Animal-Bench: Benchmarking Multimodal Video Models for Animal-centric Video Understanding

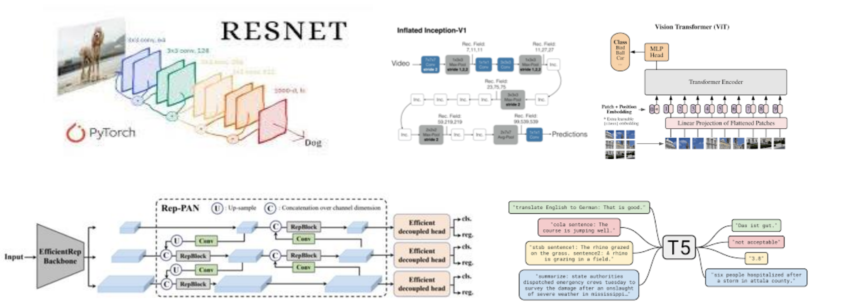
Yinuo Jing[1], Ruxu Zhang[1], Kongming Liang[1]✉, Yongxiang Li[2], Zhongjiang He[2], Zhanyu Ma[1], Jun Guo[1]

[1]School of Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]China Telecom Artificial Intelligence Technology Co. Ltd

Outline

Outline

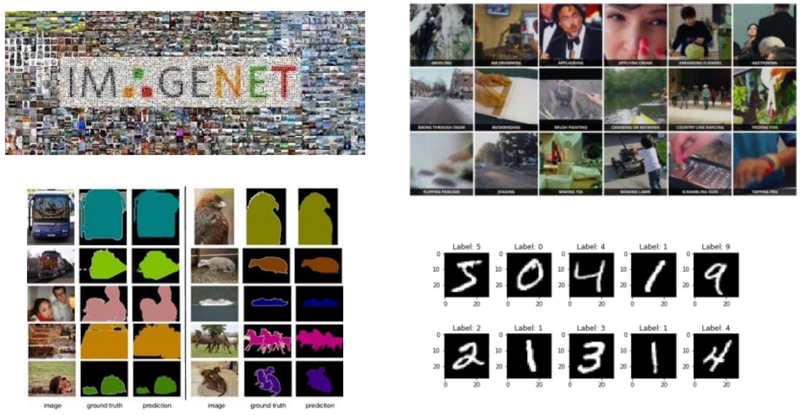
- Technological development: from **basic singular** capabilities to **multimodal comprehensive** abilities



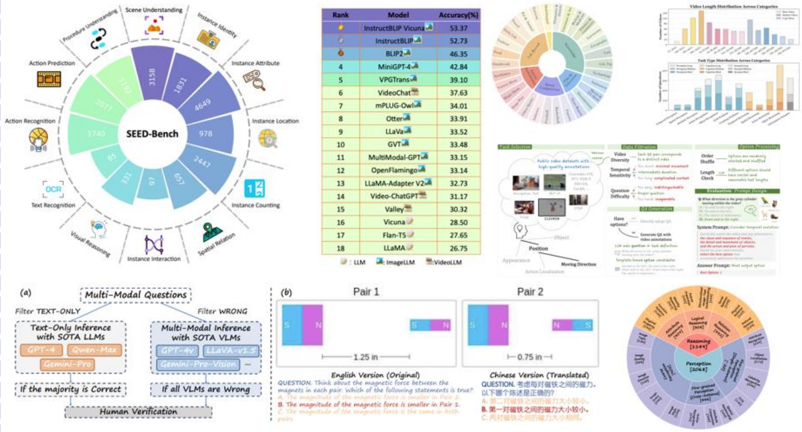
Single-modal pre-trained models

The **rapid explosion** of multi-modal models

- Benchmark developments: from **single-task** to **multi-dimensional** evaluation



