



---

# 多模态通用大模型

---

魏明强、宫丽娜

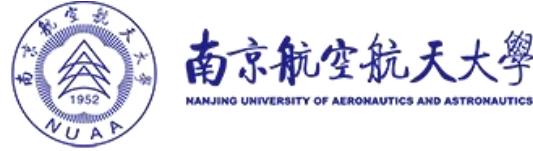
计算机科学与技术学院

---

智周万物·道济天下

---

# 目录



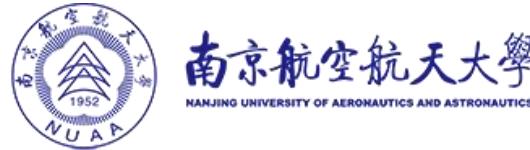
- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V
- ImageBind
- 3D-LLM
- Sora

# 目录

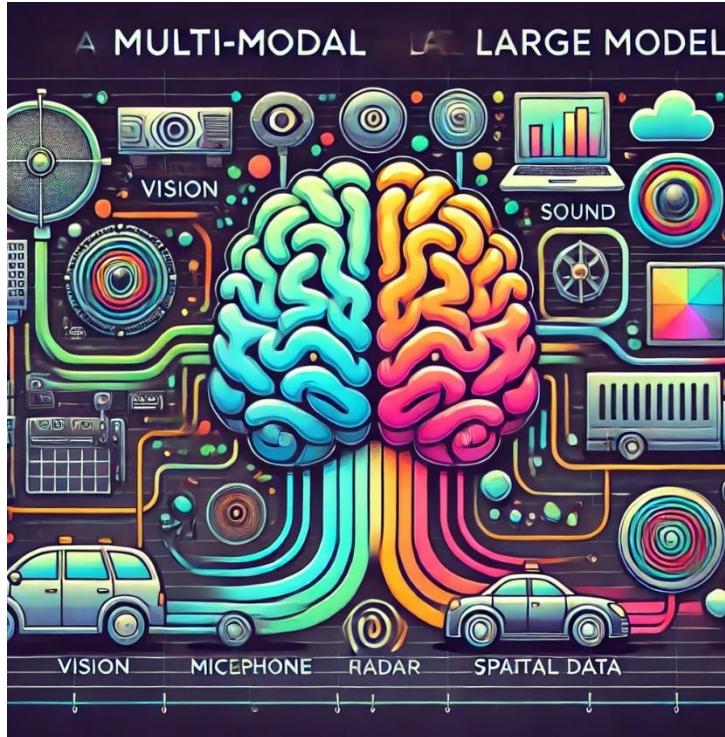
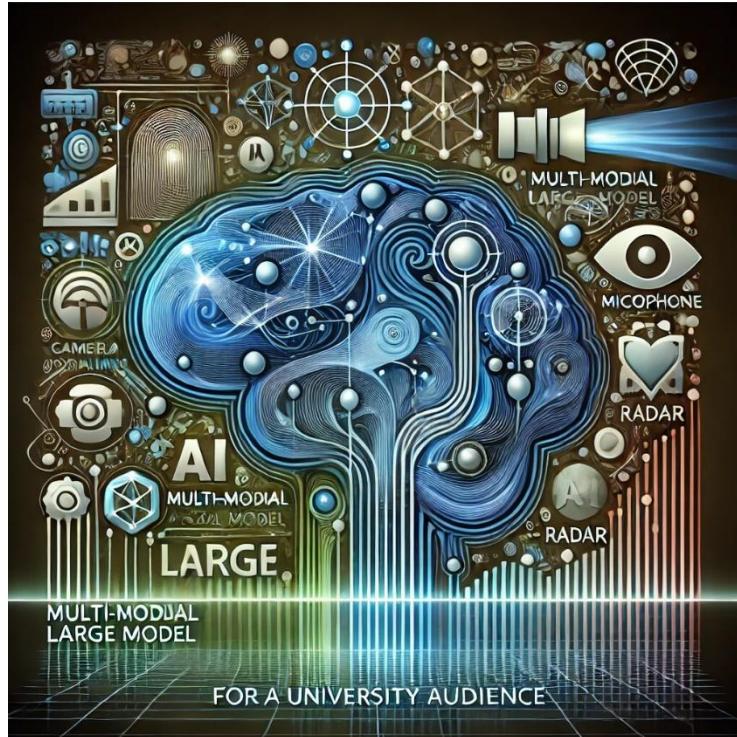


- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V
- ImageBind
- 3D-LLM
- Sora

# 引言和目标



- **多模态大模型**是一种能够同时处理和理解多种类型数据的人工智能模型，旨在**模拟人类多感官的信息处理方式**。
- 本章将介绍多模态大模型的基础概念、代表性模型（如CLIP和GPT-4V）、关键技术、实际应用，以及未来的发展趋势。



学习目标：掌握多模态大模型的关键技术和代表性模型的架构；

# 目录



- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V
- ImageBind
- 3D-LLM
- Sora

# 多模态数据集介绍

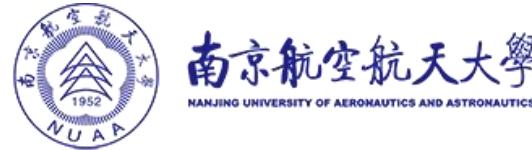
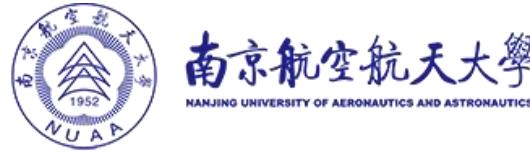


表10.1 常用多模态数据集

数据集	年份	模态			描述
		图像	文本	点云	
GQA	2019年	✓	✓		大规模视觉推理和问答数据集
HowTo100M	2019年	✓	✓		大规模叙事视频数据集
Conceptual-12M	2021年	✓	✓		拓展于Conceptual Captions数据集
YT-Temporal-180M	2021年	✓	✓		覆盖多种主题的大规模多样化数据集
WebVid-2M	2021年	✓	✓		大规模视频-文本对数据集
ModelNet	2015年		✓	✓	广泛使用的合成3D模型数据集
ShapeNet	2015年		✓	✓	大规模的合成3D模型数据集
ScanObjectNN	2019年		✓	✓	真实世界中扫描得到的3D模型数据集
KITTI	2012年	✓	✓	✓	广泛使用的真实自动驾驶基准数据集
nuScenes	2019年	✓	✓	✓	标注细致的真实驾驶场景数据集
Waymo	2019年	✓	✓	✓	大规模跨城市的真实自动驾驶数据集

