UNIVERSITY OF
WEST LONDON
The Career University

ASSIGNMENT

Student Name: FATHIMA SAFVA OVINAKATH KAMMUKKAKATH
Student ID: 33151133
Programme of Study: BSc COMPUTER SCIENCE, LEVEL 5

TABLE OF CONTENTS

1. Table Of Contents..... 2
2. Q1: Research Summary Analysis 3
    - 2.1 Paper 1 Summary 3
    - 2.2 Paper 2 Summary 4
    - 2.3 Paper 3 Summary 5
    - 2.4 Paper 4 Summary 6
    - 2.5 Paper 5 Summary 7
3. PPT and Video Presentation Link..... 8
4. Comparative Insights 10
5. Comparative Analysis 11
6. Q2: Autonomous Search-and-Rescue Drone 13
    - 5.1 (a) PEAS Framework 13
    - 5.2 (b) Drone's Working Environment 14
    - 5.3 (c) Agent Design & Critical Comparison 14
7. Reference List 16

Q1) 1. RESEARCH SUMMARY ANALYSIS

Paper 1

- **Full reference (APA): Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021).** Learning Transferable Visual Models From Natural Language Supervision. *ArXiv*. https://arxiv.org/abs/2103.00020
- **Year:** 2021
- **Source:** arXiv / MLR proceedings
- **Technology/Topic:** Zero-shot vision using contrastive image-text pretraining (CLIP)
- **Key contributions:** To achieve robust zero-shot transfer across numerous vision tasks, scalable contrastive image–text pretraining was introduced on hundreds of millions of image-caption pairs.
- **Methodology / Approach:** To match paired picture-text representations, two encoders (text and image) were trained in a contrastive manner. Zero-shot transfer was assessed on about thirty datasets.
- **Findings / Results:** Without task-specific fine-tuning, CLIP achieves solid zero-shot performance across a variety of workloads. A paradigm shift was demonstrated by using language to supervise eyesight.
- **Relevance to My Project:** Fundamental method for multimodal models; describes the emergence of language-vision alignment.
- **Limitations / Gaps Identified:** Lack of grounded thinking on complicated visual scenes, restricted open-ended instruction following, and susceptible to dataset bias (web data).
- **Summary (2–3 sentences):** Large-scale contrastive pretraining on image-text pairs produces transferable visual representations that allow zero-shot behavior, as demonstrated by CLIP. It laid the foundation for later multimodal models that link potent LLMs to visual encoders.

Paper 2

- **Full reference (APA): Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., . . . Simonyan, K. (2022).** Flamingo: A Visual Language Model for Few-Shot Learning. *ArXiv*. https://arxiv.org/abs/2204.14198
- **Year:** 2022
- **Source:** NeurIPS / DeepMind
- **Technology/Topic:** Few-shot visual language models that consume text and image sequences that are interleaved (Flamingo)
- **Key contributions:** Few-shot multimodal in-context learning is made possible by an architectural module that bridges pretrained vision and language models; it can handle text interspersed with images and videos.
- **Methodology / Approach:** In-context few-shot learning is made possible by per-image cross-attention "perceiver-style" blocks that condition the LLM on visual characteristics.
- **Findings / Results:** Excellent few-shot performance on captioning and VQA activities; a single model can do a range of visual tasks with little adjustment.
- **Relevance to My Project:** Demonstrates architectural techniques for in-context multimodal learning that are applicable to interactive systems in the real world.
- **Limitations / Gaps Identified:** High processing and data requirements; restricted interpretability; possible hallucinations when visual grounding is weak.
- **Summary (2–3 sentences):** Flamingo advanced flexible multimodal in-context learning by introducing architectural concepts that enable LLMs to execute few-shot multimodal tasks by conditioning on visual inputs.

