# Introduction

LLM – can only handle text, however we want AI to be able to handle all modalities of data.

Vision Language Models (VLM) - combine visual and textual information, showcasing a remarkable proficiency in comprehending and generating content that involves both images and text. These models excel at image captioning, querying based on image, and image generation based on prompt.

# Vision Language Models

This Paper categorizes VLMs into 3 categories:

1. Vision-Language Understanding Models
   1. Designed for the interpretation and comprehension of visual information in conjunction with language.
2. Text Generation with Multimodal input Models
   1. These models process multimodal input to generate text outputs
3. Multimodal output with multimodal input Models
   1. These models are capable of accepting multimodal inputs and generating multimodal outputs.

Comparative Analysis

They authors of the paper have conducted a through test by testing the models on 10 popular benchmark datasets. Additionally, they have evaluated the perception and cognitive abilities of these VLM using Multi Modal Evaluation (MME) benchmark.

## Vision Language Understanding

CLIP, developed by OpenAI, excels in understanding visual concepts via natural language, showing robust zero-shot performance across various image classification benchmarks, surpassing fine-tuned vision models. However, it faces challenges with abstract tasks, fine-grained classification, generalization, and sensitivity to specific wording.

AlphaCLIP enhances CLIP by adding an alpha channel for improved focus on specific regions, maintaining recognition accuracy but struggling with multiple objects and attention amplitude.

MetaCLIP improves on CLIP by refining data using metadata from CLIP's concepts, enhancing data transparency and outperforming CLIP in benchmarks with a large dataset of 400 million image-text pairs from CommonCrawl.

GLIP advances CLIP by focusing on object-level alignment in language-image representations and redefining object detection as a vision-language task. It excels in scalable pre-training on semantic-rich data, enabling robust zero/few-shot capabilities and outperforming CLIP in image captioning and object detection tasks.

Vlmo features a dual and fusion encoder setup in a modular Transformer with a Mixture-of-Modality-Experts Transformer, enhancing versatility and flexibility. It excels in both vision-language classification and image-text retrieval. The model's stage-wise pre-training strategy leverages large-scale multimodal data to achieve top performance in tasks like visual question answering and image retrieval.

ImageBind aligns embeddings from various modalities to image embeddings, utilizing web-scale image-text data and models like CLIP for zero-shot recognition across modalities. It requires minimal training, leveraging large-scale, naturally paired, self-supervised data across diverse modalities (audio, depth, thermal, IMU) for robust zero-shot classification and retrieval. It surpasses specialist models in audio benchmarks and shows versatility in compositional tasks, with potential improvements from richer alignment data and task-specific embedding adaptations.

VideoCLIP is a unified model trained for zero-shot video and text understanding using a contrastive learning framework that doesn't rely on labels for downstream tasks. It innovates by using loosely temporally overlapping video-text pairs for positives and a retrieval-based method for selecting hard negative pairs. The model uses contrastive loss and overlaps video-text clips to strengthen associations. Evaluated on various tasks, it achieves state-of-the-art results on video language datasets like Youcook2 and surpasses prior works and some supervised methods in zero-shot video-text understanding.

VideoMAE adapts the masked autoencoder framework to video by introducing a video tube masking strategy, enabling efficient learning from smaller datasets (3k-4k videos) using a Vision Transformer with joint space-time attention. Despite higher energy consumption during pre-training, it excels in tasks like action detection and offers potential for further improvement with dataset expansion and additional data streams.

## Text Generation with Multimodal Input

GPT-4V: GPT-4V is an enhancement of OpenAI's GPT-4, designed to analyze both text and image inputs by training on a large dataset of text and image data. Prior to its broad release, OpenAI conducted extensive safety evaluations, including red team assessments and implemented mitigations, to ensure the model's reliability and safety.

LLaVa: LLaVA is an open-source multimodal framework that combines a vision encoder from CLIP with language-only GPT-4, enabling it to process visual and textual information for enhanced multimodal understanding. In early tests, LLaVA achieved an 85.1% relative score on a synthetic instruction-following dataset and set a new state-of-the-art accuracy of 92.53% on Science QA tasks.

Flamingo: Flamingo utilizes an advanced architecture with interleaved cross-attention and frozen language-only self-attention layers to adeptly handle mixed visual and textual data, using a Perceiver-based approach to process inputs like videos into a fixed set of visual tokens. This model leverages large-scale multimodal web corpora for few-shot learning, significantly outperforming models fine-tuned on more data and showing great adaptability in image and video understanding tasks.

