When Art Meets AI

Yike Guo

郭毅可

Provost

The Hong Kong University of Science and Technology

Director

Hong Kong Generative AI Research and Development Centre

AI : Building a Human-like Machine

Like human to think

Prediction & Reasoning

Like human to see

Computer Vision (CV)

Like human to hear

Acoustic Speech Recognition (ASR)

Like human to talk

Text-To-Speech (TTS)

Like human to act

Robotics

Like EXPERT to teach

Smart Education

Like EXPERT to trade & invest

Smart Trading & Investing

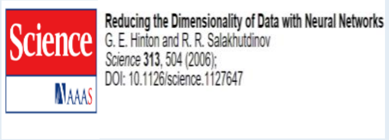
Like EXPERT to make medicine

iDrug

We Are Towards Future AGI

Specialized models:
One model solves one specific problem

DL Theory Breakthrough


Science
AAAS

Reducing the Dimensionality of Data with Neural Networks
G. E. Hinton and R. R. Salakhutdinov
Science 313, 504 (2006);
DOI: 10.1126/science.1127647

Deep Confidence Networks

2006

Large-scale speech recognition

Switchboard Errors Reduced by 9%

2011

ImageNet Competition

1000 classes， 1 million data

2012

Face Recognition

LFW recognition rate of 99%, surpassing humans

2014

Go Games

AlphaGo 4:1 Lee Sedol

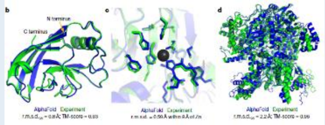
2016

Texas Hold'em

First Surpassed humans in complex multi-player games

2019

AlphaFold

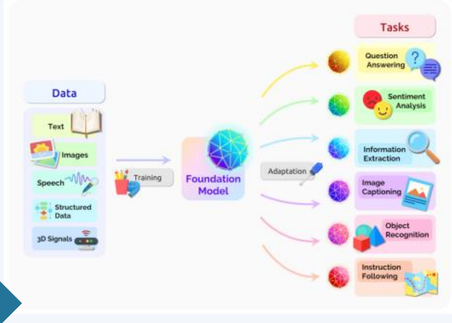


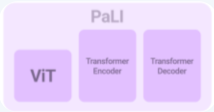
Recorded accuracy in protein structure prediction

2021

Foundation models:

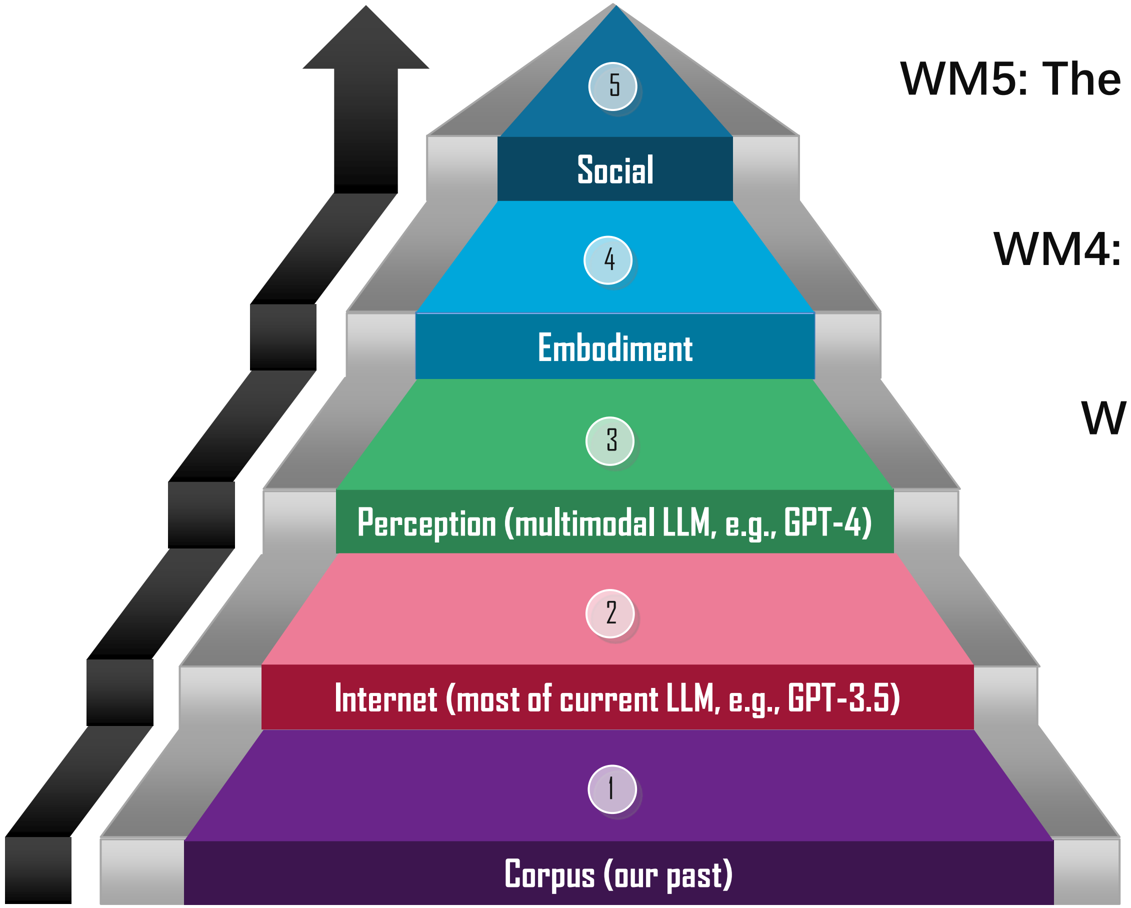
One generative model for multiple tasks and modalities





DALL·E 2
DALL·E 2 is a new AI system that can create realistic images and art from a description in natural language.

Future Development: From GPT to World Models



5

Social

4

Embodiment

3

Perception (multimodal LLM, e.g., GPT-4)

2

Internet (most of current LLM, e.g., GPT-3.5)

1

Corpus (our past)

WM5: The Social World

WM4: Embodiment and Action

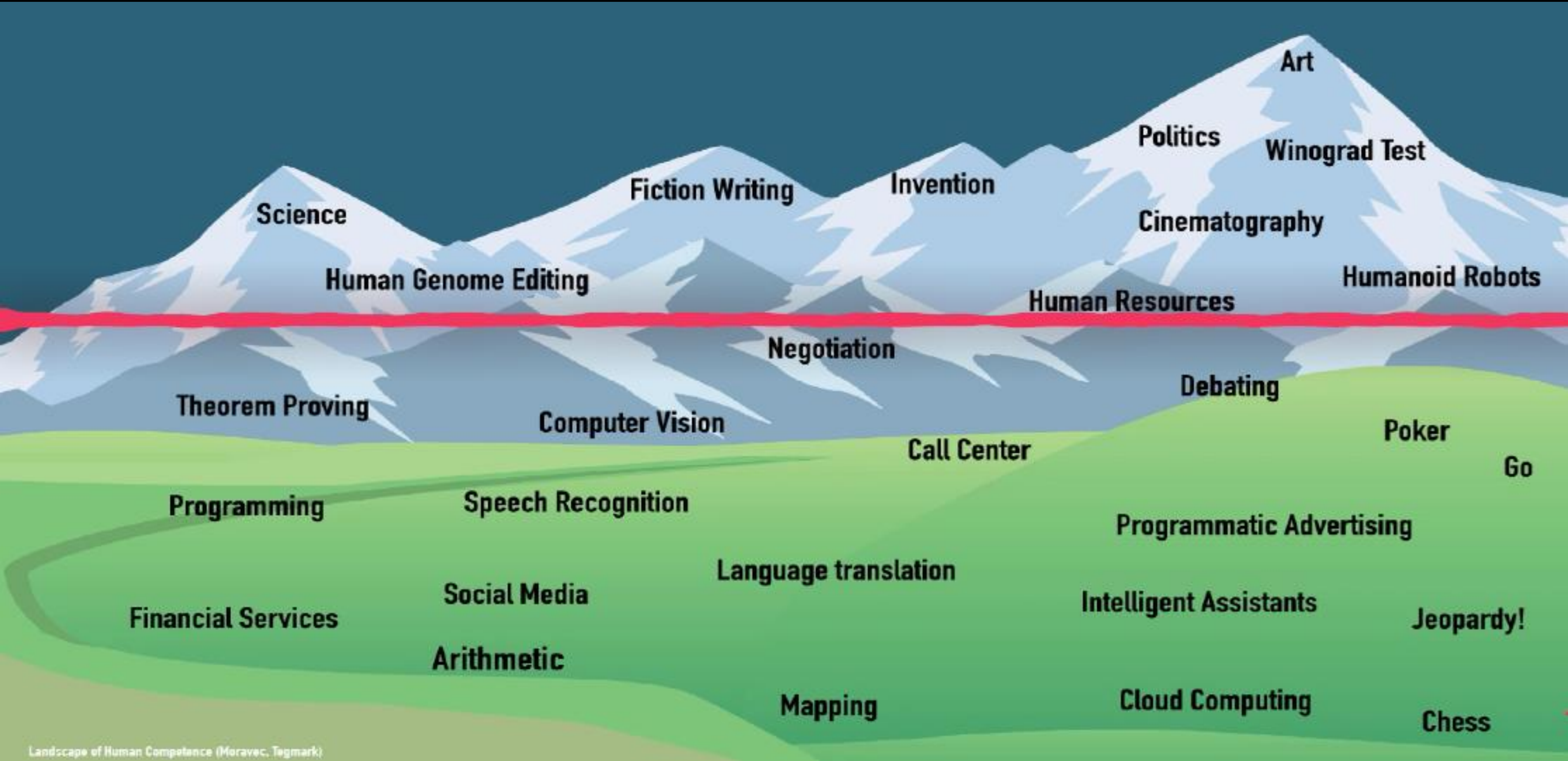
WM3: The Visual/Audio World

WM2: The Text World

WM1: Corpora & Representations

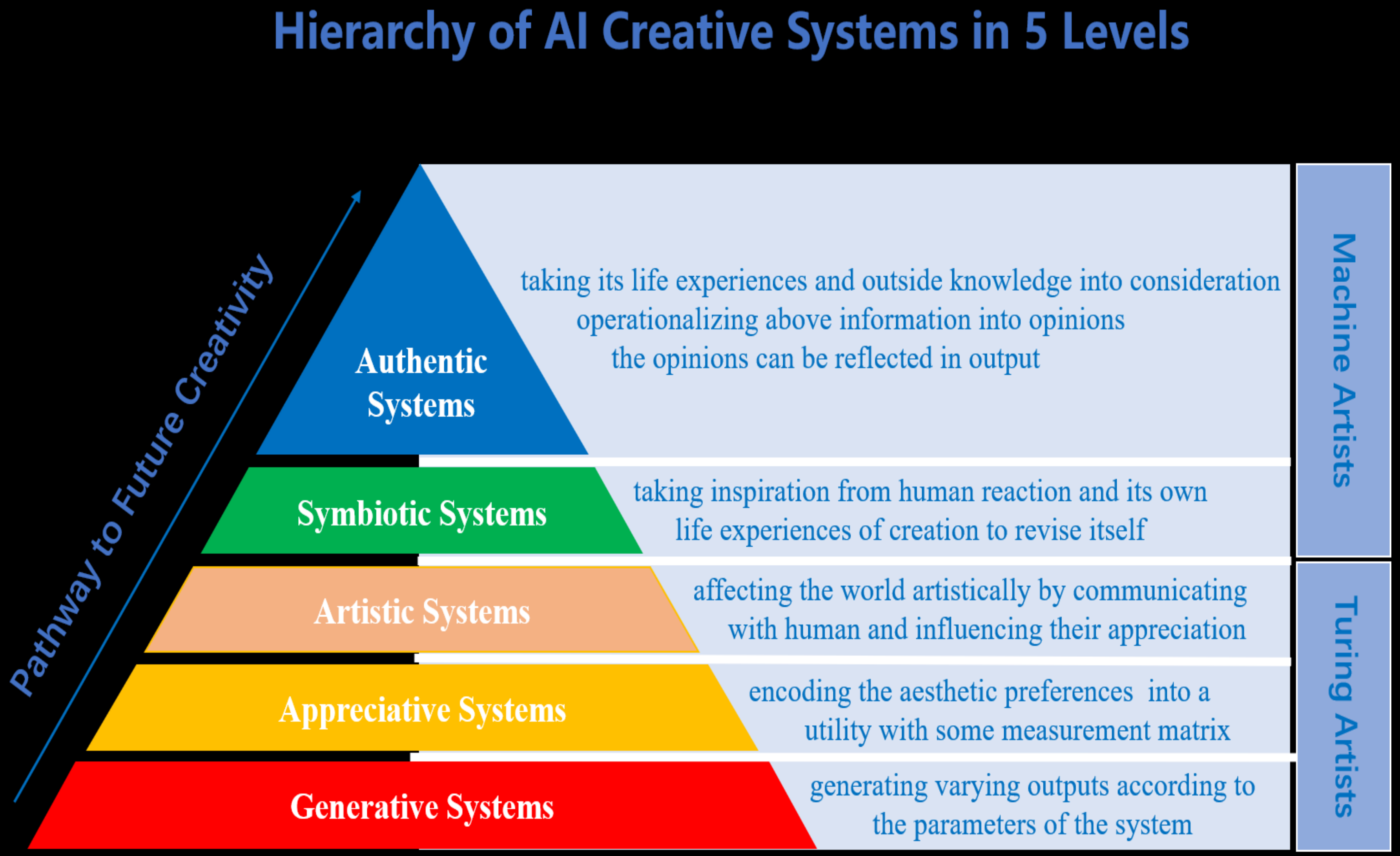
“Experience Grounds Language”, EMNLP 2020

Art : Unique to Human, Challenging to Machine

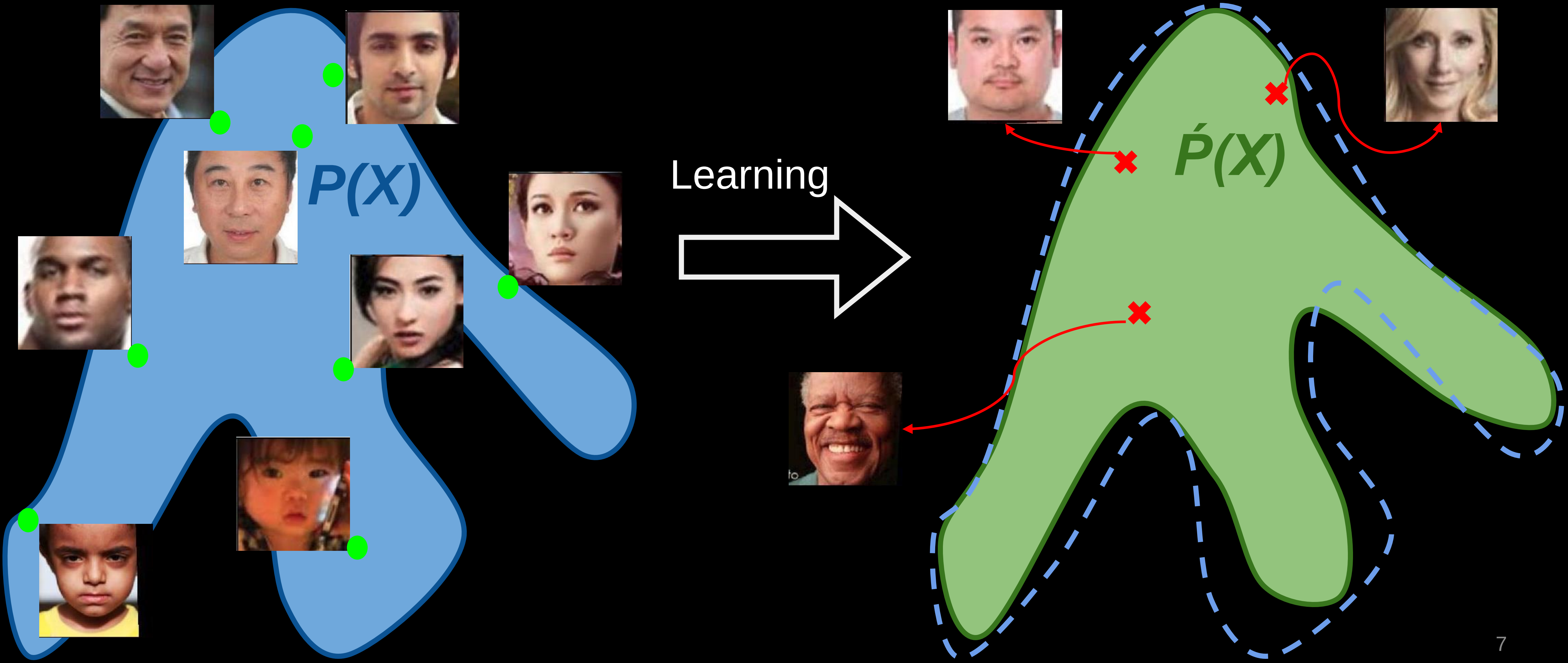


Landscape of Human Competence (Moravec, Tegmark)

Two Types of AI Artist Systems


Hierarchy of AI Creative Systems in 5 Levels
Pathway to Future Creativity
Authentic Systems
taking its life experiences and outside knowledge into consideration
operationalizing above information into opinions
the opinions can be reflected in output
Symbiotic Systems
taking inspiration from human reaction and its own life experiences of creation to revise itself
Artistic Systems
affecting the world artistically by communicating with human and influencing their appreciation
Appreciative Systems
encoding the aesthetic preferences into a utility with some measurement matrix
Generative Systems
generating varying outputs according to the parameters of the system
Machine Artists
Turing Artists

Today’s Turing Artists : Mimicking Human



Training Data

P(X)

