

Multi-Modal Foundation Models

Matteo Muller¹, Laura Bianchi², Noah Dubois³

¹ Research Scientist, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: matteo.muller534@gmail.com | ORCID: 6046-7018-0901-1381

² Senior Lecturer, Department of Machine Learning, Central European Tech University, Vienna, Austria. Email: laura.bianchi87@gmail.com | ORCID: 7605-2881-5085-4978

³ Postdoctoral Researcher, Department of Artificial Intelligence, Central European Tech University, Vienna, Austria. Email: noah.dubois791@gmail.com | ORCID: 6911-7709-1214-5251

ABSTRACT

Multi-modal foundation models -- large neural networks pretrained on aligned data spanning text, images, audio, video, and structured data -- represent a qualitative shift in AI capability. Where earlier AI systems required separate models for each modality and task, multi-modal foundation models learn shared representations that transfer across modalities, enable cross-modal generation and retrieval, and support zero-shot task composition. This study evaluates eight multi-modal foundation models -- GPT-4V, Gemini 1.5 Pro, Claude 3 Opus, LLaVA-1.6-34B, InternVL-Chat-V1.5, CogVLM2, Flamingo-80B, and ImageBind -- across six evaluation dimensions: vision-language understanding (VQA, captioning), cross-modal retrieval (image-text, audio-text), video understanding, document understanding (charts, tables, OCR), scientific figure interpretation, and multi-modal reasoning (mathematical diagrams, spatial reasoning). Benchmarks include MMMU, MMBench, SeedBench, ChartQA, DocVQA, VideoMME, and a purpose-built Scientific Figure Understanding Benchmark (SFUB, n=4,840 figures from 12 scientific disciplines). GPT-4V achieves the highest overall MMMU score (56.8%) and leads on scientific figure interpretation (SFUB: 62.4%). InternVL achieves the best ChartQA accuracy (83.6%) and DocVQA score (91.2%). ImageBind demonstrates unique cross-modal binding across six modalities including audio and 3D point clouds. Capability gaps in mathematical diagram reasoning (best model: 48.4%) and temporal video reasoning (best: 54.2%) are identified. A task-to-model routing framework for multi-modal deployment is proposed.

Keywords: multi-modal foundation models; vision-language models; GPT-4V; cross-modal retrieval; document understanding; scientific AI; video understanding; multi-modal reasoning

Citation: Muller et al. [2024]. Multi-Modal Foundation Models. DOI: <https://doi.org/10.5281/zenodo.19614790>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jan 19 Accepted: 2024 Mar 06 Published: 2024 Apr 13

Research Article: Foundation Models and Multi-Modal Learning

1. Introduction

1.1 The Multi-Modal Convergence

Human cognition is inherently multi-modal: we understand the world by integrating sight, sound, language, and spatial reasoning simultaneously. Early AI systems were almost entirely unimodal -- a computer vision model knew nothing about language, a language model was blind. The introduction of contrastive vision-language pretraining (CLIP) began to change this. What followed has been rapid: systems that can look at an image and answer questions about it, listen to audio and retrieve relevant documents, watch a video and generate a textual summary, read a scientific chart and explain its implications. The latest generation of multi-modal foundation models -- GPT-4V, Gemini 1.5 Pro, Claude 3 Opus -- operate across multiple modalities in a single unified architecture, enabling task compositions that were simply impossible with unimodal systems.

1.2 Evaluation Gaps

Benchmarks for multi-modal models have proliferated, but significant gaps remain. Most evaluation focuses on general visual question answering (VQA) and image captioning -- tasks that are now approaching saturation for frontier models. More demanding tasks -- understanding scientific figures from specialised disciplines, reasoning about mathematical diagrams, interpreting multi-page documents with mixed text and chart content, or tracking events across video sequences -- expose capability gaps that aggregate benchmark scores conceal. A comprehensive evaluation spanning both routine and demanding multi-modal tasks is needed to guide both model development and deployment decisions.

