Correspondence of high-dimensional emotion structures elicited by video clips between humans and Multimodal LLMs

Haruka Asanuma[1], Naoko Koide-Majima[4, 5], Ken Nakamura[2], Takato Horii[3], Shinji Nishimoto[4, 5, 6], and Masafumi Oizumi[1, *]

[1]The University of Tokyo, Graduate School of Arts and Sciences, Tokyo, 153-8902, Japan
[2]The University of Tokyo, Faculty of Engineering, Tokyo, 113-8656, Japan
[3]The University of Osaka, Graduate School of Engineering Science, Osaka, 565-0871, Japan
[4]Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology, Osaka, 565-0871, Japan
[5]The University of Osaka, Graduate School of Frontier Biosciences, Osaka, 565-0871, Japan
[6]The University of Osaka, Graduate School of Medicine, Osaka, 565-0871, Japan
[*]c-oizumi@g.ecc.u-tokyo.ac.jp

The PDF file includes:

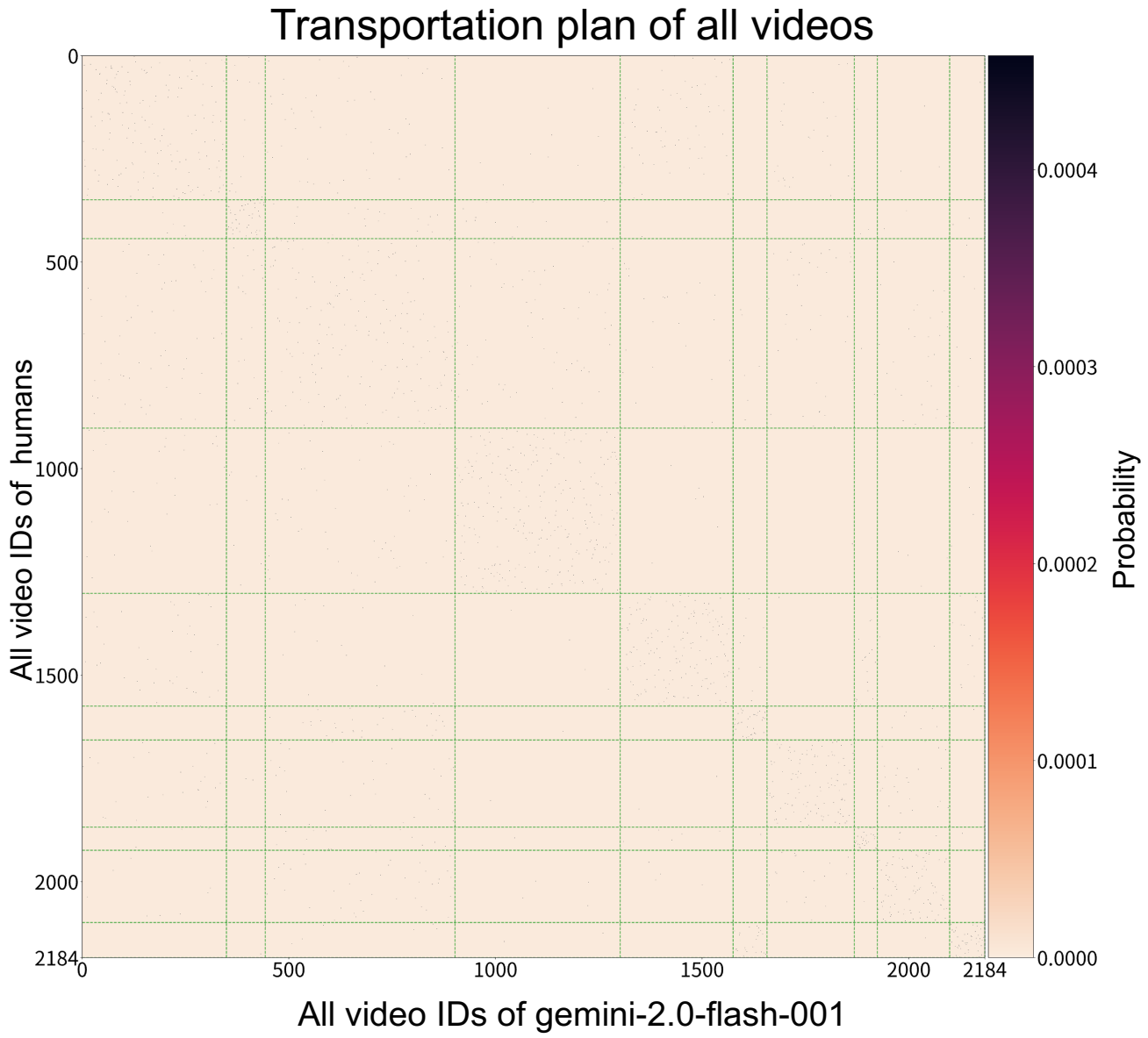

Figure S1. Comparison of the human similarity structures of videos in the Cowen & Keltner dataset with the similarity structures estimated by Gemini. Optimal transportation plan obtained for all videos by GWOT between the human and Gemini RDMs. Green lines represent the category boundaries of the videos.