JINA-VLM: SMALL MULTILINGUAL VISION LANGUAGE MODEL

Andreas Koukounas* **Georgios Mastrapas*** **Florian Hönicke** **Sedigheh Eslami†**
Guillaume Roncari **Scott Martens** **Han Xiao**

Jina AI *by* Elastic
Prinzessinnenstr. 19-20, Berlin 10969, Germany
research@jina.ai

ABSTRACT

We present jina-vlm, a token-efficient 2.4B parameter vision-language model that achieves state-of-the-art multilingual VQA performance among open 2B-scale VLMs. The model couples a SigLIP2 vision encoder with a Qwen3 language decoder and makes use of image tiling and attention-pooling for token-efficient processing of arbitrary-resolution images. To understand the contribution of different training data categories, we conduct a leave-one-out data mixture ablation study—systematically removing task, domain, modality, and language categories—to diagnose which data types are necessary versus redundant and whether task benefits transfer across domains. Model weights and code are publicly released at https://huggingface.co/jinaai/jina-vlm.

1 INTRODUCTION

Vision-language models (VLMs) combine pretrained vision encoders with large language models to tackle tasks requiring joint visual and textual understanding (Alayrac et al., 2022; Liu et al., 2023). Recent VLMs have achieved strong results on visual question answering (VQA), OCR, and multimodal reasoning. However, two challenges limit their practical deployment. First, multilingual capabilities often degrade during vision adaptation: models that perform well on English benchmarks show uneven results across other languages (Manea & Libovický, 2025). Second, high-quality VLMs remain computationally expensive to train and deploy, limiting accessibility for researchers and practitioners with constrained resources.

This work introduces jina-vlm, a 2.4B parameter VLM that addresses both challenges. The model aligns a SigLIP2-So400M/14-384 vision encoder (Tschannen et al., 2025) with Qwen3-1.7B-Base (Yang et al., 2025a) through an attention-pooling connector, trained with a two-stage pipeline that explicitly incorporates multilingual data. Among open 2B-scale VLMs, jina-vlm achieves state-of-the-art multilingual performance on MMMB and Multilingual MMBench (Sun et al., 2025). On standard English benchmarks spanning diagrams, charts, documents, and OCR, jina-vlm achieves the highest average score (72.3) across eight VQA benchmarks among 2B-scale VLMs. These results are enabled by a training recipe that incorporates high-quality English and multilingual data. Token efficiency is achieved via an arbitrary-resolution pipeline that combines overlapping tiling with attention-based token pooling to reduce visual token count by 4×.

We identify dataset mixture optimization as a key consideration in training small-scale VLMs. Prior work focuses on total data volume and quality or domain diversity (Tong et al., 2024; McKinzie et al., 2024; Li et al., 2025b), concentrating less on the fine-grained data categories (task type, domain, language) and whether their benefits transfer across tasks. To this end, we conduct an ablation study to provide preliminary insights into which data types are essential versus redundant and whether data benefits transfer across domains and tasks.

*Equal contribution.
†Work done during internship at Jina AI.

2 Related Work

VLM architecture and training. Modern VLMs follow an architectural pattern in which a pre-trained vision encoder extracts visual features, a connector projects them into the language model's embedding space, and a language model generates text conditioned on these visual tokens. This pattern was introduced by PaLI (Chen et al., 2023) using an encoder-decoder language model, and popularized in its decoder-only form by LLaVA (Liu et al., 2023). Vision Transformers (ViTs) (Dosovitskiy et al., 2021) produce patch-level representations that the language model processes alongside text embeddings. This decoder-only design is adopted by LLaVA (Liu et al., 2023; 2024a; Xu et al., 2024; Li et al., 2024d;a), QwenVL (Bai et al., 2023; Wang et al., 2024c; Bai et al., 2025), InternVL (Chen et al., 2024d;c; 2025; Zhu et al., 2025; Wang et al., 2025), and Ovis (Lu et al., 2024b; 2025). Training strategies vary: Wang et al. (2024c); Chen et al. (2025) mix multimodal and text-only data across training stages; Liu et al. (2024a) incorporate academic VQA datasets; Deitke et al. (2025), Li et al. (2024a), and Tong et al. (2024) curate large-scale, diverse data mixtures.

Efficient resolution-agnostic image processing. Standard ViTs process fixed-resolution images, requiring resizing that discards fine-grained detail. Since visual token count scales with resolution and Transformer computation scales quadratically with sequence length, naive high-resolution processing is prohibitive. Several solutions exist: Deitke et al. (2025) tile images with overlap; Wang et al. (2024c) introduce Naive Dynamic Resolution with Multimodal Rotary Positional Embedding (Su et al., 2024; Heo et al., 2024); Lu et al. (2025) use native-resolution ViTs (Dehghani et al., 2023). Orthogonally, images often contain low-information regions (e.g., sky backgrounds), making visual tokens highly redundant. Token compression methods address this (Chen et al., 2024a; Shang et al., 2025; Yang et al., 2025b; Xing et al., 2025). Chen et al. (2024c) develop Dynamic High-Resolution Tiling, and Liu et al. (2025) propose scale-then-compress strategies. Recent work on training-free token budgeting, such as HERO (Li et al., 2025a), demonstrates that inference-time pruning can achieve significant speedups while preserving accuracy; our approach differs by learning compact representations during training rather than dropping tokens at inference.

Vision-language connectors. The connector bridging vision encoders and language models significantly impacts both efficiency and performance. BLIP-2 (Li et al., 2023a) introduces Q-Former, a learnable query-based transformer that extracts fixed-length representations from visual features, reducing the number of tokens fed to the LLM. Flamingo (Alayrac et al., 2022) uses a Perceiver Resampler with cross-attention to compress visual tokens. Our attention-pooling connector reduces tokens but operates differently: rather than learning a fixed set of queries, we apply local 2×2 attention pooling that preserves spatial structure while achieving $4\times$ compression. Dense Connector (Yao et al., 2024a) concatenates features from multiple ViT layers and shows that intermediate layers carry complementary spatial and semantic information that improves VLM performance. MM1 (McKinzie et al., 2024) studies the effect of vision encoder configuration, including image resolution, token count, and connector design, on downstream accuracy. Our connector design draws on these observations, concatenating an intermediate and a near-final layer to capture both fine-grained and high-level features, combined with attention pooling for token reduction.

Small VLMs. Efficiency has become a central objective. Shao et al. (2025) combine quantization with aggressive resolution reduction for mobile deployment, matching larger models' performance. MiniCPM-V (Yao et al., 2024b) targets edge deployment while maintaining strong OCR and multilingual capabilities. Marafioti et al. (2025) systematically explore design parameters to train VLMs as small as 256M parameters.

Multilinguality and text-only performance. Many lightweight VLMs (Beyer et al., 2024; Steiner et al., 2024; Abdin et al., 2024) achieve strong English performance but degrade on other languages or on text-only capabilities. Wang et al. (2024c) and Chen et al. (2024c) improve on non-English languages through targeted multilingual training data, and Yue et al. (2025) introduce instruction-tuning data spanning 39 languages. For pure text performance, mitigation strategies include parameter-efficient fine-tuning (Laurençon et al., 2024) and staged training with progressive unfreezing (Li et al., 2024a).

