

MagicVL-2B: Empowering Vision-Language Models on Mobile Devices with Lightweight Visual Encoders via Curriculum Learning

**Yi Liu^{*†}, Xiao Xu*, Zeyu Xu*, Meng Zhang*, Yibo Li*, Haoyu Chen*,
Junkang Zhang, Qiang Wang, Jifa Sun, Siling Lin,
Shengxun Cheng, Lingshu Zhang, Kang Wang[✉]**

Honor Device Co., Ltd
wangkang12@honor.com

Abstract

Vision-Language Models (VLMs) have achieved remarkable breakthroughs in recent years, enabling a diverse array of applications in everyday life. However, the substantial computational and storage demands of VLMs pose significant challenges for their efficient deployment on mobile devices, which represent the most ubiquitous and accessible computing platforms today. In this work, we introduce **MagicVL-2B**, a novel VLM meticulously optimized for flagship smartphones. MagicVL-2B leverages a lightweight visual encoder with fewer than 100M parameters and features a redesigned dynamic resolution scheme that adaptively generates image tokens without excessive modification of image dimensions. To further enhance the performance of this compact encoder within VLMs, we propose a multimodal curriculum learning strategy that incrementally increases task difficulty and data information density throughout training. This approach substantially improves the model's performance across a variety of sub-tasks. Extensive evaluations on standard VLM benchmarks demonstrate that MagicVL-2B matches the accuracy of current state-of-the-art models while reducing on-device power consumption by **41.1%**. These results establish MagicVL-2B as a practical and robust solution for real-world mobile vision-language applications, enabling advanced multimodal intelligence to run directly on smartphones.

Introduction

In recent years, Vision-Language Models (VLMs) (Achiam et al. 2023; Chen et al. 2024c,b; Bai et al. 2023; Wang et al. 2024; McKinzie et al. 2024; Zhang et al. 2024; Tong et al. 2024b) have achieved remarkable breakthroughs, enabling a wide range of real-world applications. These advances have empowered richer human-computer interactions and deeper contextual understanding, resulting in more intuitive and intelligent user experiences. However, the substantial computational and memory demands of VLMs pose a significant barrier to their seamless deployment on mobile devices—the most ubiquitous and user-friendly computing platforms today (Ding et al. 2024; Qu et al. 2024; Hua et al. 2024; Yao et al. 2024; Xue et al. 2024).

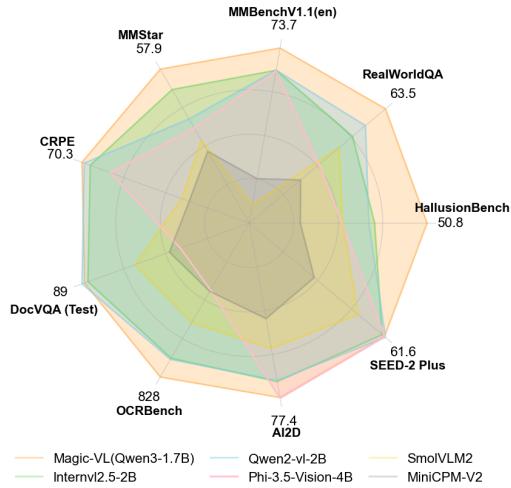
^{*}These authors contributed equally.

[†]Project Leader, [✉]Corresponding Author.

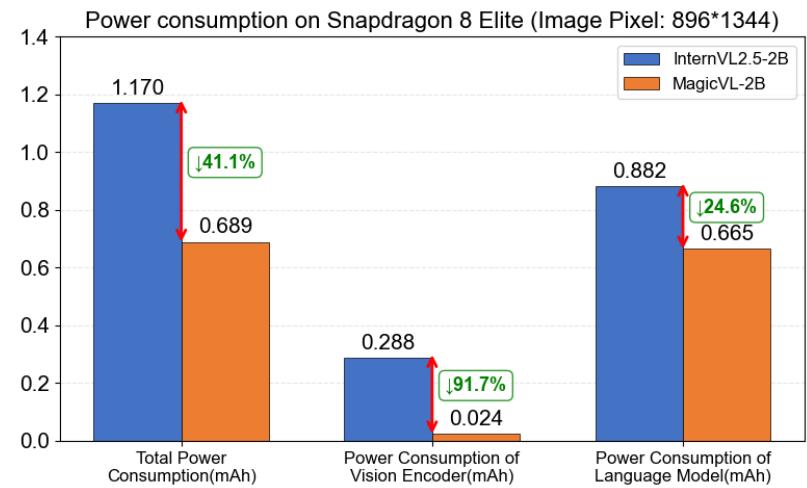
Among all computing platforms, smartphones are especially well-positioned to benefit from VLMs, as they support real-time on-device inference, enabling instant interactions and enhanced privacy (Ding et al. 2024; Qu et al. 2024). Deploying VLMs on mobile devices also greatly improves model accessibility, allowing users to conveniently access advanced multimodal models in daily scenarios such as augmented reality, real-time translation, and smart assistants (Hua et al. 2024; Chu et al. 2023).

Despite these advantages, deploying VLMs efficiently on smartphones remains challenging. First, limited memory capacity restricts the deployment of large-scale models, affecting their representational power and accuracy. Second, the constrained computational capability of mobile processors limits inference speed and energy efficiency. Third, mainstream VLMs typically adopt large Vision Transformer (ViT) encoders, which, due to hardware constraints, result in higher power consumption for visual encoding on-device compared to GPUs in the cloud. Few works leverage lightweight vision encoders to reduce on-device power consumption, likely because such encoders are more difficult to align with Large Language Model (LLM) capabilities, leading to suboptimal performance.

