Taskverse: A Benchmark Generation Engine for Multi-modal Language Model

Jieyu Zhang[1], **Weikai Huang**[1,*], **Zixian Ma**[1,*], **Oscar Michel**[2], **Dong He**[1],
Tanmay Gupta[2], **Wei-Chiu Ma**[2], **Ali Farhadi**[1,2], **Aniruddha Kembhavi**[2], **Ranjay Krishna**[1,2]
[1]University of Washington, [2]Allen Institute for Artificial Intelligence
https://www.task-me-anything.org

Abstract

Benchmarks for large multimodal language models (MLMs) now serve to simultaneously assess the general capabilities of models, rather than evaluating a specific capability. As a result, when a developer seeks to identify which models to use for their application, they are often overwhelmed by the number of benchmarks and remain uncertain about which benchmark results are most reflective of their specific use case. This paper introduces TASKVERSE, a benchmark generation engine that produces benchmarks tailored to different user needs. TASKVERSE maintains an extendable taxonomy of visual assets, including 113K images, 10K videos, 2K 3D object assets, over 365 object categories, 655 attributes, and 335 relationships, and can programmatically generate over 750 million Image/VideoQA questions. Additionally, it algorithmically addresses user queries regarding MLM performance efficiently within a relative computational budget by employing interactive learning query approximation algorithms. With TASKVERSE, we can answer specific and fine-grained user queries like: "Which model is the best VideoQA model for recognizing material within 1000 inference times?"

1 Introduction

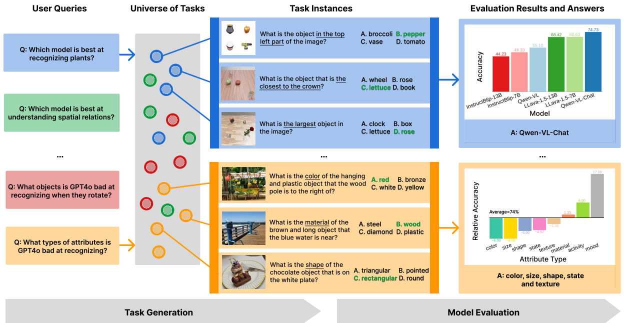

User Queries
Universe of Tasks
Task Instances
Evaluation Results and Answers
Q: Which model is best at recognizing plants?
Q: Which model is best at understanding spatial relations?
...
Q: What objects is GPT4o bad at recognizing when they rotate?
Q: What types of attributes is GPT4o bad at recognizing?
What is the object in the top left part of the image?
A. broccoli B. pepper C. vase D. tomato
What is the object that is the closest to the crown?
A. wheel B. rose C. lettuce D. book
What is the largest object in the image?
A. clock B. box C. lettuce D. rose
...
What is the color of the hanging and plastic object that the wood pole is to the right of?
A. red B. bronze C. white D. yellow
What is the material of the brown and long object that the blue water is near?
A. steel B. wood C. diamond D. plastic
What is the shape of the chocolate object that is on the white plate?
A. triangular B. pointed C. rectangular D. round
Accuracy
44.23
49.33
55.10
68.42
68.63
74.73
InstructBlip-13B
InstructBlip-7B
Qwen-VL
LLava-1.5-13B
LLava-1.5-7B
Qwen-VL-Chat
Model
A: Qwen-VL-Chat
...
Relative Accuracy
Average=74%
−8.00
−8.00
−5.00
−4.67
−2.33
1.33
6.00
17.00
color
size
shape
state
texture
material
activity
mood
Attribute Type
A: color, size, shape, state and texture
Task Generation
Model Evaluation

Figure 1: We present examples of user queries, corresponding task instances generated by TASKVERSE as well as the evaluation results on them that answer the queries.

When a developer wants to identify which models to use for their application, they remain uncertain about which benchmark results are most aligned with their specific use case.

[*] The authors contribute equally to this work.

Workshop on Video-Language Models at 38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Although MLMs are released with evaluations on benchmarks like MMBench, MMMU, BLINK, and SeedBench [29, 39, 25, 24, 9], their performance across these holistic benchmarks do not pinpoint which fine-grained capabilities are lacking.

There is a need for a principled benchmark generation process that answers task-specific user queries: "(**Q1**) Which model is the best at recognizing the shape of objects?" or "(**Q2**) what are the model's weaknesses that we can further improve on?".

To address this, we present TASKVERSE, a benchmark generation engine that curates a custom benchmark given a user query (Figure 1).

