NEURAL INFORMATION PROCESSING SYSTEMS

UVE: Are MLLMs Unified Evaluators for AI-Generated Videos?

Yuanxin Liu[§] Rui Zhu[‡] Shuhuai Ren[§] Jiacong Wang[¶]
Haoyuan Guo[‡] Xu Sun[§] Lu Jiang[‡]

[§] State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University

[‡] ByteDance Seed

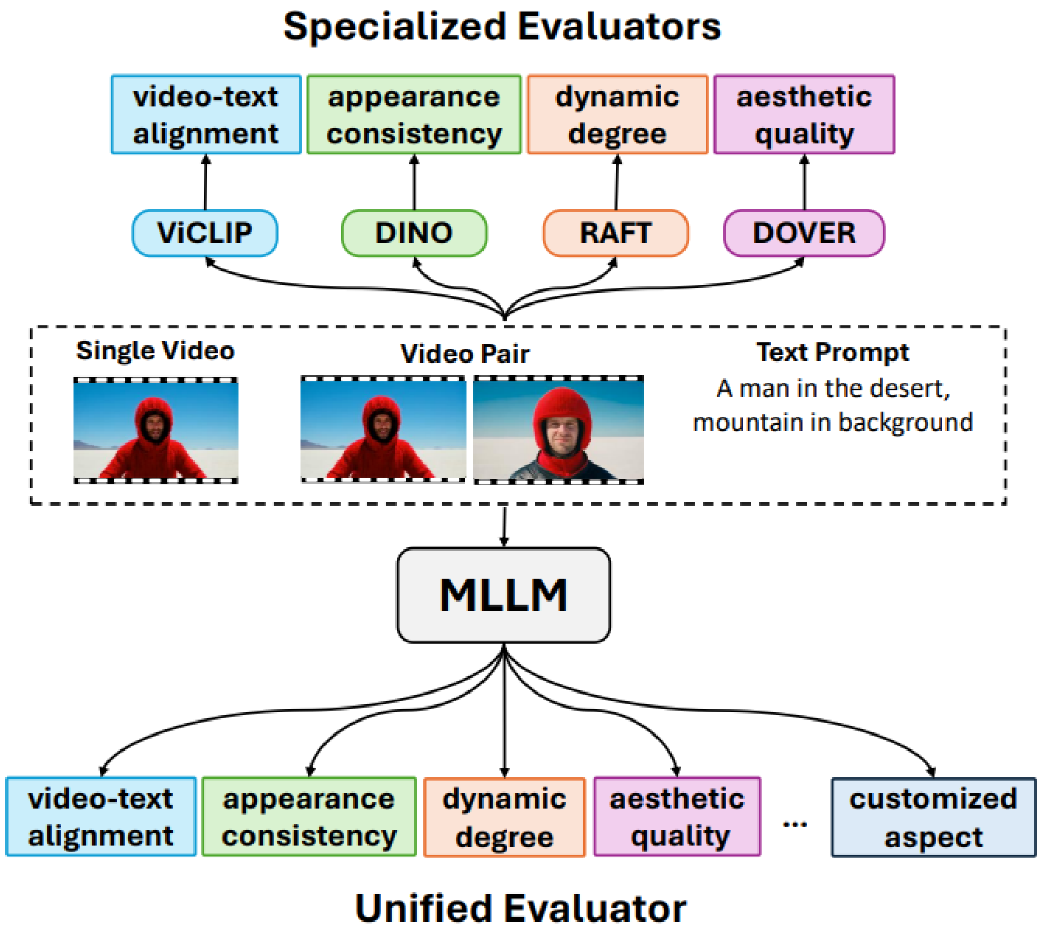
[¶] School of Artificial Intelligence, University of Chinese Academy of Sciences