# 多模态数据集介绍



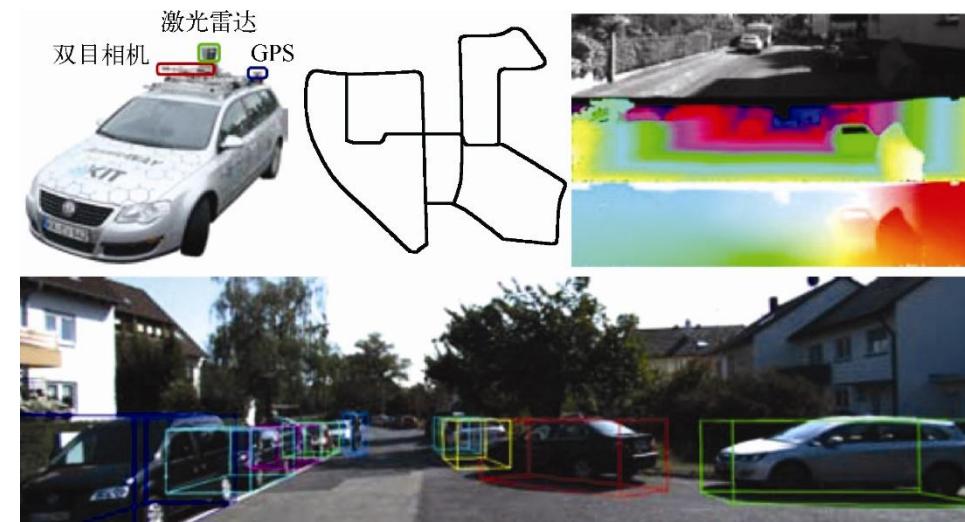
GQA数据集示例



HowTo100M数据集示例



ScanObjectNN数据集示例



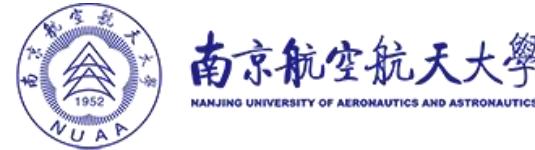
KITTI数据集示例

# 目录



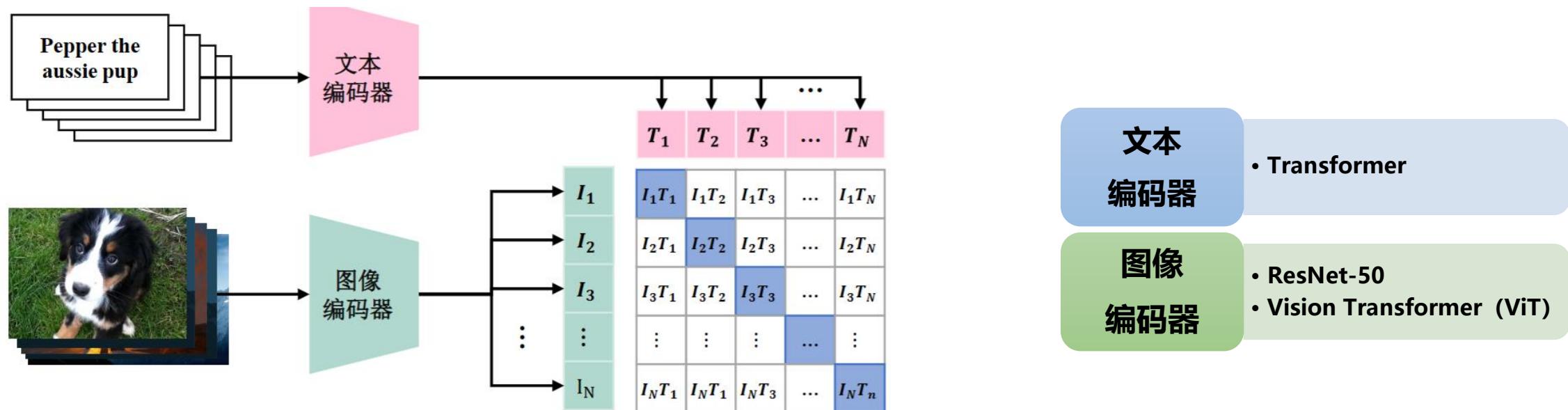
- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V
- ImageBind
- 3D-LLM
- Sora

# CLIP：探索图文结合的奥秘

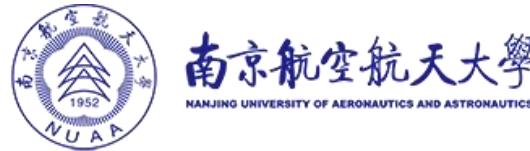


CLIP是OpenAI提出的多模态大模型，**CLIP融合了对语言和图像信息的综合处理**，不仅在图像分类和物体检测等传统视觉任务上取得了领先的性能，而且在自然语言处理等领域也表现出色。

## 对比学习预训练



# CLIP：探索图文结合的奥秘



## 训练过程

$I_1 T_1$	$I_1 T_2$	$I_1 T_3$	...	$I_1 T_N$
$I_2 T_1$	$I_2 T_2$	$I_2 T_3$	...	$I_2 T_N$
$I_3 T_1$	$I_3 T_2$	$I_3 T_3$	...	$I_3 T_N$
:	:	:	...	:
$I_N T_1$	$I_N T_2$	$I_N T_3$	...	$I_N T_n$

