SmolVLM: Redefining small and efficient multimodal models

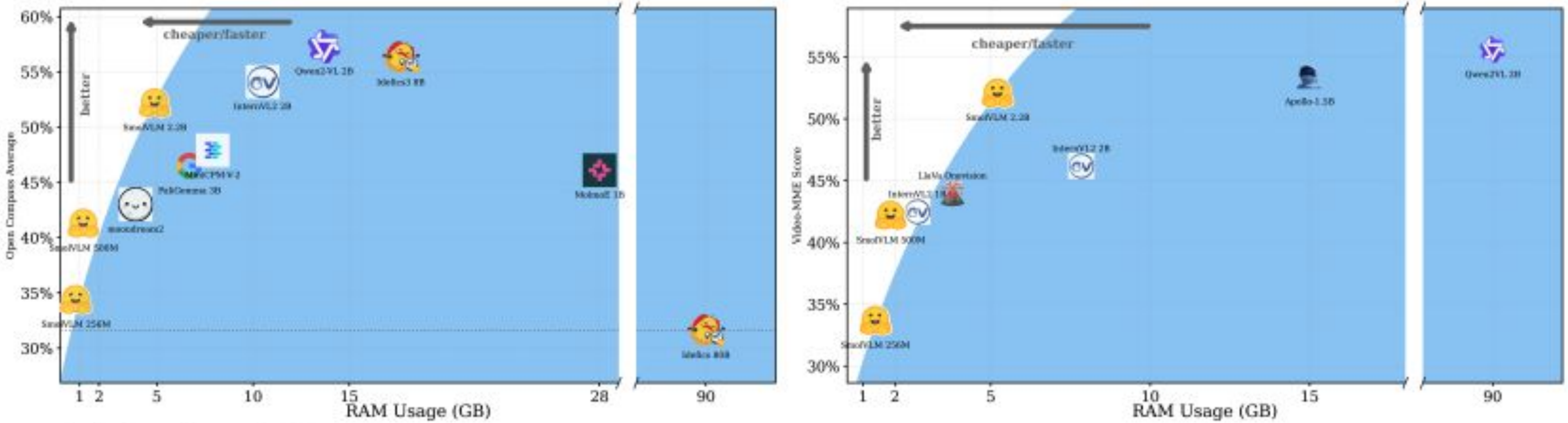


Figure 1 | **Smol yet Mighty:** comparison of SmolVLM with other state-of-the-art small VLM models. Image results are sourced from the OpenCompass OpenVLM leaderboard (Duan et al., 2024).

Introduction

- Compact yet Powerful Models
- Efficient GPU Memory Usage
  - smallest model runs inference using less than 1GB GPU RAM
- Systematic Architectural Exploration
- Robust Video Understanding on Edge Devices
- Fully Open-source Resources
  - weights, datasets, code

Architecture

- How to assign compute between vision and language towers?
  - Language Model : SmolLM2 (135M, 360M, 1.7B)
  - Vision Model : SigLIP-B/16(93M), SigLIP-SO400M(428M)

