Divid: Disentangled Spatial-Temporal Modeling within LLMs for Temporally Grounded Video Understanding

Yepeng Tang[1,2], Weining Wang[3,4], Longteng Guo[3,4], Tongtian Yue[3,4], Wenxuan Wang[3,4], Chunjie Zhang[1,2] ✉, Jing Liu[3,4] ✉

[1]Institute of Information Science, School of Computer Science and Technology, Beijing Jiaotong University,

[2]Visual Intelligence + X International Cooperation Joint Laboratory of MOE,

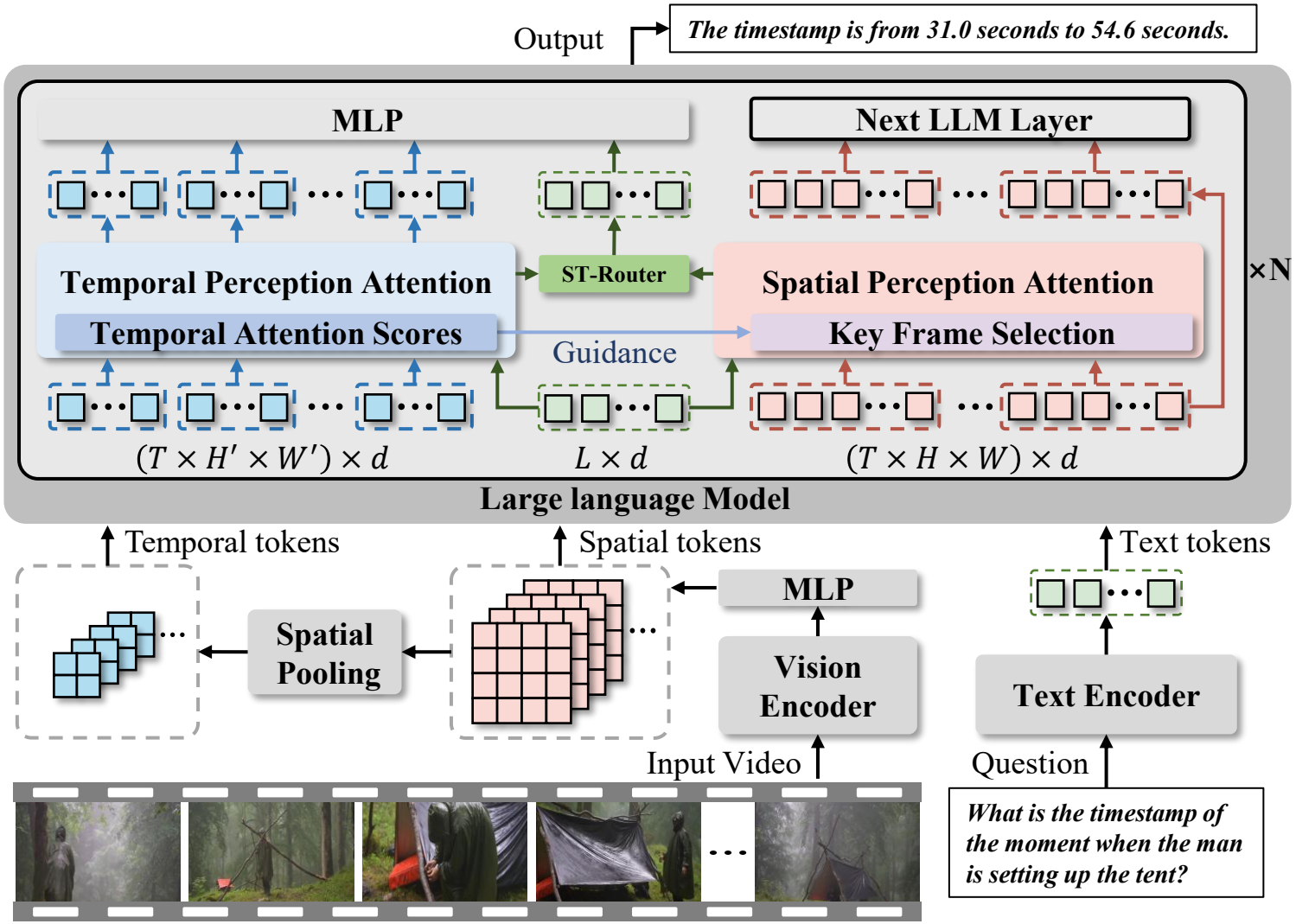
School of Computer Science and Technology, Beijing Jiaotong University,

[3]Institute of Automation, Chinese Academy of Sciences,

[4]University of Chinese Academy of Sciences

Divid: Disentangled Spatial-Temporal Modeling within LLMs for Temporally Grounded Video Understanding

- A spatiotemporal decoupling framework for efficient temporal understanding of long videos.
- By separating temporal perception from spatial perception, it effectively alleviates the computational overhead and information redundancy in long-video modeling.
- A temporal perception module is used to extract global temporal attention scores, which guide the selection of spatially critical information and improve the efficiency of spatial modeling.
- An ST-Router is introduced to enable the dynamic fusion of temporal and spatial information.