To address these challenges, we present **MagicVL-2B**, an innovative VLM specifically optimized for flagship smartphones. In terms of algorithmic design, MagicVL-2B employs a lightweight visual encoder tailored for efficient on-device inference, with fewer than 100M ViT parameters, as illustrated in Figure 2(a). This significantly reduces the power consumption of visual encoding on mobile devices. To further unleash the potential of lightweight encoders, we curate a large-scale multimodal dataset and introduce a curriculum learning strategy. By progressively increasing the information density and task difficulty during training, we substantially enhance the capabilities of VLMs with lightweight visual encoders, while maintaining fast and low-power inference. As shown in Figure 1, MagicVL-2B achieves both state-of-the-art performance and superior efficiency compared to existing lightweight VLMs. Specifically, MagicVL-2B consistently outperforms or matches other small-scale models on a wide range of challenging multimodal benchmarks, while signifi-



(a) Comparison with mainstream MLLMs



(b) Comparison with power and latency

Figure 1: **Comparison between MagicVL-2B and mainstream VLMs.** (a) MagicVL-2B demonstrates competitive performance across a wide range of multimodal benchmarks, matching or even surpassing other small-scale models. (b) MagicVL-2B achieves substantially lower inference power consumption and latency compared to InternVL2.5-2B, underscoring its efficiency and practicality for real-world deployment.

cantly reducing **41.1%** inference total power consumption. This remarkable combination of accuracy and efficiency highlights MagicVL-2B as a highly practical and scalable solution for real-world multimodal applications, where both resource constraints and model capability are critical requirements.

Our main contributions are summarized as follows:

- Efficient and Lightweight Visual Encoder:** We adopt Siglip2-Base-384/16 (Tschannen et al. 2025) as an efficient and lightweight visual encoder with fewer than 100M parameters. This encoder is capable of processing images at arbitrary resolutions while producing a compact set of tokens, without modifying the original image size.
- Curriculum Learning Strategy:** We introduce a curriculum learning strategy that systematically structures the training process by staging both information density and task difficulty. By progressively increasing the complexity of training samples and tasks, our approach enables the model to acquire foundational capabilities before addressing more challenging scenarios. This staged progression facilitates more stable convergence and yields significant improvements in overall model performance.
- Superior Performance and Efficiency:** MagicVL-2B achieves state-of-the-art results among models with similar parameter scales, demonstrating superior accuracy across a range of vision-language benchmarks. Furthermore, our model reduces power consumption on mobile devices by 41.1%, making it highly suitable for real-world on-device applications where efficiency is critical.

Related Works

Efficient Image Encoding

CLIP-pretrained (Radford et al. 2021) vision transformers (Dosovitskiy et al. 2020) remain the mainstream image encoders for VLMs, with models such as SigLIP (Zhai et al. 2023), EVA-CLIP (Sun et al. 2023), InternViT (Chen et al. 2023), and DFN-CLIP (Fang et al. 2023) widely used. Recent works (Karamcheti et al. 2024; Tong et al. 2024a; Shi et al. 2024) improve performance by ensembling visual encoders with diverse objectives, while methods like LLaVA-PruMerge (Shang et al. 2024) and Matryoshka-based token sampling (Hu et al. 2024b; Cai et al. 2024) dynamically prune visual tokens to improve encoding efficiency. Additional strategies (Dai et al. 2023; Cha et al. 2024; Chu et al. 2023, 2024) leverage perceiver-style resamplers or pooling operations to reduce token numbers. Hierarchical architectures such as ConvNeXT (Liu et al. 2022) and FastViT (Vasu et al. 2023) further decrease token counts via downsampling the input tensor at each computational stage.

Vision Language Models

LLMs (Brown 2020; Touvron et al. 2023a,b; Anil et al. 2023) have proven highly effective in tackling a wide spectrum of challenging tasks (Wei et al. 2022; Trinh et al. 2024). Vision-language models (VLMs) (OpenAI 2023; Liu et al. 2024b; Chen et al. 2024c; Zhang et al. 2023) extend LLMs to multimodal inputs via mechanisms such as linear projectors (Liu et al. 2024b; Chen et al. 2024b; Wang et al. 2024), Q-Former modules (Li et al. 2023), and perceiver resamplers (Alayrac et al. 2022; Bai et al. 2023; Yao et al. 2024). To better process high-resolution images, dynamic resolution techniques (Chen et al. 2024b; Liu et al. 2023a, 2024a) have

been introduced, enabling finer-grained visual understanding at varying resolutions (Huang et al. 2024). However, these dynamic resolution methods introduce specific challenges for mobile deployment: the proliferation of image patches can substantially slow down the visual encoder, while the resulting longer sequences of image tokens lead to increased inference latency for the language model (Lin et al. 2023).

On-Mobile-Device Large Language Models

With the expansion of application scenarios for large language models, there is growing interest in small-scale large language models (SLMs) as users prioritize efficiency and cost-effectiveness (Ashkboos et al. 2024). Recently, a range of SLMs have been developed to address these needs, covering both language-only (Hu et al. 2024a; Abdin et al. 2024; Mehta et al. 2024) and multimodal (Yao et al. 2024; Wang et al. 2024; Chen et al. 2024b; Luo et al. 2024; Li et al. 2024a) models. Thanks to their reduced parameter counts (typically 2-3B), these models are now feasible for deployment on personal devices such as PCs and smartphones. Beyond the creation of more compact yet powerful LLMs and VLMs, recent system-level research has proposed various approaches for efficiently deploying SLMs on end-user hardware, including personal computers (Wei et al. 2024) and mobile phones (Yao et al. 2024; Li et al. 2024b). Our proposed MagicVL-2B adopts a smaller visual encoder, which enables the model to achieve significantly lower power consumption on mobile devices, while still delivering strong performance across a wide range of benchmarks.

MagicVL-2B

In this section, we provide a comprehensive overview of MagicVL-2B, focusing on its model architecture and the design of a lightweight visual encoder. We also describe how a curriculum learning strategy is employed to progressively train the model.

