Sora for Social Vision With Parallel Intelligence: Social Interaction in Intelligent Vehicles

Hui Yu, *Senior Member, IEEE*, Wei Liang, Lili Fan, Yutong Wang, and Fei-Yue Wang, *Fellow, IEEE*

***Abstract*—Artificial technologies have made rapid progress and achieved various superior tasks in the past few years, including but not limited to classification, detection, image generation and data processing. Particularly, the very recent emerging Sora has demonstrated the exceptional ability of text-to-video generation lasting for 1 minute long with impressive quality. It provides a huge potential for many new applications across industries, especially social interaction in intelligent vehicles. The emergence of innovative intelligence vehicle applications has given rise to novel requirements for social and human-vehicle interaction within the associated contexts, where Sora and social vision could play an important role. In this perspective, we present a new Social Interaction framework based on Sora and parallel intelligence in intelligent vehicles and provide a novel perspective for conducting new social and human-vehicle interaction in the context of intelligent vehicles.**

***Index Terms*—Sora, parallel intelligence, social vision, social interaction, intelligent Vehicles, diffusion model, human-machine interaction.**

I. Introduction

COMPUTER vision and deep learning as a type of eminent artificial technologies (AI) have made fast progress recently in various tasks such as classification [1], [2], object detection and tracking [3], [4], [5], automation [6], [7], [8], [9], emotional expression [10], [11] and reconstruction [12]. Due to its powerful abilities for processing data of various modalities such as vision, audio and text, deep learning has also enabled a wide range of applications in social and human-machine interaction [13], [14]. Analysis and processing of human (e.g. passenger) movements, expressions and social signals and the interaction between humans and machines play an important role in intelligence vehicles. For instance, Gou et al. [15] proposed to detect the driver's facial feature points and perform head pose estimation to monitor the driver's face in a non-invasive way, thereby realizing the identification of dangerous operations of intelligence vehicles.

Recently, large language models, which predominantly rely on the Transformer architecture [21], have achieved remarkable results across various domains. These models not only provide valuable insights but also offer diverse perspectives on the interaction between humans and machines. Furthermore, the fusion of diffusion model-driven image generation models [16], [17], [18], [19], [20] further enriches the way of human-machine interaction. Notably, the recently announced AI model named Sora, shows the remarkable performance for creating realistic video scenes up to 1 minute long directly from input language description [16].

The progress and development of autonomous driving technology and intelligence vehicles significantly push forward the way of the traditional human-vehicle interaction and human-human interaction. The interaction between humans and vehicles (driving operations) as well as social interaction among humans (passengers), may be completely changed in intelligent vehicles. The communication between passengers in the vehicles will not be limited to simple language and verbal communication but in a multimodal and multidimensional way. In such scenarios, Sora, as a superior generation model, can generate corresponding visual scenes based on passengers' language and actions, extending the original verbal communication between passengers to multi-modal communication that includes visuals, see Fig. 1. Furthermore, based on the obtained human-human

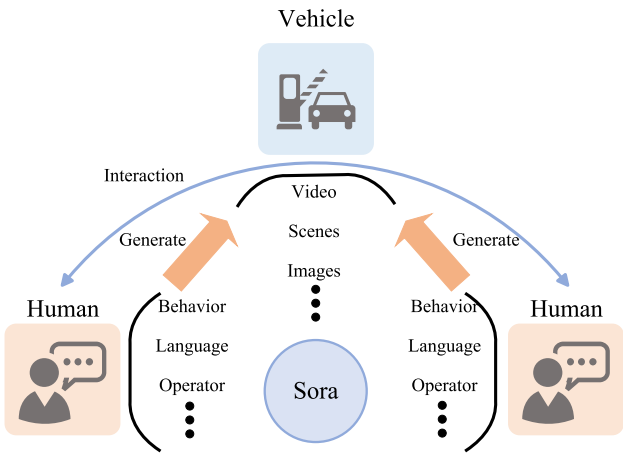

Fig. 1. Sora for Human-Machine interactions in Intelligent Vehicles.

