1. A LANAVLM: A Multimodal Embodied AI Foundation Model for Egocentric Video Understanding
2. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI
3. Generative AI in Multimodal User Interfaces: Trends, Challenges, and Cross-Platform Adaptability
4. Multimodal Generative AI for Precision Health

5. Generative Multimodal Models are In-Context Learners

6. A Survey on Evaluation of Multimodal Large Language Models

7. Generating Images with Multimodal Language Models

8. Achieving multi-modal brain disease diagnosis performance using only single-modal images through generative AI

9. Towards artificial general intelligence via a multimodal foundation model

10. SEED-Bench: Benchmarking Multimodal Large Language Models

11. Beyond Text-to-Text: An Overview of Multimodal and Generative Artificial Intelligence for Education Using Topic Modeling

12. DesignPrompt: Using Multimodal Interaction for Design Exploration with Generative AI

13. A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks

14. A Survey on Multimodal Benchmarks: In the Era of Large AI Models

15. Multi-Modal Generative AI: Multi-modal LLM, Diffusion and Beyond

16. Multimodal Foundation Models: From Specialists to General-Purpose Assistants

**Chapter 1**

**Literature Review**

**1.1 Related Work:**

**1. A LANAVLM: A Multimodal Embodied AI Foundation Model for Egocentric Video Understanding**

### Introduction

Embodied AI systems are designed to process and respond to environmental stimuli in ways that mimic human cognition and interaction. Such systems must interpret complex visual-spatial data, anticipate actions, and effectively communicate, often using nonverbal cues. This understanding is critical for tasks requiring seamless collaboration between humans and AI in domains such as robotics, wearable computing, augmented reality (AR), and virtual reality (VR).

Traditional Vision-Language Models (VLMs) have focused predominantly on third-person video data, limiting their applicability to tasks requiring an egocentric (first-person) perspective. Egocentric understanding is crucial for applications like assistive devices, such as wearable cameras or augmented smart glasses, designed to help users interpret and interact with their immediate environments.

To address this gap, this review identifies three key contributions:

1. The Egocentric Video Understanding Dataset (EVUD), specifically curated for training VLMs on video captioning and question-answering tasks unique to egocentric video data, enhances VLM training with diverse data, including Ego4D VQA (1,137 QA pairs), Ego4D VQA Gemini (96,523 QA pairs across seven categories), VSR (7,680 QA pairs), EgoClip Captioning (7,000 video-caption pairs), and HM3D Captioning (3,475 simulator-based video-caption pairs)
2. The development of A LANA VLM, a 7B parameter model trained using parameter-efficient methods that preserve the capabilities of pre-trained knowledge and to add visual understanding,used VSR (Visual Spatial Reasoning).
3. A LANAVLM achieved state-of-the-art performance on the OpenEQA benchmark, surpassing open-source models like Socratic GPT-4 by 3.6% and matching proprietary models like GPT-4V in spatial reasoning. It excelled in object recognition and functional reasoning but faced challenges in spatial reasoning and egocentric alignment.

**Challenges**

VLMs often lack egocentric understanding, crucial for real-world applications, and struggle with retaining pre-trained knowledge during fine-tuning. Limited and imbalanced datasets, such as EVUD, hinder diversity and quality. Spatial reasoning remains a challenge, with models prone to hallucinations and weak object relationships. While parameter-efficient methods like LoRa reduce costs, they may limit full data utilization. Ethical concerns, including privacy, consent, and cultural bias in datasets, further complicate development

### Limitations

LoRa fine-tuning restricts the use of datasets like EVUD, while visual encoder errors hinder attribute recognition. Alignment issues emerge due to insufficient egocentric annotations, and low-quality HM3D videos further degrade performance.

**Future Direction**

VLMs should focus on improving egocentric video analysis, spatial reasoning, and evaluation frameworks. Expanding diverse real-world datasets and reducing reliance on synthetic data will boost model performance. Efficient training methods like LoRa, along with ethical guidelines on privacy and fairness, are essential for responsible, effective deployment.

**References**

**[1]** Mark Johnson. 2015. Embodied understanding. Fron-tiers in psychology,6

**2. MMMU: A Massive Multi-discipline Multimodal**

**Understanding and Reasoning Benchmark for Expert AGI**

#### 1. Introduction to Multimodal Models (LMMs) and Language Models (LLMs)

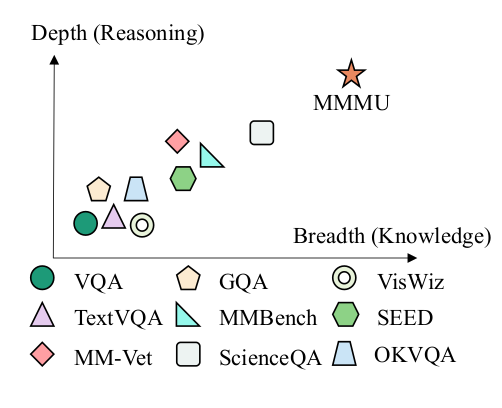
Multimodal models (LMMs) and large language models (LLMs) have made substantial progress in recent years, enabling systems to process and generate information across different modalities, such as text, images, and video. For evaluting performance of these models, a new benchmark MMMU designeon massive multi-discipline task.

* + MMMU is a comprehensive multimodal benchmark for expert-level tasks, consisting of 11.5K questions across 6 disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering.
  + The questions involve highly heterogeneous image types such as charts, diagrams, medical scans, music sheets, and more, and require advanced reasoning.

Unlike current benchmarks, which focus on either text or images, MMLU introduces two key challenges: diverse image formats, testing the perceptual abilities of LMMs, and interleaved text-image inputs, requiring models to reason across both modalities. These challenges highlight the need for MMLU to evaluate multimodal models effectively.

**2. Comparisons with Existing Benchmarks and Evalution**

The MMMU benchmark differs from existing ones in breadth and depth. It covers 30 subjects with 30 image formats, including diagrams and medical images, compared to the limited scope of prior benchmarks focused on daily knowledge. While existing benchmarks rely on commonsense or simple reasoning, MMMU requires deliberate reasoning with college-level subject knowledge, making it more challenging and comprehensive for evaluating multimodal models. (See Figure 1).

****

Figure

The MMMU benchmark evaluates various Large Multimodal Models (LMMs) and Large Language Models (LLMs) in a zero-shot setting to assess their ability to generate accurate answers without fine-tuning. The results reveal that GPT-4V outperforms other models, demonstrating significant capability in multimodal understanding. However, open-source models like LLaVA-1.5 and BLIP-2 lag behind proprietary models, with a maximum accuracy of 34%, showing considerable room for improvement. The benchmark highlights a performance gap in more complex domains like Health & Medicine, emphasizing the need for models capable of intricate reasoning with visual and textual data. Additionally, OCR and captioning enhancements do not yield substantial improvements for text-only models, suggesting that multimodal integration is crucial for success.

The error analysis of the MMMU benchmark identifies key challenges in multimodal understanding and reasoning. First, the interplay of language and vision shows that while language can help clarify visual information, it can also lead to hallucinations. Second, tasks involving grounding or referring to specific visual elements remain difficult for models, even advanced ones like GPT-4V. Lastly, complex reasoning tasks, especially those requiring long chains of thought or detailed calculations, continue to be problematic for current models. These insights highlight areas for further research in visual perception, knowledge representation, and multimodal integration.

**3. Generative AI in Multimodal User Interfaces: Trends, Challenges, and Cross-Platform Adaptability**

**Abstract:**

The paper focuses on Generative AI's role in transforming multimodal user interfaces, emphasizing cross-platform adaptability and personalized interactions. It addresses the "interface dilemma" regarding effective interactions with multimodal large language models (LLMs) and evaluates lightweight frameworks for mobile platforms. The research encompasses technical challenges, privacy concerns, and future directions in adaptive user interfaces.

**Introduction**

The introduction effectively establishes the context by highlighting the evolution of user interfaces and the transformative potential of Generative AI, particularly through multimodal LLMs. The authors present a clear problem statement centered on the "Interface Dilemma" - the challenge of determining optimal interfaces for human-AI interaction despite advances in multimodal capabilities.

**Key Points:**

- Historical development of user interfaces

- Impact of Generative AI on UI transformation

- Critical questions about ideal interface design

- Integration challenges across platforms

Based on the provided text, I can help develop a more detailed methodology section for your literature review. Here's a comprehensive analysis of the methodology used in this paper:

**Methodology**

The paper employs a systematic review methodology focusing on the historical evolution, current frameworks, and future directions of AI-driven multimodal user interfaces. The research approach can be broken down into several key components:

**1. Historical Analysis Approach**

- Systematic examination of UI evolution from early text-based interfaces to modern multimodal systems

- Chronological categorization of interface developments across different eras (1960s-2020s)

- Comparative analysis of interaction modalities across different time periods

**2. Framework Analysis**

The methodology includes detailed examination of:

- Current tech stacks and their integration with AI systems

- Server-client communication models

- Cross-platform frameworks (e.g., React Native, Flutter)

- Cloud-based AI services integration (AWS SageMaker, Google AI Platform, Microsoft Azure)

**3. System Architecture Analysis**

The research utilizes a structured approach to analyze:

- Input collection mechanisms (text, voice, video)

- Local processing systems (preprocessing, feature extraction)

- Cloud processing components (LLM inference, model updates)

- Output generation systems

- Context storage and retrieval mechanisms

**4. Comparative Assessment**

The methodology incorporates:

- Evaluation of different interface types (console-based, GUI, voice-based, immersive)

- Analysis of hardware considerations, particularly for mobile devices

- Assessment of lightweight frameworks for multimodal systems

- Comparison of processing capabilities across platforms

**5. Technical Performance Analysis**

The study examines:

- Processing power requirements

- Memory management techniques

- Energy efficiency considerations

- Model optimization approaches

- Hardware acceleration methods (e.g., Neural Processing Units)

**6. Data Processing Framework**

The research methodology includes analysis of:

- Input preprocessing techniques

- Feature extraction methods

- Context retention mechanisms

- Response generation systems

- Multimodal LLM processing workflows

**Limitations and Challenges**

**1. Technical Constraints:**

- Real-time performance issues with multimodal inputs

- Hardware limitations on mobile devices

- Processing power and memory constraints

- Latency challenges in cloud-based solutions

**2. Ethical Considerations:**

- Data privacy concerns

- Transparency issues in AI decision-making

- Potential biases in AI responses

- User trust challenges

**Future Directions**

The paper outlines several promising future developments:

1. AI-Automated Interfaces:

- Dynamic adaptation to user behavior

- Context-aware UI adjustments

- Real-time personalization

2. New Modalities:

- Brain-computer interfaces

- Gesture recognition

- Haptic feedback integration

3. Innovation Areas:

- Emotionally adaptive interfaces

- Privacy-conscious design frameworks

- Cross-platform AR UIs

- Real-time collaborative systems

**Conclusion**

The research effectively synthesizes current developments in Generative AI for multimodal UIs while acknowledging existing challenges and future opportunities. It emphasizes the need for balanced solutions that address technical constraints while maintaining user privacy and trust.

**Critical Analysis**

**Strengths:**

- Comprehensive coverage of current trends and challenges

- Strong focus on practical implementation issues

- Clear framework for future development

**Limitations:**

- Limited empirical data to support some claims

- Could benefit from more specific case studies

- Some technical details could be more thoroughly explored

4. Multimodal Generative AI for Precision Health

In this Paper, The Author mainly focused on Precision health that aims to deliver personalized care by using individual health data, allowing for targeted treatments and preventive measures. AI has become essential in advancing this approach by processing large amounts of health data, which is often multimodal (e.g., clinical notes, medical images, and test results). Early AI models, such as BERT, were limited to text data and required fine-tuning for specific tasks. However, newer models like GPT-4 have demonstrated greater versatility, handling both text and images to assist in areas like clinical documentation, diagnostics, and patient care.

