B-VLLM: A Vision Large Language Model with Balanced Spatio-Temporal Tokens

Supplementary Material

Table 1. Statistics of training datasets.

Stage	Modality	Sample	Source
LLaMA-VID [13]			
PT	Image	558K	[14, 21]
	Video	232K	[3, 13]
FT	Image	625K	[7, 9–11, 18–20, 22]
	Video	98K	[5]
	Text	40K	[1]
VideoLLaMA2 [4] - Valley [16]			
PT	Image	558K	[14, 21]
	Video	702K	[16]
FT	Image	665K	[14]
	Video	100k	[16, 17]

Table 2. Statistics of benchmark datasets. Type indicates the QA question type: **Open** represents open-ended answering and **MCQA** denotes multi-choice question answering.

Benchmark	Avg. Len (s)	# QA Pair	Type
MSVD-QA	9.8	10K	Open
MSRVTT-QA	15.2	50.5K	Open
MVBench	16	4K	MCQA
Perception-Test	23	2.7K	MCQA
Next-QA	44	4000	MCQA
VNBench	54	5.6K	MCQA
VideoMME-S	80.7	0.9K	MCQA
ActivityNet-QA	111	8K	Open
EgoSchema	180	5K	MCQA
VideoMME-M	515.9	0.9K	MCQA
VideoMME-L	2466.7	2.7K	MCQA

1. Additional Implementation Details

1.1. Training Details

We adopt a two-stage training strategy [4, 13, 14], dividing training into pretraining for modality alignment and fine-tuning for instruction tuning. During pretraining, the learning rate is set to 1e-3 with a linear learning rate schedule, and the total batch size is set to 256. Only the frame selection module and the MLP projection layer, which maps the visual token features to the LLM features, are updated. In the fine-tuning stage, the entire framework is unfrozen except for the preliminary visual encoder. Additionally, LoRA [8] is enabled with the rank set to 128 and α set to 256. The total batch size was set to 128, and the learning rate is 1e-4 with cosine scheduling. Supervision is provided by minimizing cross-entropy for masked text tokens, and the optimizer for both stages is AdamW [15], with DeepSpeed ZeRO2 [2] enabled for memory efficiency. Training is conducted with 16 Nvidia L40S GPUs. Following previous works [4, 12, 17], videos are downsampled to 1 *fps* in both training and evaluation.

1.2. Dataset for Training

The following provides details on the training datasets utilized in LLaMA-VID-Dataset [13] and VideoLLaMA2 [4]. The dataset used by LLaMA-VID [13] consists of 790K samples for pre-training stage and 763K samples for fine-tuning stage. Compared to the dataset used by LLaMA-VID, Valley [16] contains significantly more video samples for pre-training. There are 1.25M samples for pre-training stage and 765K sample for fine-tuning stage. The difference in the volume of video samples may explain the performance differences observed between LLaMA-VID and VideoLLaMA2.

1.3. Video Benchmark Summary

The details of the video benchmarks are summarised in Table 2. We selected benchmarks with diverse durations, avoiding a focus on a fixed average video length. For example, the MSVD-QA [24] benchmark consists of short videos with an average duration of 9.8 seconds. In contrast, the VideoMME-L [6] is a representative of long video benchmark, with an average duration of 2466.7 seconds and a maximum video length capped at 1 hour. We argue that our selection of video benchmarks reflects the diversity of video durations in real-world scenarios, demonstrating the applicability and robustness of our model in practical environments.

1.4. B-VLLM Experimental Setup

Our B-VLLM is implemented based on VideoLLaMA2 [4], where Qwen2 [25] is selected as the backbone LLM. Each frame is encoded as 768 preliminary visual tokens using EVA-CLIP [23], Subsequently, 32 frames are selected using the frame selection module and temporal frame token merging module. The visual tokens corresponding to the selected frames are encoded into 32 spatial visual tokens by our spatial visual token sampling module.

