Atmospheric Composition Analysis: Evaluating and Utilising Large Language Model's Ability in Recognising Physical Phenomena

香港科技大學 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

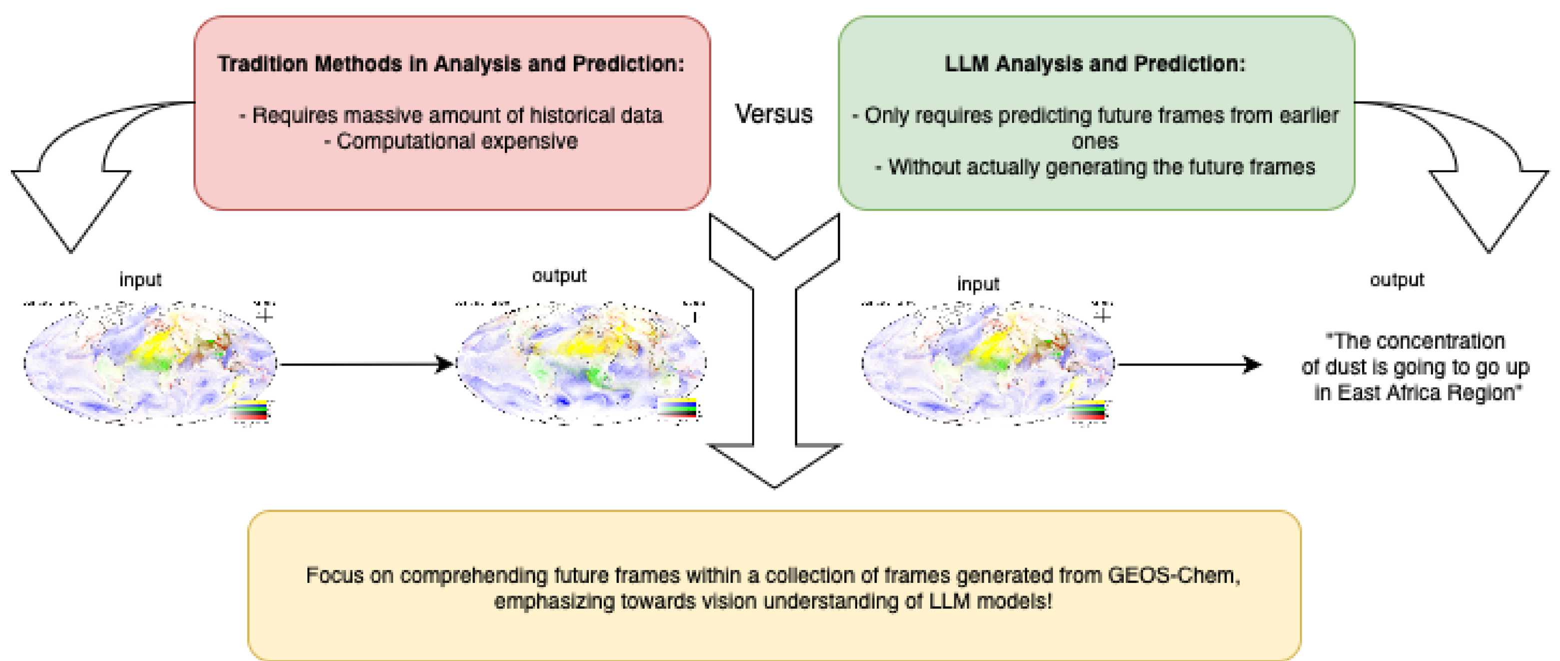
Washington University in St. Louis

Ching Ting LEUNG [1] Nick CRISPINO [2] Chenguang WANG [2]

[1]Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology

[2]Department of Computer Science and Engineering, Washington University in St. Louis

Introduction



Current Methods in Prediction

Mathematical Modelling

Puff model dispersion:

$$< C > (x, y, 0, t) = \frac{Q_m^*}{\sqrt{2}\pi^{3/2}\sigma_x\sigma_y\sigma_z} exp[-\frac{1}{2}[(\frac{x-ut}{\sigma_x})^2 + \frac{y^2}{\sigma_y^2}]]$$

Plume model dispersion:

$$< C > (x, y, 0) = \frac{Q_m}{\pi\sigma_y\sigma_z u} exp[-\frac{1}{2}[\frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2}]]$$

