Enhancing Temporal Modeling of Video LLMs via Time Gating

Zi-Yuan Hu, Yiwu Zhong, Shijia Huang, Michael R. Lyu, Liwei Wang[*]
Department of Computer Science and Engineering, The Chinese University of Hong Kong
{zyhu22,sjhuang,lyu,lwwang}@cse.cuhk.edu.hk   yiwuzhong@cse.cuhk.edu.hk

Abstract

Video Large Language Models (Video LLMs) have achieved impressive performance on video-and-language tasks, such as video question answering. However, most existing Video LLMs neglect temporal information in video data, leading to struggles with temporal-aware video understanding. To address this gap, we propose a **T**ime **G**ating **Vid**eo LLM (**TG-Vid**) designed to enhance temporal modeling through a novel **T**ime **G**ating module (**TG**). The TG module employs a time gating mechanism on its sub-modules, comprising gating spatial attention, gating temporal attention, and gating MLP. This architecture enables our model to achieve a robust understanding of temporal information within videos. Extensive evaluation of temporal-sensitive video benchmarks (*i.e.*, MVBench, TempCompass, and NExT-QA) demonstrates that our TG-Vid model significantly outperforms the existing Video LLMs. Further, comprehensive ablation studies validate that the performance gains are attributed to the designs of our TG module. Our code is available at https://github.com/LaVi-Lab/TG-Vid.

1 Introduction

The advancement of Large Language Models (LLMs) (Touvron et al., 2023; Chiang et al., 2023) has greatly fueled multi-modal research, such as Image LLMs (Liu et al., 2024b; Bai et al., 2023; Dai et al., 2023; Liu et al., 2024a) which have achieved success on image-and-language downstream tasks (Goyal et al., 2017). Inspired by Image LLMs, many recent efforts manage to empower LLMs to understand video data (Maaz et al., 2023; Li et al., 2023c; Liu et al., 2024d). The typical architecture of these Video LLMs comprises a pretrained vision encoder (Radford et al., 2021; Sun et al., 2023), a pretrained LLM (Chiang et al.,

2023), and a connection module in between (Zhu et al., 2023; Dai et al., 2023).

Despite the impressive performance demonstrated by Video LLMs (Maaz et al., 2023; Xu et al., 2017; Yu et al., 2019; Jang et al., 2017), a recent study (Liu et al., 2024e) reveals that most Video LLMs perform comparably to, or even worse than, Image LLMs. This discrepancy arises because existing video benchmarks can often be adequately addressed by single-frame bias (Lei et al., 2022; Buch et al., 2022), without the need for capturing the temporal dynamics of videos. To better evaluate the temporal modeling capability, multiple temporal-sensitive benchmarks have been developed (Liu et al., 2024e; Li et al., 2023c; Xiao et al., 2021) that cannot be solved by simply relying on single-frame bias as a shortcut.

In this paper, we aim to enhance the temporal modeling ability of Video LLMs and evaluate our model on the temporal-sensitive benchmarks. Specifically, we propose a temporal-aware Video LLM (**TG-Vid**) in this work, featuring a novel **T**ime **G**ating module (**TG**) to enhance temporal modeling. This TG module comprises three sub-modules, gating spatial attention, gating temporal attention, and gating MLP, simultaneously capturing spatial and temporal information. A recent relevant work ST-LLM (Liu et al., 2024d) also tries to enhance temporal modeling, by directly utilizing BT-Adapter (Liu et al., 2024c) which applies spatio-temporal attention in parallel to the vision encoder. In contrast, our work builds gating spatio-temporal attention on top of the vision encoder, and our gating mechanism imposes effective module-specific control over each sub-module of the TG module. As validated by experiments, our design achieves better performance on temporal-sensitive benchmarks.

We conduct comprehensive experiments on three temporal-sensitive video benchmarks (*i.e.*, MVBench (Li et al., 2023c), TempCompass (Liu
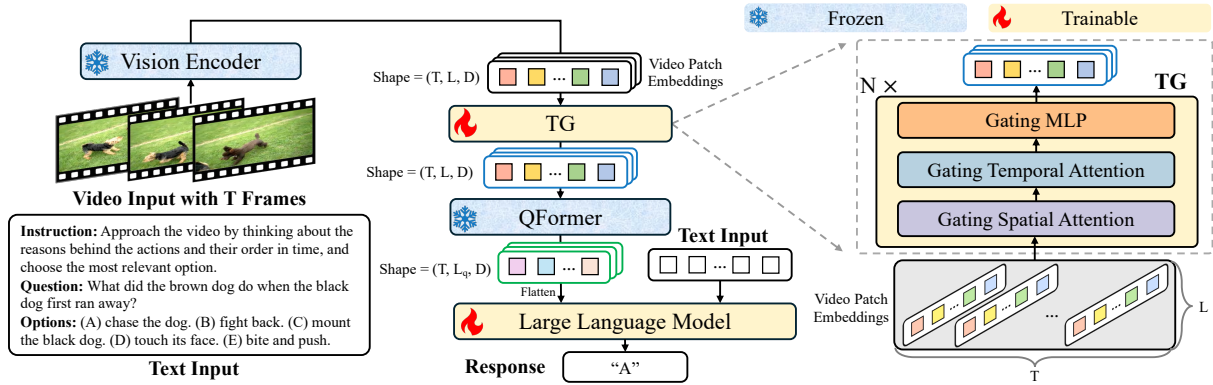
---

[*]Corresponding author.