$n \times n$  的相似矩阵

**$n$ 个文本-图像对，预测出 $n^2$ 个可能得文本图像对的相似度**

- **$n$ 个正样本**
- **$n^2 - n$ 个负样本**

伪代码

```
# image_encoder - ResNet或者Vision Transformer
# text_encoder - CBOW或者Text Transformer
# I[n, h, w, c] - 对齐图像的minibatch
# T[n, l] - 对齐文本的minibatch
# W_i[d_i, d_e] - 图像特征
# W_t[d_t, d_e] - 文本特征
# t - 可学习温度参数

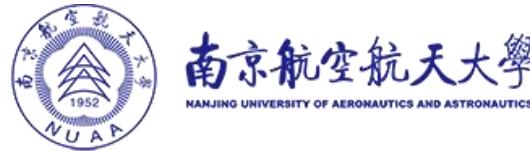
# 分别提取图像特征和文本特征
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) #[n, d_t]

# 对两个特征进行线性投射，得到相同维度的特征，并进行l2归一化
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

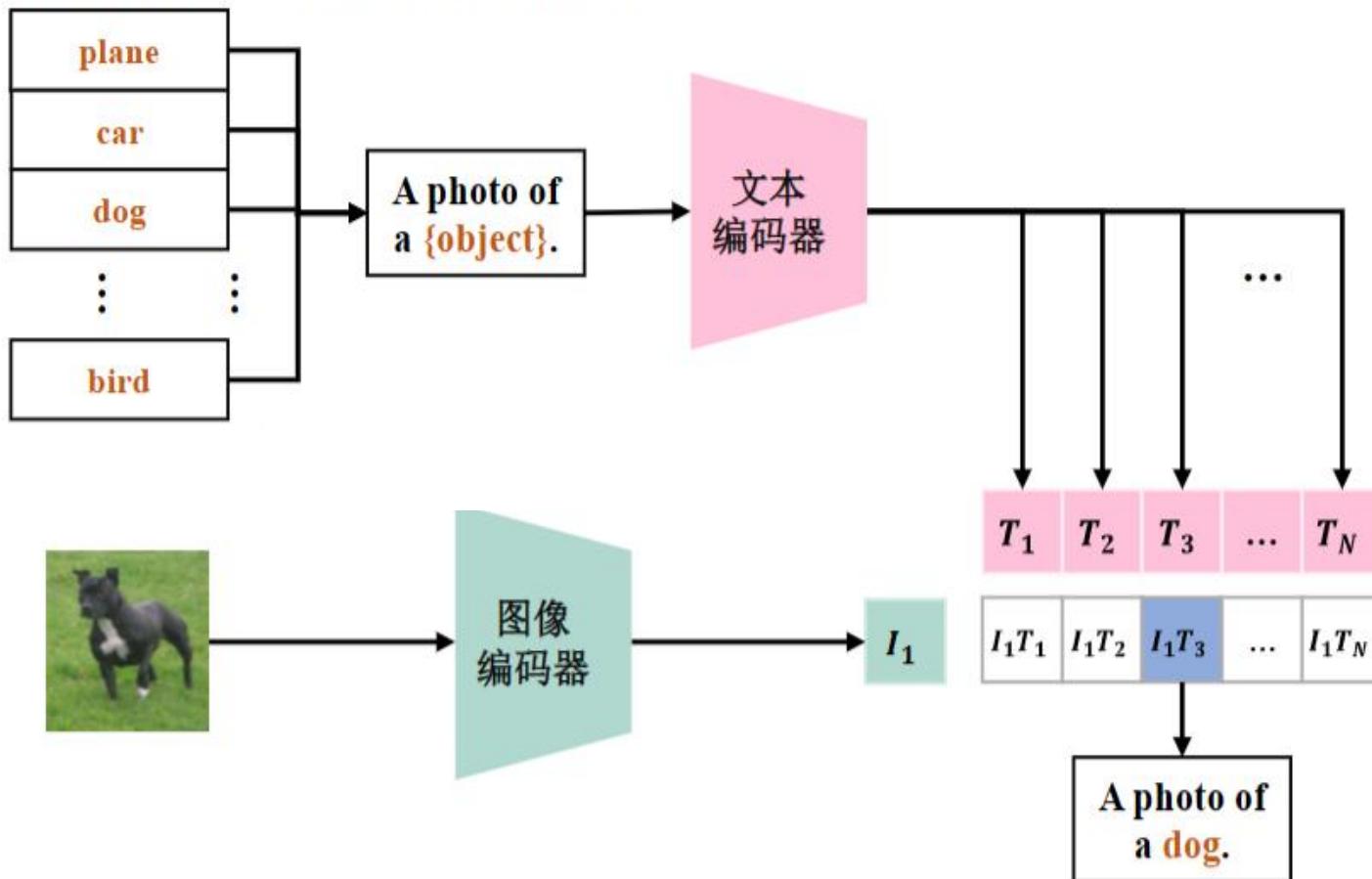
# 计算缩放的余弦相似度：[n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# 对称的对比学习损失：等价于N个类别的cross_entropy_loss
labels = np.arange(n) # 对角线元素的labels
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

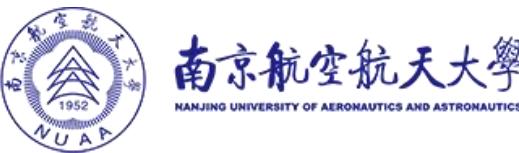
# CLIP：探索图文结合的奥秘



## CLIP的零样本预测能力



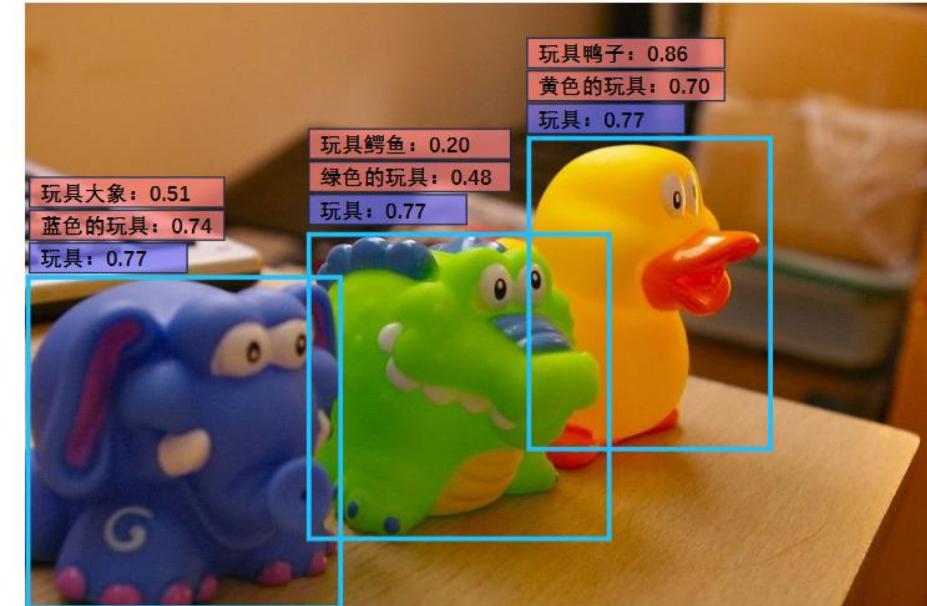
# CLIP：探索图文结合的奥秘



## 相关应用

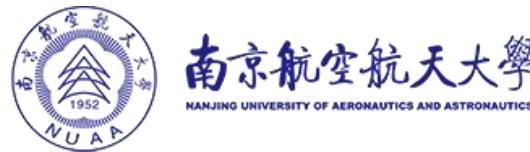


图像编辑



开放词汇目标检测

# CLIP：探索图文结合的奥秘



## 相关应用



Salvador Dalí的生动肖像绘画，其中有一半是机械化的面部。



一个戴着贝雷帽和黑色高领毛衣的shiba inu



一个长着叶子的棕榈树的特写



一种用人类灵魂制作咖啡的浓缩咖啡机，艺术站



土星上一只穿着太空服的海豚，艺术站



一张宣传海报，画的是一只猫打扮成法国皇帝拿破仑，手里拿着一块奶酪

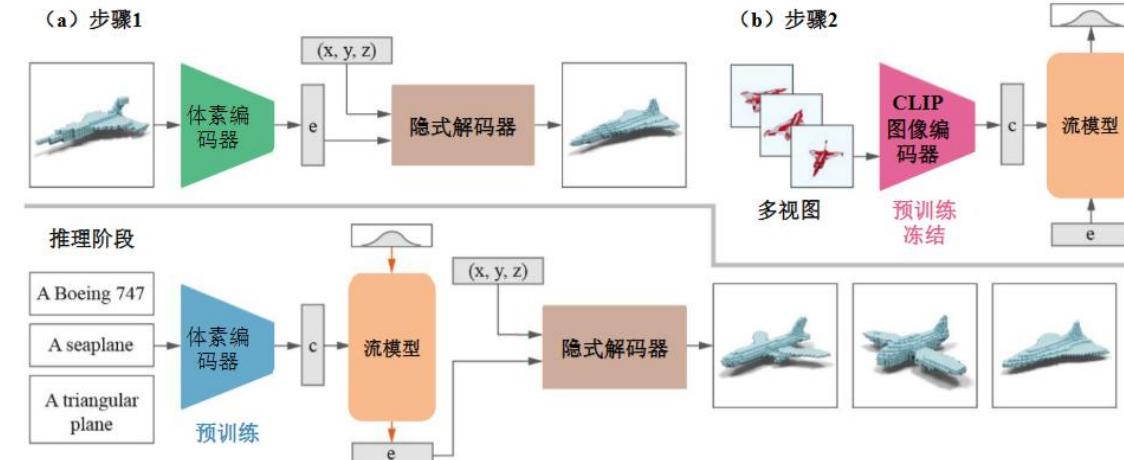


时代广场滑板上的泰迪熊



熊猫疯狂科学家混合闪闪发光的化学物质，艺术站

## 文本生成图像



## 文本生成三维形状

# 目录



- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V**
- ImageBind
- 3D-LLM
- Sora

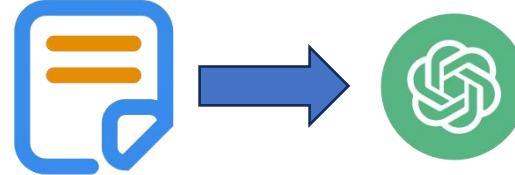
# GPT-4V：大模型视觉能力新篇章



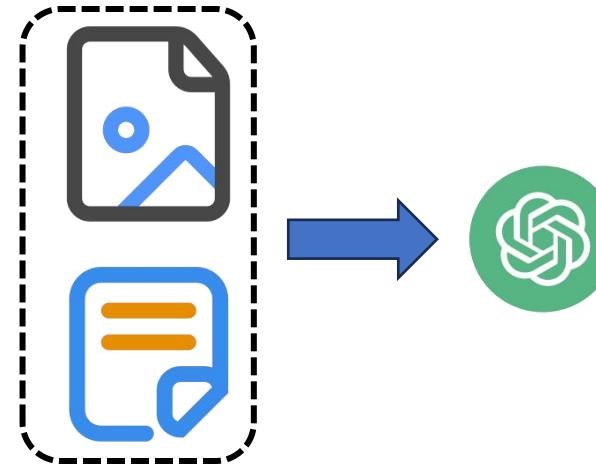
南京航空航天大學  
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

GPT-4V是OpenAI推出的多模态语言模型，它的主要特点是能够处理**图像和文本输入**。这使得它可以理解并生成与图像相关的文本描述，回答视觉相关的问题，甚至在更复杂的多模态任务中表现出色。在本节中，我们将重点介绍GPT-4V 的输入模式及其所具备的部分视觉语言能力。

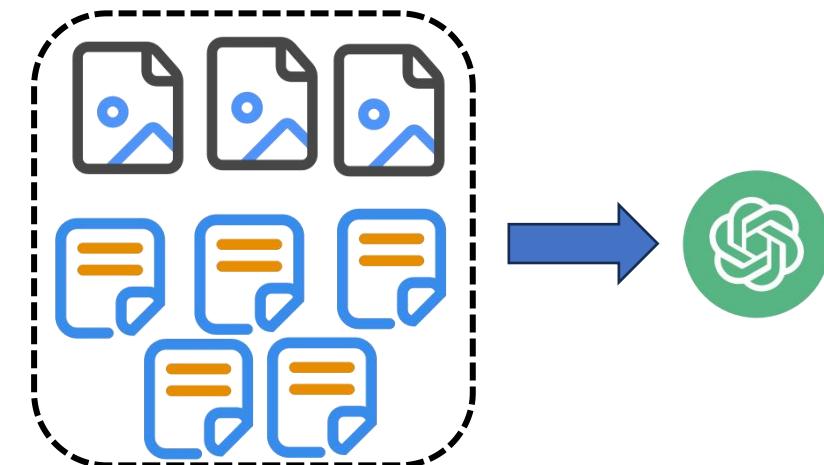
GPT-4V能够处理三种主要类型的输入：



纯文本输入

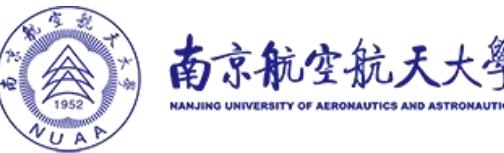


单一图像-文本对输入



交错图像-文本对输入

# GPT-4V：大模型视觉能力新篇章



遵循文本指令

工作方式

视觉指向和视觉参考提示

视觉文本组合提示

上下文少样本示例学习

视觉参考提示

视觉引导和视觉参考提示

(1) 提示词:

描述图中所框选的区域。



GPT-4V回答:

图像中指向的区域是一排挂在电线上的灯。这些灯小而圆，它们沿着电线均匀分布。电线被拉伸在两根杆子之间，位于一个带有桌子和伞的庭院区域上方。这些灯看起来是装饰性的，可能用于在傍晚提供环境照明。

(2) 提示词:

描述图中所框选的区域。

Method	Validation set									Test set									Test set								
	in.			near.			out.			overall			in.			near.			out.			overall					
	C	S	-	C	S	-	C	S	-	C	S	-	C	S	-	C	S	-	C	S	-	C	S	-			
OSCAR	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4	84.8	12.1	82.1	11.5	73.8	9.7	80.9	11.3	84.4	11.3	84.4	11.3	84.4	11.3	84.4	11.3			
Human	84.4	14.3	85.0	12.8	85.7	14.0	87.1	14.2	86.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6	84.4	11.3	84.4	11.3	84.4	11.3	84.4	11.3			