Stable Video Diffusion [1] in Machine Learning

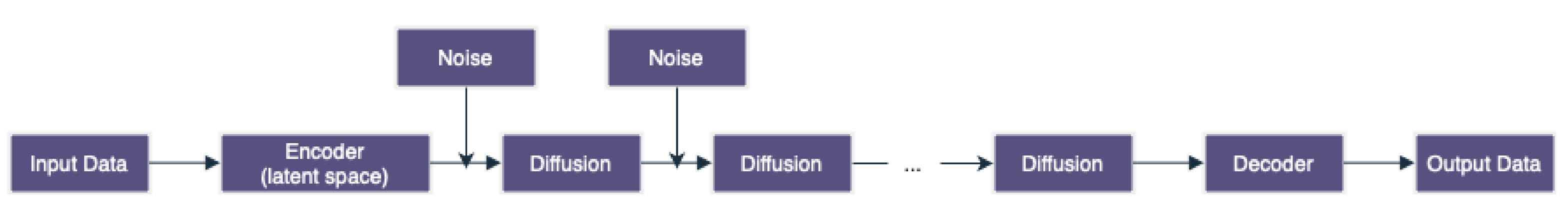


Figure 1. Architecture of a stable video diffusion model.

Bench-marking Video Generating AI Performances

Without understanding the LLMs' ability in recognising physical phenomena, similarity metrics are use to benchmark AI generated videos.

1. Detection Score [2]: $\frac{1}{M_1}\sum_{i=1}^{M_1}(\frac{1}{K}\sum_{i=1}^{K}\sigma_{t_k}^i)$
2. Colour Score [2]: $\frac{1}{M_3}\sum_{i=1}^{M_3}(\frac{1}{K}\sum_{k=1}^{K}s_{t_k}^i)$
3. Trend Score: $\frac{1}{M_8}\sum_{i=1}^{M_8}(\frac{1}{K}\sum_{k=1}^{K}\delta_{t_k}^i)$
4. Text Score: $\frac{1}{2}\sum_{i=1}^{2}(\frac{1}{K}\sum_{k=1}^{K}\varepsilon_{t_k}^i)$

