LFS: Learnable Frame Selector for Event-Aware and Temporally Diverse Video Captioning

Author Name
Affiliation
email@example.com

Abstract

Video captioning models convert frames into visual tokens and generate descriptions with large language models (LLMs). Since encoding all frames is prohibitively expensive, uniform sampling is the default choice, but it enforces equal temporal coverage while ignoring the uneven events distribution. This motivates a Learnable Frame Selector (LFS) that selects temporally diverse and event-relevant frames. LFS explicitly models temporal importance to balance temporal diversity and event relevance, and employs a stratified strategy to ensure temporal coverage while avoiding clustering. Crucially, LFS leverages caption feedback from frozen video-LLMs to learn frame selection that directly optimizes downstream caption quality. Additionally, we identify the gap between existing benchmark and human's cognition. Thus, we introduce ICH-CC built from carefully designed questions by annotators that reflect human-consistent understanding of video. Experiments indicate that LFS consistently improves detailed video captioning across two representative community benchmarks and ICH-CC, achieving up to 2.0% gains on VDC and over 4% gains on ICH-CC. Moreover, we observe that enhanced captions with LFS leads to improved performance on video question answering. Overall, LFS provides an effective and easy-to-integrate solution for detailed video captioning.

1 Introduction

Recent advances in multimodal large language models (MLLMs) have greatly improved visual-to-text generation by encoding visual inputs as tokens processed by LLMs [Wu *et al.*, 2023]. Compared with image-related tasks, video understanding tasks requires jointly modeling spatial content and temporal dynamics [Zheng *et al.*, 2023; Song *et al.*, 2024; Chen *et al.*, 2024], and are commonly categorized into video question answering (QA) [Tian *et al.*, 2025], short captioning [He *et al.*, 2025], and detailed captioning [Chai *et al.*, 2025]. Among them, the latter targets fine-grained and temporally coherent descriptions, providing richer information than the former two and better supporting downstream tasks such as video retrieval and multimodal QA. Accordingly, this work focuses on detailed video captioning, and Fig. 1 demonstrates consistent improvements over baselines across nine benchmarks spanning both detailed captioning and QA tasks.

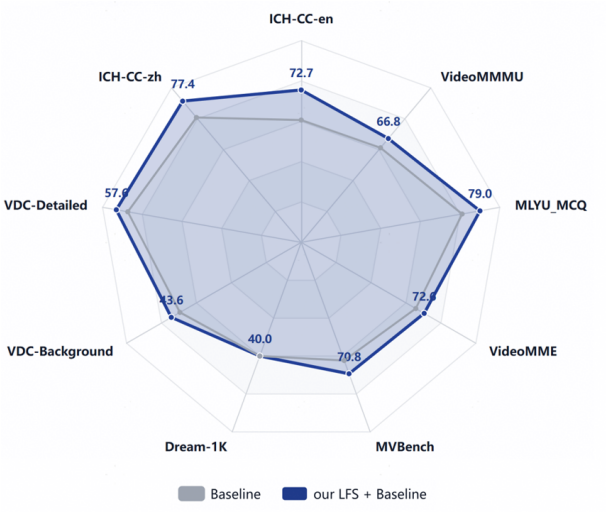

Figure 1: Performance comparison between baselines and the LFS-enhanced counterparts across nine benchmarks. Baselines include AuroraCap-7B, Tarsier2-7B, Qwen2.5-VL-7B, and Qwen3-VL-8B.

Due to computational constraints, most video-LLMs adopt uniform sampling to select a fixed number of frames [Maaz *et al.*, 2024; Li *et al.*, 2024; Lin *et al.*, 2023]. While this strategy ensures temporal coverage, it ignores the uneven distribution of informative events and often underrepresents short but critical actions. In contrast, query-driven

frame selection methods for video QA retrieve frames relevant to a given question [Tang *et al.*, 2025b; Zhu *et al.*, 2025; Zhang *et al.*, 2025], but are unsuitable for query-agnostic tasks such as detailed video captioning.

An effective frame selection strategy for detailed captioning should satisfy two requirements: (1) event awareness, capturing salient actions and scene changes, and (2) temporal diversity, avoiding redundant frames concentrated in short intervals. However, existing approaches typically fail to meet both: Top-K selection leads to temporal clustering, while reinforcement learning-based methods introduce high variance and are difficult to integrate with frozen video-LLMs.

To address these limitations, we propose a Learnable Frame Selector (LFS) that selects temporally diverse and event-relevant frames. LFS integrates an event-aware temporal scoring network with a stratified Top-K selection mechanism to balance event relevance and temporal diversity. Furthermore, LFS leverages caption feedback from frozen video-LLMs as supervision, directly optimizing frame selection for caption quality rather than proxy objectives, and can be seamlessly integrated into existing pipelines without modifying language model parameters.

In the evaluation of detailed video captioning, representative benchmarks such as VDC and Dream-1K report relatively low performance. Prior studies further show that even experienced human annotators achieve accuracies below 50% [Wang *et al.*, 2024; Chai *et al.*, 2025], highlighting a substantial gap between benchmark evaluation and human understanding. To address this gap, we introduce ICH-CC, a benchmark for detailed video captioning on intangible cultural heritage cuisine videos, constructed with fully human-authored captions and carefully designed question–answer pairs.

