

DeepSeek-VL: 迈向现实世界的视觉-语言理解

Haoyu Lu^{*1†}, Wen Liu^{*1}, Bo Zhang^{*1‡}, Bingxuan Wang^{1†}, Kai Dong¹, Bo Liu^{1†}, Jingxiang Sun^{1†},
Tongzheng Ren^{1†}, Zhuoshu Li¹, Hao Yang^{1†}, Yaofeng Sun¹, Chengqi Deng¹, Hanwei Xu¹, Zhenda Xie¹,
Chong Ruan¹

¹DeepSeek-AI

{neal, liuwen, bo}@deepseek.com

<https://github.com/deepseek-ai/DeepSeek-VL>

Abstract

我们提出 DeepSeek-VL, 这是一个专为现实世界视觉与语言理解应用设计的开源视觉-语言 (VL) 模型。我们的方法围绕三个关键维度构建:

- **数据构建:** 我们致力于确保数据具有多样性、可扩展性, 并广泛覆盖现实世界场景, 包括网页截图、PDF、OCR、图表以及基于知识的内容 (专家知识、教科书), 旨在全面呈现实际应用场景。此外, 我们从真实用户场景中构建用例分类体系, 并据此构建指令微调数据集。使用该数据集进行微调显著提升了模型在实际应用中的用户体验。

- **模型架构:** 考虑到效率及大多数现实场景的需求, DeepSeek-VL 采用混合视觉编码器, 能够在固定的 Token 预算内高效处理高分辨率图像 (1024 x 1024), 同时保持相对较低的计算开销。这一设计选择确保了模型在各种视觉任务中捕捉关键语义和细节信息的能力。

- **训练策略:** 我们认为, 一个优秀的视觉-语言模型首先应具备强大的语言能力。为确保在预训练期间保留大语言模型 (LLM) 的能力, 我们通过从一开始就整合 LLM 训练, 并仔细管理视觉与语言模态之间观察到的竞争动态, 探索了一种有效的 VL 预训练策略。我们从专注于文本开始, 逐步调整比例以促进两种模态的均衡融合。

DeepSeek-VL 系列模型 (包含 1.3B 和 7B 版本) 在现实世界应用中作为视觉-语言聊天机器人展现出卓越的用户体验, 在相同模型规模下, 于广泛的视觉-语言基准测试中达到最先进或具有竞争力的性能, 同时在以语言为中心的基准测试中保持稳健表现。我们已公开 1.3B 和 7B 模型, 以促进基于该基础模型的创新。

* 同等贡献.

† 在 DeepSeek-AI 实习期间完成的工作.

‡ 项目负责人.

目录

1 引言	3
2 数据构建	5
2.1 视觉-语言预训练数据	5
2.2 监督微调数据	10
3 方法	10
3.1 架构	10
3.2 训练流程	12
3.2.1 阶段 1: 训练视觉-语言适配器	12
3.2.2 阶段 2: 联合视觉-语言预训练	13
3.2.3 阶段 3: 监督微调	14
3.3 超参数与基础设施	15
4 评估	15
4.1 公开多模态基准测试评估	15
4.2 Public Language Benchmarks Evaluation	17
4.3 Human Evaluation	18
4.4 Ablation Study	20
5 Conclusion, Limitation, and Future Work	23
A 附录	32

1. 引言

大型语言模型 (LLM) 的显著成功 (Anthropic, 2023; Google, 2023; OpenAI, 2022, 2023a) 激发了对能够处理语言之外多种模态的通用接口的需求。为响应这一日益增长的需求, 我们见证了 GPT-4V (OpenAI, 2023b) 和 Gemini (Team et al., 2023) 等大型多模态模型 (LMM) 的涌现, 它们作为通用助手, 能够理解并执行跨越视觉与语言的指令。这些模型在执行复杂多样的现实世界任务方面展现出巨大潜力, 使交互更加自然且贴近人类。

近期, 开源大型多模态模型层出不穷, 旨在缩小与闭源模型的差距。尽管在基准测试性能方面取得了实质性进展, 但在现实世界性能和用户体验方面, 大多数开源模型与最先进的闭源模型 (Bai et al., 2023; Bavishi et al., 2023; OpenAI, 2023b; Team et al., 2023) 之间仍存在显著差距。开源社区开发具备强大通用多模态能力以应对现实世界应用的模型仍具挑战性。

开源模型与闭源模型之间的性能差距在现实场景中尤为明显, 主要原因如下:

- 许多开源方案将大量计算资源分配给指令微调阶段。然而, 训练强大语言模型的经验强调了广泛预训练在通用智能发展中的重要性。为使多模态模型赋予丰富的世界知识, 应侧重于利用广泛视觉-语言数据的全面预训练。
- 一种常见做法是在指令微调期间合并多个学术数据集。尽管这种方法可能带来良好的基准测试结果, 但往往难以提供真实的现实世界使用体验。
- 在模型架构方面, 先前的工作大多将通常与文本对齐的视觉 Transformer 适配到预训练语言模型。然而, 这些模型大多在相对较低的分辨率下运行, e.g., 336×336 或 448×448。复杂现实场景的复杂性, 如光学字符识别或微小物体辨识, 需要高分辨率处理能力。
- 尽管部分模型 (01-ai, 2024; Lin et al., 2023a; Sun et al., 2023; Wang et al., 2023b) 已开始利用预训练, 但它们往往忽视了语言技能的保留。通常, 经过长时间的多模态训练后, 语言能力会出现退化。由于我们的目标是开发在两种模态上均具备强大能力的通用模型, 因此在开发新模态能力时, 应采用能有效保留语言能力的训练策略。

鉴于此, 我们提出 DeepSeek-VL, 这是一个基于 DeepSeek 语言模型系列构建的开源大型多模态模型。我们开发该模型旨在追求在现实场景中的卓越性能, 这涉及广泛的预训练、基于用例分类的仔细数据策划、针对高分辨率处理的模型架构设计, 以及平衡多模态的训练策略。在此基础上, 我们开发了一种指导模型从 1B 扩展到 7B 的训练方法。这些全面的探索带来了显著的性能优势, 相较于其他类似规模的大型多模态模型 (LMM), 在实际设置中表现更优。

DeepSeek-VL 的预训练数据集汇编自多种来源, 包括但不限于 Common Crawl、Web Code、电子书、教育材料和 arXiv 文章。该集合全面涵盖了网页截图、PDF、OCR、图表和基于知识的内容 (专业知识、教科书) 等现实场景, 旨在实现广泛且实用的代表性, 同时保持可扩展性。

尽管我们的预训练数据涵盖了广泛的世界知识, 但我们精心策划指令微调数据集以反映现实世界的使用场景。为此, 我们从互联网上手动收集了 GPT-4V 和 Gemini 的真实测试用例。这些用例已被系统地组织成全面的分类体系。我们使用该结构化分类体系为每张测试图像选择提

Prompt

图片描述了什么算法？

Figure 1: Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict over 1000 CLIP prompts from an image encoder and a text encoder to predict the correct prompt of a batch of images, our training process, as you can see from the flowchart, involves a cross-modal linear classifier by contrasting the vector or description of the target dataset's classes.

Response

图片展示了一个名为CLIP的算法的流程图。该算法被描述为一种新的零样本图像到图像转换方法，它使用文本描述来预测目标数据集的类别。

在流程图中，我们看到了三个主要步骤：

1. **Contrastive pre-training (对比式预训练)**：在这个阶段，模型被训练以最小化两个编码器之间的差异，一个用于图像，另一个用于文本。这通过将图像和文本嵌入到相同的潜在空间中来实现。
2. **Create dataset classifier**：从给定的标签文本（创建标签文本分类器）；这个步骤涉及从给定的标签文本中提取特征向量，并将其与图像嵌入一起输入到一个分类器中。
3. **Use for zero-shot prediction (用于零样本文本预测)**：最后，使用前面步骤中学习的模型进行零样本文本预测。这涉及到将图像嵌入和提取的特征向量传递给分类器，以预测图像属于哪个类别。

整个过程都围绕着学习一个能够理解图像和文本之间关系的模型这一核心概念。这种方法的关键在于利用大量的无标注数据，通过对比式的预训练来提高模型的泛化能力。在流程图的底部，有一个例子展示了如何使用CLIP进行零样本预测。这里，一张狗的图片和一个关于狗的短语被用来预测这张照片中的狗属于哪个品种。

总的来说，流程图清晰地描绘了CLIP算法的结构和它的工作原理，强调了它在处理大量未标记数据方面的潜力以及其对零样本任务的有效性。

Prompt

You are a home robot. Now you are at the kitchen, this is what you can see from there. To fetch something from the fridge, what should you do?

Response

To fetch something from the fridge, I would follow these steps:

- Approach the refrigerator:** The refrigerator is located on the right side of the kitchen island. It's a stainless steel model with a freezer compartment at the bottom and an ice maker dispenser above it.
- Open the refrigerator door:** To open the refrigerator door, I would use my hands to grasp the handle or pull bar located on the front edge of the refrigerator door. This action should be done carefully to avoid any accidental bumps into other items in the kitchen.
- Locate the desired item:** Once inside the refrigerator, I would scan through the shelves and compartments for the specific item I am looking for. The refrigerator has multiple shelves and drawers that are likely organized by type of food or drink.
- Retrieve the item:** After locating the desired item, I would reach out and grab it using one hand while ensuring not to disturb the organization of the contents. If the item is placed on a shelf or drawer, I would lift it up gently to remove it without knocking anything over.
- Close the refrigerator door:** After retrieving the item, I would close the refrigerator door securely using the handle or pull bar. This ensures that the cold air remains trapped within the appliance and maintains its temperature efficiency.
- Return to the task at hand:** With the item retrieved from the fridge, I can now proceed with whatever activity I was doing before, such as cooking or preparing a meal.

Prompt

Which bathroom is the bedroom A closer to?

Response

Bedroom A is closer to Bathroom A.

This can be determined by observing the floor plan provided in the image. The layout of the rooms and bathrooms are clearly marked with labels, allowing us to identify their relative positions. Bedroom A is located on the left side of the floor plan, while Bathroom A is situated directly below it, towards the bottom-left corner of the image. This positioning indicates that there is no other bathroom between them, making Bathroom A the closest one for Bedroom A.

Prompt

Tell me the main contribution of this paper.

The role of individual differences in human-automated vehicle interaction

Christopher A. Fisher, ..., Morgan S. Morris, Christopher A. Stone, & Garrett Swaney

Response

The main contribution of this paper, as stated in the abstract, is to develop a cognitive model that integrates individual differences into human-automated vehicle interaction. The authors aim to provide insights into how humans interact with new technologies and propose a framework for understanding these interactions. They also discuss potential pitfalls associated with using a one-size-fits-all model and explain how modeling individual differences can help design effective user interfaces and interventions. Additionally, they conduct a simulation study to demonstrate possible benefits from incorporating individual differences into the design process. This conclusion is drawn directly from the text under the "ABSTRACT" section, where it outlines the purpose and contributions of the research.

Prompt

Can you translate the flowchart to a python code?

```
def max_of_two(x, y):
    # Step 1: compare x and y
    if x > y:
        # Step 4: if x is greater than y, output max = x
        return x
    else:
        # Step 5: if x is not greater than y, output max = y
        return y

# Put max to output
output = max_of_two(input_value_for_x, input_value_for_y)
print(output)
```

Response

The flowchart provided is a simple decision-making process that compares two variables, x and y, to determine which one has the greater value. The steps are as follows:

1. Start with an input for variable 'x' and get its next input.
2. Repeat step 1 for variable 'y'.
3. Compare 'x' and 'y' using the "x > y" condition. If true, it means 'x' is greater than 'y', so put the maximum of these two values into the output. If false, it means 'x' is not greater than 'y', so put the maximum of 'x' and 'y' into the output.
4. Finally, end the program by putting the max value into the output.

Here is how you can translate this flowchart into Python code:

```
def max_of_two(x, y):
    # Step 1: compare x and y
    if x > y:
        # Step 4: if x is greater than y, output max = x
        return x
    else:
        # Step 5: if x is not greater than y, output max = y
        return y

# Put max to output
output = max_of_two(input_value_for_x, input_value_for_y)
print(output)
```

In this Python code, we define a function called max_of_two that takes two arguments, x and y, and returns the maximum of these two values. We then call this function with the appropriate inputs and print the result.

