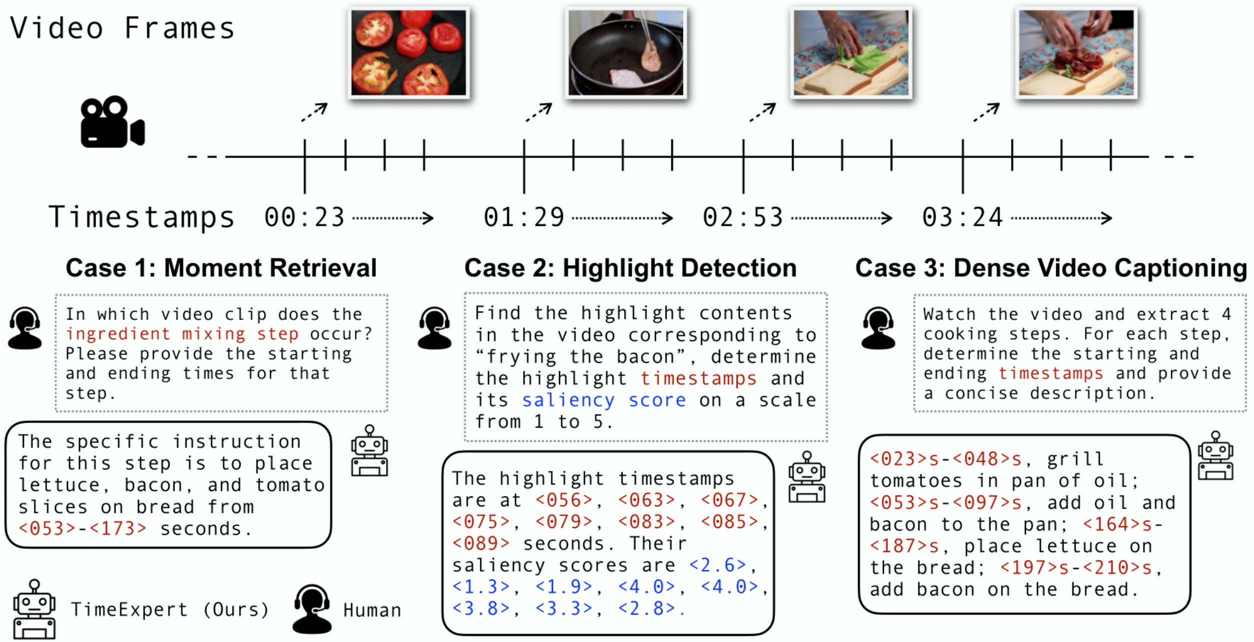
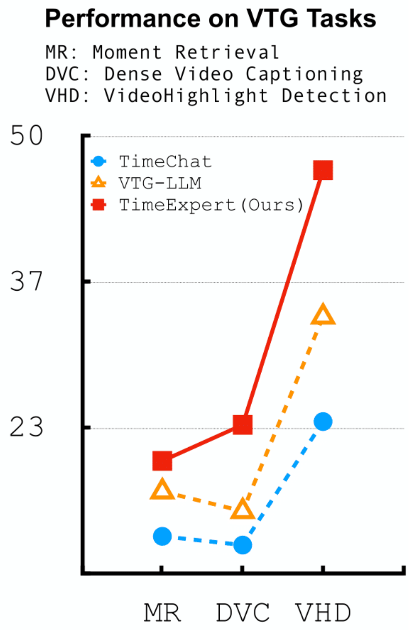
TimeExpert: An Expert-Guided Video LLM for Video Temporal Grounding

Zuhao Yang[1*] Yingchen Yu[2] Yunqing Zhao[2] Shijian Lu[1†] Song Bai[2]

[1]Nanyang Technological University [2]ByteDance Inc.
[*]Work was done while interning at ByteDance [†]Corresponding Author

Task & Performance





Video Temporal Grounding (VTG) Tasks

Moment Retrieval: localize segments matching a query.
Highlight Detection: rank and locate salient moments.
Dense Video Captioning: generate timestamped step descriptions.

Performance Comparison

Our approach demonstrates **substantial improvements** over state-of-the-art Video-LLMs on several VTG benchmarks. For example, here we visualize zero-shot F1 score for DVC on the YouCook2 dataset, $R@1_{IoU=0.7}$ for MR on the Charades-STA dataset, and HIT@1 for VHD on the QVHighlights dataset.