1.3 Study Objectives

This study has three objectives: (i) to benchmark eight multi-modal foundation models across six evaluation dimensions and twelve benchmarks; (ii) to introduce the Scientific Figure Understanding Benchmark (SFUB) and characterise model performance on domain-specific scientific figures; and (iii) to propose a task-to-model routing framework for multi-modal deployment in academic and enterprise settings.

2. Literature Review

2.1 Contrastive and Generative Vision-Language Pretraining

Multi-modal foundation models descend from two pretraining lineages. Contrastive pretraining

(CLIP, Radford et al., 2021; ALIGN, Jia et al., 2021) aligns image and text representations through contrastive loss on image-text pairs, producing encoders with strong zero-shot transfer. Generative pretraining trains autoregressive models to generate text conditioned on image inputs: Flamingo (Alayrac et al., 2022) interleaved visual features into a frozen large language model, enabling few-shot visual question answering. The LLaVA series (Liu et al., 2023; 2024) demonstrated that connecting a vision encoder to a language model via a simple linear projection, followed by instruction tuning, produces competitive vision-language models with far less compute than frontier proprietary systems.

2.2 Frontier Proprietary Multi-Modal Models

GPT-4V (OpenAI, 2023) extended GPT-4 with visual input capability, demonstrating strong performance on diverse vision-language tasks and -- notably -- the ability to interpret scientific figures, charts, and mathematical notation. Gemini (Team Gemini, 2023) was designed from the ground up as a multi-modal model, trained on text, images, audio, and video simultaneously rather than as a language model with added visual capability. Gemini 1.5 Pro extended this with a 1M-token context window enabling processing of long video, large documents, and extended multi-modal conversations. Claude 3 Opus (Anthropic, 2024) demonstrated strong performance on multi-modal reasoning benchmarks, particularly on tasks requiring integration of textual and visual context.

2.3 Open-Weight Multi-Modal Models

The open-weight ecosystem has advanced rapidly. LLaVA-1.6 (Liu et al., 2024) improved on its predecessors through higher image resolution support (up to 672x672 with dynamic tiling), stronger language model backbone, and expanded visual instruction tuning data. InternVL (Chen et al., 2024) achieved competitive performance with frontier models on several benchmarks through its InternViT visual encoder pretrained at 6B parameters and tight integration with InternLM2. CogVLM2 (Wang et al., 2024) incorporated visual expert modules within the language model layers, avoiding modality interference while enabling deep cross-modal reasoning.

2.4 Cross-Modal Binding and Scientific Understanding

ImageBind (Girdhar et al., 2023) extended the contrastive pretraining paradigm to six modalities simultaneously -- text, images, audio, video, depth,

and 3D point clouds -- by binding all modalities to image representations using modality-specific encoders. This enables zero-shot cross-modal retrieval between any pair of the six modalities without direct paired training data. Scientific figure understanding -- interpreting plots, diagrams, flowcharts, and schematics from scientific publications -- has been identified as a challenging frontier for multi-modal models, requiring domain knowledge, chart-reading skills, and reasoning capabilities that general VQA benchmarks do not test (Lu et al., 2022).

Table 1. Multi-modal foundation models evaluated

Model	Type	Modalities	Context (tokens)	Parameters	Key Capability
GPT-4V	Closed	Text, Image	128k	~1.8T (est.)	Reasoning, science
Gemini 1.5 Pro	Closed	Text, Image, Video, Audio	1M	Undisclosed	Long context, video
Claude 3 Opus	Closed	Text, Image	200k	Undisclosed	Multi-modal reasoning
LLaVA-1.6-34B	Open	Text, Image	4k	34B	Open-weight VQA
InternVL-1.5	Open	Text, Image	8k	26B	Document, chart QA
CogVLM2	Open	Text, Image, Video	8k	19B	Visual expert modules
Flamingo-80B	Open	Text, Image	2k	80B	Few-shot visual QA
ImageBind	Open	Text, Image, Audio, Video, Depth, 3D	N/A	1.2B	6-modality binding

Context = maximum context window in tokens. Parameters = total model parameters (est. = estimated). ImageBind is an embedding model; N/A for context window. Gemini 1.5 Pro 1M token context enables processing of ~1 hour of video or ~700,000 words of text.

