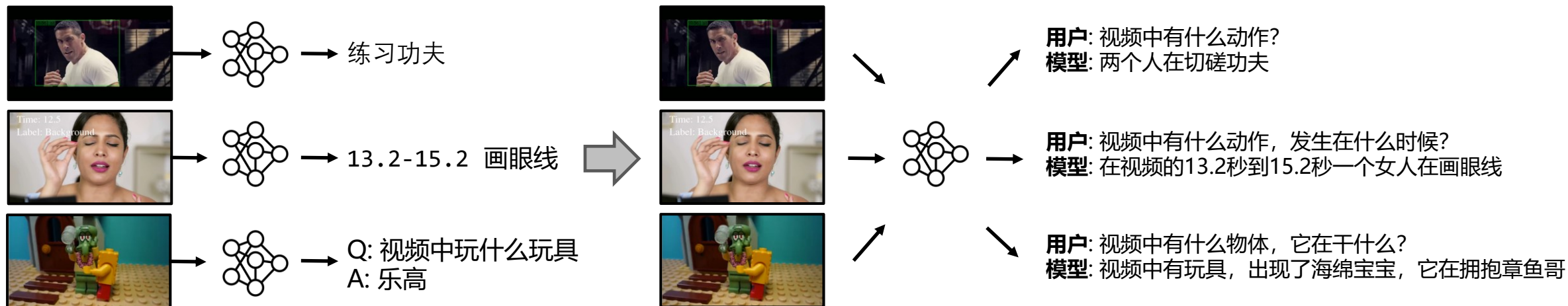


# 多模态视频理解基础模型

汇报人：王利民  
南京大学 & 上海人工智能实验室

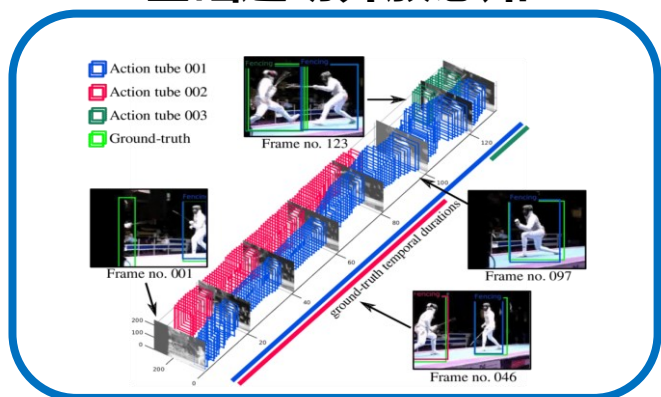


## • 一个基础模型谱系支撑多个应用需求

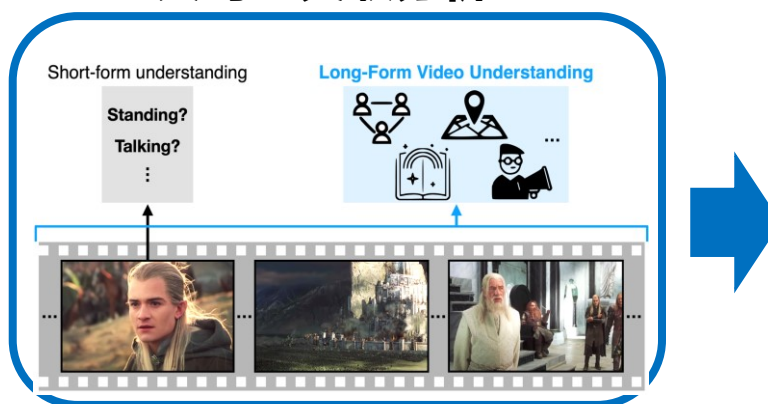


## • 视频理解: 从基本动态感知, 到复杂内容理解

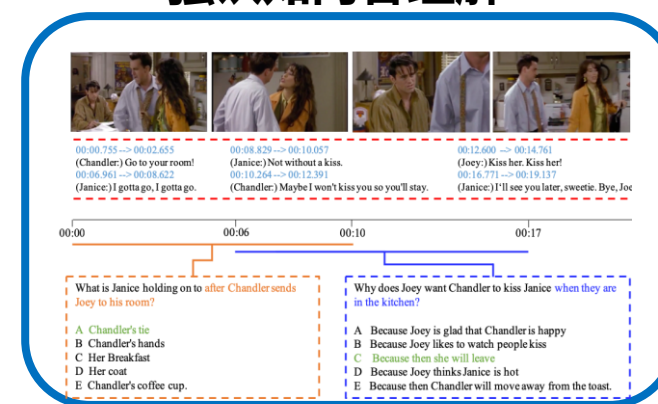
### 基础运动开放感知



### 大时空关联分析



### 强认知内容理解



## 主要研究视频理解的基础模型、解析方法、数据评测

数据与计算平台

开源：应用场景



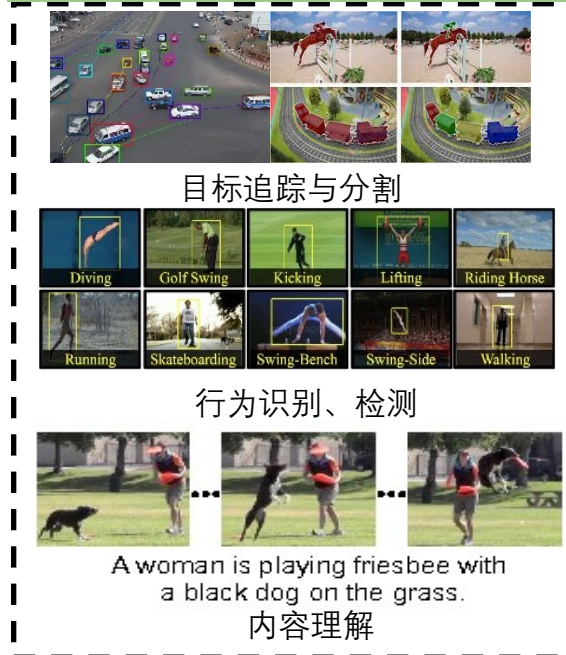
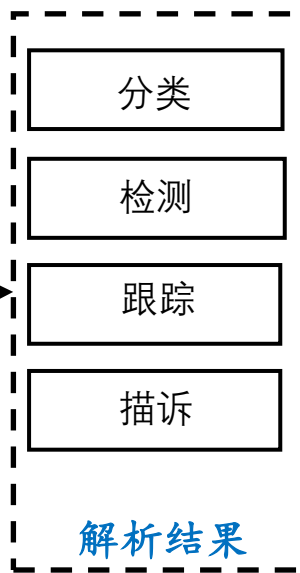
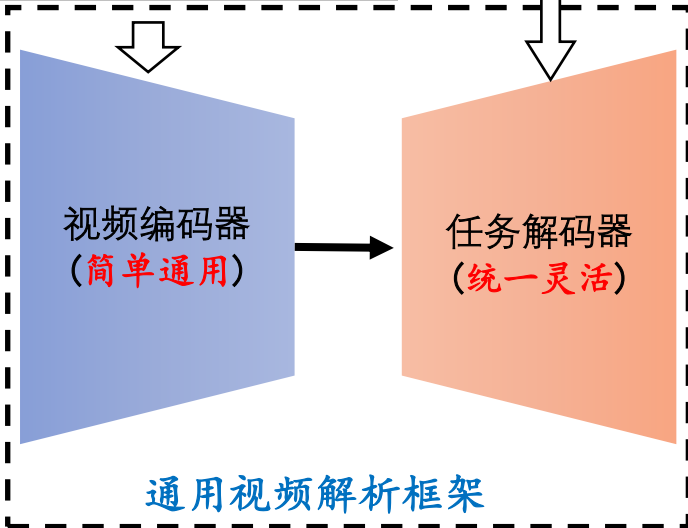
一个基础模型  
一套算法体系

视频模型预训练  
(高效普适)