Prompt

描述图片

Response

这幅图片是Minecraft官方网站的生动欢迎界面。背景是一片辽阔的蓝天，点缀着朵朵蓬松的白云。下方是一个充满活力的像素艺术世界，特色是一个村庄、一片茂密的森林和一条蜿蜒的小路。

在图片的中心，有一个绿色的按钮吸引着注意力，上面用粗体白色字母写着“获取Minecraft”。这个按钮似乎邀请观者进入游戏并开始他们的Minecraft冒险。

在图片的顶部，一段文字热情地欢迎观众来到官方Minecraft网站，它写道：“欢迎来到官方Minecraft网站！加入我们的社区，了解如何玩这款游戏，加入我们最大的社区之一，并在今天开始创造吧！”这段文字传达了游戏的包容性和社区精神，鼓励玩家探索和创造。

总的来说，这幅图片捕捉到了Minecraft世界的精髓——它的无限可能性以及与他人互动乐趣。

Prompt

What art style is evident in this image?

Response

The art style evident in this image is Impressionism. This can be determined by several visual clues present:

- Brushwork:** The painting exhibits a loose, visible brushstroke technique which is characteristic of Impressionist art. The strokes are not smooth or blended but rather they create a sense of movement and texture.
- Color Palette:** The use of bright, vibrant colors that seem to capture the light and atmosphere of the scene without precise detail is typical of Impressionist painters. The sky's blue with white clouds, the warm tones on the cliffs, and the varied hues used for the boats all contribute to this impressionistic approach.
- Subject Matter:** The subject matter itself—a coastal landscape with figures and boats—is common in Impressionist paintings, which often depicted scenes from daily life and nature.
- Lack of Detail:** There is a lack of fine detail in the depiction of objects such as the boats and figures, which suggests an emphasis on capturing the overall mood and impression of the scene rather than meticulous realism.
- Light and Shadow:** The play of light and shadow is another hallmark of Impressionist works. In this painting, we see how the artist has captured the interplay of sunlight on the water and rocks, creating a dynamic effect. These elements together strongly suggest that the artwork belongs to the Impressionist genre.

Prompt

Which direction of paper is the most likely source of the table in the picture.

Response

The image shows a graph with various lines representing different models' validation log perplexity over the length of the validation sequence, which is measured in 10^3 . The title "C4 language modeling (large model)" suggests that this graph pertains to a study or analysis related to large-scale natural language processing tasks, specifically focusing on language modeling using a C4 dataset.

To determine the most likely direction of paper from which the table originated, we should look for visual clues in the image:

- 1. Graph Style:** The graph has a clear and professional appearance, suggesting it could be part of an academic paper or technical report.
- 2. Data Representation:** The use of multiple lines with markers such as squares, diamonds, and circles indicates comparisons between different models or methods. This kind of data representation is common in research papers where different experimental results are presented side by side for comparison.
- 3. Title and Content:** The title provides context about the nature of the content—it mentions "language modeling" and "large model", indicating that the paper deals with computational linguistics or machine learning topics.
- 4. Legend:** There is a legend at the bottom left that matches the colors of the lines to their respective models. This is typical in scientific graphs to help readers understand what each line represents without having to refer back to the text.
- 5. Axis Labels:** The axes have labels that are consistent with scientific notation ("Validation sequence length" and "Validation log perplexity"). These labels suggest that the paper is intended for an audience familiar with these terms, which would include researchers or students in fields like computer science or engineering.
- 6. Font and Formatting:** The font used for the title and axis labels appears to be sans-serif, which is often used in modern publications due to its readability and clean design. Based on these observations, the most likely source of the table in the picture is a publication from a field related to artificial intelligence, machine learning, or computational linguistics. It could be a journal article, conference proceedings, a thesis, or even a presentation slide from a seminar or workshop. The style and content align well with scholarly work in those domains.

Prompt

Please generate the latex code of the formula in the image.

$$A_n = a_0 \left[1 + \frac{3}{4} \sum_{k=1}^n \left(\frac{4}{9} \right)^k \right]$$

Response

$$A_n = a_0 \left[1 + \frac{3}{4} \sum_{k=1}^n \left(\frac{4}{9} \right)^k \right]$$

图 1 | DeepSeek-VL 具备通用多模态理解能力，能够处理逻辑图表、网页、公式识别、科学文献、自然图像以及复杂场景中的具身智能。

示词，确保指令微调数据集的实用性和相关性。该分类体系还用于创建能有效评估现实世界性能的评估数据集。

视觉模块的设计旨在优化高分辨率视觉输入的利用，同时保持在固定的 Token 预算内以有效管理推理成本。为此，我们采用混合视觉编码器，将用于在 384×384 分辨率下进行粗略语义提取的文本对齐编码器与在 1024×1024 分辨率下捕捉详细视觉信息的高分辨率编码器相结合。通过融合这两种编码器，我们的混合方法高效地将 1024×1024 分辨率图像（足以满足大多数用例）压缩为 576 个 Token。该 Token 数量在丰富的视觉表示与 Token 经济性之间取得了平衡，使其适用于图文交错和多轮推理场景。

在多模态模型的预训练过程中，一个常见的挑战是当训练过程过度依赖视觉-语言数据时，语言能力可能会退化。我们的研究表明，保持较高比例的语言数据——具体而言，至少 70%——对于保留模型内语言知识的完整性至关重要。这种平衡对于实现不损害语言性能的稳健多模态能力至关重要。此外，我们引入了一种新颖的“模态预热”策略。该方法在训练期间仔细调整模态比例，逐步融入更多视觉-语言数据。模态比例的仔细调整与预热策略相结合，实现了两种模态的均衡性能。

在迭代优化模型时，我们在扩展到更大模型规模之前先进行小规模实验。然而，较小的模型（e.g., 1B 模型）无法在基准测试 (Schaeffer et al., 2024) 上展示合理的性能，也无法真实反映模型的性能。我们采用两种方法来解决这一问题。首先，我们将评估协议从多项选择题修改为比较选项的困惑度。此外，为防止指令遵循能力成为瓶颈，我们在预训练阶段混合了少量指令微调数据。通过这种方式，我们可以使用 1B 模型实现合理的性能，并在实验过程中更准确地衡量每次迭代的影响。

通过对通用视觉和语言基准的广泛评估，DeepSeek-VL 系列模型在现实世界应用中展现出卓越的用户体验，在相同模型规模下，于广泛的视觉-语言基准测试中达到最先进或具有竞争力的性能，同时保持稳健的语言中心性能。为促进创新并实现广泛的应用，我们已公开两个版本 (1.3B 和 7B)，希望满足不同的计算能力需求。

2. 数据构建

多样化且大规模的数据集是视觉语言模型训练最重要的要素。我们的数据集可分为两部分：视觉-语言预训练数据和视觉-语言监督微调数据。VL 预训练数据由来自各种来源的视觉-文本数据组成，旨在增强模型的基础跨模态理解能力；而 VL 监督微调数据规模相对较小，旨在教导模型完成特定的下游任务。按设计，VL 预训练数据用于在训练阶段 1 预热视觉-语言适配器，并在阶段 2 联合预训练视觉-语言模型，而 VL 监督微调数据则用于训练阶段 3，即视觉-语言监督微调。

表 1 | 联合视觉与语言预训练阶段所用数据集的摘要。

类别	数据集	比例
图文交错	MMC4 (Zhu et al., 2024) Wikipedia EN& CN (Foundation) Wikihow (Yang et al., 2021) 内部 PDF 和 Epub 教科书	13.1%
图像描述	Capsfusion (Yu et al., 2023a) TaiSu (Liu et al., 2022b) Detailed Caption (echo840, 2024)	11.1%
表格与图表	Chart2text (Kantharaj et al., 2022) Geo170K (Gao et al., 2023) Ureader (Ye et al., 2023) Unichart (Masry et al., 2023) M-paper (Hu et al., 2023) ScienceQA (Lu et al., 2022b) ScreenQA (Hsiao et al., 2022) SciGraphQA-295K (Li and Tajbakhsh, 2023) Paper2figure100k (Rodriguez et al., 2023) Widget Captioning (Li et al., 2020) Screen2words (Wang et al., 2021) Refexp (Mao et al., 2016)	2.1%
网页代码	Websight (HuggingFaceM4, 2024) 从 GitHub notebook 抓取的 python 绘图	0.4%
场景文本 OCR	ArT (Chng et al., 2019) MLT-17 (Nayef et al., 2017) LSVT (Sun et al., 2019) UberText (Zhang et al., 2017) Coco-text (Veit et al., 2016) RCTW-17 (Shi et al., 2017) ReCTS (Zhang et al., 2019) TextOCR (Singh et al., 2021) OpenVINO (Krylov et al., 2021) HierText (Long et al., 2022)	1.2%
文档 OCR	arXiv 渲染的 markdown (Blecher et al., 2023)	2.1%
纯文本语料	DeepSeek-LLM 2T 文本语料 (DeepSeek-AI, 2024)	70.0%

2.1. 视觉-语言预训练数据

本研究使用的预训练数据集涵盖了多种公开来源，此外还包含部分专有数据。我们在表 1 中提供了联合视觉与语言预训练阶段所用数据来源的全面概述。此类数据集有助于 LLM 理解图像中描绘的实体。

此外，我们提供了完整数据集的详细分类，组织如下：

图文交错数据使模型具备更好的多模态输入上下文学习能力，我们使用了三个公开数据集

MMC4 (Zhu et al., 2024)、Wiki (Burns et al., 2023)、Wikihow (Yang et al., 2021) 以及 Epub 教科书。

图像描述数据来自三个高质量的图文配对数据集: Capsfusion (Yu et al., 2023a)、TaiSu (Liu et al., 2022b) 和 Detailed Caption (echo840, 2024)。

表格与图表数据使模型能够学习通用的表格和图表图像理解能力。它涵盖了多种公开数据来源, 包括 Chart2text (Kantharaj et al., 2022)、Geo170K (Gao et al., 2023)、Unichart (Masry et al., 2023)、Ureader (Ye et al., 2023)、M-paper (Hu et al., 2023)、ScienceQA (Lu et al., 2022b)、ScreenQA (Hsiao et al., 2022)、SciGraphQA-295K (Li and Tajbakhsh, 2023)、Paper2figure100k (Rodriguez et al., 2023)、Widget Captioning (Li et al., 2020)、Screen2words (Wang et al., 2021) 和 Refexp (Mao et al., 2016)。

网页代码数据赋予模型从图形界面或可视化图表重建代码的能力。利用 Websight (HuggingFaceM4, 2024) 进行 UI 逆向渲染, 我们采用了类似于 MATCHA (Liu et al., 2022a) 中用于可视化图表逆向渲染的策略。这涉及处理来自 Stack 数据集 (Kocetkov et al., 2023) 的约 146 万个 Jupyter notebook。通过提取这些 notebook 并整理所有图表及其对应的上方代码段, 我们成功策划了一个包含 200 万对图像和代码的集合。为提高数据质量, 我们筛选出 110 万个实例, 每个实例包含单张图像和至少 5 行代码, 以构成我们的主要训练数据集。

文档光学字符识别 (OCR) 数据有助于在文档级别识别光学字符, 即使在具有挑战性的现实场景中也是如此。据我们所知, 目前尚无公开可用的涵盖中英文文档的大规模数据集。尽管存在公开可用的小规模数据集 Latex-OCR (Blecher, 2024), 但我们额外构建了一个全面的中英文文档 OCR 数据集。它由两部分组成: 1) **arXiv 文章**: 我们收集了 140 万篇 arXiv 文章的源代码和编译后的 PDF。利用 Nougat (Blecher et al., 2023) 的预处理工具, 我们将这些文章渲染为配对的图像和文本; 2) **电子书与教育材料**: 我们从 Anna's Archive (Anna's Archive, 2024) 清理了 86 万本英文和 18 万本中文电子书, 以及数百万道 K-12 教育考试题。随后, 我们使用 HTML 渲染工具 (Kulkarni and Truelsen) 将这些带有不同模板的 HTML 文件转换为配对的图像和文本格式。

场景文本 OCR 数据增强了模型从文本与环境融合的图像中识别和提取文本的能力。该数据集由多个公开数据集组成, 包括 ArT (Chng et al., 2019)、MLT-17 (Nayef et al., 2017)、LSVT (Sun et al., 2019)、UberText (Zhang et al., 2017)、Coco-text (Veit et al., 2016)、RCTW-17 (Shi et al., 2017)、ReCTS (Zhang et al., 2019)、TextOCR (Singh et al., 2021)、OpenVINO (Krylov et al., 2021) 和 HierText (Long et al., 2022)。