A significant challenge in health AI is creating realistic benchmarks for evaluation. Many existing benchmarks do not capture the complexity of real-world healthcare applications. Metrics like BLEU and ROUGE are often inadequate for assessing the clinical relevance of AI-generated content. This gap in evaluation methods makes it difficult to measure the progress of health AI. Researchers are exploring the use of generative AI to improve this process by developing self-improving systems that can evaluate their own outputs, ensuring more accurate benchmarks.

Multimodal learning, where AI can process different types of data (such as images and text), is a growing field in health AI. Contrastive learning methods have been used successfully to align different data types, but the scarcity of multimodal biomedical data is a bottleneck. Projects like BiomedCLIP and LLaVA-Med are addressing this by using large-scale biomedical datasets, combining text and images to improve model training. These approaches show promise in areas like pathology and radiology.

Longitudinal health data, which tracks patient health over time, is key to predicting treatment outcomes and adverse events. However, it is often fragmented and hard to access. Some initiatives, such as Microsoft's digital pathology project, are working to build models that can analyze large volumes of health data to improve disease detection and treatment modeling. The ultimate goal is to use multimodal AI to unlock insights from population-level data, predicting health outcomes more accurately and improving overall care.

In conclusion, while progress has been made, challenges remain in scaling health AI. Addressing these challenges—such as improving model evaluation, expanding multimodal datasets, and integrating longitudinal data—will be crucial to realizing the full potential of precision health.

**5. Generative Multimodal Models are In-Context Learners**

Introduction

Multimodal tasks, which involve understanding and generating content across multiple modalities such as text, image, and video, have been a significant area of research in AI. These tasks can be highly diverse, making them difficult for traditional models to scale efficiently. Typically, multimodal systems require task-specific architectures and sizable supervised training sets, making it challenging to generalize across various tasks. However, humans excel at solving new tasks with minimal demonstrations or instructions—a capacity that current multimodal models are still working to replicate.

In this context, they introduce Emu2, a generative multimodal model with 37 billion parameters, trained on large-scale multimodal sequences using a unified autoregressive objective. Unlike previous approaches, Emu2 can handle diverse multimodal tasks in an in-context learning setting, enabling it to solve tasks it has never encountered during training. By predicting the next multimodal element (e.g., visual embeddings or textual tokens), Emu2 demonstrates strong multimodal reasoning abilities, even emerging as a solution for complex tasks that require real-time reasoning, such as visual prompting and object-grounded generation.

Emu2 achieves new milestones in several few-shot multimodal understanding tasks, setting a new record in settings where only a few examples or instructions are provided. Furthermore, when instruction-tuned for specific tasks, it achieves state-of-the-art results in question answering benchmarks and subject-driven generation, two complex challenges for large multimodal models. These advancements show that Emu2 not only excels in multimodal contexts but can also serve as a general-purpose interface for a wide range of multimodal tasks. The model’s code and pre-trained versions are publicly available to facilitate future research and development in the field.

### Advances in Multimodal Systems

1. Multimodal Understanding and Generation: The seminal work by Vaswani et al. (2017) on transformers provided the backbone for many multimodal systems, especially with models like BERT and GPT. Recent works, including CLIP (Radford et al., 2021) and DALL-E (Ramesh et al., 2021), have demonstrated the potential for models that are able to understand both vision and language. These systems use large-scale data and transformers to align vision-language features, showing the power of joint multimodal representation learning.
2. Few-Shot Learning: One of the breakthrough advancements has been the emergence of few-shot learning, where models generalize well from a small number of examples. Notable models such as GPT-3 (Brown et al., 2020) and Flamingo-80B (Alayrac et al., 2022) have pushed the envelope by demonstrating how large-scale pretrained models can perform well with limited task-specific fine-tuning. These models can handle diverse tasks, from vision-language understanding to simple reasoning and inference, much like Emu2's demonstration of few-shot capabilities on tasks like visual question answering (VQA), counting, and classification.
3. Visual Prompting and Object-Grounded Generation: As with BLIP-2 (Li et al., 2023) and Flamingo (Alayrac et al., 2022), Emu2 exhibits significant advancements in visual prompting and object-grounded generation. Visual prompting techniques, which involve providing visual cues (e.g., bounding boxes, highlights, etc.), have proven to enhance model performance on multimodal reasoning tasks. Emu2 takes this further by incorporating visual cues effectively even in zero-shot conditions, an important feature for generalizing across diverse multimodal tasks.
4. Instruction Tuning and Adaptability: Models such as ChatGPT (OpenAI, 2023) and PaLM (Chowdhery et al., 2022) have demonstrated significant progress in instruction tuning, where models are refined to follow specific instructions with minimal supervision. Emu2 also employs this technique, fine-tuning its base model on conversational data and multimodal tasks, resulting in Emu2-Chat, a robust multimodal dialogue system that can follow instructions for visual question answering and even perform image generation tasks.
5. Controllable Visual Generation: In the domain of controllable visual generation, systems such as Stable Diffusion (Rombach et al., 2022) and DALL-E 2 (Ramesh et al., 2022) have focused on generating high-quality images based on textual prompts. Emu2 extends this ability by conditioning its visual generation on text, locations, and images, and demonstrating high fidelity in subject-driven image generation, even excelling in complex scenarios like visual consistency across different settings.

### Challenges in Scaling Multimodal Models

While significant strides have been made, large multimodal models still face challenges. Despite strong performance on many benchmarks, Emu2 shares some common limitations with prior models:

* Data and Computational Requirements: Training such large models requires immense computational resources, which remains a barrier for many researchers and developers.
* Robustness and Bias: Like other large-scale models, Emu2 must be rigorously tested for biases and ethical concerns related to their deployment, as discussed in the conclusion of the paper.
* Task-Specific Weaknesses: Although Emu2 performs well across several benchmarks, certain complex tasks, particularly those involving nuanced language understanding or very specific visual reasoning, can still pose challenges for multimodal models.

### Future Directions and Societal Impact

The paper discusses the broader societal impact of Emu2 and its potential for misuse. As multimodal systems continue to grow, it is critical to ensure their responsible deployment, including mitigating risks related to misinformation, privacy, and accessibility. Future work may include enhancing model robustness, reducing reliance on vast amounts of training data, and improving energy efficiency in training large multimodal models.

### Conclusion

Emu2's contribution is a pivotal step towards creating scalable, task-agnostic, and instruction-following multimodal models. By embracing a unified autoregressive framework and significantly scaling up model size, Emu2 demonstrates new potential in multimodal in-context learning. This work sets a new benchmark for the field, positioning Emu2 as a base model for general-purpose multimodal tasks.

### Key Papers and Models to Include:

* Transformers for Multimodal Learning: Vaswani et al. (2017), Radford et al. (2021), Ramesh et al. (2021)
* Large Multimodal Models: Alayrac et al. (2022), Brown et al. (2020), Li et al. (2023)
* Controllable Generation: Rombach et al. (2022), Ramesh et al. (2022)

### **6. A Survey on Evaluation of Multimodal Large Language Models**

### 1. Introduction to Multimodal Large Language Models (MLLMs)

Multimodal Large Language Models (MLLMs) are models that integrate large language models (LLMs) with sensory modality encoders (e.g., vision, audio) to replicate human-like perception and reasoning. Unlike traditional LLMs that handle only text, MLLMs process multimodal inputs such as text, images, and audio, allowing them to understand and reason over complex multimodal data. The integration of modality encoders, such as Vision Transformers (ViTs) for images or Whisper for audio, with the powerful LLMs, helps MLLMs understand and interpret various data types, making them capable of reasoning and generating outputs based on multimodal contexts.

#### Key Aspects of MLLMs:

* Human-like Perception: MLLMs aim to replicate human sensory systems, such as vision and hearing, for better reasoning and problem-solving capabilities.
* Potential for AGI: MLLMs are seen as a potential step towards achieving Artificial General Intelligence (AGI), as they combine reasoning with sensory data processing.

### 2. Survey Focus and Contributions

This paper conducts a comprehensive review of the evaluation methods used for MLLMs, presenting a taxonomy of evaluation approaches and benchmarks. It focuses on three main aspects:

1. Taxonomy of Evaluation Methods: A classification of the existing methods based on the tasks and capabilities they assess.
2. Benchmarking of MLLMs: A discussion on the benchmarking of existing MLLMs across different datasets and their performance.
3. Challenges and Future Directions: Identifies challenges and promising research directions in MLLM evaluation.

The paper emphasizes that evaluation is critical in advancing the development of MLLMs, ensuring that these models are reliable, trustworthy, and capable in different domains.

### 3. MLLM Framework

MLLMs typically consist of the following components:

1. Large Language Models (LLMs): These process textual inputs. Transformer-based models like LLaMA are common for feature extraction.
2. Modality Encoders: These are responsible for processing data from other modalities such as images (Vision Transformers), audio (Whisper), video (temporal models), and 3D data (Point-BERT).
3. Modality Projector: This component aligns features from different modalities (text, image, audio) into a unified feature space, enabling better multimodal understanding.

### 4. MLLM Training Strategy

The training of MLLMs involves several stages:

1. Alignment Pre-training: Aligns different modalities (text, image, audio) by learning their relationships, often using large-scale, text-paired data (e.g., image captions).
2. Multimodal Instruction Tuning: Fine-tunes the model using language instructions to improve the model's ability to respond to multimodal inputs.
3. Alignment for Human Preference: Fine-tunes the model based on human feedback (e.g., Reinforcement Learning with Human Feedback - RLHF) to ensure that the outputs align with human expectations.

### 5. Evaluation Framework

#### What to Evaluate?

Evaluation of MLLMs can be categorized into several tasks based on the capabilities being assessed, such as:

* General Multimodal Recognition and Perception: Basic understanding tasks like object recognition, scene understanding, etc.
* Reasoning: Tasks that involve logical reasoning and decision-making across modalities.
* Trustworthiness: Ensuring that models do not produce biased, misleading, or harmful outputs, especially in sensitive domains (e.g., healthcare, autonomous driving).
* Domain-Specific Applications: Tasks specific to certain fields, like medical analysis, socioeconomic analysis, or AI agents.

#### Where to Evaluate?

Evaluation can be done using two types of benchmarks:

1. General Benchmarks: These assess broad capabilities like recognition, reasoning, and trustworthiness. Examples include:
   * MMBench: Tests basic recognition tasks.
   * MM-Vet: Assesses robustness and generalizability.
   * Seed-Bench: Evaluates contextual relevance and coherence of model outputs.
2. Specific Benchmarks: These focus on specialized tasks or domains, such as:
   * MathVerse: Tests mathematical reasoning capabilities.
   * ScienceQA: Focuses on science-related question answering tasks.
   * GMAI-MMBench: Evaluates medical analysis capabilities.
   * CVQA: Assesses cross-cultural visual question answering.

#### How to Evaluate?

Evaluation can be done through various methods:

1. Human Evaluation: This is essential for tasks requiring high comprehension and qualitative judgment (e.g., relevance, coherence, fluency).
2. GPT-4 Evaluation: Recent studies use GPT-4 for efficient evaluation, leveraging its ability to assess outputs on several dimensions like relevance, helpfulness, and accuracy.
3. Metric Evaluation: Traditional metrics like Accuracy, Precision, BLEU, ROUGE are used for quantitative assessments across tasks such as recognition, text generation, and multimodal interaction.

### 6. Key Challenges and Future Directions

Despite the progress in MLLM development, several challenges persist:

* Complexity in Evaluation: Evaluating multimodal models involves understanding interactions across multiple modalities, which complicates performance assessment.
* Trustworthiness Issues: Models may generate hallucinated or biased content, especially in critical applications such as healthcare or law.
* Data and Benchmark Limitations: Existing datasets and benchmarks may not fully capture all the complexities of real-world multimodal tasks.

