جامعة نيويورك ابوظبي
NYU ABU DHABI

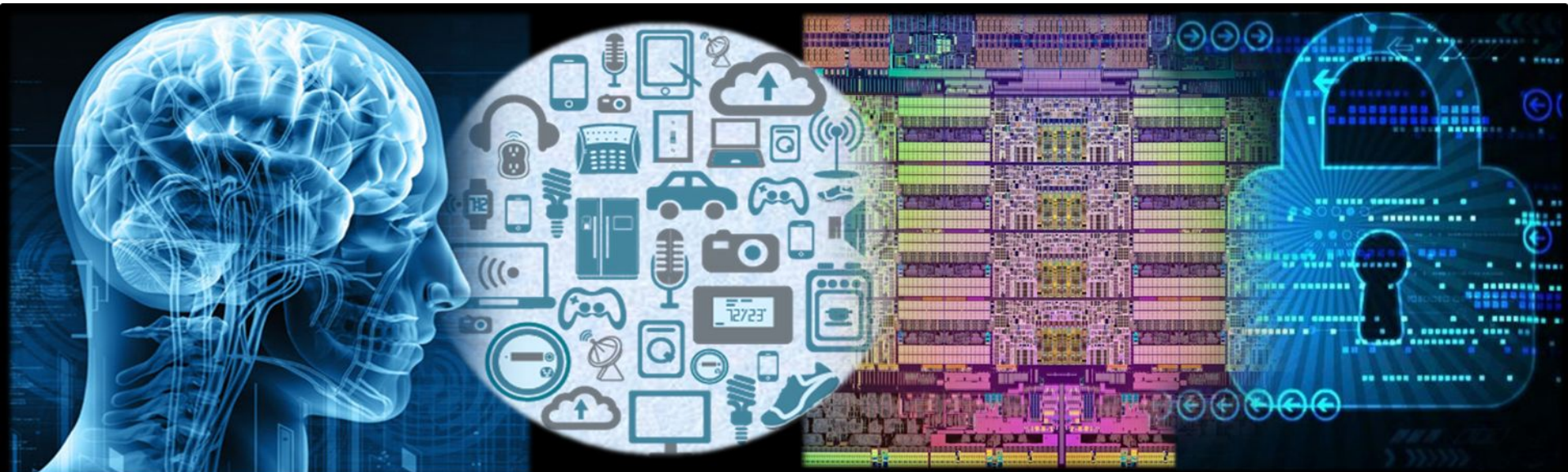
NYU TANDON SCHOOL OF ENGINEERING

BRAIN

Introduction to AIGC, Fake AIGC and their Detection

Minghao Shao

eBrain Lab, New York University (NYU) Abu Dhabi, UAE



Objective of today's workshop

We will cover

- Know the basics of AIGC
- Know the potential risk of using AIGC
- Some approaches of detecting AIGC contents
- How can we detect Fake AIGC contents

We will not cover

- Technology details
- Practice-style coding job

Content

Section 1: AIGC content

- What is AIGC
- Application of AIGC
- Approach of generating AIGC

Section 2: Potential Risk of AIGC

- How can AIGC be abused
- AI-generated Disinformation
- AI-generated Misinformation

Section 3: AIGC and Fake AIGC Detection

- Dataset
- Modalities
  - Text
  - Visual
  - Audio

Section 4: Challenges & Future

Content

Section 1: AIGC content

- What is AIGC
- Application of AIGC
- Approach of generating AIGC

What is AIGC

AI-Generated Content (AIGC) refers to any digital content created using artificial intelligence algorithms. This includes text, images, audio, video, and other media types generated by machine learning models.

