Application Talk

Computer Science Ph.D. Applicant

Wenhao Chai

University of Washington
Department of Electrical & Computer Engineering

December 9, 2024

Table of Contents

1. Introduction
   - Background
2. Research
   - Research Overview
   - Large Multi-modal Models for Video Understanding
3. Future Plan

Background

M.S.	University of Washington (UW)
EE	Advisors: Jenq-Neng Hwang
2023-2025	Thesis: *LMMs for Video Understadning*
Visiting Scholar	University of Illinois Urbana-Champaign (UIUC)
2022	National Center for Supercomputing Application
B.S.	Zhejiang University (ZJU)
2019-2023	GPA: 3.70 / 4.00
Research Intern	Pika Labs
Summer 2024	Research Intern Working on Video Captioning
Research Intern	Microsoft Research Asia
Spring/Summer 2023	Research Intern Working on Video Editing

Research Overview

Large Multi-modal Models for Video Understanding

AuroraCap [1] @ ICLR 25 for *first* video detailed caption
MovieChat [2] @ CVPR 24 for *first* long-form video

Embodied Agent @ Virtual Environment

STEVE [3] @ ECCV 24 for minecraft agent

Generative Models for Video, Image, and 3D

StableVideo [4] @ ICCV 23 for video editing

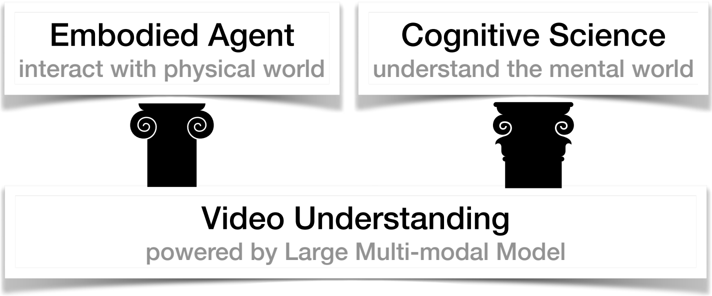
Human Pose and Motion

PoseDA [5] @ ICCV 23, RT-Pose [6] @ ECCV 24 for 3D human pose
UniAP [7] @ AAAI 24 for 2D animal pose

AI for Applied Science

structure analysis @ civil engineering [8, 9]
medical image analysis [10]

Future Research



Embodied Agent

interact with physical world

Cognitive Science

understand the mental world

Video Understanding

powered by Large Multi-modal Model

Figure: Future Research

Large Multi-modal Models for Video Understanding

Short videos, short captions — can they tell the whole story?

Figure: Video example of MSR-VTT [11], which is a widely used video question answering and captioning benchmark. Labeled caption: *Teams are playing soccer.*

Large Multi-modal Models for Video Understanding

Long videos MovieChat: From Dense Token to Sparse Memory for Long Video Understanding @ CVPR 24

MovieChat+: Question-aware Sparse Memory for Long Video Question Answering @ TPAMI *minor*

Long captions AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark @ ICLR 25

Long-form Video Understanding

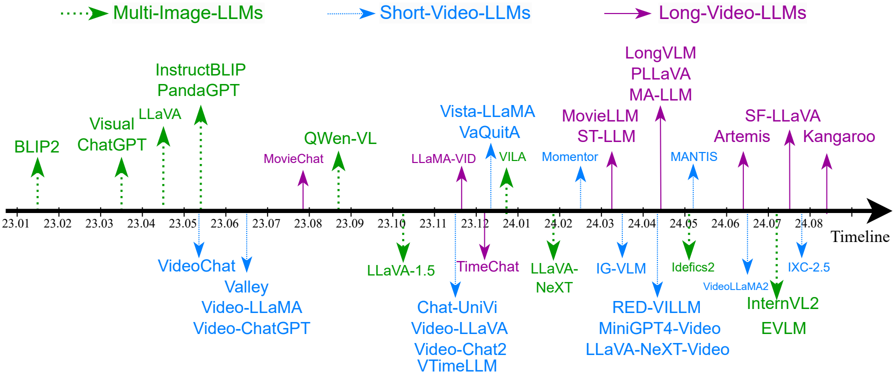


Figure: The development of LMMs for multiple images, short videos and long videos from [12].

