TABLE 7: Detection rate (DR) and false positive rate (FPR) of the implemented defense on edge agents.

Agent Type	DR(%)		FPR(%)	
	Orthogonal	Harmful	Orthogonal	Harmful
Textual	95.2	97.8	3.1	2.5
Visual	92.7	90.6	3.8	4.9
Average	**93.95**	**94.2**	**3.45**	**3.7**

TABLE 8: Overall effectiveness of the T-Guard. SBR denotes successful blocking rate, and SBL denotes successful blocking latency.

Topology	ASR(%)	SBR(%)	SBL(s)
Chain	$6.2_{\downarrow 39.8}$	93.8	1.12
Star	$4.1_{\downarrow 50.9}$	95.9	0.78
Tree	$4.9_{\downarrow 47.1}$	95.1	0.91
Ring	$8.2_{\downarrow 39.8}$	91.8	1.54
Mesh	$2.6_{\downarrow 38.4}$	97.4	0.62
Average	$\mathbf{5.2}_{\downarrow 43.2}$	**94.8**	**0.99**

Protection of edge environments. As demonstrated by TOMA, edge environments are often exploited as entry vectors for attacks. To mitigate this risk, the cross-modal validator is deployed on edge agents to detect environment injection. We evaluated textual and visual edge agents over 1,000 runs, with 50% of the runs containing environment injection attacks. Table 7 summarizes the detection results. The validator achieves consistently high detection rates, averaging 93.95% for orthogonal and 94.2% for harmful attacks, demonstrating strong capability in identifying injected environments. Textual agents perform slightly better than visual agents, particularly under harmful attacks (97.8% vs. 90.6%). False positive rates remain low across all settings, with averages of 3.45% and 3.7% respectively, demonstrating the validator's reliability and minimal disruption to benign inputs.

Overall defense effectiveness. Table 8 summarizes the overall performance of T-Guard across 1,000 runs on five MAS topologies. Compared with the baseline (Section 7.3), the attack success rate decreases by 38.4% to 50.9%, resulting in a high average successful blocking rate (SBR) of 94.8%. These results demonstrate the strong effectiveness of T-Guard in mitigating environment injection attacks. Performance varies slightly across topologies. The mesh topology achieves the best defense results, with the highest blocking rate (97.4%) and lowest ASR (2.6%), likely due to its dense inter-agent connections that enhance detection and containment. In contrast, the chain and ring topologies exhibit relatively lower blocking rates, possibly due to their limited communication paths. Excluding MAS agent processing time, the average successful blocking latency (SBL) remains below 1 second.

System overhead analysis. Using the performance at 10 queries per second (QPS) as the baseline, we evaluated the efficiency of the proposed defense framework.

TABLE 9: System overhead introduced by the defense system under different deployment settings.

Deployment	Framework	TLR(%)	CLD(%)	MD(%)	LD(ms)
CMV-Only	LangManus	2.8	3.1	4.6	15.4
	OWL	2.1	2.8	3.9	13.7
	Magentic-One	2.5	3.3	4.3	17.2
	Average	**2.5**	**3.1**	**4.3**	**15.4**
TTE-Only	LangManus	1.9	1.5	2.6	8.9
	OWL	1.6	1.2	3.2	7.5
	Magentic-One	2.2	1.8	2.9	9.2
	Average	**1.9**	**1.5**	**2.9**	**8.5**
T-Guard	LangManus	8.3	6.9	10.8	31.6
	OWL	7.4	6.1	9.4	28.9
	Magentic-One	8.9	7.2	11.5	33.2
	Average	**8.2**	**6.7**	**10.6**	**31.2**
T-Guard (50QPS)	LangManus	11.2	10.4	17.3	58.1
	OWL	10.3	9.5	16.2	53.6
	Magentic-One	12.1	11.7	18.5	62.7
	Average	**11.2**	**10.5**	**17.3**	**58.1**

CMV-Only and TEE-Only indicate deployments with only the cross-modal validator and only the topology trust evaluator with its access control manager, respectively. TLR, CLD, MD, and LD denote throughput loss ratio, CPU load delta, memory delta, and latency delta.