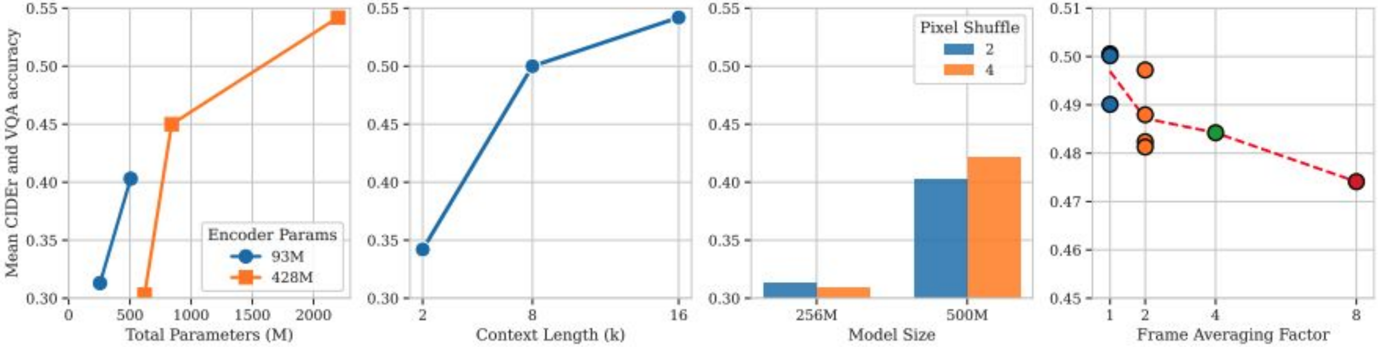


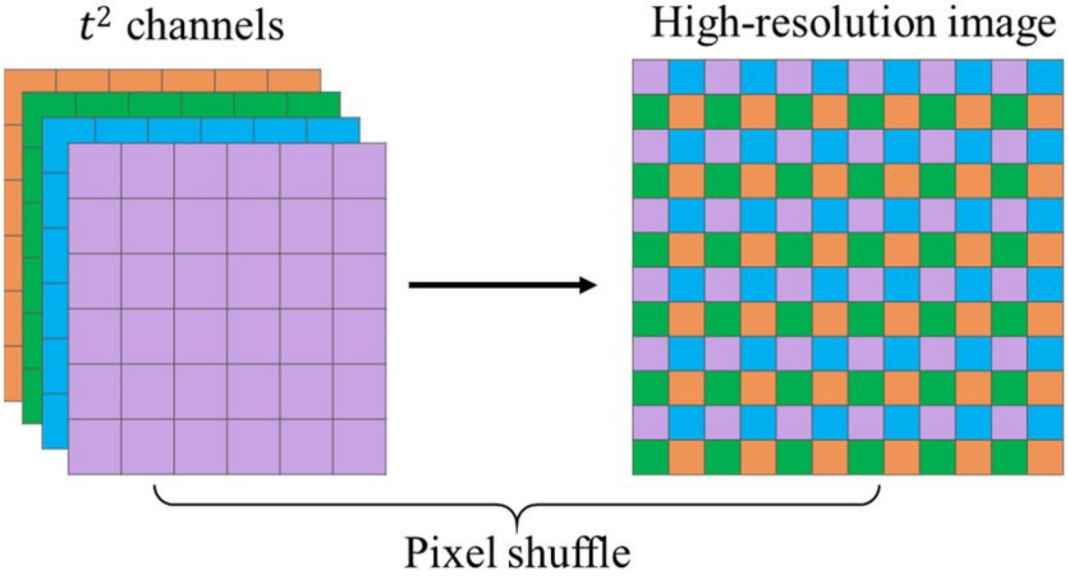
Figure 3 | Performance analysis of SmolVLM configurations. *(Left)* Impact of vision encoder and language model sizes. Smaller language models (135M) benefit less from larger vision encoders (SigLIP-SO-400M, 428M) compared to SigLIP-B/16 (93M), while larger language models gain more from powerful encoders. *(Middle-left)* Performance significantly improves with increased context lengths (2k to 16k tokens). *(Middle-right)* Optimal pixel shuffle factor (PS=2 vs. PS=4) varies by model size. *(Right)* Frame averaging reduces video performance, with a rapid decline as more frames are averaged. Metrics average CIDEr (captioning) and accuracy (visual question answering).

Architecture

- Context Length
  - RoPE
  - small model - 8k context length
  - large model - 16k context length

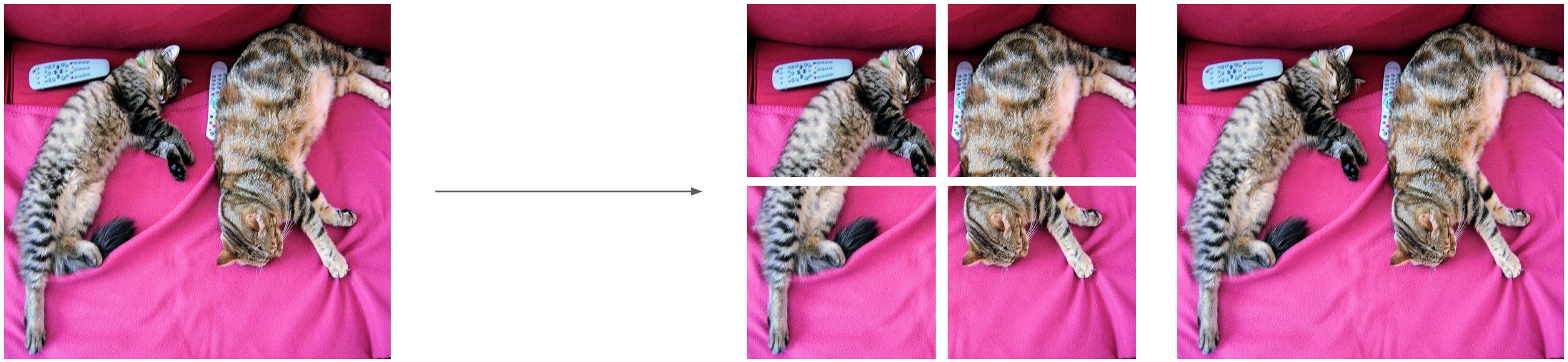
Architecture

- How can we efficiently pass the images to the Language Model?
  - Pixel Shuffle