PALM-E: PALM-E, an Embodied Multimodal Language Model with 562B parameters, integrates language comprehension with sensor inputs for applications like robotic manipulation and visual tasks. It excels in embodied reasoning and visual-language tasks by transferring knowledge across domains. However, it struggles with low-level language-conditioned policies in robotic tasks, highlighting the need for improved self-supervised entity-centric labeling to enhance performance.

BLIP: This Vision-Language Pre-training (VLP) framework excels in understanding and generating tasks, overcoming noisy training data issues with a Multimodal Mixture of Encoder-Decoder (MED) architecture. It uses a unique mix of pre-training objectives and improves data quality through Captioning and Filtering (CapFilt), leading to superior performance in image-text retrieval and captioning tasks.

BLIP-2: This model introduces a cost-effective VLP strategy with a static image encoder and large language model linked by the Querying Transformer (QFormer), which serves as a dynamic bridge enhancing the interoperability between visual and language modalities. The QFormer focuses on vision-language representation initially, followed by vision-to-language generative learning, aiming for efficient use of fewer parameters.

InstructBLIP: Focuses on instruction-aware visual feature extraction to tailor feature extraction directly to provided instructions, leading to state-of-the-art zero-shot performance. It particularly excels in specialized tasks such as ScienceQA IMG, demonstrating high accuracy and qualitative advantages in visual scene understanding and multi-turn visual conversation.

KOSMOS-1: A Microsoft VLM trained on a vast multimodal corpus, using a Transformer-based architecture to align vision with language models. KOSMOS-1 shows excellent performance in tasks requiring language understanding and generation, OCR-free NLP, and various perception-language tasks, supported by diverse data sources like The Pile and Common Crawl.

KOSMOS-2: An advancement over KOSMOS-1, this model incorporates spatial understanding by linking text spans to specific locations in images, enhancing its capabilities in object descriptions and grounding text visually. This development pushes closer towards Embodiment AI by combining multimodal perception, action, and world modeling.

MultiInstruct: Not a model but a benchmark dataset that tests the OFA pre-trained multimodal language model across 62 tasks and 10 categories, aiming to enhance zero-shot performance through extensive text-only instruction tuning. It focuses on improving model robustness across multimodal tasks and optimizing instruction-sensitive performance.

IDEFICS is an open-access model that reproduces DeepMind's Flamingo, featuring two versions with 80 billion and 9 billion parameters. It excels in image-text benchmarks like visual question answering and image captioning, utilizing few-shot learning approaches.

PaLI from Google Research integrates large encoder-decoder language models with vision transformers, achieving top results in multilingual vision-language tasks across over 100 languages. Its strength lies in its scalable and modular design, optimizing both vision and language components.

Frozen by DeepMind combines a frozen language model with a trained vision encoder, focusing on few-shot learning for rapid task acquisition. It's effective in visual question-answering and learning new objects but shows limitations in few-shot task performance compared to fully trained models.

Qwen-VL series includes models optimized for a variety of vision-centric tasks, with notable performance in image captioning, question answering, and dialogue, particularly with high-resolution and fine-grained data. Trained on multilingual datasets, Qwen-VL supports multiple languages and excels in complex scenario analysis.

Fuyu-8B by Adept AI is designed for digital agents, handling diverse image resolutions and excelling in rapid image processing tasks like graph comprehension and UI queries. It features a vanilla decoder-only transformer architecture, treating image tokens similarly to text for flexibility and efficiency.

SPHINX: An advanced VLM that unfreezes LLMs during pre-training for better vision-language alignment, utilizing weights from models trained on real-world and synthetic data. It excels in tasks like region-level understanding and human pose estimation by incorporating diverse tasks and rich visual embeddings.

Mirasol: A multimodal autoregressive model from Google that handles time-aligned (audio, video) and non-aligned (text) modalities, using segmented video-audio sequences and a Combiner for feature fusion. It emphasizes content consistency and adaptability in dynamic sequences, trained with a focus on unaligned text.

MiniGPT-4 and MiniGPT-v2: These models integrate a visual encoder (ViT) with a large language model (LLM) using a single projection layer, focusing on tasks like meme interpretation and visual question answering. MiniGPT-v2, specifically, uses task-specific identifiers and a three-stage training strategy to enhance visual grounding and multi-task learning efficiency.