Paper 3

- **Full reference (APA): Li, J., Li, D., Savarese, S., & Hoi, S. (2023).** BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *ArXiv*. https://arxiv.org/abs/2301.12597
- **Year:** 2023
- **Source:** ICML / arXiv / MLR proceedings
  Technology/Topic: Effective pretraining for vision and language using frozen image encoders and frozen LLMs connected by a tiny Querying Transformer (Q-former)
- **Key contributions:** Shown that using a small trainable module to connect frozen visual encoders and frozen LLMs produces robust performance at a far lower cost than end-to-end training.
- **Methodology / Approach:** Pretraining in two stages: (1) inquiring Transformer uses frozen visual backbone features to train concise visual queries, then (2) uses image-text data to align those queries with LLM.
- **Findings / Results:** Practical scaling of multimodal systems employing current big LLMs was made possible by competitive performance with significantly less computing power.
- **Relevance to My Project:** Explains the current trend in engineering toward multimodal systems that are modular and computationally efficient.
- **Limitations / Gaps Identified:** Continues to rely on big pretraining datasets; alignment safety issues persist; and the bridge module may still permit erroneous correlations.
- **Summary (2–3 sentences):** By bootstrapping frozen big models, BLIP-2 demonstrated a computationally efficient route to powerful V+L models, speeding up the real-world implementation of multimodal assistants.

