Multimodal Vision Language Models in Interactive and Physical Environments

Lucas Pereira , Martina Kovács , Ahmed El-Masry , Feidlimid Shyama *

Posted Date: 26 December 2025

doi: 10.20944/preprints202512.2407.v1

Keywords: multimodal learning; vision–language models; large language models; human–computer interaction; robotics; human–robot interaction; embodied AI; multimodal reasoning; grounded language; interactive systems

 doi:10.20944/preprints202512.2407.v1

Article

Multimodal Vision Language Models in Interactive and Physical Environments

Lucas Pereira [1], Martina Kovács [2], Ahmed El-Masry [3] and Feidlimid Shyama [4,*]

[1] Department of Electrical and Computer Engineering, University of Coimbra, Portugal
[2] Faculty of Information Technology, University of Pannonia, Hungary
[3] Department of Computer and Systems Engineering, Ain Shams University, Egypt
[4] National Technical University of Athens, Greece
* Correspondence: feidlimid.shyama@ntua.gr

Abstract

Multimodal Large Vision–Language Models (LVLMs) have emerged as a central paradigm in contemporary artificial intelligence, enabling machines to jointly perceive, reason, and communicate across visual and linguistic modalities at unprecedented scale. By integrating advances in large language models with powerful visual representation learning, LVLMs offer a unifying framework that bridges perception, cognition, and interaction. This capability is particularly consequential for Human–Computer Interaction (HCI) and robotic applications, where effective intelligence must be grounded in sensory input, responsive to human intent, and robust in dynamic, real-world environments. This review provides a comprehensive and in-depth examination of LVLMs from the perspective of interactive and embodied systems. We begin by situating LVLMs within the broader evolution of multimodal learning, highlighting the theoretical foundations and mathematical formulations that underpin vision–language alignment, representation fusion, and autoregressive generation. We then analyze dominant architectural paradigms, including dual-encoder models, fusion-based designs, and unified token-based transformers, discussing their respective trade-offs in terms of scalability, grounding fidelity, computational efficiency, and suitability for interaction-driven and robotic contexts. Building on these foundations, the review surveys a wide range of applications in HCI and robotics. In HCI, LVLMs enable visually grounded conversational agents, intelligent user assistance, explainable interfaces, and novel forms of human–AI co-creation that lower barriers to interaction and expand accessibility. In robotics, they support language-guided manipulation, navigation, exploration, and human–robot interaction by linking high-level natural language instructions with perceptual understanding and physical action. Across both domains, LVLMs facilitate generalization, adaptability, and more natural communication, while also exposing new challenges related to reliability, safety, and user trust. We further provide a critical analysis of current limitations and open research problems, including hallucination and weak grounding, limited temporal and causal reasoning, high computational cost, lack of interpretability, dataset bias, and insufficient evaluation methodologies for long-term interaction and embodied performance. These challenges highlight the gap between impressive benchmark results and the demands of real-world deployment. Finally, we outline key future research directions, emphasizing stronger grounding mechanisms, temporal and memory-aware modeling, efficiency and sustainability, human-centered and ethical design, and interdisciplinary evaluation and governance. By synthesizing insights across machine learning, HCI, and robotics, this review frames LVLMs not merely as technical artifacts but as interactive agents embedded in social and physical contexts. Our goal is to provide researchers and practitioners with a holistic understanding of the state of the field, clarify the opportunities and risks associated with deploying LVLMs in interactive and embodied systems, and chart a path toward multimodal AI technologies that are powerful, trustworthy, and aligned with human values.

Keywords: multimodal learning; vision–language models; large language models; human–computer interaction; robotics; human–robot interaction; embodied AI; multimodal reasoning; grounded language; interactive systems

 doi:10.20944/preprints202512.2407.v1

1. Introduction

The past decade has witnessed a profound transformation in artificial intelligence (AI), driven largely by advances in deep learning, large-scale data, and high-performance computing. Among these advances, Large Language Models (LLMs) have emerged as a cornerstone technology, demonstrating unprecedented capabilities in natural language understanding, reasoning, and generation. More recently, these models have evolved beyond unimodal text processing into *Multimodal Large Vision–Language Models* (LVLMs), which jointly reason over visual, linguistic, and increasingly other sensory modalities such as audio, depth, tactile signals, and proprioception [1]. This multimodal shift marks a critical inflection point for Human–Computer Interaction (HCI) and robotic applications, where intelligence must be grounded in perception, action, and interaction within complex real-world environments [2]. Traditional human–computer interfaces have relied on rigid input modalities such as keyboards, mice, and touchscreens, often requiring users to adapt to system constraints [3]. In parallel, classical robotic systems have typically employed modular pipelines, separating perception, planning, and control into independently optimized components [4]. While effective in constrained domains, these paradigms struggle to scale to open-ended, ambiguous, and dynamic settings that characterize natural human interaction and real-world robotics. LVLMs promise a unifying representational and computational framework that can bridge these gaps by learning holistic mappings between perception, language, and action, enabling systems that are more intuitive, adaptive, and collaborative. At their core, multimodal large vision–language models integrate visual encoders (e.g., convolutional neural networks or vision transformers) with large language backbones pretrained on massive text corpora [5]. Through alignment objectives, cross-modal attention mechanisms, and instruction tuning, these models learn shared semantic spaces in which images, videos, and text mutually inform one another. Recent generations of LVLMs have demonstrated remarkable abilities such as detailed image captioning, visual question answering, referring expression comprehension, diagram understanding, and grounded reasoning over complex scenes [6]. These capabilities are directly relevant to HCI scenarios, where systems must interpret user intent from multimodal cues, and to robotics, where perception and language must be tightly coupled to physical embodiment and action. In the context of Human–Computer Interaction, LVLMs represent a paradigm shift from interface-centric design toward *interaction-centric intelligence* [7]. By supporting natural language dialogue grounded in visual context, these models enable conversational interfaces that can explain, justify, and adapt their behavior in ways that align with human mental models. For example, an LVLM-powered system can interpret a user's spoken or typed instruction while simultaneously attending to on-screen content, physical objects, or augmented reality overlays [8]. This capability opens new avenues for accessible computing, assistive technologies, creative tools, and collaborative design systems, where users can interact with machines through fluid combinations of speech, gestures, sketches, and visual references rather than fixed command structures [9]. Robotic applications further amplify the importance of multimodal understanding. Unlike purely digital systems, robots operate in the physical world, where perception is noisy, partial, and temporally evolving [10]. Successful human–robot interaction requires shared situational awareness, grounded communication, and the ability to translate high-level language instructions into low-level motor actions [11]. LVLMs provide a promising substrate for such grounding by connecting visual observations (e.g., camera images, depth maps) with linguistic concepts (e.g., object names, spatial relations, task descriptions) [12]. Recent work has shown that vision–language models can support tasks such as object manipulation, navigation, task planning, and learning from human demonstrations, particularly when combined with reinforcement learning or classical control methods. A key advantage of LVLMs in both HCI and robotics lies in their capacity for *generalization*. Pretrained on diverse web-scale datasets, these models encode broad world knowledge that extends beyond any single task or environment [13]. This enables zero-shot and few-shot adaptation, where systems can perform new tasks or respond to

