Vision Language Models (VLMs): A Comprehensive Review

ANMOL[1], DIYA GOELL[2], DHRUV JAIN[3], KUSHAL[4], PREETI[5]
[1, 2, 3, 4, 5]*Department of Computer Science and Engineering (AI & ML), Panipat Institute of Engineering and Technology (PIET), Samalkha, Panipat, Haryana, India*

Abstract- ***Vision-Language Models (VLMs) represent one of the most significant advances in multimodal artificial intelligence, enabling systems to jointly reason over visual and textual information. This review surveys four representative VLMs — GPT-4 Vision, LLaVA, Qwen-VL, and Florence — examining their architectures, training paradigms, and downstream capabilities. We compare the models along dimensions of vision encoder design, language backbone, training data scale, and benchmark performance, and we discuss common failure modes such as hallucination, weak spatial grounding, and limited fine-grained visual reasoning. We further outline open challenges including data efficiency, evaluation standardization, and computational cost, and we propose promising directions such as unified tokenization, retrieval-augmented multimodal reasoning, and efficient adapter-based fine-tuning. This review is intended as a structured reference for researchers and practitioners entering the field of vision-language modeling.***

Index Terms—Vision-Language Models, GPT-4V, LLaVA, Qwen-VL, Florence, multimodal learning, visual instruction tuning, transformers.

I. INTRODUCTION

The convergence of computer vision and natural language processing has given rise to Vision-Language Models (VLMs), systems capable of interpreting images, videos, and text within a single unified reasoning framework. Traditional computer vision pipelines relied on task-specific architectures — separate models for classification, detection, and captioning — whereas VLMs unify these capabilities by projecting visual features into the embedding space of a large language model (LLM). This shift has been driven by three converging trends: the maturation of transformer-based vision encoders, the emergence of instruction-tuned LLMs, and the availability of large-scale image-text paired datasets scraped from the web.

Early multimodal systems such as CLIP demonstrated that contrastive pretraining on image-text pairs could produce transferable visual representations aligned with natural language. Building on this foundation, subsequent models introduced generative capabilities, allowing a system not only to recognize visual concepts but to describe, reason about, and answer open-ended questions concerning them. This review focuses on four influential VLMs — GPT-4 Vision (GPT-4V), LLaVA, Qwen-VL, and Florence — chosen because they represent distinct design philosophies: a closed, general-purpose commercial system; an open-source instruction-tuned academic model; an industrially trained multilingual and grounding-capable model; and a foundation vision model with a unified representation objective.

The practical motivation for this review stems from the rapid and somewhat fragmented growth of the VLM landscape: new models are released frequently, often accompanied by benchmark claims that are difficult to compare directly due to differing evaluation protocols, undisclosed training data, and inconsistent prompt formats. Researchers and engineers selecting a VLM for a downstream application — whether a conversational assistant, a document-processing pipeline, or a visual search engine — must currently synthesize information scattered across technical reports, blog posts, and community benchmarks. This review consolidates that information for four representative and widely cited systems, with an emphasis on architectural mechanism rather than leaderboard position alone, since benchmark scores change quickly while underlying design trade-offs remain informative over a longer horizon.

The remainder of this paper is organized as follows. Section II provides background on the architectural principles underlying VLMs. Section III surveys each

of the four models in detail. Section IV presents a comparative analysis across architecture, data, and benchmark dimensions. Section V discusses applications, Section VI addresses challenges and limitations, Section VII outlines future research directions, and Section VIII concludes the paper.

II. BACKGROUND AND RELATED WORK

Modern VLMs are typically composed of three core components: (1) a vision encoder that converts raw pixels into a sequence of patch-level embeddings, commonly a Vision Transformer (ViT); (2) a modality projection layer, often a linear layer or a small multilayer perceptron (MLP), that maps visual embeddings into the token embedding space of a language model; and (3) a large language model backbone that consumes the projected visual tokens interleaved with text tokens to generate a response autoregressively.

