

# A Brief Overview of Multimodal Large Language Models

Özlem Karabulut

University of Tübingen / Department of Computational Linguistics  
oezlem.karabulut@student.uni-tuebingen.de

## Abstract

Multimodal large language models (MLLMs) combine different modalities (text, vision, audio, and video) to enable grounded understanding, reasoning, and generative capabilities. This survey reviews core architectures, training strategies, and evaluation paradigms, with a focus on vision-language MLLMs. The rapid progress inevitably introduces challenges in robustness, efficiency, and safety. Future work aims to develop more capable and reliable multimodal systems.

## 1 Introduction

Humans integrate information from multiple senses, and MLLMs increasingly strive to mirror this ability by processing and reasoning over multimodal data. Combining modalities is motivated by (i) the fact that complementary cross-modal signals enable richer representations and robust reasoning, and (ii) many real-world tasks are multimodal (Baltrušaitis et al., 2017). This paper covers conceptual foundations, model architectures, training strategies, evaluation benchmarks, key challenges, and future directions.

## 2 Background

MLLMs perform crossmodal learning on large, diverse datasets to produce joint embedding spaces. This enables crossmodal generalization: interpreting complex inputs and reasoning about dependencies (Liang et al., 2024). It unlocks diverse tasks, from image captioning to visual question answering (VQA) (Yin et al., 2024). MLLMs typically use different approaches to fuse modalities. Contrastive objectives bring matched pairs closer and push mismatched ones apart, yielding efficient representations. In contrast, generative objectives train models to produce text conditioned on visual inputs and cause higher computational costs. Recent work

combines two objectives to balance efficiency and expressiveness (Chen et al., 2020).

**Brief History.** Key milestones such as the encoder-decoder architecture and cross-attention (Vaswani et al., 2017) paved the way for MLLMs. Earlier LLMs achieved zero-shot transfer by pre-training on raw web-scale text (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2023). However, in computer vision, it was still standard practice to pretrain on crowd-labeled datasets, e.g., ImageNet (Deng et al., 2009). Vision Transformers (ViTs) (Dosovitskiy et al., 2021) first applied transformer logic to images by splitting them into patches and processing them sequentially. This enabled CLIP (Radford et al., 2021) to unlock vision generalization capabilities on par with large language models (LLMs) by scaling contrastive learning on vast image-text data. Other ViT examples (e.g., ALIGN (Jia et al., 2021), Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023b)) emerged subsequently, followed by CNN-based and hybrid models. Recent models, e.g., GPT-4V (OpenAI, 2023) and Gemini family (Gemini et al., 2025) shows advanced perception skills. This paper covers architectural and training paradigms in Section 3 and 4, respectively.

## 3 Model Architectures

Understanding MLLM components is crucial to grasp their significance. Main parts are a modality encoder, a pretrained LLM, and a modality interface connecting them (Yin et al., 2024).

### 3.1 Modality Encoder

Multimodal encoders vary widely in design. CLIP-based models (Cherti et al., 2023; Zhu et al., 2023; Sun et al., 2024) leverage its semantic alignment capability. Alternative encoders use convolutional neural networks (CNNs) (He et al., 2015; Tan and Le, 2020). Some CLIP variants use CNN encoders:

Yuan et al. (2025) use ConvNext-L (Liu et al., 2022) for better pixel-wise understanding via high-resolution features. CNN-based CLIP showed better generalization across various input resolutions, though ViT-based variants show better zero-shot performance (Yu et al., 2023; Wu et al., 2024). Some models eliminate the encoder and use a generator: Bavishi et al. (2023) directly project image patches into the first transformer layer, bypassing the embedding lookup and supporting arbitrary resolutions. McKinzie et al. (2024) showed that image resolution is more important compared to parameter size and training data, and using higher resolution can attain notable performance gains (Liu et al., 2024b; Li et al., 2024c). Thus, MLLMs adopt two main scaling strategies to process higher-resolution images: *direct scaling*, which adapts the encoder via fine-tuning or an additional pretrained encoder (e.g., Hong et al. (2024)), or *patch-division*, which adapts the input by cutting the image into patches compatible with the original encoder resolution (e.g., Li et al. (2024c); Lin et al. (2023)). For modality-specific encoders, examples include Deshmukh et al. (2024) using CLAP-variants (Nizumi et al., 2025) for audio and Han et al. (2023) using ImageBind Girdhar et al. (2023).

### 3.2 Pretrained LLM

MLLMs use predominantly decoder-only architectures known for scalable generative and instruction-following capabilities, while earlier models (e.g., Dai et al. (2023)) used encoder-decoder setups. Open-source options, e.g., LLAMA (Touvron et al., 2023), are widely used in academia but remain limited in multilingual tasks. In contrast, models such as Qwen (Bai et al., 2023) support bilingual tasks. Recent state-of-the-art (SoTA) models (e.g., McKinzie et al. (2024); Lin et al. (2024)) also use mixture-of-experts (MoE) architecture to address the growing demand for efficiency and task specialization.

### 3.3 Modality Interface

Since LLMs accept only text inputs, interface is needed to integrate other modalities into text representations. However, it is computationally expensive to train an MLLM end-to-end. Thus, more practical ways are introduced.

