Does Your Vision-Language Model Get Lost in the Long Video Sampling Dilemma?

Supplementary Material

7. Limitations and Forecast

Our work highlights the necessity of creating high-NSD (Necessary Sampling Density) tasks on long videos, which require dense and high-quality human annotations. However, due to the current limitations in community resources and the high cost of annotations, large-scale densely annotated datasets for long videos remain scarce. In this work, we chose EGO4D as the foundation for generating our dataset because it provides high-quality annotations that align with our standards. Nevertheless, this choice imposes certain constraints on the diversity of task types and scenarios. We hope that more high-quality annotated datasets for long videos will emerge in the future. Additionally, our proposed pipeline for generating high-NSD tasks on long videos is highly adaptable to new annotated datasets and can be extended, with slight modifications, to audio-visual data.

Regarding our Reasoning-Driven Hierarchical Sampling approach, its current design is limited by the reasoning capabilities of existing vision-language models (VLMs) and our computational resources. We implemented a relatively straightforward 2-stage framework, which has already achieved impressive results. However, with advancements in reasoning capabilities and models' capacity for sustained "long thinking," there is potential to integrate sampling as an adaptive tool. This could enable dynamic, on-demand visual cue searches during reasoning in an end-to-end manner, further enhancing the framework's efficiency and flexibility.

Finally, our SGFS (Sparse-Guided Frame Selection) currently employs a vision encoder similar to CLIP, which does not explicitly account for temporal information. This limitation leads to the loss of motion-related cues, particularly in short-term temporal sequences where adjacent frames exhibit high similarity. Consequently, we restricted its use to sparse sampling scenarios. We are optimistic that future research will introduce zero-shot encoders specifically designed to encode temporal information, enabling improvements in this area.

8. Discussion on speed

We conducted experiments using Qwen2.5-VL-7B on LSD-Bench using single GPU. Ours achieved result comparable to base model with 768 frames, while requiring significantly less inference time.

Method	Qwen2.5-VL	Qwen2.5-VL	Ours
Avg. Frames	256	768	225
Acc(%)	50.1	52.5	52.2
Avg. Time	22.4s	40.3s	27.5s

9. Discussion on question-prior

We deliberately choose not to incorporate question prior in SGFS because methods relying on semantic similarity can only select frames that are directly similar to the question. Such approaches may fail in more complex scenarios involving indirect references or temporal reasoning. For instance, when faced with "What did this person do after washing the dishes?", key information after "washing dishes" will be filtered out. This limitation underscores the necessity of leveraging reasoning-capable VLMs in stage 1.

We also conducted corresponding experiments. *(i)* We completely replaced SGFS with selecting frames solely based on their semantic similarity to the question, maintaining the same number of initial sampling frames. The results showed a significant drop in accuracy. *(ii)* We attempted to incorporate question similarity into SGFS. Specifically: $W_{ij} = S_{ij} + P_{ij} + \gamma * sim(\text{question}, \text{initial\_frame}_j)$. We present the best result obtained through simple enumeration of γ. The improvement in accuracy is still minor. However, it is important to note that despite the minor improvements, existing text encoders have limited input sequence length (e.g., CLIP's 77-token, SIGLIP2's 64-token), which hinder the processing of longer questions. And in multi-turn dialogue scenarios, frames would need to be repeatedly reselected, significantly increasing computational overhead.

Method	SGFS	Only Question-Prior	SGFS + Question-Prior
Acc(%)	52.2	49.6	52.8

In summary, we conclude that incorporating a question prior into SGFS has considerable limitations.

10. Robustness to noise

To simulate noise, we introduced perturbations, including random resizing and randomly replacing sampled frames with 0.5s-neighboring frame. We conducted 5 random runs for RHS, obtaining average accuracy of 52.84, and a variance of 0.6. This result demonstrate the robustness of our method against noise.

11. Additional Experiment Results

We present in Table 4 the accuracy of the model on LSDBench when the full video is used as input under varying numbers of sampled frames. Additionally, in Table 3, we provide the accuracy performance when using SGFS and RHS under different initial sampling frame counts and SGFS target frame settings.

Initial Sampling	Acc (%)
1-FPM	51.0
1-FPM → 0.50-FPM	50.1
2-FPM → 0.25-FPM	48.7
2-FPM → 0.50-FPM	50.1
2-FPM → 1.00-FPM	51.5
4-FPM → 0.50-FPM	51.8
4-FPM → 1.00-FPM	52.2

Table 3. Ablation study on different setting (including number of initial sampling frames and the ratio of kept frames) of the sampling module.

12. Exploration of SGFS hyperparameters

We conducted ablation study, as shown in the table below. Properly setting hyperparameters within an appropriate range can improve performance.