Figure 1: **Model architecture of TG-Vid**. Given a video with T frames, the vision encoder extracts T frame-level embeddings. Our TG employs a novel time gating mechanism to enhance video temporal modeling, thereby enhancing the frame-level video modeling ability of the QFormer. Moving forward, the QFormer compresses each frame-level video embedding from L patch tokens to L_q query tokens, followed by LLMs.

et al., 2024e) and NExT-QA (Xiao et al., 2021)). The results show that our TG-Vid significantly outperforms the existing Video LLMs across all benchmarks and demonstrate the effectiveness of our TG-Vid on temporal-aware video understanding. The thorough ablation studies further emphasize that the performance gains are attributed to the designs of our TG module.

2 Related Work

Video Large Language Models. Benefited from the reasoning power of large language models (LLMs) (Zhang et al., 2022; Brown et al., 2020; Touvron et al., 2023; Chiang et al., 2023; Zhao et al., 2023), Video LLMs (Li et al., 2023b; Maaz et al., 2023; Zhang et al., 2023; Lin et al., 2023; Tang et al., 2023; Ren et al., 2024; Wang et al., 2024c; Tan et al., 2024) have shown impressive performance on video-and-language tasks, such as video question answering (Xu et al., 2017; Jang et al., 2017; Yu et al., 2019; Maaz et al., 2023; Xiao et al., 2021). However, most existing Video LLMs inherit the design of Image LLMs (Zhu et al., 2023; Liu et al., 2024b; Dai et al., 2023) and overlook the temporal modeling that is critical for video data, leading to unsatisfactory capability on temporal-aware video understanding (Li et al., 2023c; Liu et al., 2024e). For example, TempCompass (Liu et al., 2024e) reveals that the temporal understanding ability of most Video LLMs is on par with or even weaker than Image LLMs. In this work, we propose a temporal-aware Video LLM, featuring a new architecture of time gating module to enhance video temporal modeling.

Video Temporal Modeling. Modeling temporal information has been a long-standing topic in video research. Early work utilizes 3D convolutional networks (CNNs) to achieve spatio-temporal video modeling (Carreira and Zisserman, 2017; Feichtenhofer et al., 2016; Tran et al., 2015). To reduce training costs, subsequent CNN-based models explore factorizing convolutions across spatial and temporal dimensions (Sun et al., 2015; Tran et al., 2019, 2018; Xie et al., 2018; Feichtenhofer, 2020). Further, by leveraging the superiority of Transformer in processing sequences, TimesFormer (Bertasius et al., 2021) and ViViT (Arnab et al., 2021) employ Transformer-based architectures to enhance spatio-temporal modeling via spatial and temporal attention. Beyond single action, a line of work seeks to learn the temporal ordering of actions in procedural activities (Bojanowski et al., 2014; Chang et al., 2020; Zhao et al., 2022; Zhong et al., 2023). More recently, pretrained image-language models (Radford et al., 2021) are transferred to video tasks (Ni et al., 2022; Pan et al., 2022; Luo et al., 2022; Fang et al., 2021; Liu et al., 2024c), such as action recognition and video retrieval. Unlike these works, we extend the idea of spatio-temporal attention to Video LLMs, targeting at temporal-sensitive VideoQA and filling the gap of video modeling in Video LLMs.

3 Methodology

In this section, we introduce our **T**ime **G**ating **Vid**eo LLM (**TG-Vid**). Fig. 1 provides an overview of our model architecture. To enhance the temporal modeling of a Video LLM (comprising an LLM, a vision encoder, and a connection module), we

propose a **T**ime **G**ating (**TG**) module with a novel module-specific time gating mechanism.

3.1 Preliminary

Given a video input with T frames, a pretrained vision encoder (Sun et al., 2023) extracts patch embeddings for each frame and concatenates them into video embeddings $\mathbf{V} \in \mathbb{R}^{T \times L_V \times D_V}$, where L_V denotes the number of patch embeddings in each frame and D_V denotes the dimension of patch embeddings. On the other side, given the text input, we employ the text embedder of a pretrained LLM (Chiang et al., 2023) to obtain the text embeddings $\mathbf{T} \in \mathbb{R}^{L_T \times D_T}$, where L_T denotes the number of text tokens and D_T denotes the dimension of the text embeddings. This video and text encoding process is common in Video LLMs methods (Li et al., 2023c; Wang et al., 2024b; Liu et al., 2024d).

Our model design extends spatio-temporal attention from ViViT (Arnab et al., 2021) and Times-Former (Bertasius et al., 2021). We provide the background knowledge of spatio-temporal attention. For clarity, we first formulate the vanilla N-layer **S**patio-**T**emporal module (**ST**), which is placed between the vision encoder and the QFormer. Each ST layer comprises a **spatial attention**, a **temporal attention**, and a two-layer **MLP**. Given the input $\mathbf{V}^\ell \in \mathbb{R}^{T \times L_V \times D_V}$, the ℓ-th layer of ST ($\mathbf{V}^0$ is set as $\mathbf{V}$) can be formulated as:

$$\mathbf{V}_S^\ell = \text{ReshapeS}(\mathbf{V}^\ell) \tag{1}$$

$$\mathbf{Y}_S^\ell = \text{MSA}(\text{LN}(\mathbf{V}_S^\ell)) + \mathbf{V}_S^\ell \tag{2}$$

$$\mathbf{V}_T^\ell = \text{ReshapeT}(\mathbf{Y}_S^\ell) \tag{3}$$

$$\mathbf{Y}_T^\ell = \text{MSA}(\text{LN}(\mathbf{V}_T^\ell)) + \mathbf{V}_T^\ell \tag{4}$$