Promising future directions include:

* Improving Evaluation Metrics: Developing more precise metrics for assessing multimodal reasoning and trustworthiness.
* Domain-Specific Benchmark Development: More benchmarks are needed for specialized domains like 3D point cloud analysis, remote sensing, and social sciences.
* Collaboration for AGI Progress: Ongoing efforts to bridge gaps in evaluation practices to help achieve AGI through MLLMs.

### 7. Conclusion

MLLMs are promising for advancing AI systems toward more human-like capabilities, but rigorous evaluation is necessary to ensure they perform reliably across diverse tasks and domains. This paper provides a thorough review of the evaluation frameworks, tasks, and benchmarks for MLLMs, offering valuable insights into their current capabilities and future development. Effective evaluation methods are crucial for enhancing MLLMs' effectiveness, ethical standards, and applicability in real-world scenarios.

7. Generating Images with Multimodal Language Models

The paper introduces GILL (Generating Images with Large Language Models), a method that combines a frozen text-only LLM (Large Language Model) with pre-trained image encoder-decoder models to generate and retrieve images based on text and image inputs. The model focuses on three main tasks: image retrieval, novel image generation, and multimodal dialogue.

### Key points:

1. Multimodal Capabilities: GILL processes interleaved image-and-text inputs and generates both images and text, as well as retrieves images. It outperforms prior models in tasks requiring long and complex language, demonstrating improved context sensitivity.
2. LLM-to-Image Mapping: The model leverages a GILLMapper transformer module to map LLM embeddings to a visual model's embedding space, enabling it to generate images without requiring full retraining of the image generation model.
3. Image Retrieval & Generation: GILL decides whether to retrieve an image or generate a new one at inference time, based on the input context, using a learned decision model.

### Methodology:

* Training: The model is trained with a small set of parameters fine-tuned on image-caption pairs, using datasets like Conceptual Captions (3.3M pairs). It uses a combination of captioning, image token prediction, and retrieval losses.
* Image Generation: GILL uses Stable Diffusion as its image generation backbone. It processes text and image sequences, enabling the generation of images conditioned on complex multimodal inputs.
* Decision Model: At inference, GILL uses a linear classifier to decide between generating new images or retrieving existing ones based on the context provided.

### Experiments:

1. Image Generation: GILL outperforms Stable Diffusion (SD) when processing long text inputs and multimodal context (VIST and VisDial datasets). It can generate images from both single captions and full multimodal context, including images and text interwoven.
2. Image Retrieval: GILL competes favorably with prior retrieval models, showing its versatility in handling both generation and retrieval tasks.
3. Performance: GILL excels in processing dialogue and longer text, outperforming non-LLM generation models in multimodal tasks.

### Evaluation Metrics:

* CLIP Similarity: Measures how similar generated images are to real ones.
* LPIPS: Assesses perceptual similarity between images.

### Limitations:

* Hallucinations and Repetitions: GILL inherits issues common to LLMs, like generating irrelevant or repetitive content.
* Visual Model Constraints: The model is limited by the visual representation's capacity (only four visual tokens).
* Context Sensitivity: GILL performs better with longer input contexts but has some room for improvement in terms of fine-grained visual detail.

### Future Work:

* Scalability: Scaling up to larger models or datasets, or improving fine-tuning and data curation, can enhance performance.
* Bias & Safety: The model could inherit biases from pretraining, and better data curation could reduce these risks.

### Impact:

* AI Assistants: GILL could be used for advanced AI assistants that generate both images and text.
* Disinformation Risk: It could be used maliciously to generate misleading images, necessitating mitigation strategies.

### Conclusion:

GILL is an efficient and effective approach to combining LLMs and image generation models, outperforming existing systems in tasks requiring complex multimodal inputs. It offers a scalable solution for integrating image and text generation, with potential for further improvements in accuracy and efficiency.

8. Achieving multi-modal brain disease diagnosis performance using only single-modal images through generative AI

Brain disease diagnosis using medical imaging is a crucial aspect of clinical practice, with techniques like MRI widely employed for disease classification. However, due to challenges such as cost, radiation risk, and accessibility, multi-modal imaging is often not feasible in clinical settings. As a result, there is growing interest in methods that can synthesize information from available modalities to enhance classification accuracy, potentially mimicking multi-modal performance using single-modal data.

1. Multi-Modal Imaging for Brain Disease Diagnosis: Medical imaging techniques such as MRI, PET, and FLAIR are widely used to diagnose diseases like Alzheimer’s disease (AD), vascular cognitive impairment, and glioblastoma. Different imaging modalities offer complementary information, capturing both anatomical and functional aspects of the brain. However, the integration of these modalities is not always feasible in clinical settings due to logistical constraints such as cost or radiation exposure. Despite this, studies have shown that utilizing multiple modalities typically improves disease classification performance compared to relying on a single modality.

2. Image Synthesis for Multi-Modal Integration: The synthesis of missing modalities based on available ones has emerged as a viable approach for enhancing classification performance. Generative models like Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and more recently, diffusion models, are increasingly being explored for modality synthesis in medical image classification tasks. However, these models often focus on generating perceptually high-quality images or ensuring diversity in natural images, which may not guarantee clinical reliability for medical purposes.

3. Challenges in Generating Reliable Medical Images: Generating medical images, especially when they are not available, is a complex and ill-posed problem. Dense voxel-wise synthesis requires high computational resources and substantial amounts of data, and may result in inefficient learning. Furthermore, such models may fail to synthesize clinically relevant features effectively, which is a key challenge in improving medical diagnoses. Some studies attempt to address these challenges by introducing feature-level synthesis rather than voxel-wise synthesis to reduce complexity and enhance efficiency. However, ensuring the clinical relevance and accuracy of these synthesized features remains a challenge.

4. Uncertainty in Classification: In addition to improving classification accuracy, quantifying the uncertainty of predictions is a critical aspect of trustworthy AI in medical diagnosis. Uncertainty-aware models help assess the reliability of predictions, providing clinicians with confidence in the diagnosis. These models are broadly categorized into Bayesian and non-Bayesian approaches. Bayesian approaches estimate the posterior distribution of network weights and provide uncertainty by approximating predictive distributions. Non-Bayesian approaches, such as ensemble methods, provide uncertainty by estimating it directly from the model’s outputs. Recent advancements in non-Bayesian methods, like the multiview classification network based on variational Dirichlet and evidence-level fusion, have shown promise in improving uncertainty estimation for medical applications.

5. Integration of Synthesis and Uncertainty: Recent work has focused on combining uncertainty estimation with modality synthesis. This approach uses deep learning to synthesize disease-relevant features and estimates the uncertainty of predictions using evidential learning. The integration of these two techniques allows for more accurate and trustworthy classification, as it not only provides disease classification but also informs clinicians of the confidence in the diagnosis. Dempster-Shafer theory (DST) has been applied to combine belief masses from synthesized modalities and their uncertainties, providing a framework for reliable classification decisions.

6. Comparative Studies and Performance Evaluation: Several studies have compared the performance of single-modal and multi-modal models for brain disease classification. While multi-modal models generally outperform single-modal models, the proposed synthesis-empowered methods show significant promise. For instance, frameworks that synthesize disease-relevant features from available modalities achieve performance similar to multi-modal systems, even with single-modal input. This is particularly useful in clinical settings where multiple imaging modalities may not be available. In comparison to other models like DeepGuide and DSNet, the proposed frameworks exhibit significant improvements in classification metrics, including AUC and F1-scores, highlighting their potential for real-world medical applications.

7. Evaluation Across Datasets: The performance of synthesis-based frameworks has been evaluated on multiple datasets, including those for Alzheimer’s disease, subcortical vascular mild cognitive impairment (svMCI), and glioblastoma. The results demonstrate that these frameworks can consistently provide performance close to multi-modal systems. Additionally, the uncertainty estimation provided by these frameworks allows for better assessment of classification reliability, with predictions associated with lower uncertainty generally being more accurate. The ability to synthesize features from different modalities while also estimating uncertainty is a key factor in improving classification robustness.

8. Future Directions and Limitations: While the synthesis-empowered frameworks show promise, there are still limitations. The synthesis ability is highly dependent on the quality of the generative model used, and performance may degrade if the synthesized features do not align well with real ones. The development of advanced synthesis methods, such as diffusion models, could improve the synthesis quality. Additionally, the framework currently focuses on disease-specific models, though a unified model that can handle multiple diseases could further improve performance by exploiting common latent spaces between related diseases. Finally, uncertainties across different datasets may have different dynamic ranges, which could complicate cross-center comparisons. A framework designed to standardize uncertainty across datasets would be beneficial for broader clinical deployment.

Conclusion: In summary, the integration of modality synthesis and uncertainty-aware classification holds great potential for improving brain disease diagnosis using single-modal data. By leveraging the generative capabilities of deep learning and employing evidential learning for uncertainty estimation, these frameworks offer a reliable and efficient way to enhance classification performance in clinical settings. However, further advancements in synthesis methods and uncertainty standardization are needed to fully realize their potential across diverse clinical environments.

**9.** **Towards artificial general intelligence via a multimodal foundation model**

The study explores whether a large-scale, multimodal foundation model can make progress toward achieving Artificial General Intelligence (AGI). Specifically, it aims to:

1. Build a foundation model that mimics human cognitive processes across multiple tasks and domains.
2. Develop a pre-training strategy based on self-supervised learning with weakly correlated multimodal data.
3. Demonstrate the model's adaptability, imagination, and generalization capabilities in various cognitive and cross-modal tasks.

**Key Findings and Conclusions**

1. **Model Performance**:
   * The proposed model, **BriVL (Bridging Vision-and-Language)**, achieved strong performance on cross-modal tasks (e.g., image-text retrieval) and domain-specific tasks (e.g., news classification, remote sensing classification).
   * BriVL exhibits a robust imagination ability, effectively creating coherent representations of abstract concepts and unseen scenarios.
2. **Multimodal Pre-training**:
   * Using 650 million weakly correlated image-text pairs collected from the web, the model outperformed alternatives like CLIP and ALIGN by relying on a broader and more diverse dataset.
3. **Efficient Architecture**:
   * The two-tower model design (separate encoders for image and text) enhanced efficiency and scalability, reducing GPU memory requirements compared to single-tower approaches.
4. **Imagination Capability**:
   * BriVL demonstrated a unique ability to imagine and generalize across abstract concepts, rare scenarios, and even multiple languages, marking a critical step toward AGI.

**Research Methodology**

1. **Data Collection**:
   * The model was pre-trained on 650 million weakly correlated image-text pairs collected from diverse web sources (e.g., social media, news, encyclopedias).
2. **Architecture**:
   * **Two-Tower Model**: Separate encoders for image and text, combined with momentum encoders for maintaining a dynamic queue of negative samples during contrastive learning.
   * Multi-Scale Patch Pooling (MSPP): Enabled efficient and fine-grained feature extraction without the need for object detectors.
3. **Training**:
   * The model used distributed training with 112 NVIDIA A100 GPUs over 10 days.
4. **Evaluation**:
   * The model was tested on multiple tasks such as cross-modal retrieval, visual question answering, and single-modal classification, using both zero-shot and fine-tuned approaches.

**Significant Gaps or Limitations Identified**

1. **Bias and Prejudices**:
   * The reliance on web-crawled data introduces a risk of embedding societal biases and stereotypes into the model.
2. **Limited Modalities**:
   * The study focused on vision and language; incorporating other modalities (e.g., audio, video) could further enhance AGI capabilities.
3. **Fine-Tuning Challenges**:
   * Finding optimal strategies for task-specific fine-tuning remains a challenge, as different downstream tasks benefit from distinct approaches.

**Contribution to the Field**

The study marks a transformative step toward AGI by:

1. Proposing a scalable and efficient pre-training approach for multimodal models.
2. Demonstrating how weak semantic correlation data can enhance cognitive abilities like imagination and generalization.
3. Establishing BriVL as a versatile model with applications across various AI+ fields, such as healthcare, neuroscience, and biomedicine.

