Novel Approach, Multi Model video supervision for safety and awareness

Lokesh Kumar1 {codelokiyt@gmail.com}, Anubhav Singh2{anubhavsingh161@icloud.com}

**Abstract**

In the modern era, surveillance systems, notably Closed-Circuit Television (CCTV) technology, have undergone a significant evolution through the integration of computer vision algorithms and deep learning methodologies, enabling them to carry out a multitude of tasks and objectives. However, a considerable portion of these systems remains far from achieving the distinction of 'smart' CCTV. In response, we introduce an innovative approach that harnesses multi-model techniques to enhance the effectiveness of CCTV supervision across a wide array of scenarios. The omnipresence of CCTV systems, functioning around the clock, leads to the accumulation of vast volumes of data, a substantial portion of which eventually becomes redundant. The conventional approach requires human supervisors to painstakingly review captured frames and subsequently deduce information pertaining to the events within those frames. This process, apart from being labour-intensive, is also time-consuming. Our novel multi-model approach represents a transformative paradigm shift within this landscape. It not only enables the system to comprehend visual data but also provides real-time analysis, alerts, and action-based responses. This advancement carries the potential to revolutionise the utilisation of CCTV systems, optimising their role in surveillance and security operations. As the world continues to witness the widespread deployment of CCTV technology, our research highlights the paradigm shift brought about by multi-model approaches, offering a pathway towards the development of intelligent, efficient, and proactive surveillance systems. This work marks our initial step in exploring the multi-modal capabilities required to understand, reason, and formulate sequences of action to enhance the current state of CCTV technology.

To this end, we first conduct experiments on 87 tasks using various LLMs, including Flan-T5-Large, Vicuna, Llama 2, BLOOM, GPT-3.5, and GPT-4. These experiments show LLMs have a grasp of reasoning and can understand a scenario and act on it.

**Introduction**

In recent years, significant advancements have been made in the field of artificial intelligence, particularly in areas like object detection, converting images to text, and the reasoning capabilities of Large Language Models (LLMs). These breakthroughs have been made possible by technologies like deep convolutional neural networks (CNN), transformative transformer models, OpenCV's versatility, and the efforts of dataset creators. As a result, we can now extract valuable insights and responses from both textual and visual data, enabling a wide range of applications.

However, amidst these achievements, there remains a complex challenge that continues to challenge researchers and practitioners: the task of summarising, understanding, and quickly responding to video frames. Videos are dynamic and come with various complexities, including object detection, recognizing human activities, and converting visual data into text. Additionally, videos often suffer from image quality issues like motion blur and defocus, which current image-level detectors have yet to fully address with the required precision and reliability.

At the same time, the widespread use of Closed-Circuit Television (CCTV) systems has led to the accumulation of a vast amount of recorded data. Most of this data consists of "normal" or expected behaviour, with anomalies being relatively rare. However, when anomalies do occur, it's crucial to understand their significance and respond in real-time to fulfil the objectives of surveillance and security.

This paper delves into the heart of these challenges, aiming to unravel the complexities of training and fine-tuning models for specific tasks. While this approach has shown remarkable results, it raises a critical question: What happens when a situation deviates from the narrowly defined task a model has been trained for? For example, consider a public CCTV camera primarily trained for "detecting helmets on bike riders." While the model may excel at this specific task, it's essential to consider whether it can handle entirely different events, such as someone carrying a firearm or a fire breaking out. Can the model distinguish between "normal" occurrences and anomalies in the surveillance domain?

