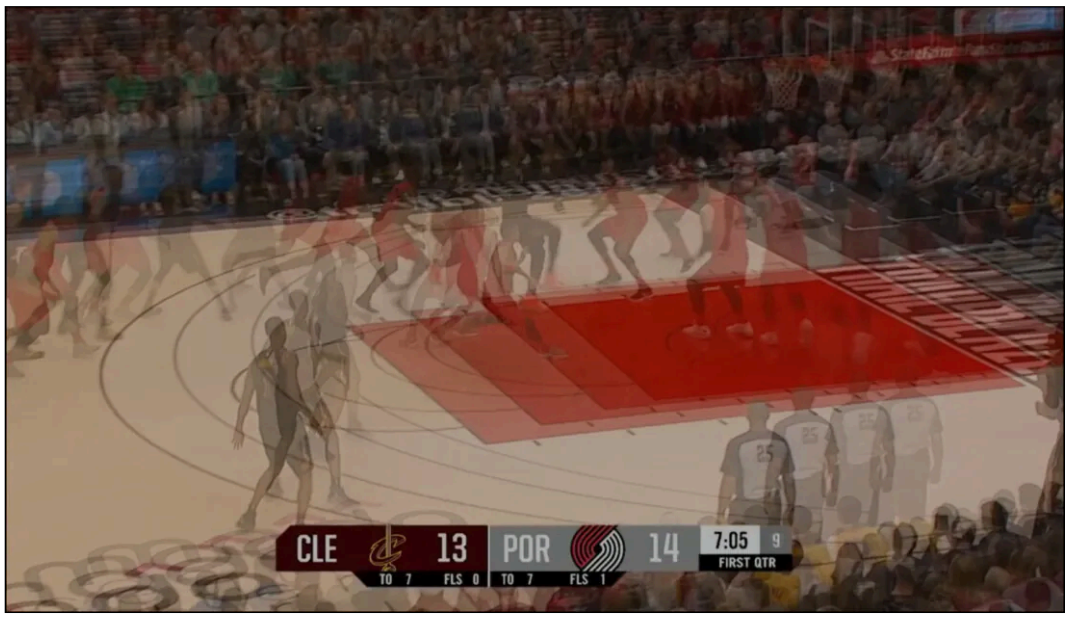
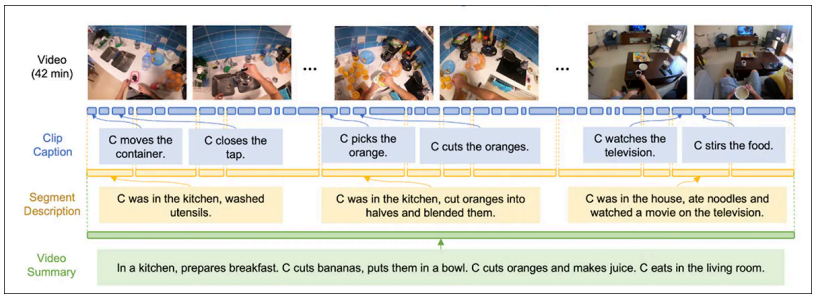
The Challenge of Captioning Video at More Than 1fps

Originally published at https://www.unite.ai/the-challenge-of-captioning-video-at-more-than-1fps/



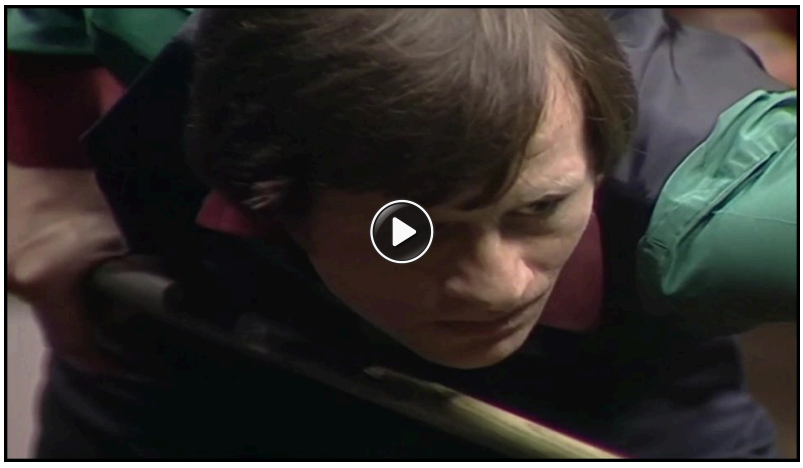
The ability for machine learning systems to recognize the events that occur inside a video is crucial to the future of AI-based video generation – not least because video datasets require accurate captions in order to produce models that adhere to a user's request, and that do not excessively hallucinate.



