CG-BENCH: CLUE-GROUNDED QUESTION ANSWERING BENCHMARK FOR LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

Contents

1 EXPERIMENTAL ANALYSIS

2 ANNOTATION

2.1 QUALITY CONTROL

During the annotation process, we implement a quality control system as illustrated in Figure 1. We use a batch increment method for data iteration, reviewing each batch of about 1,000 items.

First, a manual review checks for typos and ensures question quality. We focus on two main aspects: clarity and granularity. Questions must have a clear anchor point, such as an event or scene, to avoid confusion. The granularity should be appropriate; overly broad questions provide too many easy clues, which undermines our goal of testing the model's ability to pinpoint clues.

Next, to ensure question difficulty, we conduct tests using LLM, such as GPT4 OpenAI (2023) and Qwen2.5 Yang et al. (2024), with pure text and small MLLM, like InternVL2-2B Chen et al. (2024) and InternVL2-4B, with sparse frames. The pure text test ensures that questions and options don't reveal too much information, allowing models to answer without visual data.

Finally, the second manual review catches other remaining issues, resulting in the final test set.

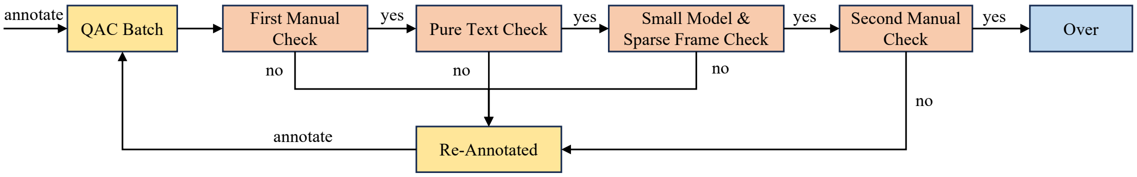


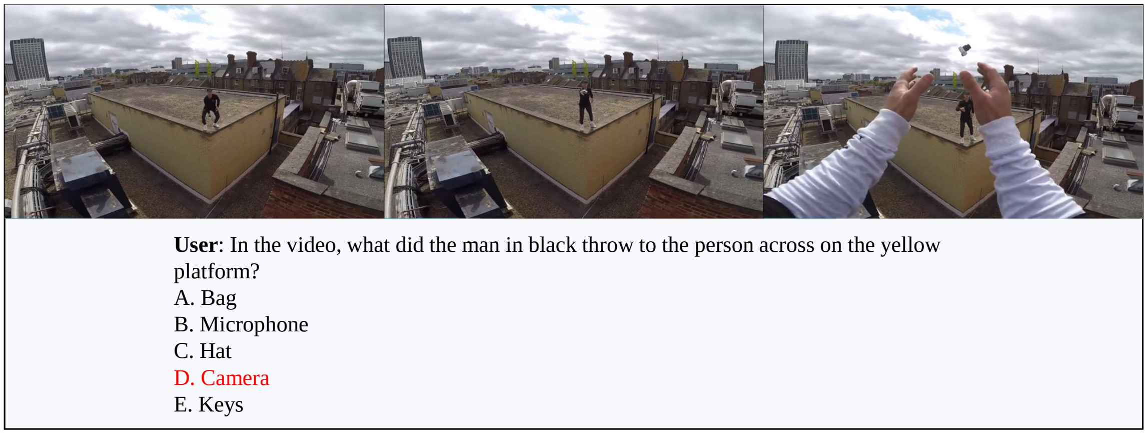
Figure 1: Annotation Quality Control Flowchart.

2.2 Statistics

Word Cloud.