ByteDance | Seed

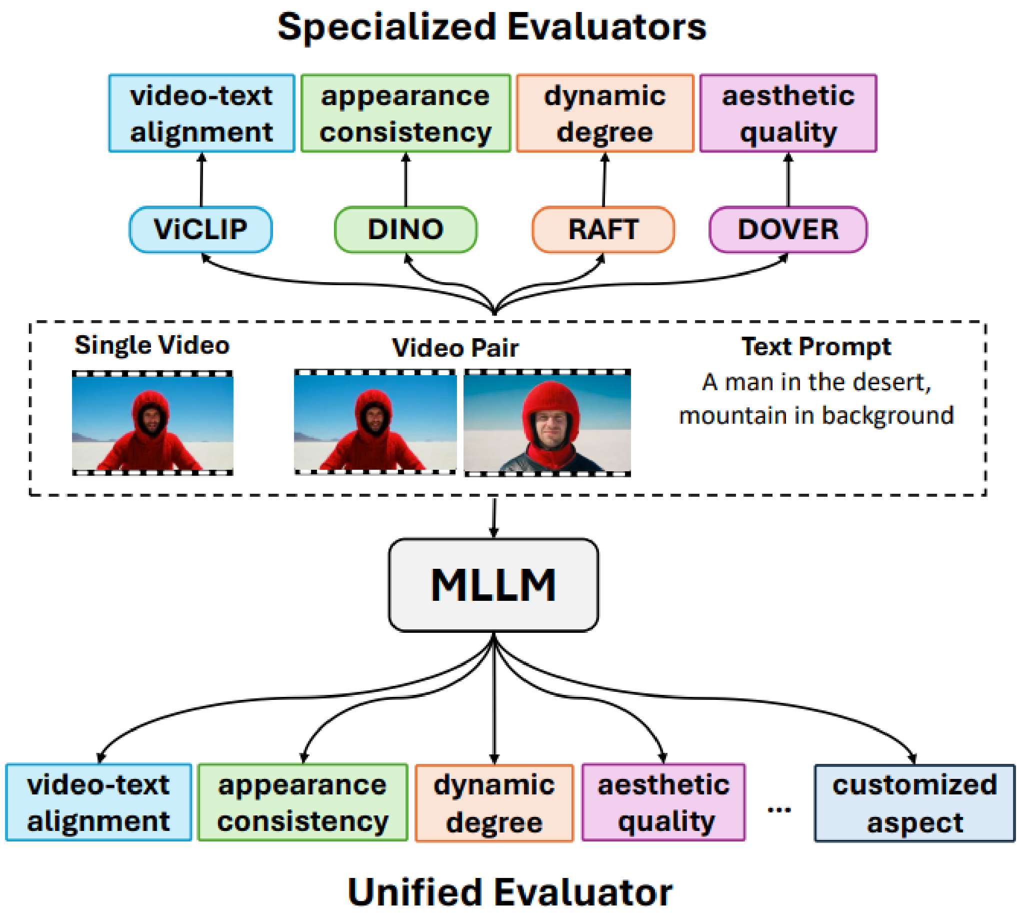
Motivation

- **Challenge of AI-Generated Video Evaluation:**
  - Current studies rely on **specialized evaluators** for individual aspects (e.g., video-text alignment, aesthetic quality), which is **incomprehensive and hard to scale**.
- **Opportunity:**
  - Modern **MLLMs** exhibit **general vision-language understanding** ability in open domain scenarios.
- **Key question:**