Two dominant training strategies have emerged. The first, exemplified by CLIP and its derivatives, uses contrastive learning to align an image encoder and a text encoder in a shared embedding space, optimizing similarity between matched image-text pairs against a large batch of negatives. The second, used by generative VLMs such as LLaVA and Qwen-VL, freezes or lightly fine-tunes a pretrained vision encoder and language model, then trains the connecting projection layer — and optionally the LLM itself — using a two-stage process: feature alignment pretraining on caption data, followed by visual instruction tuning on curated multimodal conversation data.

Foundation models such as Florence depart from the instruction-following paradigm and instead aim to learn a single, general-purpose visual representation usable across a broad taxonomy of vision tasks — classification, retrieval, detection, and video action recognition — spanning what its authors term the space-time-modality dimensions of vision.

The choice of vision encoder itself constitutes an important design axis. Plain Vision Transformers, as used in CLIP and LLaVA, divide an image into fixed-size non-overlapping patches and process them with a stack of standard transformer blocks, producing a representation that is simple to interface with a language model but that scales quadratically in compute with image resolution. Hierarchical encoders such as the Swin Transformer family, used in Florence, instead compute attention within local windows that are progressively merged across layers, yielding a multi-scale representation more analogous to convolutional feature pyramids and better suited to dense prediction tasks such as detection. Adapter-based encoders, as in Qwen-VL, retain a standard ViT backbone but insert a compression module before the language model to control the number of visual tokens independently of input resolution.

Fusion strategy — how visual and textual information are combined inside the model — also varies. Early fusion approaches concatenate projected visual tokens directly into the input sequence of a decoder-only language model, allowing full self-attention between text and image tokens at every layer; this is the approach used by LLaVA and Qwen-VL. Cross-attention fusion instead keeps a frozen or lightly modified language model and injects visual information through dedicated cross-attention layers inserted at intervals within the decoder, an approach used by some earlier multimodal systems such as Flamingo. Early fusion tends to allow richer interaction between modalities at the cost of longer effective sequence length, while cross-attention fusion is more parameter-efficient but can bottleneck the amount of visual detail available to the language model at each layer.

III. SURVEY OF PROMINENT VLM ARCHITECTURES

A. GPT-4 Vision (GPT-4V)

GPT-4V extends OpenAI's GPT-4 language model with the ability to accept image inputs alongside text. While OpenAI has not disclosed the full architectural details, the GPT-4V system card describes a model trained on large-scale web and licensed image-text data, further refined through reinforcement learning from human feedback (RLHF) to improve helpfulness and reduce harmful outputs. GPT-4V demonstrates strong zero-shot performance across a wide range of tasks, including document and chart understanding, visual question answering, optical character recognition (OCR) in natural scenes, and multi-step

visual reasoning that combines several images with textual context.

A distinguishing characteristic of GPT-4V is its ability to perform compositional reasoning — for example, comparing quantities across multiple charts, or following multi-turn instructions that reference earlier images in a conversation. Because the model shares its underlying reasoning engine with text-only GPT-4, it inherits strong general knowledge and chain-of-thought capability, which it applies to visual inputs by treating image tokens as an additional modality within the same context window. This allows GPT-4V to combine information across several images and long textual passages within a single reasoning chain, a property that has made it popular for tasks such as comparing product photographs, auditing user-interface screenshots, and interpreting scientific figures.

However, as a closed, proprietary system, GPT-4V offers no access to model weights, training data composition, or intermediate activations, and its behavior can only be studied through black-box API-level evaluation. This limits reproducibility in academic research and makes it difficult to attribute specific failure modes to particular architectural or data decisions. OpenAI's system card acknowledges known weaknesses, including occasional misreading of dense or rotated text, unreliable performance on spatial reasoning tasks such as counting overlapping objects, and susceptibility to adversarial visual prompts designed to elicit unsafe or incorrect outputs. Deployment safeguards, including refusal behavior for sensitive image categories such as faces used for identification, are layered on top of the base model through additional RLHF and classifier-based filtering.