Background & Motivation

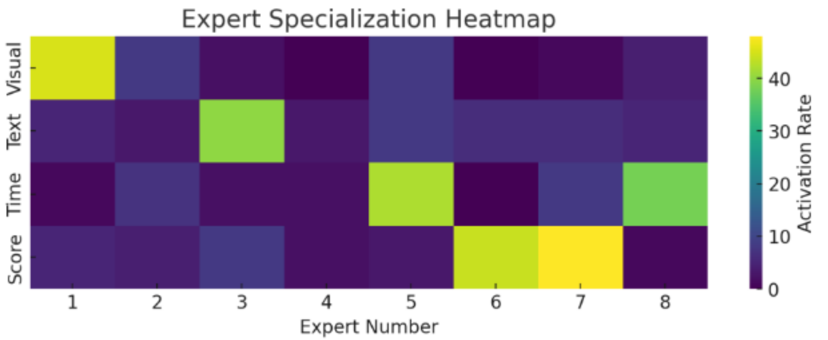


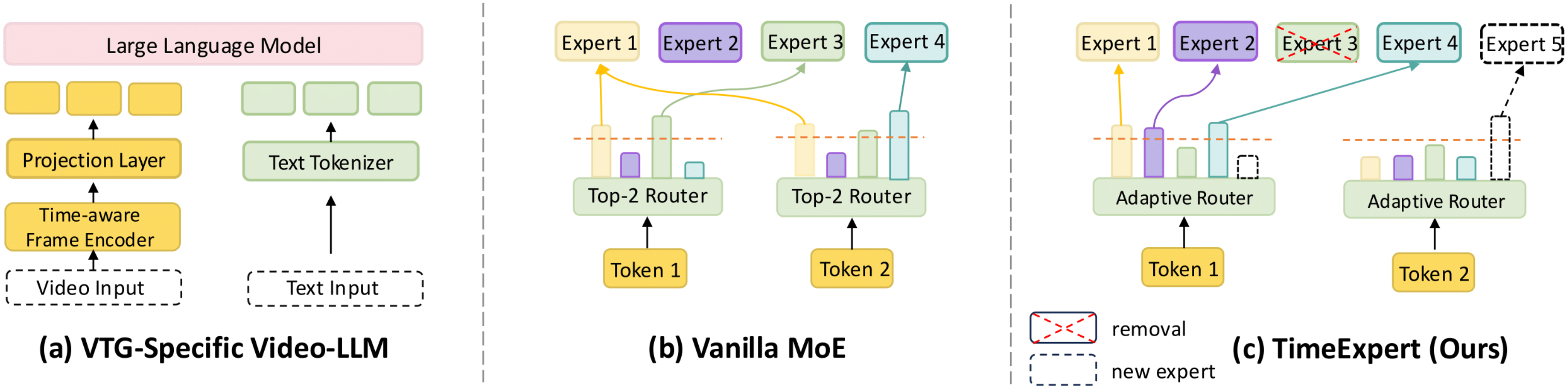
Figure 4. **Visualization of Expert Assignments on Various Task Tokens using Our Vanilla MoE Implementation.** We take layer 4 as an example and only visualized the first 8 experts out of a total of 64, due to the space limitation.

Existing Video-LLMs process all VTG tokens through the same pathway, ignoring the distinct nature of time, score and caption tokens.

Temporal boundaries and saliency rankings require different reasoning than caption generation. Without specialization, interference between tasks degrades performance.

TimeExpert proposes to recognize the importance of each task token and route them to specialised experts.

Method Comparison



VTG-specific Video-LLM

One shared model handles all tokens with limited specialization.