Manuscript received 19 March 2024; revised 20 March 2024 and 27 March 2024; accepted 1 April 2024. This work was supported by the State Scholarship Fund from the China Scholarship Council (CSC). *(Corresponding author: Fei-Yue Wang.)*

Hui Yu is with the School of Creative Technologies, University of Portsmouth, PO1 2DJ Portsmouth, U.K. (e-mail: hui.yu@port.ac.uk).

Wei Liang is with the University of Shanghai for Science and Technology, Shanghai 200093, China, and also with the School of Creative Technologies, University of Portsmouth, PO1 2DJ Portsmouth, U.K. (e-mail: 201440058@st.usst.edu.cn).

Yutong Wang and Fei-Yue Wang are with the State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: yutong.wang@ia.ac.cn; feiyue.wang@ia.ac.cn).

Lili Fan is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100811, China (e-mail: lilifan@bit.edu.cn).

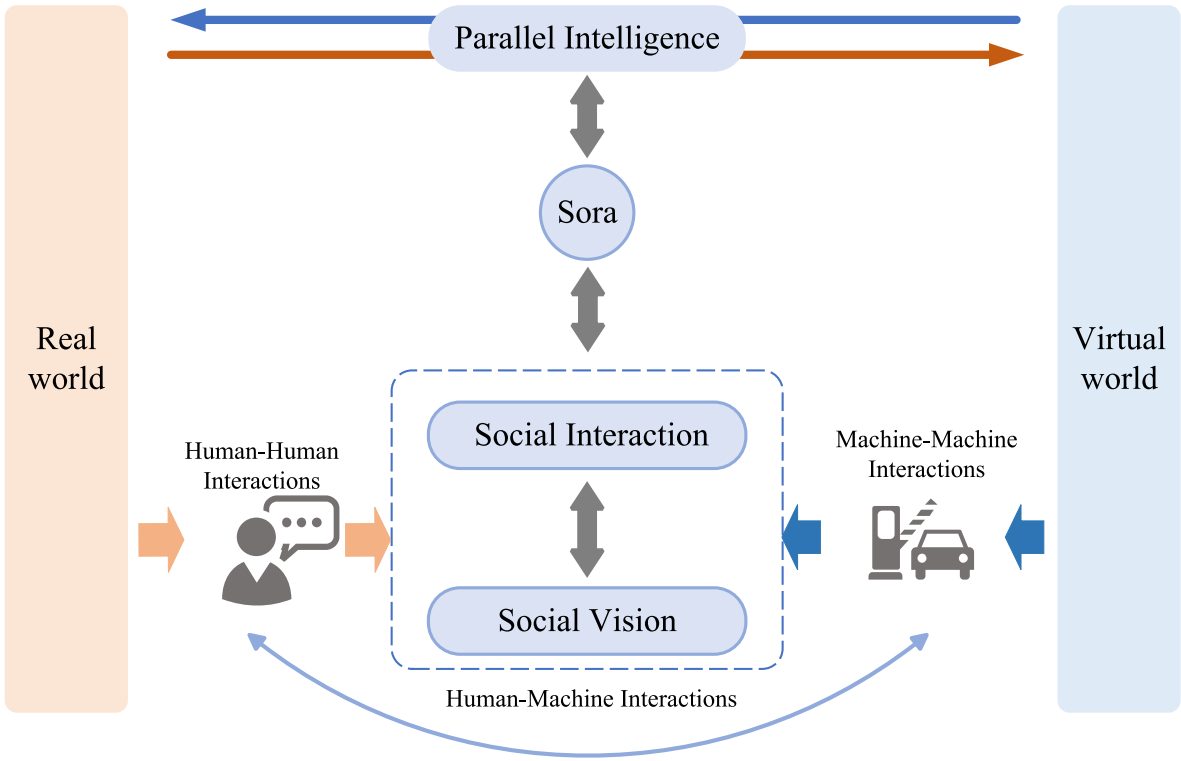

Fig. 2. Overview of Sora and Parallel Intelligence for Social Interaction and Social Vision in Intelligent Vehicles.

and human-vehicle interaction data, Sora generates a virtual scene, realizing the interaction between the real world and the virtual world through Parallel Intelligence.