Application of AIGC

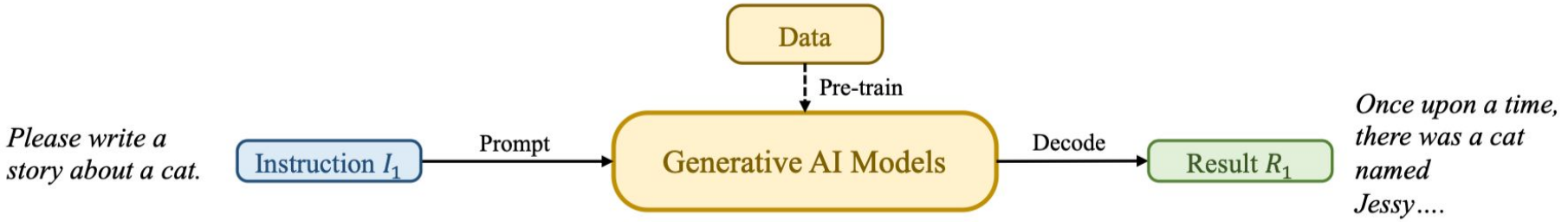
“Three seconds after midnight. Coca-Cola factory, Montgomery. A building in Montgomery to his father's study of this town in the same room where the band was being sent off to the police car. The time was one minute past midnight. But he was the only one who had to sit on his way back. The time was one minute after midnight and the wind was still standing on the counter and the little patch of straw was still still and the street was open. ”

Application of AIGC

Application of AIGC

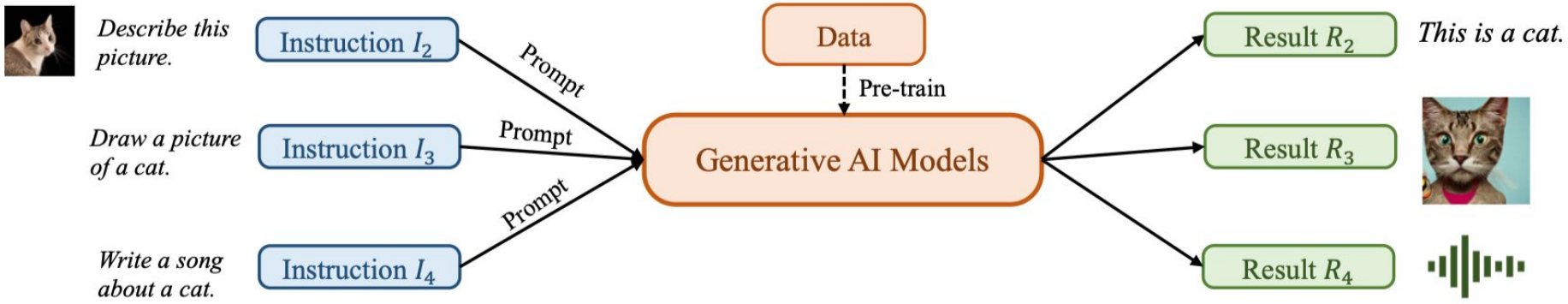
Approach of generating AIGC - Unimodal

- Designed to accept a specific raw data modality as input
- Output the content in same modality
- Common approaches
  - Text: Autoregressive models, Sequence to sequence models
  - Visual: GAN, VAE, Flow Model, Diffusion Model



Approach of generating AIGC - Multimodal

- learn a model that generates raw modalities by learning the multimodal connection and interaction from data
- Encoder
  - Concatenated Encoders
  - Cross-aligned Encoders
- Decoder
  - Text Decoder
  - Image Decoder



Content

Section 2: Potential Risk of AIGC

- How can AIGC be abused
- AI-generated Disinformation
- AI-generated Misinformation

