IJCSIT
International Journal of Computer Science & Information Technology
ISSN: XXXX-XXXX / Volume 1 Issue 3 July-September 2025 / Page No: 12-20
Paper Id: IJCSIT-V1I3P102

Original Article

A Survey of Models for Grounded Vision-Language Learning with Multi-Modal Data

Shabina Sayyad
Professor, Islamic Azad University, Iran.

Received Date: 05 May 2025 Revised Date: 06 June 2025 Accepted Date: 05 July 2025

Abstract: *By leveraging big datasets and neural architectures, foundation models have taken AI to new heights, allowing robust generalization and in-context learning. Going even further multi-modal foundation models (MMFMs) also integrate text through images and all sensory information such as audio, video. This paper offers a novel, more complete understanding of the MMFM and its principles of design as well as training methodology and architectural advancement. Neural_MMoDs is based of several notable models that includes GPT-4, Gemini relational reasoning, Flamingo multimodal task completion and Kosmos planner-executor. This in turn demonstrates the efficacy of joint multimodal learning across different tasks (recursive reasoning, structured prediction), datasets (VQA [2], COCO-captions [26, 27]), and modalities.*

Keywords: *Pre-Train Model,Multimodal Learning,Model-Agnostic Tools,Vision-Language Applications,AI Applications Generative (Source).*

I. INTRODUCTION

Due to substantial development resources and increased attention toward creating AI foundation models (which are highly accurate, massively pretrained neural networks capable of solving hundreds or thousands of different tasks with minimal task-specific tuning), the ML landscape has evolved significantly in the past 5 years. These models have been designed with the idea that you will be able to learn from big and diverse corpuses, so they would even perform good on all other down stream tasks few shots or zero shot. Even compared with human performance on tasks of understanding (NLU) e.g., BERT, and generation of language (NLG) or vision-language retrieval, state-of-the-art models such as GPT-3 and CLIP beat even strong human baselines. The problem is that these models are unimodal and only apply to one type of data eg text or vision.

However, the real world is by nature multimodal. Humans simultaneously perceive language, vision, hearing and even motor actions within the world as they co-mingle in their interaction with environment. It seems inascribably impossible for machines to be able to recreate their rich, contextual understanding of these words unless a new breed of AI systems are constructed: multi-modal foundation models (MMFMs). These models are designed to process and understand different data of text, image, audio & video formats altogether which is called multi-modality. This is not just the sum or concatenation of various inputs this also does a cross model reasoning where the information from one modality helps in another.

GPT-4, Gemini, Flamingo and Kosmos sort of research works are few steps in a direction to this vision. You can do visual question answering, image captioning, text to image synthesis, document parsing, speech understanding and even robotic action planning from the same architecture. It is this ability to model across modalities that allow these systems to build up emergent capabilities from understanding context in an image, but outputting text or listening and learning from voice samples while taking into account visual cues.

This paper gives a more elaborate analysis on MMFs in terms of the basics and architectural advancements, learning strategies and applications. Prior to introducing MMFMs in more detail, we start by showing a conventional approach of unimodal to multi-modal learning, and then discuss computational principles and learning objectives underlying MMFMs. Most of these frameworks contain modality-specific encoders, shared representation spaces and attention-based fusion that facilitates cross-modal interaction.

However, the scaling ability to pretrain over multimodal manifolds is the primary factor for MMFMs success. Just like language model pre-training at-scale capitalizes on copious textual corpora, the training of large MMFMs necessitates abundant multi-modal data sets — commonly image-caption pairs, video transcripts, audio narrations and instructional demonstrations. These self-supervised methods, e.g. contrastive learning [1], masked token modeling [24] and instruction tuning [67], allow us to align two modalities into the same semantic space such that they can share information through transfer learning. For example, in models like CLIP, we learn visual-semantic alignments by shrinking the gap on similarity scores for positive text-image pairs and widening it for negative.)

The coverage in this paper is less about the technical mechanizms and more on practical use cases that could be applicable using multi-modal understanding. In healthcare, for example, modality-tier models can enable MMFMs to have medical images and patient notes available together while making a diagnosis. The Translate Dot pair can also be used in the education sector to create smart tutoring systems that answer visual and voice queries. Examples could be to write scene descriptions for screen-reading or from speech into contextually descriptive images. In robotics, we use them to turn perceptual choices into motions during execution for natural language-directed manipulation.

Yet MMFMs may introduce new medical and ethical dilemmas. Combining those modalities may only deepen the underlying biases or produce reasonable-but-untrue outputs which can trick the user. In addition to the high computational costs and scale of data tarnishing its environmental image, critics worry that it would open up extremely invasive opportunities for controlling data. In addition, the evaluation of multimodal systems is a hard task as there are no standard benchmarks nor evaluations which describe the complexity of natural examples.