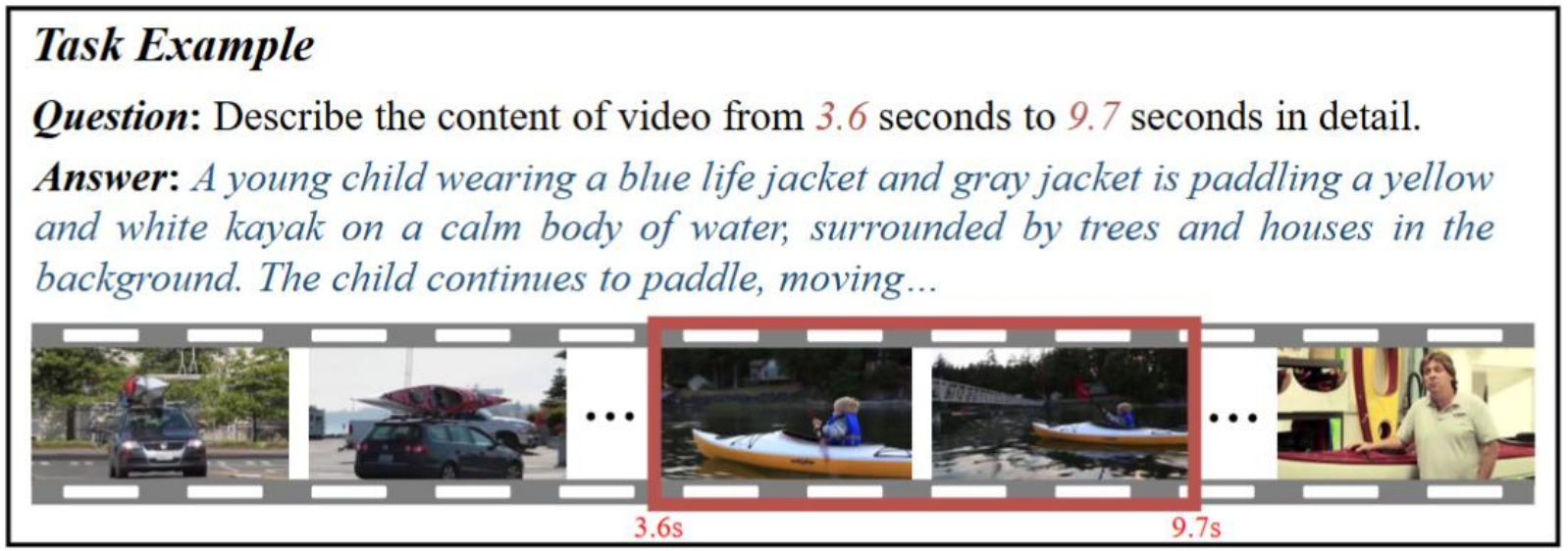
Divid: Disentangled Spatial-Temporal Modeling within LLMs for Temporally Grounded Video Understanding

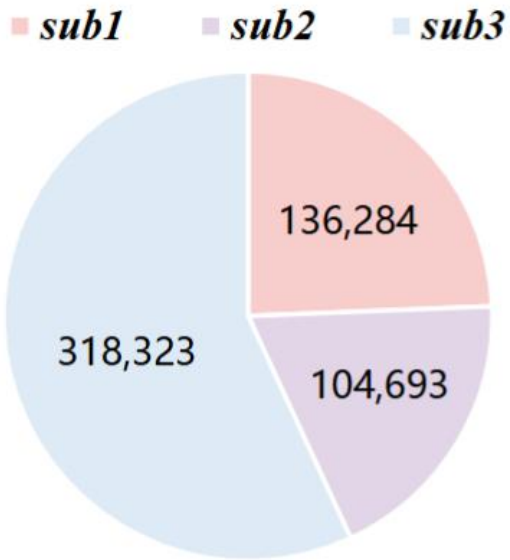
- **TempGCap-559K**: An efficient and scalable timestamp-guided captioning dataset.
- A training paradigm for temporal understanding in long videos, enabling models to learn precise temporal reasoning and cross-modal alignment.
- **Efficient construction pipeline:**

[1] Data reuse and efficient re-annotation: Reuse existing temporally annotated datasets with efficient relabeling.

[2] Context restoration and boundary refinement: Recover the original uncropped video context and refine temporal boundaries.

[3] Automatic pseudo-long-video synthesis: Concatenate multiple short video clips into long sequences automatically.