Model Architecture

Overall Architecture Our architecture is an enhanced variant built upon the InternVL2.5 framework (Chen et al. 2024b). The overall pipeline is depicted in Figure 2 and comprises the following key components. **Visual Encoder:** To handle multimodal (image and language) inputs, we employ the SigLIP2-base (Tschannen et al. 2025) Vision Transformer (ViT) with an input resolution of 384×384 , as adopted in prior works (Lin et al. 2023). This encoder contains 93M parameters. **MLP Projector:** A two-layer multilayer perceptron (MLP) is utilized to project image tokens into the token space of the large language model (LLM). **LLM:** We leverage Qwen2.5-1.5B (Yang et al. 2024) or Qwen3-1.7B (Yang et al. 2025) as the backbone language model to construct MagicVL-2B. To further enhance the model’s capability in comprehending inputs at varying resolutions, we introduce a Dynamic High Resolution module. Inspired by the limitations of excessively enlarged images observed in InternVL2.5 (Chen et al. 2024b) and LLaVA-NeXT (Liu et al. 2024a), we propose a novel solution that significantly improves both training and inference efficiency.

Lightweight Visual Encoder The visual encoder in our model is responsible for processing image-modality inputs. We systematically evaluated three candidate architectures: ViT (Dosovitskiy et al. 2020), SigLIP (Zhai et al. 2023), and SigLIP2 (Tschannen et al. 2025), and ultimately selected SigLIP2 due to its superior performance. To achieve an optimal trade-off between computational efficiency and the capability to extract fine-grained visual features for on-device VLMs, we adopt the following design choices: First, regarding model size, we employ the SigLIP2-base variant, which comprises approximately 93 million parameters, to maintain computational efficiency. Second, for image resolution, we use a relatively high input resolution (384×384) to improve the representation of global visual information. Third, we set the patch size to 16, enabling the encoder to better capture fine-grained and complex visual details, which is particularly beneficial for on-device scenarios.

Dynamic High Resolution Our visual encoder is pre-trained on a fixed input resolution of 384, which significantly constrains its adaptability to various images, especially higher resolutions. Dynamic resolution has emerged as an effective approach to address this limitation, as demonstrated by models such as InternVL (Zhu et al. 2025; Chen et al. 2024a). However, existing dynamic resolution techniques often suffer from severe distortion artifacts, typically requiring resizing the original image’s height and width to integer multiples of the pre-training resolution. As a result, when the aspect ratio of the input image differs from that of the pre-training resolution, the image content is easily distorted, leading to degraded semantic representation and the introduction of redundant tokens, which further reduces inference efficiency. This issue is particularly evident on mobile devices, where atypical aspect ratios (e.g., long screenshots) are prevalent.

To address this problem, we propose a token-level resizing strategy: instead of resizing the image to integer multiples of the pre-training resolution, we resize each dimension to the nearest integer multiple of the pixel size corresponding to a single visual token. This approach minimizes image distortion under the VLM token paradigm, ensuring that image resolution and content are almost perfectly preserved. Given an input image $V \in \mathbb{R}^{H \times W \times C}$ where H , W , and C denote the original image height, width, and number of channels, respectively, the resized image is $V' \in \mathbb{R}^{H' \times W' \times C}$ where H' and W' are computed as:

$$H' = \left\lfloor \frac{H}{N_{\text{token}}} + 0.5 \right\rfloor \times N_{\text{token}}$$

$$W' = \left\lfloor \frac{W}{N_{\text{token}}} + 0.5 \right\rfloor \times N_{\text{token}}$$

$$N_{\text{token}} = N_{\text{patchsize}} \times R_{\text{psf}}$$

Here, $\lfloor \cdot \rfloor$ denotes the floor function, $N_{\text{patchsize}}$ is the patch size (set to 16). We utilize pixel unshuffle for token compression, with a compression ratio of $R_{\text{psf}} = 2$. To accommodate images of varying sizes with an encoder that operates at a fixed input resolution, we standardize all inputs to match the encoder’s required resolution. To prevent image distortion during this unification process, we adopt a padding strategy: for any image boundary v_i that does not meet the required dimension, we pad the image with zeros until it reaches

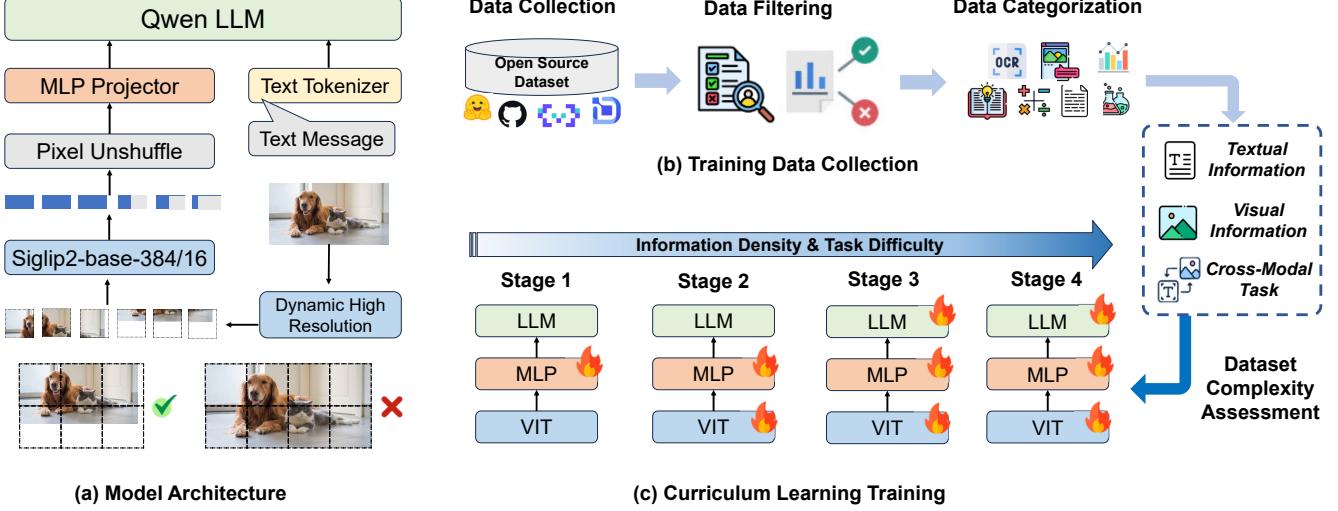


Figure 2: Overview of our MagicVL-2B. (a) **Model Architecture**: The model architecture integrates large language model using a visual encoder with dynamic high resolution and MLP projector. (b) **Training Data Collection**: Open-source datasets are filtered and categorized into sub-tasks based on data quality and task types. (c) **Curriculum Learning Training**: Dataset complexity is assessed along three dimensions. The model is trained in multiple stages, with each stage introducing tasks of increasing difficulty and more complex information.