纯文本语料用于维持语言中心任务的熟练度。在本研究中, 我们使用了与 DeepSeek-LLM (DeepSeek-AI, 2024) 相同的文本语料。

表 2 | 我们联合视觉与语言监督微调阶段所用数据的摘要。

类别	数据集	比例
内部数据	基于分类体系的 SFT 数据 (图 4)	10.5%
通用多模态	ShareGPT4V (Chen et al., 2023) LAION-GPTV (LAION, 2023) LVIS-Instruct4V (Wang et al., 2023a) textOCR-GPT4V (Carter, 2024) LLaVA1.6-GPT4V (Liu et al., 2024a) IconQA (Lu et al., 2021)	35.5%
表格与图表	Ureader (Ye et al., 2023) Geo170K (Gao et al., 2023) ScienceQA (Lu et al., 2022b)	4.1%
网页代码	Screen-to-code (Abi, 2024) ScreenQA (Hsiao et al., 2022)	2.0%
纯文本 SFT	DeepSeek-LLM (DeepSeek-AI, 2024)	47.9%

2.2. 监督微调数据

本研究使用的监督微调数据集涵盖了多种多模态和语言数据来源，包括知名的开源共享 GPT4V 数据集，如 ShareGPT4V (Chen et al., 2023)、LAION-GPTV (LAION, 2023)、LVIS-Instruct4V (Wang et al., 2023a)、textOCR-GPT4V (Carter, 2024)、LLaVA1.6-GPT4V (Liu et al., 2024a) 和 IconQA (Lu et al., 2021)。此外，我们纳入了从预训练数据集中提取的部分表格和图表数据，如 Ureader (Ye et al., 2023)、ScreenQA (Hsiao et al., 2022)、Geo170K (Gao et al., 2023) 和 ScienceQA (Lu et al., 2022b)。而且，我们整合了来自 Screen-to-code (Abi, 2024) 任务的 UI 代码数据集。为提高多模态 SFT 数据的质量，我们还策划了一部分高质量的内部多模态 SFT 数据，其中部分为中文。我们的内部指令微调数据集经过精心设计，以反映现实世界的使用场景并覆盖广泛的任務。我们首先从各种在线来源收集多样化的 GPT-4V 和 Gemini 真实测试用例。然后对这些测试用例进行仔细分析并组织成全面的分类体系，该体系涵盖多个类别，如识别、转换、分析、推理、评估和安全，详见表 4。该结构化分类体系作为为每张测试图像选择代表性提示词的指南，确保我们的指令微调数据集既实用又与现实世界应用相关。此外，该分类体系还用于构建平衡且全面的评估数据集，使我们能够有效评估模型在不同任务和类别上的性能。通过遵循这一系统方法，我们确保内部多模态 SFT 数据覆盖的类别与分类体系良好对齐，并代表现实世界的使用场景。此外，我们将 DeepSeek-LLM (DeepSeek-AI, 2024) 中使用的纯文本 SFT 数据作为我们联合视觉与语言 SFT 数据的一部分。

3. 方法

3.1. 架构

我们的系统包含三个模块：混合视觉编码器、视觉适配器和语言模型。本节将介绍各个部分。

大类	说明	二级分类	三级分类
识别 (Recognition)	该部分用例主要考察大模型对图像内容的理解和描述能力，对模型的知识储备和推理能力要求不高，且部分任务可使用传统机器学习模型完成。	全局描述 局部描述 OCR 及转录	主题描述、事件/行为描述、地点/场景描述、情感/情绪描述、风格识别、姿态识别、其他指向描述、位置描述、人物识别、物品属性描述、Logo 识别、计数、货币识别 印刷体转录、手写体转录、指定格式转录、指定语言转录
转化 (Conversion)	该类用例要求模型在能够描述和识别图像内容的基础上，运用特定知识（如代码知识、提示词工程知识）将图像内容转化为其他形式。	图转代码 图转文本	UI 转代码、图表转代码、照片转 SVG/Base64 编码、公式转代码、流程图转代码 图转提示词、文本摘要、看图创作、文本解读
分析 (Analysis)	该类用例要求模型运用特定知识和逻辑能力，基于图像内容进行合理分析与理解，并根据指令描述图像内容。	数据图表分析 专业图表分析 专业影像分析 百科知识分析	图解读、表解读 电路图、流程图、地图、乐谱、金融图表、平面图、其他传感器影像、生物及医学影像、声纹图、点云图像、红外影像、卫星影像 艺术文化知识、自然环境知识、衣食住行相关知识、娱乐相关知识、历史知识
常识推理 (Commonsense Reasoning)	该类用例主要测试模型对生活常识的理解与掌握程度，需基于对图像内容的解读与分析，并结合生活常识完成推理。	关系推理 功能推理 环境推理 异常推理 姿态推理 其他常识推理	人际关系、空间关系、尺寸关系、物种关系 硬件功能推理、软件功能推理 环境状态分析、基于环境的行为推理、具身智能识别图像中的异常、缺陷检测、事故判断 - 状态推理、原因推理、属性比较、视错觉、趣味游戏、意愿解读、行为预测
逻辑推理 (Logical Reasoning)	该类用例要求模型结合对图像的理解，综合运用领域知识与逻辑推理能力，完成相应任务。	数学推理 其他逻辑推理	代数与运算、平面几何、立体几何 物理、化学、生物、代码、智力题
评估 (Evaluation)	该类用例要求模型按照特定标准对图像内容进行评价。	-	真实性评价、相似度评价、美学评价、开放式评价、改进建议
多图能力 (Multi-graph)	该类用例考察模型对多张图片的分析与理解能力。	时序理解 多图对比	事件预测、图片排序、行为分析 属性对比、图文配对、找关联、找不同、图片辨析
安全性 (Safety)	该类用例考察模型在安全性方面的表现。	-	诱导性提问、反事实提问、提示词注入

表 3 | 基于图像理解的 AI 模型综合能力评估体系。该体系全面考察模型在图像理解方面的多项能力，涵盖识别、转化、分析、推理、评估等各个层面，对模型的知识储备、逻辑推理、常识理解等方面进行深入测试，以期为大模型在多模态场景下的应用提供参考标准。

主类别	描述	子类别	三级类别
识别	此类用例主要考察大模型对图像内容的理解与描述能力，不需要模型具备较高的知识储备和推理能力，部分任务可使用传统机器学习模型完成。	全局描述	主题描述、事件/行为描述、位置/场景描述、情感/氛围描述、风格识别、食物识别、其他
		局部描述	指代描述、位置描述、人物识别、物体属性描述、标志识别、计数、货币识别
		光学字符识别与转录	印刷体文本转录、手写体文本转录、指定格式转录、指定语言转录
转换	此类用例要求模型能够描述和识别图像内容，并利用特定知识（如代码知识、提示工程知识）将图像内容转换为其他形式。	图像转代码	UI 转代码、图表转代码、照片转 SVG/p64 编码、公式转代码、流程图转代码
		图像转文本	图像转提示词、文本摘要、基于图像的创作、文本解读
分析	此类用例要求模型运用特定知识和逻辑能力，基于图像内容进行合理的分析与理解，并按照指令描述图像。	数据图表分析	图表解读、表格解读
		专业图表分析	电路图、流程图、地图、乐谱、金融图表、平面图、其他
		专业图像分析	传感器图像、生物医学图像、声纹图像、点云图像
		百科知识分析	艺术与文化知识、自然环境知识、衣食住行相关知识、娱乐相关知识、历史知识
常识推理	此类用例主要测试模型对生活中常识的理解与掌握程度，需要结合常识对图像内容进行解读与分析后进行推理。	关系推理	人际关系、空间关系、大小关系、物种关系
		功能推理	硬件功能推理、软件功能推理
		环境推理	环境状态分析、基于环境的行为推理、具身智能
		异常推理	图像异常识别、缺陷检测、事故判定
		幽默推理	-
		其他常识推理	状态推理、因果推理、属性对比、视觉错觉、趣味游戏、意图解读、行为预测
逻辑推理	此类用例要求模型结合对图像的理解，综合运用领域知识与逻辑推理能力以完成相应任务。	数学推理	代数与运算、平面几何、立体几何
		其他逻辑推理	物理、化学、生物、代码、智力题
评估	此类用例要求模型根据特定标准对图像内容进行评估。	-	真实性评估、相似度评估、美学评估、开放式评估、改进建议
多图	此类用例考察模型分析与理解多张图像的能力。	时序序列理解	事件预测、图像排序、行为分析
		多图对比	属性对比、图文匹配、寻找关联、找不同、图像判别
安全性	此类用例考察模型在安全性方面的表现。	-	诱导性提问、反事实提问、提示词注入

表 4 | 我们针对内部 SFT 数据构建的分类体系。本分类体系全面涵盖了我们的内部多模态 SFT 数据所涉及的类别。

混合视觉编码器。我们采用 SigLIP 作为视觉编码器，从视觉输入中提取高级语义特征表示。然而，我们观察到单个 SigLIP 编码器难以全面应对所有现实世界问题。CLIP 家族的视觉编码器（包括 SigLIP）主要为语义视觉表示设计，但面临模糊编码的挑战，导致视觉上截然不同的图像因所谓的“CLIP 盲对” Tong et al. (2024) 而被编码为相似。同时，CLIP 家族模型受限于相对较低的分辨率输入（例如 224 x 224、336 x 336、384 x 384、512 x 512），这阻碍了它们处理需要更详细低级特征的任务（如密集 OCR 和视觉定位任务）的能力。

为克服这些局限性，近期研究 (Lin et al., 2023b; Tong et al., 2024; Wei et al., 2023) 提倡整合额外的纯视觉自监督编码器，以增强多模态模型的视觉定位能力。基于先前的动机，我们额外使用基于 SAM-B (Kirillov et al., 2023) 的纯视觉编码器和预训练的 ViTDet (Li et al., 2022) 图像编码器来处理低级特征，其接受高分辨率 1024 x 1024 图像输入。除了 SAM-B 编码器外，我们保留了接受低分辨率 384 x 384 图像输入的 SigLIP-L 视觉编码器。因此，我们的混合视觉编码器结合了 SAM-B 和 SigLIP-L 编码器，在高效编码高分辨率 1024 x 1024 图像的同时保留语义和详细信息。具体而言，高分辨率 SAM-B 视觉编码器首先将图像调整为 1024 x 1024，并生成 64 x 64 x 256 的特征图。

对于 SAM-B 生成的 64 x 64 x 256 高分辨率特征图，VL 适配器首先将其插值为 96 x 96 x 256。随后，它采用两个步长为 2 的卷积层，生成 24 x 24 x 1024 的特征图，并将其重塑为 576 x 1024。与此同时，SigLIP-L 生成的 576 x 1024 低分辨率特征图与高分辨率特征拼接，最终得到 576 个维度为 2048 的视觉 Token。这些视觉 Token 具备增强高级语义视觉识别和低级视觉定位任务的巨大潜力。随后它们经过 GeLU 激活，并通过嵌入层与语言模型建立连接。

视觉-语言适配器。我们采用两层混合 MLP 来桥接视觉编码器和 LLM。最初，使用不同的单层 MLP 分别处理高分辨率特征和低分辨率特征。随后，这些特征沿其维度拼接，然后通过另一层 MLP 转换到 LLM 的输入空间。

语言模型。我们的语言模型基于 DeepSeek LLM (DeepSeek-AI, 2024) 构建，其微观设计很大程度上遵循 LLaMA (Touvron et al., 2023a,b) 的设计，采用带有 RMSNorm (Zhang and Sennrich, 2019) 函数的 Pre-Norm 结构，并使用 SwiGLU (Shazeer, 2020) 作为前馈网络 (FFN) 的激活函数，中间层维度为 $\frac{8}{3}d_{model}$ 。它还结合了 Rotary Embedding (Su et al., 2024) 用于位置编码，并使用与 DeepSeek-LLM 相同的分词器。我们推出了一系列 DeepSeek-VL 模型。鉴于我们进行多模态与语言联合预训练的目标，我们选择了 DeepSeek 预训练模型中的一个中间检查点以继续预训练。

具体而言，DeepSeek-VL-1B 模型基于 DeepSeek-LLM-1B 模型构建，该模型使用约 5000 亿文本 Token 的语料库进行了训练。而 DeepSeek-VL-7B 模型则利用使用约 2 万亿文本 Token 训练的 DeepSeek-LLM-7B 模型开发。

3.2. 训练流程

我们分三个连续阶段训练 DeepSeek-VL，如图 3 所示：视觉-语言适配器预热、联合视觉-语言预训练和监督微调。我们目前专注于视觉理解能力，仅对语言部分计算下一个 Token 预测损失。



Prompt

Is the cyclist on the left or right side of the woman's handbag?



