DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding

DeepSeek-VL2 is a new series of open-source Vision-Language Models that leverages the Mixture-of-Experts (MoE) architecture to achieve substantial improvements in both performance and efficiency compared to its predecessor, DeepSeek-VL.

Dynamic Tiling Strategy

Addressing Fixed Resolution Limitations

DeepSeek-VL2 introduces a dynamic tiling strategy that efficiently processes high-resolution images of varying aspect ratios. This approach improves over DeepSeek-VL's hybrid vision encoder, which extracted features from images at two fixed resolutions (384 × 384 and 1024 × 1024).

Enhanced Visual Understanding

This approach avoids the limitations of the old fixed-size encoder and excels in tasks requiring ultra-high resolution, including visual grounding, document/table/chart analysis, and detailed feature extraction, while maintaining a manageable number of visual tokens.

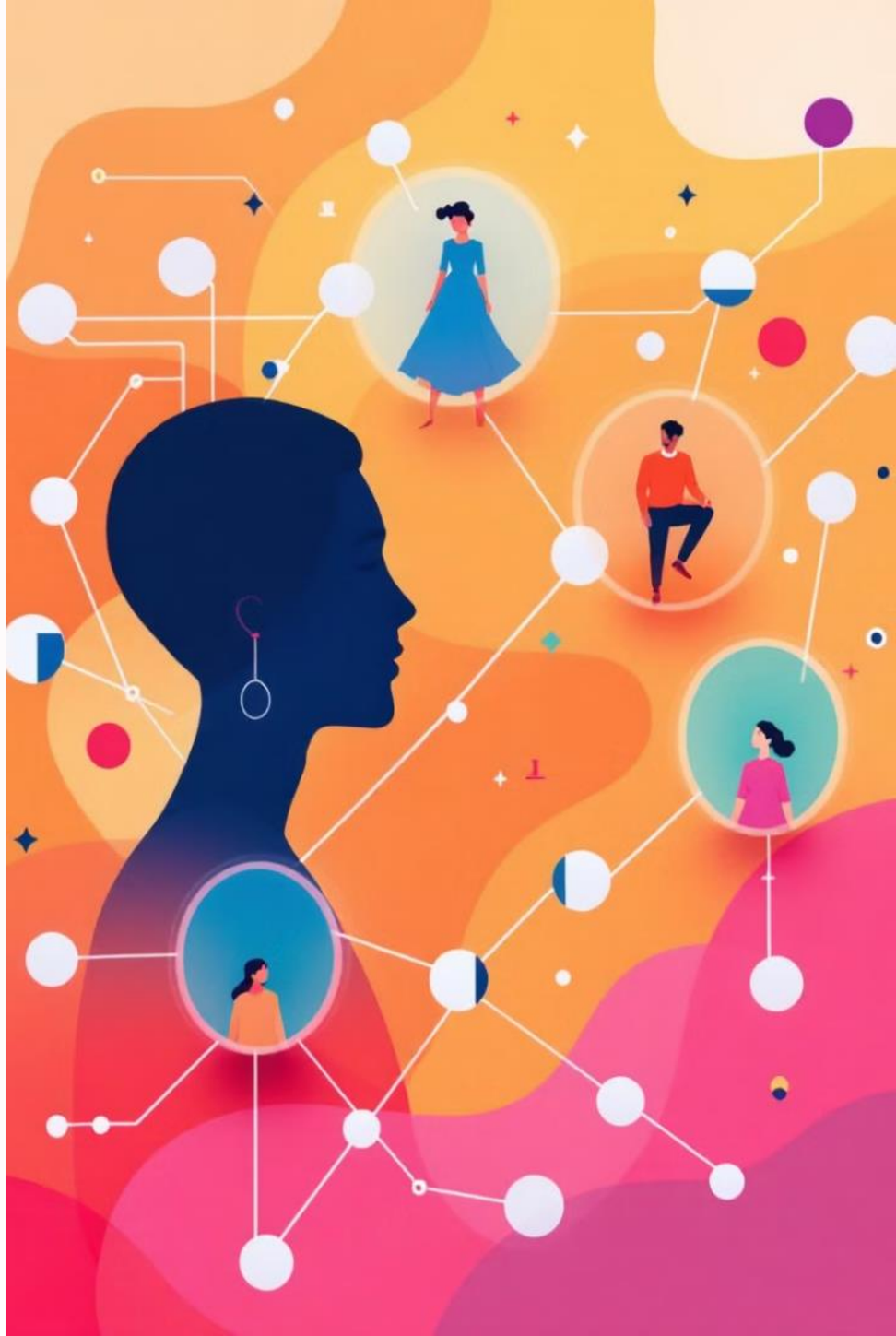
DeepSeekMoE LLM

1 Multi-head Latent Attention

This language model is based on DeepSeekMoE, which incorporates the Multi-head Latent Attention mechanism. MLA enhances inference efficiency by compressing the Key-Value cache into a latent vector, enabling increased throughput capacity.

2 MoE Architecture

The model also incorporates a MoE architecture allowing for efficient inference through sparse computation. During MoE training, they introduced a global bias term for each expert to cost-effectively improve load balancing between experts.



Vision-Language Pretraining Data

Interleaved Image-Text Data

Deepseek's data collection begins with several open-sourced datasets, including WIT, WikiHow, and 30% random samples from OBELICS. To enhance multilingual capabilities, deepseek supplemented the predominantly English datasets with Chinese content extracted from Wanjuan.

Image Captioning Data

Image captions represent fundamental data in VLM training, providing direct alignment between visual and textual information. Deepseek initially leveraged diverse open-source datasets, but our preliminary analysis revealed severe quality variations across these datasets.

Optical Character Recognition Data

To develop OCR capabilities, deepseek used open-source datasets including LaTeX OCR and 12M RenderedText. Deepseek combined these datasets with an extensive in-house OCR dataset covering diverse document types.

