Attn	QLLaMA	**128.2**	79.2	113.7

Table 10. **Comparison of zero-shot image captioning.** QLLaMA inherently possesses promising zero-shot captioning capabilities thanks to its scaled-up parameters and datasets.

method	visual encoder	glue layer	LLM	Res.	PT	SFT	train. param	image captioning			visual question answering				dialogue	
								COCO	Flickr	NoCaps	VQA^{v2}	GQA	VizWiz	VQA^{T}	MME	POPE
InstructBLIP [34]	EVA-g	QFormer	Vicuna-7B	224	129M	1.2M	188M	–	82.4	123.1	–	49.2	34.5	50.1	–	–
BLIP-2 [81]	EVA-g	QFormer	Vicuna-13B	224	129M	–	188M	–	71.6	103.9	41.0	41.0	19.6	42.5	1293.8	85.3
InstructBLIP [34]	EVA-g	QFormer	Vicuna-13B	224	129M	1.2M	188M	–	82.8	121.9	–	49.5	33.4	50.7	1212.8	78.9
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-7B	224	1.0B	4.0M	64M	141.4*	89.7	120.5	72.3*	57.7*	44.5	42.1	1298.5	85.2
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-13B	224	1.0B	4.0M	90M	142.4*	89.9	123.1	71.7*	59.5*	54.0	49.1	1317.2	85.4
Shikra [21]	CLIP-L	Linear	Vicuna-13B	224	600K	5.5M	7B	117.5*	73.9	–	77.4*	–	–	–	–	–
IDEFICS-80B [66]	CLIP-H	Cross-Attn	LLaMA-65B	224	1.6B	–	15B	91.8*	53.7	65.0	60.0	45.2	36.0	30.9	–	–
IDEFICS-80B-I [66]	CLIP-H	Cross-Attn	LLaMA-65B	224	353M	6.7M	15B	117.2*	65.3	104.5	37.4	–	26.0	–	–	–
Qwen-VL [5]	CLIP-G	VL-Adapter	Qwen-7B	448	1.4B†	50M†	9.6B	–	85.8	121.4	78.8*	59.3*	35.2	63.8	–	–
Qwen-VL-Chat [5]	CLIP-G	VL-Adapter	Qwen-7B	448	1.4B†	50M†	9.6B	–	81.0	120.2	78.2*	57.5*	38.9	**61.5**	1487.5	–
LLaVA-1.5 [91]	$CLIP\text{-}L_{336}$	MLP	Vicuna-7B	336	558K	665K	7B	–	–	–	78.5*	62.0*	50.0	58.2	1510.7	85.9
LLaVA-1.5 [91]	$CLIP\text{-}L_{336}$	MLP	Vicuna-13B	336	558K	665K	13B	–	–	–	80.0*	63.3*	53.6	61.3	1531.3	85.9
InternVL-Chat (ours)	IViT-6B	MLP	Vicuna-7B	336	558K	665K	7B	–	–	–	79.3*	62.9*	52.5	57.0	1525.1	86.4
InternVL-Chat (ours)	IViT-6B	MLP	Vicuna-13B	336	558K	665K	13B	–	–	–	80.2*	63.9*	54.6	58.7	1546.9	87.1
InternVL-Chat (ours)	IViT-6B	QLLaMA	Vicuna-13B	336	1.0B	4.0M	13B	**146.2***	**92.2**	**126.2**	**81.2***	**66.6***	**58.5**	**61.5**	**1586.4**	**87.6**

Table 9. **Comparison with SoTA methods on 9 benchmarks.** Image captioning datasets include: COCO Karpathy test [22], Flickr30K Karpathy test [116], NoCaps val [2]. VQA datasets include: VQAv2 test-dev [54], GQA test-balanced [64], VizWiz test-dev [56], and TextVQA val [127]. *The training annotations of the datasets are observed during training. "IViT-6B" represents our InternViT-6B.

