Application Talk

Computer Science Ph.D. Applicant

Wenhao Chai

University of Washington
Department of Electrical & Computer Engineering

January 11, 2025

Table of Contents

Background

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

Generative Models for Video, Image, and 3D

StableVideo [3] @ ICCV 23 for video editing

Human Pose and Motion

PoseDA [4] @ ICCV 23, RT-Pose [5] @ ECCV 24 for 3D human pose
UniAP [6] @ AAAI 24 for 2D animal pose

Embodied Agent in Virtual Environment

STEVE [7] @ ECCV 24 for minecraft agent

AI for Applied Science

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

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

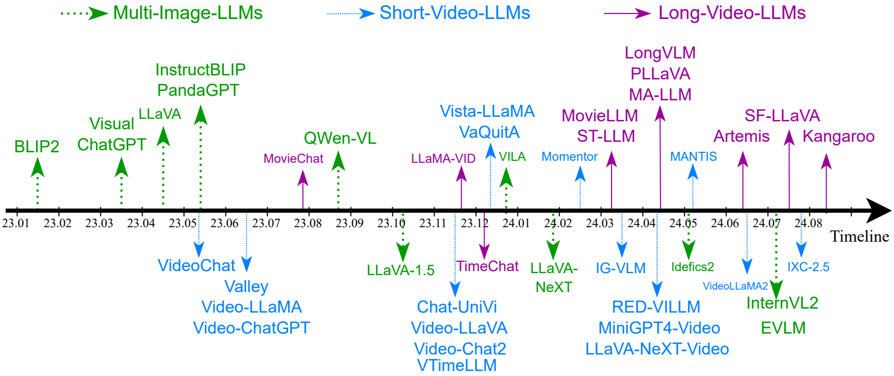

Figure: The development of LMMs for multiple images, short videos and long videos from survey paper [12].

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

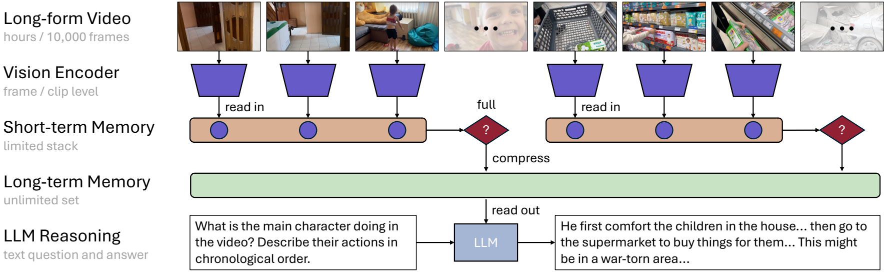

Figure: Framework of MovieChat [2].

Long-form Video Understanding

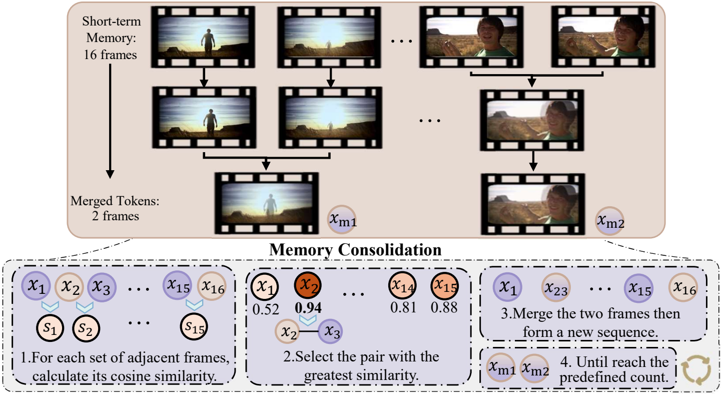

Figure: Memory compression in MovieChat [2].

Long-form Video Understanding

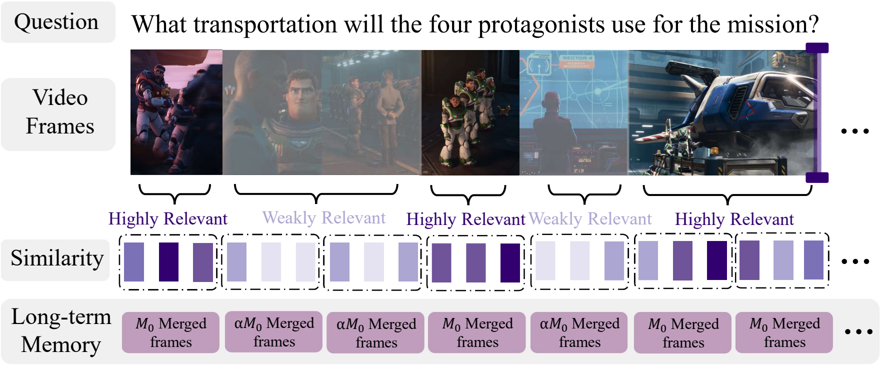

Figure: Question-aware memory selection in MovieChat+ [13].

