SmolVLM: Redefining small and efficient multimodal models

Andrés Marafioti ★ Orr Zohar ★ Miquel Farré ★

Merve Noyan Elie Bakouch Pedro Cuenca Cyril Zakka Loubna Ben Allal Anton Lozhkov Nouamane Tazi Vaibhav Srivastav Joshua Lochner Hugo Larcher Mathieu Morlon Lewis Tunstall Leandro von Werra Thomas Wolf

Hugging Face, Stanford University ★ Equal Contribution

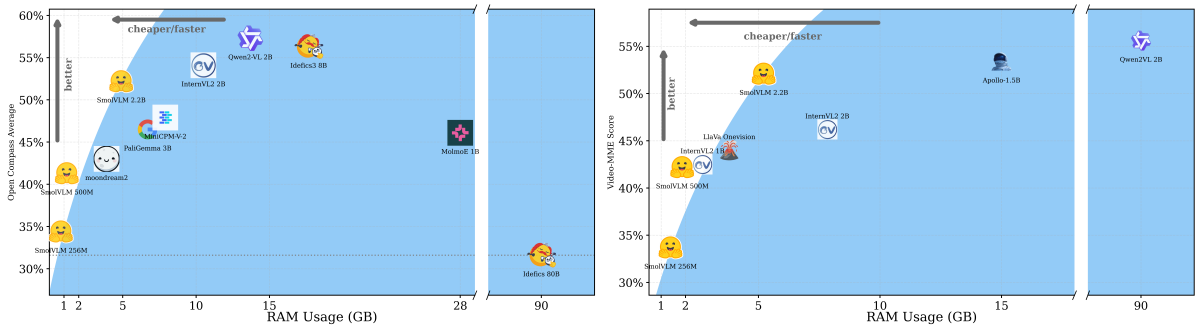

Figure 1 | Smol yet Mighty: comparison of SmolVLM with other state-of-the-art small VLM models. Image results are sourced from the OpenCompass OpenVLM leaderboard (Duan et al., 2024).

Abstract

Large Vision-Language Models (VLMs) deliver exceptional performance but require significant computational resources, limiting their deployment on mobile and edge devices. Smaller VLMs typically mirror design choices of larger models, such as extensive image tokenization, leading to inefficient GPU memory usage and constrained practicality for on-device applications.

We introduce **SmolVLM**, a series of compact multimodal models specifically engineered for resource-efficient inference. We systematically explore architectural configurations, tokenization strategies, and data curation optimized for low computational overhead. Through this, we identify key design choices that yield substantial performance gains on image and video tasks with minimal memory footprints.

Our smallest model, SmolVLM-256M, uses less than 1GB GPU memory during inference and outperforms the 300-times larger Idefics-80B model, despite an 18-month development gap. Our largest model, at 2.2B parameters, rivals state-of-the-art VLMs consuming twice the GPU memory. SmolVLM models extend beyond static images, demonstrating robust video comprehension capabilities.

Our results emphasize that strategic architectural optimizations, aggressive yet efficient tokenization, and carefully curated training data significantly enhance multimodal performance, facilitating practical, energy-efficient deployments at significantly smaller scales.

Code	gitub/huggingface	Blog	blog/smolvlm2
Weights	community/smol-research	VLM Browser	spaces/smolvlm-webgpu
Demo	spaces/smolvlm2	HuggingSnap	apple/huggingsnap

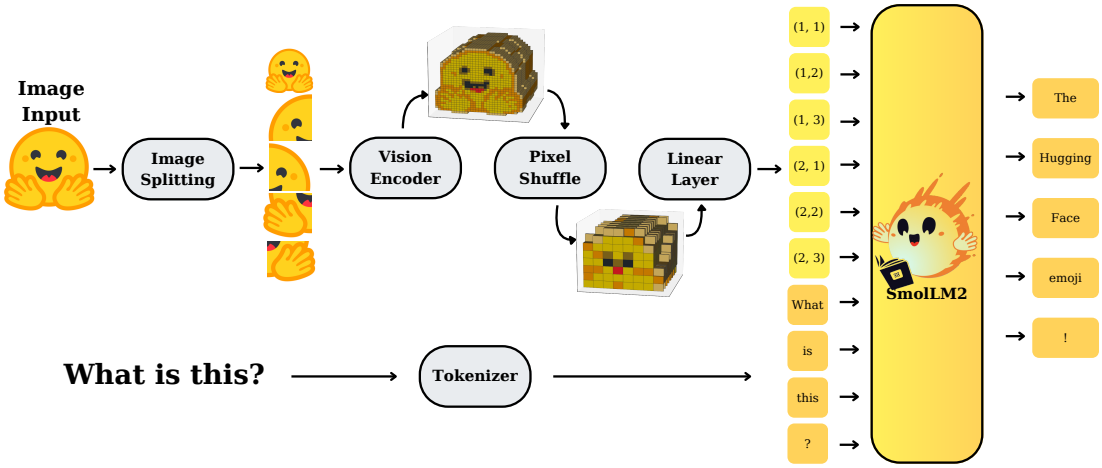

Figure 2 | SmolVLM Architecture. Images are split into subimages, frames are sampled from videos, and then encoded into visual features. These features are first rearranged via a pixel-shuffle operation, then mapped into the LLM input space as visual tokens using an MLP projection. Visual tokens are then concatenated/interleaved with text embeddings (orange/red). This combined sequence is passed to the LLM for text output.

1 Introduction

Vision-Language Models (VLMs) have rapidly advanced in capability and adoption (Achiam et al., 2023; Bai et al., 2023; Beyer et al., 2024; Chen et al., 2024c; McKinzie et al., 2024), driving breakthroughs in cross-modal reasoning (Liu et al., 2024a, 2023) and document understanding (Appalaraju et al., 2021; Faysse et al., 2024a; Livathinos et al., 2025; Nassar et al., 2025a). However, these improvements typically entail large parameter counts and high computational demands.

Since early large-scale VLMs like Flamingo (Alayrac et al., 2022a) and Idefics (Laurençon et al., 2023) demonstrated capabilities with 80B parameters, new models have slowly appeared at smaller sizes. However, these models often retain high memory demands due to architectural decisions made for their larger counterparts. For instance, Qwen2-VL (Wang et al., 2024a) and InternVL 2.5 (Chen et al., 2024b) offer smaller variants (1B-2B), but retain significant computational overhead. Conversely, models from Meta (Dubey et al., 2024) and Google (Gemma 3) reserve vision capabilities for large-scale models. Even PaliGemma (Beyer et al., 2024), initially efficiency-focused, scaled up significantly in its second release (Steiner et al., 2024). In contrast, Moondream (Korrapati, 2024) keeps focusing on improving performance while maintaining efficiency, and H2OVL-Mississippi (Galib et al., 2024) explicitly targets on-device deployment. Efficient processing is particularly critical for video understanding tasks, exemplified by Apollo (Zohar et al., 2024b), where memory management is essential. Furthermore, reasoning LLMs generate more tokens during inference, compounding computational costs (DeepSeek-AI, 2025; OpenAI et al., 2024). Therefore, efficiency per token becomes vital to ensure models remain practical for real-world use. *Our contributions are:*