Divid: Disentangled Spatial-Temporal Modeling within LLMs for Temporally Grounded Video Understanding

Table 1: TempGCap dataset statistics grouped by annotation strategy and domain..

Name	Domain Category	#Videos	#Samples	Avg. Duration	Duration	Caption Len. (w)
			Sub1: Reannotating Untrimmed Videos with Manual Temporal Annotations			
DiDeMo	Open Domain	7,678	30,158	28.2 s	60.2 h	7.5
ActivityNet Captions	Human Activities / Events	10,009	37,421	117.2 s	325.8 h	13.5
HACS	Human Activities	21,552	68,705	139.3 s	834.1 h	86.3
Sub1 Sum.	-	39,239	136,284	111.9 s	1220.1 h	48.9
			Sub2: Recovering Untrimmed Contexts from Captioned Clips			
VATEX	Multilingual & Crowd Scenes	22,422	22,425	165.7 s	1031.9 h	82.1
ActivityNet	Human Activities	9,075	32,595	115.7 s	291.6 h	97.8
Kinetics-700	Human Actions	49,660	49,673	149.6 s	2063.9 h	81.5
Sub2 Sum	-	81,157	104,693	150.3 s	3387.4 h	86.7
			Sub3: Synthesized from Short Clips without Untrimmed Videos			
Oops	Accident / Unexpected Events	7,948	7,948	39.1 s	86.3 h	80.7
SomethingSomethingV2	Object-centric Human Actions	9,996	9,996	15.0 s	41.8 h	67.5
TREC-VTT	Web Video / Diverse Topics	14,199	14,199	25.2 s	99.4 h	81.0
LSMDC	Movie / Narrative Video	108,271	108,271	16.4 s	492.8 h	72.8
WebVid	Web-scale Open-domain	177,909	177,909	71.0 s	3503.6 h	71.3
Sub3 Sum	-	318,323	318,323	47.8 s	4223.9 h	72.3
Total	-	438,719	559,300	72.5 s	8831.4 h	69.3

Divid: Disentangled Spatial-Temporal Modeling within LLMs for Temporally Grounded Video Understanding

- Divid-1.5B outperforms most 7B models.
- Divid-7B surpasses Qwen2.5-VL-72B.
- Achieves state-of-the-art performance on temporal localization video question answering.

Table 1: Charades-STA [18] Dataset.

Method	Size	R@0.5	R@0.7	mIoU
(FT) Moment-DETR [42]	–	52.1	30.6	45.5
(FT) UniVTG [43]	–	58.1	35.6	50.1
(FT) R^2-Tuning [44]	–	59.8	37.0	50.9
Grounded-VideoLLM [26]	4B	36.4	19.7	36.8
E.T. Chat [24]	4B	45.9	20.0	42.3
Qwen2.5-VL [13]	3B	-	-	38.8
VideoMind [3]	1.5B	51.1	26.0	45.2
Divid (Ours)	1.5B	**51.4**	**26.9**	**47.3**
GPT-4o [45]	–	-	-	35.7
Qwen2.5-VL [13]	72B	-	-	50.9
VTimeLLM [25]	13B	34.3	14.7	34.6
TimeChat [20]	7B	32.2	13.4	–
Momentor [10]	7B	26.6	11.8	28.5
HawkEye [29]	7B	31.4	14.5	34.7
ChatVTG [46]	7B	33.0	15.9	34.9
VideoChat-TPO [47]	7B	40.2	20.8	38.1
VideoChat-T [21]	7B	48.7	24.0	-
VideoChat-Flash [23]	7B	53.1	27.6	–
Qwen2.5-VL [13]	7B	-	-	43.6
TimeMarker [22]	8B	51.9	26.9	48.4
TimeSearch [48]	7B	52.4	24.5	48.6
VideoMind [3]	7B	59.1	31.2	50.2
Divid (Ours)	7B	**59.5**	**31.3**	**51.3**

Table 2: Grounded VideoQA on CG-Bench [40].

Method	Size	mIoU	R@IoU	A@IoU
GPT-4o [45]	–	5.62	8.30	**4.38**
GPT-4o-mini [49]	–	3.75	5.18	2.21
Gemini-1.5-Pro [50]	–	3.95	5.81	2.53
Gemini-1.5-Flash [50]	–	3.67	5.44	2.45
Claude-3.5-Sonnet	–	3.99	5.67	2.79
Qwen2-VL [33]	72B	3.58	5.32	2.54
VITA [51]	8×7B	3.06	3.53	2.33
ShareGPT4Video [52]	16B	1.85	2.65	1.01
Video-CCAM [53]	14B	2.63	3.53	1.76
Chat-UniVi-v1.5 [54]	13B	2.07	2.53	1.20
LLaVA-OV [55]	13B	1.63	1.78	1.01
Video-LLaVA [4]	7B	1.13	1.96	0.59
VideoLLaMA [56]	7B	1.21	1.87	0.84
Videochat2 [57]	7B	1.28	1.98	0.94
Qwen-VL-Chat [58]	7B	0.89	1.19	0.42
ST-LLM [59]	7B	2.23	2.86	1.13
ViLA [60]	8B	1.56	2.89	1.35
MiniCPM-v2.6 [61]	8B	2.35	2.96	1.35
LongVA [62]	7B	2.94	3.86	1.78
Kangaroo [63]	8B	2.26	3.24	1.23
InternVL2 [64]	7B	3.91	5.05	2.64
Divid (Ours)	1.5B	3.47	4.81	1.29
Divid (Ours)	7B	**5.74**	**8.36**	4.11