Learning

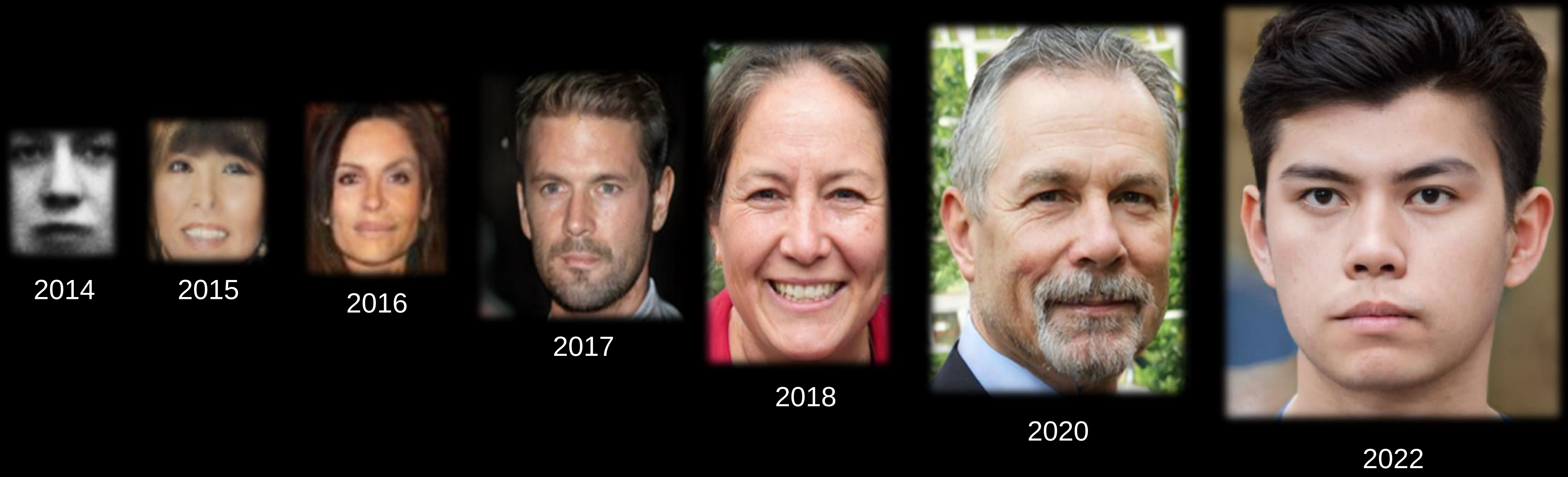
Sampling

Ṕ(X)

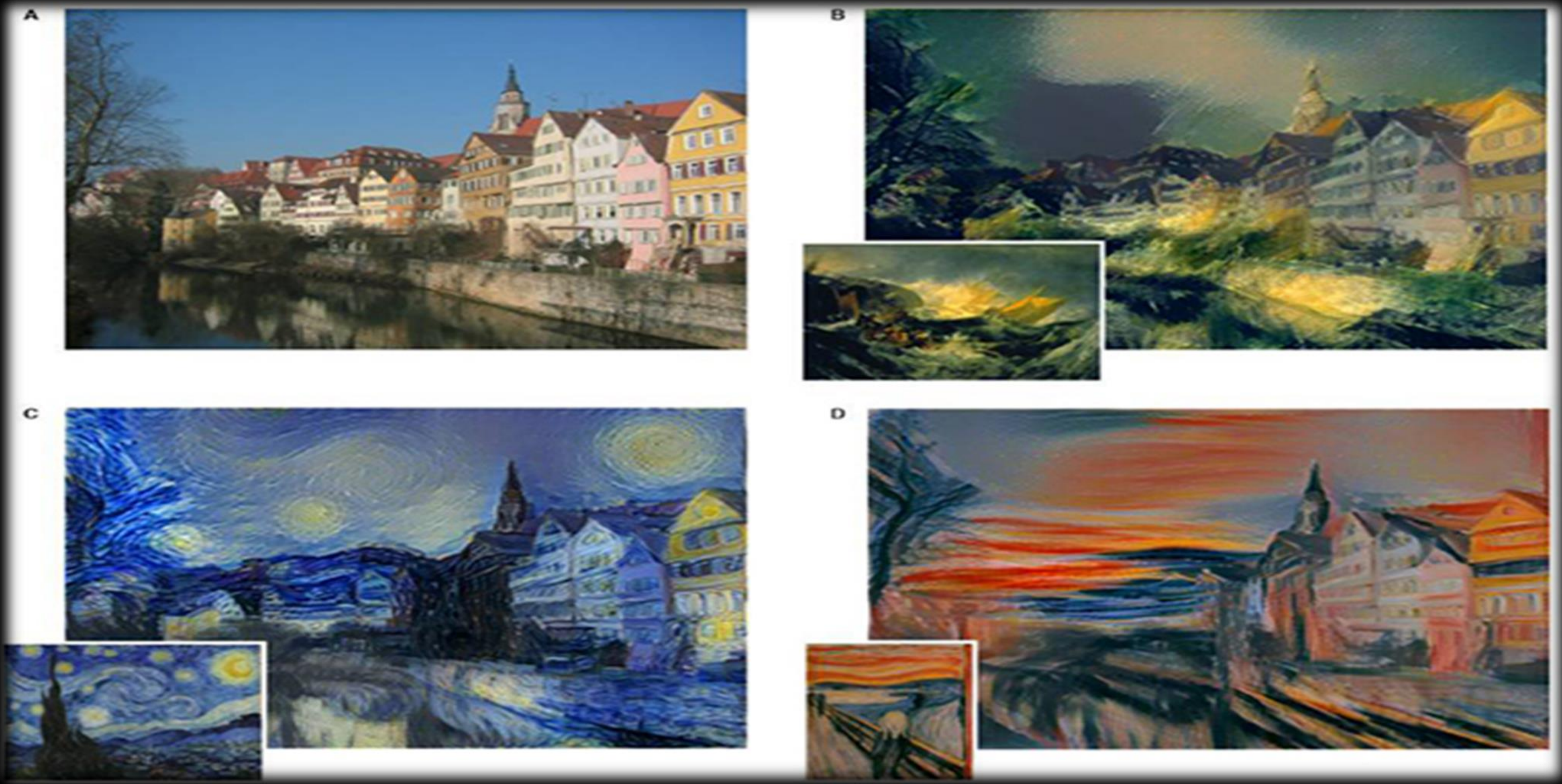
Mimicking Artefacts : Generative System

GAN PROGRESS ON FACE GENERATION

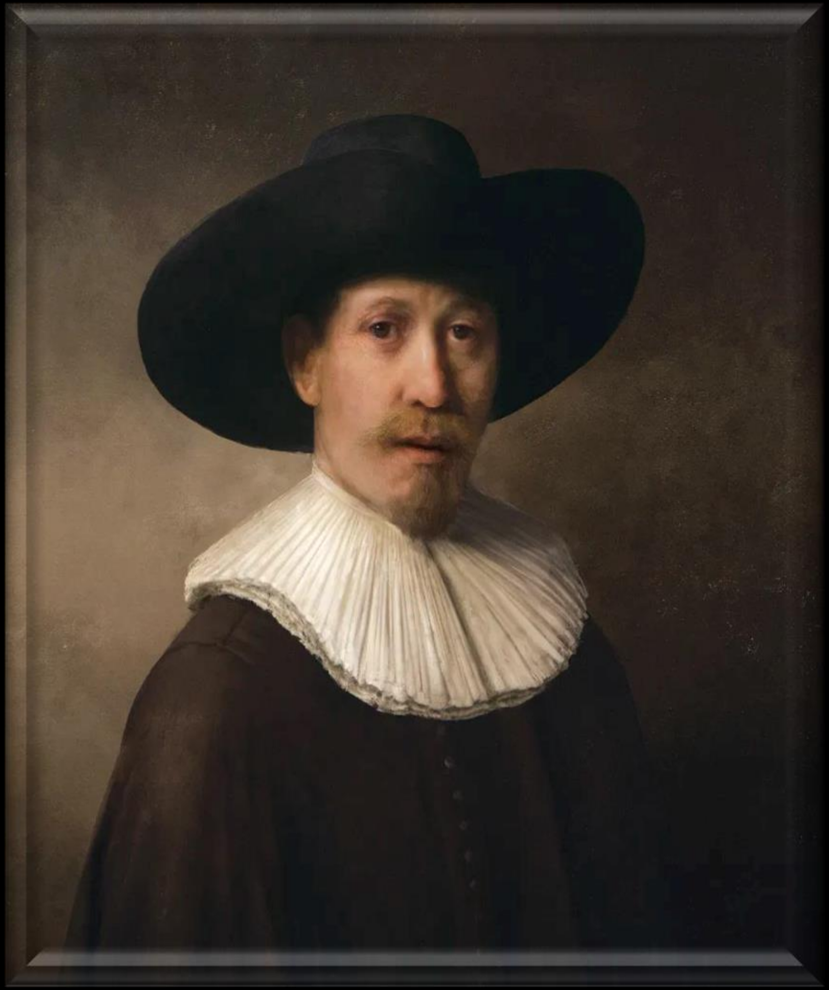
Source: Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021; Hou et al., 2022



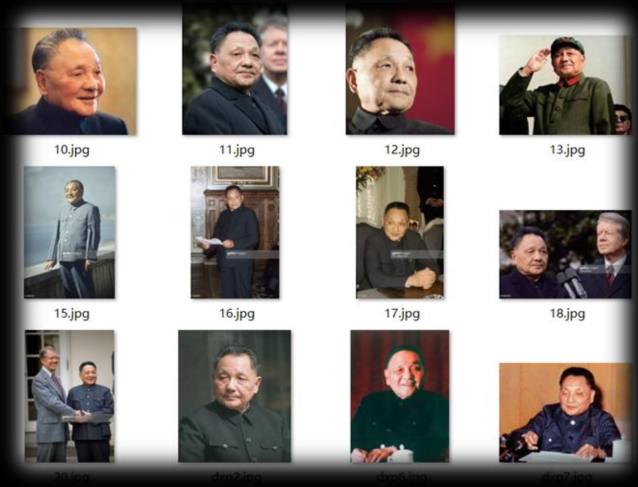
Mimicking Styles : Appreciative System

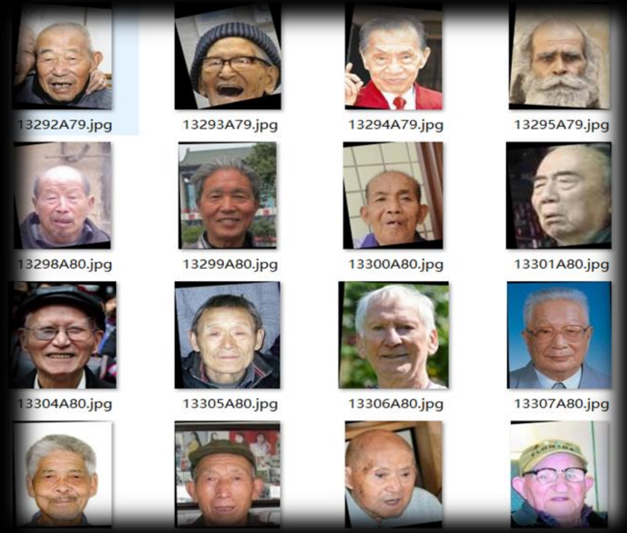


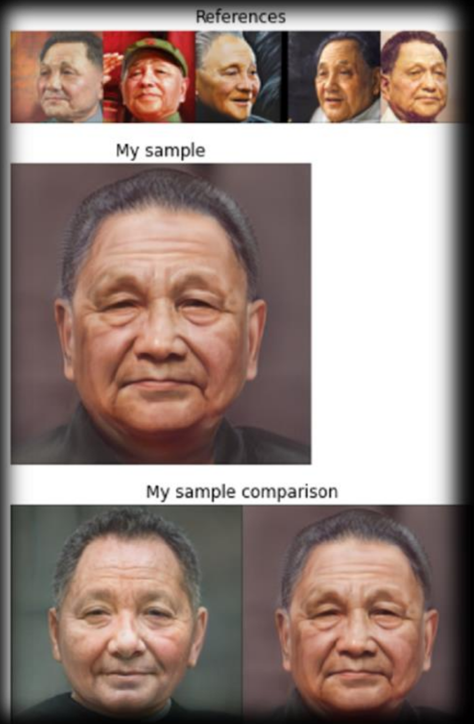
Mimicking Inspirations : Artistic System

