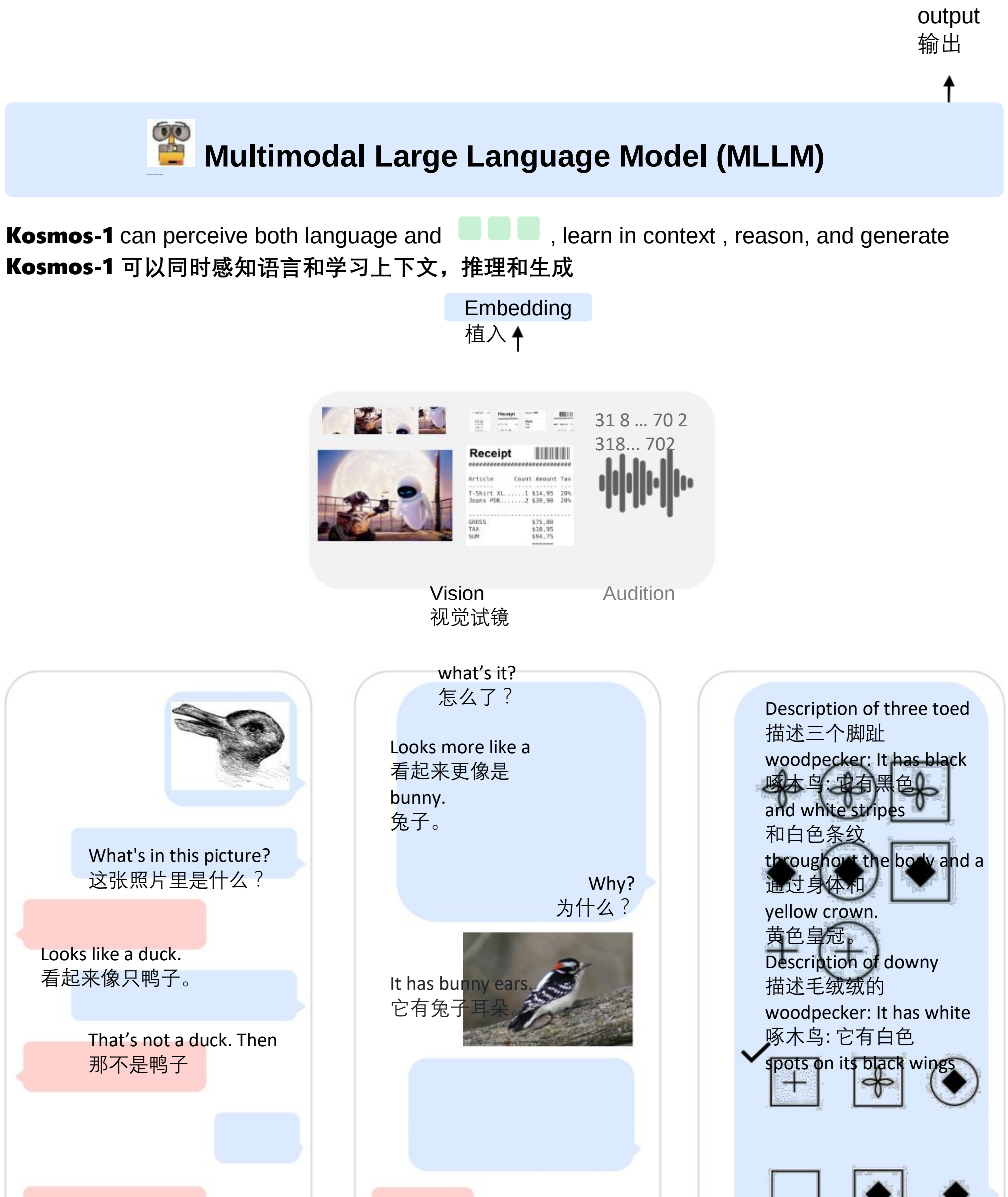
Language Is Not All You Need: Aligning Perception
语言不是你所需要的全部: 调整感知
with Language Models
语言模型

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, Kriti Aggarwal, Zewen Chi Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, Furu Wei[y]
黄绍汉，李东，王文辉，郝亚如，萨克森辛格尔，马腾朝吕淑明，雷翠，奥瓦斯可汗穆罕默德，刘强，克里蒂 · 阿加瓦尔，泽文奇约翰 · 比约克，维什拉夫 · 乔杜里，苏霍吉特森，夏松，弗鲁威利
Microsoft
微软
https://github.com/microsoft/unilm
Https://github. com/microsoft/unilm



Kosmos-1 can perceive both language and 🟩🟩🟩 , learn in context , reason, and generate
Kosmos-1 可以同时感知语言和学习上下文，推理和生成

What's in this picture?
这张照片里是什么？

Looks like a duck.
看起来像只鸭子。

That's not a duck. Then
那不是鸭子

what's it?
怎么了？

Looks more like a
看起来更像是
bunny.
兔子。

Why?
为什么？

It has bunny ears.
它有兔子耳朵

Description of three toed
描述三个脚趾
woodpecker: It has black
啄木鸟: 它有黑色
and white stripes
和白色条纹
throughout the body and a
通过身体和
yellow crown.
黄色皇冠。
Description of downy
描述毛绒绒的
woodpecker: It has white
啄木鸟: 它有白色
spots on its black wings

黑色翅膀上的斑点
and some red on its
crown.
在它的皇冠上有些红
色。

Here are eight images:
下面是 8 张图片:

Question: what is the
问题: 什么是
name of the
的名称
woodpecker in the
啄木鸟
picture?
照片？

The
followin
g image
is:
下图是:

B *C*
B c

Downy
唐尼

D *E* *F*
好了，好了

Figure 1: KOSMOS-1 is a multimodal large language model (MLLM) that is capable of perceiving multimodal input, following instructions, and performing in-context learning for not only language tasks but also multimodal tasks. In this work, we align vision with large language models (LLMs), advancing the trend of going from LLMs to MLLMs.
图 1: kosmos-1 是一个多模态大语言模型(MLLM)，它能够感知多模态输入，遵循指令，并且不仅对语言任务而且对多模态任务执行上下文学习。在这项工作中，我们将视觉与大语言模型(LLMs)结合起来，推进了从 LLMs 到 MLLMs 的发展趋势。

Equal contribution. y Corresponding author.
同等贡献通讯作者。

“ ”

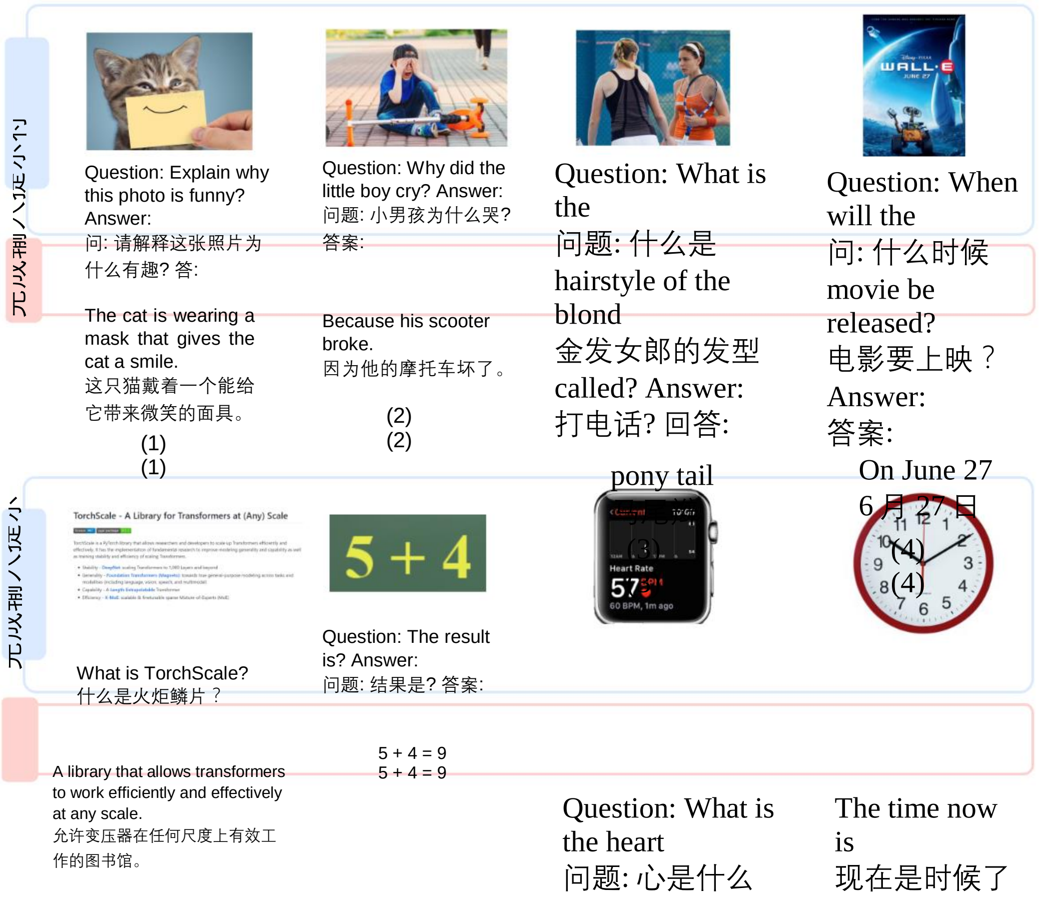
The limits of my language means the limits of my world. Ludwig Wittgenstein

路德维希 · 维特根斯坦(ludwig wittgenstein)

Abstract
摘要

A big convergence of language, multimodal perception, action, and world modeling is a key step toward artificial general intelligence. In this work, we introduce KOSMOS-1[2], a Multimodal Large Language Model (MLLM) that can perceive general modalities, learn in context (i.e., few-shot), and follow instructions (i.e., zero-shot). Specifically, we train KOSMOS-1 from scratch on web-scale multimodal corpora, including arbitrarily interleaved text and images, image-caption pairs, and text data. We evaluate various settings, including zero-shot, few-shot, and multimodal chain-of-thought prompting, on a wide range of tasks without any gradient updates or finetuning. Experimental results show that KOSMOS-1 achieves impressive performance on (i) language understanding, generation, and even OCR-free NLP (directly fed with document images), (ii) perception-language tasks, including multimodal dialogue, image captioning, visual question answering, and (iii) vision tasks, such as image recognition with descriptions (specifying classification via text instructions). We also show that MLLMs can benefit from cross-modal transfer, i.e., transfer knowledge from language to multimodal, and from multimodal to language. In addition, we introduce a dataset of Raven IQ test, which diagnoses the nonverbal reasoning capability of MLLMs.

语言、多模态知觉、动作和世界模型的大融合是迈向人工通用智能的关键一步。在这项工作中，我们介绍了 KOSMOS-12，一个多模态大语言模型(MLLM)，它可以感知一般模式，在上下文中学习(即，少镜头)，并遵循指示(即，零镜头)。具体来说，我们从头开始训练 kosmos-1 在 web 规模的多模态语料库，包括任意交织的文本和图像，图像-标题对，和文本数据。我们评估各种设置，包括零镜头，少镜头，多模式的思想链提示，在广泛的任务范围内，没有任何梯度更新或微调。实验结果表明，kosmos-1 在(i)语言理解、生成、甚至无 ocr 的 NLP (直接提供文档图像)、(ii)知觉语言任务(包括多模态对话、图像标题、视觉问题回答和(iii)视觉任务(包括带描述的图像识别(通过文本指令指定分类)等方面取得了令人瞩目的成绩。我们还表明，MLLMs 可以从跨模态转移中受益，即，将知识从语言转移到多模态，从多模态转移到语言。此外，我们还介绍了 Raven IQ 测试的数据集，它可以诊断 MLLMs 的非语言推理能力。



Question: Explain why this photo is funny? Answer:
问: 请解释这张照片为什么有趣? 答:

The cat is wearing a mask that gives the cat a smile.
这只猫戴着一个能给它带来微笑的面具。

(1)
(1)

Question: Why did the little boy cry? Answer:
问题: 小男孩为什么哭? 答案:

Because his scooter broke.
因为他的摩托车坏了。

(2)
(2)

Question: What is the
问题: 什么是
hairstyle of the blond
金发女郎的发型
called? Answer:
打电话? 回答:

pony tail
马尾辫

(3)
(3)

Question: When will the
问: 什么时候
movie be released?
电影要上映？
Answer:
答案:

On June 27
6 月 27 日

(4)
(4)

TorchScale - A Library for Transformers at (Any) Scale

TorchScale is a PyTorch library that allows researchers and developers to scale up Transformers efficiently and effectively. It has the implementation of fundamental research to improve modeling generality and capability as well as training stability and efficiency of scaling Transformers.

- Stability - DeepNet: scaling Transformers to 1,000 Layers and beyond
- Generality - Foundation Transformers (Magneto): towards true general-purpose modeling across tasks and modalities (including language, vision, speech, and multimodal)
- Capability - A Length-Extrapolatable Transformer
- Efficiency - X-MoE: scalable & finetunable sparse Mixture-of-Experts (MoE)

What is TorchScale?
什么是火炬鳞片？

A library that allows transformers to work efficiently and effectively at any scale.
允许变压器在任何尺度上有效工作的图书馆。

Question: The result is? Answer:
问题: 结果是? 答案:

5 + 4 = 9
5 + 4 = 9

Question: What is the heart
问题: 心是什么

The time now is
现在是时候了

rate in this picture?率?答案:

Answer:

在这张图片中的比

57 bpm

每分钟 57 次

10:10 on a large clock

大钟上的 10 点 10 分

(5) (6) (7) (8)

(5)(6)(7)(8)

Figure 2: Selected examples generated from KOSMOS-1. Blue boxes are input prompt and pink boxes are KOSMOS-1 output. The examples include (1)-(2) visual explanation, (3)-(4) visual question answering, (5) web page question answering, (6) simple math equation, and (7)-(8) number recognition.

图 2: kosmos-1 生成的选定示例。蓝色框是输入提示符，粉色框是 kosmos-1 输出。例如: (1)-(2)视觉解释，(3)-(4)视觉问答，(5)网页问答，(6)简单的数学方程，(7)-(8)数字识别。

[2] KOSMOS is pronounced as and means “Cosmos”.
KOSMOS 发音为“ Cosmos”，意思是“ Cosmos”。

An image of
图片来源:

a girl blowing out a candle on her birthday cake.
一个女孩吹灭了她生日蛋糕上的蜡烛。

(1)
(1)

An image of
图片来源:

a group of people posing for a wedding photo.
一群人摆姿势拍结婚照。

(2)
(2)

Question: What coffee
问题: 什么咖啡
shop is in the background?
商店是在后台？
Answer:
答案:

Starbucks
星巴克

(3)
(3)

Question: What is inside this canned good? Answer:
问: 罐头里面是什么? 答:

Corn
玉米

(4)
(4)

Question: What is on the pillowcase? Answer:
问题: 枕套上是什么? 回答:

a Texas Longhorns pillow case.
一个德克萨斯长角牛的枕套。

(5)
(5)

Question: What is the title of this book? Answer:
问: 这本书的书名是什么? 答:

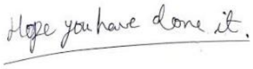
A Fine, Fine School
一所很好很好的学校

(6)
(6)

An image of
图片来源:

a parking sign that says
停车标志上写着
``Waiting? Please turn off
“等待? 请关闭。”
your engine."
你的引擎。”

(7)
(7)



An image of
图片来源:

Hope you have done it
希望你已经做到了

(8)
(8)

What's in this picture?
这张照片里是什么？

sausage roll.
香肠卷。

How to cook it?
怎么做？

Soak sausage roll in
把香肠卷泡进去
ketchup. Bake in oven for
番茄酱。放入烤箱烘烤
15 min. Cut and enjoy.
15 分钟,
切碎后享用。

Can I put cheese in the
我可以把奶酪放进
dish?
盘子？

Sure. But make sure it is
当然，但要确保它是
melted.
融化了。

(9)
(9)

What kind of animal is
什么样的动物
in the picture?
在照片里？

It's a zebra.
是斑马。

Where do they live in?
它们住在哪里？

In Africa.
在非洲。

Tell me a joke about
告诉我一个关于
zebras.
斑马。

Zebras get in trouble for
斑马有麻烦了
crossing the road.
过马路。

(10)
(十)

What's in this
picture?
这张照片里是什
么？

A screenshot of the Windows
视窗的截图
10 shutdown window
10 关闭窗口

I would like to restart
我想重新启动
my computer. Which
我的电脑
button should I click?
我应该点击哪个按钮？

Press OK.
按 OK。

Now I would not like to
现在我不想这么做
restart. What can I do?
重新开始，我能做什
么？

Click Cancel.
点击取消。

(11)
(11)

Figure 3: Selected examples generated from Kosmos-1. Blue boxes are input prompt and pink boxes are Kosmos-1 output. The examples include (1)-(2) image captioning, (3)-(6) visual question answering, (7)-(8) OCR, and (9)-(11) visual dialogue.
图 3: kosmos-1 生成的选定示例。蓝色框是输入提示符，粉色框是 kosmos-1 输出。例子包括(1)-(2)图像字幕，(3)-(6)视觉问答，(7)-(8) OCR，和(9)-(11)视觉对话。

Dataset 数据集	Task description 任务描述	Metric 米制	Zero-shot 一枪毙命	Few-shot 几发子弹
Language tasks 语言任务				
StoryCloze [MRL+17] 完形填空[MRL + 17]	Commonsense reasoning 常识推理	Accuracy 准确性	3	3
HellaSwag [ZHB+19] HellaSwag [ZHB + 19]	Commonsense NLI 常识 NLI	Accuracy 准确性	3	3
Winograd [LDM12a] Winograd [LDM12a]	Word ambiguity 词语歧义	Accuracy 准确性	3	3
Winogrande [SBBC20] Winogrande [SBBC20]	Word ambiguity 词语歧义	Accuracy 准确性	3	3
PIQA [BZB+20] PIQA [BZB + 20]	Physical commonsense 身体常识	Accuracy 准确性	3	3
BoolQ [CLC+19] 布尔克[CLC + 19]	Question answering 回答问题	Accuracy 准确性	3	3
CB [dMST19] CB [dMST19]	Textual entailment 文字蕴涵	Accuracy 准确性	3	3
COPA [RBG11] COPA [RBG11]	Causal reasoning 因果推理	Accuracy 准确性	3	3
Rendered SST-2 [RKH+21] 渲染 SST-2[RKH + 21]	OCR-free sentiment classification Ocr-自由情绪分类	Accuracy 准确性	3	
HatefulMemes [KFM+20] 可恨的文化基因 [KFM + 20]	OCR-free meme classification 无 ocr 模因分类	ROC AUC ROC AUC	3	
Cross-modal transfer 跨模式转移				
RelativeSize [BHCF16] 相对化[BHCF16]	Commonsense reasoning (object size) 常识推理(对象大小)	Accuracy 准确性	3	
MemoryColor [NHJ21] 内存颜色[NHJ21]	Commonsense reasoning (object color) 常识推理(对象颜色)	Accuracy 准确性	3	
ColorTerms [BBBT12] 色彩术语[BBBT12]	Commonsense reasoning (object color) 常识推理(对象颜色)	Accuracy 准确性	3	
Nonverbal reasoning tasks 非语言推理任务				
IQ Test 智商测试	Raven’s Progressive Matrices 雷文递进矩阵	Accuracy 准确性	3	
Perception-language tasks				