Response

The cyclist is on the left side of the woman's handbag. This can be determined by observing the relative positions of the objects in the image:

1. The woman is holding a handbag that is positioned to her right side, as we view the image.
2. There is a cyclist depicted on the road behind the woman and to her left when looking at the image.
3. The cyclist appears to be moving away from the viewpoint, given their direction and the motion blur effect on their wheels
4. The cyclist is further down the road compared to the woman, which supports the conclusion that they are on the left side of the handbag based on our perspective.

These visual clues confirm that the cyclist is indeed on the left side of the woman's handbag.

图 2 | 可视化结果。DeepSeek-VL 能够捕捉微小物体并提供条理清晰的解释。

3.2.1. 阶段 1: 训练视觉-语言适配器

本阶段的主要目标是在嵌入空间内建立视觉与语言元素的概念联系，从而促进大型语言模型 (LLM) 对图像中描绘实体的全面理解。与 LLaVA (Liu et al., 2024b) 和 Instruct-BLIP (Dai et al., 2023) 的先前研究一致，我们采用类似的方法，在此阶段保持视觉编码器和 LLM 冻结，仅允许视觉-语言适配器内的参数可训练。我们使用包含 125 万对来自 ShareGPT4V 的图文配对描述以及 250 万对文档 OCR 渲染配对的数据集来训练 VL 适配器。

然而，与大型语言模型 (LLM) 相比，视觉-语言适配器 (例如 2 层 MLP) 的参数容量显著较小。这种模型容量的限制限制了本阶段可学习的能力。一个自然的问题随之产生：**数据扩展**

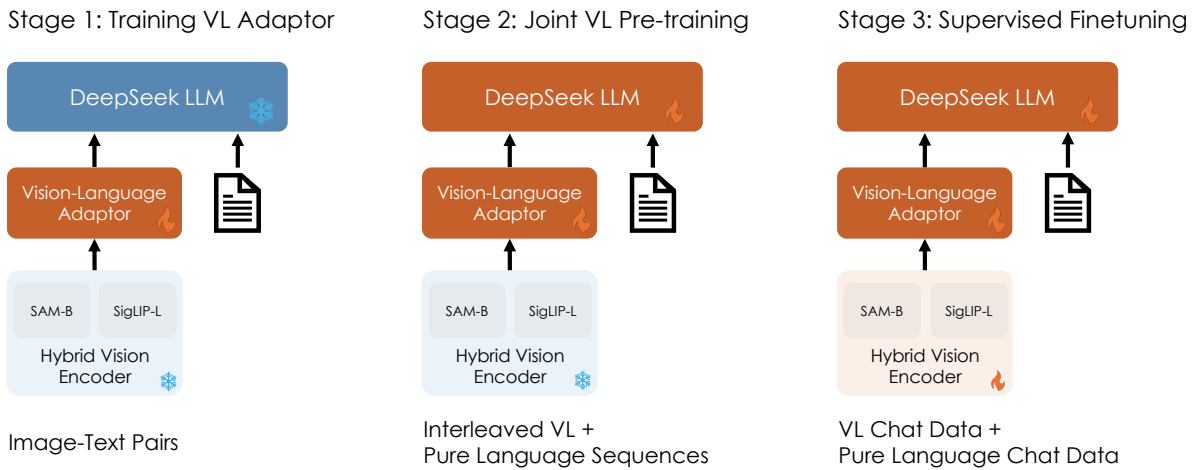


图 3 | 我们的训练流程包含三个阶段。阶段 1 涉及训练视觉-语言 (VL) 适配器，同时保持混合视觉编码器和语言模型固定。阶段 2 是联合视觉与语言预训练的关键部分，其中 VL 适配器和语言模型均可训练。阶段 3 是监督微调阶段，在此期间将训练低分辨率视觉编码器 SigLIP-L、VL 适配器和语言模型。

定律在此阶段是否有效？为回答这一问题，我们在表 9 中进行了一项简单实验。结果表明，在此阶段扩大数据规模并未带来益处，甚至可能导致性能下降。因此，我们继续解冻大型语言模型 (LLM)，并在阶段 2 探索高效的视觉-语言预训练方法。

3.2.2. 阶段 2: 联合视觉-语言预训练

在本阶段，我们探索有效的预训练策略，这可视为使大型语言模型 (LLM) 理解多模态输入的额外阶段。我们保持视觉编码器冻结，并优化语言模型和 VL 适配器。

最初，我们尝试直接使用多模态数据训练 LLM。然而，我们发现尽管多模态性能的指标逐步提升，但语言指标却出现急剧且严重的下降，如图 4 (多模态: 语言 =100%:0%) 所示。这凸显了在 LLM 基础上直接进行多模态预训练的内在挑战，揭示了增强多模态能力与保留语言熟练度之间的关键权衡。

我们假设观察到的现象源于两个主要因素：首先，大多数多模态语料过于简单，与语言数据的复杂性和分布存在显著差异。其次，多模态与语言模态之间似乎存在竞争动态，导致 LLM 内语言能力的灾难性遗忘。

联合语言-多模态训练为应对这一挑战，我们设计了一种简单而有效的联合语言-多模态训练策略。在训练期间，我们不仅进行多模态数据训练，还将大量语言数据纳入训练。该方法旨在平衡训练重点，缓解观察到的不利影响。我们在图 4 中对 DeepSeek-VL 1B 模型进行实验，以探索改变模态混合比例的影响。

对图表的分析得出以下几个关键结论：(1). 整合语言数据显著缓解了语言能力的下降，证明了模型语言性能的实质性提升。(2). 语言数据的加入并未导致多模态性能的显著损失，表明模型保留了其多模态处理能力。(3). 不同模态的性能与其在训练数据集中的比例 strongly 相关，

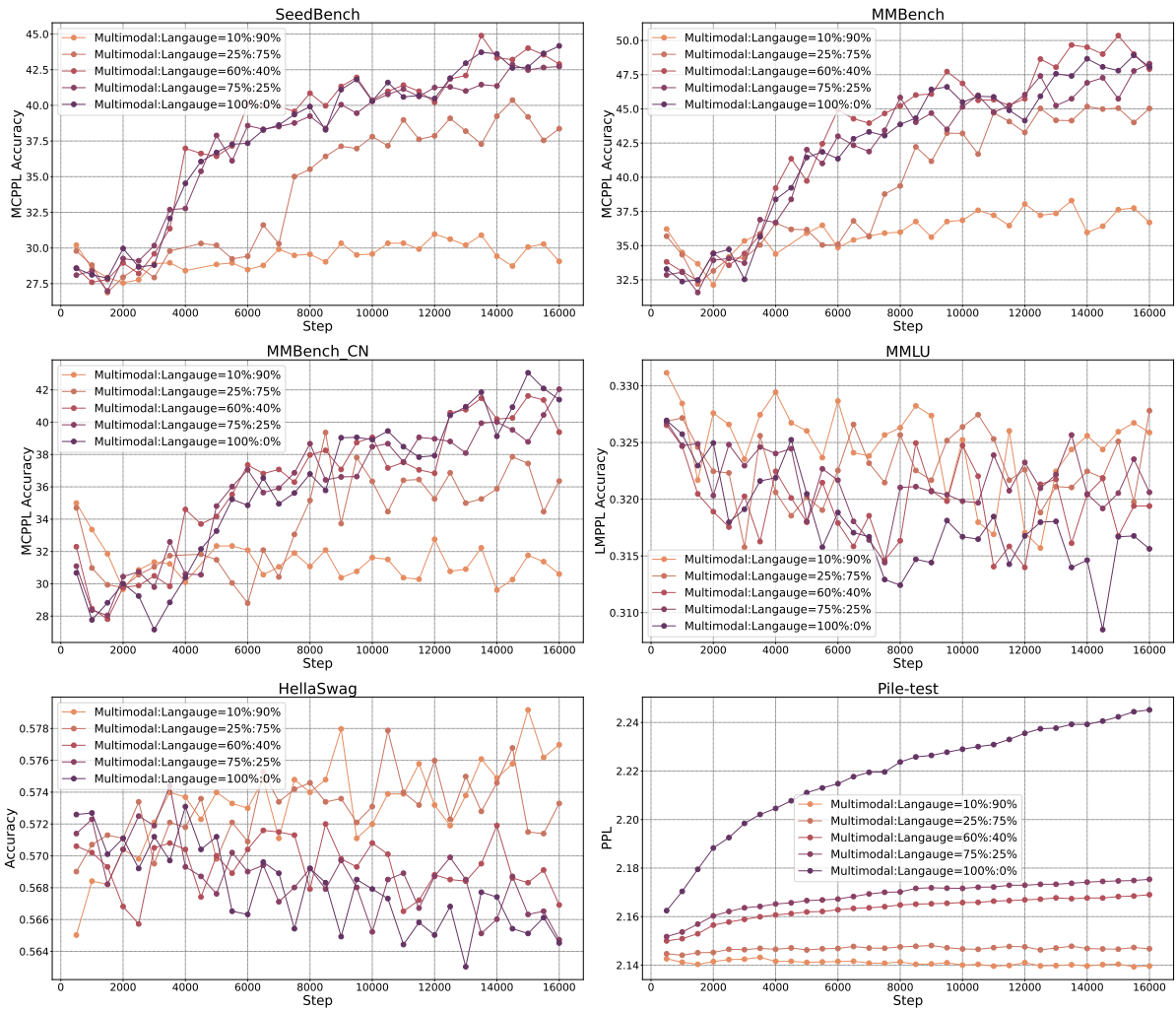


图 4 | 训练阶段 2 不同模态融合比例的对比性能结果。过大的多模态数据比例（多模态: 语言 =100%:0%）导致 LLM 语言能力的显著遗忘。合适的比例（多模态: 语言 =70%:30%）能有效缓解语言遗忘问题，同时增强模型的多模态能力。

证实了两种模态之间的竞争关系。最终，我们为最终模型选择了语言与多模态数据训练比例约为 7:3。该比例使模型在保持语言能力的同时，在多模态数据上实现更好的预训练，有效平衡了语言和多模态熟练度的发展。

扩展视觉-语言预训练然而，模型的预训练阶段会产生巨大的计算成本，在 7B 模型上进行迭代需要过多的算力和时间。一种合适的策略是在较小的模型（具体为 1.3B 模型）上进行实验，随后将其扩展到 7B 模型。幸运的是，我们观察到 1.3B 模型获得的大部分结果可以通过利用 SFT（例如编码器设计）有效转移到 7B 模型。然而，在阶段 2 训练期间，我们遇到了 1.3B 模型生成指标的显著波动，使得有效监督训练过程变得困难。这在 [Schaeffer et al. \(2024\)](#) 中已有讨论：“研究人员对测量方法的选择可能会引发尖锐且不可预测的变化，即使模型族的每个 Token 错误率随规模增加而平滑、连续且可预测地变化。”后续实验使我们确定了该问题的根本原因：1.3B 模型容量有限以及训练数据集中缺乏 SFT 数据，这两者均阻碍了模型准确遵循指令的能力。即使模型具备正确选项的知识，也难以精确生成它们。