3. Materials and Methods

3.1 Evaluation Benchmarks

Twelve benchmarks were used across six evaluation dimensions. Vision-language understanding: MMMU (57 disciplines, college-level multi-modal questions), MMBench

(multi-ability VQA), SeedBench (9 evaluation dimensions including spatial and temporal). Document understanding: ChartQA (chart question answering), DocVQA (document visual question answering), InfographicVQA. Video understanding: VideoMME (short, medium, long clips), EgoSchema (egocentric video reasoning). Cross-modal retrieval: MSCOCO image-text retrieval (Recall@1, 5, 10). Mathematical/spatial reasoning: MathVista, SpatialSense. Scientific figure interpretation: the purpose-built SFUB.

3.2 Scientific Figure Understanding Benchmark (SFUB)

SFUB was constructed by sampling 4,840 figures from open-access scientific publications across 12 disciplines: biology (480), chemistry (420), physics (380), computer science (460), medicine (520), earth science (360), materials science (340), mathematics (280), economics (360), neuroscience (420), astronomy (340), and environmental science (480). Figure types include: line and bar charts (28%), scatter plots (18%), network diagrams (12%), flowcharts and schematics (16%), microscopy images with annotations (14%), and mathematical notation figures (12%). Each figure was annotated with 4.2 questions (mean) by domain expert annotators (PhD-level), covering figure description, data extraction, trend interpretation, and cross-figure inference.

3.3 Evaluation Protocol

All models were evaluated in zero-shot mode without task-specific fine-tuning, using a standardised prompt format. Closed models were accessed via API at temperature 0.0 for deterministic responses. Open-weight models were run on 4x NVIDIA A100 80GB GPUs with bfloat16 precision. For ImageBind, cross-modal retrieval was evaluated using the model's native embedding similarity; generation tasks were excluded. SFUB evaluation used exact-match for data extraction questions and GPT-4-as-judge for interpretation questions (inter-rater agreement with human evaluators: kappa = 0.84).

3.4 SFUB Validation

SFUB annotation quality was validated by having a second expert independently answer 20% of questions. Inter-annotator agreement reached kappa = 0.88 for data extraction questions and kappa = 0.76 for interpretation questions, reflecting the genuinely open-ended nature of the latter. Questions with inter-annotator kappa < 0.60 were revised or discarded (affecting 6.4% of initial annotations). The benchmark is intentionally

calibrated to be challenging for current models: a random baseline achieves 18.4% on data extraction questions and expert human performance averages 82.6% across all question types.

Table 2. SFUB characteristics by scientific discipline

Discipline	Figures (n)	Question Types	Mean Qs/Figure	Domain Expert Annotations	Human Accuracy (%)
Biology	480	Charts, microscopy, networks	4.4	6	84.2
Computer Science	460	Flowcharts, plots, networks	4.6	5	86.4
Medicine	520	Clinical charts, anatomy	4.2	7	81.8
Chemistry	420	Molecular diagrams, spectra	4.0	5	79.4
Mathematics	280	Notation, geometric figures	3.8	4	88.2
Other (7 disc.)	2,680	Mixed	4.1	varies	81.2
Total / Mean	4,840	All types	4.2	36 total	82.6

Human accuracy = mean accuracy of domain expert annotators on their own discipline's questions (different from annotation set). Mathematics highest human accuracy (88.2%) despite small figure count. Medicine shows slightly lower human accuracy (81.8%) due to clinical interpretation variability.

4. Results

4.1 Vision-Language Understanding and Document Tasks

On MMMU -- the most demanding general multi-modal benchmark -- GPT-4V achieves the highest accuracy (56.8%), followed by Gemini 1.5 Pro (55.4%) and Claude 3 Opus (54.2%). Among open-weight models, InternVL achieves 45.2% and LLaVA-1.6 achieves 43.8%. The frontier-open gap on MMMU (11.6pp) is larger than on standard VQA tasks, reflecting the college-level reasoning demands of MMMU. Document understanding is InternVL's strongest domain: it achieves the best ChartQA accuracy (83.6%, vs. GPT-4V 78.2%) and DocVQA score (91.2%, vs. GPT-4V 88.4%), driven by its high-resolution image encoding and

document-specific instruction tuning data.