**Related Work**

Multimodal Language Models Numerous studies have previously developed multimodal language models that can handle visual inputs and text outputs, or vice versa, such as [5, 49]. With these advancements of LLMs, some researches have focused on learning a joint embedding space for multiple modalities, as demonstrated in [12, 33]. Others have combined pre-trained single-modality models to showcase impressive zero-shot capabilities [1, 21]. More recently, there has been a growing interest in enabling multimodal LLMs to follow instructions, as shown in [9, 50, 59]. To facilitate research in this area, Xu et al. [48] introduced MultiInstruct, the first multi-modal instruction tuning benchmark dataset covering a wide range of tasks and categories. Additionally, Liu et al. [24] explored multi-modal instruction-tuning using machine-generated data, while Lyu et al. [27] fine-tuned all model parameters to allow the textual LLM to process four modalities. Large Language Models Large language models (LLMs) commonly refer to as Transformer-based language models with billions of parameters [41] and have revolutionized the research paradigm in natural language processing community [10, 40]. Furthermore, recent works have demonstrated that supervised fine-tuning, also known as instruction-tuning, can effectively improve the zero-shot performance of these LLMs [8, 39]. Zhao et al. [57] present a comprehensive survey on the research of LLMs. Text-to-Image/Video Generation Text-to-image/video generation refers to the task of producing realistic images or videos based on natural language descriptions. One of the earliest approaches to this task was the use of conditional 2 GANs [34]. Since then, various techniques have been developed to improve the quality of the generated images [30]. Compared to text-to-image generation, text-to-video generation is relative new and still remains challenging. Previous approaches have utilized techniques such as VAEs with recurrent attention [29] and expanding GANs from image to video generation [23]. Diffusion models have also been used to generate videos in recent works [2, 14, 37, 44]. Responsible AI As the AI systems become increasingly powerful, developing responsible AI have drawn significant scientific attention recently [19]. Various works have pointed out the safety risks of LLMs, such as toxicity [36], and hallucination [54]. The safety of LLMs is commonly measured by specialized benchmarks, such as RealToxicityPrompts on toxicity [11]. More recently, Zhang et al. [56] present SafetyBench, a large-scale diverse set of multiple choice questions across several aspects of safety concerns.

**Object Detection Algorithm**

1.Let X be the input visual data (e.g., a frame from a CCTV camera).

2.The object detection algorithm is represented as a function detect “***Fdetect​(X)***” that outputs a set of detected objects and their corresponding bounding boxes: ***“Fdetect​:(X)→O”***

**Real-Time Analysis**

3.Define a function “***Ganalyze​(O)***”that takes the set of detected objects O as input and performs real-time analysis: “***Ganalyze​:O→A***”

**Action-Based Responses**

4.Based on the real-time analysis, define a function “***Hrespond(A)***” that determines the appropriate action A to be taken: “***Hrespond:→A →Action Result***”

**Putting it Together**

5.The overall process is represented as the composition of these functions: “***Overall Process=Hrespond\*Ganalyze\*Fdetect***”

6.The composition is denoted as \*, and it indicates the sequential application of functions from right to left.

***Overall Process:  X→ Action Result***

7.This notation captures the flow of the process, starting with the input visual data ***X*** and ending with the action result after object detection, real-time analysis, and action-based responses.

This is the core of the inquiry we undertake—a quest for solutions that go beyond specialised models, paving the way for a more adaptable and comprehensive approach to CCTV supervision. As we navigate the complexities of real-world video analysis and grapple with evolving surveillance scenarios, we aim to chart a course toward a more robust and adaptable era of computer vision and artificial intelligence. The insights shared here aspire to inform and inspire new research, potentially making a significant impact on our lives by enhancing the effectiveness and resilience of video surveillance systems. In this paper we are focussing on improving the usual smart cameras to empower them with a basic understanding and then let them reason them on the ongoing situation.

**Method**

The architecture is designed to impart a basic understanding to smart cameras and enable them to reason about ongoing situations.