Experiments demonstrate that LFS significantly improves performance on ICH-CC, boosting Qwen3-VL by 3.18% and 4.47% on the Chinese and English subsets, respectively. On community benchmarks such as VDC, LFS consistently improves multiple video-LLM backbones under saturated evaluation settings. Moreover, gains in detailed captioning achieved by LFS reliably transfer to downstream video QA.

In summary, frame selection is a critical bottleneck for detailed video captioning. By explicitly modeling event relevance and temporal diversity, LFS provides an effective and easily deployable solution that improves captioning quality. Our main contributions are:

- We propose LFS, a learnable frame selector that balances event awareness and temporal diversity via stratified Top-K selection and caption-guided supervision.
- We introduce ICH-CC, a benchmark aligned with human cognition for evaluating detailed video captioning in Chinese and English.
- We conduct extensive experiments and show that LFS consistently improves detailed video captioning and zero-shot video QA with multiple video-LLMs and benchmarks.

2 Related Work

2.1 Video-LLMs for Question Answering

Early Video-LLMs, such as Video-Chat [Li *et al.*, 2024] and Video-LLaMA [Zhang *et al.*, 2023], concatenate frame features with text prompts and fine-tune LLMs on instruction-tuning datasets, enabling open-ended dialogue and basic temporal reasoning but suffering from long-context inefficiency. Subsequent works improve scalability through spatio-temporal pooling, Q-former adapters [Qasim *et al.*, 2025; Diba *et al.*, 2023; Kim *et al.*, 2024], and token dropping, while specialized pre-training objectives enhance motion and event understanding. Despite these advances, most Video-LLMs remain optimized for query-driven reasoning, leaving frame selection for query-agnostic tasks relatively underexplored.

2.2 Video-LLMs for Detailed Captioning

Recent work extends video captioning from brief summaries to detailed descriptions [Kim *et al.*, 2025; Wu *et al.*, 2025; Ge *et al.*, 2024]. Models such as Video-LLaVA [Lin *et al.*, 2023], VideoChat2 [Li *et al.*, 2024], and PLLaVA [Xu *et al.*, 2024] are fine-tuned on large-scale video–text datasets and evaluated using standard captioning metrics, but often generalize poorly to fine-grained and temporally structured descriptions. To mitigate this issue, benchmarks such as Dream-1K [Wang *et al.*, 2024] and VDC [Chai *et al.*, 2025] have been introduced, along with methods like AuroraCap that reduce visual tokens via token merging. Nevertheless, most approaches still rely on uniform frame sampling, which frequently misses short yet informative events in long videos.

2.3 Frame Selection before Video-LLMs

To reduce computational cost, prior frame selection methods aim to retain task-relevant frames. Rule-based approaches use uniform subsampling or shot boundary detection, while model-based methods typically select Top-K frames based on query relevance using CLIP-like models [Guo *et al.*, 2025; Hu *et al.*, 2025; Tang *et al.*, 2025a]. Token-level selectors further prune redundant visual tokens within frames. While effective for query-driven tasks such as video QA, these methods often overlook temporal diversity and are not designed for query-agnostic settings [Li *et al.*, 2025a;