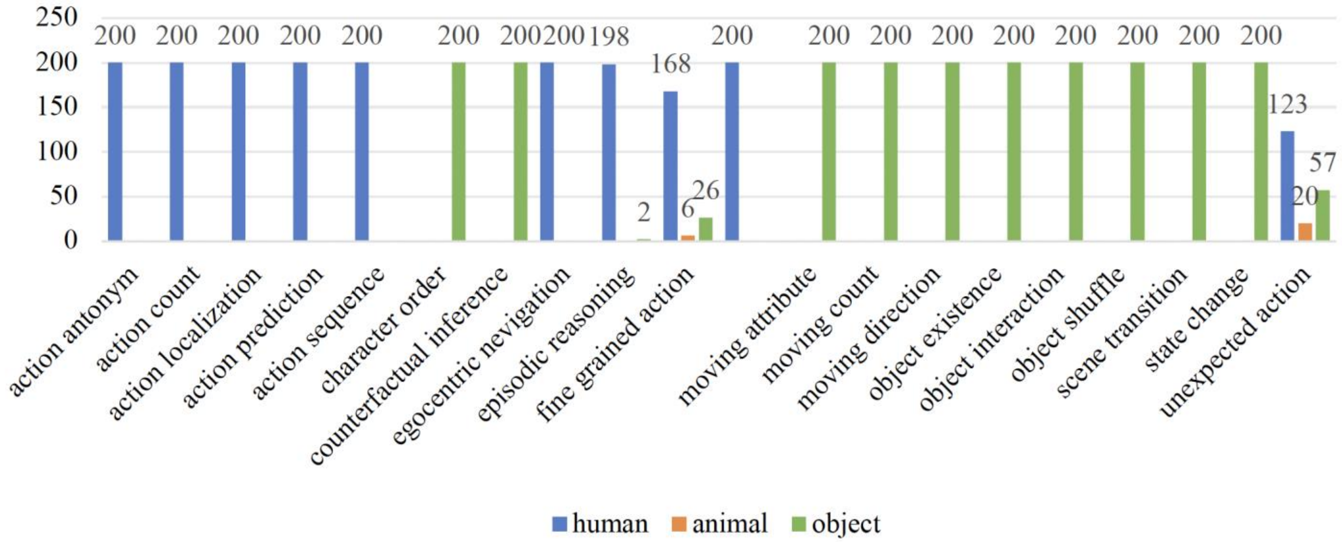
Task-specific data



Multidimensional multi-task data

Outline

Motivation



The data quantity for different agents in each task of MVBench

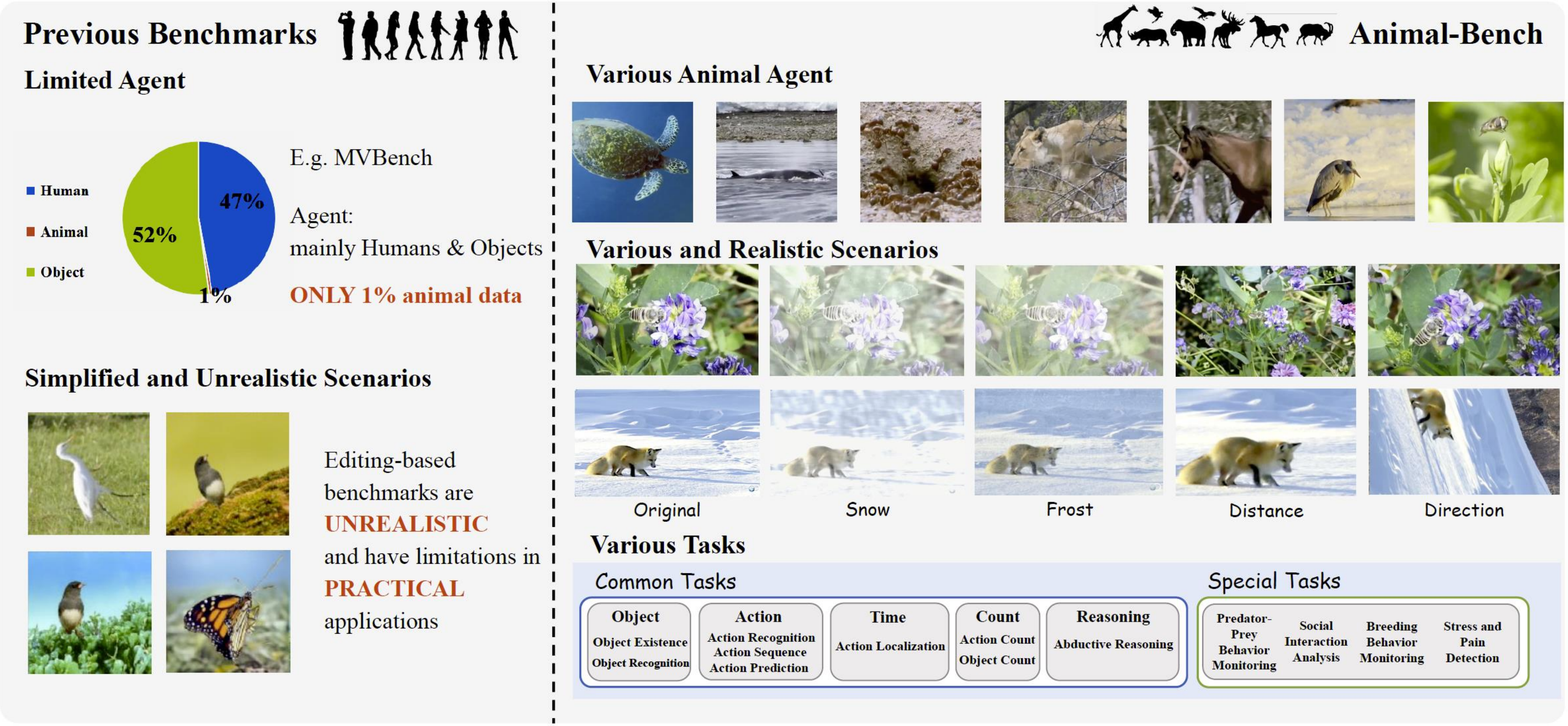
Benchmarks	Dataset Properties			Tasks
	Label	QA Size	Agent(main)	
Video-MME [36]	Multi-Choice QA	2700 QAs	Human & Object	object, action, attribute, position, count, time, reasoning, summarization, etc.
Video-Bench [11]	Multi-Choice QA	15,033 QAs	Human & Object	action, object, attribute, position, count, time, reasoning, etc.
MVBench [3]	Multi-Choice QA	4000 QAs	Human & Object	action, object, position, count, scene, pose, attribute, character and cognition
Animal Kingdom [37]	Classification	N/A	Animal	object, action, time
MammalNet [38]	Classification	N/A	Animal	object, action

Comparison Table of Existing Video Understanding Evaluation Benchmarks

- Existing benchmarks focus on **humans or objects**, while animal-centric datasets assess only **limited model capabilities**.
- **Animal-related tasks are challenging** due to species diversity and environmental complexity, with low data leakage risk, making them ideal for **testing model robustness**.
- Animal-centric evaluation research is vital for monitoring ecosystem health and supporting timely interventions, which is of significant importance for **wildlife conservation**.

Outline

Previous: limited agent & simplified and unrealistic scenarios
Animal-Bench: various animal agent & scenarios & tasks

