上海大学未来技术学院 | 上海大学人工智能研究院
SCHOOL OF FUTURE TECHNOLOGY, SHANGHAI UNIVERSITY | INSTITUTE OF ARTIFICIAL INTELLIGENCE, SHANGHAI UNIVERSITY

人工智能导论

——第4课：生成式AI

叶林奇

未来技术学院（人工智能研究院）

2024秋季学期

提纲

一、生成式模型

二、角色动画生成

三、机器人动作生成

菜鸟

咏卷心菜小鸡

卷心菜里藏玄机，
神鸡孕育其中栖。
眼似星辰闪橙光，
嘴如朝阳染霞霓。
天赐灵物人间现，
地育仙禽世间稀。
莫道此物无用处，
寻常百姓盘中餐。

叶随风舞舞翩翩，
林间鸟语声声甜。
奇峰异石入眼帘，
人间仙境在心田。

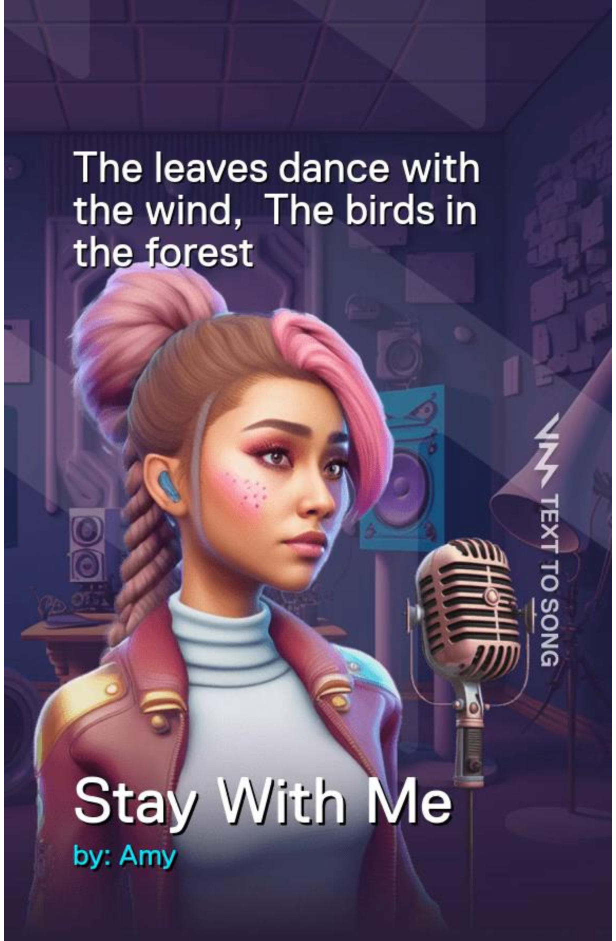
- 普通话 🔊
- 广东话 🔊
- 东北话 🔊
- 山东话 🔊
- 香港话 🔊
- 上海话 🔊
- 河南话 🔊
- 陕西话 🔊
- 四川话 🔊
- 台湾话 🔊

https://www.text-to-speech.cn/

https://png3d.com/

https://pika.art/

The leaves are dancing with the wind.



The leaves dance with the wind, The birds in the forest

https://tuna.voicemod.net/text-to-song

俯首甘为孺子牛

横眉冷对千夫指

鲁迅先生

大家好我是沐沐

所以让我们可以这样

ARAG
ARAG – Partner des Sports
ARAG
www.sports
You Tube
SOLOMONJAGWE
BUTTERFLY
IEBHERR
ARAG

https://wonderdynamics.com/

TEXT-TO-VIDEO

PIKA

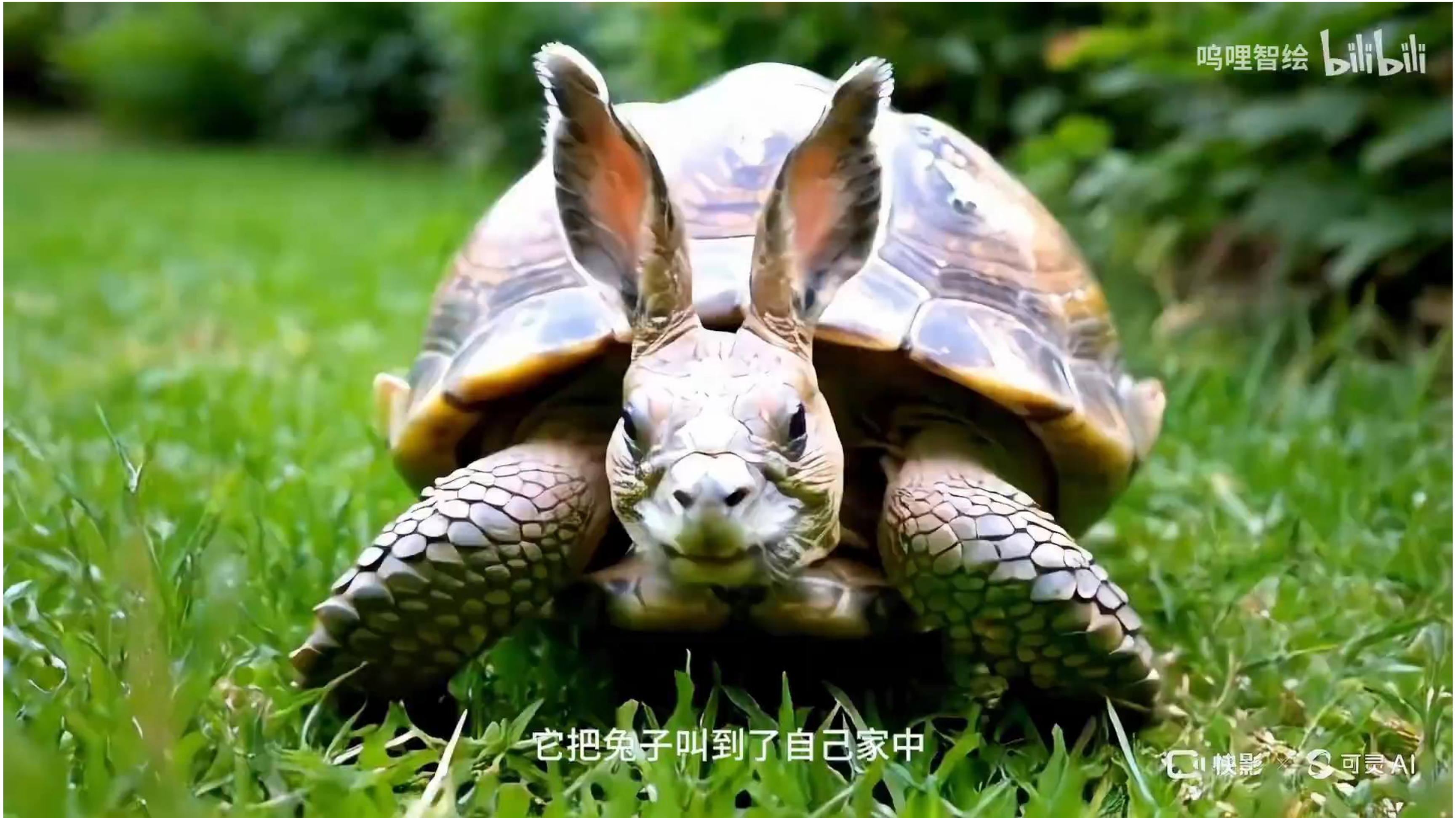
https://pika.art/

https://research.runwayml.com/gen2

上海大学首部AIGC宣传片：《智绘上大·未来可期》

可灵AI x 快影「灵感迸发」创作大赛【一等奖】作品

可灵AI 视频生成



它把兔子叫到了自己家中

可灵AI 视频生成 《夜莺与玫瑰》

文心一言

https://yiyan.baidu.com/

通义千问

https://tongyi.aliyun.com/qianwen/

讯飞星火

https://xinghuo.xfyun.cn/

Kimi

https://kimi.moonshot.cn/

天工

https://www.tiangong.cn/

豆包

https://www.doubao.com/

智谱

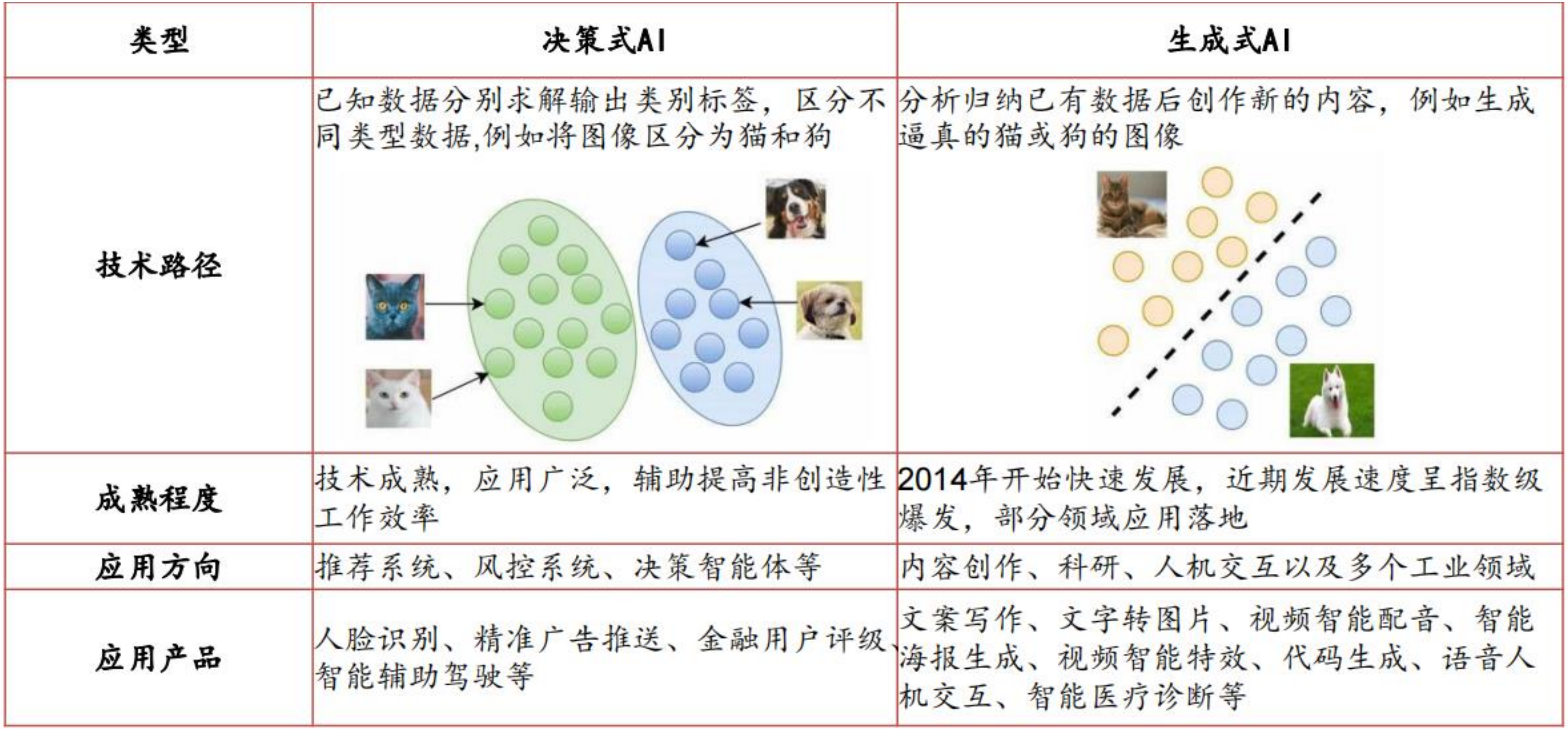
https://chatglm.cn/

可灵

https://klingai.kuaishou.com/

生成式AI

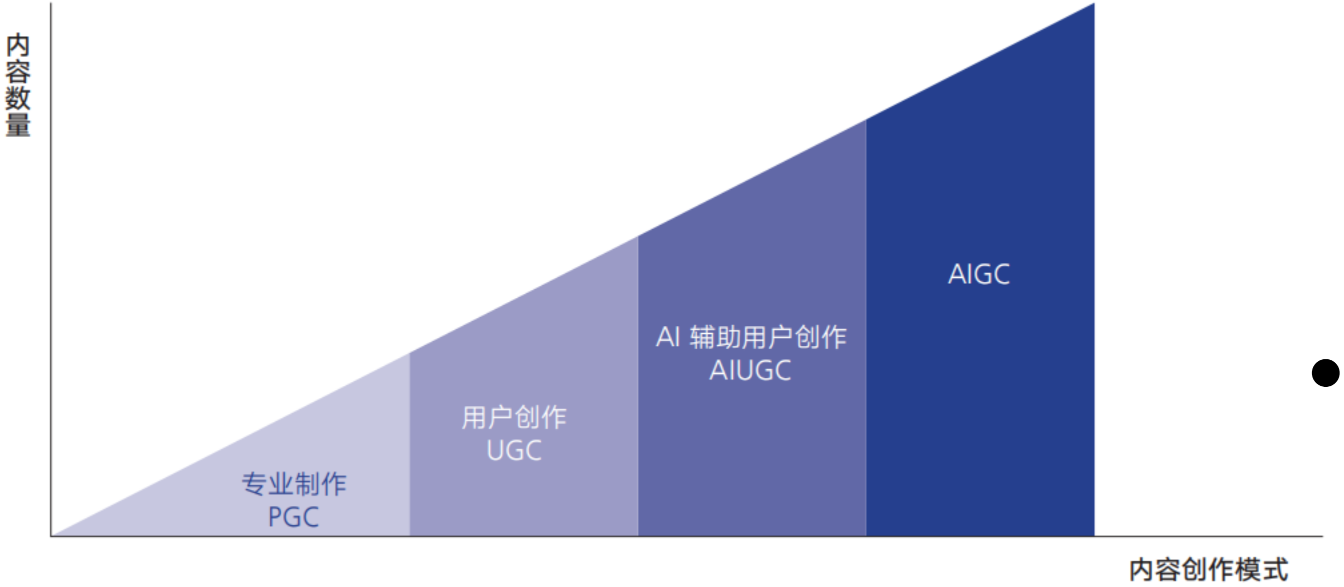
AI模型可大致分为决策式AI（Discriminant AI）和生成式AI （Generative AI）两类。



类型	决策式AI	生成式AI
技术路径	已知数据分别求解输出类别标签，区分不同类型数据,例如将图像区分为猫和狗	分析归纳已有数据后创作新的内容，例如生成逼真的猫或狗的图像
成熟程度	技术成熟，应用广泛，辅助提高非创造性工作效率	2014年开始快速发展，近期发展速度呈指数级爆发，部分领域应用落地
应用方向	推荐系统、风控系统、决策智能体等	内容创作、科研、人机交互以及多个工业领域
应用产品	人脸识别、精准广告推送、金融用户评级、智能辅助驾驶等	文案写作、文字转图片、视频智能配音、智能海报生成、视频智能特效、代码生成、语音人机交互、智能医疗诊断等