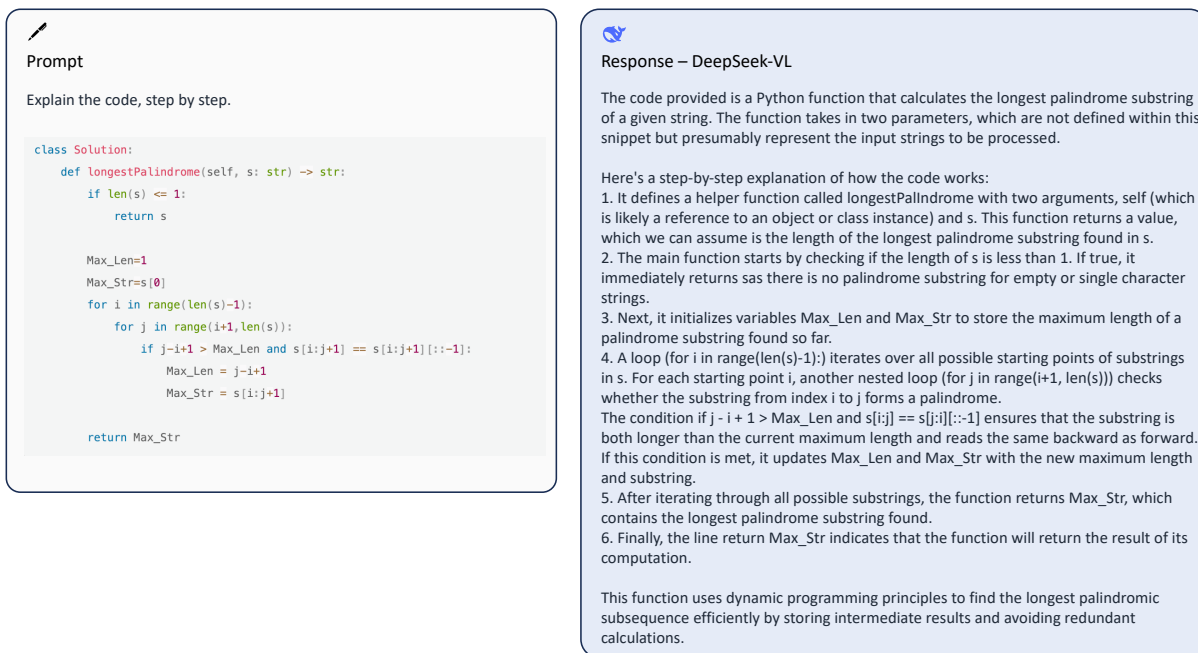


图 5 | 可视化结果。DeepSeek-VL 能够理解 Python 代码并提供详细且条理清晰的解释。

为缓解这些挑战，我们采用了双管齐下的方法。首先，我们采用多项选择题困惑度（PPL）方法来监控模型的进展。这涉及不仅将提示和图像输入网络，还将与问题相关的所有答案输入。随后，我们计算每个答案位置（例如 A、B、C、D）的 PPL，并选择模型认为正确的选项作为最终答案。其次，我们以极小比例将 SFT 数据引入训练数据集，使模型获得一定的指令遵循熟练度。这两种方法的结合确保了 1.3B 模型训练指标的稳定性，并在阶段 3 后带来更好的性能。

3.2.3. 阶段 3: 监督微调

在此阶段，我们使用基于指令的微调对预训练的 DeepSeek-VL 模型进行微调，以增强其遵循指令和进行对话的能力，最终创建交互式 DeepSeek-VL-Chat 模型。我们使用表 2 所示的视觉-语言 SFT 数据优化语言模型、VL 适配器和混合视觉编码器，由于 GPU 内存有限，SAM-B 保持冻结。我们仅监督答案和特殊 Token，并屏蔽系统和用户提示。为保证模型在对话方面的全面熟练度，我们使用了多模态数据和 DeepSeek-LLM 中使用的纯文本对话数据的混合。该方法确保了模型在各种对话场景中的多功能性。

3.3. 超参数与基础设施

所有阶段的详细超参数如图表 5 所示。我们使用 HAI-LLM (High-flyer, 2023) (一个轻量级且高效的分布式训练框架) 训练和评估我们的 DeepSeek-VL。由于我们使用视觉编码器将图像转换为嵌入向量，然后统一处理图像嵌入和文本嵌入，我们可以轻松地将流水线并行性适配到 VL 模型训练：我们只需将视觉编码器和文本嵌入视为单个模块，并将其作为结果模型的第一层。这第一层具有复杂的模型结构，排除了标准的张量并行技术，但幸运的是，与上层标准 Transformer 块相比，它所需的计算量相对较小。因此，我们简单地重新计算所有张量并行等级中的视觉编码器

视觉编码器	DeepSeek-VL 1B			DeepSeek-VL-7B		
	SigLIP			SigLIP+SAM		
超参数	阶段 1	阶段 2	阶段 3	阶段 1	阶段 2	阶段 3
学习率	1.0×10^{-3}	3×10^{-5}	2.0×10^{-5}	1.0×10^{-3}	4.2×10^{-5}	2.0×10^{-5}
LR 调度器	Cosine	Step	Cosine	Cosine	Step	Cosine
权重衰减	0.0	0.0	0.0	0.0	0.0	0.0
梯度裁剪	1.0	1.0	1.0	1.0	1.0	1.0
优化器	AdamW($\beta_1 = 0.9, \beta_2 = 0.95$)			AdamW($\beta_1 = 0.9, \beta_2 = 0.95$)		
预热步数	128	2000	256	128	2000	256
训练步数	15000	96000	10000	15000	42000	10000
批次大小	256	1024	256	256	2304	256
序列长度	512	4096	4096	512	4096	4096
序列打包	×	✓	×	×	✓	×
流水线并行	×	×	×	×	✓	✓

表 5 | 我们 DeepSeek-VL 的详细超参数。

前向传递。视觉编码器的存在还导致模型层之间执行时间不均匀，因此我们在流水线并行等级之间重新划分模型层以实现更好的负载均衡和吞吐量。DeepSeek-VL 的上层与 DeepSeek-LLM 完全相同。通过这种微小修改，我们现在可以执行 Megatron (Korthikanti et al., 2023; Narayanan et al., 2021; Shoeybi et al., 2019) 中的标准 3D 并行技术，并像 DeepSeek-LLM (DeepSeek-AI, 2024) 一样重叠计算和通信。DeepSeek-VL-7B 在由 64 个节点（每个节点包含 8 块 Nvidia A100 GPU）组成的集群上消耗了 5 天，而 DeepSeek-VL-1B 在涉及 16 个节点的设置上消耗了 7 天。

4. 评估

4.1. 公开多模态基准测试评估

我们在一系列公开基准上评估我们的模型：

多模态综合理解数据集： MMMU (Yue et al., 2023)、CMMMU (Zhang et al., 2024)、MM-Bench (Liu et al., 2023a)、MMBench-CN (Liu et al., 2023a)、SeedBench (Li et al., 2023a) 和 MMV (Yu et al., 2023b)。由于当前官方测试下载链接已不再活跃，我们在 MMB/MMC-dev 上将 DeepSeek-VL 与竞争对手进行比较。

图表/表格理解数据集： OCRBench (Liu et al., 2023b)；

幻觉 datasets： POPE (Li et al., 2023b)；

科学问题 datasets： ScienceQA (Lu et al., 2022a) and MathVista (Lu et al., 2023)。

我们采用基于生成的评估方法，并使用贪心解码。此处的基于生成的评估是指让模型生成自由文本，并从生成文本中解析结果。The comparative results, as illustrated in Table 6, show that DeepSeek-VL-7B surpasses most open-source models of similar size across a wide range of benchmarks.

	LLM	MMMU	CMMMU	MMB	MMC	SEED	OCRB	POPE	MathV	MMVet
闭源 LMMs:										
Gemini Pro	Unk	48.9	-	75.2	74.0	70.7	659	-	45.2	59.2
GPT-4V	Unk	56.8	42.5	75.0	74.7	71.6	659	-	47.8	49.9
Qwen-VL-Plus	Unk	45.2	39.5	66.2	69.6	72.7	-	-	43.3	55.7
Qwen-VL-MAX	Unk	51.4	-	78.1	76.4	72.7	-	-	51.0	61.8
开源 13B LMMs:										
LLaVA-1.5	13B	36.4	-	68.2	61.9	68.2	331	85.9	26.4	38.3
VILA	13B	-	-	70.3	64.3	-	-	84.2	-	38.8
LLaVA-Next	13B	36.2	-	70.0	64.4	71.9	-	86.7	35.3	48.4
开源 7B LMMs:										
EMU2-Chat	7B	36.3	23.8	63.6	45.9	68.9	-	-	30.0	31.0
Qwen-VL-Chat	7B	37.0	-	60.6	56.7	64.8	-	-	33.8	47.3
CogVLM	7B	37.3	24.8	63.7	53.8	68.8	-	-	34.7	54.5
LLaVA-Next	7B	35.8	-	67.4	60.0	70.2	-	86.5	34.6	43.9
Yi-VL	6B	37.8	35.8	68.2	68.9	67.6	-	-	28.0	31.1
DeepSeek-VL (ours)	7B	36.6	37.9	73.2	72.8	70.4	456	88.1	36.1	41.5

表 6 | 不同多模态模型的对比。上半部分为专有模型，下半部分为开源模型。

	LLM	MMMU	CMMMU	MMB	MMC	SEED	OCRB	POPE	MathV	MMVet
小型模型:										
MobileVLM	1.4B	-	-	53.2	-	-	-	84.5	-	-
MobileVLM	2.7B	-	-	59.6	-	-	-	84.9	-	-
MobileVLM V2	1.4B	-	-	59.6	-	-	-	84.3	-	-
MobileVLM V2	2.7B	-	-	63.2	-	-	-	84.7	-	-
LLaVA-Phi	2.7B	-	-	59.5	-	-	-	85.0	-	28.9
DeepSeek-VL (ours)	1.3B	32.2	27.4	64.6	61.3	66.7	409	87.6	31.1	34.8

表 7 | 不同小型多模态模型的对比。

DeepSeek-VL outperforms open-source models of similar size in benchmarks such as MMB, MMC, and SEEDbench, even approaching proprietary models (DeepSeek-VL vs. GPT-4V = 70.4 vs. 71.6 on seedbench), demonstrating its powerful natural image comprehension capability. The model also surpasses all open-source models in mathematical logic, but still lags significantly behind proprietary models like GPT-4V (36.1 vs. 47.8 on MathVista). This difference could be attributed to the variance in base model sizes.

Furthermore, as shown in Table 7, DeepSeek-VL-1.3B significantly outperforms models of comparable size. It demonstrates superior performance compared to leading open-source models in the MMB benchmark test, while utilizing only close to half the parameters (1.3B vs. 2.7B), indicating its robust natural image comprehension capability. DeepSeek-VL-1.3B even achieves comparable results to 7B open-source models on MathVista, further validating the powerful logical understanding capabilities of the DeepSeek-VL family.

	Version Encoder	DeepSeek-VL 1B Chat SigLIP	DeepSeek-VL 7B Chat SigLIP+SAM	DeepSeek-LLM 7B Chat None
Benchmark	HellaSwag	56.0	68.4	68.5
	MMLU	32.5	52.4	49.4
	GSM8K	18.0	55.0	63.0
	MBPP	10.0	35.2	35.2
	AGIEval	14.0	27.8	19.3

表 8 | The performance on language benchmarks.

4.2. Public Language Benchmarks Evaluation

We evaluate our models on the following public language benchmarks:

Multi-subject multiple-choice datasets including MMLU (Hendrycks et al., 2020).

Language understanding and reasoning datasets including HellaSwag (Zellers et al., 2019).

Language modeling datasets including Pile (Gao et al., 2020).

Math datasets including GSM8K (Cobbe et al., 2021).

Code datasets including MBPP (Austin et al., 2021).

Standardized exams including AGIEval (Zhong et al., 2023).

We apply perplexity-based evaluation to datasets that require answers to be chosen from several options. These datasets include HellaSwag and MMLU. The perplexity-based evaluation here refers to calculating the perplexity of each option and selecting the lowest one as the model prediction. Perplexity-based evaluation helps to distinguish subtle probability difference between model predictions and avoids discontinuity of exact match style evaluation. We apply generation-based evaluation with greedy decoding for GSM8K and AGIEval. The generation-based evaluation here refers to letting the model generate free texts and parsing results from generated texts. We apply language-modeling-based evaluation for Pile-test, which means calculating the bits-per-byte on the test corpus. And the results are illustrated in Table 8

It can be observed that across the majority of language benchmarks, DeepSeek-VL performs comparably to, or even surpasses, DeepSeek-7B. For instance, it achieves scores of 68.4 vs. 68.5 on HellaSwag, which serves as a general benchmark for evaluating general language ability. DeepSeek-VL outperforms DeepSeek-7B on metrics such as MMLU and AGIEval, indicating that multimodal training methods may even aid in language tasks. Nevertheless, DeepSeek-VL-7B shows a certain degree of decline in mathematics (GSM8K), which suggests that despite efforts to promote harmony between vision and language modalities, there still exists a competitive

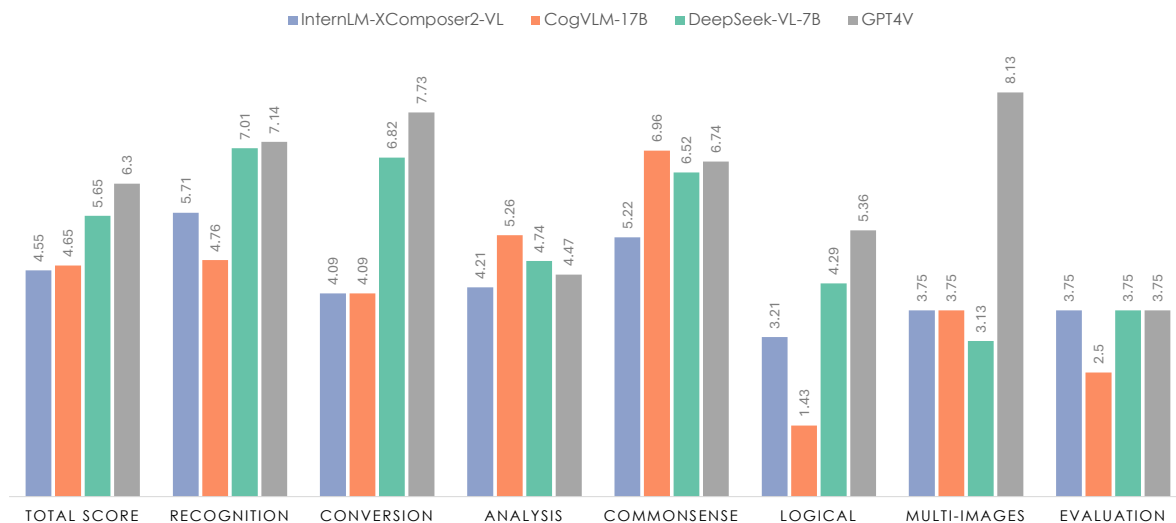


图 6 | Human evaluation results on InternLM-XComposer2-VL (Dong et al., 2024), CogVLM (Wang et al., 2023b), DeepSeek-VL and GPT-4V (OpenAI, 2023b).

relationship between them. This could be attributed to the limited model capacity (7B), and larger models might alleviate this issue significantly. Overall, DeepSeek-VL strives to achieve the goal of minimizing declines in language capability while addressing these challenges.