- **Compact yet Powerful Models**: We introduce SmolVLM, a family of powerful small-scale multimodal models, demonstrating that careful architectural design can substantially reduce resource requirements without sacrificing capability.
- **Efficient GPU Memory Usage**: Our smallest model runs inference using less than 1GB GPU RAM, significantly lowering the barrier to on-device deployment.
- **Systematic Architectural Exploration**: We comprehensively investigate the impact of architectural choices, including encoder-LM parameter balance, tokenization methods, positional encoding, and training data composition, identifying critical factors that maximize performance in compact VLMs.
- **Robust Video Understanding on Edge Devices**: We demonstrate that SmolVLM models generalize effectively to video tasks, achieving competitive scores on challenging benchmarks like Video-MME, highlighting their suitability for diverse multimodal scenarios and real-time, on-device applications.
- **Fully Open-source Resources**: To promote reproducibility and facilitate further research, we release all model weights, datasets, code, and a mobile application showcasing inference on a smartphone.

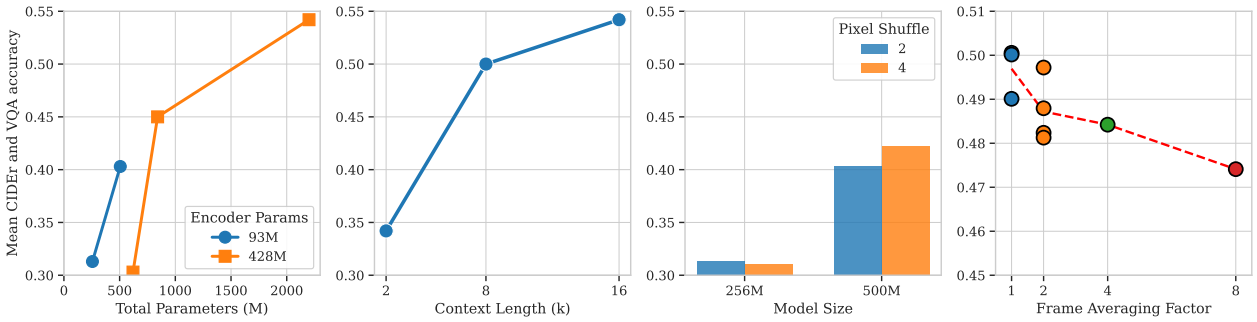

Figure 3 | Performance analysis of SmolVLM configurations. *(Left)* Impact of vision encoder and language model sizes. Smaller language models (135M) benefit less from larger vision encoders (SigLIP-SO-400M, 428M) compared to SigLIP-B/16 (93M), while larger language models gain more from powerful encoders. *(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).

2 Smoller Model Architecture

We systematically explore design choices for small multimodal models based on the architecture in Figure 2, where encoded images are pooled and projected into a SmolLM2 backbone. We first analyze optimal compute allocation, showing smaller vision encoders complement compact LMs (§2.1). Extending context length enables higher image resolutions at minimal overhead (§2.2), and pixel shuffling reduces visual tokens further. Finally, we efficiently handle high-resolution images and videos via document-specific image splitting and targeted token compression (§2.3). Together, these approaches yield a unified, performant, and cost-effective recipe for tiny LMMs.

2.1 How to assign compute between vision and language towers?

VLMs utilize vision encoders (see Figure 2) to generate 'vision tokens' that are then fed into an LM. We investigate optimal capacity allocation between vision encoders and language models (LMs) in compact VLMs. Specifically, we pair three SmolLM2 variants (135M, 360M, and 1.7B parameters) with two SigLIP encoders: a compact 93M SigLIP-B/16 and a larger 428M SigLIP-SO400M. Typically, larger VLMs disproportionately allocate parameters to the LM; however, as the LM is scaled down, this is no longer the case.

Figure 3 (left) confirms that performance declines significantly when using a large encoder with the smallest LM (135M), highlighting an inefficient encoder-LM balance. At an intermediate LM scale (360M), the larger encoder improves performance by 11.6%, yet this comes with a substantial 66% increase in parameters, making the compact encoder preferable. Only at the largest LM scale (1.7B), the larger encoder represents just a 10% parameter increase.

Finding 1. Compact multimodal models benefit from a balanced encoder-LM parameter allocation, making smaller vision encoders preferable for efficiency.

2.2 How can we efficiently pass the images to the Language Model?

Following Laurençon et al. (2024), we adopt a self-attention architecture in which visual tokens from the vision encoder are concatenated with textual tokens and jointly processed by a language model (e.g., FROMAGe (Koh et al., 2023), BLIP-2 (Li et al., 2023a)). This design requires significantly more context than the 2k-token limit used in SmolLM2, as a single 512×512 image encoded with SigLIP-B/16 requires 1024 tokens. To address this, we extended the context capacity by increasing the RoPE base from 10k to 273k, following Liu et al. (2024c), and fine-tuned the model on a mix of long-context data (Dolma books (Soldaini et al., 2024), The Stack (Kocetkov et al., 2022)) and short-context sources (FineWeb-Edu (Penedo et al., 2024), DCLM (Li et al., 2024a), and math from SmolLM2).

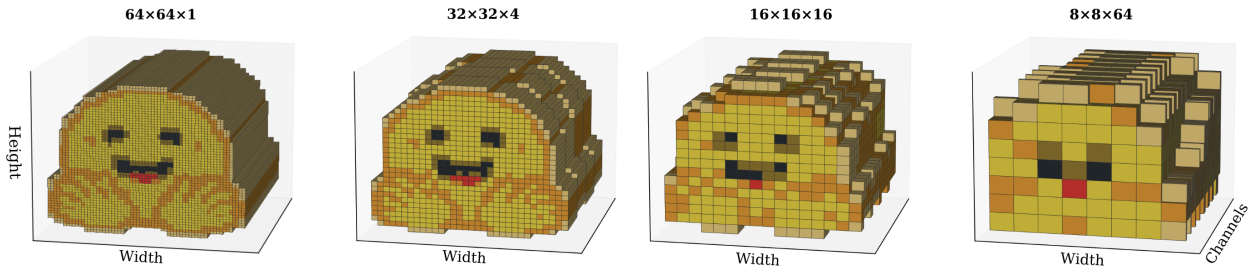

Figure 4 | Pixel shuffle. Rearranges encoded images, trading spatial resolution for increased channel depth. This reduces visual token count while preserving information density.

While fine-tuning was stable at 16k tokens for the 1.7B LM, smaller models (135M, 360M) struggled beyond 8k. Experiments with our 2.2B SmolVLM confirmed consistent performance gains up to 16k tokens (Figure 3, middle). Accordingly, we adopt a 16k-token context for SmolVLM and an 8k-token limit for smaller variants.

Finding 2. Compact VLMs significantly benefit from extended context lengths.