2. Additional Experiment & Discussion

2.1. Additional Discussion on Different Frame Selection Features.

As reported in Table 3, we additionally report the performance of selecting frames by using three other feature extraction methods: (a) *max pooling* (b) *mean pooling* and (c) *Qformer*. To investigate the impact of these variants in B-VLLM's fine-grained spatial perception, we also report the performance of **OCR** and **Counting** in VMME, which are more relevant to spatial capability. Though [CLS] token may not be the optimal choice in terms of performance, as discussed in Section 3.2, the rationale of using [CLS] token lies in balancing computational cost and effectiveness.

Table 3. Additional ablation study on frame selection features.

Model	Train.	MVB	EgoS	VMME	OCR	Counting
Mean	13.1h	51.3	51.9	54.4	48.2	36.6
Max	13.1h	49.5	51.6	50.7	45.3	38.1
Qformer	14.0h	52.4	52.9	51.6	41.7	35.8
CLS	10.9h	50.8	51.9	52.9	46.0	34.0

2.2. The Impact of $L*$.

As shown in Fig 1, we report additional experiments to investigate the impact of $L*$ on Short ($\leq$ 3min), Medium ($\leq$ 15min) and Long video ($\leq$ 60 min) settings. As shown below, B-VLLM's performance generally saturated at $L* = 28$, though it was trained with $L* = 32$. This indicates that there could be an optimal setting for $L*$. We argue that this is due to most tasks typically rely on a small portion of frame regardless of the duration of the video. However, we acknowledge that increasing $L*$ is potentially beneficial for more extreme setting and this review inspired us to develop a $L*$-free design as one of our future work.

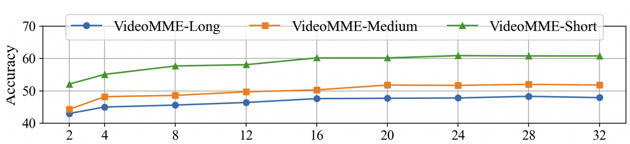


Figure 1. Additional experiment on the impact of $L*$.

2.3. Model's Parameter Size Comparison

In Table 4, we report our B-VLLM's parameter size below along with TFLOPs and GAMCs, which are similar to most existing methods.

2.4. Additional Comparisons with Recent SoTA.

We provide additional comparisons with several advanced VLLMs in Table 5. Note that these advanced VLLMs are generally trained with at least 5 times more data than our B-VLLM, and these data are often private. This indicates the great potential of our B-VLLM when it is trained with large scale data.

Table 4. Model size compared with existing methods

Model	TFLOPs	GMACs	Param.
InternVL2	1.83	910	7+1.1B
InternVideo2	1.83	915	7+1.5B
QwenVL	1.92	915	7.7+1.9B
Qwen2VL	1.82	905	7.6+0.6B
VideoLLaMA2	1.82	910	7.6+0.8B
B-VLLM (Ours)	1.81	908	7.6+1.2B

Table 5. Additional Comparison with Recent SOTA methods

Model	#Train.Data	MVB	EgoS	Perp.	VMME
InternVideo2	39M	60.3	55.8	53.0	41.9
InternVL2	20M+	66.4	-	-	54.0
LLaVA-OV	9.3M	56.7	60.1	57.1	58.2
Qwen2VL	-	**67.0**	**66.7**	**62.3**	**66.3**
B-VLLM (Ours)	2M	50.8	51.9	52.1	52.9

2.5. Inference Speed & Latency

In Table 6, we compare the inference speed between Qwen2VL and B-VLLM on three types of videos: 4-minute video, 8-minute video, and 12-minute video. As reported below, B-VLLM demonstrates clear advantage on inference speed. Specifically, it takes 5.6s for B-VLLM to generate a response for an 8-minute video while taking 24.6s for Qwen2-VL to respond.