感知-语言任务				
COCO Caption [LMB+14] 可可标题[LMB + 14]	Image captioning 图像标题	CIDEr, etc. 苹果酒等。	3	3
Flicker30k [YLHH14] Flicker30k [YLHH14]	Image captioning 图像标题	CIDEr, etc. 苹果酒等。	3	3
VQAv2 [GKSS+17] VQAv2[GKSS + 17]	Visual question answering 可视化问题回答	VQA acc. VQA acc.	3	3
VizWiz [GLS+18] VizWiz [GLS + 18]	Visual question answering 可视化问题回答	VQA acc. VQA acc.	3	3
WebSRC [CZC+21] WebSRC [CZC + 21]	Web page question answering 网页问答	F1 score F1 成绩	3	
Vision tasks 视觉任务				
ImageNet [DDS+09] ImageNet [DDS + 09]	Zero-shot image classification 零拍图像分类	Top-1 acc. Top-1 acc 排名前 1 位的 acc。	3	
CUB [WBW+11] CUB [WBW + 11]	Zero-shot image classification with descriptions 带描述的零拍图像分类	Accuracy 准确性	3	

Table 1: We evaluate the capabilities of Kosmos-1 on language, perception-language, and vision tasks under both zero- and few-shot learning settings.
表 1: 我们评估了 kosmos-1 在零镜头和少镜头学习环境下语言、感知语言和视觉任务的能力。

1 Introduction: From LLMs to MLLMs
1 简介: 从 LLMs 到 MLLMs

Large language models (LLMs) have successfully served as a general-purpose interface across various natural language tasks [BMR+20]. The LLM-based interface can be adapted to a task as long as we are able to transform the input and output into texts. For example, the input of the summarization task is a document and the output is its summary. So we can feed the input document into the language model and then produce the generated summary.
大型语言模型(LLMs)已经成功地作为跨越各种自然语言任务的通用接口[BMR + 20]。只要我们能够将输入和输出转换为文本，基于 llm 的界面就可以适应任务。例如，摘要任务的输入是一个文档，输出是它的摘要。所以我们可以将输入文档输入到语言模型中，然后生成生成的摘要。

Despite the successful applications in natural language processing, it is still struggling to natively use LLMs for multimodal data, such as image, and audio. Being a basic part of intelligence, multimodal perception is a necessity to achieve artificial general intelligence, in terms of knowledge acquisition and grounding to the real world. More importantly, unlocking multimodal input [TMC+21, HSD+22, WBD+22, ADL+22, AHR+22, LLSH23] greatly widens the applications of language models to more high-value areas, such as multimodal machine learning, document intelligence, and robotics.

尽管在自然语言处理中已经有了成功的应用，但是它仍然在努力使用 LLMs 来处理多模态数据，比如图像和音频。作为智能的一个基本组成部分，多模态知觉是实现人工通用智能的必要条件，包括知识获取和现实世界的基础。更重要的是，解锁多模态输入[TMC + 21，HSD + 22，WBD + 22，ADL + 22，AHR + 22，LLSH23]极大地扩展了语言模型在更高价值领域的应用，如多模态机器学习，文档智能和机器人技术。

In this work, we introduce Kosmos-1, a Multimodal Large Language Model (MLLM) that can perceive general modalities, follow instructions (i.e., zero-shot learning), and learn in context (i.e., few-shot learning). The goal is to align perception with LLMs, so that the models are able to see and talk. To be specific, we follow MetaLM [HSD+22] to train the Kosmos-1 model from scratch. As shown in Figure 1, a Transformer-based language model is regarded as the general-purpose interface, and perception modules are docked with the language model. We train the model on web-scale multimodal corpora, i.e., text data, arbitrarily interleaved images and texts, and image-caption pairs. In addition, we calibrate the instruction-following capability across modalities by transferring language-only data.
在这项工作中，我们介绍了 KOSMOS-1，一个多模态大语言模型(MLLM)，可以感知一般模式，遵循指令(即零镜头学习)，并在上下文中学习(即少镜头学习)。我们的目标是将感知与 LLMs 结合起来，使模型能够看到和说话。具体来说，我们遵循 METALM [HSD + 22]从头开始训练 kosmos-1 模型。如图 1 所示，基于 transformer 的语言模型被视为通用接口，感知模块与语言模型停靠在一起。我们在网络规模的多模态语料库，即文本数据，任意交织的图像和文本，以及图像-标题对上训练模型。此外，我们通过传输只有语言的数据来校准跨模式的指令跟随能力。

As shown in Table 1, the Kosmos-1 model natively supports language, perception-language, and vision tasks. We also present some generated examples in Figure 2 and 3. In addition to various natural language tasks, the Kosmos-1 models natively handle a wide range of perception-intensive tasks, spanning visual dialogue, visual explanation, visual question answering, image captioning, simple math equation, OCR, and zero-shot image classification with descriptions. We also build
如表 1 所示，kosmos-1 模型本身支持语言、感知语言和视觉任务。我们还在图 2 和图 3 中提供了一些生成的示例。除了各种自然语言任务，kosmos-1 模型本身处理广泛的感知密集型任务，包括视觉对话、视觉解释、视觉问题回答、图像标题、简单数学方程、 OCR 和带描述的零镜头图像分类。我们还构建了

an IQ test benchmark following Raven's Progressive Matrices [JR03, CJS90], which evaluates the capability of nonverbal reasoning for MLLMs. The examples show that the native support of multimodal perception enables new opportunities to apply LLMs to new tasks. Moreover, we show that MLLMs achieve better commonsense reasoning performance compared with LLMs, which indicates cross-modal transfer helps knowledge acquisition.
在 Raven's Progressive Matrices [JR03，CJS90]之后建立了一个 IQ 测试基准，该测试评估了 MLLMs 的非语言推理能力。这些例子表明，多模态知觉的本地支持使得将 LLMs 应用于新任务的新机会成为可能。此外，我们表明，与 LLMs 相比，MLLMs 获得了更好的常识推理性能，这表明跨模态转移有助于知识获取。

The key takeaways are as follows:
主要结论如下:

From LLMs to MLLMs. Properly handling perception is a necessary step toward artificial general intelligence. The capability of perceiving multimodal input is critical to LLMs. First, multimodal perception enables LLMs to acquire commonsense knowledge beyond text descriptions. Second, aligning perception with LLMs opens the door to new tasks, such as robotics, and document intelli-gence. Third, the capability of perception unifies various APIs, as graphical user interfaces are the most natural and unified way to interact with. For example, MLLMs can directly read the screen or extract numbers from receipts. We train the Kosmos-1 models on web-scale multimodal corpora, which ensures that the model robustly learns from diverse sources. We not only use a large-scale text corpus but also mine high-quality image-caption pairs and arbitrarily interleaved image and text documents from the web.
从 LLMs 到 MLLMs。正确处理知觉是迈向人工通用智能的必要一步。感知多模态输入的能力对 LLMs 至关重要。首先，多模态感知使 LLMs 能够获得文本描述之外的常识知识。其次，将知觉与 LLMs 结合起来，为新的任务打开了大门，比如机器人技术和文档智能。第三，感知能力统一了各种 api，因为图形用户界面是最自然和统一的交互方式。例如，MLLMs 可以直接读取屏幕或从收据中提取数字。我们在网络规模的多模态语料库上训练 kosmos-1 模型，这确保了模型能够从不同的来源进行有力的学习。我们不仅使用大规模的文本语料库，而且还从 web 上挖掘高质量的图像标题对和任意交织的图像和文本文档。

Language models as general-purpose interfaces. Following the philosophy proposed in MetaLM [HSD+22], we regard language models as a universal task layer. Because of the open-ended output space, we are able to unify various task predictions as texts. Moreover, natural-language instructions and action sequences (such as programming language) can be well handled by language models. LLMs also serve as basic reasoners [WWS+22], which is complementary to perception modules on complex tasks. So it is natural to align world, action, and multimodal perception with the general-purpose interface, i.e., language models.
作为通用接口的语言模型。遵循 METALM [HSD + 22]中提出的哲学，我们认为语言模型是一个通用的任务层。由于开放式的输出空间，我们能够将各种任务预测统一为文本。此外，自然语言的指令和动作序列(如编程语言)可以被语言模型很好地处理。LLMs 也可以作为基本的推理器[WWS + 22] ，这是对复杂任务的感知模块的补充。因此，将世界、行动和多模态感知与通用界面(即语言模型)结合起来是很自然的。

New capabilities of MLLMs. As shown in Table 1, apart from the capabilities found in previous LLMs [BMR+20, CND+22], MLLMs enable new usages and possibilities. First, we can conduct zero- and few-shot multimodal learning by using natural language instructions and demonstration examples. Second, we observe promising signals of nonverbal reasoning by evaluating the Raven IQ test, which measures the fluid reasoning ability of humans. Third, MLLMs naturally support multi-turn interactions for general modalities, such as multimodal dialogue.
MLLMs 的新功能。如表 1 所示，除了在以前的 LLMs [BMR + 20，CND + 22]中发现的功能之外，MLLMs 还支持新的用法和可能性。首先，我们可以通过使用自然语言指令和演示示例来进行零和少镜头多模式学习。其次，我们通过评估 Raven IQ 测试来观察有希望的非语言推理信号，Raven IQ 测试测量人类的流体推理能力。第三，MLLMs 自然支持一般模式的多回合交互，如多模式对话。

2 Kosmos-1: A Multimodal Large Language Model

KOSMOS-1: 一个多模态大语言模型

As shown in Figure 1, Kosmos-1 is a multimodal language model that can perceive general modalities, follow instructions, learn in context, and generate outputs. Given the previous context, the model learns to

generate texts in an auto-regressive manner. Specifically, the backbone of Kosmos-1 is a Transformer-based causal language model. Apart from text, other modalities are embedded and fed into the language model. The Transformer decoder serves as a general-purpose interface to multimodal input. We train Kosmos-1 on multimodal corpora, including monomodal data, cross-modal paired data, and interleaved multimodal data. Once the models are trained, we can directly evaluate the models in zero-shot and few-shot settings on both language tasks and multimodal tasks.
如图 1 所示，kosmos-1 是一个多模态语言模型，它可以感知一般模式，遵循指令，在上下文中学习，并产生输出。鉴于以前的情况，该模型学习以自动回归的方式生成文本。具体来说，kosmos-1 的主干是一个基于 transformer 的因果语言模型。除了文本，其他模式被嵌入并输入到语言模型中。Transformer 解码器作为多模态输入的通用接口。我们对 kosmos-1 进行多模态语料库培训，包括单模态数据，跨模态配对数据和交错多模态数据。一旦模型得到训练，我们可以直接评估语言任务和多模态任务的零镜头和少镜头设置的模型。

2.1 Input Representation
2.1 输入表示法

The Transformer decoder perceives general modalities in a unified way. For input format, we flatten input as a sequence decorated with special tokens. Specifically, we use <s> and </s> to denote start-and end-of-sequence. The special tokens <image> and </image> indicate the beginning and end of encoded image embeddings. For example, “<s> document </s>” is a text input, and “<s> paragraph <image> Image Embedding </image> paragraph </s>” is an interleaved image-text input. Table 21 in Appendix shows some examples of input format.
Transformer 解码器以统一的方式感知一般模式。对于输入格式，我们将输入平滑为一个用特殊标记装饰的序列。具体来说，我们使用 < s > 和 </s > 来表示序列的开始和结束。特殊标记 < image > 和 </image > 表示编码图像嵌入的开始和结束。例如，“ < s > document </s >”是一个文本输入，“ < s > paragraph < Image > Image embedded </Image > paragraph </s >”是一个交错的 Image-text 输入。附录中的表 21 显示了一些输入格式的例子。

An embedding module is used to encode both text tokens and other input modalities into vectors. Then the embeddings are fed into the decoder. For input tokens, we use a lookup table to map them into embeddings. For the modalities of continuous signals (e.g., image, and audio), it is also feasible to represent inputs as discrete code and then regard them as “foreign languages” [WBD+22, WCW+23]. In this work, following [HSD+22], we employ a vision encoder as the embedding module for input
嵌入模块用于将文本标记和其他输入模式编码为向量。然后将嵌入信息输入解码器。对于输入标记，我们使用查找表将它们映射到嵌入。对于连续信号(如图像和音频)的模式，也可以将输入表示为离散代码，然后将它们视为“外语”[WBD + 22，WCW + 23]。在这项工作中，在[HSD + 22]之后，我们采用视觉编码器作为输入的嵌入模块

images. In addition, Resampler [ADL+22] is used as an attentive pooling mechanism to reduce the number of image embeddings.
此外，Resampler [ADL + 22]被用作一个专注的池机制，以减少图像嵌入的数量。

2.2 Multimodal Large Language Models (MLLMs)
2.2 多模态大语言模型

After obtaining the embeddings of an input sequence, we feed them into the Transformer-based decoder. The left-to-right causal model processes the sequence in an auto-regressive manner, which produces the next token by conditioning on past timesteps. The causal masking is used to mask out future information. A softmax classifier upon Transformer is used to generate tokens over the vocabulary.
在获得输入序列的嵌入后，我们将它们输入到基于 transformer 的解码器中。从左到右的因果模型以自动回归的方式处理序列，通过对过去的时间步骤进行条件化来产生下一个令牌。因果掩蔽被用来掩盖未来的信息。Transformer 上的 softmax 分类器用于在词汇表上生成标记。

MLLMs serve as general-purpose interfaces [HSD+22] that can perform interactions with both natural language and multimodal input. The framework is flexible to handle various data types, as long as we can represent input as vectors. MLLMs combine the best of two worlds. First, the language models naturally inherit the capabilities of in-context learning and instruction following. Second, perception is aligned with language models by training on multimodal corpora.
MLLMs 作为通用接口[HSD + 22] ，可以执行与自然语言和多模态输入的交互。框架灵活地处理各种数据类型，只要我们能够将输入表示为向量。MLLMs 结合了两个世界中最好的一个。首先，语言模型自然地继承了上下文学习和指令跟随的能力。第二，通过多模态语料库的训练，知觉与语言模型是一致的。

The implementation is based on the library TorchScale[3] [MWH+22], which is designed for large-scale model training. Compared with the standard Transformer architecture, we include the following modifications:
该实现基于图书馆 TorchScale3[MWH + 22] ，这是为大规模模型培训设计的。与标准的 Transformer 体系结构相比，我们包括以下修改:

MAGNETO We use MAGNETO [WMH+22], a Transformer variant, as the backbone architecture. MAGNETO has better training stability and superior performance across modalities. It introduces an extra LayerNorm to each sublayer (i.e., multi-head self-attention, and feed-forward network). The method has a theoretically derived initialization method [WMD+22] to improve the optimization fundamentally, which allows us to effectively scale up the models without pain.
我们使用 MAGNETO [WMH + 22] ，一种变压器变体，作为骨干结构。MAGNETO 具有更好的训练稳定性和跨模式的优越性能。它为每个子层引入了一个额外的层次范数(例如，多头自我关注和前馈网络)。该方法有一个理论推导的初始化方法[WMD + 22] ，从根本上改进了优化，使我们能够有效地放大模型没有痛苦。

xPos We employ xPos [SDP+22] relative position encoding for better long-context modeling. The method can better generalize to different lengths, i.e., training on short while testing on longer sequences. Moreover, xPos optimizes attention resolution so that the position information can be captured more precisely. The method xPos is efficient and effective in both interpolation and extrapolation settings.
XPOS 我们使用 XPOS [SDP + 22]相对位置编码来更好地进行长期上下文建模。该方法可以更好地推广到不同的长度，即在短时间内进行训练，而在长序列上进行测试。此外，XPOS 优化了注意力分辨率，使位置信息能够被更精确地捕捉。XPOS 方法在插值和外推设置上都是有效的。

2.3 Training Objective
2.3 培训目标

The KOSMOS-1 training is conducted on web-scale multimodal corpora, including monomodal data (e.g., text corpus), cross-modal paired data (e.g., image-caption pairs), and interleaved multimodal data (e.g., documents of arbitrarily interleaved images and texts). To be specific, we use monomodal data for representation learning. For example, language modeling with text data pretrains instruction following, in-context learning, and various language tasks. Moreover, cross-modal pairs and inter-leaved data learn to align the perception of general

modalities with language models. Interleaved data also naturally fit in the multimodal language modeling task. We present more details of training data collection in Section 3.1.
Kosmos-1 的训练是在网络规模的多模态语料库中进行的，包括单模态数据(例如文本语料库)、跨模态配对数据(例如图像-标题对)和交错的多模态数据(例如任意交错的图像和文本文档)。具体来说，我们使用单模态数据进行表征学习。例如，使用文本数据的语言建模预先训练了跟随指令、上下文学习和各种语言任务。此外，跨模态对和交叉数据学习将一般模式的感知与语言模型结合起来。交错数据也自然适合多模态语言建模任务。我们在第 3.1 节中提供了更多关于训练数据收集的细节。

The models are trained with the next-token prediction task, i.e., learning to generate the next token depending on the previous context. The training objective is to maximize the log-likelihood of tokens in examples. Notice that only discrete tokens, such as text tokens, are accounted for in the training loss. Multimodal language modeling is a scalable way to train the models. More importantly, the emergence of various capabilities makes the training task favorable for downstream applications.
模型通过下一个令牌预测任务进行训练，即根据上下文学习生成下一个令牌。训练的目标是最大化示例中令牌的对数可能性。请注意，只有离散的令牌，如文本令牌，被考虑在训练损失。多模态语言建模是一种可扩展的训练模型的方法。更重要的是，各种能力的出现使得培训任务有利于下游应用程序。