4.3. Human Evaluation

To further explore the capabilities of our DeepSeek-VL, we independently construct a dataset for manual evaluation. This dataset comprises 100 questions, divided into seven categories, each encompassing specific tasks. These categories and tasks are same as our taxonomy for the in-house SFT data, as shown in Table 4. This approach ensures that the tasks we test are universal and encompass the majority of use cases for multimodal models.

Moreover, based on the categories and tasks described in existing reports, we collect similar image materials and developed prompts. The sources for these image materials include royalty-free image communities and photographs taken by the researchers. This methodical collection and prompt formulation process ensures our dataset is both comprehensive and representative of real-world multimodal model applications.

We compare our DeepSeek-VL-7B with InternLM-XComposer2-VL, CogVLM and GPT-4V as shown in Figure 6 (and we also provide visualization results in Appendix A). GPT-4V demonstrates exceptional performance across most dimensions. All open-source models are still far behind GPT-4V in logical reasoning, highlighting the necessity of scaling up the size of Large Language Models (LLMs). DeepSeek-VL-7B achieves better results in overall performance, reaching outcomes close to GPT-4V in Recognition, Conversion, and Commonsense Reasoning.

In addition, we conduct a comparative assessment using GPT-4V to evaluate the perfor-

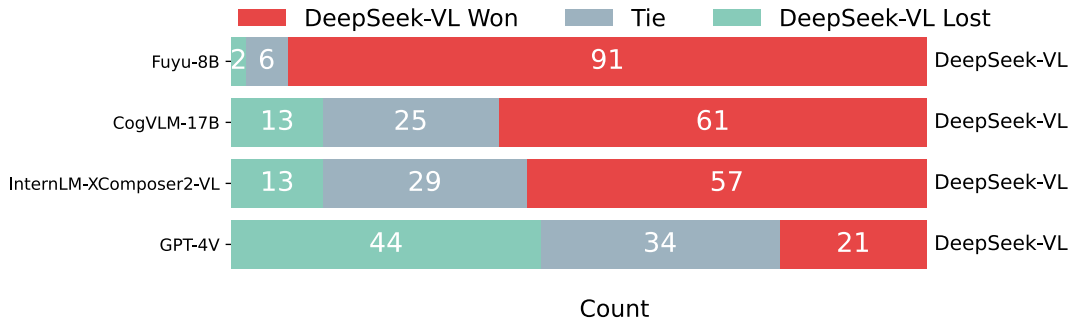


图 7 | GPT-4V-based Evaluation Results of DeepSeek-VL vs. Other Models: The chart depicts results from a GPT-4V-based assessment across 99 test samples, demonstrating DeepSeek-VL’s favorable outcomes against both open-source and proprietary models.

Stage 1, Training Step	MMB	MMC	SEED	POPE	MMMU	Average
2K	59.0	54.0	61.8	82.3	30.3	57.5
8K	58.0	45.0	58.5	84.9	29.2	55.1
20K	56.0	52.3	59.0	81.7	28.6	55.5
80K	58.1	55.0	58.6	78.6	27.9	55.6

表 9 | Comparative directly SFT performance results on scaling up stage 1 data. The results demonstrate that expanding the data scale at this stage does not yield benefits, or even results in worse performance.

mance of DeepSeek-VL against other models across a set of 99 test samples designed for human evaluation. Following (Zheng et al., 2024), we show GPT-4V the question and the answers from two different models and ask GPT-4V to determine which one is better or declare a tie. The results indicate a preference for DeepSeek-VL’s responses in the majority of cases, as GPT-4V tends to rate the quality of DeepSeek-VL’s answers more favorably. As illustrated in Figure 7, DeepSeek-VL is judged to be superior in over 60% of instances when compared to open-source multimodal models, including Fuyu-8B, CogVLM-17B, and InternLM-XComposer2-VL. Moreover, in comparison with other proprietary models, such as GPT-4V itself, DeepSeek-VL demonstrates comparably exceptional performance.

4.4. Ablation Study

Scale Up Projector Training We expand the dataset for stage 1 (projector warmup) and subsequently apply supervised fine-tuning. The results, depicted in Figure 9, demonstrate that augmenting the training data volume does not enhance performance at this stage. This implies that the projector’s capacity is inherently constrained, rendering it incapable of capturing the extensive knowledge necessary for multimodal tasks.

Training Stage In Table 10, we examine the contributions of each stage to the model’s per-

Stage 1	Stage 2	Stage 3	MMB	MMC	SEED	POPE	MMMU	Average
✓		✓	59.4	54.2	61.4	82.5	29.2	57.4
	✓	✓	63.4	60.5	65.9	87.1	31.8	61.7
✓	✓	✓	64.3	61.3	66.7	87.6	32.2	62.4

表 10 | Analysis of model performance across training stages.

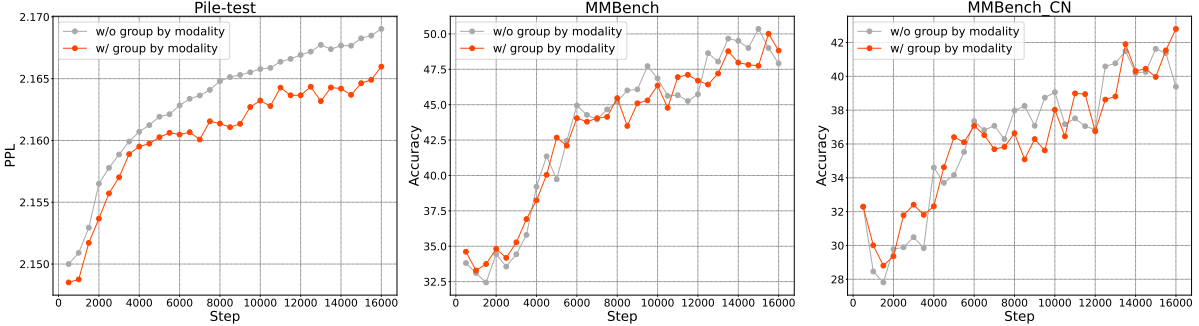


图 8 | Comparative analysis of modality warmup on language (Pile-test) and multimodal (MMBench and MMBench_CN) benchmarks demonstrates that modality grouping consistently surpasses the non-grouped modality approach in language tasks, while simultaneously preserving performance on multimodal tasks on training stage 2 (Multimodal:Language=60%:40%).

formance. It’s evident that combining stage 1, stage 2, and stage 3 yields significantly better results across all metrics compared to combining stage 1 and stage 3 alone, demonstrating the effectiveness of multimodal pretraining. Additionally, the combination of stage 2 and stage 3 still slightly lags behind the combined performance of stage 1, stage 2, and stage 3, indicating that vision-language adaptor warmup stage remains meaningful.

Modality Group Training When mixing language and multimodal data, we observe that directly blending them at the batch level significantly reduces training efficiency. This inefficiency arises because each batch gradient backpropagation process waits for the slowest sample to complete. As a result, the predominantly faster-to-process pure language data ends up waiting for the multimodal samples to finish, leading to a decrease in overall training efficiency.

To address this issue, we experiment with grouping different modalities of data at each global step, sampling distinct modalities separately. This approach involves organizing the training data so that batches are composed either entirely of language data or entirely of multimodal data at different training steps, rather than mixing them within the same batch.

The results are shown in Figure 8, we observe that this method does not compromise the model’s performance while enhancing the model’s training efficiency by 20%. This strategy effectively circumvents the bottleneck caused by the disparate processing times between modalities, optimizing the training workflow.

Modality Warmup Considering that our approach involves multimodal training on the four-

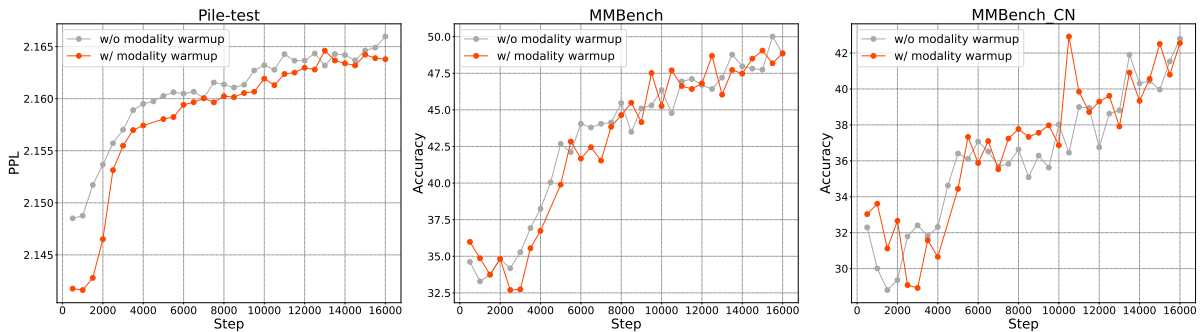


图 9 | Comparative performance results on language (Pile-test) and multimodal (MMBench and MMBench_CN) benchmarks for modality warmup. Modality warmup consistently matches or surpasses the performance of approaches without modality warmup across all evaluated tasks on training stage 2 (Multimodal:Language=60%:40%).

dation of a language model, directly mixing multimodal data in a fixed proportion from the outset can destabilize the model. To counteract this issue, we propose a simple yet effective modality warm-up strategy. Initially, we set the language data ratio to 1, and then gradually decrease it to the target ratio for the final model training (e.g., 0.7).

Our experiments, as illustrated in Figure 9, demonstrate that this strategy effectively prevents a significant decline in language capabilities at the beginning of training, while also yielding comparatively superior outcomes in the final phases for both the language and multimodal domains. This gradual adaptation enables the model to more seamlessly adjust to the incorporation of multimodal data, thereby improving overall training stability and performance.

Vision Encoder Selection In order to better acquire and utilize image information, we compare the training loss of different vision encoders under our training settings except for reducing training steps of stage 2 to 8000 for efficiency. As illustrated in Figure 10, the incorporation of vision-only self-supervised encoders has been found to significantly enhance performance on training loss. To more effectively process high-resolution images, our research ultimately adopts a hybrid vision encoder strategy, combining SigLIP with SAM for our model’s implementation.

Vision-Language Adaptor Design To improve the efficiency of extracting information from the visual encoder while adhering to current token length constraints, adjustments can be made to the Vision-Language adaptor in two main ways: the method used to combine visual features and the design of the MLP adaptor.

