

Demo-ICL: In-Context Learning for Procedural Video Knowledge Acquisition

Yuhao Dong^{1*}, Shulin Tian^{1*}, Shuai Liu¹, Shuangrui Ding^{2,3},

Yuhang Zang², Xiaoyi Dong², Yuhang Cao², Jiaqi Wang², Ziwei Liu¹

¹S-Lab, Nanyang Technological University, ²Shanghai AI Lab, ³CUHK-MMLab

*Equal Contributions.

{yuhao013, ziwei.liu}@ntu.edu.sg

Abstract

Despite the growing video understanding capabilities of recent Multimodal Large Language Models (MLLMs), existing video benchmarks primarily assess understanding based on models’ static, internal knowledge, rather than their ability to learn and adapt from dynamic, novel contexts from few examples. To bridge this gap, we present **Demo-driven Video In-Context Learning**, a novel task focused on learning from in-context demonstrations to answer questions about the target videos. Alongside this, we propose **Demo-ICL-Bench**, a challenging benchmark designed to evaluate demo-driven video in-context learning capabilities. Demo-ICL-Bench is constructed from 1200 instructional YouTube videos with associated questions, from which two types of demonstrations are derived: (i) summarizing video subtitles for text demonstration; and (ii) corresponding instructional videos as video demonstrations. To effectively tackle this new challenge, we develop **Demo-ICL**, an MLLM with a two-stage training strategy: video supervised fine-tuning and information-assisted direct preference optimization, jointly enhancing the model’s ability to learn from in-context examples. Extensive experiments with state-of-the-art MLLMs confirm the difficulty of Demo-ICL-Bench, demonstrate the effectiveness of Demo-ICL, and thus reveal future research directions for video in-context learning.

1 Introduction

Video understanding remains challenging. Recent Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Awadalla et al., 2023; Huang et al., 2023; Zhao et al., 2023; Peng et al., 2023; Wang et al., 2024a; Chen et al., 2024c; Li et al., 2024a; Dong et al., 2025) show progress from short-clip recognition (Fabian Caba Heilbron and Niebles, 2015; Mangalam et al., 2023; Goyal et al., 2017; Zhang et al., 2025d) to long videos analysis (Chandrasegaran

et al., 2024) from daily-life settings (Lei et al., 2019; Mangalam et al., 2023) to instructional videos (Miech et al., 2019; Tang et al., 2019; Zhukov et al., 2019). However, existing video benchmarks typically pose questions that rely on either internal pre-trained knowledge (e.g., asking “what is a whisk?”) or by visible facts in the target video (e.g., “where is the whisk?”). This is fundamentally different from a more challenging scenario where a model must learn a new process or skill from demonstrations (e.g., a video tutorial that teaches the model how to cook Mexican Rice) and then apply that learned knowledge to answer questions based on a new, related target video sequence. This scenario reflects human learning and is crucial for downstream applications like robotics, where robots can learn from demonstrations to tackle new tasks. For example, in Fig 1, after seeing only the first step of heating oil for Mexican Rice, the model is asked “what should you do next?” based on *in-context* text instructions or video demonstrations. This question requires knowing the specific sequence of steps for this particular version of Mexican Rice that the model is presumably meant to understand or follow, such as based on the in-context video demonstrations. To better encourage models to learn new skills from context and adapt to novel tasks, we propose a challenging video understanding task called **Demo-driven Video In-Context Learning (Demo-driven ICL)**. Our task embodies this by presenting target videos and questions alongside *in-context* text guidelines or video demonstrations. As shown in Fig 1, Demo-driven ICL has three sub-settings: (1) text-demo in-context learning, (2) video-demo in-context learning, and (3) demonstration selection. These questions explicitly require models to use the knowledge provided within in-context examples, rather than just their static internal knowledge. A key difference between Demo-driven ICL

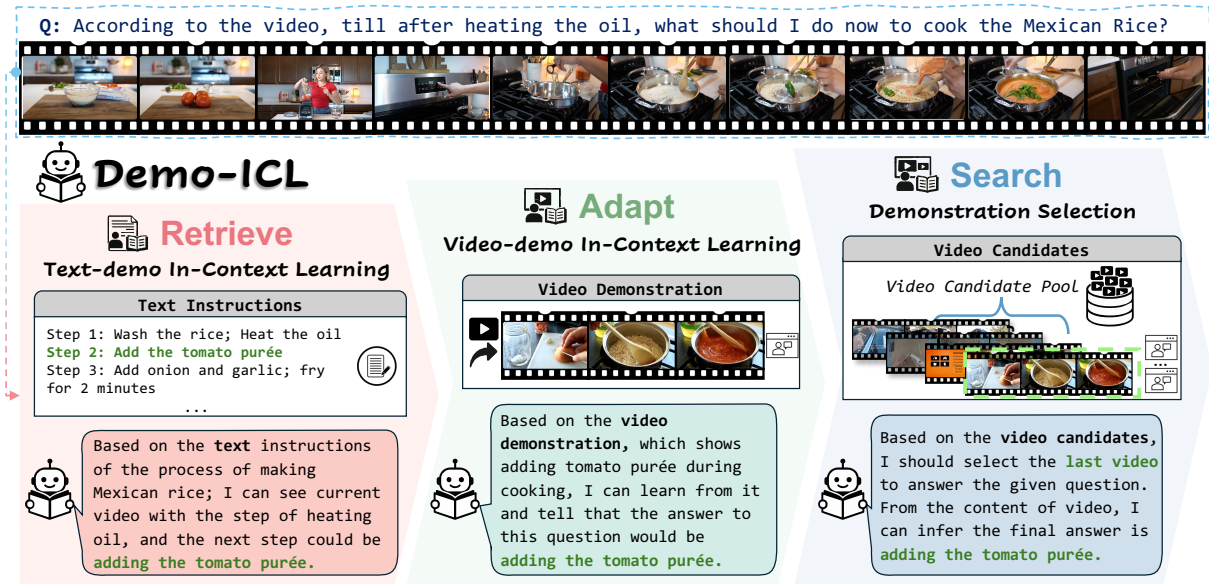


Figure 1: **Overview of the Demo-driven Video In-Context Learning Task** with three distinct settings: (1) *Text-demo in-context learning*, where text instructions act as the demonstrations; (2) *Video-demo in-context learning*, where a video demonstration acts as the reference; and (3) *Demonstration Selection*, which requires identifying the relevant video demonstrations among the video candidate pool and using them to guide video in-context learning.

and previous in-context learning paradigms lies in the data modality and interaction type: our in-context demonstrations can be videos, and the task involves choosing from a candidate pool to identify suitable in-context examples by the model itself. The Demo-driven ICL tasks mirror how one might learn a complex skill, such as cooking, by searching for related video demonstrations and watching them while also consulting supplementary visual or textual guides. To evaluate the proposed Demo-driven ICL task, we present **Demo-ICL-Bench**, a benchmark that consists of text/video demonstrations, target videos, questions, and answers. We collected instructional YouTube videos from the HowTo100M dataset (Miech et al., 2019), ensuring subtitles and timestamps. Subsequently, we used an LLM to summarize these subtitles, generating text demonstrations to serve as in-context examples. Additionally, we employ video search ranking methods and an LLM to identify and select videos similar to the target video to serve as in-context video demonstrations, and we also construct a video candidate pool for the model to select from and learn, in order to mimic real-world scenarios. Demo-ICL-Bench is complex and challenging: answering every question demands an accurate understanding of the demonstrations, resulting in frontier models like Gemini-2.5-Pro achieving merely 46.6% and 32.0% accuracy when processing text and video demonstrations, respectively. To address

Demo-driven ICL, we present **Demo-ICL** with a two-stage training strategy: video supervised fine-tuning, and information-assisted Direct Preference Optimization (DPO) for demo-driven video ICL. We design an information-assisted DPO data generation pipeline to produce high-quality chosen responses by simplifying the task with additional contextual information. Demo-ICL outperforms existing MLLMs on the proposed Demo-driven ICL, VideoMMU for video knowledge acquisition (Hu et al., 2025), and VideoMME (Fu et al., 2024a) for general video understanding. Our key contributions are: (i) **New Challenging Tasks**: We design three Demo-driven Video In-Context Learning (Demo-driven ICL) tasks, which enable models to answer questions by learning from text or video demonstrations, representing a significant step towards more human-like learning and decision-making processes in video understanding tasks. (ii) **New Benchmark and Evaluation**: We establish a new Demo-ICL-Bench that is specifically designed for evaluating demo-driven video in-context learning capabilities. Based on Demo-ICL-Bench, we conduct comprehensive evaluations of cutting-edge baselines, showcasing various challenges of our proposed task. (iii) **New Demo-ICL Model**: We present a new model, Demo-ICL, along with a customized two-stage training strategy that enhances a model’s ability to learn and adapt from in-context demonstrations. Compared with

SOTA models, Demo-ICL shows competitive performance across existing benchmarks (Fu et al., 2024a; Wu et al., 2024; Hu et al., 2025), demonstrating its superior video comprehension and in-context knowledge acquisition capabilities.