384 × 384. The influence of these padded regions is eliminated by applying an attention mask. All tokens generated from the padded regions are discarded, and only tokens corresponding to the original image content are retained for subsequent LLM computation. This approach maximizes the preservation of the original image information while minimizing the introduction of redundant information during the size unification process.

Training Data Collection

Data Collection For pre-training, we curated a large-scale collection of open-source image-text datasets, comprising approximately *150 million image-text pairs*. We prioritized datasets that offer both high data quality and diverse visual content. For datasets that feature substantial visual diversity but only moderate quality, we applied rigorous filtering and data cleaning procedures to improve their overall reliability. Due to the limited availability of open-source Chinese image-text datasets, we leveraged large-scale LLMs to translate a subset of English datasets into Chinese, thereby enhancing the model’s bilingual capabilities. A comprehensive description of our datasets collection can be found in the supplementary materials.

Data Filtering We employ a multi-stage data filtering pipeline to ensure high-quality image-text pairs for pre-training. First, a *heuristic rule-based filtering system* is applied to remove samples containing excessive abnormal characters or synthetic data with anomalous keywords. To mitigate the issue of repetitive content observed in InternVL2.5 (Chen et al. 2024a), we further design a *rule-based duplication detection system* that eliminates entries with large repeated segments or frequent occurrences of short phrases.

Finally, we introduce an *LLM prompt-based filtering system*, which leverages large language models to evaluate the *logical coherence* of each entry and to detect potential *hallucinations* in the descriptions. Representative examples of excluded data can be found in the supplementary materials.

Data Categorization We systematically organize the collected open-source datasets into the following task categories: *reasoning, GUI, OCR, text-only, chart, caption, visual question answering, and grounding*. To ensure accurate and efficient categorization, we first inspect each dataset to determine whether explicit task-type labels are provided. If such labels exist, we directly categorize the dataset accordingly. For datasets lacking explicit task-type labels, we conduct manual verification by randomly sampling a subset of data points and performing human inspection to assess whether the dataset corresponds to a single task category. If multiple task types are present within a dataset, we further employ a large-scale vision-language model (VLM) to automatically classify individual samples, thereby splitting heterogeneous datasets into several task-specific subsets.

Curriculum Learning Training

Dataset Complexity Assessment Given a dataset \mathcal{D} , where v denotes the input image, p the input prompt, x the corresponding response, and n the total number of samples, we represent the dataset as

$$\mathcal{D} = \{(v_i, p_i, x_i) \mid i = 1, \dots, n\}.$$

As depicted in Figure 2(c), we introduce a rigorous and multifaceted evaluation protocol for characterizing dataset complexity. Our framework systematically dissects the dataset along three different dimensions, yielding a holistic complex-

ity score S that encapsulates the textual information, visual information, and cross-modal task complexity.

Textual Information Complexity This dimension evaluates the complexity of a dataset based on the diversity and linguistic complexity of the textual content.

1. *Token length*: We define the normalized average token length of response L as an indicator of the level of detail and informativeness:

$$L = \frac{1}{n} \sum_{i=1}^n \text{len}(x_i),$$

where $\text{len}(x_i)$ denotes the token length of x_i .

2. *Type-token ratio (TTR)*: To capture lexical diversity, we calculate the average type-token ratio for the concatenation of prompt and response:

$$T = \frac{1}{n} \sum_{i=1}^n \text{TTR}(p_i + x_i),$$

where a higher TTR indicates greater lexical diversity, reflecting increased linguistic complexity (Kettunen 2014).

3. *Perplexity*: Besides the statistical methods, we also utilize the language model in our model architecture, Qwen2.5-1.5B, to compute the average perplexity of the response x_i conditioned on prompt p_i :

$$P = \frac{1}{n} \sum_{i=1}^n \text{PPL}(x_i | p_i).$$

The perplexity produced by the language model not only reflect the intrinsic linguistic complexity of the textual data (Ankner et al. 2024), but also implicitly capture the degree to which the text content is dependent on or grounded in the image data, providing an additional perspective for measuring dataset complexity.

The overall text-based complexity score S_{text} is then given by the arithmetic mean after normalization:

$$S_{\text{text}} = \frac{1}{3} (\hat{L} + \hat{T} + \hat{P}),$$

where the notation $\hat{\cdot}$ denotes normalization.

Visual Information Complexity This dimension assesses the complexity of a dataset by quantifying the richness of visual contents in the images.