VIV	80.2	13.7	12.9	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8	12.8		
ViuVL	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	78.0	11.5	92.5	13.1	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	103.7	13.7	
UFO	103.9	14.5	95.5	13.8	83.8	12.3	94.3	13.6	98.9	14.3	94.7	13.9	77.9	12.1	92.3	13.6	103.9	14.5	95.5	13.8	83.8	12.3	94.3	13.6	103.9	14.5	
mPLUG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
SimVLM	113.7	-	110.9	-	115.2	-	115.2	-	113.7	-	110.9	-	115.2	-	115.2	-	113.7	-	115.2	-	115.2	-	115.2	-	115.2	-	
LEMON	118.0	15.4	116.3	15.3	120.2	14.5	117.3	15.0	112.8	15.2	115.5	15.1	110.1	13.7	113.3	14.9	118.0	15.4	116.3	15.3	120.2	14.5	117.3	15.0	112.8	15.2	
UniversalCap	122.0	10.0	121.5	10.0	120.4	11.4	120.4	15.0	118.9	15.4	120.6	15.0	114.5	14.1	119.3	15.1	122.0	10.0	121.5	10.0	120.4	11.4	120.4	15.0	114.5	14.1	
CoCa	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	
GIT1	100.7	13.8	97.7	13.5	89.6	12.5	96.6	13.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
GIT1a	107.7	14.9	107.8	14.5	102.5	13.7	106.9	14.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		
GIT	129.8	16.3	124.1	16.0	127.1	15.7	125.5	16.0	122.4	16.2	123.9	16.0	122.0	15.7	123.4	15.9	129.8	16.3	124.1	16.0	127.1	15.7	123.4	15.9	122.0	15.7	
GIT2	126.9	16.1	125.8	16.2	130.6	15.8	126.0	16.1	124.2	16.4	125.5	16.1	122.3	15.6	124.8	16.1	126.9	16.1	125.8	16.2	130.6	15.8	126.0	16.1	124.2	16.4	

(2) Prompt:

找出第一列的模式，并用它推断第二列的缺失图形。



例如:



因此，给图2加一条线



将变成一个带线的菱形



这里有三个子图像排列成一个2乘2的矩阵。图像3和图像1之间有什么模式？下一个图像将是什么，以完成这个矩阵？



GPT-4V回答:

图像3和图像1之间的模式是图像3中心有一个点，而图像1中没有。用于完成矩阵的下一个图像将是选项C，即一个中心有点的星星。这是因为模式似乎是每隔一个图像中心有一个点，而前一个图像（图像3）中心有一个点。此外，图像在菱形和星星形状之间交替，所以下一个图像应该是一个星星。

视觉+文本组合提示

多模态指令

(1) 提示词:

找出第一列中的模式，并用它推断出第二列的缺失数字。



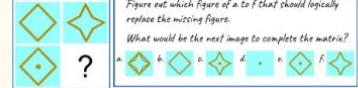
例如，左侧第一列中圈起来的区域显示了通过它添加一条线的模式。因此，答案是一个菱形。



并且带有线穿过，就像下面这样。



根据下面这个图像回答问题



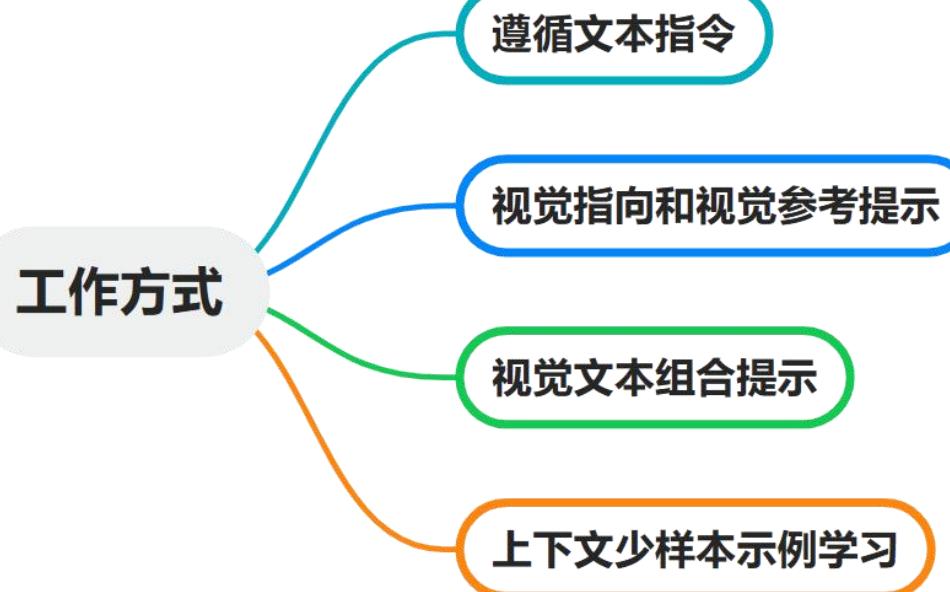
GPT-4V回答:

第一列的模式是在图形中心添加了一个点。因此，第二列缺失的图形将是一个中心有点的星星，即选项C。

# GPT-4V：大模型视觉能力新篇章



南京航空航天大學  
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS



## 上下文少样本学习

### 上下文少样本（1-Shot）

(1) 提示器：  
速度表的读数是多少？



黄色指针位于20和40英里/小时之间的长刻度之间。在20和40之间中间的长刻度是30。速度大约在29英里/小时左右，因为指针接近30英里/小时但还没有完全到达。



### GPT-4V回答：

黄色指针位于40英里/小时和60英里/小时之间的长刻度之间。在40和60之间中间的长刻度是50。速度大约在49英里/小时左右，因为指针接近50英里/小时但还没有完全到达。

### (2) 提示词：

速度表的读数是多少？



黄色指针大致在80英里/小时和100英里/小时之间的中间位置。80和100之间中间位置是90。速度大约在91英里/小时左右，因为指针刚刚超过了90英里/小时。