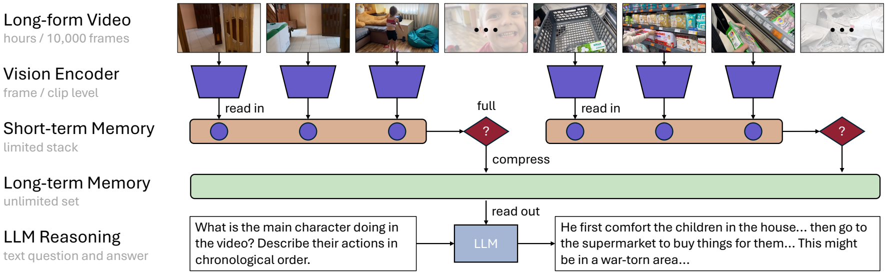
Long-form Video Understanding

Why we need long-form video understanding?
Temporal Complexity and Granularity, Narrative Comprehension, Real-World Applications, *etc*

What are the current challenges?
Efficiency, Training Data, *etc*

Can we do that with current LMMs?
Yes! We found that the LMMs trained on images and short videos can be adapted to long-form video tasks even without further fine-tuning.

Long-form Video Understanding



What is the main character doing in the video? Describe their actions in chronological order.

He first comfort the children in the house... then go to the supermarket to buy things for them... This might be in a war-torn area...

Figure: Framework of MovieChat [2].

Long-form Video Understanding

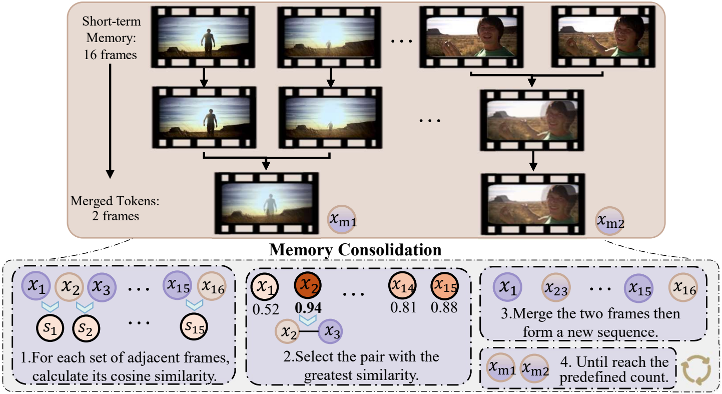
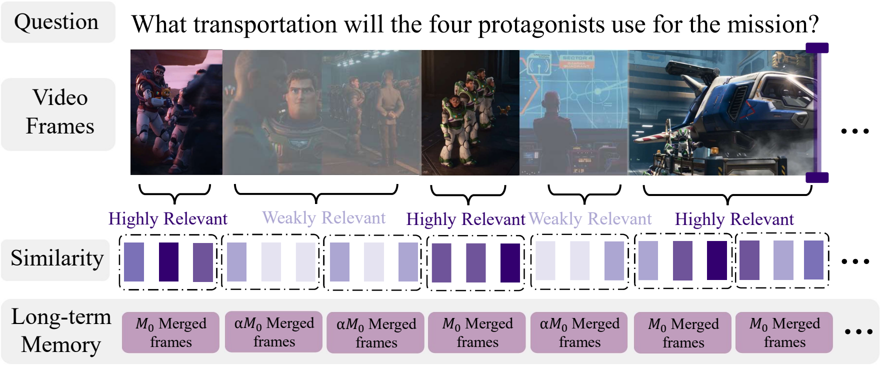

Short-term Memory: 16 frames
Merged Tokens: 2 frames
x_{m1}
x_{m2}
Memory Consolidation
x_1 x_2 x_3 … x_{15} x_{16}
s_1 s_2 … s_{15}
1.For each set of adjacent frames, calculate its cosine similarity.
x_1 0.52 x_2 0.94 … x_{14} 0.81 x_{15} 0.88
x_2 x_3
2.Select the pair with the greatest similarity.
x_1 x_{23} … x_{15} x_{16}
3.Merge the two frames then form a new sequence.
x_{m1} x_{m2}
4. Until reach the predefined count.

