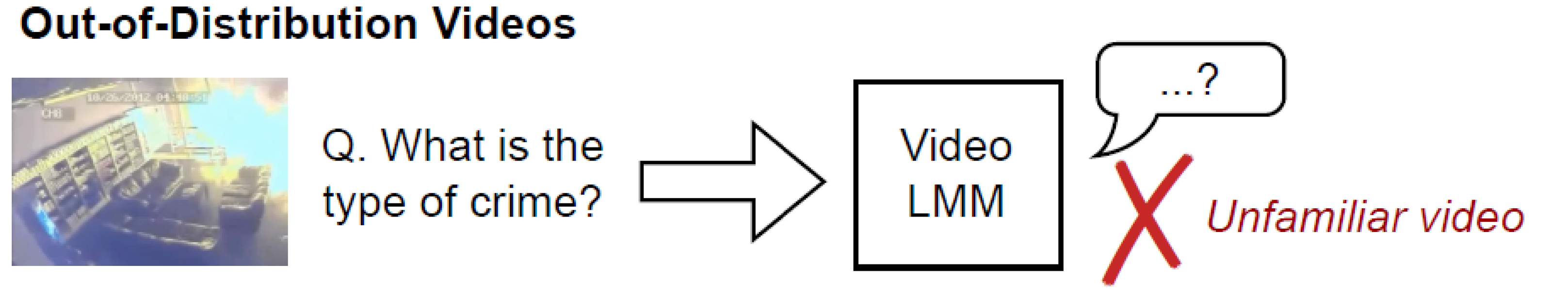
VideoICL: Confidence-based Iterative In-context Learning for Out-of-Distribution Video Understanding

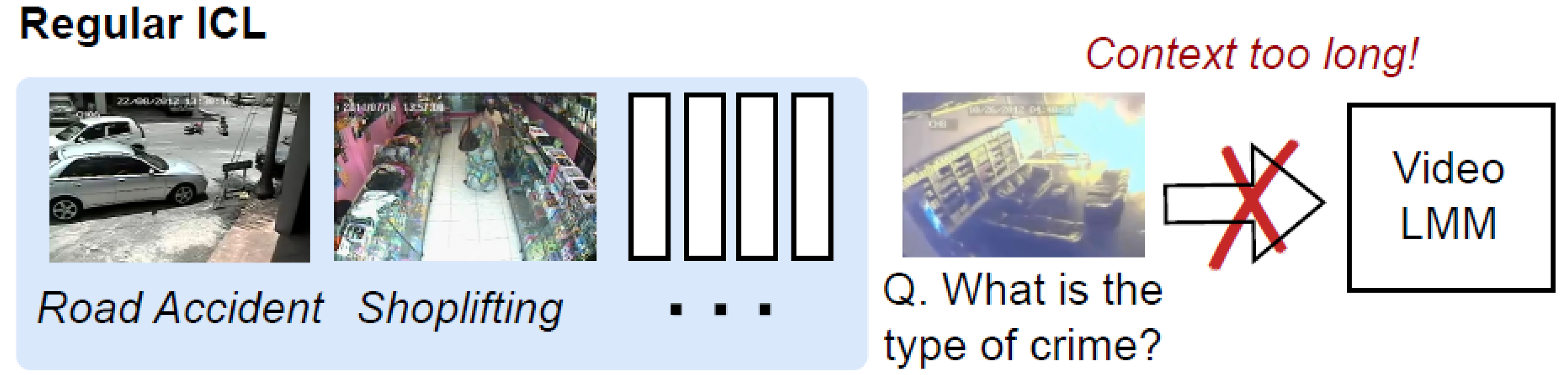
Kangsan Kim[1*], Geon Park[1*], Youngwan Lee[1,3], Woongyeong Yeo[1], Sung Ju Hwang[1,2]
[1]KAIST [2]DeepAuto.ai [3]ETRI (*Equal contribution.)