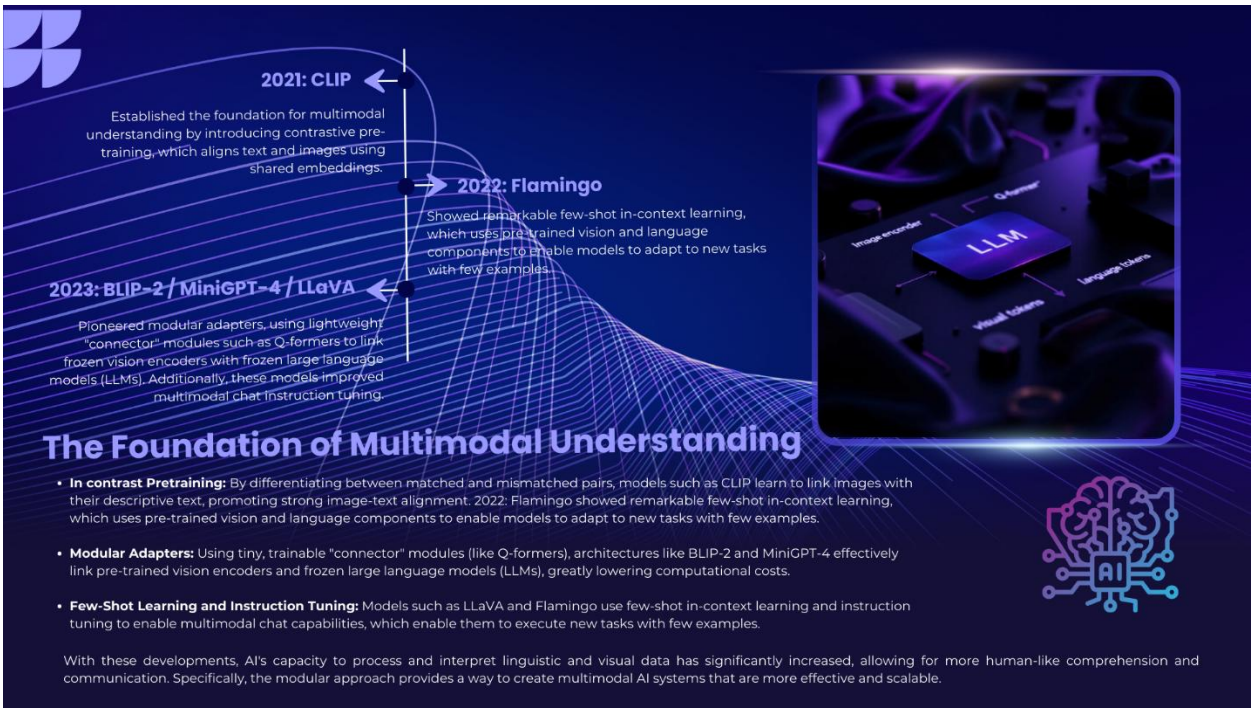
Paper 4

- **Full reference (APA): Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023).** Visual Instruction Tuning. *ArXiv*. https://arxiv.org/abs/2304.08485
- **Year:** 2023
- **Source:** arXiv / open repo (LLaVA)
- **Technology/Topic:** LLaVA multimodal assistant, visual instruction tuning (producing multimodal instruction data via GPT-4)
- **Key contributions:** GPT-4's ability to produce high-quality multimodal instruction data for the purpose of instruction-tuning multimodal models and producing powerful instruction-following multimodal assistants is first demonstrated.
- **Methodology / Approach:** Using GPT-4, create artificial multimodal instruction-response pairings. Then, use that data to refine vision+LLM models.
- **Findings / Results:** LLaVA demonstrates outstanding multimodal conversation capability, narrowing most of the gap to GPT-4 multimodal performance on various benchmarks.
- **Relevance to My Project:** Demonstrates the effectiveness of instruction tuning and synthetic instruction data for safety/alignment and capability enhancements.
- **Limitations / Gaps Identified:** The quality of synthetic data is dependent on the LLM of the teacher; it may not represent uncommon real-world situations and may spread teacher biases.
- **Summary (2–3 sentences):** A significant step toward interactive multimodal assistants, LLaVA popularized visual instruction tweaking utilizing GPT-4 to synthesize data.

Paper 5

- **Full reference (APA): Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023).** MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *ArXiv*. https://arxiv.org/abs/2304.10592
- **Year:** 2023
- **Source:** arXiv / GitHub demos
- **Technology/Topic:** Integrating a visual encoder with an open LLM (Vicuna) through a singular projection layer to facilitate GPT-like multimodal functionalities.
- **Key contributions:** Proved that a lightweight alignment layer can enable advanced open LLMs for multimodal tasks, revealing unexpected emergent capabilities (comprehensive descriptions, reasoning).
- **Methodology / Approach:** Stabilize the visual encoder and LLM, and train a linear projection layer to translate visual features into the LLM token space, followed by instruction-based fine-tuning.
- **Findings / Results:** MiniGPT-4 exhibits numerous multimodal behaviors akin to GPT-4 with limited computational resources, underscoring significant emergent capabilities resulting from alignment.
- **Relevance to My Project:** Demonstrates feasible, economical approaches to multimodal assistants and the potential hazards of emergent behaviors resulting from minimal alignment.
- **Limitations / Gaps Identified:** Safety and alignment control remain challenging; risk of hallucination persists; limited robustness in complex visual reasoning.
- **Summary (2–3 sentences):** MiniGPT-4 demonstrated that compact alignment modules can elicit robust multimodal behavior from static LLMs, facilitating the democratization of multimodal systems while also heightening safety concerns.

PPT AND VIDEO PRESENTATION LINK



https://youtu.be/0_3aupqtQ58?si=qoF7Ao-iglmoQtTQ

Comparative Insights

- **Common themes:** Language supervision + contrastive pretraining (CLIP) -> modular alignment of frozen LLMs and vision encoders (BLIP-2, MiniGPT-4) -> instruction tuning (LLaVA), and few-shot in-context methods (Flamingo).

- **Differences in approaches:** MiniGPT-4 uses a minimal projection/alignment to an LLM; LLaVA emphasizes synthetic instruction tuning; BLIP-2 emphasizes a small querying transformer and frozen backbones; Flamingo uses cross-attention to combine pretrained models for few-shot learning; and CLIP uses contrastive pretraining from scratch.