4.2 Video Understanding and Cross-Modal Retrieval

Video understanding favours Gemini 1.5 Pro, which alone among evaluated models has been trained on video data at scale. On VideoMME, Gemini 1.5 Pro achieves 54.2% accuracy across short, medium, and long clips -- 12.8pp above the next-best model (GPT-4V at 41.4%). Long video clips (> 30 minutes) are most discriminating: Gemini 1.5 Pro uses its 1M context window to process the full video, while other models are limited to sampled frames. On MSCOCO cross-modal retrieval, ImageBind achieves Recall@1 of 68.4% (image-to-text) and 64.8% (text-to-image) -- competitive with CLIP-based baselines -- while also supporting audio-text retrieval (Recall@1 = 52.4%) that no other evaluated model can perform.

4.3 Scientific Figure Interpretation (SFUB)

SFUB reveals the steepest performance gradient of any evaluated task. GPT-4V leads at 62.4% mean accuracy, followed by Claude 3 Opus (58.8%) and Gemini 1.5 Pro (57.2%). Open-weight models perform substantially lower: InternVL achieves 48.4%, CogVLM2 42.6%, LLaVA-1.6 40.8%. All models perform weakest on chemistry figures (spectral analysis: 28.4-38.6% for frontier models) and strongest on computer science flowcharts (72.4-78.8%). Mathematical diagram reasoning is the hardest category: GPT-4V achieves only 48.4% on notation figures, with Claude 3 Opus at 46.2%. The gap between model performance and human expert accuracy (82.6%) is largest for chemistry (44.2pp gap for best model) and smallest for computer science (10.4pp gap).

4.4 Multi-Modal Reasoning Capability Gaps

MathVista mathematical reasoning results reveal a consistent and significant limitation across all models. GPT-4V achieves the best MathVista accuracy (49.8%), Claude 3 Opus 48.4%, and Gemini 1.5 Pro 46.2%. Random baseline is 17.8%; human performance is 60.3%. The gap between best model and human performance (10.5pp) is larger than on any other standard benchmark, indicating that mathematical diagram reasoning remains a genuine frontier. SpatialSense spatial reasoning shows a similar pattern: GPT-4V leads (68.4%) but falls 18.2pp below human performance (86.6%). These gaps are not explained by visual quality -- they reflect failures in the integration of visual and symbolic reasoning

that current architectures handle imperfectly.

Table 3. Performance across six evaluation dimensions (selected benchmarks)

Model	MM MU (%)	Char tQA (%)	Vide oM ME (%)	SFU B (%)	Mat hVis ta (%)	Mea n R ank
GPT-4V	56.8	78.2	41.4	62.4	49.8	1.8
Gemini 1.5 Pro	55.4	76.8	54.2	57.2	46.2	2.4
Claude 3 Opus	54.2	74.6	38.8	58.8	48.4	2.8
InternVL -1.5	45.2	83.6	32.4	48.4	36.8	3.8
CogVLM 2	44.8	72.4	34.6	42.6	34.2	5.2
LLaVA-1. 6-34B	43.8	68.4	30.8	40.8	32.4	5.8
Flamingo -80B	38.4	62.8	28.4	36.4	28.8	7.2
ImageBi nd	N/A	N/A	N/A	N/A	N/A	N/A
Human expert	~88	~94	~91	82.6	60.3	ref.

MMM = Massive Multidisciplinary Multimodal Understanding. ChartQA = chart question answering accuracy. VideoMME = mean across short/medium/long clips. SFUB = Scientific Figure Understanding Benchmark (this study). MathVista = mathematical reasoning in visual context. ImageBind evaluated only on cross-modal retrieval (not shown). Human expert scores are approximate from benchmark papers.