CVPR Nashville JUNE 11-15, 2025

KAIST DeepAuto.ai ETRI

Motivation





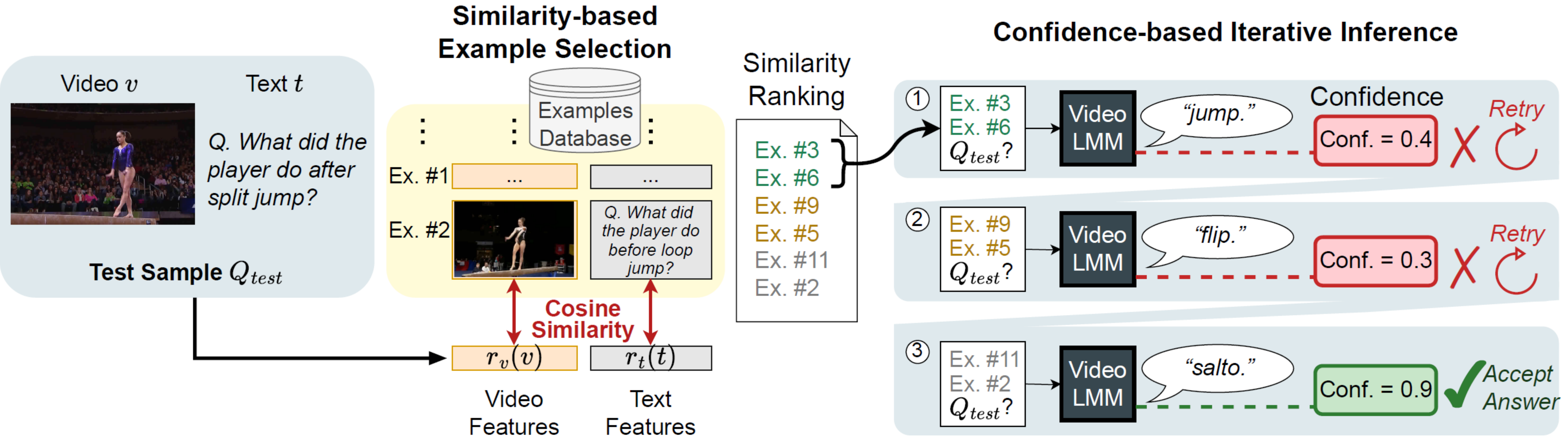
A key challenge with ICL in the video domain is that **video tokens are significantly longer** than image or text tokens, **limiting the number of video examples** in a single context.

Main Results

	n	k	Multiple Choice QA	Open-ended QA		Video Classification		Video Captioning		
								CapERA		
			Animal Kingdom	Sports-QA	Pit-VQA	UCF-Crime	Drive &Act	BLEU-4	METEOR	ROUGE-L
GPT-4o [46]	-	0	58.2	-	6.9	58.0	-	0.023	0.142	0.173
Gemini-1.5 Pro [45]	-	0	72.9	-	14.7	55.1	-	0.019	0.134	0.176
Otter-7B [27]	1	8	19.4	-	21.8	6.8	-	0.059	0.169	0.167
LLaVA-Video-7B	-	0	68.0	25.5	6.7	39.3	20.2	0.027	0.149	**0.181**
LoRA FT	-	0	70.2	-	40.5	51.9	-	0.227	0.271	0.181
MMICES [8]	1	2	69.3	43.0	46.4	50.7	51.3	0.160	0.245	0.178
SimRankOnce	1	2	69.3	41.8	54.0	50.7	52.0	0.160	0.245	0.178
RandExVote	4	8	69.6	21.5	11.5	36.6	19.9	0.116	0.189	0.153
SimRankVote	4	8	70.9	36.3	57.6	50.6	50.6	0.165	0.242	0.175
VideoICL (Ours)	4	8	**72.3**	**47.6**	**61.3**	**53.3**	**53.4**	**0.170**	**0.252**	0.178
Δ			+4.3	+22.1	+54.6	+14.0	+33.2	+0.143	+0.104	-0.003

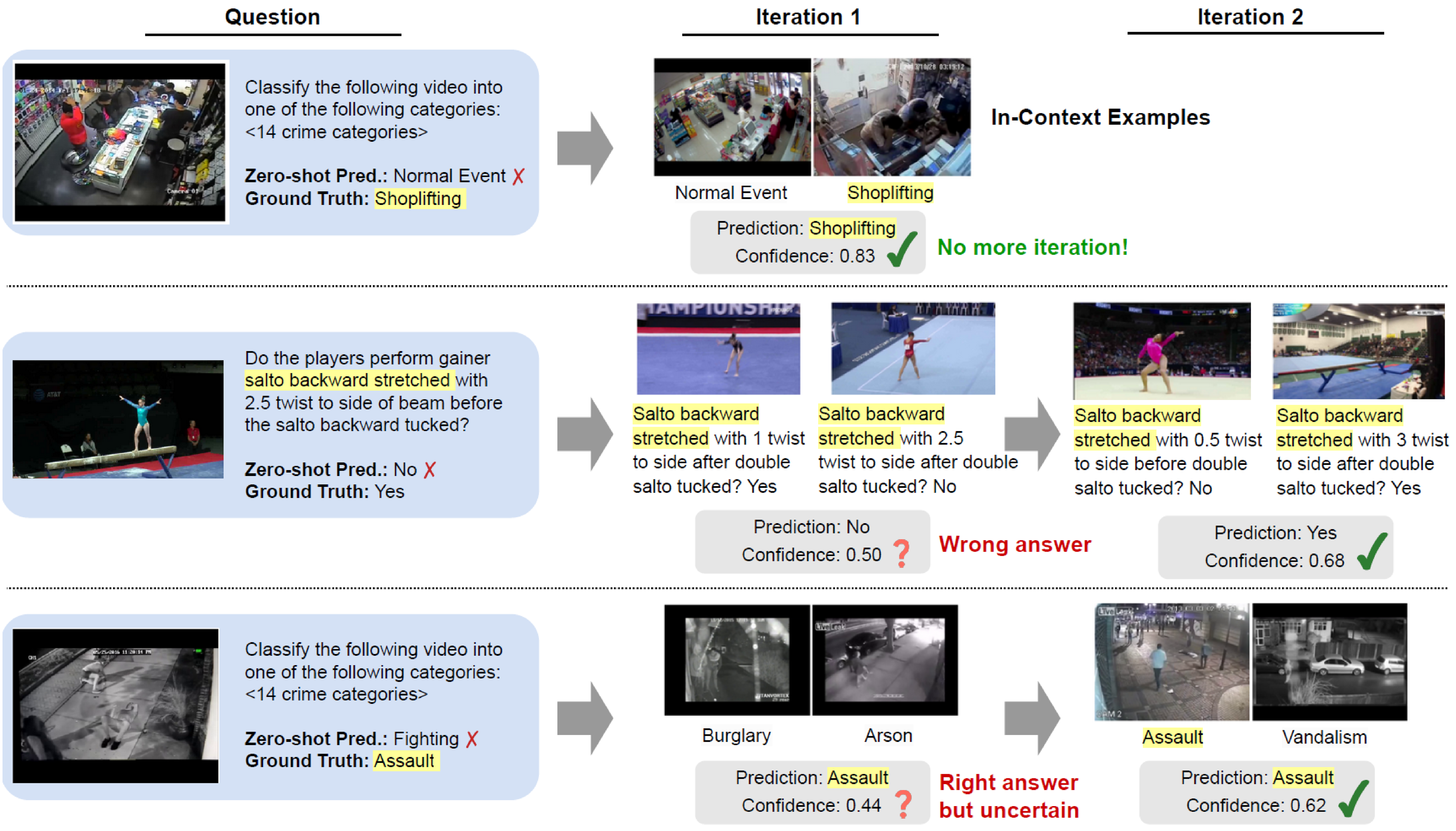
VideoICL achieves **state-of-the-art results on six diverse OOD video-language datasets**, with an average improvement of 25.6%p and up to 54.6%p in QA and classification tasks, along with a gain of 0.143 BLEU-4 points in video captioning, significantly outperforming zero-shot and baseline methods.