Programmatic and scalable task generation. The task generation process of TASKVERSE includes 2 key components: taxonomy and task generators. Task generators are Python codes that take in taxonomy and generate ImageQA/VideoQA tasks.

First, TASKVERSE maintains an extendable taxonomy with corresponding visual assets (*e.g.* images with scene graphs [23], 3D object assets [7], videos with spatio-temporal annotations [18], rendering softwares [5], *etc.*.). It is implemented as an extendable library where new concepts and their corresponding assets and annotations can be easily added.

Second, TASKVERSE contains programmatic task generators that sub-select from the taxonomy to curate a large number of ImageQA/VideoQA questions. TASKVERSE contains 28 different task generators focused on different capabilities of ImageQA/VideoQA models. With our current taxonomy, TASKVERSE can generate over 750 million questions with real or synthetic images/videos.

Efficient and fine-grained evaluation. TASKVERSE support supports 4 types of user query: Top-K query, Threshold query, Model comparison query, Model debugging query. For evaluation efficiency, we also implement 3 different ML algorithms to approximate the results of user queries via predicting the model performance across numerous questions. For example, With *Active* algorithm, we can achieve a 70% hit rate in Top-K query with under 10% of model inference times compared with evaluation on all questions.

What can we do with TASKVERSE. With TASKVERSE, we answer several critical user queries about MLMs, like what are the best VideoQA models in specific capability (e.g. recognizing, reasoning)? We also released useful resources including TASKVERSE-RANDOM benchmark, TASKVERSE-2024 benchmark, TASKVERSE-DB, TASKVERSE-UI.

2 TASKVERSE

2.1 TASKVERSE paradigm

Consider a user who wants to know "Which open-sourced MLM is best at recognizing objects even if the object is rotating?". TASKVERSE provides an interface for the user to pose such questions and provides them with an answer (Figure 2). It contains a taxonomy to symbolically represent visual content. A query identifies the relevant portion of the Taxonomy required to answer the query. It also contains task generators that create input-output pairs that test for a specific capability. The Taxonomy subset is used to select the appropriate task generator. We adopt the common input-output format used in existing benchmarks, *i.e.*, all the task instances in TASKVERSE contain an image/video, a question, and multiple options with one ground truth answer. MLMs will be evaluated on these generated task instances and the results will be returned back to the user. Finally, it also supports efficient approximation algorithms to estimate model performance without evaluating all the task space. Unlike most existing procedural data systems, we design TASKVERSE so that the generation space of tasks can be expanded by adding new source data and/or task generator code. More details about taxonomy, task generation, supported user queries, and results approximation algorithms are in Appendix B and C.

2.2 Stats of TASKVERSE

TASKVERSE can enable evaluations with up to 113K+ realistic images (from Visual Genome), 9K+ videos (from Charades), and infinite synthetic images and videos.

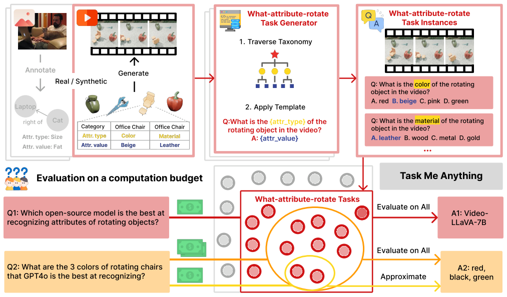
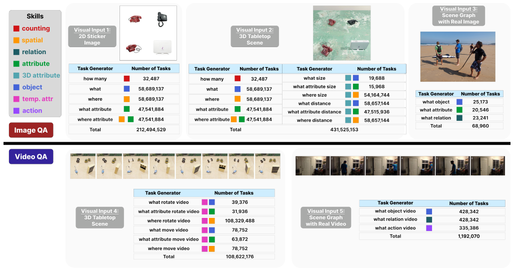

Annotate
Laptop
right of
Cat
Attr. type: Size
Attr. value: Fat
Real / Synthetic
Generate
Category | Office Chair | Office Chair
Attr. type | Color | Material
Attr. value | Beige | Leather
What-attribute-rotate Task Generator
1. Traverse Taxonomy
2. Apply Template
Q:What is the {attr_type} of the rotating object in the video?
A: {attr_value}
What-attribute-rotate Task Instances
Q: What is the color of the rotating object in the video?
A. red B. beige C. pink D. green
Q: What is the material of the rotating object in the video?
A. leather B. wood C. metal D. gold
...
Evaluation on a computation budget
Task Me Anything
Q1: Which open-source model is the best at recognizing attributes of rotating objects?
Q2: What are the 3 colors of rotating chairs that GPT4o is the best at recognizing?
What-attribute-rotate Tasks
Evaluate on All
Evaluate on All
Approximate
A1: Video-LLaVA-7B
A2: red, black, green

Figure 2: The top part illustrates the task generation process with an example video synthesized with 3D objects and their annotations. The bottom part depicts the model evaluation process, which selects the relevant tasks based on the user's query and their budget and performs either full evaluation or results approximation to answer the query.


