

# LLAVA-NEXT-INTERLEAVE: TACKLING MULTI-IMAGE, VIDEO, AND 3D IN LARGE MULTIMODAL MODELS

**Anonymous authors**

Paper under double-blind review

## A DATA STATISTICS

The detailed data statistics of M4-Instruct is shown in Table 3.

The detailed data statistics of LLaVA-Interleave Bench is shown in Table 4.

## B ABLATION STUDY

### B.1 POOL VS NOT POOL VISION TOKENS FOR VIDEO TASKS.

Similar to LLaVA-NEXT-Video, we adopt a "Pooling to 1/4" strategy for which we pool the width and height of feature maps to 1/2 therefore reducing the number of tokens to 1/4. We study the impact of image token pooling. We train and infer our model under two settings: pooling to 1/4 and not pooling with ShareGPTVideo-Caption+QA(255K) data. Pooling to a 1/4 setting is similar to LLaVA-NEXT-Video, which uses the pooling technique to trade-off between the number of image tokens and the number of frames. In our experiment, we find that not pooling yields better performance under similar #image tokens. During training, we sample 10 frames for videos. In this table, we also observe that adding more frames (from 10 to 16) during inference improves performance.

### B.2 IMPACT OF VIDEO DPO TRAINING ON OTHER TASKS.

In Table 2, we compare the results of doing video DPO on other tasks. Though DPO significantly improves the video performance as shown in Table ??, it slightly impacts the performance of other tasks.

Table 1: Ablation to compare pooling and not pooling.

Training	Inference	#frames	# Image tokens	Act	Avg	VDD	VideoChatGPT				
							CI	DO	CU	TU	CO
Pooling 1/4	Pooling 1/4	40	40x729x1/4=10x729	52.8/3.53	3.35	3.38	3.46	3.25	3.87	2.59	3.57
Pooling 1/4	Pooling 1/4	64	64x729x1/4=16x729	52.7/3.53	3.33	3.38	3.45	3.23	3.86	2.49	3.55
Not Pooling	Not Pooling	10	10x729	52.9/3.48	3.38	3.46	3.43	3.26	3.85	2.64	3.61
Not Pooling	Not Pooling	16	16x729	<b>54.4/3.51</b>	<b>3.41</b>	<b>3.46</b>	<b>3.48</b>	<b>3.28</b>	<b>3.87</b>	<b>2.74</b>	<b>3.62</b>

Table 2: Ablation on the impact of video dpo on the performance of other tasks. QB: Q-Bench, Act: ActivityNet-QA, MVB: MVBench, VDD: Video Detailed Description, MME\*: Throughout our paper, we convert MME’s score to accuracy by summing up the perception and cognition scores and dividing 2800, SQA: Scienceqa-IMG.

Setting	Multi-image				Multi-view	Single-image					
	Mantis	BLINK	QB	NLVR2	ScanQA	AI2D	ChartQA	DocVQA	MME*	POPE	SQA
Before Video-DPO	62.7	52.7	73	89.1	32.2	73.9	67.2	75.7	63.5	85.4	72.6
After Video-DPO	60.8	51.7	86.8	87.7	25.5	72.2	56.1	73.1	62.6	86.6	71.7

Table 3: M4-Instruct detailed datasets.

Task	Dataset	Scenario	# Samples
<b>Multi-image Scenarios</b>			
Spot the Difference(42.6K)	Real-world Difference	Realistic	6.7K
	Synthetic Difference	Synthetic	7.0K
	Spot-the-Diff	Surveillance	10.8K
	Birds-to-Words	Birds	14.2K
	CLEVR-Change	Solids	3.9K
Image Edit Instruction(67.7K)	HQ-Edit	Synthetic	50K
	MagicBrush	Realistic	14.2K
	IEEdit	Realistic	3.5K
Visual Story Telling(67.5K)	AESOP	Cartoon	6.9K
	FlintstonesSV	Cartoon	22.3K
	PororoSV	Cartoon	12.3K
	VIST	Realistic	26K
Text-rich VQA(21.3K)	WebQA	Webpage	9.3K
	TQA	Textbook	8.2K
	OCR-VQA	OCR	1.9K
	DocVQA	Document	1.9K
Multi-image VQA(153.5K)	NLVR2	Realistic	86.4K
	MIT-States_StateCoherence	General	1.9K
	MIT-States_PropertyCoherence	General	1.9K
	RecipeQA_ImageCoherence	Recipe	8.7K
	VISION	Industrial	9.9K
	Multi-VQA	General	5K
	IconQA	General	34.6K
Low-level Comparison(65.9K)	Coinstruct	Low-level	50K
	Dreamsim	Low-level	15.9K
Image-caption Comprehension (41.8K)	ImageCoDe	General	16.6K
	Contrast-Caption	General	25.2K
Daily Scenarios (5.7K)	MMChat_Twitter_Post	General	5.7K
Multi-image Puzzle (35K)	Raven	Abstract	35K
<b>Multi-frame (Video) Scenarios</b>			
Video QA(246.9K)	NExT-QA	General	3.9K
	STAR	General	3K
	ShareGPTVideo-VQA	General	240K
Video Detailed Captioning (15K)	ShareGPTVideo-Caption	General	15K
<b>Multi-view (3D) Scenarios</b>			
Scene VQA(45.4K)	Nuscenes	Outdoor	9.8K
	ScanQA	Indoor Realistic	25.6k
	3D-LLM-Scene	Indoor Realistic	10K
Embodied VQA(62.5K)	ALFRED	Indoor Synthetic	22.6K
	3D-LLM-Dialogue	Indoor Realistic	20K
	3D-LLM-Planning	Indoor Realistic	19.9K
<b>Single-image Scenarios</b>			
Single-image Tasks(307K)	Randomly sampling 40% SFT data of LLaVA-NeXT	General	307K