  - **Can MLLMs be utilized as a unified evaluator for AI-Generated videos?**

Method: Unified Video Evaluation Framework



Method: Unified Video Evaluation Framework



Single Video Rating
<video>
Watch the above frames of an AI-generated video and evaluate <aspect-specific description>

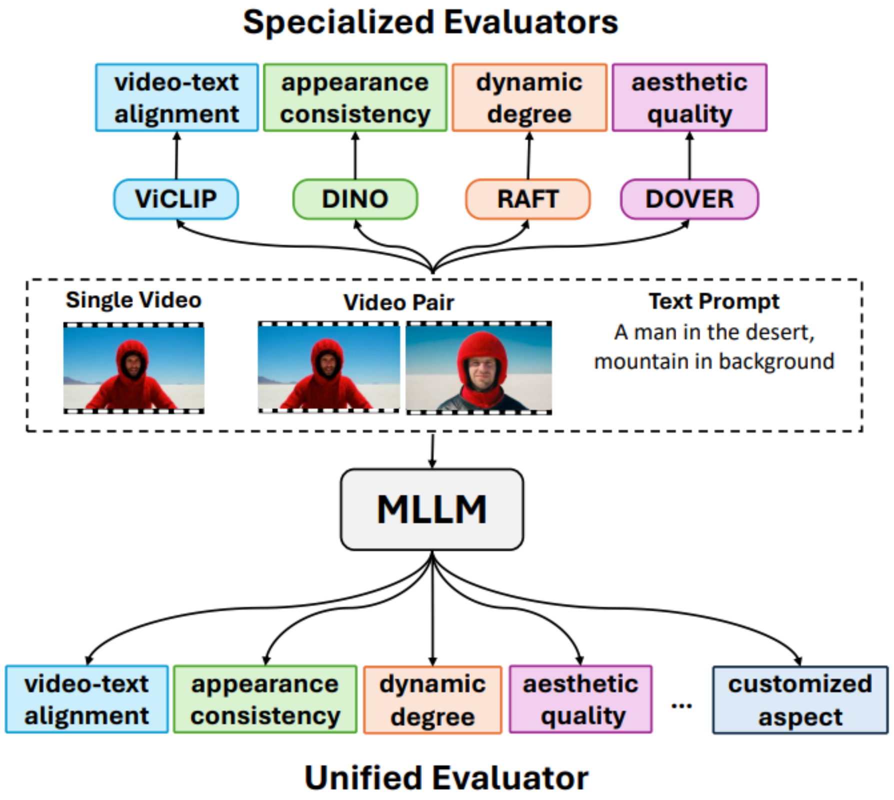
Complete your evaluation by answering this question:
<aspect-specific question>?
<answer prompt>