As shown in Table 9, all deployment configurations exhibit low overhead across all metrics, including throughput loss ratio (TLR), CPU load delta (CLD), memory delta (MD), and latency delta (LD). Both the TTE-Only and CMV-Only deployments introduce minimal impact, with average TLR and CLD below 3%, MD around 4%, and LD under 20 ms. The complete T-Guard implementation also maintains low overhead under normal load, with average TLR and CLD of 8.2% and 6.7%, respectively, and latency increase of about 31 ms. Under high-load conditions (50 QPS), the overhead of T-Guard increases moderately (TLR 11.2%, CLD 10.5%, LD 58 ms) but remains within acceptable operational limits, demonstrating the scalability and practicality of the proposed defense framework.

Answer to RQ4: The deployed defense achieved a high average attack blocking rate of 94.8% while maintaining low system overhead, demonstrating the strong effectiveness and scalability of the proposed T-Guard framework.

8. Conclusion

In this paper, we propose a topology-aware multi-hop attack scheme targeting multi-agent systems. By modeling the dynamics of agent compromise propagation and designing multi-hop attack routes, our method effectively compromises MAS across diverse configurations. Experiments show a success rate of up to 78% across five MAS network topologies and three SOTA architectures. We further propose a conceptual defense framework, which achieves an average blocking rate of 94.8% with minimal overhead in prototype evaluations.

References

[1] H. Wu, Z. He, X. Zhang, X. Yao, S. Zheng, H. Zheng, and B. Yu, "Chateda: A large language model powered autonomous agent for eda," *Trans. Comp.-Aided Des. Integ. Cir. Sys.*, vol. 43, 2024.

[2] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, and S. Rass, "PentestGPT: Evaluating and harnessing large language models for automated penetration testing," in *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2024.

[3] Z. Zhao, D. Tang, C. Liu, L. Wang, Z. Zhang, H. Zhu, K. Chen, Q. Nie, and Y. Ji, "A large language model-based multi-agent manufacturing system for intelligent shopfloors," *Advanced Engineering Informatics*, vol. 69, 2026.

[4] A. Ghafarollahi and M. J. Buehler, "Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning," *Digital Discovery*, vol. 3, 2024.

[5] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, and H. W. Park, "Mdagents: An adaptive collaboration of llms for medical decision-making," *Advances in Neural Information Processing Systems*, vol. 37, 2024.

[6] E. Debenedetti, J. Zhang, M. Balunovic, L. Beurer-Kellner, M. Fischer, and F. Tramèr, "Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents," in *Proceedings of the Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2024.

[7] C. H. Wu, R. R. Shah, J. Y. Koh, R. Salakhutdinov, D. Fried, and A. Raghunathan, "Dissecting adversarial robustness of multimodal LM agents," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[8] Y. Zhang, T. Yu, and D. Yang, "Attacking vision-language computer agents via pop-ups," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[9] M. Yu, S. Wang, G. Zhang, J. Mao, C. Yin, Q. Liu, K. Wang, Q. Wen, and Y. Wang, "NetSafe: Exploring the topological safety of multi-agent system," in *Proceedings of the Findings of the Association for Computational Linguistics (Findings of ACL)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., 2025.

[10] J. tse Huang, J. Zhou, T. Jin, X. Zhou, Z. Chen, W. Wang, Y. Yuan, M. Lyu, and M. Sap, "On the resilience of LLM-based multi-agent collaboration with faulty agents," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

[11] P. He, Y. Lin, S. Dong, H. Xu, Y. Xing, and H. Liu, "Red-teaming llm multi-agent systems via communication attacks," in *Proceedings of the Findings of the Association for Computational Linguistics (Findings of ACL)*, 2025.

[12] L. Wang, W. Wang, S. Wang, Z. Li, Z. Ji, Z. Lyu, D. Wu, and S.-C. Cheung, "Ip leakage attacks targeting llm-based multi-agent systems," *arXiv preprint arXiv:2505.12442*, 2025.

[13] T. Sternak, D. Runje, D. Granoša, and C. Wang, "Automating prompt leakage attacks on large language models using agentic approach," *arXiv preprint arXiv:2502.12630*, 2025.

[14] Y. Wang, M. Zhang, J. Sun, C. Wang, M. Yang, H. Xue, J. Tao, R. Duan, and J. Liu, "Mirage in the eyes: hallucination attack on multi-modal large language models with only attention sink," in *Proceedings of the USENIX Conference on Security Symposium (USENIX Security)*, 2025.

[15] D. Kong, H. Peng, Y. Zhang, L. Zhao, Z. Xu, S. Lin, C. Lin, and M. Han, "Web fraud attacks against llm-driven multi-agent systems," *arXiv preprint arXiv:2509.01211*, 2025.