$$\mathbf{V}_M^\ell = \text{ReshapeM}(\mathbf{Y}_T^\ell) \tag{5}$$

$$\mathbf{V}^{\ell+1} = \mathbf{Y}_M^\ell = \text{MLP}(\text{LN}(\mathbf{V}_M^\ell)) + \mathbf{V}_M^\ell \tag{6}$$

where $\text{LN}(\cdot)$ denotes layer normalization (Ba et al., 2016), $\text{MSA}(\cdot)$ denotes multi-head self-attention, and $\text{MLP}(\cdot)$ denotes a two-layer MLP. $\text{ReshapeS}(\cdot)$ reshapes $\mathbf{V}^\ell \in \mathbb{R}^{T \times L_V \times D_V}$ as $\mathbf{V}_S^\ell \in \mathbb{R}^{T \times L_V \cdot D_V}$, $\text{ReshapeT}(\cdot)$ reshapes $\mathbf{Y}_S^\ell \in \mathbb{R}^{T \times L_V \cdot D_V}$ as $\mathbf{V}_T^\ell \in \mathbb{R}^{L_V \times T \cdot D_V}$, and $\text{ReshapeM}(\cdot)$ reshapes $\mathbf{Y}_T^\ell \in \mathbb{R}^{L_V \times T \cdot D_V}$ as $\mathbf{V}_M^\ell \in \mathbb{R}^{T \times L_V \times D_V}$.

3.2 Time Gating Module (TG)

The vanilla ST module can model the spatio-temporal information in video inputs. However, directly inserting a randomly initialized ST module into Video LLM results in unstable training and sub-optimal performance. To address this issue,

we propose a novel **T**ime **G**ating Module (**TG**), featuring a **time gating mechanism** to impose constraints on each sub-module (*i.e.*, a **gating spatial attention**, a **gating temporal attention**, and a **gating MLP**) of the TG module. These gating sub-modules allow our TG to focus dynamically on relevant information in both spatial and temporal aspects, enhancing the temporal modeling ability of Video LLM.

Unlike previous research works (Sung et al., 2022; Liu et al., 2024c) that utilize gating mechanism conditioned solely on a trainable but module-agnostic scalar (*e.g.*, $\alpha \in \mathbb{R}^1$) or vector (*e.g.*, $\boldsymbol{b} \in \mathbb{R}^{D_V}$), the gating function $\text{Gating}(\cdot)$ in our TG is module-specific and conditioned on both the input and output of the sub-module. Specifically, **gating spatial attention** is implemented as:

$$\hat{\mathbf{Y}}_S^\ell = \text{MSA}(\text{LN}(\mathbf{V}_S^\ell))$$
$$\mathbf{Y}_S^\ell = \text{Gating}(\mathbf{V}_S^\ell, \hat{\mathbf{Y}}_S^\ell) + \mathbf{V}_S^\ell \tag{7}$$
$$= \sigma(\text{Cat}(\mathbf{V}_S^\ell, \hat{\mathbf{Y}}_S^\ell)\mathbf{W}_S) \odot \hat{\mathbf{Y}}_S^\ell + \mathbf{V}_S^\ell$$

where $\sigma(\cdot)$ is a sigmoid function, $\text{Cat}(\cdot)$ denotes concatenate operation, $\mathbf{W}_S \in \mathbb{R}^{2 \cdot D_V \times D_V}$ is a linear projection, and $\odot$ denotes element-wise product. Similarly, **gating temporal attention** and **gating MLP** are implemented as follows:

$$\hat{\mathbf{Y}}_T^\ell = \text{MSA}(\text{LN}(\mathbf{V}_T^\ell))$$
$$\mathbf{Y}_T^\ell = \sigma(\text{Cat}(\mathbf{V}_T^\ell, \hat{\mathbf{Y}}_T^\ell)\mathbf{W}_T) \odot \hat{\mathbf{Y}}_T^\ell + \mathbf{V}_T^\ell \tag{8}$$

$$\hat{\mathbf{Y}}_M^\ell = \text{MLP}(\text{LN}(\mathbf{V}_M^\ell))$$
$$\mathbf{Y}_M^\ell = \sigma(\text{Cat}(\mathbf{V}_M^\ell, \hat{\mathbf{Y}}_M^\ell)\mathbf{W}_M) \odot \hat{\mathbf{Y}}_M^\ell + \mathbf{V}_M^\ell \tag{9}$$

where $\mathbf{W}_T \in \mathbb{R}^{2 \cdot D_V \times D_V}$ and $\mathbf{W}_M \in \mathbb{R}^{2 \cdot D_V \times D_V}$.

3.3 Time Gating Video LLM

By inserting the N-layer TG module between the frozen vision encoder and the frozen QFormer, we propose **TG-Vid**, a **T**ime **G**ating **Vid**eo LLM. The output video embeddings of the pretrained QFormer are flattened as $\mathbf{V}_Q \in \mathbb{R}^{T \cdot L_q \times D_V}$, where L_q denotes the length of query tokens for each frame. Subsequently, $\mathbf{V}_Q$ is projected into the text embedding space and concatenated with the text embedding $\mathbf{T}$ as follows:

$$\mathbf{VT} = [\mathbf{V}_Q \mathbf{W}_{VT}, \mathbf{T}] \tag{10}$$

where $\mathbf{W}_{VT} \in \mathbb{R}^{D_V \times D_T}$ is a trainable linear projection, and $\mathbf{VT} \in \mathbb{R}^{(T \cdot L_q + L_T) \times D_T}$ is the input into the LLM. Same as previous Video LLMs, our TG-Vid model is trained on next token prediction.