生成式AI

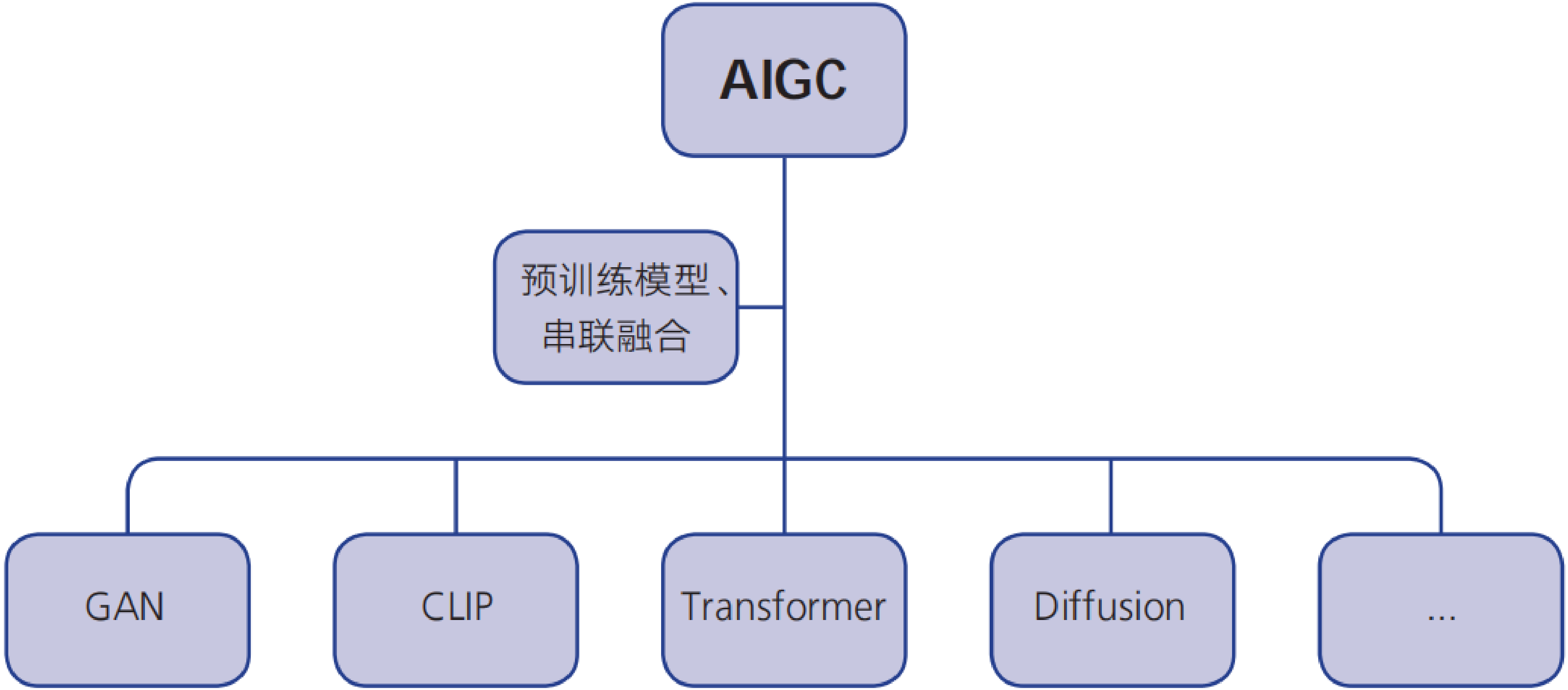
AIGC（AI-Generated Content，人工智能生成内容）



图：内容创作模式的四个发展阶段

- 1957 年莱杰伦·希勒(Leiaren Hiller)和伦纳德·艾萨克森(Leonard Isaacson)完成了人类历史上第一支由计算机创作的音乐作品就可以看作是AIGC 的开端。
- 2022 年才真正算是 AIGC 的爆发之年，人们看到了AIGC 无限的创造潜力和未来应用可能性。

生成式AI

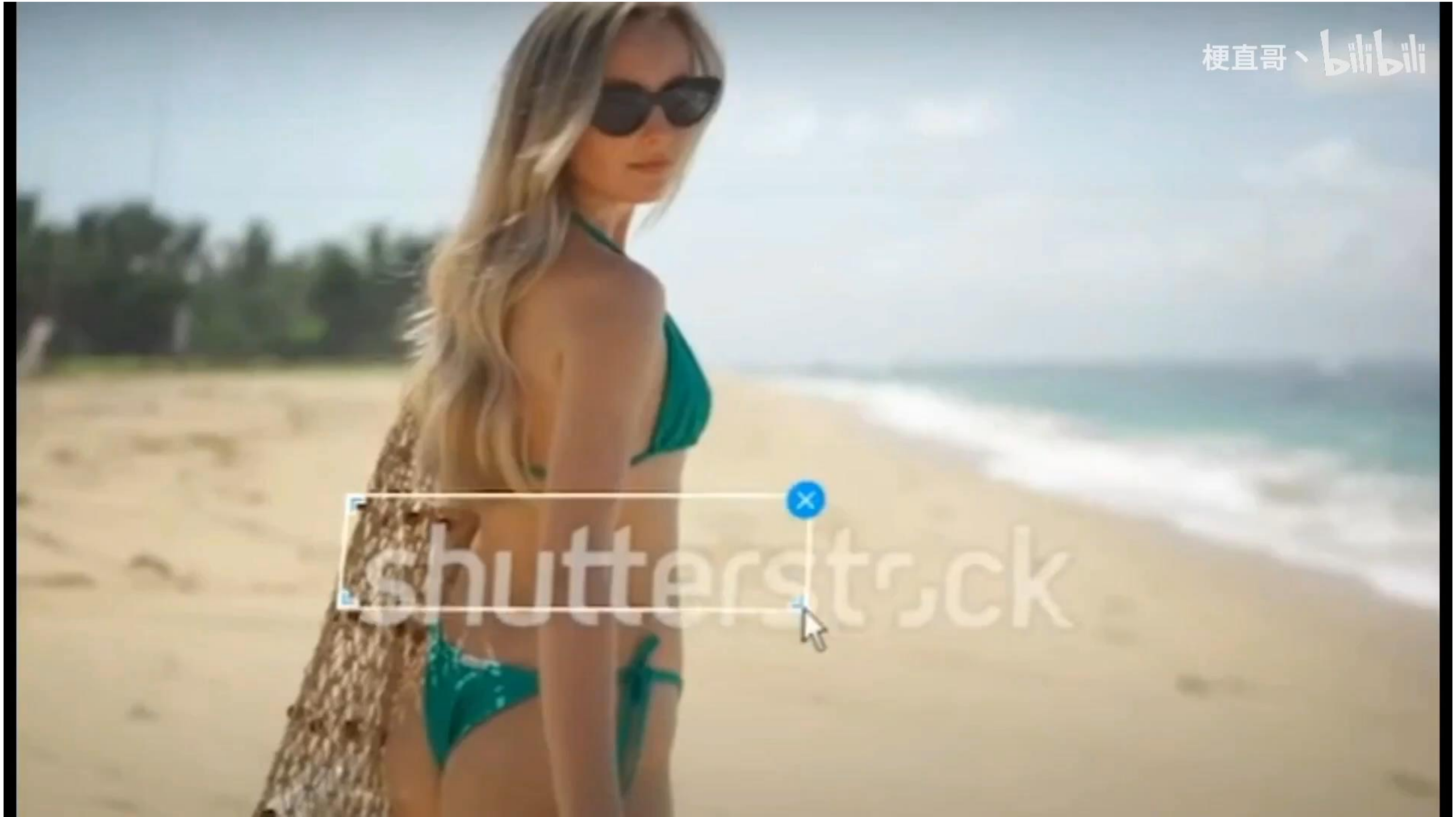


图：AIGC 技术累积融合

生成式AI

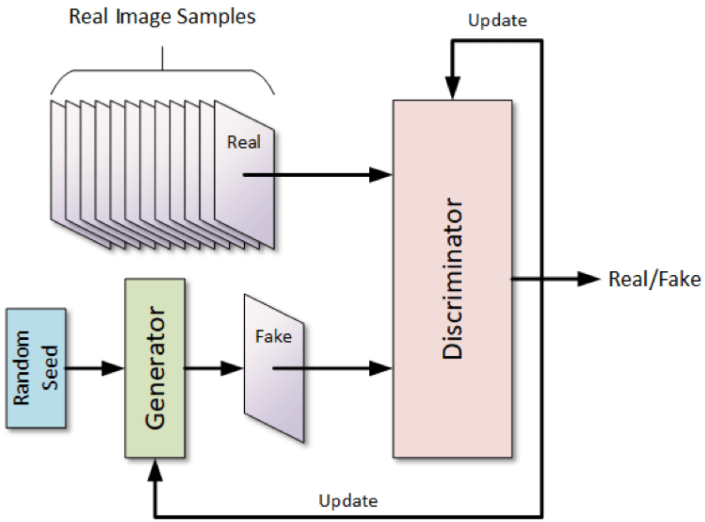
模型	提出时间	模型描述
变分自动编码（Variational Autoencoders，VAE）	2014年	基于变分下界约束得到的Encoder-Decoder模型对
生成对抗网络（GAN）	2014年	基于对抗的Generator-Discriminator模型对
基于流的生成模型（Flow-based models）	2015年	学习一个非线性双射转换（bijective transformation），其将训练数据映射到另一个空间，在该空间上分布是可以因子化的，整个模型架构依靠直接最大化log-likelihood来完成

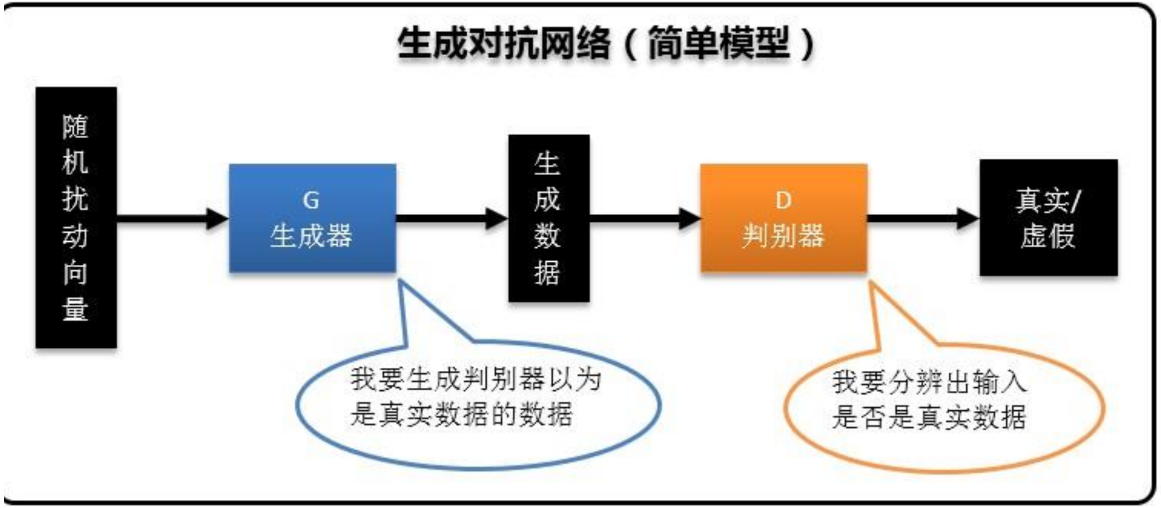
VAE （变分自编码器）原理



GAN（生成对抗网络）原理

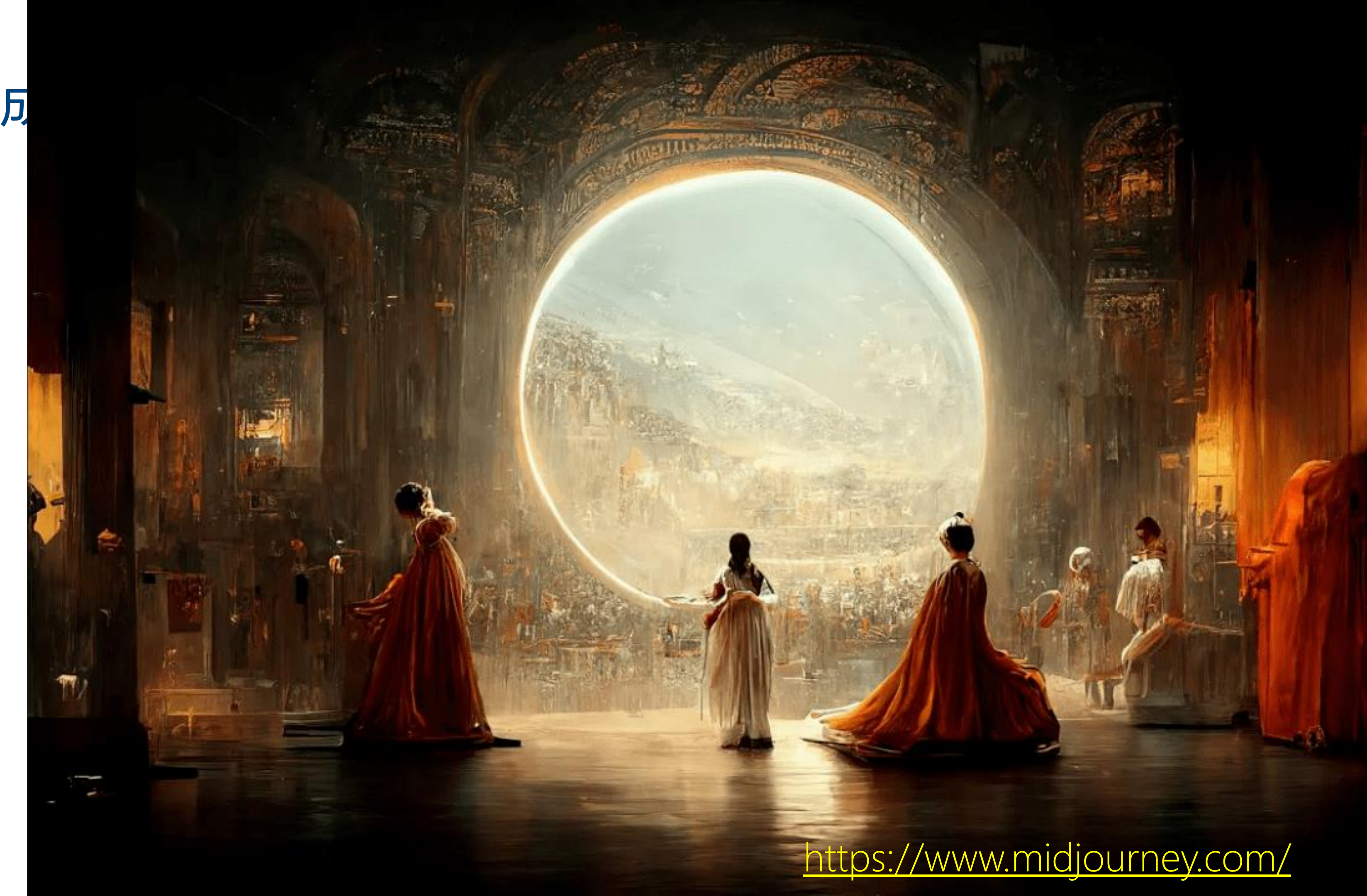
2014 年，伊恩·古德费洛(Ian Goodfellow)提出的生成对抗网络(Generative Adversarial Network，GAN) 成为早期最为著名的生成模型。GAN 使用合作的零和博弈框架来学习，被广泛用于生成图像、视频、语音和三维物体模型等。GAN 也产生了许多流行的架构或变种，如DCGAN，StyleGAN，BigGAN，StackGANPix2pix，Age-cGAN，CycleGAN、对抗自编码器(Adversarial Autoencoders，AAE)、对抗推断学习 (Adversarially Learned Inference，ALI)等