 doi:10.20944/preprints202512.2407.v1

novel instructions with minimal additional training [14]. In HCI, this translates into interfaces that are more flexible and personalized, capable of understanding user intent even when expressed in unconventional ways. In robotics, it enables more robust deployment across varied environments, reducing the need for extensive task-specific data collection and engineering. Despite their promise, the adoption of multimodal LVLMs in HCI and robotics also introduces significant challenges [15]. From a technical perspective, issues such as hallucination, grounding errors, data bias, and limited temporal reasoning remain active research problems [16]. From an interaction standpoint, the opacity of large models raises concerns about transparency, trust, and user control. In robotics, safety, real-time constraints, and sim-to-real transfer pose additional hurdles that are not fully addressed by current LVLM architectures. Furthermore, ethical and societal considerations—including privacy, accessibility, labor impact, and environmental cost—become increasingly salient as these models are integrated into everyday interactive systems and autonomous agents. Given the rapid pace of progress and the breadth of emerging applications, a comprehensive review of multimodal large vision–language models in the context of HCI and robotics is both timely and necessary. While existing surveys often focus on architectural innovations or benchmark performance, fewer works systematically examine how these models reshape interaction paradigms, enable new forms of human–machine collaboration, and alter the design space of intelligent robotic systems. This review aims to fill that gap by synthesizing developments across machine learning, human–computer interaction, and robotics, highlighting both shared foundations and domain-specific considerations [17]. Specifically, this review seeks to (i) trace the evolution of vision–language models toward large-scale multimodal systems, (ii) analyze their core architectural and training principles with an emphasis on interaction and embodiment, (iii) survey representative applications in HCI and robotics, and (iv) identify open challenges and future research directions at the intersection of multimodal intelligence, human-centered design, and autonomous systems. By framing LVLMs not merely as technical artifacts but as interactive agents embedded in human and physical contexts, this review provides a holistic perspective on their current capabilities and long-term potential. In summary, multimodal large vision–language models are redefining how machines perceive, reason, and interact. Their impact on human–computer interaction and robotic applications extends beyond performance gains, challenging long-standing assumptions about interfaces, autonomy, and collaboration [18]. Understanding this transformation requires an interdisciplinary lens that accounts for learning algorithms, system integration, human factors, and societal implications [19]. This introduction sets the stage for such an examination, positioning LVLMs as a central technology in the next generation of intelligent interactive and robotic systems.

2. Foundations of Multimodal Large Vision–Language Models

Multimodal Large Vision–Language Models (LVLMs) are built upon a confluence of theoretical ideas and practical techniques from representation learning, probabilistic modeling, information theory, and deep neural network optimization. This section provides a detailed foundation of LVLMs, emphasizing their mathematical formulation, architectural components, and learning paradigms that enable joint reasoning over vision and language. We focus on abstractions that are particularly relevant to human–computer interaction and robotic systems, where grounding, alignment, and decision-making under uncertainty are central concerns.

2.1. Problem Formulation and Multimodal Representation

At a high level, an LVLM seeks to model the joint distribution over multiple modalities, most commonly visual inputs and natural language [20]. Let $\mathcal{X}^v$ denote the space of visual inputs (e.g., images or videos), and let $\mathcal{X}^l$ denote the space of linguistic inputs (e.g., tokenized text). Given a visual observation $x^v \in \mathcal{X}^v$ and a language sequence $x^l = (w_1, w_2, \ldots, w_T)$ with $w_t \in \mathcal{V}$ (a vocabulary of size $|\mathcal{V}|$), the objective of an LVLM is to learn a parametric model

$$p_\theta(x^l \mid x^v) \quad \text{or more generally} \quad p_\theta(x^v, x^l),$$

 doi:10.20944/preprints202512.2407.v1

where θ denotes the model parameters. In interactive and robotic settings, this formulation is often extended to include actions $a \in \mathcal{A}$ and possibly additional modalities such as audio or proprioception [21]. A more general joint distribution can be written as

$$p_\theta(x^v, x^l, a, s),$$

where s represents latent world states. Learning such distributions enables reasoning that connects perception (x^v), communication (x^l), and embodiment (a), which is essential for grounded human–robot interaction. To make learning tractable, LVLMs rely on high-dimensional continuous representations [22]. Visual inputs are mapped to embeddings $z^v \in \mathbb{R}^{d_v}$ via a vision encoder $f_v(\cdot)$, while language inputs are mapped to embeddings $z^l \in \mathbb{R}^{d_l}$ via a language encoder $f_l(\cdot)$. Formally,

$$z^v = f_v(x^v), \quad z^l = f_l(x^l).$$

A central design goal is to learn a shared or aligned latent space $\mathcal{Z}$ such that semantic similarity across modalities corresponds to geometric proximity [23]:

$$\|z^v - z^l\| \text{ small} \iff (x^v, x^l) \text{ semantically aligned.}$$

2.2. Vision Encoders

The vision component of LVLMs is responsible for transforming raw pixel observations into structured feature representations [24]. Early vision–language systems relied on convolutional neural networks (CNNs), but modern LVLMs predominantly employ Vision Transformers (ViTs) [25]. Given an image $x^v \in \mathbb{R}^{H \times W \times C}$, it is first partitioned into N patches, each flattened and linearly projected into an embedding [26]:

$$x^v \to \{p_1, p_2, \ldots, p_N\}, \quad p_i \in \mathbb{R}^d$$

. These patch embeddings are augmented with positional encodings and processed through a stack of self-attention layers:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V,$$

where Q, K, V are learned linear projections of the input embeddings. The output is a sequence of visual tokens that encode both local and global spatial relationships [27]. For robotics, visual encoders may also incorporate temporal structure when processing video streams. In such cases, the visual input is $x^v = \{x_t^v\}_{t=1}^T$, and the encoder learns representations that capture spatiotemporal dynamics:

$$z^v = f_v(x_{1:T}^v),$$

which is crucial for action prediction, trajectory planning, and anticipation of human intent.

2.3. Language Models and Autoregressive Generation

The language backbone of an LVLM is typically a large autoregressive transformer trained to model the conditional distribution of tokens[28]:

$$p_\theta(x^l) = \prod_{t=1}^{T} p_\theta(w_t \mid w_{<t}).$$

Each token w_t is embedded into a vector $e_t \in \mathbb{R}^d$ and processed through multi-head self-attention and feed-forward layers [29]. The hidden state h_t at time t is used to parameterize a categorical distribution over the vocabulary:

$$p_\theta(w_t \mid w_{<t}) = \text{softmax}(Wh_t + b).$$

 doi:10.20944/preprints202512.2407.v1

In multimodal settings, the conditional distribution becomes

$$p_\theta(w_t \mid w_{<t}, z^v),$$

where visual embeddings z^v are injected into the language model via cross-attention, prefix tuning, or token concatenation. This conditioning allows generated language to be grounded in perceptual input, enabling tasks such as visual question answering and instruction following.