3 Model Training
3 模型训练

3.1 Multimodal Training Data
3.1 多模式培训数据

The models are trained on web-scale multimodal corpora. The training datasets consist of text corpora, image-caption pairs, and interleaved data of images and texts.
这些模型是在网络规模的多模态语料库上进行训练的。训练数据集包括文本语料库、图像标题对以及图像和文本的交织数据。

[3] https://github.com/microsoft/torchscale
Https://github. com/microsoft/torchscale

Text Corpora We train our model with The Pile [GBB+20] and Common Crawl (CC). The Pile is a massive English text dataset built for training large-scale language models, which is produced from a variety of data sources. We exclude data splits from GitHub, arXiv, Stack Exchange, and PubMed Central. We also include the Common Crawl snapshots (2020-50 and 2021-04) datasets, CC-Stories, and RealNews datasets [SPP+19, SPN+22]. The entire datasets have been purged of duplicate and near-duplicate documents, as well as filtered to exclude downstream task data. Refer to Appendix B.1.1 for detailed descriptions of training text corpora.
我们用 The Pile [GBB + 20]和 Common Crawl (CC)来训练我们的模型。Pile 是一个巨大的英文文本数据集，用于训练大规模的语言模型，它是由各种数据源产生的。我们排除了 GitHub，arXiv，Stack Exchange 和 PubMed Central 的数据分割。我们还包括 Common Crawl 快照(2020-50 和 2021-04)数据集、 CC-Stories 和 RealNews 数据集[SPP + 19，SPN + 22]。整个数据集已被清除重复和近似重复的文档，以及过滤排除下游任务数据。有关培训文本语料库的详细描述，请参阅附录 b. 1.1。

Image-Caption Pairs The image-caption pairs are constructed from several datasets, including English LAION-2B [SBV+22], LAION-400M [SVB+21], COYO-700M [BPK+22], and Conceptual Captions [SDGS18, CSDS21]. English LAION-2B, LAION-400M, and COYO-700M are collected from web pages of the Common Crawl web data by extracting image sources and the corresponding alt-text. Conceptual Captions are also from internet web pages. More details can be found in Appendix B.1.2.
图像-标题对图像-标题对由多个数据集构成，包括英文 LAION-2B [SBV + 22]、 LAION-400M [SVB + 21]、 COYO-700M [BPK + 22]和概念性标题[SDGS18，CSDS21]。通过提取图像源和相应的 alt 文本，从 Common Crawl web 数据的网页中收集了英文 LAION-2B、 LAION-400M 和 COYO-700M。概念说明也来自互联网网页。更多细节可以在附录 b. 1.2 中找到。

Interleaved Image-Text Data We collect interleaved multimodal data from the Common Crawl snapshot, which is a publicly available archive of web pages. We use a filtering process to select about 71M web pages from the original 2B web pages in the snapshot. We then extract the text and images from the HTML of each selected web page. For each document, we limit the number of images to five to reduce noise and redundancy. We also randomly discard half of the documents that only have one image to increase the diversity. We provide more details about the data collection process in Appendix B.1.3. By using this corpus, we enable Kosmos-1 to handle interleaved text and image and improve its few-shot ability.
交错图像-文本数据我们从 Common Crawl snapshot 收集交错的多模态数据，这是一个公开的网页存档。我们使用一个过滤过程从快照中的原始 2b 网页中选择大约 7100 万个网页。然后我们从每个选定的网页的 HTML 中提取文本和图片。对于每个文档，我们将图片数量限制在 5 个以减少噪音和冗余。我们还随机丢弃一半只有一张图片的文档，以增加多样性。我们在附录 b. 1.3 中提供了有关数据收集过程的更多细节。通过使用这个语料库，我们使 kosmos-1 能够处理交错的文本和图像，并提高其少镜头能力。

3.2 Training Setup

3.2 培训架构

The MLLM component has 24 layers with 2,048 hidden dimensions, 8,192 FFN intermediate size, and 32 attention heads, resulting in about 1.3B parameters. We use Magneto's initialization for optimization stability. For faster convergence, the image representation is obtained from a pretrained CLIP ViT-L/14 model with 1,024 feature dimensions. The images are preprocessed into 224 224 resolution during training. We freeze the parameters of the CLIP model except for the last layer during training. The total number of parameters of Kosmos-1 is about 1.6B. More details about hyperparameters can be found in Appendix A.
MLLM 组件有 24 层，2,048 个隐藏尺寸，8,192 个 FFN 中间尺寸，32 个注意头，约 1.3 b 参数。我们使用 Magneto 的初始化来优化稳定性。为了更快地收敛，图像表示是从具有 1024 个特征尺寸的预训练 CLIP ViT-L/14 模型获得的。在训练期间，图像被预处理为 224224 分辨率。我们冻结 CLIP 模型的参数，除了训练期间的最后一层。Kosmos-1 的参数总数约为 1.6 b。有关超参数的更多细节可以在附录 a 中找到。

We use a batch size of 1.2 million tokens (0.5 million tokens from text corpora, 0.5 million tokens from image-caption pairs, and 0.2 million tokens from interleaved data) and train Kosmos-1 for 300k steps, corresponding to about 360 billion tokens. We adopt the AdamW optimizer with

我们使用 120 万个令牌(来自文本语料库的 50 万个令牌，来自图像标题对的 50 万个令牌，以及来自交错数据的 20 万个令牌)，并将 kosmos-1 训练为 30 万步，相当于约 3600 亿个令牌。我们使用 AdamW 优化器

= (0:9; 0:98). We set the weight decay to 0.01 and the dropout rate to 0.1. The learning rate increases to 2e-4 for the first 375 warming-up steps and decays linearly to 0 for the rest of the training steps. We use SentencePiece [KR18] to tokenize the text. We preprocess the data in the “full-sentence” format [LOG+19], which packs each input sequence with full sentences that are sampled continuously from one or more documents.

= (0:9; 0:98).我们将重量衰减设置为 0.01，辍学率设置为 0.1。在前 375 个热身步骤中，学习速率增加到 2e-4，在其余的训练步骤中，学习速率线性衰减到 0。我们使用 SentencePiece [KR18]来标记文本。我们以“完整句子”格式[LOG + 19]对数据进行预处理，这种格式将从一个或多个文档中连续抽样的完整句子打包在每个输入序列中。

3.3 Language-Only Instruction Tuning
3.3 纯语言指令调优

In order to better align Kosmos-1 with human instructions, we perform language-only instruction tuning [LHV+23, HSLS22]. Specifically, we continue-train the model with the instruction data in the format of (instructions, inputs, and outputs). The instruction data is language-only, which is mixed with training corpora. The tuning process is conducted as language modeling. Notice that instructions and inputs are not accounted for in the loss. Section 4.9.1 shows that the improvements in the instruction-following capability can transfer across modalities.

为了更好地将 kosmos-1 与人类指令对齐，我们执行仅语言指令调整[LHV + 23，HSLS22]。具体而言，我们继续以(指令，输入和输出)的格式用指令数据对模型进行训练。指令数据是仅语言的，与培训语料库混合在一起。调优过程以语言建模的形式进行。请注意，指令和输入没有计入损失。第 4.9.1 节显示指令跟随能力的改进可以跨模式转移。

We combine Unnatural Instructions [HSLS22] and FLANv2 [LHV+23] as our instruction dataset. Unnatural Instructions is a dataset that was created by using a large language model to generate instructions for various natural language processing tasks. It has 68,478 instruction-input-output triplets in its core dataset. FLANv2 is a collection of datasets that cover diverse types of language understanding tasks, such as reading comprehension, commonsense reasoning, and closed-book question answering. We randomly select 54k examples of instructions from FLANv2 to augment our instruction dataset. Details of the training hyperparameter settings are described in Appendix A.2.

我们结合非自然指令[HSLS22]和 FLANv2[LHV + 23]作为指令数据集。Unnatural Instructions 是通过使用大型语言模型创建的数据集，用于为各种自然语言处理任务生成指令。它的核心数据集中有 68,478 个指令输入输出三元组。Flanv2 是一个涵盖不同类型语言理解任务的数据集合，例如阅读理解、常识推理和闭书问答。我们从 flanv2 中随机选择 54k 的指令例子来增加我们的指令数据集。训练超参数设置的细节在附录 a. 2 中描述。

4 Evaluation
4 评估

MLLMs can handle both language tasks and perception-intensive tasks. We evaluate Kosmos-1 on various types of tasks as follows:
MLLMs 可以同时处理语言任务和知觉密集型任务。我们评估 kosmos-1 的各种任务类型如下:

- Language tasks
  语言任务
  - – Language understanding
    - 语言理解
  - – Language generation
    一语言的产生
  - – OCR-free text classification
    - ocr-自由文本分类
- Cross-modal transfer
  跨模式转移
  - – Commonsense reasoning
    常识推理
- Nonverbal reasoning
  非语言推理
  - – IQ Test (Raven’s Progressive Matrices)
    - 智商测验(瑞文递进矩阵)
- Perception-language tasks
  感知-语言任务
  - – Image captioning
    - 图像字幕
  - – Visual question answering
    视觉问题回答
  - – Web page question answering
    - 网页问答
- Vision tasks
  视觉任务
  - – Zero-shot image classification
    - 零摄影图像分类
  - – Zero-shot image classification with descriptions
    - 带描述的零摄影图像分类

4.1 Perception-Language Tasks
4.1 感知-语言任务

We evaluate the perception-language capability of Kosmos-1 under vision-language settings. Specif-ically, we conduct zero-shot and few-shot experiments on two widely used tasks, including image captioning and visual question answering. Image captioning involves generating a natural language description of an image, while visual question answering aims to answer a natural language question with respect to an image.
我们评估了 kosmos-1 在视觉语言环境下的感知语言能力。具体来说，我们在两个广泛使用的任务上进行零镜头和少镜头实验，包括图像标题和视觉问题回答。图像字幕包括生成图像的自然语言描述，而视觉问题回答则旨在回答与图像相关的自然语言问题。

4.1.1 Evaluation Setup
4.1.1 评估设置

We evaluate the caption generation on MS COCO Caption [LMB+14], and Flickr30k [YLHH14]. We use the test set of COCO Karpathy split [KFF17], which re-partitions the train2014 and val2014 images [LMB+14] into 113,287, 5,000, and 5,000 for the training set, validation set, and test set, respectively. We conduct an evaluation on Flickr30k's Karpathy split test set. The image resolution is 224 224. We use beam search to generate the captions, and the beam size is 5. In the few-shot settings, we randomly sample demonstrations from the training set. We use COCOEvalCap[4] to compute CIDEr [VLZP15] and SPICE [AFJG16] scores as the evaluation metrics. We prompt Kosmos-1 with "An image of " for zero-shot and few-shot caption generation experiments.

我们评估 MS COCO 标题[LMB + 14]和 Flickr30k [YLHH14]上的标题生成。我们使用 COCO Karpathy split [KFF17]的测试集，它将 train2014 和 val2014 图像[LMB + 14]分别重新划分为训练集、验证集和测试集，分别为 113、287、5,000 和 5,000。我们对 Flickr30k 的 Karpathy 分割测试集进行评估。图像分辨率为 224224。我们使用光束搜索来生成标题，光束大小为 5。在少量拍摄的情况下，我们从训练集中随机抽取样本进行演示。我们使用 cocoevalcap4 计算苹果酒[VLZP15]和 SPICE [AFJG16]得分作为评估指标。我们提示 kosmos-1 用“ a image of”进行零拍和少拍字幕生成实验。

For visual question-answering tasks, we evaluate zero-shot and few-shot results on test-dev set of VQAv2 [GKSS+17] and test-dev set of VizWiz [GLS+18], respectively. The resolution of images is 224 224. We use greedy search for the decoding. We follow the normalization rules of the VQAv2 evaluation code[5] when computing the VQA accuracy. We evaluate the performance of VQA in an open-ended setting that Kosmos-1 generates answers and stops at the </s> ("end of sequence") token. The prompt is "Question: {question} Answer: {answer}" for visual question answering tasks.

对于视觉问答任务，我们分别在 vqav2 的 test-dev 集[GKSS + 17]和 VizWiz 的 test-dev 集[GLS + 18]上评价零镜头和零镜头结果。图像的分辨率是 224224。我们使用贪婪搜索来解码。在计算 VQA 准确性时，我们遵循 vqav2 评估代码 5 的规范化规则。我们评估了 VQA 在开放式设置下的性能，kosmos-1 生成答案并在 </s > (“序列结束”)标记处停止。视觉问题回答任务的提示是“ Question: { Question } Answer: { Answer }”。

4.1.2 Results

4.1.2 结果

Image Captioning Table 2 shows the zero-shot captioning performance on COCO Karpathy test split and Flickr30k test set. Kosmos-1 achieves remarkable results in zero-shot setting on two image captioning datasets. Specifically, our model achieves a CIDEr score of 67.1 on the Flickr30k dataset, compared to 60.6 and 61.5 for the Flamingo-3B and Flamingo-9B models, respectively. Notably, our model is able to accomplish this feat with a smaller size of 1.6B, compared to Flamingo models. This demonstrates our model's superiority in zero-shot image captioning.

图像字幕表 2 显示了 COCO Karpathy 测试 split 和 Flickr30k 测试集的零镜头字幕性能。Kosmos-1 在两个图像字幕数据集上实现了显著的零拍设置结果。具体来说，我们的模型在 Flickr30k 数据集上的苹果酒得分为 67.1，而 Flamingo-3B 和 Flamingo-9B 模型分别为 60.6 和 61.5。值得注意的是，与 Flamingo 模型相比，我们的模型能够以 1.6 b 的较小尺寸完成这一壮举。这证明了我们的模型在零拍图片字幕方面的优势。

[4] https://github.com/salaniz/pycocoevalcap
图片来源: http://github. com/salaniz/pycocoevalcap

[5] https://github.com/GT-Vision-Lab/VQA
5https://github. com/gt-vision-lab/VQA

Model 模特	COCO COCO		Flickr30k Flickr30k	
	CIDEr 苹果酒	SPICE 香料	CIDEr 苹果 酒	SPICE 香料
ZeroCap 零帽	14.6	5.5	-	-
VLKD VLKD	58.3	13.4	-	-
FewVLM 很少	-	-	31.0	10.0
METALM 金属	82.2	15.7	43.4	11.7
Flamingo-3B 火烈鸟 -3b	73.0	-	60.6	-
Flamingo-9B 火烈鸟 -9b	79.4	-	61.5	-
KOSMOS-1 (1.6B) KOSMOS-1(1.6 b)	84.7	16.8	67.1	14.5

Table 2: Zero-shot image captioning results on COCO caption Karpathy test and Flickr30k test.
表 2: COCO 字幕 Karpathy 测试和 Flickr30k 测试的零摄影图像字幕结果。

Flamingo [ADL+22] prompts with two examples from the downstream tasks while removing their corresponding images (i.e., similar to few-shot text prompts). The other models do not include any examples in the prompt.

Flamingo [ADL + 22]提示了两个来自下游任务的示例，同时删除了它们对应的图像(即，类似于少镜头文本提示)。其他模型在提示中不包含任何示例。

Table 3 reports the results of the few-shot (k = 2; 4; 8) settings. The overall performance improves as the number of shots increases from two to four. The trends are consistent across the two datasets. Moreover, the few-shot results outperform zero-shot captioning in Table 2.
表 3 报告了少镜头(k = 2; 4; 8)设置的结果。随着拍摄次数从 2 次增加到 4 次，总体性能提高。两个数据集的趋势是一致的。此外，少镜头结果优于表 2 中的零镜头字幕。

Model 模特	COCO COCO			Flickr30k Flickr30k		
	k = 2 K = 2	k = 4 k = 4	k = 8 k = 8	k = 2 k = 2	k = 4 k = 4	k = 8 k = 8
Flamingo-3B 火烈鸟 -3b	-	85.0	90.6	-	72.0	71.7
Flamingo-9B 火烈鸟 -9b	-	93.1	99.0	-	72.6	73.4
KOSMOS-1 (1.6B) KOSMOS-1(1.6 b)	99.6	101.7	96.7	70.0	75.3	68.0

Table 3: Few-shot image captioning results on COCO caption Karpathy test and Flickr30k test.
表 3: COCO 标题 Karpathy 测试和 Flickr30k 测试的少量图像标题结果。
CIDEr scores are reported.
苹果酒分数报告。

Visual Question Answering Table 4 reports the zero-shot visual question answering results on VQAv2 and VizWiz. We show that Kosmos-1 can better handle the diversity and complexity of the

VizWiz dataset. Kosmos-1 achieves higher accuracy and robustness than Flamingo-3B and Flamingo-9B models. In addition, our model is competitive with Flamingo on the VQAv2 dataset.
Visual Question Answering 表 4 报告了 vqav2 和 VizWiz 上的零镜头视觉问题回答结果。我们证明 kosmos-1 可以更好地处理 VizWiz 数据集的多样性和复杂性。Kosmos-1 实现了比 Flamingo-3B 和 Flamingo-9B 模型更高的准确性和鲁棒性。此外，我们的模型在 vqav2 数据集上与 Flamingo 竞争。

Model 模特	VQAv2 VQAv2	VizWiz VizWiz
Frozen 冰冻	29.5	-
VLKDViT-B/16 VLKDViT-B/16	38.6	-
METALM 金属	41.1	-
Flamingo-3B 火烈鸟 -3b	49.2	28.9
Flamingo-9B 火烈鸟 -9b	51.8	28.8
KOSMOS-1 (1.6B) KOSMOS-1(1.6 b)	51.0	29.2

Table 4: Zero-shot visual question answering results on VQAv2 and VizWiz. We present VQA
表 4: vqav2 和 VizWiz 上的零镜头视觉问题回答结果。我们呈现 VQA

accuracy scores. “ ”: Flamingo [ADL+22] builds the zero-shot prompt with two examples from the downstream tasks where their corresponding images are removed (i.e., similar to few-shot text prompts) while the others evaluate true zero-shot learning.
准确度分数。“”: Flamingo [ADL + 22]构建了零镜头提示符，其中两个示例来自下游任务，在这两个示例中，它们相应的图像被删除(即，类似于少镜头文本提示符) ，而其他示例则评估真正的零镜头学习。

