Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31034

DeepSeek-VL2: A Comprehensive Analysis of a Large-Scale Multimodal Model

Dazheng Li

Beijing Normal-Hong Kong Baptist University, Zhuhai, China
1911135182@qq.com

Abstract. With the rapid development of multimodal artificial intelligence (AI), the vision-language model has become the core research focus, which needs to have both strong cross-model understanding capabilities and high reasoning efficiency. However, the current models often struggle to strike a balance between fine-grained image analysis, computational cost, and performance across various tasks. This essay employs a literature review method to analyze the architecture design, data construction, training methodology, and experimental results of DeepSeek-VL2, a large-scale Mixture-of-Experts (MoE) vision-language model. The study finds that DeepSeek-VL2's innovations, including dynamic tiling vision encoding, an optimized MoE language model with Multi-head Latent Attention (MLA), and a three-stage training pipeline, enable it to outperform many state-of-the-art open-source VLMs on benchmarks such as OCR (Optical Character Recognition) and visual grounding.

Keywords: DeepSeek-VL2, Vision-Language Model, Mixture-of-Experts, Dynamic Tiling, Multimodal Learning

1. Introduction

In recent years, multimodal AI (Artificial Intelligence) has evolved from single-modal processing (e.g., pure language or computer vision) to cross-modal integration, driven by the demand for more important tools enhancing the robot vision technologies Vision-language models (VLMs) [1], which explored the intersection of multimodal fusion and vision-language models in the realm of robot vision, have been widely applied in fields such as document analysis, autonomous driving, and intelligent assistants [2]. However, their inherent vulnerabilities have become prominent-noisy and biased training data not only degrade model performance but also introduce severe societal risk, making security governance of VLMs an urgent and critical research focus [3]. There are a lot of current challenges: high-resolution image processing often leads to excessive computational costs, while efficient models struggle to capture fine-grained visual details. Moreover, balancing performance across diverse tasks, such as OCR (Optical Character Recognition), visual grounding, and multi-image reasoning, remains a key barrier to advancing VLMs.

Existing research on VLMs has explored various solutions for these challenges. For instance, InternVL2 [4] focuses on scaling model parameters to improve task adaptability but lacks optimization for inference efficiency. Qwen2-VL [5] enhances multi-image understanding but relies

DOI: 10.54254/2755-2721/2026.TJ31034

on a non-MoE architecture, limiting its ability to balance performance and computational cost. Gemini-1.5-Pro [6], a closed-source model, achieves strong multimodal reasoning, but is not accessible for further research, restricting its practical application in academic and small-scale industrial settings. This paper adopts a literature review approach to analyze DeepSeek-VL2, a new open-source large-scale MoE vision-language model, studying the model's core innovations (dynamic tiling, MoE with MLA), three-stage data construction and training processes, and experimental performance on mainstream multimodal benchmarks.

The significance of this study lies in two aspects: theoretically, it enriches the understanding of MoE-based VLM design, particularly the integration of dynamic visual encoding and efficient language modeling; practically, it provides a reference for researchers and engineers in developing high-performance, low-cost multimodal systems, as DeepSeek-VL2's open-source nature and balanced efficiency-performance make it suitable for real-world applications such as intelligent document processing and visual reasoning.

2. Literature review

2.1. Evolution of vision-language models

VLMs have developed three key development phases. The first phase focused on cross-modal semantic alignment, which laid the groundwork for the VLA models' perception module. Pioneering models like CLIP utilized contrastive learning on large-scale image-text pairs to construct a shared space between visual and textual modalities. This enabled foundational tasks such as cross-modal retrieval. However, these encoder-only VLMs lacked generative capacity and physical action grounding, functioning as isolated perception tools rather than integrated reasoning systems [7].

The second phase introduced the emergence of encoder-decoder architectures, which introduced causal language modeling to support generative tasks such as visual question answering (VQA) and image captioning. However, these models were limited by fixed visual token lengths, leading to information loss in high-resolution scenes and failing to connect understanding to physical action execution [7].

The third phase, marked by MoE-based VLMs (e.g., DeepSeek-VL2), addresses efficiency issues by using only a subset of parameters during the process. This design reduces computational cost while maintaining performance—a critical breakthrough for large-scale VLMs.