生成

美国科罗拉多州上月举办艺术博览会，一幅名为《太空歌剧院》的画作最终获得数字艺术类别冠军。该作品先由AI制图工具Midjourney生成，再经Photoshop润色而來。

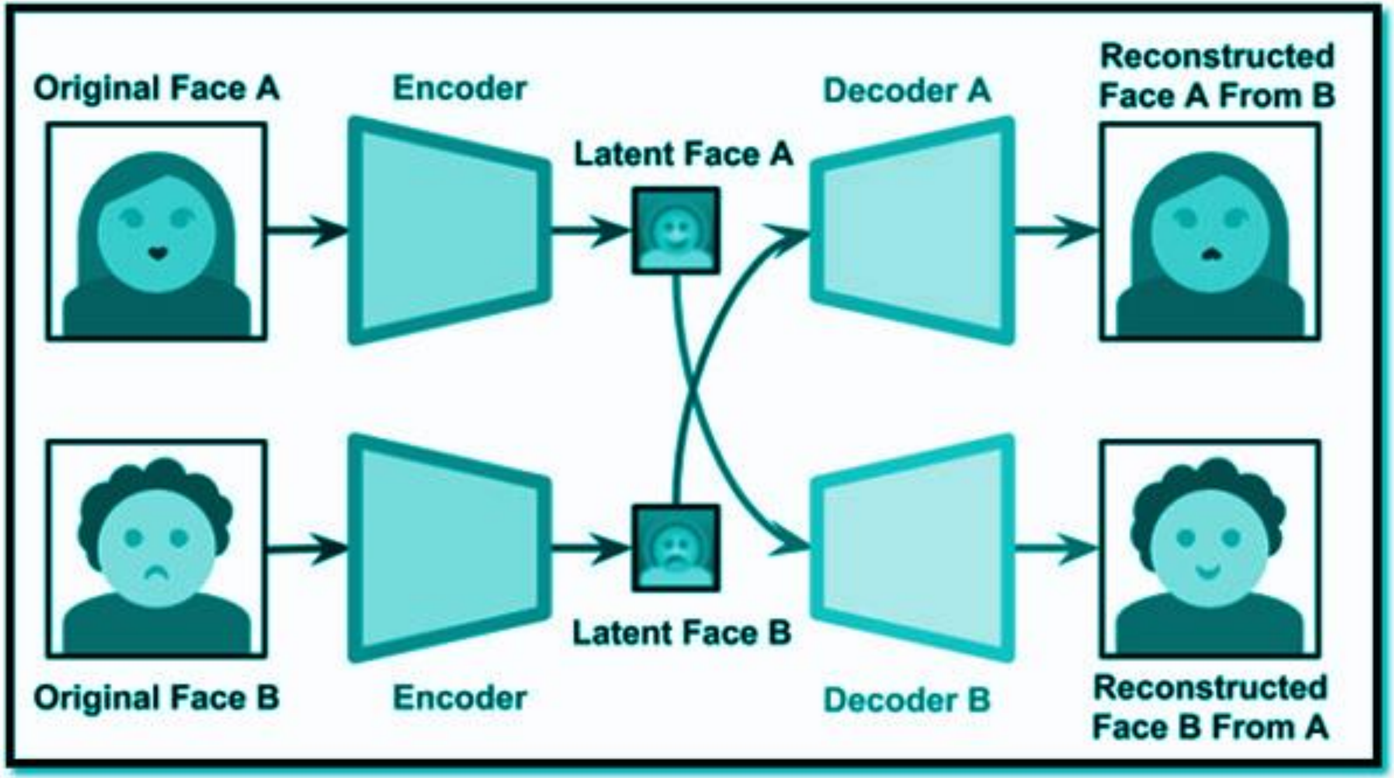


https://www.midjourney.com/

https://www.midjourney.com/

生成式AI

GAN网络应用：AI换脸 是指用另一个人脸来替换一张图片或视频中的一个人脸，合成新的媒体物，它是Deepfake技术最广为人知的一种应用形式。

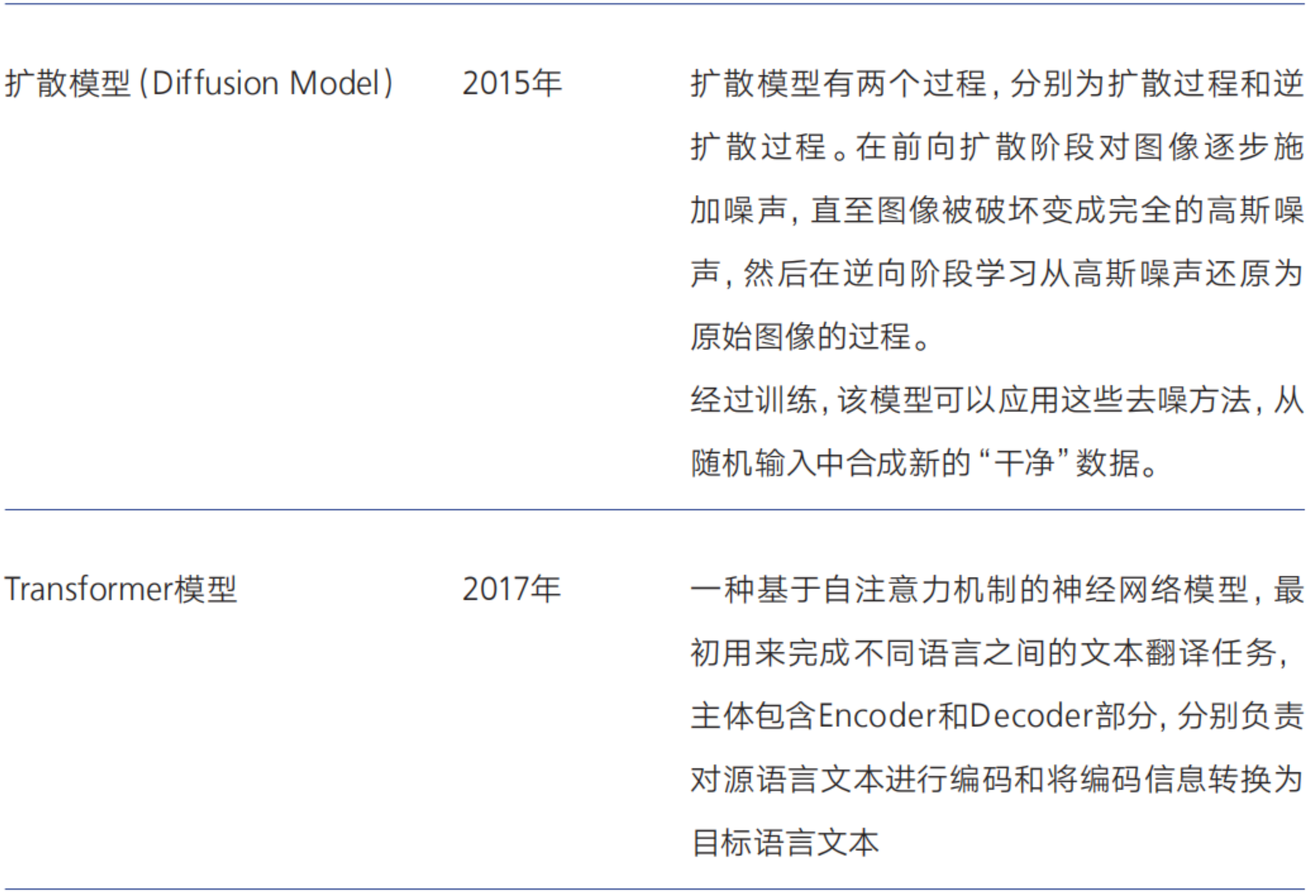


TITANIC

生成式AI

Transformer、基于流的生成模型（Flow-based models)、扩散模型(Diffusion Model)等深度学习的生成算法相继涌现。

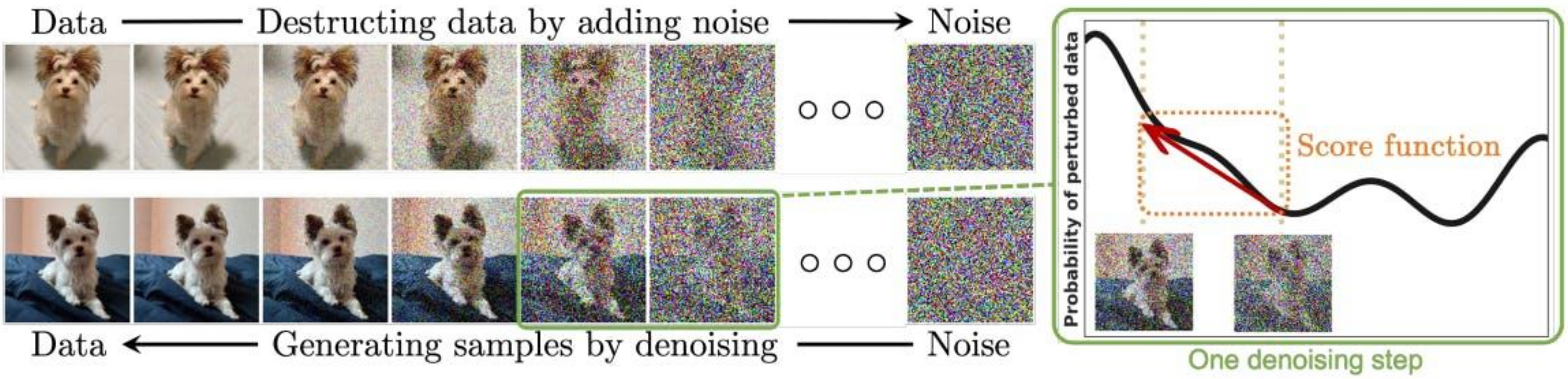
从最优化模型性能的角度出发，扩散模型相对GAN来说具有更加灵活的模型架构和精确的对数似然计算，已经取代GAN成为最先进的图像生成器。2021年6月，OpenAI发表论文已经明确了这个结论和发展趋势。



扩散模型（Diffusion Model）	2015年	扩散模型有两个过程，分别为扩散过程和逆扩散过程。在前向扩散阶段对图像逐步施加噪声，直至图像被破坏变成完全的高斯噪声，然后在逆向阶段学习从高斯噪声还原为原始图像的过程。 经过训练，该模型可以应用这些去噪方法，从随机输入中合成新的“干净”数据。
Transformer模型	2017年	一种基于自注意力机制的神经网络模型，最初用来完成不同语言之间的文本翻译任务，主体包含Encoder和Decoder部分，分别负责对源语言文本进行编码和将编码信息转换为目标语言文本

Difussion Model （扩散模型）原理

扩散模型(Diffusion Model)是受非平衡热力学的启发，定义一个扩散步骤的马尔可夫链，逐渐向数据添加随机噪声，然后学习逆扩散过程，从噪声中构建所需的数据样本。扩散模型最初设计用于去除图像中的噪声。随着降噪系统的训练时间越来越长并且越来越好，它们最终可以从纯噪声作为唯一输入生成逼真的图片。


Data
Destructing data by adding noise
Noise
Data
Generating samples by denoising
Noise
Probability of perturbed data
Score function
One denoising step

https://stablediffusionweb.com/

提纲

一、生成式模型

二、角色动画生成

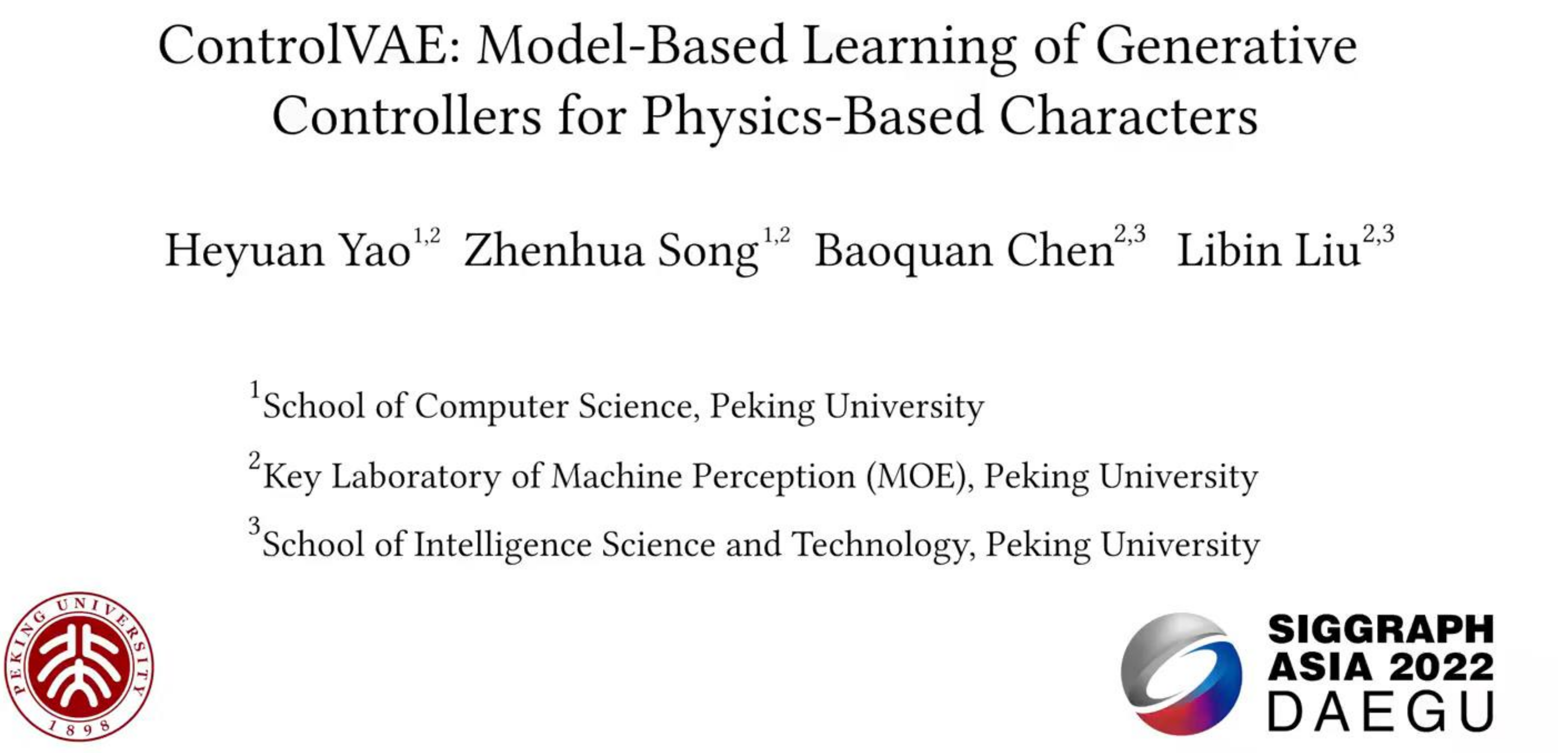
三、机器人动作生成

角色动画生成

http://motiongen.cn/

角色动画生成

Yao, Heyuan, et al. "Controlvae: Model-based learning of generative controllers for physics-based characters." *ACM Transactions on Graphics (TOG)* 41.6 (2022): 1-16.