λ	1	2	10	40	160	640	10	10	10	10	10	10
β	0.3	0.3	0.3	0.3	0.3	0.3	0.05	0.1	0.3	0.5	0.7	0.9
Acc(%)	49.9	50.3	**52.2**	**52.2**	52.0	51.7	51.9	51.5	**52.2**	51.2	49.9	50.9

13. Visualizations of SGFS

We visualized the sampling results of SGFS and compared the sampled frames obtained through uniform sampling, SGFS, and SGFS without the length penalty under the condition of retaining the same target number of frames. From the results, we observed that the frames sampled by SGFS contain significantly less redundancy, enabling more diverse and informative visual content with the same number of frames.

Model	Sampling	Frames	Accuracy
Gemini-2.0-Flash	1-FPS	2700	56.2
Gemini-2.0-Flash Oracle	1-FPS	180	64.8
Gemini-2.0-Flash	Text	0	29.1
LongVA	Fixed	2	31.4
LongVA	Fixed	4	32.4
LongVA	Fixed	8	31.7
LongVA	Fixed	16	32.6
LongVA	Fixed	32	30.9
LongVA	Fixed	64	30.4
LongVA	Fixed	128	31.5
LongVA	Fixed	256	31.3
LongVA	Fixed	512	33.0
LongVA	Fixed	1024	32.5
Qwen2-VL	text	0	30.5
Qwen2-VL	Fixed	2	40.4
Qwen2-VL	Fixed	4	42.6
Qwen2-VL	Fixed	8	42.9
Qwen2-VL	Fixed	16	44.9
Qwen2-VL	Fixed	32	45.5
Qwen2-VL	Fixed	64	47.7
Qwen2-VL	Fixed	128	49.1
Qwen2-VL	Fixed	256	48.0
Qwen2-VL	Fixed	768	45.4
LongVila	Text	0	36.9
LongVila	Fixed	2	36.6
LongVila	Fixed	4	39.6
LongVila	Fixed	8	43.0
LongVila	Fixed	16	43.8
LongVila	Fixed	32	45.4
LongVila	Fixed	64	46.9
LongVila	Fixed	128	48.3
LongVila	Fixed	256	49.8
Qwen2.5-VL	Only-Text	0	30.2
Qwen2.5-VL	Fixed	2	41.2
Qwen2.5-VL	Fixed	4	42.7
Qwen2.5-VL	Fixed	8	44.2
Qwen2.5-VL	Fixed	16	45.6
Qwen2.5-VL	Fixed	32	45.2
Qwen2.5-VL	Fixed	64	48.3
Qwen2.5-VL	Fixed	128	51.0
Qwen2.5-VL	Fixed	256	50.1
Qwen2.5-VL	Fixed	768	52.5
Qwen2.5-VL	2-Stage	45+180	52.2
InternVideo2.5	Text	0	36.1
InternVideo2.5	Text	4	43.3
InternVideo2.5	Fixed	4	43.3
InternVideo2.5	Fixed	8	46.9
InternVideo2.5	Fixed	16	47.2
InternVideo2.5	Fixed	32	49.3
InternVideo2.5	Fixed	64	49.2
InternVideo2.5	Fixed	128	50.1
InternVideo2.5	Fixed	256	50.1

Table 4. Performance comparison of different models and sampling settings.

