FLEXIVIDEO: Variation-Aware Temporal Dynamics Modeling for Efficient Video Understanding

Supplementary Material

1. Sub-Metric Definitions for FavorBench and LongVideoBench

In **Introduction**, we provide a performance comparison against similar-sized models on **FavorBench** [8] and **LongVideoBench** [9]. The radar chart is designed to offer a multi-faceted evaluation of each model's capability in understanding the complex dynamics of long-form videos. Each axis on the chart represents a specific evaluation subset, where a higher score indicates superior performance on that task.

For **FavorBench**, the subsets focus on the fine-grained understanding of actions, details, and dynamic elements within the video:

- **AS (Action Sequence)**: Focus on understanding temporal dynamics, requiring the model to compare which action occurs first or to describe the complete action sequence of a subject.
- **HAC (Holistic Action Classification)**: Require the model to identify the core action of the subjects, focusing on global action summarization.
- **SAD (Single Action Detail)**: Examine moment-specific detail recognition, such as the state of subjects at a specific moment or their interaction with an object.
- **MAD (Multiple Action Details)**: Evaluate the ability to compare and analyze details across multiple moments, such as changes in a subject's actions and states over time.
- **CM (Camera Motion)**: Test the understanding of viewpoint dynamics and focus shifts, and their coordination with subject actions.
- **NSM (Non-Subject Motion)**: Assess environmental context awareness, such as understanding the movements of non-subject elements like background objects.

For **LongVideoBench**, the subsets focus on the temporal and relational understanding of events and objects:

- **E2O (Event-Referred Object)**: Given an action or event as the question, require the model to identify the participating people or objects.
- **O2E (Object-Referred Event)**: Given a person or object as the question, require the model to identify the event that occurred or the action they took upon their appearance.
- **E2E (Event Before/After Event)**: Require the model to determine the sequential order of two or more adjacent actions or events.
- **SSS (Sequence of Scenes)**: Require the model to answer questions about the chronological order of multiple scenes in the video.
- **T2E (Event Before/After Text)**: Given a segment of subtitles as a reference, requires the model to identify the action/event that happened before or after it.
- **T2O (Object Before/After Text)**: Given a segment of subtitles as a reference, require the model to identify which specific person/object first appeared before or after it.

Through this comprehensive, multi-dimensional comparison, the figure clearly illustrates our model's significant advantages in processing complex action sequences and understanding long-range temporal dependencies.

2. Detailed Implementation of Pilot Experiment

To ensure that our pilot experiment covers the entire temporal scope of each video, we implemented a structured frame selection strategy. For each 2-3 minute video, we first sample it at 1 FPS to generate a representative frame sequence. This sequence is then divided into 8 uniform temporal segments to span the video's full duration.

Within each of these 8 segments, we compute the inter-frame difference for all possible frame pairs using the metric defined in **Pilot Experiment** and rank them accordingly. To construct the final input for each strategy, we select 4 frame pairs from each segment that correspond to that strategy's criteria (*e.g.*, low, uniform, or high dynamic). This procedure, applied for each strategy, yields a total of 32 frame pairs (64 frames), ensuring that the input sequence comprehensively represents the dynamic variations across the entire video, rather than just a single moment.

As mentioned in the main text, we leverage GPT-4o [4] for a quantitative evaluation of the generated captions. The detailed prompt and evaluation metrics [12] used for this assessment are provided in Tab. 1 for reference.

3. Can Dynamic Temporal Modeling Emerge Without Additional Training?

To transfer knowledge from existing models without retraining from scratch, DSTE module employs a temporal extension initialization procedure that periodically expands the temporal window size, enabling the model to capture longer-range temporal dependencies. Importantly, both ATS and DSTE operate without updating any trainable parameters; all transformations are fixed and introduce no training burden. Given this property, a natural question arises: Can the model preserve its original behavior while benefiting from more efficient temporal dynamics modeling without retraining?

Table 1. The prompt utilized by GPT-4o to evaluate caption generation performance in the pilot experiment.