Extending the context window alone is not sufficient. Recent VLMs (e.g., MM1 (McKinzie et al., 2024), MiniCPM-V (Yao et al., 2024), InternVL (Chen et al., 2024c)) combine the self-attention architecture with token compression techniques (Zohar et al., 2024b; Laurençon et al., 2024) to fit longer sequences efficiently and reduce computational overhead.

One particularly effective compression method is *pixel shuffle* (space-to-depth), initially proposed for super-resolution tasks (Shi et al., 2016) and recently adopted by Idefics3. Pixel shuffle rearranges spatial features into additional channels, reducing spatial resolution but increasing representational density (Figure 4). This reduces the total number of visual tokens by a factor of r^2, where r is the shuffle ratio. However, higher ratios collapse larger spatial regions into single tokens, impairing tasks requiring precise localization, such as OCR. Models like InternVL and Idefics3 use $r = 2$ to balance compression and spatial fidelity. In contrast, our experiments (Figure 3, right) show that smaller VLMs benefit from more aggressive compression ($r = 4$) as the reduced token count eases attention overhead and improves long-context modeling.

Finding 3. Small VLMs benefit from more aggressive visual token compression.

2.3 How can we efficiently encode images and videos?

Balancing token allocation between images and videos is crucial for efficient multimodal modeling: images benefit from higher resolution and more tokens to retain fidelity, whereas videos typically require fewer tokens per frame to handle longer sequences efficiently.

To achieve this, we successfully adopted an image-splitting strategy inspired by UReader (Ye et al., 2023) and SPHINX (Lin et al., 2023b), where high-resolution images are divided into multiple sub-images along with a downsized version of the original. This approach proved effective in maintaining image quality without excessive computational overhead. For videos, however, we found that strategies such as frame averaging, inspired by Liu et al. (2024f), negatively impacted performance. As shown in Figure 3 (right), combining multiple frames significantly degraded OpenCompass-Video results, particularly at higher averaging factors (2, 4, 8). Consequently, frame averaging was excluded from SmolVLM's final design, and video frames were instead rescaled to the resolution of the image encoder.

Finding 4. For small models, image splitting enhances performance for vision tasks, whereas video frame averaging does not.

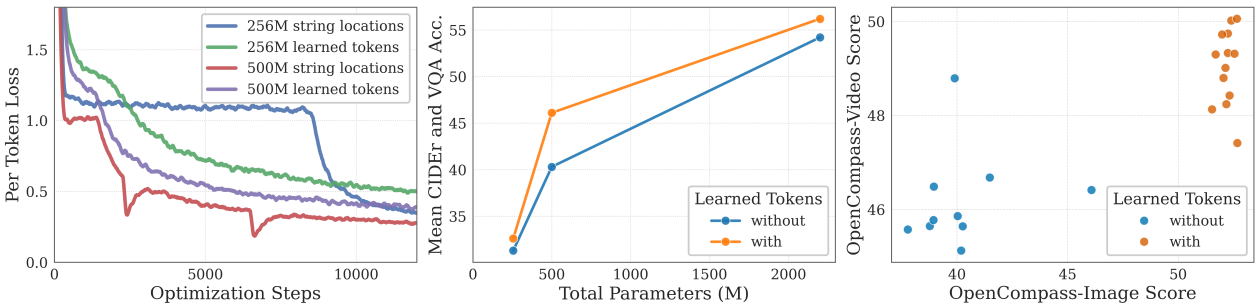

Figure 5 | Tokenization Strategy Comparisons. *(Left)* Training loss curves illustrating the "OCR loss plague" when using string-based tokens in smaller models. *(Center)* Aggregated evaluation metrics showing consistently higher scores with learned tokens (orange). *(Right)* Scatter plot of OpenCompass-Image vs. OpenCompass-Video: learned tokens dominate the higher-scoring region, especially in image-intensive tasks.

3 Smol Instruction Tuning

Smol instruction tuning requires careful vision (§3.1) and text tokenization (§3.2), alongside unified methods for multimodal modeling under tight compute constraints. Learned positional tokens and structured prompts stabilize training and improve OCR, but data composition remains crucial: reusing LLM instruction datasets negatively impacts small VLMs (§3.3), excessive Chain-of-Thought data overwhelms limited capacity (§3.4), and moderate video sequence lengths balance efficiency and performance (§3.5). Collectively, these insights highlight targeted strategies essential for effectively scaling multimodal instruction tuning to SmolVLMs.

3.1 Learned Tokens vs. String

A primary design consideration in SmolVLM involves encoding split sub-image positions effectively. Initially, we attempted to use simple string tokens (e.g., `<row_1_col_2>`), which caused early training plateaus—termed the "OCR loss plague"—characterized by sudden loss drops without corresponding improvements in OCR performance (Figure 5, left and middle).

To address instability during training, we introduced positional tokens, significantly improving training convergence and reducing stalls. Although larger models were relatively robust to using raw string positions, smaller models benefited substantially from positional tokens, achieving notably higher OCR accuracy and improved generalization across tasks. Figure 5 (center) shows that learned positional tokens consistently outperform naive string positions on multiple image and text benchmarks. Additionally, Figure 5 (right) illustrates that models leveraging learned tokens consistently score higher in both OpenCompass-Image and OpenCompass-Video evaluations, underscoring the effectiveness of structured positional tokenization in compact multimodal models.

Finding 5. Learned positional tokens outperform raw text tokens for compact VLMs.

3.2 Structured Text Prompts and Media Segmentation

We evaluated how system prompts and explicit media intro/outro prefixes incrementally improve SmolVLM's performance on image (left) and video (right) benchmarks, as shown in Figure 6. Each violin plot represents three checkpoints for a given configuration.

System Prompts. We prepend concise instructions to clarify task objectives and reduce ambiguity during zero-shot inference. For example, conversational datasets utilize prompts like *"You are a useful conversational assistant,"* whereas vision-focused tasks employ *"You are a visual agent and should provide concise answers."* The second violin plot in each subplot (Fig. 6) illustrates clear performance improvements from incorporating these system prompts, particularly evident in image-centric tasks.

