Review | Received 8 July 2025; Revised 13 October 2025; Accepted 22 October 2025; Published 6 November 2025
https://doi.org/10.55092/aiplus20260001

A survey on omni-modal language models

Lu Chen[1], Jiajie Mu[1], Jiarui Wang[1], Xiao Kang[2], Xiaoming Xi[1] and Zheyun Qin[3,*]

[1] School of Computer and Artificial Intelligence, Shandong Jianzhu University, Jinan, China
[2] School of Software, Shandong University, Jinan, China
[3] School of Computer Science and Technology, Shandong University, Qingdao, China

* Corresponding author; E-mail: zheyun.qin@sdu.edu.cn.

Highlights:

- Provides a comprehensive review of omni-modal language models (OMLMs) from architecture to deployment.
- Summarizes modality alignment, semantic fusion, and multimodal collaborative learning strategies.
- Introduces modality pruning and lightweight adaptation for real-time industrial and medical applications.
- Discusses domain-specific customization of OMLMs in healthcare, education, and quality inspection.
- Proposes future research directions on structural flexibility, semantic plasticity, and deployment efficiency.

Abstract: This paper provides a comprehensive review of Omni-Modal Language Models (OMLMs), focusing on their evolution, technical challenges, application scenarios, and evaluation frameworks. OMLMs represent a significant leap from traditional unimodal and multimodal models by unifying modalities like text, images, audio, and video into a cohesive architecture. These models aim to simulate human-like multimodal perception, achieving semantic alignment and dynamic interaction between diverse data sources. Key topics covered include modality alignment, semantic fusion, and joint representation learning, alongside their application in fields such as healthcare, education, and industrial quality inspection. The paper also examines vertical adaptation paths, knowledge injection mechanisms, real-time optimization strategies, and a multi-dimensional evaluation system. Finally, future research directions are proposed, including improvements in generalization, task adaptability, energy efficiency, and ethical considerations, all critical for the widespread deployment of OMLMs in complex, real-world scenarios.

Keywords: omni-modal language models; semantic fusion; modality alignment; joint representation learning; cross-modal interaction

1. Introduction

Omni-Modal Language Models are at the forefront of current artificial intelligence research, focusing on building a unified architecture that processes multimodal information from text, images, audio, video, and three-dimensional sensor data. Their goal is to overcome the limitations of traditional unimodal models

(e.g., BERT or Vision Transformer) by achieving cross-modal semantic alignment and dynamic interaction mechanisms, enabling joint representation and generation of multi-source heterogeneous data [1]. These models are not only required to deeply understand each modality but also to fuse semantics across modalities and manage dynamic information flow, supporting complex tasks such as text-to-image generation and text-visual decision-making. Essentially, omni-modal language models simulate human multimodal perception and cognitive mechanisms, making them one of the critical pathways for exploring Artificial General Intelligence (AGI) by unifying perception channels such as language, vision, and hearing through a single architecture, thus enabling human-like cross-modal understanding and expression capabilities [2].

Compared to traditional multimodal models (e.g., VisualBERT, ViLBERT), which only cover specific combinations of modalities, omni-modal language models place more emphasis on "modality agnosticism" and "task-driven dynamic interaction." Their architecture typically adopts a modular design or unified attention mechanism, supporting flexible integration of new modalities through parameter sharing or adapters, ensuring good scalability [3]. In inference, omni-modal models break away from early or late fusion static paradigms, transitioning to dynamic interaction mechanisms between modalities. For instance, in generative tasks, the language modality may first generate preliminary results, which are then refined by the visual modality [4]. This task-phase-aware information flow design enhances the model's adaptability in cross-modal generation and complex scene understanding [5].

In contrast to the current mainstream large language models (LLMs) and multimodal large language models (MLLMs), omni-modal language models differ fundamentally in both design philosophy and implementation paths. LLMs (e.g., GPT-4, PaLM) focus on text as the core modality and use self-supervised learning to capture linguistic patterns. Their capabilities are limited to language understanding and generation, and they cannot natively process other modalities like images or audio [6]. While MLLMs (e.g., Flamingo, Kosmos-1) introduce modalities like vision or audio, they still prioritize text, with other modalities mainly serving as supplementary inputs. The cross-modal interaction in these models largely relies on simple concatenation or attention mechanisms, lacking deep semantic fusion. For example, in text-visual tasks, image features are often only added as supplementary inputs concatenated to the text sequence, failing to achieve true bidirectional semantic alignment [7]. On the other hand, omni-modal language models eliminate the concept of a "core modality" and pursue equality and symmetry across modalities at the architectural level, emphasizing multimodal parameter sharing, unified attention mechanisms, and collaborative training strategies. For instance, in medical scenarios, the model can simultaneously process CT images, medical records, and audio descriptions from patients, improving diagnostic accuracy through complementary information between modalities, instead of relying on a dominant analysis from a single modality [8]. Additionally, during pretraining, omni-modal models emphasize joint optimization across multimodal tasks, differing from the traditional paradigm of unimodal pretraining followed by fine-tuning in LLMs and MLLMs [9], thus showing better generalization capabilities in low-resource modalities and complex tasks [10].

Although omni-modal models show great potential for advancing towards AGI, their development still faces numerous challenges, particularly in areas such as the ambiguity of feature space mappings in modality alignment, dynamic balancing of information weights in semantic fusion, and the high dependence

of joint representation learning on the construction of shared semantic spaces. At the same time, the demands for multimodal input vary greatly across industries, and how to achieve model customization in vertical applications such as medical diagnosis, intelligent education, and industrial quality inspection has become a critical issue for model deployment [11]. Current evaluation systems, mainly derived from unimodal tasks, have not yet formed a systematic framework for evaluating general perception, spatial understanding, visual quality, multimodal reasoning, and generation abilities. Furthermore, future development will need to address practical constraints such as how cross-modal alignment methods can integrate with LLM semantic abilities, how to enhance models' interactive agent capabilities in real applications, and considerations around resource consumption, data privacy, and ethical risks during deployment [12].

Figure 1 below illustrates the evolution and development paths of omni-modal large models, depicting key milestones, architectural progress, and notable models released in recent years. Based on this, the paper systematically reviews the research trajectory and engineering evolution of omni-modal large models from an academic perspective. It explores four key dimensions: technical challenges, scene adaptation, evaluation systems, and future trends, aiming to provide a systematic reference framework for theoretical development and practical deployment in this field.

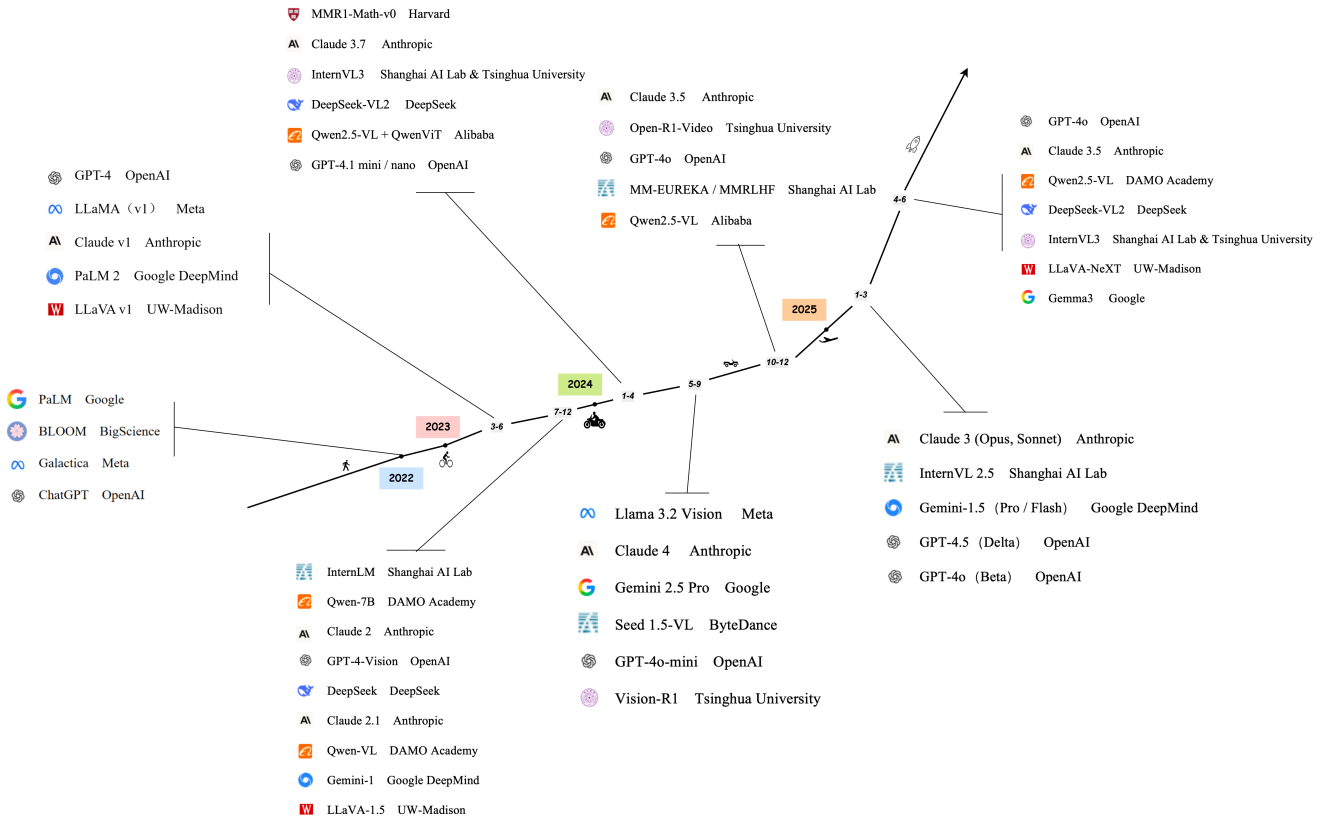

Figure 1. The evolution and development path diagram of omni-modal large models.