2 Related Work

Multimodal Video Understanding for Knowledge Acquisition. Multimodal video understanding is evolving from low-level perception toward *knowledge acquisition* - the ability to extract, structure, and apply information from complex instructional videos, where large-scale instructional datasets play a pivotal role in this shift. HowTo100M (Miech et al., 2019) introduced 1.2 million narrated videos with 136 million clip-caption pairs for procedure recognition and cross-task transfer. Other instruction-based datasets (Tang et al., 2019; Zhukov et al., 2019) provide fine-grained task annotations or support weakly supervised step parsing across diverse procedures. More recently, benchmarks like Video-MMMU (Hu et al., 2025), Video-MMLU (Song et al., 2025), and VideoMathQA (Rasheed et al., 2025) start to probe models’ capabilities to learn from educational videos, shifting emphasis from perception to knowledge uptake and application. To further benchmark the video understanding task for knowledge acquisition that is closer to real-world settings, we introduce Demo-driven Video In-Context Learning and propose Demo-ICL-Bench, to evaluate models’ capabilities of acquiring new concepts via given video demonstrations.

Multimodal In-Context Learning. In-context learning (ICL) enables models to perform new tasks by conditioning on a few examples at inference. Initially developed for large language models (LLMs), ICL has been extended to multimodal settings and shows consistent performance gains across language and image tasks (Dou et al., 2026; Brown et al., 2020; Min et al., 2021; Luo et al., 2024; Zhou et al., 2024b). However, video-based ICL remains underexplored: current video MLLMs (Lin et al., 2024a; Maaz et al., 2024; Zhang et al., 2025e) mainly emphasize zero-shot performance through curated video instruction datasets for open-ended QA, captioning, and dialog capabilities. Emerging works add chain-of-thought methods for video understanding and reasoning tasks (Wang et al., 2024c; Han et al., 2025; Arnab et al., 2025; Tian et al., 2025a; Zhang et al.,

2025c; Ghazanfari et al., 2025) encourage step-wise evidence aggregation and explicit explanation, while some retrieval-based methods like VideoRAG (Tevissen et al., 2024; Ren et al., 2025) establish a new paradigm of retrieving video moments and ground answers in cited segments. However, the aforementioned works merely use the context as a reference, rather than adapting to and learning from the provided context. In our work, we address this gap by introducing the Demo-driven ICL task on instructional video datasets, supported by an optimized training pipeline, to enhance the model’s capability to learn from the in-context video demonstrations.

3 Demo-ICL

3.1 Demo-driven Video In-Context Learning

Learning from demonstrations is central to human procedural knowledge acquisition: people can watch a few examples, infer the task logic, and adapt it to a new situation. In contrast, current video MLLMs are usually evaluated on fixed skills learned during pre-training or supervised fine-tuning, rather than on whether they can acquire a new procedure from the input context. This setting is important for realistic applications such as instructional assistance and robotics, where an agent may need to watch a tutorial, identify the current progress of a new execution, and decide the next action. We therefore study whether MLLMs can learn task-specific procedures from in-context examples and transfer them to a related target video.

We define **Demo-driven Video In-Context Learning** with three settings. Given either an instructional video V_D or a text demonstration T_D , the model must first interpret how the demonstrated task should be completed. It then observes a target video prefix $V_{Test}[0, t_1]$ and predicts the subsequent action $A_{[t_1, t_2]}$ by using the demonstration as procedural context. The task therefore requires more than recognizing objects or actions in the target video: the model must align the current state with the demonstrated procedure and infer what should happen next.

1) Text-demo In-Context Learning: The model uses step-by-step textual instructions as the demonstration and answers what action should follow the current target video state. For example, after observing that the target video has reached the step of heating oil for Mexican Rice, the model should consult the textual procedure and predict

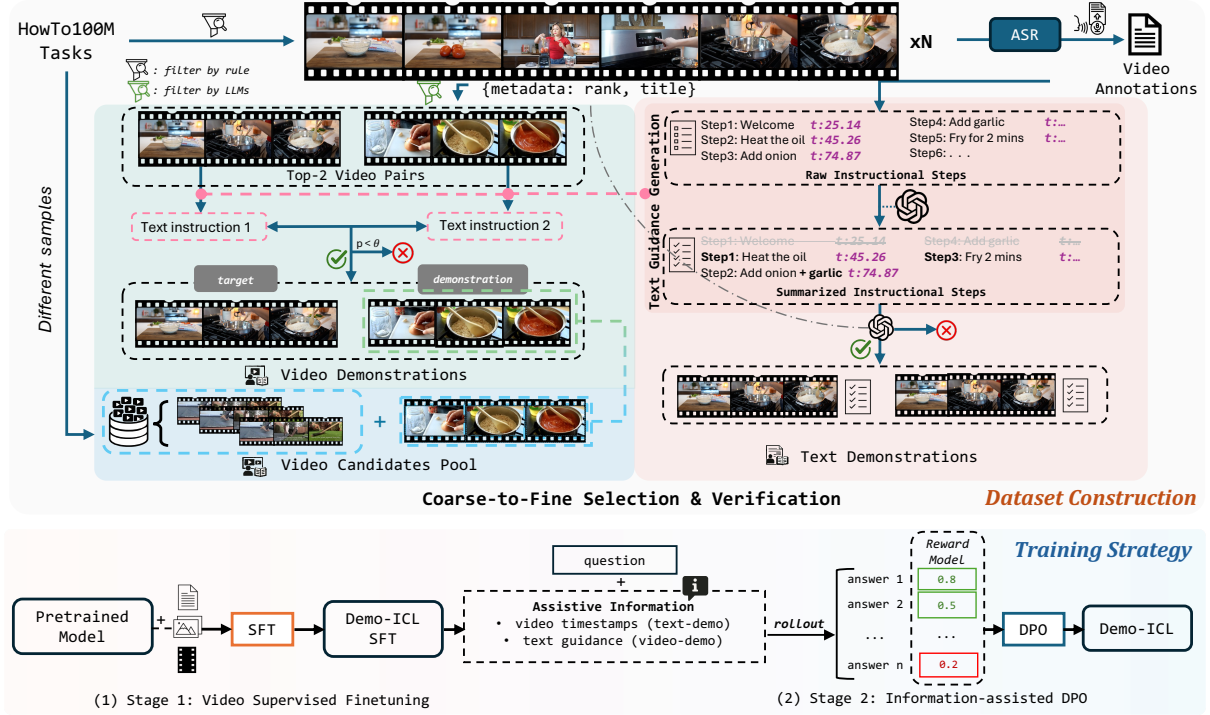


Figure 2: **Overview of Data Construction and Training Strategy.** (i) illustrates our coarse-to-fine dataset collection pipeline; (ii) presents the tailored two-stage training strategy for training the Demo-ICL model.

the following step. This setting provides an oracle textual procedure and tests whether the model can align the target video with the correct step.

2) Video-demo In-Context Learning: The model receives a related instructional video as the demonstration and must infer the transferable procedure directly from visual-temporal evidence. Unlike text-demo ICL, the procedural structure is implicit and must be extracted from another video, requiring temporal alignment between two executions of a similar task.

3) Demonstration Selection: The model is given the input video and a pool of video candidates (e.g., a pool containing “Mexican rice,” “fried rice,” and “pasta,”). The model must first **select the most relevant demonstration** (e.g., “Mexican rice”) from the pool and then use it to answer the question, simulating a scenario where a perfectly aligned demonstration is not provided.

These settings form a progression from oracle textual guidance to raw video demonstrations and finally to selecting the useful demonstration under distractors, covering both idealized and practical demo-driven video ICL scenarios.

3.2 Dataset Construction

Video Collection and Annotation. We construct our benchmark using HowTo100M (Miech et al.,

2019), a large-scale corpus of over 100 million narrated YouTube instructional clips spanning 23,000 complex activities. Its step-by-step demonstrations are highly effective for demo-driven video in-context learning and procedural knowledge acquisition. Following an initial filtering by video length, language, and title availability, human annotators curate the dataset to ensure suitable difficulty and visual relevance to knowledge acquisition. To obtain detailed activity descriptions, we use ASR transcripts as they offer more detailed descriptions of demonstrated activities compared with video captions. We utilize HTM-AA (Han et al., 2022) to extract precise sentence and word-level timestamps via WhisperX (Bain et al., 2023).

Text Demonstration Generation. As shown in Fig 2, we generate step-by-step textual demonstrations through a coarse-to-fine pipeline. First, Qwen2.5-72B (Bai et al., 2025b) parses ASR transcripts to extract the essential instructional steps required for task completion. Then, to maintain a coherent, task-focused sequence, we filter out irrelevant actions and merge redundant steps into adjacent relevant segments. Finally, Qwen2.5-VL-72B (Bai et al., 2025b) refines these demonstrations by jointly analyzing the textual descriptions alongside their corresponding video clips, ensuring precise visual grounding and contextual accuracy.

By successfully eliminating verbose and irrelevant summaries, this multistage refinement yields reliable procedural guidance and achieves a 94% human satisfaction rate, substantially outperforming the 56% rate of a direct ASR to LLM baseline.