LLaVA-Plus and BakLLaVA: LLaVA-Plus is a multimodal assistant that surpasses its predecessor by integrating diverse vision and language models, excelling in real-life multimodal tasks. BakLLaVA, on the other hand, offers a faster, less resource-intensive alternative to vision-augmented LLMs, combining Mistral 7B with LLaVA architecture.

CogVLM: Developed by Tsinghua University, this open-source foundation model uses a Vision Transformer encoder and deep fusion techniques to integrate visual and language features, excelling in tasks requiring free-form instructions and complex alignments.

FERRET: Focuses on spatial referring and grounding in images, blending discrete coordinates with continuous visual features. It's trained on the GRIT dataset to enhance multimodal chatting capabilities and mitigate object hallucination issues.

BARD: Developed by Google, BARD uses a reinforcement learning framework to automate machine learning model design, architecture search, and hyperparameter tuning. It's designed to be user-friendly for non-experts and supports tasks such as resume writing and itinerary planning. BARD emphasizes safety and quality in response generation and improves over time through human feedback and reinforcement learning.

LLaMA-VID: This model introduces a dual-token strategy (context and content tokens) for encoding video frames efficiently. LLaMA-VID uses a hybrid architecture combining text and visual encoders to handle videos effectively while preserving detail in images. Its training is optimized for speed and efficiency, demonstrating strong performance in zero-shot video QA tasks.

CoVLM: CoVLM enhances compositional reasoning in large language models by integrating vision-language communicative decoding. It uses communication tokens to dynamically compose visual entities and relationships, aiming to improve language generation and performance in tasks like visual question answering and referring expression comprehension.

Emu2: A 37 billion-parameter multimodal model that excels in in-context learning for multimodal sequences. Emu2 integrates visual and textual tokens with a unified autoregressive objective, showcasing strong capabilities in vision-language tasks, instruction tuning, and controllable visual generation.

Video-LLaMA: This model is designed to understand both visual and auditory content in videos, integrating pre-trained visual and audio encoders with frozen LLMs. It uses specialized Q-formers for temporal and audio information, effectively aligning audio-visual data with textual content and enhancing comprehension in multimedia contexts.

Video-ChatGPT: A novel multimodal model that improves video understanding by integrating a video-adapted visual encoder with a large language model. It focuses on capturing temporal relationships and contextual understanding, using a dataset of video-instruction pairs for fine-tuning, and performs well in zero-shot question-answering tasks. The model faces challenges in handling subtle temporal relationships and small visual details.

LAVIN: Utilizes a Mixture of Modality Adaptation (MMA) approach with lightweight adapters to efficiently adapt LLMs to vision-language tasks. LAVIN achieves competitive performance with minimal training resources, requiring only 1.4 hours and 3.8M trainable parameters, particularly excelling in tasks like science question answering and dialogue. Despite its efficiency and reduced costs, LAVIN faces challenges with incorrect responses and capturing fine-grained image details.

BEiT-3: A multimodal foundation model that excels in both vision and vision-language tasks through a significant convergence of modalities. BEiT-3 uses Multiway Transformers within a modular architecture that enables deep fusion and modality-specific encoding, performing state-of-the-art across various tasks including object detection, visual reasoning, and cross-modal retrieval. It uses a unified approach to mask "language" modeling on images, texts, and image-text pairs.

mPLUG-2: Introduces a multi-module composition network that enhances modality collaboration and reduces modality entanglement. mPLUG-2 is flexible, supporting diverse modules across text, image, and video modalities for a variety of tasks, achieving top results in over 30 downstream tasks, including new records in video QA and video captioning. Its design supports robust zero-shot transferability across vision-language and video-language tasks.

X2-VLM: Features a flexible modular architecture that integrates image-text and video-text pre-training in a unified framework. X2-VLM balances performance with model scale and enhances transferability, allowing adaptation across various languages or domains without specific multilingual pre-training. Its ability to substitute text encoders for better performance positions it as a versatile player in multimodal pre-training.

Lyrics: An extension of BLIP-2, Lyrics enhances vision-language alignment through fine-grained cross-modal collaboration, integrating local visual features from a visual refiner (including image tagging, object detection, and semantic segmentation) with language inputs. It employs a two-stage training approach that establishes comprehensive vision-language alignment targets during pre-training, followed by instruction fine-tuning to extract semantic-aware visual features. This model demonstrates robust performance across multiple vision-language tasks.

X-FM: A general foundation model featuring a language encoder, a vision encoder, and a fusion encoder. X-FM's unique training method includes halting gradients from vision-language training during language-encoder learning and using vision-language training to guide vision-encoder learning. Although computationally demanding, it outperforms general foundation models and competes with or surpasses specialized models in language, vision, or vision-language tasks.

