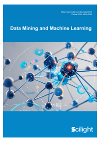
Data Mining and Machine Learning
https://www.sciltp.com/journals/jmlis



Article

A Survey of Multimodal Models on Language and Vision: A Unified Modeling Perspective

Zhongfen Deng [1,*,†], Yibo Wang [1], Yueqing Liang [2], Jiangshu Du [1,†], Yuyao Yang [1,‡], Liancheng Fang [1,‡], Langzhou He [1,‡], Yuwei Han [1,‡], Yuanjie Zhu [1,‡], Chunyu Miao [1,‡], Weizhi Zhang [1], Jiahua Chen [1], Yinghui Li [3], Wenting Zhao [4] and Philip S. Yu [1]

[1] Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607, USA
[2] Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616, USA
[3] Department of Computer Science and Technology, Tsinghua University, Beijing 100190, China
[4] Salesforce Research, Palo Alto, CA 94301, USA
* Correspondence: zdeng21@uic.edu
† Work done prior to Amazon.
‡ These authors contributed equally to this work.

How To Cite: Deng, Z.; Wang, Y.; Liang, Y.; et al. A Survey of Multimodal Models on Language and Vision: A Unified Modeling Perspective. *Data Mining and Machine Learning* **2025**, *1*(1), 100001.

Received: 26 August 2025
Revised: 27 October 2025
Accepted: 28 October 2025
Published: 3 December 2025

Abstract: In recent years, significant progress has been made in developing AI systems capable of processing multimodal data—such as text, image, and videos—to perform complex tasks. With the advent of Large Language Models (LLMs), there has been a surge of interest in building multimodal models based on LLMs. Most current approaches employ a heterogeneous architecture to process text and image separately before bridging them together, leading to a critical bridge bottleneck. Modeling multimodal data such as text and image in a unified manner can help overcome this limitation. Therefore, in this survey, we investigate the current research landscape of multimodality modeling from three perspectives. The first group of multimodal models adopts a heterogeneous architecture to bridge different modality data. The second line of research leverages LLM for multimodality modeling via a unified language modeling objective. The third group represents multimodal data entirely within a single visual representation. The latter two groups can offer a more unified treatment of modalities, helping to alleviate the bridge bottleneck and paving the way for more capable multimodal systems.

Keywords: large language models; multimodal modeling; unified modeling

1. Introduction

Since the introduction of Transformer [1] in 2017, there has been rapid progress in Natural Language Processing (NLP), particularly with the rise of Large Language Models (LLMs) capable of handling a wide range of NLP tasks.

Following the success of transformer-based LLMs in NLP, the multimodality research community has also made significant strides, leading to the development of many large Vision-Language Models. These models represent the mainstream of multimodal LLMs, and typically employ a heterogeneous architecture consisting of a backbone LLM, a vision encoder, and a vision-language adapter. However, Tong et al. [2] identify a bottleneck in the visual encoder within this line of research, noting that simply scaling the visual encoder does not yield performance improvements. In light of this, we aim in this survey to revisit the foundation for modeling the multimodal world. We examine three key research directions, and highlight recent progress to help guide future advancements in the field.

In the past two years, several surveys of multimodal large language models (MLLMs) have been published. Zhang et al. [3] focus on analyzing specific components of MLLMs designed with heterogeneous architecture and categorize those models by function and tool usage. Caffagni et al. [4] also focus on MLLMs with heterogeneous architecture, training methods, and vision-language tasks. Li et al. [5] analyze the generalist and specialist multimodal foundation models. Bai et al. [6] examine MLLMs through a data-centric lens, Qin et al. [7] explore them from a data–model co-development framework, and Jin et al. [8] together with Mai et al. [9] survey the architectural innovations and optimization strategies that underpin efficient and high-performance multimodal language models.

Different from previous surveys that primarily focus on existing models built with an LLM, a visual encoder, and an adapter, our work explores three distinct approaches for multimodal modeling as illustrated in Figure 1. We investigate current research on multimodal modeling from three different perspectives: (1) heterogeneous architectures with separate visual and language encoders, (2) unified multimodal modeling using a language-modeling objective, and (3) unified modeling within a single visual view. Models in the first category typically combine off-the-shelf visual and language models and devise methods to align their feature spaces for multimodal tasks. While these models dominate multimodal research, they face limitations as noted by Tong et al. [2] as mentioned before.

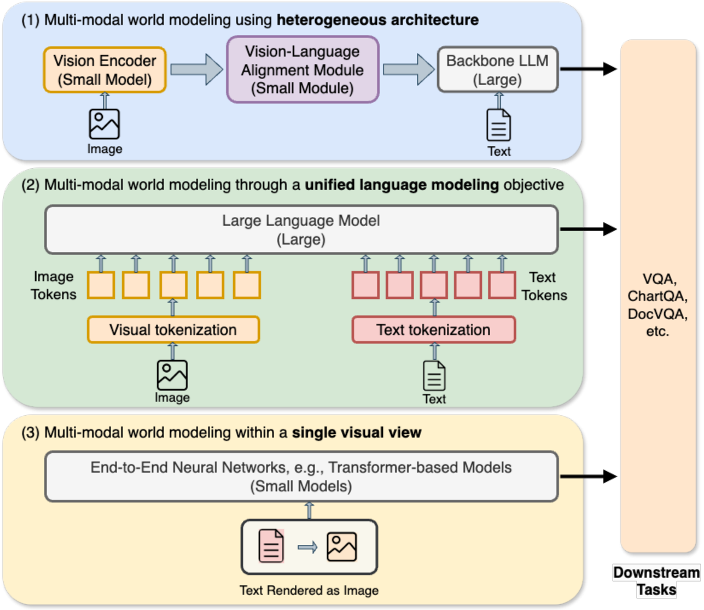

(1) Multi-modal world modeling using heterogeneous architecture
Vision Encoder (Small Model)
Vision-Language Alignment Module (Small Module)
Backbone LLM (Large)
Image
Text
(2) Multi-modal world modeling through a unified language modeling objective
Large Language Model (Large)
Image Tokens
Text Tokens
Visual tokenization
Text tokenization
Image
Text
(3) Multi-modal world modeling within a single visual view
End-to-End Neural Networks, e.g., Transformer-based Models (Small Models)
Text Rendered as Image
VQA, ChartQA, DocVQA, etc.
Downstream Tasks

Figure 1. Comparison of three different ways for multimodal modeling.

Humans perceive the multimodal world through various channels, such as language, vision, and process images and text in an integrated manner. This naturally prompts a meaningful question: Can we model multimodal data within a unified view? Additionally, Huh et al. [10] highlight that visual and text feature spaces will eventually converge. The intuition of how humans perceive the multimodal world, along with the statement by Huh et al. [10], inspired us to explore the unified multimodal modeling perspectives—specifically, unified multimodal modeling based on a language-modeling objective and unified multimodal modeling within a single visual view (In this survey, we focus on two modalities—image and text—since the majority of the research work only covers the image and text, and this taxonomy can be easily extended to broader modalities. Another point is that we also do not cover papers for vision generation tasks). Our work's main contributions are as follows:

- We analyze multimodal modeling research from a broad, high-level perspective, synthesizing diverse architecture designs and unified training objectives into a coherent framework that maps the current landscape of the field.
- We present the first systematic taxonomy of existing multimodal models, classifying them into three distinct categories—heterogeneous architectures, unified language-modeling approaches, and single-view unified models—and examine the key design trade-offs within each category and their real-world application contexts.
- We compare the models of these three approaches and highlight future directions for improving multimodal performance. Notably, the first approach faces challenges, such as limitations in visual encoders, while the other two approaches remain in their early stages. We encourage further exploration of the latter two approaches to enhance the multimodal capabilities of AI in real-world applications.

2. Preliminaries

Humans communicate with each other via natural languages and interact with the world through vision channel, it is natural to build AI agents with both capabilities. Natural Language Processing (NLP) and Computer Vision (CV) are the two fundamental research areas for building such abilities.

2.1. Language Modeling

Language modeling assigns probabilities to word sequences and has a long history in information theory. It is a core part of many NLP tasks, such as speech recognition [11], machine translation [12], and information retrieval [13]. The field has witnessed a significant transformation, particularly with the advent of large-scale neural network models [14]. This evolution can be broadly characterized as progressing through four distinct phases: statistical language models (SLMs), neural language models (NLMs), pre-trained language models (PLMs), and the currently dominant LLMs.

2.1.1. Statistical Language Models

The earliest computational approaches to language modeling were dominated by SLMs, which conceptualize text as a sequence of words and estimate the probability of a sequence by multiplying the probabilities of its constituent words. The most prevalent form of SLMs relies on the Markov assumption [15], leading to n-gram models. These models approximate the probability of a word given its entire history by conditioning only on the preceding n-1 words. The probabilities are estimated based on the frequency of n-grams observed in large text corpora. Despite their simplicity and effectiveness in certain applications, SLMs suffered from a critical limitation: data sparsity [16]. Many plausible word sequences might not appear in the training corpus, leading to zero probability estimates for unseen n-grams.

2.1.2. Neural Language Models

Neural language models emerged as a powerful alternative, directly addressing the data sparsity issue inherent in SLMs [17–19]. The core innovation of NLMs was the concept of distributed representations, where words are mapped to low-dimensional, continuous vectors known as word embeddings. Instead of relying on discrete word counts, NLMs use neural networks—initially feedforward networks and later Recurrent Neural Networks (RNNs) [19]—to process these dense embedding vectors of preceding words and predict the probability distribution of the next word.

2.1.3. Pre-Trained Language Models

The development of pre-trained language models marked a paradigm shift in NLP [20]. Instead of training models from scratch for each specific task, the PLM approach involves two stages: pre-training and fine-tuning [21–24]. In the pre-training stage, a language model is trained on vast amounts of unlabeled text data using self-supervised objectives, such as masked language modeling [22] or next-token prediction [25]. This allows the model to learn general-purpose language representations. Subsequently, the pre-trained model is fine-tuned on smaller, task-specific labeled datasets to adapt its learned knowledge to downstream applications like text classification, question answering, or machine translation.

The advent of the Transformer architecture [1] was pivotal for the success of PLMs. Its self-attention mechanism allows the model to weigh the importance to different words in the input sequence, effectively capturing long-range dependencies [1]. Crucially, self-attention computations can be heavily parallelized, enabling efficient training of much larger models on web-scale datasets compared to sequence processing architectures like RNNs [26]. This led to the development of influential PLMs based on different configurations of the Transformer architecture:

- Encoder-only models (e.g., BERT [22] , RoBERTa [23]): These models primarily use the Transformer encoder stack and are often pre-trained using masked language modeling. They excel at language understanding tasks.
- Decoder-only models (e.g., GPT-1 [25]): These models utilize the Transformer decoder stack, and are typically pre-trained for next-token prediction, making them well-suited for text generation.
- Encoder-decoder models (e.g., BART [27], T5 [28]): These models employ both encoder and decoder stacks, and are often pre-trained with sequence-to-sequence objectives, making them suitable for tasks like summarization and translation.

2.1.4. Large Language Models

LLMs represent the current state of the art in language modeling, primarily characterized by their massive scale, typically involving tens to hundreds of billions of parameters and training on massive text corpora. This significant increase in scale, often guided by scaling laws that predict performance improvements with increased model size, data volume, and compute, has led to models with substantially enhanced language understanding and generation capabilities compared to earlier PLMs [29,30].

The development of LLMs has been marked by the emergence of several prominent model families that have continually pushed the boundaries of scale and capability. These include OpenAI's influential GPT series, which evolved from early models like GPT-1 and GPT-2 to the much larger and more capable GPT-3 [31] and the multimodal GPT-4 [32]. Other major families are Google's PaLM [33], known for its efficient scaling, and Meta's LLaMA [34] series, which has been pivotal in the open-source community. Building on these foundational models, subsequent research has produced models like Alpaca [35] and Vicuna [36], which have been fine-tuned using instruction-following data and human feedback to further enhance their capabilities and better align them with user intent.

The current landscape of LLMs can be categorized based on their level of specialization and functionality. This provides a practical framework for their development and deployment, which can be broken down as follows.

- General-purpose foundational models: These models are the bedrock of the LLM hierarchy, trained on extensive and diverse text corpora to build a broad base of world knowledge. Notable examples in this category include GPT-3 [31], PaLM [33], and LLaMA [34].
- Instruction-tuned and chat models: Representing a significant evolution, these models are refined from foundational models to better align with human interaction. Through techniques such as Reinforcement Learning from Human Feedback (RLHF), they are optimized to follow instructions and sustain dialogue. Prominent models in this group are ChatGPT [37] and Vicuna [36].
- Domain-specific models: Further specialization has led to the development of models that exhibit expert-level capabilities in particular fields. These include models for code generation like Codex [38] and Code Llama [39], in medicine with Med-PaLM [40], and for scientific purposes such as Galactica [41].
- Reasoning-enhanced models: A key frontier in LLM development is the enhancement of their reasoning abilities to address complex, multi-step problems. Models like the DeepSeek [42] series and the latest iterations of the GPT and PaLM families are trained on vast datasets of code and mathematical problems. They employ techniques such as chain-of-thought reasoning to enhance their logical deduction and problem-solving performance on challenging benchmarks.

2.1.5. Non-Transformer-Based Large Language Models

While Transformer models are prevalent, alternative Large Language Model architectures are emerging to address limitations such as quadratic scaling [43]. Mamba [44], a selective State Space Model, offers near-linear scaling for long sequences by dynamically filtering information—differing from Transformers' global attention. xLSTM [45] extends traditional LSTMs with enhanced memory structures and scalability, retaining RNN-like linear complexity for long contexts, in contrast to Transformer's parallel but more computationally intensive attention. Finally, RWKV [43] presents an attention-free model that combines RNN-like efficient inference with Transformer-like parallelizable training, achieving linear complexity through novel mixing mechanisms, thereby completely avoiding the self-attention bottleneck.

2.2. Visual Modeling

Visual modeling empowers AI systems to interpret and understand information from images and videos, a critical capability for interacting with the physical world. Early approaches relied heavily on handcrafted features, such as Scale-Invariant Feature Transform (SIFT) [46] and Histograms of Oriented Gradients (HOG) [47]. These methods involved designing algorithms to extract specific types of information (e.g., edges, corners, textures) from images. While effective for specific, constrained tasks, handcrafted features often struggled with the variability and complexity of real-world visual data due to their limited representational power and inability to adapt to diverse visual patterns [48].

2.2.1. Convolutional Neural Networks

The field of computer vision underwent a dramatic transformation with the advent of deep learning, particularly Convolutional Neural Networks (CNNs). Pioneering work like LeNet [49] laid the groundwork, but the success of AlexNet [50] on the ImageNet challenge in 2012 marked a turning point. Unlike traditional methods, CNNs learn hierarchical feature representations directly from pixel data. They use layers of convolutions—applying learnable filters across the image to detect patterns—followed by pooling layers that reduce dimensionality and introduce spatial invariance. This hierarchical structure allows CNNs to learn simple features in early layers and progressively combine them to represent more complex objects and patterns in deeper layers.

Subsequent research led to increasingly deep and sophisticated CNN architectures, such as VGGNet [51], which demonstrated the benefits of deeper stacks of small convolutional filters, and ResNet [52], which introduced residual connections. These connections allowed gradients to flow more easily during training, enabling the effective training of networks hundreds or even thousands of layers deep, pushing the boundaries of image recognition performance. Despite their success, a potential limitation of standard CNNs lies in their inherent locality bias due to the nature of convolutional filters; capturing long-range dependencies and global context within an image can sometimes be challenging for pure CNN architectures [53].

2.2.2. Vision Transformers

Inspired by the success of Transformers in NLP, the Vision Transformer (ViT) [54] adapted the Transformer architecture for image recognition tasks, offering a way to overcome the locality bias of CNNs. The core idea of ViT is to treat an image as a sequence of patches. The image is divided into a grid of non-overlapping patches, each of which is linearly embedded into a vector. These patch embeddings, along with positional embeddings to retain spatial information, are then fed into a standard Transformer encoder. The self-attention mechanism within the Transformer allows the model to weigh the importance of all pairs of patches when representing a specific patch, thereby enabling the modeling of long-range dependencies and global context across the entire image directly.

Initial Vision Transformers (ViTs) demonstrated competitive performance but often required more data than contemporary CNNs due to their lack of strong inductive biases [55]. Subsequent research has aimed to improve the efficiency, data requirements, and representational power of ViTs through various architectural innovations. These efforts can be broadly categorized into two main themes: the development of hierarchical and efficient architectures that modify the transformer's structure for better feature scaling and performance, and the creation of hybrid architectures that explicitly incorporate convolutional principles to introduce beneficial inductive biases.

A primary research direction has focused on adapting the flat, monolithic structure of the original ViT into more efficient, hierarchical architectures that produce multi-scale features, often without relying on convolutions. For instance, Swin Transformer [56] and Pyramid Vision Transformer (PVT) [57] both implement a progressive shrinking pyramid to generate feature maps at different scales. Swin Transformer achieves this with its novel shifted window attention mechanism, which reduces complexity by confining self-attention to local windows. PVT adopts a spatial-reduction attention (SRA) mechanism to downsample key and value tensors. Multiscale Vision Transformer (MViT) [58] builds a feature pyramid by hierarchically increasing channel capacity while reducing spatial resolution, enabled by a flexible pooling attention operator. Other works refine the tokenization process to better capture local structure; Tokens-to-Token ViT (T2T-ViT) [59] uses a progressive tokenization process to recursively aggregate neighboring tokens, while Transformer-in-Transformer (TNT) [60] introduces a nested transformer architecture to model relationships at both the patch and sub-patch levels. CrossFormer [61] integrates a cross-scale embedding module with long-short distance attention to fuse information across its hierarchy.

To more directly leverage the proven strengths of CNNs, another line of research has developed hybrid models that explicitly incorporate convolutional layers to introduce beneficial inductive biases for locality and scale invariance. ViTAE (Vision Transformer Advanced by Exploring intrinsic IB) [62] and Convolution-enhanced image Transformer (CeiT) [63] both integrate convolutions directly within the transformer structure. ViTAE features a parallel integration of a convolution block alongside the self-attention module in each layer, allowing the model to collaboratively learn local and global features. CeiT employs a convolutional stem for low-level feature tokenization and replaces the standard feed-forward network with a Locally-enhanced Feed-Forward (LeFF) layer that uses depth-wise convolution to enhance local context. In contrast, LeViT [64] is a hybrid architecture optimized for inference speed, which uses a cascade of convolutional layers as a stem to rapidly generate a feature-rich representation before passing it to a hierarchical transformer, combining the strengths of both paradigms.

3. Multimodality Modeling

The preceding sections have detailed the significant advancements in the discrete domains of language and vision modeling, culminating in highly capable unimodal architectures. Building upon these robust foundations, the next research frontier lies in the effective integration of these disparate modalities. The primary motivation for this synthesis is the development of artificial intelligence systems with a more holistic and nuanced understanding of the world, akin to the human ability to process visual and linguistic information simultaneously. The central technical challenge is to develop methods that align representations from these heterogeneous data sources and bridge the semantic gap between symbolic language and digital visual data. Accordingly, the following section provides a systematic review and categorization of the dominant architectural paradigms designed to achieve this multimodal fusion.

We classify existing approaches into three main categories as shown in Figure 2 (the technical details of some representative models in the three categories can be found in Table 1): (i) Heterogeneous Architectures (Section 3.1) which process visual and language inputs using separate encoders and integrate them via an adapter module; (ii) Multimodal Modeling through a Unified Language-Modeling Objective (Section 3.2), which unifies multimodal data by converting different modalities into discrete tokens and jointly training with language modeling objectives; and (iii) Unified Multimodal Modeling within a Single Visual View (Section 3.3). Previous surveys [3,4] have primarily focused on heterogeneous architectures. In contrast, this work provides an overview of heterogeneous architectures, while placing greater emphasis on the latter two categories.