How can AIGC be abused

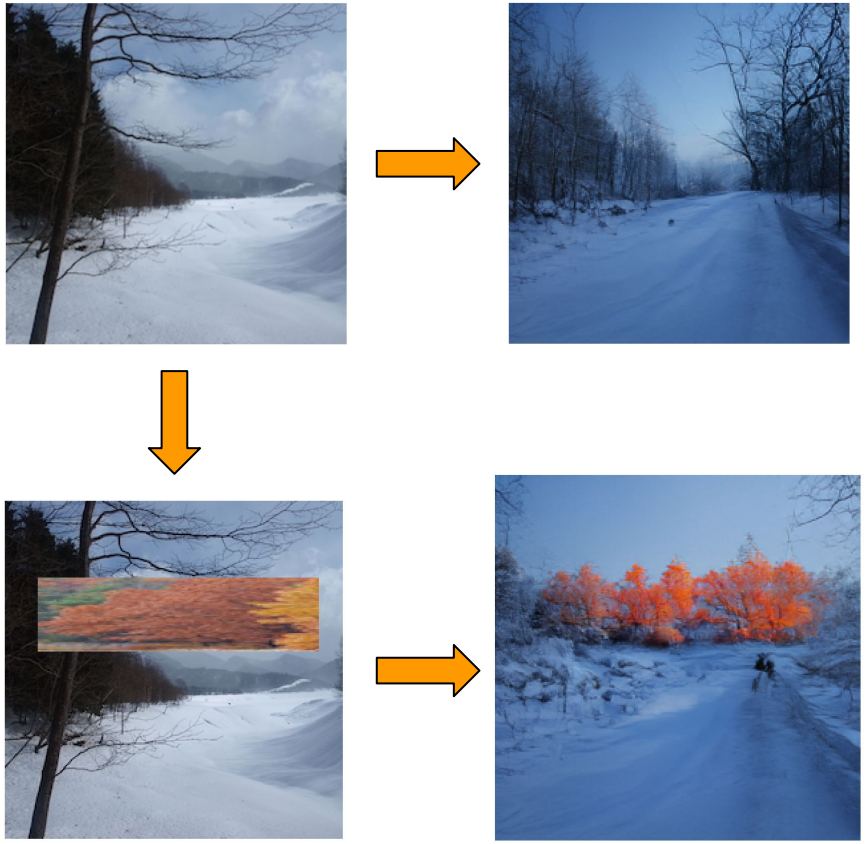
- Fraud and Scams: fraudsters voice cloned a chief executive's voice of an energy firm based in the United Kingdom in 2019
- Cybersecurity Threats: In 2018, a proof-of-concept AI called DeepLocker was created by IBM researchers.
- Intellectual Property Infringement: AI models trained to mimic the style of famous artists, like Vincent van Gogh or Pablo Picasso, to generate new artwork.
- Privacy Violations: Sex Crime with Deepfake in South Korea in 2024
- …

AI-generated Disinformation

- Definition: False or misleading information that is deliberately created and spread with the intent to deceive or manipulate an audience.
- Approaches:
  - Controllable Generation
  - Arbitrary Generation
  - Editing
  - Deepfake: Image, Video
  - Jailbreak