Architecture

- How can we efficiently pass the images to the Language Model?
    - Sub image
        - resize to (512, 512), long edge
        - divide to 4 x 4
        - use whole resized image
        - one image to 17 images



Architecture

```python
# Create input messages
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image"},
            {"type": "image"},
            {"type": "text", "text": "Can you describe the two images?"}
        ]
    },
    {
        'role': 'Assistant',
        "content": [
            {'type': 'text', 'text': 'blah blah'},
        ],
    },
]
```

Architecture

<|im_start|>User:<fake_token_around_image><row_1_col_1><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_1_col_2><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_1_col_3><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_1_col_4><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image>

<fake_token_around_image><row_2_col_1><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_2_col_2><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_2_col_3><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_2_col_4><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image>

<fake_token_around_image><row_3_col_1><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_3_col_2><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_3_col_3><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_3_col_4><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image>

<fake_token_around_image><global-img><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><fake_token_around_image><row_1_col_1><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_1_col_2><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_1_col_3><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_1_col_4><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image>

<fake_token_around_image><row_2_col_1><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_2_col_2><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_2_col_3><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_2_col_4><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image>

<fake_token_around_image><row_3_col_1><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_3_col_2><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_3_col_3><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image><row_3_col_4><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image>

<fake_token_around_image><global-img><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><image><fake_token_around_image>Can you describe the two images?<end_of_utterance>

Assistant: blah blah<end_of_utterance>

Dataset

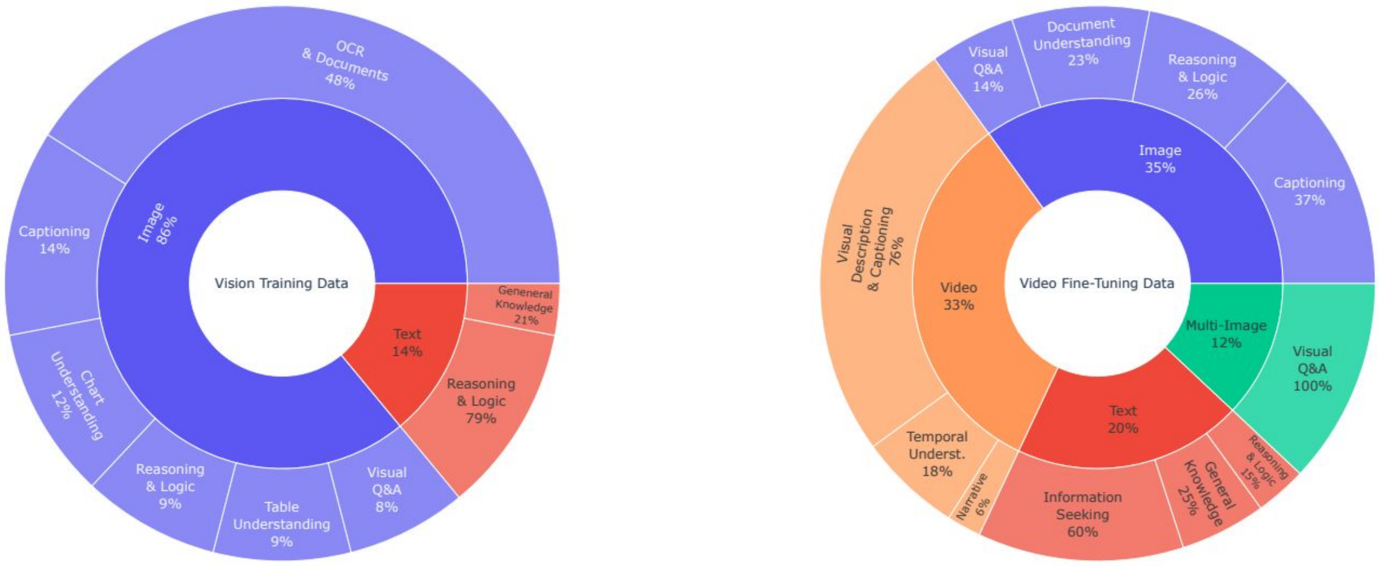


Figure 8 | Data Details. Training dataset details for Vision *(Left)* and video *(Right)*, broken down by modality and sub-categories.