Data mixture analysis. Several recent works investigate the role of training data composition in VLM performance. MM1 (McKinzie et al., 2024) conducts ablations over image-caption, interleaved, and text-only data ratios, finding that mixing all three modalities and carefully tuning their proportions is critical for strong few-shot performance; they also study the effect of image resolution

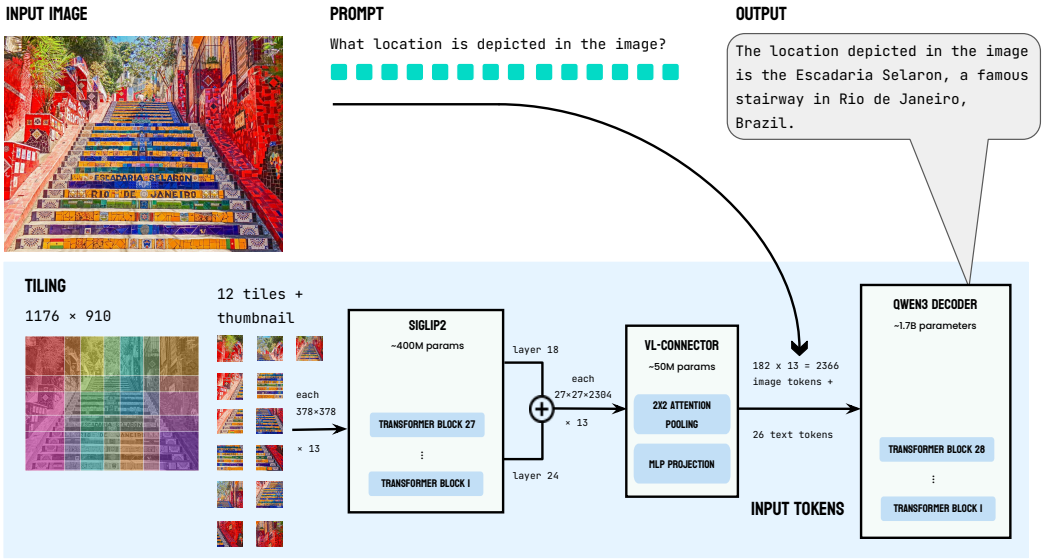


Figure 1: Architecture of `jina-vlm`. Images are resized to fit a grid of up to 12 overlapping tiles, plus a global thumbnail. Each tile is a square 378×378 crop; adjacent tiles overlap by 112 pixels with a stride of 266 pixels between tile origins. A 4×3 grid therefore spans 1176×910 pixels, and images exceeding this effective resolution are downscaled to fit the tile budget. Each tile produces 729 patches via SigLIP2 (Tschannen et al., 2025). The VL connector concatenates features from layers 24 and 18, the third- and ninth-to-last layers, then applies 2×2 attention pooling to reduce 729 tokens to 182 before projecting to the decoder dimension. Visual tokens are combined with text embeddings for the Qwen3 decoder (Yang et al., 2025a).

and token count on downstream accuracy. Cambrian-1 (Tong et al., 2024) introduces a large-scale data curation effort and evaluates the effect of data source diversity on visual grounding, reporting that balanced sampling across sources outperforms naive concatenation. Molmo/PixMo (Deitke et al., 2025) demonstrate that high-quality, targeted caption data can substitute for much larger but noisier mixtures, and provide insights on how caption style and granularity affect downstream task transfer. Eagle 2 (Li et al., 2025b) proposes a post-training data strategy built on diversity-first collection and quality filtering, using K-means clustering for balanced representation and rule-based removal of low-quality samples, showing gains from careful data curation rather than data scaling. Our work is complementary: rather than optimizing data proportions or sample weights, we perform *leave-one-out* ablations that remove entire data categories to measure their marginal contribution and cross-domain transfer effects. This provides a different diagnostic lens, identifying which categories are necessary versus redundant, that is particularly relevant for practitioners operating under fixed compute budgets.

3 Model Architecture

Figure 1 illustrates the architecture of `jina-vlm`. The model uses overlapping image tiling following Deitke et al. (2025), combined with attention-based token pooling to reduce sequence length while preserving spatial information.

3.1 Vision Encoder

The vision encoder, `SigLIP2-So400M/14-384`[1], is a 27-layer Vision Transformer with 400M parameters that processes 378×378 pixel inputs as 27×27 grids of 14×14 patches. To handle arbitrary resolutions, we decompose each image into overlapping tiles of this size and process each tile independently through the encoder. A global thumbnail, the full image resized to 378×378,

[1] `https://huggingface.co/google/siglip2-so400m-patch14-384`

provides context alongside the tile representations. We use a default of 12 tiles during training; this limit can be increased at inference or during continued training to handle higher resolutions, with memory scaling linearly with tile count. The tiling algorithm is detailed in Appendix A.1.

3.2 Vision-Language Connector

Rather than using the final ViT output, `jina-vlm` concatenates features from two intermediate layers. This design is motivated by prior findings that intermediate ViT layers retain spatial detail that is progressively abstracted in later layers (Yao et al., 2024a). Following Deitke et al. (2025), we select the third-to-last and ninth-to-last layers (24th and 18th layer), as they span the transition from mid-level spatial to high-level semantic representations in this 27-layer encoder. The connector then applies attention pooling over 2×2 patch neighborhoods, using mean-pooled features as queries, a design inspired by Q-Former (Li et al., 2023a) and Perceiver Resampler (Alayrac et al., 2022). This reduces the token count by 4× while preserving local structure. A SwiGLU projection layer maps the pooled representations to the language model's embedding dimension.

In more formal terms, let $\mathbf{H}^{(\ell)} \in \mathbb{R}^{N \times d_v}$ denote the hidden states from ViT layer ℓ, where N is the number of patches, d_v is the vision encoder hidden size, and negative indices count from the final layer (e.g., $\ell = -1$ is the last layer). We concatenate features from two internal layers:

$$\mathbf{H}_{\text{concat}} = [\mathbf{H}^{(-3)}; \mathbf{H}^{(-9)}] \in \mathbb{R}^{N \times 2d_v} \tag{1}$$

For each 2×2 patch neighborhood $\mathcal{N}_i$, we compute a query vector as the mean of the neighborhood features:

$$\mathbf{q}_i = \frac{1}{4} \sum_{j \in \mathcal{N}_i} \mathbf{h}_j, \quad \mathbf{Q} = [\mathbf{q}_1; \ldots; \mathbf{q}_M] \in \mathbb{R}^{M \times 2d_v} \tag{2}$$

where $\mathcal{N}_i$ contains the four patches at positions $(2i_x, 2i_y)$, $(2i_x + 1, 2i_y)$, $(2i_x, 2i_y + 1)$, and $(2i_x + 1, 2i_y + 1)$, and $M = N/4$.

Attention pooling is then computed as:

$$\mathbf{H}_{\text{pooled}} = (\text{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_Q(\mathbf{H}_{\text{concat}}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{H}_{\text{concat}}\mathbf{W}_V)\mathbf{W}_O \in \mathbb{R}^{M \times d_v} \tag{3}$$

where $d_k = d_v$ and $\mathbf{W}_Q \in \mathbb{R}^{2d_v \times d_k}$, $\mathbf{W}_K \in \mathbb{R}^{2d_v \times d_k}$, $\mathbf{W}_V \in \mathbb{R}^{2d_v \times 2d_v}$ and $\mathbf{W}_O \in \mathbb{R}^{2d_v \times d_v}$ are learnable weight matrices. Finally, the pooled visual features are projected to the language model embedding dimension via a SwiGLU (Shazeer, 2020) layer:

$$\mathbf{H}_{\text{proj}} = (\text{Swish}(\mathbf{H}_{\text{pooled}}\mathbf{W}_1) \odot (\mathbf{H}_{\text{pooled}}\mathbf{W}_2))\,\mathbf{W}_3 \in \mathbb{R}^{M \times d_l} \tag{4}$$

where $\text{Swish}(x) = x \cdot \sigma(x)$, σ is the sigmoid function, $\odot$ denotes element-wise multiplication, $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_v \times 3d_l}$, $\mathbf{W}_3 \in \mathbb{R}^{3d_l \times d_l}$ are learnable parameters, and d_l is the language model embedding size.

3.3 Language Decoder

The language decoder is initialized from `Qwen3-1.7B-Base`[2], which empirically outperformed the instruction-tuned variant in our setting. We introduce three special tokens to structure visual inputs: `<im_start>` and `<im_end>` delimit image and thumbnail sequences, while `<im_col>` marks row boundaries within the patch grid, where tokens are arranged left-to-right and top-to-bottom. Input and output embedding weights are not tied.