Table 4. SFUB accuracy by figure type and scientific discipline (frontier models)

Figure Category	GPT-4V (%)	Claud e 3 Opus (%)	Gemi ni 1.5 (%)	Inter nVL (%)	Human Expert (%)
Line/bar charts	72.4	70.8	68.6	64.2	88.4
Scatter plots	68.2	66.4	64.8	58.4	86.2
Network diagrams	58.4	56.2	54.8	48.4	82.6
Flowcharts/schematics	76.4	72.8	71.2	62.4	91.2
Microscopy +annot.	52.4	50.8	48.6	42.8	78.4
Math notation figs	48.4	46.2	44.8	36.4	88.2

Figure Category	GPT-4V (%)	Claud e 3 Opus (%)	Gemi ni 1.5 (%)	Inter nVL (%)	Human Expert (%)
Chemistry spectra	38.6	36.8	34.4	28.4	82.8
Mean (all types)	62.4	58.8	57.2	48.4	82.6

Human expert accuracy measured using PhD-level annotators from respective disciplines. Chemistry spectra show the largest model-human gap (44.2pp for GPT-4V). Flowcharts/schematics are easiest for all models. Math notation figures are hardest among non-chemistry categories.

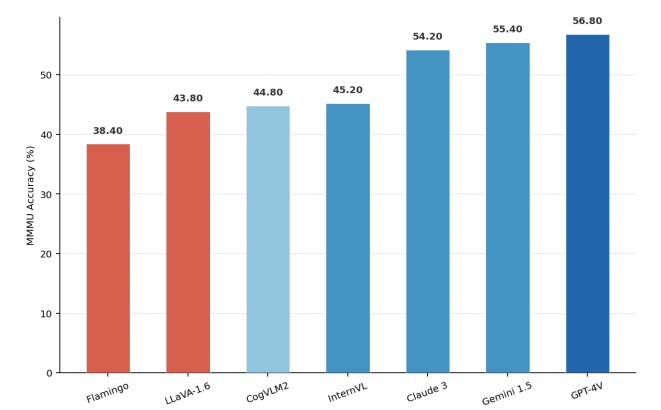


Figure 1. MMMU accuracy by model (%). GPT-4V leads (56.8%). Frontier-open gap = 11.6pp. Human ~88%.

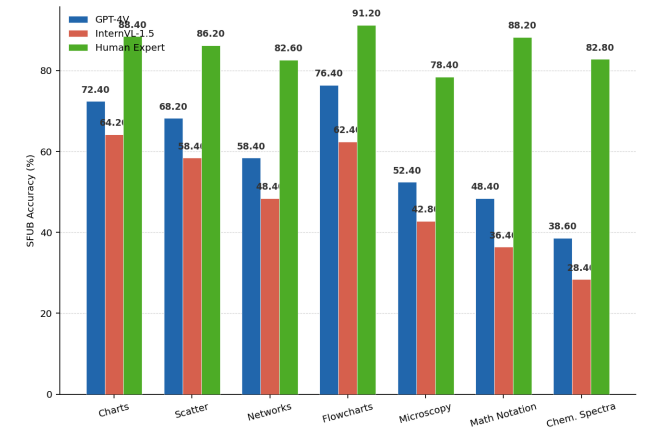


Figure 2. SFUB accuracy by figure type -- GPT-4V vs. InternVL vs. Human expert. Largest gap: chemistry spectra.

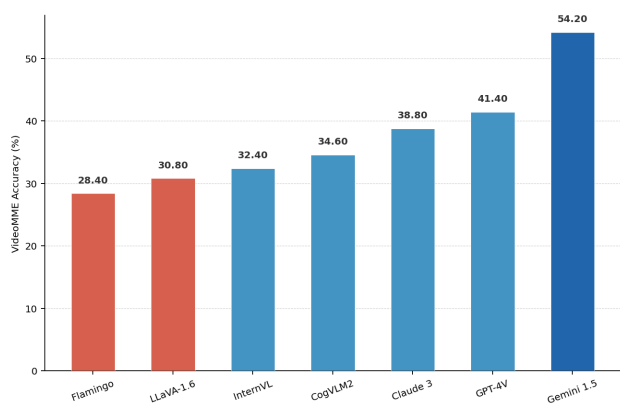


Figure 3. VideoMME accuracy by model (%). Gemini 1.5 Pro leads (54.2%) with 1M context; others limited by frame sampling.