视频通用理解能力评测体系

场景通用  
模态通用  
任务通用  
能力通用

输入视频

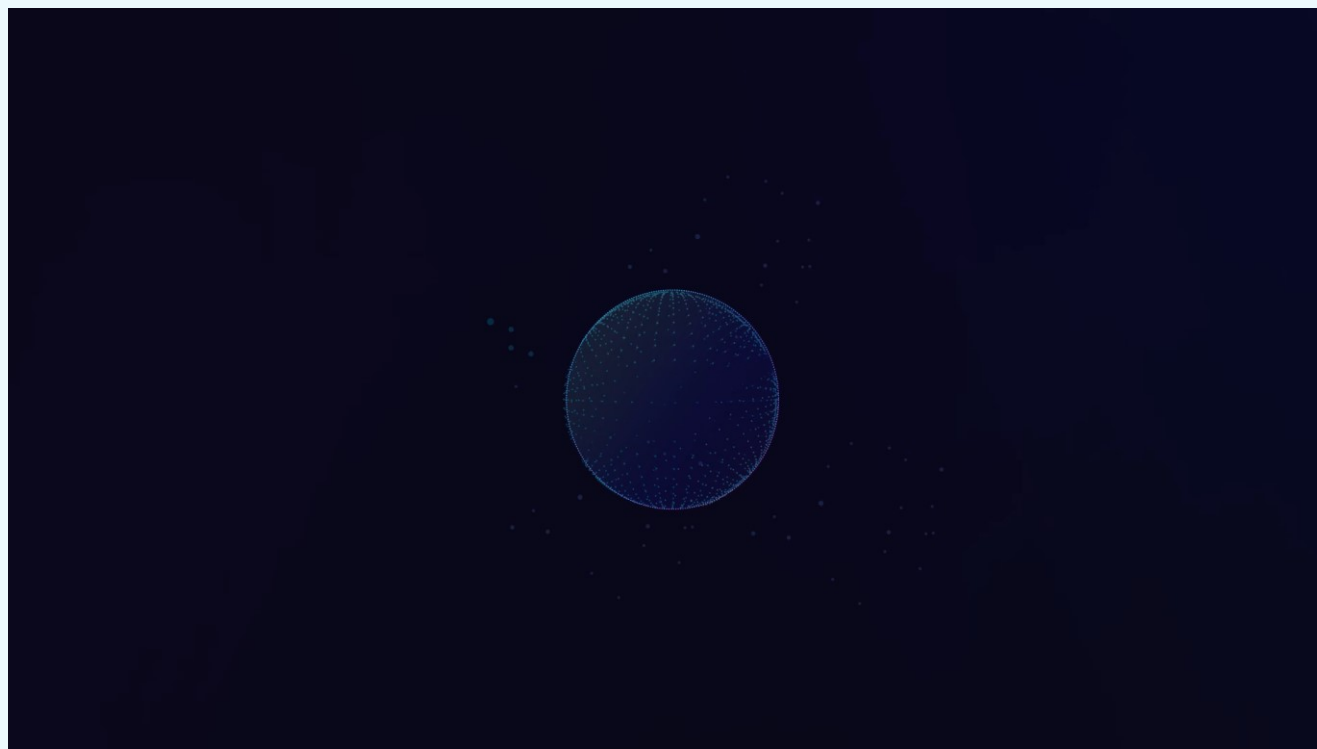


单任务性能突破  
多任务协同演化  
新任务智能涌现

## 书生2.0-通用视频模型：InternVideo

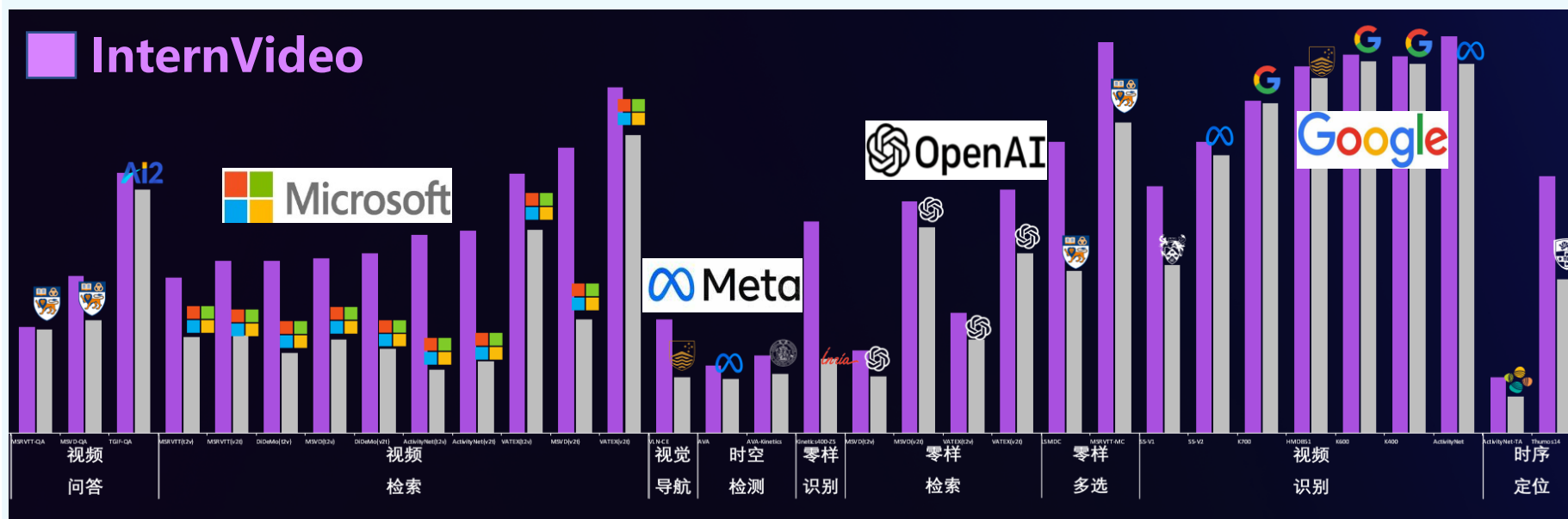
发布于2023世界人工智能大会

- 首个具备**体系化视频感知能力**的通用视频模型
- 全面覆盖**基础视频识别、开放视频感知、时空语义解析**三大任务

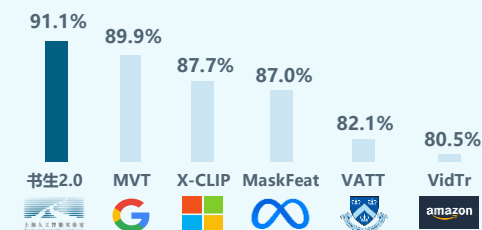


## 书生2.0-通用视频模型：InternVideo

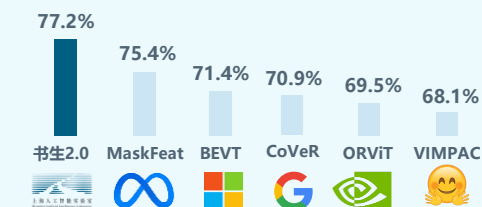
- 在视频识别、视频时序定位、视频检索等39种视频任务上精度世界领先 (截止2022年9月1日)
- 在视频识别标杆基准库Kinetics 400上首次取得超越90%的Top1准确率，达到91%



### Kinetics识别准确率首次> 90%



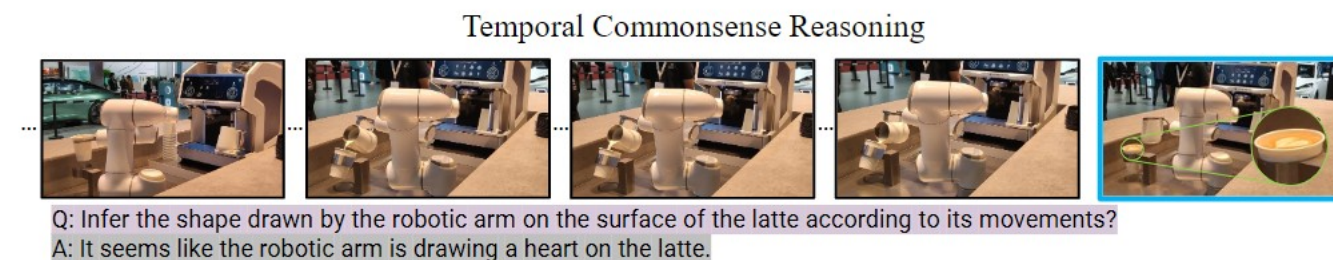
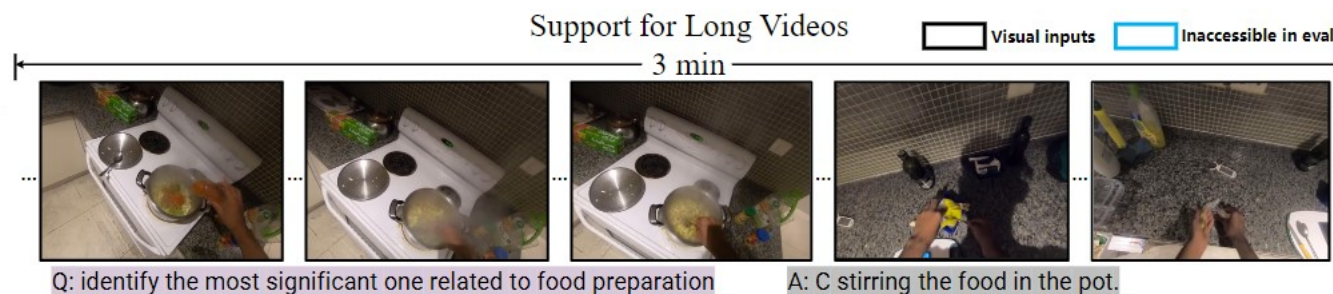
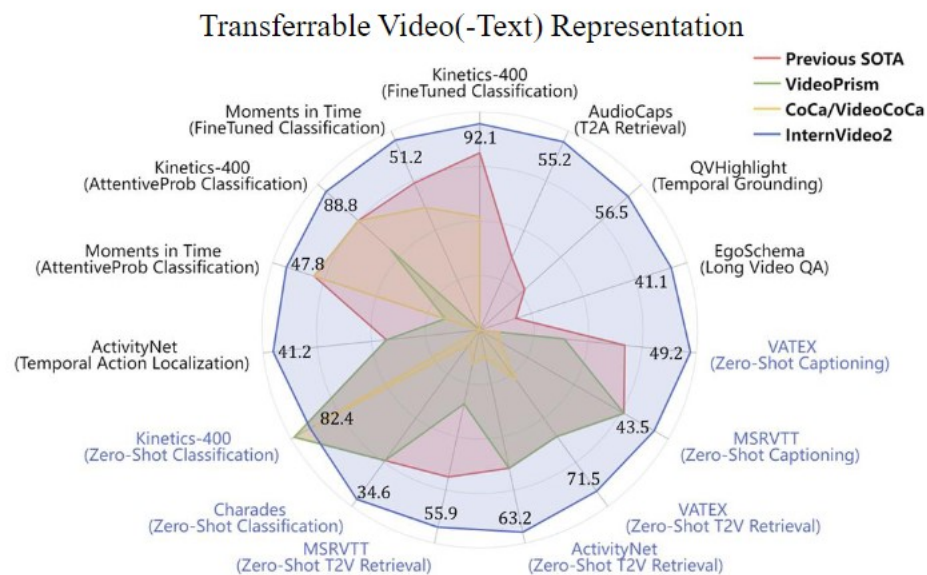
### Something-Something V2