1. *Image entropy*: we calculate the average image entropy as a statistical measure of the pixel-level information content. Let e denote the normalization constant for entropy, and we define our average *image entropy* E of the dataset as:

$$E = \frac{1}{n} \sum_{i=1}^n \text{Entropy}(v_i).$$

2. *Text density*: we utilize a state-of-the-art OCR model to compute the average text density within images in the dataset (Cui et al. 2025). Specifically, Let $t(v_i)$ denote the number of text tokens recognized by the OCR model in

image v_i , and $a_i(v_i)$ denote the area (in pixels) of image v_i . The average text density per image area is calculated as

$$D_{\text{text}} = \frac{1}{n} \sum_{i=1}^n \frac{t(v_i)}{a_i(v_i)}.$$

3. *Object density*: similarly, we also leverage an open-domain object detection model to estimate the average object density of each image (Liu et al. 2023b). Let $\text{obj}(v_i)$ denote the number of objects detected in image v_i , then the average object density per image area is defined as:

$$D_{\text{obj}} = \frac{1}{n} \sum_{i=1}^n \frac{\text{obj}(v_i)}{a_i(v_i)}.$$

We normalize each metric and compute the overall image-based complexity score S_{image} as:

$$S_{\text{image}} = \frac{1}{3} (\hat{E} + \hat{D}_{\text{text}} + \hat{D}_{\text{obj}}).$$

Cross-Modal Task Complexity This dimension assesses dataset complexity in terms of cross-modal information integration and reasoning. Following NVILA (Liu et al. 2024c), we propose a loss-based framework that uses VLMs of different scales to quantify task complexity. The key idea is that samples demanding more advanced cross-modal reasoning will yield a larger gap in autoregressive loss between smaller and larger models.

Loss-based comparative evaluation: Let M_s and M_l denote a small and a large VLM, respectively. For each data point, we compute the autoregressive loss of VLM models, $M_s(x | (v, p))$ and $M_l(x | (v, p))$. We then define the complexity score as the proportion of data points where the loss gap between M_s and M_l exceeds a controlled margin:

$$\mathcal{C}(M_s, M_l) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{M_s(x_i | (v_i, p_i)) > \beta M_l(x_i | (v_i, p_i)) > \delta\}},$$

where β is a scaling hyperparameter determining the required loss ratio, and δ is a threshold to filter out trivial cases with low absolute loss. For a more comprehensive and robust evaluation, we calculate the autoregressive loss for three models with different sizes: Qwen2VL-2B, Qwen2VL-7B, and Qwen2VL-72B (Wang et al. 2024), and define our cross-model task complexity score S_{task} as:

$$S_{\text{task}} = \frac{1}{2} (\mathcal{C}(M_{2B}, M_{7B}) + \mathcal{C}(M_{7B}, M_{72B})).$$

With the three complexity scores S_{text} , S_{image} , S_{task} obtained from the aforementioned metrics, we compute the overall dataset complexity score S as a weighted sum:

$$S = \lambda_1 S_{\text{text}} + \lambda_2 S_{\text{image}} + \lambda_3 S_{\text{task}},$$

where the weights $\{\lambda_i\}_{i=1}^3$ are adaptively selected according to the task category of each dataset, as defined in the previous section. Detailed weight configurations for each task category are provided in supplementary materials.

Model	HB	CR	MMB	RQA	MME_R	MMS	DocV	OB	A2D	SEED
Model parameters > 7B										
MiniCPM-V-2.6 (Yao et al. 2024)	48.1	-	75.1	62.8	-	57.5	90.8	852	82.1	65.7
Qwen2-VL-7B (Wang et al. 2024)	50.1	74.4	78.0	67.0	56.5	60.7	94.5	856	83.0	69.0
InternVL2-5-7B (Chen et al. 2024a)	52.9	79.6	83.0	70.3	57.4	63.9	95.7	864	83.9	70.4
InternVL2-8B (Chen et al. 2024b)	45.2	71.6	78.1	66.1	59.1	61.6	91.6	794	83.0	69.7
InternVL2.5-8B (Chen et al. 2024a)	50.1	78.4	83.1	71.0	59.1	62.8	93.0	822	84.5	69.7
2B < Model parameters ≤ 4B										
Qwen2.5-VL-3B (Bai et al. 2025)	46.3	73.6	77.4	65.4	53.1	55.9	93.9	797	81.6	67.6
BlueLM-V-3B (Lu et al. 2024)	48.1	-	78.1	66.7	-	62.3	87.8	829	85.3	-
Phi3.5-Vision-4B (Abdin et al. 2024)	40.5	68.5	72.1	59.7	35.2	47.5	69.3	599	77.8	62.2
InternVL2-4B (Chen et al. 2024b)	41.9	71.1	75.8	60.7	52.1	54.3	89.2	788	78.9	63.9
InternVL2.5-4B (Chen et al. 2024a)	46.3	75.5	79.3	64.4	55.3	54.3	91.6	828	81.4	66.9
Model parameters ≤ 2B										
LLaVA-OV-0.5B (Li et al. 2024a)	27.9	-	59.6	55.6	-	37.7	70.0	565	57.1	-
InternVL2-1B (Chen et al. 2024b)	34.0	-	59.7	61.6	40.2	45.7	81.7	754	64.1	54.3
InternVL2.5-1B (Chen et al. 2024a)	39.0	60.9	68.4	57.5	44.2	50.1	84.8	785	69.3	59.0
SmolVLM2 (Marafioti et al. 2025)	40.6	-	61.1	57.5	-	46.0	80.0	725	69.7	60.5
Qwen2-VL-2B (Wang et al. 2024)	41.7	-	72.2	<u>62.6</u>	-	48.0	90.1	<u>809</u>	74.7	<u>62.4</u>
Aquila-VL-2B (Gu et al. 2024)	43.0	-	75.2	-	-	48.0	85.0	772	75.0	63.0
InternVL2-2B (Chen et al. 2024b)	37.9	66.3	70.2	57.3	47.3	50.1	86.9	784	74.1	60.0
InternVL2.5-2B (Chen et al. 2024a)	42.6	70.0	<u>73.4</u>	60.0	48.8	<u>53.7</u>	88.7	804	74.9	60.9
MagicVL-2B (Qwen2.5-1.5B)	<u>47.7</u>	70.9	71.8	61.4	49.8	52.7	87.7	775	<u>76.7</u>	61.0
MagicVL-2B (Qwen3-1.7B)	50.8	<u>70.3</u>	73.7	63.5	<u>49.1</u>	57.9	<u>89.0</u>	828	<u>77.4</u>	61.6

Table 1: Benchmark results of various VLMs. **HB**: HallusionBench, **CR**: CRPE, **MMB**: MMBench_V11_en, **RQA**: RealworldQA, **MME_R**: MME_Realworld, **MMS**: MMStar, **DocV**: DocVQA, **OB**: OCRBench, **A2D**: AI2D, **SEED**: SEED-2 Plus.