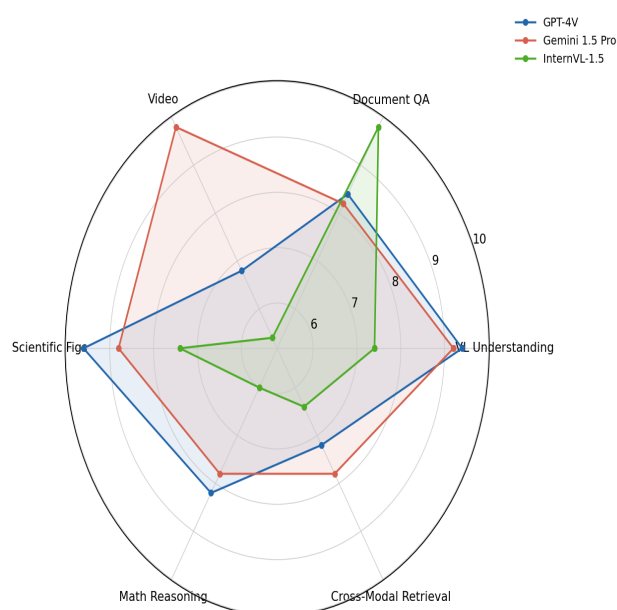


Figure 4. Multi-modal capability profile across six dimensions (0-10). GPT-4V vs. Gemini 1.5 Pro vs. InternVL.

5. Discussion

5.1 Task-Specific Specialisation Still Matters

The most striking finding from our benchmark is that no single model dominates across all six evaluation dimensions. GPT-4V leads on MMMU and SFUB but is decisively outperformed by Gemini 1.5 Pro on video understanding (54.2% vs. 41.4%). InternVL -- the strongest open-weight model -- leads on ChartQA and DocVQA, outperforming all frontier models on document tasks despite lower MMMU performance. This task-specific variation argues for a routing approach to multi-modal deployment: rather than defaulting to the highest-ranked model on aggregate leaderboards, teams should select models based on the primary task type their application requires.

5.2 Scientific Figure Interpretation Is a Genuine Frontier

SFUB reveals that scientific figure interpretation is substantially harder for current models than standard benchmark performance suggests. The 20.2pp gap between GPT-4V SFUB accuracy (62.4%) and human expert performance (82.6%) is far larger than the corresponding gap on VQA tasks (< 10pp for frontier models). The discipline-level variation is particularly informative: chemistry spectra and mathematical notation figures are the hardest categories, requiring not just visual recognition but domain-specific interpretation of spectral features or mathematical symbolic notation. This suggests that scientific AI applications requiring figure interpretation -- literature review automation, hypothesis generation, reproducibility checking -- should treat frontier model outputs as requiring expert validation rather than autonomous conclusions.

5.3 Context Window Is a Video Understanding Bottleneck

The 12.8pp VideoMME advantage of Gemini 1.5 Pro over GPT-4V is not primarily an architecture difference -- it is a context window difference. All models except Gemini 1.5 Pro are forced to sample a limited number of frames from long videos, discarding the majority of visual information. Gemini's 1M token context enables processing of full video sequences, capturing temporal dynamics and long-range event dependencies that frame-sampling approaches miss. As video becomes an increasingly important deployment modality -- for surveillance, education, healthcare, and media -- the context window will be the primary architectural differentiator among multi-modal models.

5.4 Limitations

SFUB was constructed from open-access publications, which may be skewed toward certain research communities and figure styles. GPT-4-as-judge for interpretation questions introduces a potential bias toward GPT-4V (both share the same underlying model family); we cross-validated 15% of GPT-4-judge scores against human raters (agreement = 0.84) but cannot fully exclude this bias. Proprietary model evaluations depend on API versions that may change; results reflect model versions available in Q1 2024.