截止2022年9月1日

# InternVideo2视频基础模型 (2024)

- 提出了一种渐进式预训练范式，逐步对齐多粒度语义，形成多任务处理能力
- 在Scaling Law的引导下，扩大训练数据与模型规模：400M数据 & 6B参数量
- 在超过60种的视频理解任务上面取得世界领先水平，涵盖多种视频理解任务



# InternVideo2 视频基础模型 (2024)

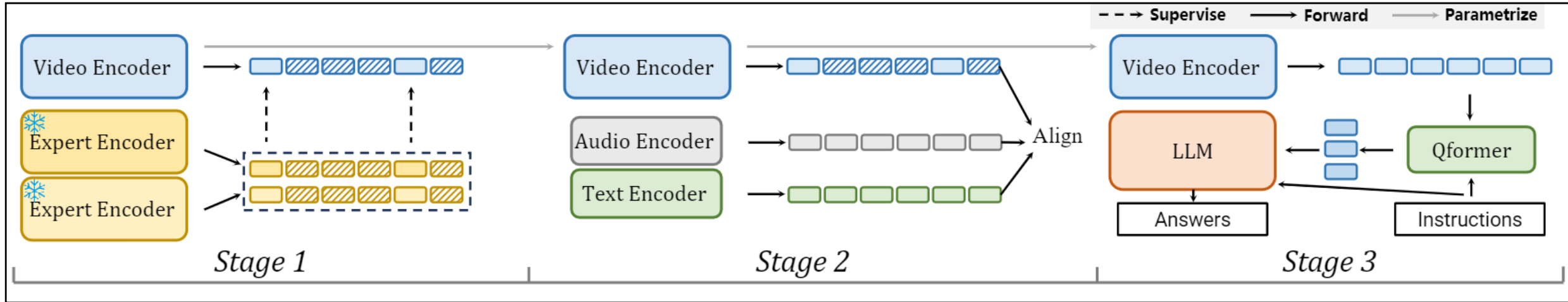
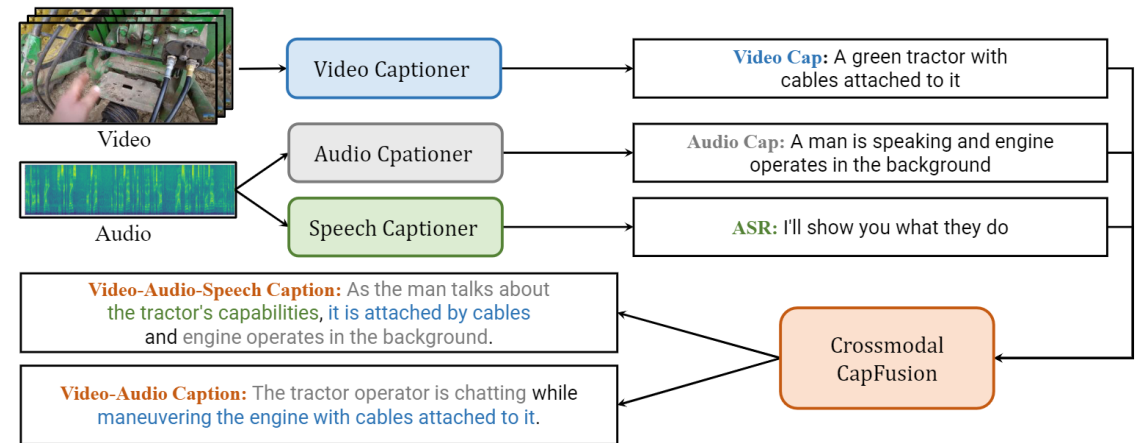
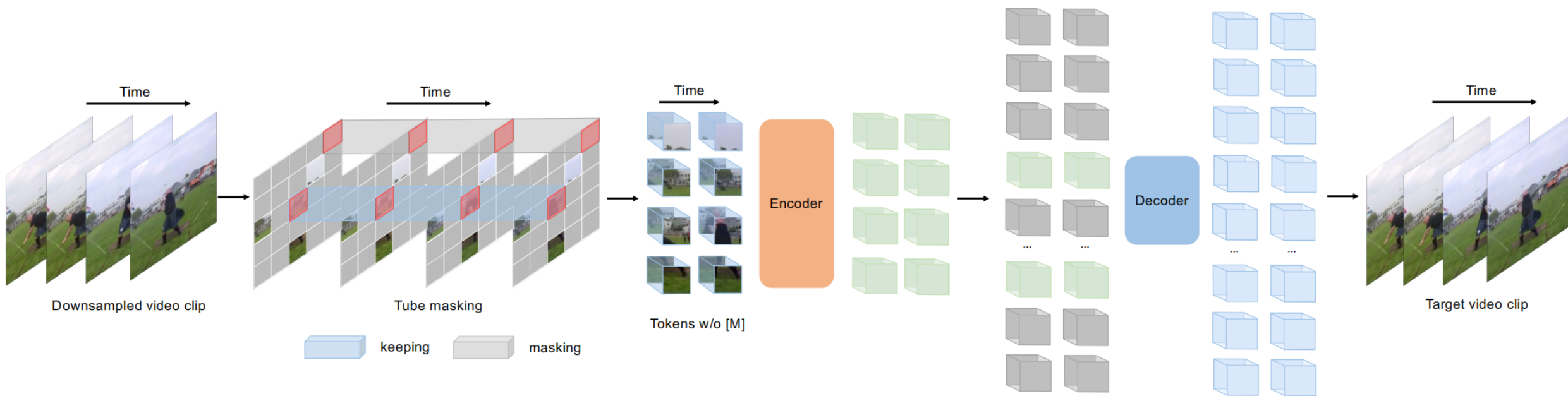


Table 1: Summary of datasets used in **InternVideo2** pretraining process.

Pretraining Stage	Dataset	Domain	# of clips	Annotation
Stage 1	KMash	Web Video	2M	Action Label
Stage 2 (img-txt)	LAION, etc	Web Image	300M	Alt-text / Generated Caps
	WebVid2M	Web Video	250k	Alt-text
	WebVid10M	Web Video	9.7M	Alt-text
	InternVid	Youtube Video	40M	Generated Caption
Stage 2 (vid-txt)	Self-collected	Youtube Video	50M	Generated Caption
	LLaVA, etc	Web Image/Video	2.1M	Conversation, QA



## • 提出了视频掩码自编码器VideoMAE，突破无监督表征学习性能瓶颈



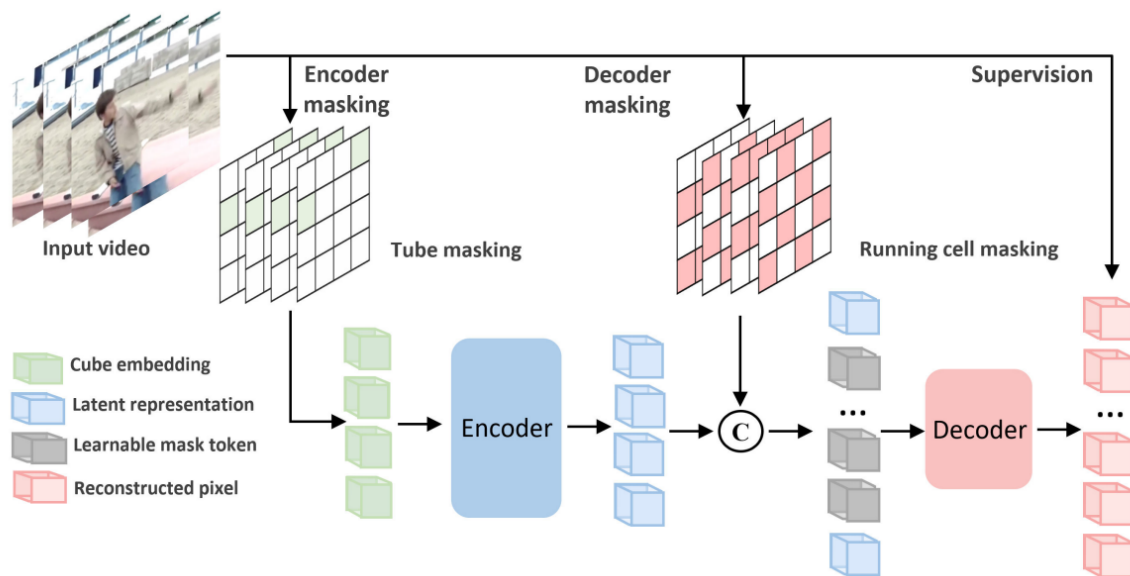
- 提出一种极高掩码率的通道掩码策略，并且验证了数据高效性
- 验证了VideoMAE是无监督视频表征学习性能
- 发现了VideoMAE学习特征具有较强时序鉴别能力

VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training, in NeurIPS 2022 (Spotlight)



# 视频基础模成果二：视频自编码器VideoMAE V2

## • 提出了VideoMAE V2，首次训练出十亿参数量视频基础模型



- 提出一种高效双通道通道掩码策略
- 验证了VideoMAE是可以Scalable & Generable 视频表征学习器
- 在下游10多种任务数据集上面验证基础模型的性能

