This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding

Kangsan Kim[1*] Geon Park[1*] Youngwan Lee[1,3] Woongyeong Yeo[1] Sung Ju Hwang[1,2]
[1]KAIST [2]DeepAuto.ai [3]ETRI
{kksan07, geon.park, ywlee88, wgcyeo, sungju.hwang}@kaist.ac.kr

Abstract

Recent advancements in video large multimodal models (LMMs) have significantly improved their video understanding and reasoning capabilities. However, their performance drops on out-of-distribution (OOD) tasks that are underrepresented in training data. Traditional methods like fine-tuning on OOD datasets are impractical due to high computational costs. While In-context learning (ICL) with demonstration examples has shown promising generalization performance in language tasks and image-language tasks without fine-tuning, applying ICL to video-language tasks faces challenges due to the limited context length in Video LMMs, as videos require longer token lengths. To address these issues, we propose VideoICL, a novel video in-context learning framework for OOD tasks that introduces a similarity-based relevant example selection strategy and a confidence-based iterative inference approach. This allows to select the most relevant examples and rank them based on similarity, to be used for inference. If the generated response has low confidence, our framework selects new examples and performs inference again, iteratively refining the results until a high-confidence response is obtained. This approach improves OOD video understanding performance by extending effective context length without incurring high costs. The experimental results on multiple benchmarks demonstrate significant performance gains, especially in domain-specific scenarios, laying the groundwork for broader video comprehension applications. Code is released at https://github.com/KangsanKim07/VideoICL

1. Introduction

Recent video large multimodal models (LMMs) [11, 34, 47, 57] have shown notable improvement in video understanding and reasoning tasks, *e.g.*, enabling these models to comprehend natural scene videos and answer causal questions. However, as new types of data continue to emerge, these models are expected to handle previously unseen, out-of-distribution (OOD) videos [5, 17, 28, 37, 38, 44]. Such OOD videos are rarely encountered during training due to their specialized nature or the need for domain-specific knowledge, such as gymnastic or surgical videos. Consequently, the performance of video LMMs on OOD videos is poor compared to in-distribution videos [23, 36], primarily due to the limited representation of OOD content in training datasets. For instance, video LMMs can readily distinguish between well-represented actions like "dancing" and "exercising" but struggle to differentiate actions like "abuse" from "assault", as crime videos are seldom included in training datasets. Given that the model faces numerous unseen situations in real-world scenarios, improving OOD video understanding with video LMMs remains a significant challenge.

Fine-tuning the model on the target videos is a straightforward method to enhance its performance. However, fine-tuning the model for each OOD scenario is time-consuming and often impractical, as it requires substantial amounts of training data to avoid overfitting and incurs significant training costs. In contrast, in the realm of large language models (LLMs), in-context learning (ICL) [7]—which involves providing example inputs alongside a test sample during inference—has been actively explored as an efficient alternative for handling OOD tasks. ICL has shown strong generalization on unseen tasks in LLMs, making it advantageous as it bypasses the need for fine-tuning. Many studies have demonstrated the effectiveness of ICL in language-only and image-language tasks [2, 16, 27, 55, 56]; however, ICL's potential in video-language tasks has not been fully explored.

A key challenge with ICL in the video domain is that video tokens are significantly longer than image or text tokens, limiting the number of video examples in a single context. For instance, LLaVA-Video [57] can process up to 32K tokens (*i.e.*, context-length), but a 30-second, 384×384 resolution video sampled in 32 frames is converted to approximately 5.5K tokens through the vision encoder of the model with a patch size of 14. This allows for a maximum of only four video samples within the token limit. Considering that

*Equal contribution.

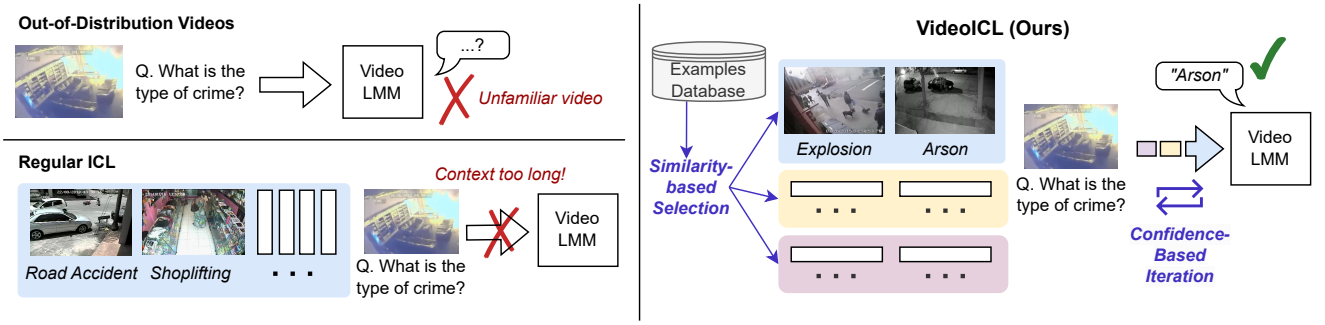

Figure 1. **Motivation. Top left:** Video LMMs perform poorly in out-of-distribution videos, such as crime videos. **Bottom left:** In-Context Learning (ICL), which is usually employed to solve this problem, is infeasible for video tasks, since the in-context demonstrations are too long. **Right:** VIDEOICL alleviates this problem by selecting the most relevant demonstrations (*e.g.*, 2-shot) by *similarity-based example selection*, and iteratively performing inference with different sets of demonstrations at each step (*confidence-based iterative inference*).

recent text-only and image-text ICL research is advancing toward many-shot learning [1, 6, 21], often utilizing over 1K examples, the number of video examples here is notably constrained in comparison.

This challenge has been approached from two main directions in multimodal ICL research. One approach aims to reduce the token length of each example [14, 59], while the other focuses on selecting a minimal set of highly effective examples [8, 48]. However, the first approach potentially may lead to a loss of crucial information, while the second relies on a limited number of in-context examples, making the model performance highly sensitive to the relevance of these examples to the given query prompt.