Table 5 shows the few-shot performance on visual question answering tasks. Kosmos-1 outperforms other models in few-shot (k = 2; 4) settings on the VizWiz dataset. We also observe a positive correlation between the number of shots and the quality of the results on the VizWiz dataset. Moreover, the few-shot results are better than the zero-shot numbers as reported in Table 4.
表 5 显示了视觉问答任务的少镜头表现。Kosmos-1 在 VizWiz 数据集的少镜头(k = 2; 4)设置中优于其他模型。我们还观察到 VizWiz 数据集上的镜头数量与结果质量之间呈正相关。此外，少量拍摄结果优于表 4 中报告的零拍数。

Model 模特	VQAv2 VQAv2			VizWiz VizWiz		
	k = 2 K = 2	k = 4 k = 4	k = 8 k = 8	k = 2 k = 2	k = 4 k = 4	k = 8 k = 8
Frozen 冰冻	-	38.2	-	-	-	-
METALM 金属	-	45.3	-	-	-	-
Flamingo-3B 火烈鸟 -3b	-	53.2	55.4	-	34.4	38.4
Flamingo-9B 火烈鸟 -9b	-	56.3	58.0	-	34.9	39.4
KOSMOS-1 (1.6B) KOSMOS-1(1.6 b)	51.4	51.8	51.4	31.4	35.3	39.0

Table 5: Few-shot visual question answering results on VQAv2 and VizWiz.VQA accuracy scores are reported.
表 5: vqav2 和 VizWiz.VQA 准确性分数报告的少量视觉问答结果。

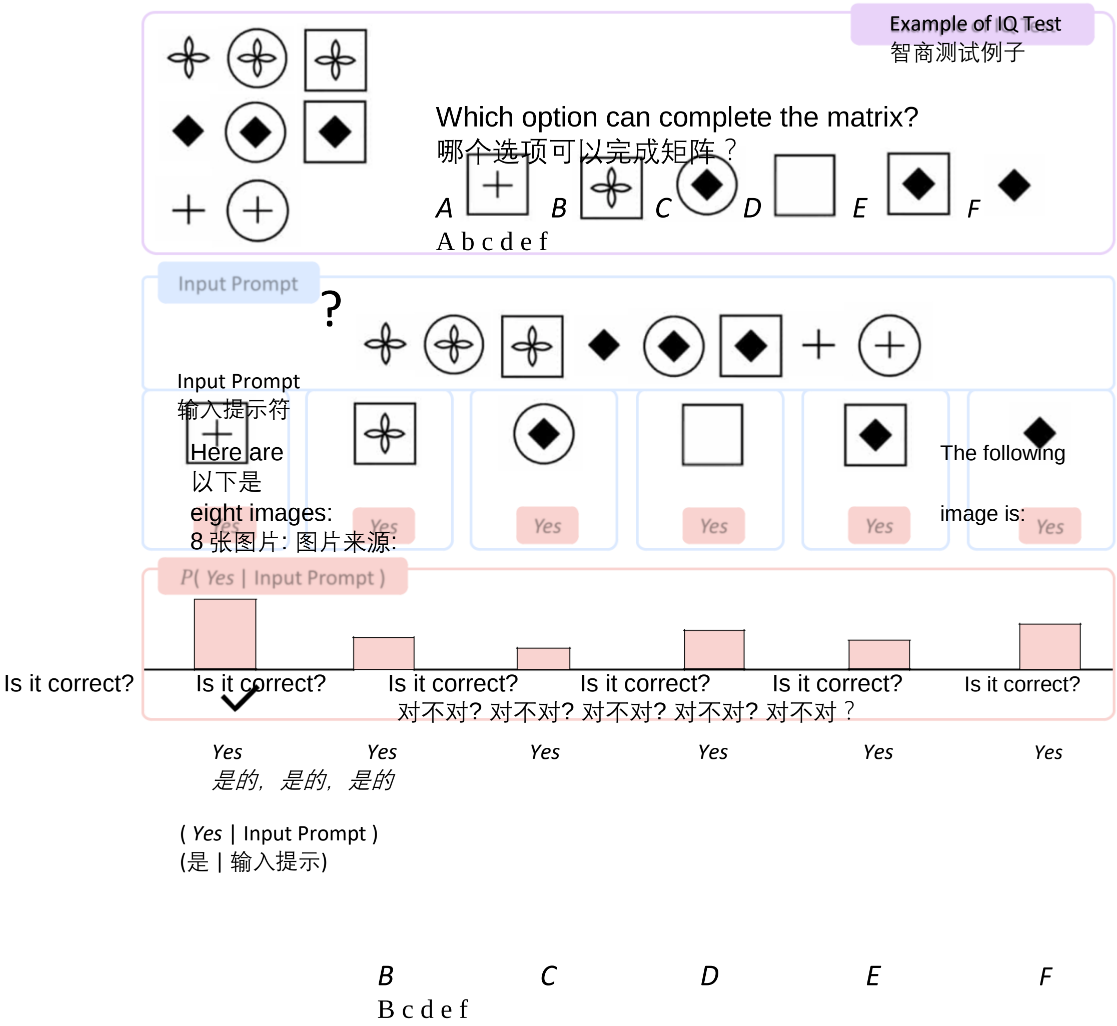


Figure 4: Top: An example of Raven IQ test. Bottom: Evaluate Kosmos-1 on Raven IQ test. The input prompt consists of the flattened image matrix and verbal instruction. We append

each candidate image to the prompt separately and query the model if it is correct. The final prediction is the candidate that motivates the model to yield the highest probability of “Yes”.
图 4: Top: Raven IQ 测试的一个例子。下图: 在 Raven IQ 测试中评估 KOSMOS-1。输入提示包括平面图像矩阵和语言指令。我们将每个候选图像分别附加到提示符上，然后查询模型是否正确。最终的预测是激励模型产生“ Yes”的最高概率的候选者。

4.2 IQ Test: Nonverbal Reasoning

4.2 IQ 测试: 非语言推理

Raven’s Progressive Matrices [CJS90, JR03] is one of the most common tests to evaluate nonverbal reasoning. The capability of nonverbal reasoning is typically a reflection of an individual’s intelligence quotient (IQ). Figure 4 shows an example. Given eight images presented in a 3 3 matrix, the task is to identify the following element from six similar candidates.
Raven’s Progressive Matrices [CJS90，JR03]是评估非语言推理最常用的测试之一。非语言推理的能力通常是个人智商(IQ)的反映。图 4 显示了一个例子。给定 8 张图片呈现在一个 33 矩阵中，任务是从 6 个相似的候选者中识别出下面的元素。

The models need to conduct zero-shot nonverbal reasoning without explicitly fine-tuning. The Raven IQ test is analogous to in-context learning of language models, where the difference is whether the context is nonverbal or verbal. In order to infer the answers, the models have to recognize abstract concepts and identify the underlying patterns of given images. So the IQ task is a good testbed to benchmark the nonverbal in-context learning capability.
模型需要在没有明确微调的情况下进行零次非语言推理。Raven IQ 测试类似于语言模型的上下文学习，区别在于上下文是非语言的还是语言的。为了推断答案，模型必须识别抽象概念，并识别给定图像的潜在模式。因此，IQ 任务是一个很好的测试基准，非语言环境下的学习能力。

4.2.1 Evaluation Setup
4.2.1 评估设置

To evaluate the Kosmos-1 on zero-shot nonverbal reasoning, we construct a dataset of the Raven IQ test. It consists of 50 examples collected from different websites[6789]. Each example has three (i.e., 2 2 matrix), four, or eight (i.e., 3 3 matrix) given images. The goal is to predict the next one. Each instance has six candidate images with a unique correct completion. We measure accuracy scores to evaluate the models. The evaluation dataset is available at https://aka.ms/kosmos-iq50.
为了评估 kosmos-1 在零次非语言推理方面的能力，我们构建了 Raven IQ 测试的数据集。它由从不同网站收集的 50 个例子组成 6789。每个例子都有三个(即 2 个矩阵)，4 个或 8 个(即 3 个矩阵)给定的图像。目标是预测下一个。每个实例都有六个候选图像，每个候选图像都有一个唯一正确的完成。我们测量准确性分数来评估模型。评估数据集可在 https://aka.ms/kosmos-iq50 获得。

Figure 4 illustrates how to evaluate Kosmos-1 on the Raven IQ test. The matrix-style images are flattened and fed into the models one-by-one. To enable the model to better understand the desired task, we also use a textual instruction “Here are three/four/eight images:”, “The following image is:”, and “Is it correct?” for conditioning. We append each possible candidate to the context separately and compare the probability that the model outputs “Yes” in a close-ended setting. The candidate that yields the largest probability is regarded as the prediction.
图 4 说明了如何在 Raven IQ 测试中评估 KOSMOS-1。矩阵样式的图像被压平并一个接一个地输入模型。为了让模型更好地理解所需的任务，我们还使用了一个文本指令“ Here are three/four/eight images:”，“ The following image Is:”，“ Is it right?”用于条件反射。我们将每个可能的候选人分别附加到上下文中，并比较模型在封闭环境中输出“ Yes”的概率。产生最大概率的候选者被认为是预测。

4.2.2 Results
4.2.2 结果

Table 6 shows the evaluation results on the IQ test dataset. Both Kosmos-1 with and without language-only instruction tuning achieve 5.3% and 9.3% improvement respectively over the random baseline. The results indicate that Kosmos-1 is able to perceive abstract conceptual patterns in a nonverbal context, and then deduce the following element across multiple choices. To the best of our knowledge, it is the first time that a model can perform such zero-shot Raven IQ tests. Although there is still a large performance gap between the current model and the average level of adults, Kosmos-1 demonstrates the potential of MLLMs to perform zero-shot nonverbal reasoning by aligning perception with language models.
表 6 显示了 IQ 测试数据集的评估结果。使用和不使用语言指令调整的 kosmos-1 分别比随机基线分别提高 5.3% 和 9.3% 。结果表明，kosmos-1 能够在非语言环境中感知抽象的概念模式，然后在多个选择中推导出以下元素。据我们所知，这是第一次一个模型可以执行这样的零镜头 Raven IQ 测试。尽管目前的模型与成年人的平均水平之间仍然存在很大的性能差距，kosmos-1 证明了 MLLMs 通过将感知与语言模型结合起来进行零次非语言推理的潜力。

Method 方法	Accuracy 准确性
Random Choice	17%
随机选择	17%
KOSMOS-1	22%
宇宙 1 号	22%
w/o language-only instruction tuning	26%
W/o 语言指令调优	26%

Table 6: Zero-shot generalization on Raven IQ test.
表 6: 乌鸦智商测验的零点概括。

4.3 OCR-Free Language Understanding
4.3 无 ocr 语言理解

OCR-free language understanding is a task that focuses on understanding text and images without relying on Optical Character Recognition (OCR). For example, during the Rendered SST-2 task, sentences from the Stanford Sentiment Treebank [SPW+13] dataset are rendered as images. The model is asked to predict the sentiment of the text within the images. The task evaluates a model’s ability to read and comprehend the meaning of words and sentences directly from the images.
不依赖光学字符识别(OCR)的语言理解是一项集中于理解文本和图像而不依赖光学字符识别(OCR)的任务。例如，在渲染 sst-2 任务中，来自 Stanford Sentiment tree bank [SPW + 13]数据集的句子被渲染为图像。该模型被要求预测图像中文本的情绪。这个任务评估模型直接从图像中阅读和理解单词和句子的意思的能力。

4.3.1 Evaluation Setup

4.3.1 评估设置

We evaluate OCR-free language understanding on the Rendered SST-2 [RKH+21] test set and HatefulMemes [KFM+20] validation set. We use accuracy as the metric for the Rendered SST-2 and report ROC AUC for the HatefulMemes dataset. We use the prompt “Question: what is the sentiment of the opinion? Answer: {answer}”, where the answer is either positive or negative for the Rendered SST-2. For the HatefulMemes task, the prompt is “Question: does this picture contain real hate speech? Answer: {answer}”, where the answer is either yes or no.
我们在渲染的 SST-2[RKH + 21]测试集和 HatefulMemes [KFM + 20]验证集上评估无 ocr 语言理解。我们使用准确性作为呈现 sst-2 的度量标准，并报告 HatefulMemes 数据集的 ROC AUC。我们使用提示“问题: 意见的情绪是什么？答案: { Answer }”，这里的答案对于呈现的 sst-2 要么是正面的，要么是负面的。对于 HatefulMemes 任务，提示符是“ Question: 这张图片是否包含真正的仇恨言论？答案: {回答}”，其中的答案要么是肯定的，要么不是。

4.3.2 Results

4.3.2 结果

As shown in Table 7, Kosmos-1 achieves a ROC AUC of 63.9% for the HatefulMemes validation set and a test accuracy of 67.1% for the Rendered SST-2 test set. It outperforms CLIP ViT-L
如表 7 所示，kosmos-1 对于 HatefulMemes 验证集的 ROC AUC 为 63.9%，对于渲染的 sst-2 测试集的测试准确率为 67.1% 。它优于 CLIP ViT-L

[6] https://en.testometrika.com/intellectual/iq-test/
6https://en.testometrika. com/intellectual/iq-test/

[7] https://en.testometrika.com/intellectual/iq-test-for-kids-7-to-16-year-old/
Https://en.testometrika. com/intellectual/iq-test-for-kids-7-to-16-year-old/

[8] https://iqpro.org/
8https://iqpro.org/

[9] https://iqhaven.com/matrix-g
9https://iqhaven. com/matrix-g

and Flamingo-9B, which achieve AUCs of 63.3% and 57.0% on the HatefulMemes task. Note that Flamingo explicitly provides OCR text into the prompt, while Kosmos-1 does not access any external tools or resources. This indicates that Kosmos-1 has built-in abilities to read and comprehend the text in the rendered images.

和 Flamingo-9B，在 HatefulMemes 任务上分别获得了 63.3% 和 57.0% 的 auc。请注意，Flamingo 明确地在提示符中提供了 OCR 文本，而 kosmos-1 没有访问任何外部工具或资源。这表明 kosmos-1 具有内置的读取和理解渲染图像中文本的能力。

Model 模特	HatefulMemes 可恨的文化基因	Rendered SST-2 渲染 SST-2
CLIP ViT-B/32 CLIP ViT-B/32 剪辑 ViT-B/32	57.6	59.6
CLIP ViT-B/16 CLIP ViT-B/16	61.7	59.8
CLIP ViT-L/14 CLIP vit-1/14 剪辑 vit-1/14	63.3	64.0
Flamingo-3B 火烈鸟 -3b	53.7	-
Flamingo-9B 火烈鸟 -9b	57.0	-
KOSMOS-1 (1.6B) KOSMOS-1(1.6 b)	63.9	67.1

Table 7: Zero-shot generalization on OCR-free language understanding. We report accuracy scores.
表 7: 无 ocr 语言理解的零镜头概括。我们报告准确性得分。

4.4 Web Page Question Answering
4.4 网页问答

Web page question answering aims at finding answers to questions from web pages. It requires the model to comprehend both the semantics and the structure of texts. The structure of the web page (such as tables, lists, and HTML layout) plays a key role in how the information is arranged and displayed. The task can help us evaluate our model's ability to understand the semantics and the structure of web pages.

网页问答旨在从网页中寻找问题的答案。它需要模型同时理解文本的语义和结构。网页的结构(如表格、列表和 HTML 布局)在信息的排列和显示方式中起着关键作用。这个任务可以帮助我们评估我们的模型理解语义和网页结构的能力。

4.4.1 Evaluation Setup
4.4.1 评估设置

We compare the performance on the Web-based Structural Reading Comprehension (WebSRC) dataset [CZC+21]. For comparisons, we train a language model (LLM) on the same text corpora with the same training setup as in Kosmos-1. The LLM takes the text extracted from the web page as input. Its template of the prompt is "Given the context below from web page, extract the answer from the given text like this: Qusestion: Who is the publisher of this book? Answer: Penguin Books Ltd. Context: {WebText} Q: {question} A: {answer} ", where the {WebText} presents the text extracted from the web page. Besides using the same prompt, Kosmos-1 prepends the image before the prompt. Two example images from WebSRC are shown in Appendix C.3. Following the original paper [CZC+21], we use exact match (EM) and F1 scores as our evaluation metrics.

我们比较了基于 web 的结构化阅读理解(WebSRC)数据集[CZC + 21]的性能。为了进行比较，我们在与 kosmos-1 相同的训练设置的相同文本语料库上训练语言模型(LLM)。LLM 将从网页中提取的文本作为输入。它的提示模板是“根据网页上的上下文，从给定的文本中提取答案，如下所示：提问：谁

是这本书的出版商？答案：Penguin Books 有限公司。Context: { WebText } q: { question } a: { answer }”，其中{ WebText }显示从网页中提取的文本。除了使用相同的提示符，kosmos-1 在提示符前面加上图片。来自 WebSRC 的两个示例图像显示在附录 c. 3 中。在原始论文[CZC + 21]之后，我们使用精确匹配(EM)和 f1 分数作为我们的评估指标。

4.4.2 Results
4.4.2 结果

The experimental results are summarized in Table 8. We observe that Kosmos-1 outperforms the LLM, indicating that Kosmos-1 can benefit from the layout and style information of web pages in images. In addition, we evaluate the performance of Kosmos-1 without the extracted text in the prompt. It shows that extracted text has a contribution of +12.0/20.7 EM/F1 to Kosmos-1, indicating that the benefit from modeling images does not sacrifice its language abilities.