Video Demonstration Selection. To enable video-based in-context learning, we construct paired video demonstrations of similar tasks. This explicit visual guidance allows models to observe alternative task executions and acquire reliable procedural knowledge. We employ a coarse-to-fine pair selection strategy, which first leverages HowTo100M metadata, specifically YouTube search rankings, to filter out tasks lacking sufficient relevant videos. Next, Qwen2.5-72B evaluates top-ranked video titles to identify the pair with the highest semantic similarity for each task. Finally, we rigorously validate these candidates by generating textual instructions for both videos via our established pipeline. The Qwen2.5-72B LLM then compares these instructions to confirm that the tasks are genuinely similar and transferable, ensuring the accuracy and reliability of the selected demonstrations.

Demo-driven Question Generation. With curated videos and corresponding instructions, we construct questions to evaluate demo-driven video in-context learning. For text-based in-context learning, we exclude tasks with fewer than six steps to ensure sufficient complexity. This strict filtering ensures that the resulting questions maintain a high standard of difficulty, requiring a genuine procedural comprehension. From each valid sequence, one intermediate step (excluding the first and last) is randomly selected, and a question is generated from this step. The model is then required to predict the next action.

To evaluate whether models effectively leverage visual demonstrations for in-context learning, we employ Qwen2.5-72B to analyze the generated instructions of paired videos and determine their suitability for question generation. Once validated, the LLM identifies the target step and its corresponding timestamp within the pair. Human annotators then assess the generated questions and paired videos to ensure the visual demonstrations provide meaningful contextual evidence. This human-in-the-loop verification guarantees the logical soundness and high fidelity of the visual context. Finally, we filter out overly similar video pairs to prevent trivial solutions, ensuring the benchmark rigorously tests model adaptability and generalization.

For the demonstration selection task, we treat

validated video pairs as ground truth and augment them with three carefully chosen irrelevant videos. Rather than maintaining a static set, we dynamically sample candidates to form a unique pool for every question. This construction requires models to distinguish informative demonstrations from distractors, which is an essential capability for real-world deployment. By calibrating the difficulty of these distractors, the benchmark faithfully simulates complex in-context learning environments.

Dataset Partition and Benchmark Statistics. Following established protocols, we first generate 5,000/2,000/1,000 questions for text-demo/video-demo ICL and demonstration selection settings. To construct Demo-ICL-Bench, we then manually curate representative video demonstrations that highlight demo-driven video in-context learning. Specifically, we sample 500 questions for both the text-demo and video-demo tasks and 200 questions for the demonstration selection task, resulting in a balanced benchmark of 1,200 questions in total.

3.3 Learning from In-Context Demonstration

3.3.1 Video Supervised Fine-tuning

To equip Demo-ICL with fine-grained video understanding and general in-context reasoning, we compile a large-scale dataset comprising millions of samples from diverse academic sources. For image-text data, we rely on resources such as LLaVA-OneVision (Li et al., 2024a), VisualWebInstruct (Jia et al., 2025), and other widely used collections. For video data, we incorporate material from open-source projects including LLaVA-Video (Zhang et al., 2024c), Oryx (Liu et al., 2024), and Ola (Liu et al., 2025). To further enhance the model’s ability for instructional video understanding, we additionally incorporate datasets such as COIN (Tang et al., 2019) and Cross-Task (Zhukov et al., 2019). To prevent data leakage and ensure fair evaluation, we strictly exclude any videos overlapping with Demo-ICL-Bench. Finally, as detailed in 3.2, we sample the constructed dataset to explicitly introduce demo-driven in-context learning capabilities. Together, these curated resources establish the robust foundational capabilities necessary for subsequent training stages.

3.3.2 Information-Assisted Preference Optimization

Current models struggle with demo-driven video in-context learning, making it difficult to generate high-quality preferred responses with a con-

Table 1: **Evaluation results on Demo-ICL-Bench.** The benchmark assesses models across three tasks. For Text-demo ICL and Video-demo ICL, we report two types of accuracy: Demo. Acc (with demos) and w/o Demo (without demos), and the improvement (Δ_{ICL}) attributed to demonstrations. DS refers to Demonstration Selection task. S.Acc refers to demonstration selection accuracy.

Model	Size	Frame	Text-demo ICL			Video-demo ICL			DS		Avg
			Demo. Acc	w/o Demo.	Δ_{ICL}	Demo. Acc	w/o Demo.	Δ_{ICL}	S.Acc	Acc	
Human	-	-	84.0	-	-	80.4	-	-	88.0	76.0	80.1
<i>Proprietary MLLMs</i>											
Gemini-2.5-Pro (Gemini Team, 2024)	-	-	54.4	-	-	36.2	-	-	-	26.0	38.9
GPT-4o (OpenAI, 2024)	-	-	48.8	-	-	31.4	-	-	-	24.5	34.9
<i>Open-Source Video MLLMs</i>											
Qwen2-VL (Wang et al., 2024a)	7B	32	29.0	21.8	+7.2	22.4	24.0	-1.6	38.0	14.5	22.0
Ola (Liu et al., 2025)	7B	32	32.2	23.0	+9.2	24.6	26.4	-1.8	43.0	16.0	24.3
LLaVA-Video (Zhang et al., 2024c)	7B	32	31.0	25.4	+5.6	30.2	29.0	+1.2	44.5	20.5	27.2
Qwen2.5-VL (Bai et al., 2025b)	7B	32	32.8	26.0	+6.8	28.0	26.2	+1.8	46.0	18.0	26.3
InternVL-3 (Zhu et al., 2025)	8B	32	31.4	26.6	+4.8	27.0	26.4	+0.6	44.0	16.5	25.0
Qwen3-VL-Instruct (Bai et al., 2025a)	8B	32	44.8	34.6	+10.2	33.2	29.8	+3.4	54.5	24.5	34.2
Qwen3-VL-Thinking (Bai et al., 2025a)	8B	32	45.8	35.2	+10.6	36.0	31.8	+4.2	57.5	25.5	35.8
Video-R1 (Feng et al., 2025)	7B	32	33.6	27.6	+6.0	27.4	26.6	+0.8	48.0	17.5	26.2
VideoChat-R1 (Li et al., 2025)	7B	32	34.4	27.0	+7.4	28.2	26.8	+1.4	52.0	18.5	27.0
Qwen2.5-VL (Bai et al., 2025b)	72B	32	45.0	24.2	+20.8	25.6	25.2	+0.4	54.0	18.0	29.5
Qwen3-VL-Instruct (Bai et al., 2025a)	235B	32	58.4	38.6	+19.8	40.2	34.4	+5.8	66.5	28.5	42.4
Qwen3-VL-Thinking (Bai et al., 2025a)	235B	32	61.2	39.4	+21.8	42.8	35.6	+7.2	69.0	30.0	44.7
Ola-Video (Liu et al., 2025) (Base)	7B	32	31.4	22.8	+8.6	26.4	26.0	+0.4	48.0	18.0	24.8
Demo-ICL (SFT)	7B	32	38.4	27.8	+10.6	29.4	26.2	+3.2	54.5	21.5	29.8
Demo-ICL	7B	32	43.4	29.4	+14.0	32.0	27.6	+4.4	58.0	24.0	33.1
Demo-ICL(Qwen3-VL)	8B	32	52.4	37.8	+14.6	41.4	34.6	+6.8	70.5	32.0	41.9

ventional self-sampling pipeline. To address this issue, we propose an information-assisted DPO pipeline that integrates automatically generated assistive information during preference-data construction, without requiring manual annotation. For text-demo ICL, we supply video timestamps to align visual inputs with textual instructions. For video-demo ICL, we pair video demonstrations with corresponding textual guidance to help the model produce responses grounded in the demonstrated procedure. These responses are used as preferred outputs, while normal model responses are treated as rejected samples.

Formally, we define a preference dataset as $\mathcal{P} = \{(x^{(i)}, R_c^{(i)}, R_r^{(i)})\}_{i=1}^{|\mathcal{P}|}$, where $x^{(i)}$ denotes the user request, and $R_c^{(i)}$ and $R_r^{(i)}$ represent the preferred and less preferred responses. We use a reward model $r^*(x, y)$ to approximate preferences, where a higher score denotes a stronger preference. Let $r_I(y) = r^*((x, I), y)$ and $r_x(y) = r^*(x, y)$, where I denotes the assistive information. Following the Bradley-Terry model (Bradley and Terry, 1952), the preference probability is:

$$p^*(y_1 \succ y_2 | x) = \frac{\exp(r_I(y_1))}{\exp(r_I(y_1)) + \exp(r_x(y_2))} = \sigma(r_I(y_1) - r_x(y_2)). \quad (1)$$

We estimate the reward model by minimizing the

negative log-likelihood:

$$\mathcal{L}_R = -\mathbb{E}_{\mathcal{P}} [\log \sigma(r_\phi(x, R_c) - r_\phi(x, R_r))]. \quad (2)$$

Using the resulting preference data, we perform multiple training rounds to obtain models $\mathcal{M}_1, \dots, \mathcal{M}_T$, where $\mathcal{M}_{t+1}$ is trained with preference data $\mathcal{P}_t$ generated by $\mathcal{M}_t$. Through this information-assisted and iterative preference optimization strategy, we progressively strengthen Demo-ICL’s demo-driven video in-context learning capability.