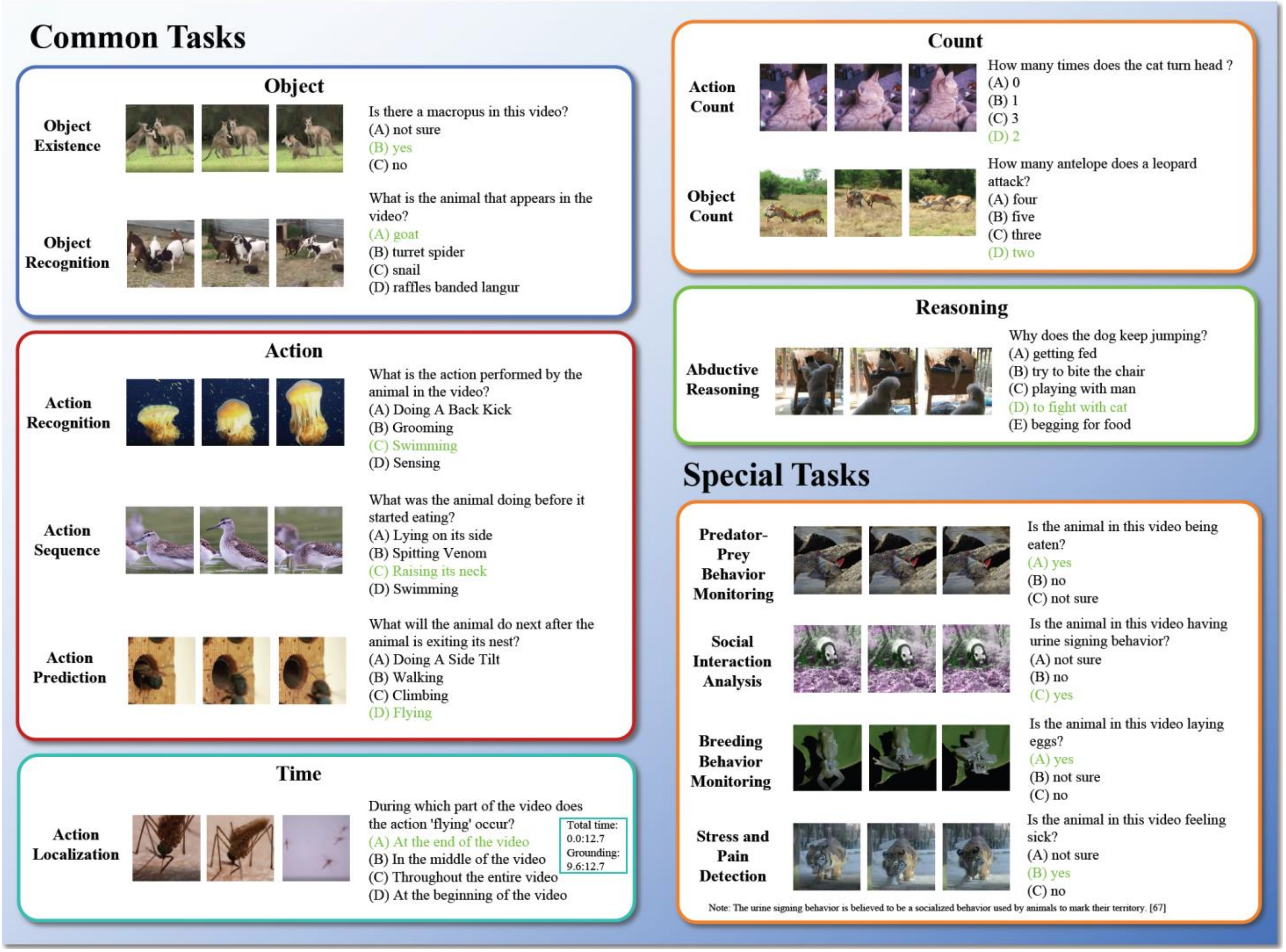


Animal-Bench Tasks and Data Composition

- Model optimization
- Animal conservation

- 13 Tasks
- 7 Categories, 819 Species
- 4 Senarios
- 8 VideoLLMs

Animal-Centric Tasks System



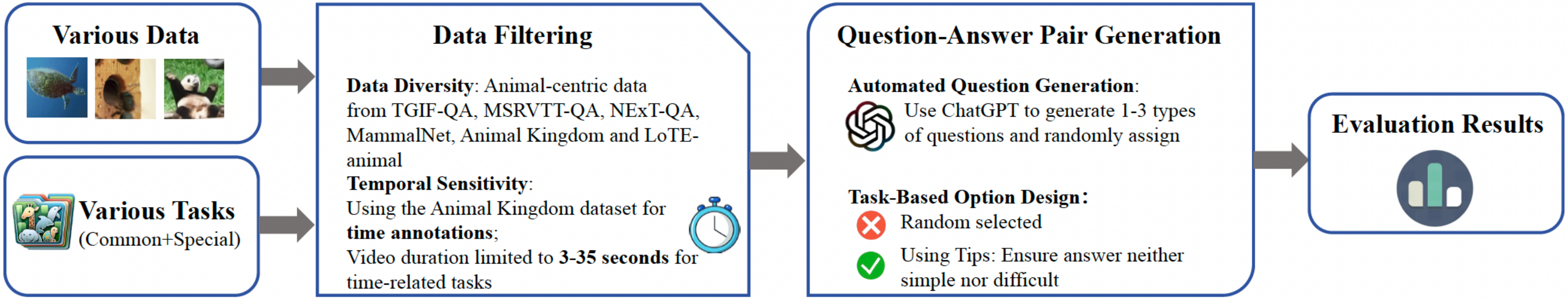
Animal-Bench Task Examples

- **Common tasks** shared with human video benchmarks, such as object detection, action recognition, counting, and reasoning
- **Specific tasks** related to **wildlife conservation**

