Beyond Scaling: Rethinking Architectures, Training, and Evaluation for Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have advanced rapidly, progressing from early vision–language alignment systems to increasingly general-purpose multimodal reasoning models. This position paper presents a systematic examination of the MLLM landscape, introduces a unified framework for model architectures, and critically analyzes prevailing training paradigms, evaluation protocols, and benchmark suites. Furthermore, we highlight the fundamental limitations of existing approaches and outline emerging research directions, proposing a roadmap for next-generation MLLMs that extends beyond scale.

1 Introduction

Large Language Models (LLMs) have driven significant advances in natural language processing, becoming a cornerstone of modern AI systems. However, many real-world tasks require reasoning expressed in language but grounded across multiple modalities, including vision, audio, and temporal signals. This recognition motivated early efforts to extend language models beyond text, leading to vision–language architectures such as ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019). These models introduced explicit mechanisms for cross-modal alignment between text and images, demonstrating the feasibility of joint multimodal representation learning. Although limited in scale and scope, they established the foundations for what are now referred to as Multimodal Large Language Models (MLLMs).

Recent years have witnessed rapid progress in MLLMs, with models evolving from task-specific vision–language systems into increasingly general-purpose multimodal foundation models. The scope of multimodality has expanded beyond images to include audio and video, enabled by large-scale pretraining (Radford et al., 2021; Jia et al., 2021a), unified transformer-based backbones (Lu et al., 2023), and instruction-tuned multimodal interfaces (Dai et al., 2023). As a result, contemporary MLLMs are capable of supporting a wide range of downstream tasks, from multimodal question answering to interactive and generative applications (Li et al., 2023a; Shen et al., 2023).

Despite these advances, research on MLLMs remains fragmented. Vision–Language Models (VLMs), Audio–Language Models (ALMs), and Video–Language Models (VideoLLMs) are often developed and evaluated in isolation, each with distinct architectural assumptions, training objectives, and benchmark suites. Although this modality-specific focus has facilitated rapid progress within individual domains, it complicates efforts to develop unified principles for multimodal modeling. As MLLMs increasingly integrate multiple modalities and are deployed in real-world settings, the absence of a modality-agnostic conceptual framework limits systematic understanding and comparison.

In addition, recent empirical advances have obscured several unresolved challenges. Current MLLMs exhibit limitations in temporal reasoning (Fu et al., 2025; Xiao et al., 2021), fine-grained cross-modal grounding (Rohrbach et al., 2018; Liu et al., 2024a), underperformance under distribution shifts due to shortcut learning (Geirhos et al., 2020), and the interpretability of multimodal decisions (Dang et al., 2024). Evaluation practices further exacerbate these issues: many benchmarks are task-driven and modality-specific, making them vulnerable to data leakage by inclusion in the training corpus, resulting in an incomplete picture of multimodal reasoning and generalization. These observations, along with the current trend of scaling models to induce capabilities, suggest that scaling alone is insufficient to address the persistent limitation in temporal reasoning, grounding, and generalization.

This paper takes the position that the advancement of MLLMs requires a shift from modality-

specific optimization toward a unified perspective on architecture, training, and evaluation. Rather than providing an exhaustive survey of existing models, we focus on identifying shared structural patterns, highlighting fundamental limitations, and articulating open research questions that shape the future of multimodal language modeling.

> **Position Statement.** *We argue that MLLMs should be conceptualized and evaluated as modality-agnostic reasoning systems rather than as collections of modality-paired extensions of language models. Progress toward truly general multimodal intelligence is likely to depend on unified architectural principles, training objectives that extend beyond alignment, and evaluation protocols that emphasize transferable multimodal capabilities rather than task-specific performance.*

2 A Unified Architectural Framework for MLLMs

MLLMs extend language-centric reasoning to heterogeneous modalities such as images, audio, and video. While recent models differ widely in implementation, we argue that they share a common computational structure: heterogeneous inputs are first normalized into a shared representational space, then interact through modality-aware mechanisms, and finally support language-driven reasoning and generation. This observation motivates a unifying architectural perspective that abstracts away modality-specific details while preserving the essential structure of multimodal computation.

We formalize this perspective through an *Alignment–Fusion–Reasoning (AFR)* framework. AFR is not intended as a descriptive survey taxonomy, but rather a unified abstraction for understanding how modern MLLMs are built, why they fail, and how design choices propagate.

unified abstraction for understanding how modern MLLMs are built, why they fail, and how design choices propagate.

2.1 Why Alignment–Fusion–Reasoning?

We argue that AFR captures the recurrent structural components that appear necessary across all MLLM architectures and provides a sufficient abstraction for analyzing design tradeoffs and failure modes.

Necessity. Any MLLM must (i) transform heterogeneous modality inputs into comparable latent representations (Alignment), (ii) enable cross-modal interaction (Fusion), and (iii) produce task-conditioned outputs through a reasoning mechanism (Reasoning). Removing any stage yields a degenerate system.

Sufficiency. Widely used architectures, including CLIP (Radford et al., 2021), BLIP-2 (Li et al., 2023a), Flamingo (Alayrac et al., 2022), Video-LLaMA (Zhang et al., 2023a), Unified-IO (Lu et al., 2023), and Gemini (Team et al., 2023), can all be mapped to specific instantiations of AFR. Differences across these models primarily reflect where fusion occurs and how reasoning capacity is distributed, rather than fundamentally distinct computational principles.

2.2 Alignment: Normalizing Heterogeneous Modalities

Alignment transforms modality-specific signals into embeddings that are compatible for interaction. In practice, this is achieved through modality-specific encoders (e.g., ViTs for images, CNN- or transformer-based encoders for audio, and spatiotemporal models for video), followed by projection layers that map embeddings into a shared latent space.

Models such as CLIP and ALIGN (Jia et al., 2021a) exemplify alignment-driven design, in which image and text encoders are jointly trained with contrastive objectives, to learn a shared embedding space that supports zero-shot tasks. Although highly scalable, this approach sacrifices token-level grounding, i.e weaker associations between individual image regions and temporal events with corresponding language elements, limiting its performance on tasks that require fine-grained spatial or temporal reasoning (Yuksekgonul et al., 2023).

Alignment becomes substantially more challenging for time-varying modalities. Video-language models, such as Frozen-in-Time (Bain et al., 2021a) and Video-LLaMA (Zhang et al., 2023a), must reconcile temporal mismatches across frames, audio segments, and text tokens. Empirically, temporal alignment failure often manifests as drift in long-horizon reasoning, highlighting alignment as a primary bottleneck for video-centric MLLMs (Fu et al., 2025; Xiao et al., 2021).

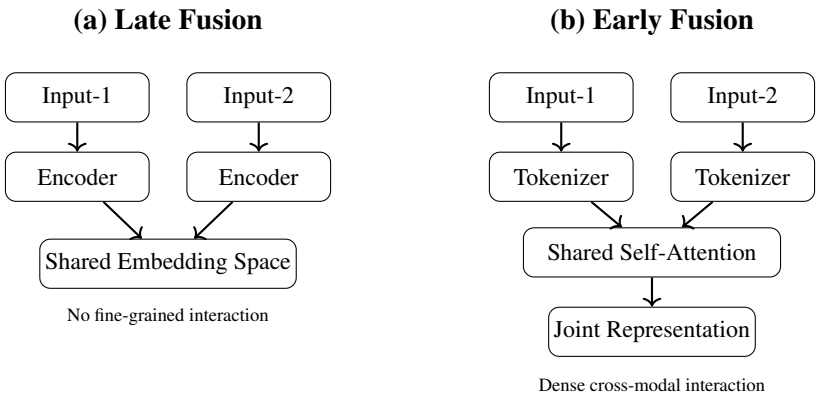

Figure 1: Comparison of fusion patterns in multimodal models.