ControlVAE: Model-Based Learning of Generative Controllers for Physics-Based Characters

Heyuan Yao[1,2] Zhenhua Song[1,2] Baoquan Chen[2,3] Libin Liu[2,3]

[1] School of Computer Science, Peking University

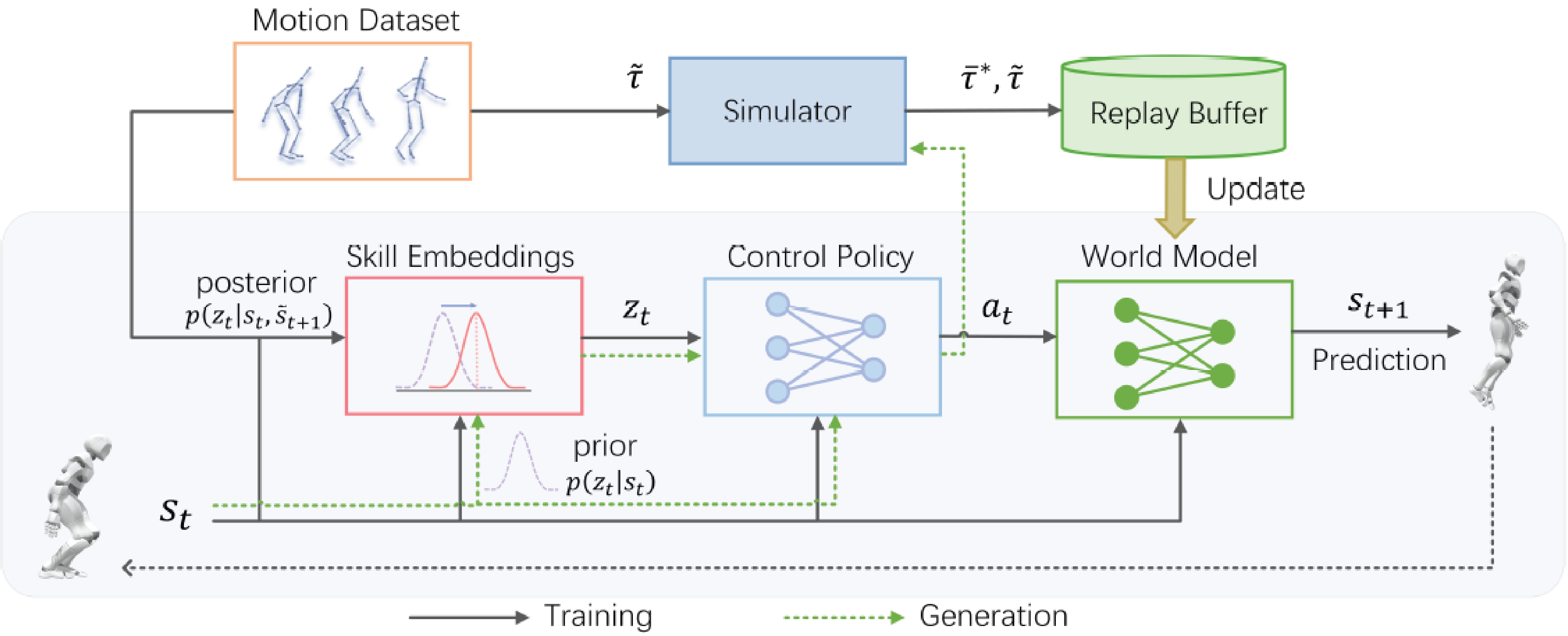
[2] Key Laboratory of Machine Perception (MOE), Peking University

[3] School of Intelligence Science and Technology, Peking University

PEKING UNIVERSITY
1898
SIGGRAPH
ASIA 2022
DAEGU

角色动画生成

Yao, Heyuan, et al. "Controlvae: Model-based learning of generative controllers for physics-based characters." *ACM Transactions on Graphics (TOG)* 41.6 (2022): 1-16.



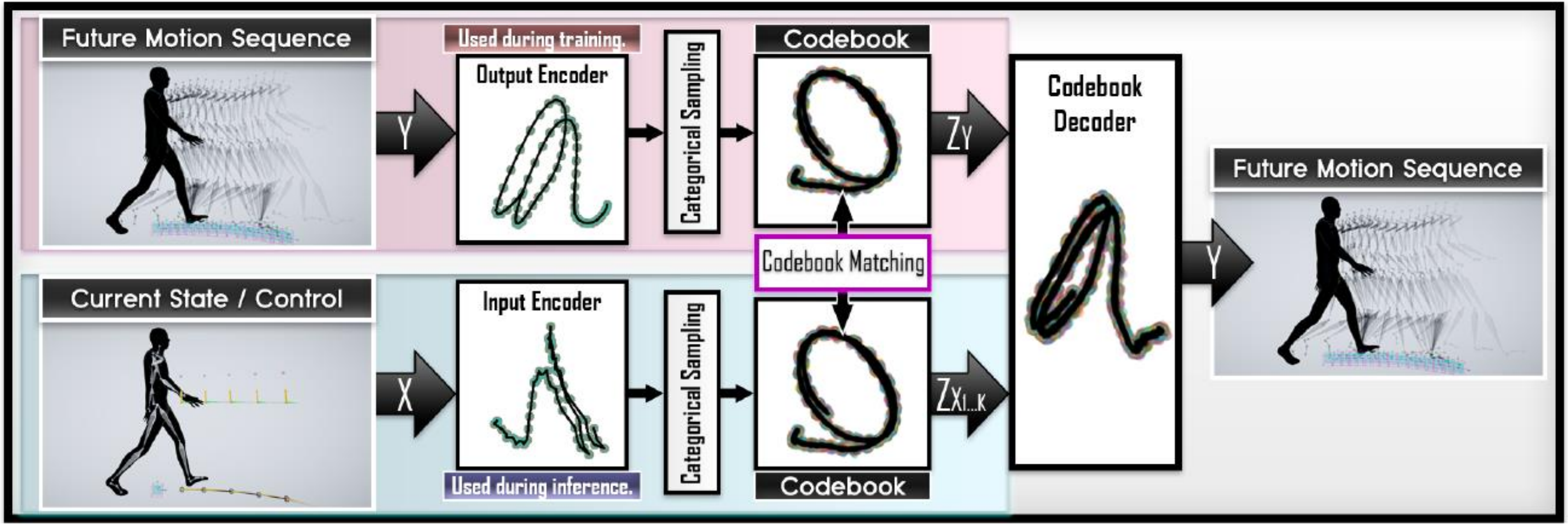
角色动画生成

Starke, Sebastian, et al. "Categorical Codebook Matching for Embodied Character Controllers." *ACM Transactions on Graphics (TOG)* 43.4 (2024): 1-14.

CATEGORICAL CODEBOOK MATCHING FOR
EMBODIED CHARACTER CONTROLLERS
Sebastian Starke[1], Paul Starke[1], Nicky He[1], Taku Komura[2], Yuting Ye[1]
Meta Reality Labs[1], University of Hong Kong[2]

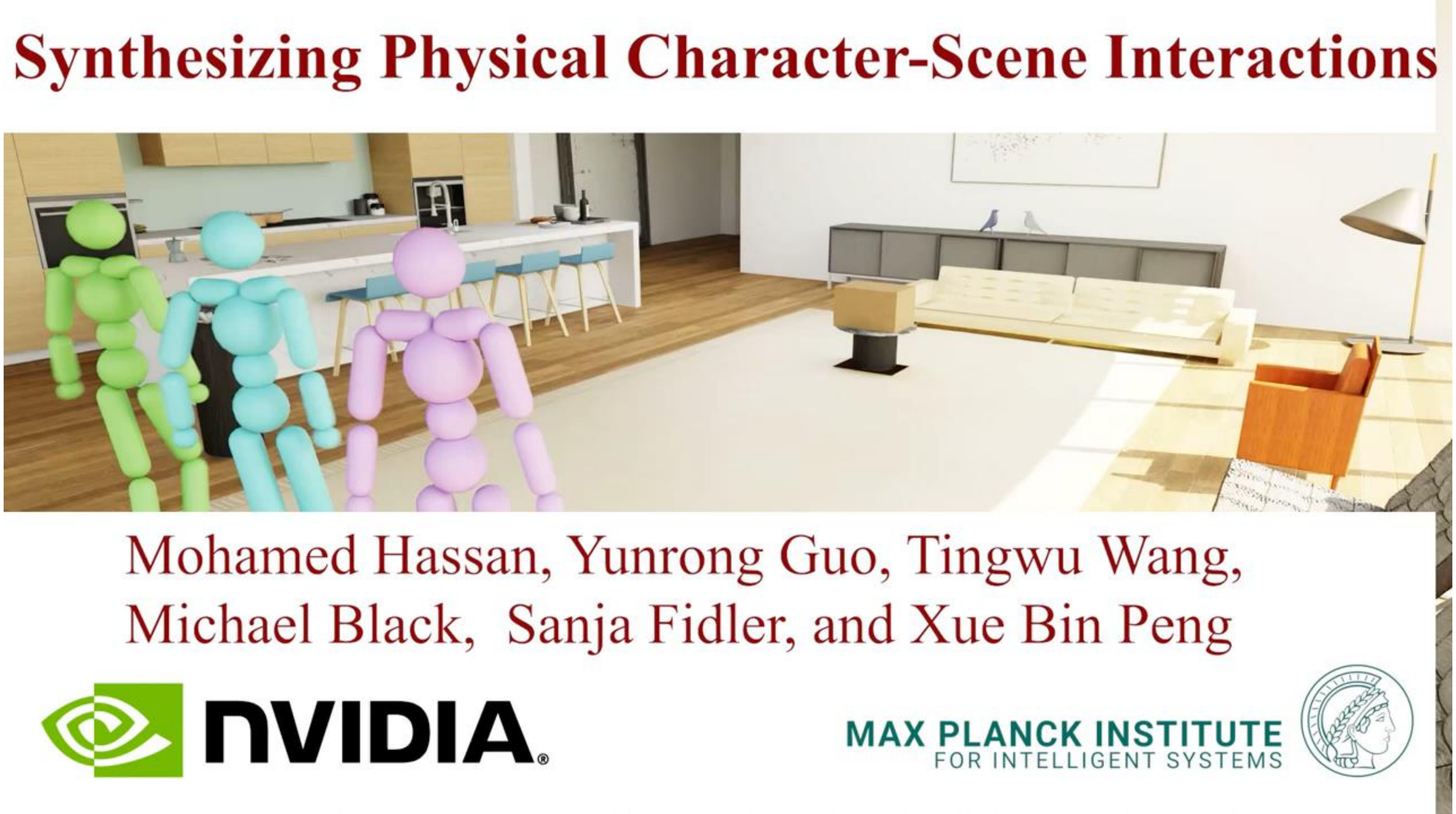
角色动画生成

Starke, Sebastian, et al. "Categorical Codebook Matching for Embodied Character Controllers." *ACM Transactions on Graphics (TOG)* 43.4 (2024): 1-14.



角色动画生成

Hassan, Mohamed, et al. "Synthesizing physical character-scene interactions." *ACM SIGGRAPH 2023 Conference Proceedings*. 2023.



Synthesizing Physical Character-Scene Interactions

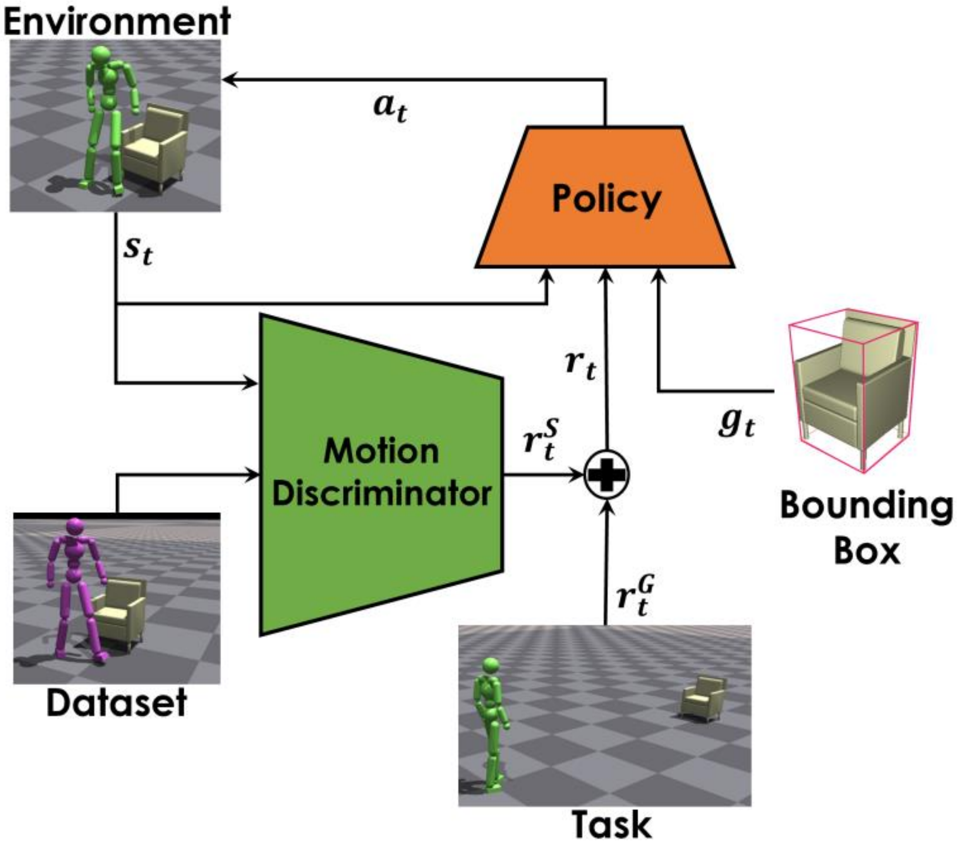
Mohamed Hassan, Yunrong Guo, Tingwu Wang,
Michael Black, Sanja Fidler, and Xue Bin Peng

NVIDIA.