2.4. Cross-Modal Alignment and Fusion

A defining characteristic of LVLMs is the mechanism by which visual and linguistic representations are aligned [30]. One common approach is contrastive learning, where aligned image–text pairs (x^v, x^l) are encouraged to have high similarity in the embedding space, while mismatched pairs are pushed apart [31]. Given a batch of B paired samples, the contrastive loss can be written as

$$\mathcal{L}_{\text{CL}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(z_i^v, z_i^l)/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(z_i^v, z_j^l)/\tau)},$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter. Beyond contrastive objectives, modern LVLMs employ deep fusion strategies. Cross-attention layers allow language tokens to attend to visual tokens:

$$\text{Attention}(Q_l, K_v, V_v),$$

where queries Q_l come from language embeddings and keys/values (K_v, V_v) come from visual embeddings. This design enables fine-grained grounding, such as resolving referring expressions or reasoning about spatial relationships [32].

2.5. Training Objectives and Optimization

The overall training objective of an LVLM is typically a weighted sum of multiple losses:

$$\mathcal{L} = \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_{\text{CL}}\mathcal{L}_{\text{CL}} + \lambda_{\text{IT}}\mathcal{L}_{\text{IT}},$$

where $\mathcal{L}_{\text{LM}}$ is the language modeling loss, $\mathcal{L}_{\text{CL}}$ is a contrastive alignment loss, and $\mathcal{L}_{\text{IT}}$ denotes instruction-tuning or supervised grounding losses [33]. The coefficients λ_i balance competing objectives. Optimization is performed via stochastic gradient descent variants, typically Adam or AdamW, over extremely large datasets:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(x^v, x^l)\sim\mathcal{D}}[\mathcal{L}(\theta)].$$

From an HCI perspective, the scale of training has direct implications for usability and bias, while in robotics it raises concerns about data efficiency and transferability to real-world environments.

2.6. Grounding, Action, and Decision-Making

For robotic applications, LVLMs are often extended to incorporate actions and control. Let a_t denote an action at time t [34]. A grounded policy can be modeled as

$$\pi_\theta(a_t \mid x_t^v, x_t^l) = p_\theta(a_t \mid z_t^v, z_t^l),$$

where the embeddings are produced by the LVLM. This formulation bridges perception and language with control, allowing robots to follow natural language instructions such as "pick up the red cup on the table."

In HCI, grounding plays a different but related role: the system must map user utterances to interface states or operations. Mathematically, this can be viewed as inferring a latent intent variable y:

$$p(y \mid x^l, x^v),$$

 doi:10.20944/preprints202512.2407.v1

which then determines system behavior. LVLMs provide a powerful approximation to this inference by leveraging shared multimodal representations.

2.7. Discussion

The mathematical and architectural foundations of multimodal LVLMs reveal why they are particularly well-suited to interactive and embodied intelligence. By jointly modeling vision, language, and potentially action within a unified optimization framework, these models blur the boundaries between perception, cognition, and interaction [35]. However, this unification also introduces challenges related to scalability, interpretability, and control, which become especially pronounced in safety-critical robotic systems and user-facing HCI applications [36]. Understanding these foundations is therefore essential for both advancing LVLM research and responsibly deploying such models in real-world interactive settings.

3. Architectural Paradigms and System Design of LVLMs

The architectural design of Multimodal Large Vision–Language Models (LVLMs) reflects an ongoing convergence of ideas from deep learning, cognitive science, and systems engineering. Unlike earlier multimodal systems that relied on loosely coupled modules, contemporary LVLMs emphasize end-to-end trainable architectures that integrate perception and language reasoning at scale. This shift has profound implications for both human–computer interaction (HCI) and robotics, as architectural choices directly influence grounding fidelity, interaction latency, explainability, and robustness in real-world deployments [37,38]. In this section, we examine dominant architectural paradigms, discuss their trade-offs, and analyze how these designs support interactive and embodied intelligence. One of the most influential paradigms is the *dual-encoder architecture*, in which vision and language are processed by separate encoders whose outputs are aligned in a shared embedding space. This design emphasizes scalability and retrieval efficiency, making it particularly attractive for HCI applications such as image search, content recommendation, and multimodal information retrieval [39]. Dual-encoder LVLMs benefit from modularity: vision encoders can be pretrained on large image datasets, while language encoders leverage extensive text corpora. However, because interaction between modalities is limited to embedding-level similarity, such architectures often struggle with fine-grained reasoning, compositional queries, and multi-step interaction, all of which are critical in dialog-driven interfaces and robotic task execution. To address these limitations, *fusion-based architectures* have emerged as a dominant alternative. In these models, visual and linguistic representations are fused at intermediate or late stages through cross-attention mechanisms. Early fusion integrates raw or low-level features, enabling deep entanglement of modalities but at significant computational cost. Late fusion preserves modality-specific processing longer, improving efficiency while still enabling contextual grounding [40]. From an HCI perspective, fusion-based LVLMs enable richer interaction patterns, such as resolving ambiguous references in a graphical user interface or explaining system actions using visual evidence. In robotics, fusion allows the system to reason jointly about spatial layouts, object affordances, and linguistic constraints, which is essential for tasks such as manipulation in cluttered environments. A more recent and increasingly prevalent paradigm treats visual inputs as *tokens* within a language model [41]. In this design, outputs from a vision encoder are projected into the same embedding space as text tokens and concatenated with the language sequence. The entire multimodal input is then processed by a single large transformer. This unifying architecture simplifies system design and leverages the strong reasoning capabilities of large language models [42]. For interactive systems, this paradigm enables a seamless conversational interface in which visual context is treated as part of the dialogue state. For robots, it supports instruction following and reasoning by allowing perceptual observations to directly condition long-horizon planning expressed in natural language [43]. However, this approach raises challenges related to sequence length, real-time responsiveness, and the interpretability of internal representations. Beyond static architectures, system-level considerations play a crucial role in practical deployment [44]. LVLMs used in HCI often operate in loop with users, requiring low latency, incremental updates, and graceful handling of ambiguity

 doi:10.20944/preprints202512.2407.v1

or error [45]. This necessitates architectural support for streaming inputs, partial observability, and interactive clarification. In robotics, system design must additionally account for sensor noise, actuation delays, and safety constraints [46]. As a result, LVLMs are frequently embedded within hybrid systems, where high-level reasoning is performed by the LVLM while low-level perception and control are handled by specialized modules [47]. The architecture thus becomes not only a neural network design problem but also a question of how to orchestrate components within a larger interactive system. Figure 1 illustrates a simplified yet representative architectural view of an LVLM deployed in HCI and robotic contexts. The figure highlights the flow of information from visual perception and language input through multimodal fusion to downstream interaction or action modules [48]. While abstract, this representation captures a key insight: LVLMs function as a semantic bridge between human communication and machine perception, mediating understanding across fundamentally different representational domains.