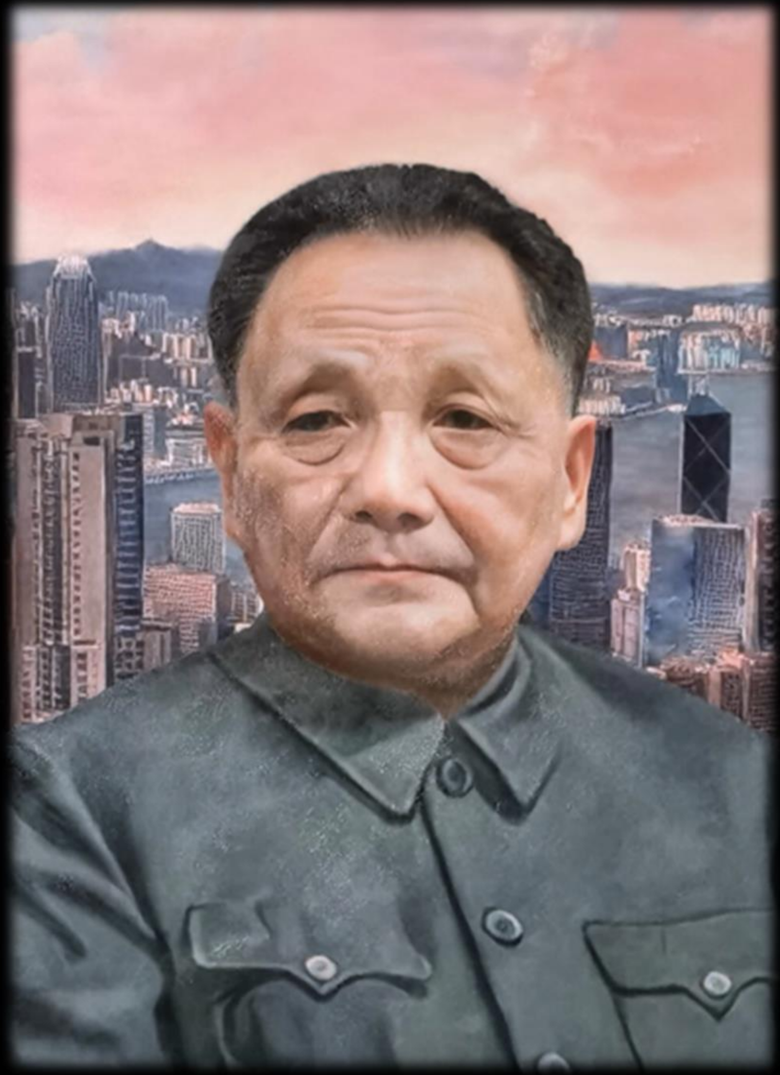


Created by Our Turing Painter

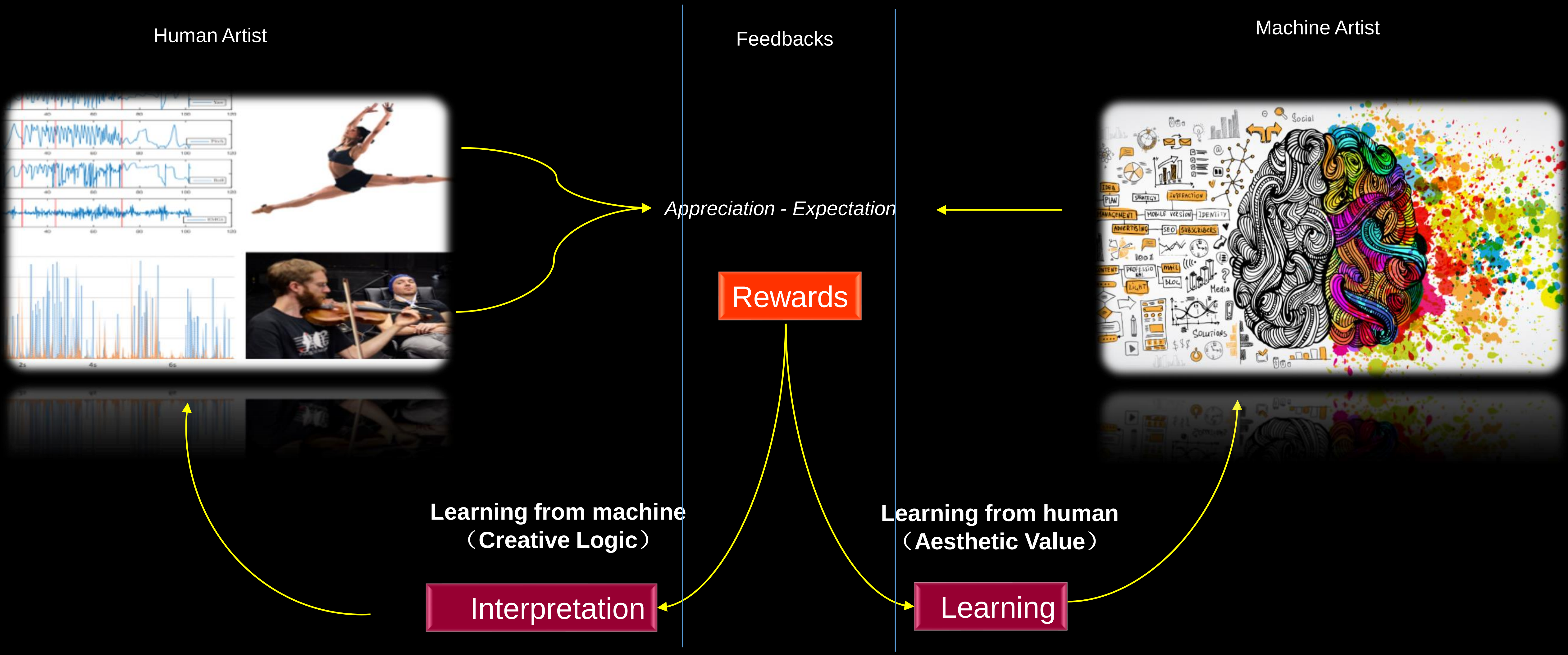








Machine Artists : Building A Creative Machine



Human Artist

Feedbacks

Machine Artist

Appreciation - Expectation

Rewards

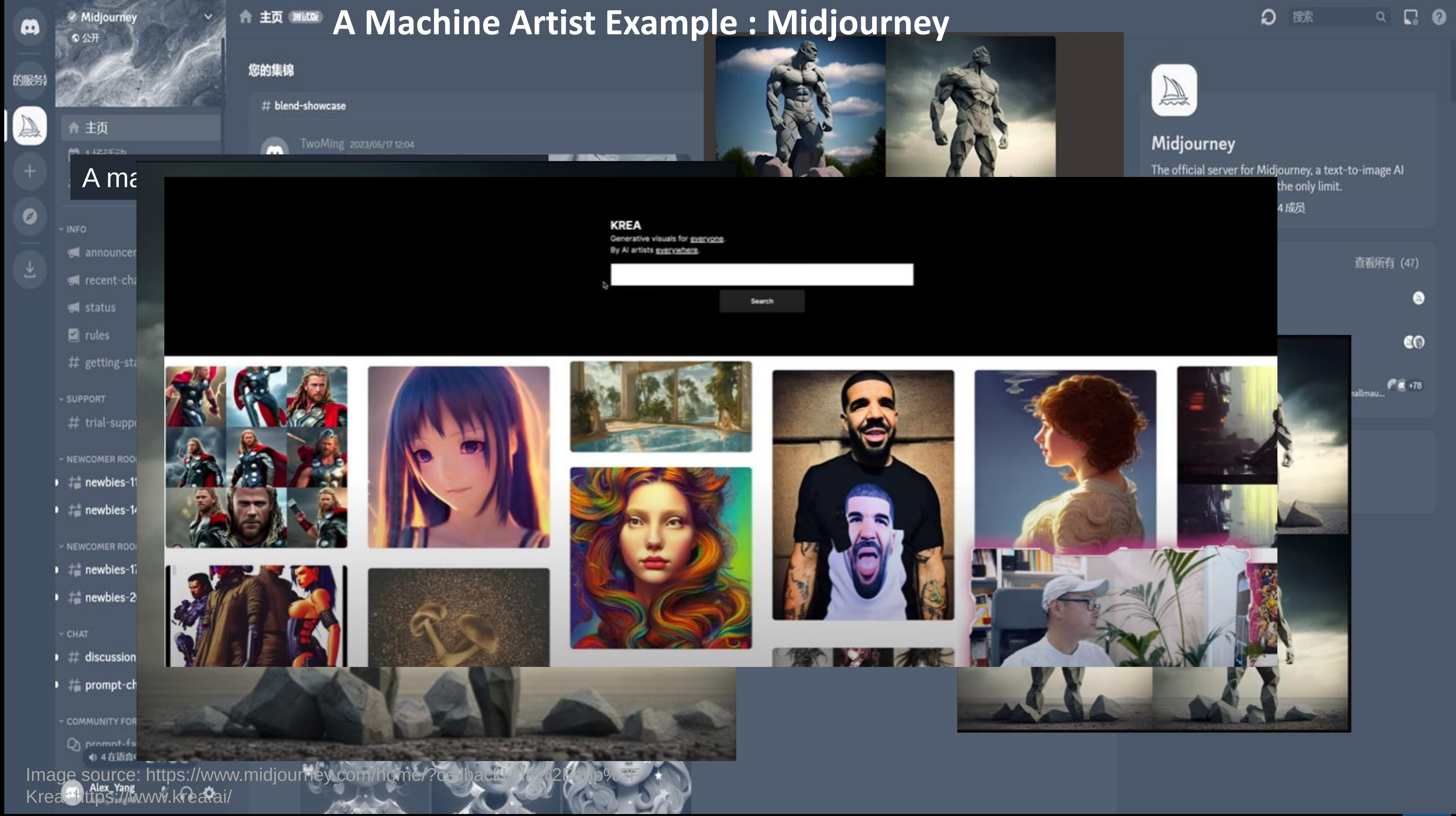
Learning from machine
（Creative Logic）

Learning from human
（Aesthetic Value）

Interpretation

Learning

A Machine Artist Example : Midjourney



Midjourney

The official server for Midjourney, a text-to-image AI ... the only limit.

4 成员

KREA

Generative visuals for everyone.

By AI artists everywhere.

Image source: https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F

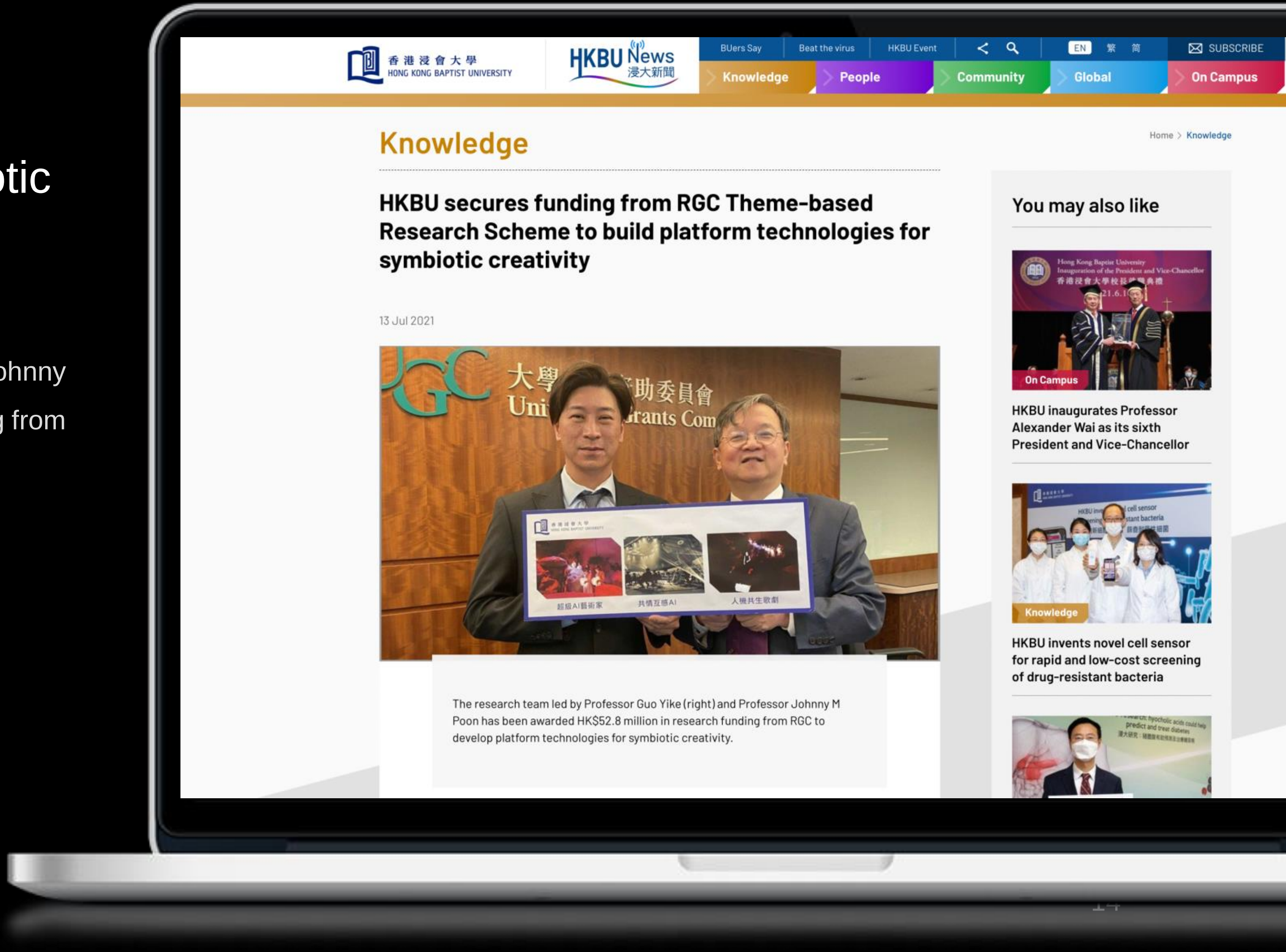
Krea: https://www.krea.ai/

Our Journey Starts from Here

HKBU secures funding from RGC Theme-based Research Scheme to build platform technologies for symbiotic creativity

The research team led by Professor Guo Yike and Professor Johnny M Poon has been awarded HK$52.8 million in research funding from RGC to develop platform technologies for symbiotic creativity

HK $52.8 million



香港浸會大學 HONG KONG BAPTIST UNIVERSITY | HKBU News 浸大新聞

BUers Say | Beat the virus | HKBU Event | EN 繁 简 | SUBSCRIBE

Knowledge | People | Community | Global | On Campus

Home > Knowledge

Knowledge

HKBU secures funding from RGC Theme-based Research Scheme to build platform technologies for symbiotic creativity

13 Jul 2021

The research team led by Professor Guo Yike (right) and Professor Johnny M Poon has been awarded HK$52.8 million in research funding from RGC to develop platform technologies for symbiotic creativity.

You may also like

On Campus

HKBU inaugurates Professor Alexander Wai as its sixth President and Vice-Chancellor

Knowledge

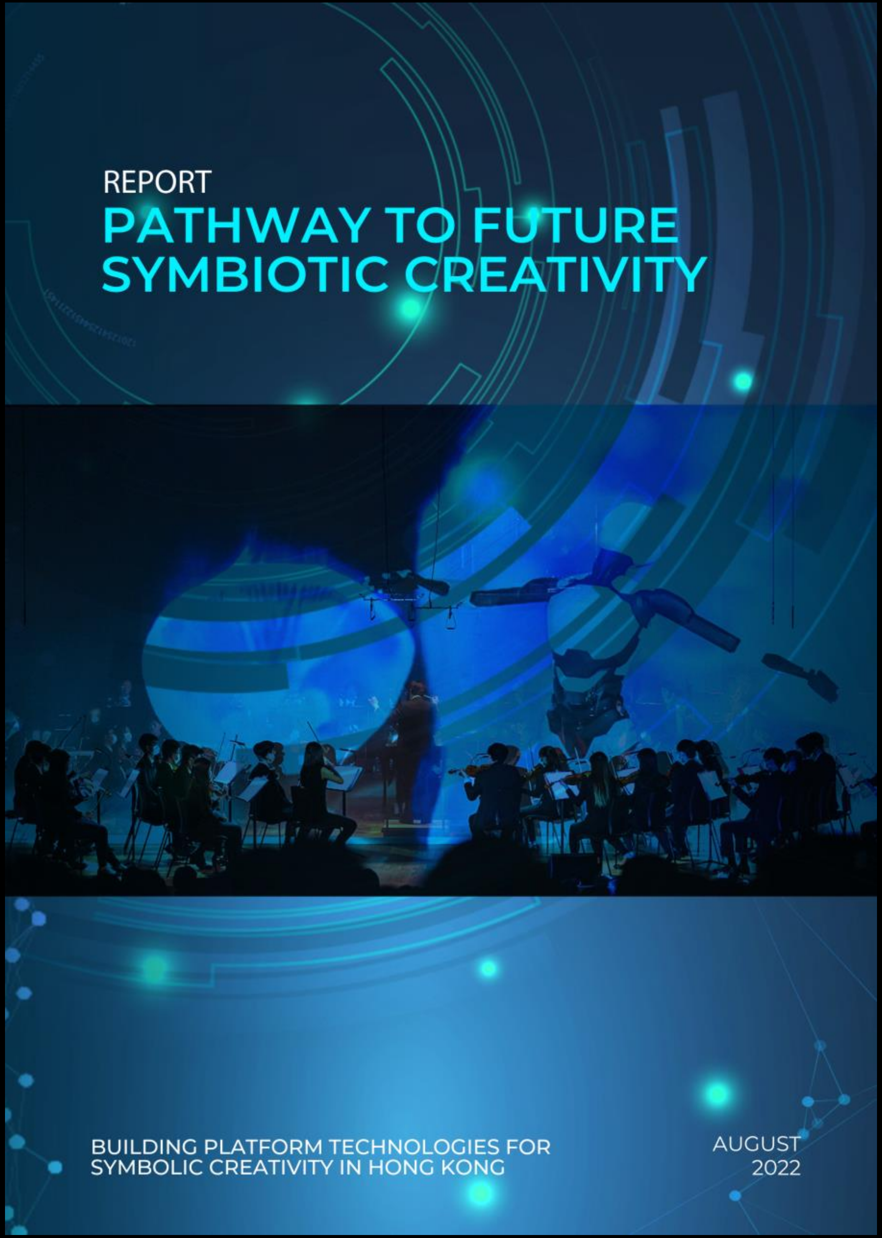
HKBU invents novel cell sensor for rapid and low-cost screening of drug-resistant bacteria

AI Meets Art : Our Ambition and Mission

“I want to paint humanity, humanity and again humanity.”

— Vincent van Gogh

- ✓ AI enables new forms of art
- ✓ Art brings new innovation to AI



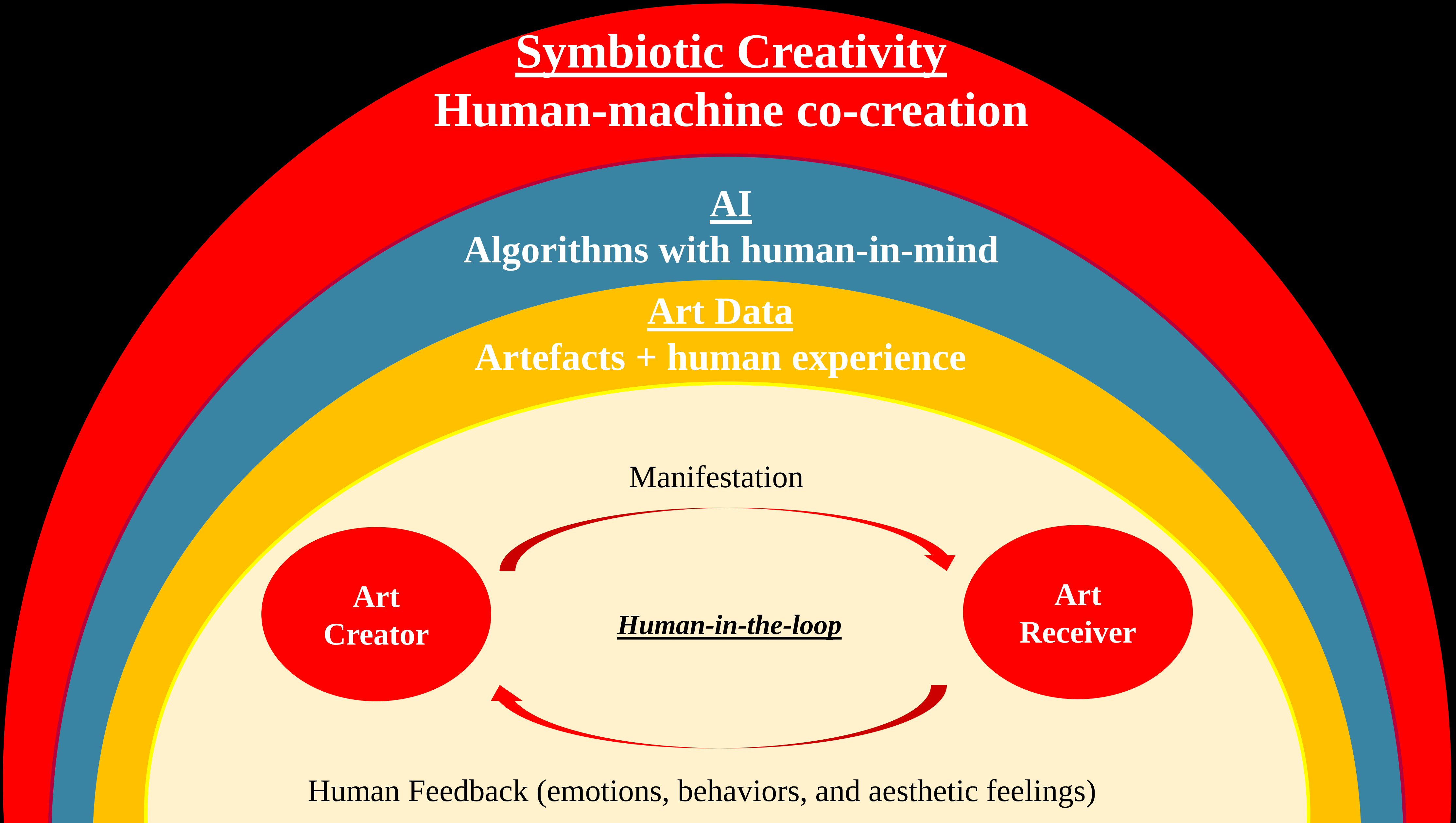
REPORT

PATHWAY TO FUTURE SYMBIOTIC CREATIVITY

BUILDING PLATFORM TECHNOLOGIES FOR
SYMBOLIC CREATIVITY IN HONG KONG

AUGUST
2022

Ambitious AI-Art: Building Unique Art Data Sets



Symbiotic Creativity

Human-machine co-creation

AI

Algorithms with human-in-mind

Art Data

Artefacts + human experience

Manifestation

Art Creator

Human-in-the-loop

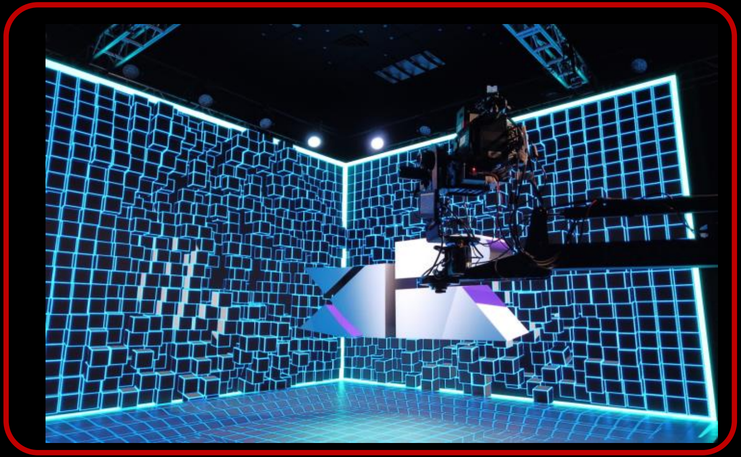
Art Receiver

Human Feedback (emotions, behaviors, and aesthetic feelings)