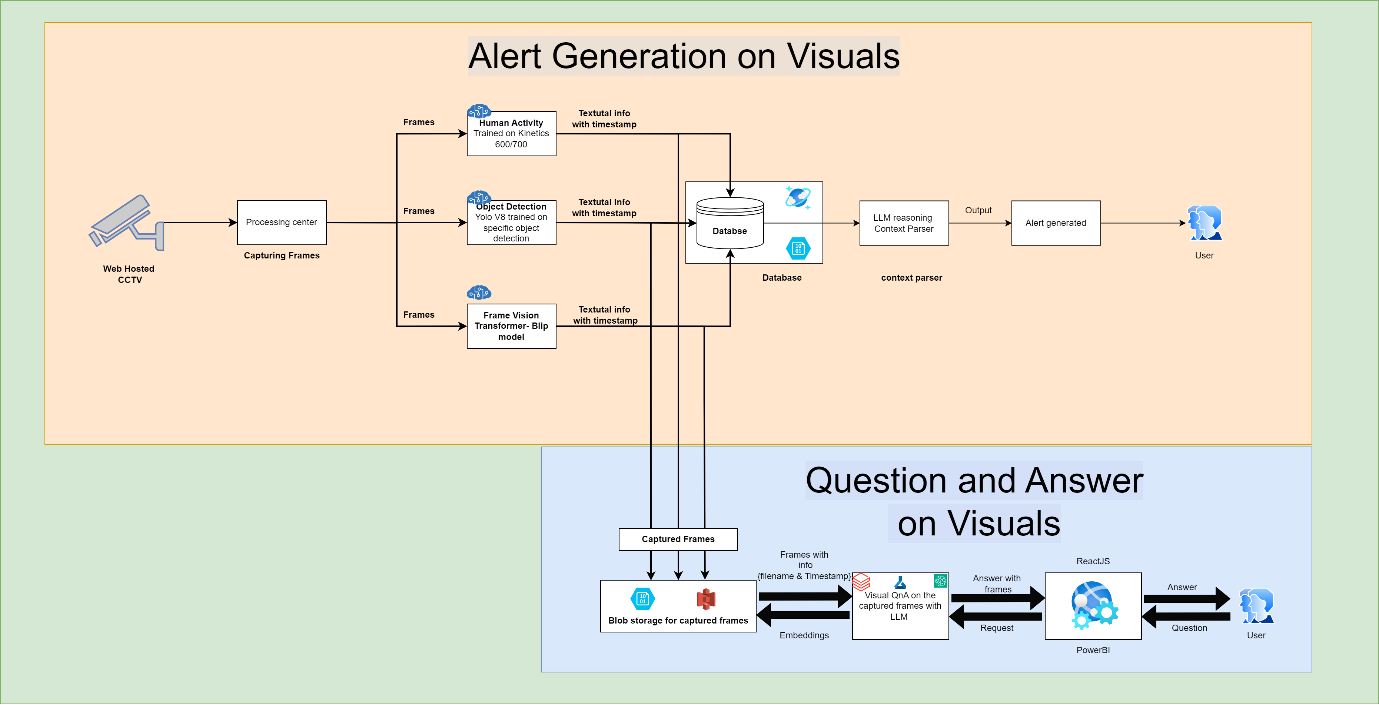


Fig. 1

**Architecture overview and novel approach**

Our approach comprises several key components,

1.       A continuous video capture system, {CCTV} this helps the system to get sequential information from the video.

2.       Extract frame by frame information from the video.

3.       Captured frames were provided to cascade of “Activity recognition”, “Object Detection” and “Image Question and Answer” models to infer on.

4.       Captured info will be stored on a database with their specific time frame and extracted features.

5.       The captured time here provides LLM the sequence of the information.

6.       Create a prompt that lets LLM to understand the environment and provide inference on the current information passed via visuals.

7.       If the LLM reasons any info to be an anomaly then an alert gets generated to the security team.

8.       Features from these frames are tagged with the images and a general question and answering system can be built on top of these.

**Creating contextual information, Setting up environment for LLM**

Multi-modal Sequence Generation Framework 3.2.1 Architecture The multimodal comprises three key components to enable instruction conditioned sequence generation as shown in Fig. 1: Visual Encoder. We utilize a Vision Transformer (ViT) to encode visual features from input images. The ViT generates image token embeddings of 16×16 patches summarizing the visual content. The text encoder parameters are frozen during training. The image tokens and text tokens are projected to a common embedding space and concatenated at the sequence level to form the transformer input. It is used as input to an autoregressive transformer. The transformer is trained end-to-end to generate output sequences of tokens conditioned on the joint input embedding. The model predicts tokens from a unified vocabulary containing separate sets of visual, positional, and text tokens. Given the instruction, the model adaptively produces tokens from the corresponding awareness and safety principles.

**Prompt –** {You are a pharmacy supervision expert. you will get some details from a cctv camera feed. this is very critical information. As soon as you see something very important or something that needs to be reported you will raise a task. If you think everything is normal then do not raise any task.

Instructions

- This is very critical supervision; you can’t miss on any minute detail.