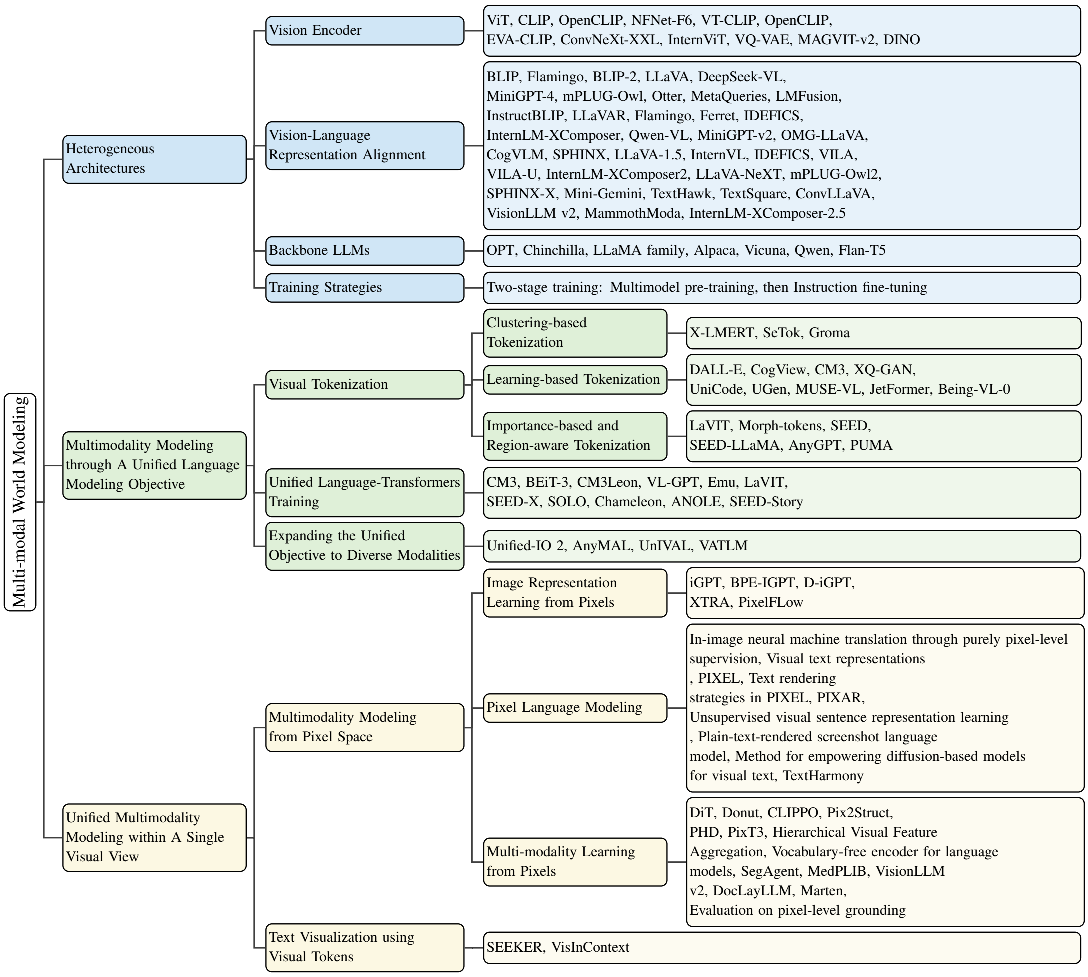

Multi-modal World Modeling
Heterogeneous Architectures
Vision Encoder
ViT, CLIP, OpenCLIP, NFNet-F6, VT-CLIP, OpenCLIP, EVA-CLIP, ConvNeXt-XXL, InternViT, VQ-VAE, MAGVIT-v2, DINO
Vision-Language Representation Alignment
BLIP, Flamingo, BLIP-2, LLaVA, DeepSeek-VL, MiniGPT-4, mPLUG-Owl, Otter, MetaQueries, LMFusion, InstructBLIP, LLaVAR, Flamingo, Ferret, IDEFICS, InternLM-XComposer, Qwen-VL, MiniGPT-v2, OMG-LLaVA, CogVLM, SPHINX, LLaVA-1.5, InternVL, IDEFICS, VILA, VILA-U, InternLM-XComposer2, LLaVA-NeXT, mPLUG-Owl2, SPHINX-X, Mini-Gemini, TextHawk, TextSquare, ConvLLaVA, VisionLLM v2, MammothModa, InternLM-XComposer-2.5
Backbone LLMs
OPT, Chinchilla, LLaMA family, Alpaca, Vicuna, Qwen, Flan-T5
Training Strategies
Two-stage training: Multimodel pre-training, then Instruction fine-tuning
Multimodality Modeling through A Unified Language Modeling Objective
Visual Tokenization
Clustering-based Tokenization
X-LMERT, SeTok, Groma
Learning-based Tokenization
DALL-E, CogView, CM3, XQ-GAN, UniCode, UGen, MUSE-VL, JetFormer, Being-VL-0
Importance-based and Region-aware Tokenization
LaVIT, Morph-tokens, SEED, SEED-LLaMA, AnyGPT, PUMA
Unified Language-Transformers Training
CM3, BEiT-3, CM3Leon, VL-GPT, Emu, LaVIT, SEED-X, SOLO, Chameleon, ANOLE, SEED-Story
Expanding the Unified Objective to Diverse Modalities
Unified-IO 2, AnyMAL, UnIVAL, VATLM
Unified Multimodality Modeling within A Single Visual View
Multimodality Modeling from Pixel Space
Image Representation Learning from Pixels
iGPT, BPE-IGPT, D-iGPT, XTRA, PixelFLow
Pixel Language Modeling
In-image neural machine translation through purely pixel-level supervision, Visual text representations , PIXEL, Text rendering strategies in PIXEL, PIXAR, Unsupervised visual sentence representation learning , Plain-text-rendered screenshot language model, Method for empowering diffusion-based models for visual text, TextHarmony
Multi-modality Learning from Pixels
DiT, Donut, CLIPPO, Pix2Struct, PHD, PixT3, Hierarchical Visual Feature Aggregation, Vocabulary-free encoder for language models, SegAgent, MedPLIB, VisionLLM v2, DocLayLLM, Marten, Evaluation on pixel-level grounding
Text Visualization using Visual Tokens
SEEKER, VisInContext

Figure 2. Overview of multimodal modeling methods: ViT [54], CLIP [65], OpenCLIP [66], NFNet-F6 [67], VT-CLIP [68], OpenCLIP [66], EVA-CLIP [69], ConvNeXt-XXL [70], InternViT [71], VQ-VAE [72], MAGVIT-v2 [73], DINO [74], BLIP [75], Flamingo [76], BLIP-2 [77], LLaVA [78,79], DeepSeek-VL [80,81], MiniGPT-4 [82], mPLUG-Owl [83], Otter [84], MetaQueries [85], LMFusion [86], InstructBLIP [87], LLaVAR [88], Flamingo [76], Ferret [89], IDEFICS [90], InternLM-XComposer [91], Qwen-VL [92], MiniGPT-v2 [93], OMG-LLaVA [94], CogVLM [95], SPHINX [96], LLaVA-1.5 [79], InternVL [71], IDEFICS [90], VILA [97], VILA-U [98] InternLM-XComposer2 [99], LLaVA-NeXT [100], mPLUG-Owl2 [101], SPHINX-X [102], Mini-Gemini [103], TextHawk [104], TextSquare [105], ConvLLaVA [106], VisionLLM v2 [107], MammothModa [108], InternLM-XComposer-2.5 [109], OPT [110], Chinchilla [111], LLaMA family [34, 112], Alpaca [35], Vicuna [36], Qwen [113], Flan-T5 [114], X-LMERT [115], SeTok [116], Groma [117], DALL-E [118], CogView [119], CM3 [120], XQ-GAN [121], UniCode [122], UGen [123], MUSE-VL [124], JetFormer [125], Being-VL-0 [126], LaVIT [127], Morph-tokens [128], SEED [129], SEED-LLaMA [130], AnyGPT [131], PUMA [132], CM3 [120], BEiT-3 [133], CM3Leon [134], VL-GPT [135], Emu [136], LaVIT [127], SEED-X [137], SOLO [138], Chameleon [139], ANOLE [140], SEED-Story [141], Unified-IO 2 [142], AnyMAL [143], UnIVAL [144], VATLM [145], iGPT [146], BPE-IGPT [147], D-iGPT [148], XTRA [149], PixelFLow [150], In-image neural machine translation through purely pixel-level supervision [151], Visual text representations[152], PIXEL [153], Text rendering strategies in PIXEL [154], PIXAR [155], Unsupervised visual sentence representation learning[156], Plain-text-rendered screenshot language model [157], Method for empowering diffusion-based models for visual text [158], TextHarmony [159], DiT [160], Donut [161], CLIPPO [162], Pix2Struct [163], PHD [164], PixT3 [165], Hierarchical Visual Feature Aggregation [166], Vocabulary-free encoder for language models [167], SegAgent [168], MedPLIB [169], VisionLLM v2 [107], DocLayLLM [170], Marten [171], Evaluation on pixel-level grounding [172], SEEKER [173], VisInContext [174].

Table 1. Details of some representative models in three categories for multimodal modeling in recent years. For details of all the models, refer to Appendix A.

Category	Model	Model Architecture	Model Size	Main Tasks/Capabilities
Heterogeneous Architecture	Flamingo [76]	NFNet-F6 → Perceiver Resampler → Gated Cross-Attention + Dense → Chinchilla-3B/9B/80B	3B, 9B, 80B	Few-shot learning, VQA, image captioning, visual dialogues, video understanding, open-ended text generation
	LLaVA [78]	CLIP ViT-L/14 → Linear Projector → Vicuna-7B/13B	7B, 13B	Visual QA, image captioning, multimodal conversations, Science QA, general-purpose vision-language tasks
	InstructBLIP [87]	ViT-G/14 → Q-Former w/ Linear Projector → Flan-T5/Vicuna-13B	4B–13B	VQA, image captioning, science QA, visual dialogues, knowledge-grounded image description
	Qwen-VL [92]	OpenCLIP ViT-bigG → Q-Former → Qwen-7B	9.6B	VQA, image captioning, text reading, visual dialogues, OCR, multilingual support
	IDEFICS [90]	OpenCLIP → Cross-attention → LLaMA-65B	9B, 80B	VQA, image captioning, visual dialogues, OCR
	SPHINX-X [102]	DINOv2 + CLIP-ConvNeXt → MoE → LLaMA2-13B/Mixtral-8 × 7B	1.1B–8 × 7B	OCR, VQA, object detection, image captioning, visual programming, bilingual support
	Mini-Gemini [103]	Dual Vision Encoder → Guided Generation → VLM	2B-34B	VQA, image-text generation, vision-language tasks
	VisionLLM v2 [107]	CLIP-L/14 → Super Link → Vicuna-7B + Task-specific decoders	7B	Vision-Language tasks, VQA, image generation, image editing, pose estimation, object detection
Unified Language View	BEiT-3 [133]	multiway transformer	1.9B	Vision, Vision-Language benchmarks
	LaVIT [127]	ViT → Cross-attention → LLaMA-7B	7B	multimodal understanding and generation
	AnyGPT [131]	SEED+SpeechTokenizer+Encodec → LLaMA-2 7B	7B	image, speech, music understanding and generation
	Morph-tokens [128]	morph-tokenizer+text tokenizer → MLLM → decoders	7B	multimodal comprehension and generation
	Emu2 [175]	EVA-02-CLIP-E-plus → LLaMA-33B → SDXL	37B	in context multimodal tasks
	Chameleon [139]	VQ-SEG+BPE → transformers	7B, 34B	text understanding, text generation, and multimodal comprehension
	ANOLE [140]	VQ-SEG+BPE → transformers	7B	multimodal comprehension and generation
Unified Visual View	CLIPPO [162]	ViT-B/16, ViT-L/16	93M, 316M	Vision and language capabilities: VQA, GLUE; Multilingual capabilities
	PHD [164]	ViT-MAE	86M	Visual SQuAD, Historical QA, GLUE
	PixT3 [165]	Pix2Struct-Base	282M	Visual data-to-text generation: Logic2Text, ToTTo
	PIXAR [155]	Decoder-only Transformer with 12 layers	85M (classification), 113M (generation)	Discriminative (GLUE) + Generative (QA) tasks
	Visual sentence learning [156]	MAE(ViT encoder + lightweight decoder)	86M	Natural language semantics learning (e.g., STS)
	SEEKER [173]	SigLIP-L & SAM-B+DeepSeek-VL [80]	1.3B/7B	Multimodal Understanding, long-form multi-image input, long-form text output
	VisInContext [174]	Flamingo	1.4B/7B/70B	Long-context multimodal understanding & sequential multimodal retrieval

3.1. Heterogeneous Architectures

Heterogeneous architectures for MLLMs have been the mainstream in recent years. These models typically consist of three main components: a vision encoder, a vision-language adaptor for representation alignment, and a backbone LLM. Given a trained LLM and vision encoder, heterogeneous architectures utilize the vision-language adapter to bridge the gap in representations between text and vision input, enabling better fusion in the following transformer blocks. We will introduce key components of heterogeneous architectures.

3.1.1. Vision Encoder

The vision encoder plays a pivotal role in visual instruction fine-tuning, responsible for extracting salient features from visual inputs. Foundational to this is the ViT [54], which adapted the transformer architecture for image processing by embedding fixed-size patches and feeding them into a transformer encoder. A notable variant, InternViT [71], has been specifically optimized for enhanced integration with Large Language Models (LLMs). Another cornerstone, CLIP (Contrastive Language-Image Pre-training) [65], typically employs a vision encoder based on ResNet or ViT, pre-trained on an extensive dataset of 400 million image-text pairs. This has spurred the development of derivatives such as VT-CLIP [68], OpenCLIP [66], and EVA-CLIP [69], all of which offer highly transferable visual representations. Furthermore, architectures like NFNets-F6 [67] have demonstrated state-of-the-art (SoTA) performance without relying on batch normalization, and their efficiency has led to adoption in influential models such as Flamingo [76].

Recent instruction-tuned Vision-Language Models (VLMs) exhibit a significant diversity in their vision encoder design choices. A prevalent strategy involves adopting CLIP-style vision transformers (e.g., ViT-L/14), as demonstrated in models like MetaMorph [176], Janus-Pro [177], DeepSeek-VL2 [81], and Qwen2-VL [178]. Commonly, these encoders are kept frozen during instruction tuning, with lightweight projection layers introduced to align their patch embeddings with the language model's input space. Innovations within this paradigm include dynamic resolution support [178,179] or tiling mechanisms [180,181], enabling the processing of high-resolution visual inputs without necessitating alterations to the core ViT backbone.

A contrasting yet significant trend is the utilization of discrete visual tokenizers, such as VQ-VAE [72] or MAGVIT-v2 [73]. These methods compress images into sequences of latent codes, enabling a unified modeling approach for both visual understanding and generation. Models such as Janus [182], ANOLE [140], and Show-o [183] leverage this technique to facilitate autoregressive image generation directly from the language model. VILA-U [98] further advances this by incorporating residual quantization, which better aligns discrete image tokens with the language modality during the pretraining phase.

For tasks demanding dense grounding and sophisticated spatial reasoning, models including Ferret [89] and OMG-LLaVA [94] employ region-aware encoders. These encoders pool features from arbitrary image regions and can optionally integrate coordinate embeddings, making them particularly well-suited for fine-grained referring expression comprehension and segmentation.

Another distinct line of research focuses on fusing multiple vision encoders to harness their complementary strengths. For instance, Cambrian-1 aggregates patch tokens from diverse encoders like CLIP [184], DINOv2 [74], and ConvNeXt-XXL [70], subsequently merging their outputs using cross-attention mechanisms. Similarly, models such as Eagle [185] and InternVL [181] explore hybrid backbones and multi-resolution inputs to effectively capture both global semantic information and fine-grained visual details.

3.1.2. Vision-Language Representation Alignment

Aligning representations across vision and language modalities remains a central challenge for multimodal large language models (MLLMs), particularly under the visual instruction-tuning setting. Current approaches can be broadly categorized into three families. One prominent approach relies on shallow projection-based adapters that map the visual features into the language embedding space, enabling seamless integration with frozen language models. For instance, LLaVA [78] employs a lightweight linear projection layer as the vision-language connector. Specifically, the image is first encoded into a sequence of patch features using a pretrained CLIP-ViT encoder. These features are then passed through a single-layer linear projection that maps them into the same dimensional space as the language model's word embeddings. The resulting projected visual tokens are prepended as soft prompts to the LLM's input, allowing the LLM to interpret them as part of the dialogue context. This minimal yet effective design enables low training cost and efficient instruction tuning. The projection-based approach is widely adopted in follow-up models such as LLaVA-1.5 [79], MiniGPT-4 [82], and DeepSeek-VL series [80,81]. For instance, MiniGPT-4 [82] uses a linear projection to align CLIP image embeddings with the Vicuna LLM embedding space, feeding the transformed visual tokens as soft prompts into the LLM. DeepSeek-VL series employs a two-layer hybrid MLP to bridge the vision encoder and the LLM. Despite their architectural simplicity and compatibility with frozen language backbones, such shallow adapters often struggle with fine-grained semantic alignment and complex interleaved image-text reasoning. These limitations have motivated the development of deeper, more expressive vision-language alignment mechanisms in subsequent models.

The second family introduces transformer-based visual adapters that serve as intermediate alignment modules between the vision encoder and the language model. These adapters process visual features into a language-compatible representation before being passed into the LLM, enabling more expressive vision-language fusion than shallow projection layers. A representative design is the Q-Former [77] in BLIP-2, which introduces a set of learnable query tokens that attend to image embeddings from a frozen vision encoder via cross-attention. The resulting query outputs—now condensed, task-aware visual representations—are then projected and fed as input tokens to a frozen language model. This design decouples vision encoding and language generation, enhancing modularity and training efficiency. This paradigm is further adopted and extended by models such as mPLUG-Owl [83], InternVL [71,181], and Qwen-VL [92,178,179]. For example, Qwen-VL employs a single-layer cross-attention module as the visual adapter, where trainable queries interact with visual features (from a ViT encoder) serving as keys and values. To preserve spatial information, 2D absolute positional encodings are injected into the query-key pairs before computing attention. The output of this adapter is prepended to the language tokens and fed into the LLM. Recent advancements, such as Meta-Queries [85] and LMFusion [186], further extend this design by inserting trainable query vectors or vision-aware feed-forward modules directly into multiple transformer layers of the language model. These hybrid designs enable layer-wise adaptation, tighter modality coupling, and improved sample efficiency—particularly beneficial when the vision input contains complex or compositional elements.

The third family embeds cross-modal fusion directly into the language model's transformer layers, allowing language tokens to dynamically attend to visual features throughout generation. Unlike the previous family, which uses an adapter to produce fixed visual prompts, this approach integrates vision-language fusion more deeply and pervasively within the language model. This architecture was pioneered by Flamingo [76], which inserts gated cross-attention blocks into a frozen LLM (e.g., Chinchilla or Gopher) at selected transformer layers. Each block allows textual tokens to attend to frozen vision embeddings, mediated by a learnable gating mechanism that controls the flow of visual information during generation. CogVLM [95] generalizes this idea by integrating vision-language fusion modules across all layers of the LLM backbone. Specifically, it injects a cross-attention layer after each self-attention block, allowing dynamic interleaving of vision and language processing at every stage. IDEFICS [90] adopts a similar architecture based on the OPT-IML [187] language model, using multimodal fusion blocks to enable direct grounding. InternLM-XComposer [91] and its successors (e.g., OMG-LLaVA [94] and Ferret [89]) further enhance this design with dedicated grounding modules. These modules specialize in spatial reasoning and referring expression comprehension by incorporating additional supervision (e.g., object-level bounding boxes or segmentation maps) into the attention pathways. For instance, Ferret enhances grounded response generation by adding both spatial-aware attention maps and hierarchical reasoning heads. Overall, this family emphasizes fine-grained, context-aware fusion by treating vision as a first-class citizen in every layer of the language model. This enables deeper compositional understanding and makes the model more capable in tasks such as visual question answering, grounding, and multimodal dialogue.

3.1.3. Backbone LLMs

The backbone Large Language Model is the core component of a Multimodal Large Language Model (MLLM), responsible for processing, reasoning, and generating language in response to multimodal inputs. Its architecture, capacity, and pretraining scheme critically influence the MLLM's capability in alignment, generalization, and instruction following. Modern MLLMs typically adopt decoder-only transformer architectures, which simplify autoregressive generation and integrate seamlessly with token-based multimodal prompts.

A dominant family of backbones is the LLaMA series from Meta AI. The original LLaMA [34] and its successor LLaMA 2 [112] are decoder-only transformers trained on trillions of tokens with a causal language-modeling objective. Their architecture follows the GPT-style design: multi-head self-attention, feed-forward layers, rotary positional embeddings (RoPE), and RMSNorm. Their open-source release in various sizes (7B, 13B, 65B) has made them popular in many MLLMs due to their strong performance and accessibility. Derivatives such as Alpaca [35] and Vicuna [36] further instruction-tune LLaMA with curated instruction datasets or human-like dialogues, yielding improved instruction-following and conversational ability. These models are often used as the backbone in systems like MiniGPT-4 [82], LLaVA [78], InternVL [71], and Qwen-VL [92].

Beyond LLaMA, several alternative backbones have been explored for their unique architectural or training advantages. The OPT series [110], also from Meta, offers decoder-only transformers trained with a transparent and reproducible methodology. Although less performant than LLaMA at the same scale, OPT has been adopted in early MLLMs such as Flamingo [76] and IDEFICS [90], owing to its public release and ease of integration.

Chinchilla [111] introduced a key insight into compute-optimal training, showing that medium-sized models (e.g., 10–20 B) can outperform much larger ones when trained on sufficiently large datasets. Its architectural foundation resembles GPT-3 but with better scaling laws and data efficiency, influencing later MLLMs that prioritize training dynamics and performance-per-flop.

From a multilingual and instruction-tuned perspective, Qwen [113] and its successors—Qwen2 [178], Qwen3 [188]—offer powerful LLMs pretrained on mixed-language corpora, featuring enhanced tokenizer design and extended context lengths (up to 64K in Qwen2). Architecturally, Qwen adopts a GPT-style decoder with SwiGLU [189] activation, rotary position embeddings, and optimized attention scaling. These models serve as backbones in the Qwen-VL multimodal series, demonstrating strong multilingual capabilities and long-context instruction-following performance.

Another influential backbone is Flan-T5 [114], which differs from the others by employing an encoder-decoder architecture based on the original T5. Flan-T5 is trained on a broad collection of instruction-tuning tasks using a mixture of supervised and zero-shot objectives. Its encoder-decoder setup makes it particularly suitable for tasks involving structured input-output mappings, such as captioning or question answering. While less common in dialogue-based MLLMs, Flan-T5 serves as the foundation in some vision-language tasks and zero-shot NLP evaluations.