B. LLaVA

Large Language and Vision Assistant (LLaVA), introduced by Liu et al., is among the first open-source models to apply visual instruction tuning to a general-purpose LLM. LLaVA combines a frozen CLIP ViT-L/14 vision encoder with a Vicuna or LLaMA language backbone, connected through a trainable linear (later MLP, in LLaVA-1.5) projection layer. Training proceeds in two stages: first, the projection layer is pretrained on image-caption pairs drawn from a filtered subset of CC3M to align visual and textual embedding spaces while the vision encoder and LLM remain frozen; second, the projection layer and LLM are jointly fine-tuned on a GPT-4-generated instruction-following dataset consisting of multi-turn visual conversations, detailed descriptions, and complex reasoning questions synthesized from COCO image captions and bounding-box annotations.

LLaVA's chief contribution is demonstrating that competitive multimodal chat capability can be achieved with a comparatively small amount of curated instruction data — roughly 150,000 conversation samples in the original release — and modest compute, typically a single node of eight GPUs trained over one to two days, making it accessible for academic replication. This stands in contrast to the substantially larger, undisclosed training budgets associated with proprietary systems. Later versions, including LLaVA-1.5 and LLaVA-NeXT (also known as LLaVA-1.6), improved resolution handling through an any-resolution image-splitting scheme, added higher-quality academic task data covering document and chart understanding, and scaled the language backbone up to 34 billion parameters, closing much of the performance gap with proprietary systems on public benchmarks such as MMBench and SEED-Bench.

Because LLaVA's weights, training code, and instruction dataset are publicly released, it has become the de facto baseline for academic VLM research, spawning numerous derivative works that modify its projection layer, apply parameter-efficient fine-tuning, or extend it to video and medical imaging domains.

C. Qwen-VL

Qwen-VL, developed by Alibaba, pairs a Qwen-7B language model with an OpenCLIP ViT-bigG vision encoder and introduces a position-aware vision-language adapter that compresses long sequences of image features using cross-attention with a fixed number of trainable query embeddings, typically 256. This design allows Qwen-VL to process higher-resolution images (up to 448x448 pixels, later extended in follow-up variants) efficiently while retaining fine-grained spatial detail necessary for grounding tasks, without incurring the quadratic cost

of feeding every patch token directly into the language model.

A notable capability of Qwen-VL is visual grounding: the model can output bounding-box coordinates in text form when asked to localize objects described in natural language, using a specialized text format that encodes normalized pixel coordinates as token sequences. It supports multilingual (Chinese and English) OCR and dense text reading from documents, tables, and receipts, a capability reinforced through dedicated pretraining on synthetic and scanned document datasets. Qwen-VL is trained in a multi-stage pipeline comprising large-scale weakly labeled pretraining on approximately 1.4 billion image-text pairs, multi-task pretraining on grounding, OCR, and VQA datasets, and a final supervised instruction-tuning stage (Qwen-VL-Chat) that aligns the model with conversational formats and human preferences.

This staged pretraining strategy gives Qwen-VL strong performance on both open-ended conversational tasks and structured perception tasks, such as extracting key-value pairs from forms, that are less emphasized in the LLaVA training recipe.

D. Florence

Florence, developed by Microsoft, differs from the three models above in that it is primarily a foundation vision model rather than a conversational assistant. Florence is trained on FLD-900M, a large-scale noisy image-text dataset curated from web sources, using a unified contrastive learning objective and a hierarchical Vision Transformer (CoSwin) encoder combined with a unified contrastive learning (UniCL) objective, with the explicit goal of producing a single representation transferable across the space (scene to object level), time (static image to video), and modality (RGB to depth, or image to language) dimensions — a framework the original authors describe as the space-time-modality (STM) design space.