2.3 Fusion: Cross-Modal Interaction Mechanisms

Fusion governs how modality representations align and exchange information. Existing MLLMs primarily instantiate one of the three fusion patterns. **Late fusion** (see Figure 1a), used in CLIP, combines two independently encoded modalities, i.e., the input has passed through separate encoding stages. This design is computationally efficient and modular, but it limits fine-grain interaction between modalities. On the contrary, **Early fusion** (see Figure 1b), used in models like Unified-IO (Lu et al., 2023), performs interaction among modality tokens often through self-attention at the cost of quadratic complexity. Apart from the above fusion mechanisms, some models like BLIP-2 (Li et al., 2023a), Flamingo (Alayrac et al., 2022), and VideoChat (Li et al., 2025) use **Cross-Attention fusion**, which strikes a middle ground by allowing one modality (typically language) to selectively attend to another. It is neither purely late nor early fusion.

Importantly, We argue that fusion mechanisms do more than mediate information flow: they shape the balance between perceptual fidelity, that is, how closely the model relies on what it actually sees or hears, and linguistic abstraction, the model's tendency to rely on language patterns and prior knowledge encoded in text, rather than on the raw perceptual input (Goyal et al., 2017).

2.4 Reasoning: Capacity Balance and Grounding

Reasoning in MLLMs is typically performed by a large language model, either as a decoder-only backbone or as part of a unified encoder-decoder. Architectural choices differ in the way the LLM interacts directly with other modality tokens. BLIP-2 (Li et al., 2023a) and Flamingo (Alayrac et al., 2022) interpose lightweight adapters (e.g., Q-Former or gated cross-attention) between the frozen encoders and the frozen LLMs, whereas one-head models feed modality tokens directly into the shared Transformer.

A recurring challenge is balancing multimodal perception with language understanding, as an imbalance in either can undermine grounded multimodal reasoning. Empirical studies show that MLLMs with weak visual grounding often exhibit hallucination and increased reliance on language shortcuts (Hamidullah et al., 2025), particularly in spatial and temporal reasoning tasks. We support the idea that increased multimodal perceptual capacity may improve grounding, but excessively strong modality conditioning might reduce the flexibility of language generation. SmolVLM finds that a larger encoder paired with a 360M-parameter LLM improves performance by 11.6% but at disproportionate parameter cost, only at 1.7B does the encoder's overhead become acceptable approx. 10%. This suggests that the encoder and LLM size have an empirical relationship. We call for systematic studies investigating how encoder and LLM capacities should be jointly balanced to maintain stable multimodal grounding, although such studies validating these scaling relationships remain limited to nonexistent. We extend the argument to say that models that pair small modality encoders with very large LLMs often exhibit hallucination and over-reliance on language priors.

2.5 Architectural Families Through AFR

The AFR framework discusses *what* an MLLM can accomplish, i.e., align heterogeneous modalities, enable cross-modal integration, and perform reasoning tasks. Viewed through AFR, most existing MLLMs that achieve these properties fall into four architectural families:

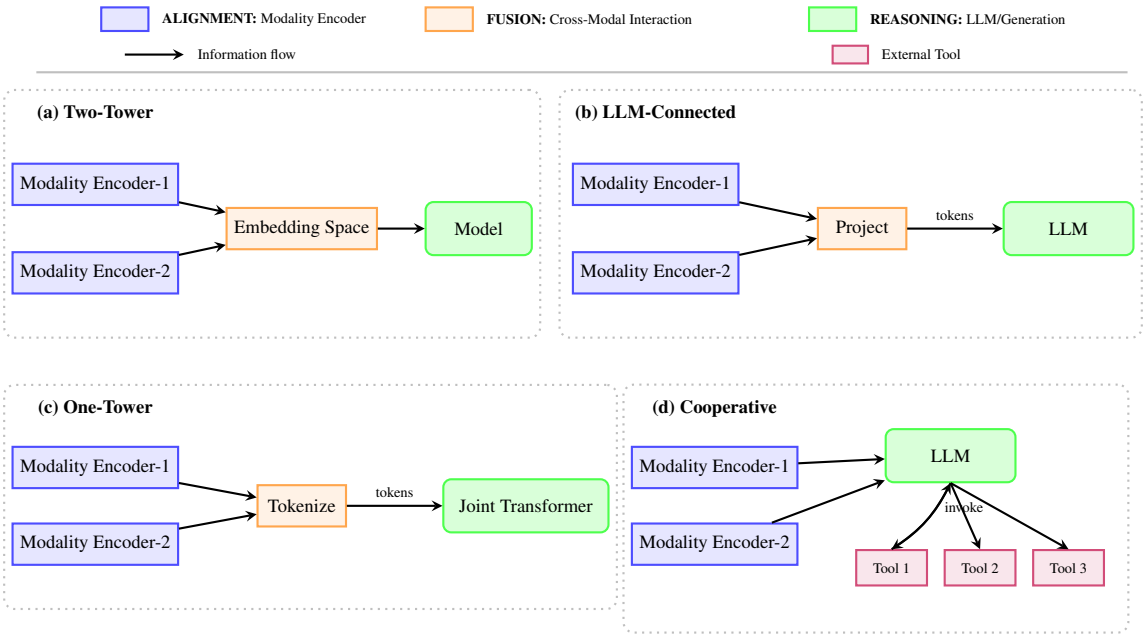

Figure 2: Architectural instances of the AFR framework

Failure Mode	AFR Stage	Root Cause	Architecture	Mitigation
Hallucination	Reasoning	LLM-encoder imbalance	LLM-connected	Grounding loss (Zhang et al., 2024)
Temporal drift	Alignment	Temporal compression	Video two-tower	Multi-scale temporal alignment (Zhang et al., 2023a)
Weak grounding	Fusion	Late pooling	Two-tower	Cross-attention
Error compounding	Reasoning	Discrete tools	Cooperative	Soft tool weighting (Schick et al., 2023)
Context overflow	Fusion	Quadratic attention	One-tower	Sparse fusion

Table 1: AFR-linked failure modes across architectural families.

- **Two-tower models** (CLIP, ALIGN, CLAP (Elizalde et al., 2023)), where different modalities are processed by an independent modality-specific encoder known as tower, that transforms the input into a latent representation before interaction with other modalities in a shared embedding space, and interact only at the representation level, enabling scalability and retrieval, but with limited reasoning depth as there is absence of a central LLM doing the reasoning (see Figure 2a).
- **LLM-connected Model** (BLIP-2 (Li et al., 2023a), Flamingo (Alayrac et al., 2022), Video-LLaMA (Zhang et al., 2023a), Qwen-Audio (Chu et al., 2023)), where modality-specific encoders are combined with a language model via learned projection layers, which unify reasoning through an LLM while retaining modular encoders (see Figure 2b).
- **One-tower models** (Unified-IO (Lu et al., 2023), SpeechT5 (Ao et al., 2022)), where all modalities are represented as a unified token sequence and processed jointly within a single transformer, enabling cross-modal interaction but at a higher computational cost (see Figure 2c).
- **Cooperative systems** (HuggingGPT (Shen et al., 2023)) where multimodal processing is decomposed into multiple specialized models or tools coordinated by a central LLM. Rather than relying on a single end-to-end architecture, these systems perform task planning, model selection, execution, and produce responses across heterogeneous components. (see Figure 2d).