Figure: Memory compression in MovieChat [2].

Long-form Video Understanding



Question

What transportation will the four protagonists use for the mission?

Video Frames

Highly Relevant

Weakly Relevant

Highly Relevant

Weakly Relevant

Highly Relevant

Similarity

Long-term Memory

M_0 Merged frames

αM_0 Merged frames

αM_0 Merged frames

M_0 Merged frames

αM_0 Merged frames

M_0 Merged frames

M_0 Merged frames

Figure: Question-aware memory selection in MovieChat+ [13].

Long-form Video Understanding

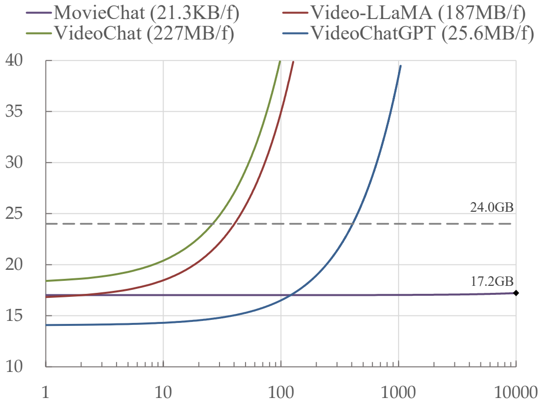


Figure: Video random-access memory (VRAM) cost under gigabyte (GB) (y-axis) v.s. frame number (x-axis) comparison.

Long-form Video Understanding

Table: The popular benchmarks for video question answering.

Benchmark	*Labels*	*#Eval Videos*	*#Eval QAs*	*Avg Duration (s)*	*Released Year*
MSVD-QA [14]	Auto	520	13,157	10	2011
MSRVTT-QA [15]	Auto	2,990	72,821	15	2017
ActivityNet-QA [16]	Human	800	8,000	180	2019
NeXT-QA [17]	Human	1,000	8,564	44	2021
MovieChat-1K [2]	Human	130	1,950	**564**	**2023.7**
EgoSchema [18]	Auto	5,031	5,031	180	2023.8
MVBench [19]	Auto	4,000	4,000	16	2023.11
LongVideoBench [20]	Human	3,763	6,678	473	2024.7

Long-form Video Understanding

Table: Quantitative evaluation for short video question answering.

Method	**MSVD-QA**		**MSRVTT-QA**		**ActivityNet-QA**		**NExT-QA**	
	Acc.	**Sco.**	**Acc.**	**Sco.**	**Acc.**	**Sco.**	**Acc.**	**Sco.**
FrozenBiLM	2.2	–	16.8	–	24.7	–	–	–
Video Chat	56.3	2.8	45.0	2.5	26.5	2.2	**56.6**	**3.2**
LLaMA Adapter	54.9	3.1	43.8	2.7	34.2	2.7	–	–
Video LLaMA	51.6	2.5	29.6	1.8	12.4	1.1	–	–
Video-ChatGPT	64.9	3.3	49.3	**2.8**	35.2	2.7	54.6	**3.2**
MovieChat	75.2	3.8	52.7	2.6	45.7	**3.4**	49.9	2.7
MovieChat+	**76.5**	**3.9**	**53.9**	2.7	**48.1**	**3.4**	54.8	3.0

Long-form Video Understanding

Table: Quantitative evaluation for long video question answering on MovieChat-1K test set.

Method	Text Decoder	# Frames	Global Mode		Breakpoint Mode	
			Acc.	Sco.	Acc.	Sco.
GIT	non-LLM based	6	28.8	1.83	29.2	1.98
mPLUG-2	non-LLM based	8	31.7	2.13	30.8	1.83
Video Chat	LLM based	32	57.8	3.00	46.1	2.29
Video LLaMA	LLM based	32	51.7	2.67	39.1	2.04
Video-ChatGPT	LLM based	100	47.6	2.55	48.0	2.45
MovieChat	LLM based	2048	62.3	3.23	48.3	2.57
MovieChat+	LLM based	2048	**71.2**	**3.51**	**49.6**	**2.62**

Long-form Video Understanding

Winners
Hierarchical Memory for Long Video QA
Yiqin Wang1,*, Haoji Zhang1,*, Yansong Tang1,†, Yong Liu1, Jifeng Dai2, Jiashi Feng3, Xiaojie Jin
1 Shenzhen International Graduate School, Tsinghua University
2 Department of Electronic Engineering, Tsinghua University 3 ByteDance Inc.
{yq-wang23@mails.,haoji-zh20@mails.,tang.yansong@sz.}tsinghua.edu.cn
jinxiaojie@bytedance.com

Figure: Photos with workshop competition winner @ CVPR 2024, Seattle.