While recent surveys such as Caffagni *et al.* [13] and Li *et al.* [14] have systematically reviewed multimodal large language models from architectural and data-centric perspectives respectively, this work differs by offering a unified framework that spans not only core technical dimensions (e.g., modality alignment, semantic fusion, and joint representation learning), but also addresses real-world deployment

issues such as energy efficiency, human-AI collaboration, and trustworthy system design.

Paper Organization: The remainder of this paper is organized as follows: Section 2 clarifies key terminology and presents the survey methodology. Section 3 compares related surveys and positions this work within the research landscape. Section 4 reviews core architectures of general OMLMs, including modality alignment, semantic fusion, and joint representation, followed by critical insights. Section 5 discusses deployment and vertical adaptation across typical domains. Section 6 introduces the evaluation framework and benchmark analysis. Section 7 outlines emerging trends and open problems. Finally, Section 8 concludes the paper.

2. Terminology and survey methodology

To ensure conceptual clarity and consistency, this section provides concise definitions of several key terms frequently used throughout the paper. Since the research on omni-modal intelligence often overlaps with multi-modal and cross-modal learning, clearly distinguishing these concepts is essential for coherent discussion.

Omni-modal: Omni-modal models are designed to process and understand information from all major sensory modalities—including text, image, audio, video, and 3D data—within a unified framework. Unlike traditional multi-modal systems, omni-modal models can dynamically adapt to arbitrary combinations of modalities without predefined input configurations.

Multi-modal: The term multi-modal refers to systems that jointly learn from multiple but fixed modalities (e.g., text + image). Such systems typically rely on cross-attention mechanisms or shared encoders, but the types of modalities they handle are limited and predefined.

Cross-modal: Cross-modal learning focuses on interaction and translation between two distinct modalities, such as image–text alignment or cross-modal retrieval. It emphasizes transfer and mapping between heterogeneous representation spaces.

Modality-agnostic architecture: A modality-agnostic architecture encodes different modalities into a shared latent space using unified parameters. This enables seamless generalization across modality types and allows the model to process unseen modality combinations with minimal adaptation.

Semantic plasticity: Semantic plasticity describes the ability of a model to adapt and reinterpret semantics across different modalities or tasks while maintaining conceptual consistency. It captures the flexibility of representation learning under modality shifts or contextual variations.

Green AI evaluation: Green AI evaluation expands the conventional notion of model assessment beyond accuracy. It considers computational efficiency, energy consumption, and carbon footprint, promoting sustainable and responsible AI development.

By defining these terms, the paper establishes a consistent conceptual foundation for subsequent sections, preventing ambiguity between overlapping notions such as omni-modal, multi-modal, and cross-modal systems.

2.1. Survey methodology

To ensure a comprehensive and unbiased review, this paper follows a systematic literature analysis process. We collected publications related to omni-modal and multimodal large models from 2020 to 2025 across

databases such as IEEE Xplore, ACM Digital Library, SpringerLink, and arXiv. Figure 2 presents the annual distribution of relevant publications, showing a steady increase in research interest and indicating the rapid expansion of this field in recent years.

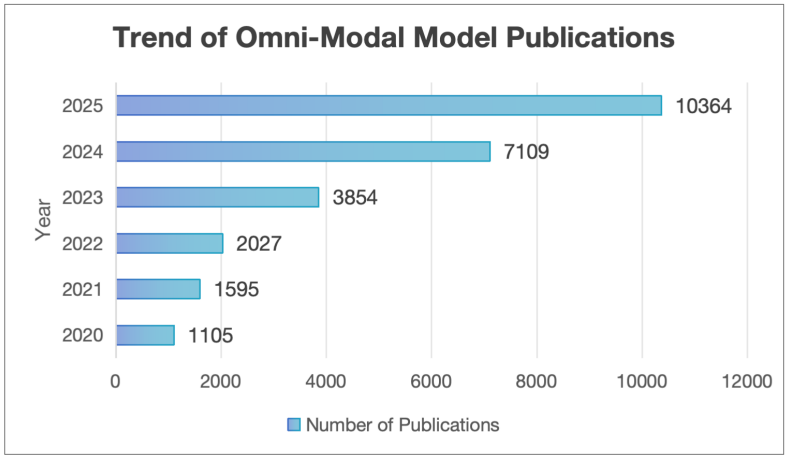

Figure 2. Number of omni-modal model publications (2020–2025). Data collected from IEEE, ACM, and arXiv.

Inclusion Criteria: (1) Peer-reviewed papers or high-impact preprints focusing on architecture, training paradigms, or evaluation of OMLMs; (2) Studies introducing new benchmarks or surveys with methodological insights.

Exclusion Criteria: (1) Non-academic reports or incomplete technical documents; (2) Papers lacking methodological or evaluative contributions.

All included works were categorized into three analytical dimensions: (a) architecture and alignment mechanisms, (b) deployment and vertical adaptation, and (c) evaluation and benchmarking.

This methodological framework ensures both coverage and reproducibility of our review process.

3. Related work and comparative analysis

Recent surveys on multimodal and foundation models have provided valuable insights into the integration of vision, language, and audio modalities. However, most existing reviews treat these modalities separately or within loosely connected frameworks, without establishing a unified perspective on omni-modality. In contrast, our work systematically reviews omni-modal language models (OMLMs) by connecting architectural design, deployment optimization, and evaluation evolution within a single analytical framework. Table 1 summarizes representative recent studies and highlights their main focuses, methodologies, and limitations.

3.1. Recent representative studies

Several recent works have advanced the omni-modal paradigm from distinct angles:

- OmnixR [15] (arXiv:2410.12219) proposes a comprehensive benchmark suite for evaluating omni-modal language models on reasoning tasks across modalities, emphasizing logical inference and compositional generalization.

- Ola [16] (arXiv:2502.04328) pushes the boundaries of omni-modality through scalable modality adapters and joint reasoning pathways, highlighting end-to-end training efficiency.
- Capybara-OMNI [17] (arXiv:2504.12315) introduces an efficient construction paradigm featuring modular design and compression strategies, aiming for lightweight yet expressive omni-modal modeling.

While these works make significant contributions toward omni-modal reasoning, unified architectures, and efficient learning, each primarily focuses on a single aspect of the broader OMLM ecosystem. In contrast, this paper offers a holistic synthesis that integrates architectural, deployment, and evaluative perspectives, bridging theoretical foundations and practical implementations.

Table 1. Comparison of representative surveys and benchmark studies on omni-modal models.

Work	Focus	Methodology	Limitations
OmnixR (2024)	Cross-modal reasoning evaluation	Unified benchmark for multi-domain reasoning tasks	Limited discussion on deployment and scalability
Ola (2025)	Unified training and architecture design	Dynamic modality adapters with scalable fusion	Focuses on architecture; lacks comprehensive evaluation synthesis
Capybara-OMNI (2025)	Efficient paradigm for OMLM construction	Modular composition and compression strategies	Limited benchmark coverage and comparative grounding
This Work (2025)	Integrated OMLM framework	Synergy across Architecture, Deployment, and Evaluation	Provides critical synthesis and forward-looking insights

3.2. Distinct contributions of this survey

Unlike prior surveys that focus narrowly on either model architectures or benchmarking, this work provides a unified analytical framework that: (1) bridges architectural design, deployment optimization, and evaluation methodology; (2) critically analyzes trade-offs among alignment precision, fusion flexibility, and scalability; (3) anchors evaluation dimensions to real-world tasks and datasets; and (4) offers a curated supplementary repository to support reproducibility and future research.