Building Platform Technologies for Symbiotic Creativity in Hong Kong (TRS T45-205/21-N)

Yaw
Pitch
Roll
EMG1
1.0
0.8
0.6
0.4
0.2
0.0
2s
4s
6s

AI Technology for Artefact Creation
– Building **ALGORITHMS**

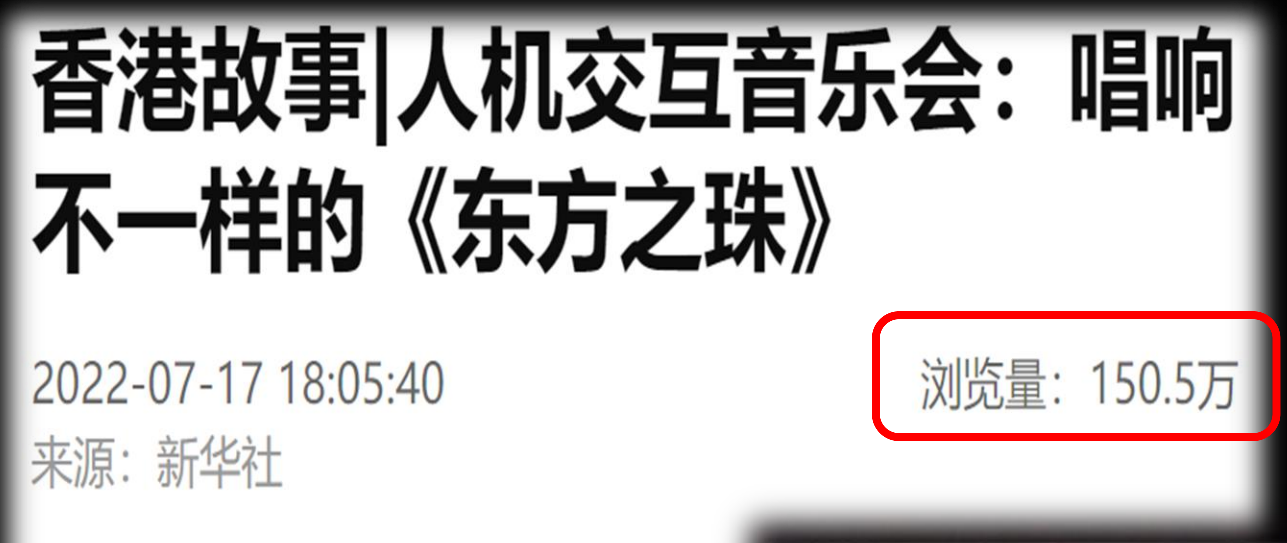


Art Tech for Manifestation and Delivery
– Building **ENVIRONMENT**

Platform Deployment and Applications
– Building **APPLICATIONS**

First Major Deliverable –

AI Choir



香港故事|人机交互音乐会：唱响不一样的《东方之珠》

2022-07-17 18:05:40

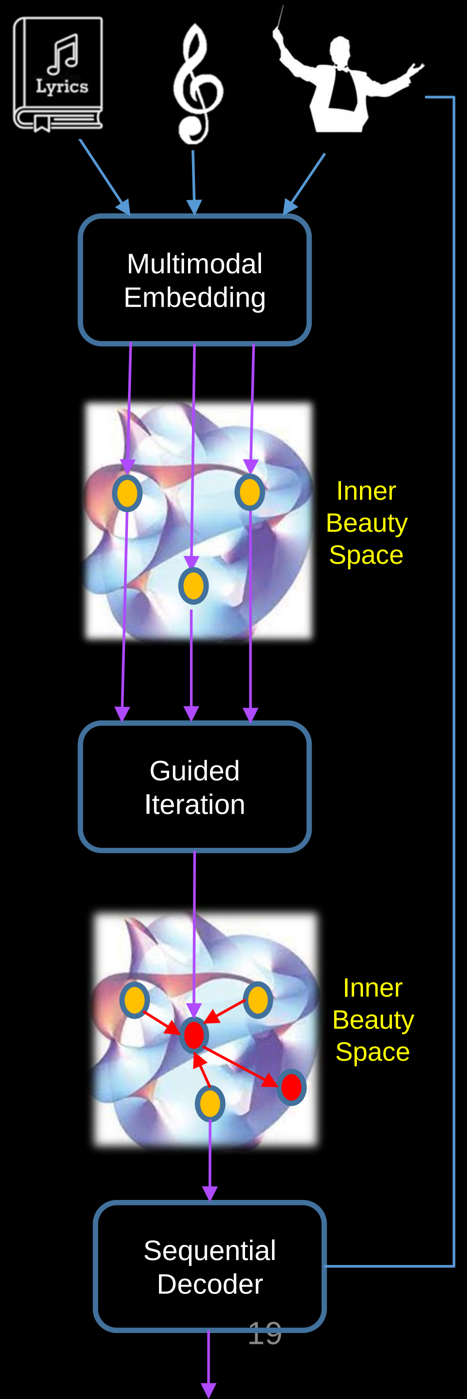
浏览量：150.5万

来源：新华社

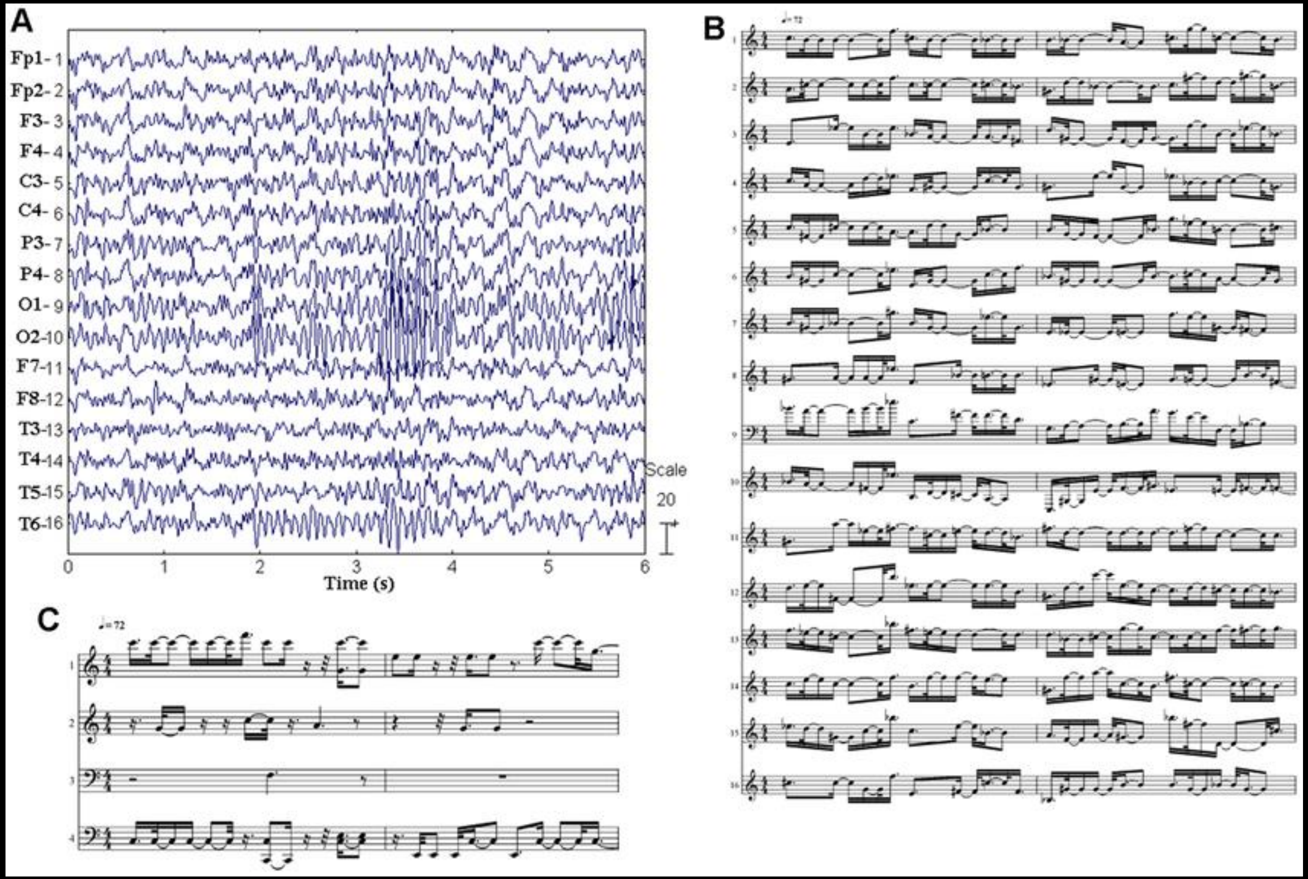
新华港澳台

Visual Art Generation

Cloud ⟶ Flower



Singing Foundation Model



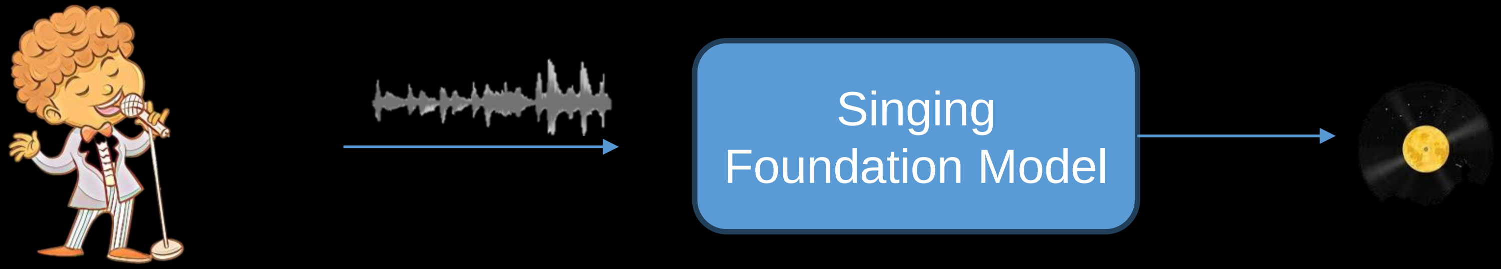
Humans create emotional and beautiful audios to by making sequential melody activate and satisfy our brain.

The emotions and characters we express in speaking and singing are controlled by our brains.

When we build the foundation model of audios, we are also modeling ourselves.

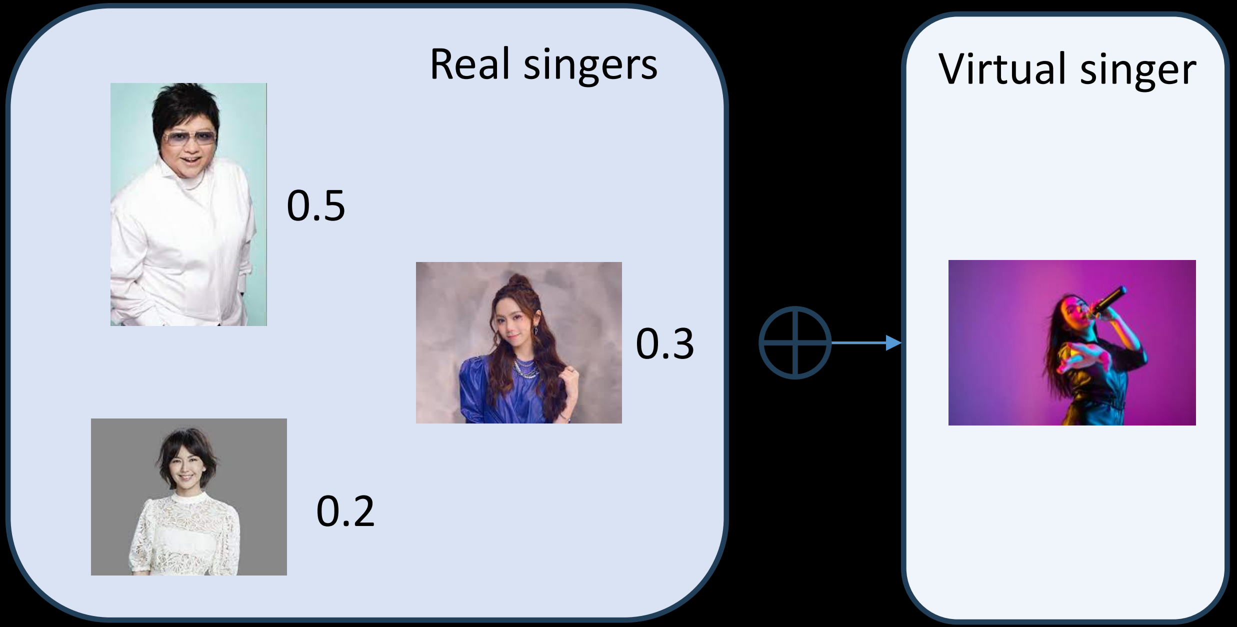
Targets

Create professional songs （歌） just by singing（唱）. (“唱”歌)



New Singer Creation (Singer Algebra)

- The model would learn an existing professional singer, and create new virtual singers which can be a combination of existing singers in terms of
  - a) **timbre**
  - b) **singing ranges**
  - c) **singing style**
- The virtual singers sing with the human singer, or replace the human singer, in the final production.

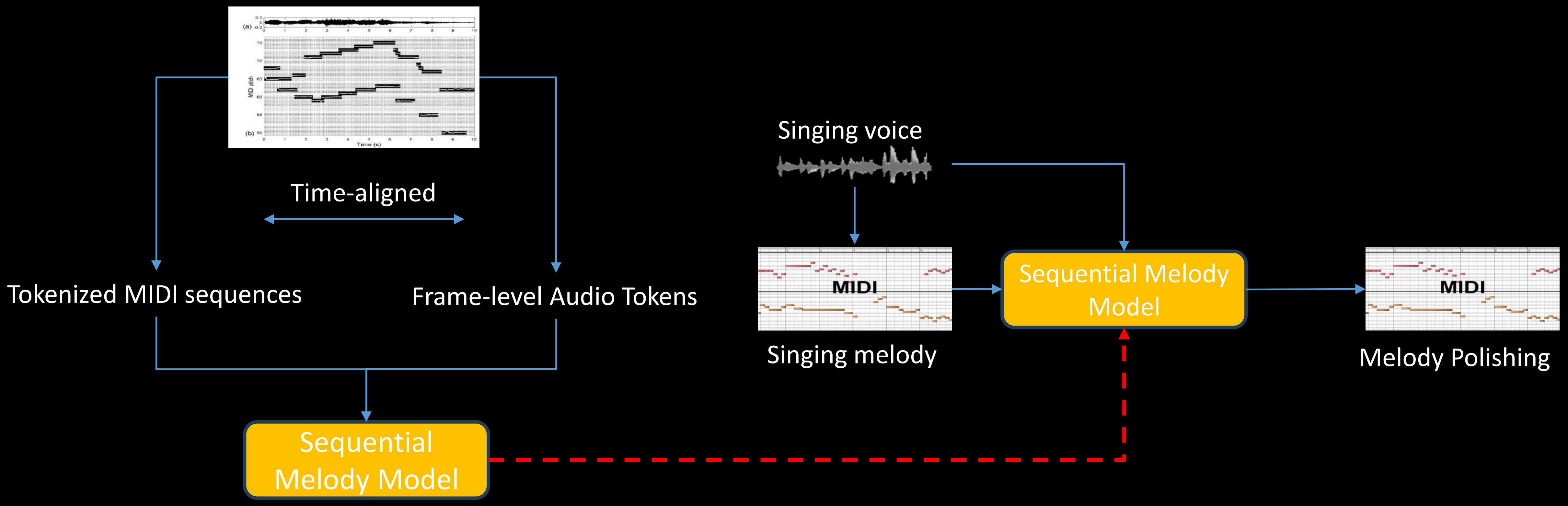


Producing an existing singer is a special case when one weight becomes 1.0

Targets

Melody Perfection

- The model can tune the input melody (in audio and MIDI) to a professional level while keeping the original structure. The singing voice is reproduced with original lyrics and finetuned melodies.