Skills
🟥 counting
🟧 spatial
⬛ relation
🟩 attribute
🔲 3D attribute
🟦 object
▨ temp. attr
🟪 action
Image QA
Video QA
Visual Input 1: 2D Sticker Image
Visual Input 2: 3D Tabletop Scene
Visual Input 3: Scene Graph with Real Image
Visual Input 4: 3D Tabletop Scene
Visual Input 5: Scene Graph with Real Video

Task Generator		Number of Tasks
how many	🟥	32,487
what	🟦	58,689,137
where	🟧	58,689,137
what attribute	🟩	47,541,884
where attribute	🟧 🟩	47,541,884
Total		212,494,529

Task Generator		Number of Tasks	Task Generator		Number of Tasks
how many	🟥	32,487	what size	🔲 🟦	19,688
what	🟦	58,689,137	what attribute size	🔲 🟩	15,968
where	🟧	58,689,137	where size	🔲 🟧	54,164,744
what attribute	🟩	47,541,884	what distance	🔲 🟦	58,657,144
where attribute	🟧 🟩	47,541,884	what attribute distance	🔲 🟩	47,515,936
			where distance	🔲 🟧	58,657,144
Total					431,525,153

Task Generator		Number of Tasks
what object	🟦	25,173
what attribute	🟩	20,546
what relation	⬛	23,241
Total		68,960

Task Generator		Number of Tasks
what rotate video	▨ 🟦	39,376
what attribute rotate video	▨ 🟩	31,936
where rotate video	▨ 🟧	108,329,488
what move video	▨ 🟦	78,752
what attribute move video	▨ 🟩	63,872
where move video	▨ 🟧	78,752
Total		108,622,176

Task Generator		Number of Tasks
what object video	🟦	428,342
what relation video	⬛	428,342
what action video	🟪	335,386
Total		1,192,070

Figure 3: The statistics of generatable tasks of each task generator and example image/video in TASKVERSE. Each task generator uses high-level perceptual skills, and this collection of task generators can collectively generate over 750M VQA tasks.

We include 28 different task generators across 5 types of visual inputs: 2D sticker images (2D), 3D tabletop scene images/videos (3D), and real images/videos with manually-annotated scene graphs. More stats of TASKVERSE are in Appendix D.

3 Analysing Video-Language Models with TASKVERSE

Example user query 1: what is the best video-language model for each specific skill? From Figure 4, VIDEO-LLAVA-7B and CHAT-UNIVI-7B are relatively well-rounded, positioning in the top 3 models across all skills except for Attribute understanding. On the other hand, while VIDEOCHAT2-7B specializes in object, attribute, and temporal attribute understanding.

Example user query 2: how do current video-language models perform in recognizing different types of concepts? From Figure 5, we can see that all models perform poorly at recognizing concepts related to orientation and materials, but perform well with activities, mood, etc. This may indicate a data bias in current model training, highlighting the need for more balanced data curation.

More fine-grained user queries about MLMs are in Appendix H.

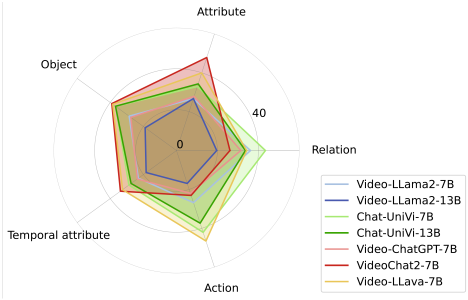


Figure 4: **Video-language model, high-level skills**

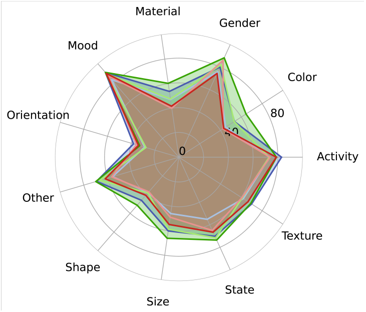


Figure 5: **Video-language model, fine-grained object and attribute skills**

4 Conclusion

In this work, we introduce TASKVERSE, a task generation and evaluation system designed to address user queries with different evaluation objectives.