II. Sola for Intelligent Vehicles

Sora is a new extension of the diffusion model [16] in generating realistic and imaginative scenes from text descriptions, combining transformer architecture [21]. At the core of Sora, Stable Diffusion [22] is a diffusion model (DM) inspired by non-equilibrium thermodynamics. They defined a Markov chain of diffusion steps to slowly add random noises to the data, and then learned to invert the diffusion process to build the desired data sample from the noises. Drawing inspiration from text tokens that unify different modalities data such as code, math and natural languages [23], Sora extracts visual patches as highly scalable and effective representations for videos. Taking spacetime latent patches as transformer token and inserting randomly initialized patches in the inputs, Sora achieves control of video including resolutions, durations and aspect ratios. Benefiting from Transformer's superior scaling features across multiple domains, including large language models [24], computer vision [25], and image generation [26], [27], Sora can generate high-quality videos according to specific languages. It potentially provides a novel way of human-human interaction, machine-machine interaction, and human-machine interaction [28], [29], [30].

Parallel Intelligence is a virtual/real interactive computing framework that integrates computer vision, deep learning, computational intelligence and virtual reality [31], [32], [33], which is the framework of artificial societies, computational experiments, and the parallel execution (ACP) theory [34], [35], [36]. It involves building realistic artificial scenarios and virtual simulation systems to model and represent complex real-world scenes in the real world. Effective machine learning and foundation vision models are trained by simulation computing experiments in a combination of large-scale artificial scene datasets and real-world scene datasets, which can achieve perception and understanding of complex environments for optimizing the vision system by parallel execution [13]. Together with the digital twin, Parallel Intelligence plays a significant role in the emerging development of intelligent vehicles [37], [38], [39], [40], [41] synergizing to promote the evolution of intelligent vehicle systems [42], [43].

Social vision is originally a psychological concept [44] based on the phenomenon that humans can effectively interpret the mental and emotional states of others and make quick judgments about their personalities and personalities simply by observing them [45]. It reflects a close relationship between vision and social interaction and plays a critical role in the development and maintenance of social exchange. Humans as social subjects have a profound impact on the visual system, and at the same time, they can also perform human-computer interaction functions through the visual system [46]. Social vision in psychology can be combined with computer vision in machines [47], extending to cutting-edge and broadly interdisciplinary research that is currently at its forefront.

In this perspective, following the introduction in extending the social vision concept to computer vision for intelligent vehicles, we present a new paradigm of human-machine and

human-human interaction with the concept of Sora, which can build controllable virtual data (e.g. virtual images and videos) with high-fidelity virtual humans [48] according to captured human language communication, physical behaviours, emotions and so on. Furthermore, human social data and machine virtual data can be integrated to establish a comprehensive virtual/real interactive framework between the virtual and real world for intelligent vehicles by Parallel Intelligence. Fig. 2 presents the overall architecture of the Sora, parallel intelligence and social vision for social interaction in Intelligent Vehicles. In the Intelligent Vehicles scenario, the behaviour and people (e.g. passengers) in the real world are obtained and processed by social interaction and social vision. Based on the obtained human-machine interaction or human-human social interaction data, Sora or other new generation models can generate virtual scenarios in real-time, realizing the interaction between the real world and the virtual world through parallel intelligence.

III. Conclusion

Overall, this perspective presents a social interaction and social vision framework based on Sora and parallel intelligence in Intelligent Vehicles. This framework takes the first step towards realizing a new social interaction and human-vehicle interaction paradigm in the context of intelligent vehicles, combining generation model Sora and virtual/real interactive vision computing framework parallel intelligence. It shows that the new generation models such as Sora along with parallel intelligence can push forward the application paradigm of connecting people and machines in the scenarios of intelligent vehicles.

References

[1] X. Cai, M. Giallorenzo, and K. Sarabandi, "Machine learning-based target classification for MMW radar in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 6, no. 4, pp. 678–689, Dec. 2021.