An example of a captioning schema from Google's VidReCap project. Source: https://sites.google.com/view/vidrecap

Manually captioning the scale of videos needed for effective training datasets is an unconscionable prospect. Although it is possible to train AI systems to auto-caption videos, a great many human-generated examples are still needed as ground truth, for variety and coverage.

More importantly, almost every current AI-based video-captioning model *operates at 1fps*, which is not a dense enough capture rate to discern variations in a great many scenarios: sudden micro-expression changes for emotion-recognition systems; rapid events in high-speed sports such as basketball; violent movements; rapid cuts in dramatic movies, where systems such as PySceneDetect may fail to identify them (or are not being used); and many other scenarios where the window of attention clearly needs to be more intense.



CLICK TO PLAY VIDEO ON SITE

Click to play. *Rapid but life-changing action in what can otherwise be one of the slowest sports in the world, as Alex Higgins clinches the world championship against Ray Reardon in 1982.* Source: https://www.youtube.com/watch?v=_1PuqKno_Ok

Move Fast and Break Logic

This low rate is the standard for various logistical reasons. For one, video-captioning is a resource-intensive activity, whether the system is studying one sequential frame at a time, or else using various methods to semantically cohere a string of frames into an interpretable caption sequence. In either case, the context window is inevitably limited by hardware constraints.

Another reason for 1fps being the current standard is that videos are not generally stuffed with rapid events; it is therefore redundant to give 300 frames of static snooker table the same attention as the split-second in which a potted black ball wins the championship (see example above).

It is possible to use broader secondary cues to identify pivotal moments in a sports video, such as the sustained crowd reaction to a rapid slam-dunk in a basketball game. However, such clues may occur for other reasons (such as unexpected player injuries), and can't be relied on. This is one example of how a mislabeled video dataset can lead to a generative video model that hallucinates or misinterprets instructions, i.e., because the model might show a player injury when it was asked to generate a slam-dunk (because the 'secondary clue' of crowd-agitation was not exclusive to one specific type of event).

This is in many ways a 'budgetary' problem, and in other ways a procedural problem. Frameworks to date have operated on the principle that sparse keyframes can effectively capture essential information, but this is more effective in establishing genre and other facets of a video's subject matter, since evidence, in that case, persists over multiple frames.

F-16

A new paper from China is offering a solution, in the form of the first multimodal large language model (MLLM, or simply LLM) that can analyze video *at 16fps* instead of the standard 1fps, while avoiding the major pitfalls of increasing the analysis rate.

In tests, the authors claim that the new system, titled *F-16*, outperforms proprietary state-of-the-art models such as GPT-4o and Google's Gemini-1.5 pro. While other current models were able to match or exceed F-16's results in trials, the competing models were far larger and unwieldier.

Though F-16 was trained on some serious hardware (as we'll examine shortly), inference is usually far less demanding than training. Therefore we can hope that the code (promised for a near-future release) will be capable of running on medium or high-level domestic GPUs.

What's needed for the vitality of the hobbyist scene (and that includes the professional VFX scene, most of the time) is a video-captioning model of this kind that can operate, perhaps quantized, on consumer systems, so that the entire generative video scene does not migrate to API-based commercial systems, or force consumers to hook local frameworks up to commercial online GPU services.

Beyond Scaling Up

The authors observe that this kind of approach is a practical alternative to scaling up datasets. One can infer also that if you were going to throw more data at the problem, this is still the kind of approach that could be preferable, because the new system distinguishes events in a more granular way.