(a) Kinetics 400				
Method	Top 1	Top 5	Views	TFLOPs
I3D NL [74]	77.7	93.3	10 × 3	10.77
TDN [70]	79.4	94.4	10 × 3	5.94
SlowFast R101-NL [19]	79.8	93.9	10 × 3	7.02
TimeSformer-L [4]	80.7	94.7	1 × 3	7.14
MTV-B (320 <sup>2</sup> ) [82]	82.4	95.2	4 × 3	11.16
Video Swin-L (384 <sup>2</sup> ) [47]	84.9	96.7	10 × 5	105.35
ViViT-L FE [1]	81.7	93.8	1 × 3	11.94
MViTv2-L (312 <sup>2</sup> ) [38]	86.1	97.0	40 × 3	42.42
MaskFeat [76]	87.0	97.4	4 × 3	45.48
MAE-ST [18]	86.8	97.2	4 × 3	25.05
VideoMAE [63]	86.6	97.1	5 × 3	17.88
<b>VideoMAE V2-H</b>	88.6	97.9	5 × 3	17.88
<b>VideoMAE V2-g</b>	88.5	98.1	5 × 3	38.16
<b>VideoMAE V2-g (64 × 266<sup>2</sup>)</b>	<b>90.0</b>	<b>98.4</b>	2 × 3	160.30
<i>Methods using in-house labeled data</i>				
CoVeR (JFT-3B) [85]	87.2	-	1 × 3	-
MTV-H (WTS 280 <sup>2</sup> ) [82]	89.9	98.3	4 × 3	73.57

(b) Kinetics 600				
Method	Top 1	Top 5	Views	TFLOPs
SlowFast R101-NL [19]	81.8	95.1	10 × 3	7.02
TimeSformer-L [4]	82.2	95.6	1 × 3	7.14
MTV-B (320 <sup>2</sup> ) [82]	84.0	96.2	4 × 3	11.16
ViViT-L FE [1]	82.9	94.6	1 × 3	11.94
MViTv2-L (352 <sup>2</sup> ) [38]	87.9	97.9	40 × 3	45.48
MaskFeat [76]	86.4	97.4	1 × 10	3.77
<b>VideoMAE V2-H</b>	88.3	98.1	5 × 3	17.88
<b>VideoMAE V2-g</b>	88.8	98.2	5 × 3	38.16
<b>VideoMAE V2-g (64 × 266<sup>2</sup>)</b>	<b>89.9</b>	<b>98.5</b>	2 × 3	160.30
<i>Methods using in-house labeled data</i>				
CoVeR (JFT-3B) [85]	87.9	97.8	1 × 3	-
MTV-H (WTS 280 <sup>2</sup> ) [82]	90.3	98.5	4 × 3	73.57

(c) Something-Something V2		
Method	Top 1	Top 5
SlowFast [19]	63.1	87.6
TEINet [46]	66.5	-
TEA [37]	65.1	89.9
TDN [70]	69.6	92.2
TimeSformer-L [4]	62.4	-
MFormer-HR [53]	68.1	91.2
ViViT-L FE [1]	65.9	89.9
Video Swin-B [47]	69.6	92.7
MViTv2-B [38]	72.1	93.4
MTV-B [82]	67.6	90.1
BEVT [72]	70.6	-
VIMPAC [60]	68.1	-
UniFormer [35]	71.2	92.8
MaskFeat [76]	75.0	95.0
MAE-ST [18]	75.5	95.0
VideoMAE [63]	75.4	95.2
<b>VideoMAE V2-H</b>	76.8	95.8
<b>VideoMAE V2-g</b>	<b>77.0</b>	<b>95.9</b>

(d) Something-Something V1		
Method	Top 1	Top 5
I3D [7]	41.6	72.2
NL I3D+GCN [75]	46.1	76.8
TSM [40]	49.7	78.5
V4D [87]	50.4	-
TANet [48]	50.6	79.3
TEINet [46]	52.5	-
TEA [37]	51.9	80.3
CorrNet [68]	53.3	-
GSM [58]	55.2	-
TDN [70]	56.8	84.1
UniFormer [35]	61.0	87.6
<b>VideoMAE V2-H</b>	66.6	90.8
<b>VideoMAE V2-g</b>	<b>68.7</b>	<b>91.9</b>

(e) AVA		
Method	Long Feature	mAP
SlowFast [19]	✗	29.0
TubeR [88]	✓	33.4
MaskFeat [76]	✗	38.8
MAE-ST [18]	✗	39.0
VideoMAE [63]	✗	39.5
<b>VideoMAE V2</b>	✗	<b>42.6</b>

(f) AVA Kinetics		
Method	Ensembled	mAP
AIA++ [78]	✓	29.0
MSF [89]	✓	33.4
ACAR [51]	✓	40.5
<b>VideoMAE V2</b>	✗	<b>43.9</b>

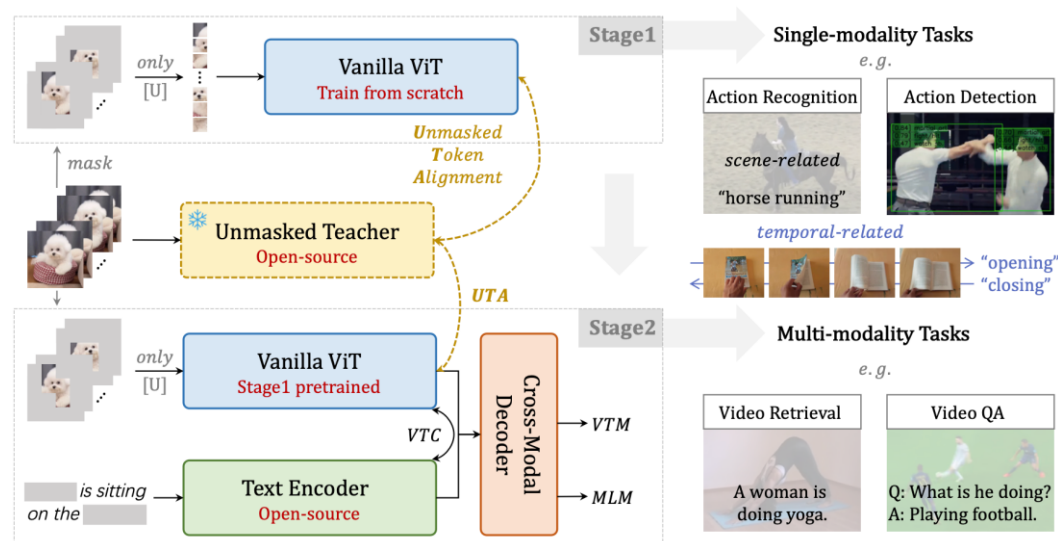
(g) THUMOS14		
Method	Optical Flow	mAP
RTD-Net [61]	✓	43.6
DaoTAD [67]	✗	50.0
AFSD [39]	✓	52.0
DCAN [8]	✓	52.3
TadTR [42]	✓	54.2
TALLFormer [12]	✗	59.2
BasicTAD [83]	✓	59.6
ActionFormer [86]	✓	66.8
<b>VideoMAE V2</b>	✗	<b>69.6</b>

(h) FineAction		
Method	Optical Flow	mAP
BMN [41]	✓	9.25
G-TAD [81]	✓	9.06
BasicTAD [83]	✗	12.2
ActionFormer [86]	✗	13.2
<b>VideoMAE V2</b>	✗	<b>18.2</b>

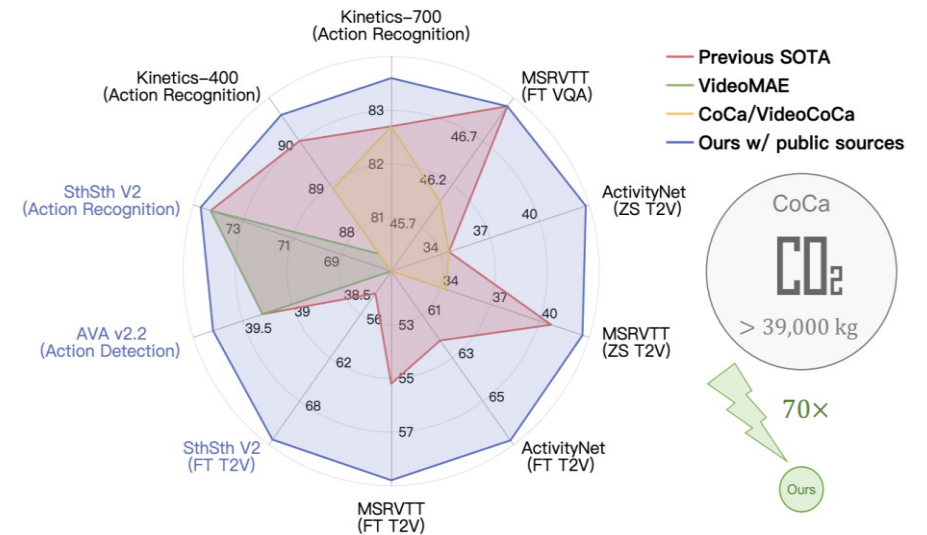
VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking, in CVPR 2023, 被引用128次