Table 4: LLaVA-Interleave Bench detailed datasets.

Task	Dataset	Scenario	# Samples
<b>In-domain Evaluation - Newly Curated Benchmarks</b>			
Spot the Difference(0.3K)	Spot-the-Diff	Surveillance	0.1K
	Birds-to-Words	Birds	0.1K
	CLEVR-Change	Solids	0.1K
Image Edit Instruction(2K)	HQ-Edit	Sythetic	1K
	MagicBrush	Realistic	0.9K
	IEdit	Realistic	0.1K
Visual Story Telling(0.4K)	AESOP	Cartoon	0.1K
	FlintstonesSV	Cartoon	0.1K
	PororoSV	Cartoon	0.1K
	VIST	Realistic	0.1K
Text-rich VQA(0.4K)	WebQA	Webpage	0.1K
	TQA	Textbook	0.1K
	OCR-VQA	OCR	0.1K
	DocVQA	Document	0.1K
Multi-image VQA(0.4K)	MIT-States_StateCoherence	General	0.1K
	MIT-States_PropertyCoherence	General	0.1K
	RecipeQA_ImageCoherence	Recipe	0.1K
	VISION	Industrial	0.1K
Puzzle (1.4K)	Raven	Abstract	1.4K
<b>In-domain Evaluation - Existing Benchmarks</b>			
NLVR2 (7K)	NLVR2	Realistic	7K
Q-Bench (1K)	Q-Bench	Low-level	1K
<b>Out-domain Evaluation - Newly Curated Benchmarks</b>			
MathVerse-mv (0.8K)	MathVerse (Vision Dominant)	Math Diagram	0.8K
SciVerse-mv (0.4K)	SciVerse (Vision Dominant)	Scientific Diagram	0.4K
<b>Out-domain Evaluation - Existing Benchmarks</b>			
Mantis-Eval (0.2K)	Mantis-Eval	General	0.2K
BLINK (1.9K)	BLINK	General	1.9k
MMMU-mv (test) (0.8K)	MMMU	Scientific Diagram	0.8K

C MORE REAL-WORLD APPLICATIONS

Table 5: An example of emerging capability: multi-doc VQA.

Real-World Applications	
User	Hi, there are two candidates, can you provide a brief description for each of them for me?
Ours	<p>Certainly! Here's a brief description of each candidate:</p> <ol style="list-style-type: none"> <li><b>Oprah Winfrey:</b> <ul style="list-style-type: none"> <li><b>Professional Experience:</b> Known for her talk show, "The Oprah Winfrey Show," and her work on "COVID-19" as a presenter on Apple TV.</li> <li><b>Skills:</b> Emphasized on emotional intelligence, interpersonal skills, and creativity.</li> <li><b>Honors &amp; Awards:</b> Notable achievements include the Golden Globe Award for Best Television Series and the Golden Globe Award for Best Television Special.</li> </ul> </li> <li><b>Steve Jobs:</b> <ul style="list-style-type: none"> <li><b>Professional Experience:</b> Co-founder of Apple Inc., where he led the development of iconic products like the iMac, iPod, and iPhone.</li> <li><b>Skills:</b> Highlighted as a thinker with a focus on innovation and problem-solving.</li> <li><b>Honors &amp; Awards:</b> Recognized with numerous accolades, including the MacArthur Fellowship and the NeXT Computer Science Award.</li> </ul> </li> </ol> <p>Please note that these descriptions are based on the information provided in the images and do not include any personal details or specific accomplishments mentioned in the images.</p>
User	Which of them better suits the role of the Congressman of the US? Why?
Ours	<p>The first resume, featuring Oprah Winfrey, would better suit the role of a Congressman because it highlights skills and experiences that are relevant to political office, such as leadership, public speaking, and media experience. These qualities are essential for a politician to effectively represent their constituents and engage in legislative processes. The second resume, while it could be suitable for a creative or entrepreneurial role, does not specifically mention skills or experiences that align with the responsibilities of a Congressman.</p>