To address these issues, we propose VIDEOICL, a training-free video in-context learning framework for OOD video understanding, which efficiently handles multiple examples within a limited context length while preserving example quality. Specifically, we first introduce a similarity-based relevant example selection strategy, ranking demonstrating examples based on both video and text feature similarity to construct an ordered set of relevant examples for inference. Then, we present a confidence-based iterative inference mechanism; VIDEOICL performs in-context learning iteratively, with each iteration leveraging a new subset of highly relevant examples from the top of the similarity ranking. At each iteration, the model calculates a confidence score based on token probabilities and stops iterating once it reaches a sufficient confidence level to generate an accurate answer. This iterative approach enables the model to effectively utilize a larger pool of examples, maintaining informational richness without exceeding context limits.

Extensive experiments on several OOD video understanding benchmarks demonstrate the superiority of our method. Specifically, we validate our approach across six datasets, spanning four distinct video-language tasks: multiple-choice question answering, open-ended question answering, video classification, and video captioning. The results show that our framework, tested with LLaVA-Video-7B [57], outperforms zero-shot baselines, achieving an average improvement of 25.6%p and up to 54.6%p in QA and classification tasks, alongside a gain of 0.143 BLEU-4 points in video captioning. Notably, our approach using a 7B model outperforms a 72B model in zero-shot settings and on some datasets, VIDEOICL achieves better results than LoRA fine-tuned counterparts. These results underscore the effectiveness of our video ICL approach, showing that even smaller models can outperform larger and fine-tuned models on OOD tasks with a robust ICL framework.

Furthermore, VIDEOICL outperforms the state-of-the-art video ICL model, Otter [27], which is pre-trained in an in-context manner, by a substantial margin. These results suggest that the training-free nature of VIDEOICL facilitates greater scalability and generalization to diverse OOD videos compared to Otter. We believe that our findings will inspire further research into video in-context learning and drive advancements in enhancing the generalization capabilities of video LMMs for OOD videos.

Our main contributions are summarized below:

- We introduce VIDEOICL, a novel training-free framework for video in-context learning that enhances out-of-distribution (OOD) video understanding without requiring model fine-tuning or compromising input quality.
- We propose a confidence-based iterative in-context learning approach that effectively leverages multiple examples, addressing token length limitations of video LMMs.
- VIDEOICL achieves state-of-the-art results on six diverse OOD video-language datasets, with an average improvement of 25.6%p and up to 54.6%p in QA and classification tasks, along with a gain of 0.143 BLEU-4 points in video captioning, significantly outperforming zero-shot and baseline methods.

2. Related works

2.1. Video large multimodal models

The impressive capabilities of LMMs have motivated substantial research into video LMMs [11, 34, 47, 57], demonstrating significant performance on video understanding benchmarks. However, the benchmarks commonly used for

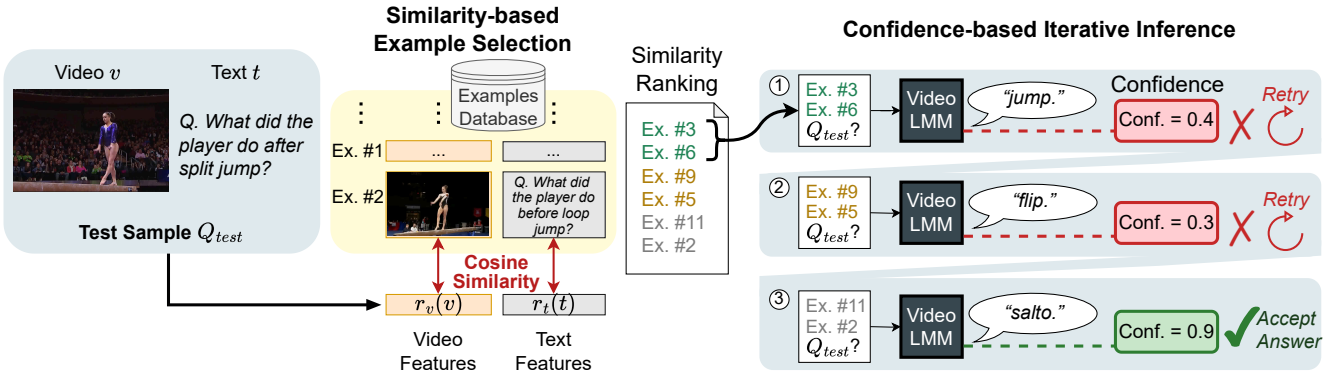

Figure 2. **Our Methodology.** Given a test query Q_{test} consisting of a video and some text, each are embedded into a vector. **Similarity-based Example Selection:** Based on the cosine similarity between the query vector and the embeddings in the database of pre-encoded examples, we retrieve top-k most similar examples. This stage takes negligible time cost since it only generates features from test samples and calculates the similarities with pre-encoded features. **Confidence-Based Iterative Inference:** Starting from the top of the list, each set of m examples are used as in-context examples for the query Q_{test}, until the confidence for the generated answer exceeds the threshold.

evaluating video LMMs, such as NExT-QA [50], Perception Test [40], and MVBench [30], primarily evaluate models on general tasks that require only basic skills for typical videos. Therefore, the strong performance of video LMMs is largely limited to in-domain videos, which resemble the videos they were trained on.

On the other hand, studies have shown that video LMMs perform poorly when tested on OOD videos. Khattak et al. [23] test different video LMMs using unusual and physically anomalous activities. Additionally, Marcu et al. [36] reveal that video LMMs struggle with autonomous driving videos. Given that the model encounters numerous unseen situations in real-world, these works underscore the need for further research to enhance OOD video understanding within video LMMs.