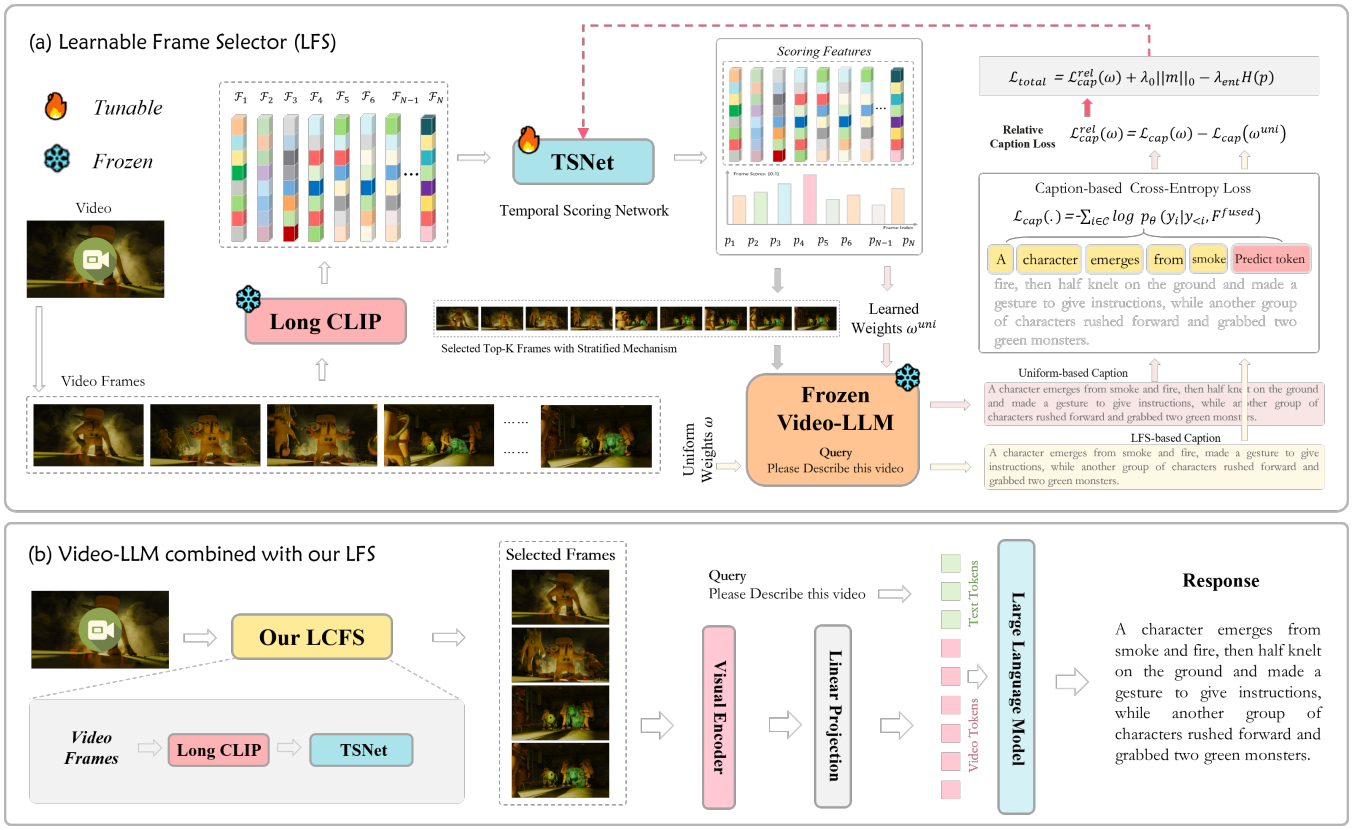

Figure 2: The overall scheme of the proposed learnable frame selector (LFS). (a) and (b) are the training and inference process of LFS, respectively.

Buch *et al.*, 2025]. In contrast, detailed video captioning requires selecting event-aware and temporally-diverse frames that capture distinct events across the entire video, a requirement largely unaddressed by existing frame selection approaches.

3 Method

3.1 Overall Framework

We propose Learnable Frame Selector (LFS), a lightweight and plug-and-play module that selects a small set of informative and temporally diverse frames for detailed video captioning, as shown in Fig.2.

Given frozen frame embeddings extracted by a vision encoder Long-CLIP [Zhang *et al.*, 2024], LFS predicts a continuous temporal importance field over all frames using a lightweight temporal scoring network (TSNet). Unlike uniform sampling or query-dependent retrieval, LFS models frame importance in a query-agnostic manner. At inference, a stratified Top-K strategy enforces temporal coverage and avoids frame clustering. During training, LFS is optimized with caption-guided supervision from a frozen video-LLM as captioner, directly aligning frame selection with caption quality.

3.2 Temporal Importance Modeling

Let $X = \{x_t\}_{t=1}^{N}$, $x_t \in \mathbb{R}^d$ denote frame embeddings from the frozen Long-CLIP. LFS predicts an importance logit $s(t)$ for each frame using TSNet shown in Fig.3.

TSNet first applies a temporal convolution to capture local transitions:

$$H_1 = \mathrm{GELU}(\mathrm{Conv}_1(X)), \qquad H_1 \in \mathbb{R}^{\mathrm{hid} \times N}. \quad (1)$$

A global summary $g = \mathrm{Mean}(H_1)$ is used for gated modulation:

$$H_1 \leftarrow H_1 \odot (1 + \alpha \tanh(\mathrm{MLP}(g)))$$

which adaptively rescales temporal features while preserving near-identity behavior at initialization.

A second temporal convolution aggregates local responses:

$$H_2 = \mathrm{GELU}(\mathrm{Conv}_2(H_1)) \quad (2)$$

followed by a pointwise projection:

$$s(t) = \mathrm{Conv}_{1\times1}(H_2)_t \quad (3)$$

Finally, logits are normalized within each video:

$$\hat{s}(t) = \frac{s(t) - \mu_s}{\sqrt{\sigma_s^2 + \epsilon}} \tag{4}$$

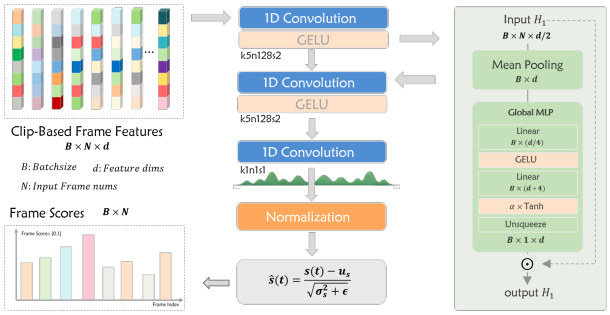

Figure 3: The architecture of temporal scoring network (TSNet).

3.3 Stratified Top-K Frame Selection

The normalized logits $\hat{s}(t)$ are converted into a soft importance distribution:

$$p(t) = \frac{\exp(\hat{s}(t)/\tau)}{\sum_j \exp(\hat{s}(j)/\tau)} \tag{5}$$

where τ is a temperature parameter annealed during training.