MAX PLANCK INSTITUTE
FOR INTELLIGENT SYSTEMS

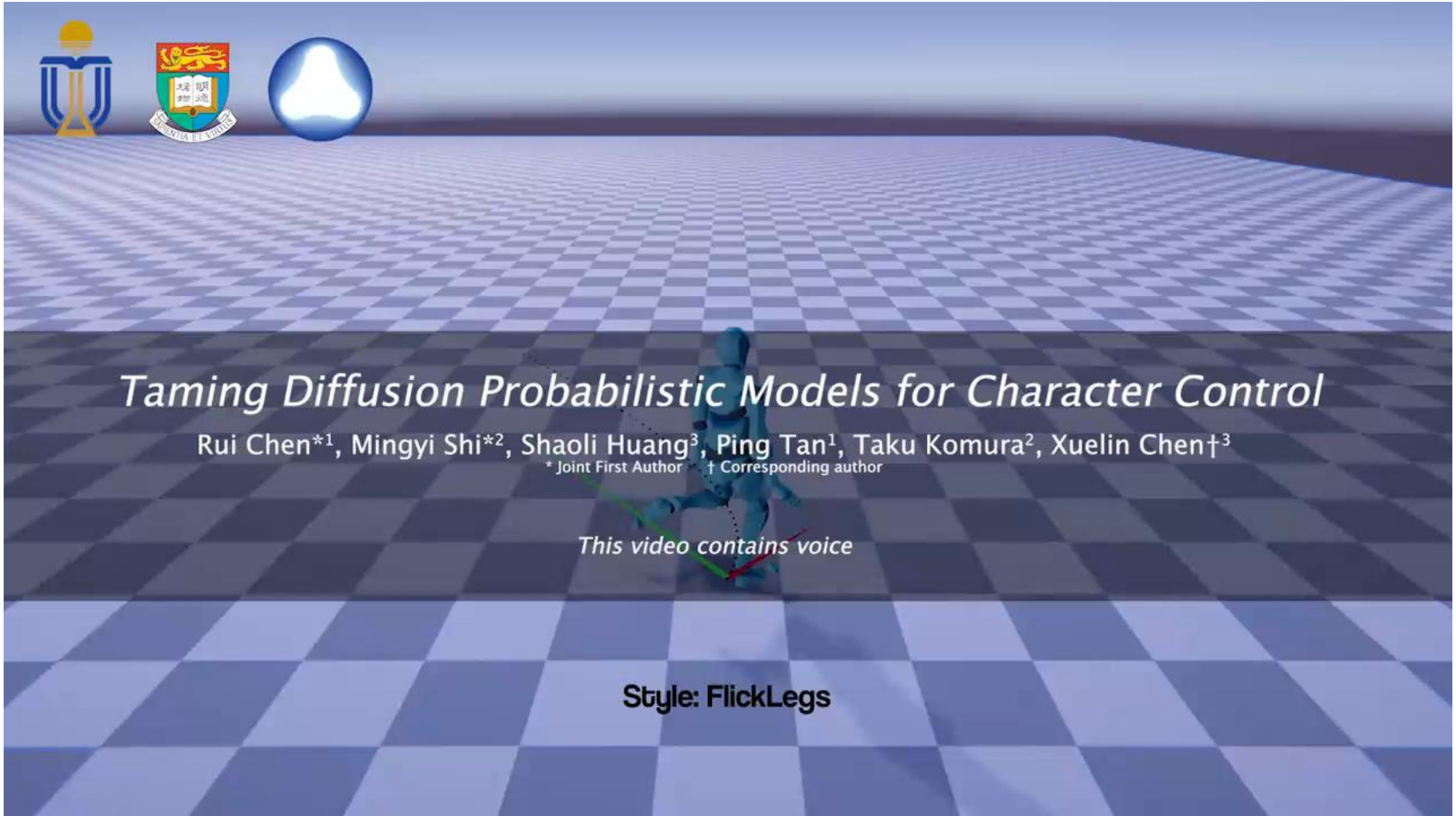
角色动画生成

Hassan, Mohamed, et al. "Synthesizing physical character-scene interactions." *ACM SIGGRAPH 2023 Conference Proceedings*. 2023.



角色动画生成

https://aiganimation.github.io/CAMDM/



Taming Diffusion Probabilistic Models for Character Control

Rui Chen*[1], Mingyi Shi*[2], Shaoli Huang[3], Ping Tan[1], Taku Komura[2], Xuelin Chen†[3]

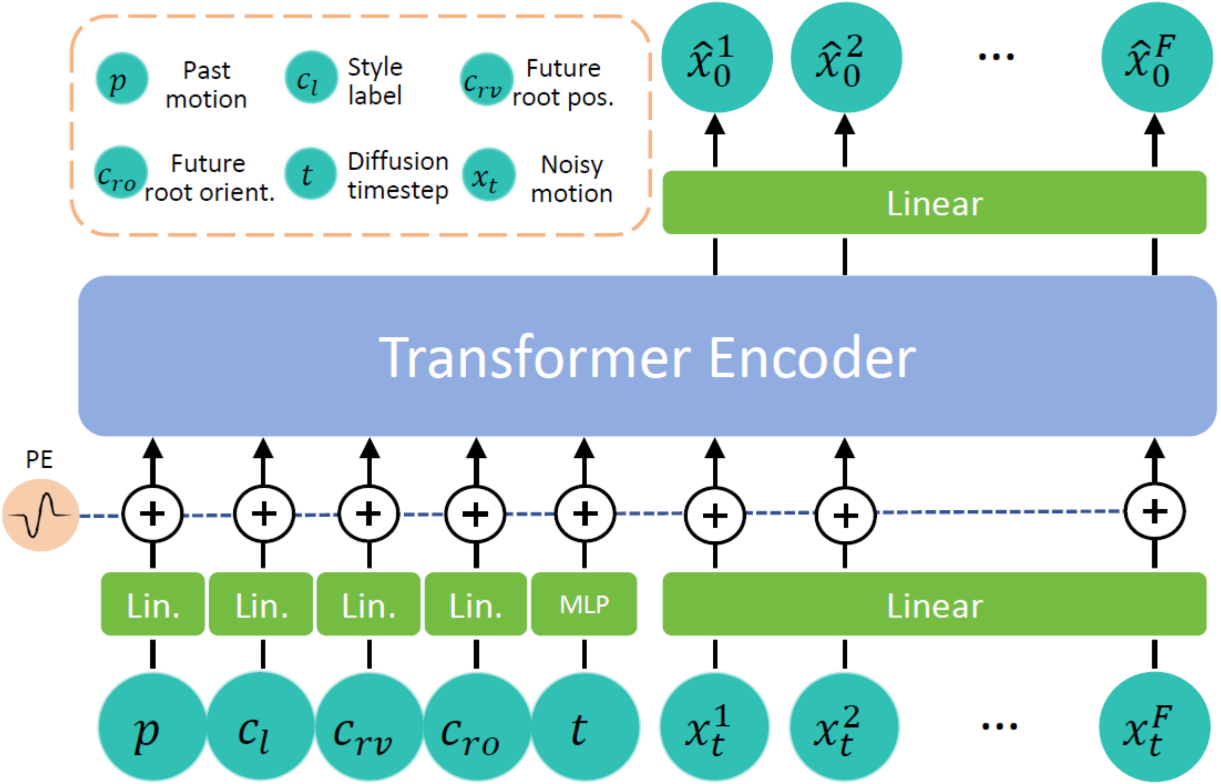
* Joint First Author † Corresponding author

This video contains voice

Style: FlickLegs

角色动画生成

https://aiganimation.github.io/CAMDM/

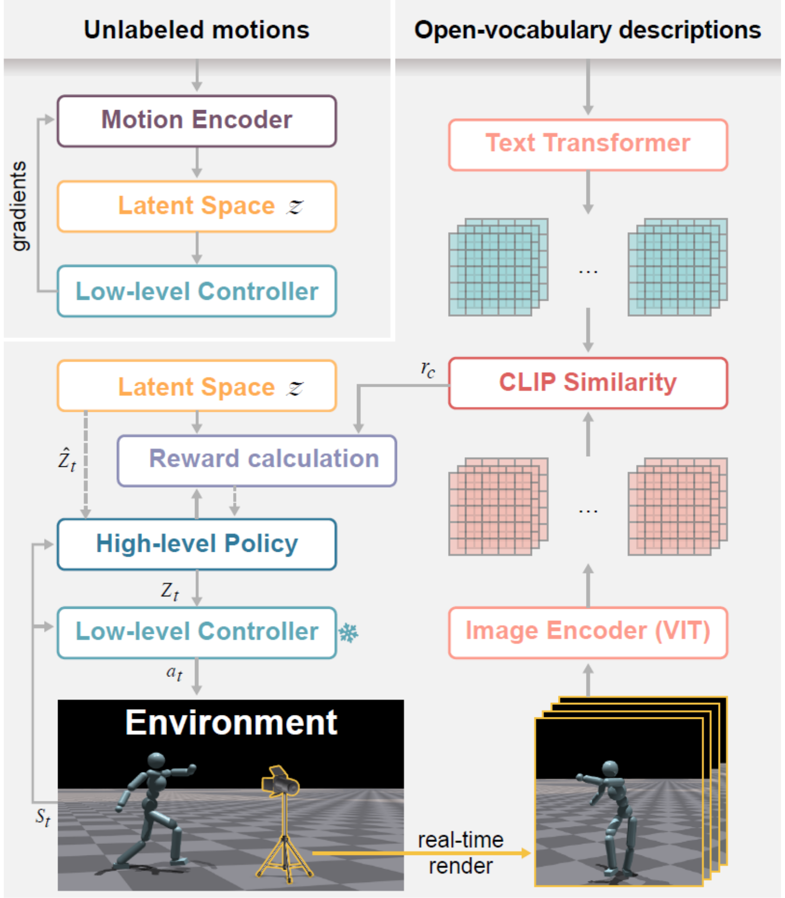


角色动画生成

https://anyskill.github.io/

角色动画生成

https://anyskill.github.io/



提纲

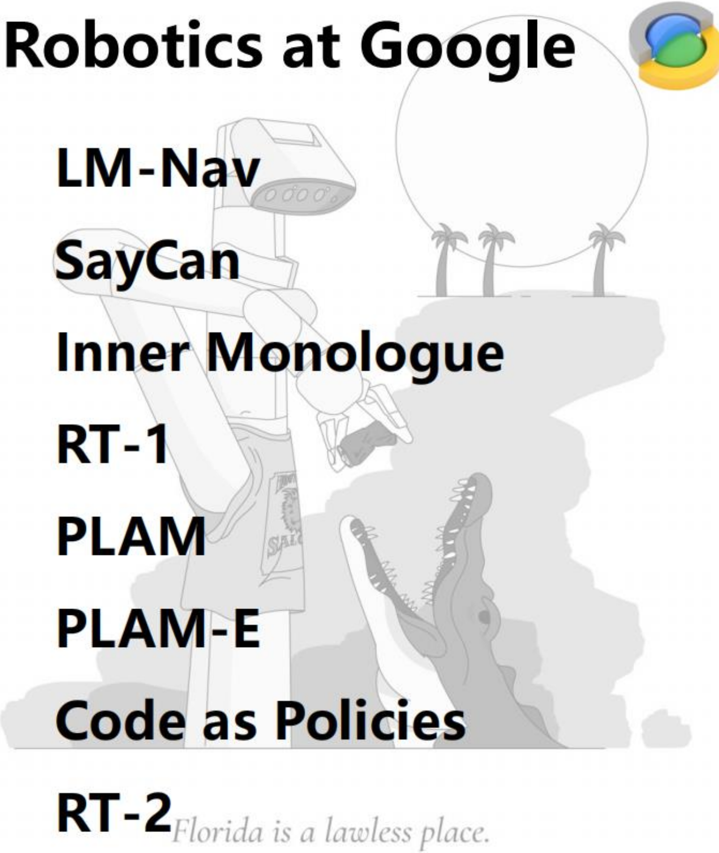
一、生成式模型

二、角色动画生成

三、机器人动作生成

机器人动作生成

Robotics at Google



- LM-Nav
- SayCan
- Inner Monologue
- RT-1
- PLAM
- PLAM-E
- Code as Policies
- RT-2

Other Reseachers

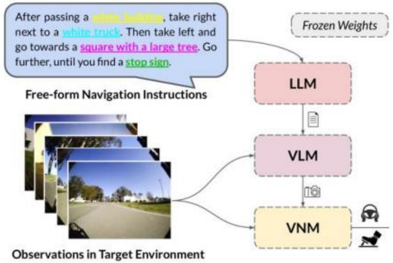
- VoxPoser - Fei-Fei Li
- Diffusion Policy - Shuran Song
- 3D-LLM - Chuang Gan
- etc.

机器人动作生成

Timeline

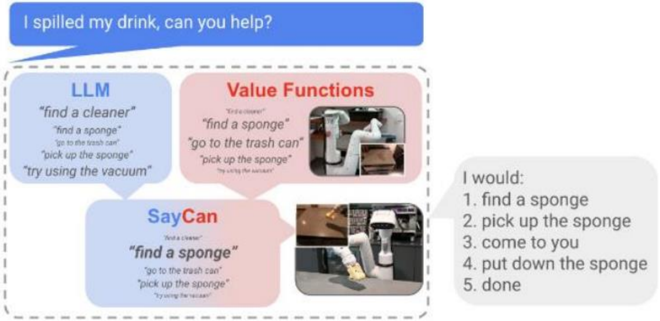
LM-Nav

Navigation based LM



After passing a white building, take right next to a white truck. Then take left and go towards a square with a large tree. Go further, until you find a stop sign.

Free-form Navigation Instructions
Frozen Weights
LLM
VLM
VNM
Observations in Target Environment



I spilled my drink, can you help?

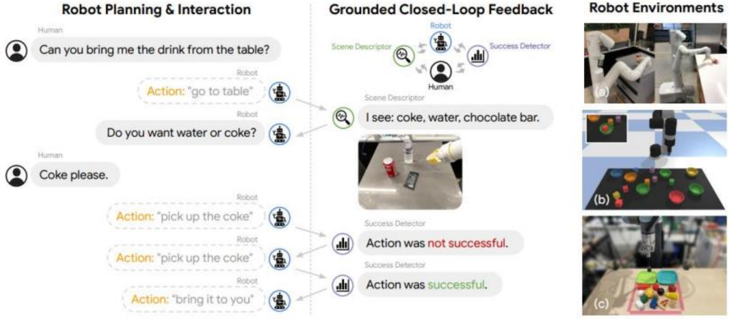
I would:
1. find a sponge
2. pick up the sponge
3. come to you
4. put down the sponge
5. done

LLM Planner
+
BC-Z Skills

SayCan

Inner Monologue

Build Feedback
System for SayCan



Human: Can you bring me the drink from the table?

Robot: Action: "go to table"

Scene Descriptor: I see: coke, water, chocolate bar.

Robot: Do you want water or coke?

Human: Coke please.

Robot: Action: "pick up the coke"

Success Detector: Action was not successful.

Robot: Action: "pick up the coke"

Success Detector: Action was successful.

Robot: Action: "bring it to you"

*cite from Robotics at Google

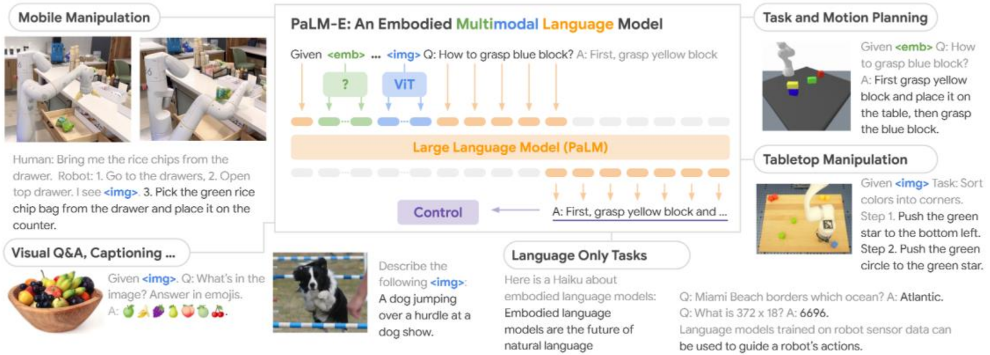
机器人动作生成

RT-1

Transformer for low-level skills

PLAM

Google self-designed LLM



PaLM-E: An Embodied Multimodal Language Model

Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block

A: First, grasp yellow block and ...