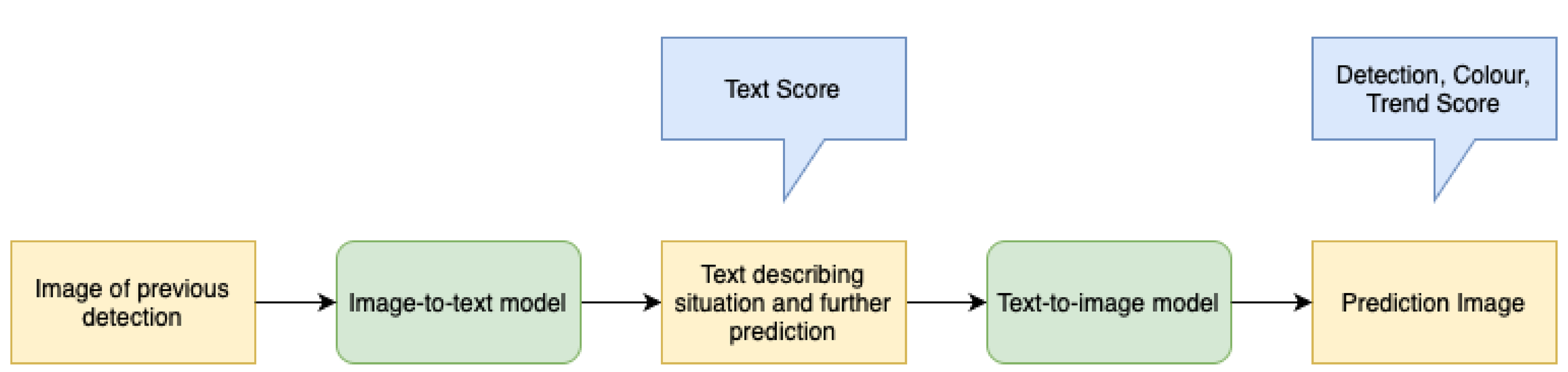
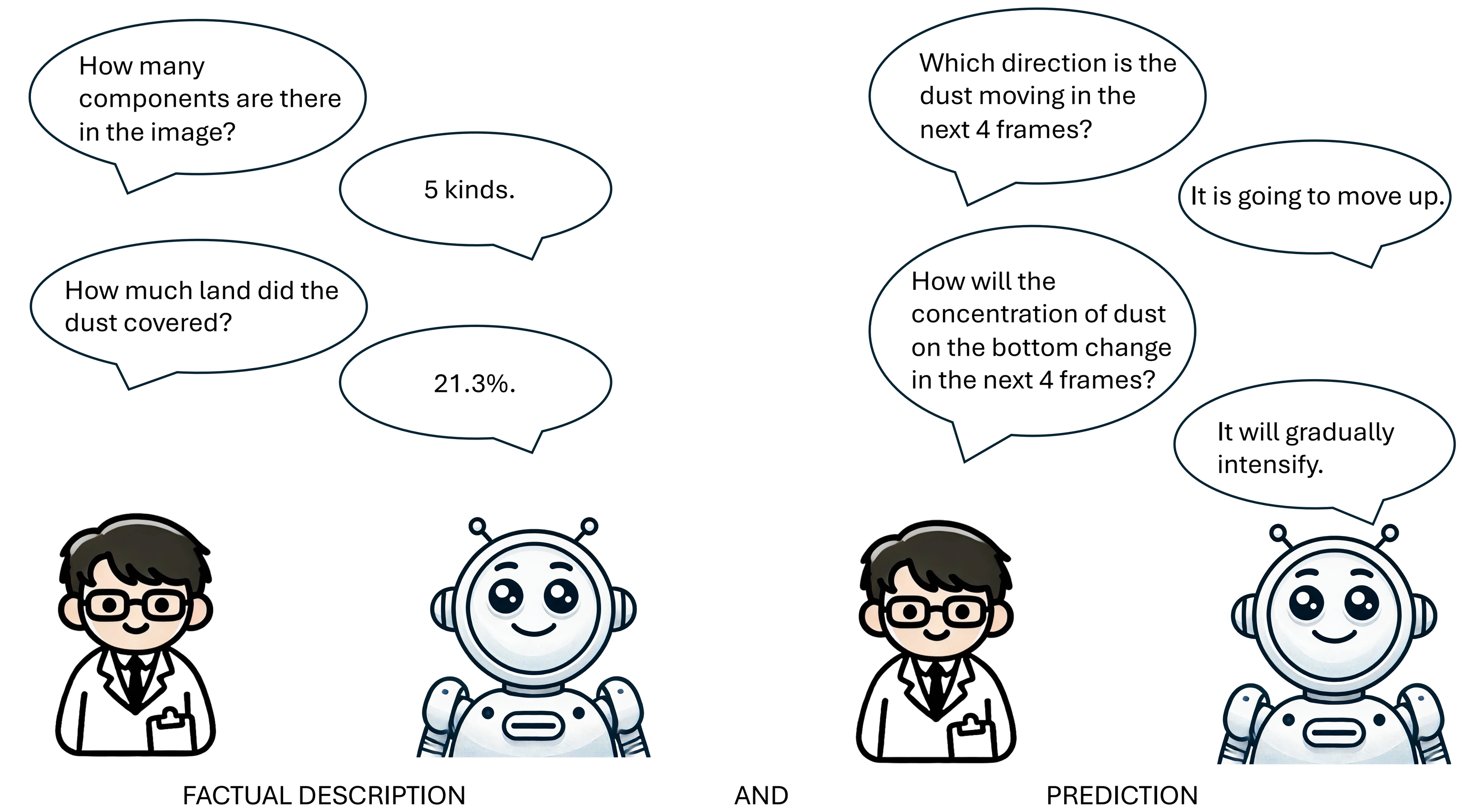


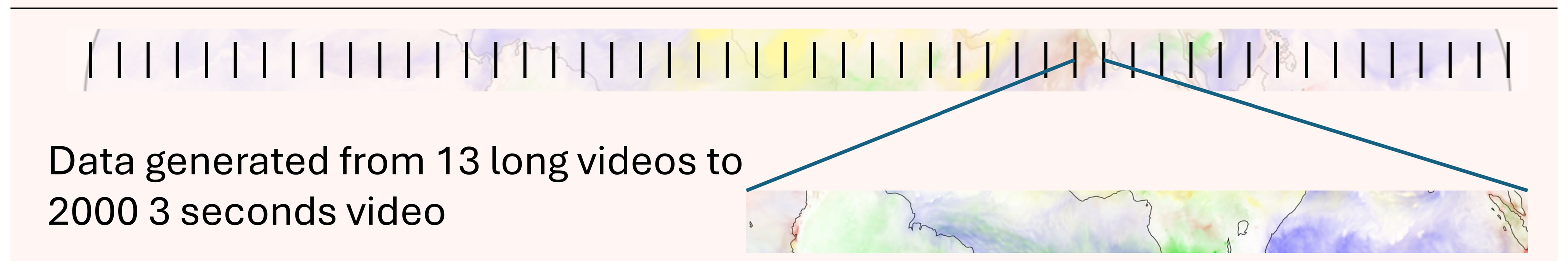
Figure 2. Overall pipeline for image processing and prediction.