Ablation Experiment

name	width	depth	MLP	#heads	#param	FLOPs	throughput	zs IN
variant 1	3968	32	15872	62	6051M	1571G	35.5 / 66.0	65.8
variant 2	3200	48	12800	50	5903M	1536G	28.1 / 64.9	66.1
variant 3	3200	48	12800	25	5903M	1536G	28.0 / 64.6	66.2
variant 4	2496	48	19968	39	5985M	1553G	28.3 / 65.3	65.9
variant 5	2816	64	11264	44	6095M	1589G	21.6 / 61.4	66.2
variant 6	2496	80	9984	39	5985M	1564G	16.9 / 60.1	66.2

Table 11. **Comparison of hyperparameters in InternViT-6B.** The throughput (img/s) and GFLOPs are measured at 224×224 input resolution, with a batch size of 1 or 128 on a single A100 GPU. Flash Attention [35] and bf16 precision are used during testing. "zs IN" denotes the zero-shot top-1 accuracy on the ImageNet-1K validation set [38]. The final selected model is marked in gray.

visual encoder	glue layer	LLM	visual question answering				dialogue	
			VQA^{v2}	GQA	VizWiz	VQA^{T}	MME	POPE
IViT-6B	MLP	Vicuna-7B	79.3	62.9	52.5	57.0	1525.1	86.4
IViT-6B	MLP	InternLM-7B	79.7	63.2	53.1	58.0	1532.8	86.4

Table 18. **Compatibility with other LLM.** Here we use InternLM [135] as an example to verify the compatibility of InternVL with LLMs other than Vicuna [184]. The experimental settings used here are the same as in Table 9 of the main paper.

visual encoder	glue layer	LLM	dataset	dialogue	caption	visual question answering		
				MME	NoCaps	OKVQA	$VizWiz_{val}$	GQA
EVA-E	MLP	V-7B	665K [91]	970.5	75.1	40.1	25.5	41.3
IViT-6B	MLP	V-7B	665K [91]	1022.3	80.8	42.9	28.3	45.8
IViT-6B	QLLaMA	V-7B	665K [91]	1227.5	94.5	51.0	38.4	57.4
IViT-6B	QLLaMA	V-7B	Ours	1298.5	120.5	51.8	44.9	57.7
IViT-6B	QLLaMA	V-13B	Ours	1317.2	123.1	55.5	55.7	59.5

Table 12. **Ablation studies of using InternVL to build multi-modal dialogue system.** V-7B and V-13B denote Vicuna-7B/13B [184], respectively. "IViT-6B" represents our InternViT-6B.

method	image size	encode image (ms)		encode text (ms)	total time	FPS
		InternViT-6B	QLLaMA	QLLaMA		
InternVL-C	224	15.5	–	4.9	20.4	48.9
InternVL-C	336	35.2	–	4.9	40.1	24.9
InternVL-C	448	66.9	–	4.9	71.8	13.9
InternVL-G	224	15.5	8.2	4.9	28.6	35.0
InternVL-G	336	35.2	10.3	4.9	50.4	19.8
InternVL-G	448	66.9	12.8	4.9	84.6	11.8

Table 19. **Efficiency analysis of InternVL for encoding image-text pairs.** The total time to encode an image-text pair includes both the image encoding part and the text encoding part. We measure the time cost with a batch size of 128 on a single A100 GPU. Flash Attention [35] and bf16 precision are used during testing.

总结

- 多模态大语言模型在视觉和语言特征融合方面还有很大的提升空间
- 视觉特征本身的升级，vision foundation model可能算力很难支持，但是用更好的视觉模型可能在刷点上是一个比较可行的方案
- 视觉特征向文本特征对齐的阶段，在SFT等操作的时候不可避免地很难让视觉特征很好地对齐文本特征，以及对LLM自身能力的灾难性遗忘。一个自身就比较对齐LLM特征空间的vision encoder或许是一个比较有效的解决方案。