**Practical Implications**

1. **Real-World Applications**:
   * The model can be adapted to specific tasks like medical imaging, natural disaster prediction, and cross-domain content creation.
2. **Future Research**:
   * Researchers can build upon BriVL's GPU-efficient framework to develop larger, more versatile models involving additional modalities.
3. **Ethical Considerations**:
   * The study underscores the need to address ethical risks like content manipulation and biased decision-making in future AGI systems.

**10. SEED-Bench: Benchmarking Multimodal Large Language Models**

**Introduction**

Multimodal Large Language Models (MLLMs) are advanced AI systems that can process and generate both text and images. However, existing evaluation methods mainly assess how well these models understand a single image-text pair. This limits our ability to measure their full potential. SEED-Bench is introduced to overcome these limitations by evaluating MLLMs across different levels of capability.

The paper classifies MLLM capabilities into **five hierarchical levels (L0 to L4)**, where each higher level includes all the capabilities of the levels below it. SEED-Bench focuses on evaluating models up to **L3**, assessing their ability to understand and generate both text and images.

**Related Work**

Several benchmarks have been developed to evaluate MLLMs, but they primarily focus on **single-image-text comprehension**. Some well-known benchmarks include:

* **LLaVA-Bench** – Evaluates MLLMs using free-form text responses.
* **MMBench** – Uses multiple-choice questions to improve evaluation objectivity.
* **LVLM-eHub** – Compares model responses through human judgment.

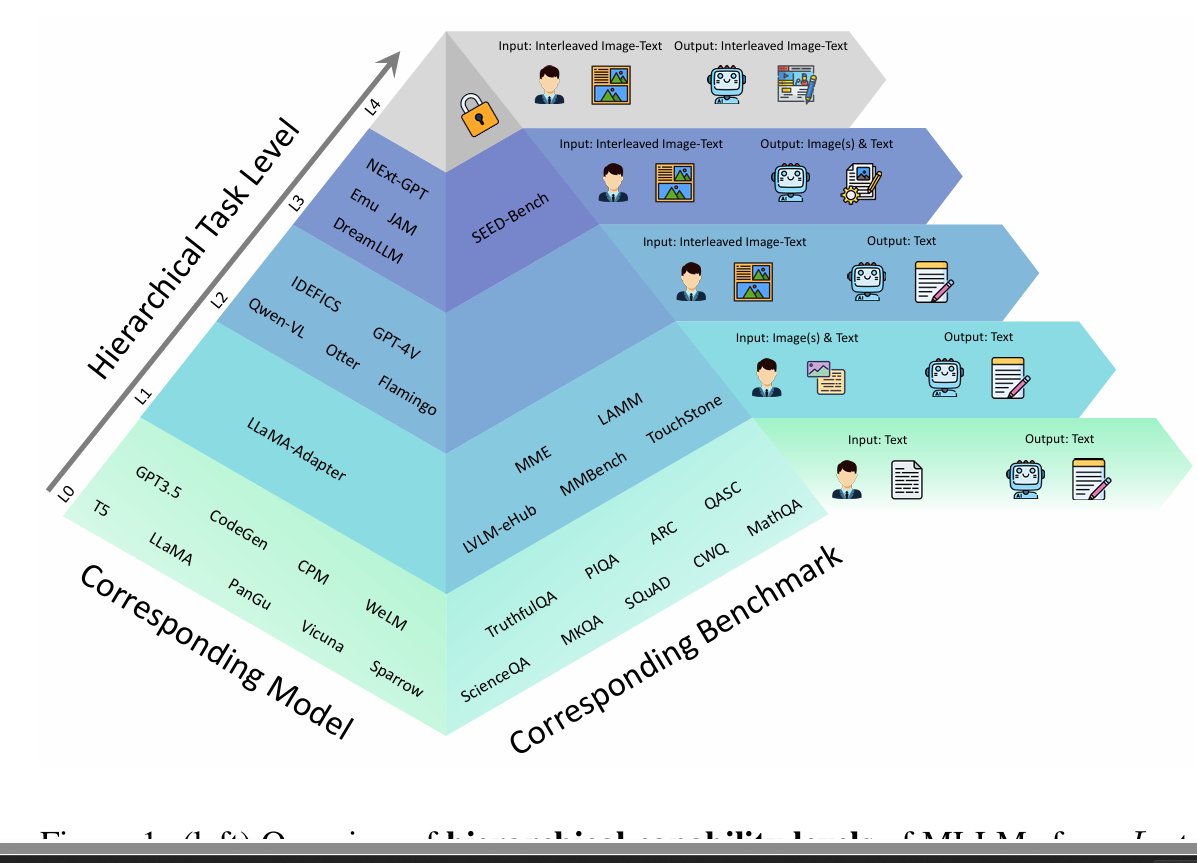
However, these benchmarks have limitations, such as relying on human annotators or lacking a clear framework for assessing multimodal outputs. SEED-Bench addresses these gaps by introducing **hierarchical evaluation levels**.

**Hierarchical Capability Levels**

SEED-Bench defines **five levels (L0 to L4)** of MLLM capabilities:

* **L0:** The ability to generate text from text input (like traditional language models).
* **L1:** Understanding structured image-text inputs (basic image comprehension).
* **L2:** Understanding complex interleaved image-text inputs (real-world-like multimodal content).
* **L3:** Generating both text and images as output.
* **L4:** Producing open-ended interleaved image-text responses, aiming towards **General AI**.

The SEED-Bench-2 benchmark focuses on evaluating models **up to L3**.



[Figure 1: Diagram: Hierarchical Capability Levels]

The diagram illustrates how higher levels include the capabilities of lower ones. SEED-Bench-2 tests MLLMs up to **L3**, meaning it assesses their ability to comprehend multimodal inputs and generate text and images.

**Evaluation Dimensions**

SEED-Bench evaluates MLLMs based on **27 dimensions** divided into three parts:

**Part 1 (L1 - Image and Text Comprehension)**

This level tests whether a model can understand structured image-text data. It includes:

* **Scene Understanding** (recognizing objects in an image)
* **Instance Identification** (identifying specific objects)
* **Spatial Relationships** (understanding object locations)
* **Visual Reasoning** (interpreting visual content)
* **Text Recognition** (reading text within images)
* **Chart and Landmark Recognition**
* **Emotion Recognition** (detecting emotions in images)

**Part 2 (L2 - Interleaved Image-Text Comprehension)**

At this level, MLLMs must process **complex, interleaved image-text data** similar to real-world applications. It assesses:

* **In-Context Captioning** (generating a description of an image)
* **Interleaved Image-Text Analysis** (answering questions using mixed image-text inputs)

**Part 3 (L3 - Image and Text Generation)**

This level tests whether an MLLM can generate images in addition to text:

* **Text-to-Image Generation** (creating an image based on a text prompt)
* **Next Image Prediction** (predicting a future image)
* **Text-Image Creation** (answering a question with both text and a generated image)

**Construction of Multiple-Choice Questions**

SEED-Bench includes **24,000 multiple-choice questions** created using:

1. **Automatic pipelines** – AI models generate and verify questions.
2. **Existing datasets** – Reformatting previous datasets into multiple-choice format.
3. **Human-GPT collaboration** – Human experts create questions with GPT assistance.

Unlike other benchmarks that use human judges or free-form text answers, SEED-Bench provides multiple-choice questions with **definitive correct answers**, improving evaluation accuracy.

**Evaluation Strategy**

SEED-Bench evaluates model outputs through:

* **Text Output Ranking** – Checking which answer choice a model finds most likely.
* **Image Output Matching** – Using **CLIP similarity scores** to compare generated images with answer choices.
* **Text-Image Evaluation** – Models must generate both the correct text and image for full credit.

This ensures a **fair and automated evaluation**, reducing human bias.

**Evaluation Results**

**Models Evaluated**

The paper tests **22 open-source MLLMs**, including:

* **BLIP-2, InstructBLIP, LLaVA, MiniGPT-4, Qwen-VL-Chat, Emu, NExt-GPT** and others.

**Key Findings**

1. **MLLMs Struggle at Even Basic Levels (L1).**
   * The best models only achieve **~60% accuracy** on structured image-text comprehension.
   * Models struggle with **charts, visual mathematics, and complex spatial reasoning**.
2. **Interleaved Image-Text Comprehension (L2) is Harder.**
   * Most models perform worse at **L2** than at L1.
   * This suggests that **real-world multimodal understanding remains a challenge**.
3. **Few MLLMs Can Generate Images (L3).**
   * Only **two** models could generate both text and images.
   * A **universal model** that excels in both tasks **is still underdeveloped**.
4. **Trade-off Between Comprehension and Generation.**
   * **NExt-GPT**, one of the few models reaching L3, performed **poorly** on L1 comprehension.
   * **Adding image generation often weakens overall model accuracy**.

**Observations**

* **No MLLM has achieved perfect L1 comprehension yet.**
* **Interleaved image-text comprehension (L2) is still an unsolved problem.**
* **Text and image generation (L3) needs more exploration.**
* **Current MLLMs either excel at comprehension or generation, but not both.**

SEED-Bench provides a **clear roadmap for improving MLLMs**, highlighting the areas that need more research.

**Final thought**

SEED-Bench is the **first comprehensive benchmark** that systematically evaluates MLLMs across **multiple hierarchical levels**. The study reveals **critical gaps** in existing models and sets the foundation for future research.

The authors plan to **launch a public leaderboard**, allowing researchers to continuously evaluate new MLLMs.

**Conclusion:**

This literature review simplifies the core ideas of SEED-Bench. It explains why the benchmark is **important**, how it works, and what it reveals about MLLMs. The study shows that while MLLMs have made progress, they still **struggle with multimodal reasoning and image generation**. SEED-Bench will play a crucial role in **guiding AI development toward General AI.**

-----------------or--------------------------------------------------------------------------

**Detailed Literature Review of the SEED-Bench Research Paper**

**1. Introduction** The SEED-Bench paper introduces a comprehensive benchmark for evaluating Multimodal Large Language Models (MLLMs). MLLMs integrate the capabilities of traditional large language models (LLMs) with multimodal inputs and outputs, encompassing text and images. Existing benchmarks primarily assess single-image text comprehension, failing to address the full spectrum of multimodal abilities. SEED-Bench bridges this gap by categorizing MLLM capabilities hierarchically from Level 0 (L0) to Level 4 (L4). The paper highlights the importance of achieving L4 capabilities, which signify advanced interleaved image-text comprehension and generation—a critical step towards general artificial intelligence (AGI).

**2. Related Work** The research contextualizes SEED-Bench by comparing it to existing benchmarks like LLaVA-Bench, MMBench, and Touchstone. These benchmarks focus on isolated tasks like visual question answering (VQA) or image captioning. In contrast, SEED-Bench evaluates MLLMs across 27 diverse dimensions, emphasizing hierarchical progression. By combining visual and textual modalities, SEED-Bench sets a new standard for comprehensive evaluation.

**3. SEED-Bench Framework**

**3.1 Hierarchical Capability Levels**

SEED-Bench categorizes MLLM capabilities into five levels:

* **L0:** Basic text generation from text inputs (traditional LLM capability).
* **L1:** Comprehension of fixed-format multimodal inputs (e.g., single image-text pairs).
* **L2:** Understanding open-form interleaved image-text data (real-world multimodal scenarios).
* **L3:** Generating both text and images as outputs.
* **L4:** Open-form interleaved image-text input and output—a future goal for AGI. SEED-Bench currently evaluates up to L3, showcasing hierarchical task complexity.