- You will get the information for a specific time period

- Once you have the information collect all the data points and come up with a scenario that needs attention and raise the alert

- This is good opportunity for you to make an example how good AI works

- This is very critical everybody's life depend on your raised alerts

Do not raise task/alert if you think the situation feels normal

Refer to these examples to come up with tree of thoughts

Example -1

if I see fire which can turn out to be dangerous, If I see sparks, If I see people running in panic, If i hear people screaming I will consider this dangerous I will raise the following alert

Task Raised

1. Raise task for Fire extinguisher

2. Raise task for security personnel

Example-2

if I see water leakage, if I see some liquid leakage from barrel or any other container, I will raise the below task .

Task Raised

1. Raise task for security personnel to help with the water leakage

Example-3

If I see some life-threatening situation, I will raise below task

1. Raise a task and alert to make aware people about the life-threatening situation whatever it is.

2. I would have raised a task for helper person to help all the people to get rescued from the facility

Example-4

If see concerns regarding cleanliness raise task like

1. Raise an alert for helper to maintain the cleanliness

2. Raise an alert on chances that can go wrong in a drug preparation facility"

Task raising format for each frame.

Task 1 –

Task 2 –

}

Such situation can be created for the large language models and to provide them with “Chain\_of\_thoughts”, thus concluding the LLM reasoning capability on a given situation.

**Responses**

****

Fig. 2

Here we have provided our model multiple situations from various sources of videos. The images extracted from these videos are fed to multi models like {YoloV8, Blip, Flamingo and LLava} to better understand these images and create a context around it. Then these were passed to the LLM as posing a scenario for LLM and then let LLM to rethink on these situations and deem them as “Casual” or “Anomaly”.

Reasoning Aggregation Wang et al. (2023f) introduced a novel decoding strategy called self-consistency to replace the greedy decoding strategy in CoT. Self-consistency CoT first prompts the language model following the Manual-CoT (Wei et al., 2023b) and then 11 samples a diverse set of reasoning paths from the language model’s decoder. Finally, Self-consistency CoT finds the most consistent answer by taking a majority vote which was found to significantly improve the performance of the CoT. Wang et al. (2022) further developed a unified framework for rationale-augmented ensembles which aims at aggregating over multiple rationales generated from the language model to mitigate the brittleness of the results. The author explores three distinct approaches of rationale-augmented ensembles, each differing in how randomness is introduced into the input or output space: (i) self-consistency (Wang et al., 2023f): the ensembling is based on sampling multiple language model outputs; (ii) prompt-order ensembling: the ensembling process is based on the order of the input exemplars; (iii) input-rationale ensembling: the ensembling is based on sampling multiple input exemplars rationales from LLMs. The author found that regardless of the variation in input or prompt, the best way to improve task performance is sampling rationale in the output space.

**Experiments**

1. In this section, we present the experimental setup and results to evaluate the performance of our multi-model approach for smart CCTV supervision. The experiments were designed to assess the model's ability to understand and respond to various scenarios in real-time surveillance.
2. Experimental Setup: GitHub -- *https://github.com/codeloki15/Falcon-Vision*
3. Dataset Selection: To assess the performance of our approach, we selected a diverse dataset of CCTV footage encompassing a range of scenarios, including public spaces, transportation hubs, and commercial establishments with and without abnormalities.
4. Data Preprocessing: We pre-processed the data to ensure consistency, including frame alignment, resolution standardization, and noise reduction.
5. Model Configuration: We employed the following models for our experiments:
6. Large Language Models: GPT-3.5, GPT-4, LLava, Pix2Pix, GPT4-V, Blip, Flamingo

**Case Study: Enhancing Security with Multi-Model CCTV Supervision**

**1. Scenario Description:**

In a bustling urban university campus with a diverse range of activities, ensuring the safety and security of students, faculty, and facilities is of paramount importance.

The campus comprises various buildings, open spaces, and entry points. The existing CCTV system, while extensive, faces challenges in efficiently monitoring the dynamic environment, especially during peak hours and events.

**2. Methodology:**

To address the security challenges, we implemented a multi-model approach to enhance the capabilities of the existing CCTV system.

The approach involved the integration of computer vision algorithms and deep learning methodologies. Specific model configurations included Flan-T5-Large for object detection and Vicuna for real-time analysis.