4. Core architectures of general OMLMs

From a technical implementation perspective, the core challenges of omni-modal language models primarily focus on three dimensions: modality alignment, semantic fusion, and joint representation learning. Modality alignment aims to establish semantic mappings between the feature spaces of different modalities. For example, aligning abstract concepts in text (such as “happiness”) with visual elements in images (such as smiling faces or celebratory scenes) [18]. This process typically relies on weak supervision signals from large-scale multimodal data, constructing training samples through text-image co-occurrence, audiovisual synchronization, and other methods, and achieving implicit alignment through contrastive learning or generative objectives (such as text-to-image generation) [19]. However, the inherent differences between modalities, such as discrete systems in text and continuous signals in images/audio, lead to inconsistencies in semantic granularity, causing distortions and expression errors in the alignment space [20]. To alleviate this issue, mainstream methods introduce cross-modal projection modules (such as projection matrices or cross-attention mechanisms) or shared latent space modeling [21] (such as

latent variable models). However, a trade-off still exists between alignment precision and computational overhead.While contrastive alignment remains the dominant paradigm for mapping multimodal feature spaces, it struggles to capture semantic ambiguity and abstraction in high-level concepts (e.g., "hope," "emotion," or "aesthetic style"). Most alignment strategies are heavily dependent on paired data and thus lack robustness when encountering imperfect correspondence between modalities. Moreover, fixed projection modules often constrain generalization across unseen domains. Future research should move toward probabilistic or adaptive alignment mechanisms, where the mapping between modalities is dynamically inferred based on context or task intent. Promising directions include Bayesian cross-modal mapping, latent space uncertainty modeling, and task-aware contrastive objectives, which can better preserve semantic diversity while reducing overfitting to specific modality pairs. While contrastive alignment dominates current practice, it struggles with semantic ambiguity and limited generalization across unseen domains. Most models depend on paired data, making them sensitive to imperfect modality correspondence. Future work should emphasize probabilistic or context-adaptive alignment, enabling dynamic and interpretable mappings between modalities.

Semantic fusion aims to dynamically integrate information and adjust the weights of different modalities in multimodal collaborative tasks. For example, in image question answering, the model needs to adjust the attention proportion between the image and language modalities depending on the attributes of the question (such as "object color" or "scene atmosphere") [22]. If the fusion mechanism overly depends on a single modality, it can lead to inference bias or information loss. Researchers have proposed gating mechanisms and dynamic routing networks [23] to achieve adaptive information flow scheduling between modalities. However, in tasks like multi-turn dialogue or complex cognitive reasoning, the robustness and generalization ability of these methods still require further improvement. Dynamic routing and gating networks improve multimodal fusion but still rely on task-specific tuning and offer limited interpretability. A promising direction is graph-structured or reinforcement-driven fusion, allowing adaptive information flow while maintaining explainable reasoning paths.

Joint representation learning seeks to construct a unified cross-modal semantic space, enabling the feature representations of different modalities to perform inference, matching, and generation within the same vector space. This goal is typically achieved through self-supervised learning tasks (such as masked language modeling, cross-modal contrastive learning, and text-image joint generation) [24]. However, due to significant differences in semantic expression granularity between modalities (e.g., word-level *vs.* pixel-level), directly merging encoder structures may lead to information distortion or semantic dilution. Therefore, some approaches introduce hierarchical modeling structures or modality-specific encoders [25] to preserve modality specificity within shared representations. However, finding the optimal balance between specificity and generalization remains a key challenge.

The Table 2 below compares the core characteristics and technical paths of the models mentioned.Unified representations often face a trade-off between shared generalization and modality specificity. Future work may focus on hierarchical disentanglement and continual cross-modal learning, enabling flexible adaptation without semantic loss.

Table 2. Comparison of core features and technical paths of various omni-modal large models.

Model	Modality alignment	Semantic fusion	Joint representation	Core advantages	Potential limitations
ChatGPT-4o	Implicit alignment	Attention-weight distribution	Shared Transformer architecture	Efficient parameter use; strong dynamic interaction	Limited modality extensibility; dependent on pretraining
Gemini-2.5	Contrastive learning	Gating + dynamic fusion	Modular adapter framework	Lightweight; easy modality expansion	Info synchronization delay; weaker real-time performance
Qwen2.5-VL	Hierarchical alignment	Dynamic routing network	Hierarchical joint representation	Balances specificity and generality	Requires hierarchical annotations; high training cost
InternVL3	Graph-structure modeling	GNN message passing	Cross-modal graph representation	Strong logical consistency in reasoning	High computational overhead; slower inference
Gemini-2.0	Multimodal mask prediction	Task-driven fusion	Multi-task shared semantic space	High alignment accuracy; good modality complementarity	Multi-task conflicts; single-task performance degradation

Note: The comparison highlights the architectural diversity, fusion mechanisms, and trade-offs in current omni-modal large models, providing insight into design evolution and optimization trends.

5. Deployment and vertical adaptation

Overall, current general omni-modal large models have formed differentiated paths in addressing the challenges of modality alignment, semantic fusion, and joint representation learning. ChatGPT-4o and Gemini-2.5 focus on parameter efficiency and modality extensibility, Qwen2.5-VL emphasizes semantic layering and specificity retention, InternVL3 focuses on logical consistency modeling, and Gemini-2.0 stresses task unity and modality complementarity. Future research could further explore lightweight designs for modality adapters, cross-modal generalization mechanisms for unified semantic spaces, and efficient modeling methods for large-scale cross-modal semantic reasoning tasks, to promote the deep integration and practical deployment of omni-modal large models in both general and vertical domains.

In cross-industry applications, omni-modal large models face bottlenecks in the generalization capabilities of their universal architecture and urgently require the de-velopment of scene-specific adaptation mechanisms to address the demands of specific industries. Vertical domains such as healthcare, education, and industry exhibit high heterogeneity in modality structure, task types, deployment environments, and resource constraints, making it difficult for general models to be directly transferred to meet practical needs. To achieve practical deployment, omni-modal models need to be custom-designed around three key dimensions: modality pruning, domain knowledge injection, and re-al-time optimization. This involves constructing efficient adaptation paths tailored to business objectives. This paper will systematically explore the application strategies and evolutionary trends of omni-modal models in typical vertical scenarios from three aspects: scene perception constraints, technical path construction, and multi-level optimization mechanisms.

5.1. Vertical domain characteristics and model adaptation constraints

Vertical domains such as healthcare, education, and industry impose much higher adaptation thresholds on models compared to general tasks. Healthcare tasks emphasize multi-source information analysis, strong logical reasoning, and explainability modeling. For example, in lung nodule detection, the model needs to establish causal relationships between CT images and the symptoms and medication information in medical records, while using significance modeling and medical knowledge injection mechanisms to enhance diagnostic performance [23]. Additionally, mechanisms such as Grad-CAM and Chain-of-Thought are widely used to improve model explainability [24].

In the smart education scenario, the focus is on dynamic modeling of students' cognitive states and emotional feedback, involving multi-source data fusion such as text, video, facial expressions, and behavior [25]. Research introduces multimodal emotion recognition, Transformer-CRF structures, and Cognitive Diagnostic Models (CDM) for personalized modeling [26]. Edge deployment and high-frequency interaction characteristics also drive the development of model distillation and computational power-aware optimization [27].

In the industrial quality inspection scenario, omni-modal models must process heterogeneous modalities such as images, text, acoustics, and vibration signals in high-concurrency and low-latency assembly line environments, achieving defect detection and root cause analysis [28, 29]. To meet deployment constraints, research focuses on introducing MM-GNN, modality attention fusion modules, and FPGA/streaming inference techniques [30].

In summary, different industries have differentiated requirements for models in terms of modality adaptability, knowledge integration capabilities, and resource sensitivity, driving omni-modal models to evolve toward structural flexibility, semantic plasticity, and deployment-friendly designs.

5.2. Modality pruning: lightweight design for scenario-based architectures

Resource constraints and response latency have become key bottlenecks in the deployment of models in vertical domains. Modality pruning has gradually become a mainstream optimization strategy. This mechanism includes both static modality reduction and dynamic modality scheduling. The latter activates modality paths based on attention gating, confidence scoring, or reinforcement learning strategies, realizing the principle of "enable when necessary, prune when non-critical" [31,32].

In healthcare scenarios, modality selection must match the diagnostic stage. For example, in skin disease diagnosis, the image modality is retained while the speech path is removed; in tumor follow-up, a gating mechanism dynamically weighs the importance of image and text information [33]. Some studies further propose task-aware pruning strategies, optimizing redundant modality path weights through data distribution fine-tuning to adapt to scenarios such as portable terminal devices [34].

In educational tasks, modalities are adjusted according to the teaching phase. Researchers have constructed a Phase-Aware scheduling mechanism that activates different modalities during the lecture, practice, and feedback stages, and introduces a modality buffer pool to preload high-frequency modality features [35].

In industrial quality inspection, the focus is more on balancing accuracy and efficiency. A typical approach is to construct a phase-based inference structure, such as using lightweight computer vision (CV) algorithms in the preprocessing stage, enabling full-modal fusion during detection, and ultimately

completing the judgment with a low-complexity structure [36]. Additionally, Modular Configurable Models (such as MoE structures) have been employed to retain multimodal capabilities during training, with on-demand loading during inference to achieve sparse activation and path compression [37,38].

5.3. Domain knowledge injection: explicit and implicit fusion mechanisms

General pre-trained models struggle to meet the specialized knowledge requirements of fields such as healthcare, education, and industry. Domain knowledge injection has become a key path for enhancing model performance and interpretability, and can be divided into two main mechanisms: explicit modeling and implicit regularization.

In medical diagnosis, research widely utilizes medical knowledge graphs, graph neural networks, and knowledge distillation techniques to improve semantic modeling and causal reasoning capabilities [39,40]. By aligning "nodule morphology" with entities in clinical text, models can jointly focus on key lesion areas in both image and text [41]. Knowledge-enhanced encoders have also been used to embed ICD codes and clinical pathways into the multimodal semantic space, facilitating the tri-modal collaborative modeling of language, images, and knowledge [42,43].