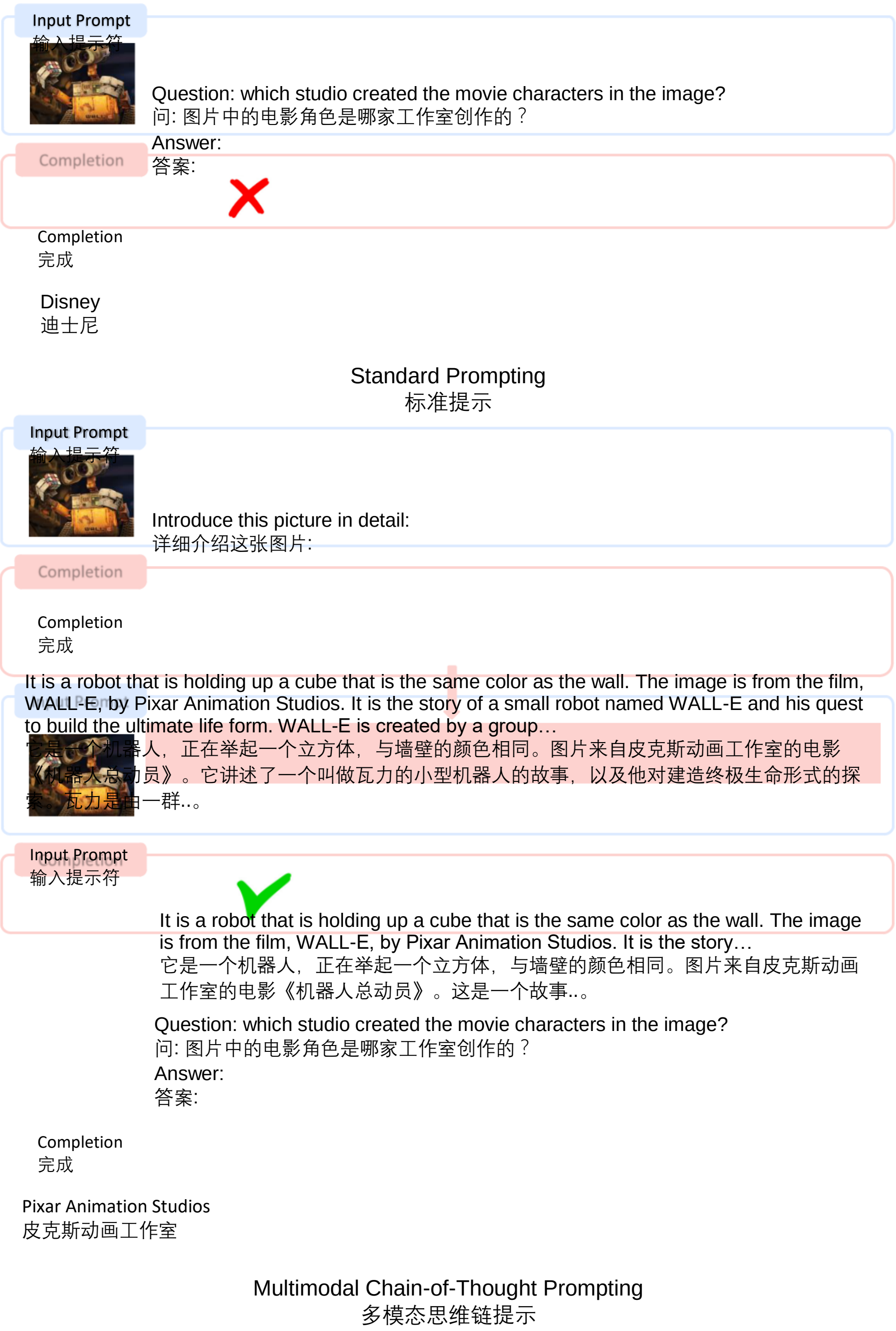
实验结果总结在表 8 中。我们观察到 kosmos-1 优于 LLM，表明 kosmos-1 可以从图像中网页的布局和样式信息中受益。此外，我们评估了 kosmos-1 在没有提示中提取文本的情况下的性能。实验结果表明，提取的文本对 kosmos-1 的贡献为 + 12.0/20.7 EM/F1，说明图像建模的好处并不影响其语言能力。

Models 模型 EM F1	EM	F1
Using extracted text 使用提取的文本		
LLM LLM 7.617.9	7.6	17.9
Kosmos-1 宇宙 -115.831.3	15.8	31.3
Without using extracted text 不使用提取的文本		
Kosmos-1 宇宙 -13.810.6	3.8	10.6

Table 8: Zero-shot performance on WebSRC task. We report exact match (EM) and F1 scores.
表 8: WebSRC 任务的零成绩。我们报告精确匹配(EM)和 f1 成绩。

4.5 Multimodal Chain-of-Thought Prompting
4.5 多模态思维链提示

Chain-of-thought prompting [WWS+22] allows large language models to generate a series of reasoning steps and decompose a multi-step problem into intermediate steps, which can significantly
思想链提示[WWS + 22]允许大型语言模型生成一系列推理步骤，并将多步问题分解为中间步骤，这可以显著提高



Question: which studio created the movie characters in the image?
问: 图片中的电影角色是哪家工作室创作的？
Answer:
答案:

Disney
迪士尼

Standard Prompting
标准提示

Introduce this picture in detail:
详细介绍这张图片:

It is a robot that is holding up a cube that is the same color as the wall. The image is from the film, WALL-E, by Pixar Animation Studios. It is the story of a small robot named WALL-E and his quest to build the ultimate life form. WALL-E is created by a group…
它是一个机器人，正在举起一个立方体，与墙壁的颜色相同。图片来自皮克斯动画工作室的电影《机器人总动员》。它讲述了一个叫做瓦力的小型机器人的故事，以及他对建造终极生命形式的探索。瓦力是由一群..。

It is a robot that is holding up a cube that is the same color as the wall. The image is from the film, WALL-E, by Pixar Animation Studios. It is the story…
它是一个机器人，正在举起一个立方体，与墙壁的颜色相同。图片来自皮克斯动画工作室的电影《机器人总动员》。这是一个故事..。

Question: which studio created the movie characters in the image?
问: 图片中的电影角色是哪家工作室创作的？
Answer:
答案:

Pixar Animation Studios
皮克斯动画工作室

Multimodal Chain-of-Thought Prompting
多模态思维链提示

Figure 5: Multimodal Chain-of-Thought prompting enables KOSMOS-1 to generate a rationale first, then to tackle complex question-answering and reasoning tasks.
图 5: 多模态思维链提示使 kosmos-1 能够首先生成一个基本原理，然后处理复杂的问题回答和推理任务。

improve the performance in complex tasks. Motivated by chain-of-thought prompting, we investigate a multimodal chain-of-thought prompting using KOSMOS-1. As illustrated in Figure 5, we break down perception-language tasks into two steps. In the first stage, given an image, we use a prompt to guide the model to generate a rationale. The model is then fed the rationale and a task-aware prompt to produce the final results.
提高在复杂任务中的表现。在思维链提示的激励下，我们使用 kosmos-1 研究了一种多模态思维链提示。如图 5 所示，我们将感知语言任务分为两个步骤。在第一阶段，给定一个图像，我们使用提示来引导模型生成一个基本原理。然后给模型提供基本原理和任务感知提示，以产生最终结果。

4.5.1 Evaluation Setup

4.5.1 评估设置

We evaluate the ability of multimodal chain-of-thought prompting on the Rendered SST-2. We use the prompt “Introduce this picture in detail:” to generate the content in the picture as the rationale. Then, we use the prompt “{rationale} Question: what is the sentiment of the opinion? Answer: {answer}” to predict the sentiment, where the answer is either positive or negative.

我们评估呈现的 sst-2 上的多模态思维链提示的能力。我们使用提示“介绍这幅图片的细节:”来生成图片中的内容作为基本原理。然后，我们使用提示符“{ rationale } Question: 观点的情绪是什么？回答: { Answer }”来预测情绪，答案是积极的还是消极的。

4.5.2 Results

4.5.2 结果

We conduct experiments to evaluate the performance of the multimodal chain-of-thought prompting. Table 9 shows that multimodal chain-of-thought prompting achieves a score of 72.9, which is 5.8 points higher than the standard prompting. By generating intermediate content, the model can recognize the text in the images and infer the sentiment of the sentences more correctly.

我们进行实验来评估多模态思维链提示的性能。表 9 显示，多模态思维链提示得分为 72.9 分，比标准提示高出 5.8 分。通过生成中间内容，模型可以识别图像中的文本，并更准确地推断句子的情绪。

4.6 Zero-Shot Image Classification

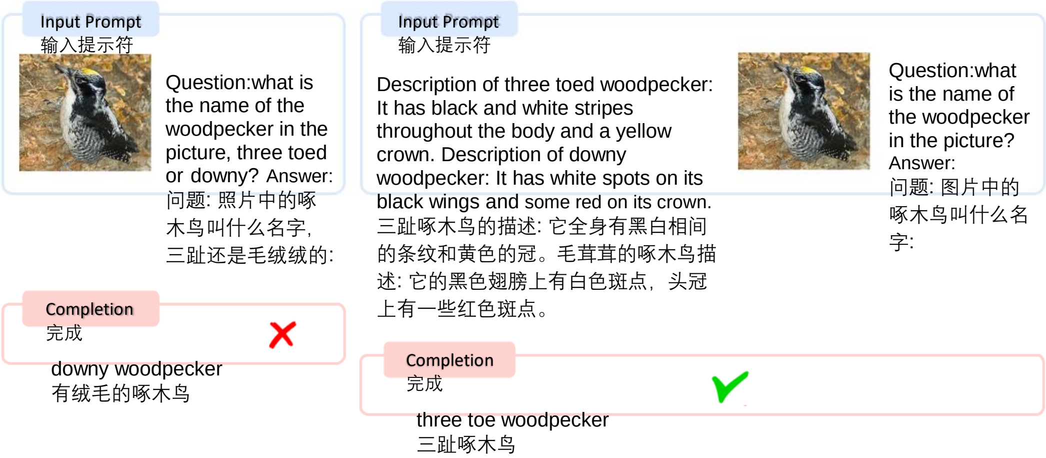
4.6 零摄影影像分类

We report the zero-shot image classification performance on ImageNet [DDS+09]. Image classifica-tion comprehends an entire image as a whole and aims to assign a label to the image. We map each label to its category name in natural language. The model is prompted to predict the category name to perform zero-shot image classification.

我们报告了 ImageNet [DDS + 09]上的零拍图像分类性能。图像分类将整个图像作为一个整体来理解，并旨在为图像分配一个标签。我们用自然语言将每个标签映射到它的分类名称。模型被提示预测类别名称来执行零镜头图像分类。

Models 模特	Accuracy 准确性
CLIP ViT-B/32 CLIP ViT-B/32 剪辑 ViT-B/32	59.6
CLIP ViT-B/16 CLIP ViT-B/16	59.8
CLIP ViT-L/14 CLIP vit-1/14 剪辑 vit-1/14	64.0
KOSMOS-1 宇宙 1 号	67.1
w/ multimodal CoT prompting 多式联运婴儿床提示	72.9

Table 9: Multimodal chain-of-thought (CoT) prompting on Rendered SST-2 task.
表 9: 呈现 sst-2 任务的多模态思维链(CoT)提示。



Question:what is the name of the woodpecker in the picture, three toed or downy? Answer:
问题: 照片中的啄木鸟叫什么名字,三趾还是毛绒绒的:

downy woodpecker
有绒毛的啄木鸟

Description of three toed woodpecker: It has black and white stripes throughout the body and a yellow crown. Description of downy woodpecker: It has white spots on its black wings and some red on its crown.
三趾啄木鸟的描述: 它全身有黑白相间的条纹和黄色的冠。毛茸茸的啄木鸟描述: 它的黑色翅膀上有白色斑点，头冠上有一些红色斑点。

Question:what is the name of the woodpecker in the picture? Answer:
问题: 图片中的啄木鸟叫什么名字:

three toe woodpecker
三趾啄木鸟

Zero-Shot Classification
带描述的零镜头分类

Zero-Shot Classification with Descriptions

Figure 6: In-context verbal descriptions can help Kosmos-1 recognize visual categories better.
图 6: 上下文中的语言描述可以帮助 kosmos-1 更好地识别视觉类别。

4.6.1 Evaluation Setup
4.6.1 评估设置

Given an input image, we concatenate the image with the prompt “The photo of the”. The input is then fed into the model to obtain the category name of the image. We evaluate the model on ImageNet [DDS+09], which contains 1.28M training images and 50k validation images in 1k object categories. The prediction is classified as correct if it is exactly the same as the ground-truth category name. The image resolution used for evaluation is 224 224. We use beam search to generate the category names and the beam size is 2.
给定一个输入图像，我们将图像与提示符“ The photo of The”连接起来。然后输入到模型中，获得图像的类别名称。我们评估 ImageNet [DDS + 09]上的模型，该模型包含 1.28 m 训练图像和 1k 对象类别中的 50k 验证图像。如果预测与地面真相类别名称完全相同，则将其分类为正确。用于评估的图像分辨率是 224224。我们使用光束搜索来生成类别名称，光束大小为 2。

4.6.2 Results
4.6.2 结果

As shown in Table 10, we report zero-shot results in both constrained and unconstrained settings. The difference between the two settings is whether we use the 1k object category

names to constrain the decoding. Kosmos-1 significantly outperforms GIT [WYH+22] by 4.6% under the constrained setting and 2.1% under the unconstrained setting.
如表 10 所示，我们报告了约束和无约束设置下的零拍结果。这两个设置的区别在于我们是否使用 1k 对象类别名来约束解码。Kosmos-1 在约束设置下显着优于 GIT [WYH + 22]4.6% ， 在无约束设置下为 2.1% 。

Model 模特	Without Constraints 没有约束	With Constraints 带有约束
GIT [WYH+22] GIT [WYH + 22]	1.9	33.5
KOSMOS-1 宇宙 1 号	4.0	38.1

Table 10: Zero-shot image classification on ImageNet. For the results with constraints, we use the 1k ImageNet object category names for constrained decoding. We report top-1 accuracy scores.
表 10: ImageNet 上的零摄影图像分类。对于带有约束的结果，我们使用 1k ImageNet 对象类别名称进行约束解码。我们报告前 1 名的准确性得分。

4.7 Zero-Shot Image Classification with Descriptions
4.7 带描述的零拍图像分类

The standard approach of image classification as above is to prompt the model for the specific name of the object depicted in the image. However, there are also some classification rules customized for different users and scenarios, such as the refined classification of complex animal subspecies. We can utilize natural language descriptions to guide Kosmos-1 to distinguish images in the zero-shot setting, which makes the decision process more interpretable.
上述图像分类的标准方法是提示模型为图像中描述的对象的特定名称。然而，也有一些分类规则为不同的用户和场景定制，如复杂的动物亚种的精确分类。我们可以使用自然语言描述来指导 kosmos-1 在零拍摄环境中区分图像，这使得决策过程更具可解释性。

Category 1 第一类	Category 2 第二类
three toed woodpecker 三趾啄木鸟	downy woodpecker 有绒毛的啄木鸟
 It has black and white stripes throughout the body and a yellow crown. 它全身有黑白相间的条纹，头上戴着黄色的皇冠。	 It has white spots on its black wings and some red on its crown. 它黑色的翅膀上有白色的斑点，头顶上有一些红色的斑点。
Gentoo penguin Gentoo 企鹅(Gentoo penguin)皇家企鹅	royal penguin
 It has a black head and white patch above its eyes. 它的头是黑色的，眼睛上方有一块白色的斑块。	 It has a white face and a yellow crown. 它有一张白色的脸和一个黄色的皇冠。
black throated sparrow 黑喉麻雀狐狸麻雀	fox sparrow
 It has white underparts and a distinctive black bib on the throat. 它的腹部是白色的，喉咙上有一个与众不同的黑色围嘴。	 It has a reddish-brown plumage and a streaked breast. 它有红褐色的羽毛和有条纹的胸部。

Table 11: The detailed descriptions of different categories for in-context image classification.
表 11: 上下文图像分类的不同类别的详细描述。

4.7.1 Evaluation Setup
4.7.1 评估设置

Following CUB [WBW+11], we construct a bird classification dataset that contains images and natural-language descriptions of categories. The dataset has three groups of binary image classifica-tion. Each group contains two animal categories with similar appearances. Our goal is to classify images given the categories' descriptions. Table 11 presents the data samples. The first group is from [WBW+11], while the other two groups are collected from the website. Each category contains twenty images.
遵循 CUB [WBW + 11] ，我们构建了一个鸟类分类数据集，其中包含图像和自然语言的类别描述。数据集有三组二进制图像分类。每组包含两个外观相似的动物类别。我们的目标是根据类别的描述对图像进行分类。表 11 展示了数据样本。第一组来自[WBW + 11] ，而另外两组来自网站。每个类别包含二十个图像。

The evaluation procedure is illustrated in Figure 6. For the zero-shot setting, we provide detailed descriptions of two specific categories and use the template "Question:what is the name of {general category} in the picture? Answer:" to prompt the model for the specific category name in an open-ended manner. To evaluate the effect of providing verbal descriptions in context, we also implement a zero-shot baseline without prompting descriptions. Instead, we provide the corresponding specific names in the prompt.
评估过程如图 6 所示。对于零镜头设置，我们提供了两个特定类别的详细描述，并使用模板"问题：图片中{一般类别}的名称是什么？回答:"以开放式方式提示特定类别名称的模型。为了评估在上下文中提供语言描述的效果，我们还实现了一个零拍基线，而不提示描述。相反，我们在提示中提供相应的具体名称。

4.7.2 Results

4.7.2 结果

The evaluation results are shown in Table 12. We observe that providing descriptions in context can significantly improve the accuracy of image classification. The consistent improvements indicate that KOSMOS-1 can perceive the intentions of instructions and well align the concepts in language modality with visual features in vision modality.
评估结果如表 12 所示。我们观察到，在上下文中提供描述可以显着提高图像分类的准确性。一致的改进表明 kosmos-1 可以感知指令的意图，并将语言模式中的概念与视觉模式中的视觉特征进行良好的对齐。

Settings 设置	Accuracy 准确性
Without Descriptions 无需描述	61.7
With Descriptions 附有描述	90.0

Table 12: Results of zero-shot image classification without and with verbal descriptions.
表 12: 无文字描述和有文字描述的零镜头图像分类结果。

4.8 Language Tasks
4.8 语文任务

The models are evaluated on the language tasks given task instructions (i.e., zero-shot) or several demonstration examples (i.e., few-shot). Text inputs are directly fed into the models as in vanilla language models.
这些模型是根据给定的任务指令(即零镜头)或几个示例(即少镜头)对语言任务进行评估的。文本输入像普通语言模型一样直接输入到模型中。

4.8.1 Evaluation Setup
4.8.1 评估设置

We train a language model (LLM) baseline with the same text corpora and training setup. We evaluate Kosmos-1 and the LLM baseline on eight language tasks, including cloze and completion tasks (i.e, StoryCloze, HellaSwag), Winograd-style tasks (i.e, Winograd, Winogrande), commonsense reasoning (i.e, PIQA), and three datasets BoolQ, CB, and COPA from the SuperGLUE benchmark [WPN+19]. The detailed descriptions of these datasets are provided in Appendix C.2. We conduct experiments under zero-shot and few-shot settings. We evaluate each test example by randomly sampling examples from the training set as demonstrations. We set the number of shots to 0, 1, and 4 in our experiments.