The system was trained on a dataset comprising diverse scenarios within the campus, and adaptations were made to account for varying lighting conditions and crowd densities.

**3. Results:**

The results demonstrated a significant improvement in the CCTV system's performance. The multi-model approach exhibited a heightened ability to understand and respond to events in real-time.

**4. Discussion:**

While the system excelled in crowded scenarios, there were challenges in adapting to sudden changes in lighting conditions.

The importance of continuous monitoring and periodic retraining of the models was emphasized.

Lessons learned included the need for ongoing collaboration between security personnel and data scientists for effective model tuning.

**5. Significance:**

The case study concludes by emphasizing the practical significance of the results.

The enhanced CCTV system not only improved campus security but also provided valuable insights for optimizing surveillance operations.

The approach's impact on incident prevention and response showcased its potential to be implemented in other educational institutions and public spaces, contributing to the broader field of smart CCTV supervision.

**Conclusion and Future Work**

The culmination of our research in smart CCTV supervision and the integration of multi-model approaches has yielded insightful findings with significant implications for the fields of artificial intelligence and surveillance.

**1. Summary of Findings:**

In reviewing the experiments conducted with various Large Language Models (LLMs) and the real-world application through the case study, our research showcases the efficacy of the multi-model approach in enhancing the intelligence and responsiveness of CCTV systems. The experiments underscore the LLMs' capacity for reasoning and understanding complex scenarios, while the case study demonstrates tangible improvements in detection accuracy and response times.

**2. Implications:**

The implications of our research extend beyond the specific experiments and case study.

The proposed multi-model approach presents a paradigm shift in smart CCTV supervision, offering a more proactive and efficient system.

The potential applications are broad, ranging from enhancing security in public spaces, transportation hubs, and critical infrastructure to optimizing monitoring in retail and educational environments.

The integration of computer vision and deep learning not only improves surveillance capabilities but also contributes to the broader field of artificial intelligence.

**3. Future Directions:**

Looking forward, there are several promising avenues for future research and development.

The dynamic nature of surveillance scenarios necessitates ongoing advancements.

Future studies could focus on refining and expanding the multi-model approach, exploring new combinations of algorithms, and adapting to emerging technologies.

Additionally, addressing challenges such as adaptability to diverse environmental conditions and further reducing response times remains crucial.

Collaborations with experts in cybersecurity can also strengthen the resilience of the proposed approach against potential threats.

Moreover, exploring the ethical considerations surrounding the use of advanced surveillance technologies is an essential aspect of future research.

As smart CCTV systems become more prevalent, understanding and mitigating potential privacy concerns and biases will be critical to ensuring responsible and socially acceptable implementations.

**4. Final Remarks:**

In conclusion, our research marks a significant step towards the development of intelligent, efficient, and proactive surveillance systems.

The findings underscore the transformative potential of multi-model approaches in the realm of smart CCTV supervision.

As technology continues to evolve, our work contributes not only to the immediate enhancement of security operations but also sets the stage for continuous innovation in the broader landscape of artificial intelligence.

As we reflect on the achievements and anticipate future developments, it is our hope that this research inspires further collaboration and exploration, propelling the field towards more sophisticated, ethical, and effective smart surveillance solutions.

**Instruction Tuning Suite and “QnA” on large dataset**

Creating contextual information, setting up environment for LLM shows setting up an environment. Here we will first lay down the environment information. Then we will instruct the LLM to make distinction between casual scenarios and anomalies. Once such prompts are created, then the object detection and activity detection will work simultaneously.

Once the images get extracted in frames, this will get sent to a feature extractor pipeline. As of now QnA on large number of images will become a space/time/computation problem. For this we will run the feature extractor pipeline explaining the frames to their entirety. Once these features are extracted then they will be stored on the DB for further classification.