2.3 QA Examples

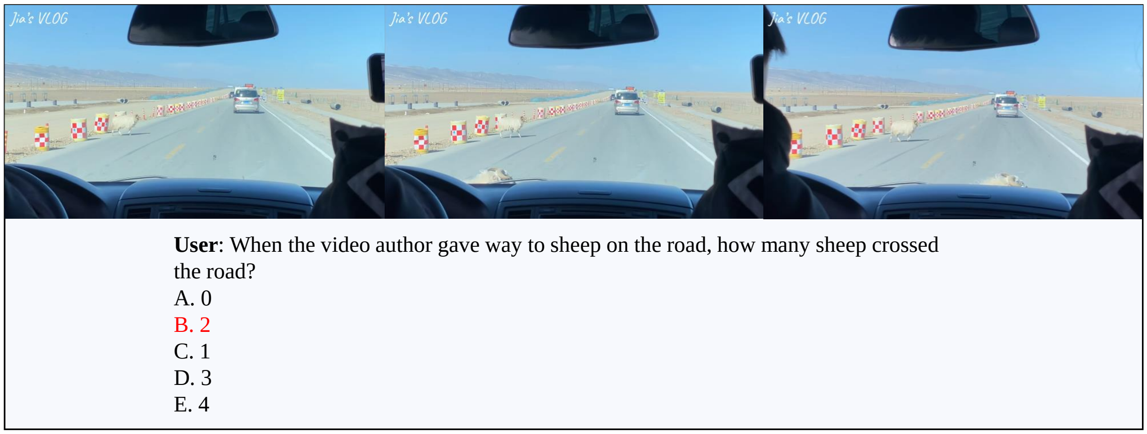
We list some QA examples in Figures 2 and 3.



User: In the video, what did the man in black throw to the person across on the yellow platform?

A. Bag
B. Microphone
C. Hat
D. Camera
E. Keys

Figure 2: An example of QA in CG-Bench of a first-person Parkour video.



User: When the video author gave way to sheep on the road, how many sheep crossed the road?

A. 0
B. 2
C. 1
D. 3
E. 4

Figure 3: An example of QA in CG-Bench of a how-to video.

2.4 QAC Examples

3 Model Inference and Evaluation

In this section, we list the prompt we use in inference and evaluating existing models.

3.1 Common Prompts

Subtitle Prompt (Denoted as <Sub>):

```
The subtitles of the video are as follows:
<Subtitles>
```

Subtitle Time Prompt

```
<Subtitle> -> [start, end]: <Subtitle> (Optional)
```

Frame Time Prompt (Denoted as <FT>)

```
A total of <n> frames are uniformly sampled from the video, and their
corresponding timestamps are <frame_time1>, <frame_time2>, ..., <frame_timen>
```

Choices Prompt (Denoted as <Choices>)

```
A. ChoiceA
B. ChoiceB
...
E/H. ChoiceE/ChoiceH (5˜8 choices)
```

3.2 Inference Pormpts

Long-Video-MCQ & Clue-based-MCQ

```
Task description:

You will watch a video and read a multiple-choice question based on the video
content. You need to choose an answer that best matches the video content
from five to eight
options.

<Frame1>, <Frame2>, ..., <Framen>
<Sub> (Optional)
<FT> (Optional)

Multiple-choice question:

<Question>
<Choices>

Important:
- You must only output the uppercase letter corresponding to the correct
answer.
- Do not include any additional text, punctuation, or explanations in your
response.
```

```
Your output is:
```

Blind-MCQ

```
Task description:

You will be read a multiple-choice question related to a visual task.
However, no visual context or information will be given. Please do your best
to answer the question based solely on the textual information. Choose the
most likely answer from the given options, even if the question appears to
require visual input.

Multiple-choice question:

<Question>
<Choices>

Important:

- You must only output the uppercase letter corresponding to the correct
answer.
- Do not include any additional text, punctuation, or explanations in your
response.

Your output is:
```

Question-Clue Grounding

```
Task description:

You will watch a video and read a multiple-choice question based on the video
content. You need to output each clue interval that can answer this question
in a nested list format.

<Frame1>, <Frame2>, ..., <Framen>
<Sub> (Optional)
<FT> (Optional)

Multiple-choice question:

<Question>

<Choices>

Important:

- The output must strictly follow the format: [[start1, end1], [start2,
end2], ...]
where start and end are the timestamps in seconds.
- Any output that does not conform to this nested array format will be
considered incorrect.

Your output is:
```

Open-Ended QA

```
Task description:

You will watch a video and read a question based on the video content. Please
answer this question directly based on the frames sampled from the video.

<Frame1>, <Frame2>, ..., <Framen>
<Sub> (Option)
<FT> (Option)

Question:
<Question>

Important:
- You must provide an answer. If explicit clues are lacking, make an
inference. Do your best based on the given frames.
- Failure to provide an inferred answer will be considered incorrect.

Your output is:
```