3.4 Efficiency Analysis

Table 1 quantifies the computational benefits of attention pooling. With the default 12-tile configuration (plus thumbnail), the unpooled baseline would produce 9,477 visual tokens per image, while our 2×2 pooling reduces this to 2,366 tokens. Since the ViT processes each tile identically regardless of pooling, the savings apply exclusively to the LLM: we observe a 3.9× reduction in prefill FLOPs and a 4× reduction in KV-cache memory. The overall FLOPs reduction is 2.3× when including the shared ViT cost.

[2] `https://huggingface.co/Qwen/Qwen3-1.7B-Base`

Table 1: Efficiency comparison with and without 2×2 attention pooling for the default 12-tile configuration. FLOPs are computed for LLM prefill; KV-cache assumes fp16 precision.

Metric	No Pooling	With Pooling	Reduction
Visual tokens	9,477	2,366	4.0×
LLM prefill FLOPs	27.2 TFLOPs	6.9 TFLOPs	3.9×
KV-cache memory	2.12 GB	0.53 GB	4.0×

Table 2: Model training hyperparameters across pre-training and fine-tuning stages.

Hyperparameter	Pre-Training	Fine-Tuning
Warmup ViT	10%	10%
Warmup Con.	1%	10%
Warmup LLM	10%	10%
LR ViT	6e-6	5e-6
LR Con.	2e-4	5e-6
LR LLM	2e-5	1e-5
Cosine Decay	0.1	0.1
Eps.	1e-6	1e-6
Betas	0.9, 0.95	0.9, 0.95
Batch Size	128	256
Steps	25K	60K
Samples	3.2M	15.3M
Tokens	10B	37B
GPU Hours	296	1,000

4 TRAINING

Training proceeds in two stages, with both stages updating all model components (encoder, connector, and decoder) without freezing, following Deitke et al. (2025). The combined data comprises samples in multiple languages including English, Chinese, Arabic, German, Spanish, French, Italian, Japanese, Korean, Portuguese, Russian, Turkish, Vietnamese, Thai, Indonesian, Hindi, and Bengali. Table 2 summarizes hyperparameters for both stages.

4.1 STAGE 1: ALIGNMENT TRAINING

The first stage focuses on cross-language semantic grounding rather than task-specific objectives. Training data consists primarily of caption datasets (`PixmoCap` (Deitke et al., 2025), `PangeaIns` (Yue et al., 2025)) spanning diverse visual domains: natural scenes, documents, infographics, and diagrams. Following McKinzie et al. (2024), who show that mixing text-only data during multimodal training preserves language capabilities, we include 15% text-only data from `PleiAS/common_corpus` (Langlais et al., 2025) to mitigate degradation on text-only tasks. The connector uses a higher learning rate and shorter warmup than the encoder and decoder.

4.2 STAGE 2: INSTRUCTION FINE-TUNING

The second stage trains instruction-following for VQA and reasoning tasks. We combine public dataset collections, including `LLaVA OneVision` (Li et al., 2024a), `Cauldron` (Laurençon et al., 2024), `Cambrian` (Tong et al., 2024), `PangeaIns` (Yue et al., 2025), and `FineVision` (Wiedmann et al., 2025), with text-only instruction data from Singh et al. (2024). The mixture covers academic VQA, document understanding, OCR, mathematics, and reasoning. Appendix A.2 shows representative examples.

Given the diversity of instruction data, we found single-source batches more effective initially, likely due to the heterogeneous data mixture. We train for 30K steps with single-source batches, then 30K steps with mixed-source batches.

Table 3: Comparison of general visual question answering performance.

Model	AI2D	ChartQA (test avg)	TextVQA (val)	DocVQA (val)	InfoVQA (val)	OCR Bench	SEED-2 Plus	CharXiv (RQ / DQ)	Overall
jina-vlm	**82.0**	**81.9**	**83.2**	90.6	71.6	778	67.2	**32.3** / 63.5	**72.3**
Qwen2-VL-2B	74.7	73.5	79.7	89.2*	64.0*	809	62.4	23.3 / 55.0*	66.4
Qwen3-VL-2B	76.9	77.2	79.5	**92.3***	**71.9***	**858**	67.3*	28.8 / 62.3	71.6
InternVL3-2B	78.6	80.2	77.0	87.4*	67.1*	835	64.6	28.3 / 54.7	69.2
InternVL3.5-2B	78.8	80.7	76.5	88.5*	69.3*	836	**68.0**	31.6 / **65.0**	71.6

Results for models other than jina-vlm are from their respective papers (Wang et al., 2025; Zhu et al., 2025; Wang et al., 2024c), except those marked with * which were computed using VLMEvalKit. All scores represent accuracy (%) except OCRBench which uses a 0–1000 scale; for overall average computation, OCRBench scores are divided by 10 to align with the 0–100 scale of other benchmarks.

Table 4: Comparison of generic multimodal understanding and real-world understanding performance.

Model	MME (sum)	MMB v1.1 (EN)	MMStar	Overall (MM)	RealWorld QA	MME-RW (EN)	R-Bench (dis)	Overall (RW)
jina-vlm	1965.8	75.8	56.2	67.4	**68.2**	50.7	66.7	61.9
Qwen2-VL-2B	1872.0	72.2	48.0	62.4	62.9	38.7*	63.2	55.0*
Qwen3-VL-2B	2000.8*	77.8	58.3	69.2	63.9	**57.9***	67.3*	**63.0**
InternVL3-2B	**2221.2**	**78.6**	60.7	**72.9**	64.3	53.8	**67.5**	61.9
InternVL3.5-2B	2123.3	76.6	**62.7**	71.7	62.0	49.7	62.4	58.0

Results for models other than jina-vlm are from their respective papers (Wang et al., 2025; Zhu et al., 2025; Wang et al., 2024c), except those marked with * which are computed using VLMEvalKit. All scores represent accuracy (%) except MME which uses a 0–2800 scale; for overall average computation, MME scores are divided by 28 to align with the 0–100 scale of other benchmarks.

5 EVALUATION

We compare jina-vlm against lightweight VLMs across six capability areas: general VQA, multimodal comprehension, multi-image reasoning and hallucination control, mathematical reasoning, text-only performance, and multilingual understanding. All evaluations use VLMEvalKit[3] (Duan et al., 2024) with English prompts matching our training format (e.g., "Return only the letter of the best answer option" for multiple-choice, "Respond very briefly" for open-ended questions).

5.1 GENERAL VQA TASKS

Table 3 reports results on eight VQA benchmarks covering diagrams (AI2D (Kembhavi et al., 2016)), charts (ChartQA (Masry et al., 2022), CharXiv (Wang et al., 2024e)), scene text (TextVQA (Singh et al., 2019)), documents (DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022)), OCR (OCRBench (Liu et al., 2024c)), and diverse scenes (SEED-Bench-2-Plus (Li et al., 2024b)). jina-vlm achieves the highest average (72.3), with particularly strong performance on diagram interpretation and text extraction.

5.2 MULTIMODAL AND REAL-WORLD UNDERSTANDING

Table 4 shows results on multimodal comprehension (MME (Fu et al., 2025), MMB v1.1 (Liu et al., 2024b), MMStar (Chen et al., 2024b)) and real-world understanding (RealWorldQA (xAI, 2024), MME-RealWorld (Zhang et al., 2025), R-Bench (Li et al., 2024c)). jina-vlm scores 67.4 on multimodal tasks and 61.9 on real-world tasks, achieving the best RealWorldQA result (68.2).

5.3 MULTI-IMAGE REASONING AND HALLUCINATION

Table 5 reports multi-image reasoning (BLINK (Fu et al., 2024), MuirBench (Wang et al., 2024a), MMT (Ying et al., 2024)) and hallucination benchmarks that measure the tendency to fabricate

[3] https://github.com/open-compass/VLMEvalKit

Table 5: Comparison of multi-image and hallucination performance.

