

APPENDIX

We provide supplementary documents to support our research. The details of Large Language Model usage are presented in Section A. Implementation details are outlined in Section B. Additional visualization results are presented in Section C followed by further experimental analysis in Section D. We also provide a more comprehensive discussion of related work in Section E. Finally, we discuss the limitations of our work in Section F.

A LARGE LANGUAGE MODEL USAGE

In this paper, we clarify that large language models (LLMs) are employed solely to support and refine the writing process. Specifically, we use LLMs to provide sentence-level suggestions and to enhance the overall fluency of the text.

B IMPLEMENTATION DETAILS

B.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 A800 GPUs.

B.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.

C VISUALIZATIONS

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

D MORE ANALYSIS EXPERIMENTS

D.1 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 subfields, with durations ranging from 11 seconds

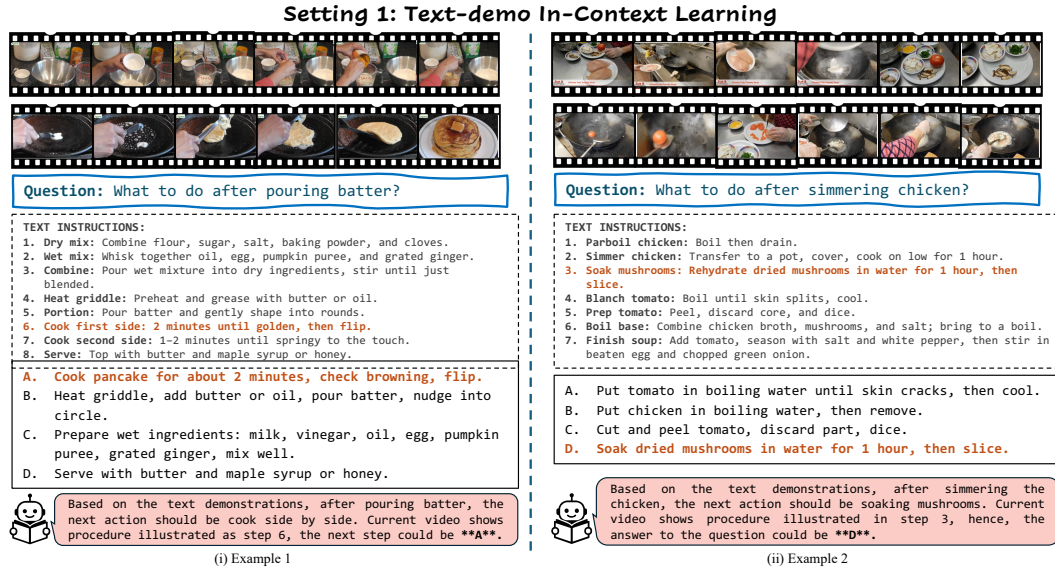


Figure 3: Visualization of Text-demo In-Context Learning. This figure provides 2 examples to illustrate the text-demo in-context learning task, where the text instructions will be provided along with the target video as the inputs.

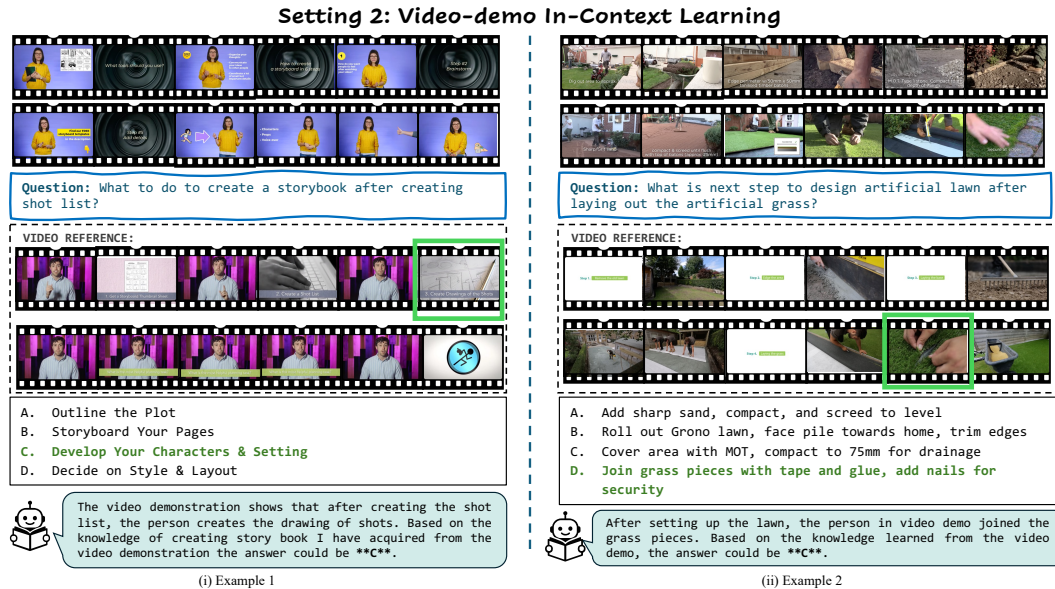


Figure 4: Visualization of Video-demo In-Context Learning. This figure provides 2 examples to illustrate the video-demo in-context learning task, where a video demonstration will be provided together with the target video input.

Table 5: 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
Commercial MLLMs									
GPT-4V (OpenAI, 2023a)	-	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

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 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 5 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.

E 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, as mentioned in 2, 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 (Koupaee & 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 & 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 pseudo 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.

Table 6: **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 & 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
VideoMMU (Fu et al., 2025)	Lectures	300	900	✗	Manual
Demo-ICL-Bench	Instructional Videos	1200	1200	✓	Mixed

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. Video-CoT (Wang et al., 2024b) 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.

F LIMITATIONS AND FUTURE DIRECTIONS

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.