Table 6. Inference Speed and Latency

Model	4 Min	8 Min	12 Min
Qwen2VL	6.94s	24.6s	26.2s
B-VLLM (Ours)	3.16s	5.6s	7.9s

2.6. Additional Discussion on Limitations

Limitation. On top of the discussed limitation in the main manuscript, we acknowledge that our proposed method has limitations in handling multi-round conversations based on the same video. In the video-based multi-round conversation, B-VLLM repeats the frame selection and spatial visual token procedure as the context shifts during conversation, resulting in additional computation cost. Moreover, as discussed in manuscript, the spatial visual tokens in B-VLLM are underutilized, and the full potential of our method for image data has yet to be fully explored. Additionally, the Gumbel-Softmax sampling technique may disrupt the temporal order of selected frames, potentially compromising our method's performance in temporal perception. Addressing these limitations could further enhance the capabilities of VLLMs. We recognize that using [CLS] tokens for frame selection has its limitations since pooled visual tokens are more informative, as an alternative strategy for exploring more accurate frame selection. More limitations are discussed in supplementary material. Additionally, the fixed value L^* can limit B-VLLM's performance on extreme long videos, especially when the number of relevant frames surpasses L^*.

3. More B-VLLM Qualitative Examples

This section provides additional conversation examples with B-VLLM based on videos. Note that the presented videos are accessible only through the specified source. The examples are depicted in Figures 2 to 6. As shown in Figure 2, when a game trailer is fed into B-VLLM, it accurately identifies the game's title displayed at the beginning of the video and effectively summarizes the gameplay content, demonstrating B-VLLM's comprehensive understanding of both temporal and spatial information. In the example shown in Figure 5, during the conversation, B-VLLM accurately identifies the man's actions. Even though the basketball appears only briefly at the beginning of the video, B-VLLM successfully captures and localizes the nuanced spatial details. As depicted in Figure 6, beyond accurately describing the content of the given video, B-VLLM demonstrates the ability to estimate the age of the man featured in the video, highlighting the spatial perception capabilities of our method.

Source: https://www.youtube.com/watch?v=u83VdXAVq08

Describe this video in detail.

The video begins with a black screen displaying the words "Game Science" and "Black Myth: Wu Kong." The scene then transitions to a forest where a man is seen walking through the trees. As he walks, he encounters a group of monsters that attack him. Despite being outnumbered, the man manages to fend off the monsters and eventually defeats them all.

Throughout the video, the man's movements are fluid and graceful, showcasing his martial arts skills. The monsters are depicted as fierce and aggressive, adding an element of danger and excitement to the scene. The forest setting adds a sense of mystery and adventure to the video, making it an engaging watch for anyone interested in action-packed scenes or martial arts. Overall, the video is a thrilling display of combat and skill, leaving viewers on the edge of their seats.

Figure 2. Visualization on understanding a 2-minutes game trailer containing intensive fighting scenes.

Source: https://openai.com/index/video-generation-models-as-world-simulators/

Can you describe this video?

Sure, the video features a woman walking down a busy street at night. She is wearing a black leather jacket and carrying a black purse. As she walks, she looks off into the distance and smiles to the camera. The street is bustling with activity, with cars passing by and people walking around her. Despite the noise and commotion of the city, the woman appears calm and collected as she makes her way down the street. Her outfit and demeanor suggest that she may be a fashionista or someone who enjoys urban living. Overall, the video captures a moment in the life of this stylish woman as she navigates the busy streets of a bustling city at night.

Which country do you think it is?

Based on the video, it appears to be Japan.

Figure 3. Visualization on understanding a 1-minute AI generated video of a woman in Japan.

Source: https://www.youtube.com/watch?v=dQw4w9WgXcQ

What do you think this video is for?

The video seems to be a promotional or advertisement for a product, as it features a man in a suit singing into a microphone and a woman in a white dress dancing. The video also includes shots of the man walking down a street and standing in front of a building with a sign that says "the best."