No single family dominates across all tasks. LLM-connected architectures currently represent a practical sweet spot for many applications, while cooperative systems dominate general-purpose deployments. However, we hypothesize that one-tower architectures, combined with early fusion,

offer the most principled path toward truly native multimodal intelligence, provided their efficiency and data requirements can be addressed. While the proposed families capture the current models' landscape to a great extent, there are certain aspects that don't fit in the current framework, which are discussed in the appendix sections A and B

2.6 Implications

The AFR framework reframes MLLM design as a problem of managing information flow through the alignment, fusion, and reasoning stages. It explains why incremental scaling alone often yields diminishing results, failing to resolve grounding, compositional reasoning, and robustness (Thrush et al., 2022; Fu et al., 2025), and why advances in multimodal intelligence require architectural innovation, such as improved fusion and unified modeling (Alayrac et al., 2022; Li et al., 2023a; Lu et al., 2023), rather than task-specific optimization such as improving performance for a particular benchmark or task. In the following sections, we discuss how current training objectives and evaluation protocols interact with these architectural constraints.

3 Training Objectives and Data: Capabilities, Biases, and Trade-offs

3.1 Pre-training Data: What We Optimize For Is What We Learn

Most contemporary MLLMs rely on large-scale paired datasets curated from the web. Image–text corpora such as Conceptual Captions (Sharma et al., 2018), CC12M (Changpinyo et al., 2021), LAION-400M/5B (Schuhmann et al., 2021, 2022), and COYO-700M (Byeon et al., 2022) have been instrumental in enabling scalable cross-modal alignment. However, these datasets are largely collected from the web, and they primarily encourage broad image–text associations learning, rather than precise grounding between specific visual details and language. Specialized datasets such as TextCaps (Sidorov et al., 2020) and DocVQA (Mathew et al., 2021) introduce OCR-awareness and document reasoning, while VizWiz (Gurari et al., 2018) exposes models to noisy real-world visual conditions, partially mitigating dataset bias.

Audio–text alignment is typically learned from datasets such as AudioSet (Gemmeke et al., 2017), Clotho (Drossos et al., 2020), AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), LibriSpeech (Panayotov et al., 2015), and GigaSpeech (Chen et al., 2021), which support event recognition and speech grounding. Video–text learning relies on large-scale paired corpora such as HowTo100M (Miech et al., 2019), WebVid-2.5M (Bain et al., 2021b), HD-VILA-100M (Xue et al., 2022), YT-Temporal-1B (Zellers et al., 2022), and InternVid (Wang et al., 2024b), which emphasize temporal correspondence but often lack causal or intentional annotations.

Despite their scale, these datasets share two structural limitations. First, they primarily encode *correlational* relationships rather than *causal* relationships, i.e., the model prioritizes capturing the correlation in observed data over learning the underlying mechanism governing how changes in input propagate to the prediction. Second, they privilege static or weakly temporal alignment, leaving long-horizon reasoning (the ability to integrate information over extended sequences of time or events) (Chen et al., 2024b) and cross-modal state tracking (tracking which objects have moved and which sounds correspond to which events) underconstrained (Grauman et al., 2022). We argue that future MLLMs will need data that capture how things interact, what actions are being pursued, and how situations evolve over time, rather than just isolated pairs of inputs and descriptions.

3.2 Training Objectives as Inductive Biases

Although MLLMs operate across diverse modalities, their training objectives can be categorized into three broad paradigms: contrastive, masked modeling, and generative objectives. These objectives closely correspond to different stages of the proposed AFR framework introduced earlier. Critically, each objective imposes distinct inductive biases that shape what models learn and, consequently, where and why they fail.

3.2.1 Contrastive Objectives: Alignment Without Reasoning

Contrastive learning, popularized by CLIP and ALIGN, remains the dominant paradigm for large-scale multimodal pretraining. Given paired modality–text samples, the objective encourages aligned pairs to have high similarity in a shared embedding space while pushing apart mismatched pairs. Variants such as Image–Text Contrastive (CLIP, ALIGN), Image–Text Matching (BLIP (Li et al., 2022), ALBEF (Li et al., 2021)), region-level alignment (OSCAR (Li et al., 2020), VinVL (Zhang

et al., 2021)), and global–regional objectives (UNITER (Chen et al., 2020)) improve ability to capture finer details, but remain fundamentally alignment-driven. Audio–text contrastive learning (CLAP (Elizalde et al., 2023), AudioCLIP (Guzhov et al., 2022)) and video–text contrastive learning (VideoCLIP (Xu et al., 2021), Frozen-in-Time (Bain et al., 2021a)) extend this paradigm into temporal data.

We believe that, although contrastive objectives scale exceptionally well and produce transferable representation, they mainly optimize representational alignment than compositional reasoning or causal relation. However, models trained predominantly with contrastive loss often excel at retrieval and zero-shot tasks.

3.2.2 Masked Modeling: Contextual Reconstruction Across Modalities

Masked modeling objectives train models to reconstruct missing portions of one modality using the remaining context and aligned modalities. Originally introduced as Masked Language Modeling in BERT (Devlin et al., 2019), this paradigm has been extended to images (ViLBERT (Lu et al., 2019), LXMERT (Tan and Bansal, 2019), VL-BERT (Su et al., 2020)), video (VideoBERT (Sun et al., 2019), ActBERT (Zhu and Yang, 2020)), and audio (BEATs (Chen et al., 2023), M2D-CLAP (Elizalde et al., 2023)). Masked modeling encourages localized reasoning, where the model's reasoning is confined to a short context window and leverages cross-modal contexts, partially addressing the limitations of pure contrastive learning.

However, we argue that masked objectives primarily focus on reconstructing missing information. Although they produce rich contextual representations, they provide little to no training for long-range planning, abstraction, or interactive reasoning, where reasoning unfolds over multiple interactions and model updates its understanding based on feedback or new information.

3.2.3 Generative Objectives: Toward Multimodal Reasoning

Generative objectives train MLLMs to predict one modality conditioned on another, typically in an autoregressive fashion. This paradigm underlies image captioning, visual question answering, speech recognition and translation, and video summarization. Recent LLM-centric models increasingly rely on generative training to unify perception and language reasoning. Unlike contrastive objectives, generative losses directly couple perception with symbolic output, enabling multi-step reasoning and compositional generation.

We argue that generative objectives are computationally expensive (Henighan et al., 2020) and sometimes sensitive to data noise, since noisy target data distributions can mislead the model.

3.3 Hybrid Objectives and Open Challenges

Most state-of-the-art MLLMs are trained using a weighted combination of these objectives. Although this hybrid strategy balances alignment, contextual modeling, and generation, it also introduces tension in optimization. In practice, alignment-heavy objectives dominate early training, often inducing bias in models toward shallow cross-modal correlations. We argue that future MLLMs must move beyond static objective mixtures toward more adaptive, capability-aware training curricula that explicitly target temporal reasoning, grounding fidelity, and interpretability.

4 Evaluation: What Is Measured and What Is Missed

Despite rapid advances in architectural and training, evaluation remains one of the least mature components of the MLLM pipeline. In this section, we argue that existing evaluation practices systematically overestimate multimodal intelligence by mistaking correlation with understanding.

4.1 Benchmarks as Probes of Capability

Benchmarks are often treated as task collections, but in practice they act as *probes* into specific model capabilities. Building on this view, we organize benchmarks along four functional dimensions (Li et al., 2024d), while emphasizing that most existing benchmarks disproportionately focus on the first and third.