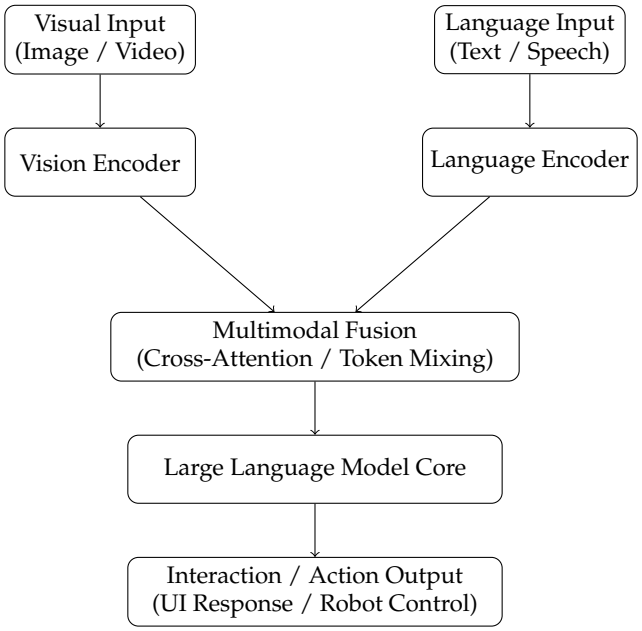

Figure 1. A high-level architectural illustration of a Multimodal Large Vision–Language Model (LVLM) for human–computer interaction and robotic applications [49]. Visual and language inputs are encoded separately, fused through multimodal mechanisms, and processed by a large language model core to produce interactive responses or physical actions.

From a broader perspective, architectural paradigms in LVLMs reflect deeper assumptions about intelligence and interaction. Dual-encoder models emphasize semantic similarity and retrieval, fusion models prioritize grounded reasoning, and token-unified models pursue architectural simplicity and generality [50]. In practice, many state-of-the-art systems combine elements of all three, adapting architectural choices to the constraints of specific applications [51]. As LVLMs continue to scale and migrate into everyday interactive systems and autonomous robots, architectural decisions will increasingly shape not only technical performance but also user experience, trust, and societal impact [52].

4. Applications in Human–Computer Interaction and Robotics

Multimodal Large Vision–Language Models (LVLMs) have rapidly transitioned from research prototypes to enabling technologies that fundamentally reshape applications in both Human–Computer Interaction (HCI) and robotics. Their ability to jointly interpret visual context and natural language allows systems to move beyond predefined interaction scripts toward more fluid, adaptive, and human-centered behaviors [53]. In this section, we examine major application domains, emphasizing how LVLM capabilities translate into practical functionality, how interaction paradigms evolve as a

 doi:10.20944/preprints202512.2407.v1

result, and what new challenges emerge when these models are embedded in real-world systems. In HCI, one of the most immediate and impactful applications of LVLMs is the development of *multimodal conversational interfaces* [54]. Traditional conversational agents typically operate in a text-only or speech-only regime, limiting their ability to reason about on-screen content or physical surroundings. LVLM-based interfaces, in contrast, can condition dialogue on visual state, enabling interactions such as "Why is this chart showing a spike here?" or "Move the icon next to the red button." This form of visually grounded conversation significantly lowers the cognitive burden on users by aligning system responses with shared perceptual context. As a result, LVLMs are increasingly explored in domains such as data analysis tools, design software, educational platforms, and accessibility technologies for users with visual or motor impairments. Another important HCI application lies in *intelligent user assistance and explanation* [55]. LVLMs can generate natural language explanations that reference specific visual elements, supporting transparency and trust in complex systems. For example, an LVLM integrated into a medical imaging interface can highlight salient regions of an image while explaining diagnostic suggestions, or a software assistant can guide users through complex workflows by visually indicating relevant interface components. This capability aligns closely with long-standing goals in HCI related to explainable and accountable systems, as it allows users to interrogate system behavior using the same multimodal cues they rely on in human–human communication. In creative and collaborative settings, LVLMs enable new forms of *co-creation* between humans and machines [56]. Designers can sketch rough visuals and describe desired modifications in natural language, while the system iteratively refines the output. Similarly, in content creation and game design, LVLMs can reason about scenes, characters, and layouts based on both textual descriptions and visual references [57]. These applications blur the boundary between tool and collaborator, raising important questions about authorship, control, and the role of AI in creative practice—central themes in contemporary HCI research. Robotic applications of LVLMs build on similar multimodal foundations but operate under stricter physical and safety constraints. A central use case is *language-guided manipulation*, where robots interpret high-level natural language instructions grounded in visual perception [58]. Tasks such as "pick up the mug on the left of the laptop" require resolving object identities, spatial relations, and task intent, all of which are naturally expressed in language but must be grounded in visual observations and translated into motor commands. LVLMs provide a powerful abstraction layer that connects these domains, enabling robots to generalize across tasks and environments without exhaustive task-specific programming. Another major robotic application domain is *navigation and exploration*. Robots equipped with LVLMs can follow language-based navigation instructions while reasoning about visual landmarks, obstacles, and goals [59]. For example, instructions like "go past the door and stop near the window" require semantic understanding of visual scenes and spatial reasoning over time [60]. By leveraging pretrained multimodal knowledge, LVLMs can interpret such instructions even in novel environments, significantly improving robustness and adaptability. This capability is particularly relevant for service robots, search-and-rescue operations, and human–robot collaboration in shared spaces [15]. Human–robot interaction (HRI) represents a natural intersection of HCI and robotics, and LVLMs play a pivotal role in enabling more natural and effective communication. In HRI scenarios, robots must not only execute tasks but also communicate their intentions, uncertainties, and limitations to human partners [61]. LVLMs support bidirectional interaction: they can interpret multimodal human input and generate explanations or clarifying questions grounded in the shared environment. This interactive loop is critical for building trust, ensuring safety, and supporting long-term collaboration between humans and robots. Table 1 summarizes representative application domains of LVLMs in HCI and robotics, highlighting key capabilities and associated challenges. While the table is necessarily high-level, it underscores the breadth of impact these models have across interactive and embodied systems [62].

 doi:10.20944/preprints202512.2407.v1

Table 1. Representative application domains of Multimodal Large Vision–Language Models (LVLMs) in Human–Computer Interaction and robotics, highlighting core capabilities and open challenges.

Application Domain	Key LVLM Capabilities	Primary Challenges
Multimodal Conversational Interfaces	Visually grounded dialogue, context-aware responses, intent inference	Latency, hallucination, user trust, evaluation of interaction quality
Intelligent User Assistance	Visual referencing, explanation generation, workflow guidance	Transparency, bias in explanations, integration with legacy systems
Creative and Collaborative Tools	Joint reasoning over sketches and text, iterative co-creation	Control, authorship, intellectual property, usability
Language-Guided Manipulation	Object grounding, spatial reasoning, task generalization	Safety, perception errors, sim-to-real transfer
Navigation and Exploration	Landmark-based reasoning, instruction following, environment generalization	Temporal consistency, localization errors, real-time constraints
Human–Robot Interaction	Multimodal communication, intent explanation, adaptive behavior	Social appropriateness, trust calibration, ethical deployment

Across both HCI and robotics, a recurring theme is the tension between generality and reliability. LVLMs excel at flexible, open-ended reasoning, yet their probabilistic nature introduces uncertainty that must be carefully managed in interactive and embodied settings [63]. As these models become more deeply integrated into user-facing interfaces and autonomous systems, application design must balance expressiveness with safeguards that ensure predictable and safe behavior. Understanding how LVLMs function across diverse application domains is therefore essential not only for advancing technical performance but also for shaping the future of human-centered and responsible AI systems.