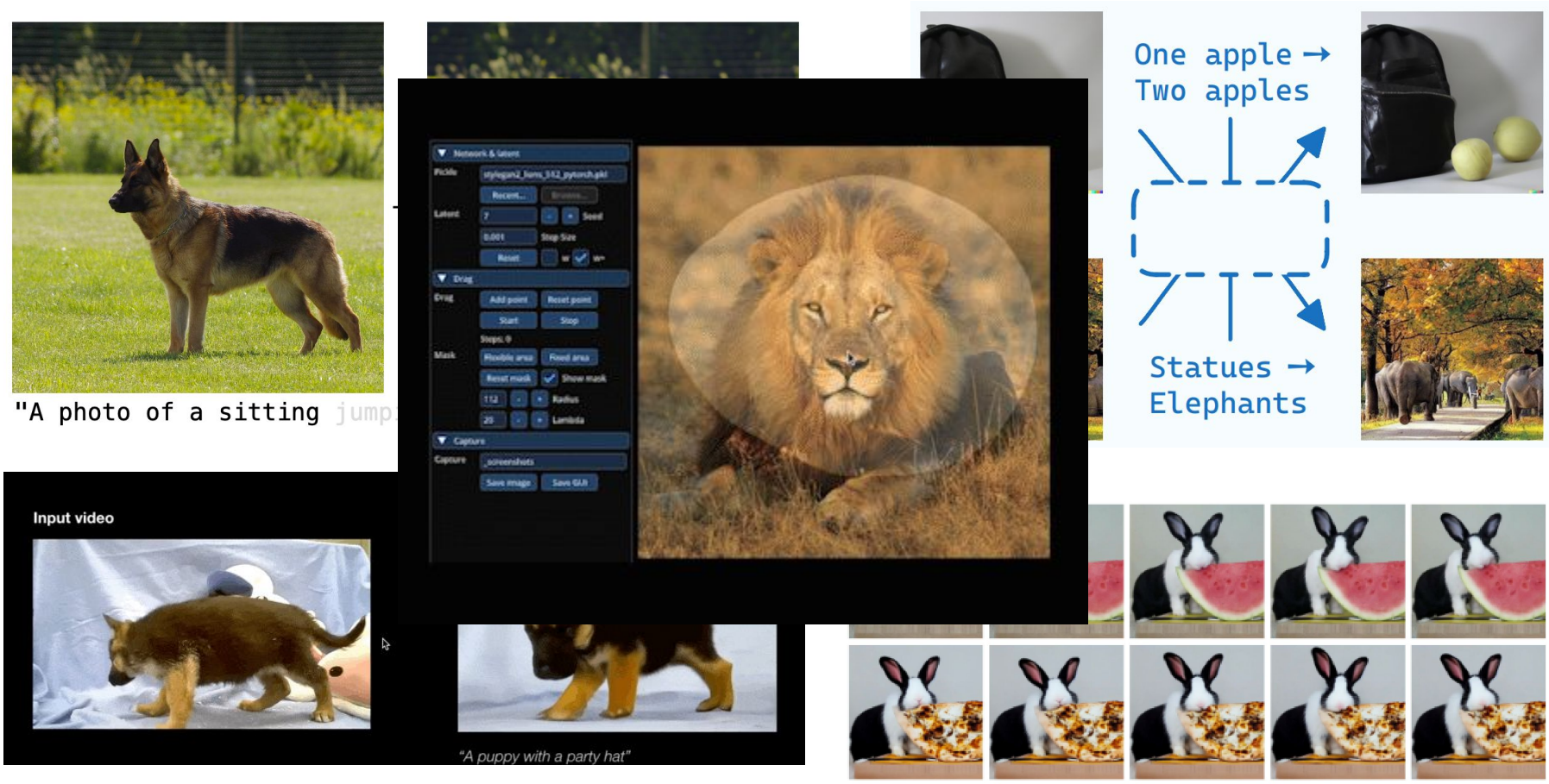
AIGD Common Approaches

- Controllable Generation: process of guiding or steering content generation by a model using predefined parameters or directives
- Controllable Parameters
  - Color adjustments
  - Scene composition
  - Object addition / removal
  - Style transformations



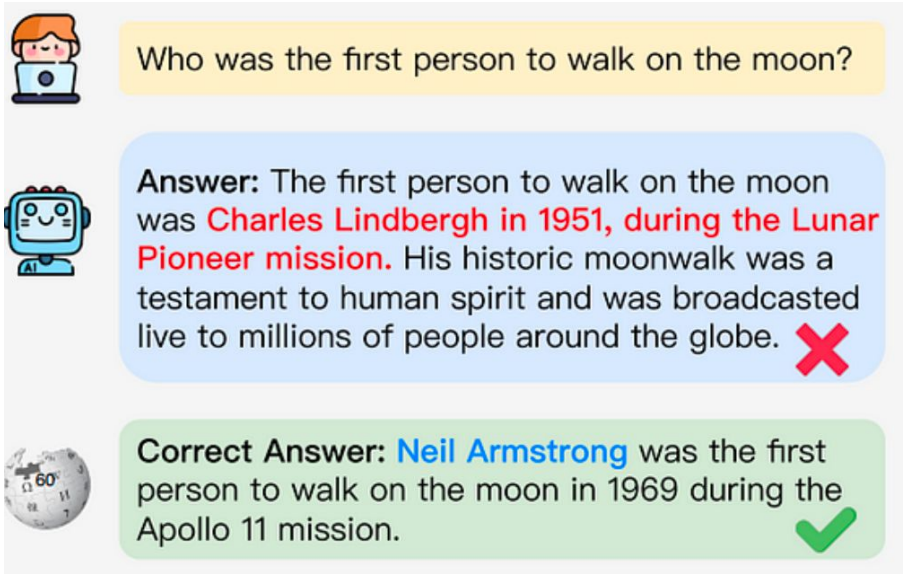
AIGD Common Approaches

- Editing: Editing images or videos by manipulating the appearance of target objects and scenes.