Method

Animal-Centric Data Processing Pipeline

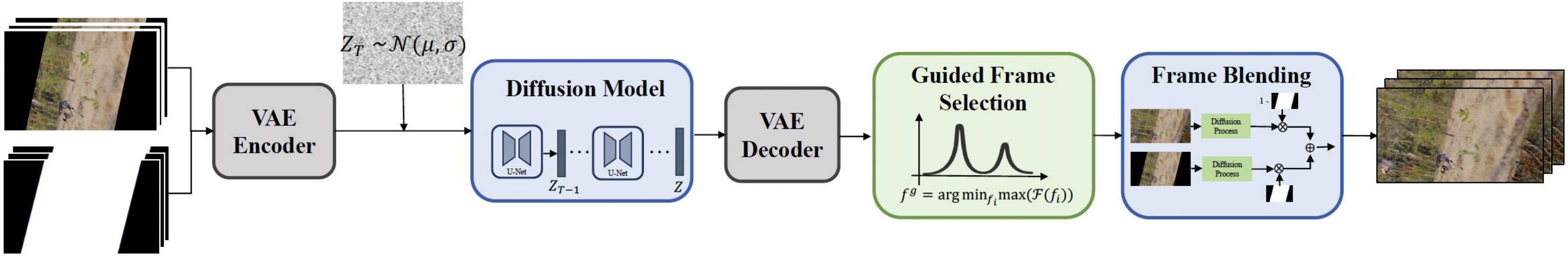
- **Data Filtering:** *ensuring data diversity and temporal sensitivity*
- **QA Pair Generation:** *ensuring moderate difficulty of questions and options*



Automated Data Filtering and QA Generation Pipeline

Realistic Simulation based on Video Editing

- **Weather conditions,** *including snowy and frosty*
- **Shooting parameters,** *including shooting distance and shooting direction*



Simulation process for shooting parameters

Outline

Effectiveness evaluation results

Task	Random 95% confidence interval	mPLUG -Owl [44]	Video Chat [62]	Video -ChatGPT [63]	Video -LLaMA [8]	Valley [64]	Chat -UniVi [65]	Video -LLaVA [4]	Video Chat2 [3]	Avg
		Multimodal Video Model								
OE	33.32 ± 0.46	42.20	49.40	44.65	49.20	41.70	44.65	45.90	50.00	45.96
OR	24.96 ± 0.19	33.62	51.61	24.31	60.23	25.06	43.25	40.55	86.75	45.67
AR	25.26 ± 0.17	27.00	32.54	24.28	35.34	24.56	32.98	31.71	66.27	34.34
AS	26.12 ± 1.23	25.86	32.76	22.41	29.74	27.16	33.19	25.86	54.31	31.41
AP	24.16 ± 1.31	25.48	27.88	24.52	29.81	28.37	27.88	28.37	50.00	30.29
AL	25.49 ± 0.39	24.49	23.25	21.22	24.67	25.45	24.14	24.32	21.22	23.60
OC	25.17 ± 1.18	24.14	27.59	24.71	26.44	25.29	31.61	31.03	64.94	31.97
AC	25.06 ± 0.37	24.43	25.51	22.92	24.99	23.78	24.34	22.49	29.16	24.70
RS	19.46 ± 0.71	22.38	27.07	25.69	35.08	22.65	36.46	21.27	68.23	32.35
				Special Task						
PM	33.58 ± 0.41	43.19	48.00	44.88	50.68	40.28	49.70	45.37	52.44	46.82
BM	33.63 ± 1.21	39.31	50.29	43.35	47.98	44.80	48.84	42.20	47.69	45.56
SA	33.22 ± 0.54	41.08	48.87	47.23	49.47	42.96	48.18	44.16	52.42	46.80
PD	33.15 ± 1.55	40.56	47.55	46.85	50.35	38.46	44.06	45.80	54.20	45.98
				Overall Performance						
Avg	27.89 ± 2.90	31.83	37.87	32.08	39.54	31.58	37.64	34.54	53.66	37.34