4 Experiments

4.1 Demo-ICL-Bench

Setup. We evaluate representative proprietary MLLMs, open-source video MLLMs, video reasoning models, and our trained variants across all three Demo-ICL-Bench tasks. For text-demo and video-demo ICL, we additionally report performance without demonstrations to isolate the benefit of in-context procedural knowledge. The baseline pool covers GPT-4o, Gemini-2.5-Pro, Qwen-series models, InternVL, LLaVA-Video, Ola, and specialized video reasoning models. We include both compact and large baselines where available, allowing us to examine how demo-driven ICL scales with model capacity and whether targeted training can compensate for limited model size.

Table 2: **General Video Understanding.** Demo-ICL achieves superior performance on both general temporal understanding and knowledge acquisition tasks, highlighting the effectiveness of the proposed demo-driven video in-context learning framework. These results also demonstrate the robustness and generalizability of our training strategy across diverse evaluation settings. † represents results reproduced by ourselves.

Model	Size	General Temporal Understanding				Knowledge
		VideoMME (wo / w sub)	MVBench	LongVideoBench	MLVU	Video MMMU
GPT-4o (OpenAI, 2024)	-	71.9/77.2	-	66.7	66.7	61.2
Gemini-2.5-Pro (Team., 2025)	-	84.3/86.9	-	-	-	83.6
LLaVA-OneVision (Li et al., 2024a)	7B	58.2 / 61.5	56.7	56.3	64.7	33.9
VideoLLaMA3 (Zhang et al., 2025a)	7B	66.2 / 70.3	69.7	59.8	73.0	47.0
LLaVA-Video (Zhang et al., 2025e)	7B	63.3 / 69.7	58.6	58.2	70.8	-
Qwen2.5-VL (Bai et al., 2025b)	7B	65.1 / 71.6	69.6	56.0	-	47.4
InternVL3.5 (Wang and et al., 2025)	8B	66.0 / 68.6	72.1	62.1	70.2	-
Qwen3-VL (Bai et al., 2025a)	8B	68.4†/72.6†	68.7	62.1†	78.1	63.3
Ola-Video (Base)	7B	63.0 / 68.9	66.9	60.7	69.1	46.2
Demo-ICL (SFT)	7B	65.6 / 70.2	69.4	61.6	70.4	48.8
Demo-ICL	7B	65.2 / 69.7	68.6	61.2	70.4	52.6
Demo-ICL(Qwen3-VL)	8B	69.6 / 73.4	70.8	62.4	78.8	67.1

Text-demo In-context Learning. As shown in Table 1, text demonstrations consistently improve over the no-demo setting, showing that the questions require task-specific context. The gains are larger for stronger models: Qwen2.5-VL-72B and Qwen3-VL-235B obtain much larger improvements than their compact counterparts, suggesting that effective use of procedural text still depends on general reasoning and video grounding ability. Targeted training also helps smaller models: Demo-ICL improves over its base and SFT variants, and Demo-ICL(Qwen3-VL) further strengthens the performance. Importantly, the no-demo results remain far below the demo-conditioned results, indicating that the benchmark is not mainly solvable by parametric recall or generic commonsense. These results establish text-demo ICL as a upper-bound-like setting for studying whether models can align a target video with an explicit procedural description.

Video-demo In-context Learning. Video demonstrations are substantially harder to use than text demonstrations. Several baselines gain little or even degrade, indicating weak extraction and transfer of temporal-visual procedure cues from a separate demonstration video. This gap is expected: the model must parse the reference video, recover the latent step structure, align it to the current state of the target video, and then predict the next action. In contrast, Demo-ICL consistently benefits from video demonstrations, though the gains remain smaller than in the text-demo setting. Demo-ICL(Qwen3-VL) obtains a 6.8% gain, narrowing the gap to much larger models on this track. This contrast highlights the central challenge of our task: models must not only perceive videos, but also abstract procedures and apply them to a new execu-

tion. It also suggests that future video MLLMs need stronger mechanisms for temporal abstraction and cross-video alignment.

Demonstration Selection. The demonstration selection task adds the need to identify the relevant reference before answering, approximating real-world scenarios where a perfectly aligned demonstration is not directly provided. We report both selection accuracy and final answer accuracy. Current models show a large gap from human performance on both metrics, suggesting that they struggle to capture global task semantics and are easily distracted by irrelevant candidate videos. This failure compounds the difficulty of video-demo ICL: even when a useful demonstration exists in the context, the model must first allocate attention to it before performing procedural transfer. Demo-ICL improves this setting, with the Qwen3-VL variant achieving the best model accuracy in 1. These results show that demonstration selection is not merely a retrieval add-on, but a necessary capability for realistic demo-driven video ICL. Further analysis is provided in Sec 4.3.

4.2 General Video Understanding

Setup. We also evaluate whether demo-driven training preserves general video ability beyond Demo-ICL-Bench. We test knowledge acquisition on VideoMMMU (Hu et al., 2025), where models must learn and apply information from videos, and temporal understanding on VideoMME (Fu et al., 2024a), MVBench (Li et al., 2024b), LongVideoBench (Wu et al., 2024), and MLVU (Zhou et al., 2024a). These benchmarks cover both knowledge-oriented video comprehension and general temporal reasoning, allowing us to

check whether the proposed training improves procedural knowledge acquisition without degrading broader video understanding.

Results. As shown in Table 2, Demo-ICL remains competitive with open-source models of similar size on general temporal understanding while improving knowledge-oriented evaluation on VideoMMU. This suggests that the training does not simply overfit to Demo-ICL-Bench, but also benefits related video knowledge acquisition settings. The Qwen3-VL variant further improves both knowledge acquisition and several temporal understanding scores, showing that the proposed training can be applied to stronger backbones. These results indicate that rather than trading off with general video perception, the model maintains comparable performance on standard video benchmarks while gaining stronger knowledge acquisition ability.

4.3 Analysis Experiments

The Model Learns or Recalls? To investigate whether the tasks in Demo-ICL-Bench can be bypassed using only pre-trained internal knowledge, we evaluate Gemini-2.5-Pro in a zero-shot setting without any contextual demonstrations. With GPT-4o as the judge, the model achieves a mere 7% success rate. This minimal performance confirms that Demo-ICL-Bench effectively resists simple parametric recall. Because the videos capture highly specific action sequences, inferring the correct subsequent steps without direct visual evidence is practically impossible. Consequently, successful performance on our benchmark strictly requires models actively extract, interpret, and apply procedural knowledge from the provided in-context demonstrations rather than relying on memorization.

What Makes the Video-demo ICL Challenging? We provide deeper insights into why the Video-demo ICL task poses significant challenges for current MLLMs, with results summarized in Table 3. First, we test Demo-ICL with more densely sampled frames, and the improvements demonstrate that fine-grained visual cues are critical for demo-driven video in-context learning. Furthermore, providing the full query video as the reference using 128 frames significantly increases performance. This indicates that direct visual grounding is substantially easier than transferring knowledge from separate demonstrations, highlighting a struggle to adapt perceived information to novel scenarios. Providing only clips of the immediate next-step action as context ex-

poses another severe limitation, as models fail to align temporal evidence across distinct videos. Replacing these video clips with ASR transcripts and captions yields further gains, proving that existing MLLMs lack robust video comprehension and struggle to abstract visual segments into actionable knowledge. Ultimately,

these results demonstrate that video-demo ICL demands much more than basic perception. It requires complex temporal alignment, abstraction, and flexible knowledge transfer, emphasizing the critical need for models to utilize dynamic visual demonstrations for advanced reasoning.

How to Train a Good Demo-driven Video ICL Model? We perform ablation studies to assess the effectiveness of our proposed training strategies, with detailed results summarized in Table 4.

Table 4: Training setting ablations on Demo-ICL.

Settings	Text-ICL	Video-ICL	DS	Avg
w/o Instructional Videos	34.0	26.2	19.0	26.4
Demo-ICL(SFT)	38.4	29.4	21.5	29.8
Vanilla DPO	40.0	30.0	22.0	30.7
Demo-ICL(DPO) 1-round	41.8	30.8	22.5	31.7
Demo-ICL(DPO)	43.4	32.0	24.0	33.1

The results show that instructional-video data, information-assisted DPO, and iterative training strategy each improve Demo-ICL on Demo-ICL-Bench. Compared with vanilla DPO, our method provides cleaner chosen responses and yields stronger demo-driven ICL performance.

5 Conclusion

In this paper, we introduce a novel task, Demo-driven Video In-Context Learning, which focuses on learning from in-context instructional demonstrations. To facilitate evaluation, we construct Demo-ICL-Bench, a benchmark consisting of 1,200 challenging questions designed to assess demo-driven video in-context learning capabilities. To effectively address this task, we further propose Demo-ICL, a video MLLM equipped with enhanced in-context learning abilities. Extensive experiments show that existing MLLMs struggle with this task, while Demo-ICL achieves stronger video understanding and in-context knowledge acquisition, highlighting the promise of demo-driven ICL for future multimodal learning.

Table 3: Ablation study on evaluation settings.

Settings	Video-demo ICL
Base(32 frames)	29.4
128 frames	30.4
+Repeat Video	38.6
+Reference Clips	35.8
+ASR & Captions	45.4

6 Limitations

In this section, we discuss the limitations of our work. The Demo-ICL model does not include a specialized architecture for demo-driven video in-context learning. Instead, we employ a customized training strategy to achieve this functionality. Our goal is to equip current MLLMs with demo-driven video in-context learning capability without requiring architectural modifications, thereby simplifying the integration of these new capabilities and the maintenance of previous multimodal understanding. Additionally, we did not explore how models can effectively learn from diverse contexts, such as different modalities or resources. This ability is more similar to the natural human learning process, where individuals can draw on a wide range of resources, such as text instructions and instructional videos, to enhance understanding simultaneously. Combining various types of contextual information to improve in-context learning and ultimately enhance a model’s performance on new tasks remains a significant challenge.