Media Intro/Outro Tokens. To clearly demarcate visual content, we introduce textual markers around image

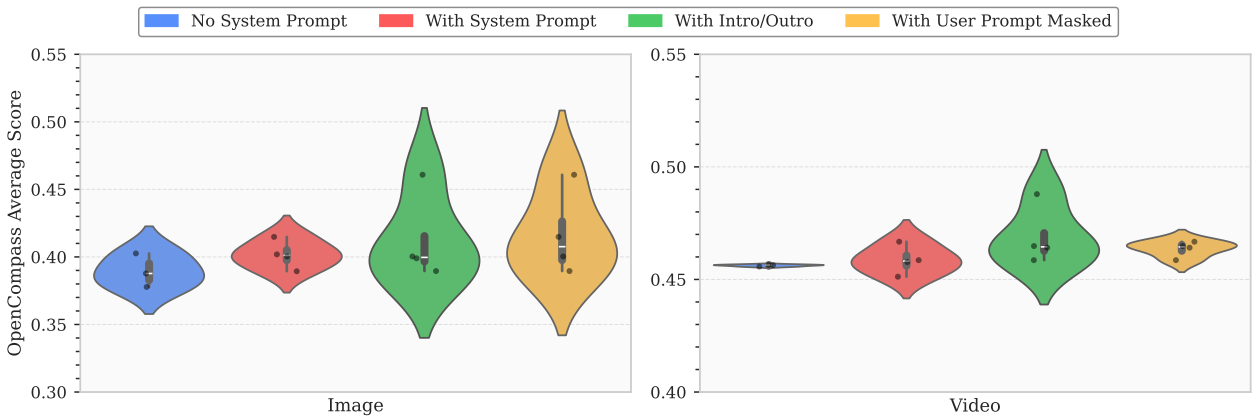

Figure 6 | Cumulative Effect of Training Strategies on SmolVLM Performance. The visualization shows the progression of performance improvements as different tokenization and prompt engineering strategies are applied sequentially to the SmolVLM base model. *(Left)* Image benchmark results show consistent improvements with each added strategy. *(Right)* Video benchmark results reveal similar patterns with more pronounced gains.

and video segments (e.g., "*Here is an image...*" and "*Here are N frames sampled from a video...*"). The outro tokens then transition back to textual instructions (e.g., "*Given this image/video...*"). The third violin indicates that this strategy substantially boosts performance on video tasks—where confusion between multiple frames is more likely—and still yields measurable improvements on image tasks.

Masking User Prompts Drawing on techniques from Allal et al. (2025), we explore user-prompt masking during supervised fine-tuning as a way to reduce overfitting. The right violin plot in Figure 6 shows that masking user queries (orange) yields improved performance in both image and video tasks, compared to the unmasked baseline (blue). This effect is significantly pronounced in multimodal QA, where questions are often repetitive and can be trivially memorized by the model. Masking thus forces SmolVLM to rely on task-related content rather than superficial repetition, promoting better generalization.

Finding 6. System prompts and media intro/outro tokens significantly improve compact VLM performance, particularly for video tasks. During SFT, only train on completions.

3.3 Impact of Text Data Reuse from LLM-SFT

A seemingly intuitive practice is to reuse text data from the final supervised fine-tuning stages of large language models, anticipating in-distribution prompts and higher-quality linguistic inputs. However, Figure 7 (left) shows that incorporating LLM-SFT text data (*SmolTalk*) can degrade performance in smaller multimodal architectures by as much as 3.7% in video tasks and 6.5% in image tasks. We attribute this negative transfer to reduced data diversity, which outweighs any benefits of reusing text. In keeping with Zohar et al. (2024b), we therefore maintain a strict 14% text proportion in our training mix. These findings highlight the importance of a carefully balanced data pipeline, rather than direct adoption of large-scale SFT text for small-scale multimodal models.

Finding 7. Adding text from SFT blend proved worse than new text SFT data.

3.4 Optimizing Chain-of-Thought Integration for Compact Models

Chain-of-Thought (CoT) prompting, which exposes models to explicit reasoning steps during training, generally enhances reasoning capabilities in large models. However, its effect on smaller multimodal architectures remains unclear. To investigate this, we varied the proportion of CoT data integrated into the Mammoth dataset (Yue et al., 2024b), covering text, image, and video tasks. Figure 7 (middle) shows that incorporating a minimal fraction (0.02–0.05%) of CoT examples slightly improved performance, but higher proportions

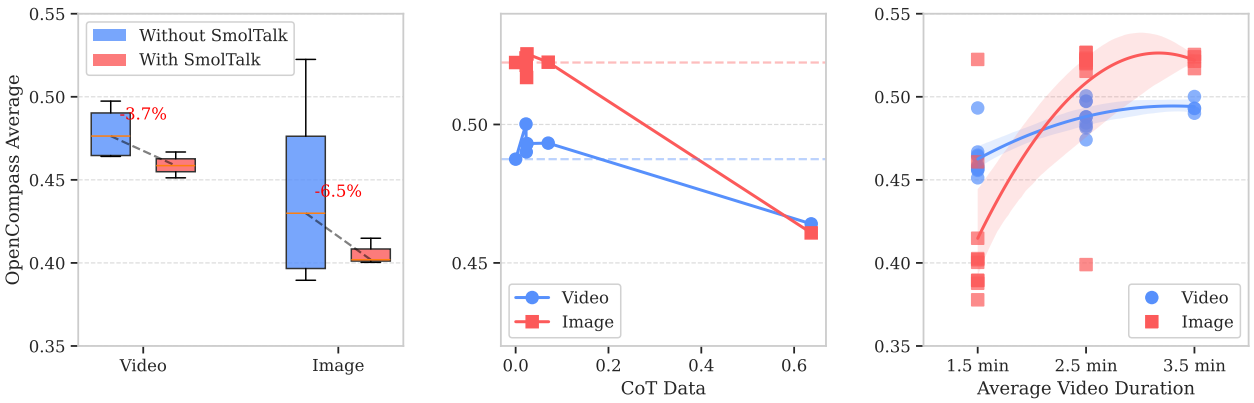

Figure 7 | Impact of Training Strategies on Smol-Scale Multimodal Models. *(Left)* Reusing text data from LLM-SFT (*SmolTalk*) reduces both image and video scores in smaller models. *(Middle)* A minimal fraction (0.02%–0.05%) of Chain-of-Thought (CoT) data yields optimal results, while heavier CoT usage degrades performance. *(Right)* Increasing average video duration beyond 3.5 min leads to diminished returns for both image and video tasks.

markedly degraded results, especially in image tasks. These observations suggest that excessive reasoning-oriented textual data can overwhelm the limited capacity of smaller VLMs, thereby compromising their visual representation capabilities. Consequently, compact models benefit most from very sparse inclusion of CoT data rather than the extensive use typically beneficial in larger-scale architectures.

Finding 8. Excessive CoT data harms compact model performance.

3.5 Impact of Video Sequence Length on Model Performance

Increasing video duration during training offers richer temporal context but comes at a greater computational cost. To identify an optimal duration, we trained SmolVLM on average video lengths ranging from 1.5 to 3.5 minutes. Figure 7 (right) demonstrates clear performance improvements for both video and image benchmarks as video durations approached approximately 3.5 minutes, likely due to more effective cross-modal feature learning. Extending video duration beyond 3.5 minutes yielded minimal further gains, indicating diminishing returns relative to the added computational expense. Thus, moderately extending video sequences enhances performance significantly in smaller models, whereas overly long sequences do not proportionally justify their computational cost.

Finding 9. Moderately increasing video duration during training improves both video and image task performance in compact VLMs.