[2] N. Kadria, A. Ellouze, M. Ksantini, and S. H. Turki, "New LSTM deep learning algorithm for driving behavior classification," *Cybern. Syst.*, vol. 54, no. 4, pp. 387–405, 2023.

[3] I. Ahmed, S. Din, G. Jeon, F. Piccialli, and G. Fortino, "Towards collaborative robotics in top view surveillance: A framework for multiple object tracking by detection using deep learning," *IEEE/CAA J. Automatica Sinica*, vol. 8, no. 7, pp. 1253–1270, Jul. 2021.

[4] W. Liang, D. Ding, and G. Wei, "Siamese visual tracking combining granular level multi-scale features and global information," *Knowl.-Based Syst.*, vol. 252, 2022, Art. no. 109435.

[5] B. Tian, Y. Li, B. Li, and D. Wen, "Rear-view vehicle detection and tracking by combining multiple parts for complex urban surveillance," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 2, pp. 597–606, Apr. 2014.

[6] L. Chen et al., "Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 6, pp. 3234–3246, Jun. 2021.

[7] B. Li et al., "Integrating large language models and metaverse in autonomous racing: An education-oriented perspective," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 59–64, Jan. 2024.

[8] S. Ge et al., "Making standards for smart mining operations: Intelligent vehicles for autonomous mining transportation," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 413–416, Sep. 2022.

[9] X. Hu, S. Li, T. Huang, B. Tang, R. Huai, and L. Chen, "How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 593–612, Jan. 2024.

[10] H. Yu, O. G. Garrod, and P. G. Schyns, "Perception-driven facial expression synthesis," *Comput. Graph.*, vol. 36, no. 3, pp. 152–162, 2012.

[11] Y. Xia, W. Zheng, Y. Wang, H. Yu, J. Dong, and F.-Y. Wang, "Local and global perception generative adversarial network for facial expression synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1443–1452, Mar. 2022.

[12] J. Lou et al., "Realistic facial expression reconstruction for VR HMD users," *IEEE Trans. Multimedia*, vol. 22, no. 3, pp. 730–743, Mar. 2020.

[13] H. Yu, Y. Wang, Y. Tian, H. Zhang, W. Zheng, and F.-Y. Wang, "Social vision for intelligent vehicles: From computer vision to foundation vision," *IEEE Trans. Intell. Veh.*, vol. 8, no. 11, pp. 4474–4476, Nov. 2023.

[14] M. I. Khedher, H. Jmila, and M. El-Yacoubi, "On the formal evaluation of the robustness of neural networks and its pivotal relevance for AI-based safety-critical domains," *Int. J. Netw. Dyn. Intell.*, vol. 2, no. 4, 2023, Art. no. 100018.

[15] C. Gou, Y. Zhou, Y. Xiao, X. Wang, and H. Yu, "Cascade learning for driver facial monitoring," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 404–412, Jan. 2023.

[16] J. Ho, J. Ajay, and A. Pieter, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020,vol. 33, pp. 6840–6851.

[17] A. Q. Nichol and D. Prafulla, "Improved denoising diffusion probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8162–8171.

[18] D. Prafulla and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 8780–8794.

[19] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 26565–26577.

[20] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4195–4205.

[21] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5998–6008.

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10674–10685.

[23] K. He, X Chen, S Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[24] J. Devlin, M. -W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[25] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021.

[26] N. Parmar et al., "Image transformer," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, vol. 80, pp. 4055–4064.

[27] A. Arnab et al., "ViVit: A video vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 6836–6846.

[28] F.-Y. Wang et al., "When does sora show: The beginning of TAO to imaginative intelligence and scenarios engineering," *IEEE/CAA J. Automatica Sinica*, vol. 11, no. 4, pp. 809–815, Apr. 2024.

[29] H. Yu, X. Liu, Y. Tian, Y. Wang, C. Gou, and F.-Y. Wang, "Sora-based parallel vision for smart sensing of intelligent vehicles: From foundation models to foundation intelligence," *IEEE Trans. Intell. Veh.*, early access, Mar. 12, 2024, doi: 10.1109/TIV.2024.3376575.