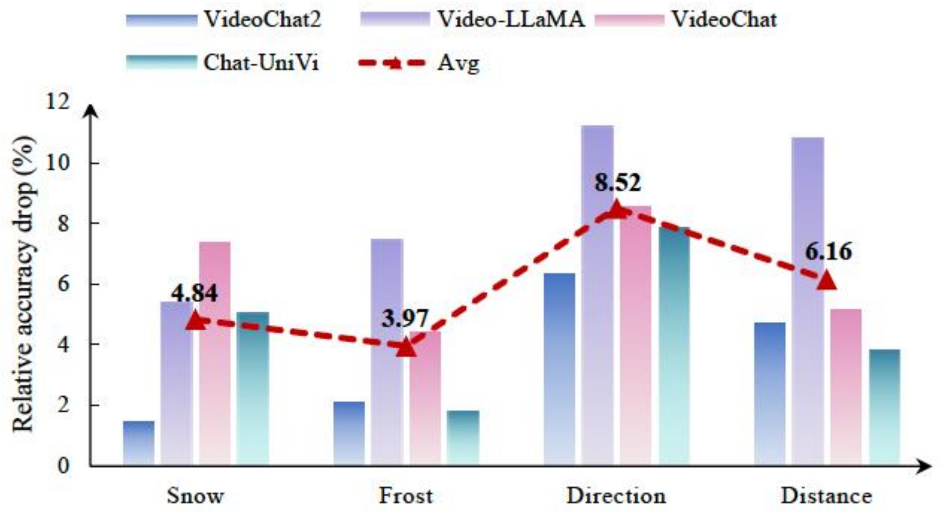
The evaluation results of 8 multimodal video models on our Animal-Bench

- Recently released models like VideoChat2 surpass previous methods in most tasks.
- Existing models need to enhance their **temporal modeling capabilities**.

Robustness evaluation results

Models	Weather condition		Shooting parameters		Overall
	Snow	Frost	Distance	Direction	
VideoChat2	1.49	2.17	4.76	6.39	3.70
Video-LLaMA	5.41	7.46	10.82	11.19	8.72
VideoChat	7.43	4.41	5.22	8.63	6.42
Chat-UniVi	5.04	1.81	3.83	7.86	4.64

Sensitivity of multimodal video models to different variations(relative accuracy drop(%))



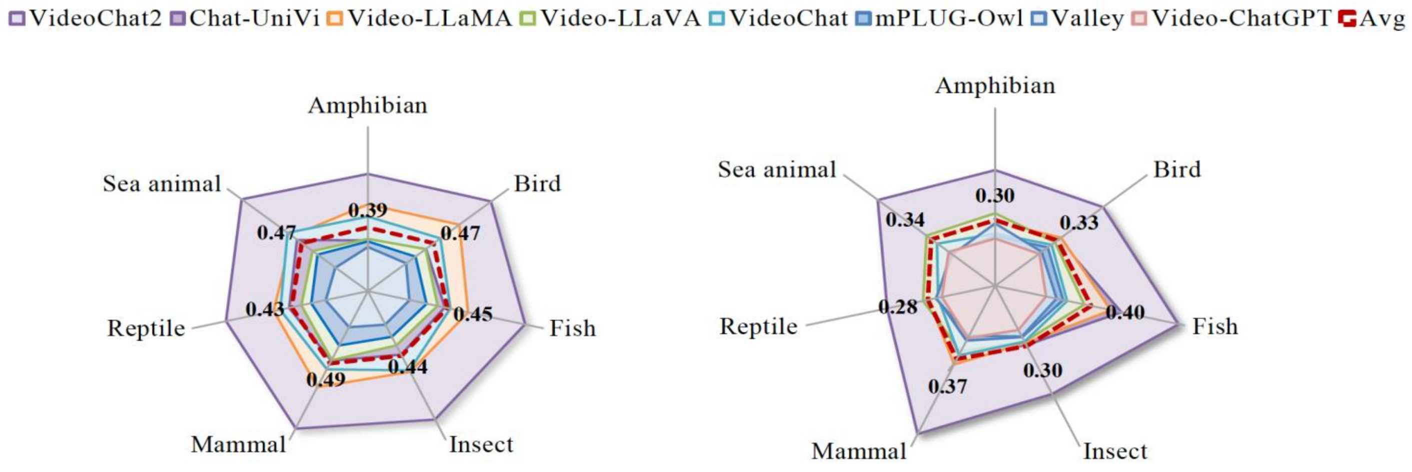
Average decrease in model accuracy(%) across four types of variations

