Dataset Statistics of motion description dataset on Kinetics400, HMDB51 and UCF101 is given below.

Kinetics-400 Dataset:

Number of videos	Number of unique motion descriptions	Average number of verbs per motion description	Number of classes	Average number of words per motion description	Number of verbs in the motion descriptions
246000	400	3.4	400	19	1371

UCF101 Dataset:

Number of videos	Number of unique motion descriptions	Number of classes	Average number of verbs per motion description	Average number of words per motion description	Number of verbs in the motion descriptions
13320	101	101	3.2	19	325

HMDB51 :Dataset

Number of videos	Number of unique motion descriptions	Number of classes	Average number of verbs per motion description	Average number of words per motion description	Number of verbs in the motion descriptions
6849	51	51	3.2	17	164

User study to evaluate the quality of GPT-4 generated motion descriptions.

We conducted a user study to evaluate the quality of generated motion descriptions. Following [1], which questions using Amazon Mechanical Turk for such studies, we conducted this study with expert graduate student volunteers who have taken courses in computer vision and NLP. All the evaluators were trained with different questions and explained the project's overall goal and their contribution to it.

We asked two volunteer graduate students of diverse ages and ethnicities to participate in this study.

The following definitions were given.

Conciseness: Conciseness is related to the length and non-redundancy of the generated text.
Hallucinations: Hallucinations are related to generating physically non-plausible motion descriptions.
Relevance: Relevance refers to how much correspondence there is between the objects, action, and motion description.

Correctness: Correctness refers to how accurate the motion description is.
Harmfulness: Is there any objectionable or harmful content in the generated text?

Each of the above attributes is evaluated on a 5-point Likert scale. For each of the motion descriptions in each dataset, we asked the volunteers to rate the generated motion description with the above attributes. The table below shows the results of the aggregated numbers across the datasets. We report mean 5-point Likert score and IAA% which is the inter agreement between annotators that measures the percentage of descriptions where annotators gave same rating.

Conciseness		Hallucinations		Relevance		Correctness		Harmfulness	
Mean	IAA%	Mean	IAA%	Mean	IAA%	Mean	IAA%	Mean	IAA%
3.86	47.5	1.12	87	3.4	54	3.92	72	1	100

Total Number of volunteer hours spent on this user study is 8 hours.

References

1. Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In Proceedings of the 2021 EMNLP.
2. Chiang, C.H. and Lee, H.Y., 2023. Can large language models be an alternative to human evaluations?. arXiv preprint arXiv:2305.01937.