Mobile Manipulation

Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Visual Q&A, Captioning ...

Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

Describe the following <img>: A dog jumping over a hurdle at a dog show.

Language Only Tasks

Here is a Haiku about embodied language models: Embodied language models are the future of natural language

Q: Miami Beach borders which ocean? A: Atlantic.
Q: What is 372 x 18? A: 6696.
Language models trained on robot sensor data can be used to guide a robot's actions.

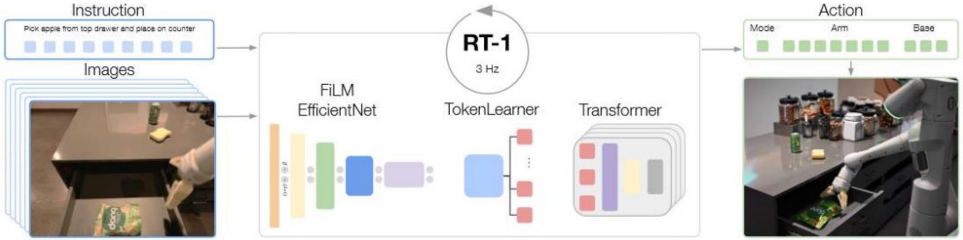
Task and Motion Planning

Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.

Tabletop Manipulation

Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

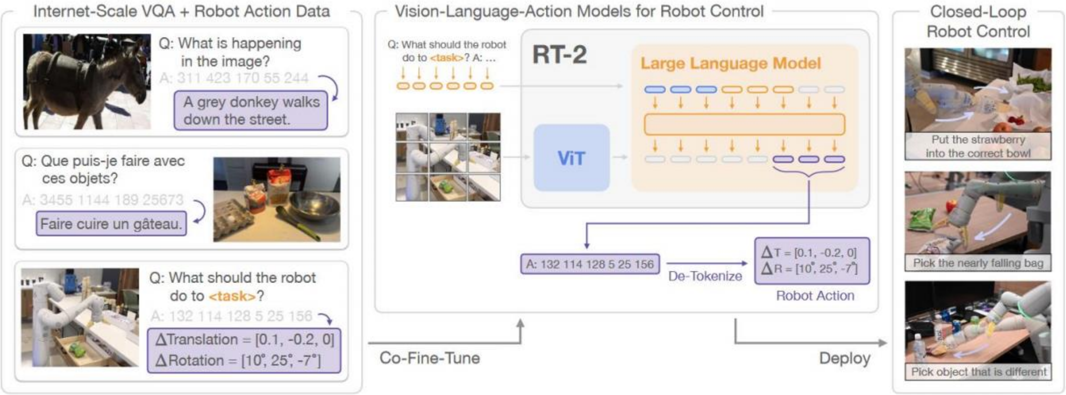
Embodied Multi-modal LM



PLAM-E

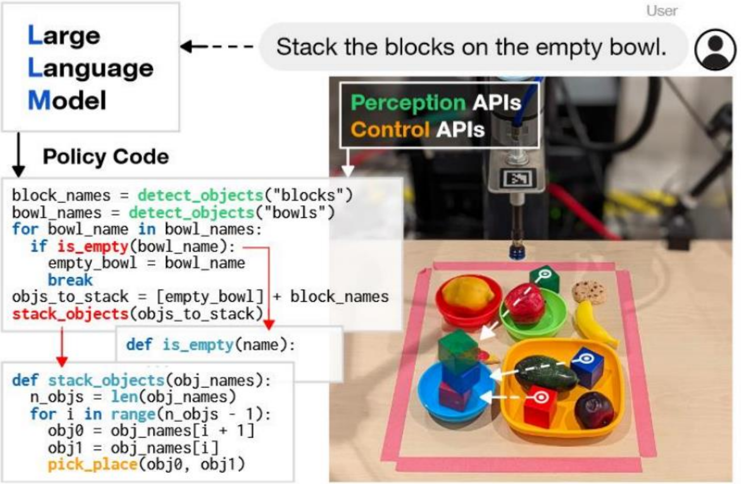
*cite from Robotics at Google

机器人动作生成


Internet-Scale VQA + Robot Action Data
Q: What is happening in the image?
A: 311 423 170 55 244
A grey donkey walks down the street.
Q: Que puis-je faire avec ces objets?
A: 3455 1144 189 25673
Faire cuire un gâteau.
Q: What should the robot do to <task>?
A: 132 114 128 5 25 156
ΔTranslation = [0.1, -0.2, 0]
ΔRotation = [10°, 25°, -7°]
Vision-Language-Action Models for Robot Control
Q: What should the robot do to <task>? A: ...
RT-2
Large Language Model
ViT
A: 132 114 128 5 25 156
De-Tokenize
ΔT = [0.1, -0.2, 0]
ΔR = [10°, 25°, -7°]
Robot Action
Co-Fine-Tune
Deploy
Closed-Loop Robot Control
Put the strawberry into the correct bowl
Pick the nearly falling bag
Pick object that is different

Code as Policies

Prompt to generate codes



Stack the blocks on the empty bowl.

```
block_names = detect_objects("blocks")
bowl_names = detect_objects("bowls")
for bowl_name in bowl_names:
  if is_empty(bowl_name):
    empty_bowl = bowl_name
    break
objs_to_stack = [empty_bowl] + block_names
stack_objects(objs_to_stack)
```

```
def is_empty(name):
```

```
def stack_objects(obj_names):
  n_objs = len(obj_names)
  for i in range(n_objs - 1):
    obj0 = obj_names[i + 1]
    obj1 = obj_names[i]
    pick_place(obj0, obj1)
```

end-to-end low-level controller

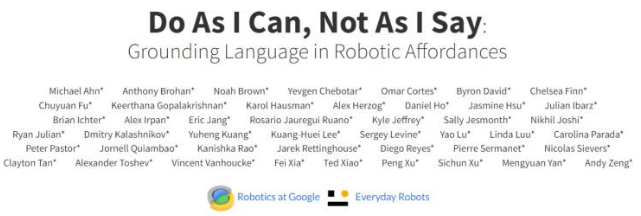
RT-2

*cite from Robotics at Google

机器人动作生成

SayCan

Paper

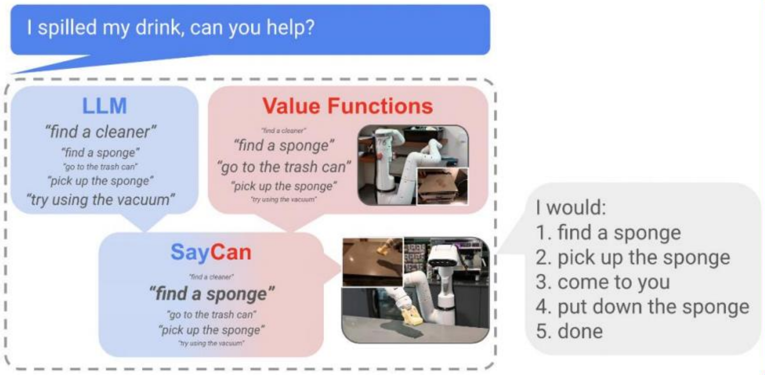


Do As I Can, Not As I Say:
Grounding Language in Robotic Affordances

Michael Ahn* Anthony Brohan* Noah Brown* Yevgen Chebotar* Omar Cortes* Byron David* Chelsea Finn*
Chuyuan Fu* Keerthana Gopalakrishnan* Karol Hausman* Alex Herzog* Daniel Ho* Jasmine Hsu* Julian Ibarz*
Brian Ichter* Alex Irpan* Eric Jang* Rosario Jauregui Ruano* Kyle Jeffrey* Sally Jesmonth* Nikhil Joshi*
Ryan Julian* Dmitry Kalashnikov* Yuheng Kuang* Kuang-Huei Lee* Sergey Levine* Yao Lu* Linda Luu* Carolina Parada*
Peter Pastor* Jornell Quiambao* Kanishka Rao* Jarek Rettinghouse* Diego Reyes* Pierre Sermanet* Nicolas Sievers*
Clayton Tan* Alexander Toshev* Vincent Vanhoucke* Fei Xia* Ted Xiao* Peng Xu* Sichun Xu* Mengyuan Yan* Andy Zeng*

Robotics at Google Everyday Robots

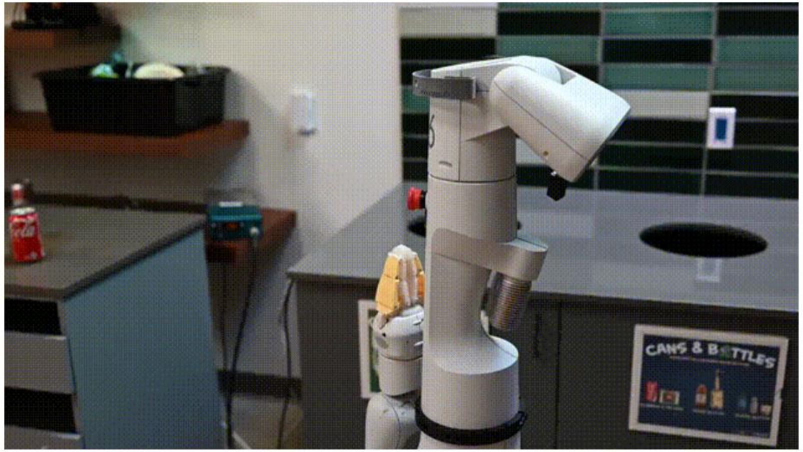
Overview



I spilled my drink, can you help?

I would:
1. find a sponge
2. pick up the sponge
3. come to you
4. put down the sponge
5. done

Demo



*cite from Robotics at Google

EXIT
076
076

机器人动作生成

RT-1

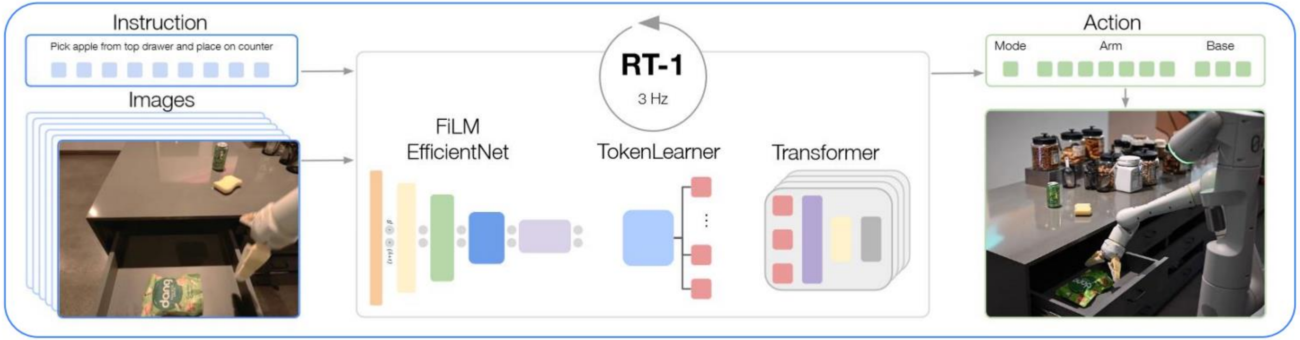
Paper

RT-1: ROBOTICS TRANSFORMER FOR REAL-WORLD CONTROL AT SCALE

Anthony Brohan*, Noah Brown*, Justice Carbajal*, Yevgen Chebotar*, Joseph Dabis*, Chelsea Finn*, Keerthana Gopalakrishnan*, Karol Hausman*, Alex Herzog†, Jasmine Hsu*, Julian Ibarz*, Brian Ichter*, Alex Irpan*, Tomas Jackson*, Sally Jesmonth*, Nikhil J Joshi*, Ryan Julian*, Dmitry Kalashnikov*, Yuheng Kuang*, Isabel Leal*, Kuang-Huei Lee‡, Sergey Levine*, Yao Lu*, Utsav Malla*, Deeksha Manjunath*, Igor Mordatch‡, Ofir Nachum‡, Carolina Parada*, Jodilyn Peralta*, Emily Perez*, Karl Pertsch*, Jornell Quiambao*, Kanishka Rao*, Michael Ryoo*, Grecia Salazar*, Pannag Sanketi*, Kevin Sayed*, Jaspiar Singh*, Sumedh Sontakke‡, Austin Stone*, Clayton Tan*, Huong Tran*, Vincent Vanhoucke*, Steve Vega*, Quan Vuong*, Fei Xia*, Ted Xiao*, Peng Xu*, Sichun Xu*, Tianhe Yu*, Brianna Zitkovich*

*Robotics at Google, †Everyday Robots, ‡Google Research, Brain Team

Overview



Details

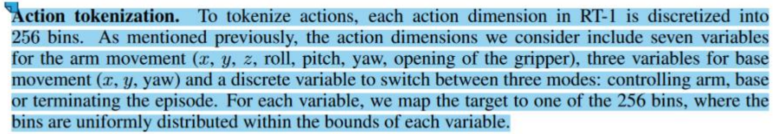
1. Action Spaces

arm：x，y，z，roll，pitch，yaw，opening of grasper

base：x，y，yaw

switch mode：{control the arm，the base，termination}

control frequency = 3Hz



Action tokenization. To tokenize actions, each action dimension in RT-1 is discretized into 256 bins. As mentioned previously, the action dimensions we consider include seven variables for the arm movement (x, y, z, roll, pitch, yaw, opening of the gripper), three variables for base movement (x, y, yaw) and a discrete variable to switch between three modes: controlling arm, base or terminating the episode. For each variable, we map the target to one of the 256 bins, where the bins are uniformly distributed within the bounds of each variable.