3.3 Evaluation Prompts

Heuristic Evaluation Method for Open-ended QA: Step 1

```
Task Description:

You are a judge. You will read a question, a model’s prediction, and the
ground truth answer to this question. You need to judge whether the model’s
prediction is correct. In most cases, this judgment can be made by
determining whether the meaning of the two texts is consistent. That is, if
the meaning of the model’s prediction is consistent with the meaning of the
ground truth answer, the prediction is considered correct; otherwise, it is
considered incorrect. However, there are some special cases among the
incorrect ones, where inconsistencies may just focus on different details of
the same visual scene and don’t have fundamental differences. In this case,
the problem cannot be judged only by text, and additional visual information
needs to be introduced.

Therefore, I hope you:
Output "yes" if the meaning of the two texts of the model’s prediction and
the ground truth answer is consistent.
Output "no" if the model’s prediction and the ground truth answer are not
consistent, and their meanings are fundamentally different.
Output "need visual clue" if the model’s prediction and the ground truth
answer are not consistent but the model’s prediction does not appear to be
fundamentally different from the ground truth answer. It is possible that the
two focus on different details of the same visual scene. Visual information
is needed for further judgment. You are required to give an explanation as to
why they might focus on different details.

Question:
<Question>
```

```
The ground truth answer is: "<Answer>"
The model’s prediction is: "<Prediction>"

Important:

- The "model’s prediction" has already been made based on visual information.
So "need visual clue" means that you need visual information to make the next
judgment, not that the model needs it.
- The "ground truth answer" is annotated by a human, so it is ABSOLUTELY
RIGHT.
Therefore, for relatively simple problems such as counting, if the model’s
prediction is different from the ground truth, just output "no" directly and
don’t need additional visual information. The only difference between the
"ground truth answer" and the
"model’s prediction" that requires further judgment based on visual
information is maybe the different details of the same visual scene they focus
on.

Your output is:
```

Heuristic Evaluation Method for Open-ended QA: Step 2

```
Task description:

You are a judge. You will read a question, a model’s prediction, and the
sampling frames of the clue intervals of this question. You need to determine
whether the model answered the question correctly based on the visual
information.
I hope you:
- Output "yes", if the model’s prediction answers this question correctly.
- Output "no", if the model’s prediction doesn’t answer this question
correctly.

Question:
<Question>

The model’s prediction is: "<Prediction>"

<Frame1>, <Frame2>, ..., <Framen>
<Sub> (Option)
<FT> (Option)

Your output is:
```

Pure Text Evaluation Method for Open-ended QA

```
Task description:

You are a judge. You will read a question, a model’s prediction and the
ground truth answer to this question. You need to determine whether the model
answered the question correctly.
I hope you:
- Output "yes", if the model’s prediction answers this question correctly.
- Output "no", if the model’s prediction doesn’t answer this question
```

```
correctly.

Question:
<Question>

The ground truth answer is: "<Answer>"
The model’s prediction is: "<Prediction>"
Your output is:
```

Full Vision-aided Evaluation Method for Open-ended QA: With Ground Truth Answer

```
Task description:

You are a judge. You will read a question, a model’s prediction, the ground
truth answer to this question, and the sampling frames of the clue intervals
of this question. You need to judge whether the model has answered the
question correctly based on the sampling frames of the clue intervals.

<Frame1>, <Frame2>, ..., <Framen>
<Sub> (Option)
<FT> (Option)

Question:
<Question>

The ground truth answer is: "<Answer>"
The model’s prediction is: "<Prediction>"
Your output is:
```

Full Vision-aided Evaluation Method for Open-ended QA: Without Ground Truth Answer

```
Task description:

You are a judge. You will read a question, a model’s prediction, and the
sampling frames of the clue intervals of this question. You need to judge
whether the model has answered the question correctly based on the sampling
frames of the clue intervals.

<Frame1>, <Frame2>, ..., <Framen>
<Sub> (Option)
<FT> (Option)

Question:
<Question>

The model’s prediction is: "<Prediction>"
Your output is:
```

4 VIDEO TAGS

We collected 1219 videos on the two platforms, of which 570 videos were collected on YouTube, accounting for 46.8%; and 649 videos were collected on Bilibli, accounting for 53.2%. 32.5% of the videos have subtitles. In addition, we assigned a level 2 or 3 tag to each video, of which there are 171 level 2 tags and 638 level 3 tags. The specific categories and quantities of tag-2 and tag-3 are shown in Tables 1 and 2.