4 Experimental Results

We construct three variants of SmolVLM, tailored to different computational environments:

- **SmolVLM-256M**: Our smallest model, combining the 93M SigLIP-B/16 and the SmolLM2-135M (Allal et al., 2025). Operating on < 1GB GRAM makes it ideal for resource-constrained edge applications.
- **SmolVLM-500M**: A mid-range model with the same 93M SigLIP-B/16 paired with the larger SmolLM2-360M. Balancing memory efficiency and performance, it is suitable for moderate-resource edge devices.
- **SmolVLM-2.2B**: The largest variant, with a 400M SigLIP-SO400M and a 1.7B-parameter SmolLM2 backbone. This model maximizes performance while remaining deployable on higher-end edge systems.

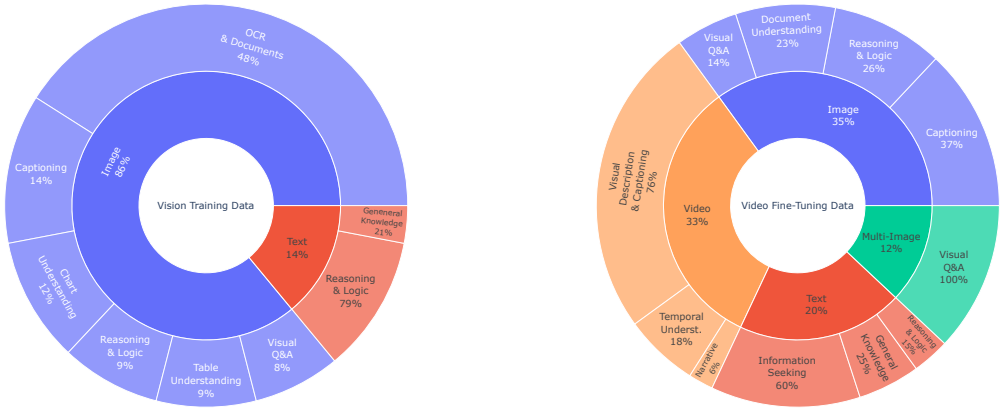

Figure 8 | Data Details. Training dataset details for Vision *(Left)* and video *(Right)*, broken down by modality and sub-categories.

4.1 Training Data

Model training proceeds in two stages: (1) a vision stage, and (2) a video stage. The vision training stage uses a new mixture of the datasets used in Laurençon et al. (2024), to which we added MathWriting (Gervais et al., 2024). The mixture was balanced to emphasize visual and structured data interpretation while maintaining the focus on reasoning and problem-solving capabilities. 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.

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).

For a more granular description, Figure 8 provides a detailed overview of the training data distribution used in both our vision and video fine-tuning stages.

4.2 Evaluation details

We evaluated SmolVLM using VLMEvalKit (Duan et al., 2024) to ensure reproducibility. The full results are available online[1]. Currently, the OpenVLM Leaderboard covers 239 different VLMs and 31 different multi-modal benchmarks. Further, we plot the performance against the RAM required to run the evaluations. We argue that model size is usually used as a proxy for the computational cost required to run a model. This is misleading for VLMs because the architecture strongly influences how expensive it is to run the model; in our opinion, RAM usage is a better proxy. For SmolVLM, this resizes the longest edge of images to 1920 in the 256M and 500M models and 1536 in the 2.2B.

4.3 Strong Performance at a Tiny Scale

We evaluate SmolVLM's performance relative to model size, comparing three variants (256M, 500M, and 2.2B) against efficient state-of-the-art open-source models. Table 1 summarizes results across nine demanding vision-language benchmarks and five video benchmarks. We highlight in the table MolmoE 7B with 1B

[1] OpenVLM Leaderboard

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.

activated parameters (Deitke et al., 2024)(MolmoE-A1B-7B) for vision tasks and InternVL2-2B Chen et al. (2024c) for video tasks. A broader array of competing models are shown in Fig. 1.

Efficiency and Memory Footprint. SmolVLM demonstrates remarkable computational efficiency compared to significantly larger models. Single-image inference requires only 0.8GB of VRAM for the 256M variant, 1.2GB for the 500M, and 4.9GB for the 2.2B—dramatically lower than the 27.7GB required by MolmoE-A1B-7B. Even compared to models of similar parameter scales, SmolVLM is notably more efficient: Qwen2VL-2B requires 13.7GB VRAM and InternVL2-2B requires 10.5GB VRAM, highlighting that parameter count alone does not dictate compute requirements. At batch size 64, memory usage for SmolVLM remains practical: 15.0GB (256M), 16.0GB (500M), and 49.9GB (2.2B). These results highlight SmolVLM's substantial advantages for deployment in GPU-constrained environments.

Overall Gains from Scaling. Increasing SmolVLM's parameter count consistently yields substantial performance improvements across all evaluated benchmarks. The largest model (2.2B) achieves the highest overall score at 59.8%, followed by the intermediate 500M variant (51.0%) and the smallest 256M variant (44.0%). Notably, even the smallest SmolVLM-256M significantly surpasses the much larger Idefics 80B model (see Fig. 1) on nearly all benchmarks, emphasizing effective vision capabilities at modest scales. The few exceptions—particularly MMMU (29.0% vs. 42.3%) and AI2D (46.4% vs. 56.3%)—highlight benchmarks where strong linguistic reasoning from a large language backbone remains crucial. Intriguingly, visually oriented tasks such as OCRBench also benefit markedly from scaling language model capacity, with a nearly 10-point improvement when moving from 256M (52.6%) to 500M (61.0%). These results underscore that