7 Ethics

Data provenance. The training data for Demo-ICL, as well as the collected video data for Demo-ICL-Bench construction, are used in accordance with their original licenses and intended research use. We do not crawl additional information and inherit the upstream curation practices for any identifiable individuals or sensitive scenes. Our collected data will be released as public resources, and we make sure no sensitive information will be included.

Human annotation. Human involvement in this work is limited to verifying the quality of the benchmark data. Volunteers are informed of the research purpose, work only with the collected public videos; we do not collect demographic information. The instructions for the volunteers are limited to quality check and do not involve additional annotations.

8 Acknowledgement

This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223-0002). This research is also supported by cash and in-kind funding from NTU S-Lab and industry partner(s).

References

- Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Wang Huiyu Nagarajan, and Lorenzo Torresani. 2023. Ht-step: Aligning instructional articles with how-to videos. In *Neural Information Processing Systems*.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.
- Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. 2025. [Temporal chain of thought: Long-video understanding by thinking in frames](#). *Preprint*, arXiv:2507.02001.
- Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.
- Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyza-guirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. 2024. HourVideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197.

- Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024a. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*.
- Yuxiao Chen, Kai Li, Wentao Bao, Deep Patel, Yu Kong, Martin Renqiang Min, and Dimitris N. Metaxas. 2024b. Learning to localize actions in instructional videos with llm-based multi-pathway text-video alignment. *Preprint*, arXiv:2409.16145.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.
- Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkan Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2025. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072.
- Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, Huaibing Xie, Jianglu Hu, Shaolei Wang, Weichao Wang, Yanling Xiao, Yiting Liu, Zenan Xu, Zhen Guo, Pluto Zhou, and 8 others. 2026. Cl-bench: A benchmark for context learning. *Preprint*, arXiv:2602.03587.
- Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G. Derpanis, Animesh Garg, Richard P. Wildes, and Allan D. Jepson. 2023. Stepformer: Self-supervised step discovery and localization in instructional videos. *Preprint*, arXiv:2304.13265.
- Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970.
- Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiaowu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2024b. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.
- Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiaowu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Preprint*, arXiv:2501.01957.
- Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10826–10834.
- Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.
- Sara Ghazanfari, Francesco Croce, Nicolas Flammarion, Prashanth Krishnamurthy, Farshad Khorrami, and Siddharth Garg. 2025. Chain-of-frames: Advancing video understanding in multimodal llms via frame-aware reasoning. *arXiv preprint arXiv:2506.00318*.
- Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thureau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. *Preprint*, arXiv:1706.04261.
- Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2025. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26181–26191.
- Tengda Han, Weidi Xie, and Andrew Zisserman. 2022. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916.
- Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *Preprint*, arXiv:2501.13826.
- Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, and 1 others. 2023. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109.
- Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, and Wenhui Chen. 2025. Visualwebinstruct: Scaling up multimodal instruction data through web search. *arXiv preprint arXiv:2503.10582*.

- Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.
- Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. [Less is more: Clipbert for video-and-language learning via sparse sampling](#). *Preprint*, arXiv:2102.06183.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. [Tvqa: Localized, compositional video question answering](#). *Preprint*, arXiv:1809.01696.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. [Llava-onevision: Easy visual task transfer](#). *Preprint*, arXiv:2408.03326.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.
- Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*.
- Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.
- Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. [Video-llava: Learning united visual representation by alignment before projection](#). *Preprint*, arXiv:2311.10122.
- Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoenybi, and Song Han. 2024b. [Vila: On pre-training for visual language models](#). *Preprint*, arXiv:2312.07533.
- Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2024. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*.
- Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. 2025. [Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment](#). *Preprint*, arXiv:2502.04328.
- Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. 2024. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. [Video-chatgpt: Towards detailed video understanding via large vision and language models](#). *Preprint*, arXiv:2306.05424.
- Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. [Egoschema: A diagnostic benchmark for very long-form video language understanding](#). *Preprint*, arXiv:2308.09126.
- Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.
- Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hananeh Hajishirzi. 2021. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*.
- OpenAI. 2023. Gpt-4 with vision. <https://openai.com/research/gpt-4v-system-card>. Accessed: 2025-05-21.
- OpenAI. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.
- Hanoona Rasheed, Abdelrahman Shaker, Anqi Tang, Muhammad Maaz, Ming-Hsuan Yang, Salman Khan, and Fahad Shahbaz Khan. 2025. Videomathqa: Benchmarking mathematical reasoning via multi-modal understanding in videos. *arXiv preprint arXiv:2506.05349*.
- Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. 2025. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*.
- Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevalia, and Aniruddha Kembhavi. 2021. [Visual semantic role labeling for video understanding](#). *Preprint*, arXiv:2104.00990.
- Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. 2025. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*.
- Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. [Videobert: A joint model for video and language representation learning](#). *Preprint*, arXiv:1904.01766.
- Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. [Coin: A large-scale dataset for comprehensive instructional video analysis](#). *Preprint*, arXiv:1903.02874.

- Gemini Team. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Yannis Tevissen, Khalil Guetari, and Frédéric Petitpont. 2024. Towards retrieval augmented generation over large video libraries. In *2024 16th International Conference on Human System Interaction (HSI)*, pages 1–4. IEEE.
- Shulin Tian, Ziqi Huang, Fan Zhang, Hongyuan Zhu, Yu Qiao, and Ziwei Liu. 2026. Open evaluation agent: Efficient and promptable evaluation of visual generative models. *arXiv preprint arXiv:2608.09666*.
- Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkan Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025a. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*.
- Shulin Tian, Ziniu Zhang, Liang-Yu Chen, and Ziwei Liu. 2025b. Mmina: Benchmarking multihop multimodal internet agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 13682–13697.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *Preprint*, arXiv:2409.12191.
- Weiyun Wang and et al. 2025. [Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency](#). *Preprint*, arXiv:2508.18265.
- Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024b. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.
- Yan Wang, Yawen Zeng, Jingsheng Zheng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. 2024c. Videocot: A video chain-of-thought dataset with active annotation tool. *arXiv preprint arXiv:2407.05355*.
- Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.
- Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. [Next-qa:next phase of question-answering to explaining temporal actions](#). *Preprint*, arXiv:2105.08276.
- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.
- Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*.
- Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. [Videollama 3: Frontier multimodal foundation models for image and video understanding](#). *Preprint*, arXiv:2501.13106.
- Fan Zhang, Shulin Tian, Ziqi Huang, Yu Qiao, and Ziwei Liu. 2025b. Evaluation agent: Efficient and promptable evaluation framework for visual generative models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7561–7582.
- Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkan Yang, Yuanhan Zhang, Ziyue Wang, Hao-ran Tan, Chunyuan Li, and Ziwei Liu. 2024a. [Long context transfer from language to vision](#). *Preprint*, arXiv:2406.16852.
- Yongheng Zhang, Xu Liu, Ruihan Tao, Qiguang Chen, Hao Fei, Wanxiang Che, and Libo Qin. 2025c. Vitcot: Video-text interleaved chain-of-thought for boosting video understanding in large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5267–5276.
- Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. 2025d. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20626–20636.
- Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. [Llava-next: A strong zero-shot video understanding model](#).
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024c. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.
- Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2025e. [Llava-video: Video instruction tuning with synthetic data](#). *Preprint*, arXiv:2410.02713.
- Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with

multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. 2023. [Procedure-aware pretraining for instructional video understanding](#). *Preprint*, arXiv:2303.18230.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024a. *MLvu: A comprehensive benchmark for multi-task long video understanding*. *arXiv e-prints*, pages arXiv–2406.

Yucheng Zhou, Xiang Li, Qianing Wang, and Jianbing Shen. 2024b. *Visual in-context learning for large vision-language models*. *arXiv preprint arXiv:2402.11574*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. *Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models*. *arXiv preprint arXiv:2504.10479*.

Linchao Zhu and Yi Yang. 2020. [Actbert: Learning global-local video-text representations](#). *Preprint*, arXiv:2011.07231.

Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. 2019. *Cross-task weakly supervised learning from instructional videos*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545.

Appendix

We provide supplementary documents to support our research. Implementation details are outlined in Section A. Additional visualization results are presented in Section B, followed by further experimental analysis in Section C. We also provide a more comprehensive discussion of related work in Section D. Finally, we discuss the potential risks of our work in Section E, and how AI contributes to this work F.

A Implementation Details

A.1 Experiment Details

In this section, we detail the implementation of Demo-ICL. The Demo-ICL model is built upon Ola-Video, a highly pretrained multimodal understanding model that integrates OryxViT as its visual encoder to process native arbitrary-resolution visual inputs, alongside Qwen2.5 as the language model. For the training process, we construct a customized dataset to establish foundational image

and video understanding capabilities. For image data, resolutions range from 768 to 1536, while for video data, the number of frames is capped at 64, with frame resolutions varying between 288×288 pixels and 480×480 pixels. During training, the maximum token length is set to 16,384, and a learning rate of 1e-5 is used throughout both stages. In the DPO (Direct Preference Optimization) training phase, we curate 5,000 samples using the specified pipeline and apply a learning rate of 5e-7. A batch size of 256 is maintained across both fine-tuning stages and the DPO phase, with experiments conducted using 64 NVIDIA A100 80G GPUs with around 10,000 GPU hours in total.

