GIFT: Global Irreplaceability Frame Targeting for Efficient Video Understanding

Supplementary Material

In the appendix, we provide more baseline, benchmark and model details in Experiments.

A. Baseline Details

We compare our method against the following training-free token compression strategies:

- **BOLT** [29] addresses the limitations of uniform sampling in long-form videos by introducing a training-free frame selection strategy based on Inverse Transform Sampling (ITS). By leveraging a lightweight proxy model (e.g., CLIP) to compute fine-grained frame-query similarity, ITS constructs a cumulative distribution function based on these similarity scores to sample frames probabilistically. This mechanism effectively prioritizes query-relevant frames while preserving temporal diversity, thereby mitigating the redundancy issues inherent in deterministic Top-K selection. Following the original settings, we set the sharpness hyperparameter α to 2.5.
- **AKS** [39] addresses the inefficiency of uniform sampling by introducing Adaptive Keyframe Sampling strategy. By formulating frame selection as an optimization problem that balances frame-query relevance with temporal coverage, AKS employs a recursive "judge-and-split" algorithm to select informative frames. This mechanism effectively maximizes useful visual context within a limited token budget while minimizing redundancy. Following the original configuration, we set the max depth L to 3 and the threshold s_{thr} to 0.2.

B. Benchmark Details

We evaluate GIFT on various video understanding benchmarks detailed as follows:

- **MVBench** [24] presents 20 distinct video understanding tasks, each comprising 200 QA pairs. It is specifically formulated to evaluate temporal comprehension capabilities that extend beyond static single-frame analysis, thereby providing a holistic assessment of model performance.
- **LongVideoBench** [46] comprises 3,763 videos paired with 6,678 multiple-choice questions. Spanning diverse domains such as movies and news, it rigorously assesses a model's capacity for temporal information retrieval and complex analysis within extended contexts.
- **MLVU** [59] is tailored for long-form video understanding, featuring content with durations ranging from 3 minutes to over 2 hours (averaging 12 minutes). It encompasses diverse genres, including documentaries, movies, and TV series, and evaluates models across nine distinct tasks, such as video summarization, topic reasoning, and needle-in-a-haystack question answering.
- **VideoMME** [10] consists of 900 videos and 2,700 QA pairs, with durations varying from 11 seconds to 1 hour. The benchmark is stratified into three temporal subsets (short, medium, and long-term) and covers six primary visual domains, including knowledge and life record.

C. Model Details

We evaluate our method against several VLMs:

- **LLaVA-Video** [58] builds upon the single-image stage checkpoint of LLaVA-OneVision and is subsequently fine-tuned on a large synthetic video instruction dataset (LLaVA-Video-178K). Architecturally, it synergizes the SigLIP visual encoder with the Qwen2 Large Language Model and introduces newline tokens for each frame to distinguish spatial and temporal positions, ensuring robust video comprehension across diverse benchmarks.
- **LLaVA-OneVision** [18] integrates single-image, multi-image, and video tasks within a unified architecture by treating video input as a continuous sequence of visual tokens. This paradigm facilitates seamless transfer learning from image to video tasks, thereby bolstering zero-shot video understanding capabilities. Furthermore, the model leverages bilinear interpolation to optimize token usage, enabling the processing of a larger number of frames by compressing the token per frame.
- **Qwen2.5-VL** [1] incorporates window attention within its native dynamic resolution ViT to efficiently process visual inputs of arbitrary resolutions and aspect ratios. It upgrades the M-RoPE to align with absolute time, which allows the model to perceive real-world temporal dynamics rather than frame indices. This temporal alignment, combined with dynamic FPS sampling, enables the model to achieve precise event localization and effectively comprehend long-form videos.
- **VideoLLaMA3** [54] prioritizes video understanding through a vision-centric training paradigm that leverages large-scale image-text datasets before extending to video data. It employs Any-resolution Vision Tokenization (AVT) to process images and videos of dynamic resolutions by introducing 2D-RoPE. Furthermore, a Differential Frame Pruner (DiffFP) is integrated to compress video representations by discarding redundant visual tokens based on inter-frame similarity in pixel space, thereby achieving efficient video understanding.