DynFocus: Dynamic Cooperative Network Empowers LLM with Video Understanding

Yudong Han[1], Qingpei Guo[2], Liyuan Pan[1], Liu Liu, Yu Guan[3], Ming Yang[2]
[1]Beijing Institute of Technology, [2]Ant Group, [3]University of Warwick

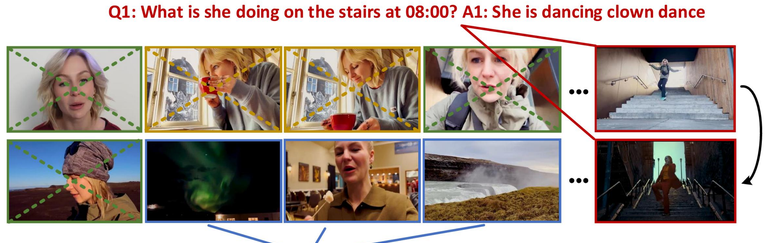
Introduction

➢ Task Definition of Video Understanding

Preserving visual and semantic information in long videos while maintaining a memory-affordable token count

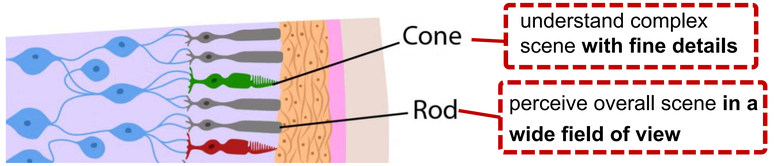
➢ Oversevation

(i) **Redundancy:** significant redundancy among frames, with only a few meaningful frames directly contributing to question answering
(ii) **Correspondence:** answering different questions generally requires focusing on different parts of the frame



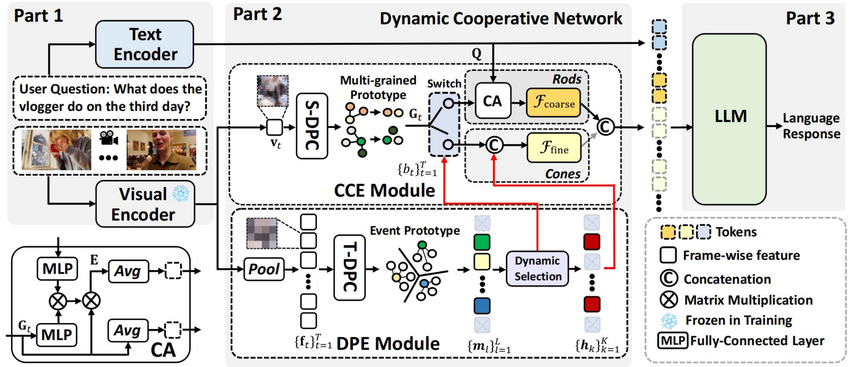
➢ Insight

(i) **dynamically identifying meaningful frames**
(ii) **adopting a dynamic encoding strategy**



Detaild Designs: (i) Which cell is activated depends on whether the current input frame is meaningful or not. (ii) The meaningful frames are encoded with fine-grained tokens as key detailed clues, akin to Cones, whereas the marginal frames are condensed into low-resolution tokens, ensuring better temporal consistency, similar to Rods.

Methods



CCE module: DPE serves as the dynamic selector to accurately discern the meaningful frames in a differential manner

DPE module: CCE dynamically encodes frames, where key frames retain fine-grained details, while redundant ones are compressed into fewer tokens, allowing LLMs to grasp broader temporal context within fixed limits

Experimental Results

MLVU Benchmark

Methods	Input	Holistic TR	Holistic AR	Holistic VS	Single Detail NQA	Single Detail ER	Single Detail PQA	Single Detail SSC	Multi Detail AO	Multi Detail AC	M-Avg	G-Avg
Short Video MLLMs												
VideoChat [31]	16 f	33.0	32.0	2.31	27.0	32.1	27.6	5.01	24.3	28.6	29.2	3.66
Video-ChatGPT [ACL24] [45]	100 f	26.9	24.0	2.31	40.3	42.0	29.9	5.48	25.1	31.1	31.3	3.90
Video-LLaMA2 [9]	16 f	54.5	41.5	2.34	39.4	33.5	35.4	5.22	18.5	25.7	35.5	3.78
VideoChat2 [ICVPR 24] [32]	16 f	74.6	51.5	2.57	42.0	47.4	43.8	5.04	22.8	29.6	44.5	3.81
Video-LLaVA [36]	8 f	71.6	57.0	2.43	53.2	45.2	48.4	5.25	20.1	35.9	47.3	3.84
Long Video MLLMs												
MovieChat [CVPR 24] [60]	2048 f	29.5	25.0	2.33	24.2	24.7	25.8	3.23	28.6	22.8	25.8	2.78
Movie-LLM [62]	1 fps	30.0	29.0	2.88	29.6	24.7	24.1	5.00	20.5	24.8	26.1	3.94
TimeChat [CVPR 24] [55]	96 f	23.1	27.0	2.54	24.5	28.4	25.8	4.29	24.7	32.0	30.9	3.42
LLaMA-VID [ECCV 24] [35]	1 fps	50.8	34.5	3.22	30.1	32.7	32.5	5.22	23.9	27.8	33.2	4.22
MA-LMM [CVPR 24] [19]	1000 f	51.9	35.5	2.12	43.1	38.9	35.8	4.80	25.1	24.3	36.4	3.46
MiniGPT4-Video [3]	90 f	70.9	52.5	2.64	49.0	48.6	44.5	4.07	23.2	23.0	44.5	3.36
DynFocus ($L=25, K/L=0.8$)	16 f	75.4	60.5	3.36	50.6	42.3	50.5	5.34	26.2	32.6	48.3	4.35
DynFocus ($L=25, K/L=0.8$)	32 f	76.2	60.9	3.36	55.5	41.5	54.0	5.39	26.8	32.8	49.6	4.38
GPT-4o† [51]	0.5 fps	87.4	74.5	4.90	64.8	57.1	65.1	6.69	56.7	46.3	64.6	5.80