Florence's representations are adapted to downstream tasks via lightweight task-specific heads — for classification, retrieval, object detection, and video action recognition — rather than through instruction tuning or autoregressive generation. Extension modules such as Florence-CLIP for zero-shot classification, ELEVATER for few-shot transfer, and dynamic detection heads allow the same backbone to be repurposed across dozens of downstream benchmarks with minimal additional training. This makes Florence less suited to open-ended dialogue but highly effective as a backbone that can be fine-tuned quickly for narrow, high-accuracy production tasks such as retail image search, industrial defect detection, and satellite image classification, where inference efficiency and label precision matter more than conversational flexibility.

E. Summary of Architectural Trends

Taken together, the four architectures trace a broader trend in the field: an initial period dominated by contrastive dual-encoder models such as CLIP, followed by a wave of generative, instruction-tuned VLMs that repurpose pretrained LLMs as the reasoning core, and a parallel track of foundation vision models that prioritize transferable representation quality over conversational ability. GPT-4V represents the current ceiling of general-purpose closed-source capability; LLaVA represents the accessible open baseline against which new open models are typically measured; Qwen-VL represents an industrially motivated design optimized for structured perception tasks such as grounding and OCR; and Florence represents an alternative branch of the field oriented toward efficient, task-agnostic representation learning rather than dialogue. Understanding these four points in the design space provides a useful map for situating newer models that continue to emerge, since most subsequent VLM releases can be understood as recombinations or refinements of these underlying strategies rather than fundamentally new paradigms.

IV. TRAINING DATA AND OPTIMIZATION STRATEGIES

The four models illustrate distinct philosophies toward data curation and optimization. GPT-4V and Florence rely on very large, broadly scraped web corpora, trading annotation precision for scale and diversity, with quality controlled largely through downstream filtering and RLHF rather than upstream curation. LLaVA takes the opposite approach, relying on a small but carefully synthesized instruction dataset generated by prompting a stronger LLM (GPT-4) with

structured image annotations, prioritizing instruction diversity and conversational realism over raw scale. Qwen-VL sits between these extremes, combining large-scale weakly supervised pretraining with targeted, task-specific fine-tuning stages for grounding and OCR.

Optimization strategies also differ in which components are updated at each stage. Both LLaVA and Qwen-VL freeze the vision encoder during early alignment training to preserve the semantic structure learned during large-scale contrastive pretraining, updating only the projection layer, and later unfreeze part or all of the language model for instruction tuning. This staged, curriculum-style training reduces catastrophic forgetting of pretrained visual knowledge while still allowing the model to acquire task-specific behavior. Florence, in contrast, trains its encoder end-to-end from the outset under the UniCL objective, since its goal is to produce a general-purpose representation rather than a fixed encoder wrapped by a separately trained LLM.

V. EVALUATION BENCHMARKS

Standard evaluation of VLMs spans several benchmark families. Visual question answering benchmarks such as VQAv2 and OK-VQA test general scene understanding and commonsense reasoning. Text-reading benchmarks such as TextVQA and DocVQA test OCR-grounded comprehension of signs, documents, and charts. Instruction-following and conversational quality are assessed using LLaVA-Bench and MMBench, which use either human judges or a stronger LLM as an automated evaluator to score open-ended responses. Foundation-model benchmarks such as ImageNet-1k classification accuracy and COCO retrieval recall are used primarily to evaluate Florence-style representation models rather than conversational systems.

Reported results across these benchmarks should be interpreted cautiously. Differences in prompt phrasing, few-shot versus zero-shot evaluation protocol, and answer-parsing heuristics can shift reported accuracy by several points even for the same underlying model, and some benchmark test sets have been shown to overlap with the web-scraped pretraining corpora of large models, raising contamination concerns that complicate fair comparison.

VI. COMPARATIVE ANALYSIS

Table I summarizes the four models across vision encoder, language backbone, and key strength. GPT-4V leads in open-ended, general-purpose reasoning but is closed-source; LLaVA offers the most accessible open replication path; Qwen-VL provides the strongest grounding and multilingual OCR support; and Florence provides the most efficient, broadly transferable representation for narrow downstream tasks rather than conversational use.