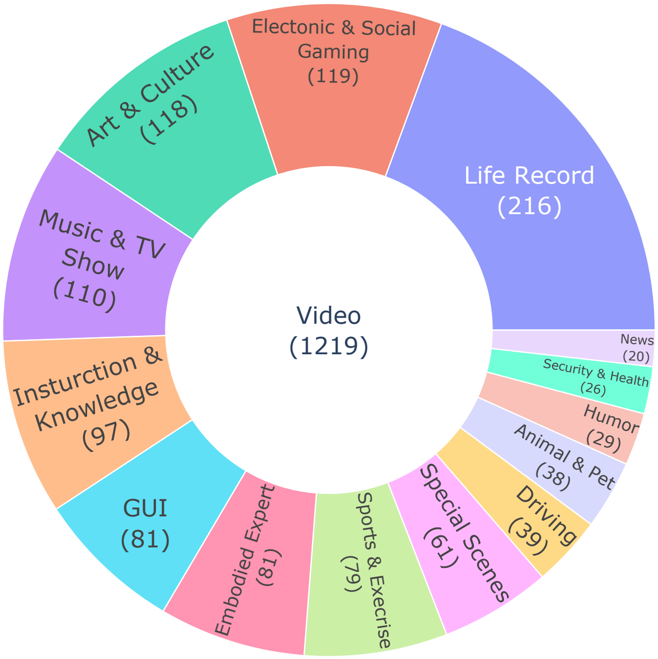


Figure 4: Distribution of video root categories, displaying the number of videos within each category.

4.1 TAG-1

The categories and quantities of Tag-1 (root categories) are shown in Figure 4.

4.2 TAG-2

The specific categories and quantities of Tag-2 are shown in Tables 1.

Table 1: Categories and counts of the level-2 video tags.

Category	#	Category	#	Category	#	Category	#	Category	#
Diverse life	66	Beach	3	Diet	47	Knowledge sharing	3	Variety shows	46
First-person work	40	Forest	3	Traditional sports	41	Board games	3	Travel	34
Extreme sports	28	Pet care	2	Simulation games	37	Russian cuisine	2	Movies/TV dramas	29
Software demonstration	28	Racing games	2	Wildlife	24	MOBA games	2	Social games	25
Festivals	22	Driver's license test	2	Documentary	21	Waterside living	2	Play	24
Coding	22	InDesign	2	Humor/Comedy	20	Designbuilder	2	Learning	21
Working	18	Illustrator	2	Makeup	17	ZBrush	2	Eating	16
Traditional crafts	16	Bus	2	RPG games	15	Digital product reviews	2	Shopping sharing	16
Pets	14	Reality challenge games	2	Public safety	13	Karting	2	Fitness	13
Cooking	12	Excavator	2	Housekeeping services	12	Social news	2	Animation	12
Strategy games	11	Helicopter	2	Renovation	11	Motorcycle	2	Handicraft	10

Category	#	Category	#	Category	#	Category	#	Category	#
Funny videos	10	Efficiency tool software	2	Shopping	10	Ruins	2	Underwater	9
Music	9	House tour	2	Architecture	9	Political news	2	Humanities	9
Fashion	9	Insects	2	Dance	8	First-person augmented reality experience	2	Technology	8
Real battlefield/Counter-terrorism	7	Business news	2	School	7	Chemistry	2	Open world games	7
First aid	6	Antarctica	2	Shooting games	7	Debate competition	2	In the cave	7
Medical care	6	Human-animal relationship	2	Art	7	Auction	2	Stage performance	6
Real-time strategy games	6	First-person live-action CS	1	Board games	6	Prison	2	Note-taking software	6
Desert	6	Raccoon	1	Clothing	6	Primates	2	Test drive	5
First-person sports	5	Battlefield	1	Packing	5	Chinese dim sum	1	First-person cooking	5
Aquatic animals	5	Installation	1	Storage	5	Robots	1	Cave	4
Trucks	4	Texas Hold'em	1	Graphic design software	5	Laboratory	1	Train	4
Cars	4	Game: Cities Skylines	1	First-person driving	4	Driver's license	1	Space	4
Knowledge management software	4	Photography	1	Electric vehicles	4	Tea culture	1	Comprehensive	4
Snow	4	Comic convention	1	Sailing	4	Tennis	1	Religion	4
Health and wellness	3	First-person adventure	1	Airplane	4	Motorcycle maintenance	1	Repair	3
Street photography	3	Wild	1	Selection	4	Canyoning	1	Beach	3
Video editing software	3	Cycling	1	Animation and image generation software	4	First-person homework	1	Street interviews	1
Economic news	1	First-person work: Coffee shop	1	Entertainment news	4	Driving	1	Diet and wellness	1
Rescue and disaster relief	1	First-person virtual reality experience	1	Environmental news	4	Sports games	1	Music production software	1
Jade carving	1	First-person work: Burger shop	1	Detective	1	Military news	1	Drawing techniques	1
International news	1	Polar animals	1			First-person games	1		