Understanding benchmarks evaluate whether a model can perceive and associate multimodal inputs with semantic labels. Image-based benchmarks such as MMBench (Liu et al., 2024b), SEED-Bench (Li et al., 2024a), OCRBench (Liu et al., 2024c), and MMUBench (Li et al., 2024b) primarily test visual perception, recognition, and basic scene comprehension. In the audio domain, benchmarks such as SUPERB (Yang et al., 2021), AudioSet (Gemmeke et al., 2017), and AudioCaps (Kim et al., 2019) evaluate event detection,

speech understanding, and low-level audio semantics. Video benchmarks such as MVBench (Li et al., 2024c), VideoMME (Fu et al., 2025), and EgoSchema (Mangalam et al., 2023) extend these evaluations to spatio-temporal perception and long-video understanding. Although strong performance on these benchmarks indicates effective alignment, it does not guarantee a grounded or compositional understanding.

Reasoning benchmarks aim to assess logical inference, relational reasoning, and temporal integration across modalities. Image-based reasoning benchmarks such as SPIQA (Pramanick et al., 2024), M3Exam (Zhang et al., 2023b), MathVision (Wang et al., 2024a), MathV360K (Shi et al., 2024), MMSci (Li et al., 2024e), and CMMU (He et al., 2024) focus on domain-specific, mathematical, and scientific reasoning grounded in visual input. Benchmarks including MLLM-CompBench (Kil et al., 2024), VCog-Bench (Cao et al., 2025), MARVEL (Jiang et al., 2024), and VisualCoT (Shao et al., 2024) emphasize relational and multi-step reasoning, often requiring models to compose multiple visual and linguistic cues. Audio reasoning benchmarks, such as SpeechR (Yang et al., 2025), MMAR (Ma et al., 2025), MMAU (Sakshi et al., 2025), AVH-Bench (Sung-Bin et al., 2025), ClothoAQA (Lipping et al., 2022), and Daily-Omni (Zhou et al., 2025), assess contextual and cross-modal reasoning over speech and acoustic signals, evaluating whether models can effectively integrate temporal audio cues with language. In the video domain, benchmarks such as VideoCoT (Zhang et al., 2025c), CLEVRER-Human (Mao et al., 2022), TGIF-QA (Jang et al., 2017), Video-Bench (Ning et al., 2025), and MVBench (Li et al., 2024c) assess temporal reasoning, event ordering, and causal relationships across sequences, while MovieQA (Tapaswi et al., 2016), STAR (Wu et al., 2021), and NExT-QA (Xiao et al., 2021) target commonsense and multi-step reasoning over longer narratives. In practice, many reasoning benchmarks reduce to pattern matching in short contexts, particularly when answer spaces are limited. As a result, models trained with large contrastive corpora often perform well without explicitly modeling causality or temporal dependencies.

Generation benchmarks measure a model's ability to produce coherent outputs, such as captions, answers, or summaries, conditioned on multimodal inputs. Image-based generation benchmarks, such as MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017b), Flickr30K (Plummer et al., 2015), VIST (Huang et al., 2016), VisIT-Bench (Bitton et al., 2023), and MIA-Bench (Qian et al., 2025), focus on image captioning and image-conditioned text generation, emphasizing descriptive completeness and fluency. In the audio domain, benchmarks such as LibriTTS (Zen et al., 2019), InstructTTSEval (Huang et al., 2025b), Clotho (Drossos et al., 2020), and AudioCaps (Kim et al., 2019) evaluate speech synthesis and audio captioning, measuring naturalness and semantic relevance. Video generation benchmarks such as VATEX (Wang et al., 2019), MSVD (Chen and Dolan, 2011), ActivityNet-Captions (Krishna et al., 2017a), VideoInstruct-100K (Maaz et al., 2024), and MSR-VTT (Xu et al., 2016) extend these evaluations to spatiotemporal inputs, testing whether models can generate temporally coherent descriptions or instructions. These benchmarks primarily evaluate linguistic fluency and relevance, while only weakly constraining factual grounding or faithfulness to perceptual evidence.

Application benchmarks test end-to-end task performance in settings such as document understanding, assistive perception, or multimodal search. Although these benchmarks are closer to real-world deployment, they often inherit the same evaluation limitations as their constituent subtasks. Audio-focused application benchmarks, such as MusicBench (Melechovsky et al., 2024) and MuChoMusic (Weck et al., 2024), evaluate music understanding, analysis and composition combining both perceptual and generative capabilities. In image-based and video-based benchmarks such as PubMedVision (Chen et al., 2024c), Med-VidQA (Gupta et al., 2023), and M3D-Bench (Li et al., 2023b), multimodal reasoning is evaluated in medical and scientific contexts.

A key observation is that few benchmarks explicitly isolate failure modes at the Alignment, Fusion, and Reasoning stages. Instead, existing analyses reveal these failures only indirectly, such as many VLM benchmarks admit unimodal shortcuts (Huang et al., 2025a), one modality can dominate or sabotage another (Zhang et al., 2025a), and recent reasoning-oriented benchmarks expose gaps in causal and cross-modal reasoning without disentangling whether errors arise from misalignment, insufficient fusion, or flawed reasoning (Zhang et al., 2025b; Chen et al., 2025). Consequently, it often remains unclear which stage is responsible

for a given model failure. Further discussion on evaluation metrics is provided in the Appendix D

5 Toward Next-Generation MLLMs

The limitations discussed suggest that incremental scaling of existing MLLMs is unlikely to yield general multimodal intelligence. Instead, progress will require rethinking the core design choices that span architecture, reasoning, learning dynamics, and deployment. We outline several principled directions that we believe are both necessary and underexplored.

From Modular Pipelines to Native Multimodal Architectures. Most current MLLMs rely on modality-specific encoders mapped into a shared embedding space, constraining cross-modal interaction and encouraging late fusion. A key research direction is the development of native multimodal architectures in which all modalities are represented as a unified token stream and processed jointly within a single transformer, enabling early fusion, tighter alignment, and more expressive multimodal reasoning. However, their success depends on the availability of robust audio and video backbones, which currently lag behind those of image and text encoders.

Explicit Reasoning, Concept Discovery, and Interpretability. Despite impressive generative fluency, current MLLMs exhibit limited capabilities in multimodal concept discovery, causal inference, and multi-step reasoning. These shortcomings are not merely data-related but reflect the absence of an explicit reasoning structure. Promising directions include hybrid neuro-symbolic approaches (Yi* et al., 2020; Yi et al., 2018; Choi et al., 2024) that introduce intermediate abstractions and structured reasoning over multimodal evidence. Crucially, such approaches also improve interpretability by producing step-by-step explanations grounded in perceptual inputs. We argue that reasoning and interpretability should be treated as first-class objectives, rather than post-hoc analysis tools.

Learning Beyond Static Pretraining. MLLMs today are largely static systems: adapting to new tasks, domains, or modalities often requires costly retraining and risks catastrophic forgetting. Continual and lifelong learning (van de Ven et al., 2024; Yu et al., 2024), therefore, represents a key frontier. Mechanisms such as modality-aware adapters, replay buffers, and selective parameter isolation offer partial solutions, but a deeper challenge remains: enabling MLLMs to update knowledge while preserving cross-modal consistency. Addressing this will be essential for real-world deployment in dynamic environments.