Prompt Used by GPT-4o for Evaluating Caption Generation Performance

System Prompt:

##TASK DESCRIPTION: You are required to evaluate the performance of the respondent in the video summarization task based on the standard answer and the respondent's answer. You should provide two scores. The first is the COMPLETENESS score, which should range from 1 to 5. The second is the RELIABILITY score, which should also range from 1 to 5. Below are the criteria for each scoring category:

##COMPLETENESS Scoring Criteria:
The completeness score focuses on whether the summary covers all key points and main information from the video.
Score 1: The summary hardly covers any of the main content or key points of the video.
Score 2: The summary covers some of the main content and key points but misses many.
Score 3: The summary covers most of the main content and key points.
Score 4: The summary is very comprehensive, covering most to nearly all of the main content and key points.
Score 5: The summary completely covers all the main content and key points of the video.

##RELIABILITY Scoring Criteria:
The reliability score evaluates the correctness and clarity of the video summary. It checks for factual errors, misleading statements, and contradictions with the video content. If the respondent's answer includes details that are not present in the standard answer, as long as these details do not conflict with the correct answer and are reasonable, points should not be deducted.
Score 1: Contains multiple factual errors and contradictions; presentation is confusing.
Score 2: Includes several errors and some contradictions; needs clearer presentation.
Score 3: Generally accurate with minor errors; minimal contradictions; reasonably clear presentation.
Score 4: Very accurate with negligible inaccuracies; no contradictions; clear and fluent presentation.
Score 5: Completely accurate with no errors or contradictions; presentation is clear and easy to understand.

—

##INSTRUCTION:
1. Evaluate COMPLETENESS: First, analyze the respondent's answer according to the scoring criteria, then provide an integer score between 1 and 5 based on sufficient evidence.
2. Evaluate RELIABILITY: First, analyze the respondent's answer according to the scoring criteria, then provide an integer score between 1 and 5 based on sufficient evidence.
3. Output Scores in JSON Format: Present the scores in JSON format as follows:
{'score_completeness': score_comp, 'score_reliability': score_reli, 'total_score': score_comp + score_reli}
Please ensure your response ends with the JSON object containing the scores.

User Prompt:
Please score the respondent's answer according to the steps in the Instructions. You must end with a JSON dict to store the scores.
Standard Answer: {*standard_answer*}
Respondent's Answer: {*prediction*}

Table 2. Performance of the base model, the temporal extension initialized Model, and FLEXIVIDEO (all without training) on mainstream video benchmarks.

Model	Qwen2.5-VL	Qwen2.5-VL	FLEXIVIDEO w/o training
Perception Window	2	6	Dynamic
Video-MME$_{\text{w/o sub.}}$	61.5	59.2	61.3
MLVU$_{\text{M-Avg}}$	68.2	61.9	67.6
LVBench	43.3	40.5	44.1
MotionBench	55.8	53.4	55.3

To answer this, we evaluate the FLEXIVIDEO with temporal extension initialized models that adopt fixed multi-frame encoding scheme, which means severe visual confusion. All parameters of each model are not trained. We evaluate using a fixed total frame budget, and report the results in Tab. 2.

Directly enlarging the temporal window exacerbates visual confusion, and the naively extended backbone (perception window set to 6) exhibits a clear performance drop. In contrast, FLEXIVIDEO behaves strikingly differently. Owing to the feature-preserving nature of the transformation matrices, it sustains performance comparable to-and in some cases slightly exceeding the original Qwen2.5-VL model (perception window set to 2), despite operating entirely without finetuning. Meanwhile, it achieves notable improvements in computational efficiency due to the reduced number of visual tokens.

These findings underscore the strong transferability and efficiency of our framework, demonstrating that the proposed **Variation-Aware Temporal Dynamics Modeling**

Table 3. Main results on mainstream video benchmarks of FlexiVideo-7B.