Supervised Fine-Tuning Data

General Visual Question-Answering

While public visual QA datasets are diverse, they often suffer from three main limitations: (1) short responses, (2) poor OCR quality, and (3) hallucinated content. To address these issues, Deepseek regenerate responses by jointly considering the original questions, images, and OCR information.

OCR and Document Understanding

DeepSeek-VL2 already demonstrates superior OCR capabilities compared to other state-of-the-art VLMs. Therefore, rather than further enhancing OCR performance during the SFT stage, Deepseek focused on cleaning existing open-source datasets by removing samples with poor OCR quality.

Table and Chart Understanding

Deepseek enhanced table-based QA data by regenerating responses for all public datasets based on their original questions except Cauldron, which already exhibits high quality. Similar to its OCR capabilities developed during VL pretraining, deepseek's model demonstrated strong performance in chart understanding without requiring additional efforts.

Training Methodology

Vision-Language Alignment

1

Building upon pre-trained language models (DeepSeekMoE 3B/16B/27B), Deepseek’s primary objective is to establish robust connections between visual features and language features. This alignment enables the pre-trained language model to effectively handle visual inputs.

Vision-Language Pre-training

2

After establishing the vision-language alignment in the embedding space, Deepseek dedicate the majority of their computational resources to vision-language pre-training. This stage focuses on developing comprehensive joint vision-language knowledge across diverse tasks.

Supervised Fine-Tuning

3

In the final stage, Deepseek enhance the pre-trained model's instruction-following and conversational capabilities through supervised fine-tuning. Using deepseek’s in-house vision-language SFT data, it optimized all parameters while supervising only the answers and special tokens, masking both system and user prompts.

Multimodal Performance

1

Benchmarks

Holistic evaluation of DeepSeek-VL2 across a collection of commonly used benchmarks, including DocVQA, ChartQA, InfoVQA, TextVQA, RealWorldQA, OCRBench, AI2D, MMMU, MMStar, MathVista, MME, MMBench, MMBench-V1.1 and MMT-Bench.