## 两阶段训练范式 (支持单模态与多模态任务)



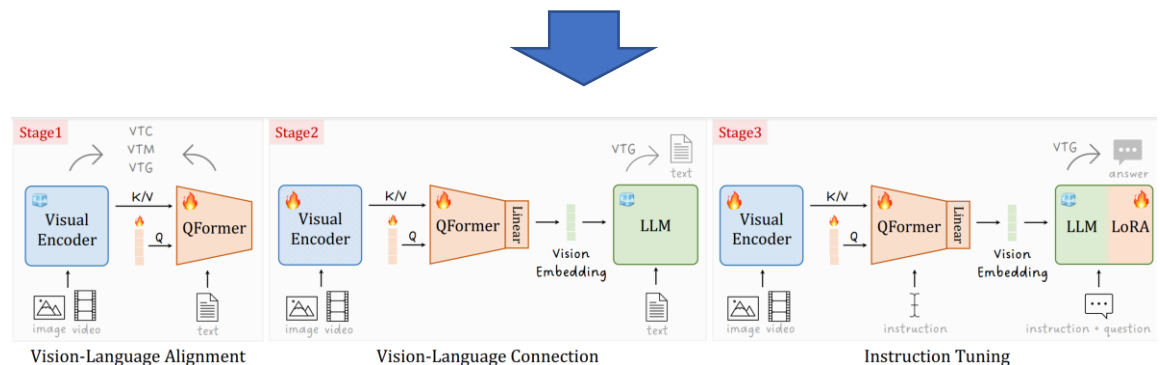
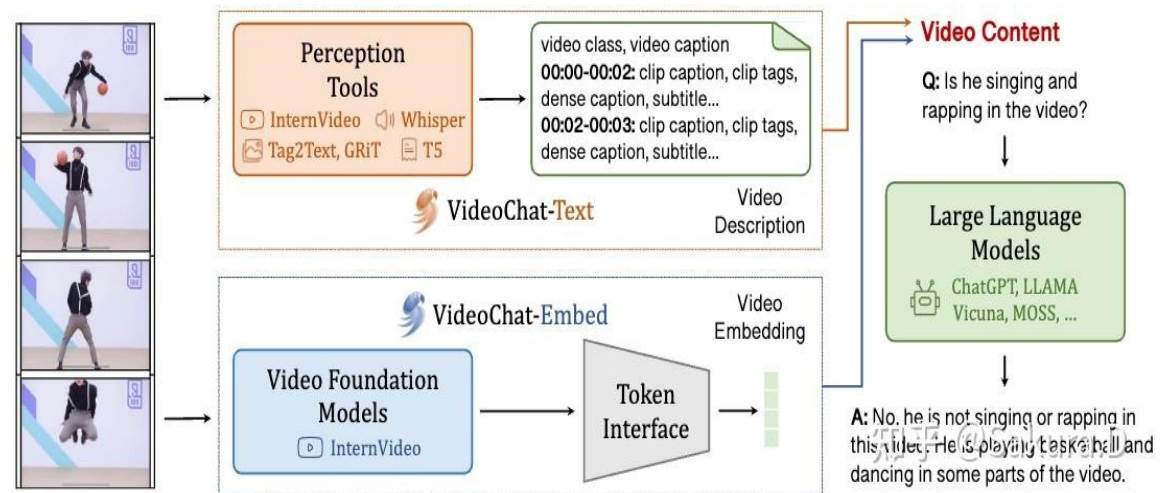
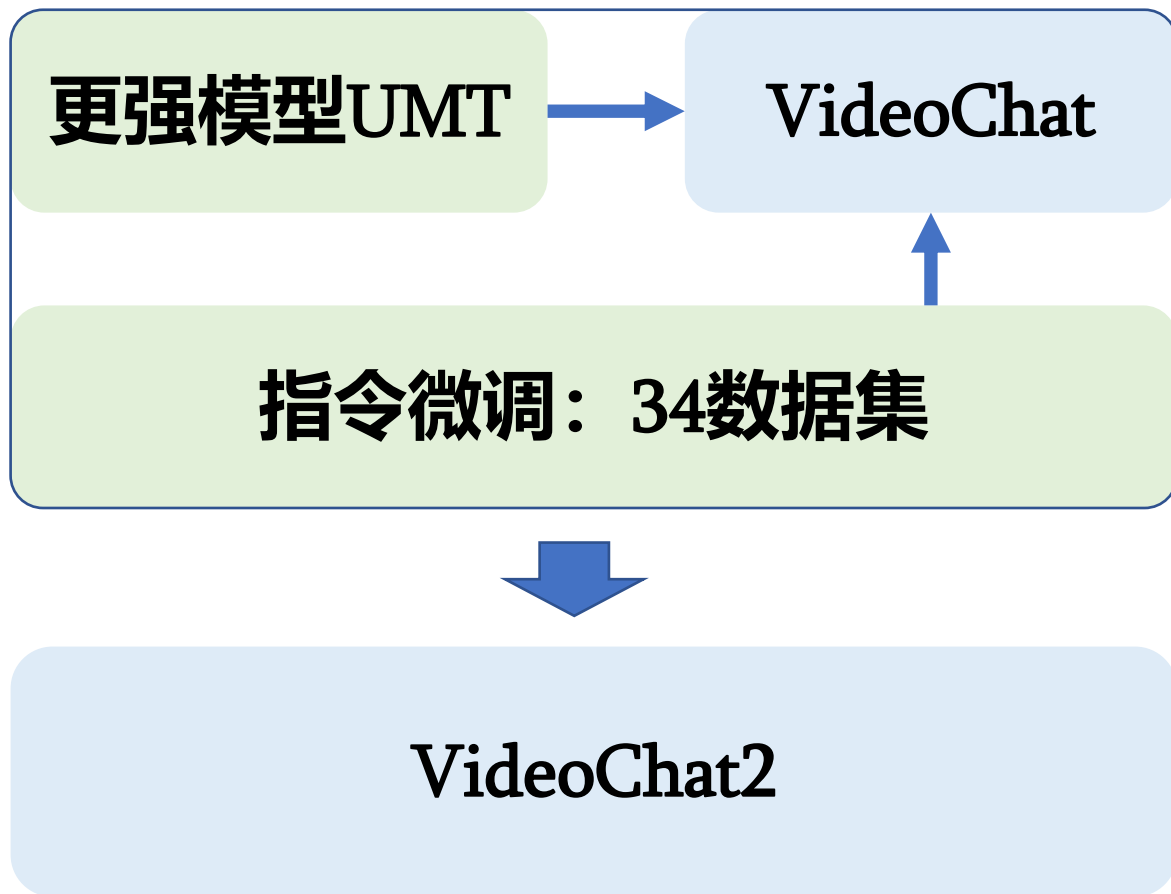
## UMT vs 谷歌CoCa

(节省近70倍碳排放)



# 视频基础模成果四：多模态视频对话模型VideoChat

## 进化成**时序理解能力最强**的视频对话系统：**20种**时序问答任务最佳



## 诗词问答



输入视频

画面描述的是哪句诗词?



这段视频中所描绘的景象与唐代诗人李白的《望庐山瀑布》非常相似。诗中写道：“日照香炉生紫烟，遥看瀑布挂前川。飞流直下三千尺，疑是银河落九天。”这句诗表现了瀑布的壮丽和气势磅礴，同时也展示了作者对自然美景的赞美和敬畏之情。

你是怎么判断的?



作为一个拥有强大视频理解能力的AI助手，我可以分析和解读视频中的各个元素，并将其与已有的知识库进行匹配和比对。在这个例子中，我分析视频中的场景和细节，包括山峰、瀑布、水流、峡谷等元素，结合诗句的意义和语境，最终确定这句话正是对这幅画的恰当描述。

如果我是《中国诗词大会》的编导，请按步骤给出这个问答环节脚本



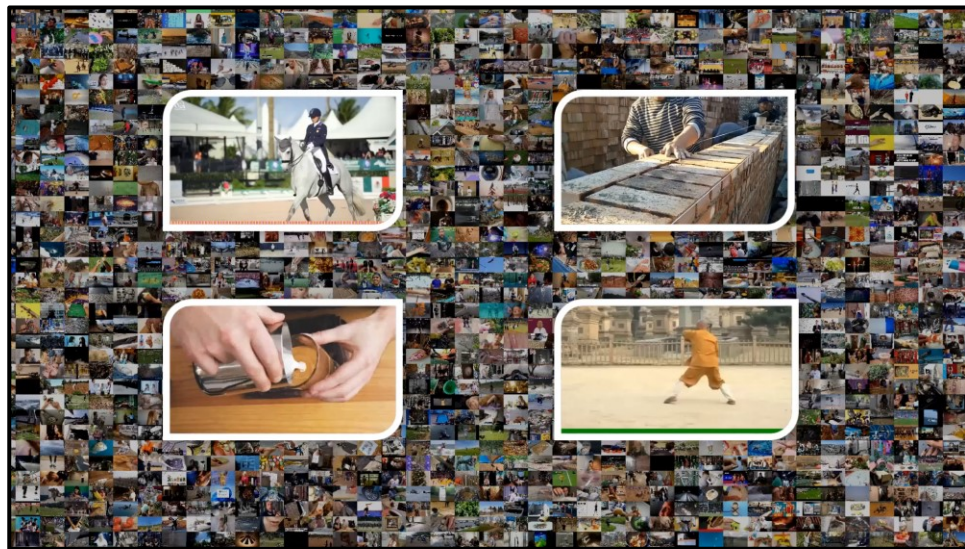
好的，以下是一份可能适合该节目的答题环节脚本：

- 1.主持人介绍游戏规则并引入题目背景(例如：今天我们将欣赏到一幅美丽的画卷，它描绘了一条巨大瀑布从山顶飞流直下，气势磅礴，令人惊叹。)
- 2.播放视频或展示图片，让选手观察几分钟。
- 3.提问选手：这幅画面让你想起了哪首经典的唐诗？
- 4.给予选手几秒钟思考时间，然后邀请他/她作答。
- 5.判断选手的答案是否正确，并给出相应的奖励分数。
- 6.解释诗句的含义及与画面的联系，以便观众更好地理解 and 欣赏。
- 7.结束本次问答环节，进入下一轮比赛。

# 视频数据基准成果一：大规模数据集InternVid

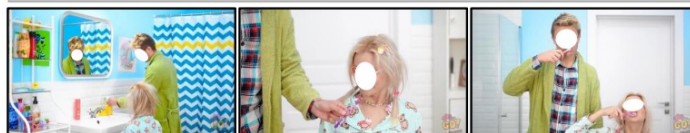
## • 建立了大规模以视频为中心的多模态数据集：**生成和理解**

- 2.34亿秒级短视频，总时长超76万小时，文字描述超41亿单词
- 11个国家，20个主题，6000余种人类日常行为活动



T a woman in a red wagon with a child in it in a store as she walks through the store.

drive a car out.