Model	Video-MME$_{\text{w/o sub.}}$	Video-MME$_{\text{w/ sub.}}$	LongVideoBench	MLVU$_{\text{M-Avg}}$	LVBench	MotionBench	FavorBench
LLaVA-OneVision-7B [5]	58.2	-	56.3	64.7	-	-	-
LLaVA-Video-7B [11]	63.3	69.7	58.2	70.8	41.5	-	38.6
LongVILA-7B [3]	60.1	65.1	57.1	-	-	-	-
Oryx-1.5-8B [6]	58.8	64.2	56.3	67.5	-	-	-
Apollo-7B [14]	61.3	63.3	58.5	70.9	-	-	-
VideoLLaMA3-7B [10]	66.2	70.3	59.8	73.0	45.3	-	41.5
InternVL3-8B [13]	66.3	68.9	58.8	71.4	44.1	58.1	45.3
NVILA-8B [7]	64.2	70.0	57.7	70.1	-	-	-
Qwen2.5-VL-7B [1]	65.1	71.6	56.0	70.2	45.3	-	40.8
FlexiVideo-7B (Ours)	66.1	71.3	60.9	70.8	46.2	59.7	48.8

supports effective video understanding with substantially improved efficiency, even when applied directly without additional training.

4. Detailed Implementation of Case Analysis

Feature map calculations, as described in the **Experiments** section, begin by extracting embeddings from the Vision Encoder of the model. For each frame group, embeddings are spatially concatenated in order, with each video and model corresponding to a unique set of embeddings. To reduce the dimensionality of these high-dimensional feature representations, we apply Uniform Manifold Approximation and Projection (UMAP), mapping them into a two-dimensional plane for further analysis.

It is important to note that while the embeddings from different models are reduced to the same 2D space for comparative visualization, we plot them separately. This approach allows us to maintain a clear distinction between the feature spaces of each model, enabling a more intuitive comparison.

5. Detailed Derivation of PI-Resize

The following derivation is adapted from [2]. While originally derived for spatial dimensions, we apply it here to the temporal domain, which follows the same mathematical structure.

We begin by rewriting the objective function defined in **Adaptive Temporal Encoding** as follows:

$$\begin{aligned}
&\mathbb{E}_{u\sim\mathcal{U}}\left[(\langle u, w\rangle - \langle Bu, w_{new}\rangle)^2\right] \\
=&\mathbb{E}_{u\sim\mathcal{U}}\left[(u^T(w - B^T w_{new}))^2\right] \\
=&\mathbb{E}_{u\sim\mathcal{U}}\left[((w - B^T w_{new})^T u)(u^T(w - B^T w_{new}))\right] \\
=&(w - B^T w_{new})^T \mathbb{E}_{u\sim\mathcal{U}}\left[uu^T\right](w - B^T w_{new}) \\
=&||w - B^T w_{new}||_\Sigma^2,
\end{aligned} \tag{1}$$

where $||x||_\Sigma^2 = x^T\Sigma x$ and $\Sigma = \mathbb{E}_{u\sim\mathcal{U}} uu^T$ is the (uncentered) covariance matrix of $\mathcal{U}$. When $\mathcal{U} = \mathcal{N}(0, I)$, this recovers the standard Euclidean norm $||w - B^T w_{new}||^2$.

Next, recall that the pseudoinverse yields the least-squares solution for a linear system of equations:

$$(B^T)^+ w \in \arg\min_{w_{new}} ||w - B^T w_{new}||^2. \tag{2}$$

We can also derive an analytic solution for an arbitrary covariance matrix $\Sigma = \mathbb{E}_{u\sim\mathcal{U}} uu^T$. Noting that $||x||_\Sigma^2 = x^T\Sigma x = (\sqrt{\Sigma}x)^T(\sqrt{\Sigma}x) = ||\sqrt{\Sigma}x||^2$, we can rewrite the objective as:

$$||w - B^T w_{new}||_\Sigma^2 = ||\sqrt{\Sigma}w - \sqrt{\Sigma}B^T w_{new}||^2. \tag{3}$$

The optimal solution is then given by:

$$(\sqrt{\Sigma}B^T)^+\sqrt{\Sigma}w \in \arg\min_{w_{new}} ||w - B^T w_{new}||^2. \tag{4}$$

6. Can FlexiVideo Exhibit Consistent Scaling Behavior?

To evaluate the scalability of our variation-aware temporal dynamics modeling approach, we trained a larger version of FlexiVideo, specifically FlexiVideo-7B, using the same underlying architecture and data. This allowed us to examine the model's performance with increased parameters.