Method



We propose a **confidence-based iterative in-context learning approach** that effectively leverages multiple examples, addressing token length limitations of video LMMs.

Case Study



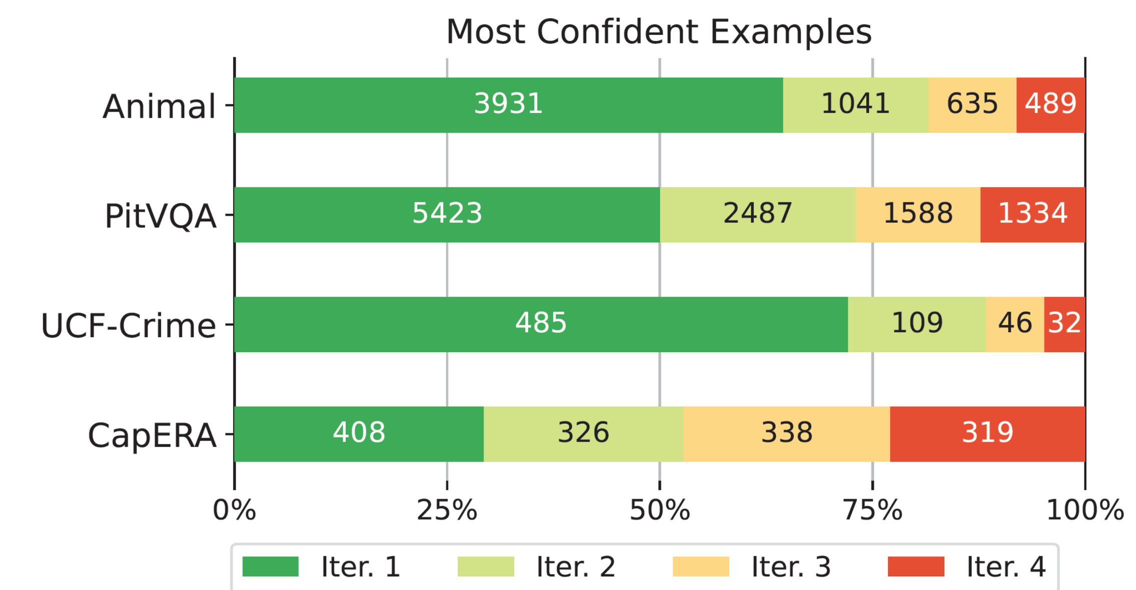
Analysis

	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

Using **more examples** lead to **better results**.

	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)

Both textual and visual features impact similarity-based selection.



Most confident examples emerge **after first round**.

	Animal Kingdom	Pit-VQA	UCF-Crime	CapERA	
				BLEU-4	METEOR
Verbalization	69.7	54.6	51.8	0.160	0.245
Trained Probe	71.7	42.5	52.7	0.162	0.250
Token Prob.	**72.3**	**61.3**	**53.3**	**0.170**	**0.253**

Token probability outperforms other confidence estimation methods.