Model	Otter-V (Li et al., 2023a)	mPLUG-Owl (Ye et al., 2023b)	Video-ChatGPT (Maaz et al., 2023)	Video-LLaMA (Zhang et al., 2023)	VideoChat (Li et al., 2023b)	VideoChat2 (Li et al., 2023c)	HawkEye (Wang et al., 2024b)	ST-LLM (Liu et al., 2024d)	TG-Vid	TG-Vid
LLM	LLaMA-7B	LLaMA-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B
#IT	-	-	-	-	-	1.9M	2.2M	220K	197K	220K
Avg	26.8	29.7	32.7	34.1	35.5	51.1	47.6	54.9	<u>56.0</u>	**56.4**

Table 1: **MVBench benchmark experiments.** Comprehensive results are provided in the Appendix Tab. 9. #IT denotes instruction tuning samples. "Vi-" denotes "Vicuna-". **Bold**/<u>underline</u> denotes the best/second-best result.

Model	V-LLaVA (Lin et al., 2023)	LLaMA-VID (Li et al., 2023d)	mPLUG-Owl (Ye et al., 2023b)	PandaGPT (Su et al., 2023)	Valley (Luo et al., 2023)	VideoChat2 (Li et al., 2023c)	V-ChatGPT (Maaz et al., 2023)	V-LLaMA (Zhang et al., 2023)	ST-LLM♣ (Liu et al., 2024d)	TG-Vid	TG-Vid
LLM	Vi-7B	Vi-7B	LLaMA-7B	Vi-13B	StableVi-7B	Vi-7B	Vi-7B	Vi-13B	Vi-7B	Vi-7B	Vi-7B
#IT	-	-	-	-	-	1.9M	-	-	220K	197K	220K
Avg(Caption Matching)	63.7	53.6	49.3	51.3	22.0	55.6	51.8	53.5	64.8	**67.6**	<u>67.5</u>
Avg(Yes/No QA)	56.4	53.0	54.4	51.8	53.5	<u>58.0</u>	50.7	53.7	54.0	**58.1**	56.8
Avg(Multi-Choice QA)	44.7	35.3	40.0	31.1	31.8	51.1	35.2	33.9	<u>53.7</u>	52.9	**54.4**
Avg(ALL)	54.9	47.3	47.9	44.7	35.8	54.9	45.9	47.0	57.5	<u>59.5</u>	**59.6**

Table 2: **TempCompass benchmark experiments.** Comprehensive results are provided in the Appendix Tab. 10. #IT denotes instruction tuning samples. "V-" denotes "Video-" and "Vi-" denotes "Vicuna-". Avg(ALL) is the overall average result, calculated as the average of Avg(Caption Matching), Avg(Yes/No QA), and Avg(Multi-Choice QA). **Bold**/<u>underline</u> denotes the best/second-best average result. ♣: We reproduce the training and inference.

4 Experiments

Compared with existing Video LLMs, we evaluate our TG-Vid on three temporal-sensitive video understanding benchmarks (*i.e.*, MVBench (Li et al., 2023c), TempCompass (Liu et al., 2024e), and NExT-QA (Xiao et al., 2021; Buch et al., 2022)). More details of datasets, implementation, experiment results and visualization are provided in the Appendix.

4.1 Main Results

Tab. 1, Tab. 2, and Tab. 3 show our main results on MVBench, TempCompass, and NExT-QA, respectively. Our TG-Vid model achieves the best performance and surpasses previous methods by a large margin across all benchmarks. For example, compared with the closest competitor ST-LLM, our TG-Vid-220K achieves +1.5 on MVBench, +2.1 on TempCompass, +3.2 on NExT-QA ATP-hard, and +3.2 on NExT-QA Val. These impressive results demonstrate a consistent finding that our TG-Vid model can capture temporal information more effectively, attributed to the TG designs.

4.2 Ablation Studies

Given the comparable performance of TG-Vid-220K and TG-Vid-197K, the ablation studies are based on the latter for efficiency consideration.

TG Module. In Tab. 4, row 3 significantly outperforms row 1 by a large margin (+3.0), demonstrating the effectiveness of our TG module in empowering temporal-aware video understanding.

Model	#IT	NExT-QA ATP-hard			NExT-QA Val			
		Acc@C	Acc@T	Acc@All	Acc@C	Acc@T	Acc@D	Acc@All
VFC♠ (Yang et al., 2021)	-	32.2	30.0	31.4	49.6	51.5	63.2	52.3
ATP (Buch et al., 2022)	-	38.4	36.5	38.8	53.1	50.2	66.8	54.3
GF (Bai et al., 2024)	-	48.7	50.3	49.3	56.9	57.1	70.5	58.8
SeViT (Kim et al., 2023)	-	43.3	46.5	-	54.0	54.1	71.3	56.7
HiTeA (Ye et al., 2023a)	-	47.8	48.6	-	62.4	58.3	75.6	63.1
VideoAgent♠ (Wang et al., 2024a)	-	57.8	58.8	58.4	72.7	64.5	81.1	71.3
SEVILA (Yu et al., 2024)	-	-	-	-	74.2	69.4	<u>81.3</u>	73.8
VideoChat2 (Li et al., 2023c)	1.9M	-	-	-	68.7	64.7	76.1	68.6
HawkEye (Wang et al., 2024b)	2.2M	-	-	-	-	-	-	67.9
ST-LLM♣ (Liu et al., 2024d)	220k	65.5	61.9	64.0	74.3	70.0	<u>81.3</u>	74.0
TG-Vid	197k	<u>68.4</u>	66.3	67.5	**77.4**	**73.8**	**84.3**	**77.3**
TG-Vid	220K	**68.5**	<u>65.2</u>	<u>67.2</u>	<u>77.3</u>	<u>73.5</u>	**84.3**	<u>77.2</u>

Table 3: **Experiments on NExT-QA ATP-hard subset and NExT-QA validation dataset.** C, T, and D are causal, temporal, and descriptive subsets, respectively. **Bold**/<u>underline</u> denotes the best/second-best result. ♣: We reproduce the training and inference. ♠: Zero-shot.