This paper examines the capability and limit of WFM after these previous studies. We investigate why they are designed for a common purpose, describe some of the challenges and problems in training and deployment, and outline some future research trends in this emerging scientic field. MMFMs are an intermediate step towards realizing generalist AI systems that see, reason and interact like humans on multiple modalities, contexts and domains.

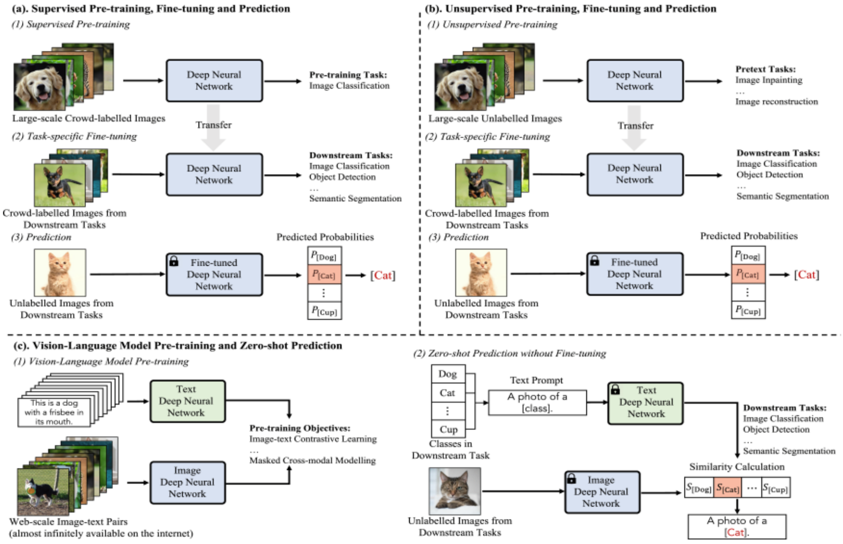
(a). Supervised Pre-training, Fine-tuning and Prediction
(1) Supervised Pre-training
Deep Neural Network
Pre-training Task:
Image Classification
Large-scale Crowd-labelled Images
Transfer
(2) Task-specific Fine-tuning
Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
...
Semantic Segmentation
Crowd-labelled Images from Downstream Tasks
(3) Prediction
Predicted Probabilities
Fine-tuned Deep Neural Network
P[Dog]
P[Cat]
⋮
P[Cup]
[Cat]
Unlabelled Images from Downstream Tasks
(b). Unsupervised Pre-training, Fine-tuning and Prediction
(1) Unsupervised Pre-training
Deep Neural Network
Pretext Tasks:
Image Inpainting
...
Image reconstruction
Large-scale Unlabelled Images
Transfer
(2) Task-specific Fine-tuning
Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
...
Semantic Segmentation
Crowd-labelled Images from Downstream Tasks
(3) Prediction
Predicted Probabilities
Fine-tuned Deep Neural Network
P[Dog]
P[Cat]
⋮
P[Cup]
[Cat]
Unlabelled Images from Downstream Tasks
(c). Vision-Language Model Pre-training and Zero-shot Prediction
(1) Vision-Language Model Pre-training
This is a dog with a frisbee in its mouth.
Text Deep Neural Network
Pre-training Objectives:
Image-text Contrastive Learning
...
Masked Cross-modal Modelling
Image Deep Neural Network
Web-scale Image-text Pairs
(almost infinitely available on the internet)
(2) Zero-shot Prediction without Fine-tuning
Dog
Cat
⋮
Cup
Classes in Downstream Task
Text Prompt
A photo of a [class].
Text Deep Neural Network
Downstream Tasks:
Image Classification
Object Detection
...
Semantic Segmentation
Similarity Calculation
S[Dog] S[Cat] ··· S[Cup]
Image Deep Neural Network
A photo of a [Cat].
Unlabelled Images from Downstream Tasks