3.1.4. Training Strategies

MLLM training typically comprises two major stages: multimodal pretraining and instruction tuning, where the latter includes both supervised fine-tuning (SFT) and reinforcement learning (RL) approaches. Throughout this process, the backbone language model is assumed to be already pretrained on large-scale text corpora.

In the multimodal pretraining stage, the goal is to align the visual encoder with the frozen language model by introducing trainable components such as projection layers, cross-attention modules, or query-based adapters. These components are optimized using paired image-text data to bridge the modality gap, enabling the language model to effectively interpret visual representations. During this phase, the language model is kept frozen to preserve its foundational capabilities, while only the newly introduced multimodal components are trained.

The subsequent instruction tuning stage enhances the model's ability to follow complex, multimodal instructions presented in a conversational format. This stage can be further categorized into two subtypes: Supervised Fine-Tuning (SFT): The model is trained on curated instruction-following datasets via standard teacher-forcing objectives. This phase improves the model's ability to generate accurate and grounded responses. Reinforcement Learning (RL): Approaches such as PPO [190], GRPO [191], DAPO [192] are used to optimize the model based on feedback signals like response quality, helpfulness, or adherence to user intent. RL-based tuning can further refine alignment and reasoning beyond what is achievable with supervised signals alone.

SFT

Depending on the model design, instruction tuning may update only the multimodal adapters or partially unfreeze the vision encoder to incorporate task-specific visual semantics, while the language model typically remains frozen to preserve linguistic fluency. Together, these stages progressively adapt a pretrained LLM into a capable multimodal conversational agent with strong visual understanding and instruction-following capabilities.

Recent MLLMs differ significantly in how they train multimodal alignment and instruction-following capabilities. Broadly, training strategies can be categorized into three schemes based on the number of stages and the scope of component updates, as summarized below (see the training strategies of representative MLLMs in Table 2):

1. Two-stage training with pre-alignment and instruction tuning. A more common and effective approach involves two stages: first aligning vision and language modalities using image-caption pairs, followed by instruction-tuning on conversational multimodal tasks. In this setting, the LLM is typically frozen in both stages, while the adapter and, optionally, parts of the vision encoder are updated. MiniGPT-4 [82] first performs vision-language pre-alignment using paired image-text data, then tunes the model on multimodal dialogue. LLaVA [78] follows a similar design, training a linear projection layer during both stages. mPLUG-Owl [83] improves flexibility by partially unfreezing the vision encoder during instruction-tuning to refine visual representations.
2. Full multistage joint tuning. Some models pursue a more integrated strategy, combining contrastive pretraining, instruction tuning, and full model fine-tuning. Janus [182] adopts a two-stage scheme: in the first stage, it trains the visual encoder and a projection head using contrastive learning while keeping the LLM frozen; in the

second stage, it jointly fine-tunes the entire model—including the vision encoder, adapter, and LLM—using instruction datasets. Similarly, SPHINX-X [102] performs full-parameter training from the outset, updating all components—vision encoder, adapter, and LLM—on multimodal instruction data. These strategies offer strong performance on vision-language reasoning tasks but require significantly more compute and data.

3. Extended instruction-tuning pipelines. Some models refine the two-stage paradigm by adding intermediate phases or task-specific supervision. Qwen-VL [92] begins with image-caption alignment, followed by instruction tuning on visual question answering and OCR data. IDEFICS [90] and InternLM-XComposer [91] incorporate grounding-specific instruction datasets to improve spatial and object-level reasoning. These extended pipelines improve generalization and multimodal understanding in complex visual environments.

In practice, the choice of training scheme reflects trade-offs between computational efficiency, alignment quality, and the desired level of multimodal reasoning. Lightweight single-stage tuning is suitable for scalable deployment, while full-model tuning and multistage pipelines offer stronger task performance, particularly in compositional reasoning and grounding-intensive scenarios.

Table 2. Comparison of training strategies across representative MLLMs. We categorize models by the number of training stages and the update scope for each component: language model (LLM), vision encoder (VE), and multimodal adapter (Adapter).

Model	Stages	LLM	VE	Adapter
LLaMA-Adapter [193]	1	Frozen	Frozen	Trainable
Flamingo [76]	1	Frozen	Frozen	Trainable
SPHINX-X [102]	1	Tuned	Tuned	Tuned
LLaVA [78]	2	Frozen	Frozen	Trainable
MiniGPT-4 [82]	2	Frozen	Frozen	Trainable
mPLUG-Owl [83]	2	Frozen	Partial	Trainable
Janus [182]	3	Tuned	Tuned	Tuned
Qwen-VL [92]	3	Frozen	Partial	Trainable
IDEFICS [90]	3	Tuned	Tuned	Tuned
InternLM-X [91]	3	Tuned	Tuned	Tuned

RL-Based Fine-Tuning

A growing line of work augments traditional supervised instruction tuning with reinforcement learning (RL) to directly optimize multimodal models for reasoning performance and generalization. Inspired by successes in text-only LLMs (e.g., DeepSeek R1 [42]), these methods adapt rule-based RL paradigms to the vision-language setting, enabling models to improve through trial-and-error based on task-specific reward signals. This trend introduces a new axis in the MLLM training design space: rather than solely maximizing the likelihood over instruction data, models are guided by structured rewards tied to visual reasoning accuracy, alignment quality, or behavioral outcomes in interactive environments. Most of these models fine-tune variants of the Qwen [92,179,194] series as their base language model, leveraging its strong general-purpose capabilities. They commonly adopt Group Relative Policy Optimization (GRPO) [191], a reinforcement learning algorithm that updates model policies based on group-wise relative comparisons of sampled outputs, rather than relying on absolute reward magnitudes. GRPO enhances stability and sample efficiency by ranking generations within mini-batches and assigning gradients based on their relative performance, making it well-suited for the high-variance reward landscapes in vision-language tasks.

VLM-R1 [195] extends the R1-style GRPO framework to vision-language models, applying it to tasks such as referring expression comprehension (REC) and open-vocabulary detection. By freezing the core LLM and vision encoder while tuning a lightweight projection head, VLM-R1 stabilizes RL training and demonstrates strong generalization to out-of-domain examples. R1-Zero [196] explores the phenomenon of reward hacking in object detection tasks and uncovers the “aha moment” in visual reasoning where the model’s attention sharpens abruptly during RL fine-tuning.

Other models like Vision-R1 [197] and LMM-R1 [198] focus on enhancing multimodal mathematical reasoning. Vision-R1 adopts a two-phase RL pipeline: it first bootstraps the model with chain-of-thought style supervised fine-tuning, then applies rule-based RL to refine multimodal reasoning behaviors, leading to state-of-the-art performance on MathVista and similar benchmarks. Similarly, LMM-R1 shows that even mid-sized MLLMs (3B) can achieve strong visual reasoning capabilities through a combination of curriculum RL and rule-based reward optimization.

G1 [199] pushes this paradigm further by introducing a suite of visual games (e.g., 2048, matching puzzles) and training MLLMs via RL in interactive multimodal environments. G1 reveals that perception and reasoning can bootstrap each other, and that reinforcement learning, when initialized from a perceptually grounded model, leads to robust and generalizable agents. R1-Onevision [200] extends this idea by reformulating visual tasks into structured text-based reasoning problems, enabling LLM-style learning on multimodal inputs and yielding strong results on educational and formal reasoning datasets.

Together, these RL-augmented training methods shift MLLM development toward explicit reasoning incentives, enhancing generalization, stability, and task alignment. While they often require careful reward design and remain compute-intensive, they represent a promising complement to conventional instruction tuning—particularly for compositional reasoning, spatial understanding, and interactive decision-making.

3.2. Multimodality Modeling through a Unified Language-Modeling Objective

While heterogeneous architectures effectively bridge the representation gap between vision and language modalities using separate encoders and adapters, they inherently face limitations such as alignment bottlenecks and integration complexity. To address these challenges, the research discussed in this section adopts a unified multimodal modeling approach from a language-modeling perspective. The core idea is to transform diverse modalities into a common representational space—analogous to words or tokens in natural language—thus enabling large language models (LLMs) to efficiently perform multimodal tasks. Initially, various tokenization strategies are employed to discretize visual inputs and potentially other modalities into token sequences. These token sequences are then processed by language models pre-trained with unified language-modeling objectives. This design inherently unlocks powerful generative abilities—allowing the model to weave together information from different modalities and carry out coherent, cross-modal reasoning over interleaved inputs. By bringing every data stream under one training objective and architecture, these models aim not merely to replicate isolated tasks but to approximate the richness and fluid adaptability of human multimodal perception and inference in real-world settings.

3.2.1. Visual Tokenization

Vision involves continuous, regression-based features. Inspired by the success of transformers in NLP, visual tokenization aims to bridge this gap by breaking images into discrete tokens representing meaningful parts, segments, or concepts within the visual data. This approach aligns visual data with established NLP methods, enabling better integration of vision and text in multimodal systems and allowing LLMs to ”read” images. Various methods have been developed to convert high-dimensional continuous visual data to a structured set of discrete tokens, each capturing different aspects of the visual input and catering to different model architectures and task requirements.

Clustering-based tokenization transforms visual data into discrete tokens by grouping similar regions or features of an image using clustering algorithms rather than relying on predefined grids or patches. Each cluster serves as a token representing a part of the image, enabling more semantically relevant tokenization. X-LMERT [115], for example, constructs a visual vocabulary through K-means clustering on object region features and then approximates target visual features using a nearest-neighbor search from this vocabulary during training. This approach allows the model to learn a fixed set of visual ”words” capable of representing diverse image content.

Furthering the idea of semantic relevance, the Semantic-Equivalent Tokenizer (SeTok) [116] automatically groups visual features from input images, determining the number of tokens based on image complexity through a dynamic clustering mechanism rather than a fixed vocabulary size. This adaptability allows for a more nuanced representation that varies with the content of the image. Recent work continues to explore the concept of semantic equivalence in tokenization for multimodal LLMs, aiming to ensure that visual tokens capture the same semantic meaning as their corresponding textual concepts, thereby improving alignment and understanding across modalities. Another approach, Groma [117], focuses on localized visual tokenization specifically for grounding MLLMs. By generating visual tokens that correspond to specific image regions identified through a localization module, Groma aims to enhance the model's ability to connect textual descriptions to their visual counterparts with greater precision, which is crucial for tasks requiring fine-grained understanding and grounding.

Learning-based tokenization relies on neural networks, particularly autoencoder-like structures, to adaptively extract meaningful, compact, and task-relevant tokens directly from the input image data. These methods learn a discrete latent space, often with a codebook, to represent visual information. CogView [119] uses a vector quantized variational autoencoder (VQ-VAE) to learn such a discrete latent space by encoding visual inputs with a codebook of discrete tokens, enabling high-quality text-to-image generation. Similarly, CM3 [120] utilizes vector quantized generative adversarial networks (VQGAN), which builds on the principles of VQ-VAE but introduces adversarial learning through GANs, to tokenize images into sequences of discrete codes for its masked multimodal modeling approach. Unlike VQ-VAE, which often uses a deterministic nearest-neighbor search for quantization, discrete variational autoencoders (dVAE) employ probabilistic methods to learn the distribution of discrete latent variables. In DALL-E [118], each image is first split into a fixed 32 × 32 grid and then each grid cell is encoded by a learned dVAE tokenizer. This design marries simplicity—the uniform grid makes implementation straightforward—with expressiveness, since the dVAE's learned codebook produces compact, content-aware patch embeddings that faithfully capture semantically rich image details.

Building on these foundational techniques, several recent models have advanced learning-based tokenization. XQ-GAN [121] provides an open-source image tokenization framework designed for autoregressive generation, offering robust and efficient image-to-token conversion. It aims to standardize and improve the quality of visual tokens for generative MLLMs. The pursuit of a UniCode [122] framework focuses on learning a unified codebook that can represent multiple modalities, including vision, within a single MLLM. This shared vocabulary simplifies the architecture and potentially improves cross-modal understanding by mapping diverse inputs into a common discrete space. UGen [123] introduces a unified autoregressive multimodal model featuring progressive vocabulary learning. This approach may involve dynamically expanding or refining the visual codebook during training to better capture the nuances of visual data as the model learns. MUSE-VL [124] models a unified VLM through semantic discrete encoding, emphasizing the importance of learned tokens that carry rich semantic meaning, facilitating better alignment between visual and textual modalities for both understanding and generation tasks. JetFormer [125] addresses the challenge of tokenizing raw image inputs for autoregressive models by introducing a unified generative framework that processes pixel arrays alongside text. It first partitions an image into patches, maps each patch through a learned codebook into discrete visual tokens, and then concatenates these visual codes with text tokens into a single sequence for a standard autoregressive Transformer—enabling efficient, end-to-end joint modeling of images and language.

A particularly interesting development is the application of techniques from NLP directly to visual tokenization. Being-VL-0 [126] explores adapting byte-pair-encoding (BPE), a widely used subword tokenization algorithm in NLP, to compress sequences of quantized visual tokens. This can lead to shorter sequence lengths and more efficient processing by the LLM, analogous to how BPE handles rare words in text.

Importance-based and region-aware tokenization dynamically focuses on image regions that are more salient or important for the task at hand, effectively selecting and encoding tokens in a content-aware, flexible manner. LaVIT [127] developed a dynamic visual tokenizer comprising a token selector and a token merger, to identify and combine the most informative and semantically meaningful patches from an image. This enables the model to allocate its representational capacity more effectively. Morph-tokens [128] utilize a Q-former, which employs learned queries to dynamically extract relevant visual tokens from image features, abstracting and quantizing them into discrete "morph-tokens" that can adapt based on context. The SEED series (e.g., SEED [129], SEED-LLAMA [130]) learn a causal Q-former to encode a fixed number of image causal embeddings and a VQ codebook to discretize these embeddings into quantized visual codes. This causal approach aims to capture dependencies within the visual information more effectively. Similarly, AnyGPT [131] employs the SEED tokenizer [129] to map images (alongside other modalities) into a shared discrete token space, unifying all inputs into a single sequence and enabling coherent cross-modal reasoning and generation.

Furthering this direction, PUMA [132] enhances unified MLLMs with multi-granular visual generation capabilities, employing a tokenization strategy that captures visual information at different levels of detail. This enables both coarse and fine-grained image synthesis. Groma [117], as mentioned earlier, also fits here due to its emphasis on localized tokens corresponding to specific regions, which are inherently importance-based. The idea of focusing on salient regions is also central to models designed for fine-grained understanding and grounding.

Unified tokenization methods for different tasks and modalities are versatile tokenization schemes that enable the development of more powerful MLLMs. Show-o [183] proposes a unified transformer architecture for multimodal understanding and generation, necessitating a robust visual tokenization method that can support both discriminative and generative tasks using the same set of visual tokens. ANOLE [140], an open, autoregressive, native large multimodal model, handles interleaved image-text generation, where visual tokenization is crucial for seamlessly weaving image representations into textual sequences. Its tokenization must be efficient and expressive enough to maintain coherence during generation. Janus [182] and its successor Janus-Pro [177] focus on decoupling the visual encoding from the LLM for unified multimodal understanding and generation. Janus-Pro takes this idea further, showing that expanding both the breadth of multimodal pretraining data and the model's parameter count yields even greater gains in seamless cross-modal understanding and high-quality generative performance across images and text.

Some models aim for extreme efficiency in tokenization. LLaVA-Mini [201] explores efficient image and video large multimodal models by representing each image or video frame with as few as a single “vision token” or a very small number of tokens. This highly compressed representation challenges the tokenizer to capture the most salient information in an extremely compact form.

The scope of unified tokenization is also expanding beyond static images. OmniMamba [202] proposes an efficient and unified approach for multimodal understanding and generation using state space models (SSMs) instead of transformers for the backbone. The visual tokenization for such an architecture must be compatible with the sequential processing nature of SSMs. In dynamic and interactive scenarios, OmniJarvis [203] introduces a unified vision-language-action tokenization scheme. This enables an agent to perceive its environment (vision), understand language, and decide on actions—all represented as sequences of discrete tokens—facilitating open-world instruction following.

Even highly specialized domains are being considered. For instance, Zhou and Poczos [204] suggest that complex scientific data, such as molecular structures, can be tokenized. While their work focuses on molecule properties, the use of a VAE for representation learning is analogous to visual tokenization, where the ”visual” input is a molecular graph or 3D structure. The ambition of visual tokenization also extends to understanding complex scientific data from other fields. For example, research on astrophysical phenomena, such as that detailed in On the formation of super-Jupiters [205], while not a tokenization method itself, underscores the complexity of visual data (e.g., from simulations, telescopic imagery) that future MLLMs might need to ingest. Developing tokenization techniques capable of faithfully representing such nuanced scientific information for LLM-based analysis remains a significant frontier. UniFashion [206] tailors unified vision–language modeling for fashion by using a specialized visual tokenizer that encodes fine-grained textures, patterns, and silhouettes. This domain-aware tokenization boosts retrieval accuracy and empowers coherent generation of new garment designs, delivering substantial improvements on industry-specific retrieval and creative synthesis tasks.

The overarching goal of these diverse visual tokenization strategies is to create a lingua franca between the visual world and powerful language models, enabling a new generation of MLLMs that can see, understand, and reason about visual information with increasing sophistication and efficiency.

3.2.2. Unified Transformers Training

Once visual (and other modal) information is tokenized, the next crucial step is to train the transformer-based LLM to understand and generate multimodal content. Many multimodal models employ various training objectives derived from or inspired by language models, such as Masked Language Modeling (MLM), Autoregressive/Causal Language Modeling (CLM), or a combination of both. In addition to these objectives, they incorporate several techniques from the field of language modeling, like retrieval augmentation, instruction tuning, and supervised fine-tuning, to enhance their capabilities and align them with human expectations.

Tasks and Objectives

Early works like BEIT-3 [133] adapted the Masked Language Modeling (MLM) task—popularized by models like BERT [22]—as a unified multimodal pretraining objective. In BEIT-3, patches of images (visual tokens) and text tokens are masked, and the model is trained to predict these masked tokens based on the surrounding unmasked multimodal context. This encourages the model to learn rich, bidirectional representations across modalities.

Later, with the rise of powerful generative LLMs like GPT-3 [31], most multimodal works adopted the autoregressive or causal language modeling (CLM) approach. In this paradigm, the model learns to predict the next token in a sequence, where the sequence can be composed of interleaved text and visual tokens. This naturally lends itself to generation tasks, as the model can continue generating tokens (either text or visual) to complete a sequence. Emu [136] integrates video and image data interleaved with text into a unified format of visual embeddings and text tokens. It uses a Causal Transformer to map visual embeddings into a causal latent space for unified autoregressive prediction of both modalities, though it employed different losses for text tokens and visual embeddings initially. Emu3 [207] further champions this, suggesting that ”Next-Token Prediction is All You Need” for building powerful multimodal models, simplifying the training objective to a single autoregressive one across modalities.

Unlike Emu’s initial dual-loss approach, LaVIT [127] fully unifies image and text processing by converting images into discrete tokens, enabling a single generative (autoregressive) objective for both modalities within a unified multimodal modeling. This streamlined approach has been adopted by many contemporary methods. VL-GPT [135], the SEED series (SEED [129], SEED-LLAMA [130], SEED-Story [141], SEED-X [137]), Chameleon [139], ANOLE [140], MetaMorph [176], and Orthus [208] all utilize this mainstream autoregressive fashion to train their models on interleaved sequences of text and visual tokens, predicting the next token whether it is textual or visual. LMFusion [86] specifically focuses on adapting pretrained language models for multimodal generation, likely using autoregressive objectives on combined visual and textual token streams. The work by QLIP [209] leverages text-aligned visual tokenization to unify autoregressive multimodal understanding and generation, highlighting the tight coupling between tokenization quality and the effectiveness of unified training. Similarly, TokenFlow [210] proposes a unified image tokenizer designed to support both multimodal understanding and generation, trained likely through autoregressive objectives. ILLUME+ [211] combines dual visual tokenization with diffusion refinement within its unified MLLM framework, suggesting a hybrid training approach where autoregressive generation might be enhanced by diffusion-based refinement stages for visual outputs.

The combination of MLM and autoregressive objectives aims to integrate the bidirectional understanding capabilities of MLM with the sequential prediction strengths of CLM. This enables the model to understand context from all directions (like MLM) while also generating coherent, context-aware sequences (as in CLM). The CM3 [120] is a prominent example. It generates tokens sequentially from left to right (autoregressive) but also masks out some tokens during training, which are then predicted at the end of the sequence rather than in their original positions. This allows for efficient training on very long sequences of interleaved document data from the web.

A novel direction is explored in Sugar [212]. This approach seeks to combine the strengths of both generative (e.g., next-token prediction) and discriminative (e.g., contrastive learning, matching scores) training objectives within a single framework. Such co-training can potentially lead to models that are not only good at generating multimodal content but also possess a strong discriminative understanding of cross-modal relationships, which is beneficial for tasks like retrieval and VQA. The UniFashion model [206], designed for fashion retrieval and generation, likely employs such mixed training objectives to excel at both tasks.

Strategies

Several training strategies, often borrowed and adapted from LLMs, are employed to enhance the performance and capabilities of these unified multimodal models.