T a man and a woman brushing their teeth in the bathroom.

分かってるわよ待ってティッシュはどこシャワーを買って閃いた  
まずはこんな風に居るの (I understand, just wait. Where are the  
tissues? I had an idea to buy a shower, but for now I'll stay like this.)



T women dressed in traditional costumes standing in a room.

[музыка] [аплодисменты]  
([Music] [Applause])



T older women sitting at a table eating a bowl of food with chopsticks.

老祖婆就是这点好，不挑食。啥子都吃得。老祖婆坐车哪里都能去。  
(The old granny is good in this aspect, she is not picky and can eat  
anything. She can go anywhere by car.)

Huggingface 下载：724次， Berkeley, NAVER, Meta, Allen Institute for AI, Google...

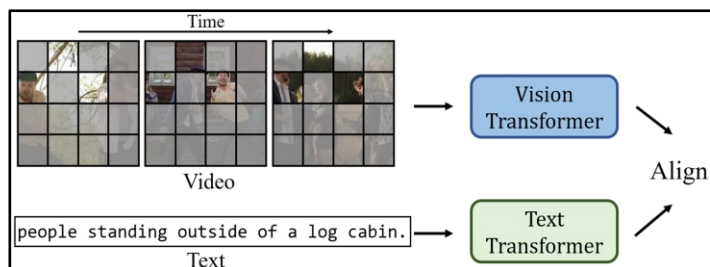
Opendatalab 浏览量：上架30天，浏览2.4K，下载219次

# 视频数据与基准成果一：大规模数据集InternVid

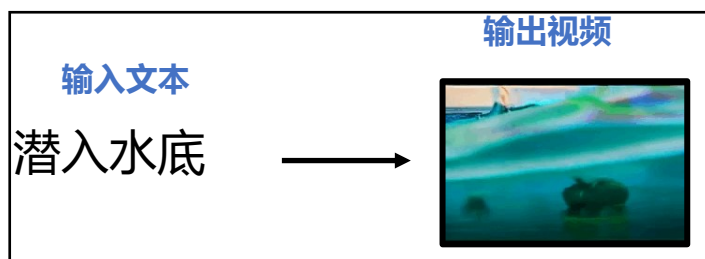
## • 建立了大规模以视频为中心的多模态数据集：生成和理解

支撑丰富的生成及感知任务

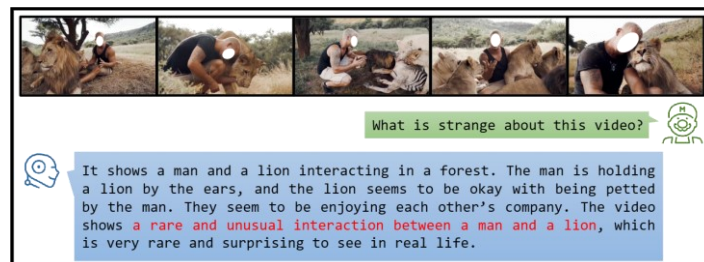
可迁移多模态表征



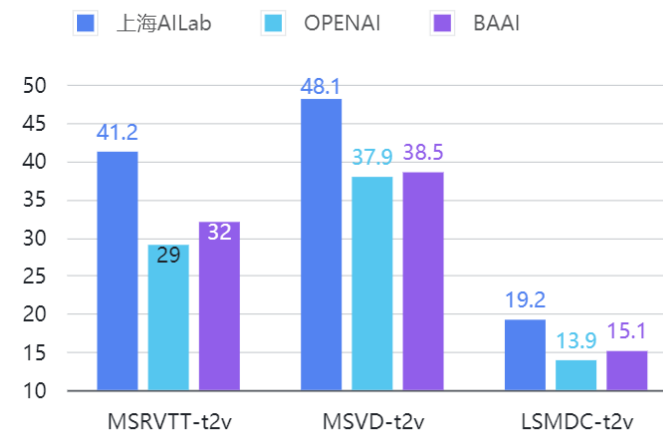
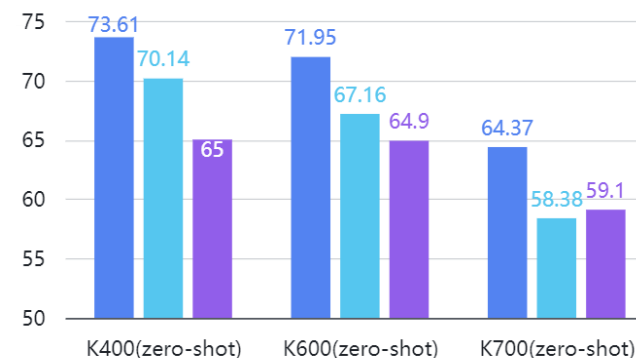
多模态内容生成



多模态对话系统



可迁移视频表征性能世界一流



- InternVid 数据和标定方法收到世界著名企业关注



## Distilling Vision-Language Models on Millions of Videos

Yue Zhao<sup>1,2\*</sup> Long Zhao<sup>1</sup> Xingyi Zhou<sup>1</sup> Jialin Wu<sup>1</sup>  
Chun-Te Chu<sup>1</sup> Hui Miao<sup>1</sup> Florian Schroff<sup>1</sup> Hartwig Adam<sup>1</sup>  
Ting Liu<sup>1</sup> Boqing Gong<sup>1</sup> Philipp Krähenbühl<sup>2</sup> Liangzhe Yuan<sup>1</sup>  
<sup>1</sup>Google <sup>2</sup>University of Texas, Austin

**基于我们的数据，改进标定方法**



## Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets

Andreas Blattmann\* Tim Dockhorn\* Sumith Kulal\* Daniel Mendelevitch  
Maciej Kilian Dominik Lorenz Yam Levi Zion English Vikram Voleti  
Adam Letts Varun Jampani Robin Rombach  
Stability AI

**基于我们的标定方法，重新收集标记数据**

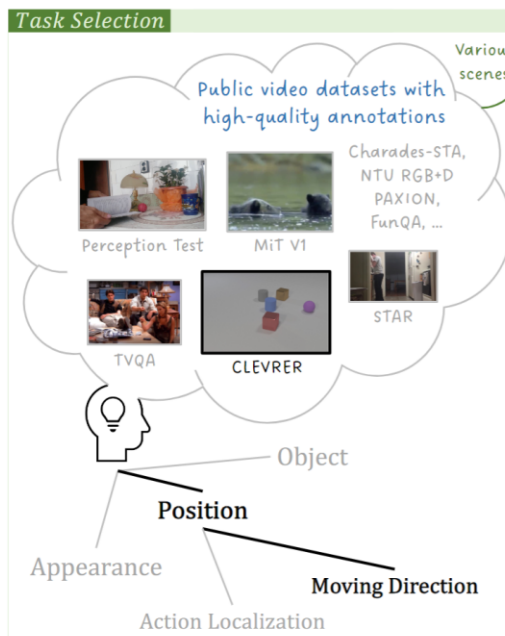
# 视频理解模型评估: MVBench

## • 提出了多模态视频理解评测基准MVBench



### Temporal Understanding: Reasoning based on entire video

- |                       |                       |                            |
|-----------------------|-----------------------|----------------------------|
| ① Action              | ③ Position            | ⑦ Attribute                |
| - Action Sequence     | - Moving Direction    | - State Change             |
| - Action Antonym      | - Action Localization | - Moving Attribute         |
| - Action Prediction   | ④ Count               | ⑧ Character                |
| - Unexpected Action   | - Action Count        | - Character Order          |
| - Fine-grained Action | - Moving Count        | ⑨ Cognition                |
| ② Object              | ⑤ Scene               | - Episodic Reasoning       |
| - Object Shuffle      | - Scene Transition    | - Egocentric Navigation    |
| - Object Existence    | ⑥ Pose                | - Counterfactual Inference |
| - Object Interaction  | - Fine-grained Pose   |                            |



### Data Filtration

Video Diversity: Each QA pair corresponds to a distinct video

Temporal Sensitivity:
 

- Too short: minimal movement
- Intermediate duration
- Too long: complicated context

Question Difficulty:
 

- Too easy: indistinguishable
- Proper question
- Too hard: inseparable

### QA Generation

Have options?
 

- yes: Directly adopt QA
- no: Generate QA with video annotations

LLM asks question  $\leftarrow$  task definition

Template-based option candidates

Up and to the left; Up and to the right; Down and to the left; Down and to the right; The object is stationary.

### Option Processing

Order Shuffle: Options are randomly selected and shuffled

Length Check: LLM have similar and reasonable text lengths

### Evaluation: Prompt Design

Q: What direction is the gray cylinder moving within the video?

(A) Up and to the right. (B) Up and to the left. (C) The object is stationary. (D) Down and to the right.

System Prompt: Consider temporal evolution

Carefully watch the video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons. Based on your observations, select the best option that accurately addresses the question.

Answer Prompt: Must output option

Best Option: (

Video MLLMs: All models take 16 frames as input, with the exception of VideoChatGPT, which uses 100 frames.