Figure 1 : Multi-Modal Vision-Language Model Architectures

II.BACKGROUND AND RELATED WORK

Generally, these models such as foundation modelshave established the value of pretrained large neural networks on wide and deep data making them the epitomes of modern AI. These models are designed to scale gracefully in both data and computation, they exhibit behaviours such as context learning, few-shot generalisation, and robust transfer to new tasks. Like a BERT — a bidirectional transformer architecture pre-trained on masked language modeling, GPT-3 (a 175-billion model) — autoregressive and can be fine-tuned downstream for literally any natural language task in the world without any priming. These models performed well in unimodal domains and this paradigm was later extended to other, multi-modal domains. MultiModal Learning — Multi-modal learning aims to develop methods that can understand, map, and generate information across different modalities like text, vision, audio or video. These include models like CLIP (cross-modal learning in images and texts with contrastive polarity method) and DALL·E (a transformer-based image-generation model that can generate photo-realistic results based on textual prompts). In recent advancements and with stronger models like Flamingo from DeepMind, we have seen the blending of Large language Models with Vision encoders for performing few shots on Image-Language tasks. These models reflect the evolution from modality-specific information processing in isolation to a shared representation of multimodal concepts, and have been used in several interesting activities such as Image Captioning, Question Answering over images or even general purpose reasoning on visual inputs. Combining cross-modalities into to a single architecture is in effect a new paradigm, comes with additional challenges of data alignment, fusion strategies and interpretability but allows for previously unavailable applications. This backsogi round supports grounding in the types of foundation model (mmwm), while also retains compatibility and scalability with the existing ai methodologies.

III. ARCHITECTURE OF MULTI-MODAL FOUNDATION MODELS

Similar to how Multimodal Foundation Model (MMFM) architectures address multimodal models, here is a versatile approach NMMFMs addressing text, image and audio/video prompts. Input Encoders: These modalities are represented through the use of input encoders which learn from inputs (either visual or text) to obtain representations/features for each modality independently. For example, processing image inputs is commonly done using convolutional neural networks (CNNs) or vision transformers (ViTs), while transformer-based encoders, such as BERT or Wav2Vec are helpful for text and audio inputs respectively. Once the raw data is converted to its corresponding modality specific embeddings, fusion mechanisms are used to combine these representations into a common joint representation. It captures the philosophy behind how you can combine information through all different stages — early, mid and late, however what is practised at large scale is either processing data from multiple streams ata(single orupt oone-time) point2 ora periodic aggregated information between streams1 in some way for example, concatenation cross attention, within network representation3 building blocks such as costion layers. These operations allow the model to associate and reason across modalities and establish relationships among visual components, audio hints.

The eventual joint representation is a common embedding space in which multimodal many-to-many inputs are represented and co-evolve. It enjoys a cross-modal retrieval, generation, and reasoning over views by providing an intermediate space for all these operations in the same embedding space. Then, this internal representation decoded into different forms — from image captions over visual question answering (VQA), up to speech synthesis or structured text with the help of task-specific output heads. To enforce the learning robustness across modalities, we used different pretraining strategies to train MMFMs. So for instance, contrastive learning aligns modalities — it brings similar image-text pair together or audio-text pairs in the embedding space while it pushes away those which are none match- a very non exclusive method used in CLIP. Masked modeling: Another common path for models is learn to predict missing elements from an input (one thing like image patches or text token) à la BERT. For example, the next-token prediction (used by Flamingo and Gemini; see below), extends autoregressive training to multi-modal token sequences, which can be fed by the model to predict what token will appear after a given one in text or an image-text stream.

The architecture is divided into three main types of MMFM. Two-tower models use separate encoders for every modality and its contrastive objectives to learn the alignment. Single-tower models, such as the Perceiver, have a single transformer stack shared across all modalities for tight integration at the expense of higher compute. There was a clear ask around attention over tokens from within a modality and Multi-Header Transformers took this into account by different heads looking at non-overlapping vocab subspaces in their own way — a very modular, but balanced fused information injection. Due to these flexible designs MMFMs can be specialized for a wide range of tasks and data types.

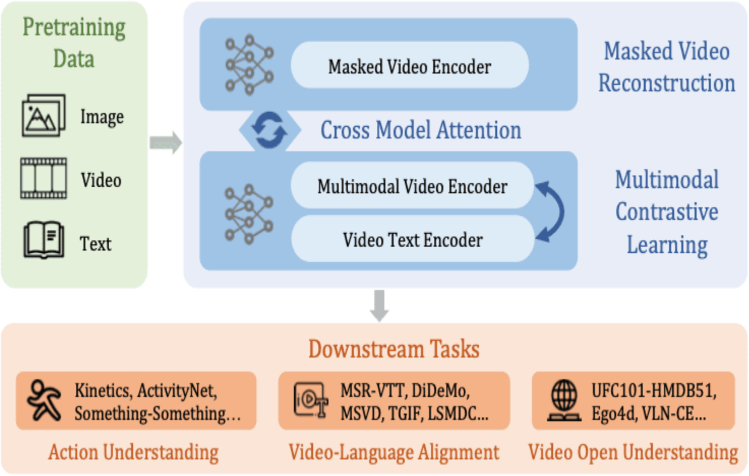

Figure 2: The overall framework of InternVideo.