To avoid premature collapse of the importance distribution during early training, we include an entropy regularization term $H(p)$ that encourages exploration over frames when the selector is still uncertain.

At inference time, LFS selects exactly K frames using a stratified Top-K strategy. The video timeline is divided into K equal temporal segments, and one frame is selected from each segment:

$$t_i = \arg \max_{t \in \text{segment}(i)} \hat{s}(t), \quad i = 1, \ldots, K \tag{6}$$

This strategy guarantees full temporal coverage and prevents temporal clustering. When applicable, the first and last frames are always retained.

During training, gradients are propagated through the soft distribution $p(t)$, while the stratified Top-K operator is applied only for hard frame selection at inference.

3.4 Caption-Guided Optimization

To align frame importance with detailed captioning quality, LFS is trained under the guidance of a frozen video captioner. Given temporal importance scores $p(t)$, we select a truncated set of top-ranked frames to control computation and normalize their weights to obtain w with $\sum_t w_t = 1$. This truncated-and-renormalized design enables efficient optimization under a fixed frame budget while preserving differentiability.

To encourage compact frame selection, we impose an ℓ_1 regularization on w. Although w is normalized, the ℓ_1 penalty operates on the truncated candidate set prior to renormalization, discouraging mass spreading and promoting concentration on representative frames. In addition, an entropy regularizer is applied to $p(t)$ to encourage early exploration and is gradually annealed during training.

We inject the frame weights w into the captioner through a differentiable weighting hook at a designated visual module. Let $F_t \in \mathbb{R}^{L \times D}$ denote per-frame visual features at the hooked layer. The hook computes a fused representation:

$$F^{\text{fused}} = \sum_{t=1}^{T} w_t F_t, \tag{7}$$

which is broadcast back to match the original tensor shape. This operation acts as a global modulation signal, while the captioner's internal temporal and linguistic modeling remains unchanged.

The captioner input consists of a textual prompt concatenated with the ground-truth caption. We compute token-level cross-entropy only on caption tokens, masking prompt and padding tokens:

$$\mathcal{L}_{\text{cap}}(w) = -\sum_{i \in \mathcal{C}} \log p_\theta\big(y_i \mid y_{<i}, F^{\text{fused}}\big) \tag{8}$$

where $\mathcal{C}$ indexes caption tokens. The captioner parameters θ are frozen, so gradients propagate only to the temporal scoring network through w.

To reduce captioner bias and stabilize optimization, we adopt a relative caption objective by subtracting a uniform baseline computed over the same truncated frame set:

$$\mathcal{L}_{\text{cap}}^{\text{rel}}(w) = \mathcal{L}_{\text{cap}}(w) - \mathcal{L}_{\text{cap}}(w^{\text{uni}}) \tag{9}$$

where w^{uni} assigns equal weights to the selected frames.

The final training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{cap}}^{\text{rel}}(w) + \lambda_0 \|w\|_1 - \lambda_{\text{ent}} H(p) \tag{10}$$

where the ℓ_1 term promotes compact frame weighting, the entropy regularizer encourages exploration, and the relative caption loss aligns frame selection with captioning quality.

3.5 Training Details

LFS is trained for five epochs using AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. We use mixed-precision training with a batch size of 1. The temperature parameter is annealed from $\tau_{\text{start}} = 2.0$ to $\tau_{\text{end}} = 1.0$. Sparsity and entropy regularization weights are set to $\lambda_0 = 0.01$ and $\lambda_{\text{ent}} = 0.01$, respectively, and the number of selected frames is capped at 16 per video. The frozen captioning model used for caption-guided supervision is Qwen3-VL-8B. In the training process of LFS, we employed three NVIDIA A800 GPUs.

The training set consists of 1,588 videos of 2–3 minutes collected from WebVid-10 [Bain *et al.*, 2021], TGIF [Li *et al.*, 2016], Charades [Sigurdsson *et al.*, 2016], YouCook2 [Zhou *et al.*, 2018], and TREC-VTT [Awad *et al.*, 2023]. Training captions are obtained via distillation from Qwen3-VL-253B-A22B.

4 ICH-CC Benchmark

Existing video understanding benchmarks, such as VDC and Dream-1K, primarily prioritize evaluation challenges and thus leads to misalignment with human cognitive understanding ability. Therefore, we present a ICH-CC benchmark to reflect human cognitive ability.

As shown in Fig.4, five experienced annotators each carefully labeled 20 videos only, producing the detailed captions and 100 QA counterparts, and conducted five rounds of cross-checking to ensure that every item was reviewed through all annotators. All videos are sourced from real-world business scenarios in the context of intangible cultural heritage Chinese cuisine.