Large Language Models on Physical Phenomena



Both scenarios necessitate the vision capabilities of large language models; however, prediction demands a more profound comprehension of the model's interpretation of the map. This includes an understanding of the typical movement of particles, variations in concentration, and the interactions between different atmospheric components.

Large Language Models' Prediction Evaluation



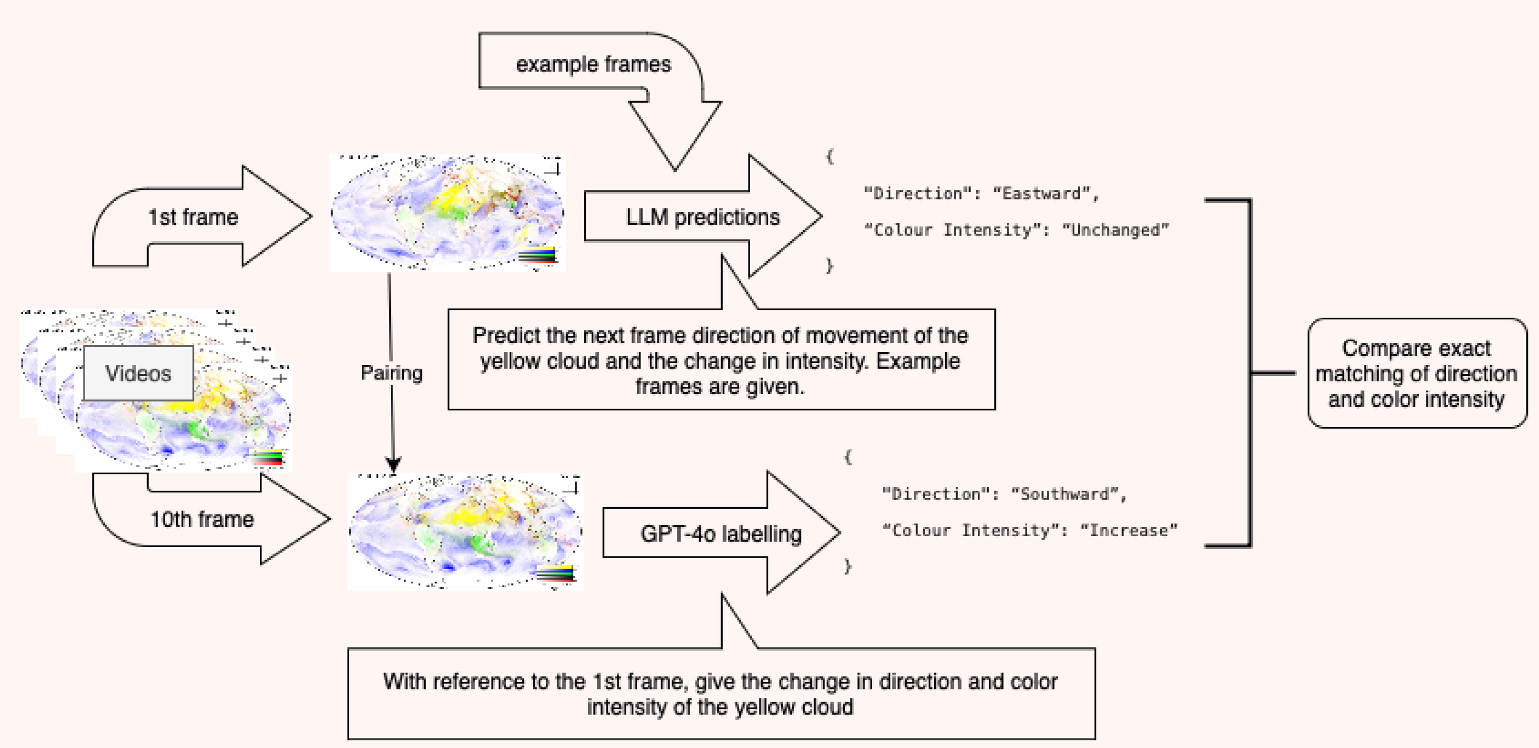


Figure 3. Experimental Procedures of Evaluating the Predictive Ability of Large Language Models

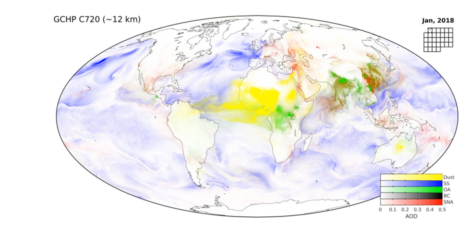
Current popular LLMs are evaluated, such as ChatGPT-4o, Claude-instant-100k and Gemini-1.5-Flash. To investigate the sensitivity of language models on the complexity of images, or the results are randomly generated, we propose the **complexity factor**.