Long-form Video Understanding

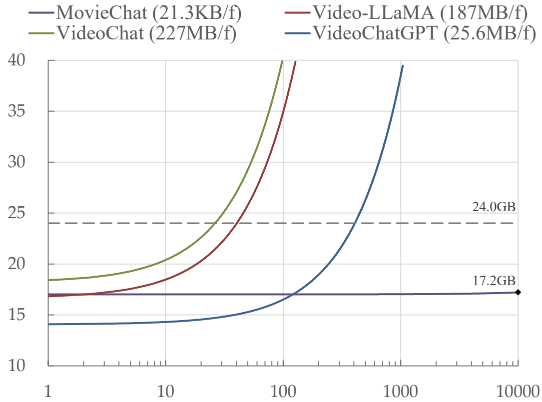

Figure: Video random-access memory (VRAM) cost under gigabyte (GB) (y-axis) v.s. frame number (x-axis) comparison.

Long-form Video Understanding

Table: The popular benchmarks for video question answering.

Benchmark	*Labels*	*#Eval Videos*	*#Eval QAs*	*Avg Duration (s)*	*Released Date*
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	–	–	–
VideoChat	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	–	–
VideoLLaMA	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
VideoChat	LLM based	32	57.8	3.00	46.1	2.29
VideoLLaMA	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

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

Long-form Video Understanding

MovieChat (+)	https://arxiv.org/abs/2307.16449
(≈200 citations)	https://arxiv.org/abs/2404.17176
GitHub (567⋆)	https://github.com/rese1f/MovieChat
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 submission with 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

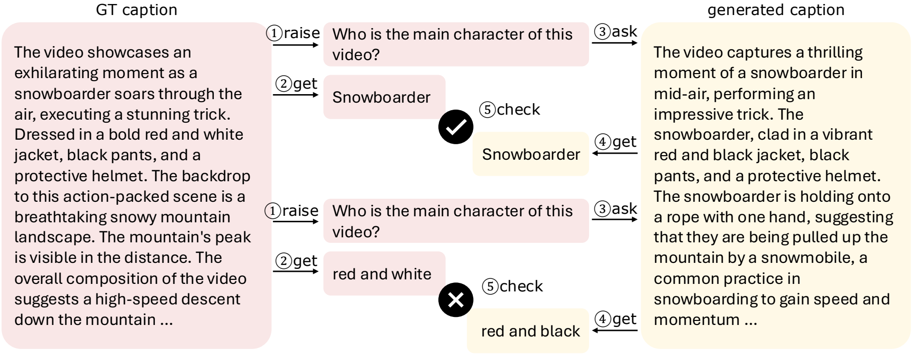

Figure: Evaluation pipeline with VDCscore. Like when humans take reading comprehension tests, we transform the matching between two paragraphs into a set of question-answer pairings.

Other Features (Current Thinking)

the interesting exploration for bridging AR and diffusion models in text (and image) generation: **Jacobi Decoding**

AR pre-training $\mapsto$ diffusion-style inference

Jacobi Decoding [21]

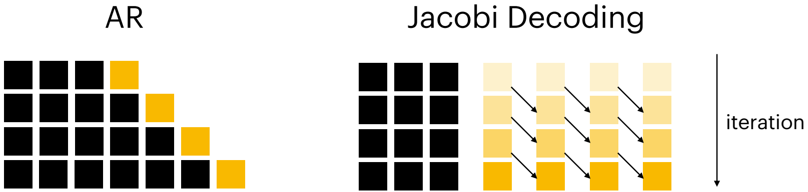

Figure: Jacobi decoding uses AR model as a diffusion-like way.

Jacobi Decoding [21]

Iteration from j-th to $j+1$-th.

$$\begin{cases} y_1^{(j+1)} & = \text{argmax}_y p(y|x) \\ y_2^{(j+1)} & = \text{argmax}_y p(y|y_1^{(j)}, x) \\ y_3^{(j+1)} & = \text{argmax}_y p(y|y_{:3}^{(j)}, x) \\ & \vdots \\ y_n^{(j+1)} & = \text{argmax}_y p(y|y_{:n}^{(j)}, x) \end{cases}$$

(cont'd) However...

1. Gap between training and inference
2. No close-form guarantee for optimization via iteration
3. Mathematically not a standard diffusion process (cold diffusion)
4. Not continue in text space (large concept model)

   Blog Link https://rese1f.github.io/blogs.html

Future Plan

about **research** - video understanding, generative models, embodied intelligence and (maybe) cognitive science with high quality papers

about **career** - faculty job in the university

References I

Chai, Wenhao, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-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.

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.

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.

References II

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.

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.

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.

Yang Song, Chenlin Meng, Renjie Liao, and Stefano Ermon.
Accelerating feedforward computation via parallel nonlinear equation solving.
In *International Conference on Machine Learning*, pages 9791–9800. PMLR, 2021.