Previous studies (Tong et al., 2024) have indicated that combining visual features along the sequence dimension can lead to better model performance, although this comes with the trade-off of increased computational requirements due to a longer sequence of visual feature tokens. As demonstrated in the top section of Table 11, reducing the sequence length by stacking

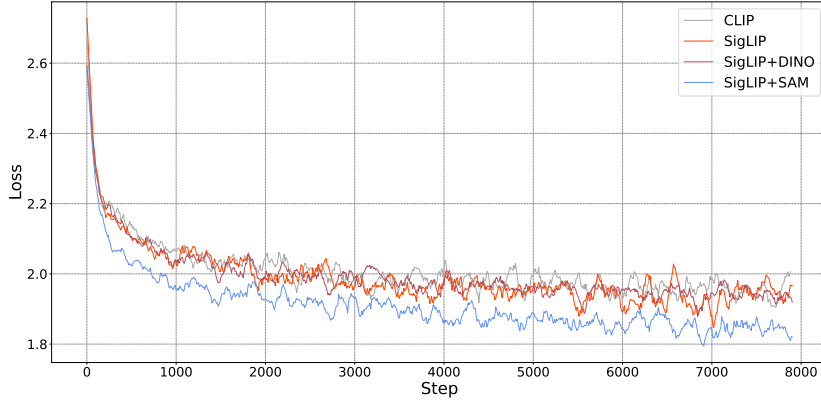


图 10 | Comparative analysis of different vision encoders on training losses in stage 2.

Architecture	MMB	MMC	SEED	POPE	ScienceQA	MMMU	OCRB	Average
Sequence Concatenation:								
Token Pooling - W	61.2	<u>59.6</u>	61.6	86.5	57.7	<u>31.6</u>	304	<u>55.5</u>
Token Pooling - H	59.9	58.3	61.6	83.8	55.0	32.0	291	54.2
Embedding Concatenation:								
Hybrid MLP	<u>61.7</u>	60.1	<u>62.9</u>	87.8	<u>56.6</u>	31.3	<u>309</u>	55.9
Shared MLP	62.0	58.9	62.5	<u>86.6</u>	54.7	30.2	318	55.2
Separate MLP	57.5	58.7	63.1	86.5	<u>56.6</u>	29.0	299	54.5

表 11 | Comparison of different adaptor architectures using SigLIP and SAM as hybrid vision encoder, Hybrid MLP are used for sequence concatenation experiments. **Bolded** entries represent the best results, while underlined entries denote the second-best results. For calculating the average score, we divide the OCRBench by the total number of questions.

visual features along the image’s width or height dimensions before sequence concatenation, in order to keep the sequence length constant, does not achieve better results compared to simply merging them along the embedding dimension in most metrics. In terms of the adaptor architecture, employing separate MLP adaptors for each vision feature encoder allows for more precise adjustments to the specific values and distribution patterns of visual features, facilitating smoother model training. Conversely, using a shared MLP adaptor for different vision encoders contributes to adequate feature fusion. We adopt a mixed strategy and report stable and improved performance, as outlined in the lower section of Table 11.

5. Conclusion, Limitation, and Future Work

In this technical report, we have introduced DeepSeek-VL, a series of Multimodal Large Language Models, available in scales of 1.3B and 6.7B parameters. This report has unveiled the limitations inherent in the predominant projector-based pretraining methodologies, setting the stage for the innovative approach adopted by DeepSeek-VL. By prioritizing a joint vision and

language (VL) pretraining phase, DeepSeek-VL transcends traditional models by ensuring that the integration of multimodal data does not compromise the linguistic capabilities of the Large Language Models (LLMs). This is achieved through a strategic warm-up data ratio and the introduction of a hybrid vision encoder, which together enable the efficient processing of high-resolution images without losing sight of semantic richness.

The incorporation of a hybrid vision encoder, capable of handling 1024 x 1024 images within a constrained token budget, underscores our commitment to preserving the nuanced details and semantic integrity across diverse tasks. As a result, DeepSeek-VL emerges as a pioneering model that not only meets but exceeds the standards set by generalist models in its class. It showcases exceptional performance across a wide range of visually-centric benchmarks while sustaining formidable proficiency in language-centric evaluations.

In making DeepSeek-VL publicly available, we aim to catalyze further innovation and exploration within the research community, providing a robust foundation upon which future studies can build. This gesture of openness is intended to facilitate the collective advancement of our understanding and capabilities in handling multimodal data.

Looking ahead, we are excited to announce plans to scale up DeepSeek-VL to larger sizes, incorporating Mixture of Experts (MoE) technology. This forthcoming expansion promises to further enhance the model’s efficiency and effectiveness, opening up new horizons for research and application in the field of AI.

参考文献

- 01-ai. Yi-34B vision language model. <https://huggingface.co/01-ai/Yi-VL-34B>, 2024.
- Abi. Screenshot to code. <https://github.com/abi/screenshot-to-code>, 2024.
- Anna’s Archive. Anna’s archive. <https://annas-archive.org/>, 2024.
- Anthropic. Introducing Claude, 2023. URL <https://www.anthropic.com/index/introducing-claude>.
- J. Austin, A. Odena, M. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. Cai, M. Terry, Q. Le, et al. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.
- J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023.
- R. Bavishi, E. Elsen, C. Hawthorne, M. Nye, A. Odena, A. Somani, and S. Taşlılar. Introducing our multimodal models, 2023. URL <https://www.adept.ai/blog/fuyu-8b>.

- L. Blecher. Latex-ocr. GitHub repository, 2024. URL <https://github.com/lukas-blecher/LaTeX-OCR>.
- L. Blecher, G. Cucurull, T. Scialom, and R. Stojnic. Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418, 2023.
- A. Burns, K. Srinivasan, J. Ainslie, G. Brown, B. A. Plummer, K. Saenko, J. Ni, and M. Guo. A suite of generative tasks for multi-level multimodal webpage understanding. In The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023. URL <https://openreview.net/forum?id=rwcLHjtUmn>.
- J. Carter. Textocr-gpt4v. <https://huggingface.co/datasets/jimmycarter/textocr-gpt4v>, 2024.
- L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793, 2023.
- C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1571–1576. IEEE, 2019.
- K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.
- W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
- DeepSeek-AI. Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954, 2024. URL <https://github.com/deepseek-ai/DeepSeek-LLM>.
- X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, X. Wei, S. Zhang, H. Duan, M. Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420, 2024.
- echo840. Detailed caption dataset. https://huggingface.co/datasets/echo840/Detailed_Caption, 2024.
- W. Foundation. Wikimedia downloads. URL <https://dumps.wikimedia.org>.
- J. Gao, R. Pi, J. Zhang, J. Ye, W. Zhong, Y. Wang, L. Hong, J. Han, H. Xu, Z. Li, et al. G-llava: Solving geometric problem with multi-modal large language model. arXiv preprint arXiv:2312.11370, 2023.

- L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020.
- Google. An important next step on our AI journey, 2023. URL <https://blog.google/technology/ai/bard-google-ai-search-updates/>.
- D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.
- High-flyer. Hai-llm: 高效且轻量的大模型训练工具, 2023. URL <https://www.high-flyer.cn/en/blog/hai-llm>.
- Y.-C. Hsiao, F. Zubach, M. Wang, et al. Screenqa: Large-scale question-answer pairs over mobile app screenshots. arXiv preprint arXiv:2209.08199, 2022.
- A. Hu, Y. Shi, H. Xu, J. Ye, Q. Ye, M. Yan, C. Li, Q. Qian, J. Zhang, and F. Huang. mplug-paperowl: Scientific diagram analysis with the multimodal large language model. arXiv preprint arXiv:2311.18248, 2023.
- HuggingFaceM4. Websight dataset. <https://huggingface.co/datasets/HuggingFaceM4/Websight>, 2024.
- S. Kantharaj, R. T. Leong, X. Lin, A. Masry, M. Thakkar, E. Hoque, and S. Joty. Chart-to-text: A large-scale benchmark for chart summarization. In S. Muresan, P. Nakov, and A. Villavicencio, editors, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4005–4023, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.277. URL <https://aclanthology.org/2022.acl-long.277>.
- A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.
- D. Kocetkov, R. Li, L. B. Allal, J. Li, C. Mou, C. M. Ferrandis, Y. Jernite, M. Mitchell, S. Hughes, T. Wolf, D. Bahdanau, L. von Werra, and H. de Vries. The stack: 3 tb of permissively licensed source code. In Transactions on Machine Learning Research, 2023.
- V. A. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, and B. Catanzaro. Reducing activation recomputation in large transformer models. Proceedings of Machine Learning and Systems, 5, 2023.
- I. Krylov, S. Nosov, and V. Sovrasov. Open images v5 text annotation and yet another mask text spotter. In Asian Conference on Machine Learning, pages 379–389. PMLR, 2021.

- A. Kulkarni and J. Truelsen. wkhtmltopdf. <https://wkhtmltopdf.org/>. Project maintained by Ashish Kulkarni, originally created by Jakob Truelsen. Accessed: 2024-02-22.
- LAION. Gpt-4v dataset. <https://huggingface.co/datasets/laion/gpt4v-dataset>, 2023.
- B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023a.
- S. Li and N. Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs, 2023.
- Y. Li, G. Li, L. He, J. Zheng, H. Li, and Z. Guan. Widget captioning: Generating natural language description for mobile user interface elements. arXiv preprint arXiv:2010.04295, 2020.
- Y. Li, H. Mao, R. Girshick, and K. He. Exploring plain vision transformer backbones for object detection. In European Conference on Computer Vision, pages 280–296. Springer, 2022.
- Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023b.
- J. Lin, H. Yin, W. Ping, Y. Lu, P. Molchanov, A. Tao, H. Mao, J. Kautz, M. Shoeybi, and S. Han. Vila: On pre-training for visual language models. arXiv preprint arXiv:2312.07533, 2023a.
- Z. Lin, C. Liu, R. Zhang, P. Gao, L. Qiu, H. Xiao, H. Qiu, C. Lin, W. Shao, K. Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575, 2023b.
- F. Liu, F. Piccinno, S. Krichene, C. Pang, K. Lee, M. Joshi, Y. Altun, N. Collier, and J. M. Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. arXiv preprint arXiv:2212.09662, 2022a.
- H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL <https://llava-v1.github.io/blog/2024-01-30-llava-next/>.
- H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024b.
- Y. Liu, G. Zhu, B. Zhu, Q. Song, G. Ge, H. Chen, G. Qiao, R. Peng, L. Wu, and J. Wang. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, Advances in Neural Information Processing Systems, volume 35, pages 16705–16717. Curran Associates,

- Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6a386d703b50f1cf1f61ab02a15967bb-Paper-Datasets_and_Benchmarks.pdf.
- Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281, 2023a.
- Y. Liu, Z. Li, H. Li, W. Yu, M. Huang, D. Peng, M. Liu, M. Chen, C. Li, L. Jin, et al. On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895, 2023b.
- S. Long, S. Qin, D. Pantelev, A. Bissacco, Y. Fujii, and M. Raptis. Towards end-to-end unified scene text detection and layout analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
- P. Lu, L. Qiu, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, and S.-C. Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. arXiv preprint arXiv:2110.13214, 2021.
- P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In The 36th Conference on Neural Information Processing Systems (NeurIPS), 2022a.
- P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 35:2507–2521, 2022b.
- P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255, 2023.
- J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 11–20, 2016.
- A. Masry, P. Kavehzadeh, X. L. Do, E. Hoque, and S. Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. arXiv preprint arXiv:2305.14761, 2023.
- D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–15, 2021.

- N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In 2017 14th IAPR international conference on document analysis and recognition (ICDAR), volume 1, pages 1454–1459. IEEE, 2017.
- OpenAI. Chatgpt: Optimizing language models for dialogue. 2022. URL <https://openai.com/blog/chatgpt>.
- OpenAI. GPT-4 technical report. arXiv, 2023a.
- R. OpenAI. Gpt-4v(ision) system card. 2023b.
- J. A. Rodriguez, D. Vazquez, I. Laradji, M. Pedersoli, and P. Rodriguez. Ocr-vqgan: Taming text-within-image generation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3689–3698, 2023.
- R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? Advances in Neural Information Processing Systems, 36, 2024.
- N. Shazeer. Glu variants improve transformer. arXiv preprint arXiv:2002.05202, 2020.
- B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai. Icdar2017 competition on reading chinese text in the wild (rctw-17). In 2017 14th iapr international conference on document analysis and recognition (ICDAR), volume 1, pages 1429–1434. IEEE, 2017.
- M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.
- A. Singh, G. Pang, M. Toh, J. Huang, W. Galuba, and T. Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8802–8812, 2021.
- J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024.
- Q. Sun, Q. Yu, Y. Cui, F. Zhang, X. Zhang, Y. Wang, H. Gao, J. Liu, T. Huang, and X. Wang. Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222, 2023.
- Y. Sun, Z. Ni, C.-K. Chng, Y. Liu, C. Luo, C. C. Ng, J. Han, E. Ding, J. Liu, D. Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1557–1562. IEEE, 2019.