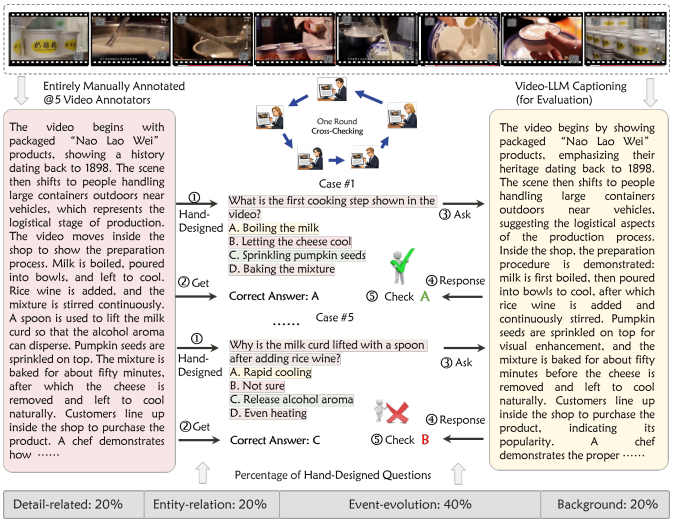

Figure 4: Overview of ICH-CC construction pipeline. The five annotators cross-checked each other's annotated data. After five rounds, each data was thoroughly examined by all annotators to eliminate human biases.

ICH-CC consists of two subsets, ICH-CC-en for English and ICH-CC-zh for Chinese, with each containing 500 evaluation questions designed from 100 videos (2-3 minutes each). ICH-CC has a balanced composition of question types: detail-related, entity-relation, event-evolution, and background questions account for 20%, 20%, 40%, and 20%, respectively.

Models are tasked with generating a detailed caption for each video, which is then assessed by answering associated objective questions derived from human-authored descriptions. A question is deemed correct if the caption provides sufficient evidence to support the correct answer. The final score is calculated based on the percentage of correctly answered questions, with a focus on both semantic coverage and temporal consistency.

5 Experimental Results

5.1 Analysis of ICH-CC benchmark

Quantitative Analysis. Table 1 reports results on the ICH-CC benchmark, where ICH-CC-en/ICH-CC-zh denote accuracy on the English and Chinese subsets. Strong baselines such as Qwen2.5-VL-7B and Qwen3-VL-8B show competitive performance, with Qwen3-VL-8B reaching 74.25% on ICH-CC-zh, outperforming prior open-source models.

Integrating LFS yields consistent gains under a fixed budget of 16 frames. LFS + Qwen2.5-VL-7B improves overall accuracy from 68.46% to 71.77% (+3.31%). LFS + Qwen3-VL-8B further achieves 75.05% (+3.82%), with gains of +4.47% on ICH-CC-en and +3.14% on ICH-CC-zh. These results demonstrate that our LFS with event awareness and temporal diversity effectively enhances detailed video captioning.

Model	ICH-CC-en	ICH-CC-zh	Overall
Vicuna-v1.5-7B	66.25	69.36	67.81
Video-ChatGPT-7B	67.88	68.25	68.07
LLaMA-VID	64.25	63.77	64.01
LongVA-7B	66.37	65.89	66.13
ShareGPT4Video-8B	62.36	65.96	64.16
MovieChat-7B	62.38	62.86	62.62
Qwen2.5-VL-7B*	67.52	69.39	68.46
Qwen3-VL-8B*	68.20	74.25	71.23
Our LFS + Qwen2.5-VL-7B*	**69.88**	**73.65**	**71.77**
ΔAcc / ΔScore	+2.36	+4.26	+3.31
Our LFS + Qwen3-VL-8B*	**72.67**	**77.43**	**75.05**
ΔAcc / ΔScore	+4.47	+3.14	+3.82

Table 1: Accuracy (%) on the ICH-CC benchmark. ICH-CC-en and ICH-CC-zh represent the English and Chinese subsets, respectively. Bold numbers indicate the best scores. * indicates the baseline before applying LFS

Qualitative Analysis. Figure 5 compares uniform sampling and LFS on an ICH-CC video (Nai-Lao-Wei) with 16 frames. Uniform sampling captures only five event-aware frames and misses key short steps, while LFS retrieves eight frames covering the main procedure. The bar plots show that uniform sampling wastes frames on redundant intervals, whereas the