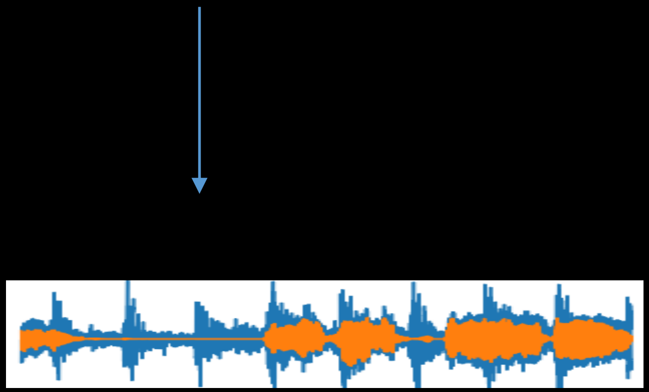


Targets

Accompaniment Generation

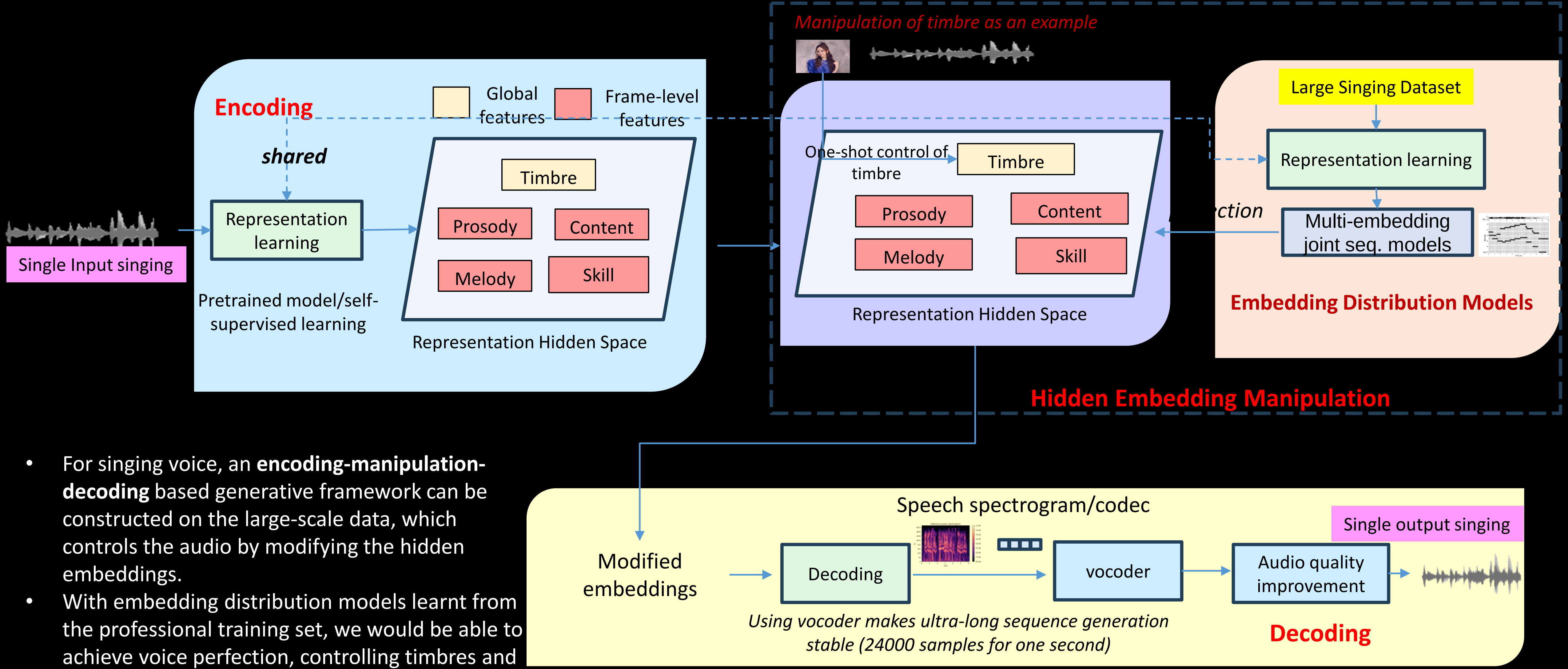
- The model can generate multi-instrument accompaniment based on the vocal audio.
  - Accompaniments are in harmony with melody of the input singing.

Singing Audio



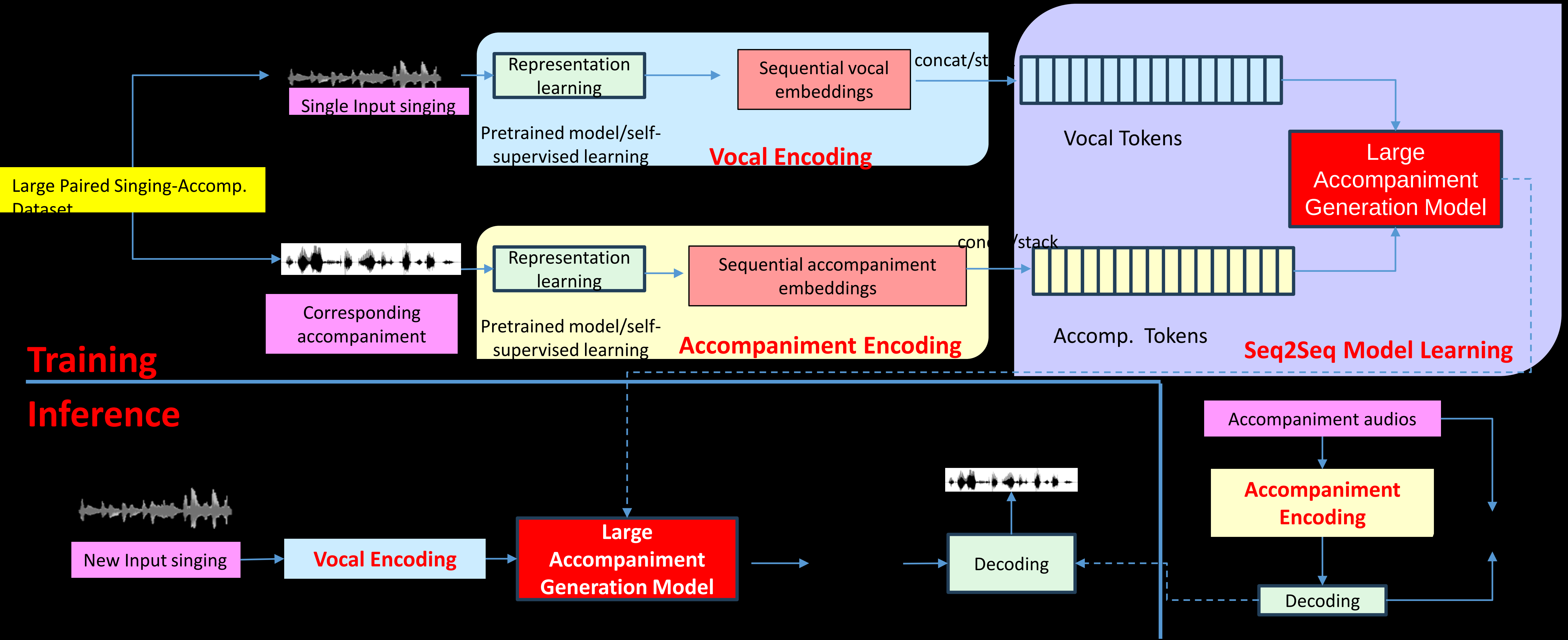
Time-Aligned
Accompaniment Audio

Overall Framework


Manipulation of timbre as an example
Encoding
Global features
Frame-level features
shared
Single Input singing
Representation learning
Pretrained model/self-supervised learning
Timbre
Prosody
Content
Melody
Skill
Representation Hidden Space
One-shot control of timbre
Timbre
Prosody
Content
Melody
Skill
Representation Hidden Space
Large Singing Dataset
Representation learning
ction
Multi-embedding joint seq. models
Embedding Distribution Models
Hidden Embedding Manipulation
Speech spectrogram/codec
Single output singing
Modified embeddings
Decoding
vocoder
Audio quality improvement
Using vocoder makes ultra-long sequence generation stable (24000 samples for one second)
Decoding

- For singing voice, an **encoding-manipulation-decoding** based generative framework can be constructed on the large-scale data, which controls the audio by modifying the hidden embeddings.
- With embedding distribution models learnt from the professional training set, we would be able to achieve voice perfection, controlling timbres and finetune melodies. Details will be explained later.

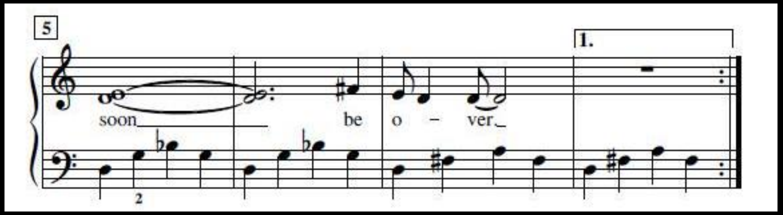
Overall Framework

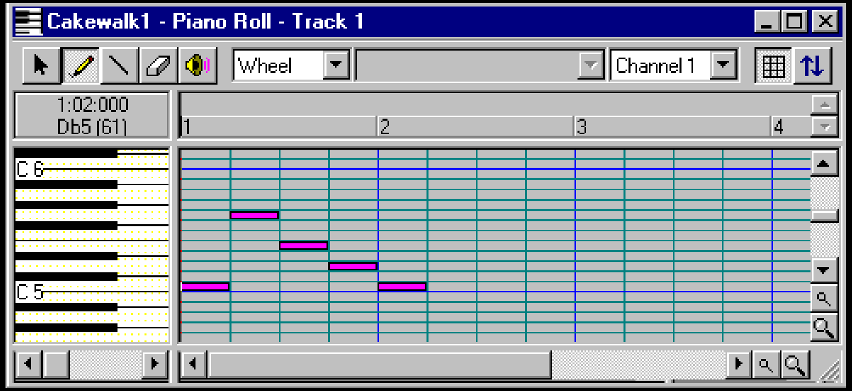

Large Paired Singing-Accomp. Dataset
Single Input singing
Representation learning
Sequential vocal embeddings
concat/st
Pretrained model/self-supervised learning
Vocal Encoding
Vocal Tokens
Large Accompaniment Generation Model
Corresponding accompaniment
Representation learning
Sequential accompaniment embeddings
con /stack
Pretrained model/self-supervised learning
Accompaniment Encoding
Accomp. Tokens
Seq2Seq Model Learning
Training
Inference
Accompaniment audios
Accompaniment Encoding
New Input singing
Vocal Encoding
Large Accompaniment Generation Model
Decoding
Decoding

- For accompaniment, given paired vocal/accompaniment audio data, a large **sequence-to-sequence model** can be trained on the token sequences of the vocal and accompaniment audios.
- Accompaniment is first generated in the token form, and finally decoded into the waveforms.

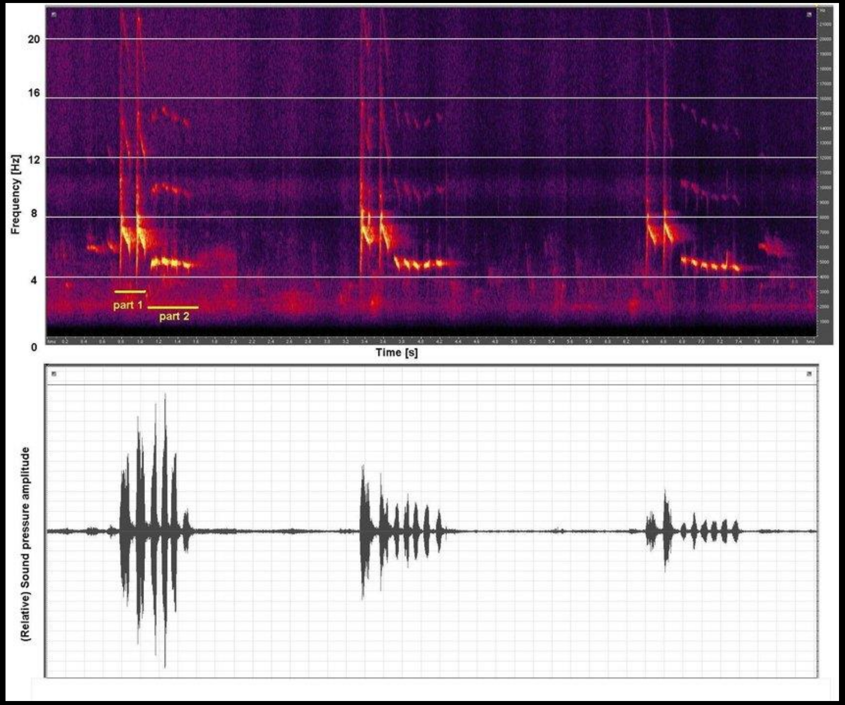
Audio Foundation Model (Singing Voice Model)

Audio is a sequence in “sound languages”.
Audio foundation model is a special case of large “language model”.





Symbolic music sequence



Audio music sequence

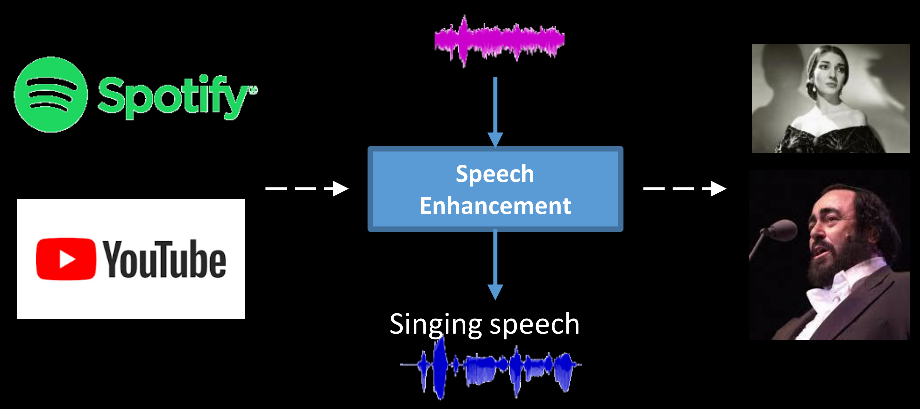
Audio Foundation Model: **Digitalize Human Voice**

Auto-encoder Voice Foundation Model

Humans use existing skills for new tasks

Four models were pretrained using public datasets for the specific tasks

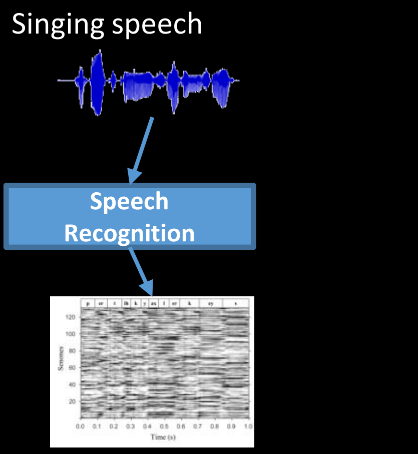
Mixture of singing speech, music and noise



How to Get Data?

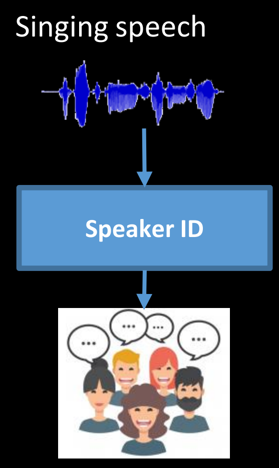
Online sources

Singing voice datasets



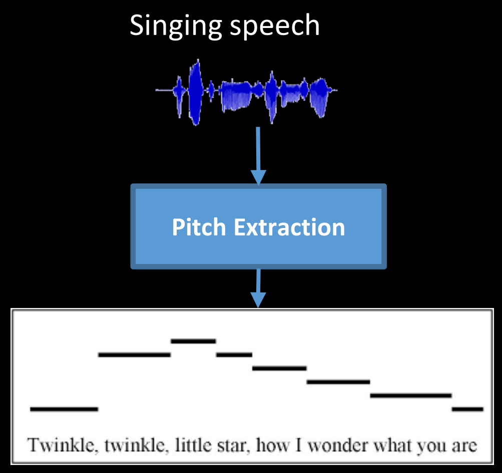
Phonetic distribution (Lyrics information)

What Lyrics (content)?



Identity Embedding

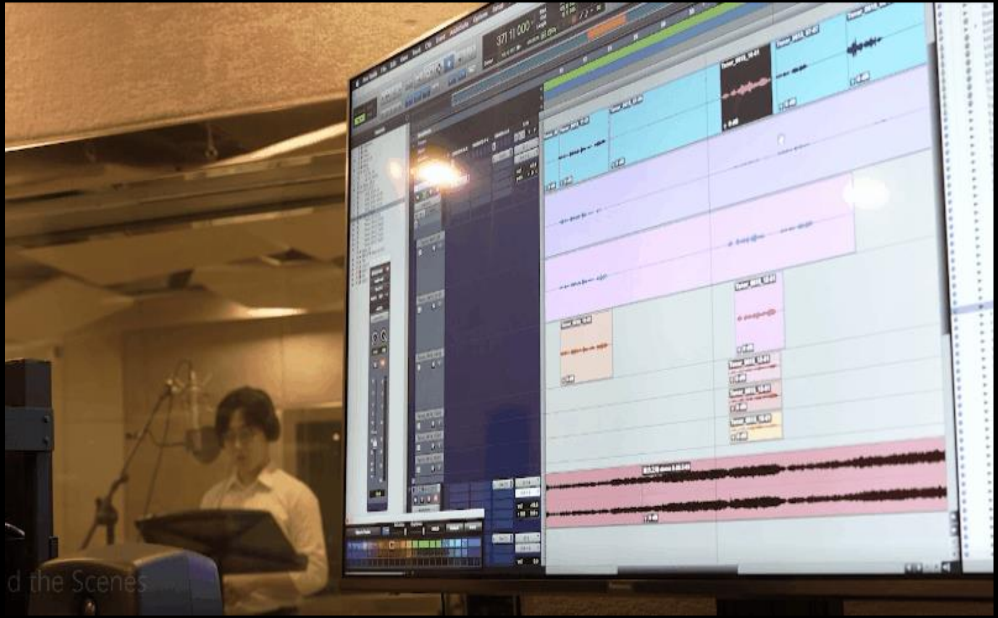
Who Sings?



Pitch Contour

What Melody?