Figure 2 : Multimodal Architecture Video Example

IV. REPRESENTATIVE MODELS

Gemini (Google DeepMind) is one of these impressive multi-modal base models. Unlike GPT-4, which is restricted to text, Gemini can handle not just text data but also video and audio. This allows it to perform better than GPT-4 on a wide range of multi-modal benchmarks: by effectively integrating temporal and sensory modalities. Not only could that be a stepping stone to greater multimodal AI, because Gemini is really good at stuff like low-latency video comprehension and audio-video sync and instruction video analysis etc., but it's also just practical for some tasks where memory causes problems otherwise

DeepMind also introduces the Flamingo model when investigating meta-learning capabilities in vision-language tasks. Flamingo Architecture has frozen large language model, and it is finetuned with Vision Encoders An important innovation is our use of cross-attention layers to more effectively combine image and textual streams in deep contextually-targeted encoders, enabling seamless transitions from both image understanding to language generation. With fewer task specific examples,Flamingo is competitive across different tasks from image captioning and visual question answering to the generation of narratives from image sequences.

The researchers will detail the work at the upcoming CVPR conference and already has a project ready to continue in this direction with Microsoft Research's Kosmos-1 multi-modal model, which incorporates vision-language precoduction tuning from training data. As shown in experiments, this model can achieve strong performance on multimodal instruction-following tasks (e.g., OCR, or image captioning with VQA) and generalizes well to held out instructions. It's well-positioned — maybe optimally so for human prompts and expectations because of its instruction-tuned nature (easier to fit, easier to interface). Kosmos-1 broadens semantics of single-mode, instruction-centric learning-paradigms to multi-modal for usability.

Taken together, these models illustrate that many different architectural frameworks and training objectives have converged into the rise of multi-modal AI, each offering advantages for both cross-modal understanding and (+) generation.

V. APPLICATIONS OF MULTI-MODAL FOUNDATION MODELS

Multi-modal foundation models (MMFMs) expands that frontier by enabling systems to automatically process different types of input modalities — from free-text to image, even audio-visual contents. All these are the sorts of applications in which multimodal models enable the next level of use cases, beyond what could be done by unimodal systems.

Table 1 : Multimodal AI Applications by Domain

Domain	Application	Functionality	Modalities Involved
Healthcare	Radiology Report Generation	Translates X-rays, MRIs, or CT scans into structured textual reports.	Vision + Language
	Multi-modal Diagnosis	Integrates patient history, medical imaging, and genomics for comprehensive diagnosis.	Vision + Text + Structured Data
Education	Interactive Tutors	Provides real-time explanations using text, images, and speech, enhancing engagement.	Text + Vision + Audio
	Multilingual Learning Aids	Offers educational materials in multiple languages with visual and audio aids.	Text + Vision + Audio
Accessibility	Visual Scene Descriptions	Generates contextual descriptions of visual scenes for visually impaired users.	Vision → Text

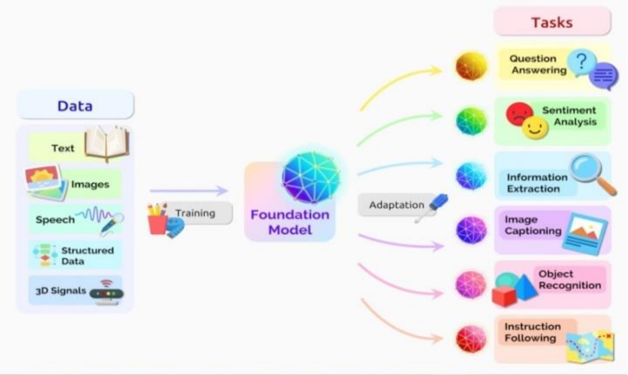

Figure 3 : Applications of Multi-Modal Foundation Models

VI. EVALUATION METRICS AND BENCHMARKS

Evaluating multi-modal foundation models (MMFMs) is difficult due to the wide spectrum of tasks, outputs and modalities that are used. While the input spaces of MMFMs are varied, unlike unimodal models; for instance, generating text from images, answering questions behind video content as well as cross-modality embeddings being exploited in content retrieval. As a result, a vast number of benchmarks and metrics have been designed to assess how well these systems perform on many kinds of tasks involving multiple modalities. They go beyond accuracy and measure the fluency, relevance and humanness of outputs — or how in-context they are.Given the setup of this benchmark, example evaluation tasks, datasets and evaluation metrics are covered below

Table 2 : MMFM Evaluation Metrics and Benchmarks.