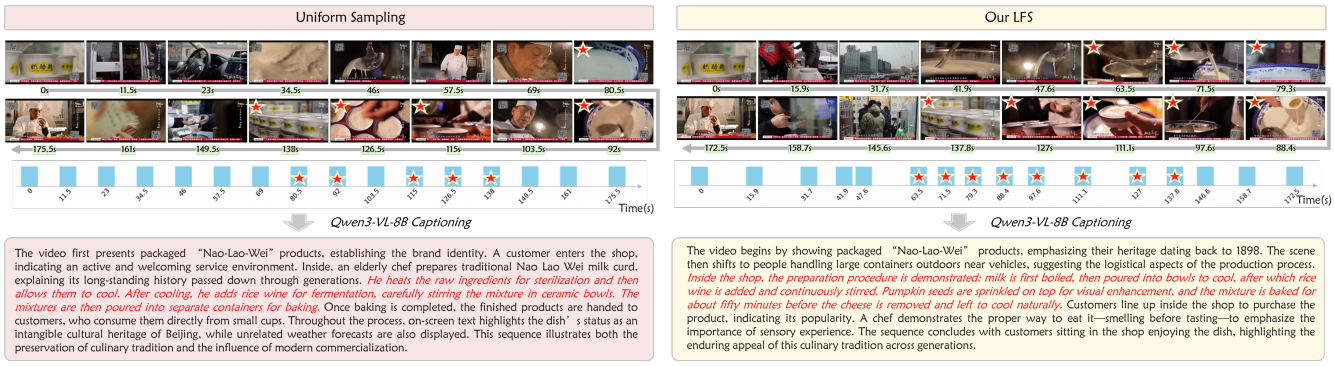

Figure 5: Qualitative results on the ICH-CC-en benchmark. The left shows Qwen3-VL-8B using uniform sampling and our LFS. ☆ indicates selected event-aware frames, and the bar chart visualizes their distribution along the timeline.

Model	Camera Acc / Sim	Short Acc / Sim	Background Acc / Sim	Main Object Acc / Sim	Detailed Acc / Sim
Gemini-1.5 Pro	38.68 / 2.05	35.71 / 1.85	43.84 / 2.23	47.32 / 2.41	43.11 / 2.22
Vicuna-v1.5-7B	21.68 / 1.12	23.06 / 1.17	22.02 / 1.15	22.64 / 1.16	23.09 / 1.20
LLaMA-VID	39.47 / 2.10	29.92 / 1.56	28.01 / 1.45	31.24 / 1.59	25.67 / 1.38
Video-ChatGPT-7B	37.46 / 2.00	29.36 / 1.56	33.68 / 1.70	30.47 / 1.60	24.61 / 1.26
MovieChat-7B	37.25 / 1.98	32.55 / 1.59	28.99 / 1.54	31.97 / 1.64	28.82 / 1.46
VILA-7B	34.33 / 1.83	30.40 / 1.55	35.15 / 1.80	33.38 / 1.72	29.78 / 1.58
Video-LLaVA-7B	37.48 / 1.97	30.67 / 1.63	32.50 / 1.70	36.01 / 1.85	27.36 / 1.43
LLaVA-1.5-7B	38.38 / 2.04	28.61 / 1.51	34.86 / 1.79	34.62 / 1.76	33.43 / 1.73
LongVA-7B	35.32 / 1.90	31.94 / 1.63	36.39 / 1.85	40.95 / 2.11	27.91 / 1.48
ShareGPT4Video-8B	33.28 / 1.76	39.08 / 1.94	35.77 / 1.81	37.12 / 1.89	35.62 / 1.84
InternVL-2-8B	39.08 / 2.11	33.02 / 1.74	37.47 / 1.89	44.16 / 2.22	34.89 / 1.82
AuroraCap-7B*	43.50 / 2.27	32.07 / 1.68	35.92 / 1.84	39.02 / 1.97	41.30 / 2.15
Qwen3-VL-8B*	47.66 / 2.69	38.28 / 1.35	39.98 / 2.16	42.36 / 2.56	55.56 / 2.59
Our LFS + AuroraCap-7B*	44.10 / 2.35	34.57 / 1.77	36.02 / 1.98	40.65 / 2.77	43.04 / 2.21
ΔAcc / ΔSim	+0.60 / 0.08	+2.50 / 0.09	+0.10 / 0.14	+1.63 / 0.80	+1.74 / 0.06
Our LFS + Qwen3-VL-8B*	**48.82 / 2.97**	**39.58 / 1.75**	**41.62 / 2.59**	**43.59 / 2.87**	**57.58 / 2.71**
ΔAcc / ΔSim	+1.16 / 0.32	+1.30 / 0.40	+1.64 / 0.43	+1.23 / 0.31	+2.02 / 0.12

Table 2: Results of VDC benchmark. * represents the baseline models, the bold numbers represent the best scores in the open-source models. Acc and Sim indicates the accuracy and the similarity between the outputs and answers, respectively.

stratified Top-K strategy distributes samples across the timeline and concentrates on high-importance segments, preserving both early context and late-stage frames. Consequently, captions from LFS-selected frames recover more detailed procedures (boiling, cooling, adding rice wine, stirring, garnishing, and baking) than uniform sampling.

5.2 Analysis of VDC Benchmark

We evaluate our approach on VDC, which measures accuracy and similarity scores. Following AuroraCap, we prompt Llama-3.1-8B to answer based on the predicted captions. As shown in Table 2, integrating LFS with AuroraCap-7B and Qwen3-VL-8B yields consistent improvements across categories. In particular, the largest gains appear in Detailed, where accuracy improves from 41.30% to 43.04% for AuroraCap-7B and from 55.56% to 57.58% for Qwen3-VL-8B. These results align with our method design: event-aware scoring plus stratified selection better captures critical moments for detailed descriptions, and the improvements generalize across architectures.