#### 3.3.1 Learnable Interface

Different fusion approaches are adopted to project information to a space that LLMs understand.

*Token-level fusion* concatenates transformed encoder output tokens with text tokens. It can be applied in two ways: (i) Query-based fusion learns some query tokens to compress visual tokens into meaningful embeddings. After BLIP-2 introduced this approach, subsequent models (Dai et al., 2023; Chen et al., 2023a; Zhang et al., 2023) adopted it. (ii) MLP-based fusion uses a simple MLP-based interface to project and align visual tokens with word embeddings (e.g., LAVA series (Liu et al., 2023b); Su et al. (2023); Pi et al. (2023); Zhang et al. (2024b)). *Feature-level fusion* inserts extra modules between transformer layers via attention mechanism. For example, Flamingo injects extra cross-attention layers, Wang et al. (2024) visual expert models, and Zhang et al. (2024a) learnable prompts. Zeng et al. (2023) found that token-level fusion yields stronger VQA performance, and feature-level fusion often needs better hyperparameter optimization to compete.

#### 3.3.2 Expert Model

This pipeline converts modalities to text without joint training. Li et al. (2024b) uses vision models enriched with a speech recognition model, but this extra conversion may cause information loss and is less flexible and efficient than learnable interfaces. Xu et al. (2025) explore retrieval-augmented generation to mitigate this problem.

## 4 Training Strategies and Data

Two main training phases are pretraining and finetuning. Each phase discussed in this section serves different purposes.

### 4.1 Pretraining and Data

**Data.** Training the model on large-scale data to build general world knowledge is essential. McKinzie et al. (2024) carefully conduct ablations on MLLM components and data, and reveal important findings, such as showing that a careful mixture of multimodal data can yield optimal performance. Data used in this mixture differ in quality and source: *coarse-grained data* is typically large-scale and noisy as it is often scraped from the web. Various cleaning methods via tools (e.g., CLIP) can improve quality (Schuhmann et al., 2022). In contrast, *fine-grained data* is obtained through methods such as prompting strong MLLMs (e.g., GPT-4V). This results in smaller but higher-quality data, though at higher costs. Chen et al. (2023b) balance

this trade-off using a pretrained captioner. Readers can find a detailed list of datasets in Table 2.

**Pretraining Objectives.** Main pretraining approaches include contrastive and masked objectives, which are usually combined in practice (Chen et al., 2020). CLIP-based models adopt *contrastive learning* for zero-shot transfer; this makes models robust to noise and highly scalable (Radford et al., 2021; Jia et al., 2021). Jiang et al. (2024) augments hallucinated captions as hard negatives to address one of the main weaknesses: hallucination, i.e., where models invent non-existent facts/objects. *Multimodal masked modeling* combines masked language modeling and masked image modeling, meaning the model (e.g., as in Chen et al. (2020)) learns to recover masked words/image regions, thereby learning joint representations. Researchers often use it with image-text matching or contrastive learning (e.g., Bugliarello et al. (2021)). *Visual and language pretraining (VLP)* combines previous approaches to perform complex tasks (e.g., Lu et al. (2019); Li et al. (2020)). A common challenge is that the models are usually trained on broad-domain data, so applying them to niche domains often requires domain-specific pretraining (as in Khare et al. (2021)).

## 4.2 Fine-Tuning

After pretraining, the model undergoes the following fine-tuning steps and methods to adapt its general world knowledge to specific downstream tasks.

**Data and Common Practices:** Models need task-specific data to adapt their knowledge, e.g., VQAv2 dataset (Agrawal et al., 2016) for VQA tasks. Even more specialized datasets exist, e.g., Singh et al. (2019) to test reading and reasoning about text in images. The second crucial step is hyperparameter optimization. Starting with a small learning rate (LR) to preserve world knowledge, using learning-rate decay schedules (Dale, 2025; Kalra and Barkeshli, 2024), and using adaptive optimizers (Loshchilov and Hutter, 2019) are example methods. Layer-wise strategies are also used: Singh et al. (2015) employs layer-wise rate decay. As for efficiency challenges, researchers often employ parameter-efficient fine-tuning (PEFT) methods (e.g., LoRA (Hu et al., 2021)), which include adding new task-specific layers, tokens, or adapter blocks on top of the frozen base.

**Multi-task Fine-tuning:** Training on multiple tasks at once allows MLLMs to generalize better across related tasks: Nguyen and Okatani (2018) and Mahabadi et al. (2021) report better generalization performance by fine-tuning on several tasks together. MFTCoder (Liu et al., 2023a) also report SoTA on HumanEval (Chen et al., 2021). This method requires careful design since tasks may have conflicting objectives and require balanced sampling/scheduling (Sener and Koltun, 2019).

**Improving Alignment on Crossmodal Tasks:** Some tasks, such as image–text retrieval, require precise alignment. SoTA approaches use additional alignment modules to address this. For instance, M2IST (Liu et al., 2025) adds side MoE adapters, showing superior results on referring expression tasks.