2.2. Multimodal in-context learning

Beyond the success of ICL on text-only tasks, it has been extended to multimodal tasks involving images [4, 8, 9, 14, 27, 43, 52, 58] and videos [2, 27, 54]. In video ICL, most studies aim to equip LMMs with the ability to comprehend multiple video examples by training them on video-text interleaved datasets. Otter [27] trains image-text ICL model using ICL instruction tuning dataset with video-language demonstrations. Yu et al. [54] propose a strategy to generate a video-text ICL training dataset that induces small-scale language models. However, these studies do not address OOD videos, and fine-tuning is impractical for handling OOD tasks, as it is both costly and time-consuming to train the model for each task.

A significant challenge in multimodal ICL is the increased length of each demonstration. One approach to address this is to compress the tokens of context examples to fit within the model's limit, using fused virtual tokens [14] or learnable parameters [9]. However, such compression methods risk losing crucial details. Another approach is selecting the most effective demonstrations [31, 48, 52]. MMICES [8] considers the varying importance of text and image information. However, since these methods still use a limited number of examples during inference, their performance can be less robust and highly sensitive to selection.

2.3. Iterative in-context learning

To enhance the robustness of ICL, some studies have employed iterative methods in language-only tasks. IDS [41] utilizes iterative refinement by selecting multiple sets of demonstrations and performing majority voting among the generated answers to identify the best response. Se^2 [33] introduces an example selection approach that considers the sequence of examples, iteratively refining the selection to find the optimal set based on previously chosen examples. Iterative retrieval [10] enhances ICL by using a stateful, policy-learning framework for iterative example selection, improving performance in semantic parsing tasks. Inspired by these works, we adopt an iterative method to address the challenge of limited example numbers in video ICL.

2.4. Factual confidence estimation

Estimating the factual confidence of LLMs has become an important research focus, as higher confidence levels are often associated with more accurate outputs and fewer hallucinations [12, 15, 19]. Xiong et al. [51] and Lin et al. [32] explore *verbalization*, where models express confidence in natural language. Kumar et al. [24] employ logit-based methods to estimate token-level confidence by assessing token probabilities as indicators, while Liu et al. [33] apply these techniques for selecting ICL examples. Sequence-level confidence is often derived by aggregating token-level confidence, typically using the minimum probability across all tokens [20]. Azaria and Mitchell [3] and Li et al. [29] utilize a *trained probe* approach, using an MLP trained on supervised data to map hidden states to confidence values.

3. Method

In this section, we describe our method to overcome the challenges of video ICL. First, we select a certain number (k) of in-context examples for a given query, constructing a list of demonstrating examples based on similarity (Sec. 3.2). Next, we sequentially fetch a small number ($m \leq k$) of examples from the list, iteratively refining the answer based on model confidence (Sec. 3.3). The complete workflow of VIDEOICL is shown in Fig. 2.

3.1. Notation

Throughout the paper, $\hat{y} = \mathsf{M}(x; \mathcal{D})$ denotes the answer a video LMM M generates for a query x using a set of in-context examples $\mathcal{D}$. $\mathrm{Conf}_{\mathsf{M}}(x, \hat{y}; \mathcal{D})$ denotes the video LMM M's estimated confidence in a generated answer $\hat{y}$, given a question x and a set of in-context examples $\mathcal{D}$.

3.2. Similarity-based example selection

At test time, we are given a query $x = (t, v)$, where t and v are the text and video in the query prompt, respectively. Based on the relevance to the query, k in-context examples are selected from the set of target task-specific example data. k is a fixed hyperparameter. For this, we use a linear combination of the cosine similarities on the vector representation of the given query and example data:

$$\begin{aligned} &S_Q\left((t, v), (\tilde{t}, \tilde{v})\right) \\ &:= \alpha S_C(r_t(t), r_t(\tilde{t})) + (1-\alpha) S_C(r_v(v), r_v(\tilde{v})), \end{aligned} \tag{1}$$

where S_C denotes the cosine similarity between two vectors. $r_t(\cdot)$ and $r_v(\cdot)$ denote text and video encoders, respectively, which map arbitrary-length text and video inputs into fixed-length vectors. α is a balancing coefficient between the text and the video similarities. From the set of example data $\mathsf{D} = \{(\tilde{t}_1, \tilde{v}_1), \ldots, (\tilde{t}_n, \tilde{v}_n)\}$, we select the top-$k$ examples that maximize S_Q:

$$\mathrm{SelectRelevant}_k(\mathsf{D}, x) := \underset{\tilde{x} \in \mathsf{D}}{\mathrm{Top}-k} \left[S_Q(x, \tilde{x})\right]. \tag{2}$$

Note that this example selection has negligible cost overhead at inference time, since all of the text and video vector representations in the example set can be pre-processed and be stored in a vector database.

3.3. Confidence-based iterative inference

Since all k in-context examples do not fit within the context size of open-source video LMMs, we instead provide the model m examples at each iteration. $m \leq k$ is a fixed hyperparameter, chosen so that m examples, along with the query, fit within the model's context length limit. Specifically, from the list of k relevant examples $\mathcal{D}$, we first generate an answer for the given query x using the first m examples: $\hat{y}_1 = \mathsf{M}(x; \mathcal{D}_{1:m})$. If $\mathrm{Conf}_{\mathsf{M}}(x, \hat{y}_1; \mathcal{D}_{1:m}) > c_{th}$, *i.e.*,

Algorithm 1 VIDEOICL Inference Process

input Query x, video LMM M, example pool D.
output Final answer $\hat{y}$.

$\mathcal{D} = \mathrm{SelectRelevant}_k(\mathsf{D}, x)$.
$i \leftarrow 1$.
repeat
  $\mathcal{D}_{cur} \leftarrow \mathcal{D}_{(i-1)m+1:im}$. ▷ *Select next batch of samples*
  $\hat{y}_i = \mathsf{M}(x; \mathcal{D}_{cur})$. ▷ *Generate answer*
  $c_i = \mathrm{Conf}_{\mathsf{M}}(x, \hat{y}_1; \mathcal{D}_{cur})$. ▷ *Evaluate confidence*
  $i \leftarrow i + 1$.
until $c_i > c_{th}$ or $im > k$.
$\hat{y} = \hat{y}_{\arg\max_i c_i}$. ▷ *Select the largest confidence answer*

the model's confidence for the generated answer exceeds a predefined confidence threshold c_{th}, then we output $\hat{y}_1$ as the final answer. Otherwise, we retry using the next m examples $\mathcal{D}_{m+1:2m}$. We repeat this process until either the confidence for the last answer exceeds c_{th} or all of $\mathcal{D}$ is exhausted. After multiple iterations are completed, we finally output the answer which had the highest confidence.