Video Pair Comparison
The first video: <video>
The second video: <video>
Watch the above two AI-generated videos and evaluate <aspect-specific description>

Complete your evaluation by answering this question:
Which video is <aspect-specific question>?

You should make your judgment based on the following rules:
<instructions on how to make the choice>
Now give your judgment:

Method: Unified Video Evaluation Framework



Single Video Rating
<video>
Watch the above frames of an AI-generated video and evaluate <aspect-specific description>

Complete your evaluation by answering this question:
<aspect-specific question>?
<answer prompt>

Video Pair Comparison
The first video: <video>
The second video: <video>
Watch the above two AI-generated videos and evaluate <aspect-specific description>

Complete your evaluation by answering this question:
Which video is <aspect-specific question>?

You should make your judgment based on the following rules:
<instructions on how to make the choice>
Now give your judgment:

single video rating

$$\mathcal{S} = \frac{P_\theta(t_{\text{pos}}|\mathcal{V},\mathcal{T},\mathcal{G}_a)}{P_\theta(t_{\text{pos}}|\mathcal{V},\mathcal{T},\mathcal{G}_a) + P_\theta(t_{\text{neg}}|\mathcal{V},\mathcal{T},\mathcal{G}_a)}$$

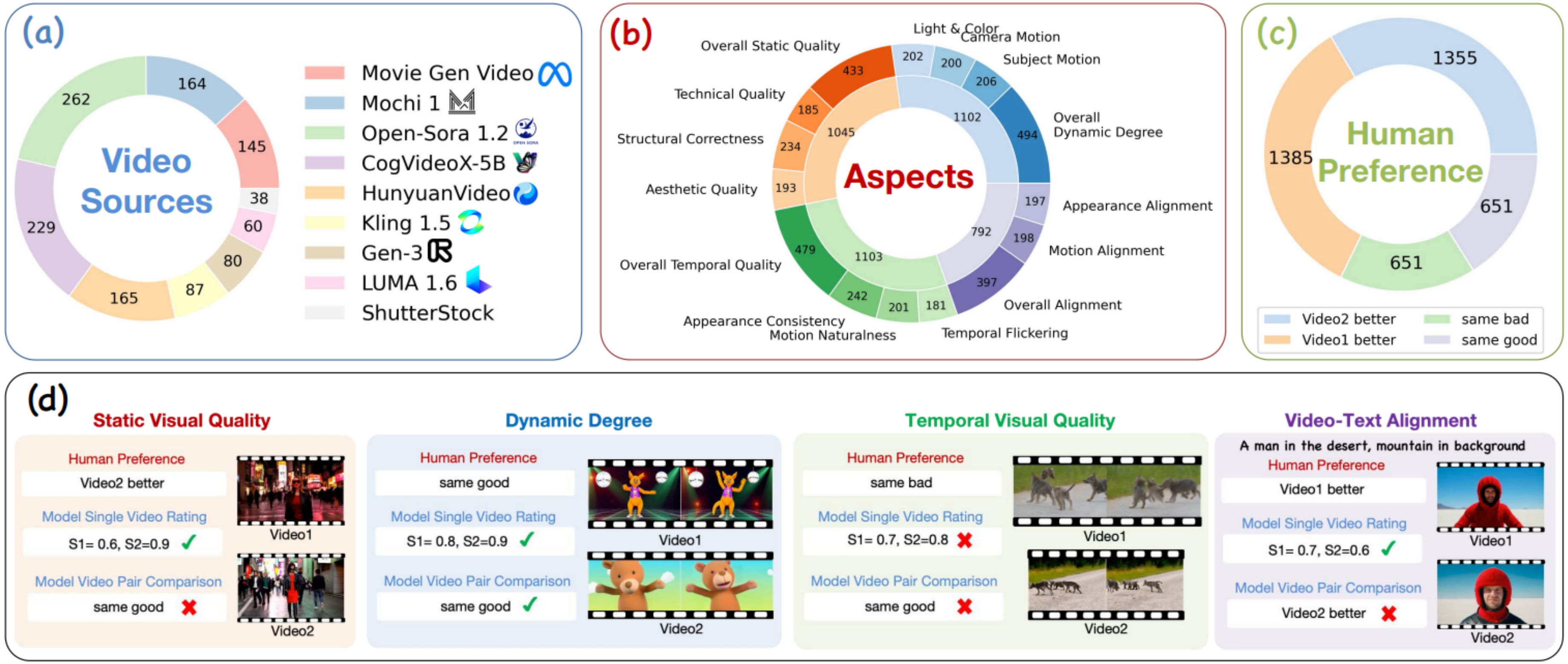
video pair comparison

$$\mathcal{C} = f_\theta(\mathcal{V}_1, \mathcal{V}_2, \mathcal{T}_1, \mathcal{T}_2, \mathcal{G}_a) \in \mathbf{O}$$

$$\{\text{“}\mathcal{V}_1\text{better”}, \text{“}\mathcal{V}_2\text{better”}, \text{“same good”}, \text{“same bad”}\},$$

UVE-Bench: A benchmark to assess AIGV evaluators

(a): 1,230 videos generated by 8 SOTA T2V models and real-world videos from ShutterStock
(b): 15 fine-grained AIGV evaluation aspects
(c): Human pairwise preference annotation as ground-truth
(d): Support of both single video rating and video pair comparison



Performance of MLLMs as Unified Video Evaluator

- **Unified evaluators outperforms specialized evaluators**
- **Good Aspects**: *Dynamic Degree, Technical/Aesthetic Quality, Appearance Alignment*
- **Bad Aspects**: *Structural Correctness, Temporal Quality, Motion Alignment*