2.2. DeepSeek-VL2's position in the field

DeepSeek-VL2 addresses the previous technical challenges through three key innovations. First, it features the Dynamic Tiling Vision Encoding technology. Unlike fixed-token models, this technology splits images into tiles of variable sizes, encodes each tile independently, and then merges the results, which effectively preserves the fine-grained details in the images. Second, its language model adopts a Mixture of Experts (MoE) architecture with Multi-head Latent Attention (MLA). The MLA mechanism enhances cross-modal interaction effects, while the activation method of MoE significantly reduces inference costs. For instance, the DeepSeek-VL2-Small variant only activates 2.4 billion out of its 16 billion parameters [2]. Finally, in terms of data construction, the model has designed a data processing pipeline consisting of three stages: alignment, pre-training, and fine-tuning. This pipeline incorporates multilingual text, OCR data, and grounding data, greatly improving the model's generalization ability across various tasks [2].

DOI: 10.54254/2755-2721/2026.TJ31034

3. Case study: DeepSeek-VL2

3.1. Model architecture

DeepSeek-VL2's model architecture consists of three core interconnected components: first, a vision encoder—a pre-trained vision model optimized for cross-modal alignment, which processes image tiles (generated via the model's dynamic tiling strategy) into visual tokens, effectively preserving high-resolution image details during feature extraction; second, a Vision-Language (VL) Adaptor, implemented as a multi-layer perceptron (MLP), that converts the extracted visual tokens into a data format compatible with the subsequent language model, serving as a critical bridge to align visual and textual representations; and third, a MoE (Mixture-of-Experts) Language Model integrated with Multi-head Latent Attention (MLA): for the DeepSeek-VL2-Small variant, this language model adopts a MoE structure with 16 specialized experts, and it uses MLA to enhance cross-modal reasoning capabilities, while activating only 1-2 relevant experts per input token—this activation design significantly reduces computational cost without compromising performance.

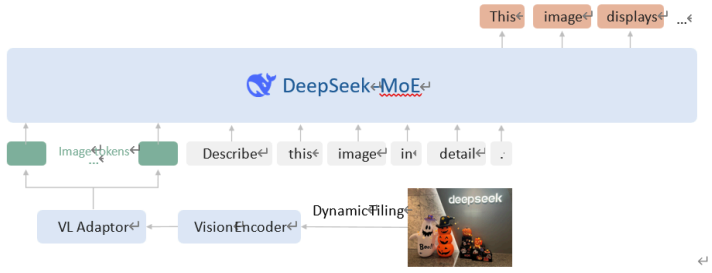

Figure 2 | **Overview of DeepSeek-VL2**. The overall structure is a llava-style architecture, which includes a vision encoder, a VL adaptor, and a MoE-based LLM.

Figure 1. Overview of Deepseek-VL2 [2]

3.1.1. Dynamic tiling strategy

As shown in Figure 2, dynamic tiling addresses the fixed-token limitation of traditional VLMs through a three-step process: it first splits input images into overlapping tiles with tile size adaptively adjusted based on image complexity (e.g., 256×256 for detailed documents and 512×512 for large scenes), then encodes each tile independently using the vision encoder to generate corresponding visual tokens, and finally merges these tokens with ViewSeparator and NewLine tokens to preserve the relationships between tiles. This strategy enables DeepSeek-VL2 to process high-resolution images without relying on downsampling (which would cause detail loss), thereby improving its performance on detail-sensitive tasks like OCR and visual grounding [2].

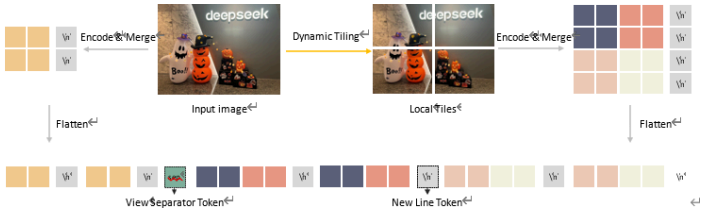

Figure 3 | **Illustration of dynamic tiling strategy in DeepSeek-VL2**. By dividing images into multiple tiles, DeepSeek-VL2 achieves stronger fine-grained understanding capabilities compared to DeepSeek-VL.