5. Challenges, Limitations, and Open Research Problems

Despite the remarkable progress and expanding application landscape of Multimodal Large Vision–Language Models (LVLMs), their deployment in human–computer interaction and robotic systems remains constrained by a range of fundamental challenges [39]. These challenges arise not only from technical limitations in model architectures and training procedures but also from deeper issues related to grounding, interaction dynamics, safety, and societal impact [64]. Understanding these limitations is essential for contextualizing current achievements and for guiding future research toward more reliable, transparent, and human-centered multimodal intelligence. A central and widely discussed challenge is the problem of *hallucination* and weak grounding. LVLMs may generate linguistically coherent and confident responses that are not supported by the underlying visual input. In HCI settings, such hallucinations can mislead users, undermine trust, and result in incorrect decisions, particularly in high-stakes domains such as healthcare, education, or data analysis. In robotics, grounding errors can have more severe consequences, potentially leading to unsafe actions or task failure [65]. These issues stem in part from the fact that LVLMs are trained to optimize likelihood over large datasets, where linguistic fluency may dominate over strict perceptual correctness [66]. Addressing hallucination requires advances in training objectives, multimodal evaluation metrics, and explicit mechanisms for uncertainty estimation and abstention [67]. Another significant limitation lies in *temporal and causal reasoning* [68]. While LVLMs can process static images and short video clips effectively, many interactive and robotic tasks require long-horizon reasoning over sequences of observations and actions [69]. For example, understanding how a scene evolves over time or predicting the consequences of a sequence of actions remains challenging for current models. In HCI, this manifests as difficulty maintaining coherent dialogue state across extended interactions, especially when visual context changes. In robotics, insufficient temporal reasoning hampers planning, recovery from errors, and adaptation to dynamic environments. These limitations point to the need for

 doi:10.20944/preprints202512.2407.v1

architectures that explicitly model time, memory, and causality rather than relying solely on implicit representations learned from static datasets. Scalability and efficiency present another major obstacle. State-of-the-art LVLMs often contain billions of parameters and require substantial computational resources for both training and inference. In interactive applications, high latency can disrupt user experience and reduce the sense of responsiveness that is critical for effective HCI [70]. In robotics, computational constraints are even more severe, as models must often run on embedded hardware with strict energy and real-time requirements [71]. Techniques such as model compression, distillation, and modularization offer partial solutions, but they often come at the cost of reduced generality or performance. Balancing scale with deployability remains an open research problem with direct implications for the accessibility and sustainability of LVLM-based systems [72]. The *opacity and interpretability* of LVLMs further complicate their use in interactive and embodied contexts. While these models can generate explanations in natural language, the extent to which such explanations reflect true internal reasoning versus post hoc rationalization is unclear. In HCI, this raises questions about transparency and accountability: users may overtrust systems that appear articulate and confident without understanding their limitations [73]. In robotics, lack of interpretability makes debugging and verification difficult, particularly in safety-critical scenarios [74]. Developing methods for faithful explanation, introspection, and human-understandable model diagnostics is therefore a key challenge for the field [75]. Data-related issues also pose significant constraints. LVLMs are typically trained on large-scale web data that may contain biases, inaccuracies, and culturally specific assumptions. When such models are deployed in HCI applications, these biases can manifest as exclusionary or inappropriate behavior, disproportionately affecting marginalized user groups. In robotics, dataset bias can limit generalization to diverse physical environments and human behaviors. Moreover, collecting high-quality multimodal data for embodied tasks is expensive and time-consuming, further exacerbating disparities between simulated benchmarks and real-world performance [76]. Addressing these issues requires not only technical solutions but also careful dataset curation, participatory design, and interdisciplinary collaboration [77]. Safety and ethical considerations represent an overarching set of challenges that cut across all applications of LVLMs [78]. In HCI, concerns include privacy, informed consent, and the potential for manipulation through persuasive or deceptive multimodal interfaces [79]. In robotics, safety is paramount, as failures can result in physical harm. LVLMs must therefore be integrated with robust safety mechanisms, including constraint enforcement, human-in-the-loop oversight, and formal verification where possible. At a broader level, the increasing autonomy and communicative ability of LVLM-powered systems raise questions about responsibility, regulation, and the long-term societal impact of delegating perception and decision-making to machines. Finally, there remains a substantial gap between current research benchmarks and real-world interactive performance. Many evaluations of LVLMs focus on static datasets and short-form tasks that do not capture the complexity of sustained interaction, social context, or embodied action [80]. In HCI, this limits our understanding of how users adapt to and appropriate these systems over time [81]. In robotics, it obscures challenges related to wear, uncertainty, and long-term autonomy. Closing this gap requires new evaluation paradigms that account for interaction dynamics, user experience, and longitudinal performance in realistic settings. In summary, while Multimodal Large Vision–Language Models offer a powerful and unifying framework for perception, language, and interaction, their current limitations highlight the need for continued research across multiple dimensions [82]. Addressing challenges in grounding, temporal reasoning, efficiency, interpretability, data quality, and safety is essential for realizing the full potential of LVLMs in human–computer interaction and robotics. These open problems not only define the technical frontier of the field but also shape its ethical and societal trajectory, underscoring the importance of a holistic and human-centered approach to future development [83].

 doi:10.20944/preprints202512.2407.v1

6. Future Directions and Research Opportunities

Looking forward, Multimodal Large Vision–Language Models (LVLMs) occupy a pivotal position in the evolution of intelligent interactive and robotic systems. While current models already demonstrate impressive multimodal reasoning capabilities, their limitations point to a rich landscape of future research opportunities that span algorithmic innovation, system integration, human-centered design, and societal governance [84]. This section outlines several promising directions that are likely to shape the next generation of LVLMs and determine their long-term impact on human–computer interaction and robotics [85]. One important research direction concerns the development of *stronger and more explicit grounding mechanisms* [86]. Future LVLMs are expected to move beyond implicit alignment learned from large-scale data toward architectures that incorporate structured representations of objects, relations, and affordances. Integrating symbolic or neuro-symbolic components with continuous multimodal embeddings may enable models to reason more reliably about spatial layouts, physical constraints, and task semantics [87]. In HCI, such grounding could support interfaces that are more predictable and controllable, enabling users to reference complex visual structures with precision. In robotics, explicit grounding is critical for safety and robustness, as it allows systems to verify that planned actions are consistent with physical reality before execution [88]. Another promising avenue lies in *temporal, memory, and causal modeling*. Future LVLMs will need to reason over extended interaction histories, track evolving world states, and understand cause–effect relationships across time. This capability is essential for sustained dialogue, long-term collaboration, and autonomous robotic operation in dynamic environments [89]. Research into external memory modules, recurrent multimodal transformers, and world-model learning offers potential solutions. By incorporating mechanisms for remembering past interactions and predicting future outcomes, LVLMs could transition from reactive responders to proactive collaborators, capable of anticipating user needs and adapting strategies over time. Efficiency and sustainability will also play a central role in shaping future LVLM research. As models continue to scale, there is growing recognition that brute-force parameter growth is neither economically nor environmentally sustainable. Future work is likely to focus on more efficient architectures, adaptive computation, and sparsity-aware training methods that reduce resource consumption without sacrificing performance [90]. For HCI applications, this could enable responsive multimodal assistants on personal devices rather than relying exclusively on cloud-based inference [53]. In robotics, improved efficiency is essential for deploying LVLMs on edge hardware, enabling real-time perception and decision-making in resource-constrained settings [91]. From a human-centered perspective, future LVLMs must be designed not only to perform tasks but also to *support meaningful and ethical interaction*. This includes developing models that can express uncertainty, ask clarifying questions, and adapt their communication style to individual users [92]. In HCI, such capabilities align with principles of usability, accessibility, and inclusive design. In robotics, they support safer and more transparent human–robot collaboration [93]. Research into affect-aware modeling, user-adaptive interaction, and culturally sensitive multimodal communication will be essential for ensuring that LVLM-powered systems are broadly usable and socially acceptable [94]. Evaluation methodologies represent another critical frontier. Current benchmarks often fail to capture the complexity of real-world interaction and embodied action [95]. Future research must develop evaluation frameworks that measure not only task success but also interaction quality, user satisfaction, learning outcomes, and long-term system behavior [96]. For robotics, this includes assessing robustness under uncertainty, recovery from failure, and safety in human-populated environments. For HCI, it requires longitudinal studies that examine how users learn, trust, and adapt to LVLM-based systems over time. Such evaluations will provide a more holistic understanding of system performance and guide more responsible deployment [97]. Finally, governance, policy, and interdisciplinary collaboration will increasingly influence the trajectory of LVLM research [98]. As these models become embedded in everyday tools and autonomous systems, questions of accountability, regulation, and societal impact will move to the forefront. Researchers and practitioners must engage with stakeholders from law, ethics, design, and the social sciences to ensure that LVLMs are developed and deployed in ways that align with human values [43]. This includes