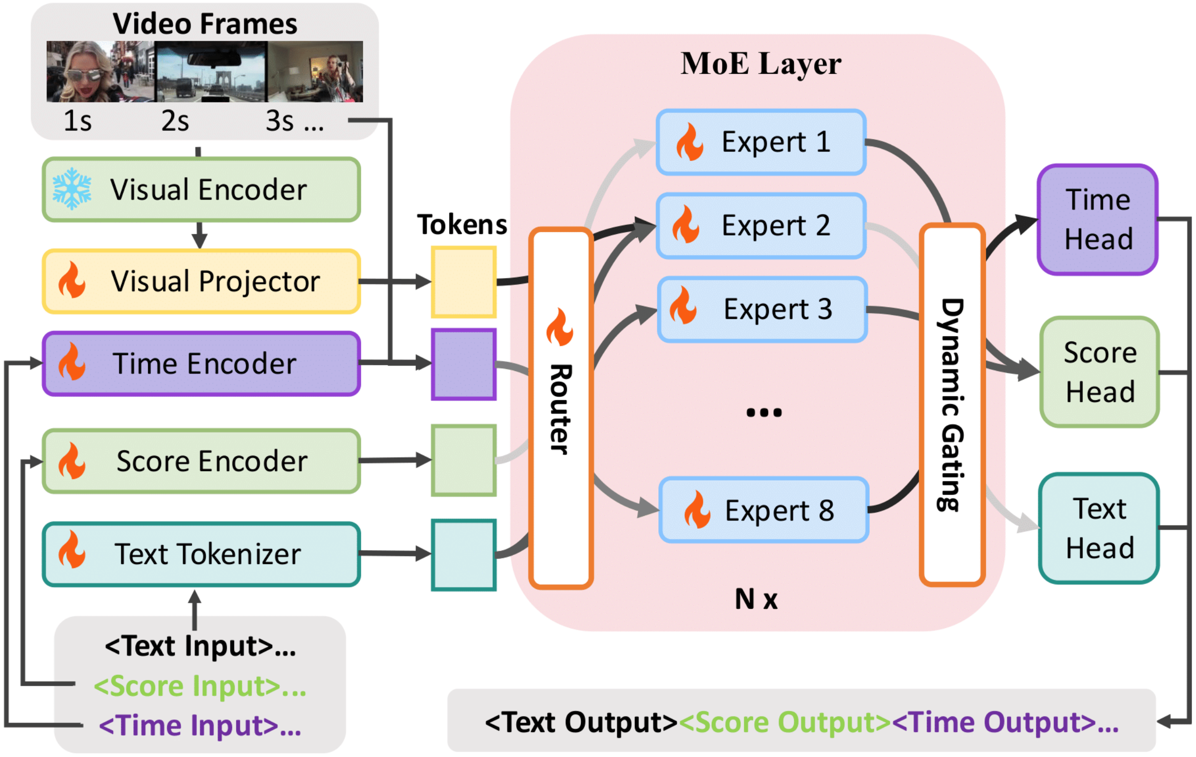
Vanilla MoE

Activates a fixed number of experts (e.g., k=2) irrespective of token type.

TimeExpert (Ours)

Dynamic routing allocates new experts when needed and prunes unused ones, adapting to token importance.

Framework Overview



Independent Encoders

Visual, time and score encoders compress frames into compact tokens.

Dynamic Gating

A router activates appropriate experts based on token importance.

Separate Heads

Separate time, score and text heads decode structured outputs.

Training Strategy

Training Stage	Datasets	No. of Samples
Stage 1: Task Module Pretraining	Valley [36], LLaVA-Image [33], TextVR [54], ShareGPT4Video [5], VTG-IT [17]	1.9M
Stage 2: MoE Decoder Pretraining	Valley*, TextVR*, ShareGPT4Video*, VTG-IT*, ActivityNet Captions [3], VideoChatGPT [37], InternVid [51], Next-QA [56]	0.9M
Stage 3: Supervised Fine-tuning	Filtered and Re-annotated Data from: Previous Stages' Data, EgoQA [40], STAR [53], Moment-10M [42], LLaVA-Video-178K [66]	2.3M

Table 1. **Training Data Recipe of *TimeExpert*.** These data are categorized into three training stages. $*$ denotes we utilize the filtered subset of the original dataset.

Task Module Pretraining
Learning core video representations with large-scale multimodal data.

→

MoE Decoder Pretraining
Aligning expert routing with task tokens to prevent expert collapse and enhance specialization.

→

Supervised Fine-tuning
Jointly optimizing task modules and MoE decoder on full-scale data.

Experimental Results

Method	No. of Activated Parameters	Dense Video Captioning (YouCook2 [67])			Moment Retrieval (Charades-STA [15])		Video Highlight Detection (QVHighlights [26])	
		SODA$_c$ (↑)	CIDEr (↑)	F1 Score (↑)	R@1$_{\text{IoU}=0.5}$ (↑)	R@1$_{\text{IoU}=0.7}$ (↑)	mAP (↑)	HIT@1 (↑)
TimeChat [43]	7B	1.2	3.4	12.6	32.2	13.4	14.5	23.9
VTimeLLM [21]	7B	–	–	–	27.5	11.4	–	–
Momentor [42]	7B	–	–	–	26.6	11.6	7.6	–
HawkEye [52]	7B	–	–	–	31.4	14.5	–	–
VTG-LLM [17]	7B	1.5	5.0	17.5	33.8	15.7	16.5	33.5
TRACE [18]	7B	2.2	<u>8.1</u>	22.4	40.3	19.4	26.8	42.7
TimeExpert (adaptive k)	≈ 5.9B / 3.5B / 4.8B	**2.5**	**8.2**	**23.6**	**42.8**	**20.3**	**29.6**	**46.9**
TimeExpert (TRACE's data)	≈ 5.2B / 3.1B / 4.0B	<u>2.4</u>	<u>8.1</u>	<u>23.3</u>	<u>41.9</u>	<u>20.1</u>	<u>29.1</u>	<u>46.3</u>

Table 2. **Zero-shot Performance Comparison of *TimeExpert* against several state-of-the-art VTG-specific Video-LLMs on Dense Video Captioning, Temporal Grounding, and Video Highlight Detection.** Some results are sourced from [18]. The best results are in **bold**. We also <u>underline</u> the second-best results.

1. Strong zero-shot performance across Dense Captioning, Moment Retrieval, and Highlight Detection.

2. Adaptive-k routing achieves the best results with fewer activated parameters, proving efficiency and expert specialization.

3. Consistent gains on all benchmarks (e.g., +4.2% HIT@1 on QVHighlights).

4. Data-efficient variant (TRACE's data) still outperforms baselines, showing strong generalization.

Conclusion

- Introduced dynamic expert routing to specialize processing of VTG tokens
- Achieved state-of-the-art results on multiple VTG benchmarks
- Scales efficiently by activating only necessary experts

Future Directions

- Incorporate audio inputs for more robust video temporal grounding
- Apply to other multimodal reasoning tasks with reinforcement fine-tuning

Thanks for watching this video!