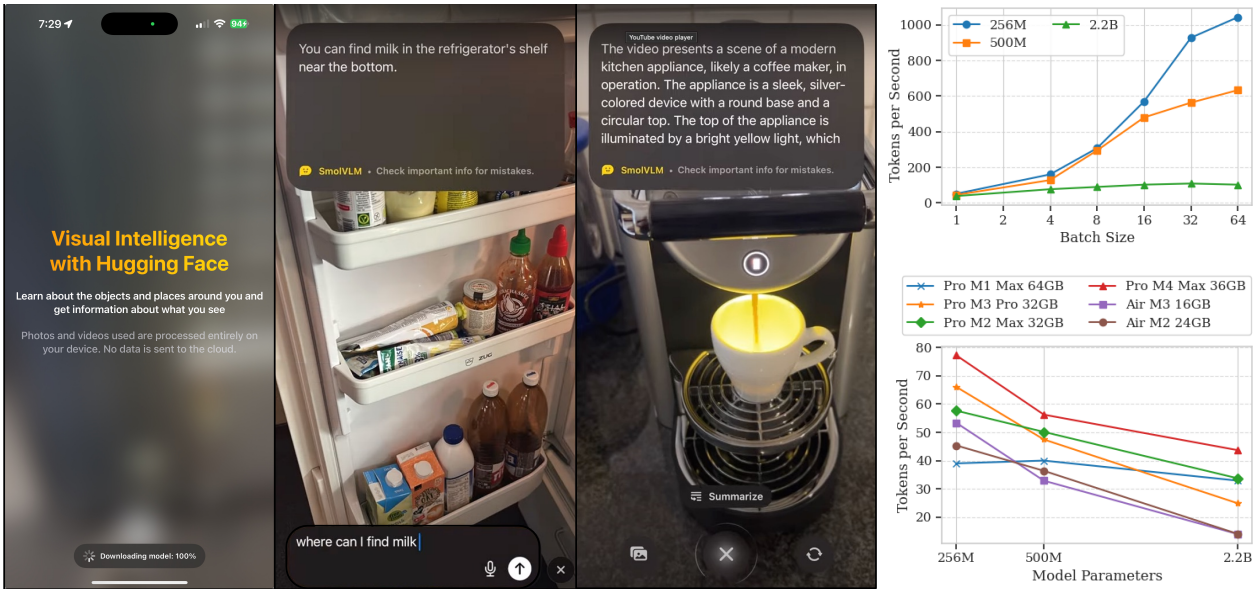

Figure 9 | SmolVLM on edge device. *(Left)* Examples of the `HuggingSnap` app, where SmolVLM can run locally, on the device, on consumer phones. For example, interactions can be done using a mobile interface to detect objects and answer questions. *(Right)* Throughput in tokens per second on NVIDIA A100 GPUs *(top)* and different consumer personal computers *(bottom)* across different batch sizes and model variants.

larger language models provide enhanced context management and improved multimodal reasoning, benefiting both language-intensive and vision-centric tasks.

Comparison with Other Compact VLMs. Figure 1 situates SmolVLM-2.2B among recent small-scale VLMs by comparing OpenCompass benchmark performance against GPU memory consumption per image. SmolVLM-2.2B achieves notably strong performance on MathVista (51.5) and ScienceQA (90.0), while maintaining exceptionally low GPU usage of just 4.9GB VRAM. In contrast, models requiring significantly more compute, such as Qwen2VL-2B and InternVL2-2B, aren't clearly better performers. Specifically, Qwen2VL-2B slightly surpasses SmolVLM-2.2B on AI2D (74.7 vs. 70.0) and ChartQA (73.5 vs. 68.8), yet falls short on MathVista (48.0 vs. 51.5) and ScienceQA (78.7 vs. 90.0). Similarly, InternVL2-2B achieves higher scores on ScienceQA (94.1 vs. 90.0) and MMStar (49.8 vs. 46.0), but at more than double the VRAM cost.

Further comparisons highlight distinct trade-offs among size, memory footprint, and task-specific performance. MiniCPM-V2 (2.8B parameters) underperforms SmolVLM-2.2B on most benchmarks. Other models such as Moondream2 and PaliGemma (both around 2˘3B parameters) exhibit significant variance across tasks: Moondream2, for instance, scores well on ChartQA (72.2) with just 3.9GB VRAM but substantially underperforms on MMMU (29.3). Conversely, PaliGemma excels at ScienceQA (94.3) yet struggles on ChartQA (33.7). This variability underscores how specialized training impacts per-task.

Video Benchmarks. Table 1 provides comprehensive results across five diverse video benchmarks: Video-MME, MLVU, MVBench, TempCompass, and WorldSense. SmolVLM-2.2B notably excels at Video-MME (52.1) and WorldSense (36.2), outperforming significantly larger models such as Qwen2 VL-7B (32.4 on WorldSense), showcasing strong capabilities in complex multimodal video comprehension tasks. The SmolVLM-500M variant also demonstrates robust performance, achieving competitive scores on TempCompass (49.0) and WorldSense (30.6), highlighting sophisticated temporal reasoning and real-world visual understanding at a scale ideal for edge-device deployment. Despite their compact parameter counts, SmolVLM variants consistently balance efficient resource use with impressive accuracy, reinforcing their suitability for resource-constrained scenarios.

4.4 On-Device Performance

To comprehensively assess the deployment practicality of SmolVLM, we benchmarked its throughput across varying batch sizes on two representative hardware platforms: NVIDIA A100 and NVIDIA L4 GPUs (see Figure 9). Our evaluations highlight SmolVLM's suitability for on-device and edge deployment scenarios.

On the A100 GPU, the smallest SmolVLM-256M variant achieves impressive throughput, scaling from 0.8 examples per second at batch size 1 to 16.3 examples per second at batch size 64. The 500M variant similarly scales from 0.7 to 9.9 examples per second, while the largest 2.2B variant demonstrates more modest scaling (0.6 to 1.7 examples per second), indicative of its higher computational demands.

Evaluations on the L4 GPU further emphasize SmolVLM's edge compatibility. Here, the 256M variant reaches peak throughput at 2.7 examples per second with batch size 8, subsequently diminishing due to memory constraints. The 500M and 2.2B variants peak at lower batch sizes (1.4 and 0.25 examples per second, respectively), underscoring their efficiency even under more restrictive hardware conditions.

Finally, we accompany the release with several optimized ONNX (Open Neural Network Exchange) exports, facilitating cross-platform compatibility and broadening deployment opportunities across consumer-grade hardware targets. Notably, we demonstrate the ability to efficiently run these models locally within a browser environment via WebGPU, with the 256M variant achieving up to 80 decode tokens per second on a 14-inch MacBook Pro (M4 Max).

4.5 Downstream Applications

Beyond our own evaluations, SmolVLM has seen adoption in various downstream applications developed by the broader research community, emphasizing its efficiency in real-world, resource-constrained scenarios.

ColSmolVLM: On-Device Multimodal Inference. ColSmolVLM utilizes the smaller SmolVLM variants (256M and 500M parameters) designed explicitly for on-device deployment, as detailed in recent work by Hugging Face (Faysse et al., 2024b). These compact models enable efficient multimodal inference directly on mobile devices, consumer laptops, and even within browser-based environments, significantly lowering computational demands and operational costs.

Smol Docling: Ultra-Compact Document Processing. Smol Docling is an ultra-compact 256M-parameter variant of SmolVLM, optimized explicitly for end-to-end multimodal document conversion tasks (Nassar et al., 2025b). By employing specialized representations known as DocTags, Smol Docling efficiently captures content, context, and spatial relationships across diverse document types, including business documents, academic papers, and patents. Its compact architecture maintains competitive performance with considerably larger VLMs, highlighting its suitability for deployment in scenarios with computational constraints.

BioVQA: Biomedical Visual Question Answering. BioVQA leverages SmolVLM's compact and efficient architecture to address visual question answering tasks within the biomedical domain (Lozano et al., 2025). Small-scale SmolVLM models have demonstrated promising capabilities in interpreting medical images, assisting healthcare professionals by providing accurate answers to clinical questions based on visual data. This capability is particularly valuable in healthcare settings where quick, reliable image interpretation is critical, yet computational resources may be limited.

5 Related Work

5.1 First-Generation Vision-Language Models

Early multimodal models achieved significant progress primarily by scaling parameters, but their high computational demands limited practical deployment. For instance, Flamingo (Alayrac et al., 2022b), an 80B-parameter Vision-Language Model (VLM), integrated a frozen 70B-parameter LM (Hoffmann et al., 2022) with a vision encoder employing gated cross-attention and a Perceiver Resampler (Jaegle et al., 2021) for

efficient token compression. Despite state-of-the-art few-shot capabilities without task-specific fine-tuning, Flamingo's large scale posed significant deployment challenges.