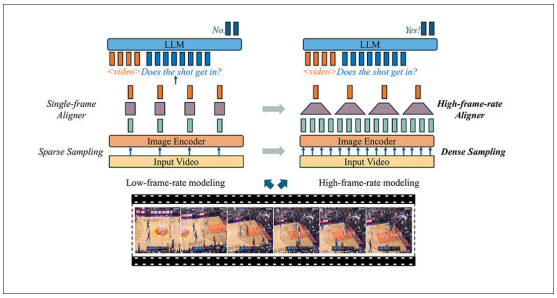
They state:

> *'Low frame rate sampling can result in critical visual information loss, particularly in videos with rapidly changing scenes, intricate details, or fast motion. Additionally, if keyframes are missed, yet the model is trained on labels that rely on keyframe information, it may struggle to align its predictions with the expected content, potentially leading to hallucinations and degraded performance…*
>
> *'… F-16 achieves SOTA performance in general video QA among models of similar size and demonstrates a clear advantage in high-frame-rate video understanding, outperforming commercial models such as GPT-4o. This work opens new directions for advancing high-frame-rate video comprehension in multimodal LLM research.'*

The new paper is titled Improving *LLM Video Understanding with 16 Frames Per Second*, and comes from eight authors across Tsinghua University and ByteDance.

Method

Since consecutive frames often contain redundant information, F-16 applies a high-frame-rate aligner to compress and encode key motion details while retaining visual semantics. Each frame is first processed by a pretrained image encoder, extracting feature representations before being passed to an aligner based on Gaussian Error Linear Units (GELUs).



F-16's architecture processes video at 16 FPS, capturing more frames than traditional low-frame-rate models, and its high-frame-rate aligner preserves visual semantics while efficiently encoding motion dynamics without adding extra visual tokens. Source: https://arxiv.org/pdf/2503.13956

To handle the increased frame count efficiently, F-16 groups frames into small processing windows, merging visual features using a three-layer Multi-Layer Perceptron (MLP), helping to retain only the most relevant motion details, and reducing unnecessary duplication, while preserving the temporal flow of actions. A spatial max-pooling layer further compresses the token count, keeping computational costs within bounds.

The processed video tokens are then fed into the Qwen2-7B LLM, which generates textual responses based on the extracted visual features and a given user prompt.

By structuring video input this way, F-16 enables, the authors assert, more precise event recognition in dynamic scenes, while still maintaining efficiency.

The Short Version

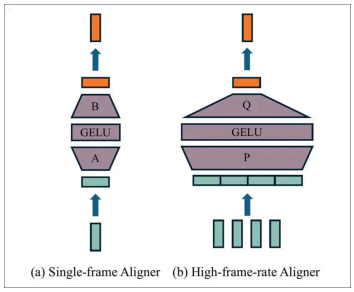
F-16 extends a pretrained image LLM, LLaVA-OneVision, to process video by transforming its visual input pipeline. While standard image LLMs handle isolated frames, F-16's high-frame-rate aligner reformats multiple frames into a form the model can more efficiently process; this avoids overwhelming the system with redundant information while preserving key motion cues necessary for accurate video understanding.

To ensure compatibility with its image-based foundation, F-16 reuses pretrained parameters by restructuring its aligner into *sub-matrices*. This approach allows it to integrate knowledge from single-frame models while adapting to sequential video input.

The aligner first compresses frame sequences into a format optimized for the LLM, preserving the most informative features while discarding unnecessary details. The architecture design enables the system to process high-frame-rate video while keeping computational demands under control, which the authors posit as evidence that scaling is not the only (or the best) way forward for video captioning.

Varying the Pace

Since processing video at 16 FPS improves motion understanding but increases computational cost, particularly during inference, F-16 introduces a *variable-frame-rate decoding* method, allowing it to adjust frame rate dynamically without retraining.



The single-frame and high frame rate aligners available to F-16.

This flexibility enables the model to operate efficiently at lower FPS when high precision isn't required, and reduces computational overhead.

At test time, when a lower frame rate is selected, F-16 reuses previously trained aligner parameters by repeating input frames to match the expected dimensions. This ensures the model can still process video effectively without modifying its architecture.

Unlike naive downsampling (i.e., simply removing frames), which risks losing critical motion details, this method preserves the aligner's learned motion representations, maintaining accuracy even at reduced frame rates. For general video comprehension, a lower FPS setting can speed up inference without significant performance loss, while high-speed motion analysis can still leverage the full 16 FPS capability.

Data and Tests

Built on Qwen2-7B, FP-16 extends LLaVA-OneVision using SigLIP as an image encoder. With video frames sampled at 16 FPS, up to 1,760 frames can be obtained from each video. For longer video clips, frames were uniformly (i.e., more sparsely) sampled.