黄色指针大致位于20英里/小时和40英里/小时之间的中间位置。20和40之间的中间位置是30。速度大约在29英里/小时左右，因为指针接近30英里/小时但还没有完全到达。



### GPT-4V回答：

黄色指针位于0英里/小时和20英里/小时之间的长刻度之间。在0和20之间中间的长刻度是10。速度大约在9英里/小时左右，因为指针接近10英里/小时但还没有完全到达。

## 上下文少样本示例学习

# GPT-4V：大模型视觉能力新篇章



南京航空航天大學  
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

理解和表达视觉信息在人类认知中扮演着至关重要的角色。GPT-4V以其卓越的视觉理解和描述能力，在解读图像信息方面展现出了极强的表现。GPT-4V展现了卓越的视觉语言能力，能够理解和表达复杂的视觉信息。这种能力体现在多个高级视觉任务中：



名人识别



地标识别



食物识别



医学影像分析



场景理解

## GPT-4V图像描述能力在不同领域的应用

### 标志识别与理解

提示词：

描述下面的图像



### GPT-4V回答：

这张图像由三个不同的图标水平排列组成，每个图标之间有一条细的垂直线分隔。左侧是一个图标，类似字母“O”，中间有一条蓝色的水平线穿过。这是汽车品牌宝马（BMW）的标志。中间是由两个重叠的三角形组成的图标，三角形的顶点朝上。这是汽车品牌英菲尼迪（Infiniti）的标志。右侧是一个由盾牌形状组成的图标，有三个箭头指向上方。这是汽车品牌玛莎拉蒂（Maserati）的标志。

### 提示词：

描述一下那些logo的细节

Note: 此处依然使用左图作为查询图像

### GPT-4V:

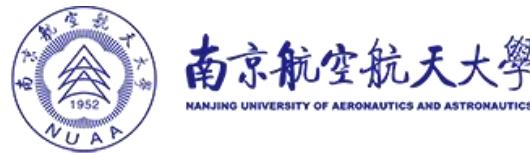
GPT-4V：

最左侧的标志是宝马（BMW）的标志。它是一个圆形图标，填充有黑色。在圆内，有一条蓝色的水平线穿过中间。蓝线上方是白色的，下方是浅蓝色的。这个标志代表飞机旋转的螺旋桨，这是对该公司在航空领域历史的致敬。

中间的标志是英菲尼迪（Infiniti）的标志。它由两个重叠的三角形组成，三角形的顶点朝上。三角形填充了黑色。该标志代表一座山或一条道路，这是对该公司力量、强度和性能理念的致敬。

最右侧的标志是玛莎拉蒂（Maserati）的标志。它由一个类似盾牌的形状组成，有三个红色的箭头指向上方。盾牌形状填充了淡白色，箭头之间的区域是黑色的。这个标志是对该公司在赛车历史上的致敬，箭头代表公司的胜利。

# GPT-4V：大模型视觉能力新篇章



## 高级视觉任务应用



### 空间关系理解



### 对象计数



### 对象定位



#### GPT-4V高级视觉任务的应用

##### 空间关系理解

提示词:

飞盘和人之间的空间关系是什么?



提示词:

这个人比车还要大吗?



GPT-4V回答:

根据提供的图像, 飞盘和人之间的空间关系是飞盘在人的上方, 略微在人的前方。它似乎还在远离人, 就好像是被他扔出去或放开的。

#### GPT-4V高级视觉任务的应用

##### 对象定位

提示词:

使用包围框对图像中的每个人进行定位。图像尺寸为(宽度, 高度) = (998, 768)



GPT-4V回答:

图中只有一个人, 所以包围框的坐标是:

- 左上方点: (254,118)
- 右下方点: (731, 693)

包围框的可视化:



#### GPT-4V高级视觉任务的应用

##### 对象计数

提示词:

图中有多少个苹果?



GPT-4V回答:

图中有7个苹果。

提示词:

图中有几个人?



GPT-4V回答:

这张图中共有4个人。

# 目录

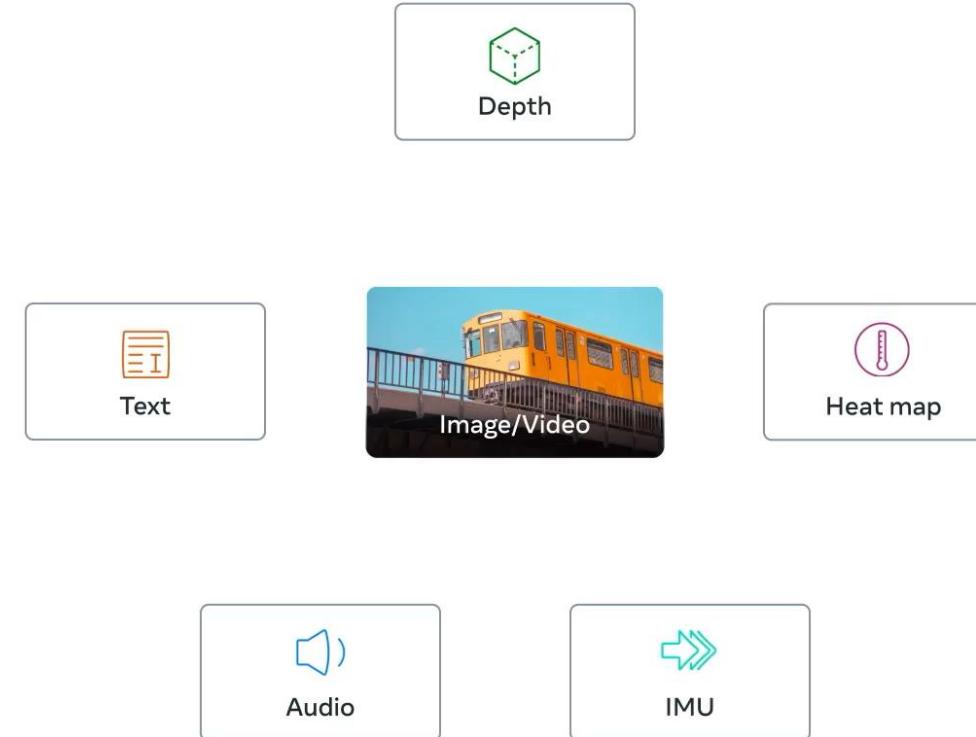


- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V
- ImageBind**
- 3D-LLM
- Sora

- 以CLIP和GPT-4V为代表的多模态大模型在音频、深度、热像等感官领域仍有一些不足。
- 在这一背景下，Meta AI提出了一种多感官统一大模型，即**ImageBind**。

## 六种感官统一到联合嵌入空间

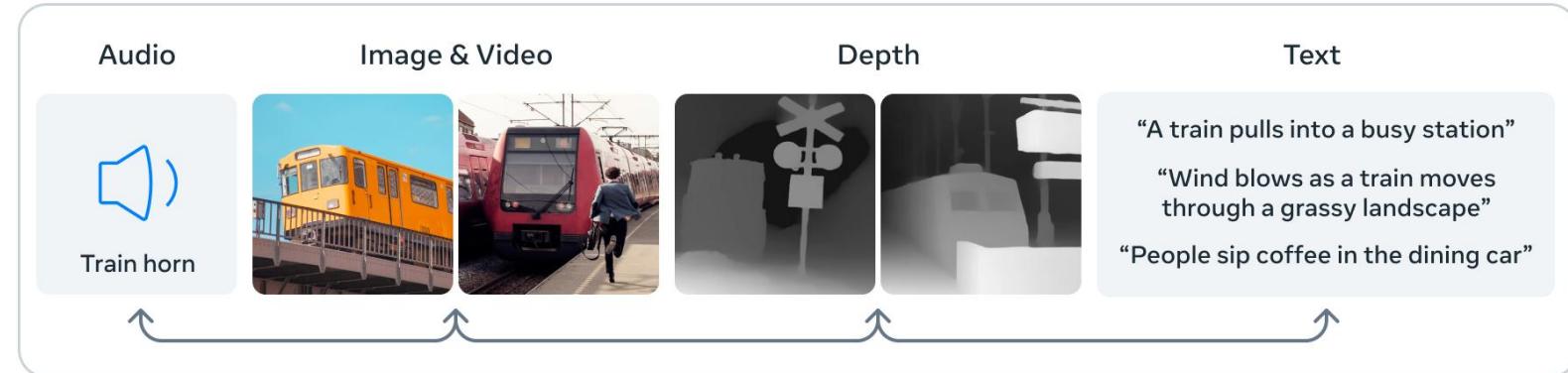
- 图像/视频
- 文本
- 音频
- 深度
- 热像
- IMU



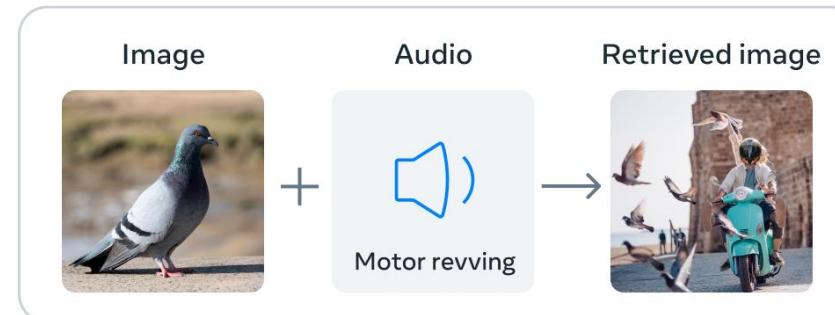
## 相关应用

- 跨模态检索
- 嵌入空间算术
- 音频生成图像
- ...

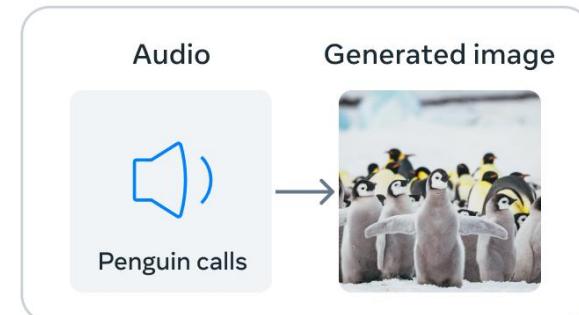
### Cross-modal retrieval



### Embedding-space arithmetic



### Audio to image generation



# ImageBind: 多感官统一



南京航空航天大學  
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

## 应用示例

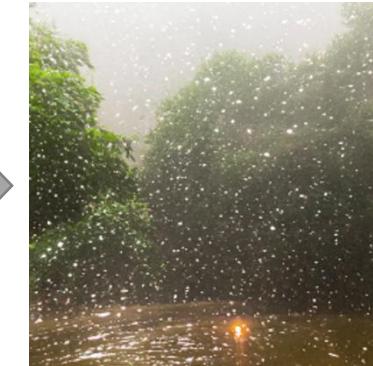
文本->音频&图像

Trains → +



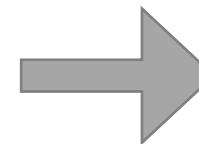
音频->图像

Rain →



音频&图像->图像

Pouring → +



# ImageBind: 多感官统一

## 使用方法 (<https://github.com/facebookresearch/ImageBind.git>)