Efficiency as a Design Constraint, Not an Afterthought. The computational and environmental costs of training MLLMs raise fundamental questions about the sustainability of these models. While parameter-efficient fine-tuning (PEFT) (Mu et al., 2024), curriculum learning (Bengio et al., 2009), and data-efficient strategies help reduce costs, they do not address the root issue: architectures that scale multimodal inputs inefficiently. We argue that treating efficiency as an important architectural constraint will be increasingly important for the sustainable development of future MLLMs. This includes modular execution, selective modality activation, and adaptive token processing, rather than uniform scaling.

Trustworthy and Robust Multimodal Deployment. Finally, the deployment of MLLMs introduces risks that are qualitatively distinct from those of unimodal LLMs. Cross-modal prompt injection, adversarial audio or visual cues, and semantic misalignment attacks (Qi et al., 2024; Wang et al., 2025; Cao et al., 2024) expose new vulnerabilities that cannot be mitigated by text-only safety mechanisms. Ensuring robustness, fairness, and privacy in MLLMs requires integrated solutions that span data curation, model architecture, and evaluation protocols. We contend that safety and ethics must be co-designed with multimodal reasoning systems, rather than retrofitted after deployment.

6 Conclusion

MLLMs have rapidly transitioned from task-specific vision–language systems to a dominant paradigm for building general-purpose AI models. While recent progress has been driven largely by scale and data, this paper argues that such trends alone are insufficient to achieve robust and general multimodal intelligence. We advocate a shift towards native multimodal architectures, explicit and interpretable reasoning mechanisms, continual learning paradigms, and efficiency-aware system design. We emphasize that safety, fairness, and robustness must be treated as core architectural concerns rather than post-hoc constraints.

7 Limitations

In the paper, the proposed Alignment—Fusion—Reasoning (AFR) framework is a unifying abstraction for analyzing Multimodal Large Language Models and provides specific directions for architecture, training, and evaluation. We acknowledge the following limitations of the framework :

- **Scope of AFR.** The framework is primarily a descriptive abstraction rather than a formally proven theory. We do not validate AFR through controlled experiments. The failure modes summarized in Table 1 are derived from prior literature and understanding, and future work would be to test whether AFR-predicted behaviors consistently emerge across architectures.
- **Coverage of Architectural Families.** The AFR-based architectural families capture dominant paradigms prevalent at the time of writing, but emerging architectures, such as diffusion-based, may not fit cleanly within this taxonomy. The framework may therefore require extension as the field evolves.
- **Modality Coverage Imbalance.** Although AFR is intended to be modality-agnostic, the paper focuses more on Vision-Language Models than on audio- and video-based models, reflecting the limited availability of literature in those domains.
- **Reliance on Published Results.** The observations discussed and the position statements drawn throughout the paper are based on prior work, rather than on new experiments conducted by the authors.
- **Time Limited.** The MLLM landscape is evolving rapidly, and architectures or training paradigms introduced after our paper may address some of the limitations identified. Although we focus on structural issues likely to persist across generations, some architecture-specific observations may become outdated over time.
- **The Scope of Work.** As a position paper, this work prioritizes breadth of synthesis over depth in any single area. The future directions proposed in Section 5 are intended as research agendas rather than fully developed technical solutions.

8 Ethics Statement

This study involves a review and synthesis of previously published research and publicly available datasets. No human or user data were collected or analyzed. All works included in this survey were appropriately cited to the best of our knowledge to acknowledge original authorship. Our review primarily covers research from ACL, EMNLP, NeurIPS, ICLR, ICML, CVPR, ECCV, and arXiv, and we may have missed some important work published in other venues. We will stay up to date on ongoing discussions in the research community and plan to revise our work in the future to include overlooked contributions.

References

Adnen Abdessaied, Ekta Sood, and Andreas Bulling. 2022. Video language co-attention with multimodal fast-learning feature fusion for videoqa. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 143–155.

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, and 1 others. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 5723–5738.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021a. Frozen in time: A joint video and

image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021b. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schimdt. 2023. Visit-bench: a benchmark for vision-language instruction following inspired by real-world use. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 26898–26922.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2024. Defending against alignment-breaking attacks via robustly aligned llm. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10542–10560.

Xu Cao, Bolin Lai, Wenqian Ye, Yunsheng Ma, Joerg Heintz, Meihuan Huang, Jintai Chen, Jianguo Cao, Aidong Zhang, and James Matthew Rehg. 2025. Benchmarking visual cognition of multimodal LLMs via matrix reasoning.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 4376–4380. International Speech Communication Association.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Frank Wang. 2024b. Rextime: A benchmark suite for reasoning-across-time in videos. *Advances in Neural Information Processing Systems*, 37:28662–28673.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024c. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, Miami, Florida, USA. Association for Computational Linguistics.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, YANDONG LI, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, and Boqing Gong. 2025. Omnixr: Evaluating omni-modality language models on reasoning across modalities. In *International Conference on Representation Learning*, volume 2025, pages 68100–68118.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Minkyu Choi, Harsh Goel, Mohammad Omama, Yunhao Yang, Sahil Shah, and Sandeep Chinchali. 2024. Towards neuro-symbolic video understanding. In *European Conference on Computer Vision*, pages 220–236. Springer.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, and 1 others. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

DeepLearning.AI. 2024. The problem with benchmark contamination in ai. *The Batch*. Accessed: 2025-12-15.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Haytham M Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2283–2294.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Wentao Ge, Shunian Chen, Guiming Hardy Chen, Junying Chen, Zhihong Chen, Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, and 1 others. 2023. Mllm-bench: Evaluating multimodal llms with per-sample criteria. *arXiv preprint arXiv:2311.13951*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE.

Yasser Hamidullah, Koel Dutta Chowdhury, Yusser Al Ghussin, Shakib Yazdani, Cennet Oguz, Josef van Genabith, and Cristina España-Bonet. 2025. Grounding or guessing? visual signals for detecting hallucinations in sign language translation. *arXiv preprint arXiv:2510.18439*.

Zheqi He, Xinya Wu, Pengfei Zhou, Richeng Xuan, Guang Liu, Xi Yang, Qiannan Zhu, and Hua Huang. 2024. Cmmu: A benchmark for chinese multimodal multi-type question understanding and reasoning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 830–838. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, and 1 others. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 7514–7528.

Jinsheng Huang, Liang Chen, Taian Guo, Fu Zeng, Yusheng Zhao, Bohan Wu, Ye Yuan, Haozhe Zhao, Zhihui Guo, Yichi Zhang, and 1 others. 2025a. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4805–4822.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. 2025b. Instructttseval: Benchmarking complex natural-language instruction following in text-to-speech systems. *arXiv preprint arXiv:2506.16381*.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, and 1 others. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision–language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1725–1735, Dubrovnik, Croatia. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021a. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021b. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Shixin Jiang, Jiafeng Liang, Jiyuan Wang, Xuan Dong, Heng Chang, Weijiang Yu, Jinhua Du, Ming Liu, and Bing Qin. 2025. From specific-MLLMs to omni-MLLMs: A survey on MLLMs aligned with multimodalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8617–8652, Vienna, Austria. Association for Computational Linguistics.

Yifan Jiang, Jiarui Zhang, Kexuan Sun, Zhivar Sourati, Kian Ahrabian, Kaixin Ma, Filip Ilievski, and Jay Pujara. 2024. Marvel: Multidimensional abstraction and reasoning through visual evaluation and learning. *Advances in Neural Information Processing Systems*, 37:46567–46592.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Jihyung Kil, Zheda Mai, Justin Lee, Arpita Chowdhury, Zihe Wang, Kerrie Cheng, Lemeng Wang, Ye Liu, and Wei-Lun Chao. 2024. Mllm-compbench: a comparative reasoning benchmark for multimodal llms. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Dohwan Ko, Joonmyung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, and Hyunwoo J Kim. 2022. Video-text representation learning via differentiable weak temporal alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5016–5025.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017b. Visual genome: Connecting language

and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seedbench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2024b. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2025. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. 2024d. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*.

Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, Fukun Yin, Gang Yu, and Tao Chen. 2023b. M3dbench: Let's instruct large models with multi-modal 3d prompts. *arXiv preprint arXiv:2312.10763*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pages 121–137. Springer.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, and 1 others. 2026. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, and 1 others. 2024e. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*.

Bin Lin, Zhenyu Tang, Yang Ye, Jinfa Huang, Junwu Zhang, Yatian Pang, Peng Jin, Munan Ning, Jiebo Luo, and Li Yuan. 2026. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024c. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2023. UNIFIED-IO: A unified model for vision, language, and multi-modal tasks. In *The Eleventh International Conference on Learning Representations*.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, and 1 others. 2024. Fairclip: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, Bangkok, Thailand. Association for Computational Linguistics.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Jiayuan Mao, Xuelin Yang, Xikun Zhang, Noah Goodman, and Jiajun Wu. 2022. Clevrer-humans: Describing physical and causal events the human way. *Advances in Neural Information Processing Systems*, 35:7755–7768.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2024. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2024. Mustango: Toward controllable text-to-music generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8286–8309.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Jie Mu, Wei Wang, Wenqi Liu, Tiantian Yan, and Guanglu Wang. 2024. Multimodal large language model with lora fine-tuning for multimodal sentiment analysis. *ACM Transactions on Intelligent Systems and Technology*.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2025. Videobench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Computational Visual Media*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume*, pages 487–503.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. 2024. Spiqa: A dataset for multimodal question answering on scientific papers. *Advances in Neural Information Processing Systems*, 37:118807–118833.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan

Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. *arXiv preprint arXiv:2503.17489*.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. 2024. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21527–21536.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2025. MIA-bench: Towards better instruction following evaluation of multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2022. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239.

Chahat Raj, Anjishnu Mukherjee, Aylin Caliskan, Antonios Anastasopoulos, and Ziwei Zhu. 2024. BiasDora: Exploring hidden biased associations in vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10439–10455, Miami, Florida, USA. Association for Computational Linguistics.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessí, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: language models can teach themselves to use tools. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *CoRR*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4663–4680, Miami, Florida, USA. Association for Computational Linguistics.

Mustafa Shukor, Enrico Fini, Victor Guilherme Turrisi da Costa, Matthieu Cord, Joshua Susskind, and Alaaeldin El-Nouby. 2025. Scaling laws for native multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12–23.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer.

Zijia Song, Zelin Zang, Yelin Wang, Guozheng Yang, Wanyu Chen, Miaoyu Wang, Stan Z Li, and 1 others. 2024. Set-clip: Exploring aligned semantic from low-alignment multimodal data through a distribution view. *arXiv preprint arXiv:2406.05766*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3645–3650.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2025. AVHBench: A cross-modal hallucination benchmark for audio-visual large language models. In *The Thirteenth International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Gido M van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. 2024. Continual learning and catastrophic forgetting. *arXiv preprint arXiv:2403.05175*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Ali Vosoughi, Luca Bondi, Ho-Hsiang Wu, and Chenliang Xu. 2024. Learning audio concepts from counterfactual natural language. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370. IEEE.

Beinan Wang. 2022. A parallel implementation of computing mean average precision. *arXiv preprint arXiv:2206.09504*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Le Wang, Zonghao Ying, Tianyuan Zhang, Siyuan Liang, Shengshan Hu, Mingchuan Zhang, Aishan Liu, and Xianglong Liu. 2025. Manipulating multimodal agents via cross-modal prompt injection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10955–10964.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024b. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *CoRR*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. 2021. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. 2022. Advancing high-resolution video-language representation with large-scale video transcriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. Superb: Speech processing universal performance benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. 2025. Speechr: A benchmark for speech reasoning in large audio-language models. *arXiv preprint arXiv:2508.02018*.

Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31.

Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S Yu, and Irwin King. 2024. Recent advances of multimodal continual learning: A comprehensive survey. *arXiv preprint arXiv:2410.05352*.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16375–16387.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *Interspeech 2019*.

Huimin Zeng, Zhenrui Yue, Yang Zhang, Lanyu Shang, and Dong Wang. 2024. Fair federated learning with biased vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10002–10017, Bangkok, Thailand. Association for Computational Linguistics.

Chenyu Zhang, Minsol Kim, Shohreh Ghorbani, Jingyao Wu, Rosalind Picard, Patricia Maes, and Paul Pu Liang. 2025a. When one modality sabotages the others: A diagnostic lens on multimodal reasoning. *arXiv preprint arXiv:2511.02794*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and Jainwei Yang. 2024. Llava-grounding: Grounded visual chat with large multimodal models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLIII*, page 19–35, Berlin, Heidelberg. Springer-Verlag.

Kaiyuan Zhang, Chenghao Yang, Zhoufutu Wen, Sihang Yuan, Qiuyue Wang, Chaoyi Huang, Guosheng Zhu, He Wang, Huawenyu Lu, Jianing Wen, and 1 others. 2025b. Mme-cc: A challenging multi-modal evaluation benchmark of cognitive capacity. *arXiv preprint arXiv:2511.03146*.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. 2025c. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 12745–12752, New York, NY, USA. Association for Computing Machinery.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Yiman Zhang, Ziheng Luo, Qiangyu Yan, Wei He, Borui Jiang, Xinghao Chen, and Kai Han. 2025d. Omnieval: A benchmark for evaluating omni-modal models with visual, auditory, and textual inputs. *arXiv preprint arXiv:2506.20960*.

Xiongtao Zhou, Jie He, Yuhua Ke, Guangyao Zhu, Víctor Gutiérrez-Basulto, and Jeff Pan. 2024. An empirical study on parameter-efficient fine-tuning for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10057–10084.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. 2025. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.

Yongshuo Zong, Oisin Mac Aodha, and Timothy M Hospedales. 2024. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5299–5318.

A Mixture-Of-Expert architecture

As current MLLMs are scaled to enormous parameter sizes, they encounter two major problems: the inability to handle diverse datasets and tasks, and high computational resource consumption during training and inference. To tackle the problem, the Mixture-of-Experts (MoE) architecture was introduced, where a submodel, called an expert, is trained to handle a specific type of data or task.

The core components of the MoE architecture are as follows: the Gating Mechanism, which routes the token to the desired expert network, the Expert Network, which handles a particular data type or task and the Routing Mechanism, which decides which expert handles a given data type or task. During training, the routing mechanism is used to route different data to emerging expert networks. We will discuss a few architectures to better understand.

MoE-LLaVA (Lin et al., 2026), extends the LLaVA architecture by including the MoE layer in the LLM backbone to improve the multimodal reasoning capabilities of the model. Similar to LLaVA, the vision encoder is used with a projection layer to align the visual embeddings with the LLM space. However instead of relying on the LLM fully-connected neural network, MoE-LLaVA replaces them with selected fully-connected layers with sparse expert layers, where a gating mechanism routes the token to specialized experts. This architecture allows the model to scale without increasing inference cost. However, according to the AFR framework, the model still qualifies as a standard alignment model, where the visual input is projected into the same space as a language-centric reasoning model.

Recent MoE models, such as **KimiVL (Team et al., 2025)**, follow the same architecture paradigm, emphasizing multimodal reasoning, long-context understanding, and strong agent capabilities.