4.3 Tag-3

The specific categories and quantities of Tag-3 are shown in Tables 2.

Table 2: Categories and counts of the level-3 video tags.

Category	#	Category	#	Category	#	Category	#	Category	#
Eight Cuisines	16	Photography Tips	2	Cat	5	Python	2	TV Series	5
Chinese Pastries	6	Raft Survival	2	Short Film	5	Psychology	2	Merchandise	5
Tea Culture	5	Portal	2	Opera	5	Drama	2	Giant Panda	2
Electric Vehicle	4	MasterChef	2	Pottery	4	Food Exploration	2	Basketball	4
Cleaning Tips	4	Action Film	2	Football	4	MatLab	2	Bullet Journal	4
Sketch	4	The Amazing Race	2	Motorcycle	4	History and Culture: Museum	2	Parenting	3
Grocery Shopping	3	Detective Chinatown	2	Public Service Short Film	3	Space Launch	2	Keep Running	3
Food Delivery	3	Unity	2	Taiwan Travel	3	Prison Documentary	2	Dog	3
Rescue and Disaster Relief	3	Kung Fu	2	Monopoly	3	Golf	2	Tennis	3
Organization Tips	3	Pandemic Response	2	Grading Homework	3	Human-Animal Symbiosis	2	Hide and Seek	3