A.2 Data Collection Details

In the data generation process, we utilize Qwen2.5-72B as our LLM and Qwen2.5-VL-72B as our MLLM within the pipeline. For generating text instructions, we first use Qwen2.5-72B to create summarized instruction steps. Then, when refining these steps with the MLLM, we forward each step along with 64 uniformly sampled frames from the corresponding video clips. For generating questions for video-demo ICL, we provide the text instructions of paired videos and ask the LLM to assess their reasonableness for question generation. Both the LLM and MLLM are deployed using four NVIDIA A800 GPUs.

A.3 Evaluation Details

For evaluation on the Demo-ICL-Bench, we design three experimental settings. For the Text-demo ICL task, we sample 32 frames from each video and provide textual demonstrations alongside the question during evaluation. For the Video-demo ICL task, we sample 32 frames for both the reference (demonstration) video and the target video. Finally, for the Demonstration Selection task, due to context-length limitations in some models, we uniformly sample 16 frames from each candidate demonstration video and 32 frames from the target video. In all three settings, video subtitles are not provided.

A.4 License

The data and benchmarks we collected strictly follow the MIT license, and we will open-source them upon acceptance.

B Visualizations

We present visualization results to clarify the task design of Demo-ICL-Bench. These results are shown in Fig. 3 and Fig. 4.

C More Analysis Experiments

C.1 Perception Bottleneck vs. Knowledge Transfer in Video-Demo ICL

To clarify whether the performance gains from providing ASR and captions reflect a fundamental improvement in video-demo in-context learning (ICL) or primarily alleviate a perception bottleneck, we conduct two additional analyses. First, we evaluate text-demo ICL on the same data, where ground-truth textual descriptions of the demonstrations are provided directly. This setting removes the video perception requirement and isolates the knowledge transfer challenge. The model achieves 48.8% accuracy, which is higher than video-demo ICL with ASR and captions but still far from the ceiling, indicating that even with perfect textual access, distilling and transferring knowledge from demonstrations remains a substantial difficulty. Second, we assess direct video understanding using the same ASR/caption outputs: when asked to answer similar questions about the original video content, the model reaches 93.4% accuracy. This near-ceiling performance confirms that once the perception bottleneck is removed, the model understands the video content well, and the remaining gap in video-demo ICL is attributable primarily to knowledge transfer rather than perception. Together, these results suggest that ASR/caption signals currently improve video-demo ICL mainly by compensating for limited multimodal perception, while the defining capability of the task—and the target for future model improvements—is the ability to organize demonstration content into transferable knowledge and apply it to new settings. As video perception continues to improve, the relative benefit of external ASR/caption pipelines is expected to diminish, leaving knowledge transfer as the core challenge.

C.2 Impact of Frame Sampling Density

To examine how the amount of visual information affects performance, we vary the number of uniformly sampled frames (32, 64, and 128) in our model and report results on Text-demo ICL, Video-demo ICL, and Demonstration Selection. Table 5 summarizes the results.

Table 5: **Performance comparison with different numbers of uniformly sampled frames.**

Frames	Text-demo ICL	Video-demo ICL	DS
32 frames	52.4	41.4	32.0
64 frames	53.6	42.8	34.5
128 frames	54.0	43.6	36.0

Increasing frame density leads to consistent improvements across all tasks. Text-demo ICL shows a modest gain (52.4 $\rightarrow$ 54.0), which is expected because denser frames provide better visual perception of the videos, which can also benefit text-demo ICL. In contrast, Video-demo ICL and Demonstration Selection exhibit more pronounced improvements (41.4 $\rightarrow$ 43.6 and 32.0 $\rightarrow$ 36.0, respectively), confirming that richer visual input helps alleviate the perception bottleneck.

However, even at 128 frames, the performance of the two video-centric tasks remains far from saturated and still lags substantially behind Text-demo ICL (43.6 vs. 54.0 for Video-demo ICL; 36.0 for Demonstration Selection). These results indicate that while increasing frame density effectively enhances visual perception, it does not address the core challenges of our benchmark, namely, distilling transferable knowledge from demonstrations and applying it to new settings. Thus, our benchmark evaluates a combination of perception and higher-order knowledge transfer, with the latter remaining a significant open challenge.

C.3 General Video Understanding on Video-MME

We further evaluate the Demo-ICL model on general video understanding tasks of varying lengths and scenarios. Specifically, we employ the VideoMME benchmark to highlight its offline video comprehension capabilities, providing a broader assessment beyond domain-specific settings.

Setup. To further evaluate the generalization ability of Demo-ICL on diverse video understanding tasks, we adopt the Video-MME benchmark (Fu et al., 2024a). The dataset consists of 900 videos (254 hours) covering 6 visual domains and 30 sub-fields, with durations ranging from 11 seconds to 1 hour, categorized into Short, Medium, and Long. In addition to visual content, VideoMME provides audio and subtitles, enabling a multimodal and comprehensive evaluation of video MLLMs. Under

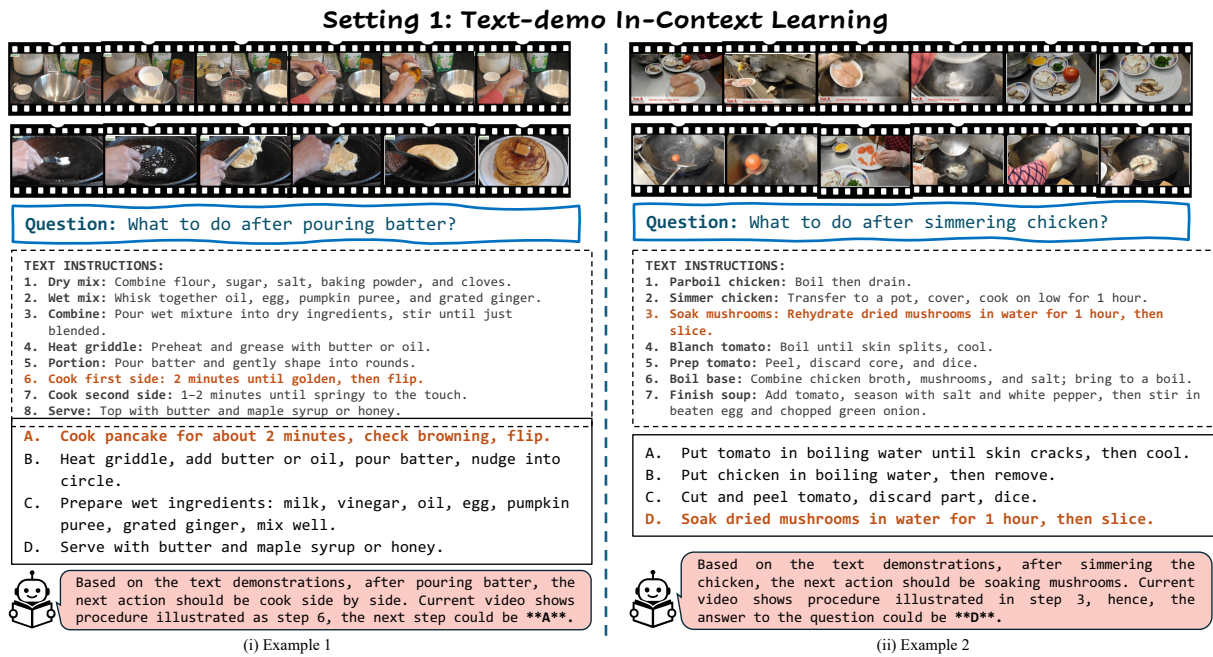


Figure 3: **Visualization of Text-demo In-Context Learning.** This figure provides 2 examples to illustrate the text-demo in-context learning task, where the text instructions will be provided along with the target video as the inputs.