In educational tasks, the focus is on cognitive modeling and adapting to student differences. Bloom's Taxonomy is often integrated into training labels to simulate the teacher's instructional logic, guiding the model to develop cognitive level perception [44]. Additionally, by integrating teacher style simulation and student profiling, systems can dynamically predict cognitive states based on interaction data, enabling strategy-based knowledge transfer [45]. Multimodal feedback mechanisms, such as speech, facial expressions, and interaction behaviors, are used to form a knowledge—behavior—feedback closed loop [46,47].

In industrial scenarios, lightweight knowledge integration methods are preferred, such as constructing process parameter embedding layers and mechanical prior filters. These methods introduce process rules in a structured manner into the visual judgment process, reducing false alarm rates [48]. Tensor Fusion Attention mechanisms have also been used to adaptively adjust modality weights based on signal quality, enabling more efficient modality collaboration [49].

In all three scenarios, knowledge injection pathways manifest as structural graphs (healthcare), cognitive constraints (education), and rule fusion (industry), providing stable support for model performance and interpretability.

5.4. Real-time optimization: engineering transformation for latency-sensitive scenarios

In scenarios such as industrial quality inspection, remote healthcare, and smart education, omni-modal models must meet low-latency requirements. Traditional end-to-end omni-modal structures often lead to computational redundancy, necessitating systematic modifications in model compression, modality scheduling, and edge-cloud collaboration.

In industrial quality inspection, structural compression techniques such as channel pruning, low-rank decomposition, and sparse activation can reduce model parameters by over 70% while maintaining accuracy [50,51]. Dynamic scheduling strategies activate high-computation modalities based on anomaly probabilities, maintaining lightweight paths during routine inspections [52]. Combined with region-based attention and sparse Transformers, updates are triggered only for significant areas in images, effectively

reducing computational load [53,54]. Deploying the system to FPGA or edge GPUs for streaming processing enables millisecond-level response times [55,56].

In healthcare scenarios, the focus is on edge deployment and bandwidth optimization. Techniques such as INT8/mixed-precision quantization and modular distributed deployment (e.g., local image encoding and cloud-based language generation) can significantly shorten response times [57, 58]. A dynamic routing mechanism can adjust the model structure based on the diagnostic stage, such as using lightweight models during initial screening and full models for critical stages [59]. Additionally, modality compression selectively transmits data from high-diagnostic-value areas to optimize remote collaboration efficiency [60]. The introduction of uncertainty estimation mechanisms enhances diagnostic reliability [61].

In educational scenarios, models adopt a layered edge-cloud and asynchronous response architecture. Edge modules handle high-frequency tasks such as student attention and emotion recognition, while the cloud performs complex reasoning, achieving a "fast feedback—slow optimization" collaborative mechanism [62,63]. By combining edge-cloud caching and lightweight parameter fine-tuning modules (e.g., LoRA, Adapter), the system can quickly adapt to scenario changes, improving interaction response and transferability [64,65].

Latency optimization has shifted from single-point compression to system-level collaboration. In the future, integrating neural architecture search (NAS) and intelligent scheduling mechanisms will likely enable the efficient deployment of large models in high-real-time scenarios.

5.5. Scenario validation and continuous iteration mechanism

The long-term performance of omni-modal models in vertical scenarios depends not only on static metrics but also on their adaptability and update mechanisms in dynamic environments. Research is shifting from offline accuracy evaluation to a closed-loop optimization model that incorporates scenario validation, online feedback, and incremental iteration.

In healthcare scenarios, multi-center data validation is used to assess the model's cross-institution generalization ability. For example, in lung nodule detection, data from different device brands is incorporated to uniformly evaluate the robustness of the model [66]. Uncertainty estimation methods such as MC-Dropout are employed to select low-confidence samples, combined with expert reviews for local retraining, thus enhancing stability [67,68]. Modal consistency validation (e.g., using CLIP cosine similarity) further enhances diagnostic reliability [69]. In system integration, the model's predictions are linked with PACS systems to establish a "micro-closed-loop feedback" mechanism, enabling product-level deployment.

In the field of intelligent education, A/B testing is used to evaluate the impact of different modality combinations on teaching effectiveness, achieving dynamic task scheduling and path recommendations [70]. Reinforcement learning and MDP methods have been employed to adjust the pace of teaching and resource allocation based on student confusion feedback [71]. Additionally, a student cognitive state transition map is built to support the model in mapping learning paths and controlling rhythm.

In industrial quality inspection systems, defect-process closed-loop feedback mechanisms are constructed based on SCM and temporal graph networks. Model outputs are used for real-time process adjustments and to feed back into model retraining [72,73]. To ensure stability, elastic rollback and staged

iterative training mechanisms are introduced to avoid false alarms that may disrupt production processes [74].

Dynamic feedback mechanisms are becoming the infrastructure for the deployment of omni-modal large models. By establishing a positive feedback loop of "model → quality control → process → retraining," the system transitions from perceptual intelligence to decision-making intelligence, enabling sustainable optimization and autonomous iteration for industrial-grade deployment capabilities.

6. Evaluation systems and benchmarks

The evaluation of omni-modal large models needs to break through the limitations of traditional single-modal benchmarks and construct a comprehensive evaluation system covering multi-level capabilities. Current research has shifted from focusing solely on task performance to multi-dimensional capability diagnostics. This shift involves designing evaluation frameworks that encompass general perception, spatial cognition, visual quality assessment, infographic understanding, multi-modal reasoning, and cross-modal generation. These frameworks systematically measure the model's generalization and robustness in complex scenarios [75]. However, the existing evaluation systems still face challenges such as insufficient metric coverage, task design homogenization, and human cognition alignment bias. It is essential to continuously improve these systems by integrating the latest developments in the field.

Table 3 summarizes the main benchmarks and metrics across five evaluation dimensions of omni-modal language models (OMLMs), covering both perception and reasoning abilities. General Perception and Spatial Cognition address visual recognition and 3D scene understanding, while Visual Quality focuses on perceptual fidelity. Multimodal Reasoning assesses cross-modal information integration, and Infographic Understanding evaluates comprehension of structured visual-textual data. This framework provides a concise basis for analyzing the comprehensiveness of OMLM evaluation.

Table 3. Representative benchmarks and metrics corresponding to each evaluation dimension.

Evaluation dimension	Representative tasks	Typical datasets	Common metrics
General perception	Image classification; object detection	ImageNet; COCO	Accuracy; mAP
Spatial cognition	3D scene understanding; depth estimation	ScanNet; SUN RGB–D	IoU; Chamfer Distance; Depth RMSE
Visual quality	Image/video quality assessment	AVA; CLIVE; LIVE	MOS; PSNR; SSIM
Multimodal reasoning	Visual question answering; causal reasoning	VQA; CLEVRER; ScienceQA	Accuracy; Consistency; Reasoning Score
Infographic understanding	Chart and diagram comprehension	InfographicVQA; ChartQA; PlotQA	EM; BLEU; F1

Note: Each evaluation dimension is anchored to publicly available datasets and standard metrics to enhance reproducibility and comparability in OMLM assessment.

6.1. Hierarchical evaluation of general perception ability

General perception ability is the foundation for building the cognitive system of omni-modal large models. Its evaluation must cover the entire process, from the extraction of raw signals to high-level

semantic reasoning, reflecting whether the model can construct a unified and semantically consistent multimodal representation space after cross-modal fusion. To achieve this, current research is constructing a layered and multi-dimensional evaluation system that encompasses basic perception, mid-level semantic alignment, high-level understanding, and dynamic robustness.

At the basic perception level, evaluation focuses on the model's foundational understanding of single-modal inputs, including image classification and object detection (e.g., ImageNet, COCO), audio recognition (e.g., AudioSet), and language parsing (e.g., SQuAD). These tasks are used to verify the model's precise ability to extract key visual features, acoustic signals, and semantic structures, laying the foundation for subsequent modality fusion.

Mid-level semantic alignment focuses on the accuracy of semantic mapping between different modalities. Evaluation tasks include image-text retrieval, cross-modal pointing, and audio-video alignment, emphasizing the model's ability to perform spatial pointing, temporal alignment, and semantic consistency modeling. Datasets such as Flickr30K, RefCOCO, and AVA are widely used to validate the model's cross-modal matching and positioning performance. Additionally, frameworks like MME and SEED-Bench provide multi-task joint testing frameworks to effectively measure the model's fusion consistency and generalization ability [76].

In the evaluation of high-level semantic understanding and reasoning abilities, the focus shifts to whether the model has the ability to perform multi-modal abstraction and causal logic modeling. Typical tasks such as VQA and ScienceQA assess the model's joint question-answering ability, while tasks like CLEVRER and VCR examine its event reasoning, common-sense transfer, and semantic consistency judgment abilities. Some evaluations also construct samples with conflicting modalities to test the model's ability to identify semantic contradictions across modalities, reflecting its level of cross-modal common-sense reasoning [77].