**3.2 Evaluation Dimensions**

SEED-Bench spans 27 evaluation dimensions, divided into three parts:

* **Part 1 (L1):** Fixed-format multimodal comprehension, including tasks like scene understanding, visual reasoning, and text recognition.
* **Part 2 (L2):** Interleaved image-text analysis and in-context captioning, reflecting real-world data complexities.
* **Part 3 (L3):** Image and text generation, including tasks like text-to-image generation and text-image creation.

**3.3 Construction of Multiple-Choice Questions**

The benchmark includes 24,000 human-annotated multiple-choice questions, designed through three approaches:

1. **Automated Pipelines:** Leveraging foundation models for visual information extraction and question generation.
2. **Adapted Datasets:** Modifying existing datasets to fit the multiple-choice format.
3. **Human and GPT Collaboration:** Crafting questions for dimensions without suitable datasets. These methods ensure high-quality, objective, and scalable evaluation.

**3.4 Evaluation Strategy**

To objectively assess model outputs, SEED-Bench employs:

* **Text Evaluation:** Ranking answer likelihoods without relying on instruction-following capabilities.
* **Image Evaluation:** Using CLIP similarity scores to compare generated images with options.
* **Text and Image Evaluation:** Combining text accuracy with image similarity for holistic evaluation.

**4. Evaluation Results**

**4.1 Models**

The study evaluates 22 open-source MLLMs, such as BLIP-2, InstructBLIP, and MiniGPT-4. Models were assessed based on their highest achieved capability level.

**4.2 Main Results**

The findings reveal:

1. MLLMs struggle to achieve even L1 ceiling performance, with top models attaining only 60% accuracy.
2. L2 tasks are notably challenging due to the complexity of interleaved image-text inputs.
3. Few models demonstrate L3 capabilities, indicating limited progress in unified multimodal generation.

**4.3 Observations**

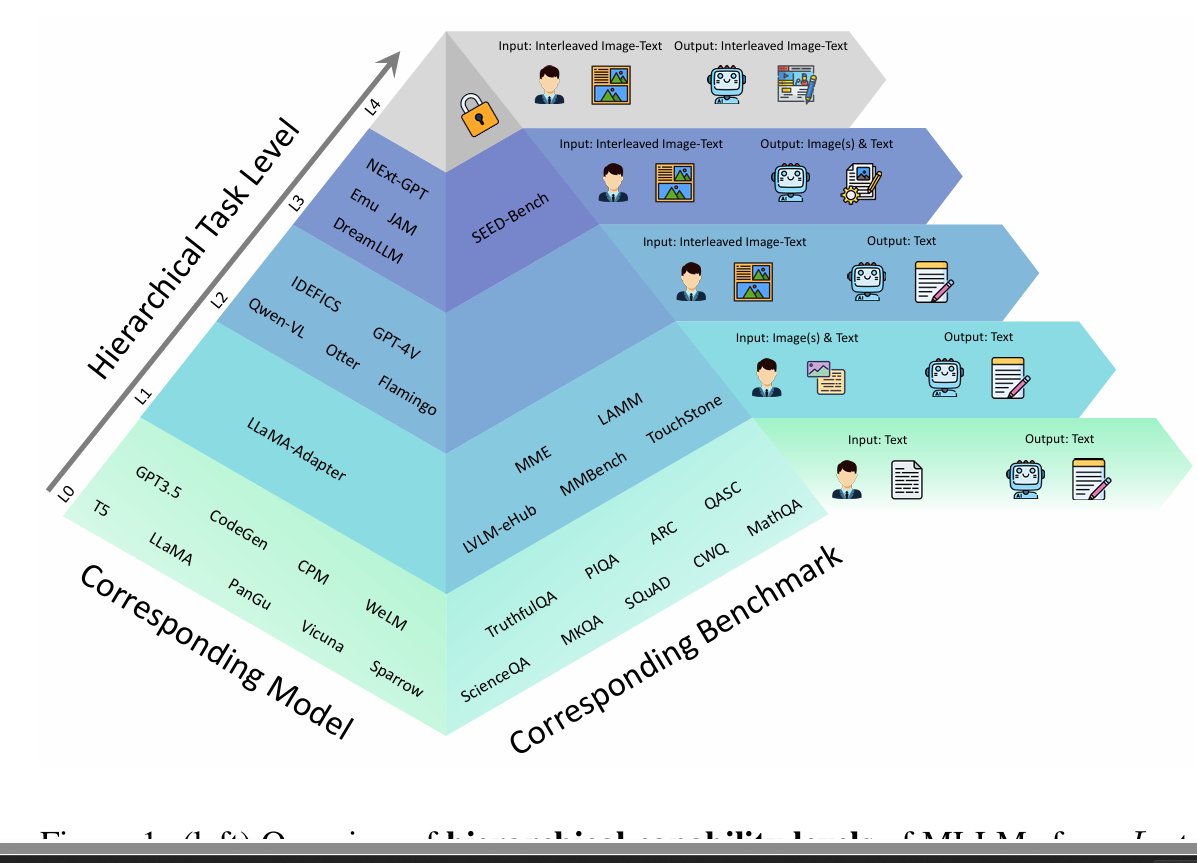
* MLLMs exhibit poor performance in specific areas like chart understanding and visual mathematics.
* Training primarily on structured image-caption pairs limits adaptability to real-world scenarios.
* Achieving simultaneous multimodal comprehension and generation remains an open challenge.

**5. Key Contributions** SEED-Bench’s innovations include:

* **Hierarchical Evaluation:** Providing a roadmap for MLLM advancement.
* **Comprehensive Scope:** Covering diverse tasks and modalities.
* **Objective Metrics:** Enhancing reliability through multiple-choice questions and automated evaluation.

**6. Conclusion** SEED-Bench establishes a robust framework for benchmarking MLLMs, setting a precedent for future research. It reveals significant gaps in current capabilities, urging advancements in multimodal comprehension and generation. By maintaining a leaderboard, SEED-Bench aims to foster continuous progress toward AGI.

**Figure Reference**



* **Figure 1 (Left):** This figure illustrates the hierarchical capability levels (L0 to L4) of MLLMs, emphasizing that higher levels encompass the skills of lower tiers. SEED-Bench-2 evaluates models up to L3, mapping models and benchmarks to each capability tier.

**Reference** The dataset and evaluation code are accessible at:

https://github.com/AILab-CVC/SEED-Bench

**11. Beyond Text-to-Text: An Overview of Multimodal and Generative Artificial Intelligence for Education Using Topic Modeling**

**1. Introduction**

The study begins by highlighting the potential of Generative Artificial Intelligence (GenAI) to transform education. It notes that while large language models (LLMs) like ChatGPT have received much attention in educational research, multimodal AI—which can generate not just text but also images, speech, and video—remains underexplored. The study seeks to map the research landscape of multimodal and generative AI in education using a topic modeling approach. The main research question is:  
*"What is the high-level research landscape of multimodal approaches and generative AI in education?"*

**2. Background**

The background section discusses how AI has long been linked to education, starting from early computational models of learning. While LLMs dominate research today, AI has also historically been used for tasks like text-to-speech and speech-to-text conversion since the 1950s. More recently, AI models have advanced to text-to-image, text-to-video, and even text-to-music applications.

Despite the growing interest in AI for education, the study highlights several key issues:

* Research is unevenly distributed, with a heavy focus on LLMs.
* Market leaders (like OpenAI, Google, and Microsoft) dominate AI research.
* There are ethical and policy concerns regarding AI in education.
* The role of multimodal AI in education remains largely unexplored.

**3. Materials and Methods**

**3.1 Data Collection**

To understand AI's role in education, the researchers conducted an extensive literature review using Dimensions.ai, an academic search engine. They collected 4,175 research papers related to generative AI and education. Their search covered different AI applications, including:

* Text-to-speech
* Text-to-image
* Speech-to-text
* Text-to-video
* Large Language Models (LLMs)

Notably, specific AI product names (e.g., ChatGPT, Bard, Claude) were excluded from the search to avoid bias.

**3.2 Topic Modeling Approach**

To analyze the large dataset, the researchers used a technique called BERTopic modeling, which identifies common research themes. The method involved:

1. Extracting topics from paper abstracts using a machine learning model.
2. Clustering similar topics using statistical methods.
3. Reviewing and interpreting the topics manually.

The final model identified 38 key topics, which were grouped into 14 broad themes.

**4. Results**

The study's results reveal several key insights about AI research in education.

**4.1 AI Modalities in Education**

* LLMs dominate educational AI research, with ChatGPT being the most studied tool.
* Among multimodal AI technologies, the most common were:
  + Text-to-speech (used for accessibility and language learning).
  + Text-to-image (used for creative education and visual learning).
  + Speech-to-text (used for note-taking and accessibility).
* Text-to-video and text-to-music are emerging but remain under-researched.

**4.2 Topic Modeling Insights**

The 38 research topics identified in the study were categorized into 14 major themes:

|  |  |  |
| --- | --- | --- |
| 1 | Domains of Education | AI applications in medical, business, engineering, and geoscience education |
| 2 | Personalized Learning | AI tools for adaptive learning, sentiment analysis, and feedback. |
| 3 | Problem Solving | AI's role in mathematics, physics, and simulations. |
| 4 | Technology Adoption | How students and educators accept and integrate AI into learning. |
| 5 | Teacher Training | AI's impact on teacher education and professional development. |
| 6 | Creativity and Art | Using text-to-image AI for design and creativity. |
| 7 | Serious Games | AI-powered educational gaming. |
| 8 | AI Tools and Content | AI-driven lesson planning, question generation, and information management. |
| 9 | Assessment and Integrity | grading, plagiarism detection, and academic honesty. |
| 10 | Ethics and Security | Concerns around data privacy, misinformation, and fairness in AI. |
| 11 | AI in Language Learning | The use of chatbots and AI-driven language instruction. |
| 12 | Chatbots in Education | AI chatbots assisting in tutoring and student engagement. |
| 13 | AI Sentiment Analysis | Analyzing student emotions and attitudes toward AI. |
| 14 | Detecting AI-Generated Content | Methods to identify AI-generated text and code. |

These themes [table 1] indicate that most research still focuses on LLMs rather than multimodal AI approaches.

**5. Discussion and Conclusion**

The study reinforces existing concerns about the dominance of LLMs in educational AI research. While LLMs are useful for writing support and feedback, they also raise concerns about plagiarism and academic integrity. At the same time, multimodal AI offers promising but underutilized tools.

* **Key Takeaways**
* LLMs like ChatGPT dominate AI research in education.
* Multimodal AI (text-to-speech, image, video, and music) remains underexplored.
* AI can personalize learning but also poses risks to academic integrity.
* More research is needed on multimodal AI and its impact beyond higher education.
* **Recommendations for Future Research**
* Develop educationally focused AI models (e.g., "EdGPTs").
* Explore multimodal AI to enhance visual and auditory learning.
* Investigate AI's impact on K-12 education, not just higher education.
* Address ethical concerns about AI bias, misinformation, and privacy.

**Final Thought:**

The study sheds light on the research gaps in AI for education. While LLMs are useful, multimodal AI has the potential to create richer, more interactive learning experiences. The authors advocate for a more balanced focus on different AI technologies, moving beyond text-to-text applications to unlock AI’s full potential in education.

**Summary of Contributions:**

This study provides a comprehensive review of AI in education and highlights major trends, gaps, and future opportunities. It is a valuable resource for educators, researchers, and policymakers seeking to understand how AI is shaping education today and where it should go next.

**12. DesignPrompt: Using Multimodal Interaction for Design Exploration with Generative AI**

**1. Introduction**

Generative AI (GenAI) has revolutionized the design field by enabling designers to create images using text prompts. Tools like DALL·E, Stable Diffusion, and MidJourney help generate novel visuals, but these tools pose challenges for non-experts. Designers often struggle with translating their intentions into effective text prompts, leading to misaligned results.