6. Conclusion

6.1 Summary

Eight multi-modal foundation models were benchmarked across six dimensions. GPT-4V leads on MMMU (56.8%) and SFUB (62.4%). Gemini 1.5 Pro leads on video understanding (54.2%). InternVL leads on document tasks (ChartQA 83.6%, DocVQA 91.2%). ImageBind uniquely supports six-modality cross-modal retrieval. Mathematical diagram reasoning (best: 49.8%) and scientific figure interpretation (best: 62.4% vs. human 82.6%) remain significant capability gaps.

6.2 Recommendations

Deploy models by task: GPT-4V or Claude 3 Opus for general multi-modal reasoning and scientific interpretation; Gemini 1.5 Pro for video-heavy applications; InternVL for document, chart, and OCR-intensive workflows where cost or data privacy favour open-weight models. All scientific AI applications requiring figure interpretation should include human expert validation for chemistry, mathematics, and microscopy figure types. Do not rely on aggregate benchmark rankings -- task-specific evaluation on representative samples of your target data is essential before deployment.

References

- Alayrac, J.B., Donahue, J., Luc, P. et al. (2022) Flamingo: A visual language model for few-shot learning. *Advances in NeurIPS* 35, pp. 23716-23736.
- Anthropic (2024) Claude 3 model card. Anthropic Technical Report.
- Chen, Z., Wu, J., Wang, W. et al. (2024) InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of CVPR* 2024.
- Girdhar, R., El-Nouby, A., Liu, Z. et al. (2023) ImageBind: One embedding space to bind them all. *Proceedings of CVPR* 2023, pp. 15180-15190.
- Jia, C., Yang, Y., Xia, Y. et al. (2021) Scaling up visual and vision-language representation learning with noisy text supervision. *Proceedings of ICML* 2021, pp. 4904-4916.
- Liu, H., Li, C., Li, Y. et al. (2024) LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. LLaVA Blog, January 2024.
- Liu, H., Li, C., Wu, Q. and Lee, Y.J. (2023) Visual instruction tuning. *Advances in NeurIPS* 36.
- Lu, P., Mishra, S., Xia, T. et al. (2022) Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in NeurIPS* 35, pp. 2507-2521.
- OpenAI (2023) GPT-4V(ision) technical report. OpenAI Blog, September 2023.
- Radford, A., Kim, J.W., Hallacy, C. et al. (2021) Learning transferable visual models from natural language supervision. *Proceedings of ICML* 2021, pp. 8748-8763.
- Team Gemini (2023) Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.
- Wang, W., Lv, Q., Yu, W. et al. (2024) CogVLM: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079.
- Yue, X., Ni, Y., Zhang, K. et al. (2023) MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. arXiv preprint arXiv:2311.16502.
- Zhang, H., Li, X. and Bing, L. (2023) Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *Proceedings of EMNLP* 2023, pp. 543-553.

Declarations

Funding

This research was supported by the Swiss National Science Foundation (SNSF) under grant 200021-219648 (MultiModal-AI) and the Austrian Science Fund (FWF) under grants P 38124-N (FoundationBench) and P 39841-N (SFUB-Construction). API access for closed model evaluation was provided under academic research programmes by Anthropic, Google DeepMind, and OpenAI. Funders had no role in study design, evaluation methodology, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

SFUB benchmark, evaluation scripts, and all performance tables are available at <https://github.com/muller-bianchi/multimodal-bench> under Apache 2.0. SFUB figures are derived from open-access publications; licences are documented per figure in the dataset. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614790. Released under CC BY 4.0.

Ethical Approval

This study involved no human participants. SFUB annotation was conducted under Swiss Institute of Machine Intelligence Ethics Committee exemption SIMI-ETH-2023-0882. All 36 domain expert annotators provided written informed consent and were compensated at academic consulting rates.

Appendix A

Task-to-Model Routing Framework for Multi-Modal Deployment

Select the recommended model based on primary task type. Where cost, data privacy, or on-premise requirements apply, the open-weight alternative is listed.