Complexity factor $= \frac{1}{M}\sum_{i=1}^{M}\alpha_i\beta_i$

α = coverage of target cloud on the map, β = number of clusters. Distance between center of clusters should be at least one third of image width, and the cluster radius is at least one third of image width.

Current Model Performances

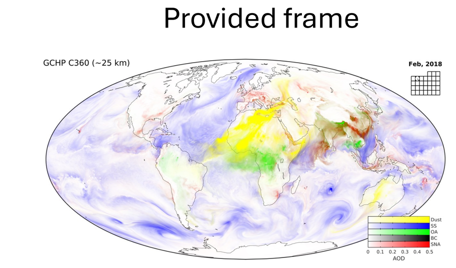
Prompt: Identify the yellow cloud on the map and give its center coordinates



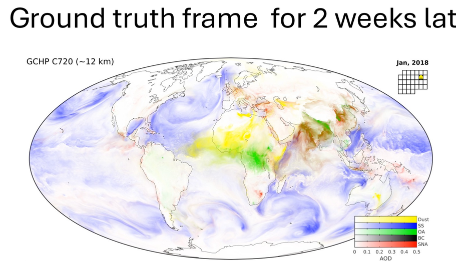
	ChatGPT 4o	Claude-instant-100k	Gemini-1.5-Flash
Coordinates	Latitude: approximately 20°N Longitude: approximately 15°E	Region around the equator at approximately 0.15 AOD	Its center coordinates are approximately 25°N, 35°E
Map recognition	Region of Northern Africa, likely around the central Sahara Desert.	It is predominantly located over central Africa.	Over the Sahara Desert
Graph understanding	The yellow cloud on the map represents dust aerosols	The yellow cloud on the map represent atmospheric dust	The yellow cloud on the map represents dust

Provided frame

Ground truth frame for 2 weeks later



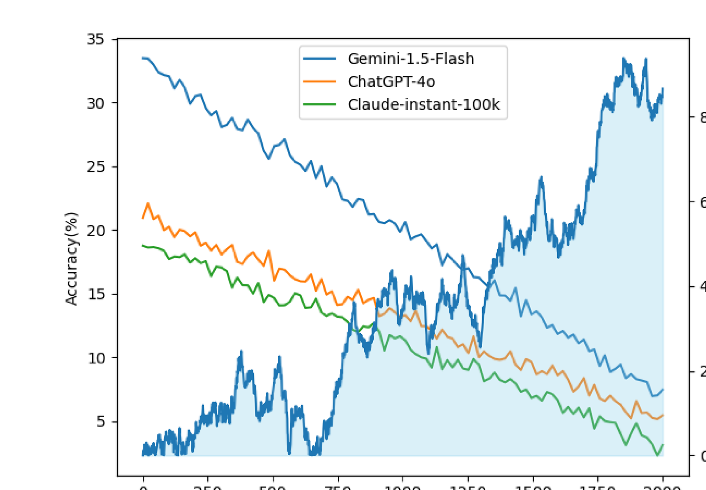
Prompt: describe what you think it will happen for the yellow cloud on the image of its position and color intensity after 2 weeks



	Ground Truth	ChatGPT 4o	Claude-instant-100k	Gemini-1.5-Flash
Position movement	Southward Movements	Eastward movements as dust clouds from the Sahara Desert are often carried by trade winds towards the Atlantic Ocean.	The prevailing trade winds would carry the dust gradually across the Atlantic ocean.	The dust cloud will likely move further west, following the trade winds across the Atlantic Ocean.
Color intensity	Decrease in color intensity	Depends on the activity	Without more specific meteorological data from the intervening time period, dust interaction can depend on sunlight, wind currents and precipitation	The color intensity of the yellow cloud is likely to decrease.

General conclusions: (Only the performance on dust is evaluated in this metric)

- LLMs do not have good predictive ability on target tasks.
- LLMs are not sensitive to the complexity of images
- Their predictive ability cannot be scaled at this point.



Perspective work

Our next goal is to apply these models to atmospheric composition analysis and prediction, aiming to improve accuracy and efficiency in interpreting atmospheric data and enhancing predictive models for better decision-making in atmospheric sciences.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

[2] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. 2023.

ctleungaf@connect.ust.hk