2

Comparison with State-of-the-Art Models

On the multimodal understanding benchmarks, they compared DeepSeek-VL2 with state-of-the-art models, including LLaVA-OV, InternVL2, DeepSeek-VL, Qwen2-VL, Phi-3.5-Vision, Molmo, Pixtral, MM1.5 and Aria-MoE.

3

Superior Performance

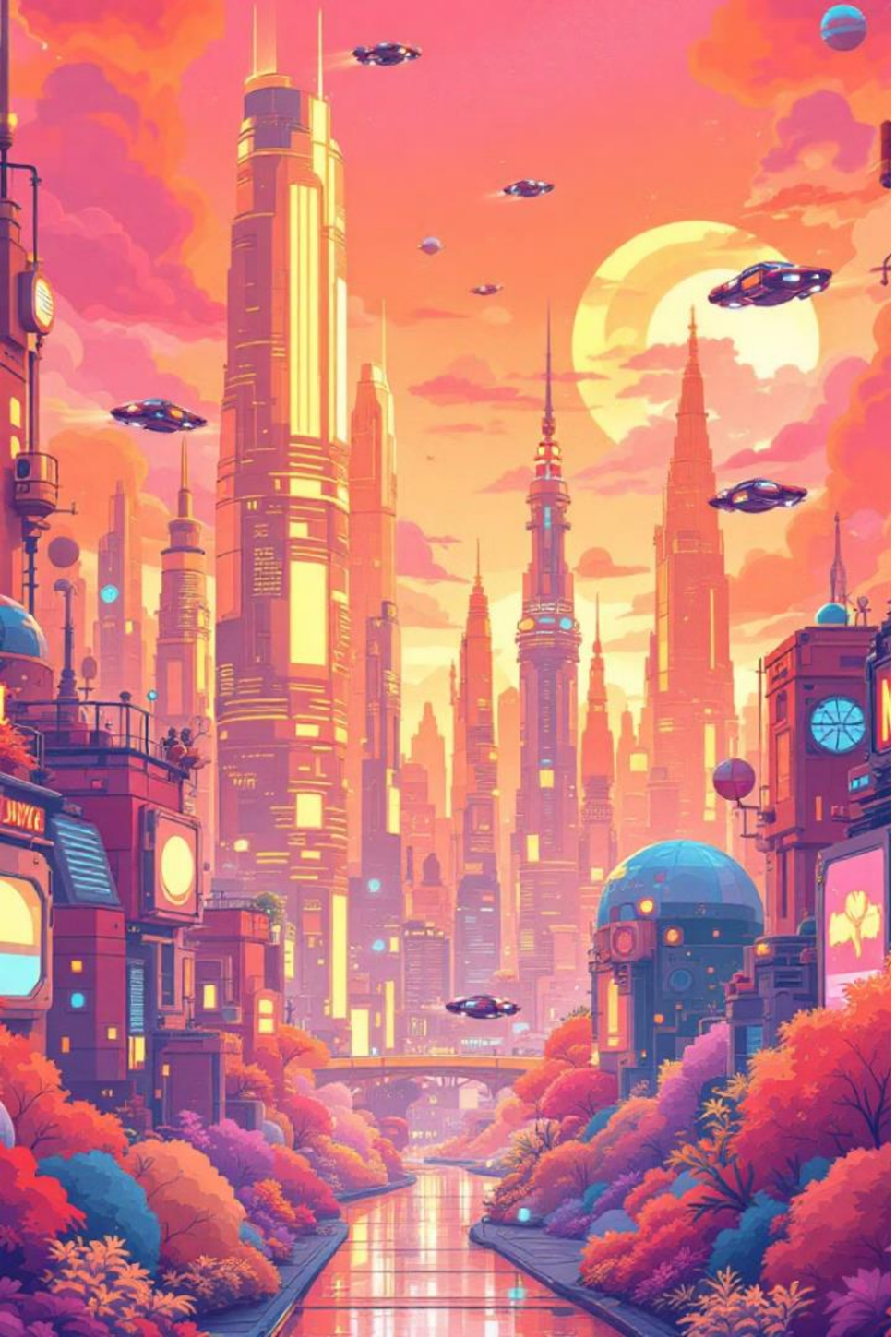
Benefited from MoE architecture, DeepSeek-VL2 achieves similar or better performance with fewer activated parameters. On the grounding benchmarks, they compared DeepSeek-VL2 with Groudning DINO, UNINEXT, ONE-PEACE, mPLUG-2, Florence-2, InternVL2, Shikra, TextHawk2, Ferret-v2, MM1.5 and Qwen2. Deepseek’s models outperforms the other VLMs at similar scales.