Vocal Illusion : Voice Modelling



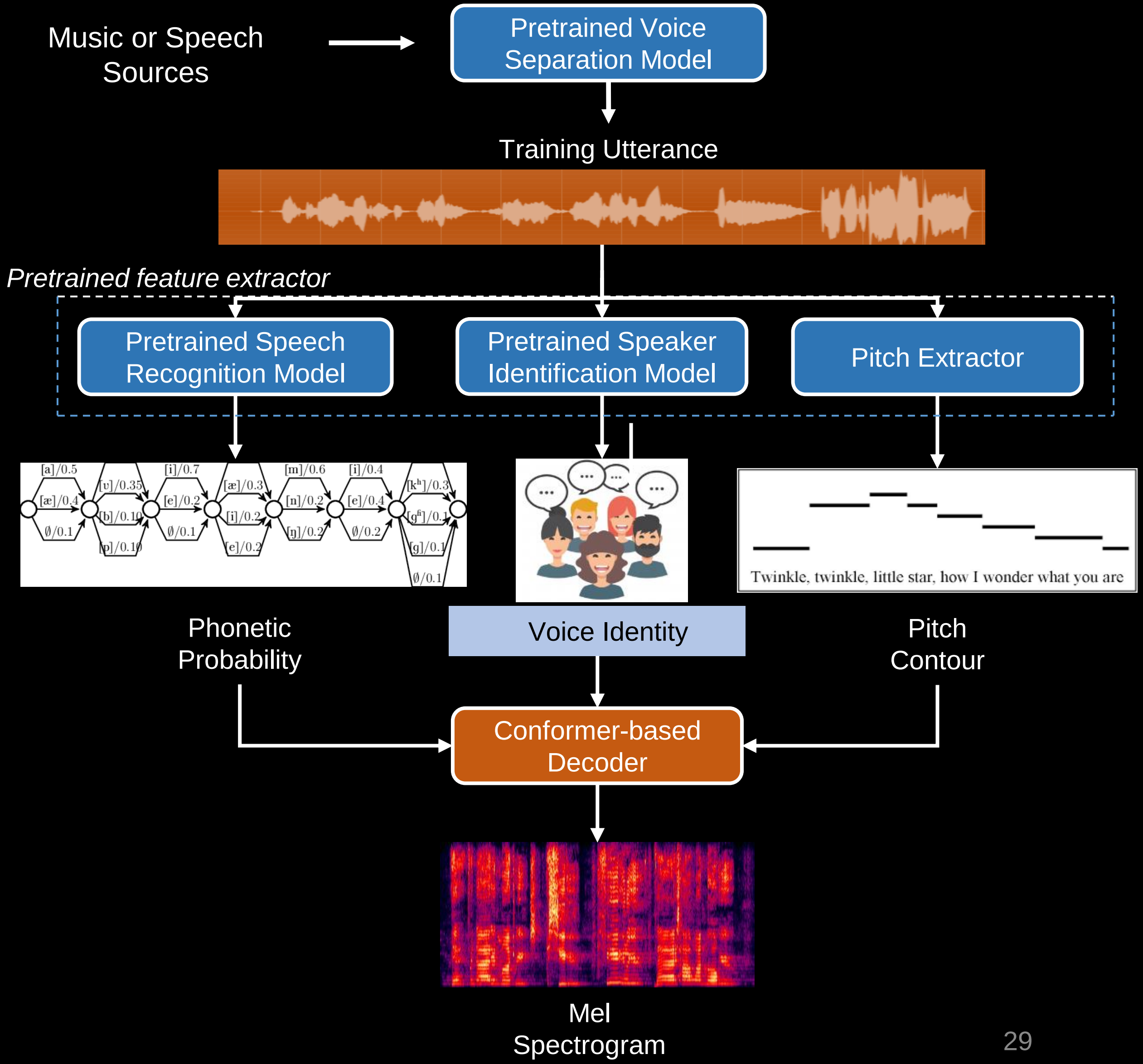
說話→唱歌 & 英文→普通話

Yike講話樣本

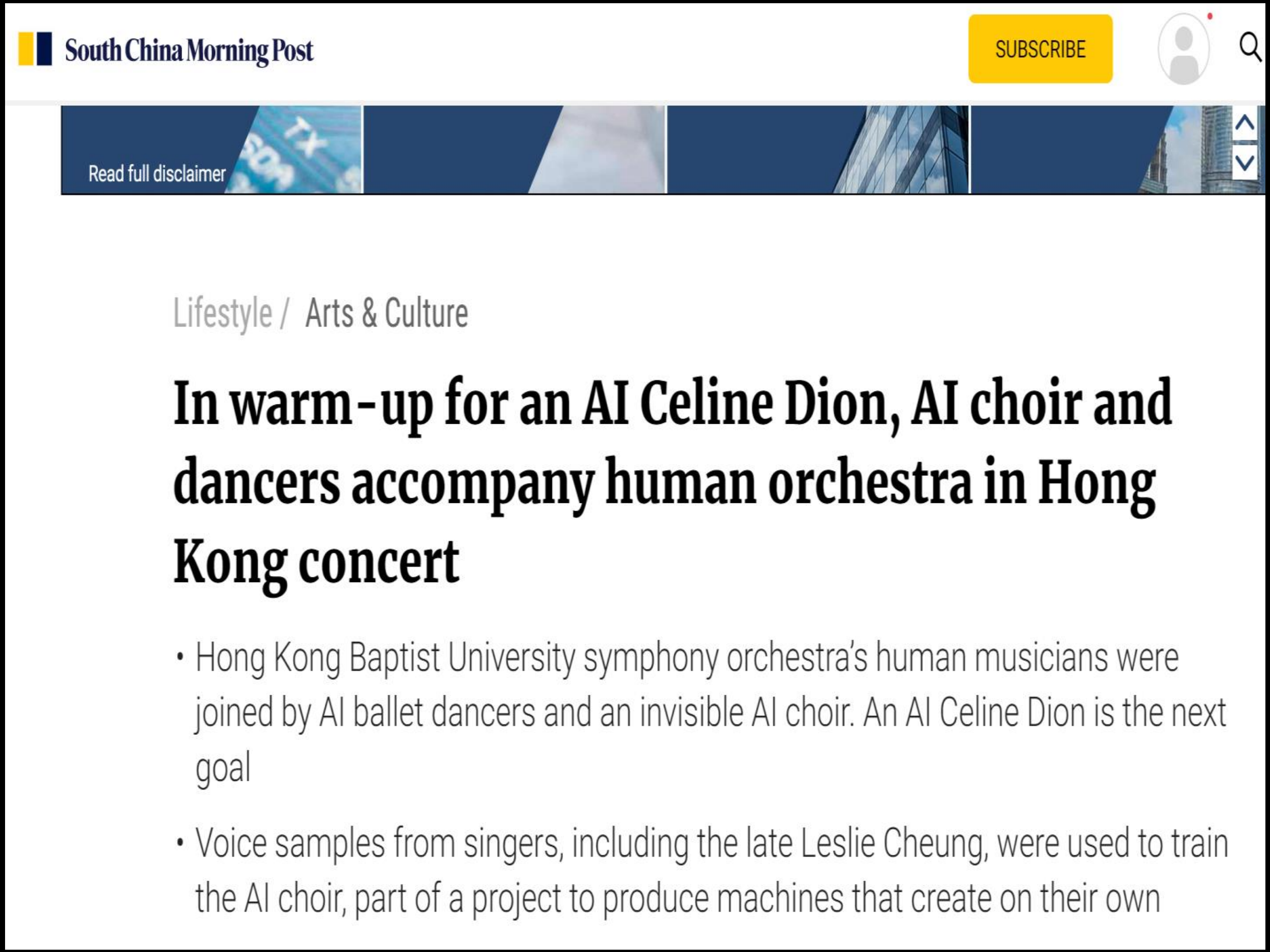
Yike唱歌生成樣本

某男歌手訓練樣本

某男歌手生成樣本


Music or Speech Sources
Pretrained Voice Separation Model
Training Utterance
Pretrained feature extractor
Pretrained Speech Recognition Model
Pretrained Speaker Identification Model
Pitch Extractor
Twinkle, twinkle, little star, how I wonder what you are
Phonetic Probability
Voice Identity
Pitch Contour
Conformer-based Decoder
Mel Spectrogram

Learning the Style of a Singer

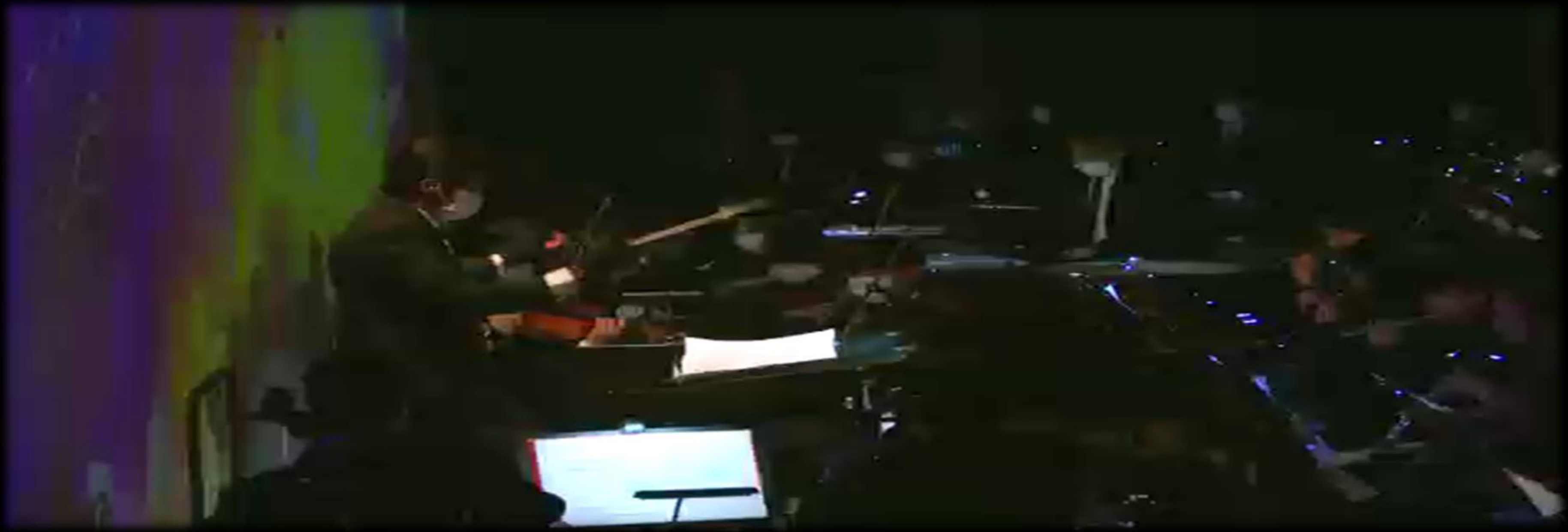


Lifestyle / Arts & Culture

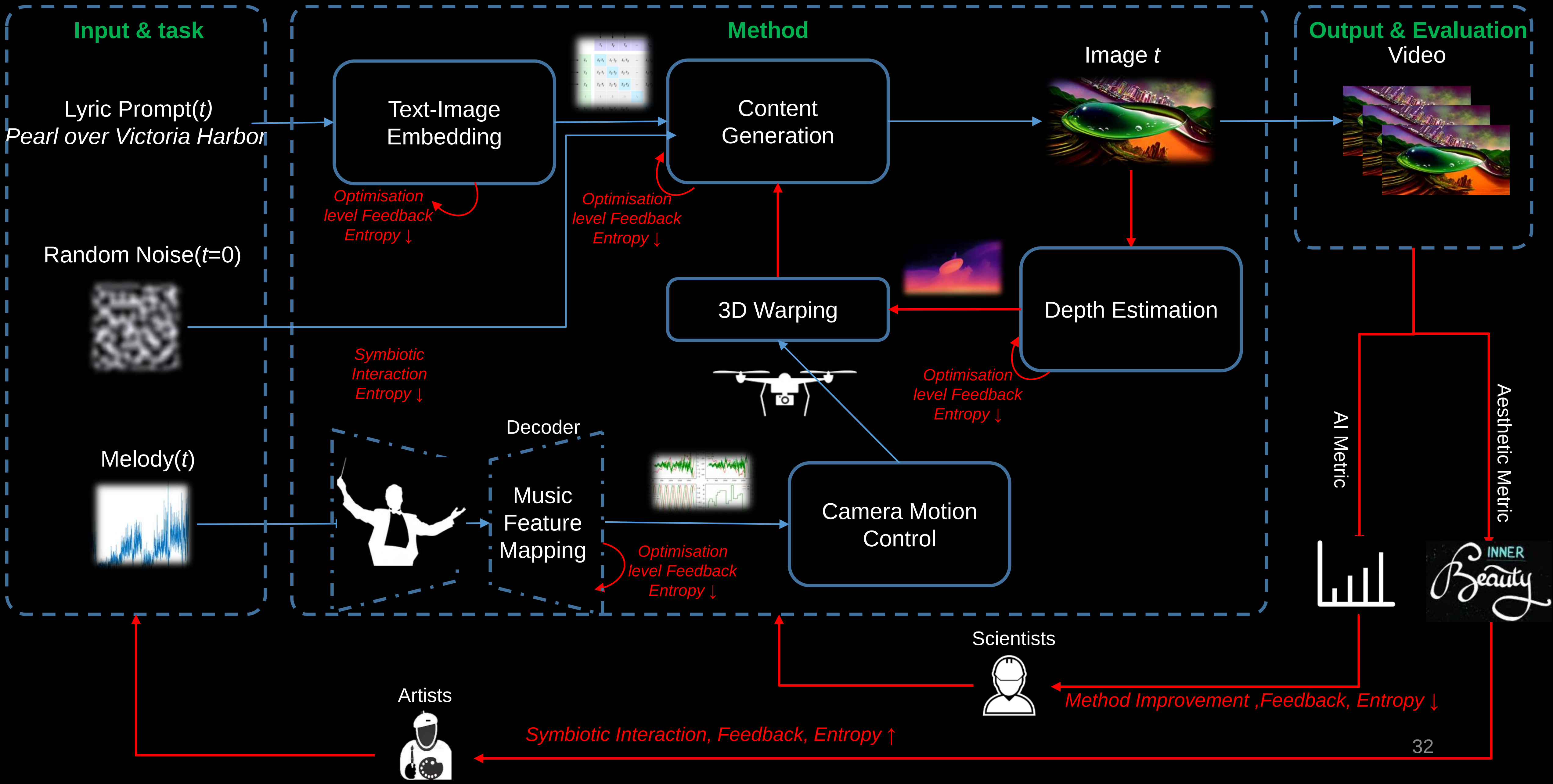
In warm-up for an AI Celine Dion, AI choir and dancers accompany human orchestra in Hong Kong concert

- Hong Kong Baptist University symphony orchestra's human musicians were joined by AI ballet dancers and an invisible AI choir. An AI Celine Dion is the next goal
- Voice samples from singers, including the late Leslie Cheung, were used to train the AI choir, part of a project to produce machines that create on their own

Conducting Machine



AI-based Cross-media Visual Storytelling System



Input & task

Lyric Prompt(*t*)
Pearl over Victoria Harbor

Random Noise(*t*=0)

Melody(*t*)

Method

Text-Image Embedding

Optimisation level Feedback Entropy ↓

Content Generation

Optimisation level Feedback Entropy ↓

Image *t*

Depth Estimation

Optimisation level Feedback Entropy ↓

3D Warping

Symbiotic Interaction Entropy ↓

Decoder

Music Feature Mapping

Optimisation level Feedback Entropy ↓

Camera Motion Control

Output & Evaluation

Video

AI Metric

Aesthetic Metric

INNER Beauty

Scientists

Method Improvement ,Feedback, Entropy ↓

Artists

Symbiotic Interaction, Feedback, Entropy ↑

Understanding : From Lyrics to Themes

MADE WITH GIFOX

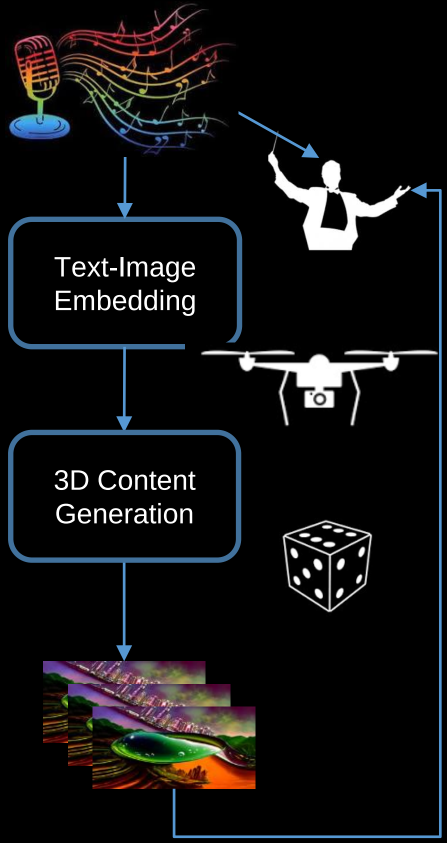
The stream meanders to the south
小河彎彎向南流

MADE WITH GIFOX

The pearl of the east over night
東方之珠整夜未眠

MADE WITH GIFOX

The see wind blows for 5000 years
讓海風吹拂了五千年



there in a garden of roses while moonbeams calmly shine

There the musk roses are sighing

On banks by the Ganges tide

We woo the power of bright dreams to shed their heavenly charms

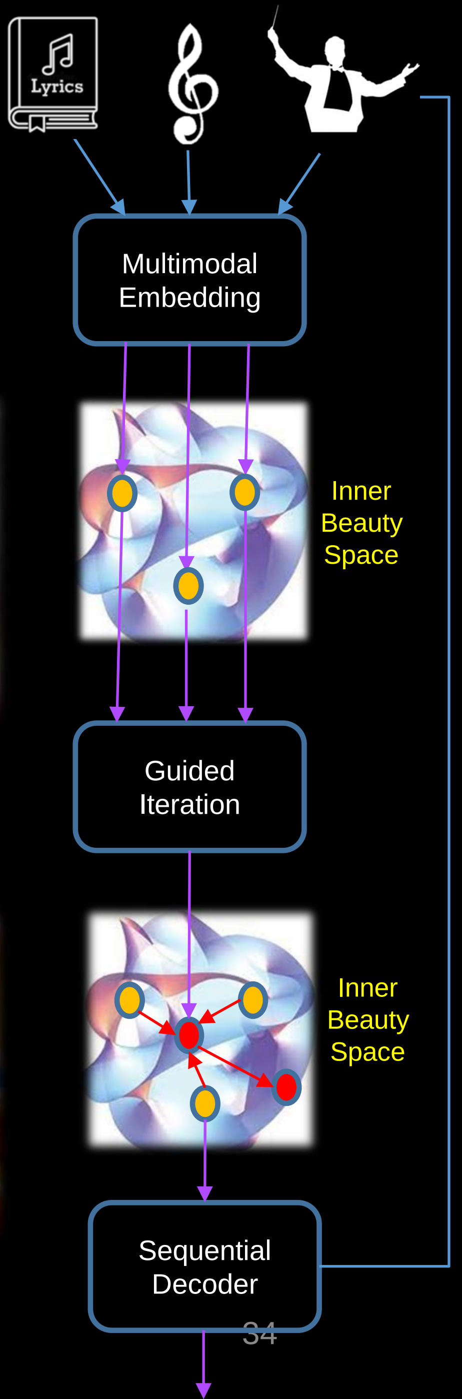
Machine Artist : Creation as Decoder

Cloud ⟶ Flower

Cloud ⟶ Butterfly ⟶ Many butterflies

Cloud ⟶ Cupid wings with love

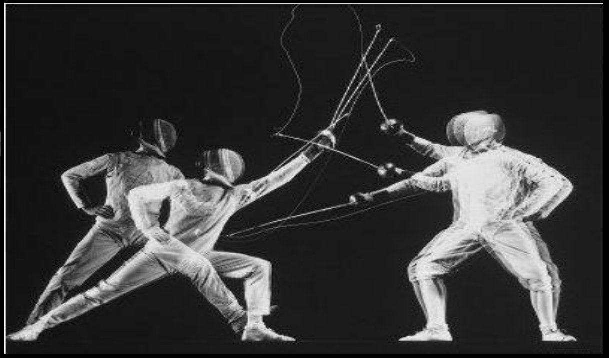
Treble clef ⟶ Flower petals



Motion Foundation Model

Motion Foundation Model aims at generating 3D trajectories that animate the human body in response to various physical world, human, or virtual context.