VALOR: A trimodal pretraining model that integrates vision, audio, and language, designed to excel in understanding and generating tasks across these modalities. It utilizes tasks like Multimodal Grouping Alignment and Captioning, and is trained on datasets specifically designed for trimodal pretraining. VALOR achieves state-of-the-art performance in tasks such as retrieval and captioning across vision, audio, and audiovisual modalities.

Prismer: A data- and parameter-efficient vision-language model that uses a frozen ensemble of domain experts to minimize the need for extensive training data. It adapts to different vision-language tasks by inheriting and freezing weights from pre-trained domain experts, offering competitive performance with less data. However, Prismer struggles with zero-shot in-context generalization and adapting to new experts or partial expert ensembles during inference.

MM-REACT: MM-REACT introduces a novel textual prompt design that allows language models to process multimodal information effectively, including text descriptions, spatial coordinates, and file names. It shows promise in zero-shot experiments for advanced visual understanding, though it faces limitations such as the absence of annotated benchmarks for systematic evaluation and issues with the integration of vision experts.

PICa: Utilizes image captions to prompt GPT-3 for visual question answering, converting images into captions or tags for processing by GPT-3. This method achieves notable performance with a few-shot learning approach on the OK-VQA dataset but may miss crucial visual details due to the abstraction of images as text.

PNP-VQA: A modular framework for zero-shot visual question answering that leverages natural language and network interpretation as intermediaries, connecting pre-trained models without additional training. It utilizes informative image captions to aid pre-trained language models in answering questions, achieving state-of-the-art results on zero-shot VQAv2 and GQA datasets.

Img2LLM: Facilitates zero-shot visual question answering with LLMs by developing LLM-agnostic models that use exemplar question-answer pairs as effective prompts. It matches or surpasses end-to-end trained methods, offers flexibility with various LLMs, and incurs additional computational time which can be mitigated by optimizing prompt length.

SimVLM: A streamlined pretraining framework that adopts a minimalist approach by using large-scale weak supervision and a singular prefix language modeling objective. It surpasses previous models in various vision-language tasks without needing additional data or task-specific tailoring, demonstrating robust generalization and transfer capabilities.

VideoCOCA: An adaptation of the CoCa model for video-text tasks, utilizing generative and contrastive attentional pooling layers to achieve state-of-the-art results in zero-shot video classification and text-to-video retrieval. It processes video frames uniformly through an image encoder and employs attention-pooling layers for task modeling, excelling in video reasoning and action recognition but facing challenges in capturing subtle temporal relationships.

TinyGPT-V: A compact and efficient multimodal model designed to provide high performance with lower computational requirements. It integrates pre-trained vision modules from BLIP-2 or CLIP and performs well in visual tasks like question-answering and referring expression comprehension. TinyGPT-V is optimized for use on more accessible hardware, making it a practical solution for deploying high-performance multimodal language models in diverse applications.

ChatBridge: A multimodal language model that uses language as a bridge to connect different real-world modalities such as text, image, video, and audio. It undergoes a two-stage training process that includes aligning each modality with language and fine-tuning on a new multimodal instruction dataset. ChatBridge extends zero-shot capabilities but faces challenges with long-range video and audio content, indicating a need for improved temporal modeling.

Macaw LLM: Integrates visual, audio, and textual data through a modality module, a cognitive module using pre-trained LLMs, and an alignment module that harmonizes these diverse representations. Macaw LLM is designed for multi-turn dialogue with a new multimodal instruction dataset but is not optimized for this format and faces potential issues such as hallucination and toxicity.

GPT4Tools: Designed to enable open-source LLMs like LLaMA and OPT to efficiently use multimodal tools, addressing the high computational costs and data access issues of proprietary models like GPT-4. GPT4Tools uses self-instruction to generate datasets for visual problem-solving and employs Low-Rank Adaptation for improved tool invocation accuracy, exploring more efficient invocation methods.

PandaGPT: Enhances large language models with visual and auditory instruction-following capabilities, excelling in tasks such as image description and audio-related question answering. PandaGPT integrates ImageBind’s multimodal encoders and Vicuna’s language models, requiring only aligned image-text pairs for training. It suggests further improvements for richer multimedia content generation and addressing common language model deficiencies.

mPLUG-Owl: A training paradigm that equips LLMs with multimodal abilities through a two-stage training process, integrating a foundation LLM, a visual knowledge module, and a visual abstractor module. mPLUG-Owl excels in visual understanding, multi-turn conversation, and knowledge reasoning, demonstrating capabilities like multi-image correlation and multilingual understanding. However, it faces challenges in multi-image correlation and complex scene OCR, showing mixed performance and potential in vision-only tasks like document comprehension.