Model	BLINK (val)	Muir Bench	MMT (val)	Overall (MI)	HallBench (avg)	POPE (avg)	Overall (Hall)
`jina-vlm`	50.1	34.7	57.2	47.3	39.1	**90.3**	64.7
Qwen2-VL-2B	44.4	25.5*	55.1	41.7	41.7	87.9*	64.8
Qwen3-VL-2B	**53.8**	**47.4**	**60.0***	**53.7**	44.5	88.9*	66.7
InternVL3-2B	50.3	38.8	59.5	49.5	42.5	89.6	66.1
InternVL3.5-2B	51.3	44.0	58.5	51.3	**48.6**	87.2	**67.9**

Results for models other than `jina-vlm` are from their respective papers, (Wang et al., 2025; Zhu et al., 2025; Wang et al., 2024c), except those marked with * which are computed using VLMEvalKit. All scores represent accuracy (%).

Table 6: Comparison of multimodal reasoning and mathematical problem-solving performance.

Model	MMMU	MathVista	MathVision	MathVerse (Vision Only)	WeMath	LogicVista	Overall
`jina-vlm`	45.6	59.5	19.2	23.9	17.1	33.3	33.1
Qwen2-VL-2B	41.1	43.0	12.4	17.3*	10.9*	27.3*	25.3
Qwen3-VL-2B	53.4	61.3	31.6	22.7*	28.0*	35.4*	38.7
InternVL3-2B	48.6	57.0	21.7	25.3	22.4	36.9	35.3
InternVL3.5-2B	**59.0**	**71.8** / 61.5†	**42.8** / 26.5†	**53.4** / 35.3†	**48.5** / 19.1†	**47.7** / 41.4†	**50.7**

Results for models other than `jina-vlm` are from their respective papers, (Wang et al., 2025; Zhu et al., 2025; Wang et al., 2024c), except those marked with * which are computed using VLMEvalKit. † indicates scores for `InternVL3.5-2B` without thinking mode, evaluated using VLMEvalKit. All scores represent accuracy (%).

Table 7: Comparison of text-only benchmarks.

Model	MMLU	MMLU-Pro	GSM-8k	ARC-C	HellaSwag	Overall
`jina-vlm`	56.1	30.3	71.3	**77.3**	**59.4**	58.9
Qwen3-1.7B	**62.6**	**46.4**	**75.3**	73.4	59.0	**63.3**

Results are collected using our evaluation code. All scores represent accuracy (%).

visual details (HallBench (Guan et al., 2024), POPE (Li et al., 2023b)). `jina-vlm` scores 47.3 on multi-image tasks, which is expected given limited multi-image training data, but achieves the best POPE score (90.3), indicating low hallucination rates.

5.4 Mathematical Reasoning

Table 6 reports structured reasoning benchmarks: multidisciplinary comprehension (MMMU (Yue et al., 2024)), visual mathematics (MathVista (Lu et al., 2024a), MathVision (Wang et al., 2024b), MathVerse (Zhang et al., 2024), WeMath (Qiao et al., 2025)), and logical reasoning (LogicVista (Xiao et al., 2024)). `jina-vlm` performs comparably to InternVL3-2B and outperforms Qwen2-VL-2B.

5.5 Text-Only Performance

Table 7 compares `jina-vlm` against the backbone Qwen3-1.7B on text-only benchmarks: MMLU (Hendrycks et al., 2021), MMLU-Pro (Wang et al., 2024d), GSM-8K (Cobbe et al., 2021), ARC-C (Clark et al., 2018), and HellaSwag (Zellers et al., 2019). Results show mixed preservation of text-only capabilities: `jina-vlm` matches or exceeds the backbone on commonsense reasoning (ARC-C, HellaSwag) and retains most performance on MMLU and GSM-8K. However, MMLU-Pro shows substantial degradation (46.4 $\rightarrow$ 30.3), likely because this benchmark emphasizes extended multi-step reasoning that conflicts with our instruction-tuning toward concise visual responses.

Table 8: Comparison of multilingual multimodal understanding performance.

Benchmark		jina-vlm	Qwen2-VL-2B	Qwen3-VL-2B	InternVL3-2B	InternVL3.5-2B
MMMB	ar	**76.9**	68.3	72.7*	68.6	68.5
	cn	**80.0**	74.2	75.7*	78.3	77.7
	en	**82.0**	78.3	80.7*	81.9	80.2
	pt	**79.2**	72.6	75.0*	75.4	75.9
	ru	**79.2**	72.8	75.9*	74.6	76.3
	tr	**75.5**	61.8	68.5*	62.9	69.1
	avg	**78.8**	71.3	75.0*	73.6	74.6
Multi. MMBench	ar	**70.0**	66.7	66.2*	66.4	63.7
	cn	75.9	67.0	75.7*	**77.8**	75.9
	en	78.8	71.1	77.8*	**81.3**	78.4
	pt	74.7	72.1	71.4*	**75.9**	73.7
	ru	75.3	69.9	**75.9***	70.7	71.4
	tr	**71.1**	69.3	67.0*	59.5	62.0
	avg	**74.3**	69.4	72.3*	71.9	70.9
MTVQA		25.6	20.6	27.3*	26.7	**28.5**
Overall		**59.6**	53.8	58.2	57.4	58.0

Results for baseline models are derived from their original publications (Wang et al., 2025; Zhu et al., 2025; Wang et al., 2024c), except those marked with * which are computed using VLMEvalKit. All scores represent accuracy (%).

5.6 Multilingual Understanding

Table 8 reports multilingual multimodal benchmarks: MMMB (Sun et al., 2025), Multilingual MMBench (Sun et al., 2025), and MTVQA (Tang et al., 2025). jina-vlm achieves state-of-the-art multilingual performance among 2B-scale VLMs, with the highest averages on MMMB (78.8) and Multilingual MMBench (74.3). In our evaluation setup, the system prompt is in English (following the training format, not a specific deployment target) while the questions, answer options, and text within images are presented in the target language. Baseline models use their default prompts. This setup tests cross-lingual visual understanding but not target-language instruction-following.

6 Data Mixture Analysis

The rapid advancement of large-scale VLMs has been driven by scaling both model size and training data. Aggregate data volume and domain diversity have received significant attention (Tong et al., 2024; Deitke et al., 2025; Li et al., 2025b); however, the composition and relative importance of different data types within training mixtures remain poorly understood. This issue is particularly pressing when training small-scale models, where limited budget needs to be spent optimally. In such settings, we hypothesize that, besides data quality, dataset mixture composition is critical.

We conduct a comprehensive ablation study that systematically removes different data types from our training mixture and measures the downstream impact across diverse capabilities. By training 9 variants of our model, each excluding a specific category of data—such as OCR datasets, VQA data, or multilingual content—we can quantify both the necessity of each component and potential cross-domain transfer effects. Our approach provides diagnostic insights into data category necessity and cross-domain transfer, and offers preliminary empirical evidence for mixture optimization under resource constraints.

6.1 Experimental Design

Our training dataset consists of data across two stages, alignment and instruction fine-tuning, as described in sections 4.1 and 4.2. Besides dividing by stage, we group data along four axes: modality (multimodal vs. text-only), language (English vs. multilingual), task type (VQA, OCR, open-ended, reasoning, etc.), and domain (real-world, math, charts, documents, code, etc.). Language detection is done automatically using a language classifier, and modalities are derived from each dataset's

schema. Domain and task type categorization is done through manual inspection of examples in each dataset, after defining a high-level taxonomy. Tables 10, 11 and 12 in Appendix A.3 show the detailed data mixture composition for the instruction fine-tuning stage.

Starting from the original training mixture, we design 9 experiments, each removing one category—**Task:** `Xvqa` (VQA data), `Xocr` (OCR data); **Domain:** `Xrealworld` (real-world photos), `Xmath` (math/reasoning), `Xcharts` (charts/diagrams), `Xdocs` (documents), `Xcode` (code); **Modality & Language:** `Xtext` (text-only data), `Xmlingual` (non-English data). All ablations run against the baseline `base` that uses the original data mixture used to train `jina-vlm`.