*cite from Robotics at Google

机器人动作生成

PALM-E

Paper

PaLM-E: An Embodied Multimodal Language Model

Danny Driess[1,2] Fei Xia[1] Mehdi S. M. Sajjadi[3] Corey Lynch[1] Aakanksha Chowdhery[3]
Brian Ichter[1] Ayzaan Wahid[1] Jonathan Tompson[1] Quan Vuong[1] Tianhe Yu[1] Wenlong Huang[1]
Yevgen Chebotar[1] Pierre Sermanet[1] Daniel Duckworth[3] Sergey Levine[1] Vincent Vanhoucke[1]
Karol Hausman[1] Marc Toussaint[2] Klaus Greff[3] Andy Zeng[1] Igor Mordatch[3] Pete Florence[1]

[1] Robotics at Google [2] TECHNISCHE UNIVERSITÄT BERLIN [3] Google Research

Details

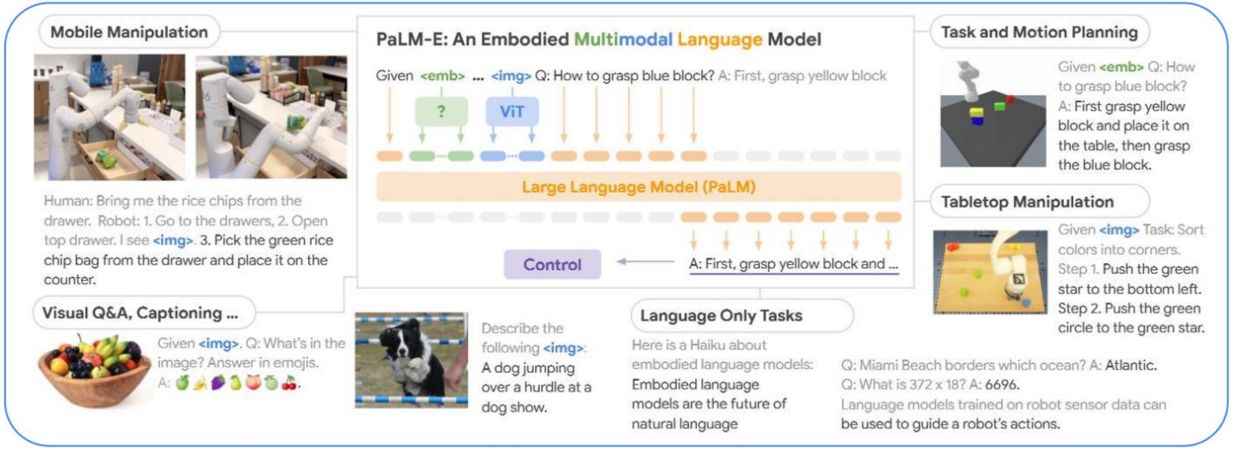
1. Multi-modal Imputs

1.1 State estimation

1.2 Images

1.3 Language

Overview


Mobile Manipulation
PaLM-E: An Embodied Multimodal Language Model
?
ViT
Large Language Model (PaLM)
Control
Task and Motion Planning
Tabletop Manipulation
Visual Q&A, Captioning ...
Language Only Tasks

Given <emb> ... <img> Q: How to grasp blue block? A: First, grasp yellow block

A: First, grasp yellow block and ...

Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see <img>. 3. Pick the green rice chip bag from the drawer and place it on the counter.

Given <emb> Q: How to grasp blue block? A: First grasp yellow block and place it on the table, then grasp the blue block.

Given <img> Task: Sort colors into corners. Step 1. Push the green star to the bottom left. Step 2. Push the green circle to the green star.

Given <img>. Q: What's in the image? Answer in emojis. A: 🍏🍌🍇🍐🍑🍈🍒.

Describe the following <img>: A dog jumping over a hurdle at a dog show.

Here is a Haiku about embodied language models: Embodied language models are the future of natural language

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is 372 x 18? A: 6696. Language models trained on robot sensor data can be used to guide a robot's actions.

4x speed
106

机器人动作生成

Code as Polices

Paper

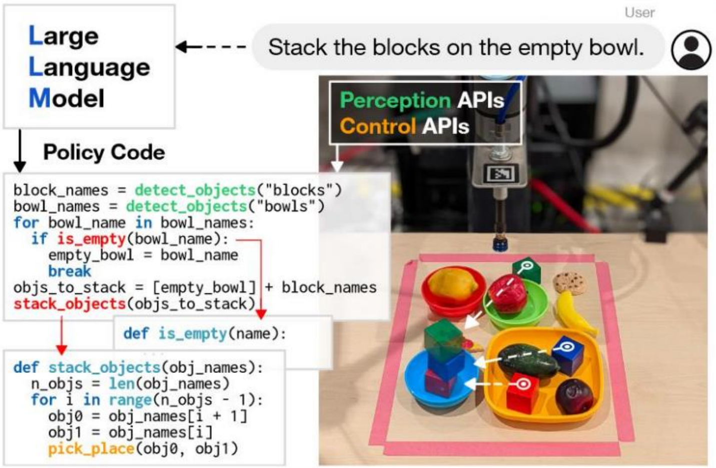
Code as Policies:

Language Model Programs for Embodied Control

Jacky Liang Wenlong Huang Fei Xia Peng Xu Karol Hausman Brian Ichter Pete Florence Andy Zeng

Robotics at Google

Overview



```
block_names = detect_objects("blocks")
bowl_names = detect_objects("bowls")
for bowl_name in bowl_names:
  if is_empty(bowl_name):
    empty_bowl = bowl_name
    break
objs_to_stack = [empty_bowl] + block_names
stack_objects(objs_to_stack)
```

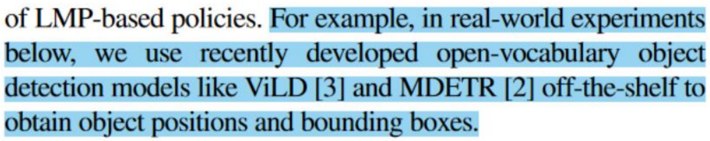
```
def is_empty(name):
```

```
def stack_objects(obj_names):
  n_objs = len(obj_names)
  for i in range(n_objs - 1):
    obj0 = obj_names[i + 1]
    obj1 = obj_names[i]
    pick_place(obj0, obj1)
```

Details

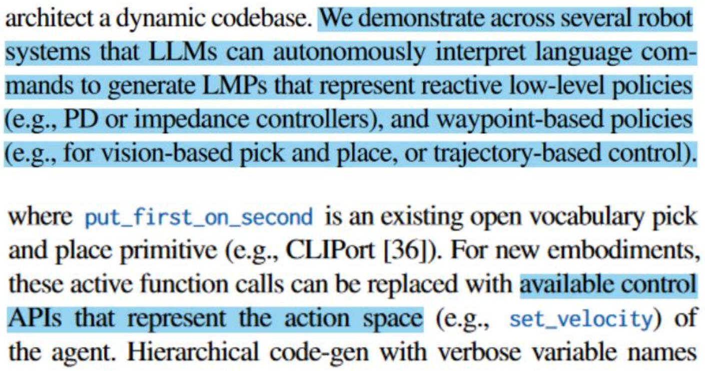
1. APIs

Perception APIs



of LMP-based policies. For example, in real-world experiments below, we use recently developed open-vocabulary object detection models like ViLD [3] and MDETR [2] off-the-shelf to obtain object positions and bounding boxes.

Control APIs



architect a dynamic codebase. We demonstrate across several robot systems that LLMs can autonomously interpret language commands to generate LMPs that represent reactive low-level policies (e.g., PD or impedance controllers), and waypoint-based policies (e.g., for vision-based pick and place, or trajectory-based control).

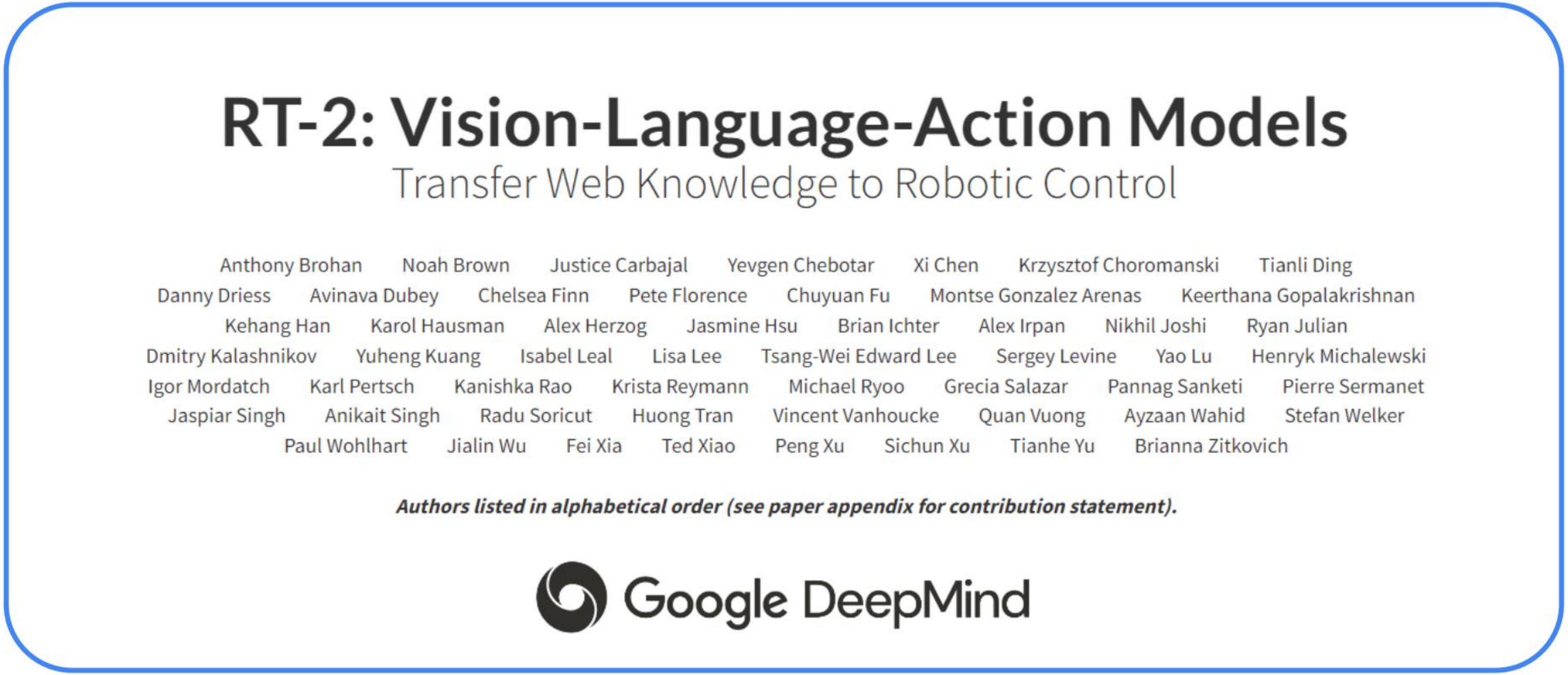
where put_first_on_second is an existing open vocabulary pick and place primitive (e.g., CLIPort [36]). For new embodiments, these active function calls can be replaced with available control APIs that represent the action space (e.g., set_velocity) of the agent. Hierarchical code-gen with verbose variable names

*cite from Robotics at Google

机器人动作生成

RT-2

Paper



RT-2: Vision-Language-Action Models
Transfer Web Knowledge to Robotic Control

Anthony Brohan Noah Brown Justice Carbajal Yevgen Chebotar Xi Chen Krzysztof Choromanski Tianli Ding
Danny Driess Avinava Dubey Chelsea Finn Pete Florence Chuyuan Fu Montse Gonzalez Arenas Keerthana Gopalakrishnan
Kehang Han Karol Hausman Alex Herzog Jasmine Hsu Brian Ichter Alex Irpan Nikhil Joshi Ryan Julian
Dmitry Kalashnikov Yuheng Kuang Isabel Leal Lisa Lee Tsang-Wei Edward Lee Sergey Levine Yao Lu Henryk Michalewski
Igor Mordatch Karl Pertsch Kanishka Rao Krista Reymann Michael Ryoo Grecia Salazar Pannag Sanketi Pierre Sermanet
Jaspiar Singh Anikait Singh Radu Soricut Huong Tran Vincent Vanhoucke Quan Vuong Ayzaan Wahid Stefan Welker
Paul Wohlhart Jialin Wu Fei Xia Ted Xiao Peng Xu Sichun Xu Tianhe Yu Brianna Zitkovich

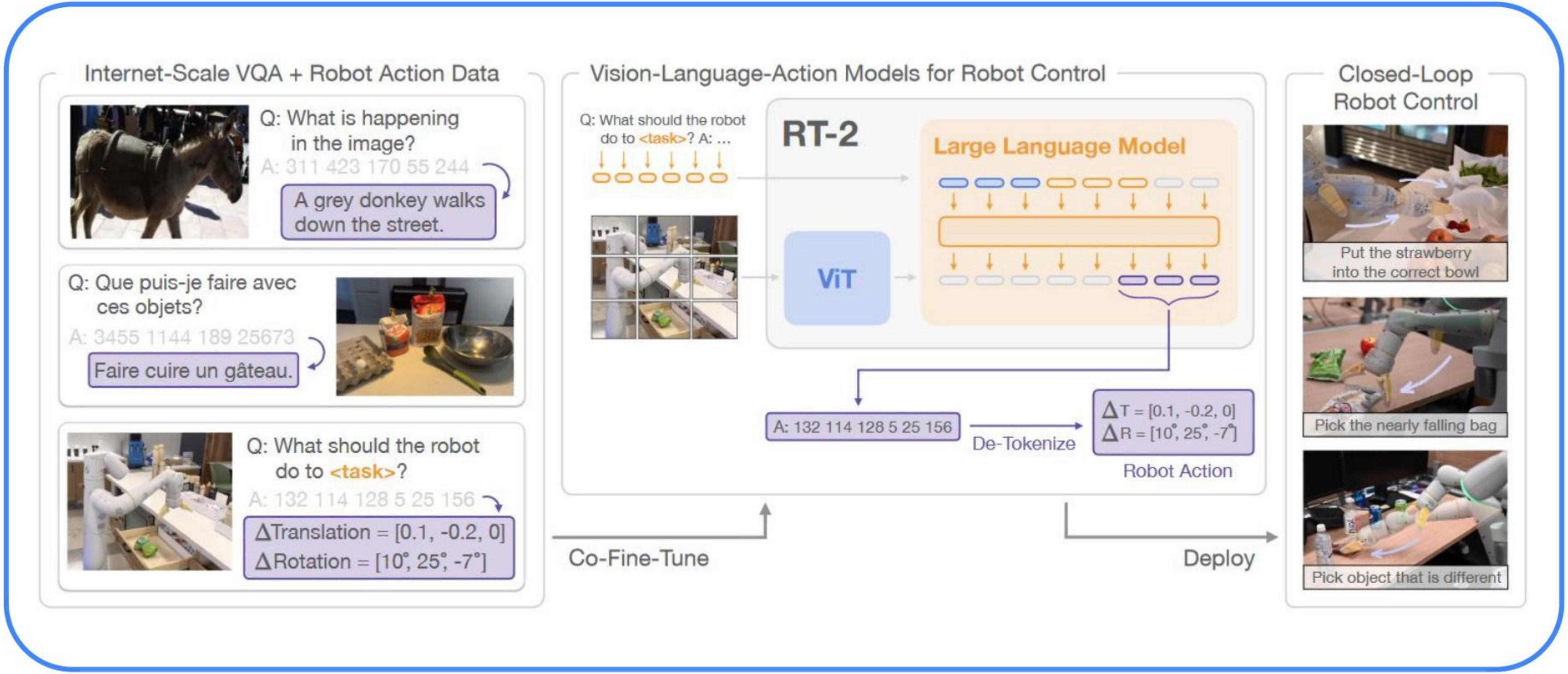
Authors listed in alphabetical order (see paper appendix for contribution statement).