Single Video Rating

Method	Model Size	Overall Dynamic	SM	CM	LC	Overall Static	TQ	SC	AQ	Overall Temporal	AC	MN	TF	Overall Alignment	MA	AA	AVG
Random	-	48.8	49.5	48.9	49.8	49.2	47.6	49.0	48.1	49.0	48.6	48.2	46.2	48.4	48.6	48.3	48.6
Specialized Evaluators																	
VideoScore-v1.1	8B	57.4	-	-	-	40.1	30.4	47.7	41.9	-	43.1	36.1	-	38.9	-	-	-
VBench	-	**87.8**	-	-	-	-	62.5	-	75.6	-	54.0	50.2	-	54.4	-	-	-
UMTScore	-	-	-	-	-	-	-	-	-	-	-	-	-	66.4	-	-	-
VIDEOCON-PHYSICS	7B	-	-	-	-	-	-	-	-	-	-	53.4	-	68.7	-	-	-
DOVER	58M	-	-	-	-	-	69.2	-	80.3	-	-	-	-	-	-	-	-
Unified Evaluators																	
Video-LLaVA	7B	52.4	74.8	63.4	47.6	47.8	66.1	44.5	63.4	44.1	51.3	55.5	48.6	59.4	54.9	66.8	54.5
LongVA-DPO	7B	59.8	76.2	69.9	57.5	62.4	74.9	56.2	72.0	56.0	47.3	40.7	54.3	68.7	63.9	71.6	61.6
ShareGPT4Video	8B	77.5	81.0	77.1	82.5	59.4	68.7	54.1	69.4	54.4	48.1	42.6	60.9	62.7	54.3	70.3	63.9
VideoLLaMA2.1	7B	72.1	80.5	67.1	77.2	61.4	78.7	47.5	72.6	46.1	50.9	50.2	61.8	72.6	65.7	77.7	64.4
mPLUG-Owl3	7B	78.6	84.8	77.8	79.9	76.0	83.1	55.6	80.0	59.8	59.5	42.0	72.0	80.4	75.2	87.6	72.6
VideoChat2-Mistral	7B	83.1	92.2	89.6	74.9	68.1	76.2	53.8	74.1	58.7	58.3	52.8	85.6	75.6	78.0	80.9	72.6
MiniCPM-V-2.6	8B	81.4	86.3	80.3	88.8	70.9	75.0	52.9	80.6	61.1	59.4	51.7	70.9	82.1	74.3	90.8	73.4
LLaVA-OneVision	7B	81.0	87.6	83.2	84.9	70.7	78.2	50.6	83.1	62.9	60.4	41.5	85.9	79.3	66.9	86.7	73.0
LLaVA-OneVision	72B	82.3	87.8	78.4	88.2	71.3	77.6	60.2	81.5	64.6	61.9	39.9	86.8	84.4	71.7	93.5	75.0
LLaVA-Video	7B	80.4	85.5	80.9	81.5	66.2	74.2	49.5	75.7	58.3	58.2	39.3	82.3	80.5	69.9	90.0	71.0
LLaVA-Video	72B	82.8	86.1	82.9	86.9	70.2	80.0	55.4	77.7	60.1	59.2	40.4	83.5	84.8	73.7	94.6	74.0
Qwen2-VL	7B	84.6	89.7	**94.2**	79.7	64.6	67.3	50.7	70.6	51.1	51.3	48.0	62.7	85.4	78.7	92.2	70.9
Qwen2-VL	72B	86.5	**92.6**	92.7	86.0	70.6	76.9	60.2	83.4	52.5	58.0	48.9	71.0	**89.0**	**81.8**	95.0	75.4
InternVL-2.5-MPO	8B	81.3	86.1	80.4	88.0	68.1	77.9	53.6	77.5	60.9	54.9	50.9	72.5	80.6	73.2	90.5	72.6
InternVL-2.5-MPO	78B	84.3	86.6	82.4	**91.6**	72.8	**84.0**	**67.2**	82.3	61.5	65.2	**61.8**	**88.4**	87.4	79.3	**95.5**	78.2
GPT-4o	-	79.0	84.3	74.9	81.8	74.0	81.6	65.2	84.2	70.0	**79.8**	54.0	58.6	80.8	77.2	89.6	75.7
Seed1.5-VL	20B Act.	83.2	91.2	83.7	89.5	**82.4**	82.9	66.8	**88.8**	**70.5**	78.6	59.5	70.1	84.2	79.0	94.2	**80.0**

Performance of MLLMs as Unified Video Evaluator

- **Good Aspects**: *Dynamic Degree, Technical/Aesthetic Quality, Appearance Alignment*
- **Bad Aspects**: *Structural Correctness, Temporal Quality, Motion Alignment*
- There is still a notable gap between MLLM evaluators and humans