As demonstrated in Tab. 3, FlexiVideo-7B exhibits a clear performance boost over other popular models of comparable size. Importantly, since the synergistic effect of ATS and DSTE is content-aware but model-agnostic, it improves the efficiency of the base model and continues to scale effectively, outperforming its counterparts as the model size grows. This consistent scaling behavior underscores the robustness of our approach, confirming its ability to sustain high performance. The results suggest that the temporal perception design allows it to effectively leverage larger model capacities and advance model capabilities.

These findings provide strong evidence that our method is effective not only at the current scale, but that its foundational approach to temporal modeling is widely applicable for effectively leveraging and advancing larger model capabilities.

7. Sensitivity to τ

We performed an ablation on the threshold τ. Notably, setting $\tau = 0.1$ can even further improve performance on Video-MME. However, the trend is clear: as τ increases, more frames are merged and efficiency improves, but performance consistently drops across benchmarks, indicating increased visual confusion from overly aggressive merging. We therefore use $\tau = 0.2$ as a practical trade-off between visual encoding quality and efficiency. As future work, we are exploring adaptive segmentation strategies that adjust segment boundaries according to video dynamics.

Table 4. Effect of the threshold τ on performance.

τ Temporal Variation Threshold	0.1	0.2	0.3	0.4	0.5
Video-MME$_{\text{w/o sub.}}$	62.9	62.5	62.5	61.8	61.6
FavorBench	46.4	46.8	45.9	45.6	45.7

8. Additional Analysis on Model Components.

The ablation results show that only applying DSTE introduces visual confusion, leading to a clear performance drop. Coupling ATS and DSTE FlexiVideo mitigates this issue and improves both performance and efficiency.

Table 5. Ablation study on model components.

Model	ATS	DSTE	Video-MME$_{\text{w/o sub.}}$	LongVideoBench	MLVU$_{\text{M-Avg}}$
Qwen2.5-VL-3B			60.4	55.6	69.0
Qwen2.5-VL-3B	✓		/	/	/
Qwen2.5-VL-3B		✓	59.6	55.4	66.7
FlexiVideo-3B	✓	✓	**62.5**	**57.0**	**69.6**

9. Hardware Efficiency.

Under the same setup on a single A100 GPU with single-sample inference, FlexiVideo achieves lower TTFT latency and peak memory than the baseline, without any hardware optimizations. As a practical proxy for batched throughput, we report end-to-end multi-GPU training time on 32×A100: FlexiVideo takes 8.5h *vs*. 10.5h for the baseline (19.0% less time).

Table 6. Hardware efficiency on MotionBench and MLVU.

Benchmark	Model	#Tokens	Mem Peak	Latency
MotionBench	Qwen2.5-VL-3B	24097.9	15.7 GB	2759.7 ms
	FlexiVideo-3B	14010.6	11.3 GB	1535.1 ms
MLVU$_{\text{M-Avg}}$	Qwen2.5-VL-3B	38245.8	20.7 GB	4206.8 ms
	FlexiVideo-3B	30722.3	16.2 GB	3466.4 ms

10. Implementation atop Other MLLM

We implement atop Qwen3-VL-4B in training-free setting. Ours achieves comparable performance (Video-MME 69.0 *vs*. 69.3/LongVideoBench 63.7 *vs*. 64.2) with 47.1%/53.4% less visual tokens. As Qwen3-VL is a concurrent work, the efficiency gains is significant.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[2] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023. 3

[3] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 3

[4] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1

[5] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[6] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 3

[7] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4122–4134, 2025. 3

[8] Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding. *arXiv preprint arXiv:2503.14935*, 2025. 1

[9] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024. 1

[10] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 3

[11] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 3

[12] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and*

Pattern Recognition Conference, pages 13691–13701, 2025. 1

[13] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 3

[14] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901, 2025. 3