TG Components			Gating Mechanism	#IT	MVBench Avg
Spatial	Temporal	MLP			
✗	✗	✗	✗	197K	53.0
✓	✓	✓	✗	197K	54.5
✓	✓	✓	✓	197K	**56.0**
✗	✓	✓	✓	197K	55.6
✓	✗	✓	✓	197K	54.7
✓	✓	✗	✓	197K	55.7

Table 4: **Ablation studies on TG module.**

Time Gating Mechanism. Row 3 significantly surpasses row 2 (+1.5), underscoring the crucial role of the time gating mechanism in enhancing video temporal modeling.

TG Components. The results in Tab. 4 indicate that each sub-module of TG module contributes to performance improvement. Notably, the proposed gating temporal attention provides the most significant enhancement (from 54.7 to 56.0), further validating the necessity of temporal modeling.

5 Conclusion

In this paper, we focus on developing a Video LLM, **TG-Vid**, to overcome the struggles of the existing Video LLMs in temporal-aware video understanding. Specifically, we propose a novel **T**ime **G**ating module (**TG**) with a time gating mechanism, to enhance the temporal modeling ability of TG-Vid. Comprehensive experiments and ablation studies conducted on three temporal-sensitive benchmarks (*i.e.*, MVBench, TempCompas, and NExT-QA) indicate that TG-Vid outperforms the existing Video LLMs by a large margin. These results demonstrate the effectiveness of our TG design in enhancing temporal modeling, thereby empowering our TG-Vid with a strong ability of temporal-aware video understanding.

Limitations. Our proposed TG-Vid model has achieved strong performance on the temporal-sensitive video understanding benchmarks. However, there are still some limitations: (1) Despite that our TG module can significantly enhance the temporal modeling of the Video LLM, integrating it into Video LLM requires additional computation; (2) Similar to the existing Video LLMs, our TG-Vid model has the potential to inherit the undesired biases from the training dataset and the pretrained LLMs; (3) The focus of this work is on temporal modeling. Whether the proposed TG-Vid model and the TG module can be generalized to other video-and-language tasks, such as long video understanding, is worth exploring in future research.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6816–6826. IEEE.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Ziyi Bai, Ruiping Wang, and Xilin Chen. 2024. Glance and focus: Memory prompting for multi-event video question answering. *Advances in Neural Information Processing Systems*, 36.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 813–824. PMLR.

Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. 2014. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2907–2917.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. 2020. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*.

Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213.

Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. 2016. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3468–3476.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Sungdong Kim, Jin-Hwa Kim, Jiyoung Lee, and Minjoon Seo. 2023. Semi-parametric video-grounded text generation. *arXiv preprint arXiv:2301.11507*.

Jie Lei, Tamara L Berg, and Mohit Bansal. 2022. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *CoRR*, abs/2305.03726.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2023c. Mvbench: A comprehensive multi-modal video understanding benchmark. *CoRR*, abs/2311.17005.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023d. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ruyang Liu, Chen Li, Yixiao Ge, Thomas H Li, Ying Shan, and Ge Li. 2024c. Bt-adapter: Video conversation is feasible without video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2024d. ST-LLM: large language models are effective temporal learners. *CoRR*, abs/2404.00308.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024e. Tempcompass: Do video llms really understand videos? *CoRR*, abs/2403.00476.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer.

Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. 2022. St-adapter: Parameter-efficient image-to-video transfer learning. *Advances in Neural Information Processing Systems*, 35:26462–26477.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. 2015. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. EVA-CLIP: improved training techniques for CLIP at scale. *CoRR*, abs/2303.15389.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Lst: Ladder side-tuning for parameter and memory efficient transfer learning. *Advances in Neural Information Processing Systems*, 35:12991–13005.

Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. 2024. Koala: Key frame-conditioned long video-llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13581–13591.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. 2019. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5552–5561.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024a. Videoagent: Long-form video understanding with large language model as agent. *arXiv preprint arXiv:2403.10517*.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. 2024b. Hawk-eye: Training video-text llms for grounding text in videos. *CoRR*, abs/2403.10228.

Yuxuan Wang, Yueqian Wang, Pengfei Wu, Jianxin Liang, Dongyan Zhao, and Zilong Zheng. 2024c. Lstp: Language-guided spatial-temporal prompt learning for long-form video-text understanding. *arXiv preprint arXiv:2402.16050*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. 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.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321.

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

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697.

Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2023a. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2024. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. 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.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 543–553. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Henghui Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G. Derpanis, Richard P. Wildes, and Allan D. Jepson. 2022. P3IV: Probabilistic procedure planning from instructional videos with weak supervision. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2928–2938.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. 2023. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14825–14835.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

In the appendix, we provide more details of: (1) the statistics of the training dataset; (2) the implementation details and hyper-parameters for training; (3) additional ablation study on the number of layers of the proposed TG module; (4) additional ablation study on the training output format; (5) additional ablation study on the designs of TG module; (6) the comprehensive results of MVBench (comprising twenty sub-tasks) and TempCompass (comprising three sub-tasks); (7) the visualization of time gating.