Furthermore, model robustness and fault tolerance have become key areas of current research. Typical methods include interference tests like modality masking, style perturbation, and semantic misalignment, which assess the model's performance stability under missing or abnormal inputs. For instance, "modality masking tests" and "crash point analysis" are used to quantify the model's tolerance threshold for redundant modalities, measuring its robustness and stability when a modality is missing.

As model application scenarios expand into dynamic environments, static task evaluations have started to show limitations. Emerging evaluation mechanisms such as real-time video semantic alignment, closed-loop task verification in simulated environments (e.g., digital twin in medical scenarios), and human-in-the-loop interaction assessments are rapidly developing. By incorporating user feedback, physiological signals, and contextual data, multimodal evaluation is moving towards a more realistic and interactive comprehensive paradigm, adapting to the trend of omni-modal systems evolving into task decision bodies.

6.2. Three-dimensional expansion of spatial cognitive ability

Spatial cognitive ability is a key foundation supporting omni-modal large models toward embodied intelligence. Early evaluations mainly relied on two-dimensional visual tasks, such as object detection, image segmentation, and scene recognition, to measure the model's ability to identify object categories,

locations, and scene attributes at the image level. However, as multi-modal systems increasingly demand a deeper understanding of the environment, relying solely on 2D perception is no longer sufficient to meet the complex requirements for structural restoration, spatial relationship modeling, and geometric reasoning [78]. The spatial cognitive evaluation system is systematically expanding from two-dimensional to three-dimensional dimensions.

In recent years, researchers have introduced three-dimensional point clouds, RGB-D images, and multi-view reconstruction to advance models' ability to represent and reconstruct three-dimensional structures. A typical example is the ScanNet benchmark dataset, which reconstructs indoor scene geometries using RGB-D input and evaluates them based on metrics such as Chamfer Distance and Normal Consistency [79]. These tasks not only assess the model's geometric modeling accuracy but also emphasize its understanding of occluded objects, topological structures, and overall spatial consistency.

Further evaluation systems are extending the focus from static structural reconstruction to task-driven spatial behavior reasoning. For example, in simulation environments like 3D-FRONT, AI2-THOR, and Gibson, the model needs to perform complex tasks such as path planning, target navigation, or object interaction. These tasks rely not only on accurate three-dimensional structure perception but also require the model to understand physical causality, motion laws, and environmental feedback, integrating spatial cognition with action generation [80].

Despite the growing number of three-dimensional evaluation tasks, widespread application still faces practical challenges. First, acquiring and annotating high-quality three-dimensional data is costly, and existing datasets still lack scale, task diversity, and cross-domain transferability. Second, current evaluations are still largely centered around the visual modality, lacking integration of non-visual sensory signals such as haptic feedback, IMU, and force feedback, which limits the model's generalization ability in real-world robotic systems. Especially in embodied intelligence scenarios, spatial understanding requires multi-modal fusion, including visual, dynamic, and contact sensing, to achieve robust environmental modeling.

Therefore, the future development of spatial cognitive evaluation systems should focus on: (1) multi-modal input fusion; (2) end-to-end task-driven verification; (3) simulation-to-real-world transfer. In particular, for tasks such as 3D navigation, manipulation, and causal reasoning, building dynamic test environments with real interactions, traceable behavioral feedback, and high semantic complexity will become the core support for evaluating the embodied cognitive abilities of omni-modal large models.

6.3. Fusion of objective and subjective visual quality evaluation

Visual quality evaluation is an important dimension in assessing the performance of omni-modal large models, directly influencing their practicality and user experience in tasks such as image generation, reconstruction, and enhancement. Traditional objective metrics like Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) primarily quantify low-level consistency in images based on pixel differences and local structural similarity. However, as the capabilities of generative models improve, the limitations of these metrics in evaluating semantic consistency, naturalness, and realism become more apparent. For instance, a generated image might be pixel-wise close to a real image, but exhibit significant deviations in semantic logic or visual plausibility, which traditional metrics fail to capture.

To bridge the gap between such evaluations and human perception, researchers have introduced perceptually driven subjective modeling methods. For example, NIMA (Neural Image Assessment) trains models using large-scale subjective scoring data to simulate human judgments on image aesthetics and naturalness. LPIPS (Learned Perceptual Image Patch Similarity) measures the distance between image patches in deep feature space, demonstrating higher perceptual relevance in semantic consistency judgments [81]. At the same time, statistical distribution matching metrics like Fréchet Inception Distance (FID) and Inception Score (IS) are widely used in open-domain image generation tasks. These metrics compare the distribution distance between generated and real images in deep embedding spaces to measure the authenticity and diversity of the generated results [82].

To further adapt to complex perceptual tasks, adversarial evaluation mechanisms are also being applied to visual quality discrimination. By training discriminators to identify perceptual differences between generated and real images, a flexible quality scoring system with scene adaptability has been established, yielding significant results in tasks such as medical image synthesis and face reconstruction.

However, current objective-subjective fusion evaluation methods still face challenges in cross-cultural contexts, style diversity, and task explainability. The insufficient coverage of subjective scoring data limits the generalization capability of perceptual models. In highly subjective tasks, such as artistic generation and emotional images, the lack of unified evaluation standards also causes issues with repeatability and comparability.

6.4. Complex structural analysis in infographic understanding

Infographic understanding is a crucial dimension in assessing the high-level semantic capabilities of omni-modal large models. It involves the modeling of associations between structured visual elements and text information, as well as the ability for cross-modal reasoning. A typical task like Infographic-VQA requires the model to not only recognize graphic components such as legends and data points but also to understand the relationships and trends between variables in the graph, supporting complex question answering tasks such as comparison and prediction [83]. Compared to earlier document-based question answering tasks, infographic understanding relies more heavily on structural recognition, semantic association, and text-to-image mapping capabilities. In particular, in dynamic infographics or interactive chart scenarios, the model also needs to have causal analysis and temporal reasoning abilities across the time dimension.

However, the current evaluation systems lack systematic coverage of professional graphics (e.g., chemical structures, industrial blueprints) and multi-step logical chains, which limits the comprehensive testing of a model's true understanding abilities. Future evaluation systems should integrate interdisciplinary knowledge graphs and multimodal temporal data to build an integrated testing platform that supports structural analysis, dynamic modeling, and causal reasoning. This will promote the development of infographic understanding tasks towards greater semantic depth and broader application scope.

In conclusion, infographic understanding is not only an extension of a model's perceptual capabilities but also a key evaluation channel for multi-modal semantic reasoning and alignment with human knowledge. Establishing a multi-level evaluation system covering structural analysis, logical modeling,

and temporal reasoning will provide a strong foundation for the reliable deployment of omni-modal large models in fields such as research, education, and healthcare.

6.5. Unified construction of multimodal generation and reasoning evaluation systems

As the capabilities of omni-modal large models continue to expand, the evaluation system is shifting from static perception and modality alignment to the evaluation of reasoning abilities, generative usefulness, and the scalability and reliability of the evaluation mechanisms. Current multimodal reasoning evaluations focus on the model's ability to model semantic relationships, causal chains, and mathematical logic. For example, NLVR2 and MathVista assess the model's spatial reasoning and text-image mathematical reasoning capabilities, respectively [84]. Although some tasks introduce confounding variables to test causal modeling abilities, overall, the evaluations still heavily rely on synthetic data, making it difficult to cover the complex biases and reasoning processes that occur in real-world contexts.

Generation capability evaluation has also expanded from low-level metrics such as image quality and text fluency to task-driven measures of content value and interaction ability, such as semantic consistency in image-text generation (CLIP-Score) or image distribution fitting (FID). However, these evaluations often rely on internal model features, making it challenging to quantify the effectiveness of generated content in specific application scenarios like education and design. Additionally, they lack an auditing system for potential risks such as false content, biases, and privacy leakage. Moreover, the response latency in generation tasks has long been overlooked, limiting the measurement of real-time interaction potential.

From a system perspective, current evaluation tasks focus on a few types such as image-text question answering and image-text pairing, and lack comprehensive coverage of small-sample generalization, interactive consistency, and multi-turn dynamic adaptation. Meanwhile, the integration of subjective and objective evaluations is still immature, leading to discrepancies in evaluation standards and poor comparability of results, making it difficult to support large-scale model rankings and capability analysis.

To improve the evaluation adaptability and practical value of multimodal models, future system development should optimize on three levels: methodologically, by transitioning from static testing to an interactive platform that supports dynamic environments and task transformation; in terms of tools, by introducing self-supervision and counterfactual mechanisms to enhance automatic evaluation capabilities; and value-oriented, by constructing a multidimensional goal metric system that covers accuracy, security, energy consumption, and ethical risks. Ultimately, the evaluation system needs to shift from "performance verification" to "task value measurement," achieving a collaborative evolution between model development and practical feedback.

6.6. Performance comparison and capability analysis of mainstream omni-modal large models

To ensure transparency and consistency, the performance results summarized in Table 4 were collected from publicly available benchmark reports and official model evaluations, rather than from independent experiments. Specifically, data were aggregated from widely used omni-modal evaluation platforms such as MME, SEED-Bench, and OpenCompass, covering multiple perception, reasoning, and generation tasks. To enable fair comparison among heterogeneous benchmarks, all reported values were normalized to a 0–100 scale, where 100 represents the best performance observed within each capability dimension.