1. Instruction tuning is a technique used to enhance multimodal models by fine-tuning them on a diverse set of tasks framed through natural language instructions. This approach helps models like VL-GPT [135], Emu [136], SeTok-based models [116] (referred to as SETOKIM in the original PDF—likely a typographical error referring to models using SeTok), SOLO [138], and SEED-X [137] more effectively handle and respond to both visual and text data within a unified framework, thereby improving their zero-shot and few-shot generalization to new tasks. MetaMorph [176] explicitly focuses on multimodal understanding and generation via instruction tuning, showcasing its importance in steering the behavior of large MLLMs.
2. Retrieval-augmented methods are another strategy to enhance multimodal models by allowing them to incorporate relevant external information retrieved from a knowledge base (which can itself be multimodal) during the generation or understanding process. This helps them better handle complex multimodal tasks that require external knowledge not present in their training data or parameters. RA-CM3 [213] incorporates the CM3 model with a dense multimodal retriever to integrate external knowledge from a large corpus of multimodal documents, significantly improving its performance on knowledge-intensive tasks. CM3Leon [134], distinct from the Chameleon [139] model, also adapts large-scale retrieval-augmented pretraining and multi-task supervised fine-tuning, showing strong results in both image and text generation.
3. Supervised fine-tuning (SFT) in text-only LLMs trains the model on labeled instruction–response pairs to boost task-specific accuracy and instill desired behaviors such as helpfulness and safety. Multimodal models similarly undergo SFT after their core pretraining phase, using annotated image–text examples to align visual and linguistic representations and enhance instruction following across modalities. In addition to CM3Leon [134], architectures like Chameleon [139], SEED-LLAMA [130], and ANOLE [140] adopt supervised fine-tuning to achieve tighter multimodal alignment, enforce more reliable instruction compliance, and improve robustness—ultimately enabling these models to deliver state-of-the-art performance on tasks ranging from VQA and image captioning to interactive multimodal generation.

The principles of unified training are also being explored in domains beyond traditional vision-language tasks, showcasing the versatility of the approach. For instance, Uni2TS [214] and UniTS [215] (a unified multi-task time series model) demonstrate how a single transformer architecture can be trained with unified objectives to handle diverse time series forecasting tasks. While focused on time series data, the underlying philosophy of unifying different data types and tasks under a common modeling framework resonates with the efforts in multimodal LLMs. These examples underscore a broader trend in AI towards building more general and adaptable models through unified training paradigms.

Choosing an effective visual tokenizer is inseparable from designing the unified training objectives and strategies that follow. A thoughtfully engineered tokenizer converts raw pixels into meaningful "visual words" that the LLM can process, while the unified training regimen imparts the rules by which these visual words and textual words merge to form coherent multimodal representations and enable robust cross-modal reasoning.

3.2.3. Expanding the Unified Objective to Diverse Modalities

While unified multimodal modeling has primarily focused on vision and language, the flexibility inherent in token-based processing and unified language-modeling objectives offers potential to include a wider array of modalities. The core idea remains consistent: transform data from various modalities into discrete token sequences that can be processed by a unified Large Language Model (LLM). Several pioneering models have explored extending these unified objectives beyond traditional vision-language scenarios.

Unified-IO 2 [142] significantly broadens autoregressive sequence-to-sequence Transformers to handle audio and action modalities alongside vision and language, demonstrating enhanced multimodal reasoning capabilities. Building on this, AnyMAL [143] introduces a modular tokenization framework, allowing new modalities to be plugged in without altering the core architecture. In contrast, OmniJarvis [203] targets embodied AI by unifying vision, language, and action tokens to support environment perception and instruction execution. More recently, UnIVAL [144] extends the unified objective to video and audio by projecting text, images, video frames, and raw audio into a shared vocabulary and pretraining a BART-based model on all combinations of modality pairs. Complementarily, VATLM [145] introduces a unified masked-prediction framework that processes visual, audio, and text inputs through a shared multipath Transformer backbone, strengthening cross-modal feature alignment—especially for speech signals. Together with the BART-based paired-modality pretraining, these two lines of work lay the foundation for truly modality-agnostic large language models—systems that can flexibly ingest any combination of images, audio, and text, align their representations, and tackle downstream tasks ranging from multimodal question answering and image captioning to audio-guided content generation, all within a single unified architecture.

3.3. Unified Multimodality Modeling within a Single Visual View

Since humans perceive text by seeing it on screens or paper, a natural question arises: can machines comprehend text in a visual context like humans, rather than processing it sequentially? Over the past four years, researchers have begun designing models to explore whether machines can understand visually-situated text similarly to human perception.

3.3.1. Modeling from Pixel Space

Visual Modeling from Pixel Space

Visual modeling from pixel space is a natural approach to learning visual information, as pixels are the fundamental components of an image. Chen et al. [146] introduced iGPT, a model that learns image representations directly from pixels. iGPT employs a sequence Transformer trained to autoregressively predict pixels without prior knowledge of the 2D image structure. Notably, even when trained on low-resolution ImageNet data without labels, a GPT-2 scale iGPT demonstrates robust visual representation learning, as evaluated by linear probing, fine-tuning, and low-data classification tasks.

The initial success of iGPT highlighted the potential of autoregressive pretraining in computer vision. However, its computational demands—particularly the quadratic complexity of attention mechanisms relative to sequence length—limited its scalability to higher-resolution images. Subsequent research has focused on overcoming these limitations and further developing iGPT-based methods.

One direction has focused on enhancing computational efficiency and representation capacity. Razzhigaev et al. [147] proposed BPE-iGPT, tackling computational challenges by adapting Byte-Pair-Encoding to compress pixel sequences. BPE-iGPT effectively reduces sequence lengths—for example, encoding a 112×112 image into approximately one thousand tokens—thereby decreasing computational requirements while maintaining expressive image representations. Another significant enhancement, D-iGPT, introduced by Ren et al. [148], extends the iGPT framework by making two critical modifications. Firstly, it shifts the prediction target from raw pixel tokens to semantic tokens derived from discriminatively trained vision-language models like CLIP [184], enabling higher-level understanding. Secondly, D-iGPT supplements the standard autoregressive objective by predicting semantic tokens of visible image regions. Amrani et al. [149] developed XTRA, an autoregressive vision model emphasizing both sample and parameter efficiency. XTRA introduces a "Block Causal Mask," enforcing causality at the block level rather than the token level. By reconstructing pixel values block-by-block, XTRA captures higher-level structural patterns, enabling more abstract and semantically meaningful representations. This approach has led XTRA models to surpass previous autoregressive benchmarks while significantly reducing training samples and parameters.

Distinct from these autoregressive approaches, an alternative generative framework was presented by Chen et al. [150] through PixelFlow. PixelFlow utilizes flow-based modeling directly in pixel space, simplifying the generative process without relying on a pre-trained variational autoencoder. It employs cascade flow modeling with a unified set of parameters across multiple resolutions, incrementally increasing image resolution via flow matching. PixelFlow demonstrated competitive performance on benchmarks such as ImageNet, representing a viable alternative for high-quality pixel-space image generation.

Text Modeling from Pixel Space

Mansimov et al. [151] are the first to propose the in-image neural machine translation task. Their work proposed an end-to-end neural model that learns to translate text embedded within an image to another language, producing a new image with the translated text. Critically, this model is trained using purely pixel-level supervision, meaning it learns the translation task by directly manipulating and generating pixel data without relying on intermediate symbolic text representations. This pixel-domain approach offers a pathway to more universal text processing models, as it inherently accommodates diverse scripts and visual text renderings by treating text as part of the image itself, thereby unifying the input and output spaces via pixels. Salesky et al. [152] utilize visual text representations for robust open-vocabulary translation. Their approach involves rendering raw, unsegmented text into images and then processing these images with sliding windows and techniques inspired by optical character recognition to create continuous vocabulary representations. This method of deriving text understanding directly from the visual domain, rather than relying on discrete unicode sequences and predefined subword vocabularies, provides substantial robustness against various text permutations and noise that typically degrade traditional models. Inspired by such improvement for translation robustness from visual text representations, PIXEL [153], which introduces a new paradigm called Pixel Language Modeling, is proposed to address the vocabulary bottleneck in language modeling and achieves promising results on standard language modeling benchmarks. Specifically, PIXEL adopts the ViT-MAE architecture [216], consisting of a ViT [54] encoder and a transformer [1] decoder. It is trained to reconstruct the pixels of masked patches, eliminating the need to predict a distribution over tokens. PIXEL's training involves two stages: pixel pretraining and pixel fine-tuning. The encoder is ultimately used for language tasks and evaluated against the BERT model on NLP tasks across multiple languages. Although PIXEL performs slightly worse than BERT on Latin Scripts (e.g., English) tasks, it demonstrates strong performance on tasks involving other non-Latin Scripts.

Motivated by PIXEL, subsequent methods have emerged to learn text representation with pixels [155–157]. Xiao et al. [156] propose to learn the textual semantics of the sentence and document as a visual representation learning process that mirrors the human cognitive process, as human understanding of the text is not only visually grounded but also tolerant of irregularities such as typos and varied word orders. They design a pixel sentence representation learning framework that leverages the perceptual continuity of vision models to capture the rich multimodal semantic signals in text, using visually grounded textual perturbations to create contrastive pairs for contrastive learning.

Based on this insight, the framework proposes a progressive monolingual alignment scheme for monolingual learning and an iterative cycle training scheme for cross-lingual transfer learning. The results presented in this work validate that modeling textual semantics in the pixel space—by leveraging the shape-based information inherent in text—is a promising approach for developing stronger and more human-like sentence encoders.

While the aforementioned works focused primarily on language understanding, the paradigm was significantly extended to include generative capabilities. PIXAR [155] was the first to pioneer generative language modeling directly in pixel space. It adopts a decoder-only transformer architecture with 12 layers to generate text by taking text rendered as images as input. As the first approach of its kind, PIXAR demonstrated the potential to create more expressive language models that rely solely on perceptual input. It not only showcased strong performance on discriminative tasks like the GLUE benchmark [217], but also achieved performance comparable to GPT-2 [218] for text generation tasks and proved more robust to orthographic attacks. Building upon PIXAR [155], Gao et al. [157] further enhanced pixel-based language models by proposing a novel pretraining objective that integrates both patch and text prediction. Their method demonstrated superiority over previous models, including PIXEL and PIXAR, on natural language understanding benchmarks like GLUE. Crucially, they also extended their objective into an autoregressive setting that processes both visual and textual inputs, mapping them into token embeddings before feeding them into the backbone transformer decoder, which can generate text by taking the visual input.

Recent contributions continue to build upon these foundations by enhancing backbone models for visual text generation and aligning comprehension with generation. Li et al. [158] focused on empowering diffusion-based models to generate accurate and aesthetically pleasing visual texts. They identified key limitations: BPE tokenization, which fragments words and increases the difficulty of generation, and insufficient learning in cross-attention modules, which hinders the model's ability to associate text tokens with their visual locations. To address these issues, they proposed a mixed granularity input strategy using OCR features for improved text representation and introduced glyph-aware training losses. These include an attention alignment loss to refine attention maps, a local MSE loss to focus on text regions, and an OCR recognition loss to ensure accuracy, effectively enhancing visual text quality while preserving general image generation capabilities. Concurrently, Zhao et al. [159] introduced TextHarmony, aiming to create a unified model proficient in both understanding and creating visual text. Addressing the challenge of modality inconsistency that often degrades performance in models handling both vision and language generation, they proposed Slide-LoRA. This mechanism dynamically aggregates modality-specific and modality-agnostic LoRA experts via a gating network, partially decoupling the generative spaces within a single model instance with minimal parameter increase. Furthermore, they developed the DetailedTextCaps-100K dataset to provide high-quality, text-focused captions, enhancing the model's training for visual text generation. TextHarmony demonstrated the viability of an integrated approach, achieving performance comparable to specialized models across various visual text comprehension and generation tasks.

In summary, modeling text within the pixel space has rapidly evolved from a niche solution for specific tasks into a comprehensive paradigm capable of both robust language understanding and high-fidelity generation, paving the way for more universal and resilient language models.

Document-Level Multimodal Modeling

Document images represent a common multimodal data format, which has led to various pixel-level modeling approaches. The goal of these methods is to analyze document content and structure directly from raw pixels, thereby reducing the reliance on separate Optical Character Recognition engines.

For instance, DiT by Li et al. [160] introduced a self-supervised pretraining strategy that leverages large-scale unlabeled text images to enhance document image understanding. DiT processes input document images by resizing them and splitting them into patches. Its pretraining is inspired by BEiT, employing a Masked Image Modeling objective in which the model learns to recover visual tokens from corrupted input images. Crucially, DiT retrains a discrete variational autoencoder specifically with a large corpus of document images, ensuring that the resulting visual tokens are more domain-relevant for Document AI tasks than those from tokenizers trained on natural images. This pretraining, conducted without any human-labeled document images, enables DiT to serve as a powerful vision backbone. It effectively supports downstream tasks such as document image classification, document layout analysis, table detection, and text detection for OCR, demonstrating notable improvements and robust multimodal capabilities.

OCR-free methods constitute an essential research direction in document image analysis, aiming to eliminate error propagation and computational overhead associated with separate OCR steps. Donut by Kim et al. [161], was a pioneering OCR-free transformer explicitly designed for visual documents. It features an end-to-end architecture consisting of a Swin Transformer-based visual encoder and a BART-based textual decoder. Donut jointly learns text prediction from the image and text contexts during its pretraining phase, which is framed as a pseudo-OCR task where the model learns to read all texts in reading order using a cross-entropy loss objective. For pretraining on diverse languages and domains, Donut can utilize a synthetic data generator called SynthDoG, which creates document images with varied backgrounds, layouts, and text content. In the fine-tuning stage, Donut is adapted to various downstream applications by generating a sequence of tokens conditioned on a task-specific prompt; this token sequence can then be converted into a structured JSON format. Its OCR-free design offers excellent generalization across multiple languages and varied domains, along with benefits in speed and reduced computational cost.

Building on such OCR-free paradigms, Park et al. [166] introduced a framework that leverages pretrained Multimodal Large Language Models (MLLMs) and multiscale visual features to effectively handle varying font sizes within document images. This approach first generates multiple subimages using shape-adaptive cropping at different scales to capture both holistic and detailed information. To manage the increased number of visual tokens resulting from these multiscale inputs, the authors proposed the hierarchical visual feature aggregation module. This module incorporates cross-attentive pooling within a feature pyramid structure, where features from finer scales are compressed and aggregated with features from coarser scales. This pooling uses the coarser scale's pooled features as queries and the finer scale's original features as keys and values, significantly reducing token inputs to the MLLM's language model component. A reconstruction loss is used during training to ensure that the module preserves essential visual details during compression. This design balances computational efficiency and information retention.

Expanding pixel-based modeling into broader contexts, PHD [164], inspired by PIXEL [153], tackles the challenges of analyzing historical documents, which often suffer from OCR inaccuracies and loss of visual context. It leverages both authentic historical newspaper scans and synthetically generated data during its pretraining phase, focusing on pixel-level reconstruction from masked image patches. This strategy significantly enhances the model's ability to reconstruct both the visual form and textual content of historical documents, ultimately leading to robust performance in downstream tasks such as historical document question-answering after fine-tuning. Lotz et al. [167] further advanced pixel-level textual modeling by introducing a vocabulary-free encoder designed to augment pretrained language models. This encoder derives input embeddings directly from text rendered as pixels, thereby circumventing the limitations of fixed vocabularies. Their approach demonstrates considerable improvements over traditional subword tokenization, vocabulary expansions, and byte-level encodings, particularly enhancing model performance on languages and scripts underrepresented in initial training data. This pixel-based fallback mechanism not only supports effective cross-lingual transfer but also enhances the multilingual capabilities of primarily monolingual models without requiring extensive retraining. Zhu et al. [168] proposed SegAgent to specifically enhance fine-grained pixel-level comprehension and interaction within MLLMs. This system enables MLLMs to simulate human segmentation workflows by interacting with images through established interactive segmentation tools. SegAgent formulates the segmentation task as a multi-step Markov Decision Process, in which the MLLM iteratively generates precise textual click points. This approach enables high-quality image segmentation without requiring architectural modifications to the MLLM or the generation of implicit, non-textual tokens, thereby preserving the model's native output space.

3.3.2. Text Visualization Using Visual Tokens for Long-Context Multimodal Learning

While previous works in multimodality have focused on relatively short image and text inputs, more recent research seeks to extend models to handle longer contexts. Lu et al. [173] proposed SEEKER, a method that improves long-context multimodal understanding by compressing text into compact visual tokens and then interleaving them with tokens from multiple images.

SEEKER concatenates the visual tokens of text and image tokens in a sequential manner and adopts an autoregressive training objective to fine-tune the model. This supervised fine-tuning strategy enables the model to generate coherent responses based on multimodal inputs. In addition, it also utilizes instruction-tuning to enhance the model's long-context multimodal capabilities. Specifically, the instruction data is collected from CC3M [78], COCO [219] images, and arXiv PDF documents for long-form multi-image input tuning. For the long-form text output instruction-tuning, Lu et al. [173] propose a challenging task that requires the model to read the text in text-rich images and generate the same text without hallucination. The experimental results showcase SEEKER's strong long-context multimodal modeling abilities.

A concurrent study by Wang et al. [174] introduces VisInContext, a multimodal learning method designed to process long text contexts by representing them as visual tokens. Built upon the Flamingo model, VisInContext's core mechanism renders in-context text into images. These text-rendered images are then processed by the same vision encoder used for raw images, followed by two learnable resamplers that output a fixed number of tokens. To bridge the semantic gap between the rendered visual tokens and the original text, the authors designed two specific mechanisms which are token masking and text-centric contrastive learning. This approach significantly reduces GPU memory consumption during both training and inference and was used to extend the in-context text length during the pretraining of a 56-billion parameter Mixture-of-Experts (MoE) model.

3.3.3. Domain-Specific Extensions and Evaluations within Visual View

While previous sections focused on general modeling strategies, recent works have significantly broadened the landscape of pixel-level multimodal modeling. This expansion includes new foundational pretraining strategies, applications in highly specialized domains, and critical evaluations of core model capabilities. For example, expanding the foundational approaches, CLIPPO [162] unifies vision and textual information by rendering all inputs into images and training a shared encoder via contrastive learning. Concurrently, numerous works have applied pixel-level modeling to specialized domains. Huang et al. [169] developed MedPLIB, a biomedical-focused MLLM capable of pixel-level grounding, responding to arbitrary pixel-level prompts, and performing VQA tasks. MedPLIB uses a Mixture-of-Experts (MoE) multi-stage training strategy, separately training visual-language and pixel-grounding experts before integration for effective multitask learning. Likewise, VisionLLM v2 by Wu et al. [107], a generalist MLLM, integrates visual perception, understanding, and generation via a flexible "super link" mechanism that enhances multitask learning by mitigating potential training conflicts. OCR-dependent approaches also remain valuable; DocLayLLM by Liao et al. [170] integrates visual patch tokens and OCR-derived text into language models, significantly boosting document comprehension. The focus on structured data is further exemplified by other works. Lee et al. [163] propose Pix2Struct, a model focused on visually-situated language understanding that learns directly from masked screenshots of web pages. Its pretraining, which includes predicting HTML structures from pixel inputs, and its variable-resolution handling of visual inputs enable robust generalization across diverse document layouts. Similarly, PixT3 [165] reframes table-to-text generation as a visual recognition challenge, utilizing a self-supervised pretraining approach to capture structural table features from images effectively. Finally, beyond domain applications, research has also focused on enhancing and evaluating core capabilities such as grounding. To this end, Wang et al. [171] introduced Marten, a multimodal LLM utilizing VQA with mask-generation pretraining to foster refined spatial alignment between textual and visual elements, thereby reducing hallucination and improving interpretability. To demonstrate the capability, evaluations by Siam [172] offer valuable insights on pixel-level grounding supervision, indicating inconsistent improvements in VQA and grounding performance on benchmarks. These findings suggest that grounding capabilities may naturally emerge even without explicit supervision.

4. Multimodal Benchmarks and Tasks

The rapid development of multimodal large language models (MLLMs) has been accompanied by the creation of a diverse range of benchmarks and datasets. These resources are critical for systematically evaluating the capabilities and limitations of both classic vision–language models and the newest generation of unified MLLMs. In this section, we summarize widely used benchmarks across four major categories as shown in Figure 3, reflecting different evaluation paradigms and the evolution of the field, and inform future design choices.

4.1. Vision-Language Benchmarks and Tasks

A central focus in multimodal learning has been the development of vision-language benchmarks, which have long served as touchstones for measuring progress in integrating visual and textual modalities. Early datasets such as MS COCO and VQA laid the groundwork for tasks like image captioning and visual question answering. Building on these foundations, subsequent benchmarks have expanded the landscape to include image referring, document understanding, chart and diagram reasoning, and science-focused multimodal QA. Collectively, these resources have played a pivotal role in both evaluating model performance and inspiring new model architectures. Vision-language models are typically assessed across diverse tasks that target specific visual capabilities, such as image captioning, image referring, visual question answering, and document question answering. Specifically, there are 4 datasets for the image captioning task, including Flickr30K [220], MS COCO [219], Flickr30K Entities [221] and NoCaps [222]. There are 2 datasets for the image-referring task, including RefCOCO [223] and RefCOCOg [224]. For visual question answering, there are various VQA datasets based on natural images (VQA [225], Taiwan-VQA [226], PDFVQA [225],E-VQA [225],KRVQA [225], Lora [225],VQAV2 [227], OK-VQA [228], GQA [229] and VizWiz [230]), charts (ChartQA [231] and PlotQA [232]), infographics (InfographicVQA [233]), diagrams (AI2D [234]), document images (DocVQA [235]), text-rich images (OCR-VQA [236] and TextVQA [237]) and multimodal QA in scientific domain (ScienceQA [238]).