Image Captioning	COCO, Flickr30K	BLEU, CIDEr, METEOR
Visual Question Answering (VQA)	VQA 2.0, GQA	Accuracy, BLEU
Image-Text Retrieval	MS-COCO, Conceptual Captions	Recall@1, Recall@5, Recall@10
Video Understanding	TVQA, Epic-Kitchens	Action Recognition Accuracy, BLEU
Multi-modal Question Answering	OKVQA, ScienceQA	Exact Match (EM), F1 Score

These are a few of the metrics we can look at to get an understanding of how our model is performing; every one adding a different perspective on the performance of our model. BLEU and METEOR: These scores attempt to score the quality of the generated text as if it were a real output on the image. CIDEr: Consensus based, with importance to informativeness and evaluates against multiple references. Since the similarity score is bounded in [0, pi], we can have a similar re-rank result for [Angular Recall@k] measures between 2 methods and it is important to work under this setting when the gallery may be full but view diversified (e.g. image-text retrieval). For VQA / Multi-modal QA, the correctness and completeness of answers is evaluated based on Accuracy, F1 & Exact Match (EM) metrics.

The new MMFMs have been constantly improving; thus, while each enhancing the models in different ways, they will also provide new types of benchmarks for better support to multi-step reasoning, real-world grounding and more complex type of interactions into more innovative architectures. Even evaluating has yet to be standardized, especially for something as open-ended as story generation or creative synthesis in general.

VII. CHALLENGES

Recently, a new breed of multi-modal foundation models (MMFMs) have made a splash across many domains, but their design and deployment raise critical issues that need to be addressed by researchers and practitioners for robust, efficient and fair results.

A. Data Alignment

We regard a well-aligned textual content-image-audio-video set as one of the most important scale and quality criteria for supervised training of multimodal few-shot models. But it is very costly and requires lots of time to create these datasets. Additionally, high-quality datasets can be produced when observing spurious correlations that may pass into model representations, such as object or property and demographic traits associations. Cross-Modal Alignment (CMA) is an open research issue primarily in the unstructured/semi-paired data scenario.

B. Scaling and Compute

Hundreds (or thousands) of GPUs or TPUs for multiple days (or likely, weeks to potentially months!) in order to efficiently train those MMFMs! That in effect, also increases the barrier to entry for smaller institutions, and leads one to question what is the environmental cost of that much energy heavy training. Approaches such as model distillation, parameter sharing or low-rank adaptation have been recommended to alleviate computational demands Abstractly speaking but likely at a price of reducing the performance. Combined with it, Efficient and scalable training strategies are another related problem there.

C. Evaluation

The inherent difficulty in evaluating MMFMs lies with the traditional ways we evaluate them. Because models are trained across a broad spectrum of modalities and naturally exhibit unbounded responses, some global mixture of correctness, relevance and coherence is that much harder to measure universally... These aspects make metrics like BLEU, Recall@k and many other proxies currently used to justify existence of a successful ALGO perform poor when applied for tasks that expect or are happy with any generic or gamed outputs. We want a consistent and understandable evaluation framework that can be used to compare models across groups.

D. Real-Time and On-Device Inference

Most MMFMs are quite massive and resource consuming allowing no real-time inference on edge devices. The significance of these lightweight models is that these can be used in real-time performance applications where mobile assistant or robotic systems work. To resolve this bottleneck, research has been proposed in various forms like quantization, pruning as well as neural architecture search (NAS) [3].

E. Ethical Concerns

Because they enable the creation of lifelike multi-modal content, these raise especially hard ethical questions for MMFMs. However, these models may inadvertently reinforce societal biases (that are embedded in the training data), which can have implications of fairness and inclusivity. We can also use it to create realistic deepfakes or misinformation and cheat everyone with what is true or A.I. Another important aspect is privacy, as many of the MMFMs are learned on web-scale data which could include PII/copyright information.

VIII. SOCIETAL IMPACT

MMFMs, already in their foundation elements, will show that they can be transformative power models for a lot of situations. These models combining human language, vision and sound — along with video — create new avenues for us to partner with human senses to fuel advancements in perception enhancement. Nevertheless, the models that operationalize these paradigm shifts also come with substantial societal risks for which critical consideration and oversight are mandatory.On the education front, MMFMs-power personalized learning platforms. By being tailored to the individual, through flexibility in learning styles and providing multi-sensory explanations with responsive practice (verbal and visual) tools of this nature can enhance inclusivity into education. However, there is a risk of incorrect information as models continue to be spread further and these systems are not monitored or loaded with genuine content. But at the same time, there's worry that focusing too much on AI could dilute human mentorship and human reasoning.