Long-form Video Understanding

MovieChat	https://arxiv.org/abs/2307.16449
MovieChat+	https://arxiv.org/abs/2404.17176
GitHub (530⋆)	https://github.com/rese1f/MovieChat
Model	https://huggingface.co/Enxin/MovieChat-vicuna
Benchmark	https://huggingface.co/datasets/Enxin/MovieChat-1K_train (test)
Eval Code	https://github.com/EvolvingLMMs-Lab/lmms-eval
Project Page	https://rese1f.github.io/MovieChat
Workshop Page	https://sites.google.com/view/loveucvpr24/track1

Large Multi-modal Models for Video Understanding

Short videos, short captions — can they tell the whole story?

Figure: Video example of MSR-VTT [11], which is a widely used video question answering and captioning benchmark. Labeled caption: *Teams are playing soccer.*

Video Detailed Captioning

AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark

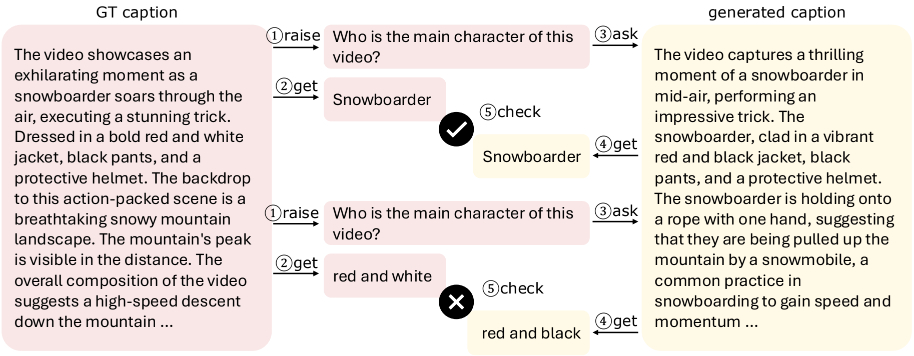
ICLR 25, score 8, 8, 6, 6, 6

Video Detailed Captioning

Table: **Benchmark comparison** for video captioning task. Ave. Length indicates the average number of words per caption.

Dataset	Theme	# Video	# Clip	# Caption	# Word	# Vocab.	Ave. Length
MSVD	Open	1,970	1,970	70,028	607,339	13,010	8.67
MSR-VTT		7,180	10,000	200,000	1,856,523	29,316	9.28
ActivityNet		20,000	100,000	100,000	1,340,000	15,564	13.40
S-MiT		515,912	515,912	515,912	5,618,064	50,570	10.89
M-VAD	Movie	92	48,986	55,905	519,933	18,269	9.30
MPII-MD		94	68,337	68,375	653,467	24,549	9.56
Youcook2	Cooking	2,000	15,400	15,400	121,418	2,583	7.88
Charades	Human	9,848	10,000	27,380	607,339	13,000	22.18
VATEX		41,300	41,300	413,000	4994,768	44,103	12.09
VDC (ours)	Open	1,027	1,027	1,027	515,441	20,419	**500.91**

VIdeo Detailed Captioning



GT caption

The video showcases an exhilarating moment as a snowboarder soars through the air, executing a stunning trick. Dressed in a bold red and white jacket, black pants, and a protective helmet. The backdrop to this action-packed scene is a breathtaking snowy mountain landscape. The mountain's peak is visible in the distance. The overall composition of the video suggests a high-speed descent down the mountain ...

generated caption

The video captures a thrilling moment of a snowboarder in mid-air, performing an impressive trick. The snowboarder, clad in a vibrant red and black jacket, black pants, and a protective helmet. The snowboarder is holding onto a rope with one hand, suggesting that they are being pulled up the mountain by a snowmobile, a common practice in snowboarding to gain speed and momentum ...