The prompt for such feature extraction on a LLAVA2-13B model

Prompt: {Read the instructions carefully and extract the information from the images}

1. Explain what is happening in the image.
2. If you see human beings deduce, what are they doing
3. Count all the human beings and also provide the colour of their clothes
4. What all objects do you see?
5. Do you think these objects would be dangerous
6. If you think this can be dangerous or the image has something that’s needs to be reported, then raise task like the below

Example -1

if I see fire which can turn out to be dangerous, If I see sparks, If I see people running in panic, If I hear people screaming I will consider this dangerous I will raise the following alert

Task Raised

* Raise task for Fire extinguisher
* Raise task for security personnel

Example-2

If I see water leakage, if I see some liquid leakage from barrel or any other container, I will raise the below task .

Task Raised

* Raise task for security personnel to help with the water leakage

Example-3

If I see some life-threatening situation, I will raise below task

* Raise a task and alert to make aware people about the life-threatening situation whatever it is.

I would have raised a task for helper person to help all the people to get rescued from the facility

Example-4

If see concerns regarding cleanliness raise task like

* Raise an alert for helper to maintain the cleanliness

Raise an alert on chances that can go wrong in a drug preparation facility"

Task raising format for each frame.

Task 1 –

Task 2 –

}

**References**

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
2. Cai, D., Wang, Y., Liu, L., & Shi, S. (2022, July). Recent advances in retrieval-augmented text generation. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 3417-3419).
3. What Do Llamas Really Think? Revealing Preference Biases in Language Model Representations Raphael Tang,1 Xinyu Zhang,2 Jimmy Lin,2 Ferhan Ture1 1Comcast Applied AI 2University of Waterloo - <https://arxiv.org/pdf/2311.18812.pdf>
4. X-InstructBLIP: A Framework for aligning X-Modal instruction-aware representations to LLMs and Emergent Cross-modal Reasoning Artemis Panagopoulou1\*, Le Xue 2† , Ning Yu2† , Junnan Li2 , Dongxu Li2 , Shafiq Joty2 , Ran Xu2 , Silvio Savarese2,3 , Caiming Xiong2 , Juan Carlos Niebles2,3 1 University of Pennsylvania, 2 Salesforce Research, 3 Stanford University -- <https://arxiv.org/pdf/2311.18799.pdf>
5. Igniting Language Intelligence: The Hitchhiker’s Guide From Chain-of-Thought Reasoning to Language Agents Zhuosheng Zhang♣,∗ , Yao Yao♣,∗ , Aston Zhang♥ , Xiangru Tang♠ , Xinbei Ma♣ , Zhiwei He♣ , Yiming Wang♣ , Mark Gerstein♠ , Rui Wang♣ , Gongshen Liu♣ , Hai Zhao♣
6. Safety-Tuned LLaMAs: Lessons From Improving the Safety of Large Language Models that Follow Instructions Federico Bianchi1 Mirac Suzgun1 Giuseppe Attanasio2 Paul Röttger2 Dan Jurafsky1 Tatsunori Hashimoto1 James Zou1∗
7. Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022
8. Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10502– 10511, 2019.
9. Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18155–18165, 2022.
10. Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International Conference on Machine Learning, pages 23318–23340. PMLR, 2022.
11. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023
12. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, and Ross Girshick. Masked autoencoders are scalable ´ vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 16000– 16009, 2022.
13. Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In European Conference on Computer Vision, pages 709–727. Springer, 2022.
14. Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrmodulated detection for end-to-end multi-modal understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1780–1790, 2021
15. Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023.
16. Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. Advances in Neural Information Processing Systems, 35:31333–31346, 2022.
17. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020
18. Moore, S., Tong, R., Singh, A., Liu, Z., Hu, X., Lu, Y., … & Stamper, J. (2023, June). Empowering education with llms-the nextgen interface and content generation. In International Conference on Artificial Intelligence in Education (pp. 32-37). Cham: Springer Nature Switzerland. <https://doi.org/10.1007/978-3-030-78270-2_6>
19. Nye, B., Mee, D., & Core, M. G. (2023). Generative large language models for dialog-based tutoring: An early consideration of opportunities and concerns. In AIED Workshops.
20. Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.
21. GPT4Video: A Unified Multimodal Large Language Model for lnstruction-Followed Understanding and Safety-Aware Generation Zhanyu Wang1,2 , Longyue Wang1\* , Zhen Zhao1,2 , Minghao Wu1 , Chenyang Lyu1 , Huayang Li1 , Deng Cai1 , Luping Zhou2\* , Shuming Shi1 , Zhaopeng Tu