Ying-VLM: Trained on the M3IT dataset, Ying-VLM excels in following human instructions and provides engaging responses with strong generalization in unseen video and Chinese tasks. The dataset includes diverse tasks and instructions, showing that increasing task numbers and instruction diversity can significantly influence performance outcomes.

BLIVA: A multimodal language learning model designed to handle text-rich visual questions more effectively than existing models like GPT-4 and Flamingo. BLIVA integrates query and patch embeddings and uses a QFormer for instruction-aware visual features, demonstrating significant improvements in OCR-VQA and visual spatial reasoning tasks, achieving a notable performance boost in multimodal benchmarks.

LLaVA-Phi: A highly efficient multimodal assistant using the compact Phi-2 language model, LLaVA-Phi is optimized for engaging complex dialogues that blend text and visuals. Despite its smaller size (2.7B parameters), it shows impressive capabilities in visual comprehension, reasoning, and knowledge-based perception, suitable for real-time interactions such as embodied agents. The model's training includes feature alignment and visual instruction tuning.

MoE-LLaVA: Developed by a collaboration of Chinese universities and tech companies, this model uses a novel MoE-tuning strategy to address performance degradation and model sparsity in multimodal learning. By activating only the top-k experts through routers, MoE-LLaVA efficiently manages 3 billion sparsely activated parameters, achieving superior performance and reducing hallucinations compared to other state-of-the-art models. Its architecture includes vision and language components with specialized MoE blocks.

Yi-VL: An open-source multimodal model from the Yi Large Language Model series, Yi-VL excels in content comprehension and multi-round conversations about images. It supports bilingual (English and Chinese) text-image interactions and fine-grained image resolution. However, it faces limitations with visual question answering and managing content generation in complex scenes.

Moondream: A 1.6 billion parameter model created by Vikhyatk, Moondream combines elements from SigLIP, Phi-1.5, and the LLaVa training dataset. It is designed for scholarly use in AI research, emphasizing non-commercial exploration and innovation in artificial intelligence.

Shikra: A Multimodal Large Language Model that focuses on enhancing human-like referential abilities within dialogue, handling spatial coordinates in natural language. Shikra integrates a vision encoder and alignment layer with a large language model, excelling in tasks like REC, PointQA, and VQA. Future updates aim to expand its multilingual capabilities and improve coordinate representations for complex visual tasks.

BuboGPT: A Vision-Language Model (VLM) with enhanced visual grounding capabilities, BuboGPT specializes in cross-modal interactions and specific object localization in images. It uses a visual grounding module and a comprehensive two-stage training scheme to improve text-image-audio understanding. Despite some limitations in grounding QA and language hallucination, BuboGPT shows promising multi-modality understanding and grounding abilities.

ChatSpot: A unified end-to-end multimodal large language model designed to enhance human-AI interaction through interactive modalities like mouse clicks, drag-and-drop, and drawing boxes. It uses precise referring instructions to focus on specific image regions, trained with a multi-grained vision-language dataset. ChatSpot shows strong performance in region referring tasks, demonstrating minimal hallucination and highlighting its potential for improving interactive accuracy in multimodal applications.

MiniGPT5: Introduces an innovative interleaved vision-and-language generation technique using "generative vokens" to synchronize image-text outputs. MiniGPT-5 employs a two-staged training strategy that eliminates the need for comprehensive image descriptions, enhancing model integrity with classifier-free guidance. It delivers superior multimodal outputs across diverse benchmarks, particularly noted in its performance on the MMDialog and VIST datasets.

DRESS: A large vision language model that uses Natural Language Feedback (NLF) from large language models to enhance alignment and interactions. DRESS categorizes NLF into critique (identifying strengths and weaknesses) and refinement (providing specific improvement suggestions), utilizing conditional reinforcement learning for training. It produces more helpful and honest responses and learns effectively from feedback during multi-turn interactions.

X-InstructBLIP: A cross-modality framework built on frozen large language models, facilitating integration across various modalities without extensive customization. X-InstructBLIP uses high-quality instruction tuning data collected automatically and demonstrates emergent cross-modal reasoning capabilities. It outperforms captioning baselines and introduces a novel evaluation task, Discriminative Crossmodal Reasoning (DisCRn), to assess cross-modal abilities.