我们使用相同的文本语料库和训练设置来训练语言模型(LLM)基线。我们评估 kosmos-1 和 LLM 基线的八个语言任务，包括完形和完成任务(即 StoryCloze，HellaSwag)，Winograd 风格的任务(即 Winograd，Winogrande)，常识推理(即 PIQA)，以及三个数据集 BoolQ，CB 和 COPA 从 SuperGLUE 基准[WPN + 19]。这些数据集的详细描述在附录 c. 2 中提供。我们在零拍和少拍设置下进行实验。我们通过从训练集中随机抽取例子作为演示来评估每个测试例子。在我们的实验中，我们将拍摄次数设置为 0,1 和 4。

4.8.2 Results
4.8.2 结果

Table 13 presents the in-context learning performance of language tasks. Kosmos-1 achieves comparable or even better performance in cloze completion and commonsense reasoning tasks when compared to LLM. In terms of the average result across all these datasets, LLM performs better in zero-shot and one-shot settings, whereas our model performs better in few-shot (k = 4) settings. The results indicate that Kosmos-1 also handles language-only tasks well and achieves favorable performance across datasets. In addition, Section 4.9.2 shows that MLLMs learn better visual commonsense knowledge compared with LLMs.

表 13 显示了语言任务在上下文中的学习表现。与 LLM 相比，kosmos-1 在完形填空和常识推理任务方面取得了可比的甚至更好的表现。就所有这些数据集的平均结果而言，LLM 在零镜头和一镜头设置中表现得更好，而我们的模型在少镜头(k = 4)设置中表现得更好。结果表明 kosmos-1 也能很好地处理只有语言的任务，并在数据集中获得良好的性能。此外，第 4.9.2 节显示，与 LLMs 相比，MLLMs 学习更好的视觉常识知识。

Task 任务	Zero-shot 一枪毙命		One-shot 一枪毙命		Few-shot (k = 4) 少量射击(k = 4)	
	LLM 法学硕 士	KOSMOS-1	LLM	KOSMOS-1	LLM	KOSMOS-
		宇宙 -1 LLM KOSMOS-1 LLM KOSMOS-1				
StoryCloze StoryCloze 故事完结	72.9	72.1	72.9	72.2	73.1	72.3
HellaSwag 地狱怪谈	50.4	50.0	50.2	50.0	50.4	50.3
Winograd Winograd 温诺格拉德	71.6	69.8	71.2	68.4	70.9	69.8
Winogrande Winogrande	56.7	54.8	56.7	54.5	57.0	55.7
PIQA PIQA	73.2	72.9	73.0	72.5	72.6	72.3
BoolQ 布尔克	56.4	56.4	55.1	57.2	58.7	59.2
CB CB	39.3	44.6	41.1	48.2	42.9	53.6
COPA COPA	68.0	63.0	69.0	64.0	69.0	64.0
Average	61.1	60.5	61.2	60.9	61.8	62.2

平均

Table 13: Performance comparisons of language tasks between Kosmos-1 and LLM. We use the same textual data and training setup to reimplement a language model. Both models do not use instruction tuning for fair comparisons.
表 13: kosmos-1 和 LLM 之间语言任务的性能比较。我们使用相同的文本数据和训练设置来重新实现语言模型。两种模型都没有使用指令调整来进行公平的比较。

4.9 Cross-modal Transfer
4.9 跨模式转移

Cross-modal transferability allows a model to learn from one modality (such as text, image, audio, etc.) and transfer the knowledge to the other modalities. This skill can enable a model to perform various tasks across different modalities. In this part, we evaluate the cross-model transferability of Kosmos-1 on several benchmarks.
跨模态可迁移性允许模型从一种模式(如文本、图像、音频等)中学习，并将知识转移到其他模式中。这种技能可以使模型在不同的模式下执行各种任务。在这一部分中，我们根据几个基准评估 kosmos-1 的跨模型可转移性。

4.9.1 Transfer from Language to Multimodal: Language-Only Instruction Tuning
4.9.1 从语言到多模态的转换: 只有语言的指令调整

To evaluate the effect of language-only instruction tuning, we conduct an ablation study using four datasets: COCO, Flickr30k, VQAv2, and VizWiz. These datasets consist of image captioning and visual questions anwsering. The evaluation metrics are: CIDEr scores for COCO/Flickr30k and VQA accuracy for VQAv2/VizWiz.
为了评估语言教学调整的效果，我们使用四个数据集进行消融研究: COCO，Flickr30k，vqav2 和 VizWiz。这些数据集由图像标题和视觉问题回答组成。评估指标是: COCO/Flickr30k 的苹果酒评分和 VQAv2/VizWiz 的 VQA 准确性。

Table 14 shows the experimental results. Language-only instruction tuning boosts our model’s performance by 1.9 points on Flickr30k, 4.3 points on VQAv2, and 1.3 points on VizWiz. Our experi-ments show that language-only instruction tuning can significantly improve the model’s instruction-following capabilities across modalities. The results also indicate that our model can transfer the instruction-following capability from language to other modalities.
表 14 显示了实验结果。只使用语言的指令调优将我们模型的性能在 Flickr30k 上提高了 1.9 分，在 vqav2 上提高了 4.3 分，在 VizWiz 上提高了 1.3 分。我们的实验表明，只使用语言的指令调优可以显著提高模型的指令跟随能力。结果还表明，我们的模型可以将指令跟随能力从语言转移到其他模式。

Model 模特	COCO COCO	Flickr30k Flickr30k	VQAv2 VQAv2	VizWiz VizWiz
KOSMOS-1 宇宙 1 号	84.7	67.1	51.0	29.2
w/o language-only instruction tuning W/o 语言指令调优	87.6	65.2	46.7	27.9

Table 14: Ablation study on language-only instruction tuning. We report CIDEr scores for COCO and Flickr30k, and VQA accuracy scores for VQAv2 and VizWiz.
表 14: 仅语言教学调整的消融研究。我们报告 COCO 和 Flickr30k 的苹果酒评分，vqav2 和 VizWiz 的 VQA 准确性评分。

4.9.2 Transfer from Multimodal to Language: Visual Commonsense Reasoning
4.9.2 从多模式转换到语言: 视觉常识推理

Visual commonsense reasoning tasks require an understanding of the properties of everyday objects in the real world, such as color, size, and shape. These tasks are challenging for language models because they may require more information about object properties than what is available in texts. To investigate the visual commonsense capabilities, we compare the zero-shot performance of Kosmos-1 and LLM on visual commonsense reasoning tasks.
视觉常识推理任务需要理解现实世界中日常物体的特性，如颜色、大小和形状。这些任务对于语言模型来说是具有挑战性的，因为它们可能需要比文本更多的关于对象属性的信息。为了研究视觉常识能力，我们比较了 kosmos-1 和 LLM 在视觉常识推理任务中的零镜头表现。

Evaluation Setup We compare Kosmos-1 and the LLM baseline on three object commonsense reasoning datasets, RelativeSize [BHCF16], MemoryColor [NHJ21] and ColorTerms [BBBT12] datasets. Table 15 shows some examples of object size and color reasoning tasks. RelativeSize contains 486 object pairs from 41 physical objects. The model is required to predict the size relation between two objects in a binary question-answering format with “Yes”/“No” answers. MemoryColor and ColorTerms require the model to predict the color of objects from a set of 11 color labels in a multiple-choice format. We use only text as our input and do not include any images. We measure the accuracy of our model on these three datasets.
我们比较 kosmos-1 和 LLM 基线在三个对象共义推理数据集，RELATIVESIZE [BHCF16]，MEMORYCOLOR [NHJ21]和 COL-ORTERMS [BBBT12]数据集。表 15 显示了对象大小和颜色推理任务的一些示例。RELATIVESIZE 包含来自 41 个物理对象的 486 对对象。该模型需要预测两个对象之间的大小关系在一个二进制问题-回答格式与“是”/“否”的答案。MEMORYCOLOR 和 COLORTERMS 要求模型以多项选择的形式从一组 11 种颜色标签中预测物体的颜色。我们只使用文本作为输入，不包括任何图像。我们在这三个数据集上测量我们模型的准确性。

Task 任务	Example Prompt 示例提示	Object / Pair 对象/对	Answer 答案

Object Size Reasoning
物体大小推理
	Is {Item1} larger than {Item2}? {Answer}
	{ Item1}比{ Item2}大吗?
	{ Answer }
		(sofa, cat)
		(沙发，猫)
			Yes
			是的

Object Color Reasoning
物体颜色推理
	The color of {Object} is? {Answer}
	{ Object }的颜色是? { Answer }
		the sky
		天空
			blue
			蓝色

Table 15: Evaluation examples of object size and color reasoning.
表 15: 物体大小和颜色推理的评估示例。

Results Table 16 presents the zero-shot performance of Kosmos-1 and LLM on visual commonsense reasoning tasks. Kosmos-1 significantly outperforms LLM by 1.5% on RelativeSize, 14.7% on MemoryColor, and 9.7% on ColorTerms dataset. The consistent improvements indicate that Kosmos-1 benefits from the visual knowledge to complete the corresponding visual commonsense reasoning. The reason for Kosmos-1's superior performance is that it has modality transferability, which enables the model to transfer visual knowledge to language tasks. On the contrary, LLM has to rely on textual knowledge and clues to answer visual commonsense questions, which limits its ability to reason about object properties.
结果表 16 展示了 kosmos-1 和 LLM 在视觉常识推理任务中的零镜头表现。Kosmos-1 在 RELATIVESIZE 上显著优于 LLM 1.5% ，MEMORYCOLOR 上显著优于 LLM 14.7% ，COLORTERMS 数据集上显著优于 LLM 9.7% 。一致的改进表明 kosmos-1 受益于视觉知识来完成相应的视觉常识推理。Kosmos-1 的优越性能的原因是它具有模态可迁移性，这使得模型能够将视觉知识转移到语言任务中。相反，LLM 必须依靠文本知识和线索来回答视觉常识问题，这限制了它对对象属性的推理能力。

Model 模特	Size Reasoning 大小推理	Color Reasoning 颜色推理	
	RELATIVESIZE 相对大小	MEMORYCOLOR 记忆的颜色	COLORTERMS 色彩术语
Using retrieved images 使用检索到的图像			
VALM [WDC+23] VALM [WDC + 23]	85.0	58.6	52.7
Language-only zero-shot evaluation 只有语言的零镜头评估			
LLM 法学硕士	92.7	61.4	63.4
KOSMOS-1 宇宙 1 号	94.2	76.1	73.1

Table 16: Zero-shot visual commonsense reasoning on RelativeSize, MemoryColor, and ColorTerms datasets. Accuracy scores are reported.
表 16: 对 RELATIVESIZE，MEMORYCOLOR 和 COLORTERMS 数据集的零镜头视觉常识推理。精确度分数报告。

5 Conclusion
5 结论

In this work, we introduce Kosmos-1, a multimodal large language model that can perceive general modalities, follow instructions, and perform in-context learning. The models trained on web-scale
在这项工作中，我们介绍了 KOSMOS-1，一个多模态大语言模型，可以感知一般模式，遵循指令，并执行在上下文学习。在网络规模上训练的模型

multimodal corpora achieve promising results across a wide range of language tasks and multimodal tasks. We show that going from LLMs to MLLMs enables new capabilities and opportunities. In the future, we would like to scale up Kosmos-1 in terms of model size [MWH+22, WMH+22, CDH+22], and integrate the speech [WCW+23] capability into Kosmos-1. In addition, Kosmos-1 can be used as a unified interface for multimodal learning, e.g., enabling using instructions and examples to control text-to-image generation.
多模态语料库在广泛的语言任务和多模态任务中取得了令人满意的结果。我们展示了从 LLMs 到 MLLMs 能够带来新的能力和机会。在未来，我们希望按照模型尺寸[MWH + 22，WMH + 22，CDH + 22]扩大 KOSMOS-1，并将语音[WCW + 23]能力整合到 kosmos-1 中。此外，kosmos-1 还可以作为多模式学习的统一接口，例如，使用指令和示例来控制文本到图像的生成。

References
参考文献

[ADL+22] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Mar-ianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Advances in Neural Information Processing Systems, 2022.

[ADL + 22] Jean-Baptiste Alayrac Jeff Donahue Pauline Luc Antoine Miech ian Barr Yana Hasson Karel Lenc Arthur Mensch Katherine Millican Malcolm Reynolds Roman Ring Eliza Rutherford Serkan Cabi Tengda Han Zhitao Gong sina Samangooei Mar-ianne Monteiro Jacob Menick Sebastian Borgeaud Andrew Brock Aida Nematzadeh Sahand Sharifzadeh Mikolaj Binkowski Ricardo Barreira Oriol Vinyals Andrew Zisserman 和 Karen Simonyan。Flamingo: 少镜头学习的视觉语言模型。在神经信息处理系统的进步，2022 年。

[AFJG16] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In ECCV, pages 382–398, 2016.

[AFJG16]彼得 · 安德森，巴苏拉 · 费尔南多，马克 · 约翰逊和斯蒂芬 · 古尔德。Spice: 语义命题图像标题评估。在 ECCV，页 382-398,2016。

[AHR+22] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Na-man Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. CM3: A causal masked multimodal model of the Internet. ArXiv, abs/2201.07520, 2022.

[AHR + 22] Armen aghajyan，Bernie Huang，Candace Ross，Vladimir Karpukhin，Hu Xu，Na-man Goyal，Dmytro Okhonko，Mandar Joshi，Gargi Ghosh，Mike Lewis，and Luke Zettlemoyer.CM3: 互联网的因果屏蔽多模态模型。ArXiv，abs/2201.07520,2022.

[BBBT12] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. Distributional semantics in technicolor. In ACL, 2012.

[BBBT12] Elia Bruni，Gemma Boleda，Marco Baroni，and Nam Khanh Tran.

[BHCF16] Hessam Bagherinezhad, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. Are ele-phants bigger than butterflies? reasoning about sizes of objects. ArXiv, abs/1602.00753, 2016.

[BHCF16] Hessam Bagherinezhad，Hannaneh Hajishirzi，Yejin Choi 和 Ali Farhadi。大象比蝴蝶大吗？对物体大小的推理。ArXiv，abs/1602.00753,2016.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sand-hini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[BMR + 20]汤姆 · 布朗，本杰明 · 曼，尼克 · 赖德，梅勒妮 · 苏比亚，贾里德 · d · 卡普兰，普拉弗拉 · 达里瓦尔，阿文德 · 尼拉坎坦，普拉纳夫 · 希亚姆,阿里尔 · 赫伯特-沃斯，格雷

琴 · 克鲁格，汤姆 · 亨尼汉，Rewon Child，Aditya Ramesh，Daniel Ziegler，Jeffrey Wu，Clemens Winter，Chris Hesse，Mark Chen，Eric Sigler，Mateusz Litwin，Scott Gray，Benjamin Chess，Jack Clark，Christopher Berner，Sam McCandlish，Alec Radford，Ilya Sutskever，and Dario amodel。语言模型是少数学习者。在神经信息处理系统的进展，卷 33，页 1877-1901。Curran Associates，inc。，2020。

[BPK+22] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.

[BPK + 22] Minwoo Byeon，Beomhee Park，Haecheon Kim，sung-jun Lee，Woonhyuk Baek，and Saehoon Kim. coyo-700 米: 图像-文本对数据集，2022。

[BZB+20] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.

[BZB + 20] Yonatan Bisk，Rowan Zellers，Ronan Le Bras，Jianfeng Gao，Yejin Choi. 皮卡: 关于自然语言中身体常识的推理。在 2020 年第 34 届 AAAI 人工智能会议上。

[CDH+22] Zewen Chi, Li Dong, Shaohan Huang, Damai Dai, Shuming Ma, Barun Patra, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. On the representation collapse of sparse mixture of experts. In Advances in Neural Information Processing Systems, 2022.

[CDH + 22]泽文池，李东，黄绍汉，戴大美，马淑明，巴伦 · 帕特拉，萨克森 · 辛格尔，巴亚尔 · 巴贾杰，夏松，毛贤灵，黄何燕和魏夫鲁。关于稀疏混合专家的代表性崩溃。In Advances In Neural Information Processing Systems，2022 在神经信息处理系统的进展，2022。

[CJS90] Patricia A Carpenter, Marcel A Just, and Peter Shell. What one intelligence test measures: a theoretical account of the processing in the raven progressive matrices test. Psychological review, 97(3):404, 1990.

[CJS90]帕特里夏是木匠，马塞尔是公正，还有彼得 · 谢尔。一个智力测试测量的内容: 乌鸦级进矩阵测试中处理过程的理论解释。心理学评论，97(3) : 404,1990。

[CLC+19] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 2924–2936, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Christopher Clark，Kenton Lee，Ming-Wei Chang，Tom Kwiatkowski，Michael Collins，and Kristina Toutanova.BoolQ: 探索自然的是/否问题令人惊讶的困难。计算机语言学协会北美分会 2019 年会议记录: 人类语言技术，第一卷(长篇和短篇)，第 2924-2936 页，明尼苏达州明尼阿波利斯，2019 年 6 月。计算机语言学协会。

[CND+22] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. ArXiv, abs/2204.02311, 2022.

[CND + 22] Aakanksha Chowdhery Sharan Narang Jacob Devlin Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi Sasha Tsvyashchenko Joshua Maynez,abhishek b Rao Parker Barnes Yi Tay Noam m. Shazeer Vinodkumar Prabhakaran Emily Reif Nan Du Benton c. Hutchinson Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari peng cheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev,henryk Michalewski Xavier García Vedant Misra Kevin Robinson Liam Fedus Denny Zhou Daphne Ippolito David Luan Hyeontaek Lim Barret Zoph Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick Andrew m. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz Erica Oliveira Moreira Rewon Child,oleksandr Polozov Katherine Lee 周宗伟，王雪志 Brennan Saeta Mark Díaz Orhan Firat Michele Catasta Jason Wei Kathleen s. Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov 和 Noah Fiedel。PaLM: 通过路径缩放语言建模。ArXiv，abs/2204.02311,2022.

[CSDS21] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3558–3568, 2021.

