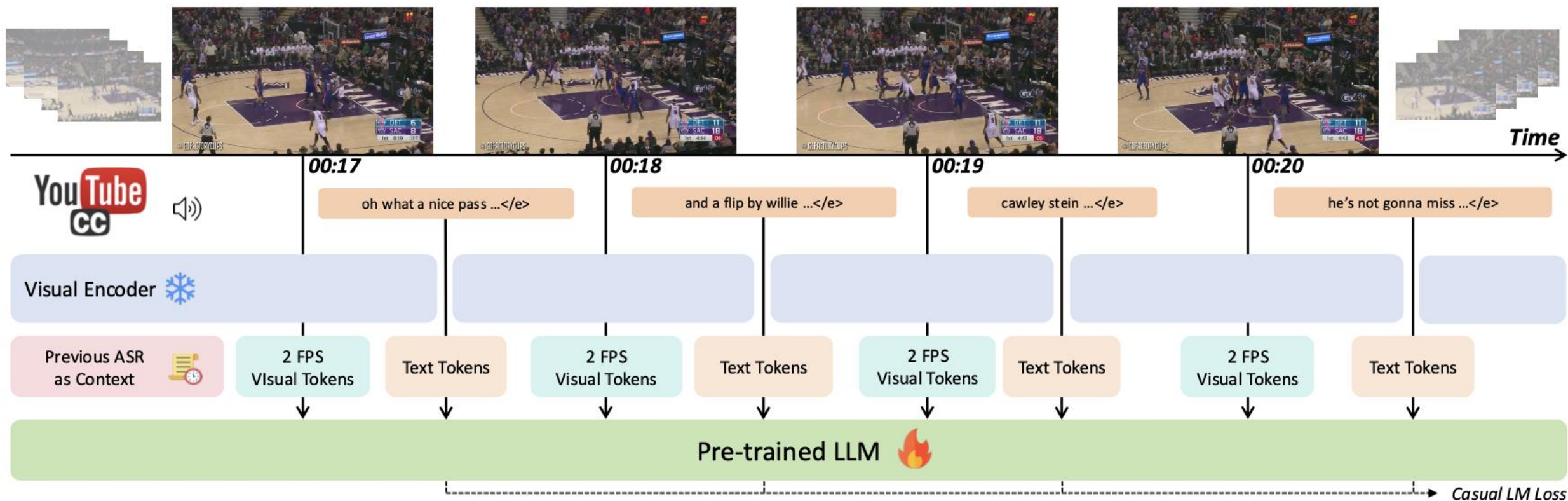


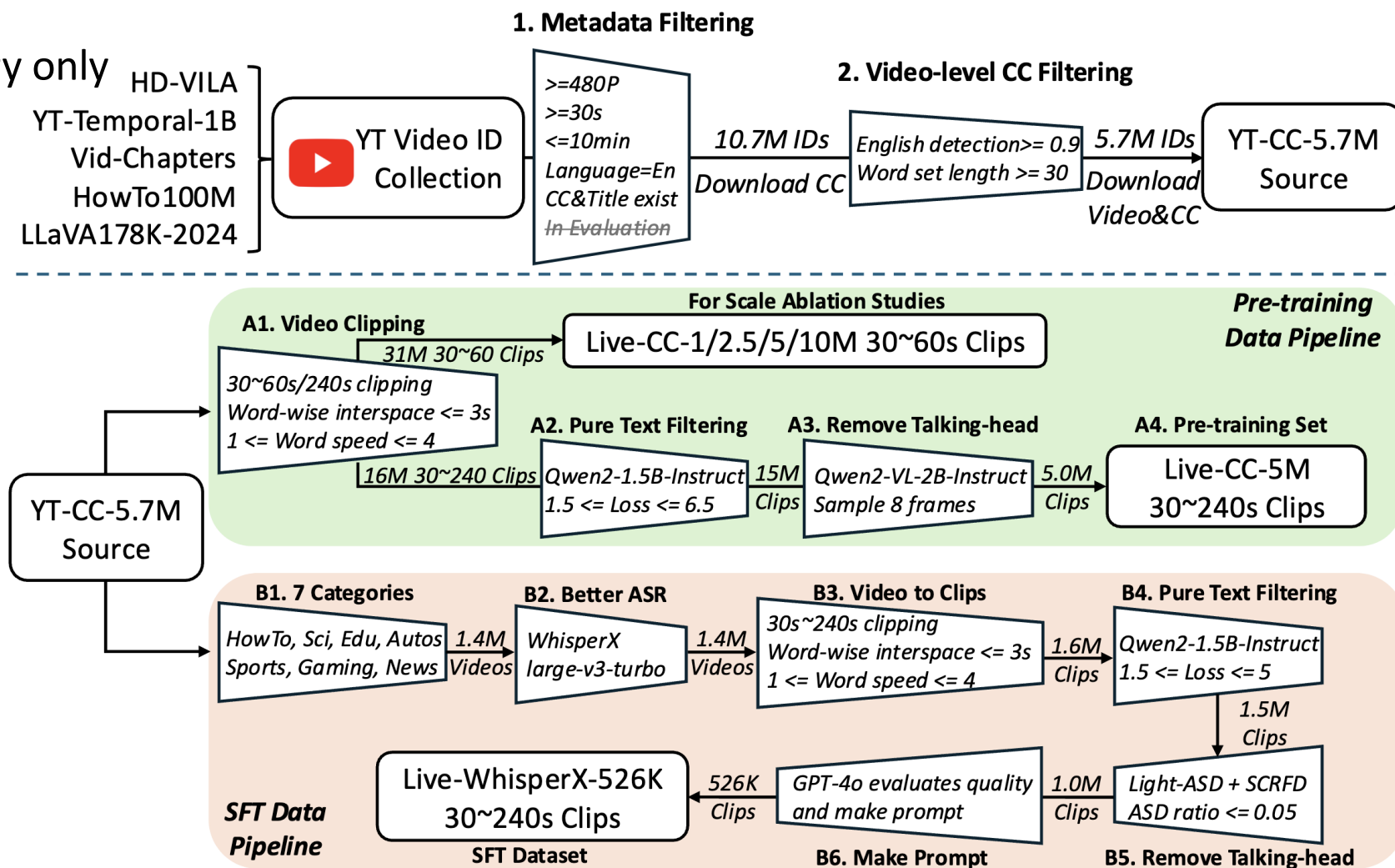
Key Points: Modeling

1. Densely interleaving frame-words according to their timestamps
2. Title or Previous ASR as context during pre-training, but not during SFT
3. Introducing '...' as the streaming EOS token, instead of reusing the common EOS token



Key Points: Data Production Pipeline

1. Language perplexity to remove low-quality ASR
2. Active speaker detection (ASD) to remove talking-head videos
3. GPT-4o for making query only



Key Points: Datasets

- 5M Pre-training Video-ASR Data with YouTube CC
- 526K SFT Video-ASR Data with WhisperX