A Training Dataset Statistics

To ensure a fair comparison with the state-of-the-art Video LLM, ST-LLM (Liu et al., 2024d), our TG-Vid-220K model utilizes the same training dataset as ST-LLM, as detailed in Tab. 5. For the training dataset of our TG-Vid-197K model, we filter out the Conversation-VideoChatGPT and VQA-WebVidQA datasets to improve training efficiency.

Moreover, the training dataset with 220K video-text pairs is also a subset of the training dataset of VideoChat2 (Li et al., 2023c), which contains 1.9M video-text pairs.

B Implementation Details

Following ST-LLM, we adopt the Vicuna-7B-v1.1 (Chiang et al., 2023) as our pretrained LLM and the EVA-ViT-g (Sun et al., 2023) as our pretrained vision encoder. The QFormer is also initialized from the pretrained InstructBLIP (Dai et al., 2023), while our TG module is randomly initialized. Following the designs of LLaMA (Touvron et al., 2023), the self-attention inside the TG module is implemented as self-attention with rotary position embeddings (RoPE) (Su et al., 2024). Similarly, the activation function of MLP inside the TG module is implemented as the SwiGLU activation function. Our TG-Vid model is subsequently trained on the video instruction tuning dataset, which is described in Appendix A.

During training, the vision encoder and the QFormer are frozen, while other modules of TG-Vid are trainable. The training of TG-Vid-197K costs about 7 hours and the training of TG-Vid-220K costs about 13 hours. Both of these trainings are conducted on 8 A100 GPUs (each GPU has 80G memory).

Category	Training Dataset	#Video-Text Pairs
Conversation	VideoChatGPT$^\diamond$	13,303
Classification	Kinetics-710	40,000
Classification	SthSthV2	40,000
Reasoning	NExTQA	34,132
Reasoning	CLEVRER_QA	40,000
Reasoning	CLEVRER_MC	42,620
VQA	WebVidQA$^\diamond$	10,000
Total	-	220055

Table 5: The statistics of the training dataset with 220K video-text pairs. $\diamond$ denotes the datasets that are filtered out in the training dataset with 197K video-text pairs.
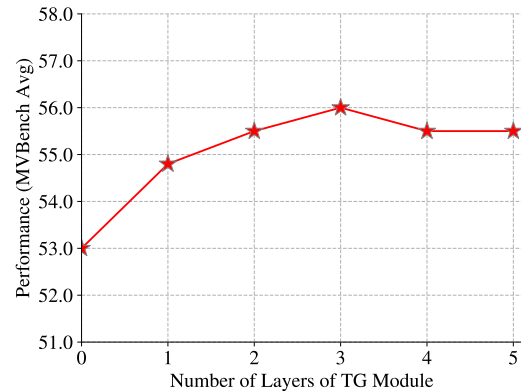
Figure 2: Ablation study on the number of layers of the TG module based on TG-Vid-197K.

Moreover, TG-Vid-197K and TG-Vid-220K share the same hyper-parameters for training, which are listed in Tab. 6.

C Number of Layers of TG

As shown in Fig. 2, we ablate the depth of the TG module. The results reveal that all of the models with the TG module significantly surpass the model without the TG module, demonstrating the effectiveness of our proposed TG module in empowering temporal-aware video understanding. Moreover, the results also indicate that the 3-layer TG module achieve the best performance. Therefore, we use $N = 3$ in Tab. 6 by default.

D Training Output Format

To improve the training efficiency for LLM decoding, we explore modifying the training output format. The original output format form VideoChat2-IT (Li et al., 2023c) is "(A) chase the dog.", while ours is modified as a direct output format: "A". Following the instruction used in LLaVA1.5 (Liu et al.,

Hyper-parameter	TG-Vid Training
Number of Layers of TG	$N = 3$
Training Epochs	2
Btach Size	64
Input Frame	16
Input Resolution	224×224
Max Text Length	256
Max Model Length	1024
Optimizer	AdamW
Optimizer Momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Weight Decay	0.0
Learning Rate Schedule	Cosine
Learning Rate	2e-5
Warmup Ratio	0.03
QFormer Query Token Length	$L_q = 32$

Table 6: Hyper-parameters for TG-Vid training.

Model	Training Output Format	MVBench Avg
TG-Vid-197K	"(A) chase the dog."	56.3
TG-Vid-197K	"A"	56.0

Table 7: Experiments on different training output formats.

Model	Self-Attention with RoPE	MLP with SwiGLU	MVBench Avg
TG-Vid-197K	✗	✓	56.3
TG-Vid-197K	✓	✗	56.0
TG-Vid-197K	✓	✓	56.0

Table 8: Experiments on the designs borrowed from LLMs.

2024a), we also add an instruction "Answer with the option's letter from the given choices directly." in the text input.

As shown in Tab. 7, the performance after modification is slightly decreased but the performance is still comparable. Therefore, we utilize the direct output format as our training output format for efficiency consideration.

E Designs Borrowed from LLaMA

As mentioned in Appendix B, we borrow two designs (*i.e.*, RoPE and SwiGLU) from LLaMA into the implementation of the TG module. In this section, we ablate these designs, as shown in Tab. 8. The results demonstrate that these designs can slightly improve the model performance. Therefore, we introduce these designs into the implementation of our TG module.

F Comprehensive Results of MVBench and TempCompass.

Due to the space limitation of the main paper, we present the comprehensive results of MVBench (comprising twenty sub-tasks) and TempCompass in (comprising three sub-tasks) in Tab. 9 and Tab. 10, respectively.

G Visualization of Time Gating.

To gain more insight into how time gating works, we provide some visualizations in xx. To be specific, we visualize the heatmap of the gate values produced by our time gating mechanism, based on the following steps:

- Step 1: Randomly select an input sample (with a T-frame video).