Hugging Face's Idefics (Laurençon et al., 2023) adopted Flamingo's architecture, offering models at both 9B and 80B parameters, further exemplifying the approach of large-scale multimodal training. In contrast, BLIP-2 (Li et al., 2023a) proposed a more parameter-efficient, modular design by freezing both the vision encoder and language model, introducing instead a lightweight Query Transformer (Q-Former) that translates visual features into language-compatible tokens. This approach significantly reduced trainable parameters, surpassing Flamingo's performance on VQA tasks (Antol et al., 2015; Goyal et al., 2017) with roughly 54 times fewer trainable parameters, thus paving the way toward more efficient multimodal architectures.

Similarly, LLaVA (Large Language-and-Vision Assistant) (Liu et al., 2023) connected a pretrained CLIP (Radford et al., 2021) ViT image encoder to a LLaMA/Vicuna language backbone (Touvron et al., 2023; Zheng et al., 2024), fine-tuning the combined model on instruction-following datasets. Resulting in a 13B-parameter multimodal chatbot with GPT-4V-like capabilities (Achiam et al., 2023), LLaVA achieved notable visual conversational performance. However, despite being smaller and faster than Flamingo, it still demands substantial GPU memory for real-time interaction and inherits the limitations of the underlying language model's context window (typically 2048 tokens).

Recent research has actively explored various design choices, training strategies, and data configurations to enhance Vision-Language Models (VLMs). For instance, Idefics2 (Laurençon et al., 2024) introduced architectural and training-data improvements compared to its predecessor, advancing open-source VLM capabilities. Concurrently, Cambrian1 (Tong et al., 2024) examined fundamental design principles and scaling behaviors, aiming for more efficient architectures. Projects like Eagle (Shi et al., 2024) and its successor Eagle2 (Li et al., 2025b) have optimized specific architectural components, targeting improved performance and efficiency. Additionally, recent efforts such as Apollo (Zohar et al., 2024b) extend multimodal architectures from static images to video understanding, further enriching the diversity of approaches.

5.2 Efficiency-Focused Vision-Language Models

Larger models, such as InternVL (Chen et al., 2024c,b) and Qwen-VL (Bai et al., 2023, 2025; Wang et al., 2024a), introduced architectural innovations for improved computational efficiency. InternVL aligns a 6B-parameter vision transformer (ViT) with an 8B-parameter language "middleware," forming a 14B-parameter model that achieves state-of-the-art results across multiple vision and multimodal tasks. This balanced architecture narrows the modality gap, enabling robust multimodal perception and generation capabilities. Similarly, Qwen-VL integrates a Qwen language model with specialized visual modules, leveraging captioned bounding-box data to enhance visual grounding and text recognition capabilities. Despite its strong multilingual and multimodal performance, Qwen-VL generates exceptionally long token sequences for high-resolution inputs, increasing memory requirements.

On the smaller end, models like PaliGemma, Moondream2, and MiniCPM-V demonstrate impressive multimodal capabilities within constrained parameter budgets. PaliGemma (Team et al., 2024), with just 3B parameters (400M vision encoder from SigLIP-So (Zhai et al., 2023) and 2B Gemma language model), effectively covers a wide range of multimodal tasks. However, its condensed visual interface can limit detailed visual analysis. Moondream2, at merely 1.8B parameters, pairs SigLIP visual features with Microsoft's Phi-1.5 language model (Li et al., 2023b), showcasing competitive performance on tasks such as image description, OCR, counting, and classification, ideal for edge and mobile applications. MiniCPM-V (Hu et al., 2024), specifically designed for on-device scenarios, integrates a 400M vision encoder and a 7.5B language model via a perceiver-style adapter. This compact model notably achieves GPT-4V-level performance on selected benchmarks. Deepseek VL and Deepseek VL2 (Lu et al., 2024a; Wu et al., 2024), spanning 2–7B and 4–27B parameters respectively, further illustrate the growing focus on efficient yet powerful multimodal models suitable for resource-constrained environments. Collectively, these models demonstrate the increasing feasibility of deploying effective, real-time multimodal AI in practical scenarios.

5.3 Multimodal Tokenization and Compression Strategies

Efficient tokenization significantly reduces computational and memory demands in Vision-Language Models (VLMs). Early methods, encoding every pixel or patch individually, resulted in lengthy sequences—196 tokens for a 224×224 image at 16×16 resolution. Recent strategies compress visual data while preserving essential details. Learned modules like Perceiver Resamplers (Jaegle et al., 2021) used by Flamingo and Idefics2 (Alayrac et al., 2022b; Laurençon et al., 2024), and BLIP-2's Q-Former (Li et al., 2023a), compress inputs into a small set of latent tokens. While effective in shortening sequences, these methods may limit performance on fine-grained tasks like OCR (Singh et al., 2019; Biten et al., 2019). Spatial compression via patch pooling and pixel shuffle is increasingly popular. InternVL v1.5 and Idefics3 (Chen et al., 2024c,b; Laurençon et al., 2023) use 2×2 pixel-shuffle, reducing token counts fourfold while maintaining OCR capability. Models like Qwen-VL-2 (Wang et al., 2024a) adopt multi-scale representations and selective token dropping via convolutional and Transformer modules. Adaptive methods, such as image tiling in UReader and DocOwl, dynamically adjust token counts based on task complexity, sacrificing some global context.

5.4 Video-Capable Vision-Language Models

Extending vision-language models (VLMs) from images to videos significantly increases complexity due to temporal dimensions, expanding token counts and computational demands. Early models, such as Video-LLaVA (Lin et al., 2023a), unified image and video training, aligning video frame features with static images and substantially outperforming predecessors like Video-ChatGPT (Maaz et al., 2023) on benchmarks including MSRVTT (Xu et al., 2016), MSVD (Chen and Dolan, 2011), TGIF (Li et al., 2016), and ActivityNet (Caba Heilbron et al., 2015). Meanwhile, Video-STaR (Zohar et al., 2024a) introduced the first self-training approach that leverages existing labeled video datasets for instruction tuning of Large Multimodal Models.

Recent models enhance efficiency and effectiveness in handling long-form video content. Temporal Preference Optimization (TPO) (Li et al., 2025a) employs self-training with localized and comprehensive temporal grounding, improving benchmarks like LongVideoBench, MLVU, and Video-MME. Oryx MLLM (Liu et al., 2024g) dynamically compresses visual tokens via its OryxViT encoder, balancing efficiency and precision across tasks. VideoAgent (Wang et al., 2024b) models long-form video understanding as a decision-making process, utilizing a large language model (LLM) as an agent to identify and compile crucial information for question answering iteratively. VideoLLaMA3 (Zhang et al., 2025) adapts its vision encoder for variable resolutions and uses multi-task fine-tuning to enhance video comprehension. Video-XL (Shu et al., 2024) introduces Visual Summarization Tokens (VST) and curriculum learning for efficient handling of hour-scale videos. Similarly, Kangaroo (Liu et al., 2024b) utilizes curriculum training to scale input resolution and frame count progressively, achieving top performance on diverse benchmarks.