- **Trends (last 3 years):** Modular architectures that link frozen backbones and tiny adapters will replace monolithic end-to-end training, with a focus on instruction tweaking and synthetic data to achieve instruction-following multimodal behaviour.

- **Gaps / open questions:** Strong long-form visual reasoning, domain adaptation for low-resource environments, ecological grounding (real-world sensors), hallucination prevention, and safety evaluation benchmarks.

- **How literature informs the assignment:** For practical demonstrations, use a modular BLIP-2 style architecture; for safety and alignment discussions, focus on instruction tuning (LLaVA); and contrast few-shot and instruction-tuned behaviors (Flamingo vs. LLaVA).

2) Comparative analysis

Paper #	Strengths	Limitations	Practical Applications	Ethical & Societal Challenges	Technical Challenges / Gaps
Paper 1	Robust multimodal alignment (CLIP). Modular plus economical scaling (BLIP-2). Following instructions (LLaVA).	Unfounded reasoning and hallucinations. Data bias and security concerns. High environmental and computational costs.	Accessibility tools and visual Q&A. HRI and robotics perception. imaging in medicine (with adjustment). E-commerce moderation and search.	Privacy issues with online data. false knowledge through delusions. Model transparency is lacking.	Weak commonsense and causal reasoning. Weak domain resilience. Better human-aligned benchmarks are required.
Paper 2	Robust confluence of multiple modalities. Enhanced few-shot learning (like Flamingo).	Poor temporal reasoning. Instability of the model with noisy images.	Assistive technology, surveillance, and video captioning.	Bias in datasets of videos. Ethical issues with the use of surveillance.	Video logic is currently quite basic. Latency when performing things in real time.

Paper #	Strengths	Limitations	Practical Applications	Ethical & Societal Challenges	Technical Challenges / Gaps
Paper 3	Effective LLM and vision encoder bridging (BLIP-2). Inexpensive training tuning.	Restricted ability to perceive small details. Reliance on LLM quality.	Triage of medical images. Industrial examination.	Misdiagnosis risk. Problems with traceability.	Requires a foundation in domain data. Restricted criteria for evaluation.
Paper 4	Effective instruction-following using visual language alignment (LLaVA).	Visual details that are hallucinated. Fails in scenes that are complicated.	Systems of explanation, teaching, and educational resources.	Inaccurate explanations raise the possibility of false information.	Better fact-checking and alignment are required.
Paper 5	Multimodal structures that are scalable. Improved generalization.	Training's environmental cost. Edge device inefficiency.	AI assistants on-device. AR and smart glasses.	Real-time capture poses privacy problems.	Hardware restrictions. Compression of the model is required.

Q2. AUTONOMOUS SEARCH-AND-RESCUE DRONE

(a) PEAS Framework

Performance:

1. **Person detection recall @ ≤30m:** Aim for ≥90% recall for survivors within 30m during the day. Crucial since it is unacceptable to miss a survivor.
2. **Accuracy of localization (position error):** 95% of the time, the target is within 3 meters of the coordinates of the rescue squad. Crucial for accurately guiding rescues.

Environment:

1. **GPS-denied or degraded surroundings (collapsed structures):** Increases uncertainty in localization decisions by forcing reliance on onboard SLAM/visual odometry & local mapping.
2. **Smoke, dust, limited visibility, and rain:** Cause visual sensors to deteriorate, which lowers detection reliability and necessitates multisensor fusion (thermal, LiDAR) and modified perception pipelines.

Actuators:

1. **Propulsion motors and rotors:** Hovering stability and safe approach speeds close to debris are determined by precision and responsiveness. Micro-adjustments during close inspections require low latency control.
2. **Gimbal + camera pan/tilt actuator:** This affects perception accuracy and the capacity to lock on to targets; it must be precise and low-jitter to generate crisp images for detection.