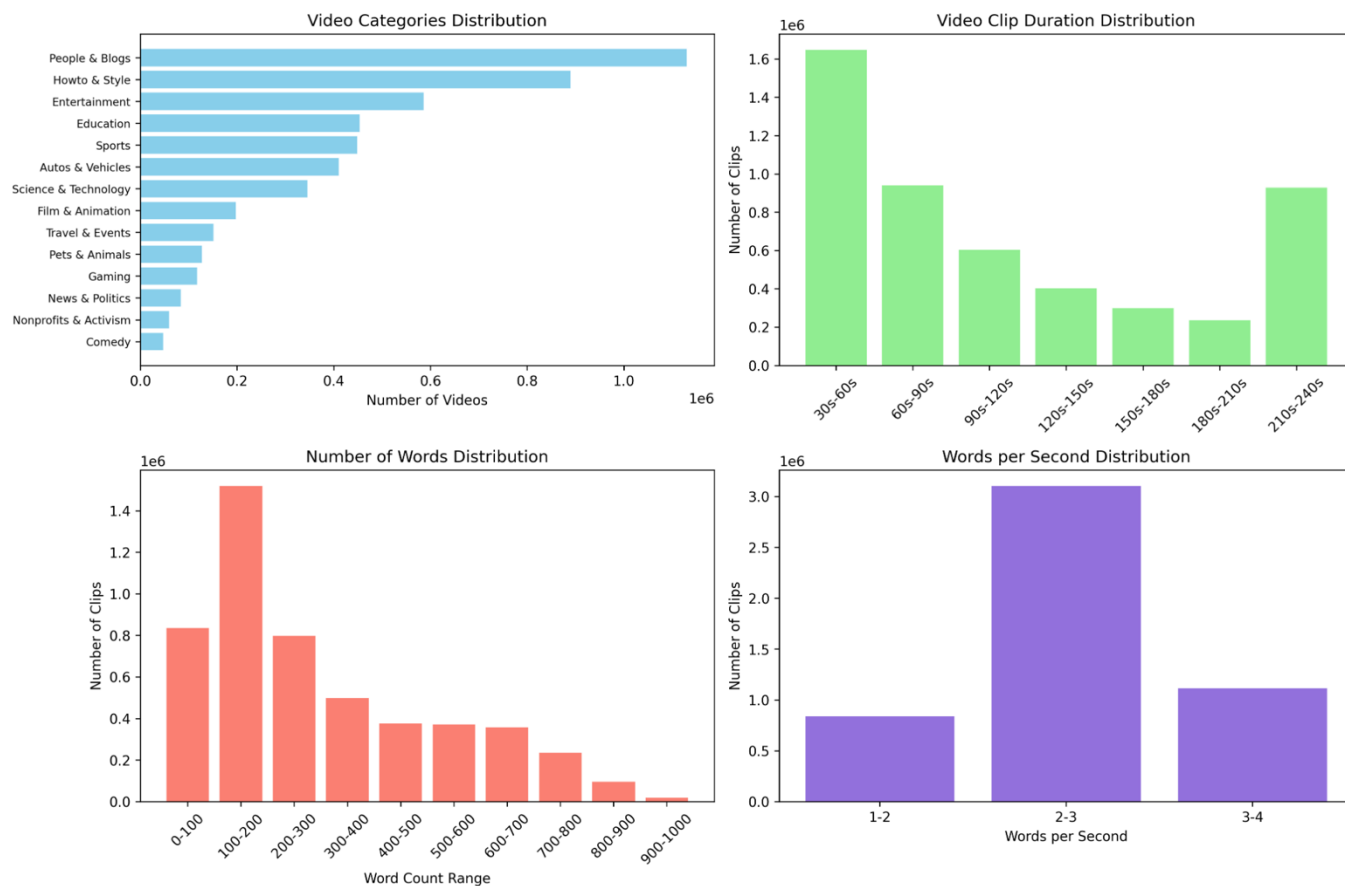
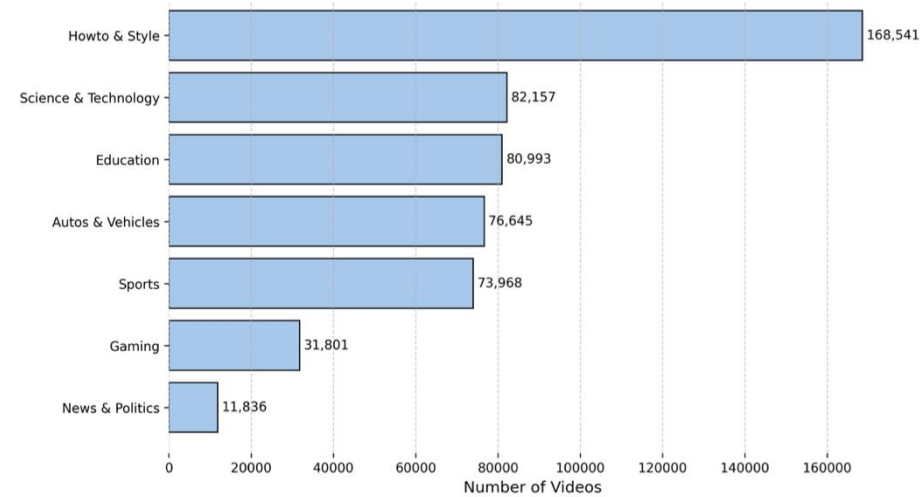
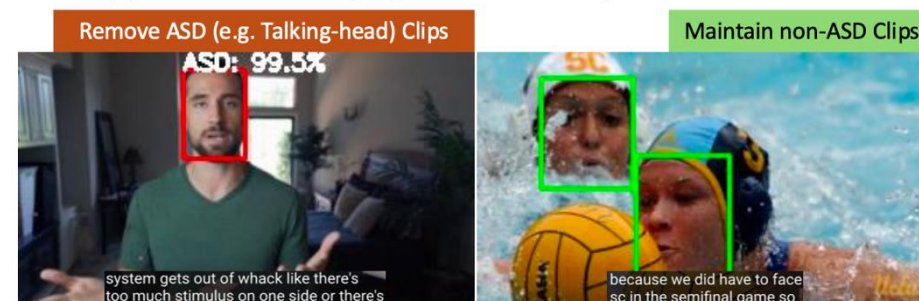