Table 3: Grounded VideoQA on NExT-GQA [19].

Method	LLM Size	IoU R@0.3	IoU R@0.5	IoU mIoU	IoP R@0.3	IoP R@0.5	IoP mIoP	Acc@GQA
FrozenBiLM NG+ [65]	890M	13.5	6.1	9.6	28.5	23.7	24.2	17.5
VIOLETv2 [66]	–	4.3	1.3	–	25.1	23.3	23.6	12.8
SeViLA [67]	4B	29.2	13.8	21.7	34.7	22.9	29.5	16.6
LangRepo [68]	8×7B	–	12.2	18.5	–	20.8	27.1	14.9
VideoStreaming [69]	8.3B	–	13.3	19.4	–	22.0	30.8	18.1
LLoVi [70]	1.8T	15.3	20.0	–	36.6	37.3	24.3	24.3
HawkEye [29]	7B	37.0	19.5	25.7	44.7	31.6	33.4	23.5
Grounded-VideoLLM [26]	4B	30.2	18.0	21.1	42.6	34.4	34.5	26.7
VideoChat-TPO [47]	7B	41.2	23.4	27.7	47.5	32.8	35.6	25.5
VideoMind [3]	1.5B	45.2	23.2	28.6	51.3	30.6	35.6	25.2
VideoMind [3]	7B	50.2	25.8	31.4	**56.0**	35.3	39.0	28.2
Divid (Ours)	1.5B	47.5	26.8	32.9	52.0	33.7	38.9	26.4
Divid (Ours)	7B	**51.3**	**27.5**	**34.5**	54.6	**38.2**	**40.8**	**29.2**

Table 4: Grounded VideoQA on ReXTime [41].

Method	LLM Size	FT	R@0.3	R@0.5	mIoU	Acc@IoU
VTimeLLM [25]	7B	✓	43.69	26.13	29.92	17.13
TimeChat [20]	7B	✓	40.13	21.42	26.29	10.92
VTimeLLM [25]	7B	–	28.84	17.41	20.14	–
TimeChat [20]	7B	–	14.42	7.61	11.65	–
LITA [6]	13B	–	29.49	16.29	21.49	–
VideoMind [3]	1.5B	–	34.31	22.69	24.83	17.26
VideoMind [3]	7B	–	38.22	25.52	27.61	20.20
Divid (Ours)	1.5B	–	40.50	26.82	29.71	18.57
Divid (Ours)	7B	–	**42.56**	**31.05**	**35.78**	**22.26**

Divid: Disentangled Spatial-Temporal Modeling within LLMs for Temporally Grounded Video Understanding

- Reduces computational cost by over 60% while maintaining performance.
- Surpasses Momentor-1M with only 559K samples, demonstrating high data efficiency.

Table 5: Comparison with existing frameworks.

Method	LLM TFLOPs	Charades-STA R1@0.5	Charades-STA mIoU	ReXTime mIoU	ReXTime Acc@IoU
Full	28.2	51.55	47.76	30.16	19.11
Slow-Fast	16.2	50.65	46.71	29.49	18.45
Divid	10.5	51.37	47.33	29.71	18.57

Table 6: The impact of temporal guidance.

Method	Charades-STA R1@0.5	Charades-STA mIoU	ReXTime mIoU	ReXTime Acc@IoU
Uniform	49.70	46.27	27.61	17.32
Weighted	48.62	45.43	26.44	16.78
Top-K	51.37	47.33	29.71	18.57

Table 7: Ablation studies of ST-Router.

Method	Charades-STA R1@0.5	Charades-STA mIoU	ReXTime mIoU	ReXTime Acc@IoU
Add	50.66	46.81	29.11	18.23
Weight	50.98	47.13	29.31	18.42
Soft-Router	51.37	47.33	29.71	18.57

Table 8: Ablation studies of datasetss.

Method	Charades-STA R1@0.5	Charades-STA mIoU	ReXTime mIoU	ReXTime Acc@IoU
Momentor-1M [10]	48.96	44.90	26.77	15.43
Momentor [10]	50.47	46.42	28.75	16.78
TempGCap	51.37	47.33	29.71	18.57