A new benchmark, ENIGMAEVAL [239], uses 1184 challenging puzzles from competitions to test advanced reasoning in language models. Current models perform very poorly on these puzzles, which require finding hidden connections for solutions, indicating a significant gap in their cognitive abilities compared to existing benchmarks. Recent works such as TokenFormer [240] scaling and Vision-RWKV [241] have also evaluated vision-language models and vision backbones on classic benchmarks including ImageNet-1K, CIFAR-10/100, ADE20K, COCO, Cityscapes, and Pascal VOC for image classification, segmentation, detection, and video tasks (YouTube-VOS). In addition, Dhariwal and Nichol [242] benchmarks generative models on ImageNet 128 × 128, LSUN Bedrooms/Cat, CelebA-HQ, and CIFAR-10, further expanding the evaluation landscape for multimodal and image synthesis tasks. The MMStar benchmark [44] further unifies many existing datasets such as VQAv2, TextVQA, DocVQA, ChartQA, and InfographicVQA, providing a holistic evaluation protocol for vision-language models across a wide range of tasks. The UnIVAL model [144] introduces a new unified benchmark spanning multiple modalities (image, video, audio, text), enabling evaluation across classic vision-language tasks but also generalizing to video and audio-text. The UnIVAL benchmark covers VQA, image captioning, video question answering, and audio-text tasks in a single framework, and has been released as a resource for evaluating truly unified MLLMs.

4.2. MLLM Benchmarks Assisted by GPT

As large language models have matured, a growing number of multimodal benchmarks now leverage the generative and evaluative capabilities of models like GPT-3 and GPT-4. In this paradigm, synthetic instruction-following data, dialogue, and reference answers are produced at scale, enabling richer and more nuanced evaluation settings. Datasets in this category are particularly valuable for stress-testing model alignment, reasoning, and generalization in scenarios where rapid, diverse, and scalable annotation is essential.

Liu et al. [78] introduce LLaVA-Bench by utilizing ChatGPT/GPT-4 to generate three types of multimodal instruction-following data from MS-COCO [219] images. Following LLaVA-Bench, Yin et al. [243] propose LAMM which utilizes GPT-API to construct more challenging visual instruction-following data by adding dense and fine-grained visual information. Bai et al. [244] propose a comprehensive visual dialogue dataset named TouchStone with the help of GPT-4 to evaluate the comprehension, multi-image analysis, and literary creation abilities of MLLMs. Farsi et al [245] propose five datasets, developed with the help of GPT-4o, including Persian-OCRQA for optical character recognition, PersianVQA for visual question answering, a Persian world-image puzzle for multimodal integration, Visual-Abstraction-Reasoning for abstract reasoning, and Iran-places for visual knowledge of Iranian figures and locations. M3Exam [43] is a new multimodal benchmark specifically designed to evaluate MLLMs on tasks that require the integration of image, audio, and text information. It supports large-scale, multilingual, and multilevel evaluation of instruction-following and reasoning abilities in modern MLLMs.

In addition, MMBench [246] provides a comprehensive grounding-focused evaluation for multimodal models, while MMT-Bench [247] targets large multimodal language models with a diverse suite of tasks. LVLM-eHub [248] offers a quantitative capability evaluation and leaderboard platform, widely used for open-world MLLM assessment. JourneyBench [249] enables the evaluation of open-world multimodal reasoning and instruction-following. MathVista [250] benchmarks mathematical and chart-based reasoning. CharXiv [251] addresses complex chart and document understanding. MangAUB [252] introduces multimodal understanding benchmarks in the anime domain.

Recent work includes VisionLLM v2 [107], which provides a generalist benchmark suite for hundreds of vision-language tasks, and MultiBench [253], which offers a wide-ranging multimodal benchmark for learning and understanding. LAMM [243] and MM-VET v2 [254] are also influential in multi-instruction and multimodal evaluation.

Open-source suites like InternLM-XComposer [91] and perception tests [255] further broaden the scope of multimodal and cross-domain evaluation.

4.3. MLLM Benchmarks Using Human Annotation

Despite the advantages of automation, human-annotated benchmarks remain essential for robust and trustworthy evaluation of MLLMs. These datasets are curated and labeled by human experts, ensuring high reliability across open-world, perception, reasoning, and multi-discipline tasks. Such benchmarks provide detailed evaluation suites and often include complex or open-ended scenarios that challenge both current models and annotation methodologies. Given the limitations—such as reduced reliability—of GPT-assisted MLLMs benchmarks, increasing human effort is being invested in developing more reliable and comprehensive open-world multimodal understanding benchmarks.

POPE [256] evaluates MLLM's hallucination, MathVista [250] and MV-Math [257] evaluates their math reasoning, DEMON [258] is designed to evaluate the capability of in-context learning, CharXiv [251] focuses on complex chart understanding with diverse styles, ConTextual [259] and SEED-Bench-2-Plus [260] evaluate MLLMs in text-rich multimodal scenarios. MMMU [261] focuses on multi-discipline multimodal understanding. TemporalVQA [262] evaluates visual temporal understanding and reasoning abilities of multimodal LLMs. MMSci [263] evaluates and improves multimodal scientific table understanding and reasoning. VLRMBench [264] is a new benchmark comprising 12 tasks across mathematical reasoning, hallucination understanding, and multi-image understanding to evaluate Vision-Language Reasoning Models (VLRMs). LVLM-eHub [248] includes 13 representative LVLMs such as InstructBLIP and LLaVA, which are thoroughly evaluated through both a quantitative capability assessment and an online arena platform. The former evaluates five categories of multimodal capabilities of LVLMs such as visual question answering and object hallucination on 42 in-domain text-related visual benchmarks, while the latter provides user-level evaluation of LVLMs in an open-world question-answering scenario. The LVLM-eHub benchmark [265] also offers a comprehensive human-annotated evaluation suite for 13 major vision-language models. It covers 42 in-domain visual-text benchmarks including VQA, text-rich image QA, hallucination detection, math reasoning, and more, offering a fine-grained comparison and public leaderboard. MMStar [44] also includes a significant portion of human-annotated data in its sub-benchmarks, unifying VQA, OCR, chart, and math tasks in a single evaluation.

Recently, new unified vision-language benchmarks have emerged. VILA-U [98] introduces a unified autoregressive model that evaluates both visual understanding and generation within a single framework. VILA-U is benchmarked on standard vision-language tasks—including image-language understanding, video-language understanding, and image/video generation—and is designed to bridge the gap between understanding and generation in multimodal models. VILA-U's benchmarks assess performance on both traditional datasets (such as VQAv2, COCO, and more) and its own unified evaluation suite. Qwen2-VL [179] presents a family of large-scale open-weight vision-language models evaluated on a comprehensive set of multimodal benchmarks. These include image-text understanding, code/math reasoning, and video analysis, demonstrating strong cross-modal generalization and robustness.

4.4. Long-Context Multimodal Benchmarks

Long-context multimodal benchmarks target the advanced ability of models to reason over extended, multimodal sequences—such as lengthy scientific papers, complex documents, or time series data. These datasets are designed to evaluate both the breadth and depth of model understanding across text, images, charts, and tables, as well as measure long-range context dependencies.

To evaluate the advanced ability of MLLMs—such as long-context understanding—Ma et al. [266] propose MMLongBench-Doc, a benchmark constructed from lengthy PDF-formatted documents containing multimodal components (i.e., text, image, chart, table, and layout structure). Similarly, Lu et al. [173] introduce ArxivQA, which requires the model to answer questions based on images of Arxiv documents. Large Concept Models (LCMs) [267] introduce the evaluation of models in a sentence embedding space (SONAR), focusing on sentence-level reasoning and zero-shot generalization for summarization and summary expansion tasks. These evaluations, though not strictly document-level, push the field toward higher-level multimodal abstraction and long-context understanding. Tong et al. [176] proposes Visual-Predictive Instruction Tuning (VPiT), which allows pretrained LLMs to generate both text and visual tokens. While not a benchmark, it is evaluated on both visual understanding and generation tasks, using established datasets like MS COCO, LLaVA-Instruct, and custom instruction-following datasets. The results show the model's ability to process and generate multi-turn, interleaved text-image sequences.

LMFusion [186] is a framework for adapting pretrained text-only LLMs (such as Llama-3) to handle interleaved text and image generation. The LMFusion paper evaluates its models across several multimodal tasks and datasets—including image generation and understanding—and demonstrates its performance on benchmarks like MS COCO and LLaVA-Instruct. LMFusion establishes new baselines for both efficiency and performance, with a focus on preserving language capabilities while incorporating vision skills. Recently, there has been a surge in biomedical MLLMs that go beyond image-level VQA to support pixel-level tasks. For instance, MedPLIB [169] introduces a comprehensive end-to-end MLLM designed for biomedical applications with pixel-level understanding. MedPLIB supports not only traditional VQA but also arbitrary pixel-level prompts (points, bounding boxes, free-form shapes) and grounding, enabling fine-grained region analysis in complex medical images. To advance this field, the authors propose the MeCoVQA dataset, covering 8 medical modalities for complex question answering and region understanding. MedPLIB achieves state-of-the-art results on several biomedical vision–language tasks and shows substantial improvements in zero-shot pixel grounding over previous models. Moreover, these capabilities position MedPLIB as a versatile tool for both research innovation and clinical decision support, accelerating translational applications in medical imaging.

To critically evaluate pixel-level vision-language grounding, PixFoundation [172] introduced two challenging new benchmarks: PixMMVP and PixCV-Bench. These benchmarks focus on segmentation and referring expression tasks that require fine-grained pixel-level understanding. Unlike many recent models trained with pixel-level supervision, PixFoundation demonstrates that MLLMs not trained for pixel-level grounding can sometimes outperform specialized models on grounding and VQA, highlighting the need for more nuanced evaluation protocols. The paper provides baseline protocols for extracting pixel-level grounding from generic MLLMs and offers insights on when such grounding abilities naturally emerge. These new benchmarks (PixMMVP, PixCV-Bench) expand the evaluation space for grounding, segmentation, and vision-centric VQA, complementing existing datasets and highlighting gaps in current pixel-level MLLMs. UNITS [215] is notable for evaluating unified models across 38 datasets—including human activity, healthcare, engineering, and finance—across forecasting, classification, imputation, and anomaly detection tasks. These datasets vary in temporal scale and complexity, making UNITS relevant as a "long-context" and cross-domain challenge, especially for multimodal foundation models seeking to integrate time series and vision-language data.

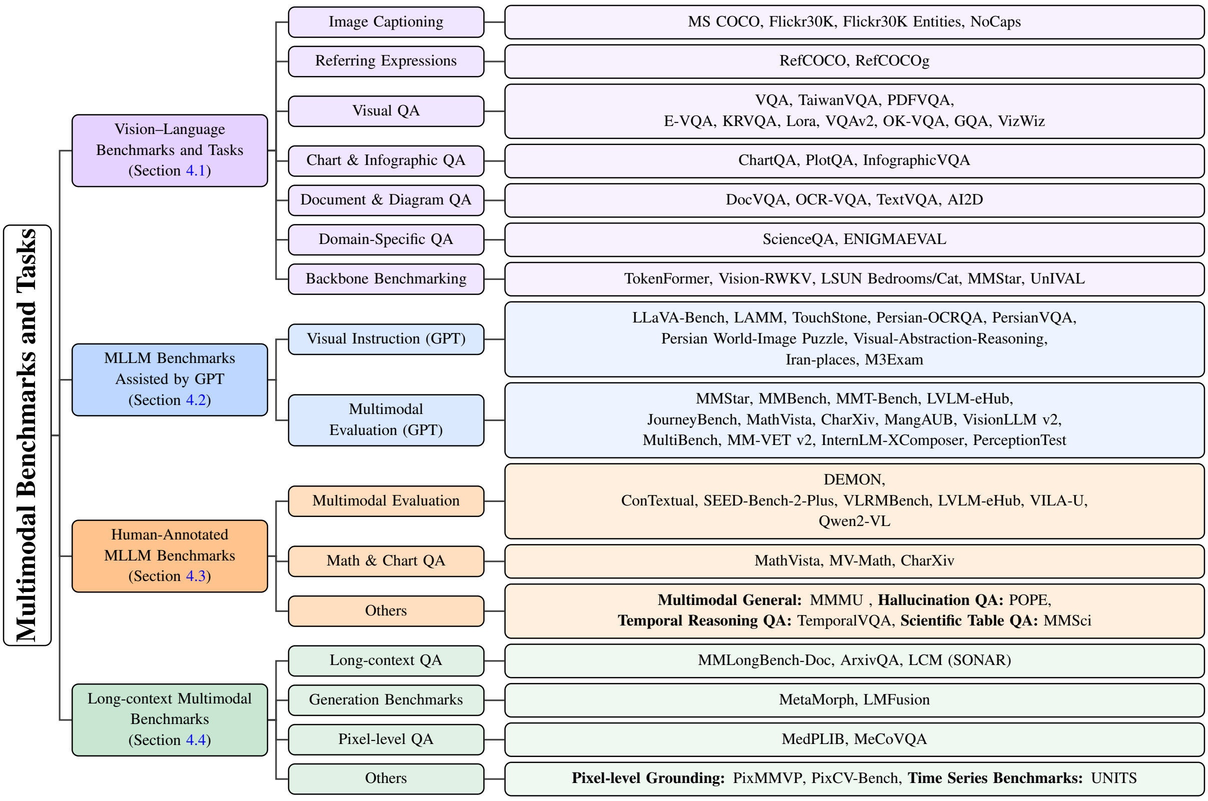

Multimodal Benchmarks and Tasks
Vision–Language Benchmarks and Tasks (Section 4.1)
Image Captioning
MS COCO, Flickr30K, Flickr30K Entities, NoCaps
Referring Expressions
RefCOCO, RefCOCOg
Visual QA
VQA, TaiwanVQA, PDFVQA, E-VQA, KRVQA, Lora, VQAv2, OK-VQA, GQA, VizWiz
Chart & Infographic QA
ChartQA, PlotQA, InfographicVQA
Document & Diagram QA
DocVQA, OCR-VQA, TextVQA, AI2D
Domain-Specific QA
ScienceQA, ENIGMAEVAL
Backbone Benchmarking
TokenFormer, Vision-RWKV, LSUN Bedrooms/Cat, MMStar, UnIVAL
MLLM Benchmarks Assisted by GPT (Section 4.2)
Visual Instruction (GPT)
LLaVA-Bench, LAMM, TouchStone, Persian-OCRQA, PersianVQA, Persian World-Image Puzzle, Visual-Abstraction-Reasoning, Iran-places, M3Exam
Multimodal Evaluation (GPT)
MMStar, MMBench, MMT-Bench, LVLM-eHub, JourneyBench, MathVista, CharXiv, MangAUB, VisionLLM v2, MultiBench, MM-VET v2, InternLM-XComposer, PerceptionTest
Human-Annotated MLLM Benchmarks (Section 4.3)
Multimodal Evaluation
DEMON, ConTextual, SEED-Bench-2-Plus, VLRMBench, LVLM-eHub, VILA-U, Qwen2-VL
Math & Chart QA
MathVista, MV-Math, CharXiv
Others
Multimodal General: MMMU , Hallucination QA: POPE, Temporal Reasoning QA: TemporalVQA, Scientific Table QA: MMSci
Long-context Multimodal Benchmarks (Section 4.4)
Long-context QA
MMLongBench-Doc, ArxivQA, LCM (SONAR)
Generation Benchmarks
MetaMorph, LMFusion
Pixel-level QA
MedPLIB, MeCoVQA
Others
Pixel-level Grounding: PixMMVP, PixCV-Bench, Time Series Benchmarks: UNITS

Figure 3. Overview of multimodal benchmarks and tasks: MS COCO [219], Flickr30K [220], Flickr30K Entities [221], NoCaps [222], RefCOCO [223], RefCOCOg [224], VQA [225], TaiwanVQA [226], PDFVQA [225], E-VQA [225], KRVQA [225], Lora [225], VQAv2 [227], OK-VQA [228], GQA [229], VizWiz [230], ChartQA [231], PlotQA [232], InfographicVQA [233], DocVQA [235], OCR-VQA [236], TextVQA [237], AI2D [234], ScienceQA [238], ENIGMAEVAL [239], TokenFormer [240], Vision-RWKV [241], LSUN Bedrooms/Cat [242], MMStar [44], UnIVAL [144], LLaVA-Bench[78], LAMM[243], TouchStone[244], Persian-OCRQA [245], PersianVQA [245], Persian World-Image Puzzle [245], Visual-Abstraction-Reasoning [245], Iran-places [245], M3Exam [43], MMBench [246], MMT-Bench [247], LVLM-eHub [248], JourneyBench [249], MathVista [250], CharXiv [251], MangAUB [252], VisionLLM v2 [107], MultiBench [253], MM-VET v2 [254], InternLM-XComposer [91], PerceptionTest [255], DEMON [258], ConTextual [259], SEED-Bench-2-Plus [260], VLRM-Bench [264], LVLM-eHub [248], VILA-U [98], Qwen2-VL [179], MathVista [250], MV-Math [257], CharXiv [251], MMMU [261], POPE [256], TemporalVQA [262], MMSci [263], MMLongBench-Doc [266], ArxivQA [173], LCM (SONAR) [267], MetaMorph [176], LMFusion [86], MedPLIB [169], MeCoVQA [169], PixMMVP [172], PixCV-Bench [172], UNITS [215].

5. Future Directions

The development of Multimodal Large Language Models (MLLMs) has largely been driven by heterogeneous architectures, which have achieved strong performance across various vision-language tasks. However, their limitations highlight the need for more integrated approaches. Unified modeling paradigms—which employ either a single language-modeling objective or a consolidated visual representation—offer a promising path to overcoming the bottlenecks of their heterogeneous counterparts. Although still in their early stages, these unified models show significant potential for handling complex real-world scenarios, such as understanding documents that contain charts, tables, and images.

To guide future research, this section outlines several key directions. These build on the analyses presented in this survey and are essential for developing the next generation of more capable, robust, and efficient multimodal systems.

5.1. Enhancing Efficiency and Scalability

The practical deployment of advanced MLLMs is inextricably tied to their computational cost. The scale of the models discussed in this survey implicitly underscores this challenge. Future research must prioritize the development of model compression techniques—such as quantization, pruning, and knowledge distillation—specifically tailored to the architectural complexities of MLLMs. At the same time, exploring inherently efficient architectures, such as state-space models, presents a promising alternative for multimodal tasks. Furthermore, standardizing and advancing parameter-efficient fine-tuning (PEFT) methods is crucial for enabling robust and data-efficient adaptation of MLLMs across diverse tasks and domains. Addressing these efficiency and scalability challenges is paramount for democratizing access to MLLMs and facilitating their deployment in resource-constrained environments.

5.2. Improving Robustness, Safety, and Explainability

Advancing AI systems requires a focus that goes beyond mere task performance to include reliability and trustworthiness. A key direction for future work is the systematic development of robust techniques to detect, quantify, and mitigate the propensity of MLLMs to generate factually incorrect or biased outputs—often referred to as hallucinations. This effort must be complemented by the deliberate promotion of algorithmic fairness to ensure equitable performance across diverse demographic groups and data distributions. Finally, advancing Multimodal Explainable AI (MXAI) is critical. Robust methods are needed to provide clear, human-interpretable insights into how MLLMs process multimodal inputs to reach their conclusions, thereby fostering greater transparency and trust.

5.3. Expanding Capabilities and Applications

This research thrust aims to leverage MLLMs for increasingly complex tasks and novel applications, moving beyond the established benchmarks and capabilities cataloged in this survey.

- Expanding to Other Modalities. A primary objective is to extend the scope of MLLMs beyond vision and language. While current research on modalities such as audio and video predominantly relies on heterogeneous architectures, unified paradigms offer a more promising foundation for seamless integration. Models like AnyGPT, which employ specialized tokenizers to unify speech, music, and images with text, represent important early steps. Sustained research is needed to fully realize the potential of unified models to cohesively handle a diverse spectrum of modalities.
- Deepening Document and Contextual Understanding. Significant challenges and open questions remain in the robust and nuanced comprehension of complex, long-form multimodal content:
  - Multimodal Document Understanding: As revealed by challenging benchmarks like CharXiv, MMMU, and SEED-Bench-2-Plus, current MLLMs struggle with text-rich documents that intricately integrate text with charts, tables, and other visual elements. Future work must focus on enhancing fine-grained perception and reasoning over these composite structures—a critical need for applications in academic, financial, and technical domains.
  - Long-context Multimodal Understanding: The limited capacity of state-of-the-art MLLMs to process long-form content, as demonstrated by benchmarks like MMLongBench-Doc and ArxivQA, presents a major obstacle to their broader real-world application. Enabling models to maintain coherence and perform complex reasoning over extended documents or videos will require fundamental innovations in attention mechanisms, memory architectures, and the efficient representation of long-sequence multimodal data.
- Fostering Advanced Multimodal Reasoning and Interaction. Beyond comprehension, the next frontier lies in enabling more sophisticated cognition and agency. This includes cultivating deeper and more complex reasoning abilities, moving beyond pattern recognition towards commonsense, causal, mathematical, and scientific inference from multimodal data. Furthermore, applying MLLMs to embodied AI and interactive systems represents a significant leap forward, with the goal of controlling robots and agents in dynamic physical environments and developing more natural and effective paradigms for human-AI collaboration in complex, real-world tasks.