VILA: A Visual Language model family that emerges from an enhanced pre-training recipe, systematically augmenting LLMs towards VLMs. VILA outperforms state-of-the-art models across major benchmarks and unveils properties like multi-image reasoning, enhanced in-context learning, and improved world knowledge, marking significant advancements in visual language modeling.

## Multimodal Output with Multimodal Input

CoDi: CoDi uses a multimodal approach employing Latent Diffusion Models (LDM) for text, image, video, and audio. It features a variational encoder with BERT and GPT-2 for text, a LDM with a VAE for images, and a VAE encoder-decoder for audio tasks. CoDi creates a shared multimodal space for cross-modal generation and is trained to achieve any-to-any generation using individual diffusion models with aligned prompt encoders.

CoDi-2: An advancement over CoDi, this model employs a multimodal encoder (ImageBind) and integrates diffusion models into a Multimodal Latent Language Model (MLLM) for detailed modality-interleaved generation. It projects multimodal data into a feature sequence processed by the MLLM, enhancing generation quality and supporting multi-round interactive conversations.

Google Gemini: Gemini features a transformative architecture with deep fusion capabilities to integrate text, image, audio, and video. It outperforms GPT-4 in most benchmarks, utilizing Google’s latest TPUs for training. The model emphasizes quality, safety, and has undergone extensive safety evaluations, including Reinforcement Learning from Human Feedback (RLHF) for bias and toxicity checks.

NExT-GPT: A multimodal model with three stages: encoding with ImageBind, understanding and reasoning with a large language model (LLM), and generation using diffusion decoders. It utilizes Multimodal Alignment Learning to align features across modalities and Modality-switching Instruction Tuning (MosIT) to improve LLM capabilities, offering enhanced interaction handling.

VideoPoet: A language model specialized in high-quality video synthesis with matching audio. It processes multimodal inputs using a decoder-only transformer architecture and employs a two-stage training protocol for zero-shot video generation. VideoPoet excels in text-to-video and video stylization tasks, supported by features like custom spatial super-resolution and a Large Language Model backbone. It's noted for its quality in text fidelity, video quality, and motion interestingness, with a focus on responsible AI practices in fairness and zero-shot editing.

## Future Direction

Modularity in Training: Traditional training methods for VLMs are often opaque and resemble "black boxes". The move towards modularity aims to increase understanding, control, and faithfulness by structuring models into understandable and controllable components. This approach helps elucidate how different parts of a model contribute to the overall outcome, making the system more transparent and easier to adjust or improve.

Incorporating Finer Modalities: Research is expanding VLMs to include more nuanced modalities like gaze and gestures, which are critical in contexts such as education where non-verbal cues play a significant role. This development seeks to make interactions more natural and intuitive, especially in automated educational tools.

Fine-grained Evaluation of VLMs: Efforts like DALL-Eval and VP-Eval are focused on evaluating VLMs more meticulously on aspects such as bias and fairness. These evaluations are crucial for ensuring that VLMs operate without perpetuating existing biases or creating unfair outcomes, particularly in sensitive applications.

Causality and Counterfactual Capabilities: Inspired by similar advancements in large language models (LLMs), there's significant interest in equipping VLMs with the ability to understand causality and counterfactuals. This involves enabling models to reason about "what if" scenarios and understand the implications of changes in input, which is vital for applications requiring deep reasoning and hypothetical analysis.

Continual Learning and Unlearning: This trend focuses on enabling VLMs to continuously learn and adapt without needing to be retrained from scratch. This approach not only makes models more adaptable to new data or changing environments but also includes mechanisms for unlearning biases or incorrect information, similar to what is seen in LLMs.

Efficiency in Training: There is a concerted effort to develop more efficient training protocols for multimodal models. Models like BLIP-2 demonstrate significant advancements by achieving higher performance with fewer parameters compared to previous models like Flamingo-80B, addressing the critical need for resource efficiency in training complex models.

Multilingual Grounding: As the utility of multilingual LLMs grows, there's an increasing push to extend these capabilities to VLMs. This would enable models to understand and generate content across different languages, broadening their applicability and enhancing their accessibility globally.

More Domain-Specific VLMs: There is a growing emphasis on developing VLMs tailored for specific sectors such as medicine and agriculture. These domain-specific models, like MedFlamingo and SkinGPT, are designed to address the unique challenges and needs of their respective fields, leveraging visual and textual data to provide specialized insights and solutions.