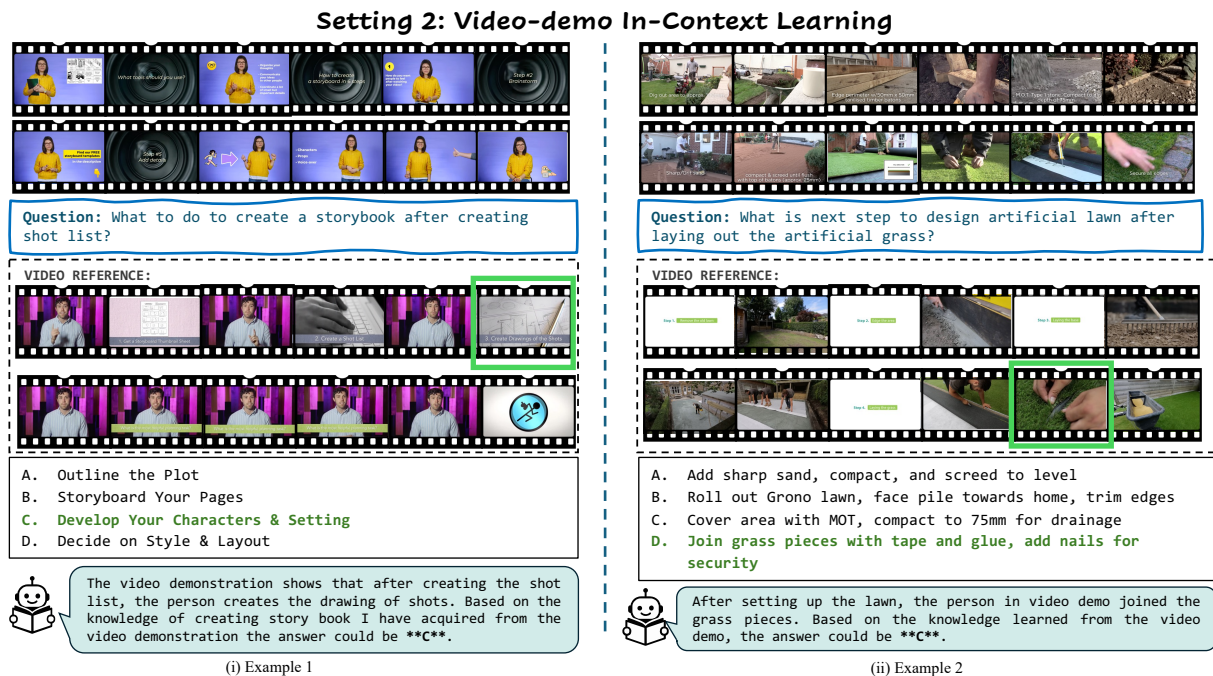


Figure 4: **Visualization of Video-demo In-Context Learning.** This figure provides 2 examples to illustrate the video-demo in-context learning task, where a video demonstration will be provided together with the target video input.

this setting, the model is required to watch an entire video and then answer corresponding questions, which allows us to systematically assess robustness across varying durations, modalities, and domains, in comparison with both open-source and commercial MLLMs.

Results. Table 6 summarizes the overall performance of Demo-ICL across short, medium, and long video tracks. Demo-ICL achieves strong results on all three tracks, demonstrating robust capabilities across different temporal lengths. It surpasses open-source video MLLMs with similar parameter sizes (7B), achieves comparable results to larger models (34B), and competes closely with some commercial MLLMs. Notably, on long-duration videos, which pose greater challenges due to extended temporal dependencies, Demo-ICL demonstrates its long video understanding capabilities, maintaining consistent performance over time.

C.4 Results on Video-MMLU Benchmark

To further validate the effectiveness and generalization capability of the proposed demo-driven video in-context learning paradigm and the info-assisted DPO training strategy, we evaluate our approach on the Video-MMLU (Song et al., 2025) benchmark, a video understanding dataset designed to assess lecture comprehension and video-based knowledge acquisition.

Setup. The Video-MMLU benchmark consists of 1,065 videos, primarily sourced from lecture recordings. The associated questions are categorized into two types: caption and reasoning questions. Since the answers are open-ended, model performance is assessed using the LLM-as-a-Judge framework, with Qwen-2.5-72B serving as the judge model.

Results. As shown in Table 7, incorporating demo-driven video in-context learning significantly enhances the model’s ability to learn from lecture videos, where the videos act as demonstrations for subsequent reasoning questions. While the model maintains competitive performance on the caption task, its performance on the Quiz track improves by a substantial margin compared to previous models. It is worth noting that we employ Qwen2.5-7B as the backbone LLM, the same base model used in Qwen2.5-VL, yet our method achieves markedly better results on the Quiz track. This

demonstrates the effectiveness of the proposed demo-driven video in-context learning and info-assisted DPO strategy in general video knowledge acquisition.

C.5 Progressive Evaluation of Video ICL

We provide in-depth analysis of the results on Demo-ICL-Bench. The three tasks in Demo-ICL-Bench establish a hierarchy toward realistic video in-context learning and systematically expose the limitations of current MLLMs. The text-demo setting serves as an oracle baseline functioning primarily as a retrieval task. Models show clear gains on this task, proving they can utilize text-based in-context learning to solve video tasks. Ideally, if a model perfectly interprets visual examples, its accuracy would approach this upper bound. However, the video-demo task introduces greater difficulty by removing textual hints while retaining the ground-truth reference to test adaptation directly from raw video. The significantly diminished gains in this setting highlight a fundamental lack of true demo-driven video ICL capabilities. Finally, the demonstration selection task mimics a real-world scenario where models must autonomously search a candidate pool for the relevant reference before adapting. The stark performance drop on this task demonstrates that current models struggle to process multiple in-context videos, let alone identify and learn from the correct one. Collectively, these cascading performance drops highlight a significant gap in model capacity and validate the reasoning behind our task design as a progressive framework to evaluate capabilities required for real-world video in-context learning scenarios.

C.6 Quality of Demo-ICL-Bench

To further evaluate the benchmark, we conduct a human review of the text-demo ICL and demonstration selection tasks, as video-demo ICL tasks are verified during generation. We random sample 200 text-demo tasks and 100 demonstration selection tasks, and the annotators confirm the accuracy and sufficiency of the textual demonstrations with a 96% pass rate. Furthermore, all candidate videos in the demonstration selection set meet the quality criteria. These strong evaluation metrics ultimately validate the high quality of our collected dataset.

C.7 Why is Demonstration Selection Difficult?

Building on the perceptual challenges of video-demo ICL task, the demonstration selection task in-

Table 6: **Performance of Demo-ICL compared to previous MLLMs on Video-MME across short, medium, and long durations, under without “subtitles” and with “subtitles” settings.**

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs
Proprietary MLLMs									
GPT-4V (OpenAI, 2023)	-	70.5	73.2	55.8	59.7	53.5	56.9	59.9	63.3
GPT-4o (OpenAI, 2024)	-	80.0	82.8	70.3	76.6	65.3	72.1	71.9	77.2
Gemini 1.5 Flash (Gemini Team, 2024)	-	79.7	83.6	68.4	74.7	61.1	68.8	70.3	75.0
Gemini 1.5 Pro (Gemini Team, 2024)	-	81.7	84.5	74.3	81.0	67.4	77.4	75.0	81.3
Open-source Video MLLMs									
LongVA (Zhang et al., 2024a)	7B	61.1	61.6	50.4	53.6	46.2	47.6	52.6	54.3
VITA 1.5 (Fu et al., 2025)	7B	67.0	69.9	54.2	55.7	47.1	50.4	56.1	58.7
mPLUG-Owl3 (Ye et al., 2024)	7B	70.0	72.8	57.7	66.9	50.1	64.5	59.3	68.1
TimeMarker (Chen et al., 2024a)	8B	71.0	75.8	54.4	60.7	46.4	51.9	57.3	62.8
MiniCPM-V 2.6 (Yao et al., 2024)	8B	71.3	73.5	59.4	61.1	51.8	56.3	60.9	63.7
VILA-1.5 (Lin et al., 2024b)	34B	68.1	68.9	58.1	57.4	50.8	52.0	59.0	59.4
Oryx-1.5 (Liu et al., 2024)	34B	77.3	80.6	65.3	74.3	59.3	69.9	67.3	74.9
Qwen2-VL (Wang et al., 2024a)	72B	80.1	82.2	71.3	76.8	62.2	74.3	71.2	77.8
LLaVA-Video (Zhang et al., 2024b)	72B	81.4	82.8	68.9	75.6	61.5	72.5	70.6	76.9
Demo-ICL	7B	78.6	79.1	63.9	68.8	53.2	61.1	65.2	69.7

Table 7: **Video Knowledge Acquisition on Video-MMLU.** Demo-ICL achieves superior performance on the Video-MMLU benchmark, showing substantial gains on the Quiz track, where the model must leverage knowledge from lecture videos to answer novel reasoning questions. These results demonstrate the effectiveness of our demo-driven video in-context learning paradigm and the strong generalization capability enabled by the info-assisted DPO training strategy.

Model	Size	Notebook	Quiz	Overall
<i>Proprietary MLLMs</i>				
Gemini-1.5-Flash	-	39.5	47.8	43.6
GPT-4o	-	53.9	44.9	49.4
Claude-3.5-sonnet	-	67.4	71.2	69.3
<i>Open-Source MLLMs</i>				
InstructBLIP-7B	7B	19.3	2.9	11.1
Cambrian-8B	8B	20.2	5.2	12.7
LLaVA-1.5	7B	22.3	9.1	15.7
Video-LLaVA-7B	7B	15.3	16.5	15.9
LLaVA-NeXT (Vicuna)	7B	18.1	24.9	21.5
Qwen-VL	7B	24.4	19.6	22.0
mPLUG-Owl3	7B	22.6	22.6	22.6
LLaVA-NeXT (LLaMA3)	8B	16.5	30.1	23.3
InternVL2-8B	7B	31.4	16.7	24.1
LLaVA-NeXT (Mistral)	7B	20.3	31.5	25.8
Qwen2-VL-7B	7B	34.2	23.4	28.8
LLaVA-NeXT-Video-7B	7B	35.8	27.4	31.6
LLaVA-OneVision-OV	7B	34.6	33.4	34.0
Qwen2.5-VL-7B	7B	42.0	32.9	37.5
InternVL2.5-8B	8B	34.5	44.5	39.5
Demo-ICL	7B	41.0	50.4	45.7

troduces the distinct difficulty of isolating informative guidance from dynamic distractor videos. Our analysis reveals that existing models achieve poor selection accuracy and exhibit fragile reasoning when navigating multiple candidates, a common challenge in real world scenarios which always lack a single isolated example. Furthermore, we uncover a critical vulnerability to context pollution by evaluating the Qwen3-VL-8B model under two

Table 8: **More experiment results on Demo-ICL-Bench.** The **T. ICL** and **V. ICL** refers to the original Text-demo ICL Demo. Acc and Video-demo ICL Demo. Acc, where **DS** refers to the original DS Acc.