Apollo (Zohar et al., 2024b) recently made an in-depth exploration of Video-LMMs and showed the architecture and training schedule that most affect performance. In so doing, it showed the remarkable efficiency gains that can be made during training and inference. Apollo achieved state-of-the-art results with modest parameter sizes on benchmarks such as LongVideoBench, MLVU, and Video-MME (Zhou et al., 2024; Fu et al., 2024).

6 Conclusion

We introduced **SmolVLM**, a family of memory-efficient Vision-Language Models ranging from 256M to 2.2B parameters. Remarkably, even our smallest variant requires less than 1GB of GPU memory yet surpasses state-of-the-art 80B-parameter models from just 18 months ago (Laurençon et al., 2023). Our findings emphasize a critical insight: scaling down large VLM architectures optimized under resource-rich conditions results in disproportionately high memory demands during inference with little advantage over specialized architectures. By contrast, SmolVLM's design philosophy explicitly prioritizes compact architectural innovations, aggressive but careful tokenization methods, and efficient training strategies, enabling powerful multimodal capabilities at a fraction of the computational cost.

All model weights, training datasets, and training code are publicly released to encourage reproducibility, transparency, and continued innovation. We hope SmolVLM will inspire the next generation of lightweight, efficient VLMs, unlocking new possibilities for real-time multimodal inference with minimal power consumption.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc., 2022a. https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022b.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. https://arxiv.org/abs/2502.02737.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 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, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. https://arxiv.org/abs/2502.13923.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluis Gomez, Marçal Rusiñol, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4290–4300, 2019. doi: 10.1109/ICCV.2019.00439.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. https://arxiv.org/abs/2501.12948.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. https://arxiv.org/abs/2409.17146.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. *arXiv preprint arXiv:2407.11691*, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. https://huggingface.co/datasets/HuggingFaceFV/finevideo, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024a.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models, 2024b. https://arxiv.org/abs/2407.01449.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Shaikat Galib, Shanshan Wang, Guanshuo Xu, Pascal Pfeiffer, Ryan Chesler, Mark Landry, and Sri Satish Ambati. H2ovl-mississippi vision language models technical report, 2024. https://arxiv.org/abs/2410.13611.

Philippe Gervais, Asya Fadeeva, and Andrii Maksai. Mathwriting: A dataset for handwritten mathematical expression recognition, 2024. https://arxiv.org/abs/2404.10690.

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms, 2025. https://arxiv.org/abs/2502.04326.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm: Unveiling the potential of small language models with scalable training strategies, 2024. https://arxiv.org/abs/2404.06395.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR, 18–24 Jul 2021. https://proceedings.mlr.press/v139/jaegle21a.html.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 235–251, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. The stack: 3 tb of permissively licensed source code. *arXiv preprint arXiv:2211.15533*, 2022.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs, 2023.

Vik Korrapati. Moondream. Online, 2024. https://moondream.ai/. Accessed: 2025-03-27.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An open web-scale filtered dataset of interleaved image-text documents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. https://openreview.net/forum?id=SKN2hflBIZ.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions, 2024. https://arxiv.org/abs/2408.12637.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-LM: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024a.

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 *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, June 2024b.

Rui Li, Xiaohan Wang, Yuhui Zhang, Zeyu Wang, and Serena Yeung-Levy. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*, 2025a.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023b.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016.

Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025b.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023a.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. https://openreview.net/forum?id=w0H2xGHlkw.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024b.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation, 2024c. https://arxiv.org/abs/2310.05209.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024d. https://arxiv.org/abs/2403.00476.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024e.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024f.

Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024g.

Nikolaos Livathinos, Christoph Auer, Maksym Lysak, Ahmed Nassar, Michele Dolfi, Panos Vagenas, Cesar Berrospi Ramis, Matteo Omenetti, Kasper Dinkla, Yusik Kim, et al. Docling: An efficient open-source toolkit for ai-driven document conversion. In *AAAI 25: Workshop on Open-Source AI for Mainstream Use*, 2025.

Alejandro Lozano, Min Woo Sun, James Burgess, Jeffrey J. Nirschl, Christopher Polzak, Yuhui Zhang, Liangyu Chen, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, Collin Chiu, Orr Zohar, Xiaohan Wang, Alfred Seunghoon Song, Chiang Chia-Chun, Robert Tibshirani, and Serena Yeung-Levy. A large-scale vision-language dataset derived from open scientific literature to advance biomedical generalist ai. *arXiv preprint arXiv:2503.22727*, 2025. https://arxiv.org/abs/2503.22727.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024a.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc., 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/11332b6b6cf4485b84afadb1352d3a9a-Paper-Conference.pdf.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024b.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

Ahmed Masry, Do 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*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. https://aclanthology.org/2022.findings-acl.177.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier Biard, Sam Dodge, Philipp Dufter, Bowen Zhang, Dhruti Shah, Xianzhi Du, Futang Peng, Haotian Zhang, Floris Weers, Anton Belyi, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu He, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis & insights from multimodal llm pre-training, 2024. https://arxiv.org/abs/2403.09611.

Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A. Said Gurbuz, Michele Dolfi, Miquel Farré, and Peter W. J. Staar. Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion, 2025a. https://arxiv.org/abs/2503.11576.

Ahmed Nassar, Andres Marafioti, Matteo Omenetti, Maksym Lysak, Nikolaos Livathinos, Christoph Auer, Lucas Morin, Rafael Teixeira de Lima, Yusik Kim, A Said Gurbuz, et al. Smoldocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion. *arXiv preprint arXiv:2503.11576*, 2025b.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. https://arxiv.org/abs/2412.16720.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Weiming Ren, Huan Yang, Jie Min, Cong Wei, and Wenhu Chen. Vista: Enhancing long-duration and high-resolution video understanding by video spatiotemporal augmentation, 2024. https://arxiv.org/abs/2412.00927.

Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, et al. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv preprint arXiv:2408.15998*, 2024.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, 2016. https://arxiv.org/abs/1609.05158.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding, 2024. https://arxiv.org/abs/2307.16449.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. Paligemma 2: A family of versatile vlms for transfer, 2024. https://arxiv.org/abs/2412.03555.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size, 2024. *URL https://arxiv. org/abs/2408.00118*, 1(3), 2024.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024b.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. https://arxiv.org/abs/2412.10302.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv preprint arXiv:2406.08464*, 2024.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words, 2024. https://arxiv.org/abs/2406.06040.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024. https://arxiv.org/abs/2408.01800.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang,

Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024a.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mammoth: Building math generalist models through hybrid instruction tuning, 2023. *URL https://arxiv. org/abs/2309.05653*, 2024b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. https://arxiv.org/abs/2410.02713.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-Levy. Video-star: Self-training enables video instruction tuning with any supervision, 2024a. https://arxiv.org/abs/2407.06189.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*, 2024b.