Progressive Training Stages As illustrated in Figure 2(c), we propose a *four-stage curriculum learning paradigm* that incrementally strengthens the model’s capability in multi-modal understanding and reasoning. Each stage is meticulously designed based on the data categorization and complexity analysis introduced in previous sections, with both training data and strategies specifically optimized for distinct learning objectives: **Stage 1: Foundational Modality Alignment.** We begin by aligning the visual and linguistic modalities. In this stage, the visual encoder and LLM are frozen, and only the MLP projector is updated. Training is conducted on *low-complexity* image-caption pairs (10M samples), enabling the model to establish fundamental cross-modal grounding within a simplified setting. **Stage 2: Enhanced Visual Representation.** Subsequently, both the visual encoder and the MLP projector are jointly optimized, while the LLM remains frozen. The training set is extended to include *high-complexity* image-caption pairs (23M samples), which encourages the model to learn richer visual features and more robust cross-modal representations. **Stage 3: Generalized Multi-Modal Ability.** At this stage, all components—the visual encoder, MLP projector, and LLM—are unfrozen for joint training. We utilize diverse multi-modal instruction-following tasks, leveraging only *low-complexity* datasets (54M samples). By gradually increasing task difficulty, this stage mitigates catastrophic forgetting and cultivates generalized reasoning ability. **Stage 4: Advanced Multi-Modal Ability.**

Finally, the model is trained on the most challenging samples (*high-complexity* data spanning all tasks, 66M samples), with all components optimized jointly. This stage consolidates advanced reasoning abilities and significantly boosts performance on both general and fine-grained tasks, particularly in real-world mobile scenarios. This progressive, complexity-aware curriculum facilitates a seamless transition from fundamental modality alignment to advanced multi-modal reasoning. As a result, the model acquires both robust generalization and strong task-specific capabilities.

Experiments

In this section, we conduct a series of experiments to validate the effectiveness of our proposed approaches and to demonstrate the capabilities of MagicVL-2B in terms of benchmark accuracy and deployment efficiency. Unless otherwise specified, MagicVL-2B refers to the model using Qwen3-1.7B as its language model.

Training Setting

The training pipeline for MagicVL-2B is structured into four progressive stages, in accordance with our curriculum learning paradigm. Several key hyperparameters remain consistent across all stages, including a packed batch with a maximum token length of 16,384 and up to 48 images, a maximum of

Model Name	Processor	Model Loading	ViT latency	LLM latency	Throughput
InternVL2.5-2B (Yao et al. 2024)	Snapdragon 8 Elite	1.04 s	0.90 s	2.0 s	14.3 token/s
MagicVL-2B	Snapdragon 8 Elite	1.01 s	0.09 s	1.7 s	23.9 token/s

Table 2: Deployment efficiency comparison with InternVL2.5-2B. MagicVL-2B achieves lower ViT and LLM inference latency as well as higher throughput compared to InternVL2.5-2B.

Models	Token	TextVQA
InternVL2.5-2B (Yao et al. 2024)	0.82 M	74.3
MagicVL-2B	0.51 M	74.5

Table 3: Ablation study of dynamic resolution with different visual tokens consumption on TextVQA.

24 dynamic patches, the AdamW optimizer (Loshchilov and Hutter 2017), and a cosine learning rate decay schedule. All training experiments are conducted on 128 NVIDIA A800 80G GPUs. Stage-specific hyperparameters are incrementally adjusted to progressively enhance the model’s capability, maintaining a balance between performance and training efficiency. Specifically, **Stage 1**: a learning rate of 2×10^{-4} , 100 warmup steps, and 65k training steps; **Stage 2**: a learning rate of 1×10^{-5} , 100 warmup steps, and 90k training steps; **Stage 3**: a learning rate of 4×10^{-5} , a warmup ratio of 0.03, and 140k training steps; **Stage 4**: a learning rate of 4×10^{-5} , a warmup ratio of 0.03, and 250k training steps.

Comparison with State-of-the-art

We evaluate the performance of our MagicVL-2B model against a comprehensive selection of state-of-the-art multimodal models across multiple benchmarks, as summarized in Table 1. For fair comparison, models are grouped according to their parameter scale. Within the $\leq 2B$ parameter regime, MagicVL-2B consistently achieves the highest scores on the majority of benchmarks, including HallusionBench (50.8), MMBench (73.7), RealworldQA (63.5), MMStar (57.9), OCRCbench (828), and AI2D (77.4), demonstrating strong capabilities in both vision-language reasoning and real-world understanding tasks. Notably, MagicVL-2B surpasses several larger models, particularly on HallusionBench and OCRCbench, where it even outperforms models with over 7B parameters. These results highlight the efficiency and effectiveness of MagicVL-2B, especially in light of its compact model size. The substantial performance improvements underscore the strength of our design in developing lightweight yet powerful multimodal models.

Ablation Study

Effectiveness of Dynamic High Resolution We compare the dynamic high-resolution method in our MagicVL-2B with InternVL on the TextVQA dataset, which is specifically designed to evaluate a model’s OCR and multi-modal reasoning capabilities in complex scenarios. As shown in Table 3, MagicVL-2B reduces the total number of tokens by approx-

Data C.	Prog T.	HB	CR	MME_R	DocV
✗	✗	49.5	69.2	48.2	87.5
✓	✗	50.3	70.0	48.4	88.1
✓	✓	50.8	70.3	49.1	89.0

Table 4: Ablation study of curriculum learning training. Data C.: data categorization, Prog T.: progressive training. **HB**: HallusionBench, **CR**: CRPE, **MME_R**: MME Realworld, **DocV**: DocVQA

imately 37.8% (0.52 M vs 0.81 M) during the evaluation, while also achieving improved accuracy (74.5% vs 74.3%). These results demonstrate that our dynamic high resolution method can significantly reduce computational costs while still preserving detailed information.