For the confidence estimation, inspired by LLM literature [20, 35, 53] that often utilizes token probability, we also quantify the confidence score by taking the minimum token probabilities from the generated response. Through empirical observations, we found that using the minimum token probability is a better measure of confidence. Given a generated sequence of logits, we first apply softmax normalization to convert logits to probabilities, obtaining a sequence of $\hat{y} = \{p_1, ..., p_T\}$. We then compute the confidence score as $c = \min_{i=1}^{T} p_i$, where p_i represents the probability of the i-th token in the generated sequence. More confidence measures are explored in Sec. 5.3.

3.4. Theoretical analysis

To substantiate our approach, we further present a theoretical analysis illustrating how our confidence-based iteration method can enhance model accuracy. Let us assume that the probability of a video LMM outputting the correct answer for a given task, given a query and a set of in-context examples, is constant: $\Pr(\mathsf{M}(x; \mathcal{D}_{\text{cur}}) = y) = p_c$, and independent across iterations. For simplicity, we further assume that confidence estimation operates independently at each iteration and that the framework outputs the final iteration's answer as the final result, rather than selecting the answer with the highest confidence score.

Proposition 1 (Asymptotic model accuracy)**.** *Let $a(n)$ be the expected accuracy of* VIDEOICL *with a maximum of n confidence-based iterations. Then,*

$$\lim_{n \to \infty} a(n) = \frac{1}{1 + \frac{\mathrm{FPR}}{\mathrm{TPR}} \cdot \frac{1-p_c}{p_c}}, \tag{3}$$

where TPR *and* FPR *stand for the true positive rate (i.e.,*

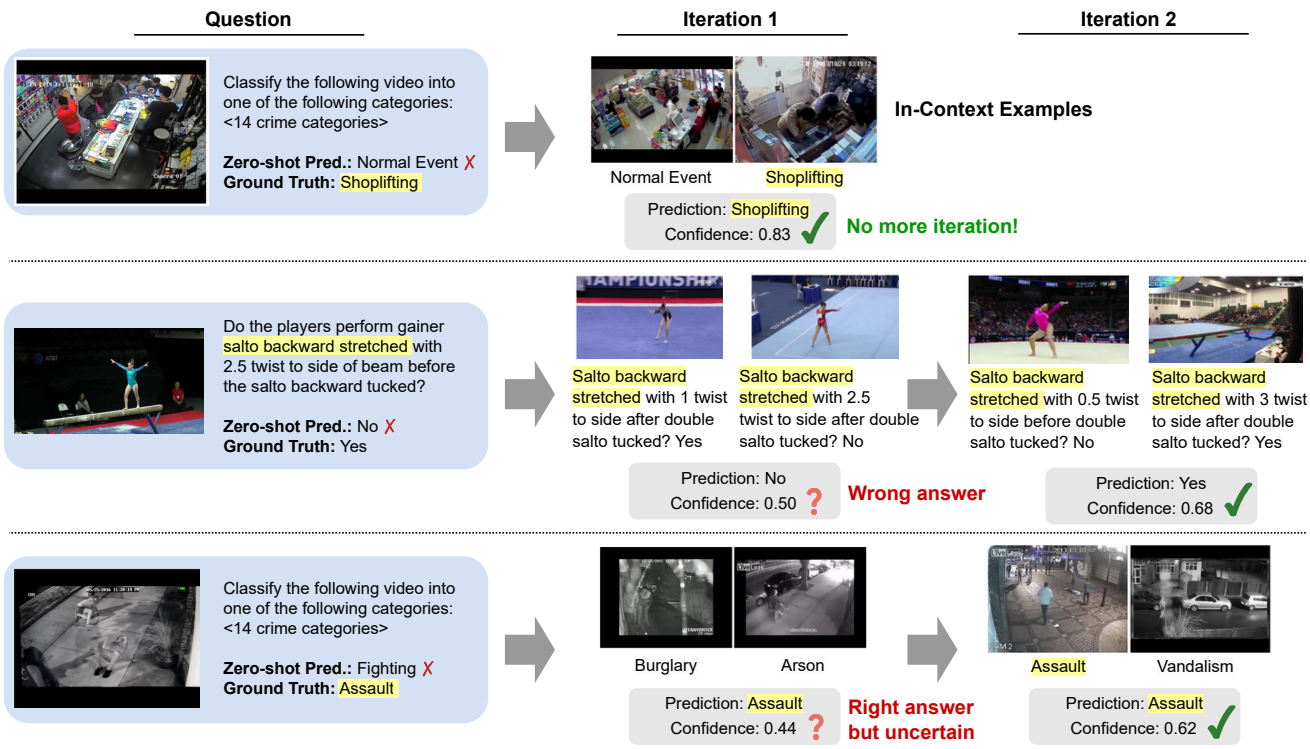

Figure 3. **Qualitative Results.** We show three real hand-picked test samples from the main benchmarks. The first and third examples are from the UCF-Crime [44] (video classification) task, and the second one is from the Sports-QA [28] (open-ended QA) task. The leftmost column shows the given question, which the vanilla (non-ICL) model makes an incorrect prediction. The second and third columns show the first and second confidence-based iterations, with the selected in-context demonstrations at each iteration.

recall) and the false positive rate of the confidence estimation method, respectively.

For example, if $p_c = 0.5$, TPR $= 0.9$, and FPR $= 0.1$, then the model's accuracy is expected to converge to 0.9 as the maximum number of iterations becomes large. In conclusion, if the confidence estimator is accurate enough, then our confidence-based iteration method is theoretically guaranteed to perform better than performing only one iteration. Please refer to Appendix C for the proof and rigorous statement of the proposition.