Table 3 | **Comparison with state-of-the-art models on OCR-related multimodal benchmarks.** †: activated parameters of MoE model.

Model	#Params (LLM)	#Params (VE)	#Params (Activated)	DocVQA (test)	ChartQA (test)	InfoVQA (test)	TextVQA (val)	OCRBench
Closed Model								
GPT-4V [69]	-	-	-	87.2	78.1	75.1	78.0	645
GPT-4o [32]	-	-	-	92.8	85.7	79.2	77.4	736
Claude 3.5 Sonnet [5]	-	-	-	95.2	90.8	74.1	74.1	788
Gemini-1.5-Pro [81]	-	-	-	93.1	87.2	80.1	78.7	754
Open-source Model (0.5B - 3B)								
LLaVA-OV 0.5B [45]	0.5B	0.4B	0.9B	70.0	61.4	41.8	-	-
InternVL2-1B [16]	-	-	0.9B	81.7	72.9	50.9	70.5	754
MM 1.5-1B [107]	-	-	1B	81.0	67.2	50.5	72.5	605
DeepSeek-VL2-Tiny	0.6B†	0.4B	1.0B†	88.9	81.0	66.1	80.7	809
MolmoE-1B [22]	1.2B†	0.3B	1.5B†	77.7	78.0	53.9	78.8	
MiniCPM-V 2.0 [99]	2.4B	0.4B	2.8B	71.9	-	-	74.1	605
InternVL2-2B [16]	1.9B	0.3B	2.2B	86.9	76.2	58.9	73.4	784
Qwen2-VL-2B [88]	1.5B	0.7B	2.2B	90.1	73.5	65.5	79.7	794
MM 1.5-3B [107]	-	-	3B	87.7	74.2	58.5	76.5	657
DeepSeek-VL2-Small	2.4B†	0.4B	2.8B†	92.3	84.5	75.8	83.4	834
Open-source Model (4B - 13B)								
Phi-3.5-Vision [1]	3.8B	0.3B	4.1B	69.3	81.8	36.6	72.0	599
InternVL2-4B [16]	3.8B	0.3B	4.1B	89.2	81.5	67.0	74.4	788
Aria-MoE [46]	3.9B†	0.4B	4.3B†	92.6	86.4	-	81.1	-
MM 1.5-7B [107]	-	-	7B	88.1	78.6	59.5	76.5	635
LLaVA-OV 7B [45]	7.6B	0.4B	8.0B	87.5	80.0	68.8	-	-
Molmo-7B-O [22]	7.3B	0.3B	7.6B	-	80.4	70.0	80.4	-
MiniCPM-V2.6 [99]	7.6B	0.4B	8.0B	90.8	82.4	-	80.1	852 (CoT)
InternVL2-8B [16]	7.7B	0.3B	8.0B	91.6	83.3	74.8	77.4	794
Qwen2-VL-7B [88]	7.6B	0.7B	8.3B	94.5	83.0	76.5	84.3	845
Pixtral-12B [3]	12.0B	0.4B	12.4B	90.7	81.8 (CoT)	50.8	75.7	
DeepSeek-VL 7B [59]	6.9B	0.4B	7.3B	-	-	-	-	456
DeepSeek-VL2	4.1B†	0.4B	4.5B†	93.3	86.0	78.1	84.2	811

Table 4 | **Comparison with state-of-the-art models on general QA and math-related multimodal benchmarks.** †: activated parameters of MoE model. *: evaluated in a different setting.

