GAIA: Rethinking Action Quality Assessment for AI-Generated Videos

NeurIPS 2024 Dataset and Benchmark Track Spotlight

Zijian Chen[1], Wei Sun[1], Yuan Tian[1], Jun Jia[1], Zicheng Zhang[1], Jiarui Wang[1], Ru Huang[2], Xiongkuo Min[1], Guangtao Zhai[1], Wenjun Zhang[1]

[1]Shanghai Jiao Tong University

[2]East China University of Science and Technology

Overview

1. Motivation
2. Construction of GAIA
3. Observations
4. Experiments

1. Motivation

➢ Background

- Action quality assessment (AQA), which aims to quantify how well actions are performed, is a growing area of research across various domains (e.g., sports event, health care, and public security)
- Assessing how well an action is presented can be difficult because of the inherent difference between real videos and generated videos.
- At minimum, a well-performed action should correctly contain all relevant objects as well as the action subject with recognizable motion presentation while conforming to the physical world dynamics.
- At present, it remains unclear to what degree any T2V model can achieve visually rational action generation that varies in action categories, much less the cognitive mechanism of action quality that affects human perception.

1. Motivation

➢ The limitations of existing AQA datasets

- Predominantly focus on ***domain-specific*** actions from real videos and collect *coarse-grained* expert-only human ratings on limited dimensions.

- The content discrepancies in those AQA videos are often subtle, as the action subjects typically perform similar actions within a consistent environment (***lacks of scene diversity***)

Table: Comparison of **GAIA** and existing AQA datasets. SS indicates the source of scores. Mix indicates that the participants in human evaluation are recruited across different backgrounds.

Dataset	Source	Action	Samples	Duration	Avg.Dur.	Resolution	FPS	SS
MIT Dive (2014) [79]	$\text{Real}_{\text{diving}}$	–	159	0.25h	6.0s	320×240	30	Judge
UNLV Dive (2017) [78]	$\text{Real}_{\text{diving}}$	–	370	0.4h	3.8s	320×240	30	Judge
AQA-7-Dive (2019) [76]	$\text{Real}_{\text{diving}}$	–	549	0.6h	4.1s	320×240	30	Judge
MTL-AQA (2019) [77]	$\text{Real}_{\text{diving}}$	–	1,412	1.5h	4.1s	1920×1080	25	Judge
Rhyth. Gym. (2020) [118]	$\text{Real}_{\text{gymnastics}}$	–	1,000	26.3h	95s	1280×720	25	Judge
FSD-10 (2020) [65]	$\text{Real}_{\text{skating}}$	10	1,484	–	3-30s	1080×720	30	Judge
Fitness-AQA (2022) [75]	$\text{Real}_{\text{workout}}$	3	13,049	14.9h	4.1s	480^2-720^2	30	Expert
FineDiving (2022) [114]	$\text{Real}_{\text{diving}}$	52	3,000	3.5h	4.2s	256×256	15	Judge
LOGO (2023) [121]	$\text{Real}_{\text{swimming}}$	12	200	11.3h	204s	1280×720	25	Judge
GAIA	**AI-generated**	**510**	**9,180**	**7.1h**	**2.8s**	**256^2-2048^2**	**4-50**	***Mixture***

1. Motivation

➢ The limitations of existing AQA methods

- Mainly follow a ***pose-based*** or ***vision-based*** feature extraction, aggregation, and score regression ternary form, which usually adopt powerful 3D backbone networks that are pre-trained on large action recognition datasets for better feature migration.

- A distinguishing characteristic of generated videos is that they may contain atypical actions with various body or object artifacts over time, such as aberrant limb count, irrational object shape, and physically implausible motion, due to the stochasticity and unstable nature of the diffusion process.

- In such cases, the model learned from real action videos may fail in AIGVs with worse prediction performance.

2. Construction of GAIA

Source content

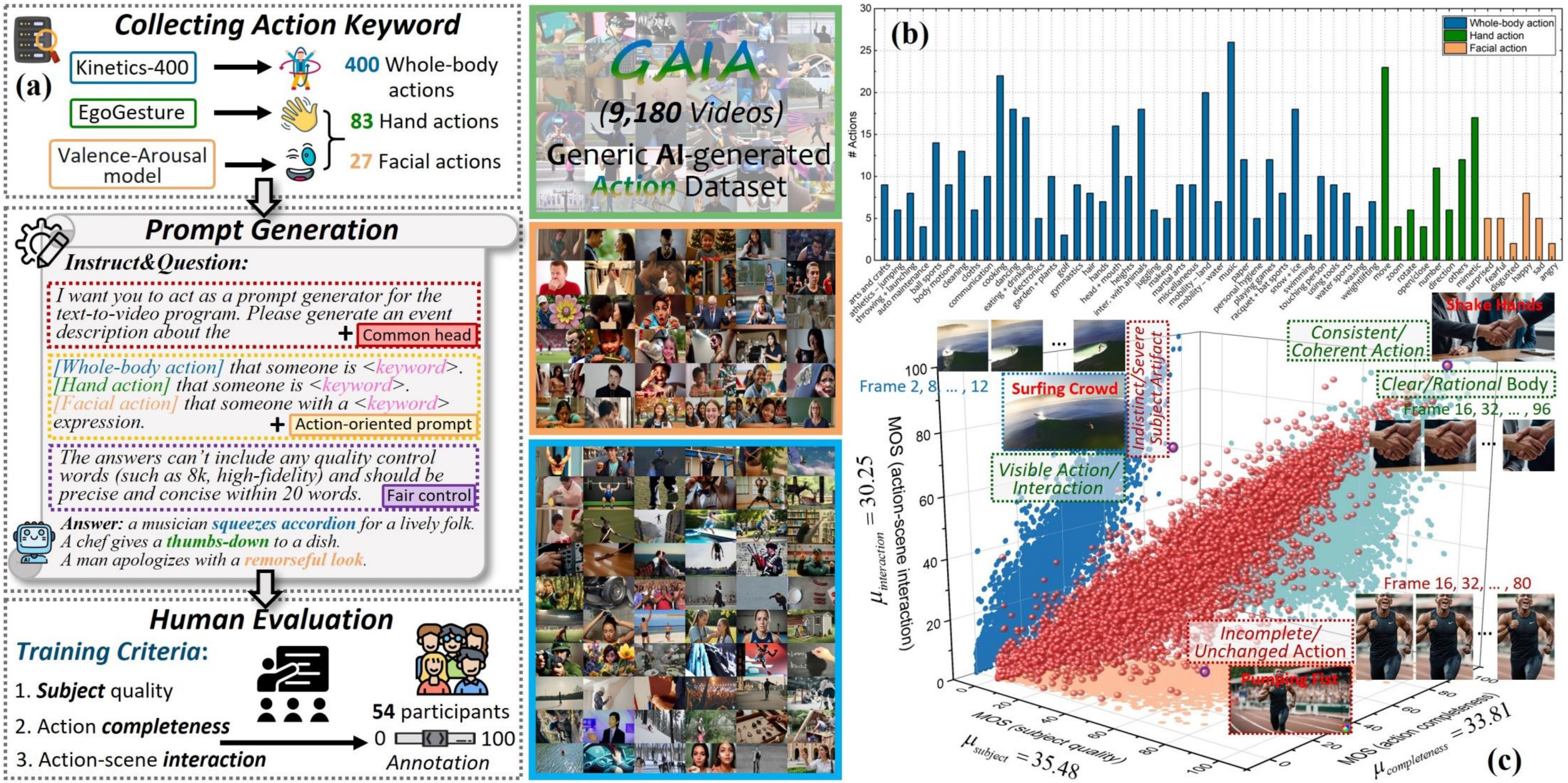

Figure: Data construction pipeline and content overview of GAIA.

Subjective study

Overall statistics

2. Construction of GAIA

11 open-source lab studies

7 popular commercial platforms

Model	Year	Mode	Resolution	FPS	Length	Speed	Feature	Open Source
CogVideo [48]	22.05	T2V	480×480	8	4s	12s	–	✓
Text2Video-Zero [52]	23.03	T2V	512×512	4	2s	21s	Pose/Edge Ctrl	✓
ModelScope [105]	23.03	T2V	256×256	8	2s	6s	–	✓
ZeroScope$_{v2\text{-}576w}$ [8]	23.06	T2V	576×320	8	3s	20s	–	✓
LaVie [109]	23.09	T2V	512×320	8	2s	14s	Interpol./Super Res.	✓
VideoCrafter1 [15]	23.10	T2V, I2V	512×320	8	2s	41s	–	✓
	23.10	T2V, I2V	1024×576	8	2s	*OOM*	–	✓
Show-1 [119]	23.10	T2V	576×320	8	4s	231s	–	✓
Hotshot-XL [70]	23.10	T2V	672×384	8	1s	14s	Personalized	✓
AnimateDiff [40]	23.12	T2V, I2V	384×256	8	2s	10s	Cam. Ctrl	✓
VideoCrafter2 [16]	24.01	T2V, I2V	512×320	8	2s	45s	–	✓
Mora [117]	24.03	T2V, I2V, V2V	1024×576	25	>12s	*OOM*	Multi-Agent	✓
Gen-1 [31]	23.02	V2V	768×448	24	4s	52s†	Style	–
Genmo [2]	23.10	T2V, I2V	2048×1536	15	4s	60s†	Style, Cam. Ctrl	–
Gen-2 [1]	23.12	T2V, I2V	1408×768	24	4s	140s†	Mot./Cam. Ctrl	–
Pika [6]	23.12	T2V, I2V, V2V	1088×640	24	3s	45s†	Mot./Cam. Ctrl, Sound	–
NeverEnds [5]	23.12	T2V, I2V	1024×576	10	3s	260s†	–	–
MoonValley [3]	24.01	T2V, I2V	1184×672	50	4s	386s†	Style, Cam. Ctrl	–
Morph Studio [4]	24.01	T2V, I2V	1920×1080	24	3s	196s†	Mot./Cam./fps Ctrl	–
Stable Video [7]	24.03	T2V, I2V	1024×576	24	4s	125s†	Style, Mot./Cam. Ctrl	✓

Table: Summary of popular video generation models: from open-source lab studies to large-scale commercial creation platforms. We tested the average generation speed (seconds/item) on an NVIDIA RTX4090 locally, except for those closed-source models. OOM is the abbreviation of out-of-memory. †We report the online generation speed under free plan.

2. Construction of GAIA

➢ Design Philosophy: the Action Syllogism

- We propose a causal reasoning-based evaluation strategy
- We decompose an action process into three parts: 1) action subject as major premise, 2) action completeness as minor premise, and 3) interaction between action and scenes as conclusion, according to the *syllogism theory*

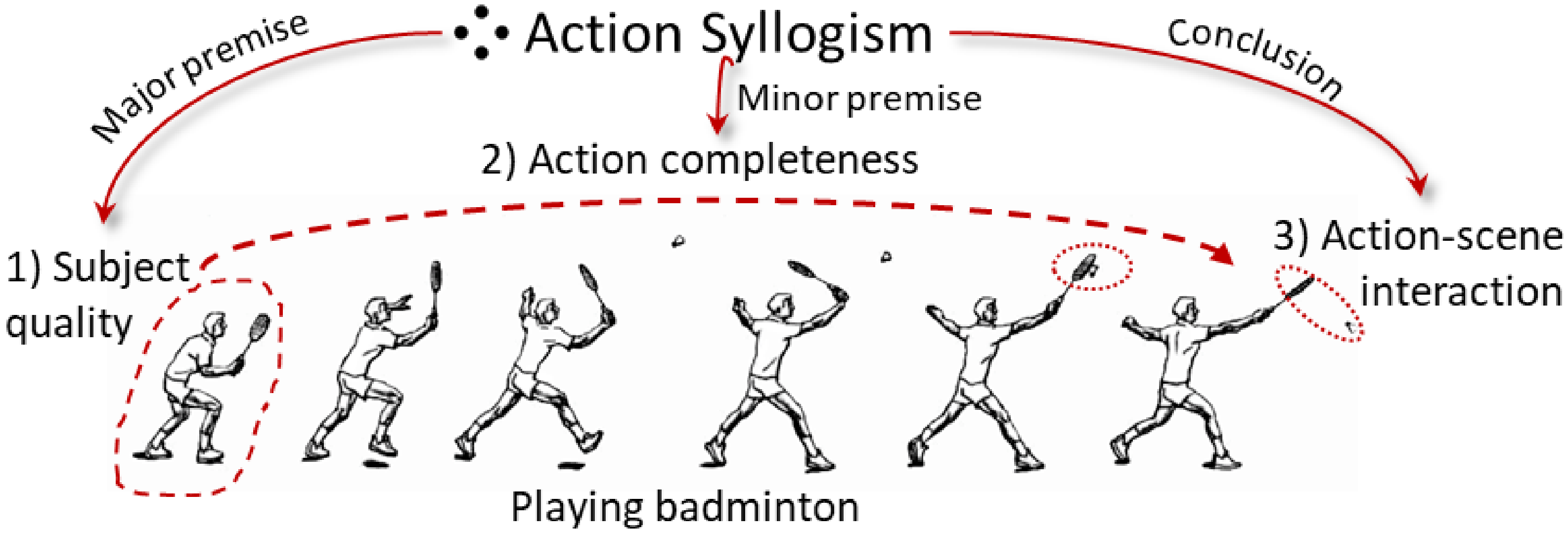

Figure: Illustration of action syllogism.

Timothy J Smiley. What is a syllogism? Journal of philosophical logic, pages 136–154, 1973.

Sangeet Khemlani and Philip N Johnson-Laird. Theories of the syllogism: A meta-analysis. Psychological bulletin, 138(3):427, 2012.

2. Construction of GAIA

➢ Rationale for the Action Syllogism:

- The visibility of the action in videos is greatly affected by the rendering quality of the action subject, which is a crucial element of visual saliency information.
- Unlike parallel-form feedbacks, the order of these three parts in action syllogism inherently aligns with the human reasoning process.

➢ Merits of the Action Syllogism:

- Can more clearly identify and analyze the specific elements that contribute to the perceived quality of the action.
- Inherently aligned with human perception and can help in understanding how different parts of action are perceived by the public, which can lead to insights into what makes AI-generated action convincing or unconvincing.
- Allows for a comparative analysis of AI-generated action against natural human action, revealing where AI excels and where it may need improvement.

2. Construction of GAIA

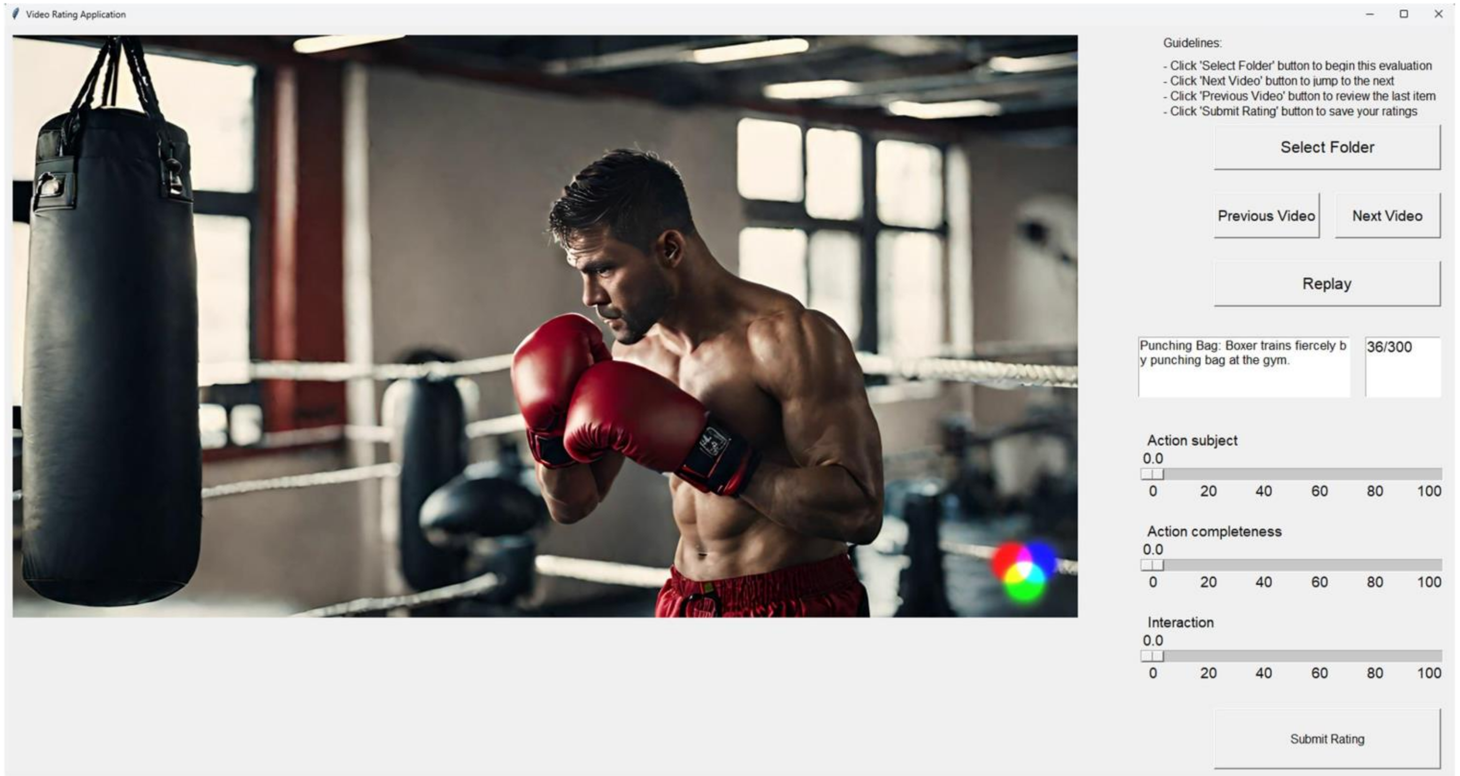

Figure: Screenshot of the rating interface for human evaluation. Participants are instructed to rate three action-related dimensions of AI-generated videos, i.e., subject quality, action completeness, and action-scene interaction, based on the given action keyword and prompt.

Category	Gender		Background		Age
	Male	Female	*w/* AIGC	*w/o* AIGC	
Number	39	15	25	29	23.4±2.6

Table: Statistics of participants. w/AIGC and w/o AIGC denote participants who have or do not have used AI generation tools, respectively.

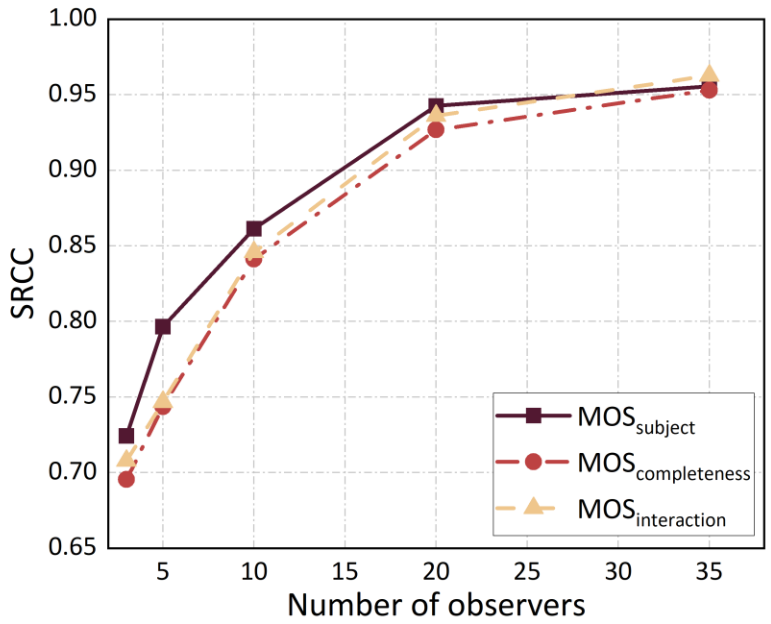

Figure: SRCC between MOSs as the observers increases

3. Observations

1. Most models exhibit ***left-skewed*** MOS distribution in all three dimensions.
2. Additionally, we can observe ***a trend of increasing performance year by year***, from the Text2Video-zero and ModelScope released in March 2023 to the VideoCrafter2 in early 2024.
3. Most models prove decent proficiency on one single dimension, i.e., ***better subject quality than action completeness and action-scene interaction***, which exposes the defects of existing models in producing temporal coherent and complete actions.

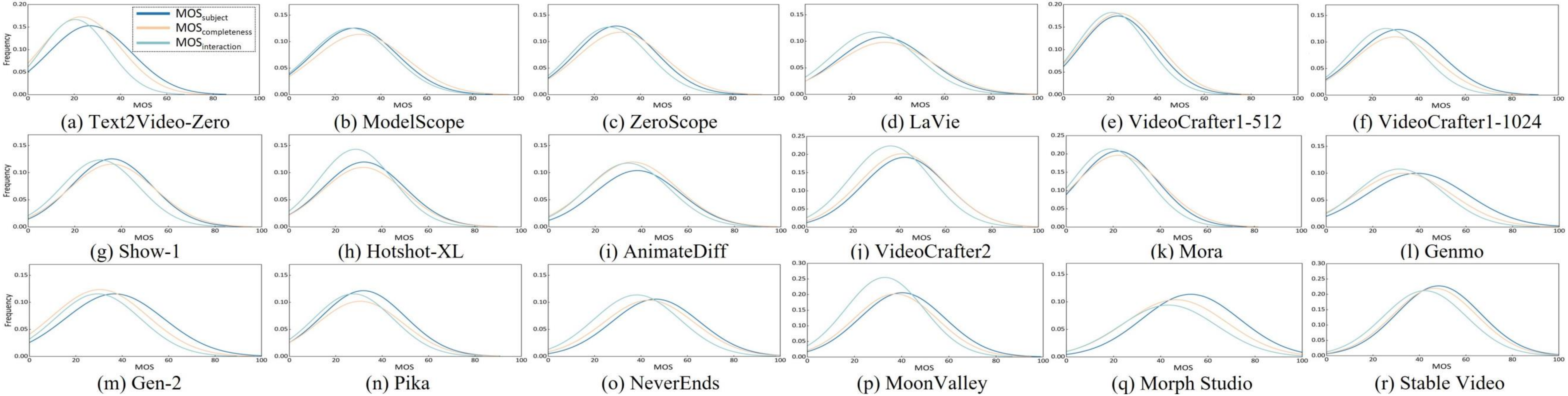

Figure: MOS distributions across different models in terms of subject quality, action completeness, and action-scene interaction. 11 Lab studies: (a)-(k); 7 Commercial applications: (l)-(r).

3. Observations

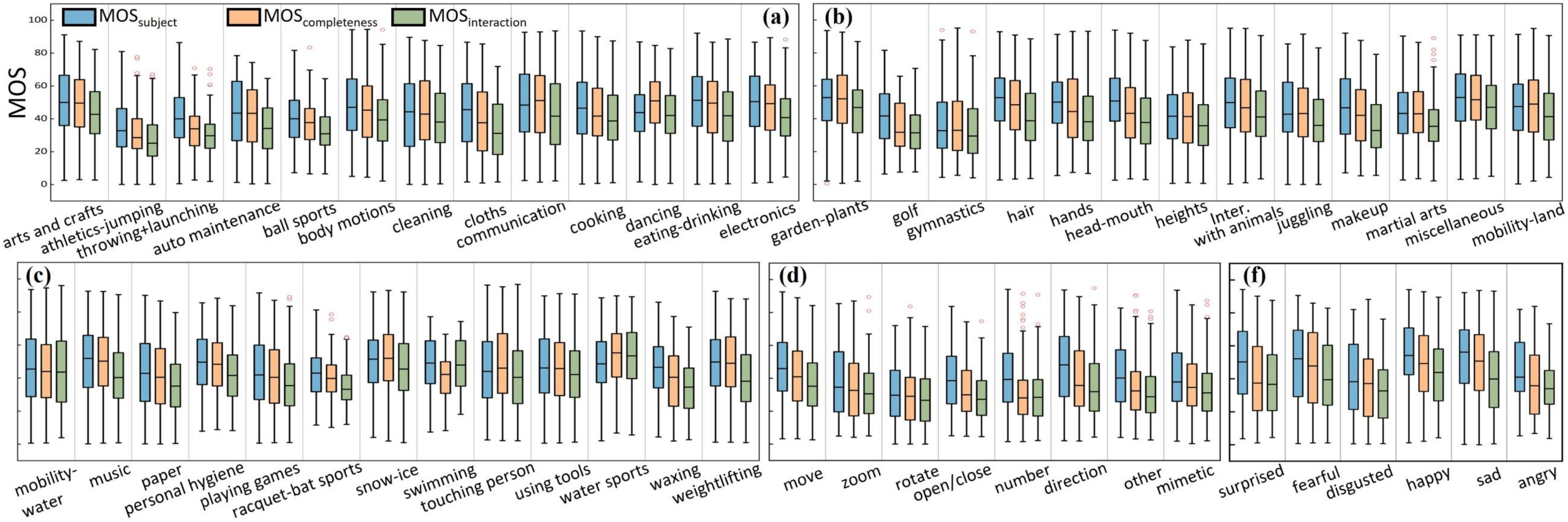

Figure: Box plots of MOSs, MOSc, and MOSi across action categories. (a), (b), and (c) show whole-body actions. (d) and (f) show hand and facial actions. For each box, median is the central box, and the edges of the box represent the 25th and 75th percentiles, while red circles denote outliers.

3. Observations

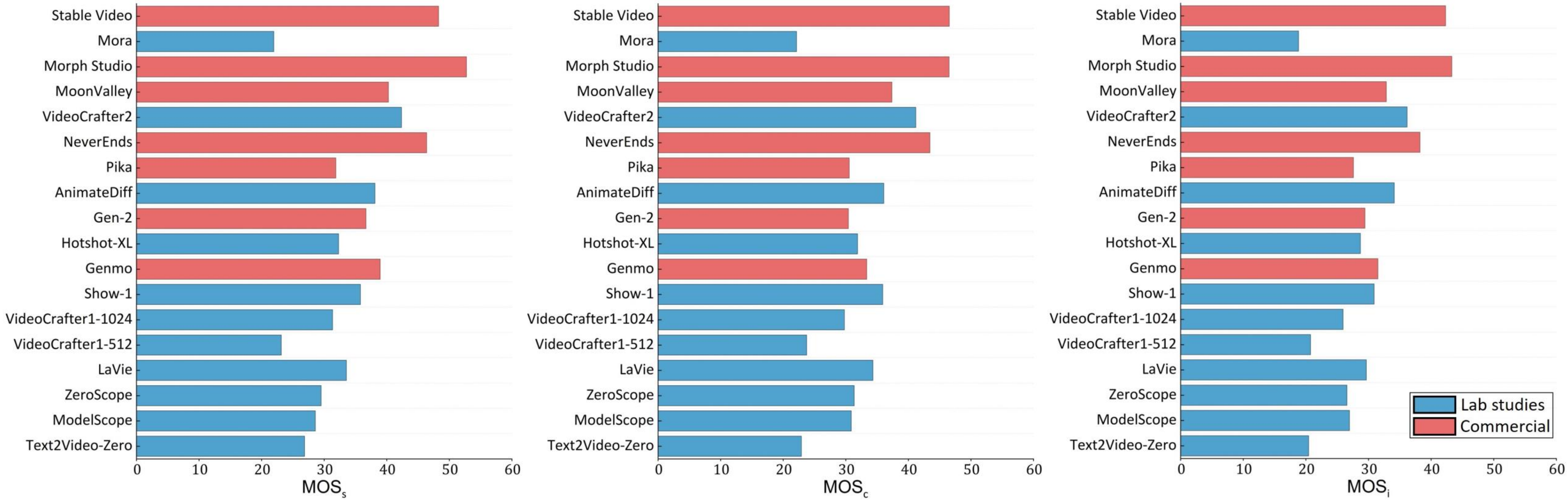

Figure: Detailed model-wise comparison in terms of MOSs, MOSc, MOSi.

3. Observations

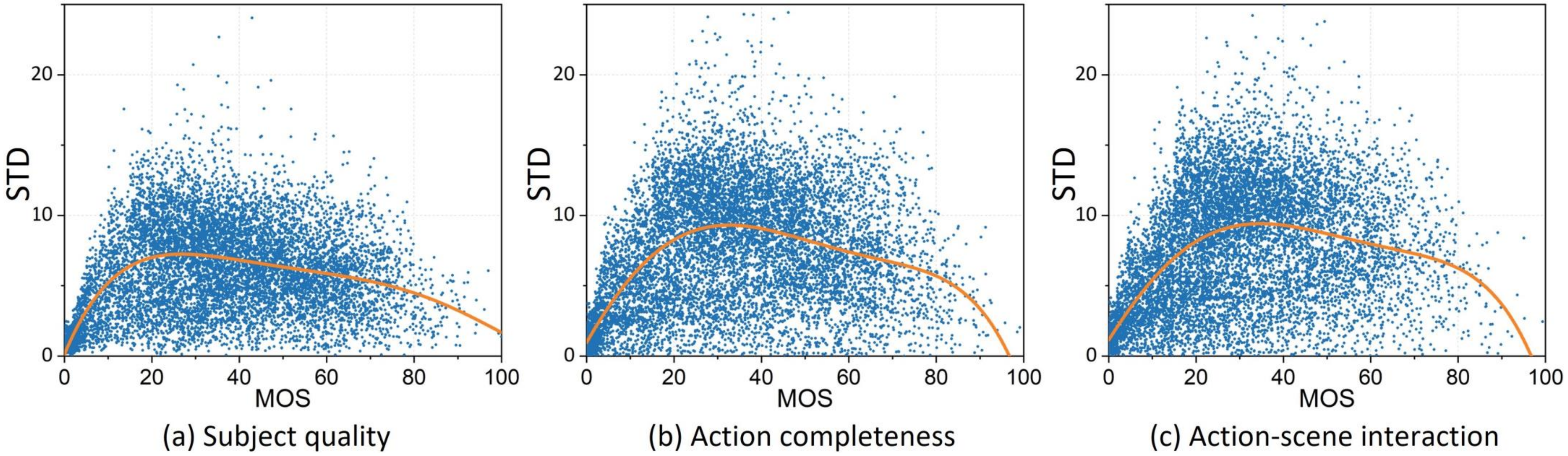

Figure: Scatter plots about MOS against its standard deviation (STD) and five-parameter polynomial fitting plots (orange line) of three perspectives of action quality.

1. Humans are more consistent in perceiving high-quality actions.
2. Medium- and low-quality actions exhibit greater diversity, leading to a more pronounced divergence among individuals.
3. The perception of spatial quality distortion in action is less divergent than the temporal consistency and rationality distortion

4. Experiments

➢ We want to figure out (main results):

- Do conventional AQA methods still work?
- Which action-related metric performs better?
- The performance of video quality assessment (VQA) methods.
- What about the video-text alignment metrics?

4. Experiments

Table: Performance benchmark on GAIA. All-Combined indicates that we sum the MOS of three dimensions and rescale it to [0, 100] as the overall action quality score. ♠, ♣, ◊, and ♡ denote the evaluated **conventional AQA method**, **action-related metrics**, **VQA methods**, and **video-text alignment metrics**, respectively. All experiments for AQA and VQA methods are retrained on each dimension under 10 random train-test splits at a ratio of 8:2.

Dimension		Subject		Completeness		Interaction		*All-Combined*	
Methods / Metrics	**Pre-training/ Initialization**	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
♠USDL (CVPR'20) [97]	Kinetics [14]	0.4197	0.4203	0.4365	0.4517	0.4289	0.4434	0.4223	0.4321
♠ACTION-NET (ACM MM'20) [118]		**0.4533**	**0.4612**	0.4722	0.4765	0.4703	0.4829	0.4587	0.4592
♠CoRe (ICCV'21) [116]		0.4301	0.4343	0.4538	0.4577	0.4521	0.4514	0.4437	0.4415
♠TSA (CVPR'22) [114]		0.4435	0.4477	**0.4963**	**0.4981**	**0.4941**	**0.4953**	**0.4861**	**0.4823**
♣Subject Consistency [50]	DINO [12]	0.2447	0.2362	0.2116	0.2056	0.2034	0.1912	0.2289	0.2273
♣Motion Smoothness [50]	AMT [62]	0.2402	0.1913	0.1474	0.1625	0.1741	0.1693	0.1957	0.1813
♣Dynamic Degree [50]	RAFT [98]	0.1285	0.0831	0.0903	0.0682	0.1141	0.0758	0.1162	0.0787
♣Human Action [50]	UMT [61]	**0.2453**	**0.2369**	**0.2895**	**0.2812**	**0.2861**	**0.2743**	**0.2831**	**0.2741**
♣Action-Score [66]	VideoMAE V2 [106]	0.2023	0.1823	0.2867	0.2623	0.2689	0.2432	0.2600	0.2377
♣Flow-Score [66]	RAFT [98]	0.1471	0.1541	0.0816	0.1273	0.1041	0.1309	0.1166	0.1430
◊TLVQM (TIP'19) [55]	NA *(handcraft)*	0.5037	0.5137	0.4127	0.4158	0.4079	0.4093	0.4655	0.4783
◊VIDEVAL (TIP'21) [101]	NA *(handcraft)*	0.5237	0.5446	0.4283	0.4375	0.4121	0.4234	0.4684	0.4801
◊VSFA (ACM MM'19) [59]	*None*	0.5594	0.5762	0.4940	0.5017	0.4709	0.4811	0.5085	0.5215
◊BVQA (TCSVT'22) [58]	*fused* [24, 35, 14, 49, 32]	0.5702	0.5888	0.4876	0.4946	0.4761	0.4825	0.5201	0.5289
◊SimpleVQA (ACM MM'22) [96]	Kinetics [14]	0.5920	0.5974	0.4981	0.5078	0.4843	0.4971	0.5219	0.5322
◊FAST-VQA (ECCV'22) [111]	Kinetics [14]	0.6015	0.6092	0.5157	0.5215	0.5154	0.5216	0.5276	0.5475
◊DOVER (ICCV'23) [112]	LSVQ [115]	**0.6173**	**0.6301**	**0.5198**	**0.5323**	**0.5164**	**0.5278**	**0.5335**	**0.5502**
♡CLIPScore (ViT-B/16) [43]	OpenAI-400M [80]	0.3360	0.3314	0.3841	0.3777	0.3753	0.3632	0.3777	0.3711
♡CLIPScore (ViT-B/32) [43]	OpenAI-400M [80]	0.3398	0.3330	0.3944	0.3871	0.3875	0.3821	0.3815	0.3826
♡- - *same as the above* - -	LAION-2B [87]	0.3179	0.3101	0.3551	0.3511	0.3504	0.3380	0.3531	0.3458
♡CLIPScore (ViT-L/14) [43]	OpenAI-400M [80]	0.3211	0.3156	0.3657	0.3574	0.3585	0.3426	0.3601	0.3515
♡BLIPScore [60]	COCO [63]	0.3453	0.3386	0.4174	0.4082	0.4044	0.3994	0.4118	0.4054
♡LLaVAScore [64]	LLaVA-PT [25]	0.3484	0.3436	0.4189	0.4133	0.4077	0.4025	0.4124	0.4086
♡InternLMScore [28]	*fused* [63, 17, 10, 90, 88]	**0.3678**	**0.3642**	**0.4324**	**0.4257**	**0.4301**	**0.4227**	**0.4314**	**0.4246**

4. Experiments

- We want to figure out (extended results):
  - Whether CLIP-based metrics excel in assessing action quality?
  - Whether the combination of different metrics can improve the perceptual consistency of action quality?

4. Experiments

Table: Performance comparison on coarse-grained actions (whole-body) and fine-grained actions (hand and facial) from GAIA dataset.

Dimension Metrics	Subset	Subject		Completeness		Interaction	
		SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
CLIPScore (ViT-B/16)	Whole-body	0.3381	0.3293	0.3732	0.3656	0.3698	0.3557
	Hand	0.3167	0.3084	0.3649	0.3564	0.3361	0.3234
	Facial	0.2221	0.2326	0.2307	0.2525	0.2711	0.2861
CLIPScore (ViT-B/32)	Whole-body	0.3848	0.3753	0.4208	0.4128	0.4168	0.4023
	Hand	0.3835	0.3788	0.4159	0.4139	0.3964	0.3910
	Facial	0.1556	0.1596	0.1747	0.1859	0.2175	0.2201
CLIPScore (ViT-L/14)	Whole-body	0.3135	0.3055	0.3499	0.3411	0.3481	0.3301
	Hand	0.3392	0.3269	0.3639	0.3499	0.3373	0.3219
	Facial	0.1743	0.1806	0.1775	0.1927	0.2294	0.2359

4. Experiments

Table: Results for the combination of different metrics on the GAIA dataset.

Dimension	Subject		Completeness		Interaction	
Metrics	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑
Human Action	**0.2453**	**0.2369**	**0.2895**	**0.2812**	**0.2861**	**0.2743**
Action-Score	0.2023	0.1823	0.2867	0.2623	0.2689	0.2432
Flow-Score	0.1471	0.1541	0.0816	0.1273	0.1041	0.1309
Human Action+Action-Score	0.1530	0.1355	0.2333	0.2098	0.2156	0.1912
Human Action+Flow-Score	0.1567	0.1550	0.0940	0.1293	0.1155	0.1324
Action-Score+Flow-Score	0.1199	0.1464	0.0439	0.1175	0.0679	0.1214
Human Action+Action-Score+Flow-Score	0.1279	0.1484	0.0530	0.1198	0.0767	0.1237
VSFA	0.1934	0.1917	0.1379	0.1322	0.1602	0.1658
VSFA+Human Action	0.0836	0.0790	0.0059	0.0142	0.0135	0.0096
VSFA+Action-Score	**0.2599**	**0.2531**	**0.3149**	**0.3046**	**0.3054**	**0.2939**
VSFA+Flow-Score	0.1309	0.1506	0.0714	0.1253	0.0914	0.1283
TSA	0.4435	0.4477	0.4963	0.4981	0.4941	0.4953
DOVER	**0.6173**	**0.6301**	**0.5198**	**0.5323**	**0.5164**	**0.5278**
TSA + DOVER	0.5744	0.5831	0.5068	0.5147	0.5081	0.5158
CLIPScore-B/16	0.3360	0.3314	0.3841	0.3777	0.3753	0.3632
CLIPScore-B/32	0.3398	0.3330	0.3944	0.3871	0.3875	0.3821
CLIPScore-L/14	0.3211	0.3156	0.3657	0.3574	0.3585	0.3426
CLIPScore-B/16+CLIPScore-B/32	0.3746	**0.3698**	**0.4234**	**0.4172**	**0.4148**	**0.4028**
CLIPScore-B/16+CLIPScore-L/14	0.3479	0.3428	0.3967	0.3893	0.3878	0.3738
CLIPScore-B/32+CLIPScore-L/14	**0.3747**	0.3687	0.4218	0.4145	0.4140	0.3998
CLIPScore-B/16+CLIPScore-B/32+CLIPScore-L/14	0.3734	0.3681	0.4227	0.4157	0.4140	0.4006
VSFA+CLIPScore-B/16	0.3782	0.3733	0.4014	0.3990	0.3984	0.3906
VSFA+CLIPScore-B/32	**0.4162**	**0.4120**	**0.4377**	**0.4355**	**0.4364**	**0.4288**
VSFA+CLIPScore-L/14	0.3651	0.3582	0.3826	0.3793	0.3821	0.3709
VSFA+CLIPScore-B/16+CLIPScore-B/32	0.4004	0.3938	0.4361	0.4303	0.4308	0.4192
CLIPScore-B/16+CLIPScore-B/32+Human Action	0.3585	0.3581	0.4041	0.4027	0.3960	0.3885

THANKS