4. Experiments

In this section, we validate our framework in four video-language tasks on a wide range of *out-of-distribution* benchmarks using six datasets: multiple-choice QA on Animal Kingdom [38], open-ended QA on Sports-QA [28] and PitVQA [17], video classification on UCF-Crime [44] and Drive&Act [37], and video captioning task on CapERA [5].

4.1. Experiment setup

We apply our framework to state-of-the-art open-source video LMMs, such as LLaVA-Video-7B [57], Qwen2-VL-7B [47], and Oryx-1.5-7B [34], selected for their strong performance in general video understanding benchmarks. For the similarity-based example selection, we use SentenceBERT [42] and InternVideo2 [49] as text and video encoders, respectively. We sample each video at 1 frame per second across benchmarks, and for videos over 32 seconds, we uniformly select 32 frames from the entire sequence. For the confidence estimation, we use minimum value among the probabilities of generated tokens, and the confidence threshold is set to $c_{th} = 0.7$ in multiple-choice QA and $c_{th} = 0.5$ in other tasks. Our reasoning for this choice and additional experiments with various c_{th} settings are provided in Appendix A. The total number of demonstrating examples is set to $k = 8$. Note that considering the limited context length of the video LMMs, we set the number of in-context examples to $m = 2$ at each iteration across all benchmarks. Therefore, the maximum number of iterations per test sample is $n := k/m = 4$.

4.2. Baselines

We compare VIDEOICL with several strong baselines to rigorously assess our method's performance: GPT-4o [46], Gemini-1.5 Pro [45], Otter-7B [27], MMICES [8], Zero-shot, SIMRANKONCE, RANDEXVOTE, and SIMRANKVOTE. Otter [27] is a LMM fine-tuned on MIMIC-IT [26], which is a training set for image and video ICL. MMICES [8] is an example selection method originally designed for image ICL, which we have extended to our video multimodal setting. In the zero-shot baseline, responses are generated without in-context examples.

To further evaluate the impact of each component

			n	k	Multiple Choice QA	Open-ended QA		Video Classification		Video Captioning					
										CapERA					
					Animal Kingdom	Sports-QA	Pit-VQA	UCF-Crime	Drive&Act	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
GPT-4o [46]			-	0	58.2	-	6.9	58.0	-	0.143	0.065	0.037	0.023	0.142	0.173
Gemini-1.5 Pro [45]			-	0	72.9	-	14.7	55.1	-	0.126	0.057	0.031	0.019	0.134	0.176
Otter-7B [27]			1	8	19.4	-	21.8	6.8	-	0.241	0.135	0.088	0.059	0.169	0.167
LLaVA-Video [57]	72B	Zero-shot	-	0	69.7	25.7	5.7	35.6	14.6	0.133	0.060	0.034	0.020	0.129	0.170
	7B	LoRA FT	-	0	70.2	-	40.5	51.9	-	0.528	0.393	0.302	0.227	0.271	0.181
		Zero-shot	-	0	68.0	25.5	6.7	39.3	20.2	0.162	0.077	0.045	0.027	0.149	**0.181**
		MMICES [8]	1	2	69.3	43.0	46.4	50.7	51.3	0.462	0.312	0.224	0.160	0.245	0.178
		SimRankOnce	1	2	69.3	41.8	54.0	50.7	52.0	0.462	0.312	0.224	0.160	0.245	0.178
		RandExVote	4	8	69.6	21.5	11.5	36.6	19.9	0.418	0.256	0.170	0.116	0.189	0.153
		SimRankVote	4	8	70.9	36.3	57.6	50.6	50.6	0.464	0.314	0.228	0.165	0.242	0.175
		VideoICL (Ours)	4	8	**72.3**	**47.6**	**61.3**	**53.3**	**53.4**	**0.465**	**0.320**	**0.235**	**0.170**	**0.252**	0.178
		Δ			+4.3	+22.1	+54.6	+14.0	+33.2	+0.302	+0.242	+0.190	+0.143	+0.104	-0.003
Qwen2-VL [47]	7B	Zero-shot	-	0	58.6	26.8	5.8	36.1	10.6	0.286	0.159	0.101	0.066	0.149	0.138
		SimRankOnce	1	2	63.8	43.2	55.3	46.3	45.4	0.457	0.310	0.223	0.158	0.249	0.189
		RandExVote	4	8	62.3	21.0	14.0	36.6	14.3	0.397	0.240	0.157	0.104	0.188	0.170
		SimRankVote	4	8	64.0	50.9	59.4	46.7	45.8	0.448	0.300	0.213	0.151	0.239	0.187
		VideoICL (Ours)	4	8	**66.3**	**51.5**	**59.6**	**48.7**	**49.3**	**0.471**	**0.329**	**0.244**	**0.176**	**0.265**	**0.189**
		Δ			+7.7	+24.7	+53.8	+12.6	+38.7	+0.185	+0.170	+0.143	+0.110	+0.116	+0.051
Oryx-1.5	7B [34]	Zero-shot	-	0	58.6	28.3	3.8	11.9	10.7	0.242	0.126	0.077	0.049	0.140	0.151
		VideoICL (Ours)	4	8	58.5	**52.0**	**58.4**	**44.0**	**57.3**	**0.327**	**0.195**	**0.128**	**0.086**	**0.188**	**0.179**
		Δ			-0.1	+23.7	+54.6	+32.1	+46.6	+0.085	+0.069	+0.052	+0.038	+0.047	+0.028

Table 1. **Main Results.** We evaluate our framework VideoICL on a wide range of *out-of-distribution* benchmarks. $n = k/m$ denotes the (maximum) number of iterations, where k is the total number of in-context examples, and m is the number of examples used in each iteration. The difference (Δ) denotes the score improvement of VideoICL over the vanilla 7B models without ICL. The LoRA FT baseline, fine-tuned on each of the downstream tasks, is shown in gray because it is not directly comparable to our method due to its additional training cost. The highest scores for each benchmark are shown in bold.

within VideoICL, we use three additional baseline methods. SimRankOnce performs ICL a single time with only m relevant examples chosen by our similarity ranking (as detailed in Sec. 3.2), without any iterative inference. RandExVote incorporates iterative inference but omits confidence-based selection, instead employing majority voting across answers generated using randomly selected in-context examples. SimRankVote also performs majority voting (*i.e.*, without confidence) but uses k relevant examples chosen by our similarity ranking instead of random examples. We also compare against the zero-shot performance of a larger 72B model. For completeness, we include the performance of LLaVA-Video-7B fine-tuned with LoRA [18] for some of the benchmark tasks. However, it is not directly comparable to ICL, since fine-tuning incurs a large training cost for each OOD task.