Dataset

- Training Dataset
  - 2 stage : vision stage -> video stage
    - vision stage
      - new mixture of the datasets https://arxiv.org/abs/2408.12637
        to which added MathWriting https://arxiv.org/abs/2404.10690
      - The visual components comprise document understanding, captioning, and visual question answering (including 2% dedicated to multi-image reasoning), chart understanding, table understanding, and visual reasoning tasks.
      - To preserve the model's performance in text-based tasks, we retained a modest amount of general knowledge Q&A and text-based reasoning & logic problems, which incorporate mathematics and coding challenges.

Dataset

- Training Dataset
  - 2 stage : vision stage -> video stage
    - video stage
      - The video fine-tuning stage maintains 14% of text data and 33% of video to achieve optimal performance, following the learnings of Zohar et al. (2024b). For video, we sample visual description and captioning from LLaVA-video-178k (Zhang et al., 2024), Video-STAR (Zohar et al., 2024a), Vript (Yang et al., 2024), and ShareGPT4Video (Chen et al., 2023), temporal understanding from Vista-400k (Ren et al., 2024), and narrative comprehension from MovieChat (Song et al., 2024) and FineVideo (Farré et al., 2024). Multi-image data was sampled from M4-Instruct (Liu et al., 2024a) and Mammoth (Guo et al., 2024). The text samples were sourced from (Xu et al., 2024).

Evaluation

Capability	Benchmark	SmolVLM 256M	SmolVLM 500M	SmolVLM 2.2B	Efficient OS
Single-Image	OCRBench (Liu et al., 2024e) Character Recognition	52.6%	61.0%	72.9%	54.7% MolmoE-A1B-7B
	AI2D (Kembhavi et al., 2016) Science Diagrams	46.4%	59.2%	70.0%	71.0% MolmoE-A1B-7B
	ChartQA (Masry et al., 2022) Chart Understanding	55.6%	62.8%	68.7%	48.0% MolmoE-A1B-7B
	TextVQA (Singh et al., 2019) Text Understanding	50.2%	60.2%	73.0%	61.5% MolmoE-A1B-7B
	DocVQA (Mathew et al., 2021) Document Understanding	58.3%	70.5%	80.0%	77.7% MolmoE-A1B-7B
	ScienceQA (Lu et al., 2022) High-school Science	73.8%	80.0%	89.6%	87.5% MolmoE-A1B-7B
Multi-task	MMMU (Yue et al., 2024a) College-level Multidiscipline	29.0%	33.7%	42.0%	33.9% MolmoE-A1B-7B
	MathVista (Lu et al., 2024b) General Math Understanding	35.9%	40.1%	51.5%	37.6% MolmoE-A1B-7B
	MMStar (Chen et al., 2024a) Multidisciplinary Reasoning	34.6%	38.3%	46.0%	43.1% MolmoE-A1B-7B
Video	Video-MME (Fu et al., 2024) General Video Understanding	33.7%	42.2%	52.1%	45.0% InternVL2-2B
	MLVU (Zhou et al., 2024) MovieQA + MSRVTT-Cap	40.6%	47.3%	55.2%	48.2% InternVL2-2B
	MVBench (Li et al., 2024b) Multiview Reasoning	32.7%	39.7%	46.3%	60.2% InternVL2-2B
	WorldSense (Hong et al., 2025) Temporal + Physics	29.7%	30.6%	36.2%	32.4% Qwen2VL-7B
	TempCompass (Liu et al., 2024d) Temporal Understanding	43.1%	49.0%	53.7%	53.4% InternVL2-2B
Average	Across Benchmarks	44.0%	51.0%	59.8%	–
RAM Usage	Batch size = 1	0.8 GB	1.2 GB	4.9 GB	27.7 GB MolmoE-A1B-7B
	batch size = 64	15.0 GB	16.0 GB	49.9 GB	–

Table 1 | Benchmark comparison of SmolVLM variants across vision-language tasks. Performance of SmolVLM models at three scales (256M, 500M, and 2.2B parameters) compared to efficient open-source models on single-image, multi-task, and video benchmarks. SmolVLM models demonstrate strong accuracy while maintaining significantly lower RAM usage, highlighting their computational efficiency for resource-constrained multimodal scenarios.