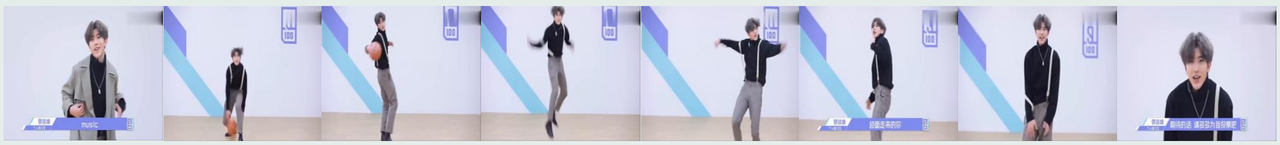
Are there any person except the man singing in this video?

Yes, there are two women standing behind the man and singing along with him..

What the man is doing in this video?

The man is seen singing and dancing in various locations, including a room with a microphone, a tunnel, and a brick building. He also wears different outfits throughout the video, such as a suit and tie, a white shirt, and sunglasses.

Figure 4. Visualization on perceiving a 4-minute wonderful music video containing scenes of singing and dancing.



Source: https://www.youtube.com/watch?v=GLu5YwiAtC4

What is the person doing in this video?

The person in the video is dancing and playing basketball.

Describe this video in detail.

The video features a man who is seen standing in front of a camera and talking. He then proceeds to dance around while holding a basketball, showcasing his impressive skills. The man's movements are fluid and graceful, as he effortlessly moves the ball from hand to hand while performing various dance moves. His passion for the sport is evident as he continues to dance with the ball, displaying his talent and love for basketball. Overall, this video showcases an impressive display of athleticism and skill, as the man combines his love for basketball with his dancing abilities to create a unique and entertaining performance.

Do you think this person is good at basketball?

Yes, the person in the video is very good at basketball. He is seen dribbling and shooting the ball with ease, and he also does some impressive moves while playing.

Figure 5. Visualization on perceiving a 1-minute dancing video a man showcasing his outstanding dancing and basketball skills.

Source: https://www.youtube.com/watch?v=K0mhNqW-mTo

What do you think this video is for?

This video is likely for educational purposes, as it showcases the man's skills in Tai Chi and provides a demonstration of how to perform the moves.

Describe the room in this video.

The room in the video is dimly lit and has a wooden floor. There are two chairs visible in the background, one on the left side of the room and another on the right side.

How old do you think this person is?

Based on the video, it is difficult to determine the exact age of the person. However, he appears to be in his 50s or 60s.

What is person doing in the video?

The man in the video is performing a series of slow and deliberate movements, which appear to be some form of martial arts or exercise routine. He moves his arms and legs in a coordinated manner, maintaining a steady pace throughout the performance.

Figure 6. Visualization of understanding a 3-minute video featuring a Kung Fu master demonstrating his martial arts expertise.

References

[1] Sharegpt.com. Accessed: 2024-09-30. 1

[2] Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, et al. Deepspeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. In *SC*, pages 1–15. IEEE, 2022. 1

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 1

[4] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. 1

[5] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *CVPR*, pages 961–970, 2015. 1

[6] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*, 2024. 1

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Eole of Image Understanding in Visual Question Answering. In *CVPR*, pages 6904–6913, 2017. 1

[8] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[9] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Eeasoning and Compositional Question Answering. In *CVPR*, pages 6700–6709, 2019. 1

[10] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics.

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 1

[12] 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 *CVPR*, pages 22195–22206, 2024. 1

[13] Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models. In *ECCV*, pages 323–340. Springer, 2025. 1

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2024. 1

[15] I Loshchilov. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[16] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video Assistant with Large Language Model Enhanced Ability. *arXiv preprint arXiv:2306.07207*, 2023. 1

[17] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *ACL*, 2024. 1

[18] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions, 2016. 1

[19] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[20] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge, 2022. 1

[21] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 1

[22] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. TextCaps: a Dataset for Image Captioning with Reading Comprehension, 2020. 1

[23] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 1

[24] 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 *MM*. 1

[25] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2024. 1