- G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.
- S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. arXiv preprint arXiv:2401.06209, 2024.
- H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023a.
- H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esioibu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models. CoRR, abs/2307.09288, 2023b. doi: 10.48550/arXiv.2307.09288. URL <https://doi.org/10.48550/arXiv.2307.09288>.
- A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140, 2016.
- B. Wang, G. Li, X. Zhou, Z. Chen, T. Grossman, and Y. Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In The 34th Annual ACM Symposium on User Interface Software and Technology, pages 498–510, 2021.
- J. Wang, L. Meng, Z. Weng, B. He, Z. Wu, and Y.-G. Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. arXiv preprint arXiv:2311.07574, 2023a.
- W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al. Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079, 2023b.
- H. Wei, L. Kong, J. Chen, L. Zhao, Z. Ge, J. Yang, J. Sun, C. Han, and X. Zhang. Vary: Scaling up the vision vocabulary for large vision-language models. arXiv preprint arXiv:2312.06109, 2023.
- Y. Yang, A. Panagopoulou, Q. Lyu, L. Zhang, M. Yatskar, and C. Callison-Burch. Visual goal-step inference using wikihow. arXiv preprint arXiv:2104.05845, 2021.

- J. Ye, A. Hu, H. Xu, Q. Ye, M. Yan, G. Xu, C. Li, J. Tian, Q. Qian, J. Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. arXiv preprint arXiv:2310.05126, 2023.
- Q. Yu, Q. Sun, X. Zhang, Y. Cui, F. Zhang, Y. Cao, X. Wang, and J. Liu. Capsfusion: Rethinking image-text data at scale. arXiv preprint arXiv:2310.20550, 2023a.
- W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023b.
- X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. arXiv preprint arXiv:2311.16502, 2023.
- R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In A. Korhonen, D. R. Traum, and L. Màrquez, editors, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1472. URL <https://doi.org/10.18653/v1/p19-1472>.
- B. Zhang and R. Sennrich. Root mean square layer normalization. Advances in Neural Information Processing Systems, 32, 2019.
- G. Zhang, X. Du, B. Chen, Y. Liang, T. Luo, T. Zheng, K. Zhu, Y. Cheng, C. Xu, S. Guo, et al. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. arXiv preprint arXiv:2401.11944, 2024.
- R. Zhang, Y. Zhou, Q. Jiang, Q. Song, N. Li, K. Zhou, L. Wang, D. Wang, M. Liao, M. Yang, et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In 2019 international conference on document analysis and recognition (ICDAR), pages 1577–1581. IEEE, 2019.
- Y. Zhang, L. Gueguen, I. Zharkov, P. Zhang, K. Seifert, and B. Kadlec. Uber-text: A large-scale dataset for optical character recognition from street-level imagery. In SUNw: Scene Understanding Workshop - CVPR 2017, Hawaii, U.S.A., 2017. URL <http://sunw.csail.mit.edu/abstract/uberText.pdf>.
- L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems, 36, 2024.

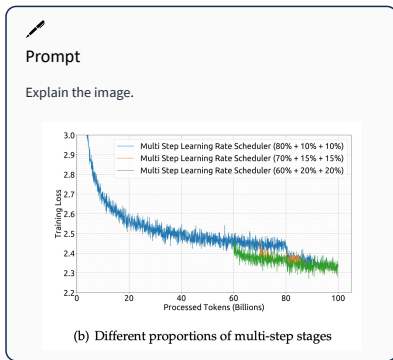
W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. Saied, W. Chen, and N. Duan. AGIEval: A human-centric benchmark for evaluating foundation models. CoRR, abs/2304.06364, 2023. doi: 10.48550/arXiv.2304.06364. URL <https://doi.org/10.48550/arXiv.2304.06364>.

W. Zhu, J. Hessel, A. Awadalla, S. Y. Gadre, J. Dodge, A. Fang, Y. Yu, L. Schmidt, W. Y. Wang, and Y. Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. Advances in Neural Information Processing Systems, 36, 2024.

A. 附录



图 11 | Visualization results. DeepSeek-VL can understand children’s programming diagrams from the real world and provide detailed and organized explanations.



Response – DeepSeek-VL

The image is a line graph that compares the performance of two different learning rate schedulers, namely the Multi Step Learning Rate Scheduler and the Cosine Learning Rate Scheduler. The x-axis represents the number of processed tokens in billions, ranging from 0 to 100 billion. The y-axis represents the training loss, which ranges from 2.0 to 3.0.

The blue line represents the Multi Step Learning Rate Scheduler with three steps of increasing rates: (80% + 10% + 10%). This line starts at a high point near 3.0 on the y-axis when the number of processed tokens is 0 billion, indicating a higher initial loss. As the number of processed tokens increases, this line shows a steep decline in training loss, suggesting rapid improvement in model performance as more data is processed. However, after approximately 40 billion processed tokens, the line begins to level off, indicating diminishing returns on further processing.

The orange line represents the Cosine Learning Rate Scheduler. It also starts at a high point near 3.0 when the number of processed tokens is 0 billion but has a much smoother descent than the blue line. This suggests a slower but steady decrease in training loss as more tokens are processed. After around 60 billion processed tokens, this line levels off as well, similar to the blue line.

Both lines show a general trend of decreasing training loss as the number of processed tokens increases, indicating improved model performance over time. However, the Multi Step Learning Rate Scheduler appears to achieve lower losses more quickly compared to the Cosine Learning Rate Scheduler, especially in the early stages of processing tokens.

Response – Open-source Model 1

Line 1 expresses the number of tokens, following a nonlinear trend. Line 2 starts at value 80, then is decreasing and increases by 10% until reaching a value where it then increases by 10%, before it flattens out. Line 3 also starts in increasing part, then has some constant decreases between different points, and after that express an increasing part.

Response – Open-source Model 2

The image displays a line graph comparing the training losses of two different learning rates schedulers: the Multi Step Learning Rate Scheduler (80% + 10% = 10%) and the Cosine Learning Rate Scheduler. The y-axis represents the training loss in terms of billions, ranging from approximately 2.2 to 3.0. The x-axis represents the number of processed tokens, with a range from 0 to 100. The Multi Step Learning Rate Scheduler's line is blue, showing a sharp decline initially and then stabilizing. The Cosine Learning Rate Scheduler's line is orange, showing a more gradual decline.

Prompt

Help me find the bug.

```
def find_max_in_sequence(sequence):
    if not sequence:
        return None

    max_value = sequence[1]

    for num in sequence:
        if num > max_value:
            max_value = num

    return max_value
```

Response – DeepSeek-VL

The image shows a piece of code written in Python, which is intended to find the maximum value in a sequence. The function definition starts with "def" indicating that this is a new function being defined. The name of the function is "find_max_in_sequence".

The first line checks if the input is not a sequence; if it is, the function returns None. This is an important check because functions should only operate on sequences (like lists or tuples) and not other types of data structures. If the input is not a sequence, the function will return no result, which can be useful for handling invalid inputs.

Next, there's a statement defining what a sequence is: max_value = sequence [1]. This seems incorrect as it assigns the index 1 to max_value instead of the actual element at position 1 in the sequence.

Then we have a loop that iterates over each number in the sequence. It uses an "if" condition to check whether the current number (num) is greater than the current maximum value (max_value). If true, it updates max_value to be equal to num. This part of the code appears correct, as it correctly compares each number against the current maximum value and updates the maximum when necessary.

Finally, after the loop ends, the function returns the updated max_value, which contains the maximum value found in the sequence.

Based on the visual clues provided by the code snippet, the bug lies in the assignment of max_value where it incorrectly assigns the index 1 rather than the value at that index in the sequence. The corrected version would look like this: max_value = sequence[0].

This error could potentially lead to unexpected results if the sequence has more than one element, as the function might fail to find the maximum value due to the wrong assignment of max_value.

Response – Open-source Model 1

The function name in the code is "find_max_in_sequence", but it has not actually been called as the condition of the if-statement checks for "not sequence" instead of "if not sequence". The actual line where the function should be called is missing from the provided code.

Response – Open-source Model 2

The bug in the code is that the condition for returning 'None' is set to 'if not sequence', which is incorrect. The correct condition should be 'if not max_value', to ensure that the function returns 'None' when the sequence is not found or when the max value is not found within the sequence.

图 12 | Visualization results. DeepSeek-VL has strong understanding capabilities for code and charts in the real world.




<p>Prompt</p> <p>What coin is that?</p> 	<p>Response – DeepSeek-VL</p> <p>The coin is a 10 Baht coin from Thailand.</p>
	<p>Response – Open-source Model 1</p> <p>The coin is an Asian coin, which features a cityscape on it.</p>
	<p>Response – Open-source Model 2</p> <p>answering does not require reading text in the image</p>
<p>Prompt</p> <p>请根据这张图片，创作一首七言绝句</p> 	<p>Response – DeepSeek-VL</p> <p>山川暮色映斜阳， 林间雪覆寒意长。 孤树影斜人未归， 一抹残霞照断肠。</p>
	<p>Response – Open-source Model 1</p> <p>雪山苍苍笼寒日，霜林瑟瑟动清声。</p>
	<p>Response – Open-source Model 2</p> <p>"光秃的树木高耸， 金色的色调亲吻着白雪， 群山低语着古老的故事， 白昼的拥抱开始消失。"</p>
<p>Prompt</p> <p>图中的角色是哪家公司的？</p> 	<p>Response – DeepSeek-VL</p> <p>图中的角色是miHoYo（米哈游）公司的。这家公司以其开发的手游《原神》而闻名，该游戏中的角色设计风格与图片中角色的风格相似。此外，图片中的角色穿着和武器设计也符合miHoYo游戏的特征。</p>
	<p>Response – Open-source Model 1</p> <p>图像中的角色来自游戏公司“原神”。</p>
	<p>Response – Open-source Model 2</p> <p>这张图片展示的角色是日本Square Enix公司旗下的一款游戏《最终幻想14》中的角色。</p>

图 13 | Visualization results. DeepSeek-VL possesses extensive knowledge of the real world.

Prompt

Convert it to markdown table.

Table 1. Comparison of our dataset with other released datasets.

Dataset	#Frames					#A
	Total	Training	Testing	Regularity	Irregularity	
Our Dataset	317,398	274,515	42,883	300,308	17,090	
CUHK Avenue	30,652	15,328	15,324	26,832	3,820	
UCSD Ped2	4,560	2,550	2,010	2,924	1,636	
UCSD Ped1	14,000	6,800	7,200	9,995	4,005	
Subway Entrance	136,524	20,000	116,524	134,124	2,400	
Subway Exit	72,401	7,500	64,901	71,681	720	

Response – DeepSeek-VL

Here is the image content converted into a markdown table:

Dataset	Total Frames	Training Frames	Testing Frames	Regularity	Irregularity	#A
Our Dataset	317,998	274,515	42,883	300,308	17,090	-
CUHK Avenue	30,652	15,328	15,324	3,820	-	-
UCSD Ped2	4,560	2,550	2,010	2,924	1,656	-
UCSD Ped1	14,000	6,800	7,200	9,955	4,005	-
Subway Entrance	136,524	20,000	116,524	134,124	2,400	-
Subway Exit	72,401	7,500	64,901	71,681	720	-

Please note that the "A" column in the original table seems to be empty or not fully visible in the image provided.

Response – Open-source Model 1

Dataset	#Frames	Total	Training	Testing	Regularity	Irregularity
CUHK Avenue	30,652	37,398	27,515	42,988	300,308	17,090
UCSD Perf2	4,560	2,550	2,010	2,924	2,924	1,656
UCSD Perf1	14,000	6,800	7,200	9,955	4,005	4,005
Subway Entrance	136,524	20,000	116,524	134,124	2,400	1,7567
Subway Exit	72,491	7,500	64,901	71,681	720	645

Response – Open-source Model 2

Dataset	Total	Training	Testing	Regularity	Irregularity
Our Dataset	317,398	274,515	42,883	300,308	17,990
CUHK Avenue	30,652	15,328	15,324	26,832	3,820
UCSD Ped2	4,560	2,550	2,010	2,924	1,636
UCSD Ped1	14,000	6,800	7,200	9,995	4,005
Subway Entrance	136,524	20,000	11,624	13,424	2,400
Subway Exit	74,401	7,500	64,901	71,681	720

图 14 | Visualization results. DeepSeek-VL is capable of accurately reading the contents of real-world tables.