Otter-V [37]	LLaMA-7B	26.8	23.0	23.0	27.5	27.0	29.5	53.0	28.0	33.0	24.5	23.5	27.5	26.0	28.5	18.0	38.5	22.0	22.0	23.5	19.0	19.5
mPLUG-Owl-V [88]	LLaMA-7B	29.7	22.0	28.0	34.0	29.0	29.0	40.5	27.0	31.5	<b>27.0</b>	23.0	29.0	31.5	27.0	40.0	44.0	24.0	31.0	26.0	20.5	29.5
VideoChatGPT [49]	Vicuna-7B	32.7	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	<b>48.5</b>	29.0	33.0	29.5	26.0	35.5
VideoLLaMA [95]	Vicuna-7B	34.1	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0
VideoChat [40]	Vicuna-7B	35.5	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0
<b>VideoChat2<sub>text</sub></b>	Vicuna-7B	34.7	24.5	27.0	49.5	27.0	38.0	53.0	28.0	40.0	25.5	27.0	38.5	41.5	27.5	32.5	46.5	26.5	36.0	33.0	32.0	40.0
<b>VideoChat2</b>	Vicuna-7B	<b>51.1</b>	<b>66.0</b>	47.5	<b>83.5</b>	<b>49.5</b>	60.0	<b>58.0</b>	<b>71.5</b>	<b>42.5</b>	23.0	23.0	<b>88.5</b>	39.0	<b>42.0</b>	<b>58.5</b>	44.0	<b>49.0</b>	36.5	<b>35.0</b>	40.5	<b>65.5</b>

GPT-4V take 16 frames as input, and the resolution is 512x512, while others use small resolution of 224x224.

GPT-4V [55]	GPT-4	43.5	55.5	<b>63.5</b>	72.0	46.5	<b>73.5</b>	18.5	59.0	29.5	12.0	<b>40.5</b>	83.5	39.0	12.0	22.5	45.0	47.5	<b>52.0</b>	31.0	<b>59.0</b>	11.0
-------------	-------	------	------	-------------	------	------	-------------	------	------	------	------	-------------	------	------	------	------	------	------	-------------	------	-------------	------



## • 提出了多模态视频理解对话模型VideoChat2

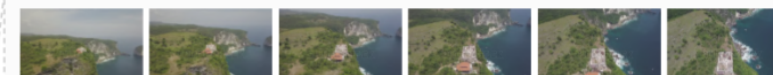
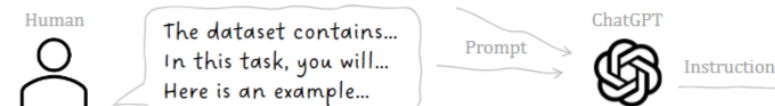
Conversation	#Num	Reasoning	#Num	VQA	#Num
 LLaVA	56,681	 LLaVA	76,643	 VQAv2	29,903
 VideoChat	13,884	CLEVR	30,000	GQA	30,001
VideoChatGPT	13,303	VisualMRC	15,000	OKVQA	8,990
Classification	#Num			A-OKVQA	17,056
 ImageNet	30,000	 NExTQA	34,132	ViQuAE	1,152
COCO-ITM	29,919	CLEVRER_QA*	40,000	OCR-VQA	11,414
 Kinetics-710	40,000	CLEVRER_MC	42,620	TextVQA	27,113
SthSthV2	40,000	Simple Caption	#Num	ST-VQA	26,074
Detailed Caption	#Num	 COCO	566,747	DocVQA	39,463
 MiniGPT-4	3,362	TextCaps	97,765	 TGIF-Frame	39,149
LLaVA	23,240	 WebVid	400,000	TGIF-Transition	52,696
Paragraph Captioning	14,575	YouCook2	8,760	WebVidQA	10,000
 VideoChat	6,905	TextVR	39,648	EgoQA	7,813

### Instruction Generation

You are professional in video understanding and instruction design. I will give you the description of video dataset and task, and one instruction example.

DATASET DESCRIPTION: {dataset\_description}  
TASK DESCRIPTION: {task\_description}  
INSTRUCTION EXAMPLE: {instruction\_example}

Based on the above message, you need to help me generate 10 instructions for handling the video tasks.





```
# video data path
'video': '023601_023650/1023815317.mp4',
# conversion tasks have multiple QA
'QA': [{
  # instruction as task guidance
  'i': "Go through the video, taking into account
key aspects, and respond to the question.",
  # no question for caption tasks
  'q': "What color cliff is the hindu temple on?",
  # short answer may be phrased
  'a': "The Hindu temple in the video is situated
on a green cliff."
}]
```

Data Example

## • 提出了多模态视频理解对话模型VideoChat2



 The video shows a large bird standing in a dry, barren landscape with a few plants in the background. The bird is a golden eagle, and it is standing on the ground with its wings spread out. The bird is not moving and appears to be resting. The landscape is dry and barren, with no signs of life or vegetation. The bird is the only living creature in the scene.

 The video shows a bird of prey, likely a hawk, standing on the ground and looking around. It then spots a dead bird and starts to eat it. The bird of prey is seen eating the dead bird and then flying away.

 The video clip shows a bird of prey attacking a small animal. The bird is seen swooping down and attacking the animal, which appears to be a rabbit. The bird is a hawk, and it is hunting for its prey in the desert.

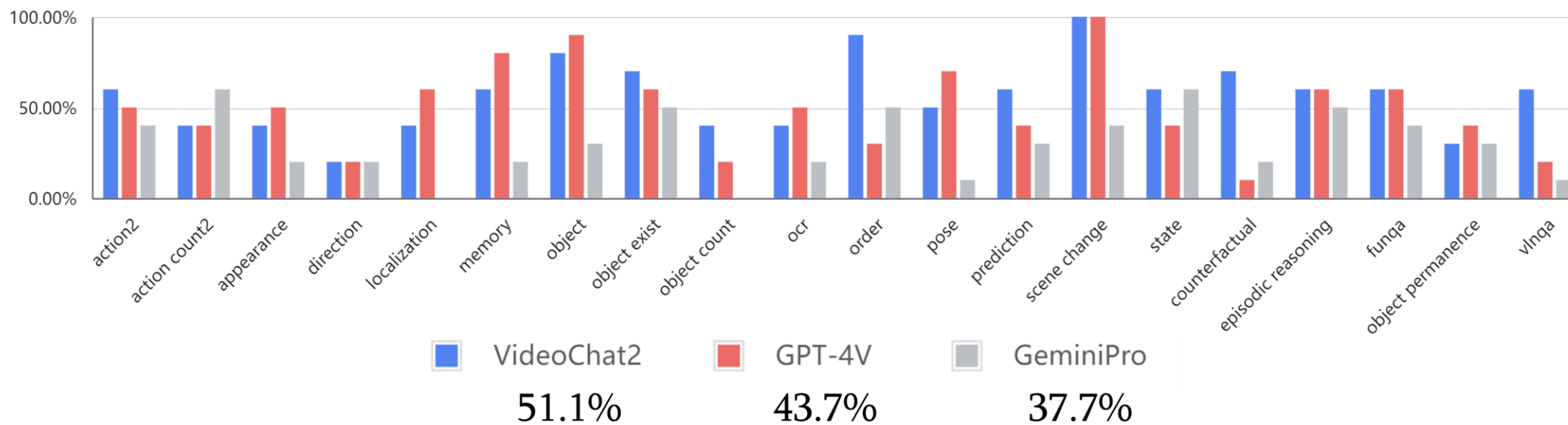
 User    VideoChat    VideoChatGPT    VideoChat2

# 视频多模态大模型：VideoChat2

首创视频理解对话新范式，基于书生多模态大模型的视频理解能力，是目前时序理解能力最强的视频对话系统。获得Amazon、微软、DeepMind、Google、Nvidia、斯坦福、AllenAI、MBZUAI等机构跟随研究\*。