Model	#Params (Activated)	MMStar	AI2D (test)	MMMU (val)	MME	MMBench (sum)	MMBench (en test)	MMBench-V1.1 (cn test)	MMT-Bench	RealWorldQA	MathVista (testmini)
Closed Model											
GPT-4V [69]	-	56.0	89.4	63.1	1,927	81	80.2	80	64.3	61.4	58.1
GPT-4o [32]	-	63.9	94.2	69.1	2,329	83.4	82.1	82.2	65.5	75.4	63.8
Claude 3.5 Sonnet [5]	-	62.2	94.7	68.3	1,920	79.7	80.7	78.5	-	60.1	67.7
Gemini-1.5-Pro [81]	-	-	94.4	62.2	-	-	-	-	64.5	70.4	63.9
Open-source Model (0.5B - 3B)											
LLaVA-OV 0.5B [45]	0.9B	37.7	57.1	31.4	1,478	61.6	55.5	59.6	-	55.6	34.8
InternVL2-1B [16]	0.9B	45.7	64.1	35.4	1,794	65.4	60.7	61.6	49.5	50.3	37.7
MM 1.5-1B [107]	1B	-	59.3	35.8	1,611	-	-	-	-	53.3	37.2
DeepSeek-VL2-Tiny	1.0B†	45.9	71.6	40.7	1,915	73.3	69.2	68.3	53.2	64.2	53.6
MolmoE-1B [22]	1.5B†	-	86.4*	34.9	-	-	-	-	-	60.4	34
MiniCPM-V 2.0 [99]	2.8B	-	-	38.2	1,809	69.6	68.1	-	-	-	38.7
InternVL2-2B [16]	2.2B	49.8	74.1	36.3	1,877	73.2	70.9	69.6	50.4	57.3	46.3
Qwen2-VL-2B [88]	2.2B	48	74.4	41.1	1,872	74.9	73.5	72.2	54.5	62.9	47.8
MM 1.5-3B [107]	3B	-	65.7	37.1	1,798	-	-	-	-	56.9	44.4
DeepSeek-VL2-Small	2.8B†	57.0	80.0	48.0	2,123	82.3	80.3	79.3	62.9	65.4	60.7
Open-source Model (4B - 13B)											
Phi-3.5-Vision [1]	4.1B	47.5	78.1	43	-	76	66.1	72.1	53.6	53.6	43.9
InternVL2-4B [16]	4.1B	54.3	78.9	47.9	2,060	78.6	73.9	75.8	55.7	60.7	58.6
Aria-MoE [46]	4.3B†	-	-	54.9	-	-	-	-	-	-	66.1
MM 1.5-7B [107]	7B	-	72.2	41.8	1,861	-	-	-	-	62.5	47.6
LLaVA-OV 7B [45]	8.0B	-	81.4	48.8	1,998	80.8	-	-	-	66.3	63.2
Molmo-7B-O [22]	7.6B†	-	90.7*	39.3	-	-	-	-	-	67.5	44.5
MiniCPM-V2.6 [99]	8.0B	57.5	82.1	49.8 (CoT)	2,348 (CoT)	81.5	79.3	78.0	60.8	65.0	60.6
InternVL2-8B [16]	8.0B	61.5	83.8	51.8	2,210	81.7	81.2	79.4	60.0	64.4	58.3
Qwen2-VL-7B [88]	8.3B	60.7	83	54.1	2,327	83	80.5	80.7	63.7	70.1	58.2
Pixtral-12B [3]	12.4B	-	-	52.5 (CoT)	-	-	-	-	-	65.4	58 (CoT)
DeepSeek-VL 7B [59]	7.3B	-	-	36.6	-	73.2	-	-	-	-	-
DeepSeek-VL2	4.5B†	61.3	81.4	51.1	2,253	83.1	79.6	79.2	63.6	68.4	62.8



Conclusion

DeepSeek-VL2 demonstrates strong capabilities across various tasks, including general question answering, visual storytelling, and visual grounding. It aim to extend the context window in its next version to enable richer multi-image interactions. Moreover, it aim to strengthen its reasoning capabilities. These identified areas guide the ongoing research directions as they continue to advance the model's capabilities.