Considered along an openness axis, LLaVA and Qwen-VL sit at the accessible end of the spectrum, releasing weights and, in LLaVA's case, training data, while GPT-4V remains fully closed and Florence's largest checkpoints are only partially released. Considered along a specialization axis, Florence is the most narrowly optimized for perception accuracy on fixed task heads, while GPT-4V is the most general-purpose, supporting open-ended natural-language interaction over arbitrary visual content without task-specific fine-tuning. Qwen-VL and LLaVA occupy an intermediate position, supporting general conversation while retaining measurable strength on specific structured tasks such as grounding and document reading, respectively.

TABLE I
Comparison of Surveyed VLM Architectures

Model	Developer	Vision Encoder	Language Backbone	Key Strength
GPT-4V	OpenAI	Proprietary ViT-based	GPT-4	General reasoning, OCR, chart QA
LLaVA	Academic (Liu et al.)	CLIP ViT-L/14	Vicuna / LLaMA	Open-source instruction tuning
Qwen-VL	Alibaba	ViT-bigG (OpenCLIP)	Qwen-7B	Grounding, multilingual OCR
Florence	Microsoft	CoSwin Transformer	Task-specific heads	Unified space-time-modality representation

TABLE II
Illustrative Benchmark Coverage by Model

Benchmark	Task Focus	Strong Performers	Weaker Performers
VQAv2	General VQA	GPT-4V, Qwen-VL	Base LLaVA
TextVQA / DocVQA	OCR & documents	Qwen-VL, GPT-4V	Florence (indirect)
LLaVA-Bench / MMBench	Conversational quality	GPT-4V, LLaVA-1.5	Florence (n/a)
ImageNet / COCO retrieval	Representation transfer	Florence	Not primary use case for others

Table II should be read qualitatively rather than as a precise ranking, since exact scores are highly sensitive to prompt format and evaluation protocol as discussed in Section V; it nonetheless illustrates that no single model dominates across all benchmark families, and that model selection should be guided by the specific task family most relevant to the target deployment.

On public benchmarks such as VQAv2, TextVQA, and MMBench, GPT-4V and Qwen-VL typically report the strongest raw accuracy, reflecting their larger training corpora and higher-resolution image processing. LLaVA-1.5, despite far smaller training cost, remains competitive on conversational benchmarks such as LLaVA-Bench, illustrating that instruction-data quality can partially substitute for scale. Florence, evaluated on classification and retrieval benchmarks such as ImageNet and COCO rather than VQA-style tasks, achieves strong transfer performance but is not directly comparable on conversational metrics.

VII. APPLICATIONS

VLMs are increasingly deployed across a wide range of real-world domains. In accessibility, they power scene description and reading assistance tools for visually impaired users, converting camera input into spoken natural-language descriptions in real time. In document intelligence, models such as Qwen-VL and GPT-4V are used to extract structured data from invoices, forms, and charts, reducing manual data-entry workload in finance and logistics workflows. In e-commerce, Florence-style embeddings support visual search and recommendation, allowing users to search a catalog using a photograph rather than a text query.

In education, VLMs assist in automatically grading handwritten or diagrammatic answers and in generating accessible descriptions of textbook figures for students with visual impairments. In robotics, VLMs are being integrated as perception front-ends

that translate visual scenes into language-conditioned action plans, bridging perception and control by allowing a robot to interpret an instruction such as "pick up the red cup" directly against its camera feed. In healthcare, early applications include VLM-assisted triage of radiology images, where a model produces a draft textual impression later reviewed by a clinician, and in security, VLMs are used to summarize surveillance footage into searchable natural-language logs.

VIII. CHALLENGES AND LIMITATIONS

Despite rapid progress, VLMs exhibit several persistent limitations. First, hallucination remains a major concern: models frequently describe objects, attributes, or relationships that are not present in the image, particularly under ambiguous or low-resolution inputs, and this tendency worsens as generated responses grow longer, since early hallucinated content can compound through the model's own autoregressive context. Second, fine-grained spatial reasoning — counting, relative positioning, and precise localization — remains weaker than task-specific detection models, even in grounding-capable systems such as Qwen-VL, because visual tokens are typically pooled or compressed in ways that discard precise pixel-level position information.