**Transfer and Few/zero-shot Learning:** Pre-trained MLLMs’ strong generalization capacities enable few/zero-shot learning, which are forms of transfer learning. It is especially valuable when labeled datasets are scarce or costly. Few-shot learning provides MLLM with only a few new-task examples (via prompts or a small dataset) (Huang et al., 2023; Tsimpoukelli et al., 2021), while zero-shot learning requires no examples. In practice, researchers often evaluate MLLMs in a zero-shot manner. However, challenges persist: generalizing to a novel task may be difficult, and a poor-quality scheme can lead to suboptimal performance. These issues can be mitigated by incorporating rich pre-training data and research into sophisticated fine-tuning methods.

**Instruction-Tuning** Multimodal instruction-tuning trains models on natural language instructions so they learn to follow complex human-like commands across modalities and tasks (Li et al., 2024a). This improves efficiency by reducing fine-tuning costs (Peng et al., 2023). Overall, instruction-tuned MLLMs can generalize to new tasks with few or zero new-task examples, reducing sensitivity to instruction variations (Xu et al., 2023b; Liu et al., 2023b).

**Chain-of-Thought (CoT) Prompting:** In this method, MLLM improves multimodal reasoning by generating intermediate steps, first producing rationales that interpret the image, then using these to infer answers (Zhang et al., 2024c). It shows gains on reasoning-heavy tasks and enhances model inter-

pretrainability, though it adds computational overhead (Wei et al., 2023).

**Adaptation on New Tasks and Domains:** A model often loses performance on prior tasks when sequentially fine-tuned on new tasks. Standard optimizers do not guard against this catastrophic forgetting problem (Chakravarthy et al., 2025). Mitigating strategies include replay buffers (Rolnick et al., 2019), distillation (Hinton et al., 2015), parameter isolation (Zeng et al., 2025), using a small LR, and freezing layers. When model encounters a niche task, domain adaptation strategies include gradual fine-tuning (Xu et al., 2021), where the model is first fine-tuned on a broad dataset and then on target domain, and domain-adversarial training (Ganin et al., 2016).

## 5 Evaluation Paradigms and Applications

This section summarizes established tasks, common metrics, and benchmarks. Readers can see Tables 1 and 3 for detailed descriptions.

### 5.1 Standard Evaluation Benchmarks and Metrics

MLLMs are typically evaluated along two complementary axes: vision–language understanding tasks and generative tasks.

#### 5.1.1 Core Vision-Language Understanding

Classical evaluation tasks include Image Captioning, VQA, Visual Commonsense Reasoning, and Crossmodal Retrieval to test vision-language fusion. Recently, LLaVA-1.5 (Liu et al., 2024b) reported SoTA, outperforming PaLI Du et al. (2022) and early LLaVA baselines on 11 benchmarks.

**Image Captioning:** MLLMs improved significantly in generating contextually accurate captions of an image (Li et al., 2020; Huang et al., 2025b). Standard automated n-gram metrics measure semantic correctness and linguistic quality of captions and are used in benchmarks such as COCO Captions (Chen et al., 2015). Recently, learned metrics (e.g., Zhang et al. (2020); Sellam et al. (2020); Hessel et al. (2022); Ruiz et al. (2025)) are preferred since n-gram metrics poorly capture semantic fluency (Pu et al., 2021) and are often complemented by human feedback due to task’s subjective nature<sup>1</sup>.

<sup>1</sup>LMArena is currently SoTA for human-preference ranking: <https://lmarena.ai/>

**Visual Question Answering (VQA):** Recent MLLMs use co-attention mechanisms (Liu et al., 2024a) or knowledge-enhanced models (Lan et al., 2023) in VQA tasks. Benchmarks measure the model’s capability to answer open-ended questions about images. Usually, accuracy is measured on VQAv2. Variants include domain-specific datasets and versions for other modalities (e.g., Yang et al. (2022); Cao et al. (2024); Xu et al. (2016)).

**Crossmodal Retrieval:** Metrics such as Recall@K, median rank (MedR), and mean reciprocal rank (MRR) measure the performance of the model for retrieving relevant images given a query and vice versa (Kiros et al., 2014; Faghri et al., 2018). ALIGN and subsequent models can serve as comparative baselines.

**Multimodal Classification and Visual Commonsense Reasoning (VCR):** Typical classification metrics include accuracy and F1-scores. Conversely, VCR (Zellers et al., 2019) requires more advanced metrics, including multiple-choice setups or annotated rationale evaluation. Because it extends beyond object recognition or scene description: it requires MLLMs to exhibit a nuanced understanding by incorporating contextual knowledge, causal relationships, and social dynamics. This capability enables models to predict an event or explain the cause of it. Benchmarks such as GQA (Hudson and Manning, 2019) test this ability.

#### 5.1.2 Generative and Creative Applications

These tasks focus on skills such as multimodal content creation and editing, rather than understanding. Evaluation is difficult, so recent practice combines learned, task-specific, or embedding-based metrics with human or LLM-assisted evaluations to judge coherence, faithfulness, quality, and safety. Composite benchmarks usually test these applications.

### 5.2 Holistic Evaluations(Composite/Meta Benchmarks)

Conventional tasks examine narrow capabilities and cannot fully characterize general-purpose MLLMs. Recent unified benchmarks, e.g., AbilityLens (Chen et al., 2025), address this gap by probing advanced tasks combined and aim to fix significant evaluation variance of first-generation composite benchmarks (e.g., Liu et al. (2024c); Li et al. (2023a)). These suites are useful to expose MLLMs’ strengths, weaknesses, and future directions.