Figure 3. Overview of our proposed YT-CC-5M dataset.



(a) Statistics of our proposed Live-WhisperX-526K dataset.



(b) An example of ASD removal in SFT data pipeline.

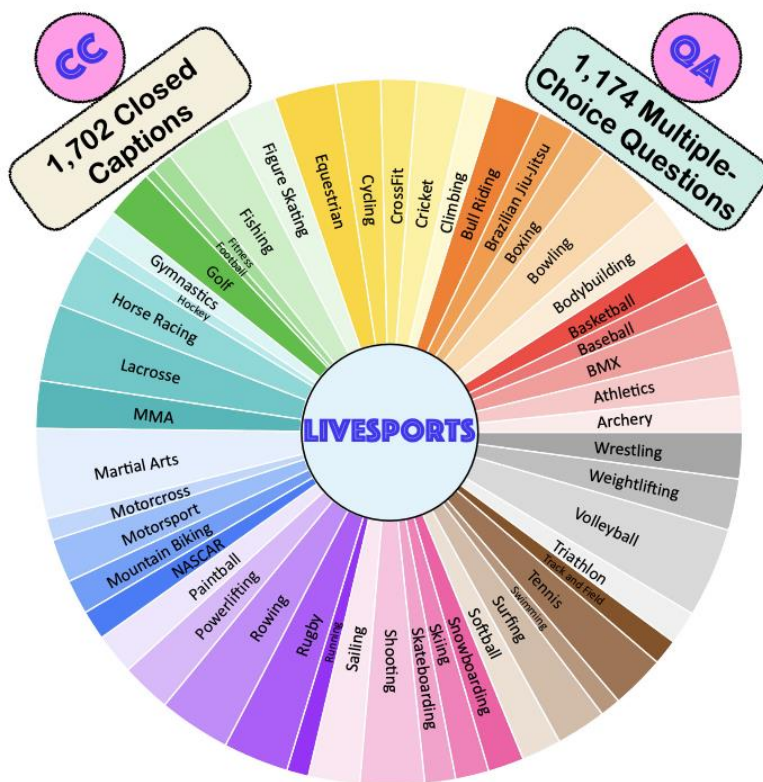


Prompt: Can you provide a step-by-step guide on how to create a special visual effect using Photoshop?
 Text Stream: ...[27.0s-27.2s, "link"], [27.2s-27.4s, "it"], [27.4s-27.9s, "in"], [27.9s-28.0s, "the"], [28.0s-28.4s, "description"], [28.4s-29.0s, "and"], [29.0s-29.1s, "I've"], [29.1s-29.3s, "also"], [29.3s-29.4s, "got"], [29.4s-29.6s, "an"], [29.6s-29.9s, "ice"], [29.9s-30.2s, "image"]...