Model	Size	T. ICL	V. ICL	DS	Overall
<i>Proprietary Methods</i>					
Gemini-2.5-Pro	-	54.4	36.2	26.0	38.9
GPT-4o	-	48.8	31.4	24.5	34.9
<i>Open-Source Methods</i>					
Qwen2-VL (Wang et al., 2024a)	7B	29.0	22.4	14.5	22.0
Ola (Liu et al., 2025)	7B	32.2	24.6	16.0	24.3
LLaVA-Video (Zhang et al., 2024c)	7B	31.0	30.2	20.5	27.2
Qwen2.5-VL (Bai et al., 2025b)	7B	32.8	28.0	18.0	26.3
InternVL-3 (Zhu et al., 2025)	8B	31.4	27.0	16.5	25.0
Video-R1 (Feng et al., 2025)	7B	33.6	27.4	17.5	26.2
VideoChat-R1 (Li et al., 2025)	7B	34.4	28.2	18.5	27.0
<i>RAG-based Methods</i>					
VideoRAG (Ren et al., 2025)	-	52.6	22.8	18.0	31.3
<i>Agent-based Methods</i>					
VideoAgent (Wang et al., 2024b)	-	36.2	23.4	17.5	25.7
Demo-ICL	7B	43.4	32.0	24.0	33.1

distinct settings on the demonstration selection subset. In the standard video-demo ICL setting, where the model receives exclusively the correct video, it achieves a 35.0% accuracy. However, when the model is explicitly told which candidate is the correct demonstration within the broader demonstration selection context, performance drops to 29.5%. This degradation indicates that the mere presence of irrelevant distractors in the context window actively impairs the capacity of the model to extract and apply procedural knowledge. Even when provided the correct target video, models struggle to allocate attention accurately over long multimodal inputs and are easily distracted by surrounding visual noise. These findings highlight the critical necessity of our information assisted DPO framework, which explicitly trains models to filter out distractors and anchor their reasoning strictly on the

Table 9: **Related Work for Demo-ICL-Bench.** Demo-ICL-Bench stands out due to its demo-driven video in-context learning settings, setting it apart from previous video benchmarks.

Benchmark	Video Domain	#Videos	#QAs	Video-ICL	Annotation
ActivityNet-QA (Fabian Caba Heilbron and Niebles, 2015)	Human Activities	800	8000	✗	Manual
How2QA (Li et al., 2020)	Instructional Videos	1166	2852	✗	Manual
KnowIT-VQA (Garcia et al., 2020)	TV Show	207	24k	✗	Manual
NExT-QA (Xiao et al., 2021)	Web Videos (Causal/Temporal)	5.4k	52k	✗	Manual
MVBench (Li et al., 2024b)	Benchmark Videos	3641	4000	✗	Auto
VideoMME (Fu et al., 2024b)	YouTube Videos	900	2700	✗	Manual
VideoMathQA (Rasheed et al., 2025)	Instructional Videos	420	420	✗	Manual
VideoMMMU (Hu et al., 2025)	Lectures	300	900	✗	Manual
Demo-ICL-Bench	Instructional Videos	1200	1200	✓	Mixed

most informative elements within complex demonstrations.

C.8 More Evaluation Results

To comprehensively evaluate existing video understanding methods on Demo-ICL-Bench, we further include additional categories of approaches, specifically RAG-based and agent-based methods. For each category, we select a representative model for evaluation: VideoRAG (Ren et al., 2025) for the RAG-based approach and VideoAgent (Wang et al., 2024b) for the agent-based approach. As shown in Table 8, the RAG-based method VideoRAG performs well on the Text-demo ICL task, primarily because its video knowledge indexing and multimodal retrieval paradigm effectively aligns videos with the provided text demonstrations. However, in the video-demo ICL and demonstration selection tasks, where the model must transfer knowledge learned from demonstrations to solve problems in a target video, VideoRAG still underperforms. This gap underscores the importance of these two task designs and highlights that beyond simple video knowledge acquisition, current video MLLMs must also develop the ability to transfer learned knowledge to new tasks. For the agent-based method VideoAgent, we observe that incorporating agentic strategies enhances performance on the Text-demo ICL task, primarily because they strengthen the grounding capabilities of current video MLLMs and improve alignment between videos and textual demonstrations. However, the method still underperforms on the other tracks, suggesting that while agentic approaches can enhance perceptual accuracy, the ability to transfer learned or perceived knowledge to new tasks remains an open challenge, which is likely crucial for advancing video understanding and reasoning.

D More Discussion on Related Works

In this section, we will include more details of related works.

Multimodal Video Understanding for Knowledge Acquisition. Recent research in video understanding has moved beyond low-level perception towards extracting structured knowledge from videos, like procedural steps, events, and concepts. Large-scale instructional datasets have been instrumental in this shift. For example, a lot of instructional datasets (Miech et al., 2019; Tang et al., 2019; Zhukov et al., 2019) have driven the development of models that seek to learn high-level knowledge from video, rather than just recognize objects or actions. Moreover, VidSitu (Sadhu et al., 2021) addresses video situation recognition by densely annotating 10-second movie clips with semantic role labels, which provides a symbolic knowledge representation of the video. By learning to predict such structured representations, models can acquire a form of event knowledge from videos. Similarly, HT-Step (Afouras et al., 2023) aligns the textual instructions from wikiHow (Koupaei and Wang, 2018) with corresponding segments in instructional videos. It provides 116k temporal segment annotations in 20k how-to videos, each labeled with a step description from wikiHow, enabling models to learn to ground declarative knowledge in procedural video footage.

To better learn from such knowledge-intensive data, early multimodal learning approaches applied language-modeling techniques to video data. For example, VideoBERT (Sun et al., 2019) quantizes video frames into discrete “visual words” and then uses a BERT-like transformer to learn joint representations of sequences of visual tokens and narration text. Following models such as ActBERT (Zhu and Yang, 2020) extended this masked language modeling paradigm to action recognition data, and ClipBERT (Lei et al., 2021) improved efficiency by

sampling sparse key frames for end-to-end video-text pretraining. By learning from millions of narrated video clips, these models demonstrate an ability to embed procedural and commonsense knowledge implicitly in their representations. (Zhou et al., 2023) proposed the model Paprika used PKG-based pre-training procedure to generate psuedo labels for instructional video to train. StepFormer (Dvornik et al., 2023) addresses the problem of discovering and localizing key procedure steps in instructional videos without human supervision. It uses video with subtitles (ASR) only, with a transformer decoder that attends to video frames via learnable queries to produce a sequence of key steps. (Chen et al., 2024b) proposes a framework, MPTVA, that aligns video segments with procedure steps derived via LLM from narration text via long-term semantic similarity and short-term fine-grained similarity.

Multimodal In-Context Learning. Inspired by the textual CoT prompting, recent works curate multimodal datasets with human-written rationales to encourage step-by-step prompting (Tian et al., 2025b; Zhang et al., 2025b; Tian et al., 2026). Video-CoT (Wang et al., 2024c) provides video QA examples paired with detailed explanations, while Video-Espresso (Han et al., 2025) scales this approach to large collections of reasoning exemplars. Beyond data-centric methods, (Arnab et al., 2025) propose Temporal Chain-of-Thought, an inference strategy for long videos where the model iteratively selects relevant clips and reasons over them, enabling efficient multi-step reasoning over extended sequences. A complementary line of work extends retrieval-augmented generation (RAG) to video. VideoRAG (Ren et al., 2025) and related work (Tevissen et al., 2024) index long videos into databases of visual and textual descriptors. At query time, relevant segments and transcripts are retrieved and passed to the language model as context, grounding answers in explicit video evidence. This improves factual accuracy, transparency, and scalability, especially for long videos where direct end-to-end processing is infeasible.

E Potential Risks

There are some potential risks of this work. First, since the model learns task behavior from demonstrations, it may unintentionally imitate undesirable or biased patterns present in the provided examples. Second, the use of video demonstrations can

raise privacy concerns, as videos may contain sensitive personal information, identities, locations, or unintended background content. This highlights the need for careful data curation, anonymization, and responsible deployment practices. Third, improving a model’s ability to infer procedures from demonstrations may increase the risk of misuse in safety-critical or harmful domains, where users could provide demonstrations of inappropriate actions and ask the model to generalize them. Finally, as demo-driven learning makes model behavior more dependent on the given context, the model may become more vulnerable to misleading, adversarial, or malicious demonstrations that induce incorrect predictions or unsafe responses.

F Usage of AI

We clarify that AI tools are used solely to assist with paper writing, including word- and sentence-level language refinement, and are not used for research design, data analysis, or the generation of scientific conclusions.