```
#提取和比较不同模态（如图像、文本和音频）的特征
from imagebind import data
import torch
from imagebind.models import imagebind_model
from imagebind.models.imagebind_model import ModalityType as MT

text_list=["A dog.", "A car", "A bird"]
image_paths=[".assets/dog_image.jpg", ".assets/car_image.jpg", ".assets/bird_image.jpg"]
audio_paths=[".assets/dog_audio.wav", ".assets/car_audio.wav", ".assets/bird_audio.wav"]
device = "cuda:0" if torch.cuda.is_available() else "cpu"
#实例化模型
model = imagebind_model.imagebind_huge(pretrained=True)
model.eval()
model.to(device)
#加载数据
inputs = {
    MT.TEXT: data.load_and_transform_text(text_list, device),
    MT.VISION: data.load_and_transform_vision_data(image_paths, device),
    MT.AUDIO: data.load_and_transform_audio_data(audio_paths, device),
}

#提取特征
with torch.no_grad():
    embeddings = model(inputs)
    #比较不同模态特征的相似性
    print(
        "Vision x Text: ",
        torch.softmax(embeds[MT.VISION] @ embeds[MT.TEXT].T, dim=-1),
    )
    print(
        "Audio x Text: ",
        torch.softmax(embeds[MT.AUDIO] @ embeds[MT.TEXT].T, dim=-1),
    )
    print(
        "Vision x Audio: ",
        torch.softmax(embeds[MT.VISION] @ embeds[MT.AUDIO].T, dim=-1),
    )
```

- 即插即用
- 灵活可拓展
  - 多模态内容创作
  - 零成本模型升级
  - 多模态图像搜索

### #预期输出

```
#图像-文本相似度矩阵
#tensor([[9.9761e-01, 2.3694e-03, 1.8612e-05],
#        [3.3836e-05, 9.9994e-01, 2.4118e-05],
#        [4.7997e-05, 1.3496e-02, 9.8646e-01]])