Category	#	Category	#	Category	#	Category	#	Category	#
Extreme Challenge	3	The Great British Bake Off	2	Dou Dizhu	3	The Life We Long For	2	Premiere Pro	3
Comedy	3	Shark	1	SketchUp	3	Puppy	1	Stable Diffusion	3
Meal Prep Tips	3	Dumplings	1	Winemaking	3	Driving Test	1	Turkish Cuisine	3
Photoshop	3	Gua Sha	1	Economy	3	Cardboard	1	Japan	3
Korea Shopping	3	VR	1	Pr	3	Japan Travel	1	Divas Hit the Road	3
Face Painting	2	Gourmet Food	1	Special Effects Makeup	2	Cream Cake	1	Everyday Makeup	2
Campus Life	2	Freediving	1	Graduation	2	Biology/Chemistry Experiments	1	Tap Dance	2
Nursing Procedures	2	Biology Experiment	1	Escape Room	2	Special Forces Training	1	Underwater Exploration	2
Racing	2	Surfing	1	Rock Climbing	2	Horizon	1	Wingsuit Flying	2
Paragliding	2	Foundation Makeup	1	Gymnastics	2	Cake	1	DOTA2	2
Civilization VI	2	Subway	1	Plants vs. Zombies	2	Pop-up Book	1	New Energy Vehicle Test Drive	2
Novice Highway Driving	2	Handmade Soap	1	CSGO	2	Milk Tea Shop	1	GTA5	2
Driver’s License	2	Solo Dining	1	Test Drive	2	Cheesecake	1	Night Market Experience	2
Housework	2	Puff Pastry	1	Work Life	2	Annual Comedy Competition	1	Craft Making	2
Music MV	2	Belly Dance	1	Symphony Orchestra	2	Trauma Care	1	Castle	2
Underwater Salvage	2	Pyramid	1	Skiing	2	Eyebrow Drawing	1	Baseball	2
Skating	2	Parrot	1	Counter-Terrorism Action	2	Subway Operations	1	Rhino	2
No Man’s Sky	2	Sushi	1	Stardew Valley	2	Nail Art	1	Supermarket Restocking	2
Amusement Park	2	Meal Prep	1	Family Feast	2	Underwater Fishing	1	Procurement	2
Magic	2	Underwater Welding	1	Where Are We Going, Dad?	2	Music Festival	1	Street Dance of China	2
Cave	2	Rabbit	1	Freediving	2	Biology	1	Cosplay Makeup	2
Velvet Flowers	2	Coffee	1	Lantern Festival	2	Medicine	1	Sailing	2
Car	2	Cultural District	1	Truck Driver’s Daily Life	2	Healthy Living Habits	1	Restaurant Waiter	2
Mountain Village	2	Baduanjin	1	Trash Picking	2	Elephant	1	Behind the Scenes	2
Latin Dance	2	Lion	1	Medical Equipment Use	2	Meerkat	1	College Entrance Exam	2
F1 Racing	2	Winter Solstice	1	Badminton	2	Mediterranean Diet	1	Long-Distance Running	2
Fitness Plan	2	Makeup Removal	1	Truth or Dare	2	Korean Makeup	1	Leather Craft	2
Hanfu	2	Shoe Making	1	Red Alert 2	2	Freelancer	1	Cooking	2
Shopping	2	Mountain Biking	1	Theme Park	2	Red Panda	1	Librarian	2
Concert	2	Brown Bear	1	Earthquake Drill	2	Wolf	1	Snowmobile	2
Cultural Relics Archaeology	2	Oolong Tea	1	Embroidery	2	Paper Cutting	1	Indian Cuisine	2
Luxury Car Test Drive	2	Collage	1	Hearthstone	2	Vanity	1	Vegetarianism	2
Microfilm	2	Mushroom Picking	1	Street Dance	2	Arab Robe	1	Emergency Evacuation	2
Rescue	2	Beading	1	Space Station Life	2	Beachcombing	1	Skateboarding	2
Diving	2	Fishing	1	Truck	2	Duck House	1	Skyline	2
Ocean Park	2	Violin	1	Rehabilitation Training	2	Dungeon	1	Real Battlefield	2
Water Splashing Festival	2	Polar Animals	1	Minecraft	2	Traditional Chinese Medicine	1	Cloud Notes	2
GoodNotes	2	Forza Horizon	1	Market Shopping	2	Delivery Service	1	Antique Market Shopping	2
Volleyball	2	Convenience Store	1	Board Games	2	Board Game: Who Are You	1	Sculpture	2