To enable efficient exploration within computational constraints, all experiments, including the baseline, run for 10% of the full training duration. This is a known limitation: ablations at 10% may not fully reflect final-training behavior, and findings should be treated as diagnostic observations rather than definitive conclusions. Ablation runs do not adjust steps to match the compute of the baseline; when a data category is removed, remaining data may be seen more frequently within the fixed budget. Training order is held fixed across all ablations; disentangling order from composition effects would require additional controlled experiments. Hyperparameters are fixed across all runs and the same two-stage training recipe is used. Each ablation is applied consistently across both stages when applicable (for example, `Xmlingual` removes non-English data from both alignment and instruction stages). Training checkpoints are evaluated against the same comprehensive evaluation suite described in Section 5.

6.2 Results

Table 9 summarizes ablation results; observations are correlational, based on 10% of full training, and should be treated as hypotheses rather than established mechanisms.

Removing VQA data (`Xvqa`) causes dramatic but highly task-specific degradation on VQA benchmarks with minimal impact on other tasks. Similarly, `Xocr` harms OCRBench while improving all other tasks. This confirms that task-specific data (VQA, OCR) primarily benefits its own tasks without broad transfer. Domain removals show predictable patterns for `Xcharts` (ChartQA -39.4%) and `Xmath` (MathVista -4.6%). More surprisingly, removing real-world (`Xrealworld`) or document data (`Xdocs`) improves several unrelated benchmarks (HellaSwag $+9.1\%$, MMMU $+7.3\%$), while code removal (`Xcode`) degrades HellaSwag (-10.6%). Removing text-only data (`Xtext`) consistently improves multimodal benchmarks but hurts HellaSwag, the only unimodal task. Removing multilingual data (`Xmlingual`) improves most tasks but, as expected, drops MTVQA by -16.1%.

These results support three conclusions for practitioners under fixed compute budgets. First, task-specific data (VQA, OCR, charts, math) is necessary for its capabilities but does not transfer broadly—consistent with MM1 (McKinzie et al., 2024) and Cambrian-1 (Tong et al., 2024). Second, real-world and document data exhibit unexpected interference: their removal improves several unrelated benchmarks, suggesting noise or gradient conflict at this scale. Third, text-only and multilingual data impose a cost on multimodal performance but are essential for their respective capabilities; practitioners who do not require them should omit them.

7 Conclusion

We presented `jina-vlm`, a 2.4B vision-language model that achieves state-of-the-art multilingual VQA and leading English VQA results among open 2B-scale VLMs.

The current approach has limitations. Multi-tile processing introduces computational overhead that scales with image resolution, and tiling can fragment global spatial context, potentially impairing performance on tasks requiring holistic scene understanding such as object counting or precise spatial reasoning across tile boundaries. While the global thumbnail partially mitigates this, native-resolution approaches (Dehghani et al., 2023) may be better suited for such tasks. We have not emphasized safety-critical training or alignment, and multi-image reasoning remains weak due to limited training data in this regime. Future work could explore more efficient resolution handling, targeted improvements for counting and spatial tasks, and investigate whether our multilingual training recipe transfers to larger model scales.

Table 9: Ablation study results.

Model	AI2D	ChartQA	DocVQA	InfoVQA	OCRBench
base	72.0 (+0.0)	64.5 (+0.0)	81.8 (+0.0)	63.2 (+0.0)	699.0 (+0.0)
Xvqa	65.1 (-9.5)	50.2 (-22.1)	75.3 (-8.0)	48.8 (-22.8)	686.0 (-1.9)
Xmlingual	74.4 (+3.3)	70.9 (+9.9)	84.0 (+2.6)	64.9 (+2.6)	**736.0 (+5.3)**
Xrealworld	74.0 (+2.8)	64.9 (+0.6)	**84.3 (+3.1)**	**65.0 (+2.8)**	717.0 (+2.6)
Xtext	72.7 (+1.1)	**70.4 (+9.1)**	84.0 (+2.7)	64.3 (+1.7)	718.0 (+2.7)
Xocr	**74.7 (+3.8)**	69.7 (+8.0)	82.3 (+0.6)	64.6 (+2.2)	622.0 (-11.0)
Xcharts	65.7 (-8.7)	39.1 (-39.4)	83.0 (+1.5)	63.8 (+0.8)	691.0 (-1.1)
Xdocs	74.4 (+3.5)	69.8 (+8.2)	80.6 (-1.4)	63.9 (+1.1)	686.0 (-1.9)
Xmath	72.9 (+1.4)	64.9 (+0.6)	83.3 (+1.8)	64.0 (+1.2)	687.0 (-1.7)
Xcode	71.6 (-0.5)	64.1 (-0.7)	82.6 (+1.0)	64.0 (+1.2)	700.0 (+0.1)
Model	**TextVQA**	**MMStar**	**MMBench1.1**	**RealWorldQA**	**MMMU**
base	71.0 (+0.0)	46.7 (+0.0)	69.1 (+0.0)	60.0 (+0.0)	41.0 (+0.0)
Xvqa	57.8 (-18.6)	44.8 (-4.0)	66.9 (-3.2)	60.7 (+1.1)	40.6 (-1.1)
Xmlingual	**75.5 (+6.4)**	48.3 (+3.4)	70.5 (+2.1)	61.7 (+2.8)	40.9 (-0.3)
Xrealworld	72.2 (+1.7)	47.7 (+2.1)	68.8 (-0.5)	59.9 (-0.2)	42.3 (+3.3)
Xtext	**75.5 (+6.3)**	**48.6 (+4.1)**	69.8 (+1.1)	**63.8 (+6.3)**	42.6 (+3.8)
Xocr	73.5 (+3.5)	48.3 (+3.4)	**71.8 (+4.0)**	60.9 (+1.5)	42.4 (+3.5)
Xcharts	72.0 (+1.5)	43.8 (-6.1)	67.5 (-2.3)	62.1 (+3.5)	41.2 (+0.5)
Xdocs	75.0 (+5.6)	47.0 (+0.7)	70.9 (+2.6)	59.6 (-0.7)	**44.0 (+7.3)**
Xmath	73.3 (+3.3)	47.2 (+1.1)	68.8 (-0.5)	61.0 (+1.7)	40.9 (-0.3)
Xcode	73.1 (+3.0)	47.2 (+1.1)	69.3 (+0.3)	60.0 (+0.0)	43.1 (+5.1)
Model	**MathVista**	**HellaSwag**	**HallBench**	**POPE**	**MTVQA**
base	45.8 (+0.0)	48.9 (+0.0)	53.8 (+0.0)	86.7 (+0.0)	**23.9 (+0.0)**
Xvqa	39.9 (-12.9)	51.7 (+5.6)	51.4 (-4.5)	87.4 (+0.8)	22.5 (-6.0)
Xmlingual	**48.9 (+6.8)**	48.4 (-1.2)	53.5 (-0.6)	81.5 (-6.0)	20.1 (-16.1)
Xrealworld	46.9 (+2.4)	**53.4 (+9.1)**	54.3 (+0.8)	84.0 (-3.1)	23.0 (-4.0)
Xtext	47.9 (+4.6)	45.4 (-7.3)	53.7 (-0.2)	87.6 (+1.0)	23.5 (-1.8)
Xocr	48.0 (+4.8)	52.8 (+8.0)	55.1 (+2.3)	86.0 (-0.8)	23.1 (-3.7)
Xcharts	40.8 (-10.9)	50.8 (+3.9)	52.4 (-2.7)	**88.8 (+2.4)**	23.9 (-0.1)
Xdocs	47.3 (+3.3)	50.9 (+4.0)	52.6 (-2.3)	86.1 (-0.7)	23.3 (-2.6)
Xmath	43.7 (-4.6)	48.9 (+0.0)	**55.4 (+2.9)**	86.2 (-0.6)	23.2 (-3.0)
Xcode	46.9 (+2.4)	43.8 (-10.6)	53.4 (-0.8)	86.0 (-0.8)	23.2 (-3.0)

Some evaluation tasks from Section 5 are omitted due to insignificant variance. Relative differences to the baseline (%) are shown in parentheses. **Bold** marks the best score per column. All scores represent accuracy (%) except OCRBench (0–1000 scale).

Our data mixture analysis offers diagnostic observations for resource-constrained training; as ablations ran at 10% of full training, findings should be interpreted as preliminary evidence motivating future systematic investigation.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, 2024.