#音频-文本相似度矩阵
#tensor([[1., 0., 0.],
#        [0., 1., 0.],
#        [0., 0., 1.]])
```

#图像-音频相似度矩阵:

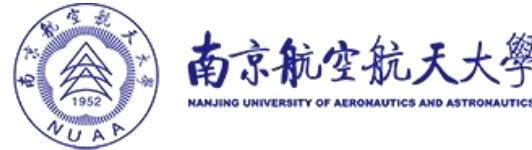
```
#tensor([[0.8070, 0.1088, 0.0842],
#        [0.1036, 0.7884, 0.1079],
#        [0.0018, 0.0022, 0.9960]])
```

# 目录



- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V
- ImageBind
- 3D-LLM**
- Sora

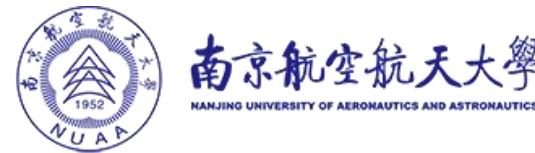
# 3D-LLM: 将3D世界注入大模型



3D-LLM是加州大学、上海交通大学等多所高校的学者联合提出的一种新的三维大模型。这一技术在机器人领域展现出了较大的应用潜力，因为它可以**将三维世界的知识注入到大模型中**，使智能体在三维环境中能够更有效地进行导航、规划和执行任务。



# 3D-LLM: 将3D世界注入大模型



## 三维数据生成

### Box-Demonstration-Instruction based Prompting

#### Bounding-Box Context:

Room 1: table: [0.4, 0.1, 0.1], chair: [0.4, 0.3, 0.4]... Room2: desk: [0.5, 0.2, 0.3]

#### Instruction (Shortened Version):

You are an AI visual assistant / conversation generator / question generator in the room. You need to generate a high-level task and its low-level actions/ 4-10 round conversations between a human and a robot assistant / question-answer pairs about the scene.

#### Demonstration: For Room 1:

<Task Decomposition> High-Level Task: make up. Low-Level Actions: 1. go to the cabinet 2. take out cosmetics from the cabinet 3. bring the cosmetics to the table with mirror 4. open the lamp 5. make up

<3D-Assisted Dialog> Human: Bring me a toothbrush. Robot: Sure! Where is it? Human: Maybe in the sink. Robot: I have arrived at the sink and find no toothbrush. I found an closed bathroom cabinet. Is it possible to contain what you need?...

<QA> Question: Is there a table between the chair and the tv? Answer: Yes. Generate similar responses for Room 2.

#### Response : For Room 2:

<Task Decomposition> High-Level Task: Get ready for work. Low-Level Actions: 1. go to the toilet 2. Wash hands in the sink cabine. 3. Gog to the mirror 4. Brush teeth and wash face 5. Turn on the lamp 6. Leave the room through the door

<3D-Assisted Dialog> Human: Turn on the light. Robot: Sure! Which light? Human: The wall lamp. Robot: I have turned it on. Is it bright enough? Human: No. Are there other lights? Robot: I see a desk lamp.

<QA> Question: Viewing the tv from the desk, is there a guitar on the right? Answer: Yes.

### ChatCaptioner based Prompting

ChatGPT  
Captioner

What can you see from the far right?

Can you tell me what's on the desk?

Aything else close to the desk?

What can you see from the near left?

What's the color of the bed?

### Revision based Prompting

Context: The white chair is near the table.



Instruction: Generate question answering pairs based on the caption.

Response: Question: What color is the chair near the table? Answer: The chair near the table is white.

### □ 基于 Boxes-Demonstration-Instruction 的提示

- 输入三维场景中物体的轴对齐包围框 (Axis-Aligned Bounding Boxes: AABB)，以提供场景的语义和空间位置信息，并通过向 GPT 模型提供具体的特定指令，从而生成多样化的数据。

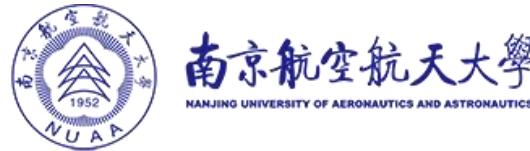
### □ 基于 ChatCaptioner 的提示

- 利用 ChatGPT 输入提示，询问一系列关于图像的信息性问题 (Informative Questions)，随后使用 BLIP-2 模型回答这些问题。

### □ 基于 Revision 的提示

- 该方法可用于将一种类型的三维数据转换为另一种类型。

# 3D-LLM: 将3D世界注入大模型



## □ Objaverse

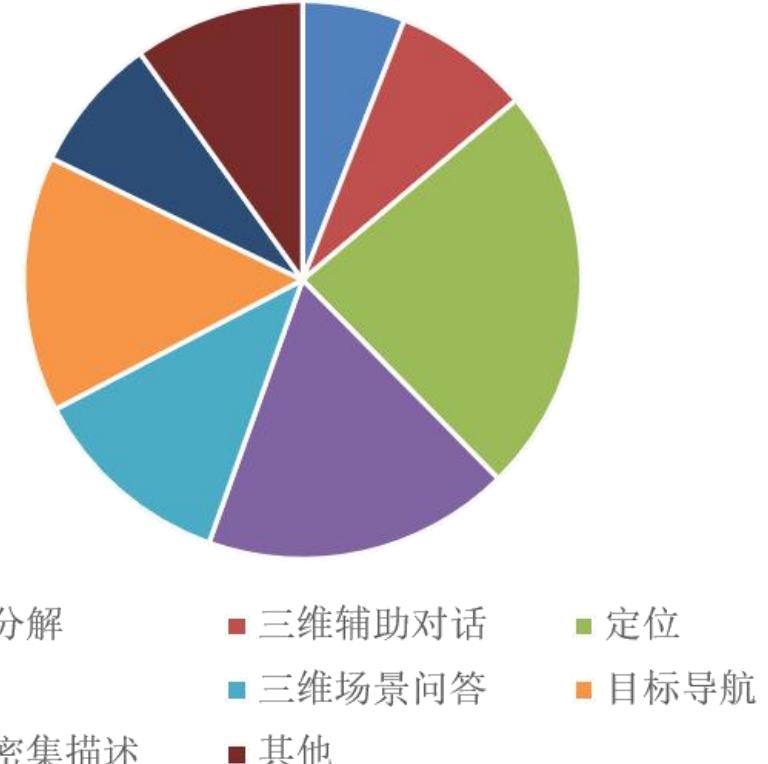
- 包含八十万个三维物体。由于这些物体的语言描述是从在线资源中提取的，未经人工校验，因此大多数描述包含大量噪声，如网址等，或根本无法生成描述。研究者们利用基于 ChatCaptioner 的提示功能，为这些场景生成高质量的三维场景相关描述。

## □ ScanNet

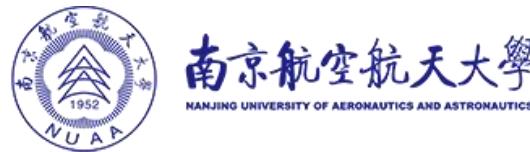
- 包含约一千个三维室内场景的丰富标注数据集，提供了场景中物体的语义和边界框信息。

## □ Habitat-Matterport (HM3D)

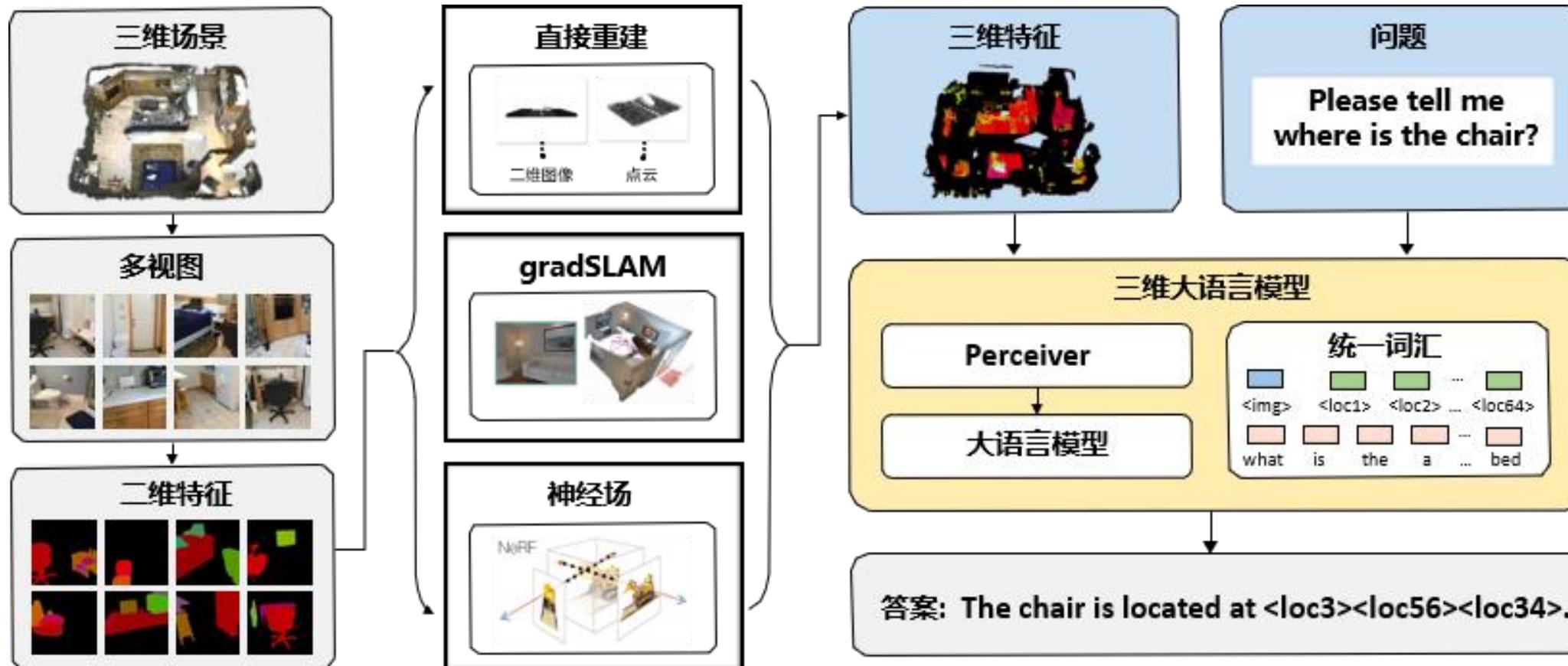
- 一个反映人工智能三维环境的数据集。HM3DSem为 HM3D 的两百多个场景进一步添加了语义注释和边界框。



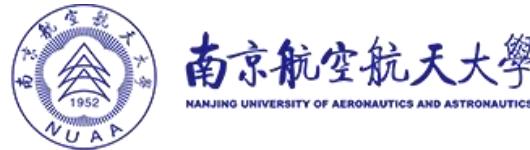
# 3D-LLM: 将3D世界注入大模型



## 3D-LLM训练方式



# 3D-LLM: 将3D世界注入大模型



## 3D-LLM安装与实现

```
# 环境安装
conda create -n lavis python=3.8
conda activate lavis

git clone https://github.com/salesforce/LAVIS.git SalesForce-LAVIS
cd SalesForce-LAVIS
pip install -e .