- VideoChat2 demonstrated relatively **good robustness**
- VideoLLMs are more sensitive to **shooting parameters** than to changes in weather changes

Experiments

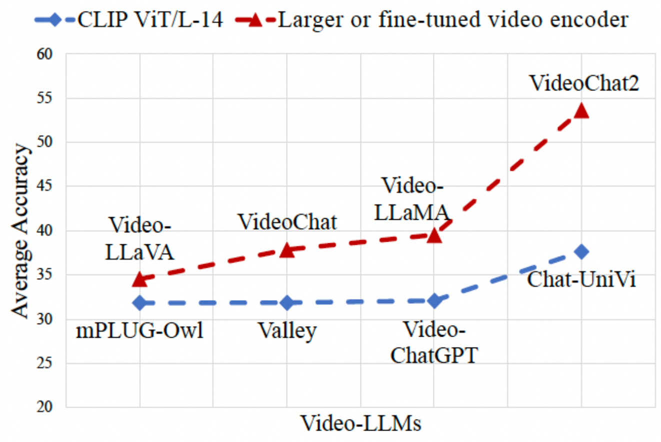
Further discussion

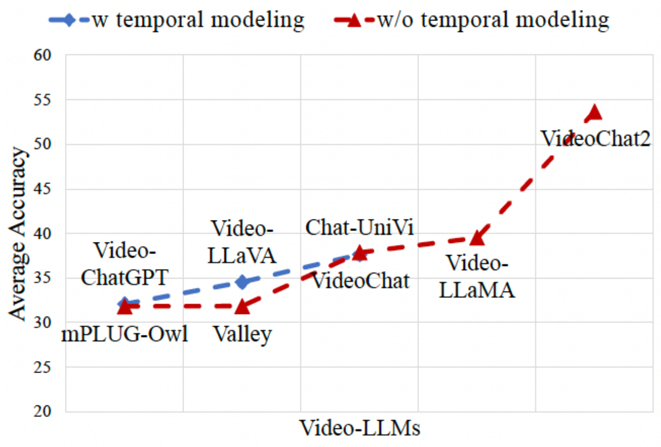
➢ Animal category bias



- Object Recognition: higher accuracy for the **“mammal” and “bird”**, while lower accuracy for **“amphibian” and “reptile”**
- Action Recognition: higher accuracy for **“fish” and “mammal”**

➢ Model structure





- Employing **more powerful video encoders** is of significant importance for the development of multi-modal video models
- The impact of **the temporal modeling module** may **not be as significant as expected**

Outline

Conclusion

- We established Animal-Bench, an **animal-centric benchmark**, to enable **comprehensive evaluation** of model capabilities in real-world contexts and overcome agent-bias in previous benchmarks.
- Animal-Bench includes **13 tasks** covering both common human-related tasks and special tasks for animal conservation, spanning 7 animal categories and 819 species, with 41,839 data entries.
- We defined a **task system** centered on animals and proposed **an automated pipeline** for processing animal-centric data.
- We applied a **video editing approach** to simulate realistic scenarios, such as weather changes and shooting parameters, caused by animal movements, to test model robustness.
- We evaluated **8 current multimodal video models** on Animal-Bench and identified **significant room for improvement**, aiming to provide insights and open new research avenues for multimodal video models.

NEURAL INFORMATION PROCESSING SYSTEMS

北京郵電大學
Beijing University of Posts and Telecommunications

Animal-Bench: Benchmarking Multimodal Video Models for Animal-centric Video Understanding

Yinuo Jing[1], Ruxu Zhang[1], Kongming Liang[1], Yongxiang Li[2], Zhongjiang He[2], Zhanyu Ma[1], Jun Guo[1]

Our data and code will be released at
https://github.com/PRIS-CV/Animal-Bench