Manoj Acharya, Kushal Kafle, and Christopher Kanan. TallyQA: Answering Complex Counting Questions. In *Proc. 33rd AAAI Conference on Artificial Intelligence*, AAAI'19, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A Visual Language Model for Few-Shot Learning. In *Proc. 36th International Conference on Neural Information Processing Systems*, NeurIPS 2022, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

Lucas Beyer, Andreas Steiner, André Susano Pinto, et al. PaliGemma: A Versatile 3B VLM for Transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Liang Chen, Haozhe Zhao, Tianyu Liu, et al. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. In *Computer Vision – ECCV 2024*, pp. 19–35, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, et al. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Proc. 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, 2024b.

Xi Chen, Xiao Wang, Soravit Changpinyo, et al. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *ICLR 2023*, 2023.

Zhe Chen, Weiyun Wang, Hao Tian, et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *Sci. China Inf. Sci.*, 67, 2024c.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024d.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, et al. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, et al. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, et al. Patch n' pack: NaViT, a vision transformer for any aspect ratio and resolution. In *Proc. 37th International Conference on Neural Information Processing Systems*, NeurIPS 2023, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations*, ICLR 2021, 2021.

Haodong Duan, Xinyu Fang, Junming Yang, et al. VLMEvalKit: An Open-Source Toolkit for Evaluating Large Multi-Modality Models. In *Proc. 32nd ACM International Conference on Multimedia*, MM '24, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, et al. BLINK: Multimodal Large Language Models Can See but Not Perceive. In *Computer Vision – ECCV 2024*, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *Int. J. Comput. Vision*, pp. 398–414, April 2019.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv preprint arXiv:2003.10286*, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2021.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary Position Embedding for Vision Transformer. In *Computer Vision – ECCV 2024*, 2024.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, et al. ScreenQA: Large-Scale Question-Answer Pairs over Mobile App Screenshots. In *Proc. 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.

Yiming Jia, Jiachen Li, Xiang Yue, et al. VisualWebInstruct: Scaling up Multimodal Instruction Data through Web Search. In *Proc. 2025 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2025, 2025.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, et al. A Diagram Is Worth A Dozen Images. In *Computer Vision – ECCV 2016*, 2016.

Pierre-Carl Langlais, Carlos Rosas Hinostroza, et al. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training. *arXiv preprint arXiv:2506.01732*, 2025.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *Proc. 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, et al. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. SEED-Bench-2-Plus: Benchmarking Multimodal Large Language Models with Text-Rich Visual Comprehension. *arXiv preprint arXiv:2404.16790*, 2024b.

Chunyi Li, Jianbo Zhang, Zicheng Zhang, et al. R-Bench: Are your Large Multimodal Model Robust to Real-world Corruptions? *arXiv preprint arXiv:2410.05474*, 2024c.

Feng Li, Renrui Zhang, Hao Zhang, et al. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*, 2024d.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proc. 40th International Conference on Machine Learning*, ICML 2023, 2023a.

Xu Li, Yuxuan Liang, Xiaolei Chen, Yi Zheng, Haotian Chen, Bin Li, and Xiangyang Xue. HERO: Rethinking Visual Token Early Dropping in High-Resolution Large Vision-Language Models. *arXiv preprint arXiv:2509.13067*, 2025a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2023, 2023b.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, Nadine Chang, Karan Sapra, Amala Sanjay Deshmukh, Tuomas Rintamaki, Matthieu Le, Ilia Karmanov, Lukas Voegtle, Philipp Fischer, De-An Huang, Timo Roman, Tong Lu, Jose M. Alvarez, Bryan Catanzaro, Jan Kautz, Andrew Tao, Guilin Liu, and Zhiding Yu. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models, 2025b. URL `https://arxiv.org/abs/2501.14818`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Proc. 37th International Conference on Neural Information Processing Systems*, NeurIPS 2023, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, et al. MMBench: Is Your Multi-modal Model an All-around Player? In *Computer Vision – ECCV 2024*, 2024b.

Yuliang Liu, Zhang Li, Mingxin Huang, et al. OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models. *Science China Information Sciences*, 67(12), 2024c.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, et al. NVILA: Efficient Frontier Visual Language Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Pan Lu, Hritik Bansal, Tony Xia, et al. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *12th International Conference on Learning Representations*, ICLR 2024, 2024a.

Shiyin Lu, Yang Li, Qing-Guo Chen, et al. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *arXiv preprint arXiv:2405.20797*, 2024b.

Shiyin Lu, Yang Li, Yu Xia, et al. Ovis2.5 Technical Report. *arXiv preprint arXiv:2508.11737*, 2025.

Andrei-Alexandru Manea and Jindřich Libovický. Multilingual Vision-Language Models, A Survey. *arXiv preprint arXiv:2509.22123*, 2025.

Andrés Marafioti, Orr Zohar, Miquel Farré, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A Dataset for VQA on Document Images. *arXiv preprint arXiv:2007.00398*, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. InfographicVQA. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods,

analysis & insights from multimodal llm pre-training, 2024. URL https://arxiv.org/abs/2403.09611.

Runqi Qiao, Qiuna Tan, Guanting Dong, et al. We-Math: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning? In *Proc. 63rd Annual Meeting of the Association for Computational Linguistics: ACL 2025*, 2025.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. In *Proc. IEEE/CVF International Conference on Computer Vision*, ICCV 2025, pp. 22857–22867, 2025.

Zhenwei Shao, Zhou Yu, Jun Yu, Xuecheng Ouyang, et al. Imp: Highly Capable Large Multimodal Models for Mobile Devices. *IEEE Transactions on Multimedia*, 27:2961–2974, 2025.

Noam Shazeer. GLU Variants Improve Transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, et al. Towards VQA Models That Can Read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Shivalika Singh et al. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proc. 62nd Annual Meeting of the Association for Computational Linguistics*, ACL 2024, 2024.

Andreas Steiner, André Susano Pinto, Michael Tschannen, et al. PaliGemma 2: A Family of Versatile VLMs for Transfer. *arXiv preprint arXiv:2412.03555*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced Transformer with Rotary Position Embedding. *Neurocomput.*, February 2024.

Hai-Long Sun, Da-Wei Zhou, Yang Li, et al. Parrot: Multilingual Visual Instruction Tuning. In *Proc. 42nd International Conference on Machine Learning*, ICML 2025, 2025.

Jingqun Tang, Qi Liu, Yongjie Ye, et al. MTVQA: Benchmarking Multilingual Text-Centric Visual Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, et al. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Proc. 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, 2024.

Michael Tschannen, Alexey Gritsenko, et al. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025.

Fei Wang, Xingyu Fu, James Y. Huang, et al. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Ke Wang, Junting Pan, Weikang Shi, et al. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In *Proc. 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, 2024b.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024c.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yubo Wang, Xueguang Ma, Ge Zhang, et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. In *Proc. 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, 2024d.

Zirui Wang, Mengzhou Xia, Luxi He, et al. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. In *Proc. 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, 2024e.

Luis Wiedmann, Orr Zohar, Amir Mahla, et al. FineVision: Open Data Is All You Need. *arXiv preprint arXiv:2510.17269*, 2025.

xAI. RealWorldQA: A Benchmark for Real-World Spatial Understanding, 2024. URL `https://x.ai/news/grok-1.5v#real-world-understanding`.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, et al. PyramidDrop: Accelerating Your Large Vision-Language Models via Pyramid Visual Redundancy Reduction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ruyi Xu, Yuan Yao, Zonghao Guo, et al. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. In *Computer Vision – ECCV 2024*, 2024.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025a.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is Better but Not Necessary in Vision Language Models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Huanjin Yao, Wenhao Wu, Taojiannan Yang, Yuxin Song, Mengxi Zhang, Haocheng Feng, Yifan Sun, Zhiheng Li, Wanli Ouyang, and Jingdong Wang. Dense Connector for MLLMs. In *Proc. 38th International Conference on Neural Information Processing Systems*, NeurIPS 2024, 2024a.

Yuan Yao, Tianyu Yu, Ao Zhang, et al. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*, 2024b.