pip install positional_encodings

# 模型推理
cd 3DLM_BLIP2-base
conda activate lavis

python inference.py # for objects
python inference.py --mode room # for scenes

# 模型微调
cd 3DLM_BLIP2-base
conda activate lavis
# use facebook/opt-2.7b:
python -m torch.distributed.run --nproc_per_node=8 train.py --cfg-path
lavis/projects/blip2/train/3dvqa_ft.yaml
# use flant5
python -m torch.distributed.run --nproc_per_node=8 train.py --cfg-path
lavis/projects/blip2/train/3dvqa_flant5_ft.yaml
```

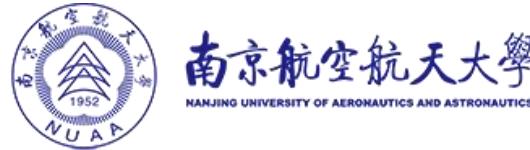
### □ 模型推理

- 下载 objaverse 子集特征
- 下载预训练检查点

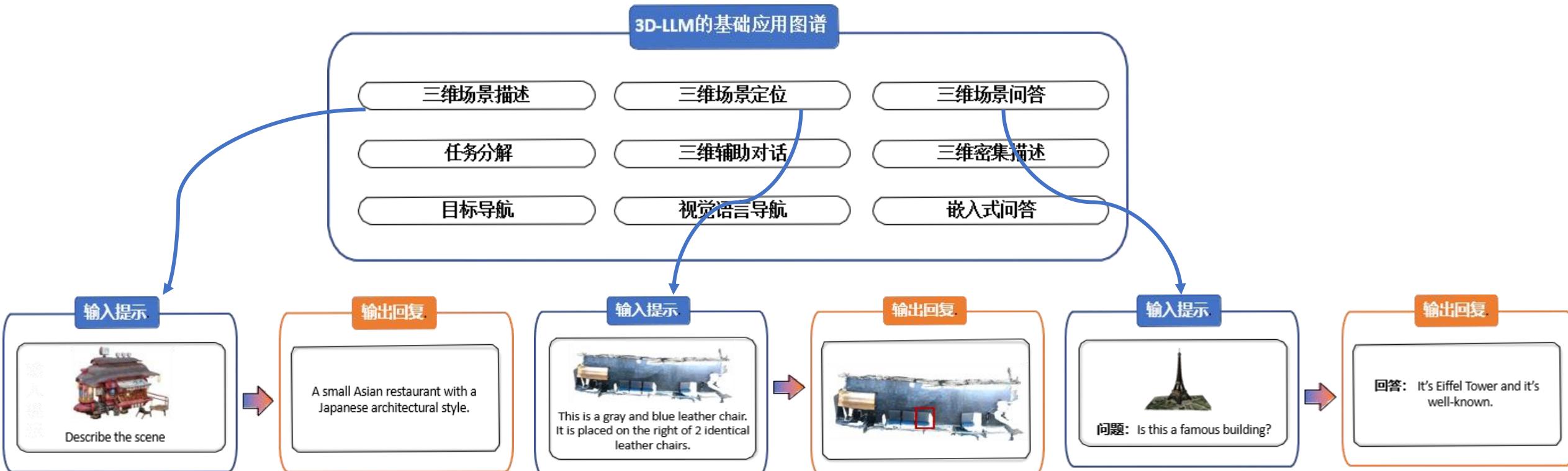
### □ 模型微调

- 下载预训练检查点：修改 yaml 配置文件中的“resume\_checkpoint\_path”路径
- 下载问题：修改 yaml 配置文件中的“annotations”路径
- 下载 scannet 特征或 3dmv-vqa 特征：修改“lavis/datasets/datasets/thirdvqa\_datasets.py”中的路径（train 和 val）

# 3D-LLM: 将3D世界注入大模型



## 3D-LLM应用图谱



# 3D-LLM: 将3D世界注入大模型



## 3D-LLM应用范围c



机器人技术



实体人工智能



智能导览和规划



.....

- 利用 3D-LLM，机器人能够更灵活地感知和理解三维环境，有效执行复杂任务，如导航、物体抓取和环境交互等。
- 3D-LLM 可用于增强实体的感知和认知能力，提高其与环境的交互效果，尤其在虚拟现实和增强现实等场景中有较多潜在应用。
- 应用于智能导览和规划中，3D-LLM 可以帮助系统更好地理解复杂的环境结构，并提供更智能、个性化的导览和规划服务。

# 目录



- 引言和目标
- 多模态数据集介绍
- CLIP
- GPT-4V
- ImageBind
- 3D-LLM
- Sora



# Sora：文生视频，理解世界



南京航空航天大學  
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Sora是OpenAI推出的多模态语言模型，它能够根据给出的提示信息有效生成高质量的、拟真的高清晰度视频。  
它通过分块编码和扩散模型结合方案，赋予了模型更低的运算成本、更高的生成自由度和更拟真的生成效果。



This close-up shot of a chameleon showcases its striking color changing capabilities. The background is blurred, drawing attention to the animal's striking appearance.

(这张变色龙的特写镜头展示了它惊人的变色能力。背景模糊，吸引人们注意这只动物引人注目的外观)



The camera rotates around a large stack of vintage televisions all showing different programs — 1950s sci-fi movies, horror movies, news, static, a 1970s sitcom, etc, set inside a large New York museum gallery.

(镜头围绕着一大堆老式电视机旋转，播放着不同的节目——20世纪50年代的科幻电影、恐怖电影、新闻、静态节目、20世纪70年代的情景喜剧等，背景是纽约的一个大型博物馆画廊。)





# Sora：文生视频，理解世界



南京航空航天大學  
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Sora是基于DiT的**隐空间扩散模型**，与原始版本的扩散模型不同，DiT使用Transformer进行逆扩散过程。

相比U-Net，Transformer架构的参数**可拓展性强**，即随着参数量的增加，Transformer架构的性能提升会更加明显。

在此基础上，Sora使用了**时空感知**模块，使其具有此前基于图像生成的扩散模型进行的精调模型所不具备的超凡的**时空感知能力**。

## 1. 时空深度融合

Sora重新训练了一套能直接压缩视频的自编码器。

相比之前的工作，Sora 的自编码器不仅能提取像素**空间**维度特征，还能提取**时间**维度特征

## 2. 处理任意分辨率、时长的视频

在训练和生成时使用的视频可以是**任何分辨率**、**任何长宽比**、**任何时长**的。这意味着视频训练数据不需要做缩放、裁剪等预处理

## 3. 真正地理解世界

Sora能够更正确认识物体在世界中客观存在的事实，能够**稳定地维持目标**；Sora能够更正确地理解真实世界中目标间**相互作用**；Sora具有一定的进行“**超视觉**”逻辑信息推理的能力

既有问题	现象	Sora	此前方法
运算成本高	架构	扩散模型 (Transformer)	扩散模型 (U-Net)
	视频时长	至多 60 秒	不足 4 秒
	高清分辨率	支持	需要使用升采样等后处理
自由度低	输出分辨率	动态可变	固定分辨率
	扩展生成视频	可生成前导和后续	只能生成后续
	进行视频编辑	支持	部分支持
理解能力不足	三维运动连贯性	强	弱
	物体一致性	强	弱
	物体稳定性	强	弱
	世界交互	强	几乎没有
	“超视觉”逻辑模拟	可以	无
	文本理解	强	一般



# Sora：文生视频，理解世界



南京航空航天大學  
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

## 局限性与争议

OpenAI于2024年2月首次公布了Sora的技术报告，但是截至2024年9月，Sora仍然未进行公开或半公开的测试，仅有极少数团队被允许在小范围内受限地试用Sora。



能力有限



根据提示生成看起来十分拟真的视频无法表明系统能够理解真实世界，通过生成像素来对世界进行建模是一种浪费，并且注定会失败。



理解世界？



效果不达标



李沐  
前Amazon首席科学家  
BosonAI 联合创始人

模型跟前作DiT可能变化不大，但是用了几百倍的算力…目前的技术报告缺失了数据、模型训练和生成1分钟视频是否需要什么新trick来保证质量



仍未完成



Patrick Cederberg  
Shy Kids 团队后期制作  
参与制作Sora短片<Air Head>

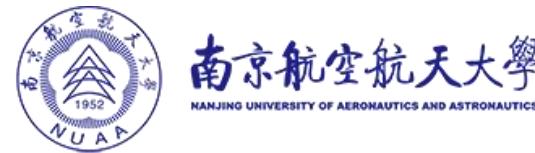
能够尝试使用 SORA 确实非常吸引人…不过，我认为无论是哪种生成式 AI 工具，控制始终是最受期待也最难以实现的部分。

在生成过程中，你可以对动作发生的具体时间点进行一定的控制，但这种控制并不精确，有点像赌博，就像是在玩老虎机，你无法确切知道它是否能在那个时刻达到预期的效果。

实际上就是获取尽可能多的镜头，尽力通过剪辑创造出符合旁白的有趣画面，数百个片段，每个片段长达10至20秒，最终影片中使用的素材大约是原始素材的300倍。



# 可灵：首个投入商业使用的视频生成大模型



Kling（可灵）是快手科技推出的多模态生成模型，包括[文生图](#)、[文生视频](#)和[图生视频](#)，在发布计划中还包括了[视频再编辑](#)等功能。且Kling是已经[公开发布](#)的产品，允许所有申请者试用生成较短的视频，对于付费用户Kling目前可以提供生成长达2分钟的视频，且帧率达到30fps。



**提示词：**一只大熊猫在湖边弹吉他



**提示词：**湖面波光粼粼，云朵在天  
际移动，山峰渐露



**提示词：**小心地将牛奶倒入杯中，  
奶流平稳，杯子渐渐被奶白色填满

谢谢!  
Thanks!

智周万物·道济天下