Table 4. Evaluation results of mainstream omni-modal large models in six capability dimensions.

Model name	Avg. score	General perception	Spatial understanding	Visual quality	Infographic understanding	Multimodal reasoning	Multimodal generation
Seed1.5-VL	59.85	81.9	57.5	50.0	75.0	46.61	61.32
Gemini-2.5-Pro	53.86	63.81	51.25	47.5	61.25	46.61	59.96
Qwen2.5-VL-72B	48.25	78.1	53.75	52.5	51.25	23.08	56.01
ChatGPT-4o-latest	47.49	60.0	53.75	62.5	55.0	19.91	61.34
GPT-4.1	46.46	55.24	51.25	67.5	51.25	20.36	59.24

Note: All scores are synthesized from publicly available benchmark reports and leaderboard results (MME, SEED-Bench, and OpenCompass, 2023–2025). Values are normalized to a 0–100 scale for comparability across benchmarks; no independent experiments were conducted by the authors.

Seed1.5-VL (ByteDance) demonstrates leading performance in general perception (81.9) and infographic understanding (75.0), reflecting its strong capabilities in low-level feature extraction and structured content analysis. However, its relatively low score in visual quality (50.0) suggests limitations in fine-grained image generation and aesthetic rendering.

Gemini-2.5-Pro (Google) maintains a balanced performance across most dimensions, particularly in multimodal reasoning and generation, making it well-suited for general-purpose tasks. While its performance in visual quality and infographic understanding slightly exceeds the average, it lacks prominent strengths in any single capability.

Qwen2.5-VL-72B (Alibaba) excels in general perception (78.1) and performs adequately in spatial understanding and multimodal generation. However, its notably low score in multimodal reasoning (23.08) indicates challenges in integrating complex semantics and modeling causal relationships across modalities.

ChatGPT-4o-latest and GPT-4.1 from OpenAI exhibit competitive performance in visual generation, achieving high scores of 62.5 and 67.5, respectively. This reflects their strong capabilities in image modeling and style retention. Nevertheless, both models score below 21 in multimodal reasoning, suggesting limitations in abstract concept modeling and reasoning consistency across modalities.

As summarized in Table 4, the current mainstream omni-modal models exhibit distinct strengths across capability dimensions. Seed1.5-VL is particularly effective in tasks requiring perception and structured comprehension. Gemini-2.5-Pro prioritizes stability and broad applicability, while Qwen2.5-VL remains strong in perception yet underperforms in reasoning. OpenAI's models are visually expressive but still face bottlenecks in cross-modal logical consistency.

Looking forward, future omni-modal model architectures should emphasize decoupled capability modules and support dynamic scheduling strategies that adapt to varying task demands. This will be essential for enabling efficient multi-task collaboration and robust adaptation in complex, real-world environments.

7. Future trends and outlook

Omni-modal large models, as an important path toward general artificial intelligence, face not only technical challenges in model architecture, task generalization, and deployment efficiency but also a pressing need to build a systematic support framework in areas such as ethical security and human-machine collaboration. This section discusses the evolutionary trends and research focuses of omni-modal large models across five key directions.

7.1. Architectural composability and modality scheduling mechanism

As application scenarios continuously diversify and data modality types expand, current omni-modal models are beginning to reveal issues of architectural rigidity and computational redundancy. To improve model adaptation efficiency across different tasks, future architectural designs will evolve toward composability, achieved through modular construction that allows flexible decoupling and on-demand collaboration between modalities. A typical approach is the "Modality-as-a-Service" operating paradigm: under a unified interface, corresponding modality processing modules are dynamically invoked based on the input content, optimizing the resource allocation for the perception-understanding-generation process [85].

Additionally, modality scheduling mechanisms will become a core enabling technology, intelligently determining the extent of modality participation based on task intentions and contextual states. For example, in multi-turn dialogue, the model could infer whether to introduce an emotion recognition modality to assist in determining user intent based on context, thus avoiding redundant computation. Moreover, in scenarios involving edge computing and multi-terminal collaborative deployment, modality caching and asynchronous update mechanisms will play a key role, allowing submodules to activate on demand, transmit with low communication overhead, and respond quickly under resource-constrained conditions.

7.2. Enhancing generalization ability and improving task adaptability

Although omni-modal large models possess cross-modal learning capabilities, they still face challenges such as performance degradation and overfitting when it comes to task transfer and domain generalization. Future model designs should emphasize the principle of "generalization first" by introducing stronger universal semantic priors and task-agnostic structures, guiding the model to quickly adapt under low-resource conditions.

On one hand, during the pre-training phase, cross-task concept mapping mechanisms or semantic consistency regularization terms can be introduced to promote implicit alignment and representation generalization across different modalities.

On the other hand, prompt engineering is increasingly becoming a key technology to enhance task adaptability. By fine-tuning instruction templates, it guides the model to activate specific reasoning paths and semantic focus mechanisms, enabling fast task transformation for zero-shot or few-shot tasks [85]. Additionally, integrating multi-task learning and meta-learning frameworks will help models share knowledge representations across tasks and improve their cross-task generalization and robustness through parameter fine-tuning.

7.3. Resource-constrained energy efficiency optimization path

Currently, omni-modal large models heavily rely on high-performance computational resources during both the training and inference stages, which severely restricts their deployment and application in edge devices, industrial environments, and embedded systems.

To optimize the model's energy efficiency ratio, future research will focus on the following paths: Firstly, at the structural level, model compression techniques will expand from traditional pruning and quantization methods to structure-aware low-rank modeling and sparse activation networks, enabling adaptive inference driven by task complexity. Secondly, at the inference strategy level, a modality

confidence prediction mechanism can be employed to construct flexible execution paths, skipping low-relevance modules while maintaining accuracy, thereby reducing latency and energy consumption. Thirdly, at the system level, the collaborative advantages of heterogeneous computing resources (such as CPU, GPU, NPU, *etc.*) should be fully leveraged, improving overall computational resource utilization through task distribution and model partitioning [86]. Additionally, a green AI evaluation system should be gradually established, guiding future system design and algorithm selection based on energy consumption, carbon emissions, and model lifecycle considerations .

7.4. Ethical and security assurance system for trustworthy intelligence

As omni-modal models are increasingly applied in content generation, contextual reasoning, and interactive decision-making, the ethical and security risks they bring cannot be overlooked. Typical issues include information hallucination, amplification of social biases, and generation of fake content. These problems not only undermine the trustworthiness of model outputs but also may have profound impacts on public opinion, security stability, and individual trust. Therefore, trustworthy intelligence should become the core concept in omni-modal system design. On the technical implementation level, models need to possess interpretability capabilities, supporting explicit labeling of reasoning paths and modality sources to enable traceable decision logic.

At the same time, auditability mechanisms should cover key behaviors such as input-output records and modality activation states, meeting the need for accountability tracing. Additionally, fine-grained controllability designs for generated content and reasoning processes can effectively constrain model behavior, preventing risks of crossing boundaries. On the governance level, efforts should be made to promote the collaborative development of cross-regional legal norms, industry standards, and ethical review mechanisms, especially in highly sensitive fields such as healthcare, justice, and education, ensuring full-cycle compliance supervision in model development, deployment, and application [87].

7.5. Interdisciplinary integration and human-machine symbiotic intelligent systems

Omni-modal large models are gradually evolving from traditional information processing systems into intelligent agents capable of autonomous perception, reasoning, and interaction. This development relies on deep integration of multiple disciplines such as cognitive science, psychology, education, and ethics [88]. For example, cognitive science provides theoretical models on human multimodal perception integration and attention allocation, which help optimize modality weight distribution and path selection mechanisms; psychological and educational theories can be used to model individual differences and interaction preferences, enhancing the system's human-centric adaptability and learning ability; while ethics and social sciences offer value norms and social responsibility constraints for model development, guiding systems toward becoming "responsible" entities.

In the future, omni-modal large models will not only serve as information processing tools but will also be embedded in critical domains such as healthcare, education, design, and art as "symbiotic intelligent agents," contributing to the creation of new intelligent ecosystems characterized by human-machine collaboration, knowledge co-creation, and emotional integration.

8. Conclusion

Omni-modal large models are emerging as a key pathway toward Artificial General Intelligence (AGI), evolving from theoretical concepts to practical systems for multimodal understanding and generation. They achieve major advances in modality alignment, semantic fusion, and representation learning, enabling end-to-end processing from perception to language generation. This paper reviews their core technologies, representative architectures, application adaptations, and evaluation systems to support future research and development.

Current representative models (such as ChatGPT-4o, Gemini series, Qwen2.5-VL) demonstrate different strategies in scalability, parameter efficiency, and task adaptability. The practical deployment outcomes depend not only on model size and pre-training strategies but also on the integration of domain knowledge, response delay control, and engineering adaptation to resource constraints. In complex scenarios like healthcare, industry, and education, the controllability and stability of models determine their potential for real-world implementation.