4.3. Datasets

We test our method on six diverse datasets across four tasks, each focusing on specialized domains that are rarely covered in the training data of current video LMMs. For the multiple-choice QA task, we use the Animal Kingdom dataset [38], which includes animal videos annotated with actions across 140 different classes. For open-ended question-answering, we apply Sports-QA [28], covering various sports videos, and PitVQA [17], designed specifically for visual QA in endonasal pituitary surgery videos. For open-ended QA, we utilize UCF-Crime [44], which categorizes types of crime in security camera footage into 13 groups, and Drive&Act [37], which identifies 34 types of activities performed by drivers in Kinect-IR videos. For the video captioning task, we evaluate models on the CapERA dataset [5], which is tailored to describe scenes from an aerial view. For more details, please see Appendix B.

4.4. Quantitative results

Tab. 1 summarizes the performance comparison with the baselines. Overall, VideoICL outperforms the baselines across a wide variety of domains. See Appendix D.1 for a more detailed discussion on the baselines' performance.

Multiple choice QA. We observe that VideoICL achieves a +4.3%p improvement in accuracy for recognizing animal actions compared to zero-shot LLaVA-Video-7B. Notably, VideoICL even surpasses the larger LLaVA-Video-72B model, despite using the smaller LLaVA-Video-7B model. This result indicates that simply increasing model size is not an effective solution for OOD video understanding. Furthermore, video ICL baselines (*e.g.*, SimRankVote) that rely solely on similarity ranking without confidence-based iteration perform less effectively than VideoICL, underscoring the advantages of our approach.

Open-ended QA. We observe the highest accuracy im-

	Animal Kingdom	PitVQA	UCF-Crime
Baseline	68.0	6.7	39.3
Random	68.4 (+0.4)	8.3 (+1.6)	38.4 (-0.9)
Text only	-	33.1 (+24.8)	-
Video only	-	29.1 (+22.4)	-
Text + Video	**72.3** (+4.3)	**61.3** (+54.6)	**53.3** (+14.0)

Table 2. **Ablation on the similarity-based example selection method.** We investigate which feature types are more helpful in selecting relevant examples.

provement in open-ended QA tasks, with VIDEOICL achieving up to +54.6%p and +22.1%p improvements on PitVQA and Sports-QA, respectively, compared to zero-shot performance. This result highlights that ICL is particularly effective for out-of-distribution tasks where answers adhere to a specific format, even when the question itself is phrased simply. For example, given the question "*What is the athlete doing?*" with a video showing a gymnast performing a move called *salto backward tucked*, the model might respond with, "*The athlete is jumping from the ground in a crowded gym.*" While this answer is not incorrect, it lacks specific domain knowledge users might expect in the context of the gymnastics domain, such as "*The athlete is performing salto backward tucked.*" ICL enables the model to align with the expected answer style and apply relevant domain knowledge, yielding a response that is both more precise and contextually accurate.

Video Classification. We observe that VIDEOICL significantly outperforms baselines by up to +14.0%p more accuracy gain in UCF-Crime and +38.7%p in Drive&Act, demonstrating the effectiveness of our method in video classification tasks. Surprisingly, VIDEOICL exceeds the LoRA fine-tuned model on UCF-Crime. As Bertsch et al. [6] discovered, VIDEOICL tends to surpass fine-tuning in a small-size dataset. These outstanding results may be due to the fact that similar in-context examples also have a high chance of being from the same class, demonstrating the usefulness of our similarity-based example selection process.

Video Captioning. Across various metrics, our method consistently surpasses the baselines, demonstrating its effectiveness in video captioning tasks. The largest improvement was observed in BLEU-1, with the gains gradually decreasing from BLEU-2 to BLEU-4. This suggests that the examples in VIDEOICL provide useful words related to the query video, enhancing the relevance of generated captions.

Comparison to proprietary models. In Tab. 1, we compare the performance of VIDEOICL with two leading proprietary APIs: GPT-4o [46] and Gemini-1.5 Pro [45]. For this analysis, we use the results of VIDEOICL with the LLaVA-Video-7B model. Overall, VIDEOICL demonstrates an average performance improvement of +17.8%p over GPT-4o and +14.2%p over Gemini-1.5 Pro. These results highlight the effectiveness and robustness

	Animal Kingdom	Pit-VQA	UCF-Crime	CapERA	
				BLEU-4	METEOR
Baseline	68.0	6.7	39.3	0.027	0.149
$k = 2$	69.3	54.0	50.7	0.160	0.245
$k = 4$	71.0	59.5	52.7	0.168	0.251
$k = 8$	72.3	**61.3**	53.3	**0.170**	**0.253**
Δ	+4.3	+54.6	+14.0	+0.143	+0.104
$k = 16$	**73.2**	61.2	**53.6**	0.169	0.250
Δ	+5.2	+54.5	+14.3	+0.142	+0.101

Table 3. **Ablation on the total number of demonstrations (k).** The highest scores for each benchmark are shown in bold.

of VIDEOICL across diverse OOD videos, even when compared to significantly larger models. Importantly, VIDEOICL achieves this performance with a 7B-parameter backbone, while GPT-4o and Gemini-1.5 Pro are much larger models. Despite this inherent size disadvantage, VIDEOICL excels due to its efficient iterative in-context learning approach, underscoring its strong capabilities in challenging scenarios.