### 5.3 Safety, Fairness and Content Moderation

Generative models introduced novel safety challenges. Web-scale data naturally contains biases, causing models to propagate stereotypes and hallucinations (Chen et al., 2023c). MLLMs should be audited for fairness and harmful content. Several frameworks have been developed to address this. For example, Chen et al. (2023c) conduct demographic analyses to identify potential biases. LLaVaShield (Huang et al., 2025a) safeguards multi-turn dialogues from malicious intent, outperforming baselines on content moderation. Raza et al. (2025) show that visual cues improve bias detection in multimodal news accuracy by 3–5%. Cui et al. (2025) focus on detecting multimodal implicit toxicity, i.e., individual modalities convey risk when combined, and provide a taxonomy of risks. Diagnostic benchmarks (e.g., H-POPE (Pham and Schott, 2024)) aim to measure hallucination, which is another critical challenge. New alignment methods can also help: Xing et al. (2025) focuses on vision-language alignment to mitigate hallucination. It achieves SoTA results on both POPE (Li et al., 2023c) and HallusionBench (Guan et al., 2024). Comprehensive benchmarks such as MM-SOC (Jin et al., 2024) are designed to evaluate many risks together.

### 5.4 Applications

Emerging applications aim to leverage the full capabilities of MLLMs to broaden their accessibility, domain coverage, and use cases. Examples include multimodal chatbots, augmented reality (AR), personalization, robotics, healthcare, assistive technologies, and smart-home assistants. Several works (Huang et al., 2023; Li et al., 2024a) demonstrate strong instruction-following capabilities. Gemini 2.5 represents a new generation of complex reasoning capabilities with SoTA performance (Comanici et al., 2025). For healthcare, Saab et al. (2024) achieved a new SoTA of 91.1% accuracy for MedQA (Jin et al., 2020), surpassing GPT-4V on NEJM (Buckley et al., 2024). For personalization, movie and e-commerce recommender systems can leverage multimodal features.

Overall, as MLLMs evolve, evaluation suites increasingly shift toward integrated, real-world, and user-centric benchmarks and provide significant insights that identify limitations and future directions through exhaustive evaluation.

## 6 Challenges and Future Directions

Despite impressive progress, several challenges remain:

**Efficiency and Scalability:** Chen et al. (2023c) note that scaling modalities improves performance, but training remains cost-intensive. Techniques such as PEFT, knowledge distillation, and model compression can reduce the cost. Lightweight encoders (e.g., LightCRL (Faye et al., 2024)) show promising reductions.

**Safety and Fairness:** Future MLLMs must better detect harmful content, improve alignment, incorporate uncertainty estimation, and respect user intent. Transparency in data and model cards should become standard.

**Multilingual and Low-Resource Modalities:** Most MLLMs still predominantly support English, due to limited multilingual multimodal data (Gao et al., 2025). Expanding coverage to low-resource languages and culturally diverse imagery will enhance inclusivity and applicability.

**Evaluation Mismatch:** Current benchmarks test narrow tasks and overlook real-world scenarios. Developing standardized evaluation suites that assess reasoning, grounding, and safety is essential.

**Interactive Multimodality and Embodied Applications:** Models that learn through user interaction can adapt to personal needs and preferences, extending accessibility. Incorporating human feedback (e.g., through reinforcement learning from human feedback (Ouyang et al., 2022)) and active learning improves robustness. Combining these with spatial and physical reasoning will enable embodied agents that operate in the real world.

## 7 Conclusion

MLLMs have transitioned from simple feature fusion to advanced general-purpose systems that reason across modalities. Recent progress signal a future where multimodality is a native capability. Despite these developments, achieving safe, reliable, and inclusive MLLMs requires responsible data practices, innovative architectures, and standardized evaluation. Addressing these challenges will help make multimodality a core component of future AI systems.

## References

- Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. [Vqa: Visual question answering](#).
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. [Flamingo: a visual language model for few-shot learning](#).
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. [Spice: Semantic propositional image caption evaluation](#).
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. [Qwen technical report](#).
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. [Multimodal machine learning: A survey and taxonomy](#).
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. [Introducing our multimodal models](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).
- Thomas Buckley, James A. Diao, Pranav Rajpurkar, Adam Rodman, and Arjun K. Manrai. 2024. [Multimodal foundation models exploit text to make medical image predictions](#).
- Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. [Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts](#).
- Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. <https://github.com/kakaobrain/coyo-dataset>.
- Yuqin Cao, Xiongkuo Min, Yixuan Gao, Wei Sun, Weisi Lin, and Guangtao Zhai. 2024. [Unqa: Unified no-reference quality assessment for audio, image, video, and audio-visual content](#).
- Anirudh S Chakravarthy, Shuai Kyle Zheng, Xin Huang, Sachithra Hemachandra, Xiao Zhang, Yuning Chai, and Zhao Chen. 2025. [Profit: A specialized optimizer for deep fine tuning](#).
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. [Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts](#).
- Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023a. [X-ilm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages](#).
- Feng Chen, Chenhui Gou, Jing Liu, Yang Yang, Zhaoyang Li, Jiyuan Zhang, Zhenbang Sun, Bohan Zhuang, and Qi Wu. 2025. [Evaluating and advancing multimodal large language models in perception ability lens](#).
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. [Sharegpt4v: Improving large multi-modal models with better captions](#).
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Heffgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#).