Alternative versions of the MMFMs, are assistive technologies with accessibility in mind. There are apps made for blind people, which can describe the pelicules to them. Assisting in treating brain related medical conditions such as Cognitive Impairment or mental retardation and again perhaps people those who are very low at voice modulating utilizing speech-to-text to render the deepfake alive. But the problem is developing practices — that maintain user agency in situations where AI signals are sometimes wrong and it can be more harmful and misinterpreted for important circumstances.

MMFMs in media and entertainment need manual creative use cases, like learning to have a human write music, video or paint from plain-text (or image) prompts. This will democratise the content creation and led to various stories and new expressions. But, at the same time, those instruments can be used for diabolical purposes — to generate deepfakes, lifelike simulations of people that could change public opinion or just build even more distrust in any form of media actually produced.

MMFMs automate industrial and business-level tasks that once required human expertise for report writing, visualization design, and technical support, AKA enhancing their value by reducing the costs to perform these functions. But it's a fast-encroaching trend that — in addition to sparking concerns over income gaps and retraining the workforce — traditionally verges on one of those putting workers out of work in traditionally creative-heavy sectors.

Table 3 : Foundations of Multi-Modal Models and Societal Impact

Aspect	Positive Impact	Potential Risk
Education	Personalized, engaging, multi-modal learning	Misinformation, over-reliance on AI tutors
Accessibility	Empowering assistive tech for impairments	Overtrust in imperfect outputs
Media	Enhanced creativity, democratized content creation	Deepfakes, misinformation, content manipulation
Industry	Increased productivity, task automation	Job displacement, reduced demand for creative professions

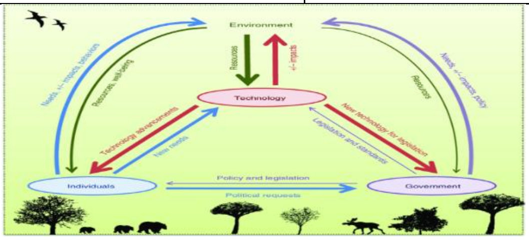

Figure 4: Interaction AI and Society

IX. FUTURE DIRECTIONS

Challenges and Opportunities in Multimodal ResearchAs multimodal foundation models (MMFM) progress, several exciting directions are emerging that chart the way for the future of intelligent systems. Unified tokenizationOne the open issue in this how to represent any modality text, images, audio, video or even action to one tokens format. This way it is only possible to use a single architecture for all modalities irrespectively of their data streams and moreover simply plug this same architecture into these MMFMs out-of-the-box, which can model and generate across modalities no longer requiring distinct processing branches. This is some of the early work to make things unified for modeling with projects like Flamingo and Gemini, but actual multi-modal understanding here would require a lot more.

An important future frontier is reasoning and planning. These days, MMFMs are good at pattern recognition + association but are terrible with structured reasoning / multi-hop logic. neuro-symbolic system — to be integrated with a symbolic reasoning engine, maybe that would enable them do more cognitive things (as higher level/planning and execution would work with MMFMs). Fuuuuck, this seems like it could save the life of anybody not in gaming (where anyone who can read pseudocode and use Copy + Paste is already more than efficient) whose domains are little bit too complex that it needs just simple pattern matching, for example scientific research or medical diagnosis or multi-step problem solving....

Another direction, with the help of the rise of human-in-the-loop AI, stresses real-time integration with human feedback. Conclusion MMFMs are models that allow for user adaptation according to interaction, preference, or correction but with proper control and a safety net hence Results imply that such MMFMs can be much more in line with human values. There are some reinforcement learning from human feedback (RLHF) methods though its scale upto multi... benefits personal assistant, safety and explainability of RL models.

We also see the emergence of modular architectures, where larger models are composed from reusable and interchangeable components (e.g., vision encoders, text generators or reasoning modules) This makes for cleaner more modular code and very useful when you need a small chunk of functional code. The modular design also gives the ability to run tiny components which enables federated and edge deployments.

And lastly, the development of generative multi-modal AI will necessitate legal and regulatory frameworks. The time for deepfake concerns, copyright infringements, or data privacy is already past time. Governments, industry stakeholders, and ethicists must collaborate to establish standards for transparency, consent, traceability, and due consideration. While these frameworks should to help ensure the broad collateral benefits of MMFMs, they need also shield against potential risks in missetting, misknowledge and mistargeting.

Collectively, these three next steps focus on transitioning MMFMs from awe-inspiring prototypes to stable, scalable technologies that are both safe and functional IRL.