For training, F-16 used the same general video datasets as LLaVA-Video, including LLaVA-Video-178K, NExT-QA, ActivityNet-QA, and PerceptionTest.

F-16 was additionally fine-tuned on the high-speed sports datasets FineGym, Diving48, and SoccerNet. The authors also curated a collection of 276 NBA games played between November 13 and November 25, 2024, focusing on whether a shot was successful (a task requiring high-frame-rate processing).

The model was evaluated using the NSVA test set, with performance measured by F1 score.

Gymnastics and diving models were evaluated based on event recognition accuracy, while soccer and basketball models tracked passes and shot outcomes.

The model was trained for 1 epoch using *128* NVIDIA H100 GPUs (and at a standard-issue 80GB of VRAM per GPU, this entailed the use of 10,24 terabytes of GPU memory; even by recent standards, this is the highest-specced GPU cluster I have personally come across in keeping up with computer vision research literature). A learning rate of 2×10^{-5} was used during training.

Additionally, a LoRA was fine-tuned on sports data used LoRA adapters with 64 GPUs for 5 epochs. Here, only the LLM was trained, leaving the image encoder frozen.

Opposing frameworks tested in the initial round for 'general video understanding' were GPT-4o; Gemini-1.5-Pro; Qwen2-VL-7B; VideoLLaMA2-7B; VideoChat2-HD-7B; LLaVA-OV-7B; MiniCPM-V2.6-8B; LLaVA-Video-7B; and NVILA-7B;

The models were evaluated on Video-MME; VideoVista; TemporalBench; MotionBench; Next-QA; MLVU; and LongVideoBench.

Model	FPS / #Frame	Video-MME				NQA ↑	TPB ↑	MB ↑	VST ↑	MLVU ↑	LVB ↑
		Avg.↑	S↑	M↑	L↑	Test	Short	Val.	Avg.	m-Avg.	Val.
GPT-4o	1FPS / 384	71.9	80.0	70.3	65.3	-	38.0	33.0	78.3	54.5	66.7
Gemini-1.5-Pro	1FPS	75.0	81.7	74.3	67.4	-	26.6	51.0	-	-	64.0
Qwen2-VL-7B	2FPS / 768	63.3	-	-	-	-	24.7	52.0	**75.6**	-	55.6
VideoLLaMA2-7B	16	47.9	-	-	-	-	-	-	60.5	48.4	-
VideoChat2-HD-7B	16	45.3	-	-	-	79.5	-	-	-	37.4	39.3
LLaVA-OV-7B	32	58.2	-	-	-	79.4	21.2	-	73.0	51.4	56.3
MiniCPM-V2.6-8B	64	60.9	71.3	59.4	51.8	-	21.4	52.0	-	-	54.9
LLaVA-Video-7B	32	63.3	-	-	-	83.2	22.9	-	-	**70.8**	**58.2**
NVILA-7B	256	64.2	75.7	62.2	**54.8**	82.2	-	-	-	70.1	57.7
F-16-7B (ours)	16FPS / 1760	**65.0**	**78.9**	**63.2**	52.8	**84.1**	**37.2**	**54.5**	74.4	70.3	57.6

Comparison of video QA results across models, showing FPS limits and performance on multiple benchmarks. F-16 achieves SOTA among 7B models on Video-MME, NQA, TPB, and MB, rivaling proprietary models such as GPT-4o and Gemini-1.5-Pro.

Of these results, the authors state:

> *'On the Video-MME Short, Medium, and NeXT-QA datasets—each designed for short video understanding—our model surpasses the previous 7B SOTA model by 3.2%, 1.0%, and 0.9% in accuracy, highlighting its strong performance on short videos.*
>
> *'For benchmarks evaluating long video understanding, such as Video-MME Long, LongVideoBench, and MLVU, the challenge is greater due to sparser frame sampling, causing frames within the processing window to exhibit more significant variations.*
>
> *'This increases the difficulty for the modality aligner to effectively encode temporal changes within the limited token representation. As a result, F-16 experiences a slight performance drop compared to [LLaVA-Video-7B], which is trained on the same video dataset.'*

F-16's high-frame-rate processing, the authors continue, also resulted in a 13.5% improvement on TemporalBench and a 2.5% gain on MotionBench, compared to existing 7B models, and performed at a similar level to commercial models such as GPT-4o and Gemini-1.5-Pro.