4.5. Qualitative results

We illustrate some representative samples from our benchmarks in Fig. 3. In the first row, we show an example from UCF-Crime [44]. Our model selects two relevant demonstrations—a *Normal Event* and a *Shoplifting* example in the first round. Using these two in-context examples, our model is able to answer the original question correctly. This demonstrates that our similarity-based example selection allows the model to select relevant and helpful examples.

In the second row, we show an example from Sports-QA [28] where the vanilla model makes an incorrect prediction. In the first iteration, the initial response of our model is incorrect too, with low confidence. By the second iteration, our model answers correctly with high confidence. This shows the effectiveness of iterative selection, especially when initial demonstrations are suboptimal.

In contrast, in the third row, even though the model predicts the right answer in the first iteration, it lacks confidence due to insufficiently relevant demonstrations, as there is no example from the same class as the ground truth. Only after observing relevant examples in the second iteration does it gain enough confidence in the correct answer, suggesting that its initial correct prediction was a fluke. We provide more qualitative results in Appendix D.2.

5. Analysis

In this section, we analyze the impact of each component and hyperparameter of our method on its performance. For our analysis, we select one representative dataset for each video-language task: Animal Kingdom [38], PitVQA [17], UCF-Crime [44], and CapERA [5].

	Animal Kingdom	Pit-VQA	UCF-Crime	CapERA	
				BLEU-4	METEOR
Baseline	68.0	6.7	39.3	0.027	0.149
Verbalization	69.7	54.6	51.8	0.160	0.245
Δ	+1.7	+47.9	+12.5	+0.133	+0.096
Trained Probe	71.7	42.5	52.7	0.162	0.250
Δ	+3.7	+35.8	+13.4	+0.135	+0.101
Token Prob.	**72.3**	**61.3**	**53.3**	**0.170**	**0.253**
Δ	+4.3	+54.6	+14.0	+0.143	+0.104

Table 4. **Ablation on the confidence estimation method.** 'Token Prob.' refers to Token Probability. The highest scores for each benchmark are shown in bold.

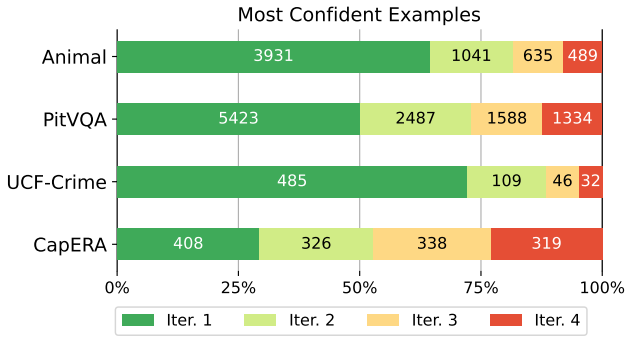

Figure 4. **Most confident examples.** The numbers on each bar indicate test samples where that iteration had the highest confidence score. The x-axis shows the proportion of each iteration.

5.1. Demonstration selection method

The impact of similarity-based example selection is studied by analyzing which query features (text or video) most influence the effectiveness of example selection. We compare our method with random demonstration selection, text feature-only selection, and video feature-only selection across three datasets. The results, shown in Tab. 2, indicate that our similarity-based selection method using both text and video features outperforms the random selection baseline by a large margin. For instance, on PitVQA, similarity-based selection leads to up to +53.0%p performance increase over random selection. Furthermore, we observe that both text and video features contribute to similarity-based selection, as using only one of the two yields lower scores.

5.2. Total number of available examples (k)

We investigate the impact of confidence-based iterative inference by comparing our method, which uses multiple iterations, with a single-iteration baseline across four datasets. The results, shown in Tab. 3, indicate that increasing the total number of demonstrations generally helps the benchmark performance. For example, compared to performing only one iteration (k = 2), we observe up to +7.2%p performance increase in PitVQA when using eight iterations (k = 16). This is consistent with the general observation that more ICL demonstrations lead to higher performance in Bertsch et al. [6]. This demonstrates that our confidence-based iterative inference can serve as an alternative to using many demonstrations at once when the context length is limited in a video ICL setting.

5.3. Confidence estimation method

We evaluate our confidence estimation method, token probability [20], against two alternatives: *verbalization* [32, 51] and *trained probe* [3, 22], across four benchmark datasets. For the trained probe, following Azaria and Mitchell [3], we pre-train a 4-layer MLP using the hidden states of the last token on a subset of diverse video-language datasets and benchmarks [13, 25, 30, 57]. The confidence score from the trained probe is used in the same way as our main method. For verbalization, we ask the model if it is confident enough to respond before generating the answer. If it answers yes, we proceed with the response; if not, we iterate with the next set of examples. The results are shown in Tab. 4.

The token probability method outperforms other approaches, with the trained probe following closely, while verbalization performs the worst. This somewhat contradicts the findings in [35], which report that the trained probe estimates model confidence most accurately. A possible explanation is that, unlike other zero-shot methods, the trained probe's reliance on pre-training makes it more dependent on its specific training data, limiting its ability to generalize, potentially reducing robustness on OOD tasks [39].

5.4. Most confident responses across iterations

We examine which iteration yields the highest confidence scores. Fig. 4 illustrates the confidence across four datasets and highlights the iteration, out of four, that produced the most confident answers. Notably, despite using the most similar examples in the first iteration, only 29% of responses in CapERA and 50% in PitVQA achieved enough confidence to surpass the threshold in the first round. This finding suggests that relying only on the first iteration may be inadequate. Multiple iterations allow for adding more examples, which may boost confidence in later rounds.

6. Conclusion

In this paper, we propose VIDEOICL, a novel framework for in-context learning for out-of-distribution video-language tasks with large multimodal models. We address the challenge of multimodal in-context examples exceeding the token length limit of LMMs by employing similarity-based demonstration selection and confidence-based iteration. Extensive experimental results highlight the effectiveness of our method for OOD videos. VIDEOICL is training-free, enabling rapid adaptation to novel tasks while offering a more efficient and feasible alternative to naive in-context learning at inference time.