5.4. Advancing Data and Benchmarking

The trajectory of MLLM research is fundamentally shaped by the quality of training data and the rigor of evaluation benchmarks. To effectively guide future work—especially for the unified paradigms central to this survey—a dual focus is essential. First, it is crucial to systematically curate high-quality, diverse datasets. This involves creating larger, more varied, and meticulously annotated multimodal datasets that span a wide range of domains and interaction types, with particular emphasis on data that can foster and evaluates advanced reasoning. Second, the field requires more diagnostic and robust benchmarks. Next-generation evaluations must move beyond simple accuracy metrics to assess specific capabilities, such as compositional understanding, causal inference, robustness to adversarial inputs, and bias mitigation. Such benchmarks are indispensable for revealing the nuanced strengths and limitations of different modeling paradigms and for driving the development of truly capable, trustworthy multimodal systems for real-world applications.

6. Conclusions

In this survey, we investigate the current research landscape for multimodal modeling and explore various research directions. The mainstream multimodal models primarily utilize heterogeneous architectures, incorporating multimodal data through different data encoders and adapters that bridge the representations of different modalities. Beyond heterogeneous architectures, two relatively new research directions focus on unified multimodal modeling: one using a unified language-modeling objective, and the other involving unified multimodal modeling within a single visual view. These two approaches remain underexplored, especially the latter. In addition, we discuss the current benchmark datasets and suggest potential future directions to advance the field of multimodal modeling. We hope that this survey encourages further research on unified multimodal modeling and serves as a valuable reference for designing new multimodal modeling methods.

Author Contributions

Z.D.: conceptualization, methodology, writing; Y.W., Y.L. (Yueqing Liang), J.D., Y.Y., L.F., L.H., Y.H., Y.Z., C.M., J.C. and W.Z. (Wenting Zhao): writing of sections; W.Z. (Weizhi Zhang), Y.L. (Yinghui Li): discussion and review; P.S.Y.: review and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. This study did not involve humans or animals.

Informed Consent Statement

Not applicable. This study did not involve humans.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-assisted Technologies

During the preparation of this work, the authors used ChatGPT to polish the grammar of the content. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Appendix A. Detailed Comparison between All Models of the Three Categories for Multimodality Modeling

Table A1 shows the architecture, model size, and training methods for the representative models in the proposed three categories. While Table A2 shows the pretraining/training data and their sizes, and the main capabilities/tasks of these representative models.

Table A1. The summary of representative models focusing on architecture, model size, and training methods.

Modeling Perspective	Model	Model Architecture	Model Size	Training Method
Heterogeneous Architecture	BLIP [75]	visual transformer as in [54] + text transformer [22]	200M	pretraining
	Flamingo [76]	NFNet-F6 → Perceiver Resampler → Gated Cross-Attention + Dense → Chinchilla-3B/9B/80B	3B, 9B, 80B	–
	BLIP-2 [77]	CLIP/Eva-CLIP ViT@224 → Q-Former w/ Linear Projector → Flan-T5/OPT	3B, 4B, 8B	pretraining
	LLaVA [78]	CLIP ViT-L/14 → Linear Projector → Vicuna-7B/13B	7B, 13B	pertaining + fine-tuning
	MiniGPT-4 [82]	Eva-CLIP ViT-G/14 → Q-Former w/ Linear Projector → Vicuna-13B	16B	pertaining + fine-tuning
	mPLUG-Owl [83]	CLIP ViT-L/14 → Cross-attention → LLaMA-7B	7.2B	pertaining + instruction tuning
	Otter [84]	CLIP ViT-L/14 → Cross-attention → LLaMA-7B	9B	finetune the Perceiver resampler module and cross-attention layers
	InstructBLIP [87]	ViT-G/14@224 → Q-Former w/ Linear Projector → Flan-T5/Vicuna-13B	4B-13B	vision-language instruction tuning
	LLaVAR [88]	CLIP ViT-L/14@224 & CLIP ViT-L/14@336 → Linear Projector → Vicuna-13B	13B	pertaining + instruction tuning
	InternLM-XComposer [91]	visual encoder(s EVA-CLIP), perceive sampler, LLM(InternLM)	7B	pre-training + sft
	Qwen-VL [92]	OpenCLIP ViT-bigG → Q-Former → Qwen-7B	9.6B	pretraing+SFT
	MiniGPT-v2 [93]	Eva-CLIP ViT@448 → Linear Projector → LLaMA-2-Chat-7B	7B	pretraining + instruction tuning
	CogVLM [95]	Eva-2-CLIP ViT → MLP → Vicuna-v1.5-7B	17B	pertaining + SFT
	SPHINX [96]	Mixture → Linear → LLaMA-2-13B	7B	pre-training + sft
	LLaVA-1.5 [79]	CLIP ViT-L@336 → MLP → Vicuna-v1.5-7B/13B	7B/13B	pre-training + instruction tuning
	InternVL [71]	InternViT-6B& LLaMA-7B → Cross-attention w/ MLP → QLLaMA-8B & Vicuna-13B	20B	pretraining+SFT
	IDEFICS [90]	OpenCLIP → Cross-attention → LLaMA-65B	9B, 80B	pretraining
	VILA [97]	ViT@336 → Linear Projector → LLaMA-2-7B/13B	7B/13B	pre-training + SFT
	VILA-U [98]	SigLIP → Linear Projector → LLaMA-2-7B/13B	7B/13B	pre-training + SFT
	Deepseek-VL [80]	SigLIP → two-layer hybrid MLP → LLaMA-2-7B/13B	1B/7B	pre-training + SFT
	Ferret [89]	CLIP-ViT-L/14 → Linear Projector → LLaMA-2-7B/13B	7B/13B	pre-training + SFT
	MetaMorph [176]	SigLIP → Linear Projector → LLaMA-3-8B	8B	pre-training + SFT
	MetaQueries [85]	SigLIP → Linear Projector → LLaVA-ov/Qwen2.5-VL	0.5B/3B/7B	pre-training + SFT
	Janus [182]	SigLIP → Linear Projector → DeepSeek-LLM 1.5B	1.5B	pre-training + SFT
	InternLM-XComposer2 [99]	ViT-Large → Partial LoRA → InternLM2-7B-Chat-SFT	7B	Pre-training + supervised fine-tuning
	mPLUG-Owl2 [101]	ViT-L/14 → visual abstractor → LLaMA-2-7B	8.2B	pre-training + visual instruction tuning
	SPHINX-X [102]	DINOv2 + CLIP-ConvNeXt → MoE → LLaMA2-13B/Mixtral-8 × 7B	1.1B–8 × 7B	visual instruction tuning
	Mini-Gemini [103]	Dual Vision Encoder → Guided Generation → VLM	2B-34B	Instruction tuning
	TextHawk [104]	ViT in SigLIP-SO → Resampler → f InternLM-XComposer	7B	Fixed and Mixed Resolution Pre-Training+Mixed Resolution Supervised Fine-Tuning
	TextSquare [105]	CLIP ViT-L14-336 → Projector → InternLM2-7B-ChatSFT	7B	Supervised fine-tuning
	ConvLLaVA [106]	ConvNeXt → two-layer MLP → Vicuna-7B	7B	Three-stage training:projector initialization+vision-language pretraining+visual instruction tuning
	VisionLLM v2 [107]	CLIP-L/14 → Super Link → Vicuna-7B + Task-specific decoders	7B	Mutimodal Training (pre-training and instruction tuning)+Multi-capacity Fine-tuning+Decoder-only Fine-tuning
	MammothModa [108]	ViT vision encoder → MLP Projector+Visual Expert → LLM	–	Three-phase training: vision-language alignment+multi-task pretraining+supervised fine-tuning
	InternLM-XComposer-2.5 [109]	CLIP ViT-L/14 → Partial LoRA → InternLM2-7B	7B	pretraining+finetuning
Unified Language View	X-LXMERT [115]	9 self-attention → 5 cross-attention → 5 self-attention	226.5M	pretraining+finetuning
	DALL-E [118]	dVAE+BPE → decoder-only transformer	12B	pretraining
	CogView [119]	SentencePiece+discrete autoencoder → 48 layers transformer	4B	pretraining+finetuning
	CM3 [120]	dense models as in [268]	medium: 2.7B; large: 13B	pretraining (causally masked objective) + finetuning
	SEED [129]	ViT image encoder → Causal Q-Former → VQ Codebook → Reverse Q-Former → UNet decoder	–	pretraining
	SEED-LLaMA [130]	–	8B, 14B	pretraining+instruction tuning
	BEiT-3 [133]	multiway transformer	1.9B	pre-training
	CM3Leon [134]	CM3	350M, 760M, 7B	CM3 training + retrieval augmentation + supervised fine-tuning
	VL-GPT [135]	CLIP-L image encoder + LLaMA 7B tokenizer → transformer → IP-Adapter + LLaMA 7B detokenizer	7.5B	image tokenizer-detokenizer pre-training + pre-training + instruction tuning
	Emu [136]	Eva-CLIP-1B → Causal Transformer → LLaMA-13B → Stable Diffusion	13B	pre-training+instruction tuning
	Fuyu-8B [269]	–	–	–
	LaVIT [127]	ViT → Cross-attention → LLaMA-7B	7B	tokenizer training + VLM pre-training
	AnyGPT [131]	SEED+SpeechTokenizer+Encodec → LLaMA-2 7B	7B	pre-training+fine-tuning
	SEED-X [137]	–	–	pre-training+instruction tuning
	Morph-tokens [128]	morph-tokenizer+text tokenizer → MLLM → decoders	7B	initialization+morph-tokenizer training+instruction tuning
	SETOKIM [116]	–	7B, 13B	tokenizer pretraining+multimodal pretraining+instruction tuning
	Emu2 [175]	EVA-02-CLIP-E-plus → LLaMA-33B → SDXL	37B	pre-training+instruction tuning
	SOLO [138]	linear projector → Mistral-7B	7B	pretraining+instruction tuning
	Chameleon [139]	VQ-SEG+BPE → transformers	7B, 34B	pre-training+supervised fine-tuning
	ANOLE [140]	VQ-SEG+BPE → transformers	7B	pre-training+supervised fine-tuning+innovative fine-tuning
	SEED-Story [141]	ViT tokenizer → MLLM → SD-XL de-tokenizer	–	Three-stage training:pre-train SD-XL-based de-tokenizer+instruction tuning+de-tokenizer adaptation
	EVE [270]	Patch embedding layer → Decoder-only architecture → Patch aligning layer	7B	LLM-guided Pre-training+Generative Pre-training+Supervised Fine-tuning
	Show-o [183]	diffusion-quantized image tokens + text tokens → autoregressive transformer	1.3B	Pretraining: autoregressive language modeling + discrete diffusion objective
	Emu3 [207]	discrete image / text / video tokens → transformer	–	next-token pure autoregressive training
	Janus [177]	vision encoder → dual token streams (understanding and generation) → autoregressive transformer	1.3B	stage-wise pretraining on mixed text, vision-understanding, generation data
	MetaMorph [176]	pretrained LLM (text tokens) + vision encoder embeddings → visual-predictive instruction tuning → outputs discrete text + continuous visual tokens	–	instruction tuning (VPiT) on interleaved image–text QA and generation data
	Orthus [208]	text tokens + continuous image features → shared AR transformer → modality-specific heads: LM head → text; diffusion head → image	7B	AR pretraining with mixed autoregressive + diffusion objectives
	JetFormer [125]	normalizing flow → “soft-token” image representation + text tokens → decoder-only Transformer → soft-tokens (images) and text tokens	–	end-to-end max-likelihood training (teacher forcing) on raw image and text data
	LMFusion [86]	frozen llama-3 text modules + parallel diffusion-based vision modules → shared self-attention → interleaved multimodal outputs	–	freeze text weights, train image modules (diffusion + modality-spec. layers) on image–text data
Unified Visual View	In-image NMT via pixel supervision [151]	2 convolutional encoder → self-attention encoder → convolutional decoder	–	End-to-End training
	Visual text representation [152]	Visual text embedder → Standard Transformer(encoder-decoder)	–	End-to-End training
	DiT [160]	Vanilla Transformer architecture	–	Pre-training(Masked Image Modeling)+Finetuning
	Donut [161]	Visual encoder(Swin-B)+decoder(BART)	–	pretraining+finetuning
	PIXEL [153]	MAE(ViT encoder+lightweight decoder)	86M	pretraining+finetuning
	Text rendering in PIXEL [154]	–	–	–
	CLIPPO [162]	ViT-B/16, ViT-L/16	93M, 316M	Image-text contrastive pretraining+text/text co-training
	Pix2Struct [163]	Image-encoder-text-decoder based on ViT	base:282M, large:1.3B	screenshot parsing pretraining+finetuning
	PHD [164]	ViT-MAE	86M	pretraining+finetuning
	PixT3 [165]	Pix2Struct-Base	282M	Self-supervised pretraining(structure learning curriculum) +finetuning
	PIXAR [155]	Decoder-only Transformer with 12 layers	85M(classification), 113M(generation)	2-stage pretraining: MLE+Adversarial
	Visual sentence learning [156]	MAE(ViT encoder+lightweight decoder)	86M	Unsupervised visual contrastive learning with iterative training+NLI supervised learning
	SEEKER [173]	SigLIP-L & SAM-B+DeepSeek-VL [80]	1.3B/7B	Supervised fine-tuning+instruction tuning
	VisInContext [174]	Flamingo	1.4B/7B/70B	pre-training:token masking & text-centric contrastive learning

Table A2. The summary of representative models focusing on pretrain data/data size, and main capabilities/tasks.

Modeling Perspective	Model	Pretrain Data Size	Main Tasks/Capabilities
Heterogeneous Architecture	BLIP [75]	Pretrain data: 129M	image-text retrieval, image captioning, and VQA
	Flamingo [76]	–	Few-shot learning, VQA, image captioning, visual dialogues, video understanding, open-ended text generation
	BLIP-2 [77]	Pretrain data: 129M	image-to-text generation, VQA, and image-text retrieval
	LLaVA [78]	Pretrain data: 595K image-text pairs from CC3M	Visual QA, image captioning, multimodal conversations, Science QA, general-purpose vision-language tasks
	MiniGPT-4 [82]	Pretrain data: 5 million image-text pairs	image captioning, advanced abilities
	mPLUG-Owl [83]	Pretrain data: 104 billion training tokens	VQA, conversation, reasoning, and joke comprehension
	Otter [84]	2.8 million multimodal instruction-response pairs	following user instructions and multi-modal in-context learning
	InstructBLIP [87]	Pretrain data: 129M, Instruction tuning data: 1.2M	VQA, image captioning, science QA, visual dialogues, knowledge-grounded image description
	LLaVAR [88]	595K pre-training data from LLaVA with 422K noisy instruction-following data	VQA
	InternLM-XComposer [91]	1.1 billion images alongside 77.7 billion text tokens	text generation, image spotting and captioning, image retrieval and selection
	Qwen-VL [92]	Pretrain data: 1.4B, Instruction tuning data: 50M	VQA, image captioning, text reading, visual dialogues, OCR, multilingual support
	MiniGPT-v2 [93]	19, 14, 18 datasets for pre-training, multi-task training and instruction tuning	grounded image caption + object parsing and grounding
	CogVLM [95]	1.5B images	VQA, image captioning, and visual grounding
	SPHINX [96]	LAION-400M, LAION-COCO, and RefinedWeb	VQA, Visual grounding
	LLaVA-1.5 [79]	Pretrain data: 0.6M, Instruction tuning data: 0.7M	instruction following vlm, VQA, Visual grounding, conversation
	InternVL [71]	4.98 billion image-text pairs	zero-shot image/video classification and zero-shot image-text retrieval
	IDEFICS [90]	OBELICS dataset	VQA, OCR
	LMFusion [186]	Instruction tuning data: 380M	VQA, image captioning, image generation, visual dialogues, OCR
	Ferret [89]	Instruction tuning data: RefCOCO, Flickr30k	VQA, image captioning, visual dialogues, OCR
	MetaMorph [176]	Instruction tuning data: ∼15M	VQA, image captioning, image generation, visual dialogues, OCR
	MetaQueries [85]	Pretrain data: 25M, Instruction tuning data: 2.4M	VQA, image captioning, image generation, visual dialogues, OCR
	Deepseek-VL [80]	Instruction tuning data: 2.75M	VQA, image captioning, visual dialogues, OCR
	VILA [97]	Pretrain data: 50M, Instruction tuning data: 1M	VQA, image captioning, visual dialogues, OCR
	VILA-U [98]	Pretrain data: 50M, Instruction tuning data: 22M	VQA, image captioning, image generation, visual dialogues, OCR
	InternLM-XComposer2 [99]	11 and 16 public datasets for pretraining and SFT respectively, 4 datasets and in-house data for instruction tuning	Multimodal understanding
	mPLUG-Owl2 [101]	400 million image-text pairs	Image Caption and VQA
	SPHINX-X [102]	–	OCR, VQA, object detection, image captioning, visual programming, bilingual support
	Mini-Gemini [103]	1.5M instruction-related conversations for image comprehension, 13K pairs for image-related generation	VQA, image-text generation, vision-language tasks
	TextHawk [104]	Conceptual and grounding captioning: 96M, 16M image-text pairs, OCR:1.28M images, Markdown:1.28M PDF pages,Instruction data, DocGemini:30K images and 195K QA pairs	Document understanding and referring tasks
	TextSquare [105]	Square-10M and in-domain datasets	General VQA and Hallucination Evaluation Benchmark
	ConvLLaVA [106]	Projector initialization: 2M+VLP:2.9M+Instruction tuning: 0.665M	Vision-language benchmarks, image grounding benchmarks (RefCOCO, RefCOCO+, RefCOCOg)
	VisionLLM v2 [107]	Instruction data for image-level and region-level VQA (ShareGPT4V, All-Seeing and VQAv2); Finetune data: COCO, RefCOCO/+/g, LAION-Aesthetics	Vision-Language tasks, VQA, image generation, image editing, pose estimation, object detection
	MammothModa [108]	Pretrain data: bilingual captions, interleaved text-image pairs, object grounding, OCR grounding, and video captions	Visual language benchmarks
	InternLM-XComposer-2.5 [109]	8+1+8 datasets	video understanding, multi-image multi-tune dialog, image-to-text generation
Unified Language View	X-LXMERT [115]	PT: 180K images and 9.18M sentences	text-to-image generation
	DALL-E [118]	250M text-image pairs	zero-shot text-to-image generation
	CogView [119]	30M Chinese text-image pairs	text-to-image generation
	CM3 [120]	45M docs from CC-News + 16M from En-Wikipedia	text-to-image generation
	SEED [129]	5M image-text pairs from CC3M, Unsplash, and COCO	image-to-text and text-to-image
	SEED-LLaMA [130]	–	image-to-text and text-to-image
	BEiT-3 [133]	21M image-text pairs, 14M images, 160GB documents	Vision, Vision-Language benchmarks
	CM3Leon (pronounced “Chameleon”) [134]	340M	text-to-image and image-to-text generation
	VL-GPT [135]	3 + 2 + 5 datasets	image-text understanding and text-to-image generation
	Emu [136]	LAION-2B, LAION-COCO, MMC4, WebVid-10M, YT-Storyboard-1B	image-to-text and text-to-image generation, in-context image and text generation
	Fuyu-8B [269]	–	–
	LaVIT [127]	93M samples from Conceptual Caption, SBU, BLIP-Capfilt; 100M image-text pairs from LAION-Aesthetics	multimodal understanding and generation
	AnyGPT [131]	–	image, speech, music understanding and generation
	SEED-X [137]	–	multimodal comprehension and generation
	Morph-tokens [128]	30M image-text pairs	multimodal comprehension and generation
	SETOKIM [116]	4 + 3 + 8 datasets	Visual Understanding, Referring Expression Segmentation, Visual Generation and Editing
	Emu2 [175]	LAION-2B, CapsFusion-120M, WebVid-10M, Multimodal-C4 (MMC4), YT-Storyboard-1B, GRIT-20M, CapsFusion-grounded-100M, Pile	in context multimodal tasks
	SOLO [138]	–	–
	Chameleon [139]	2.9T text-only tokens+1.5T text-image tokens+400B text/image interleaved tokens	text understanding, text generation, and multimodal comprehension
	ANOLE [140]	Chameleon data+5859 images from LAION-5B	multimodal comprehension and generation
	SEED-Story [141]	StoryStream: 25.79K	Multimodal story generation, long story generation
	EVE [270]	35M	Vision-language, hallucination, and open-world multi-modal understanding benchmarks
	Show-o [183]	35 M image–text pairs	VQA, text-to-image generation, text-guided inpainting/expansion, interleaved multimodal generation, video understanding, generation
	Emu3 [207]	–	image/text/video generation and perception (e.g. high-fidelity image synthesis, video prediction, visual QA)
	Janus [177]	500 B text tokens + ImageNet-k	multimodal understanding (VQA, image captioning) and generation (text ⟶ image)
	MetaMorph [176]	instruction tuning: 200 K visual-generation samples + ≫ 1 M visual-understanding pairs	visual understanding (VQA) and generation (in-context image synthesis), emergent multimodal reasoning via VPiT
	Orthus [208]	–	interleaved image–text generation, high-fidelity image synthesis, visual question answering, long-form mixed-modality content generation
	JetFormer [125]	imageNet-1K images (text-to-image) + webLI image–text pairs (image-to-text)	Text-to-image generation, image understanding (VQA), robust likelihood estimation on raw data
	LMFusion [86]	llama-3’s original text corpus	parallel visual generation and understanding, preserves LLM language performance, efficient adaptation of text-only LLM to multimodal tasks
Unified Visual View	In-image NMT via pixel supervision [151]	Train:4.5M	Image-to-Image translation
	iGPT [146]	ImageNet (low-resolution)	Autoregressive pretraining for general image representation
	Visual text representation [152]	German–English train: 4.9M, Chinese–English train:8.7M	Machine translation
	DiT [160]	42M document images	Vision-based Document AI benchmarks: table detection, document layout analysis, document image classification and text detection
	Donut [161]	2M synthetic and 11M IIT-CDIP scanned document images	Document classification and information extraction, Document Visual Question Answering
	PIXEL [153]	3.1B words(rendered into 16.8M image examples)	Discriminative tasks(POS tagging, NER, GLUE)
	BPE-iGPT [147]	Multilingual Wikipedia and BookCorpus (text rendered into images)	Efficient pixel-based language modeling by compressing pixel sequences
	Text rendering in PIXEL [154]	–	–
	CLIPPO [162]	WebLI (10 B images with 12 B corresponding alt-texts), C4	Vision and language capabilities: VQA, GLUE; Multilingual capabilities: CrossModal3600
	Pix2Struct [163]	80M(constructed from C4 corpus)	Illustrations: ChartQA, AI2D, OCR-VQA; User interfaces: RefExp, WidgetCap, Screen2Words; Natural images: TextCaps; Documents: DocVQA, InfographicVQA
	PHD [164]	Generated Pretraining Data from the BookCorpus and the English Wikipedia; Real Historical Newspaper Scans	Visual SQuAD, Historical QA, GLUE
	PixT3 [165]	Synthetic image-to-text dataset: 0.135M	Visual data-to-text generation: Logic2Text, ToTTo
	PIXAR [155]	3.1B words(rendered into 16.8M image examples)	Discriminative(GLUE)+Generative(QA) tasks
	Visual sentence learning [156]	7.6M	Natural language semantics learning(e.g., STS)
	SEEKER [173]	Instruction data:CC3M [78], COCO [219] and arXiv PDFs	Multimodal Understanding, long-form multi-image input, long-form text output
	VisInContext [174]	180M subset of DataComp1B, MMC4, OBELICS, and OCR Rendered Text	Long-context multimodal understanding & sequential multimodal retrieval
	D-iGPT [148]	ImageNet-1K	Learning strong visual representations by predicting semantic tokens
	TextHarmony [159]	General V+L datasets + fine-tuning on DetailedTextCaps-100K	Unified visual text comprehension (VQA) and generation
	VisionLLM v2 [107]	Diverse multimodal task datasets	Generalist MLLM for perception, understanding, and generation
	XTRA [149]	ImageNet	Sample and parameter-efficient autoregressive image generation
	PixelFlow [150]	ImageNet, CelebA-HQ	High-quality image generation in pixel space via flow-based modeling
	SegAgent [168]	Task-specific segmentation datasets	Interactive image segmentation via generating textual click instructions
	MedPLIB [169]	Biomedical image-text data (e.g., ROCO, PathVQA)	Biomedical VQA and pixel-level grounding
	DocLayLLM [170]	Large-scale document image datasets	Document understanding via visual layout and OCR text tokens
	Marten [171]	VQA and visual grounding datasets	Improving spatial alignment via VQA with mask-generation pre-training