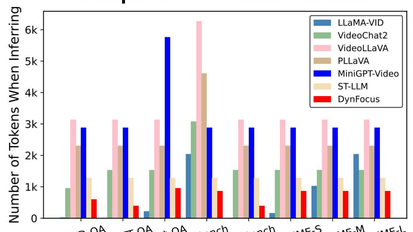
VideoMME Benchmark

Models	Input	LLM Size	Short (%) w/o subs	Short (%) w/ subs	Medium (%) w/o subs	Medium (%) w/ subs	Long (%) w/o subs	Long (%) w/ subs	Overall (%) w/o subs	Overall (%) w/ subs
LLaMA-VID [ECCV 24] [35]	1 fps	7B	-	-	-	-	-	-	25.9	-
Video-LLaVA [EMNLP 24] [37]	8 f	7B	45.3	46.1	38.0	40.7	36.2	38.1	39.9	41.6
ST-LLM [ECCV 24] [42]	16 f	7B	45.7	48.4	36.8	41.4	31.3	36.9	37.9	42.3
VideoChat2 [CVPR 24] [32]	16 f	7B	48.3	52.8	37.0	39.4	33.2	39.2	39.5	43.8
Chat-UniVi [CVPR 24] [25]	-	7B	45.7	51.2	40.3	44.6	35.8	41.8	40.6	45.9
DynFocus ($L=25, K/L=0.8$)	16 f	7B	50.9	53.7	43.7	46.0	37.7	43.6	44.1	47.8
LLaVA-NeXT† [79]	-	34B	61.7	65.1	50.1	52.2	44.3	47.2	52.0	54.9
VILA-1.5 [38]	-	34B	68.1	68.9	58.1	57.4	50.8	52.0	59.0	59.4

LVBench

Method	Size	Input	ER	EU	KIR	TG	Rea	Sum	Overall
Short Video MLLMs									
TimeChat [CVPR 24] [55]	7B	96 f	21.9	21.7	25.9	22.7	25.0	24.1	22.3
PLLaVA [73]	34B	16 f	25.0	24.9	26.2	21.4	30.0	25.9	26.1
LLaVA-NeXT [79]	34B	32 f	30.1	31.2	34.1	31.4	35.0	27.6	32.2
GPT-4o [51]	-	10 f	26.5	23.7	28.3	21.4	28.0	32.8	27.0
Long Video MLLMs									
MovieChat [CVPR 24] [60]	7B	~10k f	21.3	23.1	25.9	22.3	24.0	17.2	22.5
LLaMA-VID [ECCV 24] [35]	13B	~10k f	25.4	21.7	23.4	26.4	26.5	17.2	23.9
LWM [41]	7B	~4k f	24.7	24.8	26.5	28.6	30.5	22.4	25.5
Gemini 1.5 Pro [54]	-	~4k f	32.1	30.9	39.3	31.8	27.0	32.8	33.1
DynFocus† ($L=25, K/L=0.8$)	7B	200 f	27.9	30.3	31.2	25.4	31.8	32.8	30.4
DynFocus† ($L=50, K/L=0.8$)	7B	200 f	28.6	31.8	32.6	27.2	35.3	34.4	31.8
DynFocus† ($L=60, K/L=0.6$)	7B	200 f	29.9	33.7	35.1	25.5	33.3	26.2	32.6
DynFocus† ($L=70, K/L=0.4$)	7B	200 f	31.8	33.5	32.6	28.7	34.8	31.3	32.9
DynFocus† ($L=80, K/L=0.4$)	7B	200 f	31.1	33.5	31.6	28.6	33.8	24.1	31.8

