# Supplementary Materials: Submission4261 WorldGPT

Anonymous Authors

## 1 IMPLEMENTATION DETAIL

For WorldGPT, we have included the code implementation, trained model weights, and all configurations in the project folder. For WorldNet, the anonymous download link for the dataset and all data processing prompts are also provided in the project folder.

## 2 DATASET DETAILS IN WORLDNET

WorldNet-Crafted is collected from following datasets: YT-Temporal-180M [6] and HowTo100M [3] and More details about these datasets can be found in the Supplementary Materials.

WorldNet-Wild is built upon following datasets:

**YT-Temporal-180M** [6] originates from a diverse collection of 6 million public YouTube videos, intentionally encompassing a wide array of domains, datasets, and subjects to ensure broad coverage and variety.

**HowTo100M** [3] is a substantial dataset comprised of narrated videos, with a strong emphasis on instructional content. It features content creators teaching complex tasks, specifically designed with the explicit intention of elucidating the visual content displayed on screen.

WorldNet-Crafted is collected from following datasets: Ego4D [2], Something-Something V2 [1], YouCook2 [7], AVQA[5] and Charades [4]. More details about these datasets can be found in the Supplementary Materials.

**Ego4D** [2] is a massive-scale egocentric video dataset and benchmark suite. It offers 3,670 hours of daily-life activity video spanning hundreds of scenarios (household, outdoor, workplace, leisure, etc.) captured by 931 unique camera wearers from 74 worldwide locations and 9 different countries.

**Something-Something V2** [1] represents a vast compilation of labeled video clips showcasing humans executing pre-determined fundamental actions using everyday objects. Generated through the efforts of a broad network of crowd workers, this dataset facilitates the development of a detailed comprehension of basic actions within the physical world for machine learning models.

**YouCook2** [7] stands as one of the most extensive task-oriented, instructional video datasets within the vision community. It comprises 2,000 lengthy, unedited videos across 89 cooking recipes, with an average of 22 videos for each unique recipe. The procedural steps within each video are meticulously annotated with temporal boundaries and described using imperative English sentences.

**AVQA** [5] is an audio-visual question answering dataset comprising 57,015 videos that encapsulate daily audio-visual activities. Accompanying these videos, there are 57,335 uniquely crafted question-answer pairs. These pairs are designed to hinge on cues from both audio and visual modalities, where information from a single modality would be inadequate or ambiguous.

**Charades** [4] features recordings from hundreds of individuals in their own homes, engaging in routine daily activities. It consists of 9,848 annotated videos, each with an average duration of 30 seconds, capturing the actions of 267 people across three continents. The annotation for each video includes multiple free-text descriptions, action labels, temporal intervals for these actions, and categories of objects interacted with.

## REFERENCES

[1] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*. 5842–5850.

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.

[3] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2630–2640.

[4] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 510–526.

[5] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3480–3491.

[6] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems* 34 (2021), 23634–23651.

[7] Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.