- Step 2: For the input sample, obtain its gate values of the gating temporal attention in the first layer of TG (*i.e.*, the result of the sigmoid function in Equation (8)), denoted as $\mathbf{G} \in \mathbb{R}^{L_{\mathrm{V}} \times T \cdot D_{\mathrm{V}}}$. L_{V} denotes the number of patch embeddings in each frame and D_{V} denotes the dimension of patch embeddings.

- Step 3: Visualize the gate values. Specifically, reshape $\mathbf{G}$ as $\mathbb{R}^{L_{\mathrm{V}} \times T \times D_{\mathrm{V}}}$ and adopt average-pooling along the D_{V} dimension to obtain $\hat{\mathbf{G}} \in \mathbb{R}^{L_{\mathrm{V}} \times T}$. Finally, visualize $\hat{\mathbf{G}}$ with a heatmap visualization H.

To conduct a detailed analysis, we repeated steps 1-3 twice. Given two different input samples from MVBench, SampleA (with a T_A-frame video) and SampleB (with a T_B-frame video), we obtain $\hat{\mathbf{G}}_{\mathrm{A}}$ and $\hat{\mathbf{G}}_{\mathrm{B}}$, and visualize the corresponding heatmaps H_A and H_B. Our observations are as follows:

- For videos with different durations (*i.e.*, different numbers of frames), the shape of $\hat{\mathbf{G}}$ can adapt accordingly ($\hat{\mathbf{G}}_{\mathrm{A}} \in \mathbb{R}^{L_{\mathrm{V}} \times T_A}$ and $\hat{\mathbf{G}}_{\mathrm{B}} \in \mathbb{R}^{L_{\mathrm{V}} \times T_B}$). Therefore, for different input samples, the time gating mechanism can adapt to the content of specific input and performs temporal-sensitive control, thereby enhancing the temporal modeling ability of the model.

- For all patch embeddings at the same time (*i.e.*, the same frame), the corresponding gate values of $\hat{\mathbf{G}}$ change dynamically, revealing that the time gating mechanism can discern

Model	Otter-V	mPLUG-Owl	Video-ChatGPT	Video-LLaMA	VideoChat	VideoChat2	HawkEye	ST-LLM	TG-Vid	TG-Vid
LLM	LLaMA-7B	LLaMA-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B	Vi-7B
#IT	-	-	-	-	-	1.9M	2.2M	220K	197K	220K
Action										
Action Sequence	23.0	22.0	23.5	27.5	33.5	66.0	-	66.0	**72.5**	70.5
Action Prediction	23.0	28.0	26.0	25.5	26.5	47.5	-	53.5	**57.0**	54.5
Action Antonym	27.5	34.0	62.0	51.0	56.0	83.5	-	84.0	85.0	**87.5**
Fine-grained Action	27.0	29.0	22.5	29.0	33.5	**49.5**	-	44.0	45.0	46.0
Unexpected Action	29.5	29.0	26.5	39.0	40.5	**60.0**	-	58.5	53.5	57.5
Object										
Object Existence	53.0	40.5	54.0	48.0	53.0	58.0	-	80.5	**83.5**	83.0
Object Interaction	28.0	27.0	28.0	40.5	40.5	71.5	-	73.5	**74.5**	74.0
Object Shuffle	33.0	31.5	40.0	38.0	30.0	**42.5**	-	38.5	36.0	36.5
Position										
Moving Direction	24.5	27.0	23.0	22.5	25.5	23.0	-	42.5	**45.5**	45.0
Action Localization	23.5	23.0	20.0	22.5	27.0	23.0	-	31.0	**32.0**	29.5
Scene										
Scene Transition	27.5	29.0	31.0	43.0	48.5	**88.5**	-	86.5	82.0	85.5
Count										
Action Count	26.0	31.5	30.5	34.0	35.0	**39.0**	-	36.5	32.5	36.0
Moving Count	28.5	27.0	25.5	22.5	20.5	42.0	-	56.5	**68.5**	66.5
Attribute										
Moving Attribute	18.0	40.0	39.5	32.5	42.5	58.5	-	78.5	82.0	**85.0**
State Change	38.5	44.0	**48.5**	45.5	46.0	44.0	-	43.0	40.5	46.0
Pose										
Fine-grained Pose	22.0	24.0	29.0	32.5	26.5	**49.0**	-	44.5	47.5	42.0
Character										
Character Order	22.0	31.0	33.0	40.0	41.0	36.5	-	46.5	**47.5**	47.0
Cognition										
Egocentric Navigation	23.5	26.0	29.5	30.0	23.5	35.0	-	34.5	33.0	**37.0**
Episodic Reasoning	19.0	20.5	26.0	21.0	23.5	40.5	-	**41.5**	**41.5**	40.0
Counterfactual Inference	19.5	29.5	35.5	37.0	36.0	**65.5**	-	58.5	60.0	58.0
Avg	26.8	29.7	32.7	34.1	35.5	51.1	47.6	54.9	56.0	**56.4**

Table 9: **Comprehensive results on the MVBench benchmark.** Experiments are conducted on 20 MVBench sub-tasks. #IT denotes instruction tuning samples. "V-" in the Model names denotes "Video-" and "Vi-" in the LLM names denotes "Vicuna-". Bold/underline denotes the best/second-best result.

the information at different spatial locations and provides dynamic, fine-grained control.

- For all patch embeddings at the same spatial location, the corresponding gate values of $\hat{G}$ also change dynamically, which demonstrates that the time gating mechanism can also distinguish the temporal dynamics and provide fine-grained control.