Third, evaluation is fragmented: benchmarks differ in annotation quality, task framing, and scoring methodology, making cross-model comparison difficult and occasionally misleading, and automated LLM-based judges used in benchmarks such as MMBench can themselves introduce systematic scoring biases. Fourth, computational cost is substantial; processing high-resolution images generates long visual token sequences that strain the context windows and inference latency of the underlying LLM, and this cost scales further when multiple images or video frames are processed in a single conversation. Fifth, safety and fairness concerns are amplified in the visual domain: VLMs can inadvertently infer sensitive attributes such as approximate age, ethnicity, or location from an image even when not explicitly asked, raising privacy considerations distinct from text-only LLMs. Finally, closed models such as GPT-4V limit reproducibility and mechanistic interpretability, constraining the pace of academic progress relative to open alternatives.

IX. FUTURE RESEARCH DIRECTIONS

Several directions appear promising for advancing VLM capability and reliability. Unified tokenization schemes that represent images, video, and text within a single discrete vocabulary may simplify architectures and improve cross-modal transfer, avoiding the need for a separately trained projection module. Retrieval-augmented multimodal reasoning, in which a VLM consults an external knowledge base or image index at inference time, could reduce hallucination on knowledge-intensive visual questions by grounding generated claims in retrieved evidence rather than parametric memory alone.

Parameter-efficient adaptation methods, such as low-rank adapters applied to both the vision encoder and language backbone, may lower the cost of specializing large VLMs for narrow domains such as medical or satellite imagery, where labeled data is scarce and full fine-tuning is impractical. Improved visual tokenization — for example, adaptive-resolution patching that allocates more tokens to information-dense image regions such as text or faces — could improve fine-grained reasoning without proportionally increasing sequence length. Finally, the community would benefit from standardized, contamination-checked benchmarks that separate perception accuracy from language fluency, enabling more precise diagnosis of whether a model's errors originate from faulty visual perception or from downstream language generation.

X. ILLUSTRATIVE CASE STUDY: DOCUMENT UNDERSTANDING TASK

To make the comparative discussion concrete, consider a representative task drawn from document intelligence: given a photograph of a printed invoice, extract the vendor name, invoice date, and total amount due, and return the result as structured key-value text. This task exercises OCR accuracy, layout understanding, and instruction-following simultaneously, making it a useful lens for comparing the surveyed models.

GPT-4V typically handles this task well in a zero-shot setting, correctly parsing most printed invoices without any task-specific fine-tuning, owing to its large and diverse pretraining corpus and strong general instruction-following ability; failures are more common on handwritten or low-contrast scans. LLaVA, in its base released form, performs noticeably worse on this task, since its instruction-tuning data emphasizes natural photographs and conversational description rather than dense document text, and its relatively low input resolution in early versions limits legibility of small print; LLaVA-NeXT's higher-resolution handling partially closes this gap. Qwen-VL, having been explicitly pretrained on OCR and document datasets, performs strongly and reliably on structured extraction, often matching or exceeding GPT-4V on this narrow task despite its smaller overall parameter count. Florence, lacking an instruction-following interface, cannot perform this task directly as specified; it would instead require pairing its visual representation with a separate detection or OCR head and a downstream rule-based or learned extraction module, illustrating the practical difference between a conversational VLM and a foundation vision representation.

This case study underscores a broader pattern observed throughout this review: general-purpose reasoning strength, as in GPT-4V, and targeted pretraining for a specific perceptual skill, as in Qwen-VL's OCR emphasis, can each independently produce strong task performance, and the better choice for a given deployment depends on whether the surrounding system requires broad conversational flexibility or narrow, high-reliability extraction.