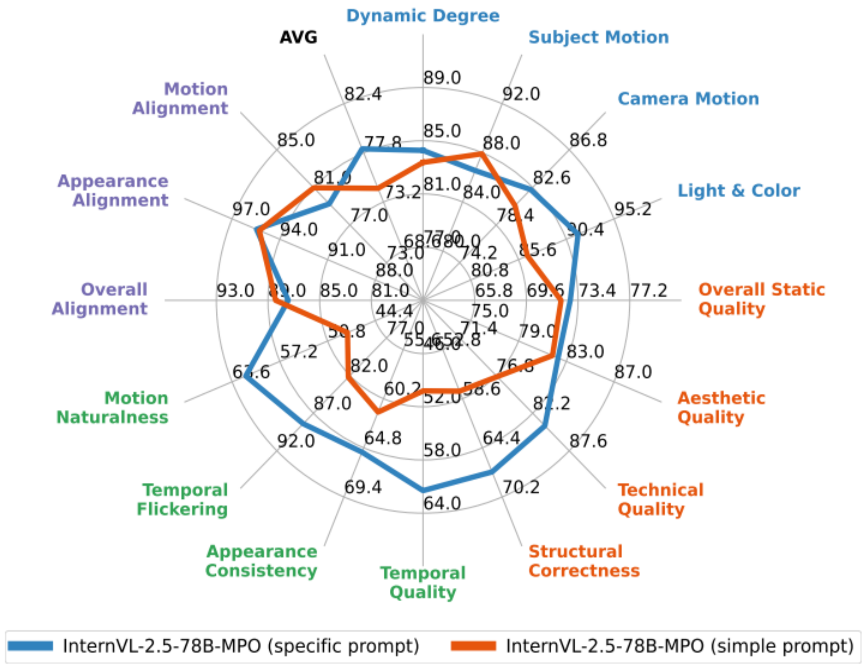
Video Pair Comparison

Method	Model Size	Overall Dynamic	SM	CM	LC	Overall Static	TQ	SC	AQ	Overall Temporal	AC	MN	TF	Overall Alignment	MA	AA	AVG
Random	-	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Unified Evaluators																	
LLaVA-OneVision	7B	44.2	52.6	40.0	38.5	35.8	35.2	25.7	51.0	37.4	23.9	29.1	52.0	43.1	39.7	43.4	38.6
LLaVA-OneVision	72B	36.5	55.1	40.0	53.8	37.0	51.4	22.2	54.1	25.6	43.3	34.5	45.7	63.0	58.9	69.7	44.3
LLaVA-Video	7B	37.0	52.6	37.5	44.9	44.4	43.8	38.9	62.2	33.7	28.9	31.1	57.5	43.1	39.7	46.9	41.1
LLaVA-Video	72B	42.5	57.7	32.5	56.4	41.6	55.2	31.2	60.2	32.6	41.1	27.7	55.9	60.3	53.0	66.2	46.1
Qwen2-VL	7B	46.4	56.4	43.8	39.7	42.8	38.1	20.8	54.1	29.8	24.4	29.7	42.5	54.9	50.3	57.9	41.1
Qwen2-VL	72B	51.4	**66.7**	**71.2**	53.8	47.7	**62.9**	19.4	60.2	41.0	40.0	31.8	40.2	**69.7**	**62.9**	**78.6**	51.6
InternVL-2.5-MPO	8B	42.5	51.3	33.8	43.6	38.3	31.4	40.3	55.1	35.1	32.2	27.0	44.9	53.5	50.3	53.8	41.8
InternVL-2.5-MPO	78B	43.1	57.7	41.2	**61.5**	54.7	**62.9**	27.1	**71.4**	45.2	**47.8**	33.8	**70.9**	67.7	55.6	76.6	53.7
GPT-4o	-	42.0	48.7	38.8	59.0	53.5	54.3	41.0	**71.4**	44.9	38.9	31.8	59.8	58.9	57.0	61.4	50.2
Seed1.5-VL	20B Act.	52.5	56.4	47.5	**61.5**	51.0	52.4	**44.4**	62.2	48.6	36.7	35.8	65.4	56.9	53.6	61.4	51.6
Gemini2.5-Flash	-	**55.8**	60.3	45.0	59.0	**55.6**	50.5	43.1	64.3	**50.8**	41.7	**38.5**	60.6	65.0	62.3	66.2	**54.6**
Human	-	87.3	85.3	87.3	85.3	90.0	92.0	88.0	91.3	88.0	90.0	78.7	92.0	88.0	84.7	92.0	88.0

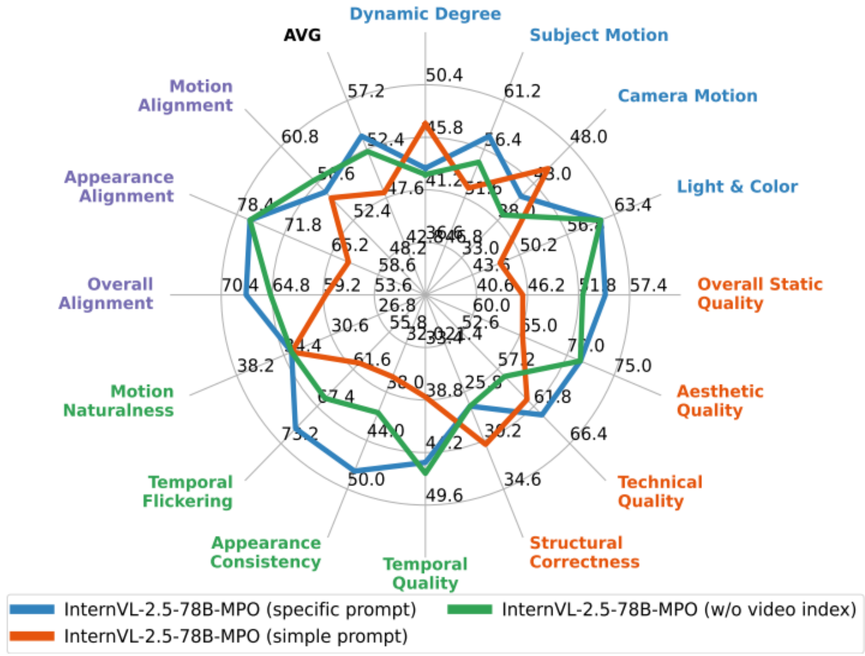
Key Design Choices: Prompting Strategy

- Aspect-specific prompting is essential
- Removing video order index does not significantly impact performance.

Single Video Rating



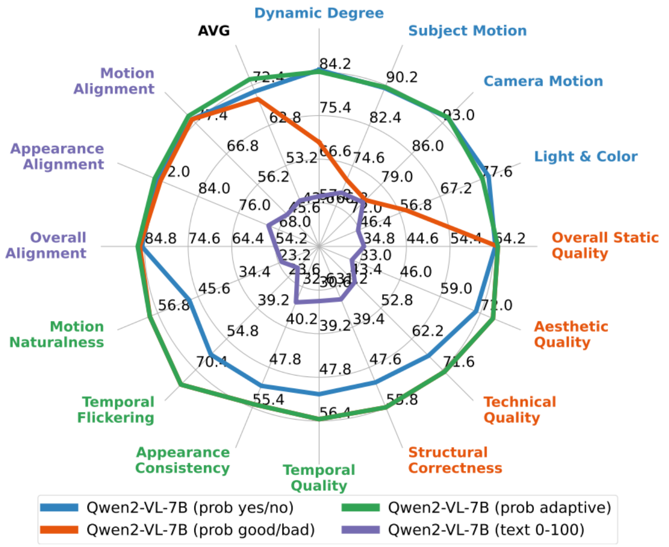
Video Pair Comparison



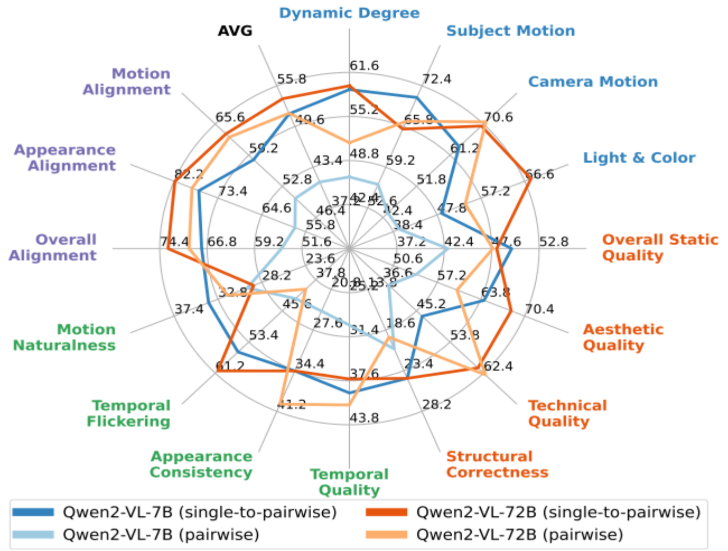
Key Design Choices: Scoring Strategy

- Single Video Rating
  - Probability-based scoring outperforms directly generating discrete rating score (0-100)
  - Yes/no and good/bad excel at different aspects, adaptive strategy is more effective
- Video Pair Comparison
  - Adapting from single video rating outperforms direct video pair comparison for 7B-scale models

Single Video Rating



Video Pair Comparison



Summary of Contributions

- **UVE Framework**
  - We introduce a unified approach to evaluate any aspect of AIGV using pre-trained MLLMs.
- **UVE-Bench**
  - We propose UVE-Bench, a comprehensive benchmark to assess the capability of unified AIGV evaluation.
- **Experiments and Analysis**
  - We demonstrate that unified MLLM evaluators substantially outperforms existing specialized evaluators.
  - We conduct in-depth analysis on the pros and cons of MLLMs in unified AIGV evaluation and the key design choices that impact their performance.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

Code: https://github.com/bytedance/UVE
Data: https://huggingface.co/datasets/lyx97/UVE-Bench

ByteDance | Seed