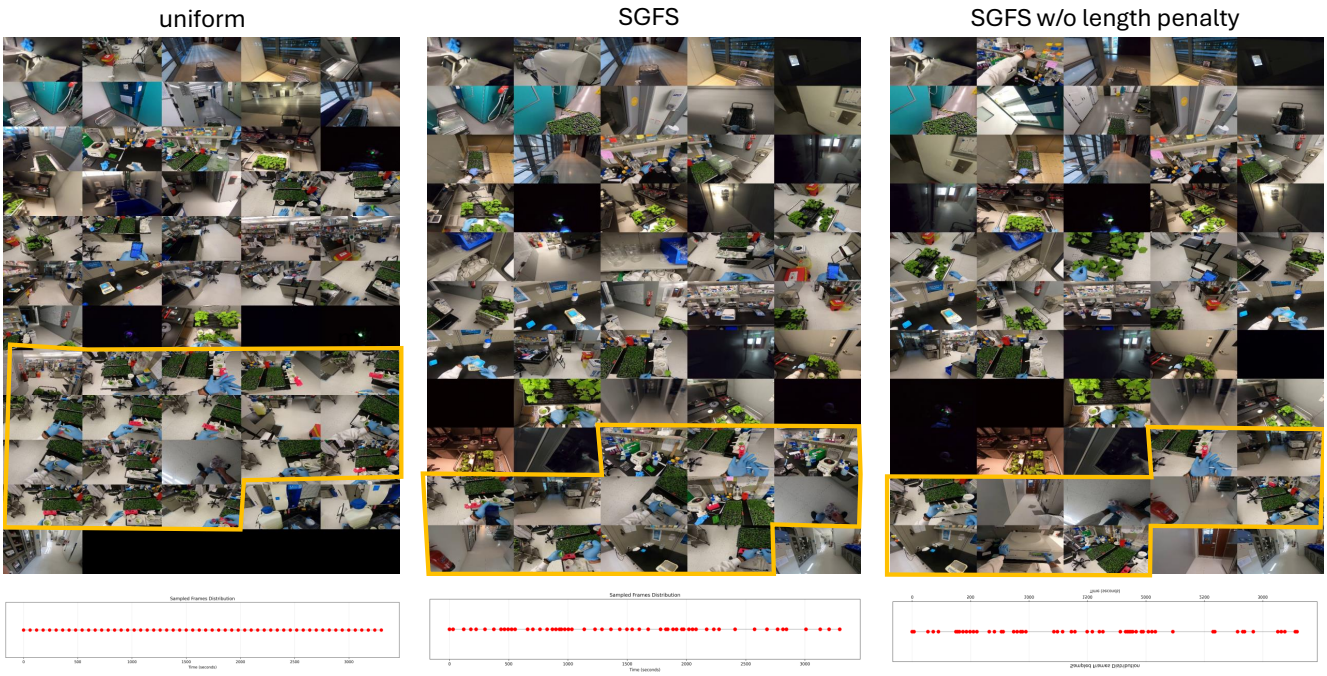

Figure 7. Sampled frames thumbnails and timeline distribution. The frames in each box are considered having similar semantic information, which shows redundancy.

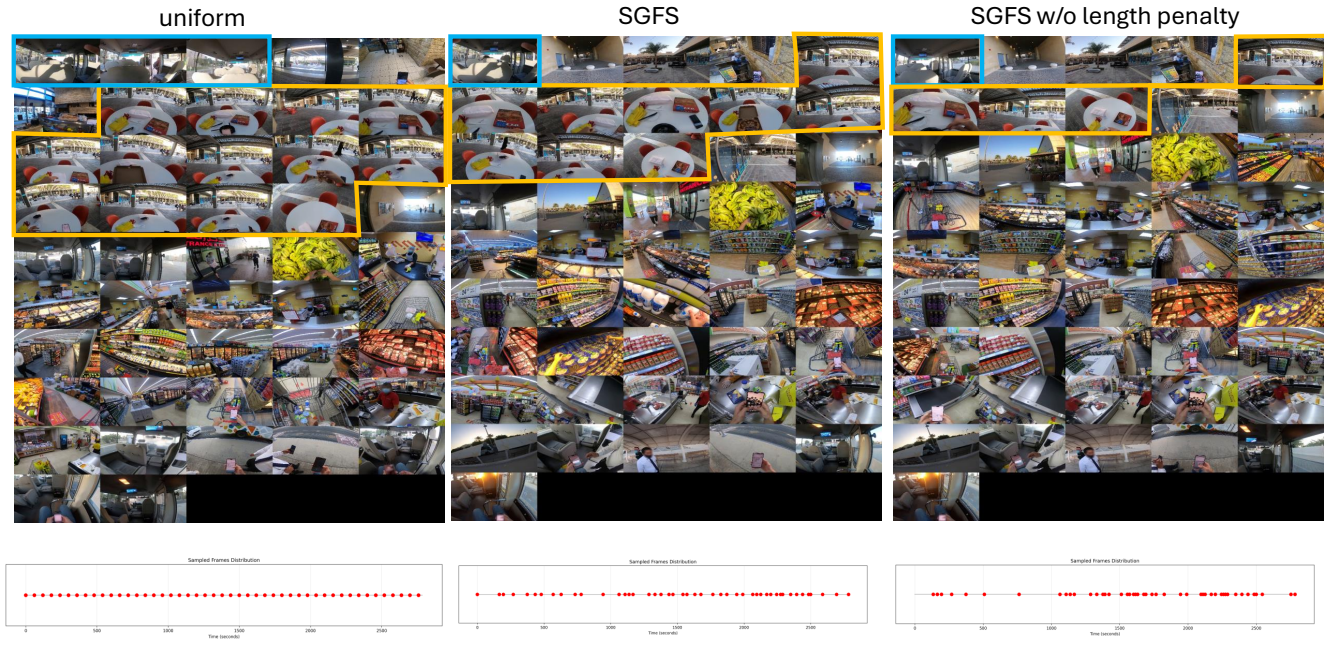

Figure 8. Sampled frames thumbnails and timeline distribution. The frames in each box are considered having similar semantic information, which shows redundancy.

Figure 9. Sampled frames thumbnails and timeline distribution. The frames in each box are considered having similar semantic information, which shows redundancy.