B Omni-Models

While recent trends in MLLMs have shifted from modality-paired modeling to omni modeling, where the model is designed to process, reason, and understand multiple modalities within a unified framework.

We can consider omni-models (O-LLMs) as an extension of modality-specific models in terms of architecture, following the same steps as any MLLM: multimodality encoding, alignment, interaction, and generation, as stated in (Jiang et al., 2025). We will discuss the omi-model architecture and training methodology to better understand. Most of the Omni-models are a combination of different encoding and alignment methods.

In O-LLMs, modalities are encoded using continuous embeddings, discrete tokens, or a hybrid representation that combines both, allowing each modality to retain its properties. The encoded modalities are then aligned with the LLM space, as in MLLMs, using a projection layer or an embedding-based alignment. Once the modalities are aligned, LLM acts as the reasoning engine, processing all modalities. We argue that current O-LLMs remain predominantly text-centric architectures, with modality adapters expanding their capabilities to handle multiple modalities. For the generation task, O-LLMs either produce textual instructions for external modality-specific generators or directly predict modality tokens that are reconstructed through dedicated decoders. Training of O-LLMs typically follows a two-stage paradigm consisting of alignment pretraining, which maps heterogeneous modalities into the LLM representation space, followed by instruction-finetuning, which enhances cross-modal understanding and reasoning using both unimodal and multimodal instruction datasets.

C Omni Models are not One-Tower Models

A common misconception is that omni-models are often mistaken for one-tower models. While both try to process the modalities uniformly, we strictly

enforce that they remain distinct models at an abstract level.

It is easier to consider Omni-modal more of a design choice than an architectural design, as is the case with one tower. A model can be omni-modal without being a unified architecture.

From the perspective of the proposed AFR framework, most O-LLMs can be classified as a co-operative system rather than a true one-tower model. Most of the O-LLMs, use modality specific encoders and connect them to LLM through projection layers. Although these systems support multimodality, their reasoning backbone is still predominantly LLMs and they are as good as reasoning in LLMs.

On contrast, the one-tower models attempt to process all the modalities in a shared transformer using a unified representation. So, that interaction between the modalities occur naturally rather than through external alignment modules or modality adapters. So we put forward the argument that O-LLMs create an illusion of multimodality. But in practice most of the O-LLMs are modular-system aligned and text dependent.

Illusion of multimodality, while multiple models are termed "natively multimodal", they can be simply interpreted as language models augmented with multimodal data rather than fundamentally a new multimodal architecture. Benchmarks such as OmniBench (Li et al., 2026) have shown the limitations of omni-modals in following instructions and reasoning capabilities when dealing with tri-modality, even when the model is provided with an alternative textual representation of the image or audio. Similarly. OmniEval (Zhang et al., 2025d), evaluate the models on tasks where strong coupling between modalities is required and force the model to leverage the perception of all modaities involved. Evaluation of models on OmniXR (Chen et al., 2025) reveals that O-LLMs show a performance drop when presented with complex multimodal inputs in tasks where cross-modal reasoning is required. We would also like to draw attention on the sparse availability of benchmarks to evaluate the current O-LLMs.

D Metrics and Their Inductive Biases

Evaluation metrics play a central role in shaping model development, yet the most commonly used metrics are often inherited from unimodal NLP, vision, or speech tasks. These metrics can be broadly grouped by the output they measure, rather than by task, revealing limitations between applications.

Metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) are widely used in image captioning, video captioning, visual question answering, and document understanding. These metrics quantify n-gram overlap or semantic similarity between the generated text and the reference text. Previous work has shown that models can achieve high scores in these metrics while hallucinating that which is not present in visual or auditory input (Rohrbach et al., 2018; Liu et al., 2024a). Several multimodal QA benchmarks also employ these metrics to evaluate the output, focusing on linguistic fluency over grounded output (Anderson et al., 2018; Abdessaied et al., 2022). Metrics such as Clipscore (Hessel et al., 2021) and BlipScore (Li et al., 2022, 2023a) measure the semantic similarity and compatibility between text and images. Accuracy and F-1 score are also commonly used in multimodal QA and reasoning benchmarks. Although simple and easy to interpret, these metrics collapse a complex multimodal reasoning task into a binary problem, obscuring whether errors arise from misalignment, insufficient fusion, or flawed reasoning (Agrawal et al., 2016; Huang et al., 2025a). As a result, models exploit unimodal shortcuts and dataset biases to score higher, without performing genuine multimodal inference.

For tasks involving alignment and retrieval, metrics such as Recall@K, Mean Reciprocal Rank, and Mean Rank (Karpathy and Fei-Fei, 2015) are commonly used as standard evaluation tools. These metrics effectively measure the structure of the shared embedding space, but primarily focus on alignment quality. For tasks involving localization and grounding, metrics such as Intersection-over-Union (IoU) and mean Average Precision (mAP) (Wang, 2022) are commonly used. These metrics capture geometric overlap between predicted and reference regions.Speech and audio metrics, such as Word Error Rate (WER), Character Error Rate (CER), and Mean Opinion Score (MOS), are used for speech recognition and synthesis.

Together, current benchmarks and metrics exhibit three systemic limitations. First, they overwhelmingly favor short-context, correlation-driven behavior, aligning closely with contrastive and masked

modeling objectives (Radford et al., 2021; Chen et al., 2020; Li et al., 2021). Second, they provide weak supervision for temporal reasoning, causal inference, and cross-modal consistency (Fu et al., 2025; Chen et al., 2025; Xiao et al., 2021). Third, they offer limited interpretability, making it difficult to diagnose whether the improvements stem from better perception, fusion, or reasoning. We argue that progress toward truly native MLLMs requires a shift from task-centric to capability-centric evaluation. The benchmarks should be designed to explicitly stress the individual components of the AFR, as listed in Table 1, to disentangle alignment from reasoning, and to measure robustness under controlled perturbations.

D.1 MLLMs as Judges for Multimodal Evaluation

MLLM-as-a-judge evaluation refers to the use of a strong multimodal model to score, compare, rank, or critique the outputs of another multimodal system while directly observing the relevant input modality. This differs from text-only LLM judging: the judge is not merely reading a caption or transcript, but can inspect the image, video, audio, document, or cross-modal context alongside the candidate answer. The motivation is clear for MLLMs because many outputs are open-ended and reference-light. Visual explanations, video summaries, audio-grounded answers, and image-conditioned instructions may all be valid in multiple forms, making exact-match and lexical-overlap metrics insufficient.

Recent work has begun to formalize this paradigm. MLLM-Bench evaluates MLLMs with per-sample criteria and uses a potent MLLM as a judge, reporting strong agreement with human evaluation when detailed criteria are available (Ge et al., 2023). The MLLM-as-a-Judge benchmark further studies three judging modes: scoring evaluation, pair comparison, and batch ranking, finding that MLLMs are more reliable in pairwise comparison than in absolute scoring or ranking (Chen et al., 2024a). Judge Anything extends the idea beyond image-text settings toward any-to-any modality evaluation, including multimodal understanding and generation tasks across images, audio, and video (Pu et al., 2025).

An MLLM judge is useful because it can evaluate dimensions that conventional metrics often miss. For image-language tasks, it can check whether the response is grounded in visible evidence, whether object attributes and spatial relations are correct, and whether the answer follows the visual instruction. For video, it can assess event order, temporal consistency, and whether the answer depends on the correct frames rather than a static scene prior. For audio and speech, it can evaluate whether the response reflects acoustic events, speaker intent, or temporal cues. In open-ended generation, an MLLM judge can also assess completeness, helpfulness, safety, and explanation quality under a task-specific rubric.