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 261–272.
2. Tong, S.; Liu, Z.; Zhai, Y.; et al. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 9568–9578.
3. Zhang, D.; Yu, Y.; Dong, J.; et al. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*; Ku, L.W.; Srikumar, V., Eds.; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 12401–12430.
4. Caffagni, D.; Cocchi, F.; Barsellotti, L.; et al. The Revolution of Multimodal Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics ACL 2024*; Ku, L.W., Srikumar, V., Eds.; Association for Computational Linguistics: Bangkok, Thailand, 2024; pp. 13590–13618.
5. Li, C.; Gan, Z.; Yang, Z.; et al. Multimodal foundation models: From specialists to general-purpose assistants. *Found. Trends Comput. Graph. Vis.* **2024**, *16*, 1–214.
6. Bai, T.; Liang, H.; Wan, B.; et al. A Survey of Multimodal Large Language Model from A Data-centric Perspective. *arXiv* **2024**, arXiv:2405.16640.
7. Qin, Z.; Chen, D.; Zhang, W.; et al. The Synergy between Data and Multi-Modal Large Language Models: A Survey from Co-Development Perspective. *arXiv* **2024**, arXiv:2407.08583.
8. Jin, Y.; Li, J.; Liu, Y.; et al. Efficient multimodal large language models: A survey. *arXiv* **2024**, arXiv:2405.10739.
9. Mai, X.; Tao, Z.; Lin, J.; et al. From Efficient Multimodal Models to World Models: A Survey. *arXiv* **2024**, arXiv:2407.00118.
10. Huh, M.; Cheung, B.; Wang, T.; et al. The platonic representation hypothesis. *arXiv* **2024**, arXiv:2405.07987.
11. Jelinek, F. *Statistical Methods for Speech Recognition*; MIT Press: Cambridge, MA, USA, 1998.
12. Manning, C.; Schutze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, USA, 1999.
13. Das, B.C.; Amini, M.H.; Wu, Y. Security and privacy challenges of large language models: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–39.
14. Zhao, W.X.; Zhou, K.; Li, J.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
15. Juang, B.; Rabiner, L. Hidden Markov Models for Speech Recognition. *Technometrics*, **1991**, *33(3)*, 251–272.
16. Chen, S.F.; Goodman, J. An Empirical Study of Smoothing Techniques for Language Modeling. *Comput. Speech Lang.* **1999**, *13*, 359–394.
17. Bengio, Y.; Ducharme, R.; Vincent, P.; et al. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
18. Schwenk, H.; Dechelotte, D.; Gauvain, J.L. Continuous Space Language Models for Statistical Machine Translation. In Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Sydney, Australia, 17–21 July 2006; Association for Computational Linguistics: Sydney, Australia, 2006; pp. 723–730.
19. Mikolov, T.; Karafiát, M.; Burget, L.; et al. Recurrent neural network based language model. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, Makuhari, Chiba, Japan, 26–30 September 2010; pp. 1045–1048.
20. Qiu, X.; Sun, T.; Xu, Y.; et al. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897.
21. Peters, M.E.; Neumann, M.; Iyyer, M.; et al. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.
22. Devlin, J.; Chang, M.W.; Lee, K.; et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MI, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Minneapolis, MI, USA, 2019; pp. 4171–4186.
23. Liu, Y.; Ott, M.; Goyal, N.; et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* **2019**.
24. He, P.; Liu, X.; Gao, J.; et al. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv* **2020**, arXiv:2006.03654.
25. Radford, A.; Narasimhan, K; Salimans, T.; et al. Improving Language Understanding by Generative Pre-Training. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 30 December 2018)
26. Dao, T.; Fu, D.; Ermon, S.; et al. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16344–16359.
27. Lewis, M.; Liu, Y.; Goyal, N.; et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, 5–10 July 2020; pp. 7871–7880.
28. Raffel, C.; Shazeer, N.M.; Roberts, A.; et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* **2019**, *21*, 140:1–140:67.
29. Henighan, T.; Kaplan, J.; Katz, M.; et al. Scaling laws for autoregressive generative modeling. *arXiv* **2020**,

arXiv:2010.14701

30. Wei, J.; Tay, Y.; Bommasani, R.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682.
31. Brown, T.B.; Mann, B.; Ryder, N.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020.
32. OpenAI. GPT-4 Technical Report. *arXiv* **2023**, arXiv:2303.08774.
33. Chowdhery, A.; Narang, S.; Devlin, J.; et al. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.* **2023**, *24*, 240:1–240:113.
34. Touvron, H.; Lavril, T.; Izacard, G.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
35. Taori, R.; Gulrajani, I.; Zhang, T.; et al. Stanford alpaca: An instruction-following llama model. Available online: https://crfm.stanford.edu/2023/03/13/alpaca.html (accessed on 13 March 2023)
36. Chiang, W.L.; Li, Z.; Lin, Z.; et al. Vicuna: An Open-Source Chatbot Impressing Gpt-4 with 90%* Chatgpt Quality. Available online: https://vicuna.lmsys.org (accessed on 14 April 2023)
37. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.; Li, Y.F.; et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *arXiv* **2023**, arXiv:2303.12712.
38. Chen, M.; Tworek, J.; Jun, H.; et al. Evaluating large language models trained on code. *arXiv* **2021**, arXiv:2107.03374.
39. Roziere, B.; Gehring, J.; Gloeckle, F.; et al. Code llama: Open foundation models for code. *arXiv* **2023**, arXiv:2308.12950.
40. Tu, T.; Azizi, S.; Driess, D.; et al. Towards generalist biomedical AI. *Nejm Ai* **2024**, *1*, AIoa2300138.
41. Taylor, R.; Kardas, M.; Cucurull, G.; et al. Galactica: A large language model for science. *arXiv* **2022**, arXiv:2211.09085.
42. Guo, D.; Yang, D.; Zhang, H.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.
43. Peng, B.; Alcaide, E.; Anthony, Q.; et al. RWKV: Reinventing RNNs for the Transformer Era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*; Bouamor, H., Pino, J., Bali, K., Eds.; Association for Computational Linguistics: Singapore, 2023; pp. 14048–14077.
44. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
45. Beck, M.; Pöppel, K.; Spanring, M.; et al. xlstm: Extended long short-term memory. *arXiv* **2024**, arXiv:2405.04517.
46. Lowe, D. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
47. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2005**, *1*, 886–893.
48. Mubarak, R.; Alsboui, T.A.A.; Alshaikh, O.; et al. A Survey on the Detection and Impacts of Deepfakes in Visual, Audio, and Textual Formats. *IEEE Access* **2023**, *11*, 144497–144529.
49. LeCun, Y.; Bottou, L.; Bengio, Y.; et al. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 3–6.
51. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
52. He, K.; Zhang, X.; Ren, S.; et al. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. d'Ascoli, S.; Touvron, H.; Leavitt, M.L.; et al. ConViT: improving vision transformers with soft convolutional inductive biases. *J. Stat. Mech. Theory Exp.* **2021**, *2022*, 2286–2296.
54. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
55. Dai, Z.; Liu, H.; Le, Q.V.; et al. CoAtNet: Marrying Convolution and Attention for All Data Sizes. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 3965–3977.
56. Liu, Z.; Lin, Y.; Cao, Y.; et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
57. Wang, W.; Xie, E.; Fan, D.P.; et al. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 548–558.
58. Fan, H.; Xiong, B.; Mangalam, K.; et al. Multiscale vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 6824–6835.
59. Yuan, L.; Chen, Y.; Wang, T.; et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 558–567.
60. Han, K.; Xiao, A.; Wu, E.; et al. Transformer in Transformer. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15908–15919.

61. Wang, W.; Chen, W.; Qiu, Q.; et al. Crossformer++: A versatile vision transformer hinging on cross-scale attention. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 3123–3136.
62. Xu, Y.; Zhang, Q.; Zhang, J.; et al. Vitae: Vision transformer advanced by exploring intrinsic inductive bias. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 28522–28535.
63. Yuan, K.; Guo, S.; Liu, Z.; et al. Incorporating convolution designs into visual transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 579–588.
64. Graham, B.; El-Nouby, A.; Touvron, H.; et al. Levit: A vision transformer in convnet's clothing for faster inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12259–12269.
65. Radford, A.; Kim, J.W.; Hallacy, C.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PmLR, Virtual,18–24 July 2021; pp. 8748–8763.
66. Wortsman, M.; Ilharco, G.; Kim, J.W.; et al. Robust fine-tuning of zero-shot models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 7959–7971.
67. Brock, A.; De, S.; Smith, S.L.; et al. High-performance large-scale image recognition without normalization. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual,18–24 July 2021; pp. 1059–1071.
68. Qiu, L.; Zhang, R.; Guo, Z.; et al. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv* **2021**, arXiv:2112.02399.
69. Fang, Y.; Wang, W.; Xie, B.; et al. Eva: Exploring the limits of masked visual representation learning at scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19358–19369.
70. Liu, Z.; Mao, H.; Wu, C.Y.; et al. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986.
71. Chen, Z.; Wu, J.; Wang, W.; et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 24185–24198.
72. Van Den Oord, A.; Vinyals, O. Neural discrete representation learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
73. Yu, Y.; Kim, Y.; Ahn, S.; et al. MAGVIT: Masked Generative Video Transformer. *arXiv* **2022**, arXiv:2206.11894.
74. Oquab, M.; Darcet, T.; Moutakanni, T.; et al. DINOv2: Learning Robust Visual Features without Supervision. *arXiv* **2023**, arXiv:2304.07193.
75. Li, J.; Li, D.; Xiong, C.; et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning. PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 12888–12900.
76. Alayrac, J.B.; Donahue, J.; Luc, P.; et al. Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736.
77. Li, J.; Li, D.; Savarese, S.; et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning. PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
78. Liu, H.; Li, C.; Wu, Q.; et al. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916.
79. Liu, H.; Li, C.; Li, Y.; et al. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 26296–26306.
80. Lu, H.; Liu, W.; Zhang, B.; et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv* **2024**, arXiv:2403.05525.
81. Wu, Z.; Chen, X.; Pan, Z.; et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv* **2024**, arXiv:2412.10302.
82. Zhu, D.; Chen, J.; Shen, X.; et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
83. Ye, Q.; Xu, H.; Xu, G.; et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv* **2023**, arXiv:2304.14178.
84. Li, B.; Zhang, Y.; Chen, L.; et al. Otter: A Multi-Modal Model with In-Context Instruction Tuning. *arXiv* **2023**, arXiv:2305.03726
85. Pan, X.; Shukla, S.N.; Singh, A.; et al. Transfer between modalities with metaqueries. *arXiv* **2025**, arXiv:2504.06256.
86. Shi, W.; Han, X.; Zhou, C.; et al. LlamaFusion: Adapting Pretrained Language Models for Multimodal Generation. *arXiv* **2024**, arXiv:2412.15188.
87. Dai, W.; Li, J.; Li, D.; et al. InstructBLIP: Towards General-purpose VisionLanguage Models with Instruction Tuning. *arXiv* **2023**, arXiv:2305.06500.
88. Zhang, Y.; Zhang, R.; Gu, J.; et al. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv* **2023**, arXiv:2306.17107.
89. You, H.; Zhang, H.; Gan, Z.; et al. Ferret: Refer and ground anything anywhere at any granularity. *arXiv* **2023**,

arXiv:2310.07704.

90. Laurençon, H.; Saulnier, L.; Tronchon, L.; et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 71683–71702.
91. Zhang, P.; Wang, X.D.B.; Cao, Y.; et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv* **2023**, arXiv:2309.15112.
92. Bai, J.; Bai, S.; Yang, S.; et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond *arXiv* **2023**, arXiv:2308.12966.
93. Chen, J.; Zhu, D.; Shen, X.; et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv* **2023**, arXiv:2310.09478.
94. Xie, Z. OMG-LLaVA: Bridging Image-level, Object-level, Pixel-level Reasoning and Understanding. *arXiv* **2024**, arXiv:2404.07143.
95. Wang, W.; Lv, Q.; Yu, W.; et al. Cogvlm: Visual expert for pretrained language models. *arXiv* **2023**, arXiv:2311.03079.
96. Lin, Z.; Liu, C.; Zhang, R.; et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv* **2023**, arXiv:2311.07575.
97. Lin, J.; Yin, H.; Ping, W.; et al. Vila: On pre-training for visual language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 26689–26699.
98. Wu, Y.; Zhang, Z.; Chen, J.; et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv* **2024**, arXiv:2409.04429.
99. Dong, X.; Zhang, P.; Zang, Y.; et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv* **2024**, arXiv:2401.16420.
100. Liu, H.; Li, C.; Li, Y.; et al. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. Available online: https://llava-vl.github.io/blog/2024-01-30-llava-next/ (accessed on 30 January 2024)
101. Ye, Q.; Xu, H.; Ye, J.; et al. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 13040–13051.
102. Gao, P.; Zhang, R.; Liu, C.; et al. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv* **2024**, arXiv:2402.05935.
103. Li, Y.; Zhang, Y.; Wang, C.; et al. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv* **2024**, arXiv:2403.18814.
104. Yu, Y.Q.; Liao, M.; Wu, J.; et al. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *arXiv* **2024**, arXiv:2404.09204.
105. Tang, J.; Lin, C.; Zhao, Z.; et al. TextSquare: Scaling up Text-Centric Visual Instruction Tuning. *arXiv* **2024**, arXiv:2404.12803.
106. Ge, C.; Cheng, S.; Wang, Z.; et al. ConvLLaVA: Hierarchical Backbones as Visual Encoder for Large Multimodal Models. *arXiv* **2024**, arXiv:2405.15738.
107. Wu, J.; Zhong, M.; Xing, S.; et al. VisionLLM v2: An End-to-End Generalist Multimodal Large Language Model for Hundreds of Vision-Language Tasks. *arXiv* **2024**, arXiv:2406.08394.
108. She, Q.; Pan, J.; Wan, X.; et al. MammothModa: Multi-Modal Large Language Model. *arXiv* **2024**, arXiv:2406.18193.
109. Zhang, P.; Dong, X.; Zang, Y.; et al. InternLM-XComposer-2.5: A Versatile Large Vision Language Model Supporting Long-Contextual Input and Output. *arXiv* **2024**, arXiv:2407.03320.
110. Zhang, S.; Roller, S.; Goyal, N.; et al. Opt: Open pre-trained transformer language models. *arXiv* **2022**, arXiv:2205.01068.
111. Hoffmann, J.; Borgeaud, S.; Mensch, A.; et al. Training compute-optimal large language models. *arXiv* **2022**, arXiv:2203.15556.
112. Touvron, H.; Martin, L.; Stone, K.; et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288.
113. Bai, J.; Bai, S.; Chu, Y.; et al. Qwen technical report. *arXiv* **2023**, arXiv:2309.16609.
114. Chung, H.W.; Hou, L.; Longpre, S.; et al. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.* **2024**, *25*, 1–53.
115. Cho, J.; Lu, J.; Schwenk, D.; et al. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 8785–8805.
116. Wu, S.; Fei, H.; Li, X.; et al. Towards Semantic Equivalence of Tokenization in Multimodal LLM. *arXiv* **2024**, arXiv:2406.05127.
117. Ma, C.; Jiang, Y.; Wu, J.; et al. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2024; pp. 417–435.
118. Ramesh, A.; Pavlov, M.; Goh, G.; et al. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning. Pmlr, Virtual,18–24 July 2021; pp. 8821–8831.
119. Ding, M.; Yang, Z.; Hong, W.; et al. Cogview: Mastering text-to-image generation via transformers. *Adv. Neural Inf.*

Process. Syst. **2021**, *34*, 19822–19835.

120. Aghajanyan, A.; Huang, B.; Ross, C.; et al. Cm3: A causal masked multimodal model of the internet. *arXiv* **2022**, arXiv:2201.07520.
121. Li, X.; Qiu, K.; Chen, H.; et al. XQ-GAN: An Open-source Image Tokenization Framework for Autoregressive Generation. *arXiv* **2024**, arXiv:2412.01762.
122. Zheng, S.; Zhou, B.; Feng, Y.; et al. Unicode: Learning a unified codebook for multimodal large language models. In *European Conference on Computer Vision*; Springer: Berlin, Germany, 2024; pp. 426–443.
123. Tang, H.; Liu, H.; Xiao, X. Ugen: Unified autoregressive multimodal model with progressive vocabulary learning. *arXiv* **2025**, arXiv:2503.21193.
124. Xie, R.; Du, C.; Song, P.; et al. MUSE-VL: Modeling Unified VLM through Semantic Discrete Encoding. *arXiv* **2024**, arXiv:2411.17762.
125. Tschannen, M.; Pinto, A.S.; Kolesnikov, A. JetFormer: An autoregressive generative model of raw images and text. *arXiv* **2024**, arXiv:2411.19722.
126. Zhang, W.; Xie, Z.; Feng, Y.; et al. From Pixels to Tokens: Byte-Pair Encoding on Quantized Visual Modalities. *arXiv* **2024**, arXiv:2410.02155.
127. Jin, Y.; Xu, K.; Xu, K.; et al. Unified Language-Vision Pretraining in LLM with Dynamic Discrete Visual Tokenization. *arXiv* **2024**, arXiv:2309.04669.
128. Pan, K.; Tang, S.; Li, J.; et al. Auto-Encoding Morph-Tokens for Multimodal LLM. In Proceedings of the Forty-First International Conference on Machine Learning, 2024.
129. Ge, Y.; Ge, Y.; Zeng, Z.; et al. Planting a seed of vision in large language model. *arXiv* **2023**, arXiv:2307.08041.
130. Ge, Y.; Zhao, S.; Zeng, Z.; et al. Making llama see and draw with seed tokenizer. *arXiv* **2023**, arXiv:2310.01218.
131. Zhan, J.; Dai, J.; Ye, J.; et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv* **2024**, arXiv:2402.12226.
132. Fang, R.; Duan, C.; Wang, K.; et al. Puma: Empowering unified mllm with multi-granular visual generation. *arXiv* **2024**, arXiv:2410.13861.
133. Wang, W.; Bao, H.; Dong, L.; et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19175–19186.
134. Yu, L.; Shi, B.; Pasunuru, R.; et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv* **2023**, arXiv:2309.02591.
135. Zhu, J.; Ding, X.; Ge, Y.; et al. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv* **2023**, arXiv:2312.09251.
136. Sun, Q.; Yu, Q.; Cui, Y.; et al. Emu: Generative pretraining in multimodality. In Proceedings of the Twelfth International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
137. Ge, Y.; Zhao, S.; Zhu, J.; et al. Seed-x: Multimodal models with unified multi-granularity comprehension and generation. *arXiv* **2024**, arXiv:2404.14396.
138. Chen, Y.; Wang, X.; Peng, H.; et al. A Single Transformer for Scalable Vision-Language Modeling. *arXiv* **2024**, arXiv:2407.06438.
139. Team, C. Chameleon: Mixed-modal early-fusion foundation models. *arXiv* **2024**, arXiv:2405.09818.
140. Chern, E.; Su, J.; Ma, Y.; et al. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv* **2024**, arXiv:2407.06135.
141. Yang, S.; Ge, Y.; Li, Y.; et al. SEED-Story: Multimodal Long Story Generation with Large Language Model. *arXiv* **2024**, arXiv:2407.08683.
142. Lu, J.; Clark, C.; Lee, S.; et al. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024, pp. 26439–26455.
143. Moon, S.; Madotto, A.; Lin, Z.; et al. Anymal: An efficient and scalable any-modality augmented language model. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, Miami, FL, USA, 12–16 November 2024, pp. 1314–1332.
144. Shukor, M.; Dancette, C.; Rame, A.; et al. UnIVAL: Unified Model for Image, Video, Audio and Language Tasks. *arXiv* **2023**, arXiv:2307.16184.
145. Zhu, Q.; Zhou, L.; Zhang, Z.; et al. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Trans. Multimed.* **2024**, *26*, 1055–1064.
146. Chen, M.; Radford, A.; Child, R.; et al. Generative pretraining from pixels. In Proceedings of the 37th International Conference on Machine Learning, Virtual, 13–18 July 2020; pp. 1691–1703.
147. Razzhigaev, A.; Voronov, A.; Kaznacheev, A.; et al. Pixel-Level BPE for Auto-Regressive Image Generation. In Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models, Gyeongju, South Korea, 12–17 October 2022; pp. 26–30.