Model	V-LLaVA	LLaMA-VID	mPLUG-Owl	PandaGPT	Valley	VideoChat2	V-ChatGPT	V-LLaMA	ST-LLM♣	TG-Vid	TG-Vid
LLM	Vi-7B	Vi-7B	LLaMA-7B	Vi-13B	StableVi-7B	Vi-7B	Vi-7B	Vi-13B	Vi-7B	Vi-7B	Vi-7B
#IT	-	-	-	-	-	1.9M	-	-	220K	197K	220K
Caption Matching											
Action	88.2	72.7	56.9	56.6	15.5	65.0	64.6	73.1	93.9	96.0	95.3
Direction	53.8	45.6	45.3	51.4	21.4	53.8	48.6	47.4	59.3	54.7	55.4
Speed	61.9	52.2	46.4	44.3	22.0	52.6	47.8	47.1	54.3	57.0	58.1
Event Order	57.0	49.0	49.3	55.0	28.3	53.0	49.3	52.0	55.0	65.3	62.0
Attribute Change	58.3	49.0	49.0	49.0	22.9	53.8	48.6	48.3	61.5	64.9	66.7
Avg	63.7	53.6	49.3	51.3	22.0	55.6	51.8	53.5	64.8	**67.6**	<u>67.5</u>
Yes/No QA											
Action	74.3	63.0	64.4	53.0	58.1	72.8	52.5	68.1	68.1	77.4	76.5
Direction	51.8	48.8	50.6	49.6	52.0	53.8	50.0	46.0	50.6	51.6	50.6
Speed	50.3	49.2	51.2	50.8	52.5	53.8	49.5	48.8	49.9	52.5	52.7
Event Order	49.2	48.4	51.3	53.7	50.3	51.3	51.0	51.8	50.0	55.0	50.0
Attribute Change	51.1	52.7	52.0	52.2	52.9	53.8	50.0	50.9	51.6	54.0	54.0
Avg	56.4	53.0	54.4	51.8	53.5	<u>58.0</u>	50.7	53.7	54.0	**58.1**	56.8
Multi-Choice QA											
Action	70.4	58.6	66.6	35.5	47.0	88.5	47.0	54.1	92.0	92.6	91.1
Direction	32.2	29.9	29.3	27.8	29.3	36.4	31.6	24.5	37.3	34.9	39.4
Speed	38.2	29.3	32.2	29.3	32.5	42.0	28.4	28.1	46.7	45.4	46.4
Event Order	41.4	30.5	34.8	31.8	18.9	40.7	37.1	32.8	42.7	41.1	43.0
Attribute Change	39.9	26.0	35.4	30.9	29.9	45.5	30.9	28.5	50.0	50.3	52.1
Avg	44.7	35.3	40.0	31.1	31.8	51.1	35.2	33.9	<u>53.7</u>	52.9	**54.4**
Avg(ALL)	54.9	47.3	47.9	44.7	35.8	54.9	45.9	47.0	57.5	<u>59.5</u>	**59.6**

Table 10: **Comprehensive results on the TempCompass benchmark.** Experiments are conducted on three TempCompass tasks: Caption Matching, Yes/No QA, and Multi-Choice QA. #IT denotes instruction tuning samples. "V-" in the Model names denotes "Video-" and "Vi-" in the LLM names denotes "Vicuna-". The overall average result **Avg(ALL)** is calculated as the average of Avg(Caption Matching), Avg(Yes/No QA), and Avg(Multi-Choice QA). Bold/underline denotes the best/second-best result. ♣: We reproduce the training and inference.
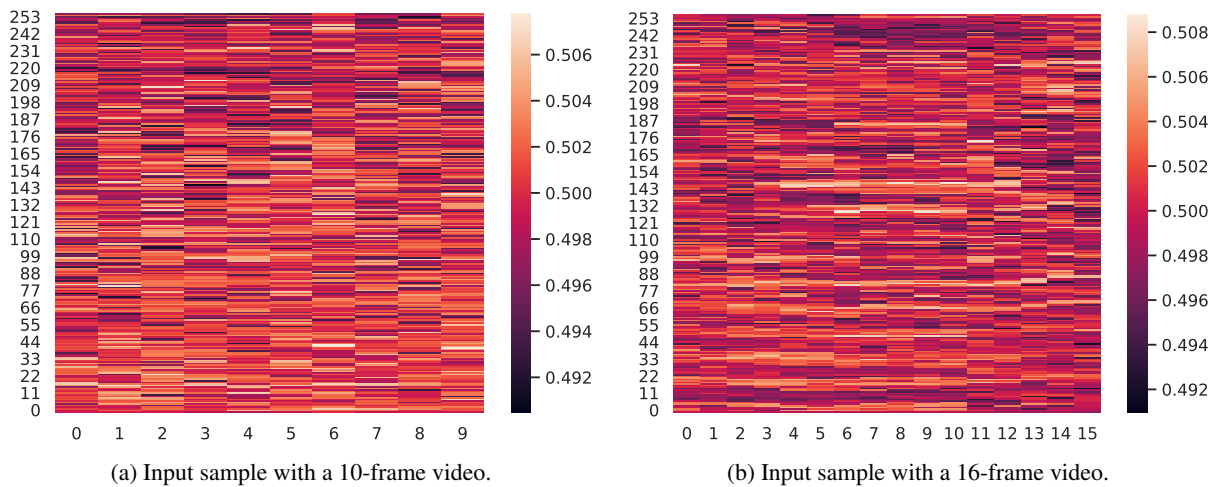

(a) Input sample with a 10-frame video.

(b) Input sample with a 16-frame video.

Figure 3: Visualization of Time Gating.