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023c. *Pali: A jointly-scaled multilingual language-image model*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. *Microsoft coco captions: Data collection and evaluation server*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. *Uniter: Universal image-text representation learning*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Itsev. 2023. *Reproducible scaling laws for contrastive language-image learning*. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2818–2829. IEEE.

Gheorghe Comanici, Eric Bieber, et al. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*.

Nick Craswell. 2009. *Mean Reciprocal Rank*, pages 1703–1703. Springer US, Boston, MA.

Shiyao Cui, Qinglin Zhang, Xuan Ouyang, Renmiao Chen, Zhixin Zhang, Yida Lu, Hongning Wang, Han Qiu, and Minlie Huang. 2025. *Shieldvlm: Safeguarding the multimodal implicit toxicity via deliberative reasoning with lvlms*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. *Instructblip: Towards general-purpose vision-language models with instruction tuning*.

Dan Dale. 2025. *Introduction to the fine-tuning scheduler*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. *Imagenet: A large-scale hierarchical image database*. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2024. *Pengi: An audio language model for audio tasks*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *Bert: Pre-training of deep bidirectional transformers for language understanding*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. *An image is worth 16x16 words: Transformers for image recognition at scale*.

Xiyang Du, Dou Hu, Jin Zhi, Lianxin Jiang, and Xiaofeng Shi. 2022. *PALI-NLP at SemEval-2022 task 6: iSarcasmEval- fine-tuning the pre-trained model for detecting intended sarcasm*. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 815–819, Seattle, United States. Association for Computational Linguistics.

Haodong Duan, Xinyu Fang, Junming Yang, Xiangyu Zhao, Yuxuan Qiao, Mo Li, Amit Agarwal, Zhe Chen, Lin Chen, Yuan Liu, Yubo Ma, Hailong Sun, Yifan Zhang, Shiyin Lu, Tack Hwa Wong, Weiyun Wang, Peiheng Zhou, Xiaozhe Li, Chaoyou Fu, Junbo Cui, Jixuan Chen, Enxin Song, Song Mao, Shengyuan Ding, Tianhao Liang, Zicheng Zhang, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2025. *Vlmevalkit: An open-source toolkit for evaluating large multi-modality models*.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. *Vse++: Improving visual-semantic embeddings with hard negatives*.

Bilal Faye, Hanane Azzag, Mustapha Lebbah, and Djamel Bouchaffra. 2024. *Lightweight cross-modal representation learning*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. *Domain-adversarial training of neural networks*.

Yufei Gao, Feijiaying, Guohang Yan, and Yunshi Lan. 2025. *Improving multimodal large language models in low-resource language contexts*. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Team Gemini et al. 2025. *Gemini: A family of highly capable multimodal models*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. *Audio set: An ontology and human-labeled dataset for audio events*. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. *Imagebind: One embedding space to bind them all*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. *Hallusionbench: An advanced*

- diagnostic suite for entangled language hallucination and visual illusion in large vision-language models.
- Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. [Imagebind-llm: Multi-modality instruction tuning](#).
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. [Deep residual learning for image recognition](#).
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. [Clipscore: A reference-free evaluation metric for image captioning](#).
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#).
- Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. [Cogagent: A visual language model for gui agents](#).
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#).
- Guolei Huang, Qinzhi Peng, Gan Xu, Yuxuan Lu, and Yongjun Shen. 2025a. [Llavashield: Safeguarding multimodal multi-turn dialogues in vision-language models](#).
- Hao Huang, Shuaihang Yuan, Yu Hao, Congcong Wen, and Yi Fang. 2025b. [A chain-of-thought subspace meta-learning for few-shot image captioning with large vision and language models](#).
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhajit Som, Xia Song, and Furu Wei. 2023. [Language is not all you need: Aligning perception with language models](#).
- Drew A. Hudson and Christopher D. Manning. 2019. [Gqa: A new dataset for real-world visual reasoning and compositional question answering](#).
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. [Scaling up visual and vision-language representation learning with noisy text supervision](#).
- Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. [Hallucination augmented contrastive learning for multimodal large language model](#).
- Di Jin, Eileen Pan, Nassim Oufattolle, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. [What disease does this patient have? a large-scale open domain question answering dataset from medical exams](#).
- Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024. [Mm-soc: Benchmarking multimodal large language models in social media platforms](#).
- Dayal Singh Kalra and Maissam Barkeshli. 2024. [Why warmup the learning rate? underlying mechanisms and improvements](#).
- Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. [Mmbert: Multimodal bert pretraining for improved medical vqa](#).
- Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. [Unifying visual-semantic embeddings with multimodal neural language models](#).
- Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weineng Qian. 2023. [Improving zero-shot visual question answering via large language models with reasoning question prompts](#).
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. [Seed-bench: Benchmarking multimodal llms with generative comprehension](#).
- Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024a. [Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models](#).
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#).
- KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024b. [Videochat: Chat-centric video understanding](#).
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. [Oscar: Object-semantics aligned pre-training for vision-language tasks](#).
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. [Evaluating object hallucination in large vision-language models](#).
- Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024c. [Monkey: Image resolution and text label are important things for large multi-modal models](#).

- Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. 2024. **A comprehensive survey and guide to multimodal large language models in vision-language tasks.**
- Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. 2024. **Moe-llava: Mixture of experts for large vision-language models.**
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries.** In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. **Microsoft coco: Common objects in context.**
- Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. **Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models.**
- Bingchang Liu, Chaoyu Chen, Cong Liao, Zi Gong, Huan Wang, Zhichao Lei, Ming Liang, Dajun Chen, Min Shen, Hailian Zhou, Hang Yu, and Jianguo Li. 2023a. **Mftcoder: Boosting code llms with multitask fine-tuning.**
- Cheng Liu, Chao Wang, and Yan Peng. 2024a. **Imcn: Improved modular co-attention networks for visual question answering.** *Applied Intelligence*, 54(6):5167–5182.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. **Improved baselines with visual instruction tuning.**
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. **Visual instruction tuning.**
- Xuyang Liu, Ting Liu, Siteng Huang, Yi Xin, Yue Hu, Quanjun Yin, Donglin Wang, Yuanyuan Wu, and Honggang Chen. 2025. **M2ist: Multi-modal interactive side-tuning for efficient referring expression comprehension.**
- Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024c. **Mmbench: Is your multi-modal model an all-around player?**
- Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. **A convnet for the 2020s.**
- Ilya Loshchilov and Frank Hutter. 2019. **Decoupled weight decay regularization.**
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. **Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.**
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. **Learn to explain: Multimodal reasoning via thought chains for science question answering.**
- Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. **Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks.**
- Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. **Improving automatic vqa evaluation using large language models.**
- Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. 2024. **Mm1: Methods, analysis insights from multimodal llm pre-training.**
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. **Howto100m: Learning a text-video embedding by watching hundred million narrated video clips.**
- Duy-Kien Nguyen and Takayuki Okatani. 2018. **Multi-task learning of hierarchical vision-language representation.**
- Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. 2025. **M2d-clap: Exploring general-purpose audio-language representations beyond clap.**
- OpenAI. 2023. **Gpt-4v(ision) system card.**
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. **Im2text: Describing images using 1 million captioned photographs.** In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. **Training language models to follow instructions with human feedback.**
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation.** In *Proceedings of the*

- 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Jay Patel. 2020. *Introduction to Common Crawl Datasets*, pages 277–324.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. *Instruction tuning with gpt-4*.
- Nhi Pham and Michael Schott. 2024. *H-pope: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models*.
- Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, Lingpeng Kong, and Tong Zhang. 2023. *Detect what you need via reasoning*.
- Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. *Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models*.
- David M. W. Powers. 2020. *Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation*.
- Amy Pu, Hyung Won Chung, Ankur P. Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. *Learning compact metrics for mt*.
- 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. 2021. *Learning transferable visual models from natural language supervision*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. *Exploring the limits of transfer learning with a unified text-to-text transformer*.
- Shaina Raza, Caesar Saleh, Azib Farooq, Emrul Hasan, Franklin Ogidi, Maximus Powers, Veronica Chattrath, Marcelo Lotif, Karanpal Sekhon, Roya Javadi, Haad Zahid, Anam Zahid, Vahid Reza Khazaie, and Zhenyu Yu. 2025. *Vilbias: Detecting and reasoning about bias in multimodal content*.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. *Object hallucination in image captioning*.
- David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. 2019. *Experience replay for continual learning*.
- Guillermo Ruiz, Tania Ramírez, and Daniela Moctezuma. 2025. *Verscore: Image captioning metric based on v&l transformers, clip, and precision-recall*.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, Si-Wai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gotweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. *Capabilities of gemini models in medicine*.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. *Laion-5b: An open large-scale dataset for training next generation image-text models*.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. *Laion-400m: Open dataset of clip-filtered 400 million image-text pairs*.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. *Bleurt: Learning robust metrics for text generation*.
- Ozan Sener and Vladlen Koltun. 2019. *Multi-task learning as multi-objective optimization*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. *Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning*. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.
- Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. *Towards vqa models that can read*.
- Bharat Singh, Soham De, Yangmuzi Zhang, Thomas Goldstein, and Gavin Taylor. 2015. *Layer-specific adaptive learning rates for deep networks*.
- Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. *Pandagpt: One model to instruction-follow them all*.

- Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. **Eva-clip-18b: Scaling clip to 18 billion parameters.**
- Mingxing Tan and Quoc V. Le. 2020. **Efficientnet: Re-thinking model scaling for convolutional neural networks.**
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwala Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Sagar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madijan Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. **Llama 2: Open foundation and fine-tuned chat models.**
- Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. **Multimodal few-shot learning with frozen language models.**
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. **Attention is all you need.**
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. **Cider: Consensus-based image description evaluation.**
- Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. **Cogvlm: Visual expert for pretrained language models.**
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. **Chain-of-thought prompting elicits reasoning in large language models.**
- Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. 2024. **Clipself: Vision transformer distills itself for open-vocabulary dense prediction.**
- Shuo Xing, Peiran Li, Yuping Wang, Ruizheng Bai, Yueqi Wang, Chan-Wei Hu, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. 2025. **Re-align: Aligning vision language models via retrieval-augmented direct preference optimization.**
- Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. 2021. **Gradual fine-tuning for low-resource domain adaptation.**
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. **Msr-vtt: A large video description dataset for bridging video and language.** In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.
- Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023a. **LvLM-ehub: A comprehensive evaluation benchmark for large vision-language models.**
- Zeyu Xu, Junkang Zhang, Qiang Wang, and Yi Liu. 2025. **E-vrag: Enhancing long video understanding with resource-efficient retrieval augmented generation.**
- Zhiyang Xu, Ying Shen, and Lifu Huang. 2023b. **Multi-instruct: Improving multi-modal zero-shot learning via instruction tuning.**
- Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. **Avqa: A dataset for audio-visual question answering on videos.** In *Proceedings of the 30th ACM International Conference on Multimedia, MM ’22*, page 3480–3491, New York, NY, USA. Association for Computing Machinery.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. **A survey on multimodal large language models.** *National Science Review*, 11(12).
- Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. 2023. **Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip.**
- Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xijie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. 2025. **Osprey: Pixel understanding with visual instruction tuning.**
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. **From recognition to cognition: Visual commonsense reasoning.**
- Biqing Zeng, Zehan Li, and Aladdin Ayesh. 2025. **Neural networks remember more: The power of parameter isolation and combination.**
- Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. 2023. **What matters in training a gpt4-style language model with multimodal inputs?**

Hang Zhang, Xin Li, and Lidong Bing. 2023. [Video-llama: An instruction-tuned audio-visual language model for video understanding.](#)

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Ao-jun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2024a. [Llama-adapter: Efficient fine-tuning of language models with zero-init attention.](#)

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert.](#)

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024b. [Pmc-vqa: Visual instruction tuning for medical visual question answering.](#)

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024c. [Multi-modal chain-of-thought reasoning in language models.](#)

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## A Detailed Evaluation Metrics

Table 1: Evaluation metrics (Section 5), categorized by task.

Metric	Primary Task(s)	Description
<b>Core Language-Vision Evaluation Metrics(Automated)</b>		
BLEU	Image Captioning	Focuses on n-gram precision ( <a href="#">Papineni et al., 2002</a> ).
METEOR	Image Captioning	Incorporates synonym matching and recall ( <a href="#">Banerjee and Lavie, 2005</a> ).
CIDEr	Image Captioning	Emphasizes consensus among human reference captions ( <a href="#">Vedantam et al., 2015</a> ).
ROUGE	Image Captioning	Measures overlap of units (e.g., n-grams or subsequences) ( <a href="#">Lin, 2004</a> ).
SPICE	Image Captioning	Used alongside other n-gram metrics. ( <a href="#">Anderson et al., 2016</a> )
<b>Core Language-Vision Evaluation Metrics (Learned/Multimodal)</b>		
BERTScore	Image Captioning / Generative	Embedding-based metric to better capture semantic similarity and fluency ( <a href="#">Zhang et al., 2020</a> ).
BLEURT	Image Captioning / Generative	Learned metric to better capture semantic equivalence and fluency aligned with human ratings ( <a href="#">Sellam et al., 2020</a> ).
CLIPScore	Image Captioning / Generative	A multimodal embedding-based score that evaluates semantic alignment ( <a href="#">Hessel et al., 2022</a> ).
VCRScore	Image Captioning	A recently proposed learned/trainable metric for caption fluency and accuracy ( <a href="#">Ruiz et al., 2025</a> ).
<b>Retrieval Metrics</b>		
Recall@K (R@1, R@5, R@10)	Crossmodal Retrieval	Measures if the correct item is retrieved within the top K results ( <a href="#">Kiros et al., 2014</a> ).
Median Rank (MedR)	Crossmodal Retrieval	The median rank of the correctly retrieved item ( <a href="#">Kiros et al., 2014</a> ).
Mean Reciprocal Rank (MRR)	Crossmodal Retrieval	The average of the reciprocal ranks of correct items ( <a href="#">Craswell, 2009</a> ).
<b>Classification &amp; VQA Metrics</b>		
Accuracy (Exact/Soft Match)	VQA / Classification	Measured as an exact or soft match against the majority human answers or gold-standard labels ( <a href="#">Mañas et al., 2024</a> ).
F1-score	Multimodal Classification	Standard metric for classification tasks ( <a href="#">Powers, 2020</a> ).