The paper introduces **DesignPrompt**, a digital moodboard tool that lets designers compose prompts using **multiple input types**—images, colors, and text. The study investigates how multimodal input enhances designers’ ability to express ideas and control AI-generated outputs.

**2. Related Work**

**2.1 Creativity Support Tools (CSTs) for Design Ideation**

Creativity Support Tools (CSTs) help designers during the ideation phase by enhancing their ability to explore concepts visually. Previous tools like **IdeaWall** and **GroupMind** have supported brainstorming with collaborative and visual mind-mapping features. However, these tools **do not integrate generative AI**, which limits their ability to assist in visual idea exploration.

**2.2 Generative AI for Design Practice**

GenAI has been widely used in fields like UI design, fashion design, and graphic design. Despite its capabilities, GenAI often replaces creative decision-making rather than enhancing it. Most AI tools focus on output generation without helping users control the creative process. Recent advancements like **DreamBooth**, **ControlNet**, and **InstructPix2Pix** have introduced methods for better AI control, but they remain **difficult for non-experts** to use effectively.

**2.3 Challenges of AI Prompting in Creative Processes**

The biggest challenge with AI-driven creativity is the difficulty of writing effective prompts. Previous research has identified three key issues:

* **Intentionality gap** – Users struggle to clearly express their design goals.
* **Capability gap** – Users don’t fully understand AI’s abilities.
* **Language gap** – Users find it difficult to communicate with AI in structured text prompts.

The study explores **multimodal prompting** as a solution to these challenges, making AI interactions **more intuitive for designers**.

**3. Preliminary Study**

**Objective**

The research team conducted a preliminary study to understand how designers use **moodboards** and interact with **GenAI tools**. The goal was to identify challenges and develop design guidelines for **DesignPrompt**.

**Method**

* **Participants**: 8 users (mix of designers and engineers).
* **Tasks**:
  1. **Moodboard Task** – Participants created a moodboard using traditional tools (e.g., Pinterest).
  2. **GenAI Task** – Participants generated images using a text-based GenAI tool.

**Key Findings**

* **Difficulty in Formulating Search Terms & Prompts** – Designers struggled to convert ideas into effective search terms or AI prompts.
* **Lack of Control Over AI Outputs** – AI-generated images often **did not match expectations**, and fine-tuning was difficult.
* **Desire for Richer Inputs** – Participants wanted to use **multiple sources** (images, text, colors) to create prompts.

**Design Implications**

The study led to four key recommendations for designing AI-powered tools for **creative professionals**:

1. Support **different levels of abstraction** in search and prompt input.
2. Help users **translate abstract intentions** into rich prompts.
3. Make the system’s **interpretation of prompts transparent**.
4. Allow users to **engage interactively** with AI-generated images.

**4. DesignPrompt: A Multimodal Moodboard System**

**DesignPrompt** was developed based on the preliminary study’s findings. It allows designers to:

* **Combine multimodal inputs** (images, colors, text).
* **Refine AI-generated results interactively**.
* **Understand how AI interprets inputs** to align expectations with outcomes.

**Features of DesignPrompt**

* **Moodboard Canvas** – Users can arrange images, colors, and text on a digital canvas.
* **Semantic Meta-Data** – Each image is automatically tagged with **colors and semantic labels** to help users search for related images.
* **Multimodal Input Composer** – Users can drag and drop elements to generate AI prompts.
* **Interactive Prompt Editor** – Users can fine-tune prompts and see how AI interprets them.
* **Prompt History & Variation** – Users can track changes in AI prompts and explore different variations.

**5. Comparative Study: Multimodal vs. Text-Only AI Prompting**

**Research Questions**

The study aimed to answer:

1. Does **multimodal prompting** help designers explore ideas better?
2. Does making the **AI’s interpretation visible** improve alignment with user expectations?
3. Does **interactive input control** make the system more transparent and useful?

**Methodology**

* **Participants**: 12 professional designers.
* **Conditions**:
  1. **Text-Only AI Prompting** (baseline).
  2. **Multimodal AI Prompting** (DesignPrompt).
* **Tasks**: Participants created moodboards for different design challenges using both systems.

**Findings**

* **Multimodal prompting improved creativity** – Designers found it easier to **express abstract ideas**.
* **Users had mixed feelings about control** – Some preferred the structured nature of text prompts, while others liked the flexibility of multimodal input.
* **AI results were often surprising** – Designers **appreciated unexpected results**, but some found them unpredictable.
* **Users adapted their workflows** – Over time, designers **learned how to use multimodal input effectively**.

**6. Discussion & Future Work**

**Key Insights**

1. **Multimodal prompting helps designers express ideas** more fluidly than text-based prompting alone.
2. **Control perception varies** – Some users preferred structured text input, while others valued **interactive control over AI interpretations**.
3. **Expectation alignment is complex** – While some designers enjoyed **unexpected AI creativity**, others wanted more **predictability**.
4. **Learning and adaptation are crucial** – Over time, designers found ways to **use AI creatively** and **develop unique workflows**.

**Limitations & Future Directions**

* The study used **a short design session (12 minutes)**, whereas real-world projects **span days or weeks**. Future studies should explore **long-term AI usage**.
* **More fine-tuned AI editing tools** (e.g., ControlNet, DreamBooth) could enhance **precision and control**.
* **Beyond moodboards**, multimodal AI prompting could be applied to **storyboarding, brainstorming, and visual concept development**.

**7. Conclusion**

The study presents **DesignPrompt** as a tool that integrates **multimodal inputs** into GenAI-powered moodboarding. The research highlights the potential of **multimodal AI interaction** to make **AI more accessible and intuitive for designers**. The findings contribute to **human-centered AI design** by exploring **how creative professionals engage with AI tools in practice**.

**Final Thoughts**

This study is significant because it shifts the focus from **text-based AI prompting** to a **more visual and intuitive approach**. By allowing designers to use **images, colors, and semantic metadata**, DesignPrompt makes AI-powered design tools **more natural and effective**. The research also opens up new questions about **how AI can be tailored to creative professionals’ needs**, paving the way for **future AI-assisted creativity tools**.

**13. A Comprehensive Review of Multimodal Large Language Models:** **Performance and Challenges Across Different Tasks**

**1. Introduction**

Multimodal Large Language Models (MLLMs) are artificial intelligence (AI) systems designed to process and integrate various types of data, including text, images, videos, audio, and physiological sequences. Unlike traditional single-modality models, MLLMs are capable of handling complex real-world applications by fusing multiple types of information. This paper systematically explores MLLMs, their performance on multimodal tasks, their limitations, and potential future research directions.

**2. Overview of Multimodal Large Language Models (MLLMs)**

**Definitions and Basic Concepts**

MLLMs extend the capabilities of traditional language models by incorporating data from multiple sources. They can understand and generate content across different modalities, making them highly useful in AI-driven applications like automated video descriptions, speech-to-text translation, and multimodal search engines.

**Main Components of MLLMs**

1. **Multimodal Input Encoder**
   * This component processes different types of input (text, images, audio, etc.) and converts them into numerical representations.
   * Examples include pre-trained vision models like CLIP for images and HuBERT for audio【8:3†2408.01319v1.pdf】.
2. **Feature Fusion Mechanism**
   * Since different modalities have different data structures, MLLMs use feature fusion techniques to align and combine them.
   * Popular fusion methods include simple linear layers, attention mechanisms (e.g., Transformer-based architectures), and contrastive learning.

**3. Classification of MLLMs Based on Tasks**

The paper categorizes MLLMs into different types based on the tasks they are designed for. These tasks can be grouped into image-related, video-related, and audio-related applications.

**A. Image-Based Tasks**

**Image Understanding**

MLLMs analyze images and provide textual descriptions or categorize visual elements. Common applications include:

* **Image Captioning**: Generating captions based on visual content (e.g., BLIP, LLaVA).
* **Visual Question Answering (VQA)**: Answering questions related to an image.
* **Object Recognition**: Identifying and labeling objects in an image.

**Image Generation**

MLLMs can generate images from text prompts, a technique commonly used in AI art and content creation. Techniques include:

* **Generative Adversarial Networks (GANs)**
* **Diffusion Models** (e.g., Stable Diffusion)

**B. Video-Based Tasks**

**Video Understanding**

MLLMs analyze video content for tasks like action recognition, event detection, and scene description. Two primary approaches exist:

* **Feature Extraction Models** (e.g., VideoCLIP, VideoLLaVA)
* **Sequence-based Learning Models** (e.g., Transformers for long-range dependencies).

**Video Generation**

AI models generate videos from text, images, or other videos. Techniques include:

* **Autoregressive models** (e.g., Video Transformer)
* **Diffusion models** (e.g., Imagen Video, Gen-2).

**C. Audio-Based Tasks**

**Audio Understanding**

MLLMs process and interpret audio signals. Applications include:

* **Speech Recognition**: Converting spoken language into text (e.g., Whisper, HuBERT).
* **Audio Classification**: Identifying different sound types (e.g., music vs. speech).
* **Emotion Recognition**: Detecting emotions from speech.

**Audio Generation**

MLLMs can generate human-like speech, music, or other audio content. Examples include:

* **Text-to-Speech (TTS) Systems** (e.g., SpeechGPT).
* **AI-generated music** (e.g., Jukebox by OpenAI).

**4. Comparison of MLLMs**

The paper compares different MLLMs based on their performance in handling various tasks. It highlights the trade-offs between accuracy, computational efficiency, and the ability to generalize across multiple modalities.

|  |  |  |  |
| --- | --- | --- | --- |
| **Task Type** | **Representative Models** | **Key Strengths** | **Limitations** |
| **Image Tasks** | CLIP, BLIP, LLaVA | Strong visual-text alignment | Struggles with fine-grained details |
| **Video Understanding** | Video-LLaVA, BuboGPT | Good action recognition | High computational cost |
| **Video Generation** | Gen-2, Imagen Video | Realistic video synthesis | Temporal consistency issues |
| **Audio Tasks** | Whisper, SpeechGPT | High speech recognition accuracy | Limited contextual understanding |

One key challenge in MLLMs is the need for massive computational resources. Training these models often requires hundreds of high-end GPUs, making them expensive to deploy.

**5. Challenges and Future Directions**

Despite their advancements, MLLMs face several limitations:

**A. Computational Costs**

* Training and inference require extensive hardware resources, limiting accessibility.
* Possible solution: Developing efficient architectures and optimizing model compression techniques.

**B. Data Limitations**

* MLLMs require large, diverse datasets to generalize well.
* Issues like dataset bias and lack of annotated multimodal data hinder performance.

**C. Fusion Complexity**

* Integrating multiple modalities while maintaining semantic coherence is difficult.
* Advanced fusion mechanisms like hierarchical attention and contrastive learning are promising areas for research.

**D. Ethical Concerns**

* Multimodal models can be used to generate misleading or biased content.
* The paper emphasizes the need for robust model interpretability and bias mitigation strategies.

**6. Conclusion**

MLLMs represent a significant advancement in AI, enabling systems to process and generate content across multiple modalities. However, their adoption is hindered by computational demands, data challenges, and ethical concerns. Future research should focus on improving efficiency, enhancing fusion mechanisms, and ensuring fairness and transparency in model outputs.

By addressing these issues, MLLMs can play a transformative role in AI applications, including education, healthcare, and creative industries.

-------------------------ORRRRRRRRR

**Introduction**

Multimodal Large Language Models (MLLMs) represent a transformative advancement in artificial intelligence (AI), designed to integrate and process data from diverse modalities such as text, images, audio, video, and physiological sequences. These models address the limitations of single-modality systems, unlocking new potential for real-world applications by providing comprehensive and nuanced information representation. This review systematically explores MLLMs' applications, performance across multimodal tasks, comparative strengths and weaknesses, and future directions for research.