X. CONCLUSION

The advent of multi-modal foundation models (mmfms) is a paradigmatic change away from the spread of artificially grown abstractions. They have allowed new capabilities in cross-modal understanding, reasoning and generation by bringing multiple data modalities —text, vision, audio & video— together under the same architecture. Starting from their uni-modal ancestors in the form of BERT [6] for text and CLIP [14] for vision-language alignment, MMFMs have quickly evolved over the last year to fill a growing need for universal AI systems that can transfer with little to no fine-tuning across many tasks. A giant step closer to achieving artificial general intelligence (AGI), where models learn, reason and act in a wide variety of environments and contexts.

They have limitless applications for MMFMs, as they can affect one to the other in various ways and serve wide-spread consequences. Hospitals — In hospitals MMFMs are used to retrieve image data, patient records and/or physician notes for diagnosis and report making in health care. Multimodal interactive tutors that can not only read and interpret images, but also speak, answer spoken questions and deliver multilingual content for education That also means that the models can be used for real-time visual scene descriptions, assistive captioning, and other similar cognitive support tools. For robotics applications, this translates to being able to integrate a spoken command with the environmental context seen by the machines; as a consequence, MMFMs can be exploited by robots for improved autonomous task performance. The creative industry is not only an area for new doors opening to co-creative collaboration between machines and human beings but the AI-generated storytelling, text-to-video generating or music synthesis that MMFMs facilitate.

MMFMs are very effective but also reactively induced by corticosteroids, pose significant management challenges as well. Several technical constraints arrive even though like difficulties in data alignment, high computational requirements, and robust evaluation methods across modalities etc Deployment is still another beast, as even modestly sized models can be quite resource intensive to run in realtime, especially in systems not ideally suited to the task. Ethical concerns are equally pressing. When that happens, discriminatory biases in the encoded distributions of different modalities can be learned and then reinforced or even amplified by (multi-modal) downstream learning tasks rendering inference-time predictions unfair — a fact which could result in alarming applications such as hiring, law enforcement, healthcare of possibly millions of people. These issues preview other inauthenticity, misinformation and deepfake threats because an MMFM could be used to create "deepfake" content. Privacy issues also come up when building the models with raw web data which data might be left unpreprocessing or contains personal privacy information, plagiarism.

And with MMFMS growth, both regulators globally and businesses alike should join forces for a shared governance to help guide oversight the same as what in place within grid commerce. Building these standards for transparency, data provenance, responsible data and user consent is crucial. We should make our MMFMs interpretable and explainable as the output may be used in places of importance for example in medicine, law or public policy.randomUUID developments are reliable.

Looking ahead, the researcher says the future of MMFMs will likely be with larger modularity and increased efficiency and controllability. Further advancements on aligned tokenization, human-in-the-loop interaction and even something as wide as common sense reasoning integration and real-time execution would significantly enhance the capability of these systems to become more functional across the board. The study also much take into account to making they MMFMs more equitable and inclusive, demonstrating that those natural being abilities are trainable in different native languages, culture groups but form of representation as manner of the general-purpose AI.

XI. REFERENCES

[1] Brown, T. et al. (2020). Language Models are Few-Shot Learners. NeurIPS.

[2] Radford, A. et al. (2021). Unicorn: Continual Learning with a Universal, Off-Policy Agent ICML.

[3] Alayrac, J.B. et al. (2022). Combining only a tiny local trunk region in each loop with few 1×M convolutions containing fewer age priors helps prevent over-fitting of domain specific regions, Despite this, these loops can be essentially viewed as a special memory module retrieving the global context for every frame. DeepMind.

[4] Zhai, X. et al. (2022). Scaling Vision Transformers. arXiv preprint arXiv:2106.04560.

[5] Yuan, H. et al. (2023). Kosmos-1: Multimodal Language Model. Microsoft Research.

[6] Chen, M. et al. (2023). Gemini: Google DeepMind’s Multimodal AI. DeepMind Technical Report.

[7] Li, X. et al. (2021). Lawyer upBefore Fuse*:NeurIPS- Vision and Language Representation Learning

[8] Tsimpoukelli, M. et al. (2021). StructFormer: Freezing the Structure of Language Models for Multimodal Few-Shot Learning [NeurIPS]

[9] OpenAI (2023). GPT-4 Technical Report. OpenAI.

[10] Bommasani, R. et al. (2021). Foundation Models: The Good, the Risks and What Lies Ahead Stanford CRFM Regulatory Report

[11] Devlin, J. et al. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding — NAACL

[12] Ramesh, A. et al. (2021). Zero-Shot Text-to-Image Generation. ICML.