However, MLLM judges should not be treated as objective ground truth. Their perceptual and reasoning failures can mirror the failures of the models being evaluated. A judge may overlook small visual details, ignore an audio cue, over-weight fluent language, or hallucinate evidence that is not present in the modality. It may also show position bias in pairwise comparisons, preference for longer responses, model-identity bias, and modality bias, where one modality dominates the judgment even when another modality contains the decisive evidence. These risks are especially important when the judge and evaluated model share similar training data or architectural assumptions, because the evaluation can become circular.

A reliable MLLM-as-a-judge protocol should therefore be explicitly multimodal and auditable. The report should specify the judge model and version, input modalities exposed to the judge, prompt, rubric, reference information, decoding settings, and aggregation method. Pairwise comparisons should randomize and swap candidate order. Scoring should use clear rubrics with capability-specific criteria rather than a single vague quality score. Human validation should be performed on a representative subset, especially for tasks requiring fine-grained grounding, temporal reasoning, or safety-sensitive decisions. Where possible, judge outputs should include short evidence-based rationales that identify the image regions, audio events, frames, or document spans supporting the score.

We view MLLM-as-a-judge as a useful complement to benchmark metrics and human evaluation, not as a replacement for them. Its strongest role is diagnostic: it can reveal groundedness, temporal, and cross-modal reasoning failures that surface metrics miss. Used carefully, MLLM judges can make multimodal evaluation more scalable and interpretable; used carelessly, they can overstate progress by rewarding fluent but weakly grounded outputs. Future evaluation protocols should there-

fore treat the judge itself as an MLLM under evaluation, measuring its alignment with human preferences, robustness to modality perturbations, and sensitivity to cross-modal evidence.

E Applications as Stress Tests for Multimodal Intelligence

Applications of MLLMs are often presented as success stories that demonstrate broad generality. We take a different stance: applications should be viewed as *stress tests* that expose the strengths, limitations, and failures ofy current multimodal systems.

E.1 Vision-Language Applications

Tasks such as image captioning, visual QA, and image-grounded dialog are frequently used to showcase multimodal capability. However, strong performance on these tasks often reflects successful image–text alignment rather than genuine visual reasoning. Models can generate fluent and plausible descriptions while hallucinating objects, attributes, or spatial relations not supported by the image. Visual grounding and region-based reasoning tasks provide a more stringent test. Applications such as document understanding, chart interpretation, and human-object interaction analysis require precise alignment between visual regions and linguistic references. Empirical evidence suggests that even SOTA MLLMs struggle with compositional queries (e.g., nested spatial relations) (Johnson et al., 2017; Zellers et al., 2019) and exhibit brittle behavior under minor visual perturbations (Hendrycks and Dietterich, 2019; Geirhos et al., 2020). These failures point to limitations in fusion mechanisms rather than encoder quality alone.

E.2 Audio-Language Applications

Audio-based applications, such as speech recognition, audio captioning, and speech question answering, highlight a different set of challenges. Although transcription accuracy has improved substantially, higher-level audio understanding remains fragile. Models often fail to distinguish overlapping sound events (Drossos et al., 2020; Kim et al., 2019; Gemmeke et al., 2017), infer causality (Vosoughi et al., 2024), or reason about temporal structure beyond short windows (Fayek and Johnson, 2020). Applications such as audio summarization and sound event reasoning expose a mismatch between training objectives and deployment requirements. Contrastive and masked objectives encourage alignment at the segment level, but do not enforce temporal coherence or semantic abstraction. As a result, MLLMs frequently produce descriptions that are temporally inconsistent or semantically vague, especially in noisy or real-world acoustic environments.

E.3 Video-Language Applications

Video-language applications, including video question answering, event localization, and summarization, are widely regarded as benchmarks for multimodal reasoning. In practice, they reveal one of the most persistent limitations of current MLLMs: their weak long-horizon temporal reasoning. Many video-language models rely on sparse frame sampling or short clip aggregation, which suffices for coarse alignment but fails for causal inference and event ordering. Applications such as surveillance analysis, instructional video understanding, and long-form video summarization require maintaining a multimodal state over an extended period. Current architectures often collapse temporal information into static representations, resulting in inconsistent or contradictory outputs.

E.4 Interactive and Embodied Applications

The most explicit applications of MLLMs are in interactive and embodied settings, including robotics, assistive technologies, and human-AI collaboration, which demands continual perception, action-conditioned reasoning, and the ability to review beliefs over time. In such settings, current MLLMs exhibit compounding errors: misaligned perception propagates through fusion, and language-driven reasoning amplifies initial mistakes. The lack of explicit mechanisms for uncertainty modeling, memory, and continuous learning leads to fragile behavior that cannot be corrected through prompting or scale alone.

Taken together, these applications illustrate a central claim of this paper: many apparent successes of MLLMs are artifacts of evaluation regimes that reward correlation rather than understanding. Real-world applications expose the need for architectures and training objectives that explicitly support temporal reasoning, grounded fusion, and adaptive learning. We argue that future MLLMs must be designed *from the application constraints upward*, rather than retrofitted to them. Applications should

not merely demonstrate model capabilities; they should actively shape the architectural and training decisions that define the next generation of multimodal systems.

F Current Challenges

Despite rapid progress, current MLLMs fall short of general multimodal intelligence. We highlight five fundamental challenges.

Data Efficiency and Imperfect Alignment. High-quality multimodal data remains a bottleneck. While image-text pairs are abundant, video-language and higher-order datasets are scarce, noisy, and temporally misaligned (Ko et al., 2022; Jia et al., 2021b; Miech et al., 2019). Existing architectures struggle to reason under partial alignment, leading to brittle grounding (Rahate et al., 2022; Song et al., 2024). Self-supervised objectives (Zong et al., 2024) improve robustness but capture global rather than fine-grained consistency.

Temporal and Cross-Modal Reasoning. MLLMs perform well on recognition tasks but struggle with temporal reasoning, causal inference, and multi-step cross-modal reasoning (Liu et al., 2024b; Fu et al., 2025; Xiao et al., 2021). These failures stem from architectures that collapse time into static embeddings. Neuro-symbolic approaches (Yi* et al., 2020; Choi et al., 2024; Yi et al., 2018) offer promise but remain difficult to integrate with large-scale pretraining.

Efficient Scaling. Computational costs raise concerns about scalability and accessibility (Patterson et al., 2021; Shukor et al., 2025; Strubell et al., 2019). Techniques such as PEFT (Mu et al., 2024; Zhou et al., 2024), adapter-based transfer (Pfeiffer et al., 2021; Sung et al., 2022), and mixture-of-experts (Cai et al., 2025) help but often trade grounding fidelity for efficiency.

Evaluation Gaps. Existing benchmarks fail to capture multimodal coherence, causal reasoning, and long-horizon understanding. Benchmark contamination (DeepLearning.AI, 2024) undermines reliability, and many benchmarks reward fluent generation over grounded correctness. We argue for adversarial, diagnostic evaluation that probes AFR-identified failure modes (Table 1).

Interpretability and Ethics. Current models offer limited cross-modal attribution (Dang et al., 2024), making it difficult to determine whether outputs are driven by perceptual evidence or linguistic bias. Training on web-scale data introduces ethical risks, amplifies biases (Luo et al., 2024; Buolamwini and Gebru, 2018; Janghorbani and De Melo, 2023) with privacy concerns. Progress in fairness-aware training (Zeng et al., 2024) and bias detection (Raj et al., 2024) remains decoupled from core model design.