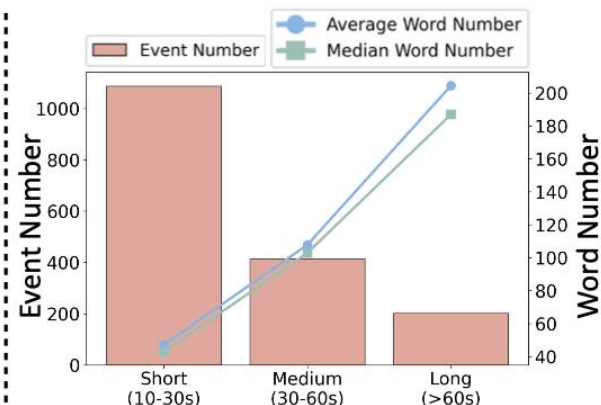
(c) An example from the Live-WhisperX-526K dataset.

Key Points: Benchmark

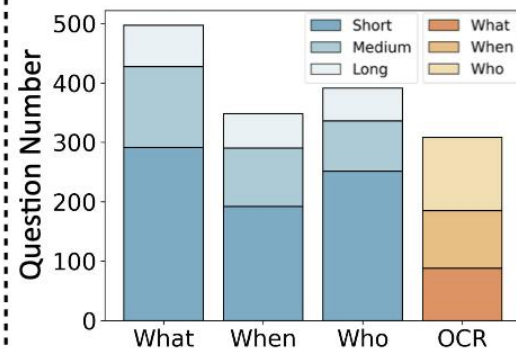
1. LiveSports-3K Benchmark, with CC and QA tracks, focusing on human instance in live sports video
2. LLM-as-a-judge to evaluate CC (video commentary) winning rate vs. GPT-4o



(a) Category Distribution of the LIVESPORTS Benchmark



(b) Event Duration and ASR Word Count in LIVESPORTS-CC



(c) Question Count by Type in LIVESPORTS-QA



Class: Archery

LIVESPORTS-CC

Context ASR Closed Caption (8.58s-13.94s):
We are not like the able-bodied game because we are not so convenient to rolling.

Groundtruth ASR Closed Caption (14.40s-23.78s):
So we will stay on the line, just give him the signal to the judge, and then the next one.

Model A Prediction:
That red hat guy is ready. He finished, now its the next one's turn.

Model B Prediction:
That man with a red hat is shooting. Then he raised his hands.

LLM Judge → A is better

LIVESPORTS-QA

Query Who, Given When and What:
Who is shooting when the player No.34B is waiting? [Options A B C D] Answer: A. The man wearing a red hat.

Query When, Given Who and What:
When did the man wearing a red hat raise his hands? [Options A B C D] Answer: C. After he finished shooting.

Query What, Given Who and When:
What did the man wearing a red hat do after he finished shooting? [Options A B C D] Answer: D. Raise hands.

(d) Sample Closed Captions in LIVESPORTS-CC and Three Query Types in LIVESPORTS-QA

Key Points: Experiments

1. Video-ASR Streaming Pre-training & SFT improve both CC and QA
2. 7B/8B Scale, SOTA on VideoMME (before CVPR submission), OVOBench, LiveSports3K-QA
3. Commentary winning rate vs. GPT-4o surpasses all 72B models

Pre-training	SFT	LiveSports-3K							VideoMME														
		CC	QA	OCR	Who	When	What	All	Duration			Perception			Recognition		Reasoning				OCR	Count	IS
									S	M	L	Te	Sp	At	Ac	Ob	Te	Sp	Ac	Ob			
Qwen2-VL-7B-Base	LV178K	16.7	67.0	66.1	70.6	57.6	71.0	62.7	74.7	62.4	51.1	74.5	61.1	73.4	64.5	70.1	49.7	80.4	49.8	57.0	76.3	45.1	76.2
	LV178K+Live526K	33.7	67.1	66.8	69.8	57.0	72.3	63.6	74.4	63.1	53.2	74.5	57.4	75.2	66.5	70.1	49.7	76.8	54.4	57.5	72.7	44.4	78.9

(a) Ablation study in the SFT data.