 doi:10.20944/preprints202512.2407.v1

addressing issues of data privacy, intellectual property, labor displacement, and environmental impact. Proactively engaging with these concerns will help shape a future in which multimodal AI technologies enhance human capabilities rather than undermine them. In conclusion, the future of Multimodal Large Vision–Language Models is defined not only by technical innovation but also by their integration into complex human and physical ecosystems. Advancing this field will require a holistic approach that combines algorithmic rigor with human-centered design and ethical foresight [99]. By addressing foundational challenges and exploring new research directions, the community can move toward LVLMs that are not only more powerful and general but also more trustworthy, efficient, and aligned with the needs of society [100].

7. Conclusions

Multimodal Large Vision–Language Models (LVLMs) represent a fundamental shift in how artificial intelligence systems perceive, reason, and interact with the world. By unifying visual perception and natural language understanding within a single large-scale modeling framework, LVLMs challenge long-standing separations between sensing, cognition, and communication that have historically structured both human–computer interaction and robotic system design [101]. Throughout this review, we have examined the conceptual foundations, architectural paradigms, application domains, and open challenges associated with these models, highlighting their transformative potential as well as the substantial work that remains to be done. From an HCI perspective, LVLMs redefine the nature of interaction itself. Rather than forcing users to adapt to rigid interfaces or predefined command languages, LVLM-powered systems move closer to interaction patterns that resemble human–human communication: grounded in shared visual context, flexible in expression, and adaptive to user intent [102]. This shift has profound implications for usability, accessibility, and creativity. Systems that can see what users see and talk about it meaningfully enable more intuitive workflows, reduce cognitive load, and open computing to broader populations [103]. At the same time, these benefits introduce new responsibilities for designers and researchers to ensure transparency, trustworthiness, and inclusivity in multimodal interfaces [104]. In robotics, the impact of LVLMs is equally significant but operates under different constraints and expectations. Robots endowed with multimodal vision–language reasoning gain a powerful abstraction layer that bridges high-level human intent and low-level physical action [105]. This capability enables generalization across tasks and environments, supports natural language instruction following, and facilitates richer human–robot collaboration [41]. However, the embodied nature of robotics also exposes the limitations of current LVLMs, particularly with respect to grounding, safety, real-time performance, and long-horizon reasoning. These challenges underscore the need for hybrid approaches that combine the flexibility of large multimodal models with the reliability of structured control and verification mechanisms. A recurring theme across all sections of this review is the tension between generality and control. LVLMs excel at open-ended reasoning and adaptation precisely because they are trained on vast, diverse datasets and operate probabilistically. Yet these same properties make their behavior difficult to predict and verify, especially in safety-critical or user-facing contexts [106]. Addressing this tension will be one of the defining challenges of future research [107]. Progress is likely to come not from a single breakthrough but from sustained advances across multiple fronts, including model architecture, training objectives, evaluation methodologies, and system-level integration [108].

Equally important are the broader societal and ethical dimensions of LVLM deployment. As these models become embedded in everyday tools and autonomous systems, their influence on human behavior, decision-making, and social structures will grow. Issues of bias, privacy, accountability, and environmental sustainability cannot be treated as secondary concerns; they are integral to the responsible development of multimodal AI. The interdisciplinary nature of HCI and robotics places these fields in a unique position to shape not only what LVLMs can do, but how and why they are used.

 doi:10.20944/preprints202512.2407.v1

In closing, Multimodal Large Vision–Language Models should be understood not merely as a new class of machine learning models, but as a foundational technology for the next generation of interactive and embodied systems. Their success will depend on our ability to align technical innovation with human values, to balance flexibility with reliability, and to design systems that augment rather than replace human capabilities. By situating LVLM research at the intersection of machine learning, human–computer interaction, and robotics, this review aims to provide a holistic perspective that can guide future work toward more intelligent, interactive, and humane AI systems.

References

1. Peng, B.; Li, C.; He, P.; Galley, M.; Gao, J. Instruction tuning with gpt-4. *arXiv:2304.03277* **2023**.
2. Overbay, K.; Ahn, J.; Park, J.; Kim, G.; et al. mRedditSum: A Multimodal Abstractive Summarization Dataset of Reddit Threads with Images. In Proceedings of the The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
3. Sun, L.; Liang, H.; Wei, J.; Sun, L.; Yu, B.; Cui, B.; Zhang, W. Efficient-Empathy: Towards Efficient and Effective Selection of Empathy Data. *arXiv preprint arXiv:2407.01937* **2024**.
4. Feng, J.; Sun, Q.; Xu, C.; Zhao, P.; Yang, Y.; Tao, C.; Zhao, D.; Lin, Q. MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 7348–7363.
5. Du, Y.; Guo, H.; Zhou, K.; Zhao, W.X.; Wang, J.; Wang, C.; Cai, M.; Song, R.; Wen, J.R. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv:2311.01487* **2023**.
6. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13. Springer, 2003, pp. 363–370.
7. Broder, A.Z. On the resemblance and containment of documents. In Proceedings of the Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171). IEEE, 1997, pp. 21–29.
8. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *NeurIPS* **2017**.
9. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 2019, pp. 3195–3204.
10. Jing, L.; Li, R.; Chen, Y.; Jia, M.; Du, X. FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models. *arXiv:2311.01477* **2023**.
11. Kothawade, S.; Beck, N.; Killamsetty, K.; Iyer, R. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems* **2021**, *34*, 18685–18697.
12. Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems* **2022**, *35*, 26418–26431.
13. Mangalam, K.; Akshulakov, R.; Malik, J. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* **2024**, *36*.
14. Wei, H.; Kong, L.; Chen, J.; Zhao, L.; Ge, Z.; Yang, J.; Sun, J.; Han, C.; Zhang, X. Vary: Scaling up the vision vocabulary for large vision-language models. *arXiv preprint arXiv:2312.06109* **2023**.
15. Grave, É.; Bojanowski, P.; Gupta, P.; Joulin, A.; Mikolov, T. Learning Word Vectors for 157 Languages. In Proceedings of the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.
16. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the CVPR, 2022.
17. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **2021**, *65*, 99–106.
18. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* **2023**.