5.3 Analysis of Dream-1K Benchmark

Table 3 summarizes results on Dream-1K. Compared with vanilla Tarsier2-7B using uniform sampling, Our LFS + Tarsier2-7B achieves the same F1 (0.40), with a slight recall improvement (0.48 vs. 0.47) and a marginal precision decrease (0.34 vs. 0.35). This similarity mainly arises from dataset characteristics: Dream-1K consists of very short clips

(typically $<$ 10s), where 8 frames already provide near-complete coverage, leaving limited headroom for selection. The modest recall gain indicates LFS can still recover additional cues, while the overall effect reflects a recall–precision trade-off. Overall, Dream-1K serves as a short-video robustness test, showing LFS can be integrated into Tarsier2-7B without degrading performance, while its advantages are more evident on longer videos (VDC and ICH-CC).

Model	Precision	Recall	F1
GPT-4V	0.30	0.41	0.34
GPT-4o	0.36	0.43	0.39
Gemini1.5 Pro	0.35	0.38	0.36
Video-LLaVA	0.16	0.28	0.20
MiniGPT-4V	0.22	0.26	0.24
LLaVA-NeXT-Video	0.21	0.36	0.26
VideoChat2	0.23	0.31	0.27
PLLaVA-34B	0.22	0.38	0.28
Tarsier2-7B*	**0.35**	0.47	**0.40**
Our LFS + Tarsier2-7B*	0.34	**0.48**	**0.40**
ΔPrecision / Recall / F1	-0.01	+0.01	+0.00

Table 3: Precision, Recall and F1 of Dream-1K benchmark. The numbers in bold indicates the best performance among the open-source models.

Model	MVBench	VideoMME w/o sub	MLYU MCQ	Video MMMU
GPT5-Nano	-	49.4	52.6	40.2
Gemini2.5-Flash-Lite	-	65.0	69.3	63.0
Qwen2.5-VL-72B	70.4	**73.3**	74.6	60.2
Qwen3-VL-4B	68.9	69.3	75.3	56.2
Qwen3-VL-8B*	68.7	71.4	78.1	65.3
Our LFS + Qwen3-VL-8B*	**70.8**	72.6	**79.0**	**66.8**
ΔAcc	+2.1	+1.2	+0.9	+1.5

Table 4: Accuracy (%) of video QA benchmark. The numbers in bold indicates the best performance.

Method (K=16)	ICH-CC-en %	ICH-CC-zh %	VDC Detailed %
Uniform Sampling	68.20	74.25	55.56
LFS w/o Stratified	67.12	71.23	56.20
LFS w/o H_2	72.58	77.21	57.30
LFS w/o $\mathcal{L}_{\text{cap}}$	71.53	77.10	57.20
LFS w/o gating	72.41	77.23	56.98
LFS w/o norm	72.34	76.36	56.42
Full LFS	**72.67**	**77.43**	**57.58**

Table 5: Ablation study of our LFS. *LFS w/o Stratified* removes the stratified mechanism, *LFS w/o* H_2 removes event-level temporal modeling, *LFS w/o* $\mathcal{L}_{\text{cap}}$ removes caption-guided supervision, *LFS w/o gating* removes global gating, and *LFS w/o norm* removes normalization.

5.4 Zero-Shot Video Question-Answering

Table 4 reports video QA results where models answer questions solely from video descriptions, without access to raw frames, directly testing whether detailed captions support downstream reasoning. LFS consistently improves Qwen3-VL-8B across all benchmarks. On MVBench, accuracy increases from 68.7% to 70.8%, indicating improved capture of complex actions. On VideoMME without subtitles, LFS yields a +1.2% gain, suggesting stronger visual grounding from captions alone. We also observe improvements on MLYU-MCQ (79.0%) and VideoMMMU (66.8%), both of which require fine-grained event understanding. These results confirm that LFS produces more informative and event-aware descriptions, enabling reliable zero-shot QA task.

5.5 Ablation Study

All ablations are conducted using Qwen3-VL-8B as the fixed captioning backbone to isolate the contribution of individual LFS components. Table 5 reports results with a fixed frame budget of K=16. The full model achieves the best performance across all benchmarks (72.67% on ICH-CC-en, 77.43% on ICH-CC-zh, and 57.58% on VDC Detailed), while removing any component degrades performance.

Removing the stratified mechanism results in the largest drop (e.g., ICH-CC-zh: 77.43 $\rightarrow$ 71.23; VDC: 57.58 $\rightarrow$ 56.20), highlighting the importance of enforcing temporal coverage and avoiding clustering for long, stage-wise events. Removing event-level modeling (H_2) or caption-guided supervision ($\mathcal{L}_{\text{cap}}$) also reduces performance, confirming the need to model coherent events and optimize selection for caption quality. Removing global gating or normalization causes smaller but consistent drops, indicating their role in stabilizing temporal scoring. Overall, these results validate that LFS benefits from the joint design of stratified selection, event-aware temporal modeling, and caption-guided supervision for accurate and temporally diverse detailed captioning.

6 Conclusion

We developed a learnable frame selector named LFS before video-LLMs by integrating event-aware temporal modeling, stratified Top-K selection, and caption-guided supervision for detailed video captioning. Experiments show that LFS consistently improves video-LLM backbones and produces more informative video descriptions, demonstrating the value of effective frame selection for scalable video understanding.

References

[Awad *et al.*, 2023] George Awad, Keith Curtis, Asad Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Eliot Godard, Lukas Diduch, Jeffrey Liu, Yvette Graham, and Georges Quenot. An overview on the evaluated video retrieval tasks at trecvid 2022, 2023.