Category	#	Category	#	Category	#	Category	#	Category	#
Bus	2	Board Game: Storytelling	1	Valorant	2	Making Small Books	1	Notion	2
City Walk	2	Eyebrow Shaping	1	Superhero Movies	2	Watch Repair	1	Train	2
Fried Chicken	2	Concealer	1	Zotero	2	Laptop	1	Duty-Free Shopping	2
Waterside Life: Beachcombing	2	Takoyaki	1	CPR	2	Creative Market	1	Free Fighting	2
Temple of Heaven	2	Variety Show	1	National Day	2	Board Game: Redemption Journey	1	Halloween	2
Dragon Boat Festival	2	Tacit Challenge	1	Acupuncture	2	Supermarket Challenge	1	Ancient Greek Temples	2
Go-Karting	2	Elephants - Wild	1	Yacht	2	Airplane	1	World of Warcraft	2
After Effects	2	Digital Product Review	1	Obsidian	2	Theme Park	1	Pixel Composer	2
Furniture Assembly	2	Digital Product Review: Smart Home	1	Digital Painting	2	Shopping in Europe	1	Digital Product Review: Tablet	2
Abandoned Buildings	2	Digital Product Review: Ergonomic Chair	1	Fat Loss Training	2	Chocolate Making	1	Ab Workout	2
Hockey	2	DIY Mini House	1	Spring Festival	2	Waterside Life: Fishing	1	Easter	2
Warcraft III	2	Digital Product Review: Smartphone	1	Wasteland Delivery	2	Drawing Techniques	1	Pizzeria	2
High-Altitude Work	2	Braised Pork Rice	1	Farming	2	Fish Pond Construction	1	Shopping in Thailand	2
Museum	2	Italy	1	Flea Market	2	Happy Old Friends	1	Art Gallery	2
Ace vs. Ace	2	Wilderness Survival	1	I Am a Singer	2	Medieval Dynasty	1	Firefighting	2
Military Exercise	2	The Witcher	1	Snow Survival	2	Planet Zoo	1	Beach Camping	2
Dumbbell Training	2	Aircraft Loading	1	Bowling	2	Real-life CS	1	Fitness Ball Training	2
Italian Cuisine	2	Car Repair	1	Japanese Cuisine	2	Pet Store Job	1	Elden Ring	2
Water Obstacle Course	2	Ergonomic Chair	1	Markdown	2	Basement	1	Word	2
CapCut	2	Glacier Climbing	1	Ruby	2	Pufferfish	1	VSCode	2
Blender	2	Jade Carving	1	Australian Travel	2	Ancient Greek Philosophy	1	Baking Techniques	2
Wedding	2	Train Driving Simulator	1	Drowning	2	Theory of Relativity	1	Ruins Exploration	2
Archery	2	Used Cars	1	Colosseum	2	Taiwan Shopping	1	Thanksgiving	2
Autonomous Driving Experience	2	AI Painting	1	Excavator	2	Fishing	1	Call of Duty	2
Adobe Acrobat Pro	2	Farm	1	Summer Outfits	2	Daily Life After Returning Home	1	Southeast Asia Travel	2
Camping	2	Home Tour	1	Disney	2	Village School	1	Massage Therapy	2
The Tonight Show Starring Jimmy Fallon	2	Desert	1	Fire Drill	2	Parkour	1	Fire Evacuation	2
Qipao	2	Buddhism	1	French Cuisine	2	Great Wall	1	Helicopter	2
Manor Lord	2	Real-Life Subway Game	1	Fallout Shelter	2	Mixed Noodles	1	Mover	2
PPT	2	Epoxy Resin	1	SQL	2	Knitting	1	Spring Outfits	2
Seafood Buffet	2	Paris	1	Studio	2	Yoga	1	North American Travel	2
Helicopter Skiing	2	Calligraphy	1	Qixi Festival	2	Thriller	1	Spanish Cuisine	2
German Cuisine	2	Real Battlefield/Counter-Terrorism	1	inZOI	2	Chinese Painting	1	Vision Pro	2
Mailing and Packaging	2	Opera	1	Making Hot Dogs	2	Luggage	1	LaTeX	2
Steam	2	Digital Product Review: Electric Toothbrush	1	Family Feud	2	Mythical Fantasy Film	1	Thai Cuisine	2
Christianity	2	Strange House	1	Kingdom of Order	2	Mahjong	1	Plants vs. Zombies Hybrid	2
Sunny and Warm	2	Cat Café	1	Grounded	2	Kimono (Japan)	1	Coffee Shop	2

Category	#	Category	#	Category	#	Category	#	Category	#
JS	2	Cleaning	1	Quicker	2	Editing Tips: Movie Commentary Editing	1	Hunting	2
Department Store Shopping	2	Chicago	1	Home Gardening	2	Market Simulator	1	Costume Drama	2
Robot Wars	2	FamiStudio	1	Movie Trailers	2	Tattoo Covering	1	Snow Mountain Adventure	2
Equestrian	2	Organic Chemistry	1	Desert Off-Roading	2	Street Food	1	Porcelain	2
Yacht Driving	2	Drawing Tips: AI Drawing	1	OBS	2	Switzerland	1	C++	2
Clothing	2	Iceland	1	Dishwashing	2	America's Got Talent	1	Olympics	2
Rugby	2	New Journey to the West	1	Korean Cuisine	2	Sand Sculpture Art	1	7 Days to Die	2
Bartender	2	Rafting	1	Radiomics	2	Battlefield	1	European Travel	2
Livehouse	2	Delivery	1	Hiking	2	Coat	1	Ping Pong	2
Christmas	2	Tea Set	1	Cat and Mouse Game	2	Thailand	1	Frostpunk	2
Black Myth: Wukong	2	Interior Design	1	First-Person Cooking	2	Hengdian	1	PC Building	2
Rainforest Survival	2	Who's the Undercover	1	High-Intensity Interval Training	2	Real-Life Hide and Seek	1	The Sinking Land	2

REFERENCES

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.