Despite significant progress, omni-modal models still face challenges such as insufficient semantic alignment, high modality heterogeneity, large training costs, and incomplete evaluation systems. Future optimization directions include: first, exploring composable architectures and task-aware scheduling mechanisms to improve structural flexibility; second, enhancing the generalization capability of semantic spaces and optimizing performance in low-resource modalities; third, constructing training-validation loops to ensure controllability in real-world applications; and fourth, establishing a multi-dimensional evaluation system that integrates perception, reasoning, and generation capabilities to achieve system-wide measurement of generality and task performance.

Overall, omni-modal large models are evolving from perceptual integration toward intelligent collaboration, driving fundamental transformations in modeling paradigms, human–machine interaction, and task execution. Building on this trend, the present survey provides a unified perspective that connects architectural design, deployment strategies, and evaluation frameworks of OMLMs. Unlike previous reviews, it critically synthesizes the trade-offs among alignment precision, fusion flexibility, and deployment scalability, offering theoretical insight and practical guidance for the development of next-generation omni-modal intelligence systems.

Data availability statement

All data analyzed in this review are from publicly available sources cited in the references. The supplementary materials, including the list of reviewed works, benchmark links, and dataset references, are available at https://github.com/Astro825/OMLM-Survey-Supplementary.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62506213), Shandong Postdoctoral Science Foundation (No. SDBX2024004), China Postdoctoral Science Foundation (Nos. 2024M761805 and 2025M771512), Postdoctoral Fellowship Program of CPSF (No. GZB20250390).

Author's contribution

Lu Chen: conceptualization, methodology, writing—original draft, project administration. Jiajie Mu: formal analysis, literature survey, writing—review & editing. Jiarui Wang: data curation, visualization, writing—review & editing. Xiao Kang: investigation, validation, resources. Xiaoming Xi: supervision, funding acquisition, writing—review & editing. Zheyun Qin: conceptualization, methodology, supervision, writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Conflicts of interests

The authors declare no conflict of interest.

References

[1] Jaegle A, Borgeaud S, Alayrac JB, Doersch C, Ionescu C, *et al.* Perceiver io: a general architecture for structured inputs & outputs. *arXiv* 2021, arXiv:2107.14795.

[2] Wiggins WF, Tejani AS. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiol. Artif. Intell.* 2022, 4(4):e220119.

[3] Li LH, Yatskar M, Yin D, Hsieh C, Chang K. Visualbert: a simple and performant baseline for vision and language. *arXiv* 2019, arXiv:1908.03557.

[4] Lu J, Batra D, Parikh D, Lee S. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, Vancouver, Canada, December 8–14, 2019.

[5] Akbari H, Yuan L, Qian R, Chuang W, Chang SF, *et al.* Vatt: transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, Los Angeles, USA, December 6–14, 2021, pp. 24206–24221.

[6] Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, *et al.* Gpt-4 technical report. *arXiv* 2023, arXiv:2303.08774.

[7] Alayrac JB, Donahue J, Luc P, Miech A, Barr I, *et al.* Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, New Orleans, USA, November 28–December 9, 2022, pp. 23716–23736.

[8] Zhang Y, Li H, Liu J, Yue X. Scaling omni-modal pretraining with multimodal context: advancing universal representation learning across modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Honolulu, USA, October 19–23, 2025, pp. 1336–1348.

[9] Guo Q, Song K, Feng Z, Ma Z, Zhang Q, *et al.* M2-omni: advancing omni-mllm for comprehensive modality support with competitive performance. *arXiv* 2025, arXiv:2502.18778.

[10] Lee JO, Zhou HY, Berzin TM, Sodickson DK, Rajpurkar P. Multimodal generative AI for interpreting 3D medical images and videos. *npj Digital Med.* 2025, 8(1):273.

[11] Jiang S, Liang J, Wang J, Dong X, Chang H, *et al.* From specific-MLLMs to omni-MLLMs: a survey on MLLMs aligned with multi-modalities. *arXiv* 2024, arXiv:2412.11694.

[12] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, *et al.* Learning transferable visual models from

natural language supervision. In *The 38th International Conference on Machine Learning (ICML 2021)*, Baltimore, USA, July 18–27, 2021, pp. 8748–8763.

[13] Caffagni D, Cocchi F, Barsellotti L, Moratelli N, Sarto S, *et al.* The revolution of multimodal large language models: a survey. *arXiv* 2024, arXiv:2402.12451.

[14] Bai T, Liang H, Wan B, Xu Y, Li X, *et al.* A survey of multimodal large language model from a data-centric perspective. *arXiv* 2024, arXiv:2405.16640.

[15] Chen L, Hu H, Zhang M, Chen Y, Wang Z, *et al.* Omnixr: evaluating omni-modality language models on reasoning across modalities. *arXiv* 2024, arXiv:2410.12219.

[16] Liu Z, Dong Y, Wang J, Liu Z, Hu W, *et al.* Ola: pushing the frontiers of omni-modal language model. *arXiv* 2025, arXiv:2502.04328.

[17] Ji X, Wang J, Zhang H, Zhang J, Zhou H, *et al.* Capybara-OMNI: an efficient paradigm for building omni-modal language models. *arXiv* 2025, arXiv:2504.12315.

[18] Tsai YHH, Bai S, Liang PP, Kolter JZ, Morency LP, *et al.* Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, Florence, Italy, July 28–August 2, 2019, pp. 6558–6569.

[19] Arevalo J, Solorio T, Montes-y Gómez M, González FA. Gated multimodal units for information fusion. *arXiv* 2017, arXiv:1702.01992.

[20] Zhao X, Bai Z, Zhou M, Ren X, Wang Y, *et al.* Integrating dynamic routing with reinforcement learning and multimodal techniques for visual question answering. In *2024 9th International Conference on Image, Vision and Computing (ICIVC)*, Suzhou, China, July 15–17, 2024, pp. 295–301.

[21] Chen Y, Li L, Yu L, El Kholy A, Ahmed F, *et al.* Uniter: universal image-text representation learning. In *European Conference on Computer Vision*, Glasgow, UK, August 23–28, 2020, pp. 104–120.

[22] Kim JH, Jun J, Zhang BT. Bilinear attention networks. In *Advances in neural information processing systems*, Montreal, Canada, December 3–18, 2018.

[23] Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, *et al.* Coca: contrastive captioners are image-text foundation models. *arXiv* 2022, arXiv:2205.01917.

[24] Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, *et al.* Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, October 22–29, 2017, pp. 618–626.

[25] Zhang Y, He N, Yang J, Li Y, Wei D, *et al.* mmformer: multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Singapore, September 18–22, 2022, pp. 107–117.

[26] Fu Z, Liu F, Xu Q, Fu X, Qi J. LMR-CBT: learning modality-fused representations with CB-transformer for multimodal emotion recognition from unaligned multimodal sequences. *Front. Comput. Sci.* 2024, 18(4):184314.

[27] Zadeh AB, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 15–20, 2018, pp. 2236–2246.

[28] Wu R, Guo X, Du J, Li J. Accelerating neural network inference on FPGA-based platforms—a survey. *Electronics* 2021, 10(9):1025.

[29] Saeed A, A Khan M, Akram U, J Obidallah W, Jawed S, *et al.* Deep learning based approaches for intelligent industrial machinery health management and fault diagnosis in resource-constrained environments. *Sci. Rep.* 2025, 15(1):1114.

[30] Zhao C, Dong Z, Chen Y, Zhang X, Chamberlain RD. GNNHLS: evaluating graph neural network inference via high-level synthesis. In *2023 IEEE 41st international conference on computer design (ICCD)*, Washington DC, USA, November 6–8, 2023, pp. 574–577.

[31] Wen Z, Gao Y, Li W, He C, Zhang L. Token pruning in multimodal large language models: are we solving the right problem? *arXiv* 2025, arXiv:2502.11501.

[32] Ye W, Wu Q, Lin W, Zhou Y. Fit and prune: fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Philadelphia, USA, February 25–March 4, 2025, pp. 22128–22136.

[33] Huang K, Zou H, Xi Y, Wang B, Xie Z, *et al.* Ivtp: instruction-guided visual token pruning for large vision-language models. In *European Conference on Computer Vision*, Milan, Italy, September 29–October 4, 2024, pp. 214–230.

[34] Xia G, Ding Y, Li F, Ren L, Chen W, *et al.* SMAR: soft modality-aware routing strategy for MoE-based multimodal large language models preserving language capabilities. *arXiv* 2025, arXiv:2506.06406.

[35] Gholami A, Kim S, Dong Z, Yao Z, Mahoney MW, *et al.* *A survey of quantization methods for efficient neural network inference*, 1st ed. Boca Raton: Chapman and Hall/CRC, 2022. pp. 291–326.

[36] Tian Y, Yang Z. SAEC: scene-aware enhanced edge-cloud collaborative industrial vision inspection with Mmultimodal LLM. *arXiv* 2025, arXiv:2509.17136.

[37] Shen L, Chen G, Shao R, Guan W, Nie L. Mome: mixture of multimodal experts for generalist multimodal large language models. In *Neural Information Processing Systems (NeurIPS) 2024*, Vancouver, Canada, December 10–15, 2024, pp. 42048–42070.

[38] Cai W, Jiang J, Wang F, Tang J, Kim S, *et al.* A survey on mixture of experts in large language models. *arXiv* 2025, arXiv:2407.06204.