Kaining Ying, Fanqing Meng, Jin Wang, et al. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. In *Proc. 41st International Conference on Machine Learning*, ICML 2024, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiang Yue, Yueqi Song, Akari Asai, et al. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages. In *13th International Conference on Learning Representations*, ICLR 2025, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proc. 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, et al. MathVerse: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems? In *Computer Vision – ECCV 2024*, 2024.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, et al. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? In *13th International Conference on Learning Representations*, ICLR 2025, 2025.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, et al. TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance. In *Proc. 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

A APPENDIX

A.1 PSEUDOCODE FOR CREATING OVERLAPPING TILES

Algorithm 1: GETALLTILESOVERLAPANDRESIZE

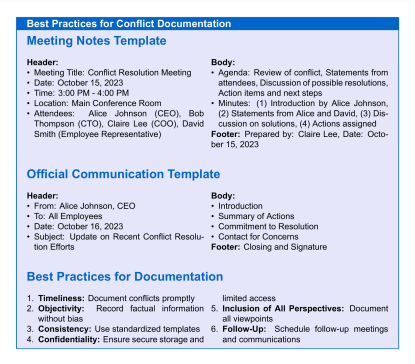
Input: Image I of size (h, w);
Base input size $\mathbf{b} = (b_h, b_w)$ ($(378, 378)$)
Patch size p (14);
Maximum number of tiles M (12 by default, configurable)
Overlap margins (m_L, m_R) in patches ($(4, 4)$)
Output: List of tiles $\mathcal{C}$ (thumbnail + grid tiles)
Tiling (t_h, t_w) = (number of rows, number of columns)

1. Compute overlap-related sizes
$m_{\text{tot}} \leftarrow p \cdot (m_L + m_R)$ `// Total overlap margin in pixels`
$s_{\text{win}} \leftarrow \big(\lfloor b_h/p \rfloor - (m_L + m_R)\big) \cdot p$ `// Tile stride in pixels`
2. Select tiling on the margin-reduced image
$(t_h, t_w) \leftarrow \text{SELECTTILINGWITHMINIMALSCALECHANGE}\big(h - m_{\text{tot}},\ w - m_{\text{tot}},\ s_{\text{win}},\ M\big)$;
3. Resize image to exactly fit the chosen tiling + margins;
$H' \leftarrow t_h \cdot s_{\text{win}} + m_{\text{tot}}$;
$W' \leftarrow t_w \cdot s_{\text{win}} + m_{\text{tot}}$;
$I_{\text{grid}} \leftarrow \text{RESIZE}(I,\ [H', W'])$;
4. Extract overlapping tiles
$\mathcal{G} \leftarrow \text{EXTRACTTILES}\big(I_{\text{grid}},\ (t_h, t_w),\ s_{\text{win}},\ b_h\big)$ `// ` b_h `is the tile height, equal to` b_w `here`
5. Build thumbnail and final tile list
$T \leftarrow \text{RESIZE}(I,\ [b_h, b_w])$ `// Global thumbnail`
$\mathcal{C} \leftarrow [T] \mathbin{+\!\!+} \mathcal{G}$ `// Concatenate thumbnail and tiles`
return $(\mathcal{C}, (t_h, t_w))$;

A.2 Training Set Examples

Captioning & Instruction

Dataset: `VisualWebInstruct` Jia et al. (2025)



Best Practices for Conflict Documentation

Meeting Notes Template

Header:
- Meeting Title: Conflict Resolution Meeting
- Date: October 15, 2023
- Time: 3:00 PM - 4:00 PM
- Location: Main Conference Room
- Attendees: Alice Johnson (CEO), Bob Thompson (CTO), Claire Lee (COO), David Smith (Employee Representative)

Body:
- Agenda: Review of conflict, Statements from attendees, Discussion of possible resolutions, Action items and next steps
- Minutes: (1) Introduction by Alice Johnson, (2) Statements from Alice and David, (3) Discussion on solutions, (4) Actions assigned

Footer: Prepared by: Claire Lee, Date: October 15, 2023

Official Communication Template

Header:
- From: Alice Johnson, CEO
- To: All Employees
- Date: October 16, 2023
- Subject: Update on Recent Conflict Resolution Efforts

Body:
- Introduction
- Summary of Actions
- Commitment to Resolution
- Contact for Concerns

Footer: Closing and Signature

Best Practices for Documentation

1. **Timeliness:** Document conflicts promptly
2. **Objectivity:** Record factual information without bias
3. **Consistency:** Use standardized templates
4. **Confidentiality:** Ensure secure storage and limited access
5. **Inclusion of All Perspectives:** Document all viewpoints
6. **Follow-Up:** Schedule follow-up meetings and communications

Question

what is the meeting title?

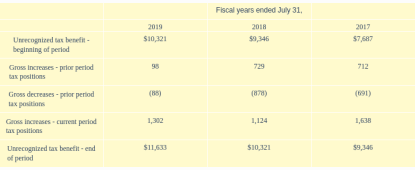
Answer

Conflict Resolution Meeting

Figure 2: Answer questions given web documents.

Charts & Tables

Dataset: `TAT-QA` Zhu et al. (2021)



	Fiscal years ended July 31,		
	2019	2018	2017
Unrecognized tax benefit - beginning of period	$10,321	$9,346	$7,687
Gross increases - prior period tax positions	98	729	712
Gross decreases - prior period tax positions	(88)	(878)	(691)
Gross increases - current period tax positions	1,302	1,124	1,638
Unrecognized tax benefit - end of period	$11,633	$10,321	$9,346

Question

Unrecognized Tax Benefits Activity related to unrecognized tax benefits is as follows (in thousands): ... As of July 31, 2019, the Company has no income tax audits in progress in the U.S. or foreign jurisdictions. What was the increase in unrecognized tax benefits in 2019?

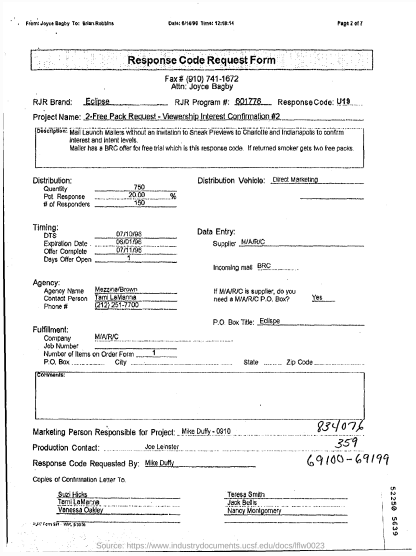
Answer

$1.3 million.

Figure 3: Financial table requiring numerical reasoning over text.

Document Understanding & Infographics

Dataset: `DocVQA` Mathew et al. (2021)



From: Joyce Bagby To: Brian Ratchke Date: 8/14/96 Time: 12:09:14 Page 2 of 7

Response Code Request Form

Fax # (910) 741-1672
Attn: Joyce Bagby

RJR Brand: Eclipse RJR Program #: 501778 Response Code: U19

Project Name: 2-Free Pack Request - Viewership Interest Confirmation #2

Description: Mail Launch Mailers without an invitation to Sneak Previews to Charlotte and Indianapolis to confirm interest and intent levels.
Mailer has a BRC offer for free trial which is this response code. If returned smoker gets two free packs.

Distribution:
Quantity 750
Pct. Response 20.00 %
# of Respondere 150

Distribution Vehicle: Direct Marketing

Timing:
DTS 07/10/96
Expiration Date 08/01/96
Offer Complete 07/11/96
Days Offer Open 1

Data Entry:
Supplier M/A/R/C
Incoming mail BRC

Agency:
Agency Name Mezzina/Brown
Contact Person Terri LaMarina
Phone # (212) 251-7700

If M/A/R/C is supplier, do you need a M/A/R/C P.O. Box? Yes

P.O. Box Title: Eclipse

Fulfillment:
Company M/A/R/C
Job Number
Number of Items on Order Form 1
P.O. Box City State Zip Code

Comments:

Marketing Person Responsible for Project Mike Duffy - 0910 834076
Production Contact: Joe Lenster 359
Response Code Requested By: Mike Duffy 69100 - 69199

Copies of Confirmation Letter To:

Sue Hicks
Terri LaMarina
Vanessa Oakley

Teresa Smith
Jack Balls
Nancy Montgomery

RJR Form #1 - Rev. 3/3/95

Source: https://www.industrydocuments.ucsf.edu/docs/lflw0023

Question
what is the response code ?