①raise → Who is the main character of this video? ③ask →

②get → Snowboarder ✓ ⑤check Snowboarder ← ④get

①raise → Who is the main character of this video? ③ask →

②get → red and white ✕ ⑤check red and black ← ④get

Figure: Evaluation pipeline with VDCscore. Like when humans take reading comprehension tests, we transform the matching between two paragraphs into a set of question-answer pairings.

Future Plan

about research - embodied agent and cognitive science with high quality papers not only CV/ML

about career - faculty job in the university

References I

Chai, Wenhao, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D. Manning.
Auroracap: Efficient, performant video detailed captioning and a new benchmark.
arXiv preprint arXiv:2410.03051, 2024.

Enxin Song, **Chai, Wenhao**†, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al.
Moviechat: From dense token to sparse memory for long video understanding.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.

Zhonghan Zhao, **Chai, Wenhao**†, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang.
See and think: Embodied agent in virtual environment.
In *European Conference on Computer Vision*, pages 187–204. Springer, 2025.

Chai, Wenhao, Xun Guo, Gaoang Wang, and Yan Lu.
Stablevideo: Text-driven consistency-aware diffusion video editing.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23040–23050, 2023.

References II

Chai, Wenhao, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang.
Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14655–14665, 2023.

Yuan-Hao Ho, Jen-Hao Cheng, Sheng Yao Kuan, Zhongyu Jiang, **Chai, Wenhao**, Hsiang-Wei Huang, Chih-Lung Lin, and Jenq-Neng Hwang.
Rt-pose: A 4d radar tensor-based 3d human pose estimation and localization benchmark.
In *European Conference on Computer Vision*. Springer, 2025.

Meiqi Sun, Zhonghan Zhao, **Chai, Wenhao**†, Hanjun Luo, Shidong Cao, Yanting Zhang, Jenq-Neng Hwang, and Gaoang Wang.
Uniap: Towards universal animal perception in vision via few-shot learning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5008–5016, 2024.

Haojia Cheng, **Chai, Wenhao**†, Jiabao Hu, Wenhao Ruan, Mingyu Shi, Hyunjun Kim, Yifan Cao, and Yasutaka Narazaki.
Random bridge generator as a platform for developing computer vision-based structural inspection algorithms.
Journal of Infrastructure Intelligence and Resilience, 3(2):100098, 2024.

References III

Yasutaka Narazaki, Wendong Pang, Gaoang Wang, and **Chai, Wenhao**.
Unsupervised domain adaptation approach for vision-based semantic understanding of bridge inspection scenes without manual annotations.
Journal of Bridge Engineering, 29(2):04023118, 2024.

Xuechen Guo, **Chai, Wenhao**, Shi-Yan Li, and Gaoang Wang.
Llava-ultra: Large chinese language and vision assistant for ultrasound.
In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8845–8854, 2024.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui.
Msr-vtt: A large video description dataset for bridging video and language.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

Heqing Zou, Tianze Luo, Guiyang Xie, Fengmao Lv, Guangcong Wang, Juanyang Chen, Zhuochen Wang, Hansheng Zhang, Huaijian Zhang, et al.
From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding.
arXiv preprint arXiv:2409.18938, 2024.

References IV

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang.
Moviechat+: Question-aware sparse memory for long video question answering.
arXiv preprint arXiv:2404.17176, 2024.

David Chen and William B Dolan.
Collecting highly parallel data for paraphrase evaluation.
In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang.
Video question answering via gradually refined attention over appearance and motion.
In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao.
Activitynet-qa: A dataset for understanding complex web videos via question answering.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

References V

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua.
Next-qa: Next phase of question-answering to explaining temporal actions.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik.
Egoschema: A diagnostic benchmark for very long-form video language understanding.
Advances in Neural Information Processing Systems, 36:46212–46244, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al.
Mvbench: A comprehensive multi-modal video understanding benchmark.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li.
Longvideobench: A benchmark for long-context interleaved video-language understanding.
arXiv preprint arXiv:2407.15754, 2024.