[39] Riquelme C, Puigcerver J, Mustafa B, Neumann M, Jenatton R, *et al.* Scaling vision with sparse mixture of experts. In *Advances in Neural Information Processing Systems*, online, December 6–14, 2021, pp. 8583–8595.

[40] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv* 2015, arXiv:1503.02531.

[41] Huang SC, Shen L, Lungren MP, Yeung S. Gloria: a multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, online, October 11–17, 2021, pp. 3942–3951.

[42] Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* 2023, 14(1):4542.

[43] Yin H, Wu W, Hao Y. DKA-RG: disease-knowledge-enhanced fine-grained image–text alignment for automatic radiology report generation. *Electronics* 2024, 13(16):3306.

[44] Li C, Wong C, Zhang S, Usuyama N, Liu H, *et al.* Llava-med: training a large language-and-vision assistant for biomedicine in one day. In *Neural Information Processing Systems 2023*, Vancouver,

Canada, December 3–9, 2023, pp. 28541–28564.

[45] Wang F, Liu Q, Chen E, Huang Z, Chen Y, *et al.* Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI Conference On Artificial Intelligence*, New York, USA, February 7–12, 2020, pp. 6153–6161.

[46] Xie Y, Yang L, Zhang M, Chen S, Li J. A review of multimodal interaction in remote education: technologies, applications, and challenges. *Appl. Sci.* 2025, 15(7):3937.

[47] Wu Y, Mi Q, Gao T. A comprehensive review of multimodal emotion recognition: techniques, challenges, and future directions. *Biomimetics* 2025, 10(7):418.

[48] Modirrousta M, Memarian A, Huang B. Causal discovery in industrial systems via physics-guided variational attention and probabilistic interventions. *Comput. Chem. Eng.* 2025, 204:109420.

[49] Zadeh A, Chen M, Poria S, Cambria E, Morency LP. Tensor fusion network for multimodal sentiment analysis. *arXiv* 2017, arXiv:1707.07250.

[50] Xie Z, Zhu L, Zhao L, Tao B, Liu L, *et al.* Localization-aware channel pruning for object detection. *Neurocomputing* 2020, 403:400–408.

[51] Han S, Pool J, Tran J, Dally W. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, Montreal, Canada, December 7–12, 2015.

[52] Howard A, Sandler M, Chu G, Chen LC, Chen B, *et al.* Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, Seoul, Republic of Korea, October 29–November 1, 2019, pp. 1314–1324.

[53] Angrist N, Beatty A, Cullen C, Matsheng M. A/B testing in education: rapid experimentation to optimise programme cost-effectiveness. 2024. Available: https://www.wwhge.org/resources/a-b-testing-in-education-rapid-experimentation-to-optimise-programme-cost-effectiveness/ (accessed on 12 May 2025).

[54] Zhu L, Wang X, Ke Z, Zhang W, Lau RW. Biformer: vision transformer with bi-level routing attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023*, Vancouver, Canada, June 18–22, 2023, pp. 10323–10333.

[55] Child R, Gray S, Radford A, Sutskever I. Generating long sequences with sparse transformers. *arXiv* 2019, arXiv:1904.10509.

[56] Montgomerie-Corcoran A, Toupas P, Yu Z, Bouganis CS. SATAY: a streaming architecture toolflow for accelerating YOLO models on FPGA devices. In *2023 International Conference on Field Programmable Technology (ICFPT)*, Yokohama, Japan, December 12–14, 2023, pp. 179–187.

[57] Wang T, Guo J, Zhang B, Yang G, Li D. Deploying AI on edge: advancement and challenges in edge intelligence. *Mathematics* 2025, 13(11):1878.

[58] Guo Y. A survey on methods and theories of quantized neural networks. *arXiv* 2018, arXiv:1808.04752.

[59] Huang Q, An Z, Zhuang N, Tao M, Zhang C, *et al.* Harder tasks need more experts: dynamic routing in moe models. *arXiv* 2024, arXiv:2403.07652.

[60] Ma Y, Yi C, Zhou Y, Wang Z, Zhao Y, *et al.* Semantic redundancy-aware implicit neural compression for multidimensional biomedical image data. *Commun. Biol.* 2024, 7(1):1081.

[61] Kurz A, Hauser K, Mehrtens HA, Krieghoff-Henning E, Hekler A, *et al.* Uncertainty estimation in

medical image classification: systematic review. *JMIR Med. Inf.* 2022, 10(8):e36427.

[62] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? In *31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, USA, December 4–7, 2017.

[63] Liu J, Du Y, Yang K, Wu J, Wang Y, *et al.* Edge-cloud collaborative computing on distributed intelligence and model optimization: a survey. *arXiv* 2025, arXiv:2505.01821.

[64] Wang Z, Ma H, Zhai J. Low-rank adaptation for edge AI. *Sci. Rep.* 2025, 15(1):33109.

[65] Matsutani H, Kondo M, Sunaga K, Marculescu R. Skip2-LoRA: a lightweight on-device DNN fine-tuning method for low-cost edge devices. In *Proceedings of the 30th Asia and South Pacific Design Automation Conference*, Tokyo, Japan, January 20–23, 2025, pp. 51–57.

[66] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, *et al.* Lora: low-rank adaptation of large language models. *arXiv* 2021, arXiv: 2106.09685.

[67] Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, *et al.* The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med. Phys.* 2011, 38(2):915–931.

[68] Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML 2016)*, New York, USA, June 19–24, 2016, pp. 1050–1059.

[69] Sensoy M, Kaplan L, Kandemir M. Evidential deep learning to quantify classification uncertainty. In *Neural Information Processing Systems (NeurIPS) 2018*, Montréal, Canada, December 2–8, 2018.

[70] Kizilcec RF. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings Of The 2016 CHI Conference On Human Factors In Computing Systems*, San Jose, USA, May 7–12, 2016, pp. 2390–2395.

[71] Rowe J, Pokorny B, Goldberg B, Mott B, Lester J. Toward simulated students for reinforcement learning-driven tutorial planning in gift. In *Proceedings of R. Sottilare (Ed.) 5th annual GIFT users symposium. Orlando, FL*, Orlando, USA, May 10–11, 2017, p. 4.

[72] Guo L, Shi H, Tan S, Song B, Tao Y. A knowledge-driven spatial-temporal graph neural network for quality-related fault detection. *Process Saf. Environ. Prot.* 2024, 184:1512–1524.

[73] Zhao Z, Xiao Z, Tao J. MSDG: multi-scale dynamic graph neural network for industrial time series anomaly detection. *Sensors* 2024, 24(22):7218.

[74] Antony J, Jalušić D, Bergweiler S, Hajnal Á, Žlabravec V, *et al.* Adapting to changes: a novel framework for continual machine learning in industrial applications. *J. Grid Comput.* 2024, 22(4):71.

[75] Semwal P. A multi-stage machine learning model for security analysis in industrial control system. In *AI-Enabled Threat Detection and Security Analysis for Industrial IoT*, 1st ed. Cham: Springer, 2021. pp. 213–236.

[76] Wei C, Chen Y, Chen H, Hu H, Zhang G, *et al.* Uniir: training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*, Milan, Italy, September 29–October 4, 2024, pp. 387–404.

[77] Li B, Ge Y, Ge Y, Wang G, Wang R, *et al.* Seed-bench: benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

Washington, USA, June 17–21, 2024, pp. 13299–13308.

[78] Fu C, Zhang Y, Yin S, Li B, Fang X, *et al.* Mme-survey: a comprehensive survey on evaluation of multimodal llms. *arXiv* 2024, arXiv:2411.15296.

[79] Dai A, Chang AX, Savva M, Halber M, Funkhouser T, *et al.* Scannet: richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Honolulu, USA, July 21–26, 2017, pp. 5828–5839.

[80] Kolve E, Mottaghi R, Han W, VanderBilt E, Weihs L, *et al.* Ai2-thor: an interactive 3D environment for visual AI. *arXiv* 2017, arXiv:1712.05474.

[81] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Utah, USA, June 18–22, 2018, pp. 586–595.

[82] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, Long Beach, USA, December 4–9, 2017.

[83] Mathew M, Bagal V, Tito R, Karatzas D, Valveny E, *et al.* Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, Hawaii, USA, January 4–8, 2022, pp. 1697–1706.

[84] Lu P, Bansal H, Xia T, Liu J, Li C, *et al.* Mathvista: evaluating mathematical reasoning of foundation models in visual contexts. *arXiv* 2023, arXiv:2310.02255.

[85] Chen Z, Xu L, Zheng H, Chen L, Tolba A, *et al.* Evolution and prospects of foundation models: from large language models to large multimodal models. *Comput. Mater. Continua* 2024, 80(2):1753.

[86] Schick T, Schütze H. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv* 2020, arXiv:2001.07676.

[87] Bouderhem R. Shaping the future of AI in healthcare through ethics and governance. *Humanit. Soc. Sci. Commun.* 2024, 11(1):1–12.

[88] Xie J, Chen Z, Zhang R, Wan X, Li G. Large multimodal agents: a survey. *arXiv* 2024, arXiv:2402.15116.