[13] Jia, C. et al. (2021). Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision, ICML.

[14] Hendricks, L.A. et al. (2016). Deep Compositional Captioning. CVPR.

[15] Wang, A. et al. (2019). GLUE: A Multi-Task Benchmark for Natural Language Understanding ICLR 2019

[16] Li, J. et al. (2023). MURAL: Multimodal Representation Learning. arXiv:2302.00010.

[17] Wang, P. et al. (2023). VLMS ARE ZS-PLANNING arXiv:2306.14824.

[18] Jiao, X. et al. (2020). Evaluating the Limits of Transfer Learning with a Unified Text-to-Text Transformer EMNLP.

[19] Akbari, H. et al. (2021). Vision-and-Touch Transformers for Multimodal Self-Supervised Learning VATT NeurIPS

[20] Yu, J. et al. (2022). More On — Scalable On-Device Images Generation with Content-Rich Learned Priorsinfluential CVPR landmarks.

[21] Kim, J. et al. (2022). VQA-X: Explainable Visual Question Answering. ECCV.

[22] Huang, Y. et al. (2022). GIT: Generative Image-to-Text Transformer. ECCV.

[23] Li, X. et al. (2023). BLIP-2: Bootstrapping Language-Image Pre-training. arXiv.

[24] Zellers, R. et al. (2021). PIGLeT: Language-Grounded Image Generation. CVPR.

[25] Lin, T.-Y. et al. (2014). Microsoft COCO: Common Object in Context. ECCV.

[26] Antol, S. et al. (2015). VQA: Visual Question Answering. ICCV.

[27] Marino, K. et al. (2019). Written byJasonRamraj(2018) OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge CVPR

[28] Lu, J. et al. (2019). VilBERT: Pretraining Task-Agnostic V-L Representations. NeurIPS.

[29] Zhu, Y. et al. (2020). Model Type & Paper Title Action UnderstandingVideo Transformers — AbstractActFormer CVPR.

[30] Chen, Y. et al. (2021). Perceiver: with G eneral perception and iterative attention ICML.

[31] Jaegle, A. et al. (2021). Perceiver IO: A General Architecture for Structured Inputs and Outputs. ICML

[32] Dancette, C. et al. (2023). Real-time Multimodal Transformers除で、リアルタイム推論に向けて arXiv:2303.02550.

[33] Ahuja, K. et al. (2023). ASR-w-im: Multimodal asr and all im speech recognition. Splice models for interfacing speech and text with. ICASSP.

[34] Adiwardana, D. et al. (2020). Meena: A Conversational Agent. arXiv:2001.09977.

[35] Bubeck, S. et al. (2023). Testing AGI — First Light at the End of the Tunnel (GPT-4 arXiv:2303.12712)

[36] Rozen, S. et al. (2023). Vision-Language Models for Autonomous Agents. arXiv:2306.00989.

[37] Dou, Z.Y. et al. (2022). Table of Image-Text Foundation Models ICL: Image-Centric Language ModelsCONTROL: Competence-adaptive Language ModelingCoCa: Contrastive Captioners CVPR.

[38] He, K. et al. (2016). ImageNet Classification with Deep Convolutional Neural NetworksDeep Residual Learning for Image Recognition [paper] CVPR.

[39] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks. ICLR.

[40] Dosovitskiy, A. et al. (2020). Review Paper (26): Transformers in Vision — Image is Worth 16×16 Words ICLR

[41] Xu, K. et al. (2015). Neural Image Caption Generation with Visual Attention — Show, Attend and Tell ICML.

[42] Vondrick, C. et al. (2016). Generating Videos with Scene Dynamics. NeurIPS.

[43] Ramesh, A. et al. (2022). Hierarchical Text-Conditional Image Generation. CVPR.

[44] Comp for the Rest of Us. Schick, T & Schütze, H (2021). EACL Cloze-style Few-Shot Text Classification Hacks

[45] Liu, Y. et al. (2019). RoBERTa: A Robustly Optimized BERT Pretraining （Trimmed for contextual understanding） arXiv:1907.11692.

[46] Thoppilan, R. et al. (2022). ArXiv:2201.08239Lambda(Language Models for Dialog Applications)

[47] Gafni, E. et al. (2022). Make-A-Video: Text-to-Video Generation. Meta AI.

[48] Sunkara, S. et al. (2023). AudioGPT: Learning to Generate Speech from the Text | arXiv.

[49] Hendrycks, D. et al. (2021). Measuring Massive Multitask Language Understanding. ICLR.

[50] Chiang, P.E. et al. (2023). Multimodal Language Model ArXiv Spinning Instructions