148. Ren, S.; Wang, Z.; Zhu, H.; et al. Rejuvenating image-GPT as Strong Visual Representation Learners. *arXiv* **2024**, arXiv:2312.02147.

149. Amrani, E.; Karlinsky, L.; Bronstein, A. Sample- and Parameter-Efficient Auto-Regressive Image Models. *arXiv* **2025**, arXiv:2411.15648.

150. Chen, S.; Ge, C.; Zhang, S.; et al. PixelFlow: Pixel-Space Generative Models with Flow. *arXiv* **2025**, arXiv:2504.07963.

151. Mansimov, E.; Stern, M.; Chen, M.; Firat, O.; Uszkoreit, J.; Jain, P. Towards end-to-end in-image neural machine translation. *arXiv* **2020**, arXiv:2010.10648.

152. Salesky, E.; Etter, D.; Post, M. Robust open-vocabulary translation from visual text representations. *arXiv* **2021**, arXiv:2104.08211.

153. Rust, P.; Lotz, J.F.; Bugliarello, E.; et al. Language modelling with pixels. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.

154. Lotz, J.; Salesky, E.; Rust, P.; et al. Text Rendering Strategies for Pixel Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 10155–10172.

155. Tai, Y.; Liao, X.; Suglia, A.; et al. PIXAR: Auto-Regressive Language Modeling in Pixel Space. *arXiv* **2024**, arXiv:2401.03321.

156. Xiao, C.; Huang, Z.; Chen, D.; et al. Pixel Sentence Representation Learning. *arXiv* **2024**, arXiv:2402.08183.

157. Gao, T.; Wang, Z.; Bhaskar, A.; et al. Improving Language Understanding from Screenshots. *arXiv* **2024**, arXiv:2402.14073.

158. Li, W.; Li, G.; Lan, Z.; et al. Empowering Backbone Models for Visual Text Generation with Input Granularity Control and Glyph-Aware Training. *arXiv* **2024**, arXiv:2410.04439.

159. Zhao, Z.; Tang, J.; Wu, B.; et al. Harmonizing Visual Text Comprehension and Generation. *arXiv* **2024**, arXiv:2407.16364.

160. Li, J.; Xu, Y.; Lv, T.; et al. Dit: Self-supervised pre-training for document image transformer. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 3530–3539.

161. Kim, G.; Hong, T.; Yim, M.; et al. Ocr-free document understanding transformer. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 498–517.

162. Tschannen, M.; Mustafa, B.; Houlsby, N. Clippo: Image-and-language understanding from pixels only. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 11006–11017.

163. Lee, K.; Joshi, M.; Turc, I.R.; et al. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 18893–18912.

164. Borenstein, N.; Rust, P.; Elliott, D.; et al. PHD: Pixel-Based Language Modeling of Historical Documents. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 87–107.

165. Alonso, I.; Agirre, E.; Lapata, M. PixT3: Pixel-based Table To Text generation. *arXiv* **2023**, arXiv:2311.09808.

166. Park, J.; Choi, J.Y.; Park, J.; et al. Hierarchical Visual Feature Aggregation for OCR-Free Document Understanding. *arXiv* **2024**, arXiv:2411.05254.

167. Lotz, J.F.; Setiawan, H.; Peitz, S.; et al. Overcoming Vocabulary Constraints with Pixel-level Fallback. *arXiv* **2025**, arXiv:2504.02122.

168. Zhu, M.; Tian, Y.; Chen, H.; et al. SegAgent: Exploring Pixel Understanding Capabilities in MLLMs by Imitating Human Annotator Trajectories. *arXiv* **2025**, arXiv:2503.08625.

169. Huang, X.; Shen, L.; Liu, J.; et al. Towards a Multimodal Large Language Model with Pixel-Level Insight for Biomedicine. *arXiv* **2025**, arXiv:2412.09278.

170. Liao, W.; Wang, J.; Li, H.; et al. DocLayLLM: An Efficient Multi-modal Extension of Large Language Models for Text-rich Document Understanding. *arXiv* **2025**, arXiv:2408.15045.

171. Wang, Z.; Guan, T.; Fu, P.; et al. Marten: Visual Question Answering with Mask Generation for Multi-modal Document Understanding. *arXiv* **2025**, arXiv:2503.14140.

172. Siam, M. PixFoundation: Are We Heading in the Right Direction with Pixel-level Vision Foundation Models? *arXiv* **2025**, arXiv:2502.04192.

173. Lu, Y.; Li, X.; Fu, T.J.; et al. From Text to Pixel: Advancing Long-Context Understanding in MLLMs. *arXiv* **2024**, arXiv:2405.14213.

174. Wang, A.J.; Li, L.; Lin, Y.; et al. Leveraging Visual Tokens for Extended Text Contexts in Multi-Modal Learning. *arXiv* **2024**, arXiv:2406.02547.

175. Sun, Q.; Cui, Y.; Zhang, X.; et al. Generative multimodal models are in-context learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 14398–14409.

176. Tong, S.; Fan, D.; Zhu, J.; et al. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv* **2024**, arXiv:2412.14164.

177. Chen, X.; Wu, Z.; Liu, X.; et al. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv* **2025**, arXiv:2501.17811.

178. Bai, S.; Chen, K.; Liu, X.; et al. Qwen2. 5-vl technical report. *arXiv* **2025**, arXiv:2502.13923.

179. Wang, P.; Bai, S.; Tan, S.; et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution.

arXiv **2024**, arXiv:2409.12191.

180. Chen, G.; Li, Z.; Wang, S.; et al. Eagle 2.5: Boosting Long-Context Post-Training for Frontier Vision-Language Models. *arXiv* **2025**, arXiv:2504.15271.
181. Zhou, C.; Hu, H.; Xu, C.; et al. InternVL 3.0 Technical Report. *arXiv* **2024**, arXiv:2405.01638.
182. Wu, C.; Chen, X.; Wu, Z.; et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 10–17 June 2025; pp. 12966–12977.
183. Xie, J.; Mao, W.; Bai, Z.; et al. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv* **2024**, arXiv:2408.12528.
184. Radford, A.; Kim, J.W.; Hallacy, C.; et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020.
185. Zhang, J.; Lin, K.; Yang, Y.; et al. Eagle: Exploring the Design Space for Multimodal LLMs with Mixture of Encoders. *arXiv* **2024**, arXiv:2404.13508.
186. Shi, W.; Han, X.; Zhou, C.; et al. LMFusion: Adapting Pretrained Language Models for Multimodal Generation. *arXiv* **2024**, arXiv:2412.15188.
187. Iyer, S.; Lin, X.V.; Pasunuru, R.; et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv* **2022**, arXiv:2212.12017.
188. Yang, A.; Li, A.; Yang, B.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388.
189. Shazeer, N. Glu variants improve transformer. *arXiv* **2020**, arXiv:2002.05202.
190. Schulman, J.; Wolski, F.; Dhariwal, P.; et al. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
191. Shao, Z.; Wang, P.; Zhu, Q.; et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv* **2024**, arXiv:2402.03300.
192. Yu, Q.; Zhang, Z.; Zhu, R.; et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv* **2025**, arXiv:2503.14476.
193. Gao, P.; Han, J.; Zhang, R.; et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv* **2023**, arXiv:2304.15010.
194. Chen, L.; Li, X.; Wang, R.; et al. Open-Qwen2VL: Compute-Efficient Pre-training of Fully-Open Multimodal LLMs on Academic Resources. *arXiv* **2024**, arXiv:2404.14074.
195. Shen, H.; Liu, P.; Li, J.; et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv* **2025**, arXiv:2504.07615.
196. Zhou, H.; Li, X.; Wang, R.; et al. R1-Zero's" Aha Moment" in Visual Reasoning on a 2B Non-SFT Model. *arXiv* **2025**, arXiv:2503.05132.
197. Huang, W.; Jia, B.; Zhai, Z.; et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv* **2025**, arXiv:2503.06749.
198. Peng, Y.; Zhang, G.; Zhang, M.; et al. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv* **2025**, arXiv:2503.07536.
199. Chen, L.; Gao, H.; Liu, T.; et al. G1: Bootstrapping Perception and Reasoning Abilities of Vision-Language Model via Reinforcement Learning. *arXiv* **2025**, arXiv:2505.13426.
200. Yang, Y.; He, X.; Pan, H.; et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv* **2025**, arXiv:2503.10615.
201. Zhang, S.; Fang, Q.; Yang, Z.; et al. LLaVA-Mini: Efficient Image and Video Large Multimodal Models with One Vision Token. *arXiv* **2025**, arXiv:2501.03895.
202. Zou, J.; Liao, B.; Zhang, Q.; et al. Omnimamba: Efficient and unified multimodal understanding and generation via state space models. *arXiv* **2025**, arXiv:2503.08686.
203. Wang, Z.; Cai, S.; Mu, Z.; et al. Omnijarvis: Unified vision-language-action tokenization enables open-world instruction following agents. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 73278–73308.
204. Zhou, C.; Poczos, B. Objective-Agnostic Enhancement of Molecule Properties via Multi-Stage VAE. *arXiv* **2023**, arXiv:2308.13066.
205. Nguyen, M.; Adibekyan, V. On the formation of super-Jupiters: Core accretion or gravitational instability? *Astrophys. Space Sci.* **2024**, *369*, 122.
206. Zhao, X.; Zhang, Y.; Zhang, W.; et al. UniFashion: A Unified Vision-Language Model for Multimodal Fashion Retrieval and Generation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024; pp. 1490–1507.
207. Wang, X.; Zhang, X.; Luo, Z.; et al. Emu3: Next-token prediction is all you need. *arXiv* **2024**, arXiv:2409.18869.
208. Kou, S.; Jin, J.; Liu, Z.; et al. Orthus: Autoregressive interleaved image-text generation with modality-specific heads. *arXiv* **2024**, arXiv:2412.00127.
209. Zhao, Y.; Xue, F.; Reed, S.; et al. Krähenbühl, P.; Huang, D.A. QLIP: Text-Aligned Visual Tokenization Unifies Auto-Regressive Multimodal Understanding and Generation. *arXiv* **2025**, arXiv:2502.05178.

210. Qu, L.; Zhang, H.; Liu, Y.; et al. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 2545–2555.
211. Huang, R.; Wang, C.; Yang, J.; et al. Illume+: Illuminating unified mllm with dual visual tokenization and diffusion refinement. *arXiv* **2025**, arXiv:2504.01934.
212. Chow, W.; Li, J.; Yu, Q.; et al. Unified Generative and Discriminative Training for Multi-modal Large Language Models. *Adv. Neural Inf. Process. Syst.* **2025**, *37*, 23155–23190.
213. Yasunaga, M.; Aghajanyan, A.; Shi, W.; et al. Retrieval-augmented multimodal language modeling. *arXiv* **2022**, arXiv:2211.12561.
214. Woo, G.; Liu, C.; Kumar, A.; et al. Unified training of universal time series forecasting transformers. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; pp. 53140–53164.
215. Gao, S.; Koker, T.; Queen, O.; et al. UniTS: A unified multi-task time series model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 140589–140631.
216. He, K.; Chen, X.; Xie, S.; et al. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
217. Wang, A.; Singh, A.; Michael, J.; et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; pp. 353–355.
218. Radford, A.; Wu, J.; Child, R.; et al. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
219. Lin, T.Y.; Maire, M.; Belongie, S.; et al. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part V 13, pp. 740–755.
220. Young, P.; Lai, A.; Hodosh, M.; et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78.
221. Plummer, B.A.; Wang, L.; Cervantes, C.M.; et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2641–2649.
222. Agrawal, H.; Desai, K.; Wang, Y.; et al. Nocaps: Novel object captioning at scale. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 27 October–2 November 2019; pp. 8948–8957.
223. Kazemzadeh, S.; Ordonez, V.; Matten, M.; et al. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 787–798.
224. Mao, J.; Huang, J.; Toshev, A.; et al. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 11–20.
225. Antol, S.; Agrawal, A.; Lu, J.; et al. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
226. Hsieh, H.Y.; Liu, S.W.; Meng, C.C.; et al. TaiwanVQA: A Benchmark for Visual Question Answering for Taiwanese Daily Life. In Proceedings of the First Workshop of Evaluation of Multi-Modal Generation, Abu Dhabi, United Arab Emirates, 20 January 2025; pp. 57–75.
227. Goyal, Y.; Khot, T.; Summers-Stay, D.; et al. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.
228. Marino, K.; Rastegari, M.; Farhadi, A.; et al. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3195–3204.
229. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6700–6709.
230. Gurari, D.; Li, Q.; Stangl, A.J.; et al. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3608–3617.
231. Masry, A.; Do, X.L.; Tan, J.Q.; et al. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 2263–2279.
232. Methani, N.; Ganguly, P.; Khapra, M.M.; et al. Plotqa: Reasoning over scientific plots. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 1527–1536.
233. Mathew, M.; Bagal, V.; Tito, R.; et al. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 1697–1706.
234. Kembhavi, A.; Salvato, M.; Kolve, E.; et al. A diagram is worth a dozen images. In Proceedings of the Computer Vision–

ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Part IV 14, pp. 235–251.

235. Mathew, M.; Karatzas, D.; Jawahar, C. DocVQA: A Dataset for VQA on Document Images. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Virtual, 5–9 January 2021, pp. 2200–2209.
236. Mishra, A.; Shekhar, S.; Singh, A.K.; et al. Ocr-vqa: Visual question answering by reading text in images. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 947–952.
237. Singh, A.; Natarajan, V.; Shah, M.; et al. Towards vqa models that can read. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8317–8326.
238. Lu, P.; Mishra, S.; Xia, T.; et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 2507–2521.
239. Wang, C.J.; Lee, D.; Menghini, C.; et al. EnigmaEval: A Benchmark of Long Multimodal Reasoning Challenges. *arXiv* **2025**, arXiv:2502.08859
240. Wang, H.; Fan, Y.; Naeem, M.F.; et al. TokenFormer: Rethinking Transformer Scaling with Tokenized Model Parameters. *arXiv* **2025**, arXiv:2410.23168.
241. Duan, Y.; Wang, W.; Chen, Z.; et al. Vision-RWKV: Efficient and Scalable Visual Perception with RWKV-Like Architectures. *arXiv* **2025**, arXiv:2403.02308.
242. Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. In Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21), Los Angeles, CA, USA, 6–14 December 2021.
243. Yin, Z.; Wang, J.; Cao, J.; et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 26650–26685.
244. Bai, S.; Yang, S.; Bai, J.; et al. Touchstone: Evaluating vision-language models by language models. *arXiv* **2023**, arXiv:2308.16890.
245. Farsi, F.; Shariati Motlagh, S.; Bali, S.; et al. Persian in a Court: Benchmarking VLMs In Persian Multi-Modal Tasks. In Proceedings of the First Workshop of Evaluation of Multi-Modal Generation, Abu Dhabi, United Arab Emirates, 20 January 2025; pp. 52–56.
246. Liu, Y.; Duan, H.; Zhang, Y.; et al. Mmbench: Is your multi-modal model an all-around player? *arXiv* **2023**, arXiv:2307.06281.
247. Ying, K.; Meng, F.; Wang, J.; et al. MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. In Proceedings of the 41st International Conference on Machine Learning (PMLR), Vienna, Austria, 21–27 July 2024; Volume 235, pp. 57116–57198.
248. Xu, P.; Shao, W.; Zhang, K.; et al. LVLM-EHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 1877–1893.
249. Wang, Z.; Liu, J.; Tang, C.W.; et al. JourneyBench: a challenging one-stop vision-language understanding benchmark of generated images. In Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24), Vancouver, BC, Canada, 10–15 December 2024.
250. Lu, P.; Bansal, H.; Xia, T.; et al. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv* **2023**, arXiv:2310.02255.
251. Wang, Z.; Xia, M.; He, L.; et al. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *arXiv* **2024**, arXiv:2406.18521.
252. Ikuta, H.; Wöhler, L.; Aizawa, K. MangaUB: A Manga Understanding Benchmark for Large Multimodal Models. *arXiv* **2024**, arXiv:2407.19034.
253. Liang, P.P.; Lyu, Y.; Fan, X.; et al. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), Virtual, 6–14 December 2021.
254. Yu, W.; Yang, Z.; Ren, L.; et al. MM-Vet v2: A Challenging Benchmark to Evaluate Large Multimodal Models for Integrated Capabilities. *arXiv* **2024**, arXiv:2408.00765.
255. Patraucean, V.; Smaira, L.; Gupta, A.; et al. Perception Test: A Diagnostic Benchmark for Multimodal Video Models. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, New Orleans, LA, USA, 10–16 December 2023.
256. Li, Y.; Du, Y.; Zhou, K.; et al. Evaluating object hallucination in large vision-language models. *arXiv* **2023**, arXiv:2305.10355.
257. Wang, P.; Li, Z.Z.; Yin, F.; et al. MV-MATH: Evaluating Multimodal Math Reasoning in Multi-Visual Contexts. *arXiv* **2025**, arXiv:2502.20808.
258. Li, J.; Pan, K.; Ge, Z.; et al. Fine-tuning multimodal llms to follow zero-shot demonstrative instructions. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7 May 2024.
259. Wadhawan, R.; Bansal, H.; Chang, K.W.; et al. ConTextual: Evaluating Context-Sensitive Text-Rich Visual Reasoning in Large Multimodal Models. *arXiv* **2024**, arXiv:2401.13311.
260. Li, B.; Ge, Y.; Chen, Y.; et al. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv* **2024**, arXiv:2404.16790.

261. Yue, X.; Ni, Y.; Zhang, K.; et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 9556–9567.
262. Imam, M.F.; Lyu, C.; Aji, A.F. Can Multimodal LLMs do Visual Temporal Understanding and Reasoning? The answer is No! *arXiv* **2025**, arXiv:2501.10674.
263. Yang, B.; Zhang, Y.; Liu, D.; et al. Does Table Source Matter? Benchmarking and Improving Multimodal Scientific Table Understanding and Reasoning. *arXiv* **2025**, arXiv:2501.13042.
264. Ruan, J.; Yuan, W.; Gao, X.; et al. VLRMBench: A Comprehensive and Challenging Benchmark for Vision-Language Reward Models. *arXiv* **2025**, arXiv:2503.07478.
265. Gemini Team Google. Gemini: A Family of Highly Capable Multimodal Models. *arXiv* **2025**, arXiv:2312.11805.
266. Ma, Y.; Zang, Y.; Chen, L.; et al. MMLongBench-Doc: Benchmarking Long-context Document Understanding with Visualizations. *arXiv* **2024**, arXiv:2407.01523.
267. team, L.; Barrault, L.; Duquenne, P.A.; et al. Large Concept Models: Language Modeling in a Sentence Representation Space. *arXiv* **2024**, arXiv:2412.08821.
268. Artetxe, M.; Bhosale, S.; Goyal, N.; et al. Efficient Large Scale Language Modeling with Mixtures of Experts. In *P*roceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Abu Dhabi, 2022; pp. 11699–11732.
269. Bavishi R.; Elsen E.; Hawthorne C.; et al. Fuyu-8B: A Multimodal Architecture for AI Agents. Available online: https://www.adept.ai/blog/fuyu-8b (accessed on 17 October 2023)
270. Diao H.; Cui Y.; Li X.; et al. Unveiling Encoder-Free Vision-Language Models. *arXiv* **2024**, arXiv:2406.11832.