## B Datasets used for pre-training or fine-tuning

Table 2: Datasets used for pre-training or fine-tuning (Section 4)

Dataset	Category	Description
<b>Supervised Vision Pre-training</b>		
ImageNet	Vision Pre-training	Crowdsourced labeled dataset for supervised vision model training ( <a href="#">Deng et al., 2009</a> ).
<b>Web-Scale Pre-training Datasets</b>		
CC-3M / CC-12M	Web Image–Text Pairs	Coarse-grained image–text pairs collected from the web with different cleaning pipelines ( <a href="#">Sharma et al., 2018</a> ; <a href="#">Changpinyo et al., 2021</a> ).
SBU Captions	Web Image–Text Pairs	Flickr-sourced coarse image–text pairs ( <a href="#">Ordonez et al., 2011</a> ).
LAION (all variants)	Large-Scale Web Image–Text Pairs	Large-scale coarse-grained image–text pairs for multimodal pre-training ( <a href="#">Schuhmann et al., 2021, 2022</a> ).
COYO-700M	Web Image–Text Pairs	Large-scale coarse-grained dataset derived from CommonCrawl ( <a href="#">Byeon et al., 2022</a> ).
CommonCrawl	Web Crawl Source	Large web scrape powering many multimodal datasets ( <a href="#">Patel, 2020</a> ).
WebLI	Multilingual Web Image–Text Pairs	Multilingual image–text pairs used by PaLI ( <a href="#">Chen et al., 2023c</a> ).
<b>Curated Image–Text Datasets</b>		
ShareGPT4V	High-Quality Captions	1.2M high-quality, fine-grained, model-generated captions ( <a href="#">Chen et al., 2023b</a> ).
MS-COCO	Curated Vision–Language	≈330K images with 5 captions each; used for image captioning & retrieval ( <a href="#">Lin et al., 2015</a> ).
COCO Captions	Caption References	Human reference captions used for evaluating caption metrics ( <a href="#">Chen et al., 2015</a> ).
Flickr30K	Curated Image–Text	Widely used dataset for crossmodal retrieval ( <a href="#">Plummer et al., 2016</a> ).
<b>Audio &amp; Video Datasets</b>		
AudioSet	Audio Clips	≈2M labeled 10s audio clips across 632 classes ( <a href="#">Gemmeke et al., 2017</a> ).
HowTo100M	Video + Narration	136M instructional video clips with narration ( <a href="#">Miech et al., 2019</a> ).

## C Evaluation Benchmarks

Table 3: Evaluation benchmarks (Section 5), categorized by task type.

Benchmark	Category	Description
<b>Core V-L Benchmarks</b>		
VQA v2.0	Visual QA	Widely used benchmark for answering natural language questions about images (Agrawal et al., 2016).
TextVQA	Text-based QA	Requires models to read and reason about text inside images (Singh et al., 2019).
GQA	VCR and VQA	Tests compositional reasoning over structured scene graphs (Hudson and Manning, 2019).
VCR	Visual Commonsense Reasoning	Evaluates commonsense inference and “why/what next” reasoning (Zellers et al., 2019).
<b>Video &amp; Temporal Benchmarks</b>		
AVQA	Audio-Visual QA	Audio + vision temporal reasoning benchmark (Yang et al., 2022).
UNQA	Video QA	Time-dependent video-based question answering (Cao et al., 2024).
MSR-VTT	Video QA / Retrieval	Benchmark for video-language understanding (Xu et al., 2016).
<b>Generative / Coding Benchmarks</b>		
HumanEval	Code Generation	Standard benchmark for evaluating code-generation abilities (Chen et al., 2021).
<b>Holistic / Meta Evaluation Suites</b>		
MMBench	Holistic Evaluation	Multilingual multi-category evaluation with CircularEval (Liu et al., 2024c).
SEED-Bench	Multimodal QA	24k human-annotated questions across 12 dimensions (Li et al., 2023a).
LLaVA-Bench	Holistic Evaluation	General-purpose MLLM evaluation suite (Liu et al., 2023b).
InstructBLIP Eval	Holistic Evaluation	Benchmark tailored to InstructBLIP-style models (Dai et al., 2023).
LVLM-eHub	Meta Evaluation Suite	All-around evaluation hub for MLLMs (Xu et al., 2023a).
VLMEvalKit	Meta Evaluation Suite	Toolkit for unified multimodal evaluation (Duan et al., 2025).
AbilityLens	Meta Evaluation Suite	A unified benchmark for evaluating six key perception abilities focusing on both accuracy and stability (Chen et al., 2025)
<b>Hallucination Benchmarks</b>		
POPE	Object Hallucination	Measures invented objects in VL model predictions (Li et al., 2023c).
HallusionBench	General Hallucination	Detects invented objects or factual inconsistencies (Guan et al., 2024).
H-POPE	Hierarchical Hallucination	Multi-level hallucination probing benchmark (Pham and Schott, 2024).
CHAIR	Caption Hallucination	Measures hallucinated objects in image captions (Rohrbach et al., 2019).
<b>Safety &amp; Bias</b>		
MM-SOC	Safety / Social Reasoning	Evaluates models on misinformation, hate speech, and social context (Jin et al., 2024).
SHIELDVLM	Toxicity Detection	Detects implicit multimodal toxicity across risk categories (Cui et al., 2025).
LLaVASHield	Safety Alignment	Guards multi-turn multimodal dialogues against malicious intents (Huang et al., 2025a).
ViLBias	Bias Detection	Measures bias in multimodal news interpretation (Raza et al., 2025).
<b>Domain-Specific Benchmarks</b>		
ScienceQA	Domain QA	Multimodal science-domain QA dataset (Lu et al., 2022).
MedQA	Medical QA	Medical-domain QA benchmark used for evaluating MedGemini (Jin et al., 2020).
NEJM Image Challenges	Medical Image QA	Clinical image-based reasoning challenge (Buckley et al., 2024).