Comparison of Token Number

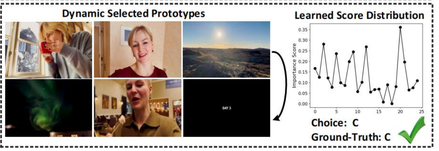


Ablation of Dynamic Encoding

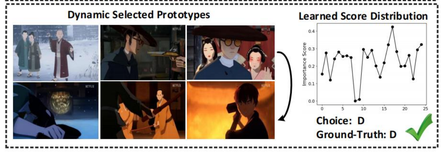
$\lvert U_{b_t=0}\rvert$	$\lvert U_{b_t=1}\rvert$	MSVD-QA Acc	MSVD-QA Score	ANet-QA Acc	ANet-QA Score	VCG-Bench Score
0	40	63.7	3.5	41.4	3.2	2.57
0	256	65.6	3.5	42.1	3.2	2.65
2	256	68.4	3.7	44.3	3.4	2.85
2	2	62.0	3.5	40.5	3.2	2.38
2	0	58.2	3.3	38.6	2.9	2.21
2	40	67.9	3.7	43.1	3.3	2.81

Dynamic Selection of DPE module

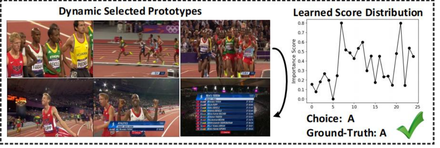
Question: What does the vlogger do on the third day?
Choice:
(A) She watches bread being made in a restaurant, visits a waterfall, eats at a restaurant, and then sees the aurora .
(B) She eats at a restaurant, watches bread being made, visits a river, and then sees the aurora .
(C) She watches bread being made by the sea, visits a waterfall, eats at a restaurant, and then sees the aurora.
(D) She enjoys the sea, visits a waterfall, eats at a restaurant and then sees the aurora.



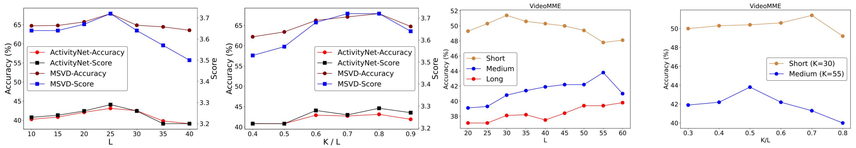
Question: What does the protagonist cut off from the man who she fights in the snow in the end?
Choice: (A) Leg (B) Cloth (C) Eye (D) Hair



Question: "What is the score of the champion?
Choice: (A) 27:30.42 (B) 27:30.90 (C) 29:32.12 (D) 27:01.17



Ablation of K and K/L



VideoHalluc Benchmark

Models	LLM Size	Object-Relation (%) Basic	Object-Relation (%) Halluc.	Object-Relation (%) Final	Temporal (%) Basic	Temporal (%) Halluc.	Temporal (%) Final	Semantic Detail (%) Basic	Semantic Detail (%) Halluc.	Semantic Detail (%) Final	Factual (%) Basic	Factual (%) Halluc.	Factual (%) Final	Non-Factual (%) Basic	Non-Factual (%) Halluc.	Non-Factual (%) Final	Overall
VideoChatGPT [31]	7B	95.5	7.0	6.0	100.0	0.0	0.0	96.5	4.0	2.0	86.5	13.5	7.0	85.5	27.5	17.0	6.4
LLaMA-VID [ECCV 24] [37]	7B	78.5	59.0	43.5	86.0	25.0	21.0	89.0	24.0	17.0	98.0	2.5	2.5	16.0	14.0	3.5	21.0
LLaMA-VID [ECCV 24] [37]	13B	87.5	55.5	44.5	78.5	35.0	27.0	90.5	30.0	25.5	85.0	17.5	12.5	84.5	46.5	36.5	23.5
Video-LLaMA2 [37]	7B	88.5	21.5	18.0	91.5	8.5	7.5	99.0	1.5	1.0	88.0	8.5	6.5	87.5	23.5	17.0	10.0
VideoChat2 [CVPR 24] [32]	7B	26.0	41.5	10.5	23.5	25.0	7.5	33.0	26.0	9.0	32.0	16.5	7.0	34.0	20.0	5.0	7.8
VideoLLaVA [EMNLP 24] [37]	7B	95.0	38.0	34.5	97.5	13.5	13.5	97.0	14.0	12.0	93.0	4.5	3.0	93.0	31.5	26.0	17.8
VideoLaVIT	-	94.5	39.0	35.5	88.5	27.0	25.5	96.5	13.0	10.5	97.5	6.0	4.0	97.5	21.5	19.0	18.9
MiniGPT4-Video [3]	7B	80.5	34.5	27.5	68.5	27.0	18.0	68.5	27.0	23.5	86.0	16.5	12.0	83.5	37.5	30.5	22.3
PLLaVA [73]	-	76.0	76.5	60.0	46.5	58.0	23.5	83.0	71.5	57.0	85.0	18.0	9.5	85.0	53.5	40.5	38.1
LLaVA-NeXT† [79]	7B	72.0	73.0	51.5	53.0	61.0	28.0	63.5	69.0	38.0	62.5	41.0	14.0	61.5	60.5	28.5	32.0
DynFocus ($L=25, K/L=0.8$)	7B	86.5	56.0	48.0	86.0	21.5	18.5	92.0	34.0	29.0	96.5	9.0	7.5	-	-	-	-
DynFocus† ($L=25, K/L=0.8$)	7B	88.0	62.0	52.5	87.0	37.5	33.5	91.5	42.0	38.5	98.5	15.0	13.0	96.5	40.0	38.5	35.1