Effectiveness of Curriculum Learning Pre-Training We conduct an ablation study on MagicVL-2B to assess the impact of curriculum learning, as presented in Table 4. The first baseline trains on all available data in the initial stages (excluding Data C. and Prog T.), resulting in the lowest accuracy among the compared methods. The second baseline introduces Data C., utilizing caption data in stages 1–2 and mixed data in stages 3–4, which improves the model’s fundamental multimodal capabilities, particularly for general vision-language understanding and hallucination tasks such as CR and HB. Our full curriculum learning strategy further enhances performance across all datasets, with notable improvements on challenging and fine-grained tasks, including multi-image, OCR, and reasoning benchmarks such as MME_R, DocV, and MMS. These findings demonstrate that curriculum learning substantially boosts cross-modal understanding and generalization in lightweight models, effectively narrowing the performance gap with larger models and enabling more efficient training and deployment.

Deployment Efficiency Evaluation We conduct a head-to-head comparison between MagicVL-2B and InternVL2.5-2B (Chen et al. 2024a) on the same Snapdragon 8 Elite processor. As summarized in Table 2, MagicVL-2B demonstrates substantial improvements in deployment efficiency across various metrics. Specifically, MagicVL-2B achieves a significantly lower ViT inference latency of 0.09s, compared to 0.90s for InternVL2.5-2B, reflecting a remarkable reduction in visual feature extraction time. Furthermore, MagicVL-2B attains a throughput of 23.9 tokens/s, approximately 1.67× higher than that of InternVL2.5-2B (14.3 tokens/s), indicating a more efficient token generation process and en-

hanced suitability for real-time applications. These results underscore the advantages of MagicVL-2B for deployment on resource-constrained edge devices, establishing it as a compelling solution for mobile and embedded AI scenarios.

Conclusion

In summary, MagicVL-2B demonstrates that it is feasible to achieve both state-of-the-art performance and outstanding efficiency within a lightweight multimodal framework. By integrating an efficient visual encoder with a curriculum learning strategy, MagicVL-2B establishes a new benchmark for small-scale MLLMs, achieving strong results on challenging benchmarks while maintaining low power consumption and latency. These advantages underscore its practical utility for real-world deployment, particularly in resource-constrained environments. We believe that MagicVL-2B paves the way for further research into scalable and efficient multimodal models, and serves as a robust foundation for deployment across diverse devices and application scenarios.

References

- Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.
- Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Milligan, K.; Reynolds, M.; et al. 2022. Flamingo: A visual language model for few-shot learning. *NeurIPS*, 35: 23716–23736.
- Anil, R.; Dai, A. M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.
- Ankner, Z.; Blakeney, C.; Sreenivasan, K.; Marion, M.; Leavitt, M. L.; and Paul, M. 2024. Perplexed by perplexity: Perplexity-based data pruning with small reference models. *arXiv preprint arXiv:2405.20541*.
- Ashkboos, S.; Mirzadeh, I.; Alizadeh, K.; Sekhavat, M. H.; Nabi, M.; Farajtabar, M.; and Faghri, F. 2024. Computational Bottlenecks of Training Small-scale Large Language Models. *arXiv preprint arXiv:2410.19456*.
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Cai, M.; Yang, J.; Gao, J.; and Lee, Y. J. 2024. Matryoshka Multimodal Models. *arXiv preprint arXiv:2405.17430*.
- Cha, J.; Kang, W.; Mun, J.; and Roh, B. 2024. Honeybee: Locality-enhanced Projector for Multimodal LLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024a. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.
- Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.
- Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.
- Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; et al. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.
- Chu, X.; Qiao, L.; Zhang, X.; Xu, S.; Wei, F.; Yang, Y.; Sun, X.; Hu, Y.; Lin, X.; Zhang, B.; et al. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *arXiv preprint arXiv:2402.03766*.
- Cui, C.; Sun, T.; Lin, M.; Gao, T.; Zhang, Y.; Liu, J.; Wang, X.; Zhang, Z.; Zhou, C.; Liu, H.; Zhang, Y.; Lv, W.; Huang, K.; Zhang, Y.; Zhang, J.; Zhang, J.; Liu, Y.; Yu, D.; and Ma, Y. 2025. PaddleOCR 3.0 Technical Report. *arXiv:2507.05595*.
- Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning.
- Ding, Y.; Niu, C.; Wu, F.; Tang, S.; Lyu, C.; and Chen, G. 2024. Enhancing On-Device LLM Inference with Historical Cloud-Based LLM Interactions. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 597–608.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Fang, A.; Jose, A. M.; Jain, A.; Schmidt, L.; Toshev, A.; and Shankar, V. 2023. Data Filtering Networks. *arXiv preprint arXiv:2309.17425*.
- Gu, S.; Zhang, J.; Zhou, S.; Yu, K.; Xing, Z.; Wang, L.; Cao, Z.; Jia, J.; Zhang, Z.; Wang, Y.; et al. 2024. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. *arXiv preprint arXiv:2410.18558*.