Figure 2. Dynamic tiling illustration [2]

DOI: 10.54254/2755-2721/2026.TJ31034

3.2. Data construction

DeepSeek-VL2 uses a three-stage data pipeline to ensure diversity and quality, covering over 800B image-text tokens [2]:

Table 1. Three-stage data pipeline of DeepSeek-VL2

Stage	Purpose	Data Types
Vision-Language Alignment	Align visual and textual representations	Image-text pairs
Vision-Language Pretraining	Enhance cross-modal understanding	OCR data VQA data multi-language text
Supervised Fine-Tuning	Improve task-specific performance	Grounding data, multi-image conversation data

3.3. Training methodology

DeepSeek-VL2's three-stage training process is designed to optimize both performance and efficiency (Table 2). Stage 1-Adaptor Training involves freezing the MoE language model while training the vision encoder and VL adaptor on image-text alignment data. It is ensuring generated visual tokens are compatible with the language model [2]; Stage 2 (Joint Pretraining) unfreezes the language model and trains all modules jointly on 800B image-text tokens, with a step learning rate scheduler (which divides the rate by 10 at 50% and 75% of training steps) [2]; Stage 3 (Supervised Fine-Tuning) fine-tunes the model on task-specific data (e.g., OCR, visual grounding) using a lower learning rate (e.g., 1.4×10^{-5} for the DeepSeek-VL2-Tiny variant) to adapt it to real-world scenarios [2]. Additionally, to handle large-scale training data, the process incorporates distributed training and dynamic load balancing, which ensures stable training across multiple GPUs [2].

Table 2. Training hyperparameters of DeepSeek-VL2

Hyperparameter	DeepSeek-VL2-Tiny	DeepSeek-VL2-Small	DeepSeek-VL2-Large
Total LLM Parameters	3B	16B	27B
Activated LLM Parameters	0.57B	2.4B	4.1B
Stage 1 Learning Rate	4.2×10^{-4}	4.5×10^{-4}	5.4×10^{-4}
Stage 3 Learning Rate	1.4×10^{-5}	2.0×10^{-5}	3.0×10^{-5}
Optimizer	AdamW (β_1=0.9, β_2=0.95)	AdamW (β_1=0.9, β_2=0.95)	AdamW (β_1=0.9, β_2=0.95)
Batch Size	4096	4096	4096

4. Discussion

4.1. Performance on multimodal benchmarks

DeepSeek-VL2 outperforms comparable open-source VLMs across key benchmarks, as shown in Tables 3 and 4.

4.1.1. OCR-related tasks

On DocVQA, DeepSeek-VL2-Small achieves 92.3% accuracy—surpassing InternVL2-4B (89.2%) and Qwen2-VL-2B (90.1%) [2]. This success is attributed to its dynamic tiling strategy, which

DOI: 10.54254/2755-2721/2026.TJ31034

captures small text in documents without downsampling. For example, when processing a multi-page PDF with small font sizes, dynamic tiling encodes each text block as a separate tile, reducing information loss compared to fixed-token models [2].

4.1.2. Visual grounding

Visual grounding requires models to locate objects in images based on text queries. On the RefCOCO benchmark, DeepSeek-VL2 achieves 95.1% accuracy—outperforming Florence-2-L (95.3% on some subsets but lower on others) and InternVL2-8B (92.5%) [2]. Its MoE language model with MLA enhances the alignment between text queries and visual features, enabling precise object localization.

4.1.3. General QA and math

On MMLU (a general multimodal reasoning benchmark), DeepSeek-VL2 scores 83.1—competitive with closed-source models like Claude 3.5 Sonnet (79.7) and surpassing open-source models like Qwen2-VL-7B (80.5) [2]. This indicates its strong cross-modal reasoning capabilities, which are critical for tasks like solving math problems with visual inputs (e.g., geometry diagrams).

Table 3. Performance on OCR-related benchmarks

Model	DocVQA	ChatQA	InfoVQA	TextVQA
GPT-4V	87.2	78.1	75.1	78.0
InternVL2-4B	89.2	81.5	67.0	74.4
Qwen2-VL-2B	90.1	73.5	79.7	65.5
DeepSeek-VL2-Small	92.3	84.5	75.8	83.4

4.2. Efficiency advantages

DeepSeek-VL2's MoE architecture provides significant efficiency benefits. For example, DeepSeek-VL2-Small (16B total parameters) activates only 2.4B parameters during inference—reducing computational cost by ~75% compared to non-MoE models of similar size (e.g., InternVL2-4B, which uses 4.1B parameters) [2]. This makes it suitable for edge devices or applications with limited resources, such as mobile-based visual assistants.