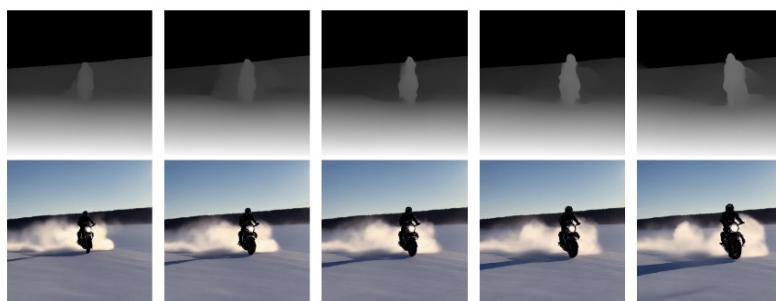
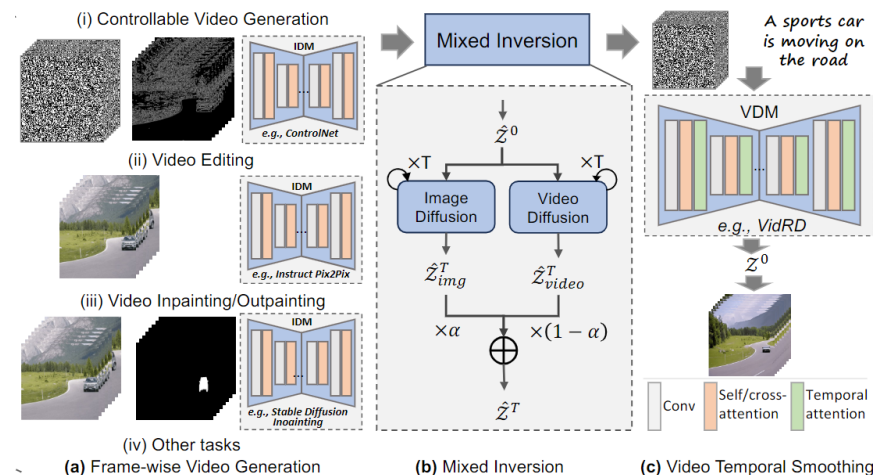
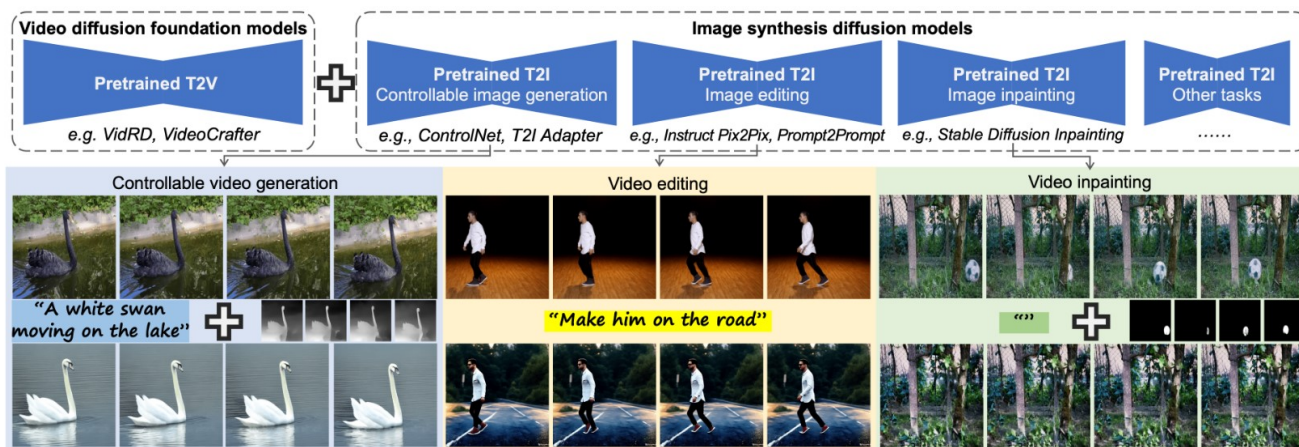
## VideoChat2 vs. GPT-4V vs. Gemini Pro

VideoChat2 vs GPT-4V vs Gemini Pro

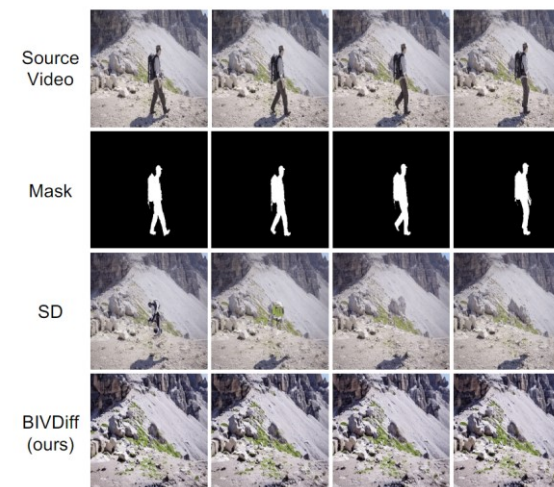


最新InternVideo2的多模态视频对话性能已经达到 60.9%

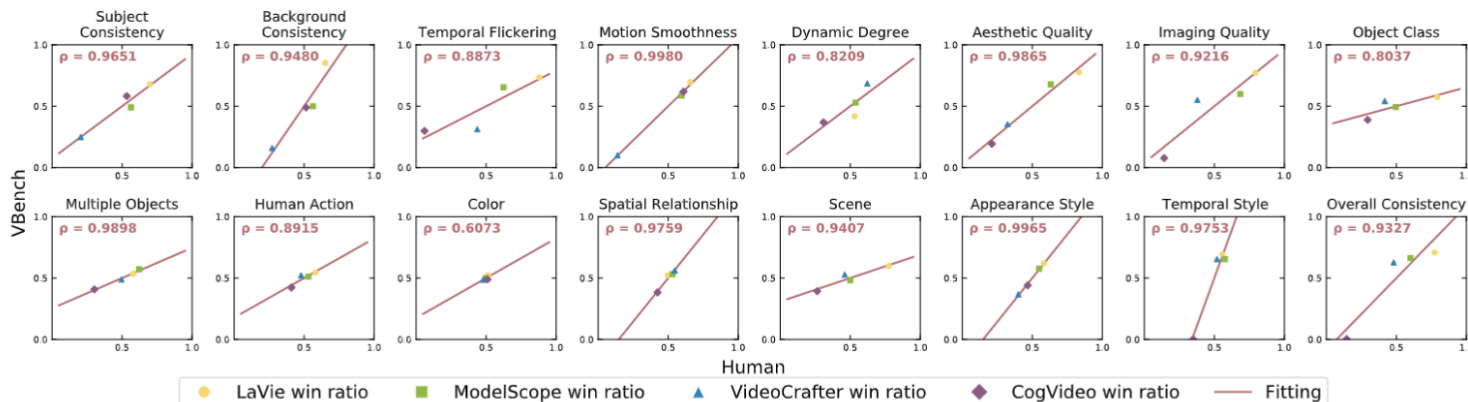
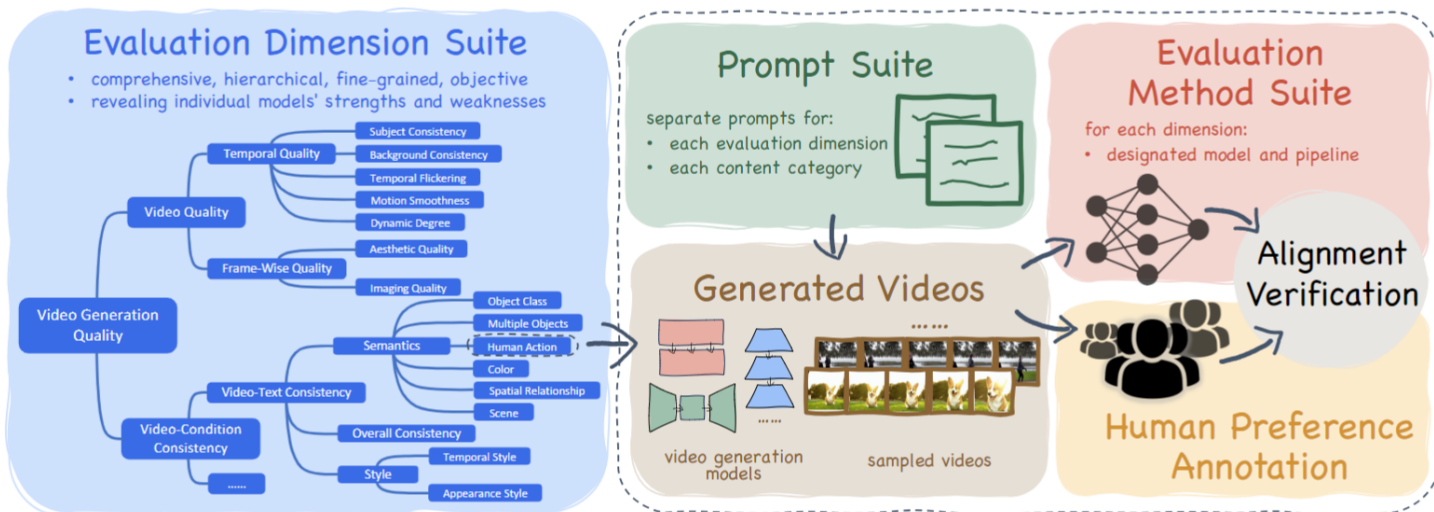
## • 提出一种Training-Free的通用视频合成方案（结合图像和视频生成模型）



(i) Depth: A person on a motorcycle does a burnout on a frozen lake.



## • 提出了全面AIGC视频评测体系与工具集



# 视频AGI研究总结

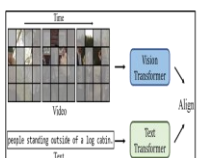
突破视频多模态表征学习的性能瓶颈，构建具有体系化动态感知能力的最强视频基础模型，形成首个以“多模态交互”为核心的通用视频理解新范式。

## 视频多模态数据：当前最大

(2.34亿多模态视频片段，76万小时)



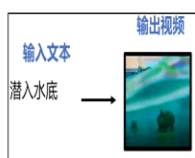
可迁移多模态表征



多模态对话系统



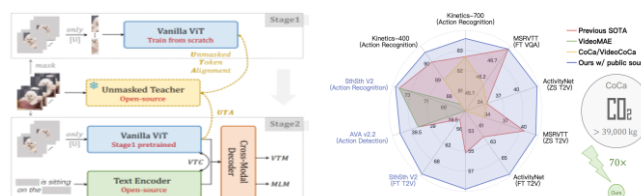
多模态内容生成



## InternVid

## 视频基础模型：当前最强

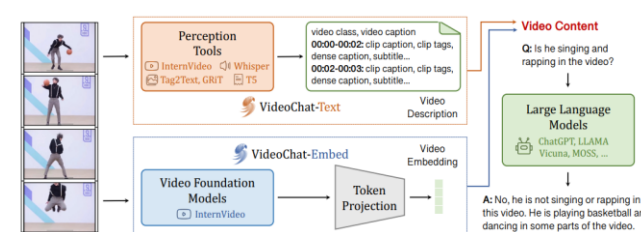
(10亿级模型参数，20+视频任务领先)



## InternVideo 系列

## 视频对话系统：国际首个

(GitHub半年2300星，支持央视应用)



## VideoChat系列