[Bain *et al.*, 2021] 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.

[Buch *et al.*, 2025] Shyamal Buch, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. Flexible frame selection for efficient video reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29071–29082, 2025.

[Chai *et al.*, 2025] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark, 2025.

[Chen *et al.*, 2024] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. Longvila: Scaling long-context visual language models for long videos, 2024.

[Diba *et al.*, 2023] Ali Diba, Vivek Sharma, Mohammad. M Arzani, and Luc Van Gool. Spatio-temporal convolution-attention video network. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 859–869, 2023.

[Ge *et al.*, 2024] Shiping Ge, Qiang Chen, Zhiwei Jiang, Yafeng Yin, Liu Qin, Ziyao Chen, and Qing Gu. Implicit location-caption alignment via complementary masking for weakly-supervised dense video captioning, 2024.

[Guo *et al.*, 2025] Weiyu Guo, Ziyang Chen, Shaoguang Wang, Jianxiang He, Yijie Xu, Jinhui Ye, Ying Sun, and Hui Xiong. Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. In *Advances in Neural Information Processing Systems*, 2025.

[He *et al.*, 2025] Xingjian He, Sihan Chen, Fan Ma, Zhicheng Huang, Xiaojie Jin, Zikang Liu, Dongmei Fu, Yi Yang, Jing Liu, and Jiashi Feng. Vlab: Enhancing video language pretraining by feature adapting and blending. *IEEE Transactions on Multimedia*, 27:2168–2180, 2025.

[Hu *et al.*, 2025] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-llm based video frame selection for efficient video understanding. 2025.

[Kim *et al.*, 2024] Sungkyung Kim, Adam Lee, Junyoung Park, Andrew Chung, Jusang Oh, and Jay-Yoon Lee. Towards efficient visual-language alignment of the q-former for visual reasoning tasks, 2024.

[Kim *et al.*, 2025] Younggun Kim, Ahmed S. Abdelrahman, and Mohamed Abdel-Aty. Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 761–771, October 2025.

[Li *et al.*, 2016] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, 2016.

[Li *et al.*, 2024] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2024.

[Li *et al.*, 2025a] Pengyi Li, Irina Abdullaeva, Alexander Gambashidze, Andrey Kuznetsov, and Ivan Oseledets. Maxinfo: A training-free key-frame selection method using maximum volume for enhanced video understanding. *arXiv preprint arXiv:2502.03183*, 2025.

[Li *et al.*, 2025b] Ping Li, Tao Wang, Xinkui Zhao, Xianghua Xu, and Mingli Song. Pseudo-labeling with keyword refining for few-supervised video captioning. *Pattern Recognition*, 159:111176, 2025.

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

[Maaz *et al.*, 2024] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

[Qasim *et al.*, 2025] Iqra Qasim, Alexander Horsch, and Dilip Prasad. Dense video captioning: A survey of tech-

niques, datasets and evaluation protocols. *ACM Comput. Surv.*, 57(6), February 2025.

[Sigurdsson *et al.*, 2016] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 510–526, Cham, 2016. Springer International Publishing.

[Song *et al.*, 2024] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232, 2024.

[Tang *et al.*, 2025a] Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Jinglin Xu, and Hao Sun. Tspo: Temporal sampling policy optimization for long-form video language understanding. *arXiv preprint arXiv:2508.04369*, 2025.

[Tang *et al.*, 2025b] Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. *arXiv preprint arXiv:2502.21271*, 2025.

[Tian *et al.*, 2025] Junrui Tian, Zexi Lin, Yi Dai, Yang Ding, Jinlei Liu, Lei Cao, and Ling Feng. Keyframes selection from multiscene videos for stress detection. *Information Processing and Management*, 62(5):104215, 2025.

[Wang *et al.*, 2024] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models, 2024.

[Wu *et al.*, 2023] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256, 2023.

[Wu *et al.*, 2025] Kangyi Wu, Pengna Li, Jingwen Fu, Yizhe Li, Yang Wu, Yuhan Liu, Jinjun Wang, and Sanping Zhou. Event-equalized dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8417–8427, June 2025.

[Xu *et al.*, 2024] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava : Parameter-free llava extension from images to videos for video dense captioning, 2024.

[Zhang *et al.*, 2023] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[Zhang *et al.*, 2024] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. *arXiv preprint arXiv:2403.15378*, 2024.

[Zhang *et al.*, 2025] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms, 2025.

[Zheng *et al.*, 2023] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[Zhou *et al.*, 2018] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L. Berg. Visual to sound: Generating natural sound for videos in the wild. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018.

[Zhou *et al.*, 2024] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18243–18252, 2024.

[Zhu *et al.*, 2025] Zi-Xuan Zhu, Hailun Xu, Yang Luo, Yong Liu, Kanchan Sarkar, Zhenheng Yang, and Yang You. Focus: Efficient keyframe selection for long video understanding. *ArXiv*, abs/2510.27280, 2025.

[Zhuang *et al.*, 2020] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. Multichannel attention refinement for video question answering. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(1s), March 2020.