- Hu, S.; Tu, Y.; Han, X.; He, C.; Cui, G.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhao, W.; et al. 2024a. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. *arXiv preprint arXiv:2404.06395*.
- Hu, W.; Dou, Z.-Y.; Li, L. H.; Kamath, A.; Peng, N.; and Chang, K.-W. 2024b. Matryoshka Query Transformer for Large Vision-Language Models.
- Hua, W.; Wan, M.; Vadrevu, S.; Nadel, R.; Zhang, Y.; and Wang, C. 2024. Interactive Speculative Planning: Enhance Agent Efficiency through Co-design of System and User Interface. *arXiv:2410.00079*.
- Huang, M.; Liu, Y.; Liang, D.; Jin, L.; and Bai, X. 2024. Mini-monkey: Multi-scale adaptive cropping for multimodal large language models. *arXiv preprint arXiv:2408.02034*.
- Karamcheti, S.; Nair, S.; Balakrishna, A.; Liang, P.; Kollar, T.; and Sadigh, D. 2024. Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models. In *International Conference on Machine Learning (ICML)*.
- Kettunen, K. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3): 223–245.
- Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Li, L.; Qian, S.; Lu, J.; Yuan, L.; Wang, R.; and Xie, Q. 2024b. Transformer-lite: High-efficiency deployment of large language models on mobile phone gpus. *arXiv preprint arXiv:2403.20041*.
- Lin, J.; Yin, H.; Ping, W.; Lu, Y.; Molchanov, P.; Tao, A.; Mao, H.; Kautz, J.; Shoeybi, M.; and Han, S. 2023. VILA: On Pre-training for Visual Language Models. *arXiv:2312.07533*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual Instruction Tuning. *NeurIPS*, 36.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.
- Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Liu, Z.; Zhu, L.; Shi, B.; Zhang, Z.; Lou, Y.; Yang, S.; Xi, H.; Cao, S.; Gu, Y.; Li, D.; et al. 2024c. NVILA: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.
- Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Lu, X.; Chen, Y.; Chen, C.; Tan, H.; Chen, B.; Xie, Y.; Hu, R.; Tan, G.; Wu, R.; Hu, Y.; et al. 2024. Bluelm-v-3b: Algorithm and system co-design for multimodal large language models on mobile devices. *arXiv preprint arXiv:2411.10640*.
- Luo, G.; Yang, X.; Dou, W.; Wang, Z.; Dai, J.; Qiao, Y.; and Zhu, X. 2024. Mono-InternVL: Pushing the Boundaries of Monolithic Multimodal Large Language Models with Endogenous Visual Pre-training. *arXiv preprint arXiv:2410.08202*.
- Marafioti, A.; Zohar, O.; Farré, M.; Noyan, M.; Bakouch, E.; Cuenca, P.; Zakka, C.; Allal, L. B.; Lozhkov, A.; Tazi, N.; et al. 2025. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.
- McKinzie, B.; Gan, Z.; Fauconnier, J.-P.; Dodge, S.; Zhang, B.; Dufter, P.; Shah, D.; Du, X.; Peng, F.; Weers, F.; et al. 2024. MM1: Methods, analysis & insights from multimodal LLM pre-training. *arXiv preprint arXiv:2403.09611*.
- Mehta, S.; Sekhavat, M. H.; Cao, Q.; Horton, M.; Jin, Y.; Sun, C.; Mirzadeh, S. I.; Najibi, M.; Belenko, D.; Zatloukal, P.; et al. 2024. Openelm: An efficient language model family with open training and inference framework. In *Workshop on Efficient Systems for Foundation Models II@ ICML2024*.
- OpenAI. 2023. GPT-4 Technical Report. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume abs/2303.08774.
- Qu, G.; Chen, Q.; Wei, W.; Lin, Z.; Chen, X.; and Huang, K. 2024. Mobile edge intelligence for large language models: A contemporary survey. *arXiv preprint arXiv:2407.18921*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Shang, Y.; Cai, M.; Xu, B.; Lee, Y. J.; and Yan, Y. 2024. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. *arXiv preprint arXiv:2403.15388*.
- Shi, M.; Liu, F.; Wang, S.; Liao, S.; Radhakrishnan, S.; Huang, D.-A.; Yin, H.; Sapra, K.; Yacoob, Y.; Shi, H.; Catanzaro, B.; Tao, A.; Kautz, J.; Yu, Z.; and Liu, G. 2024. Eagle: Exploring The Design Space for Multimodal LLMs with Mixture of Encoders. *arXiv:2408.15998*.
- Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv:2303.15389*.
- Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S. C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; Wang, A.; Fergus, R.; LeCun, Y.; and Xie, S. 2024a. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs.
- Tong, S.; Brown, E.; Wu, P.; Woo, S.; Middepogu, M.; Akula, S. C.; Yang, J.; Yang, S.; Iyer, A.; Pan, X.; et al. 2024b. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*.
- Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Trinh, T. H.; Wu, Y.; Le, Q. V.; He, H.; and Luong, T. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995): 476–482.
- Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M. F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.
- Vasu, P. K. A.; Gabriel, J.; Zhu, J.; Tuzel, O.; and Ranjan, A. 2023. FastViT: A Fast Hybrid Vision Transformer using Structural Reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.
- Wei, J.; Cao, S.; Cao, T.; Ma, L.; Wang, L.; Zhang, Y.; and Yang, M. 2024. T-mac: Cpu renaissance via table lookup for low-bit llm deployment on edge. *arXiv preprint arXiv:2407.00088*.
- Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.
- Xue, Z.; Song, Y.; Mi, Z.; Chen, L.; Xia, Y.; and Chen, H. 2024. PowerInfer-2: Fast Large Language Model Inference on a Smartphone. *arXiv preprint arXiv:2406.06282*.
- Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.
- Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. *International Conference on Computer Vision (ICCV)*.
- Zhang, H.; Gao, M.; Gan, Z.; Dufter, P.; Wenzel, N.; Huang, F.; Shah, D.; Du, X.; Zhang, B.; Li, Y.; Dodge, S.; You, K.; Yang, Z.; Timofeev, A.; Xu, M.; Chen, H.-Y.; Fauconnier, J.-P.; Lai, Z.; You, H.; Wang, Z.; Dehghan, A.; Grasch, P.; and Yang, Y. 2024. MM1.5: Methods, Analysis & Insights from Multimodal LLM Fine-tuning. *arXiv:2409.20566*.
- Zhang, R.; Han, J.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Gao, P.; and Qiao, Y. 2023. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *arXiv preprint arXiv:2303.16199*.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Duan, Y.; Tian, H.; Su, W.; Shao, J.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.