High Speed Sports Video Understanding

F-16 was tested on FineGym, Diving48, SoccerNet, and NBA datasets to evaluate its ability to understand high-speed sports actions.

Using the 10,000 manually annotated NBA clips, the training focused on ball movement and player actions, and whether the models could correctly determine if a shot was successful, using the NSVA test set evaluated with F1 score.

Model	Gym	Diving	NBA	Soccer
	%Acc.↑	%Acc.↑	%F1↑	%Acc.↑
GPT-4o	-	-	76.7	36.8
Gemini-1.5-Pro	-	-	80.6	43.1
F-16 - FPS = 1	48.5	76.0	87.1	55.4
F-16 - FPS = 16	**64.1**	**86.5**	**92.9**	**57.7**

Results of high-speed sports video analysis. F-16 with the high-frame-rate aligner performed better than its low-frame-rate counterpart across all sports tasks. GPT-4o and Gemini-1.5-Pro were also evaluated on NBA and SoccerNet QA, where in-domain training knowledge was not required.

On FineGym, which measures gymnastics action recognition, F-16 performed 13.8% better than the previous 7B SOTA model, demonstrating improved fine-grained motion understanding.

Diving48 required identifying complex movement sequences such as takeoff, *somersault*, *twist*, and *flight* phases, and F-16 showed higher accuracy in recognizing these transitions.

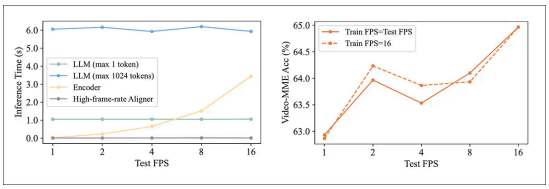
For SoccerNet, the model analyzed 10-second clips, identifying ball passes, and results showed an improvement over existing 7B models, indicating that higher FPS contributes to tracking small and rapid movements.

In the NBA dataset, F-16's ability to determine shot outcomes approached the accuracy of larger proprietary models such as GPT-4o and Gemini-1.5-Pro, further suggesting that higher frame rates enhances its ability to process dynamic motion.

Variable Frame-Rates

F-16 was tested at different frame rates to measure its adaptability. Instead of retraining, it handled lower FPS by repeating frames to match the aligner's input structure. This approach retained more performance than simply removing (prone to cause accuracy loss).

The results indicate that while reducing FPS had some impact on motion recognition, F-16 still outperformed low-frame-rate models and maintained strong results even below 16 FPS.



Left, the time consumption of different F-16 modules during inference, measured on 300 videos from the Video-MME Long set at varying test FPS and sequence lengths. Right, a comparison between Video-MME performance for models trained and tested at different FPS. The solid line represents models trained and tested at the same FPS, while the dashed line shows performance when a model trained at 16 FPS is tested at a lower frame rate.

F-16's high-frame-rate processing increased computational requirements, although its aligner helped manage these costs by compressing redundant visual tokens.

The model required more FLOPs per video than lower-FPS models, but also achieved better accuracy per token, suggesting that its frame selection and token compression strategies helped offset the added computation.

Conclusion

It is difficult to overstate either the importance or the challenges of this particular strand of research – especially this year, which is set to be the breakthrough year for generative video, throwing the shortcomings of video dataset curation and captioning quality into sharp relief.

It should also be emphasized that the challenges involved in getting accurate descriptions of internal video details cannot be solved exclusively by throwing VRAM, time, or disk space at the issue. The method by which events are isolated/extracted from otherwise long and tedious tracts of video (as with golf or snooker video clips, for instance) will benefit from a rethink of the semantic approaches and mechanisms currently dominating SOTA solutions – because some of these limitations were established in more resource-impoverished times.

(incidentally, even if 16fps seems like a very low frame rate for 2025, it is interesting to note that this is also the native training speed of video clips used in the hugely popular Wan 2.1 generative video model, and the speed at which it therefore operates with fewest issues. Hopefully the research scene will keep an eye on possible 'standards entropy' here; sometimes obsolete constraints can perpetuate future standards)

First published Wednesday, March 19, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai