Supplementary Materials for GAIA: Rethinking Action Quality Assessment for AI-Generated Videos

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
[*]Corresponding authors

1 Dataset Documentation

Here we provide documentation for our dataset in the common datasheet format [11].

1.1 Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled?

We produced the dataset to fill the gaps of current research on action quality assessment (AQA) in AI-generated videos (AIGV) and evaluate the existing automatic action-quality and video-quality metrics. To the best of our knowledge, it is the first AQA dataset that evaluates action quality from causal reasoning-based perspectives. Investigations of the action quality assessment commonly conduct for specific-domain in real-world scenarios (*e.g.*, sport events and health care). Our goal is to create an AQA dataset that focuses on AI-generated scenarios.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, or organization)?

The dataset was a joint effort by 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], and Wenjun Zhang[1]. The authors[1] are researchers affiliated with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University. The author[2] is with the School of Information Science & Engineering, East China University of Science and Technology.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was supported in part by the China Postdoctoral Science Foundation under Grant Number 2023TQ0212 and 2023M742298, in part by the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20231618, in part by the Shanghai Pujiang Program under Grant 22PJ1407400, and in part by the National Natural Science Foundation of China under Grant 62271312, 62301316, 62101325 and 62101326.

1.2 Composition

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)?

The instances represent video files with corresponding subjective action quality scores. Each video was generated by a text-to-video model using a prompt that contained the action keyword.

How many instances are there in total (of each type, if appropriate)?

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

The dataset has a total of 9,180 AI-generated videos with 971,244 reliable ratings (an average of 105.8 ratings per video (35.27 per dimension)).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)?

The dataset contains all generated instances.

What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features?

Each instance is a video file in `.mp4` format. Each subjective score is a floating-point number. Each action keyword and the corresponding prompt are in text format.

Is there a label or target associated with each instance?

Each video has three subjective scores that represent the evaluated three action quality-related perspectives (*i.e.*, *subject quality*, *action completeness*, and *action-scene interaction*) with the corresponding prompt and action keyword.

Is any information missing from individual instances?

No.

Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)?

We divided the action videos into groups on the basis of their action categories and their respective text-to-video models.

Are there recommended data splits (for example, training, development/validation, testing)?

No. In the process of benchmarking the current action quality assessment metrics, we randomly split the training-testing set at a ratio of 8:2 ten times.

Are there any errors, sources of noise, or redundancies in the dataset?

No

Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?

The dataset is self-contained.

Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?

No. All videos in our GAIA dataset are generated by text-to-video models according to the given action prompt.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

No. We have organized experts to manually check the compliance and legality of the adopted action prompts as well as the generated videos to avoid such issues.

1.3 Collection Process

How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified?

We formed the dataset using 18 popular text-to-video models including 11 lab-studies and 7 commercial applications, listed in Tab 1.

Table 1: URLs for the adopted text-to-video models.

Methods	URL
Text2Video-Zero [14]	https://github.com/Picsart-AI-Research/Text2Video-Zero
ModelScope [17]	https://modelscope.cn/models/iic/text-to-video-synthesis/summary
ZeroScope [8]	https://huggingface.co/cerspense/zeroscope_v2_576w
LaVie [18]	https://github.com/Vchitect/LaVie
Show-1 [20]	https://github.com/showlab/Show-1
Hotshot-XL [15]	https://github.com/hotshotco/Hotshot-XL
AnimateDiff [12]	https://github.com/guoyww/AnimateDiff
VideoCrafter1-512 [9]	https://github.com/AILab-CVC/VideoCrafter
VideoCrafter1-1024 [9]	https://github.com/AILab-CVC/VideoCrafter
VideoCrafter2 [10]	https://github.com/AILab-CVC/VideoCrafter
Mora [19]	https://github.com/lichao-sun/Mora
Gen-2 [1]	https://research.runwayml.com/gen2
Genmo [2]	https://www.genmo.ai
Pika [6]	https://pika.art/home
NeverEnds [5]	https://neverends.life
MoonValley [3]	https://moonvalley.ai
Morph Studio [4]	https://www.morphstudio.com
Stable Video [7]	https://www.stablevideo.com/welcome

Content types: 400 whole-body actions, 83 hand actions, and 27 facial actions. Detailed category of each action keyword in our GAIA dataset are listed in Tab. 2, Tab. 3, Tab. 4, and Tab. 5

What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

The subjective experiments (labeling) involved a training/pre-labeling and a main study processes. During the training, we instructed all participants to have a clear and consistent understanding of all evaluated aspects and tested their eligibility via a 30-video pre-labeling. In the tutorial for each dimension, participants are guided to rate 10 generated-real video pairs with the same caption. Their answer is compared with ground-truth ratings that were developed by multiple experts. Raters needed to achieve at least 75% ratings that satisfied $|ground\_truth - rating| < 1.5\sigma_{expert}$ to move on to the formal study. We adopted a single-stimulus methodology in this evaluation and asked participants to focus on the given action keyword as well as the corresponding prompt and evaluate three action-related dimensions of AI-generated videos, *i.e.*, *subject quality*, *action completeness*, and *action-scene interaction*, by dragging the slide button at a $[0, 100]$ continuous rating scale (see Fig. 1). In addition to the above pre-labeling and in-process check trial, we also performed line clickers examination and calculated the inter-annotator agreement metric (Krippendorff's α [13]) as well as the SRCC score using bootstrapping to ensure the annotation quality.

If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

Not applicable.

Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

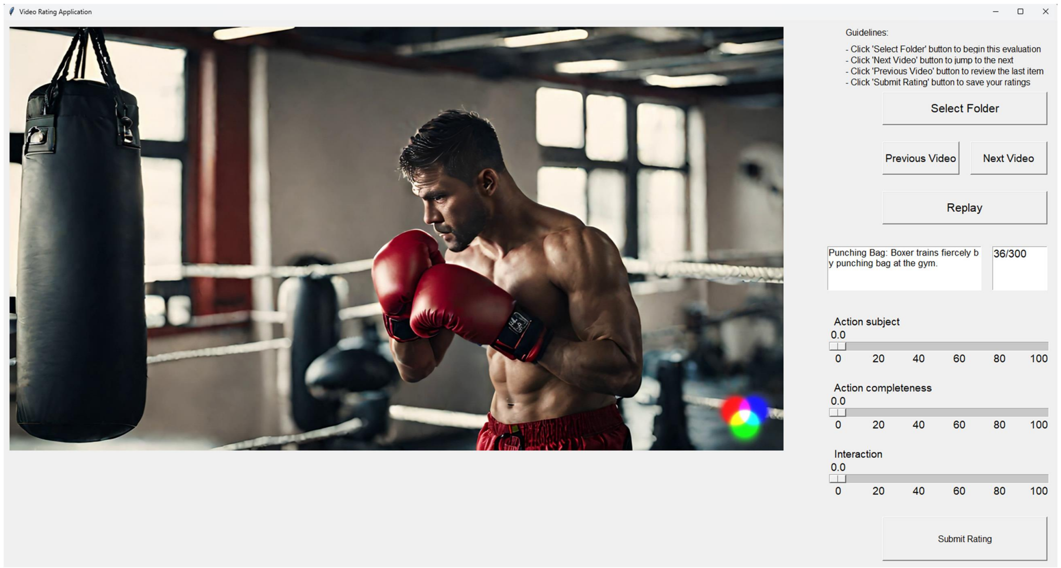
We recruit a total of 54 participants to participate in our subjective experiments. Each participant was compensated $12 for each session (rating 300 videos in three dimensions) according to the current ethical standard [16].

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)?

The collection of source content in GAIA dataset took place nearly one years (from March 2023 to March 2024). The timeframe is align with the released date of the selected text-to-video models. The subjective experiments took over a month (from March 2024 to April 2024) to complete.

Were any ethical review processes conducted (for example, by an institutional review board)?

We have organized experts to manually check the compliance and legality of the adopted action prompts as well as the generated videos to avoid ethical issues.



Guidelines:

- Click 'Select Folder' button to begin this evaluation
- Click 'Next Video' button to jump to the next
- Click 'Previous Video' button to review the last item
- Click 'Submit Rating' button to save your ratings

Punching Bag: Boxer trains fiercely b
y punching bag at the gym.

36/300

Action subject
0.0

Action completeness
0.0

Interaction
0.0

Figure 1: **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.

1.4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

We directly use the generated videos from text-to-video models as source content without any change. For the collected raw user-rated scores, we performed Z-score normalization to avoid inter-annotator scoring biases.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)?

Yes, we saved the raw user-rated scores for all dimensions.

Is the software that was used to preprocess/clean/label the data available?

We conducted the subjective experiments in-lab. All data was labeled through a self-designed program based on Python Tkinter (Fig. 1).

1.5 Uses

Has the dataset been used for any tasks already?

The dataset has served in two separate tasks:

- Measurement of action-quality metrics for AI-generated videos along with calculation of the correlation between the predicted scores and subjective scores.
- Comparing the performance of different video generation models in generating rational, artifact-free, and temporally consistent actions.

Is there a repository that links to any or all papers or systems that use the dataset?

The benchmark for action quality assessment in AI-generated videos is available through https://github.com/zijianchen98/GAIA.

What (other) tasks could the dataset be used for?

The dataset can be used to train or validate a new action quality assessment metric and determine whether it can reliably meet most action scenarios.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? Is there anything a dataset consumer could do to mitigate these risks or harms?

We applied a novel causal reasoning-based subjective evaluation strategy to collect opinions for assessing action quality. Consumers are recommended not to use the subjective score of a single dimension to compare the action quality since it is not fair.

Are there tasks for which the dataset should not be used?

No.

1.6 Distribution

Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created?

The GAIA dataset is available to everyone.

How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The GAIA dataset is accessible through `https://github.com/zijianchen98/GAIA`.

When will the dataset be distributed?

We are now preparing to release the dataset. All information will be noted at a Github repository `https://github.com/zijianchen98/GAIA`.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The **GAIA** dataset is released under the **CC BY 4.0** license, which includes all associated AIGVs, scores, and their corresponding action prompts.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?

No.

1.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

The Shanghai Jiao Tong University MultiMedia Lab hosts the dataset. The authors of this paper support the proposed GAIA dataset and the benchmark evaluated in this paper.

How can the owner/curator/manager of the dataset be contacted (for example, email address)?

Contact the first author of this paper Mr. Zijian Chen by `zijian.chen@sjtu.edu.cn`, or the corresponding author Prof. Guangtao Zhai by `zhaiguangtao@sjtu.edu.cn`.

Is there an erratum?

No.

Table 2: Categories of the 400 whole-body actions in our proposed GAIA.

Class	Action Keyword			
Arts and crafts	arranging flowers	blowing glass	brush painting	clay pottery making
	drawing	knitting	making jewelry	spray painting
	weaving basket			
Athletics – jumping	high jump	hurdling	long jump	parkour
	pole vault	triple jump		
Athletics – throwing + launching	archery	catching or throwing frisbee	disc golfing	hammer throw
	javelin throw	throwing axe	throwing ball	throwing discus
Auto maintenance	changing oil	changing wheel	checking tires	pumping gas
Ball sports	bowling	dodgeball	dribbling basketball	dunking basketball
	kicking field goal	kicking soccer ball	passing American football (in game)	passing American football (not in game)
	playing basketball	playing kickball	playing volleyball	shooting basketball
	shooting goal (soccer)	shot put		
Body motions	baby waking up	bending back	cracking neck	stretching arm
	stretching leg	swinging legs	exercising arm	exercising with an exercise ball
	lunge			
Cleaning	cleaning floor	cleaning gutters	cleaning pool	cleaning shoes
	cleaning toilet	cleaning windows	doing laundry	making bed
	mopping floor	setting table	shining shoes	sweeping floor
	washing dishes			
Cloths	bandaging	folding clothes	ironing	tying bow tie
	tying knot (not on a tie)	tying tie		
Communication	answering questions	auctioning	celebrating	crying
	giving or receiving award	laughing	news anchoring	presenting weather forecast
	sign language interpreting	testifying		
Cooking	baking cookies	barbequing	breading or breadcrumbing	cooking chicken
	cooking egg	cooking on campfire	cooking sausages	cutting pineapple
	cutting watermelon	flipping pancake	frying vegetables	grinding meat
	making a cake	making a sandwich	making pizza	making sushi
	making tea	peeling apples	peeling potatoes	picking fruit
	scrambling eggs	tossing salad		
Dancing	belly dancing	breakdancing	capoeira	cheerleading
	country line dancing	dancing ballet	dancing charleston	dancing gangnam style
	dancing macarena	jumpstyle dancing	krumping	marching
	robot dancing	salsa dancing	swing dancing	tango dancing
	tap dancing	zumba		
Eating + drinking	bartending	dining	drinking	drinking beer
	drinking shots	eating burger	eating cake	eating carrots
	eating chips	eating doughnuts	eating hotdog	eating ice cream
	eating spaghetti	eating watermelon	opening bottle	tasting beer
	tasting food			
Electronics	assembling computer	playing controller	texting	using computer
	using remote controller (not gaming)			
Garden + plants	blowing leaves	carving pumpkin	chopping wood	climbing tree
	decorating the christmas tree	egg hunting	mowing lawn	planting trees
	trimming trees	watering plants		
Golf	golf chipping	golf driving	golf putting	
Gymnastics	bouncing on trampoline	cartwheeling	gymnastics tumbling	somersaulting
	vault	bench pressing	doing aerobics	situp
	yoga			
Hair	braiding hair	brushing hair	curling hair	fixing hair
	getting a haircut	shaving head	shaving legs	washing hair
Hands	air drumming	applauding	clapping	cutting nails
	finger snapping	pumping fist	drumming fingers	
Head + mouth	balloon blowing	beatboxing	blowing nose	blowing out candles
	headbanging	headbutting	shaking head	singing
	smoking	smoking hookah	sneezing	sniffing
	sticking tongue out	whistling	yawning	gargling
Heights	abseiling	bungee jumping	climbing a rope	climbing ladder
	paragliding	rock climbing	skydiving	slacklining
	swinging on something	trapezing		
Interacting with animals	bee keeping	catching fish	feeding birds	feeding fish
	feeding goats	grooming dog	grooming horse	holding snake
	milking cow	petting animal (not cat)	petting cat	riding camel
	riding elephant	riding mule	riding or walking with horse	shearing sheep
	training dog	walking the dog		
Juggling	contact juggling	hula hooping	juggling balls	juggling fire
	juggling soccer ball	spinning poi		
Makeup	applying cream	doing nails	dying hair	filling eyebrows
	getting a tattoo			
Martial arts	arm wrestling	drop kicking	high kick	punching bag
	punching person	side kick	sword fighting	tai chi
	wrestling			
Miscellaneous	digging	extinguishing fire	garbage collecting	laying bricks
	moving furniture	spraying	stomping grapes	tapping pen
	unloading truck			
Mobility – land	crawling baby	driving car	driving tractor	faceplanting
	hoverboarding	jogging	motorcycling	pushing car
	pushing cart	pushing wheelchair	riding a bike	riding mountain bike
	riding scooter	riding unicycle	roller skating	running on treadmill
	skateboarding	surfing crowd	using segway	waiting in line
Mobility – water	crossing river	diving cliff	jumping into pool	scuba diving
	snorkeling	springboard diving	water sliding	
Music	busking	playing accordion	playing bagpipes	playing bass guitar
	playing cello	playing clarinet	playing cymbals	playing didgeridoo
	playing drums	playing flute	playing guitar	playing harmonica
	playing harp	playing keyboard	playing organ	playing piano
	playing recorder	playing saxophone	playing trombone	playing trumpet
	playing ukulele	playing violin	playing xylophone	recording music
	strumming guitar	tapping guitar		
Paper	bookbinding	counting money	folding napkins	folding paper
	opening present	reading book	reading newspaper	ripping paper
	shredding paper	unboxing	wrapping present	writing
Personal hygiene	brushing teeth	taking a shower	trimming or shaving beard	washing feet
	washing hands			

Table 3: Extension of Tab. 2.

Class	Action Keyword			
Playing games	flying kite playing paintball skipping rope	hopscotch playing poker tossing coin	playing cards riding mechanical bull playing monopoly	playing chess rock scissors paper shuffling cards
Racquet + bat sports	catching or throwing baseball playing badminton	catching or throwing softball playing cricket	hitting baseball playing squash or racquetball	hurling (sport) playing tennis
Snow + ice	biking through snow ice fishing shoveling snow skiing slalom snowmobiling	bobsledding ice skating ski jumping sled dog racing tobogganing	hockey stop making snowman skiing (not slalom or crosscountry) snowboarding	ice climbing playing ice hockey skiing crosscountry snowkiting
Swimming	swimming backstroke	swimming breast stroke	swimming butterfly stroke	
Touching person	carrying baby massaging feet slapping	hugging massaging legs tickling	kissing massaging person's head	massaging back shaking hands
Using tools	bending metal plastering welding	blasting sand sanding floor	building cabinet sharpening knives	building shed sharpening pencil
Water sports	canoeing or kayaking sailing	jetskiing surfing water	kitesurfing water skiing	parasailing windsurfing
Waxing	waxing back	waxing chest	waxing eyebrows	waxing legs
Weightlifting	pull ups front raises	push up snatch weight lifting	clean and jerk squat	deadlifting

Table 4: Categories of the 83 hand actions in our proposed GAIA.

Class	Action Keyword			
Move	Wave palm towards right Wave palm forward Move fist upward Move palm backward Move palm towards left Move fingers toward left	Wave palm towards left Wave palm backward Move fist downward Move palm forward Move palm towards right Move fingers toward right	Wave palm downward Wave finger towards left Move fist towards left Move palm upward Move fingers upward Move fingers forward	Wave palm upward Wave finger towards right Move fist towards right Move palm downward Move fingers downward
Zoom	Zoom in with two fists	Zoom out with two fists	Zoom in with two fingers	Zoom out with two fingers
Rotate	Rotate fists clockwise	Rotate fists counter-clockwise	Rotate fingers clockwise	Rotate fingers counter-clockwise
Open/close	Turn over palm Put two fingers together	Rotate with palm Take two fingers apart	Palm to fist	Fist to Palm
Number	Number 0 Number 4 Number 8	Number 1 Number 5 Number 9	Number 2 Number 6 Another number 3	Number 3 Number 7
Direction	Thumb upward Thumbs backward	Thumb downward Thumbs forward	Thumb towards right	Thumb towards left
Others	Cross index fingers OK Dual hands heart	Sweep cross Pause Bent two fingers	Sweep checkmark Shape C Bent three fingers	Static fist Hold fist in the other hand Dual fingers heart
Mimetic	Click with index finger take a picture Knock Grab (bend all five fingers) Applaud	Sweep diagonal Make a phone call Beckon Walk	Measure (distance) Wave hand Trigger with thumb Gather fingers	Sweep circle Wave finger Trigger with index finger Snap fingers
Surprised	curiosity surprise	desire	approval	realization
Fearful	confusion caring	fear	nervousness	relief
Disgusted	disgust	embarrassment		
Happy	amusement optimism	love pride	joy admiration	excitement gratitude
Sad	disappointment sadness	disapproval	grief	remorse
Angry	anger	annoyance		

Table 5: Categories of the 27 facial actions in our proposed GAIA.

Class	**Action Keyword**				
Surprised	curiosity	desire	approval	realization	surprise
Fearful	confusion	fear	nervousness	relief	caring
Disgusted	disgust	embarrassment			
Happy	amusement pride	love admiration	joy gratitude	excitement	optimism
Sad	disappointment	disapproval	grief	remorse	sadness
Angry	anger	annoyance			

Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

We plan to extend the dataset as the generative model evolves to ensure benchmark results with the highest statistical credibility and the latest observations. Such updates will be rare, as they involve subjective evaluation, a time-consuming task that requires extensive preparation. This does not affect the user's use of the current dataset. All the new information will be on the dataset website https://github.com/zijianchen98/GAIA.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)?

Not applicable.

Will older versions of the dataset continue to be supported/hosted/maintained?

We do not intend to create a version history.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? Will these contributions be validated/verified? If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers?

We encourage future researchers to share their ideas on extending our dataset to cover more action cases and model variants while providing more reliable results. We recommend people contacting us by zijian.chen@sjtu.edu.cn to collaborate on the subjective quality evaluation. It is recommended to build a more large-scale and comprehensive dataset based on our dataset.

References

[1] Gen-2. https://research.runwayml.com/gen2. Accessed: 2024-03-20.

[2] Genmo. https://www.genmo.ai/. Accessed: 2024-03-20.

[3] Moonvalley. https://moonvalley.ai/. Accessed: 2024-03-20.

[4] Morph. https://www.morphstudio.com. Accessed: 2024-03-21.

[5] Neverends. https://neverends.life. Accessed: 2024-03-21.

[6] Pika. https://pika.art/home. Accessed: 2024-03-20.

[7] Stable video. https://www.stablevideo.com. Accessed: 2024-03-21.

[8] Zeroscope-v2-576w. https://huggingface.co/cerspense/zeroscope_v2_576w. Accessed: 2024-03-20.

[9] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

[13] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

[14] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[15] John Mullan, Duncan Crawbuck, and Aakash Sastry. Hotshot-XL. https://github.com/hotshotco/hotshot-xl, October 2023.

[16] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. Responsible research with crowds: pay crowdworkers at least minimum wage. *Communications of the ACM*, 61(3):39–41, 2018.

[17] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[18] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[19] Zhengqing Yuan, Ruoxi Chen, Zhaoxu Li, Haolong Jia, Lifang He, Chi Wang, and Lichao Sun. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024.

[20] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.