[CSDS21] Soravit Changpinyo Piyush Sharma Nan Ding 还有 Radu Soricut。概念 12m: 推动网络规模的图像-文本预训练，以识别长尾视觉概念。在 IEEE/CVF 计算机视觉和模式识别会议记录中，页 3558-3568,2021。

[CZC+21] Xingyu Chen, Zihan Zhao, Lu Chen, JiaBao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. WebSRC: A dataset for web-based structural reading comprehension. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 4173–4185, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[CZC + 21]陈兴宇、赵子涵、陈陆、纪家宝、张丹阳、奥罗、熊宇轩、凯宇。WebSRC: 基于网络的结构化阅读理解数据集。2021 年自然语言处理经验方法会议记录，4173-4185 页，Online and Punta Cana，多米尼加共和国，2021 年 11 月。计算机语言学协会。

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pages 248–255. IEEE Computer Society, 2009.

[DDS + 09]贾登，魏东，理查德・索切尔，李佳丽，凯丽，李飞飞。Imagenet: 大型分层图像数据库。2009 年 ieee 计算机协会计算机视觉和模式识别会议(CVPR 2009)，2009 年 6 月 20 日至 25 日，美国佛罗里达州迈阿密，页 248-255。Ieee 计算机协会，2009。

[dMST19] Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. The Commitment-Bank: Investigating projection in naturally occurring discourse. Proceedings of Sinn und Bedeutung, 23(2):107–124, Jul. 2019.

玛丽·凯瑟琳·德·马尔内夫，曼迪·西蒙斯，朱迪斯·唐豪瑟。承诺银行: 在自然发生的话语中调查投影。Proceedings of Sinn und Bedeutung，23(2) : 107-124,2019 年 7 月。

[GBB+20] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020.

[GBB + 20] Leo Gao Stella Biderman Sid Black Laurence Golding Travis Hoppe Charles Foster Jason Phang Horace He Anish Thite Noa Nabeshima et al.这一堆: 一个 800gb 的用于语言建模的多种文本数据集。arXiv 预印本 arXiv: 2101.00027,2020。

[GKSS+17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In CVPR, pages 6325–6334, 2017.

[GKSS + 17] Yash Goyal，Tejas Khot，Douglas Summers-Stay，Dhruv Batra，和 Devi Parikh。让 vqa 中的 v 发挥作用: 提升图像理解在视觉问题回答中的作用。在 CVPR 中，页 6325-6334,2017。

[GLS+18] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3608–3617, 2018.

[GLS + 18] Danna Gurari，Qing Li，abigail j Stangl，Anhong Guo，Chi Lin，Kristen Grauman，Jiebo Luo，and Jeffrey p Bigham.Vizwiz 大挑战: 回答盲人的视觉问题。在 IEEE 关于计算机视觉和模式识别的会议记录中，页 3608-3617,2018。

[HG16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). arXiv preprint arXiv:1606.08415, 2016.

[HG16] Dan Hendrycks 和 Kevin Gimpel。高斯误差线性单位(GELUs)。 arXiv 预印 arXiv: 1606.08415,2016。

[HSD+22] Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shum-ing Ma, and Furu Wei. Language models are general-purpose interfaces. ArXiv, abs/2206.06336, 2022.

[HSD + 22]雅如浩，宋浩宇，李东，黄少汉，泽文池，王文辉，马沈明，和夫鲁伟。语言模型是通用接口。ArXiv，abs/2206.06336,2022.

[HSLS22] Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions:
霍诺维奇，托马斯·夏洛姆，奥马尔·列维和蒂莫·希克。不自然的指示:
Tuning language models with (almost) no human labor, 2022.
调优语言模型(几乎)没有人工，2022 年。

[JR03] John and Jean Raven. Raven Progressive Matrices, pages 223–237. Springer US, Boston, MA, 2003.

[JR03]约翰和让·瑞文. 瑞文进步矩阵，页 223-237. 斯普林格美国，波士顿，马萨诸塞州，2003。

[KFF17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(4):664–676, 2017.

安德烈 · 卡帕西和李菲菲。用于生成图像描述的深度视觉-语义对齐。IEEE 模式分析与机器智能交易，39(4) : 664-676,2017。

[KFM+20] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In Advances in Neural Information Processing Systems, volume 33, pages 2611–2624, 2020.

Douwe Kiela Hamed Firooz Aravind Mohan Vedanuj Goswami Amanpreet Singh Pratik Ringshia 和 Davide Testuggine。仇恨模因挑战: 在多模态模因中检测仇恨言论。In Advances In Neural Information Processing Systems，volume 33，pages 2611-2624,2020 神经信息处理系统进展，第 33 卷，2611-2624 页，2020 年。

[KR18] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In EMNLP, pages 66–71, 2018.

[KR18] Taku Kudo 和 John Richardson。句子: 一个简单和语言独立的子词标记器和解标记器，用于神经文本处理。在 EMNLP，第 66-71 页，2018 年。

[KSF23] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal generation. arXiv preprint arXiv:2301.13823, 2023.

[KSF23]金玉谷，鲁斯兰・萨拉克胡蒂诺夫，还有丹尼尔・弗莱德。用于多模式生成的图像的基础语言模型。arXiv 预印 arXiv: 2301.13823,2023。

[LDM12a] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema chal-lenge. In Thirteenth International Conference on the Principles of Knowledge Repre-sentation and Reasoning, 2012.

赫克托 · 莱维斯克，欧内斯特 · 戴维斯和莱奥拉 · 摩根斯坦。Winograd schema challenge.在 2012 年第十三届国际知识原则会议上，代表和推理。

[LDM12b] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In Principles of Knowledge Representation and Reasoning, 2012.

[LDM12b] Hector j. Levesque，Ernest Davis，和 Leora Morgenstern。Winograd 模式挑战。知识表示和推理原理，2012。

[LHV+23] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. arXiv preprint arXiv:2301.13688, 2023.

[LHV + 23] Shayne Longpre，Le Hou，Tu Vu，Albert Webson，Hyung Won Chung，Yi Tay，Denny Zhou，Quoc v Le，Barret Zoph，Jason Wei 等。果馅饼收藏: 为有效的指令调整设计数据和方法。arXiv 预印 arXiv: 2301.13688,2023。

[LLSH23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ArXiv, abs/2301.12597, 2023.

李俊南，李东旭，西尔维奥・萨瓦雷斯，还有史蒂文・海。BLIP-2: 使用冻结图像编码器和大型语言模型进行引导式语言图像预训练。ArXiv，abs/2301.12597,2023.

[LMB+14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755, 2014.

[LMB + 14]林宗怡，迈克尔 · 迈尔，塞尔吉 · 贝隆吉，詹姆斯 · 海斯，彼得罗 · 佩罗纳，蒂法 · 拉曼南，皮奥特 · 多拉，和 c · 劳伦斯 · 兹特尼克。Microsoft coco: 上下文中的常见对象。ECCV，页 740-755,2014。

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[LOG + 19]刘音涵，迈尔 · 奥特，纳曼 · 戈亚尔，杜静飞，文华 · 乔希，丹奇 · 陈，奥马尔 · 利维，迈克 · 刘易斯，卢克 · 泽特尔莫耶和韦塞林 · 斯托亚诺夫。RoBERTa: 一个强大的优化 bert 预训练方法。arXiv 预印 arXiv: 1907.11692,2019。

[MRL+17] Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. In Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 46–51, 2017.

[MRL + 17] Nasrin Mostafazadeh，Michael Roth，Annie Louis，Nathanael Chambers，and James Allen.Lsdsem 2017 分享任务: 故事完形填空测试。在关于词汇，句子和话语层次语义的链接模型的第二次研讨会的进程中，页 46-51,2017。

[MWH+22] Shuming Ma, Hongyu Wang, Shaohan Huang, Wenhui Wang, Zewen Chi, Li Dong, Alon Benhaim, Barun Patra, Vishrav Chaudhary, Xia Song, and Furu Wei. TorchScale: Transformers at scale. CoRR, abs/2211.13184, 2022.

[MWH + 22]马淑明，王红玉，黄绍汉，王文辉，泽文池，李东，阿龙本海姆，巴伦帕特拉，维什拉夫乔杜里，夏松，和夫鲁威。火炬尺度: 变形金刚尺度。CoRR，abs/2211.13184,2022.

[NHJ21] Tobias Norlund, Lovisa Hagström, and Richard Johansson. Transferring knowl-edge from vision to language: How to achieve it and how to measure it? ArXiv, abs/2109.11321, 2021.

托比亚斯 · 诺伦德，洛维萨 · 哈格斯特伦和理查德 · 约翰逊。将知识边缘从视觉转化为语言: 如何实现它，如何衡量它？ArXiv，abs/2109.11321,2021.

[RBG11] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plau-sible alternatives: An evaluation of commonsense causal reasoning. In AAAI Spring Symposium, 2011.

Melissa Roemmele，Cosmin Adrian Bejan，and Andrew s. Gordon.选择合理的替代方案: 常识性因果推理的评估。在 2011 年 AAAI 春季研讨会上。

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learn-ing transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[RKH + 21] Alec Radford Jong Wook Kim Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark et al. 从自然语言监督中学习可转换的视觉模型。在国际机器学习会议上，页 8748-8763。PMLR，2021.

[RPJ+20] Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P.

[RPJ + 20] Jack w. Rae Anna Potapenko Siddhant m. Jayakumar Chloe Hillier 和 Timothy p。

Lillicrap. Compressive transformers for long-range sequence modelling. In ICLR, 2020.

Lilicrap。用于长程序列建模的压缩变压器。在 ICLR，2020 中。

[SBBC20] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande:
坂口圭介，罗南 · 勒 · 布拉斯，钱德拉 · 巴格瓦图拉，叶金 · 蔡:
An adversarial winograd schema challenge at scale. In AAAI, pages 8732–8740, 2020.
在 AAAI 中，第 8732-8740 页，2020。

[SBV+22] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wight-man, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. arXiv preprint arXiv:2210.08402, 2022.

[SBV + 22]克里斯托弗 · 舒赫曼，罗曼 · 博蒙特，理查德 · 文库，卡德 · 戈登，罗斯 · 怀特曼，迈赫迪 · 切尔蒂，西奥 · 库姆斯，阿鲁什 · 卡塔，克莱顿 · 穆利斯，米切尔 · 沃斯曼等。Laion-5b: 一个开放的大规模数据集，用于训练下一代图像文本模型。arXiv 预印 arXiv: 2210.08402,2022。

[SDGS18] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Pro-ceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 2556–2565. Association for Computational Linguistics, 2018.

[SDGS18] Piyush Sharma Nan Ding Sebastian Goodman 还有 Radu Soricut。概念标题: 一个干净的，hypernymed，图像 alt-text 数据集，用于自动图像标题。计算机语言学协会 2018 年第 56 届年会，澳大利亚墨尔本，2018 年 7 月 15-20 日，第 1 卷: 长篇论文，2556-2565 页。计算机语言学协会，2018。

[SDP+22] Yutao Sun, Li Dong, Barun Patra, Shuming Ma, Shaohan Huang, Alon Benhaim, Vishrav Chaudhary, Xia Song, and Furu Wei. A length-extrapolatable transformer. arXiv preprint arXiv:2212.10554, 2022.

[SDP + 22]孙玉涛、李东、巴伦 · 帕特拉、马淑明、黄绍汉、阿隆 · 本海姆、维什拉夫 · 乔杜里、夏松和魏夫鲁。长度可外推变压器。arXiv 预印 arXiv: 2212.10554,2022。

[SPN+22] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhan-dari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mo-hammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, a large-scale generative language model, 2022.

[SPN + 22] Shaden Smith，Mostofa Patwary，Brandon Norick，Patrick LeGresley，Samyam Rajbhan-dari，Jared Casper，zhuliu，Shrimai Prabhumoye，George Zerveas，Vijay Korthikanti，Elton Zhang，Rewon Child，Reza Yazdani Aminabadi，Julie Bernauer，Xia Song，Mo-hammad Shoeybi，Yuxiong He，Michael Houston，Saurabh Tiwary，and Bryan Catanzaro.利用深速和威震天来训练威震天图灵 NLG 530B，一个大规模的生成语言模型，2022 年。

[SPP+19] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.

Mohammad Shoeybi，Mostofa Patwary，Raul Puri，Patrick LeGresley，Jared Casper，and Bryan Catanzaro.Megatron-lm: 使用模型并行性训练数十亿参数语言模型。arXiv 预印 arXiv: 1909.08053,2019。

[SPW+13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[SPW + 13] Richard Socher Alex Perelygin Jean Wu Jason Chuang Christopher d. Manning Andrew Ng 和 Christopher Potts。情感树库上语义组合性的递归深度模型。2013 年自然语言处理经验方法会议记录，1631-1642 页，西雅图，华盛顿，美国，2013 年 10 月。计算机语言学协会。

[SVB+21] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clay-ton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114, 2021.

[SVB + 21]克里斯托弗・舒曼，理查德・文库，罗曼・博蒙特，罗伯特・卡茨马奇克，克莱顿・穆利斯，阿鲁什・卡塔，西奥・库姆斯，耶尼亚・吉特瑟夫，和小松崎。Laion-400m: 4 亿个经过剪辑过滤的图像文本对的开放数据集。arXiv 预印 arXiv: 2111.02114,2021。

[TMC$^+$21] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In Neural Information Processing Systems, 2021.

[TMC + 21] Maria Tsimpoukelli Jacob Menick Serkan Cabi S.m. Ali Eslami Oriol Vinyals 和 Felix Hill。使用冻结语言模型的多模式少镜头学习。在神经信息处理系统，2021 年。

[VLZP15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In CVPR, pages 4566–4575, 2015.

[VLZP15] Ramakrishna Vedantam，c Lawrence Zitnick 和 Devi Parikh。苹果酒: 基于共识的图像描述评估。在 CVPR 中，页 4566-4575,2015。

[WBD$^+$22] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. ArXiv, abs/2208.10442, 2022.

[WBD + 22]王文辉，包航波，李东，约翰・比约克，彭智良，刘强，克里蒂・阿加瓦尔，奥瓦・穆罕默德，萨克森・辛格哈尔，苏霍吉特・索姆和夫鲁威。作为外语的形象: 所有视觉和视觉语言任务的贝特预训。ArXiv，abs/2208.10442,2022.

[WBW$^+$11] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie.

[WBW + 11] Catherine Wah，Steve Branson，Peter Welinder，Pietro Perona，and Serge j.

The caltech-ucsd birds-200-2011 dataset. 2011.

加州理工学院-加州大学圣地亚哥分校鸟类 -200-2011 数据集。

[WCW$^+$23] Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. ArXiv, abs/2301.02111, 2023.

[WCW + 23]王成义，陈三元，于武，张子华，龙洲，刘淑杰，卓晨，刘燕青，王华明，李金玉，李雷和，赵盛和，魏夫鲁。神经编解码语言模型是从零发射到语音合成器的文本。ArXiv，abs/2301.02111,2023.

[WDC$^+$23] Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jian-feng Gao, and Furu Wei. Visually-augmented language modeling. In International Conference on Learning Representations, 2023.

[WDC + 23]王维志、李东、郝诚、宋浩宇、刘晓东、燕西凤、高剑峰和魏夫鲁。视觉增强语言建模。In International Conference on Learning representation，20232023 年国际学习表征会议。

[WMD^{+}22] Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. DeepNet: Scaling Transformers to 1,000 layers. CoRR, abs/2203.00555, 2022.
[大规模杀伤性武器 + 22]王洪宇、马淑明、李东、黄绍汉、张东东和魏夫鲁。深网: 将变压器缩放到 1000 层。CoRR，abs/2203.00555,2022.

[WMH^{+}22] Hongyu Wang, Shuming Ma, Shaohan Huang, Li Dong, Wenhui Wang, Zhiliang Peng, Yu Wu, Payal Bajaj, Saksham Singhal, Alon Benhaim, Barun Patra, Zhun Liu, Vishrav Chaudhary, Xia Song, and Furu Wei. Foundation transformers. CoRR, abs/2210.06423, 2022.
[WMH + 22]王宏宇、马淑明、黄绍汉、李东、王文辉、彭志良、吴宇武、巴亚尔・巴贾吉、萨克森・辛格尔、阿隆・本海姆、巴伦・帕特拉、刘准、维什拉夫・乔杜里、夏松和夫鲁威。基础变压器。CoRR，abs/2210.06423,2022.

[WPN^{+}19] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. arXiv preprint arXiv:1905.00537, 2019.
[WPN + 19] Alex Wang Yada Pruksachatkun Nikita Nangia Amanpreet Singh Julian Michael Felix Hill Omer Levy 和 Samuel r Bowman。SuperGLUE: 通用语言理解系统的粘性基准。arXiv 预印本 arXiv: 1905.00537,2019。

[WWS^{+}22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903, 2022.
[WWS + 22] Jason Wei，Xuezhi Wang，Dale schurmans，Maarten Bosma，Ed Chi，Quoc Le，and Denny Zhou.思维链提示在大型语言模型中引发推理。arXiv preprint arXiv: 2201.11903,2022.

[WYH^{+}22] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. CoRR, abs/2205.14100, 2022.
[WYH + 22]王建峰，杨正元，胡晓伟，李林杰，林凯文，哲干，刘子成，刘策，王丽娟。GIT: 视觉和语言的图像到文本生成转换器。CoRR，abs/2205.14100,2022.

[YAS^{+}22] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Retrieval-augmented multimodal language modeling. ArXiv, abs/2211.12561, 2022.
[Yasunaga，Armen aghajyan，Weijia Shi，Rich James，Jure Leskovec，Percy Liang，Mike Lewis，Luke Zettlemoyer，and Wen tau Yih.检索-增强的多模态语言建模。ArXiv，abs/2211.12561,2022.

[YLHH14] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descrip-tions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2:67–78, 2014.
Peter Young Alice Lai Micah Hodosh 和 Julia Hockenmaier。从图像描述到视觉表示: 对事件描述进行语义推断的新的相似度量。TACL，2:67-78,2014.