Pre-training	SFT	LiveSports-3K							VideoMME														
		CC	QA	OCR	Who	When	What	All	Duration			Perception			Recognition		Reasoning				OCR	Count	IS
									S	M	L	Te	Sp	At	Ac	Ob	Te	Sp	Ac	Ob			
Qwen2-VL-7B-Base	-	16.3	64.0	64.8	65.2	57.9	67.3	63.4	73.2	63.2	53.9	72.7	63.0	76.1	63.9	67.2	44.6	78.6	57.5	61.5	72.7	39.6	80.2
LiveCC-7B-Base	-	43.2	57.9	61.4	59.4	50.7	61.9	61.4	68.1	58.9	57.3	65.5	63.0	64.9	60.7	61.0	50.3	80.4	56.1	61.5	61.2	42.9	82.4
Qwen2-VL-7B-Base	LV178K+Live526K	33.7	67.1	66.8	69.8	57.0	72.3	63.6	74.4	63.1	53.2	74.5	57.4	75.2	66.5	70.1	49.7	76.8	54.4	57.5	72.7	44.4	78.9
LiveCC-7B-Base	LV178K+Live526K	41.5	66.8	66.4	71.4	56.1	70.8	64.1	74.8	63.9	53.7	74.5	64.8	74.3	66.1	68.6	50.3	76.8	52.3	59.5	77.0	46.3	79.9

(b) Ablation study in the SFT model initialization.

Model (7B/8B)	VideoMME		MVBench	OVOBench			
	w/o sub	w sub	Avg.	Avg.	RTVP	BT	FAR
LongVA-7B [102]	52.6	54.3	-	-	-	-	-
InternVL2-8B [18]	54.0	56.9	66.4	50.2	60.4	43.4	46.6
LLaVA-OV-7B [42]	58.2	61.5	56.7	52.7	64.0	43.7	50.5
Oryx-7B [56]	58.3	62.6	63.9	-	-	-	-
mPLUG-Owl3-7B [93]	59.3	68.1	59.5	-	-	-	-
LongVU-7B [74]	60.6	-	66.9	46.7	57.6	35.0	47.5
MiniCPM-v2.6 [91]	60.9	63.6	-	-	-	-	-
Qwen2-VL-7B-Instruct [79]	63.3	69.0	67.0	50.4	56.0	46.5	48.7
LLaVA-Video-7B [105]	63.3	69.7	58.6	52.9	63.5	40.4	54.8
LiveCC-7B-Instruct	64.1	70.3	62.8	59.8	59.1	68.9	51.5

Size	Model	LiveSports-3K ↑						
		Live?	CC	Overall	OCR	Who	When	What
-	GPT-4o-08-06 [29]	✗	✳	72.2	74.0	75.8	63.4	75.4
	Gemini-1.5-Pro [3]	✗	52.8	61.8	61.7	59.9	51.6	70.7
72B	Qwen2-VL-72B-Instruct [80]	✗	17.0	70.8	67.8	74.6	61.2	74.6
	VideoLLaMA-2-72B [18]	✗	24.8	62.4	55.7	63.6	54.3	67.3
	LLaVA-OV-72B [41]	✗	29.2	68.7	61.7	71.1	61.5	71.8
	Qwen2.5-VL-72B-Instruct [6]	✗	30.4	73.7	70.1	75.7	69.3	75.3
	LLaVA-Video-72B [106]	✗	35.0	71.1	65.1	74.1	64.8	73.3
7B	Qwen2-VL-7B-Instruct [80]	✗	9.3	65.8	65.8	67.9	58.8	69.2
	Qwen2.5-VL-7B-Instruct [6]	✗	17.3	67.0	64.8	70.3	60.6	69.0
	InternLM-XC2.5-7B [102]	✗	17.3	59.3	56.7	60.7	54.9	61.3
	Qwen2.5-Omni-7B [87] (Thinker)	✗	17.6	66.8	66.1	70.0	60.0	69.2
	LLaVA-Video-7B [106]	✗	27.1	66.4	64.1	72.7	56.4	68.6
	LLaVA-OV-7B [41]	✗	27.7	63.4	60.7	67.4	53.7	67.1
	Qwen2-VL-7B-LiveCCInstruct	✓	33.7	67.1	66.8	69.8	57.0	72.3
	LiveCC-7B-Instruct	✓	41.5	66.8	66.4	71.4	56.1	70.8
LiveCC-7B-Base	✓	43.2	57.9	61.4	59.4	50.7	61.9	