 doi:10.20944/preprints202512.2407.v1

19. Wang, Z.; Zhang, Q.; Ding, K.; Qin, M.; Zhuang, X.; Li, X.; Chen, H. InstructProtein: Aligning Human and Protein Language via Knowledge Instruction. *arXiv preprint arXiv:2310.03269* **2023**.
20. Lai, Z.; Zhang, H.; Wu, W.; Bai, H.; Timofeev, A.; Du, X.; Gan, Z.; Shan, J.; Chuah, C.N.; Yang, Y.; et al. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699* **2023**.
21. Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J.M.; Parikh, D.; Batra, D. Visual dialog. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 326–335.
22. Gao, J.; Lin, C.Y. Introduction to the special issue on statistical language modeling, 2004.
23. Jiang, C.; Xu, H.; Dong, M.; Chen, J.; Ye, W.; Yan, M.; Ye, Q.; Zhang, J.; Huang, F.; Zhang, S. Hallucination Augmented Contrastive Learning for Multimodal Large Language Model. *arXiv:2312.06968* **2023**.
24. Liu, S.; Wang, J.; Yang, Y.; Wang, C.; Liu, L.; Guo, H.; Xiao, C. ChatGPT-powered Conversational Drug Editing Using Retrieval and Domain Feedback. *arXiv preprint arXiv:2305.18090* **2023**.
25. Elazar, Y.; Bhagia, A.; Magnusson, I.; Ravichander, A.; Schwenk, D.; Suhr, A.; Walsh, P.; Groeneveld, D.; Soldaini, L.; Singh, S.; et al. What's In My Big Data? *arXiv preprint arXiv:2310.20707* **2023**.
26. Chen, H.; Xie, W.; Afouras, T.; Nagrani, A.; Vedaldi, A.; Zisserman, A. Localizing visual sounds the hard way. In Proceedings of the CVPR, 2021.
27. Xu, M.; Yoon, S.; Fuentes, A.; Park, D.S. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition* **2023**, *137*, 109347.
28. Liu, W.; Zeng, W.; He, K.; Jiang, Y.; He, J. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
29. Stewart, G.W. On the early history of the singular value decomposition. *SIAM review* **1993**, *35*, 551–566.
30. Himakunthala, V.; Ouyang, A.; Rose, D.; He, R.; Mei, A.; Lu, Y.; Sonar, C.; Saxon, M.; Wang, W.Y. Let's Think Frame by Frame: Evaluating Video Chain of Thought with Video Infilling and Prediction. *arXiv:2305.13903* **2023**.
31. Ben Abacha, A.; Demner-Fushman, D. A question-entailment approach to question answering. *BMC bioinformatics* **2019**, *20*, 1–23.
32. Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S.N.; Weller, A. Synthetic Data–what, why and how? *arXiv preprint arXiv:2205.03257* **2022**.
33. Anonymous. NL2ProGPT: Taming Large Language Model for Conversational Protein Design, 2024.
34. Mallya, A.; Wang, T.C.; Sapra, K.; Liu, M.Y. World-consistent video-to-video synthesis. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. Springer, 2020, pp. 359–378.
35. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* **2014**.
36. Lei, J.; Berg, T.L.; Bansal, M. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems* **2021**, *34*, 11846–11858.
37. Lei, J.; Yu, L.; Berg, T.L.; Bansal, M. Tvr: A large-scale dataset for video-subtitle moment retrieval. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. Springer, 2020, pp. 447–463.
38. Pham, V.T.; Le, T.L.; Tran, T.H.; Nguyen, T.P. Hand detection and segmentation using multimodal information from Kinect. In Proceedings of the 2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR). IEEE, 2020, pp. 1–6.
39. Hernandez, D.; Brown, T.; Conerly, T.; DasSarma, N.; Drain, D.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Henighan, T.; Hume, T.; et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487* **2022**.
40. Abbas, A.K.M.; Tirumala, K.; Simig, D.; Ganguli, S.; Morcos, A.S. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. In Proceedings of the ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models, 2023.
41. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* **2023**.
42. Mnih, V.; Heess, N.; Graves, A.; et al. Recurrent models of visual attention. *Advances in neural information processing systems* **2014**, *27*.

 doi:10.20944/preprints202512.2407.v1

43. Kazemzadeh, S.; Ordonez, V.; Matten, M.; Berg, T. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 787–798.
44. Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; Smola, A. Multimodal chain-of-thought reasoning in language models. *arXiv:2302.00923* **2023**.
45. Ge, J.; Luo, H.; Qian, S.; Gan, Y.; Fu, J.; Zhan, S. Chain of Thought Prompt Tuning in Vision Language Models. *arXiv:2304.07919* **2023**.
46. Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; et al. MobileVLM: A Fast, Reproducible and Strong Vision Language Assistant for Mobile Devices. *arXiv:2312.16886* **2023**.
47. Bran, A.M.; Schwaller, P. Transformers and Large Language Models for Chemistry and Drug Discovery. *arXiv preprint arXiv:2310.06083* **2023**.
48. Ono, K.; Morita, A. Evaluating large language models: Chatgpt-4, mistral 8x7b, and google gemini benchmarked against mmlu. *Authorea Preprints* **2024**.
49. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
50. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the ACL, 2018.
51. Hu, W.; Xu, Y.; Li, Y.; Li, W.; Chen, Z.; Tu, Z. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 2256–2264.
52. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. Pre-trained models: Past, present and future. *AI Open* **2021**, *2*, 225–250.
53. Li, J.; Liu, Y.; Fan, W.; Wei, X.Y.; Liu, H.; Tang, J.; Li, Q. Empowering Molecule Discovery for Molecule-Caption Translation with Large Language Models: A ChatGPT Perspective. *arXiv preprint arXiv:2306.06615* **2023**.
54. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. *arXiv:2310.03744* **2023**.
55. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. CogVLM: Visual Expert for Pretrained Language Models, 2024, [arXiv:cs.CV/2311.03079].
56. Xu, Z.; Feng, C.; Shao, R.; Ashby, T.; Shen, Y.; Jin, D.; Cheng, Y.; Wang, Q.; Huang, L. Vision-Flan: Scaling Human-Labeled Tasks in Visual Instruction Tuning. *arXiv:2402.11690* **2024**.
57. Moon, S.; Madotto, A.; Lin, Z.; Nagarajan, T.; Smith, M.; Jain, S.; Yeh, C.F.; Murugesan, P.; Heidari, P.; Liu, Y.; et al. Anymal: An efficient and scalable any-modality augmented language model. *arXiv:2309.16058* **2023**.
58. Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* **2021**.
59. Zhao, H.; Liu, S.; Ma, C.; Xu, H.; Fu, J.; Deng, Z.H.; Kong, L.; Liu, Q. GIMLET: A Unified Graph-Text Model for Instruction-Based Molecule Zero-Shot Learning. *bioRxiv* **2023**, pp. 2023–05.
60. Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005* **2023**.
61. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971* **2023**.
62. Chen, C.; Qin, R.; Luo, F.; Mi, X.; Li, P.; Sun, M.; Liu, Y. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437* **2023**.
63. Zhang, W.; Wang, X.; Nie, W.; Eaton, J.; Rees, B.; Gu, Q. MoleculeGPT: Instruction Following Large Language Models for Molecular Property Prediction. In Proceedings of the NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development, 2023.
64. Bran, A.M.; Cox, S.; White, A.D.; Schwaller, P. ChemCrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376* **2023**.
65. Singer, P.; Flöck, F.; Meinhart, C.; Zeitfogel, E.; Strohmaier, M. Evolution of reddit: from the front page of the internet to a self-referential community? In Proceedings of the Proceedings of the 23rd international conference on world wide web, 2014, pp. 517–522.
66. Lu, Y.; Bartolo, M.; Moore, A.; Riedel, S.; Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv:2104.08786* **2021**.
67. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv preprint arXiv:2401.16822* **2024**.