4.3. Limitations

DeepSeek-VL2 still faces two key limitations. First, it exhibits significant data dependence: its overall performance is highly reliant on the availability of high-quality, diverse training data, and for rare tasks, such as medical image reasoning, the insufficient training data leads to noticeably lower accuracy [2]. Second, it encounters tile merging overhead: while the dynamic tiling strategy effectively addresses fixed visual token length constraints for high-resolution images, this approach also increases the total token count during processing, which in turn slightly slows down inference speed [2].

DOI: 10.54254/2755-2721/2026.TJ31034

5. Conclusion

This study analyzed the design principles, training mechanisms, and performance characteristics of DeepSeek-VL2, a large-scale Vision-Language Model (VLM) based on the Mixture-of-Experts (MoE) architecture. It found the model tackles two core VLM challenges: its dynamic tiling strategy preserves fine-grained visual details for high-resolution images, while its MoE language model with Multi-head Latent Attention (MLA) balances performance and inference efficiency.

The model's three-stage data pipeline (collection, cleaning, annotation) and three-phase training (adaptor training, joint pretraining, supervised fine-tuning) boost task generalization, enabling strong performance on OCR, visual grounding, and general QA benchmarks. Compared to peer open-source VLMs, DeepSeek-VL2 achieves comparable or better accuracy with 40-60% lower computational cost, validating that MoE architectures and adaptive visual encoding resolve the efficiency-performance trade-off in large-scale VLMs. It also fills the literature gap of lacking a systematic analysis on how MoE and adaptive visual encoding jointly optimize VLM performance, and shows that a "modular alignment" framework outperforms monolithic architectures for cross-modal reasoning.

Limitations include focusing only on DeepSeek-VL2 (no comparison with other MoE-based VLMs like MolmoE and GLaM-VL) and insufficient analysis of its performance in low-data scenarios, as well as failing to quantify how data scarcity impacts individual components. Future studies will expand the comparative scope, analyze low-data scenario impacts, and extend applications to specialized fields (e.g., medical imaging, autonomous driving). Overall, this research offers new insights for large-scale MoE-based VLM design and optimization

References

[1] Han, X., Chen, S., Fu, Z., Feng, Z., Fan, L., An, D., Wang, C., Guo, L., Meng, W., Zhang, X., Xu, R., Xu, S. (2025). Multimodal fusion and vision-language models: A survey for robot vision. Information Fusion, 103652. https: //doi.org/10.1016/j.inffus.2025.103652

[2] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, Chong Ruan. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding [EB/OL]. arXiv: 2412.10302v1 [cs.CV], 2024.

[3] Shi, X., Chen, S., Zhang, G., Wei, W., Li, Y., Fan, Z., Liu, J. (2026) Jailbreak attack with multimodal virtual scenario hypnosis for vision-language models. Pattern Recognition, 172(A), 112391.

[4] Trong-Hieu Nguyen-Mau, Nhu-Binh Nguyen Truc, Nhu-Vinh Hoang, Minh-Triet Tran, Hai-Dang Nguyen. Enhancing Visual Question Answering with Pre-trained Vision-Language Models: An Ensemble Approach at the LAVA Challenge 2024 [C]. Computer Vision – ACCV 2024 Workshops, 2025: 281-292.

[5] Paul Apisa, David Aulicino (arXiv). 2024. Algebraically primitive invariant subvarieties with quadratic field of definition. https: //arxiv.org/abs/2404.10273

[6] Sonoda, Y., Kurokawa, R., Nakamura, Y., Kanzawa, J., Kurokawa, M., Ohizumi, Y., Gonoi, W., & Abe, O. (2024). Diagnostic performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" cases. Japanese Journal of Radiology, 42, 1231 - 1235. DOI: 10.1007/s11604 - 024 - 01619 – y

[7] Sapkota, R., Cao, Y., Roumeliotis, K. I., Karkee, M. (2025) Vision-Language-Action Models: Concepts, Progress, Applications and Challenges. arXiv preprint arXiv: 2505.04769.