Acknowledgments This work was supported by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT,Korea)&Gwangju Metropolitan City, Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)), National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00256259), a grant of the Korea Machine Learning Ledger Orchestration for Drug Discovery Project(K-MELLODDY), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (grant number: RS-2024-12345678), and (RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-training).

References

[1] Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*, 2024. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3

[3] Amos Azaria and Tom Mitchell. The Internal State of an LLM Knows When It's Lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore, 2023. Association for Computational Linguistics. 3, 8

[4] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1539–1550, 2024. 3

[5] Laila Bashmal, Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Mansour Zuair, and Farid Melgani. Capera: Captioning events in aerial videos. *Remote Sensing*, 15(8):2139, 2023. 1, 5, 6, 7, 12

[6] Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R Gormley, and Graham Neubig. In-context learning with long-context models: An in-depth exploration. *arXiv preprint arXiv:2405.00200*, 2024. 2, 7, 8

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020. 1

[8] Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. 2, 3, 5, 6

[9] Tao Chen, Enwei Zhang, Yuting Gao, Ke Li, Xing Sun, Yan Zhang, Hui Li, and Rongrong Ji. Mmict: Boosting multimodal fine-tuning with in-context examples. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 3

[10] Yunmo Chen, Tongfei Chen, Harsh Jhamtani, Patrick Xia, Richard Shin, Jason Eisner, and Benjamin Van Durme. Learning to retrieve iteratively for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7156–7168, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3

[11] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 2

[12] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024. 3

[13] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 8

[14] Jun Gao, Qian Qiao, Ziqiang Cao, Zili Wang, and Wenjie Li. Aim: Let any multi-modal large language models embrace efficient in-context learning. *arXiv preprint arXiv:2406.07588*, 2024. 2, 3

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 3

[16] Zhongyi Han, Guanglin Zhou, Rundong He, Jindong Wang, Tailin Wu, Yilong Yin, Salman Khan, Lina Yao, Tongliang Liu, and Kun Zhang. How well does GPT-4v(ision) adapt to distribution shifts? a preliminary investigation. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. 1

[17] Runlong He, Mengya Xu, Adrito Das, Danyal Z Khan, Sophia Bano, Hani J Marcus, Danail Stoyanov, Matthew J Clarkson, and Mobarakol Islam. Pitvqa: Image-grounded text embedding llm for visual question answering in pituitary surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 488–498, 2024. 1, 5, 6, 7, 12

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, 2021. arXiv:2106.09685 [cs]. 6

[19] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. 3

[20] Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023. 3, 4, 8

[21] Yixing Jiang, Jeremy Andrew Irvin, Ji Hun Wang, Muhammad Ahmed Chaudhry, Jonathan H Chen, and Andrew Y. Ng. Many-shot in-context learning in multimodal foundation models. In *ICML 2024 Workshop on In-Context Learning*, 2024. 2

[22] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. 8

[23] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. Complex video reasoning and robustness evaluation suite for video-lmms. *arXiv preprint arXiv:2405.03690*, 2024. 1, 3

[24] Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 315–334, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[25] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 8

[26] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 5

[27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A Multi-Modal Model with In-Context Instruction Tuning, 2023. arXiv:2305.03726. 1, 2, 3, 5, 6

[28] Haopeng Li, Andong Deng, Qiuhong Ke, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, and Chen Chen. Sports-qa: A large-scale video question answering benchmark for complex and professional sports. *arXiv preprint arXiv:2401.01505*, 2024. 1, 5, 6, 7, 12

[29] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[30] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206, 2024. 3, 8

[31] Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26710–26720, 2024. 3

[32] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*, 2022. 3, 8

[33] Haoyu Liu, Jianfeng Liu, Shaohan Huang, Yuefeng Zhan, Hao Sun, Weiwei Deng, Furu Wei, and Qi Zhang. se^2: Sequential example selection for in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5262–5284, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[34] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 1, 2, 5, 6

[35] Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluis Marquez. Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4554–4570, Bangkok, Thailand, 2024. Association for Computational Linguistics. 4, 8

[36] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269, 2024. 1, 3

[37] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019. 1, 5, 6, 12

[38] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19023–19034, 2022. 1, 5, 6, 7, 12

[39] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of

llm hallucinations. *arXiv preprint arXiv:2410.02707*, 2024. 8

[40] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens Continente, Larisa Markeeva, Dylan Sunil Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3

[41] Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. *arXiv preprint arXiv:2310.09881*, 2023. 3

[42] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. 5

[43] Igor Sterner, Weizhe Lin, Jinghong Chen, and Bill Byrne. Few-shot vqa with frozen llms: A tale of two approaches. *arXiv preprint arXiv:2403.11317*, 2024. 3

[44] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 5, 6, 7, 12

[45] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 5, 6, 7

[46] OpenAI Team. GPT-4o System Card, 2024. arXiv:2410.21276. 5, 6, 7

[47] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1, 2, 5, 6

[48] Siyin Wang, Chao-Han Huck Yang, Ji Wu, and Chao Zhang. Bayesian example selection improves in-context learning for speech, text, and visual modalities. *arXiv preprint arXiv:2404.14716*, 2024. 2, 3

[49] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 5

[50] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 3

[51] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 8

[52] Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. *arXiv preprint arXiv:2407.00902*, 2024. 3

[53] Yuchen Yang, Houqiang Li, Yanfeng Wang, and Yu Wang. Improving the reliability of large language models by leveraging uncertainty-aware in-context learning. *arXiv preprint arXiv:2310.04782*, 2023. 4

[54] Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, Shane Storks, and Joyce Chai. Eliciting in-context learning in vision-language models for videos through curated data distributional properties. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20416–20431, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3

[55] Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36:58478–58507, 2023. 1

[56] Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*, 2024. 1

[57] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 2, 5, 6, 8

[58] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[59] Yufan Zhuang, Chandan Singh, Liyuan Liu, Jingbo Shang, and Jianfeng Gao. Vector-icl: In-context learning with continuous vector representations. *arXiv preprint arXiv:2410.05629*, 2024. 2