[30] X. Li et al., "Sora for senarios engineering of intelligent vehicles: V&V, C&C, and beyonds," *IEEE Trans. Intell. Veh.*, early access, Mar. 29, 2024, doi: 10.1109/TIV.2024.3379989.

[31] K. Wang et al., "Parallel vision: An ACP-based approach to intelligent vision computing," *Acta Automatica Sinica*, vol. 42, no. 10, pp. 1490–1500, 2016.

[32] J. Wang et al., "Parallel vision for long-tail regularization: Initial results from IVFC autonomous driving testing," *IEEE Trans. Intell. Veh.*, vol. 7, no. 2, pp. 286–299, Jun. 2022.

[33] F.-Y. Wang, "Parallel control: A method for data-driven and computational control," *Acta Automatica Sinica*, vol. 39, no. 4, pp. 293–302, 2013.

[34] F.-Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 630–638, Sep. 2010.

[35] F.-Y. Wang, "The DAO to metacontrol for metasystems in metaverses: The system of parallel control systems for knowledge automation and control intelligence in CPSS," *IEEE/CAA J. Automatica Sinica*, vol. 9, no. 11, pp. 1899–1908, Nov. 2022.

[36] F.-Y. Wang, "New control paradigm for industry 5.0: From big models to foundation control and management," *IEEE/CAA J. Automatica Sinica*, vol. 10, no. 8, pp. 1643–1646, Aug. 2023.
[37] Z. Wang, C. Lv, and F. -Y. Wang, "A new era of intelligent vehicles and intelligent transportation systems: Digital twins and parallel intelligence," *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 2619–2627, Apr. 2023.
[38] Z. Hu, S. Lou, Y. Xing, X. Wang, D. Cao, and C. Lv, "Review and perspectives on driver digital twin and its enabling technologies for intelligent vehicles," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 417–440, Sep. 2022.
[39] P. Ye and F.-Y. Wang, *Parallel Population and Parallel Human*, 1st ed. Hoboken, NJ, USA: Wiley-IEEE Press, Jun. 2023.
[40] P. Ye and F. -Y. Wang, "Parallel population and parallel human–a cyber-physical social approach," *IEEE Intell. Syst.*, vol. 37, no. 5, pp. 19–27, Sep./Oct. 2022.
[41] L. Chen, Y. Zhang, B. Tian, Y. Ai, D. Cao, and F. -Y. Wang, "Parallel driving OS: A ubiquitous operating system for autonomous driving in CPSS," *IEEE Trans. Intell. Veh.*, vol. 7, no. 4, pp. 886–895, Dec. 2022.
[42] R. Qin et al., "Sora for computational social systems: From counterfactual experiments to artificiofactual experiments with parallel intelligence," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 2, pp. 1531–1550, Apr. 2024, doi: 10.1109/TCSS.2024.3373928.
[43] L. Fan et al., "Sora for foundation robots with parallel intelligence: Three world models, three robotic systems," *Front. Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 1–5, 2024.
[44] R. B. Adams, *The Science of Social Vision: The Science of Social Vision*, vol. 7. Oxford, U.K.: Oxford Univ. Press, 2011.
[45] Y. Xia, H. Yu, X. Wang, M. Jian, and F.-Y. Wang, "Relation-aware facial expression recognition," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 3, pp. 1143–1154, Sep. 2022.
[46] R. B. Adams Jr, D. N. Albohn, and K. Kveraga, "Social vision: Applying a social-functional approach to face and expression perception," *Curr. Directions Psychol. Sci.*, vol. 26, no. 3, pp. 243–248, 2017.
[47] F.-Y. Wang, "Social vision for parallel vision in CPSS: A new perspective for social computing and visual reasoning," *QAII Tech. Rep.*, Qiongdao, Aug. 2020.
[48] Q. Cao, H. Yu, P. Charisse, S. Qiao, and B. Stevens, "Is high-fidelity important for human-like virtual avatars in human computer interactions?," *Int. J. Netw. Dyn. Intell.*, vol. 2, no. 1, pp. 15–23, 2023.