XI. ETHICAL AND SOCIETAL CONSIDERATIONS

The deployment of VLMs at scale raises ethical considerations distinct from, and additive to, those already documented for text-only LLMs. Because these models can jointly interpret images and generate free-form text, they introduce new vectors for misuse, including the generation of misleading captions for real photographs, automated profiling of individuals from incidental image content, and circumvention of content moderation systems that were designed around text-only inputs. Bias is also a concern: vision encoders trained on web-scraped data can inherit demographic imbalances present in the source corpora, leading to systematically lower captioning or recognition accuracy for underrepresented groups, imagery, or cultural contexts.

Responsible development practices adopted by the surveyed models vary considerably. OpenAI's GPT-4V system card documents structured red-teaming, refusal training for sensitive image categories, and staged external access prior to broad release. Open-source projects such as LLaVA and Qwen-VL instead rely on community-driven scrutiny after release, which offers transparency benefits but places the burden of safe deployment on downstream integrators. Foundation models such as Florence, being further removed from direct user interaction, shift much of the ethical responsibility to whoever fine-tunes and deploys the resulting task-specific system. These differences suggest that governance and evaluation frameworks for VLMs must account not only for a model's raw capability but for the specific deployment context and access model through which it reaches end users.

XII. CONCLUSION

This review examined four representative Vision-Language Models — GPT-4V, LLaVA, Qwen-VL, and Florence — spanning proprietary and open-source, conversational and foundation-model design philosophies. While each model advances the state of the art along different axes, all share common limitations in hallucination, fine-grained spatial reasoning, and evaluation standardization. Continued progress will likely depend on improvements in training data quality, more efficient architectures for high-resolution visual input, and the development of rigorous, standardized benchmarks that allow fair comparison across the rapidly diversifying VLM landscape.

Looking forward, the field appears to be converging toward hybrid designs that borrow the broad conversational competence of models like GPT-4V, the open reproducibility of models like LLaVA, the perceptual precision of models like Qwen-VL, and the transferable efficiency of foundation representations like Florence. As open releases continue to narrow the

gap with proprietary systems, and as evaluation practices mature toward standardized, contamination-aware benchmarks, the practical choice among VLMs is likely to shift from a question of raw capability toward a question of deployment constraints — latency, cost, openness, and task specificity — making the comparative framework offered in this review relevant well beyond the specific model versions surveyed here.

ACKNOWLEDGMENT

The author thanks the open-source community behind LLaVA and Qwen-VL for releasing model weights, training code, and datasets that make independent academic study of vision-language models possible.

REFERENCES

[1] [1] J. Achiam et al., "GPT-4 Technical Report," arXiv preprint arXiv:2303.08774, 2023.

[2] [2] OpenAI, "GPT-4V(ision) System Card," OpenAI Technical Report, 2023.

[3] [3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2023.

[4] [4] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," arXiv preprint arXiv:2310.03744, 2023.

[5] [5] J. Bai et al., "Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities," arXiv preprint arXiv:2308.12966, 2023.

[6] [6] L. Yuan et al., "Florence: A New Foundation Model for Computer Vision," arXiv preprint arXiv:2111.11432, 2021.

[7] [7] A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. Int. Conf. on Machine Learning (ICML), 2021.

[8] [8] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in Proc. ICML, 2023.

[9] [9] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proc. Int. Conf. on Learning Representations (ICLR), 2021.

[10] [10] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2021.

[11] [11] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering," in Proc. CVPR, 2017.

[12] [12] A. Singh et al., "Towards VQA Models That Can Read," in Proc. CVPR, 2019.

[13] [13] Y. Liu et al., "MMBench: Is Your Multi-modal Model an All-around Player?," arXiv preprint arXiv:2307.06281, 2023.

[14] [14] T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," in Proc. European Conf. on Computer Vision (ECCV), 2014.

[15] [15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in Proc. CVPR, 2009.

[16] [16] J.-B. Alayrac et al., "Flamingo: A Visual Language Model for Few-Shot Learning," in Proc. NeurIPS, 2022.

[17] [17] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A Dataset for VQA on Document Images," in Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV), 2021.

[18] [18] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge," in Proc. CVPR, 2019.