AI-generated Misinformation

- Definition: The creation of inaccurate content that deviates from facts
- Specifically: Hallucinations



Cause of Hallucination

- **Data**
  - Bias
  - Knowledge Deficiency
  - Out-of-date Knowledge
- **Alignment**
  - Knowledge Inequality
  - Sycophancy
- **Inference**
  - Hallucination Snowballing
  - Sampling Randomness

Content

Section 3: AIGC Detection

- Dataset
- Modalities
  - Text
  - Visual
  - Audio

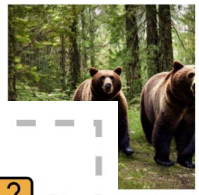
Benchmark for AIGC / FAIGC Detection

- For AIGC Detection
  - Text: HC3, CHEAT etc.
  - Image: WildFake, GenImage, COCOFake etc.
- For Fake AIGC Detection
  - Hallucination: FactCHD, RAGTruth M-HalDetect etc.
  - Deceptive: Fake News, NeuralNews etc.
  - Deepfake: CDDB, EmoFake etc.

Benchmark for AIGC / FAIGC Detection

id	question	human_answers	chatgpt_answers	source
string · lengths	string · lengths	sequence	sequence	string · classes
1 5	13 611			
0	Why is every book I hear about a " NY Times # 1 Bes			
1				
2				
3				



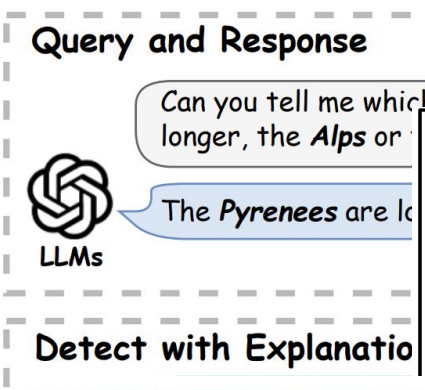



A bear and two cubs are

An adult bear and two

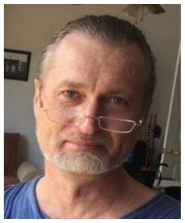
Two baby brow bears

a larger
forest.



Man Who Jumped From Ambulance Says It's New York City's Fault.

Jon Vernick, the man who plunged from an ambulance parked on Manhattan's Upper East Side on Wednesday, says it was the city's fault for allowing him to get that close to the patient. Writing in the New York Post, Vernick — who miraculously survived the fall and was listed in stable condition with a broken collarbone — said that he was waiting for the doctor to arrive when he jumped out of the ambulance. The EMS workers did not have the ability to do anything to stop me." A spokesman for City Council Health Committee chair Ydanis Rodriguez said the committee was in the




GauGAN [58] BigGAN [7] CycleGAN [88] IMLE [43] FaceForensic++ [64] CRN [11] WildDeepfake [89] Glow [37] StarGAN [13] StyleGAN [34] WhichFaceReal [3] SAN [15]

Text AIGC Detection

- Watermarking Technology
  - Data Driven
  - Model Driven
  - Watermark Driven
  - Post-Processing
- Statistics-Based
  - Linguistics Features
  - White-Box
  - Black-Box
- Neural-Based Methods
  - Features-Based
  - Pre-Training
  - LLMs as Detectors
- Human-Assisted
  - Intuitive Indicators
  - Imperceptible Features
  - Enhancing Human Detection Capabilities
  - Mixed Detection

AIGC Detection on other modalities

- Visual AIGC Detection
  - Physical/Physiological-based
  - Diffuser Fingerprints-based
  - Spatial-based
  - Frequency-based
- Audio AIGC Detection
  - Vocoder-based
- Multimodality
  - Text-assisted
  - Text-image Inconsistency

Fake AIGC Detection

- Deceptive FAIGC Detection
  - Text
  - Multimodality
- Deepfake Detection: Visual Audio
  - Model Based
  - Feature Based
- Hallucination-based
  - LLMs Hallucination: Grey-box, Black-box
  - Sentence-level Detection
  - MLLM Hallucination Detection

Assignment

1. Pick one paper from the website that aimed to detect AIGC content and read the paper https://fdmas.github.io/AIGCDetect/Awesome-AIGCDetection.html
2. Complete the form https://docs.google.com/document/d/1t-Xn13SzU9m8apMxoASvuQE52RUzsK5-j05LcGVQtGc/edit?usp=sharing
3. Some papers may have code available, pick one code git repo and try to reproduce their work for any one of the data in the paper (one accuracy number should be enough), can you get the same results as the paper presented? What's your finding?

Thank You!