 doi:10.20944/preprints202512.2407.v1

68. Huang, J.; Zhang, J.; Jiang, K.; Qiu, H.; Lu, S. Visual Instruction Tuning towards General-Purpose Multimodal Model: A Survey. *arXiv preprint arXiv:2312.16602* **2023**.
69. Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; Wang, X. Generative multimodal models are in-context learners. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14398–14409.
70. Chen, Y.; Sikka, K.; Cogswell, M.; Ji, H.; Divakaran, A. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. *arXiv preprint arXiv:2311.10081* **2023**.
71. Luo, Y.; Zhang, J.; Fan, S.; Yang, K.; Wu, Y.; Qiao, M.; Nie, Z. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442* **2023**.
72. Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* **2021**.
73. Gong, T.; Lyu, C.; Zhang, S.; Wang, Y.; Zheng, M.; Zhao, Q.; Liu, K.; Zhang, W.; Luo, P.; Chen, K. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790* **2023**.
74. Xu, H.; Ghosh, G.; Huang, P.Y.; Arora, P.; Aminzadeh, M.; Feichtenhofer, C.; Metze, F.; Zettlemoyer, L. VLM: Task-agnostic Video-Language Model Pre-training for Video Understanding. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 4227–4239.
75. Barbieri, F.; Camacho-Collados, J.; Neves, L.; Espinosa-Anke, L. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, 2020, [arXiv:cs.CL/2010.12421].
76. Luo, G.; Zhou, Y.; Zhang, Y.; Zheng, X.; Sun, X.; Ji, R. Feast Your Eyes: Mixture-of-Resolution Adaptation for Multimodal Large Language Models. *arXiv preprint arXiv:2403.03003* **2024**.
77. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV* **2017**.
78. Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Bousquet, O.; Le, Q.; Chi, E. Least-to-most prompting enables complex reasoning in large language models. *arXiv:2205.10625* **2022**.
79. Huang, S.; Dong, L.; Wang, W.; Hao, Y.; Singhal, S.; Ma, S.; Lv, T.; Cui, L.; Mohammed, O.K.; Liu, Q.; et al. Language is not all you need: Aligning perception with language models. *arXiv:2302.14045* **2023**.
80. Wang, Y.; Xiao, J.; Suzek, T.O.; Zhang, J.; Wang, J.; Bryant, S.H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research* **2009**, *37*, W623–W633.
81. Liu, Z.; Li, S.; Luo, Y.; Fei, H.; Cao, Y.; Kawaguchi, K.; Wang, X.; Chua, T.S. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798* **2023**.
82. LeCun, Y. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review* **2022**, *62*.
83. Edwards, C.; Zhai, C.; Ji, H. Text2mol: Cross-modal molecule retrieval with natural language queries. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 595–607.
84. Xue, L.; Gao, M.; Xing, C.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J.C.; Savarese, S. ULIP: Learning Unified Representation of Language, Image and Point Cloud for 3D Understanding. *arXiv preprint arXiv:2212.05171* **2022**.
85. Yu, T.; Yao, Y.; Zhang, H.; He, T.; Han, Y.; Cui, G.; Hu, J.; Liu, Z.; Zheng, H.T.; Sun, M.; et al. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. *arXiv:2312.00849* **2023**.
86. Wang, K.; Babenko, B.; Belongie, S. End-to-end scene text recognition. In Proceedings of the 2011 International conference on computer vision. IEEE, 2011, pp. 1457–1464.
87. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4195–4205.
88. LeCun, Y.; Denker, J.; Solla, S. Optimal brain damage. *Advances in neural information processing systems* **1989**, *2*.
89. Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; Zhao, R. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv:2306.15195*.
90. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **2020**, *21*, 5485–5551.
91. Xu, P.; Shao, W.; Zhang, K.; Gao, P.; Liu, S.; Lei, M.; Meng, F.; Huang, S.; Qiao, Y.; Luo, P. LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models. *arXiv:2306.09265* **2023**.

 doi:10.20944/preprints202512.2407.v1

92. Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490* **2023**.
93. Rogers, V.; Meara, P.; Barnett-Legh, T.; Curry, C.; Davie, E. Examining the LLAMA aptitude tests. *Journal of the European Second Language Association* **2017**, *1*, 49–60.
94. Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; Yuan, L. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122* **2023**.
95. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* **2023**.
96. Wang, B.; Li, G.; Zhou, X.; Chen, Z.; Grossman, T.; Li, Y. Screen2words: Automatic mobile UI summarization with multimodal learning. In Proceedings of the The 34th Annual ACM Symposium on User Interface Software and Technology, 2021, pp. 498–510.
97. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; Xie, W. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *arXiv:2305.10415* **2023**.
98. Chen, G.; Zheng, Y.D.; Wang, J.; Xu, J.; Huang, Y.; Pan, J.; Wang, Y.; Wang, Y.; Qiao, Y.; Lu, T.; et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292* **2023**.
99. Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems* **2024**, *36*.
100. Hassibi, B.; Stork, D. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems* **1992**, *5*.
101. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *NeurIPS* **2022**.
102. Wu, S.; Lu, K.; Xu, B.; Lin, J.; Su, Q.; Zhou, C. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182* **2023**.
103. Liu, F.; Wang, X.; Yao, W.; Chen, J.; Song, K.; Cho, S.; Yacoob, Y.; Yu, D. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning. *arXiv preprint arXiv:2311.10774* **2023**.
104. Chen, D.; Chen, R.; Zhang, S.; Liu, Y.; Wang, Y.; Zhou, H.; Zhang, Q.; Zhou, P.; Wan, Y.; Sun, L. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. *arXiv preprint arXiv:2402.04788* **2024**.
105. Schaeffer, R.; Miranda, B.; Koyejo, S. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems* **2024**, *36*.
106. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* **2020**.
107. Zheng, J.; Zhang, J.; Li, J.; Tang, R.; Gao, S.; Zhou, Z. Structured3D: A Large Photo-realistic Dataset for Structured 3D Modeling. In Proceedings of the Proceedings of The European Conference on Computer Vision (ECCV), 2020.
108. Jiang, J.; Shu, Y.; Wang, J.; Long, M. Transferability in deep learning: A survey. *arXiv preprint arXiv:2201.05867* **2022**.