**Overview of Multimodal Large Language Models**

**Definitions and Basic Concepts** MLLMs are AI systems engineered to interpret and generate information across multiple modalities simultaneously. By synergizing these varied data types, they deliver more accurate and enriched outputs, surpassing the limitations of single-modality systems. For instance, in tasks like image captioning or machine translation, these models utilize visual, textual, and auditory data to improve contextual understanding.

**Main Components** MLLMs comprise three critical components:

1. **Multimodal Input Encoder**: Encodes raw data from different modalities into structured feature representations. For example:
   * **Text**: Utilizes transformers and embedding layers to map textual input into vector representations.
   * **Images**: Employs Vision Transformers (ViT) or Residual Networks (ResNet) to encode visual data.
   * **Audio**: Uses specialized models like HuBERT or Whisper to process auditory inputs.
2. **Feature Fusion Mechanism**: Integrates features from multiple modalities using techniques like:
   * **Early Fusion**: Combines raw features at the input stage.
   * **Intermediate Fusion**: Merges features during the extraction phase.
   * **Late Fusion**: Combines outputs at the decision stage.
   * **Joint Fusion**: A hybrid approach leveraging all fusion stages for maximum data utility.
3. **Multimodal Output Decoder**: Converts integrated multimodal features into outputs tailored to specific tasks, such as textual summaries or visual annotations.

**Overview of Multimodal Feature Fusion in LLMs** Pre-trained large models (e.g., GPT or BERT) are often adapted for multimodal tasks by projecting features from diverse modalities into a unified feature space. This enables models to handle tasks like image-text generation or audio-visual sentiment analysis.

**Task Classification of MLLMs**

**1. Image Tasks**

MLLMs play a vital role in both **image understanding** and **image generation**:

**Image Understanding**

* These models excel in tasks such as object recognition, scene analysis, and image annotation. They integrate textual and visual data to provide detailed analyses.
* Techniques have evolved through stages:
  + **Traditional Feature Extraction**: Early methods relied on manually designed features (e.g., SIFT, HOG).
  + **Deep Learning Advances**: Convolutional Neural Networks (CNNs) enabled better feature learning.
  + **Multimodal Learning**: Cross-modal approaches integrated image and text for tasks like visual question answering.

**Image Generation**

* MLLMs generate realistic images from textual prompts, leveraging advancements like:
  + **GANs (Generative Adversarial Networks)**: Pioneered early image synthesis.
  + **Conditional GANs and StyleGANs**: Improved semantic coherence and image resolution.
  + **Multimodal Approaches**: Combined modalities for tasks such as image style transfer or text-to-image generation.

**Representative Models**

* **MiniGPT-4**: Combines visual encoders with LLMs to generate nuanced image descriptions.
* **InstructBLIP**: Integrates instruction tuning for vision-language tasks, enhancing task-specific performance.

**2. Video Tasks**

MLLMs address the complexities of video data, which integrates visual, auditory, and temporal information. Key applications include:

**Video Understanding**

* Tasks include action recognition, video question answering (VQA), and video retrieval. MLLMs like **Video-LLaMA** and **X-InstructBLIP** align video and text data to extract meaningful insights from multimedia content.

**Video Generation**

* Video synthesis from text or images relies on advanced architectures such as diffusion models, offering higher quality and semantic relevance.
* Models like **NeXT-GPT** utilize pre-trained encoders and decoders for multimodal video creation.

**3. Audio Tasks**

MLLMs have revolutionized audio processing by integrating auditory data with text and visual information.

**Audio Understanding**

* Applications include speech recognition, sentiment analysis, and audio event detection.
* Models like **Qwen-Audio** and **SALMONN** employ instruction-tuned datasets to enhance cross-modal comprehension.

**Audio Generation**

* Tasks such as text-to-speech (TTS) benefit from models like **SpeechGPT** and **AudioGPT**, which utilize multimodal datasets and architectures to produce high-quality audio outputs.

**Challenges and Comparisons of MLLMs**

**Challenges**

1. **Data Alignment**: Ensuring high-quality alignment across modalities is complex.
2. **Computational Demand**: Training MLLMs requires significant computational resources, especially for large-scale datasets.
3. **Model Robustness**: Achieving consistent performance across diverse tasks and datasets remains challenging.
4. **Real-Time Applications**: Implementing MLLMs in real-time scenarios, such as interactive systems, poses latency issues.

**Comparative Analysis**

MLLMs excel in tasks requiring contextual richness, but their performance varies depending on the fusion technique and the dataset quality. For instance:

* Models with **early fusion** often perform well on tightly coupled multimodal tasks but struggle with scalability.
* **Late fusion** models are more adaptable but may lose contextual nuance.
* Hybrid approaches like **joint fusion** balance these trade-offs, achieving superior performance in many scenarios.

**Future Directions**

1. **Enhanced Data Efficiency**: Developing techniques to reduce dependency on large labeled datasets.
2. **Universal Encoders**: Creating generalized encoders capable of processing diverse modalities with minimal customization.
3. **Scalability and Efficiency**: Optimizing architectures for real-world applications on resource-constrained devices.
4. **Ethical Considerations**: Addressing biases in multimodal datasets to ensure fairness and inclusivity.
5. **Dynamic Multimodal Interactions**: Enabling models to adapt dynamically to evolving multimodal inputs, such as real-time video streams.

**Conclusion**

MLLMs signify a paradigm shift in AI, demonstrating remarkable capabilities across diverse tasks by integrating multiple data modalities. While challenges persist, ongoing innovations in architecture, training methods, and application design promise to further expand their potential. This review provides valuable insights for researchers and practitioners, inspiring future advancements in the field of multimodal AI.

14. **A Survey on Multimodal Benchmarks: In the Era of Large AI Models**

**Introduction**

The rapid advancements in **Multimodal Large Language Models (MLLMs)** have significantly improved AI's ability to understand and generate content that combines multiple data types, such as **text, images, videos, and audio**. While much research has focused on the architecture and training of these models, the evaluation benchmarks for MLLMs have not been thoroughly analyzed.

This survey fills this gap by systematically reviewing **211 multimodal benchmarks**, categorizing them into four key domains:

1. **Understanding** – How well a model interprets multimodal data.
2. **Reasoning** – Its ability to draw logical conclusions from different data types.
3. **Generation** – The creation of text, images, and other content.
4. **Application** – The model’s real-world usability.

The paper highlights three major challenges in benchmarking:

1. **Fragmented Objectives** – Many benchmarks focus on different tasks, making comparisons difficult.
2. **Task Saturation** – Too many benchmarks exist, making it hard to determine which are most valuable.
3. **Metric Evolution & Discrepancies** – Different benchmarks use different evaluation methods, making standard comparisons difficult.

**Understanding Benchmarks**

**Understanding** refers to the ability of MLLMs to **extract information from multimodal data** and integrate features across different types (e.g., text and images). It is a foundational ability needed for reasoning, generation, and application tasks.

**Taxonomy of Understanding Benchmarks**

Understanding is divided into:

1. **Visual Perception** – The ability to extract and recognize features from images.
   * Low-Level Perception: Focuses on recognizing basic visual features (e.g., color, shape, and texture).
   * Fine-Grained Perception: Includes detailed visual recognition such as Optical Character Recognition (OCR).
   * Higher-Order Perception: Understanding emotions and aesthetics in images.
   * Comprehensive Perception: Evaluating models across multiple perception tasks.
   * Multilingual Perception: Ability to process visual content in multiple languages.
2. **Contextual Comprehension** – How well a model understands information influenced by the surrounding context.
   * Context-Dependent Understanding: Ability to identify objects within a **single image** while using contextual clues.
   * Long-Context Understanding: Processing information from **long sequences of data**.
   * Multi-Image Understanding: Comparing and interpreting multiple images.
   * Interleaved Image-Text Understanding: Understanding **text and images presented together**, such as in news articles.
3. **Specific Modality Understanding** – The ability to handle **specialized** data types.
   * Video Understanding: Capturing **motion, actions, and sequences** in videos.
   * Audio Understanding: Processing **speech and non-verbal sounds**.
   * 3D Understanding: Analyzing **depth, shape, and structure** in 3D environments.
   * Omni-Modal Understanding: Integrating **multiple modalities** simultaneously.

**Metrics for Understanding**

* Accuracy, F1-score (for classification tasks)
* BLEU and ROUGE scores (for text-based evaluation)
* Intersection over Union (IoU) for object detection

**Reasoning Benchmarks**

**Reasoning** is the ability of MLLMs to perform logical inferences and solve problems using multimodal data.

**Taxonomy of Reasoning Benchmarks**

1. **Domain-Specific Reasoning** – Applying logical processes in specialized fields like mathematics and science.
2. **Relational Reasoning** – Understanding **spatial, temporal, and logical relationships** between objects.
3. **Multi-Step Reasoning** – Solving problems that require multiple steps of logical thinking.
4. **Reflective Reasoning** – Evaluating **self-awareness and adaptability**, including counterfactual thinking and knowledge updating.

**Metrics for Reasoning**

* Logical consistency
* Step-wise accuracy (for multi-step reasoning)
* Correctness in applying domain-specific rules

**Generation Benchmarks**

**Generation** refers to the ability of MLLMs to produce coherent and relevant multimodal content.

**Taxonomy of Generation Benchmarks**

1. **Format-Centric Generation**
   * **Interleaved Image-Text Generation**: Generating visually and contextually aligned content (e.g., captions, descriptions).
   * **Code Generation**: Writing programming code from images or text instructions.
   * **Instruction Following**: Adhering to user instructions in multimodal tasks.
2. **Content-Centric Generation**
   * **Hallucination Mitigation**: Reducing incorrect content generation.
   * **Safety**: Ensuring content does not produce harmful or biased outputs.
   * **Trustworthiness**: Generating factually correct and reliable content.
   * **Robustness**: Producing stable outputs despite noisy or misleading inputs.

**Metrics for Generation**

* **Text-Only**: BLEU, ROUGE, METEOR (measuring text coherence)
* **Vision-Only**: FID (Frechet Inception Distance) for image realism
* **Cross-Modality**: Style and content consistency scores

**Application Benchmarks**

The final domain evaluates how well MLLMs perform in real-world **practical applications**.

**Taxonomy of Application Benchmarks**

1. **Visual Agents** – AI agents that can process multimodal inputs and interact with environments.
   * **Interactive Decision-Making Agents**: Used for tasks like web navigation and mobile UI interactions.
   * **Embodied Decision-Making Agents**: Robots that integrate **perception and action**.
2. **Domain-Specific Applications**
   * **Medical AI**: Analyzing medical images and text (e.g., X-ray reports).
   * **Autonomous Driving**: Understanding street images and signals.
   * **Remote Sensing**: Processing **satellite imagery** for geographic tasks.

**Metrics for Applications**

* Task completion rate
* Response accuracy in real-world scenarios
* Consistency across different environments

**Conclusion and Future Directions**

The paper concludes by emphasizing:

* The need for **standardized** evaluation metrics for better comparison of MLLMs.
* Development of **general-purpose** benchmarks that test multiple capabilities at once.
* Exploring **new multimodal interactions**, such as video + text + audio reasoning.
* Improving **fairness and robustness** in MLLM-generated content.

**Summary of Key Insights**

1. **MLLMs require new and advanced benchmarks** to assess their true capabilities.
2. **Existing benchmarks are fragmented**, making it hard to compare models effectively.
3. **Four key areas are important**: **Understanding, Reasoning, Generation, and Application**.
4. **Metrics must evolve** to assess not just accuracy but also **reliability, safety, and fairness**.

15. **Multi-Modal Generative AI: Multi-modal LLM, Diffusion and Beyond**

**1. Introduction**

Multi-modal generative AI is an emerging field that integrates multiple data types (such as text, images, and videos) to enhance AI understanding and generation capabilities. This paper discusses two main types of multi-modal models:

* **Multi-modal Large Language Models (MLLMs):** These models (e.g., GPT-4V) excel at multi-modal understanding by interpreting text and images together.
* **Diffusion Models:** These models (e.g., Sora) specialize in high-quality image and video generation from textual descriptions.