- Film production, CG (Computer Graphics) character animation
- Performing arts
- Motion science
- Mixed reality, Metaverse applications.

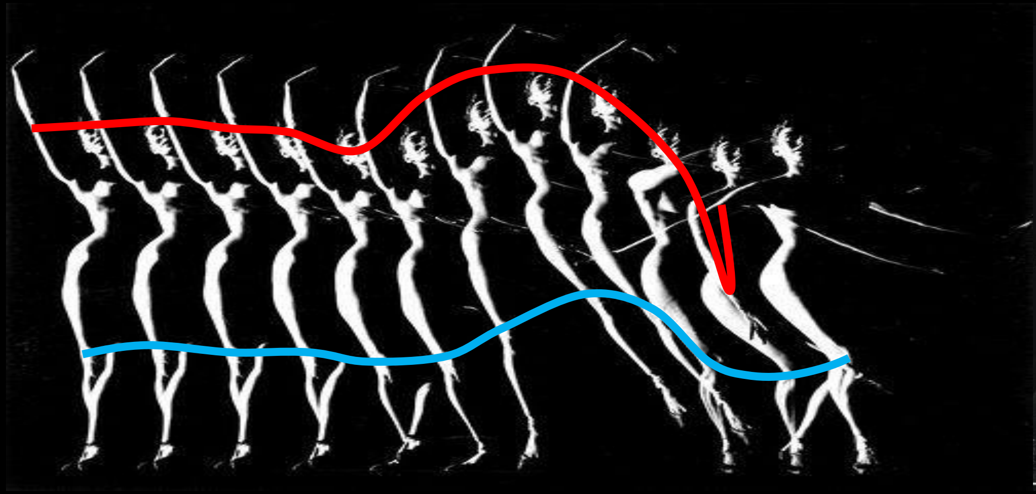


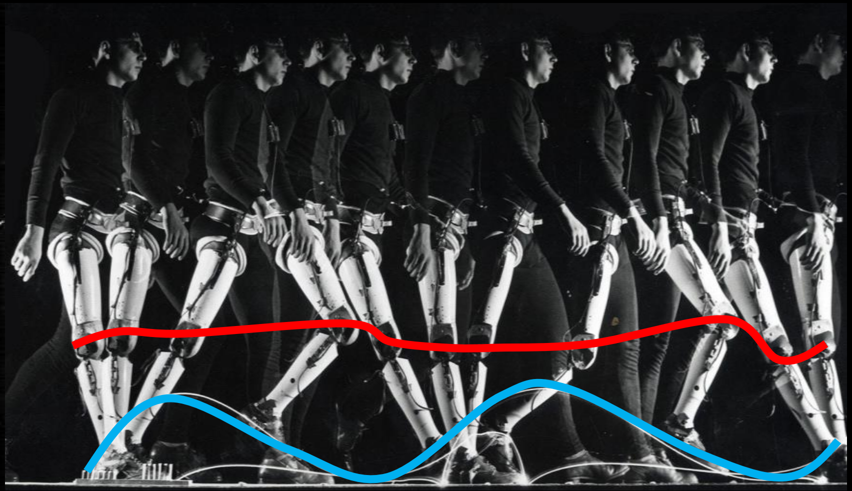
Generated Motion Qualities

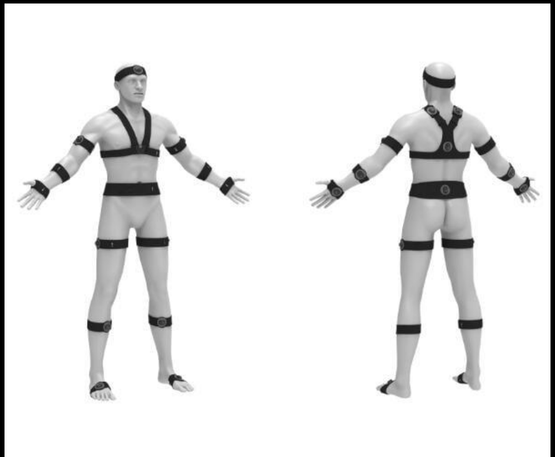
- Obey physical and kinematic laws
- Aesthetic style and messages consistent with artistic context

Motion Foundation Model

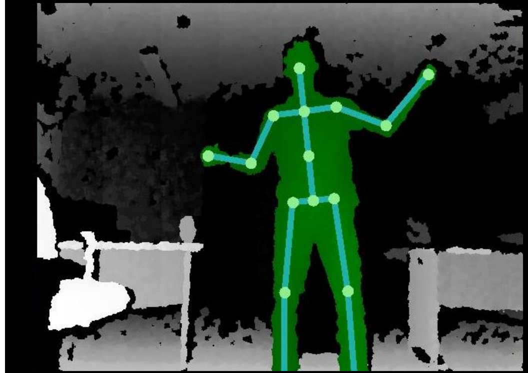
Human motion: spatial and temporal recordings of human skeletal joints captures human activity, aesthetic expression and emotion.



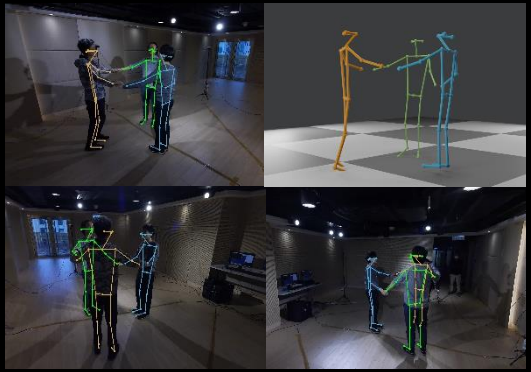




IMU Sensors



Depth Cameras

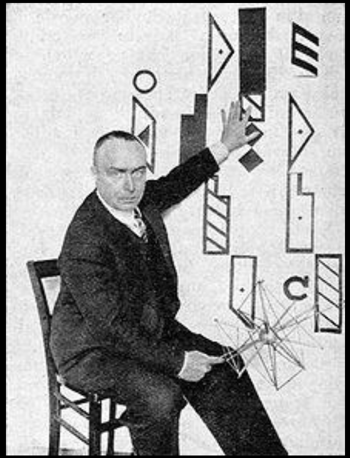


Multi-view stereo RGBs

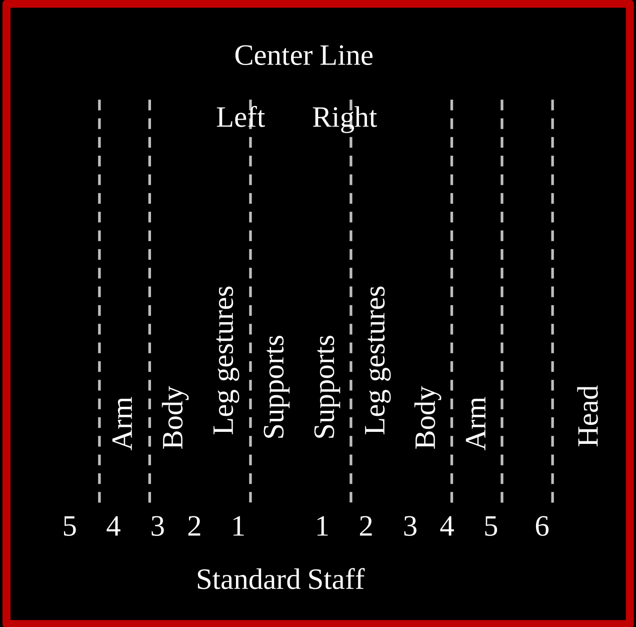
Maker-based ToF

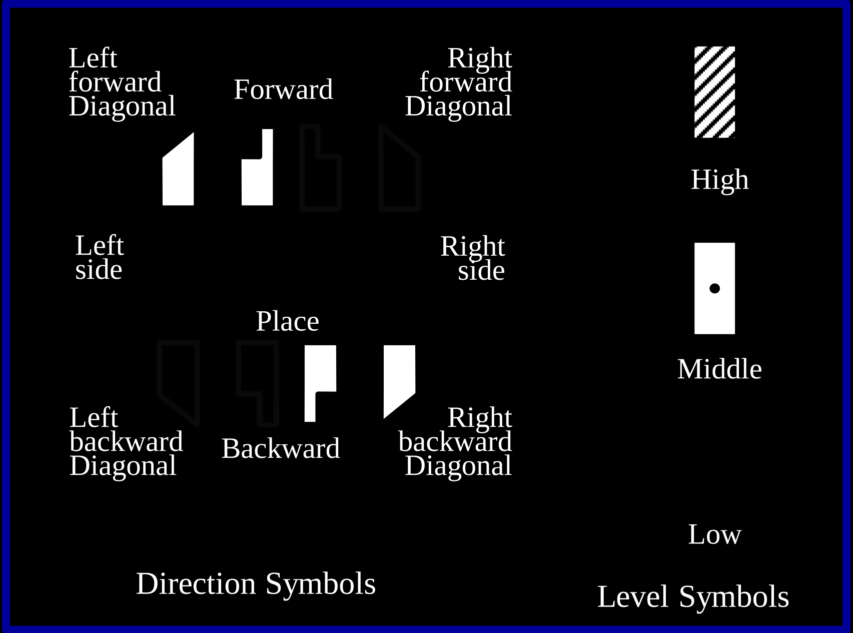
Labanotation – The Language of Motion

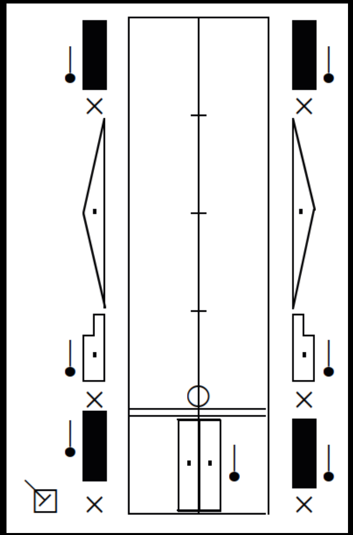
- Labanotation: a structured system for analyzing and recording movement with symbols.
  - Each symbol specifies: direction, level, timing, body part
  - Genre- and Style- independent, able to record every kind of human motion [Guest13]



Rudolf Laban presenting his notation system, circa 1929



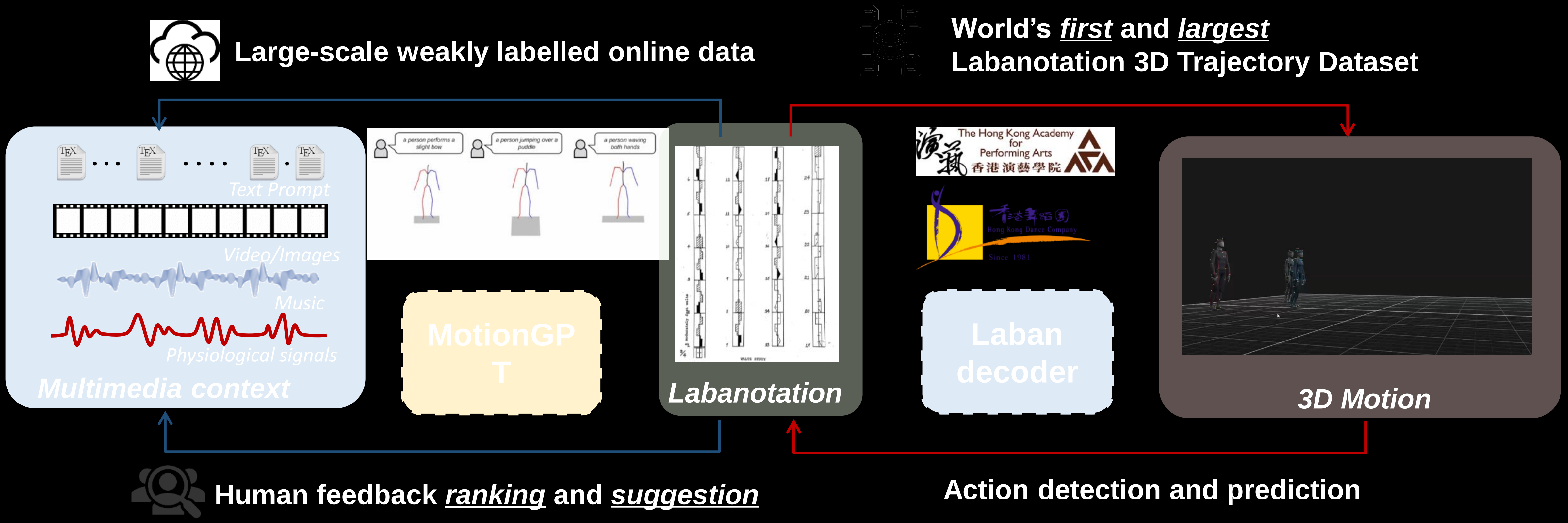




Labanotation example 1:
Ballet arm movement. Arms rest down, reach forward, outward, then back down.

MotionGPT: A Language Model for Chorography

- Enabling Cross-Modality Motion Programming and Interpretation
  - MotionGPT - a large-scale pre-trained language model for *motion concept programming*
  - Laban Decoder*: trajectory sampling* framework based on the music condition, and the Labanotation scripts.

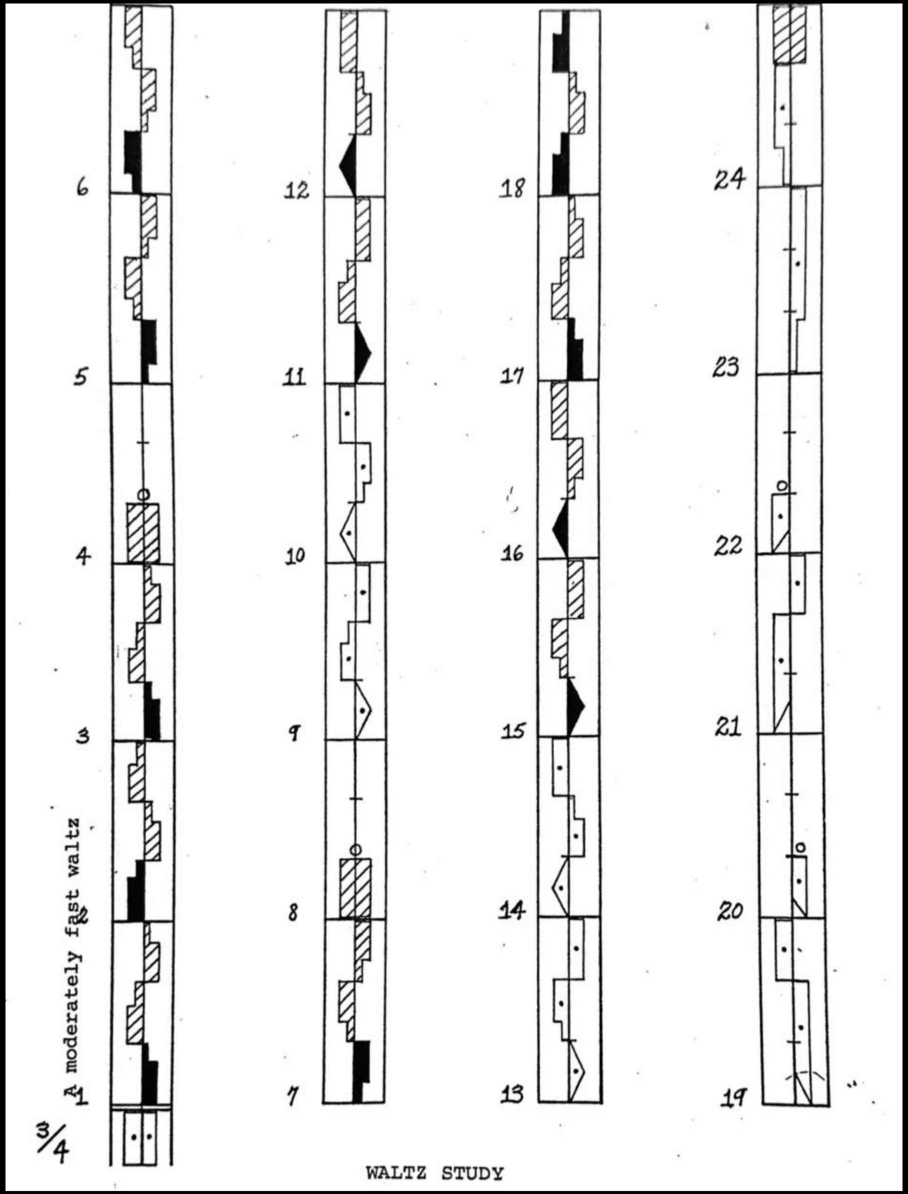

Large-scale weakly labelled online data

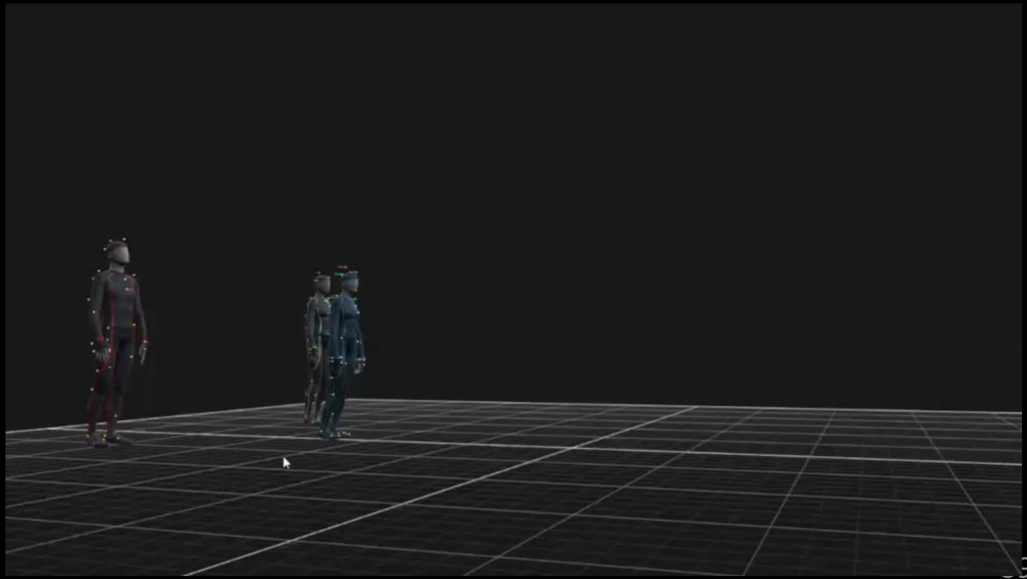
World’s *__first__* and *__largest__*
Labanotation 3D Trajectory Dataset