Sensors:

1. **RGB camera or cameras:** They offer comprehensive scene information, but they suffer from motion blur and malfunction in low light or smoke.
2. **Thermal (IR) camera:** Low spatial resolution and susceptible to being tricked by warm objects, yet useful for person detection in low light.
   Impact: To prevent false positives and negatives, strong data fusion and uncertainty modeling are necessary due to sensor flaws and different modalities.

(b) Drone's working environment

1. **Fully vs. Partially Observable:** *Partially Observable*. The drone is unable to "see" into buildings or through debris. It must make decisions based on incomplete sensor data (such as victims who are obscured).
2. **Deterministic vs. Stochastic:** *Stochastic*. Unpredictable factors include shifting debris, moving survivors, wind gusts, and battery exhaustion.
3. **Episodic vs. Sequential:** *Sequential*. Future possibilities (battery state, map knowledge) are influenced by current actions (e.g., mapping a corridor). Previous observations are important for planning.
4. **Static vs. Dynamic:** *Dynamic*. During the agent's operation, the environment (people moving, aftershocks) changes.
5. **Discrete vs. Continuous:** *Continuous*. Although internal planning can discrete-approximate, flight control, sensor readings, and locations are continuous variables.
6. **Single vs. Multi-Agent:** *Multi-Agent*. Multiple drones and human teams frequently work together; communication limitations may necessitate some autonomy.

(These support the importance of advanced planning, memory, and multi-agent cooperation.)

(c) Agent Design for the Drone - Recommended: Hybrid Model-Based + Learning Agent (Model-Based Learning Agent)

Rationale for selection

- **Model-based** elements (world model, SLAM, and planning) enable the drone to plan safe routes in crowded and GPS-denied environments and anticipate the effects of actions.
- **Learning** components (perception networks, policy refinement) manage perceptual uncertainty (person detection from noisy sensors), adjust to novel situations, and gradually enhance detection.
- Combining both promotes robust perception and safe planning, which are essential for life-saving jobs when errors can be costly.

Handling complex / uncertain conditions

- **Perception:** Employ RGB + thermal fusion deep learning detectors with calibrated uncertainty outputs, such as segmentation masks and detection confidence.
- **Planning under uncertainty:** A model-based planner that takes sensor uncertainty and dynamic impediments into account, such as the belief-space planner or POMDP approximation, enables cautious decisions when confidence is low.
- **Fallback & human-in-loop:** Low-confidence situations lead to conservative actions (hover & broadcast) or human operator review.
- **Edge adaptation:** Managing domain shifts (smoke, dust) is made possible by onboard continuous learning (e.g., lightweight online fine-tuning or adaptive thresholds).

Improvement over time

- To gradually update detectors and calibration models, the agent gathers labeled or pseudo-labeled data from missions (with human verification); experience enhances detection, false-positive rejection, and energy management strategies.

Critical comparison

- **Simple reflex agent:** Too fragile; unable to manage long-term compromises, planning, or partial observability. Not appropriate.
- **Utility-based agent:** Excellent at maximizing trade-offs, but it struggles in GPS-denied and dynamic mapping scenarios without an internal model; utility cannot account for uncertainty on its own.
- **Model-based + learning (chosen):** Safe planning plus adaptive perception is the best combination. For search and rescue, it is more reliable, secure, and effective.

REFERENCE LIST

- **Li, J., Li, D., Savarese, S., & Hoi, S. (2023).** BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *ArXiv*. https://arxiv.org/abs/2301.12597
- **Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021).** Learning Transferable Visual Models From Natural Language Supervision. *ArXiv*. https://arxiv.org/abs/2103.00020
- **Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023).** Visual Instruction Tuning. *ArXiv*. https://arxiv.org/abs/2304.08485
- **Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., . . . Simonyan, K. (2022).** Flamingo: A Visual Language Model for Few-Shot Learning. *ArXiv*. https://arxiv.org/abs/2204.14198
- **Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023).** MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *ArXiv*. https://arxiv.org/abs/2304.10592