[ZHB^{+}19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.
[ZHB + 19] Rowan Zellers，Ari Holtzman，Yonatan Bisk，Ali Farhadi，and Yejin Choi. 机器真的能把你的话说完吗？在 2019 年计算机语言学协会第 57 届年会的会议记录中。

22

A Hyperparameters
超参数

A.1 Training
答 1 训练

We report the detailed model hyperparameter settings of Kosmos-1 in Table 17 and training hyperparameters in Table 18.
我们在表 17 中报告了 kosmos-1 的详细模型超参数设置，在表 18 中报告了训练超参数。

Hyperparameters 超参数	
Number of layers	
图层数	24
Hidden size	2,048
隐藏尺寸	2,048
FFN inner hidden size	8,192
内部隐藏大小	8,192
Attention heads	
注意了	32
Dropout	
辍学	0.1
Attention dropout	
退学注意	0.1
Activation function	GeLU [HG16]
激活函数	格鲁[HG16]
Vocabulary size	64,007
词汇量	六万四千零七
Soft tokens V size	
软令牌 v 大小	64
Max length	2,048
最大长度	2,048
Relative position embedding	xPos [SDP+22]
相对位置嵌入	xPos [SDP + 22]
Initialization	Magneto [WMH+22]
初始化	万磁王[WMH + 22]

Table 17: Hyperparameters of causal language model of Kosmos-1
表 17: kosmos-1 因果语言模型的超参数

Hyperparameters 超参数	
Training steps	300,000
训练步骤	300,000
Warmup steps	
热身步骤	375
Batch size of text corpora	
文本语料库的批量	256
Max length of text corpora	2,048
文本语料库的最大长度	2,048
Batch size of image-caption pairs	6,144
图片标题对的批量大小	6,144
Batch size of interleaved data	
交织数据的批量大小	128
Optimizer	Adam
优化器	亚当

Learning rate 学习率	2e-4 第二季第四集
Learning Rate Decay Learning Rate Decay 学习速度下降	Linear 线性的
Adam 亚当	1e-6 1 e-6
Adam 亚当	(0.9, 0.98) (0.9,0.98)
Weight decay 体重下降	0.01

Table 18: Training hyperparameters of Kosmos-1
表 18: kosmos-1 的训练超参数

A.2 Language-Only Instruction Tuning
A. 2 只有语言的指令调优

The detailed instruction tuning hyperparameters are listed in Table 19.
详细的指令调优超参数列于表 19 中。

Hyperparameters 超参数	
Training steps 训练步骤	10,000 10,000
Warmup steps 热身步骤	375
Batch size of instruction data 指令数据的批量	256
Batch size of text corpora 文本语料库的批量	32
Batch size of image-caption pairs 图片标题对的批量大小	768
Batch size of interleaved data 交织数据的批量大小	16
Learning rate 学习率	2e-5 2e-5

Table 19: Instruction tuning hyperparameters of Kosmos-1
表 19: kosmos-1 的指令调优超参数

B Datasets
数据集

B.1 Pretraning
B. 1 预培训

B.1.1 Text Corpora
B. 1.1 文本语料库

Kosmos-1 is trained on The Pile [GBB+20] and Common Crawl. The Pile is an 800 GB English text corpus combining 22 diverse sources. We select a subset with seven sources from The Pile. Common Crawl is also included in training corpora. Common Crawl takes snapshots of the web, which contains massive amounts of language data. Table 20 provides a full overview of the language datasets that were used in the training of Kosmos-1 model. These data sources can be divided into the following three categories:
Kosmos-1 是在桩[GBB + 20]和普通爬行训练。Pile 是一个 800gb 的英文文本语料库，结合了 22 个不同的来源。我们从 The Pile 中选择一个包含 7 个源的子集。Common Crawl 也包含在训练语料库中。Common Crawl 拍摄了包含大量语言数据的网络快照。表 20 提供了 kosmos-1 模型训练中使用的语言数据集的全面概述。这些数据源可以分为以下三类:

- Academic: NIH Exporter
学术: NIH 出口商

- Internet: Pile-CC, OpenWebText2, Wikipedia (English), CC-2020-50, CC-2021-04, Realnews
Internet: Pile-CC，OpenWebText2，Wikipedia (English)，CC-2020-50，CC-2021-04，Realnews

- Prose: BookCorpus2, Books3, Gutenberg [RPJ+20], CC-Stories
散文: BookCorpus2，Books3，Gutenberg [RPJ + 20]，CC-Stories

Datasets 数据集	Tokens (billion) 令牌(十亿)	Weight (%) 体重(%)	Epochs 纪元
OpenWebText2 开放网页文本 2	14.8	21.8% 21.8%	1.47
CC-2021-04 CC-2021-04	82.6	17.7% 17.7%	0.21
Books3 书 3	25.7	16.2% 16.2%	0.63
CC-2020-50 CC-2020-50	68.7	14.7% 14.7%	0.21
Pile-CC 堆 -cc	49.8	10.6% 10.6%	0.21
Realnews Realnews 真实新闻	21.9	10.2% 10.2%	0.46
Wikipedia 维基百科	4.2	5.4% 5.4%	1.29
BookCorpus2 书本 2	1.5	1.1% 1.1%	0.75
Gutenberg (PG-19) 古腾堡(PG-19)	2.7	1.0% 1.0%	0.38
CC-Stories Cc-故事	5.3	1.0% 1.0%	0.19
NIH ExPorter NIH 出口商	0.3	0.2% 0.2%	0.75

Table 20: Language datasets used to train the Kosmos-1 model.
表 20: 用于训练 kosmos-1 模型的语言数据集。

B.1.2 Image-Caption Pairs
B. 1.2 图片-标题对

Kosmos-1 is trained on image-caption pairs constructed from several datasets, including English LAION-2B [SBV+22], LAION-400M [SVB+21], COYO-700M [BPK+22] and Conceptual Cap-tions [SDGS18, CSDS21]. LAION-2B, LAION-400M, and COYO-700M datasets are extracted by parsing out image URLs and alt-texts of web pages from the Common Crawl web data. LAION-2B contains about 2B English image-caption pairs, LAION-400M consists of 400M English image-caption pairs, and COYO-700M has 700M English image-caption pairs. Conceptual Captions contains 15M English image-caption pairs and consists of two datasets: CC3M and CC12M, which are also collected from internet webpages using a Flume pipeline. For Conceptual Captions, we discard pairs whose captions contain special tags such as “<PERSON>”.

Kosmos-1 接受了由若干数据集构成的图像标题对的培训，这些数据集包括英文 LAION-2B [SBV + 22]、 LAION-400M [SVB + 21]、 COYO-700M [BPK + 22]和概念上的上限[SDGS18，CSDS21]。LAION-2B、 LAION-400M 和 COYO-700M 数据集是通过从普通爬网数据中解析出网页的图像 url 和 alt 文本来提取的。LAION-2B 包含约 2b 个英文图像标题对，LAION-400M 包含 400m 个英文图像标题对，COYO-700M 包含 700m 个英文图像标题对。概念标题包含 1500 万个英文图像标题对，由两个数据集组成: CC3M 和 CC12M，这两个数据集也是通过 Flume 管道从互联网网页上收集的。对于概念标题，我们丢弃那些标题包含特殊标签如“ < person >”的对。

B.1.3 Interleaved Data
B. 1.3 交错的数据

We collect a large corpus of 2 billion web pages from the snapshots of common crawls. To ensure quality and relevance, we apply several filtering criteria. First, we discard any pages that are not written in English. Second, we discard any pages that do not have images interspersed in the text. Third, we discard any images that have a resolution lower than 64 by 64 pixels or that are single-colored. Fourth, we discard any text that is not meaningful or coherent, such as spam or gibberish. We use some heuristics to identify and remove gibberish text containing emoji symbols, hashtags, and URL links. After applying these filters, we end up with about 71 million documents for training.

我们收集了 20 亿个网页的大型语料库，这些网页来自于常见的抓取行为。为了确保质量和相关性，我们应用了几个过滤标准。首先，我们丢弃所有不是用英语写的页面。第二，我们丢弃任何没有在文本中穿插图片的页面。第三，我们丢弃任何分辨率低于 64 * 64 像素或单色的图片。第四，我们丢弃任何没有意义或不连贯的文本，如垃圾邮件或胡言乱语。我们使用一些启发式方法来识别和删除含有表情符号、标签和 URL 链接的胡言乱语文本。在应用了这些过滤器之后，我们最终得到了大约 7100 万份用于培训的文档。

B.2 Data Format
B. 2 数据格式

The training data is organized in the format as follows:

培训数据的格式如下:

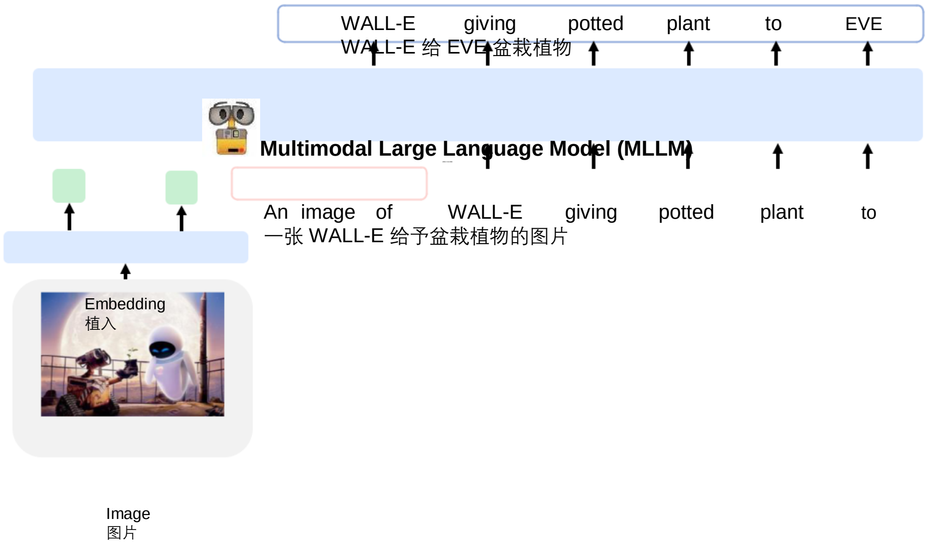
Datasets 数据集	Format Examples 格式化例子
Text 短信	<s> KOSMOS-1 can perceive multimodal input, learn in context, and gener- Kosmos-1 能够感知多模态输入，在上下文中学习，并且能够产生- ate output. </s> 食物输出
Image-Caption 图片说明	<s> <image> Image Embedding </image> WALL-E giving potted plant to 图像嵌入给盆栽植物 EVE. </s> EVE
Multimodal 多式联运	<s> <image> Image Embedding </image> This is WALL-E. <image> 这是 WALL-E Image Embedding </image> This is EVE. </s> 图像嵌入这是 EVE

Table 21: The examples of the data format to train the KOSMOS-1 model.
表 21: 训练 kosmos-1 模型的数据格式示例。

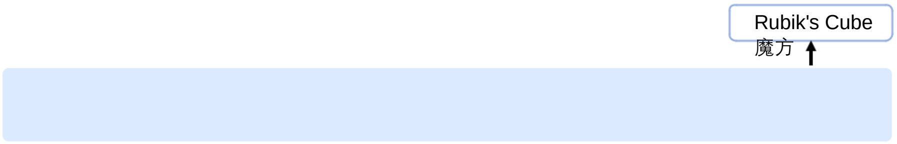
C Evaluation
评估

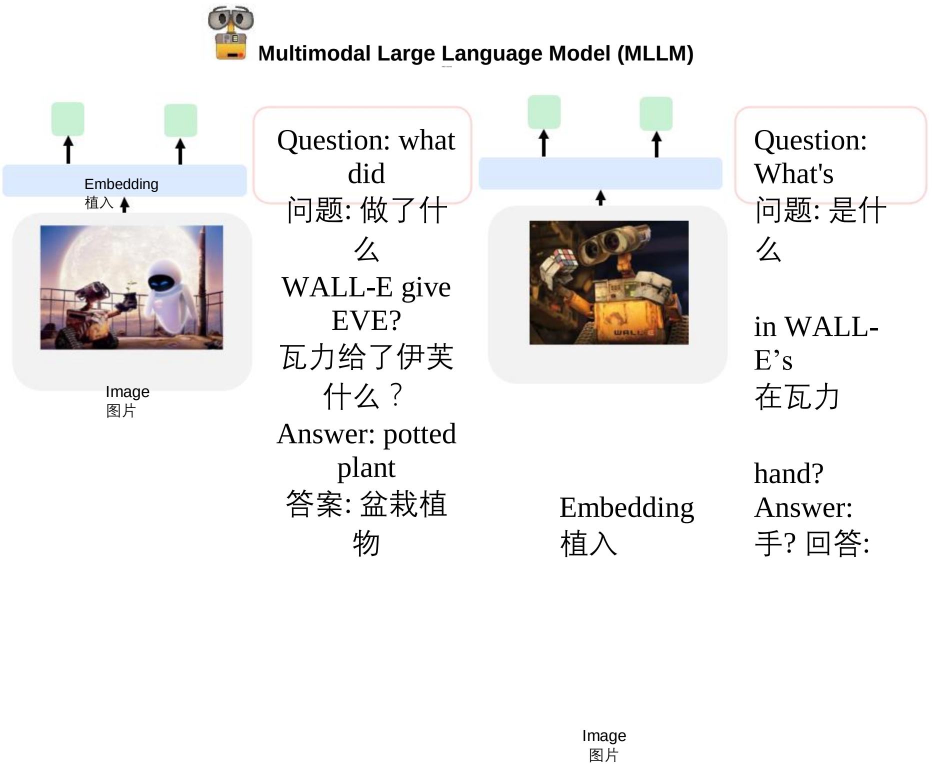
C.1 Input Format Used for Perception-Language Tasks
C. 1 用于感知的输入格式-语言任务

Figure 7 shows how we conduct zero-shot and few-shot evaluations on perception-language tasks.
图 7 显示了我们如何对感知语言任务进行零次和零次评估。


WALL-E giving potted plant to EVE
WALL-E 给 EVE 盆栽植物
Multimodal Large Language Model (MLLM)
An image of WALL-E giving potted plant to
一张 WALL-E 给予盆栽植物的图片
Embedding
植入
Image
图片

(a) Zero-shot learning
(a)零成功学习





(b) Few-shot learning
(b)少量学习

Figure 7: We evaluate Kosmos-1 on the perception-language tasks in zero- and few-shot settings. (a) Zero-shot learning, e.g., zero-shot image captioning with language prompts. (b) Few-shot learning, e.g., visual question answering with in-context learning.
图 7: 我们评估 kosmos-1 在零镜头和少镜头环境下的感知语言任务。(a)零镜头学习，例如，带有语言提示的零镜头图像字幕。(b)少镜头学习，例如，用上下文学习来回答视觉问题。

C.2 Language Tasks
C. 2 语言任务

We conduct experiments on language tasks in four categories:
我们在四个方面进行语言任务的实验:

- Cloze and completion tasks: StoryCloze [MRL+17], HellaSwag [ZHB+19]

完成任务: StoryCloze [MRL + 17]， HellaSwag [ZHB + 19]

- Winograd-style tasks: Winograd [LDM12b], Winogrande [SBBC20]

Winograd 风格的任务: Winograd [LDM12b]， Winogrande [SBBC20]

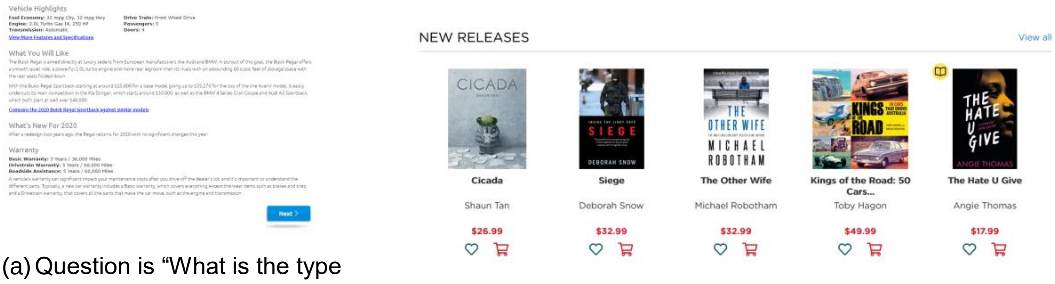
- Commonsense reasoning: PIQA [BZB+20]

常识推理: PIQA [BZB + 20]

- Three datasets from SuperGLUE benchmark [WPN+19]: BoolQ [CLC+19], CB [dMST19], COPA [RBG11]

来自 SuperGLUE 基准[WPN + 19]的三个数据集: bolq [CLC + 19]， CB [dMST19]， COPA [RBG11]

C.3 WebSRC Task Examples
C. 3 WebSRC 任务示例



Vehicle Highlights

Fuel Economy: 22 mpg City, 32 mpg Hwy
Engine: 2.0L Turbo Gas I4, 250 HP
Transmission: Automatic
Drive Train: Front Wheel Drive
Passengers: 5
Doors: 4

View More Features and Specifications

What You Will Like

The Buick Regal is aimed directly at luxury sedans from European manufacturers like Audi and BMW. In pursuit of this goal, the Buick Regal offers a smooth quiet ride, a powerful 2.0L turbo engine and more rear legroom than its rivals with an astounding 60 cubic feet of storage space with the rear seats folded down.

With the Buick Regal Sportback starting at around $25,000 for a base model going up to $35,270 for the top of the line Avenir model, it easily undercuts its main competition in the Kia Stinger, which starts around $33,000, as well as the BMW 4 Series Gran Coupe and Audi A5 Sportback, which both start at well over $40,000.

Compare the 2020 Buick Regal Sportback against similar models

What's New For 2020

After a redesign two years ago, the Regal returns for 2020 with no significant changes this year.

Warranty

Basic Warranty: 3 Years / 36,000 Miles
Drivetrain Warranty: 5 Years / 60,000 Miles
Roadside Assistance: 5 Years / 60,000 Miles

A vehicle's warranty can significant impact your maintenance costs after you drive off the dealer's lot, and it's important to understand the different parts. Typically, a new car warranty includes a Basic warranty, which covers everything except the wear items such as brakes and tires; and a Drivetrain warranty, that covers all the parts that make the car move, such as the engine and transmission.

NEW RELEASES

Cicada	Siege	The Other Wife	Kings of the Road: 50 Cars...	The Hate U Give
Shaun Tan	Deborah Snow	Michael Robotham	Toby Hagon	Angie Thomas
$26.99	$32.99	$32.99	$49.99	$17.99

(a) Question is “What is the type
问题是，什么类型的
of this drive?”
(b)问题是“谁是《蝉》的作者?”

(b) Question is “Who is the author of "Cicada"?”

Figure 8: Examples form WebSRC [CZC+21].
图 8: WebSRC [CZC + 21]的示例。

26

26