Human feedback *__ranking__* and *__suggestion__*

Action detection and prediction

Labanotation Capture Sequence





Labanotation Helps to Build a DanceGPT

Compared to direct learning 3D joint directories and existing AI choreography frameworks, ***Labanotation*** is:

Much more **compact** data representation as compared with 30 fps trajectory captures (a 1 min sequence contains 21x30x60= 37,800 data units), which will be too large to be fitted to a transformer model to calculate full self-attentions and study semantic meaning behind the variation

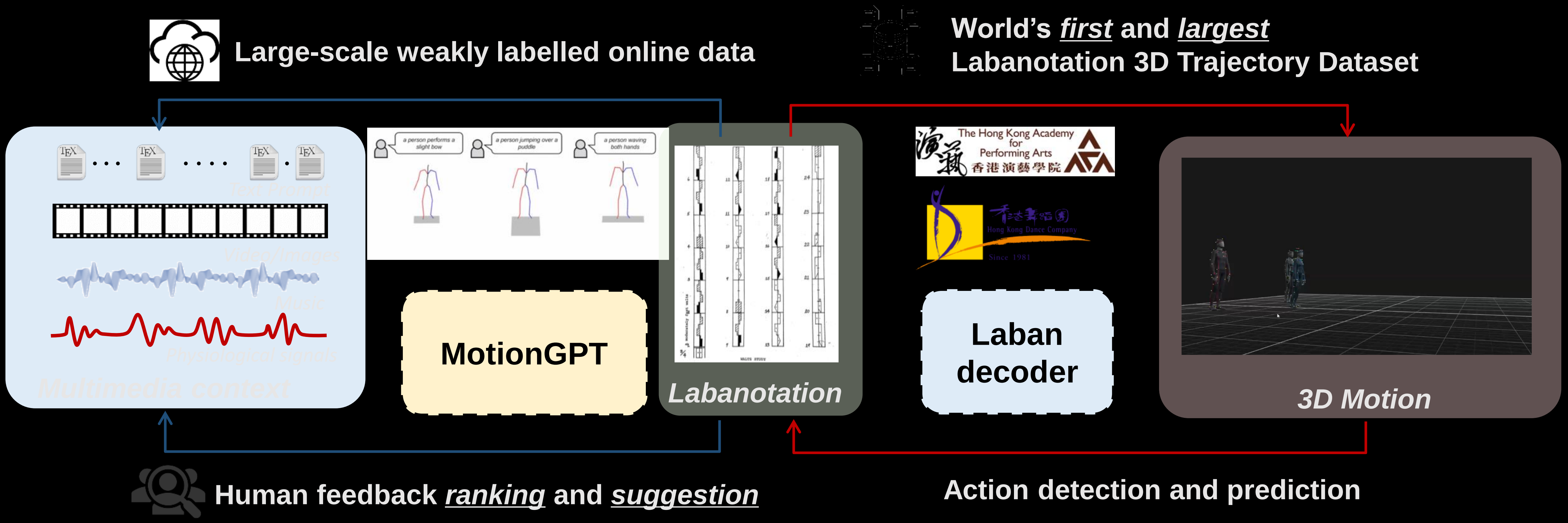
Much more **descriptive** in describing body part actions, and automatically **generalizable** as one symbol could refer to many executions

A much more **semantically meaningful** representation, that is dense, lower-dimensional, which can be easily projected to word embeddings and connect with the embedding space of other pretrained LLMs

Potentials to become a general artificial intelligence like ChatGPT

MotionGPT: A Language Model for Motion

- Enabling Cross-Modality Motion Programming and Interpretation
  - MotionGPT - a large-scale pre-trained language model for *motion concept programming*
  - Laban Decoder*: trajectory sampling* framework based on the music condition, and the Labanotation scripts.



Large-scale weakly labelled online data

World's ***first*** and ***largest***
Labanotation 3D Trajectory Dataset

MotionGPT

Labanotation

Laban decoder

3D Motion

Human feedback ***ranking*** and ***suggestion***

Action detection and prediction

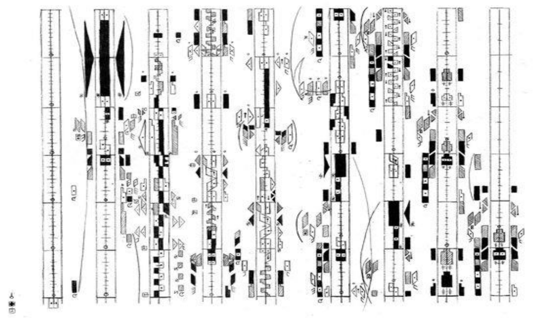
Text Prompt
Video/Images
Music
Physiological signals
Multimedia context
a person performs a slight bow
a person jumping over a puddle
a person waving both hands
The Hong Kong Academy for Performing Arts
香港演藝學院
香港舞蹈團
Hong Kong Dance Company
Since 1981

Motion Semantic Disentanglement and Style Transfer

Symbolic notation ***connects*** and ***disentangles*** kinematic concepts between dance

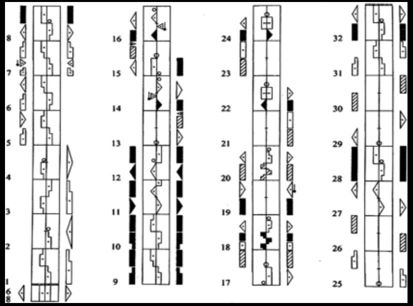
Semantic embedding for ***cross-style/cross-culture*** dance translation

Classic Western Ballet vs. **Traditional Chinese Dance**



Danse des petits cygnes
Swan Lake

Tchaikovsky

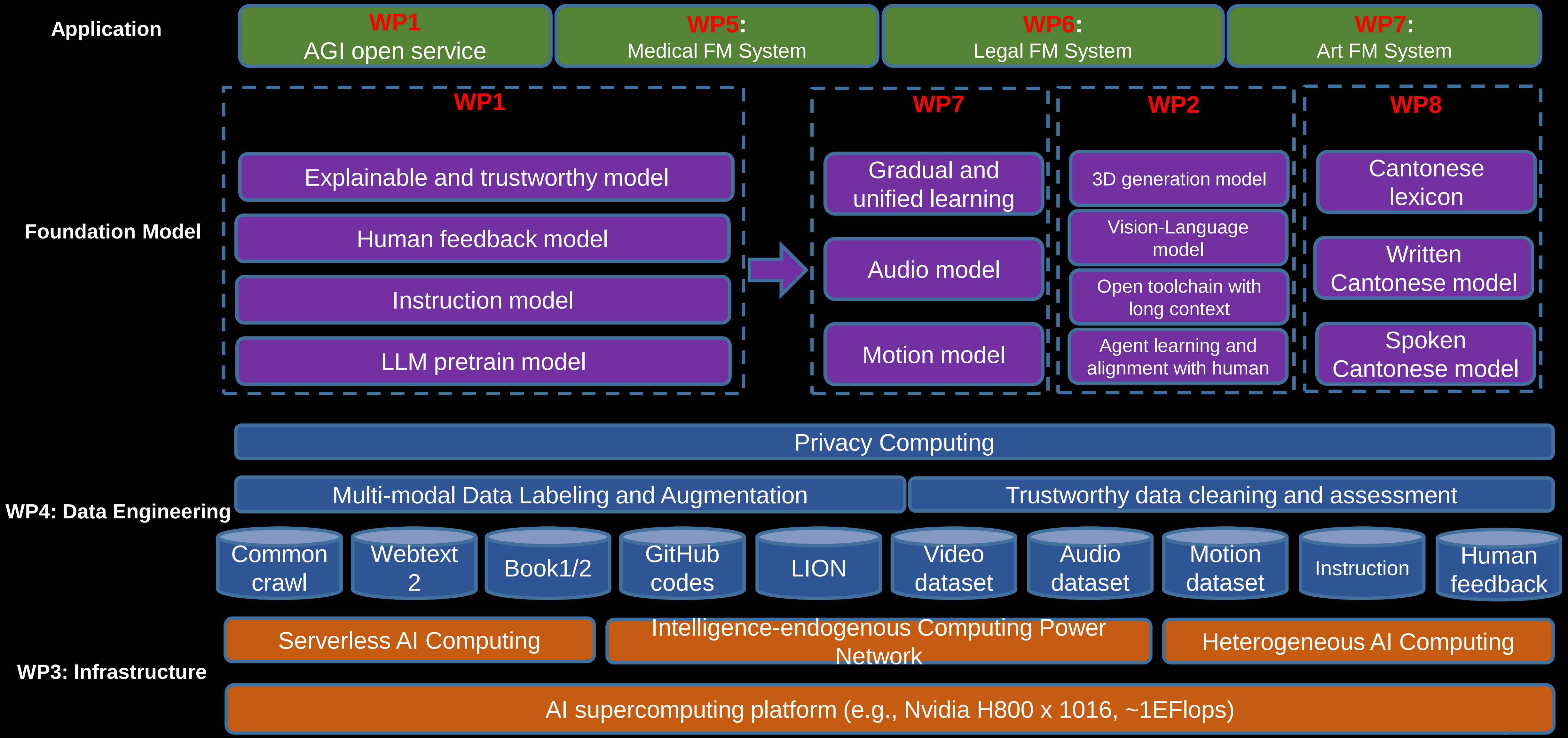


春江花月夜

张若虚（词）
鞠士林（曲）

Human and Machine Symbiotic Dancing

Hong Kong Generative AI R&D Centre



Application

WP1
AGI open service

WP5:
Medical FM System

WP6:
Legal FM System

WP7:
Art FM System

Foundation Model

WP1
- Explainable and trustworthy model
- Human feedback model
- Instruction model
- LLM pretrain model

WP7
- Gradual and unified learning
- Audio model
- Motion model

WP2
- 3D generation model
- Vision-Language model
- Open toolchain with long context
- Agent learning and alignment with human

WP8
- Cantonese lexicon
- Written Cantonese model
- Spoken Cantonese model

WP4: Data Engineering

Privacy Computing

Multi-modal Data Labeling and Augmentation

Trustworthy data cleaning and assessment

Common crawl | Webtext 2 | Book1/2 | GitHub codes | LION | Video dataset | Audio dataset | Motion dataset | Instruction | Human feedback

WP3: Infrastructure

Serverless AI Computing

Intelligence-endogenous Computing Power Network

Heterogeneous AI Computing

AI supercomputing platform (e.g., Nvidia H800 x 1016, ~1EFlops)

TURING AI ORCHESTRA

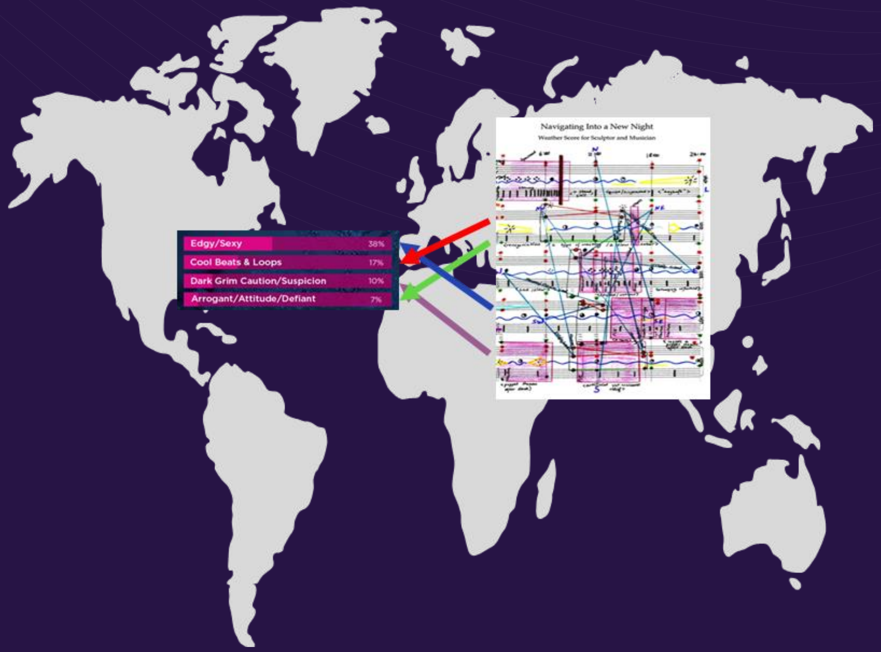
Turing AI Orchestra : A Platform of Digital Art

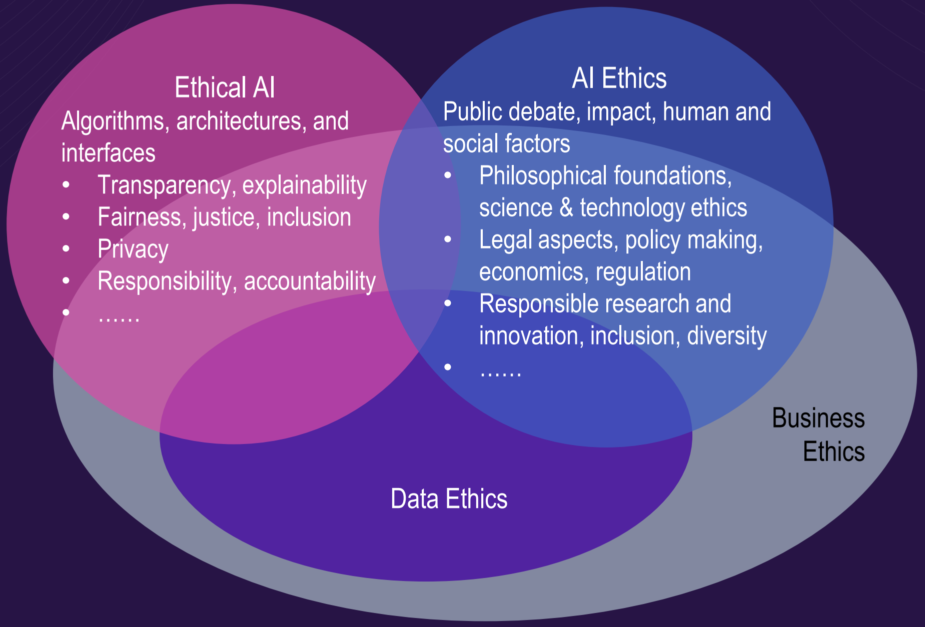
- World’s 1st Human-Machine Symbiotic Orchestra
- World’s 1st DAO for Collaborative AI-based Art Creation
- World’s 1st Platform for Human-in-the-loop Art Tech Research

A Blockchain-Based Global Art Marketplace

A Curated Art Marketplace with a global artist network

Curated marketplace means that the platform determines which NFTs are allowed to be minted, posted and sold on directly its marketplace. Compared to an open marketplace, a curated marketplace is more limited and exclusive, requiring artists to apply and be accepted before being able to mint or sell NFTs in an attempt to keep fraud down and quality high.


Edgy/Sexy 38%
Cool Beats & Loops 17%
Dark Grim Caution/Suspicion 10%
Arrogant/Attitude/Defiant 7%
Navigating Into a New Night
Weather Score for Sculptor and Musician



Ethical AI

Algorithms, architectures, and interfaces

- Transparency, explainability
- Fairness, justice, inclusion
- Privacy
- Responsibility, accountability
- ……

AI Ethics

Public debate, impact, human and social factors

- Philosophical foundations, science & technology ethics
- Legal aspects, policy making, economics, regulation
- Responsible research and innovation, inclusion, diversity
- ……

Business Ethics

Data Ethics

A NFT for the Conference: On Wings of Song

O Holy Night

AI創作音畫和舞蹈互動

O HOLY NIGHT

圖靈人工智能交響樂團