The core question explored in the paper is:  
*"Can we design a unified model that can both understand and generate multi-modal data?"*

To answer this, the paper systematically reviews MLLMs and diffusion models, exploring:

* Their **probabilistic modeling techniques** (auto-regressive vs. diffusion-based),
* Their **architectural designs** (dense models vs. Mixture of Experts),
* Strategies for building a **unified model** that balances both understanding and generation,
* **Large-scale multi-modal datasets** useful for pretraining,
* Future research directions in multi-modal generative AI.

**2. Multi-modal Large Language Models (MLLMs) for Understanding**

MLLMs focus on interpreting images and videos using a combination of deep learning techniques. This section reviews different aspects of MLLMs.

**2.1. Preliminaries**

1. **Auto-regressive Probabilistic Modeling:**
   * MLLMs use auto-regressive modeling, where each token (word or image-related feature) is predicted based on previous tokens.
   * This approach is effective for generating coherent responses based on multi-modal inputs.
2. **Vision-Language Pretraining (VLP):**
   * Early models trained on paired image-text datasets (e.g., CLIP, ALIGN) using methods like contrastive learning.
   * Two major architectures:
     + **BERT-like models**: Jointly process images and text for tasks like visual question-answering.
     + **Two-tower models**: Separate vision and language processing, useful for retrieval tasks.
3. **Visual Tokenizers:**
   * To integrate images into language models, they must be converted into tokens.
   * **VQ-VAE & VQGAN** transform images into discrete tokens, making them compatible with text-based LLMs.

**2.2. MLLM Architectures**

MLLM architectures can be categorized into:

1. **Alignment Architectures:**
   * These models use pre-trained vision encoders (e.g., CLIP) and align their outputs with an LLM.
   * Examples: LLaVA, Qwen-VL, MiniGPT-4.
   * Advantages: Leverages existing pre-trained models, requiring fewer resources.
   * Limitations: Struggles with multiple objects in an image and loses fine-grained visual details.
2. **Early-fusion Architectures:**
   * Convert images into discrete tokens and process them alongside text from the beginning.
   * Example: Chameleon, Gemini.
   * Advantages: Fully integrates visual and textual information.
   * Limitations: Requires extensive computational resources.

**2.3. Image Large Language Models (Image-LLMs)**

* **Alignment-based models**: Use pre-trained vision encoders with alignment modules (e.g., LLaVA, BLIP-2).
* **Early-fusion models**: Process text and image tokens together (e.g., Chameleon, Gemini).

Challenges:

* **Fine-grained understanding:** Some models, like AnyRef and OMG-LLaVA, aim to improve object segmentation and spatial awareness.
* **Reducing hallucination:** Strategies like better vision encoders and human feedback reduce errors in generated outputs.

**2.4. Video Large Language Models (Video-LLMs)**

* **Similar to Image-LLMs but adapted for video input.**
* Methods include:
  + Sampling frames from videos and processing them like images.
  + Using specialized alignment modules to capture temporal relationships (e.g., VideoLLaMA, VideoChatGPT).
* **Challenges:**
  + **Understanding long videos:** Models struggle with memory limitations and require efficient token representations.
  + **Segment-wise comprehension:** Some models incorporate timestamps and event detection (e.g., VTimeLLM, TimeChat).

**3. Multi-modal Diffusion Models for Generation**

Diffusion models are widely used for generating high-quality images and videos from text descriptions.

**3.1. Preliminaries**

Earlier generative models included:

1. **GANs (Generative Adversarial Networks):**
   * Generator produces images, Discriminator evaluates them.
   * Struggles with instability and mode collapse.
2. **VAEs (Variational Autoencoders):**
   * Encode images into a latent space, then decode them back.
   * Suffer from blurry images due to constraints on latent space.
3. **Diffusion Models:**
   * A new paradigm where images are progressively generated by refining noisy inputs.
   * More stable than GANs and higher quality than VAEs.

**3.2. Latent Diffusion Models (LDM)**

* Traditional diffusion models work on pixel space, which is computationally expensive.
* **LDMs (e.g., Stable Diffusion, Imagen)** operate in a compressed latent space, reducing computation while maintaining quality.

**3.3. Text-to-Image Generation**

* Early methods (e.g., GLIDE, DALL-E) used **pixel-level diffusion**.
* Later models (e.g., **Stable Diffusion, Imagen, DiT**) introduced **latent-based diffusion** and **transformer-based architectures** for better scalability.

**3.4. Controllable Image Generation**

* Users may want more **control** over generated images (e.g., specifying layout, style, or a specific object).
* **Techniques for control:**
  + **Model-based** (e.g., additional encoders).
  + **Tuning-based** (e.g., DreamBooth, Textual Inversion).
  + **Training-free** (e.g., modifying attention layers dynamically).

**4. Text-to-Video Generation**

Diffusion models have also advanced video generation.

**4.1. Text-to-Video Diffusion Models**

* **Frame-based generation**: Extends text-to-image models by generating each frame independently.
* **Temporal-aware models**: Use 3D convolutions or attention to capture motion (e.g., Make-a-Video, AnimateDiff).

**4.2. Controllable Video Generation**

* Uses **reference videos** or **motion guidance** to enhance control (e.g., Control-A-Video, Follow-Your-Pose).
* **Challenges:**
  + Maintaining **motion smoothness**.
  + Keeping **scene consistency** across frames.

**4.3. Long-Video Generation**

* **Problem**: Existing models can only generate short clips (16-24 frames).
* **Solutions**:
  + **Auto-regressive chunking**: Generate short segments sequentially.
  + **Larger datasets & models** (e.g., Sora).
  + **Memory-efficient architectures** for long-range dependencies.

**5. Towards a Unified Model for Understanding and Generation**

The paper explores how to build a **single model** capable of both **understanding and generating** multi-modal data.

**5.1. Probabilistic Modeling: Auto-Regressive vs. Diffusion**

* **Auto-regressive models** (used in MLLMs) are good for understanding but struggle with high-quality generation.
* **Diffusion models** excel at generation but are computationally expensive.
* **Hybrid approaches** aim to combine the strengths of both.

**5.2. Model Architectures**

1. **Dense Models**: Treat all modalities equally.
2. **Mixture of Experts (MoE)**: Assigns specialized experts for different modalities.

**6. Future Directions**

* **Scalability**: How to make unified models more efficient.
* **Fine-grained multi-modal reasoning**: Improve reasoning across images, videos, and text.
* **Long-form video generation**: Extend model capabilities beyond short clips.

**Conclusion**

This paper provides a comprehensive review of multi-modal generative AI, comparing MLLMs and diffusion models. It highlights the challenges in **unifying understanding and generation** and presents potential solutions. Future research should focus on **efficient hybrid models**, **better dataset curation**, and **more controllable generation techniques**.

16. Multimodal Foundation Models: From Specialists to General-Purpose Assistants

Option 1:

In this paper, the authors focus on the evolution of multimodal foundation models, tracing their journey from task-specific tools to general-purpose AI assistants capable of handling diverse vision and vision-language tasks. The paper explores visual understanding through approaches like supervised learning, contrastive language-image pretraining, and self-supervised methods, highlighting their role in building robust image backbones for tasks like classification and segmentation. It also delves into advancements in visual generation, emphasizing models like DALL-E and Stable Diffusion for text-to-image synthesis, as well as tools for spatial control, interactive editing, and concept customization. The authors discuss the shift towards unified vision models that integrate various vision and vision-language tasks into a single framework, transitioning from static, task-specific systems to dynamic, promptable ones. Further, the integration of large language models into multimodal settings is examined, showcasing end-to-end training techniques and instruction tuning to enable models like GPT-4 to handle both text and visual inputs. The paper also highlights efforts to chain multimodal tools, such as Visual ChatGPT and MM-ReAct, to create interactive agents capable of reasoning across modalities. Concluding with insights on future trends, the paper underscores the importance of aligning multimodal systems with human intents, increasing interactivity, and scaling capabilities to build versatile, general-purpose AI assistants.

Option 2:

**Multimodal Foundation Models: From Specialists to General-Purpose Assistants**

**1. Introduction**

This section introduces the concept of multimodal foundation models, which integrate capabilities in vision and vision-language tasks. The authors trace the evolution of these models from task-specific tools to general-purpose AI assistants. Inspired by the development trajectory of language models like BERT, GPT-3, and GPT-4, the paper explores how multimodal models, such as CLIP and DALL-E, aim to address the diverse requirements of computer vision and vision-language challenges.

Key Highlights:

* Vision plays a central role in AI for perceiving and interacting with the world, mimicking human abilities to recognize objects and create visual content.
* Models evolved from task-specific designs (e.g., classification) to general-purpose assistants capable of open-ended tasks (like those facilitated by GPT-4V).
* The section emphasizes the increasing need for general-purpose visual assistants that align with human intents and leverage multimodal inputs (e.g., text and images).

**2. Visual Understanding**

This section delves into methods for learning strong image backbones, crucial for visual understanding tasks such as classification and segmentation. It outlines various learning strategies:

1. **Supervised Learning**: Uses labeled datasets like ImageNet to build foundational visual models.
2. **Contrastive Language-Image Pretraining (CLIP)**: Matches image and text pairs to create models capable of zero-shot image classification.
3. **Image-Only Self-Supervised Learning**: Techniques like masked image modeling (e.g., MAE) enable the model to learn from images alone.
4. **Multimodal Fusion**: Combines vision and language to enhance tasks like object detection and pixel-level segmentation.

Key Insight:

* The integration of diverse supervision methods—ranging from label-based to self-supervised approaches—enables better transferability of models across tasks.

**3. Visual Generation**

This section discusses advancements in generating visual content (images/videos) aligned with human instructions. Key methods include:

* **Text-to-Image Generation**: Tools like DALL-E and Stable Diffusion synthesize high-quality images from textual prompts.
* **Spatial Control and Editing**: Techniques like ControlNet allow users to manipulate specific regions of an image while preserving realism.
* **Concept Customization**: Models like DreamBooth help fine-tune visual generators for specific styles or concepts.

Emerging Trend:

* Generative models are increasingly tailored to align with human creativity, enabling tasks like interactive editing and video synthesis.

**4. Unified Vision Models**

The authors discuss the push towards unified models that handle diverse vision tasks using a single architecture. This mirrors advancements in language models, which consolidate various NLP tasks.

Transitions Highlighted:

1. From **Closed-Set Models** (limited categories) to **Open-Set Models** (flexible categorization).
2. From **Task-Specific Models** to **Generic Models**: Tools like Unified-IO streamline different vision-language tasks.
3. From **Static** to **Promptable Models**: Interactive capabilities, inspired by ChatGPT, allow these models to dynamically respond to user inputs.

**5. Training Large Multimodal Models**

This section explains how large language models (LLMs) are extended to handle visual inputs. By integrating visual data into LLMs, models like Flamingo and GPT-4 achieve multimodal reasoning and understanding.

Instruction tuning is emphasized as a crucial step, ensuring models can:

* Follow multimodal instructions (e.g., text and image combined).
* Handle diverse tasks with minimal additional training.

**6. Multimodal Agents: Chaining Tools with LLM**

Here, the focus is on combining multiple specialized tools (like vision and language models) into cohesive multimodal agents. Examples include:

* **Visual ChatGPT**: Integrates ChatGPT with vision models to enable interactive image analysis and generation.
* **MM-ReAct**: Chains tools to create a pipeline for multimodal reasoning.

This approach highlights the potential for developing sophisticated AI systems that leverage the strengths of different modalities.

**7. Conclusions and Future Trends**

The final section summarizes key advancements and looks ahead to the challenges and opportunities in building general-purpose multimodal assistants. Trends like increasing interactivity, improved alignment with human intents, and enhanced scalability are emphasized.