Answer
U19

Figure 4: Document image with question about textual fields.

OCR QA (text-centric VQA)

Dataset: `TextVQA` Singh et al. (2019)

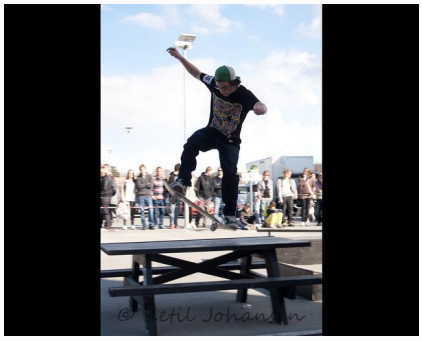
Question
what number is the cab

Answer
3G54

Figure 5: Photo with textual question needing OCR reading.

General VQA

Dataset: `VQAv2` Goyal et al. (2019)



Question
Where is he looking?

Answer
down

Figure 6: General visual question answering on natural images.

Grounding, Spatial & Counting

Dataset: `TallyQA` Acharya et al. (2019)

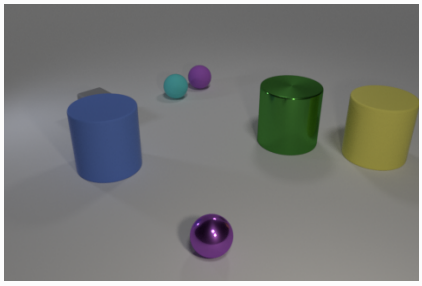
Question
How many more people can ride on the vehicle?

Answer
0

Figure 7: Scene requiring counting and spatial reasoning accuracy.

Math & Geometry (vision)

Dataset: `CLEVR` Johnson et al. (2017)



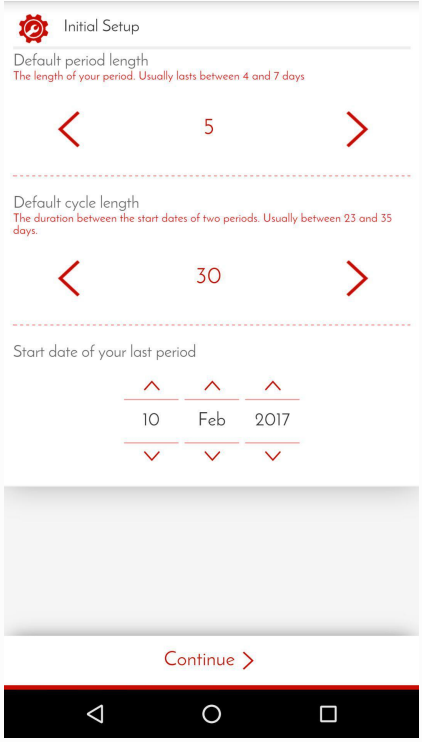
Question
There is a large shiny object; does it have the same shape as the object right of the large metallic thing? Provide a short and direct response.

Answer
Yes.

Figure 8: Synthetic shapes testing compositional spatial reasoning.

Screens, Web & GUI

Dataset: `ScreenQA` Hsiao et al. (2025)



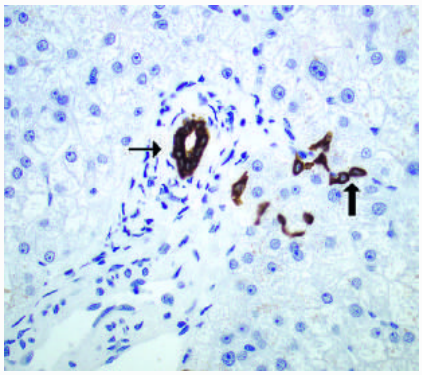
Question
What is the default period length?

Answer
5 days

Figure 9: User interface screenshot with structured textual elements.

Medical (vision)

Dataset: `PathVQA` He et al. (2020)



Question
where are liver stem cells (oval cells) located?

Answer
in the canals of hering

Figure 10: Microscopic pathology image for medical VQA.

Text-only (instruction / code / math / translation)

Dataset: `aya_dataset` Singh et al. (2024)

Question
Quels président des États-Unis ne s'est jamais marié?

Answer
James Buchanan est le seul président qui ne s'est jamais marié.

Figure 11: Text-only tasks covering multiple languages.

A.3 Instruction Fine-Tuning Stage Dataset Composition

Table 10: Instruction Fine-Tuning dataset composition

Category	Percentage (%)
Tasks	
VQA	29.08
OCR	17.92
QA	15.88
Open-ended VQA	8.57
Reasoning QA	8.49
Instruction Following	8.26
Pointing VQA	4.05
Captions	2.36
Long Captions	2.09
Counting VQA	1.08
Action/Instruction Following	0.82
Navigation	0.64
Classification	0.28
Misc	0.21
Sentiment Analysis	0.20
Multi Image Reasoning	0.07
Domains	
Real World	24.53
Math	12.46
Charts & Diagrams	12.14
Web	12.00
Documents	7.20
Code	5.93
Misc	5.15
Screens & GUI	3.97
Geometry	3.91
Tables	3.15
Handwritten	2.02
Synthetic Scenes	1.16
Landmarks	1.11
Infographics	1.02
Medical	0.92
Captcha	0.68
Academic	0.64
Science	0.64
Maps	0.49
Rendered Text	0.29
Books	0.26
Receipts	0.18
Comics	0.13
Modalities	
Image+Text	82.21
Text	17.79
Languages	
English	85.17
Multilingual	14.83

Table 11: Task type reference table

Task Type	Description
VQA	Visual question answering - questions about the image
OCR	Optical character recognition
QA	Question answering where there is no image or the image does not provide the knowledge needed to answer the question
Open-ended VQA	Open-ended questions e.g. "Describe the scene" or "What is happening in the picture?"
Reasoning QA	Question answering with a specific request to return the reasoning
Instruction Following	Following instructions given in the text
Pointing VQA	Questions requiring localization or pointing to specific regions
Captions	Short image descriptions (1-2 sentences)
Long Captions	Detailed image descriptions (multiple sentences, comprehensive)
Counting VQA	Questions requiring counting objects in images
Action/Instruction Following	Understanding and following action-based instructions
Navigation	Understanding spatial relationships and navigation
Classification	Categorizing images into predefined classes
Sentiment Analysis	Analyzing sentiment from combined image and text
Multi Image Reasoning	Reasoning across multiple images

Table 12: Domain reference table

Domain	Description
Real World	Natural photos of real-world scenes and objects (landscapes, people, animals)
Math	Math problems, equations
Charts & Diagrams	Data visualizations, flowcharts, graphs, plots
Web	Websites, memes, AI-generated images and generic web content
Documents	General document images (reports, forms)
Code	Code snippets, technical documentation
Misc	Miscellaneous content, various domains not covered elsewhere
Screens & GUI	Computer screens, mobile apps, software interfaces
Geometry	Geometric problems, trigonometry, proofs, mathematical drawings
Tables	Structured tabular data
Handwritten	Handwritten text
Synthetic Scenes	Computer-generated depictions of real-world scenes
Landmarks	Famous buildings, monuments, tourist attractions
Infographics	Graphics combining text and visuals (event posters, leaflets)
Medical	Healthcare images, anatomy, medical scans, clinical documents
Captcha	CAPTCHA verification images
Academic	Research articles and content from various academic domains
Science	Scientific images, experiments, lab setups, diagrams, educational content
Maps	Geographic maps, navigation, spatial information
Rendered Text	Computer-generated text overlays and typography
Books	Book pages, literature
Receipts	Purchase receipts, invoices, transaction records
Comics	Comic books, manga, graphic novels