Google DeepMind

机器人动作生成

RT-2

Overview


Internet-Scale VQA + Robot Action Data
Q: What is happening in the image?
A: 311 423 170 55 244
A grey donkey walks down the street.
Q: Que puis-je faire avec ces objets?
A: 3455 1144 189 25673
Faire cuire un gâteau.
Q: What should the robot do to <task>?
A: 132 114 128 5 25 156
ΔTranslation = [0.1, -0.2, 0]
ΔRotation = [10°, 25°, -7°]
Vision-Language-Action Models for Robot Control
Q: What should the robot do to <task>? A: ...
RT-2
Large Language Model
ViT
A: 132 114 128 5 25 156
De-Tokenize
ΔT = [0.1, -0.2, 0]
ΔR = [10°, 25°, -7°]
Robot Action
Co-Fine-Tune
Deploy
Closed-Loop Robot Control
Put the strawberry into the correct bowl
Pick the nearly falling bag
Pick object that is different

机器人动作生成

RT-2

Demo

put strawberry into the correct bowl

pick up the bag about to fall off the table

move apple to Denver Nuggets

pick robot

place orange in matching bowl

move redbull can to H

move soccer ball to basketball

move banana to Germany

move cup to the wine bottle

pick animal with different colour

机器人动作生成

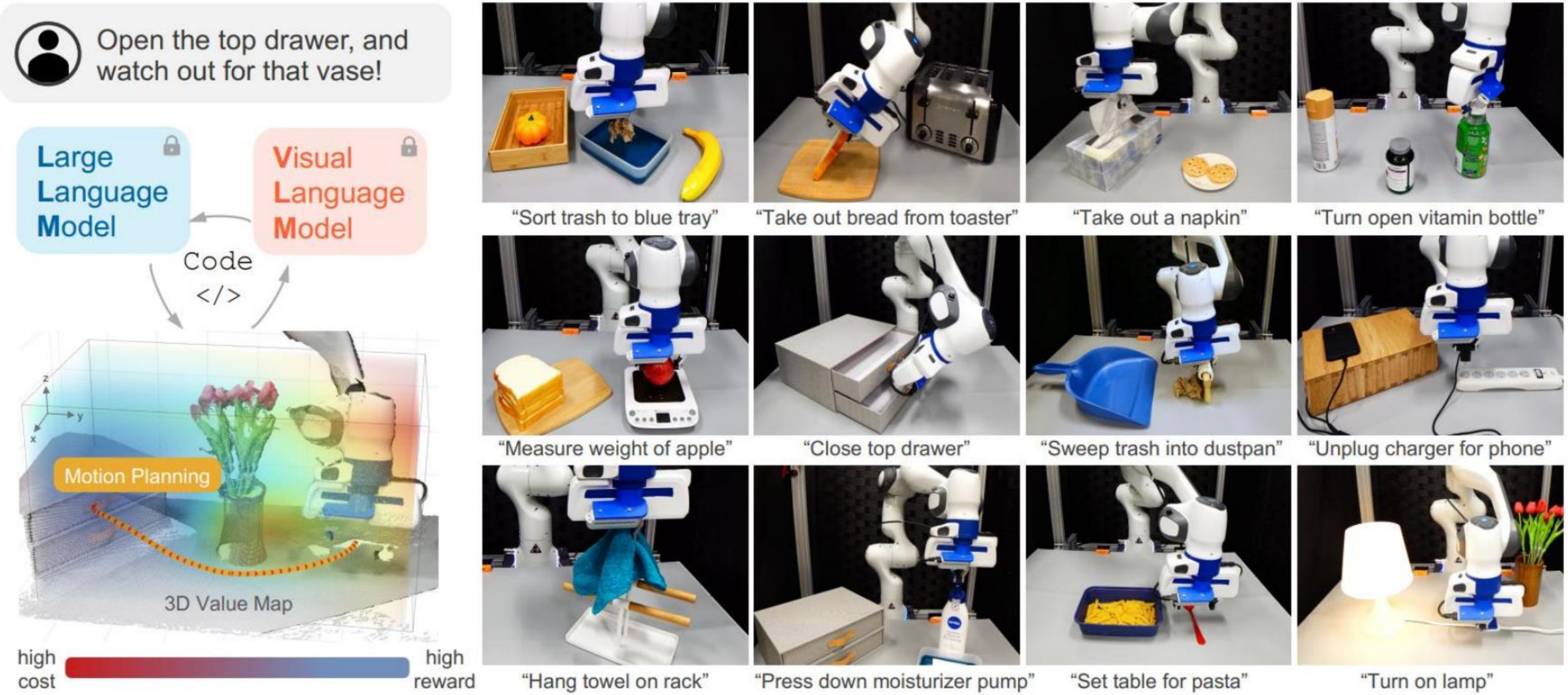

Open the top drawer, and watch out for that vase!
Large Language Model
Visual Language Model
Code
</>
Motion Planning
3D Value Map
z
y
x
high cost
high reward
"Sort trash to blue tray"
"Take out bread from toaster"
"Take out a napkin"
"Turn open vitamin bottle"
"Measure weight of apple"
"Close top drawer"
"Sweep trash into dustpan"
"Unplug charger for phone"
"Hang towel on rack"
"Press down moisturizer pump"
"Set table for pasta"
"Turn on lamp"

Figure 1: **VoxPoser** extracts language-conditioned **affordances** and **constraints** from LLMs and grounds them to the perceptual space using VLMs, using a code interface and without additional training to either component. The composed map is referred to as a 3D value map, which enables **zero-shot** synthesis of trajectories for large varieties of everyday manipulation tasks with an **open-set of instructions** and an **open-set of objects**.

This video has audio

VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, Li Fei-Fei

Stanford Artificial Intelligence

Stanford Vision & Learning

University of Illinois Urbana-Champaign

机器人动作生成

斯坦福大学的科研团队近日开发了 Mobile ALOHA，可以执行打开厨房用具柜、洗锅、炸虾、做菜、打扫卫生、整理衣物、套被套等 50 多项家务。这款家用机器人成本仅 3.2 万美元。

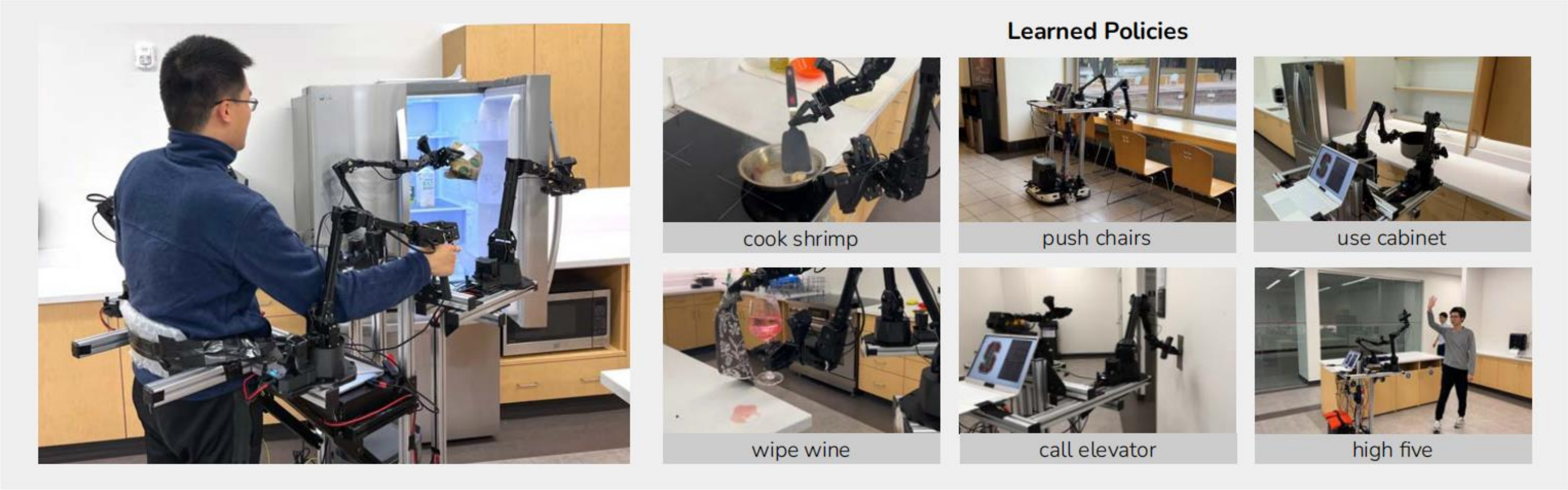


Figure 1: ***Mobile ALOHA*** 🏄 . We introduce a low-cost mobile manipulation system that is bimanual and supports whole-body teleoperation. The system costs $32k including onboard power and compute. *Left:* A user teleoperates to obtain food from the fridge. *Right: Mobile ALOHA* can perform complex long-horizon tasks with imitation learning.

机器人动作生成

其算法Action Chunking with Transformers （ACT）采用了神经网络模型Transformers，因此**具备模仿学习能力。只需要15分钟的演示，机械臂就可以学会一个动作**——直接从真实演示中执行端到端模仿学习，并通过自定义远程操作界面收集。

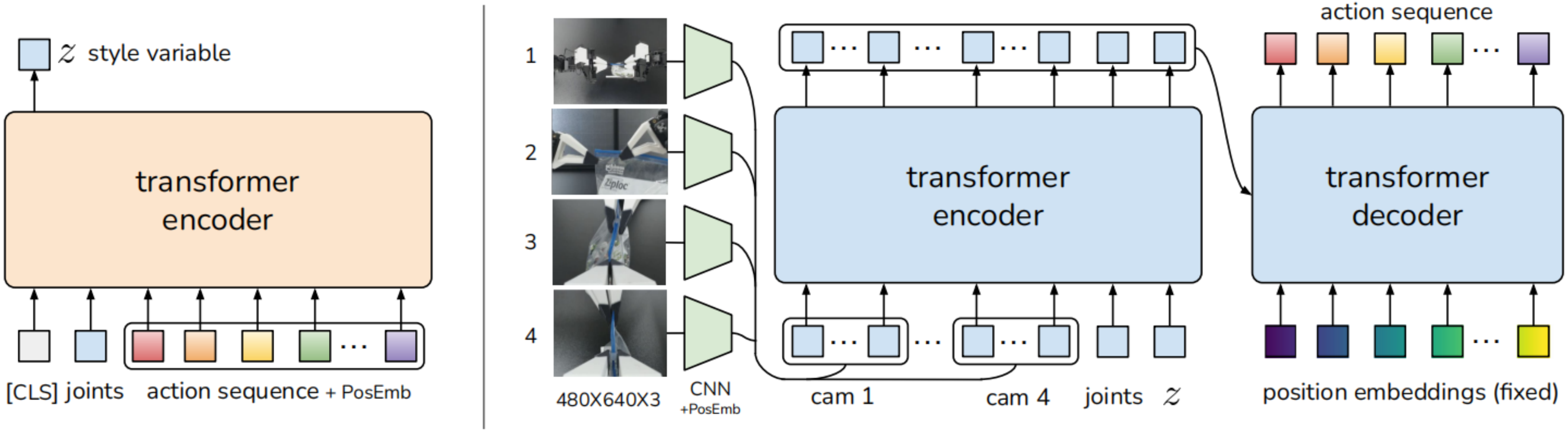
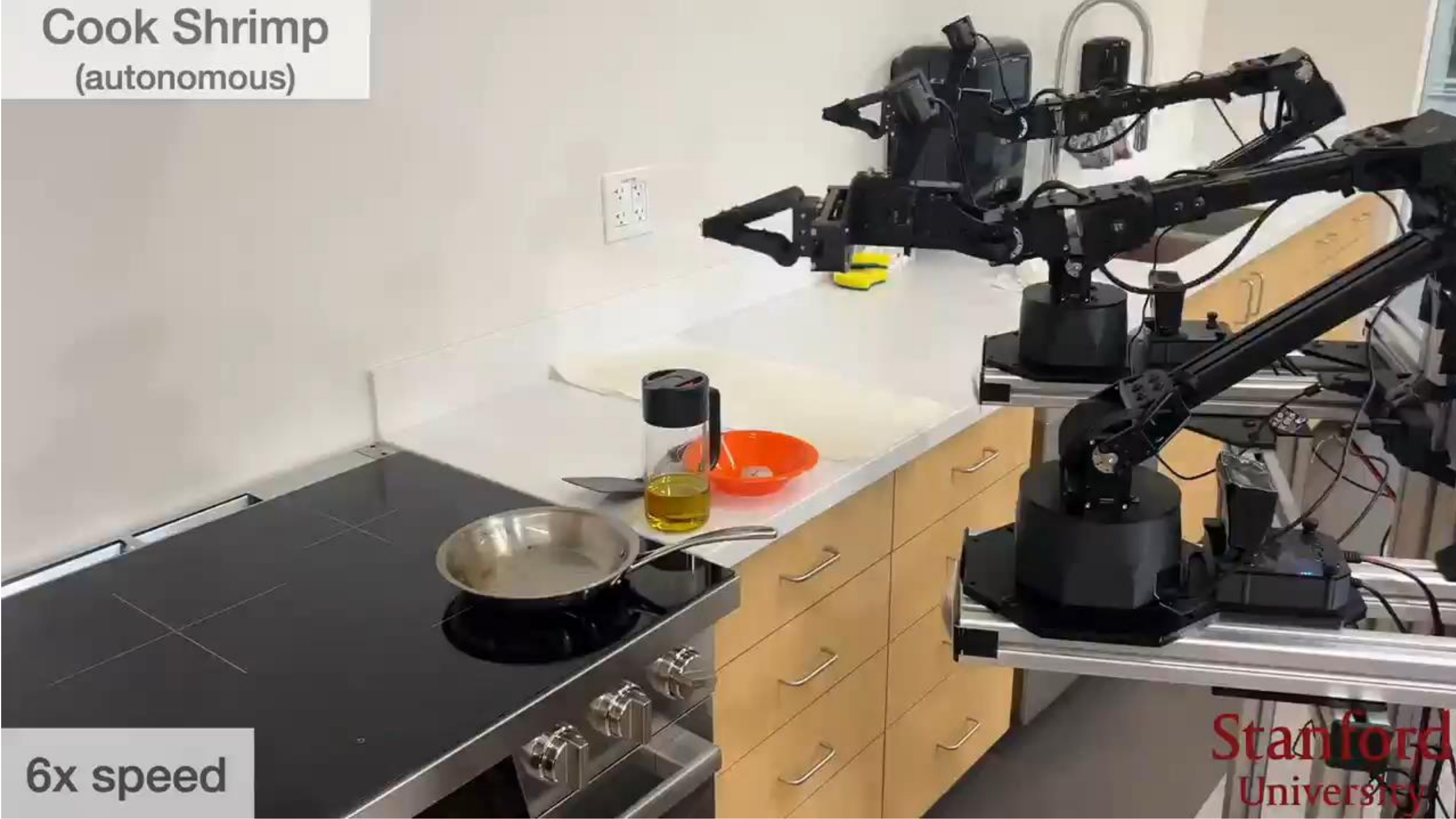


Fig. 4: *Architecture of Action Chunking with Transformers (ACT).* We train ACT as a Conditional VAE (CVAE), which has an encoder and a decoder. *Left:* The encoder of the CVAE compresses action sequence and joint observation into z, the style variable. The encoder is discarded at test time. *Right:* The decoder or policy of ACT synthesizes images from multiple viewpoints, joint positions, and z with a transformer encoder, and predicts a sequence of actions with a transformer decoder. z is simply set to the mean of the prior (i.e. zero) at test time.



Cook Shrimp
(autonomous)

6x speed

谢谢大家