[16] K. Gao, T. Pang, C. Du, Y. Yang, S.-T. Xia, and M. Lin, "Denial-of-service poisoning attacks against large language models," *arXiv preprint arXiv:2410.10760*, 2024.

[17] S. Wang, G. Zhang, M. Yu, G. Wan, F. Meng, C. Guo, K. Wang, and Y. Wang, "G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems," in *Proceedings of the Findings of the Association for Computational Linguistics (Findings of ACL)*, 2025.

[18] S. Dong, S. Xu, P. He, Y. Li, J. Tang, T. Liu, H. Liu, and Z. Xiang, "A practical memory injection attack against llm agents," *arXiv preprint arXiv:2503.03704*, 2025.

[19] Z. Chen, Z. Xiang, C. Xiao, D. Song, and B. Li, "Agentpoison: red-teaming llm agents via poisoning memory or knowledge bases," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[20] B. Wang, W. He, S. Zeng, Z. Xiang, Y. Xing, J. Tang, and P. He, "Unveiling privacy risks in LLM agent memory," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., 2025.

[21] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian *et al.*, "Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[22] Y. Kong, J. Ruan, Y. Chen, B. Zhang, T. Bao, S. Shiwei, X. Hu, H. Mao, Z. Li, X. Zeng *et al.*, "Tptu-v2: Boosting task planning and tool usage of large language model-based agents in real-world industry systems," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[23] T. Schick, J. Dwivedi-Yu, R. Dessí, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: language models can teach themselves to use tools," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[24] Y. Yue, G. Zhang, B. Liu, G. Wan, K. Wang, D. Cheng, and Y. Qi, "Masrouter: Learning to route llms for multi-agent systems," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[25] MetaGPT, "Metagpt: The multi-agent framework," 2025. [Online]. Available: https://www.deepwisdom.ai/

[26] Darwin-lfl, "Langmanus," 2025. [Online]. Available: https://github.com/Darwin-lfl/langmanus?tab=readme-ov-file

[27] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu *et al.*, "Autogen: Enabling next-gen llm applications via multi-agent conversations," in *Proceedings of the First Conference on Language Modeling (COLM)*, 2024.

[28] CAMEL-AI, "Camel," 2025. [Online]. Available: https://www.camel-ai.org/

[29] G. Zhang, Y. Yue, Z. Li, S. Yun, G. Wan, K. Wang, D. Cheng, J. X. Yu, and T. Chen, "Cut the crap: An economical communication pipeline for LLM-based multi-agent systems," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[30] H. Zhou, X. Wan, R. Sun, H. Palangi, S. Iqbal, I. Vulić, A. Korhonen, and S. Ö. Arık, "Multi-agent design: Optimizing agents with better prompts and topologies," *arXiv preprint arXiv:2502.02533*, 2025.

[31] Z. Wang, Y. Wang, X. Liu, L. Ding, M. Zhang, J. Liu, and M. Zhang, "Agentdropout: Dynamic agent elimination for token-efficient and high-performance llm-based multi-agent collaboration," 2025.

[32] J. Yang, M. Zhang, Y. Jin, H. Chen, Q. Wen, L. Lin, Y. He, W. Xu, J. Evans, and J. Wang, "Topological structure learning should be a research priority for llm-based multi-agent systems," *arXiv preprint arXiv:2505.22467*, 2025.

[33] X. Shen, Y. Liu, Y. Dai, Y. Wang, R. Miao, Y. Tan, S. Pan, and X. Wang, "Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds., 2025.

[34] K. Yang, Y. Liu, S. Chaudhary, R. Fakoor, P. Chaudhari, G. Karypis, and H. Rangwala, "Agentoccam: A simple yet strong baseline for LLM-based web agents," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[35] Z. Liao, L. Mo, C. Xu, M. Kang, J. Zhang, C. Xiao, Y. Tian, B. Li, and H. Sun, "EIA: ENVIRONMENTAL INJECTION ATTACK ON GENERALIST WEB AGENTS FOR PRIVACY LEAKAGE," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

[36] W. Luo, S. Dai, X. Liu, S. Banerjee, H. Sun, M. Chen, and C. Xiao, "Agrail: A lifelong agent guardrail with effective and adaptive safety detection," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

[37] X. Gu, X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin, "Agent smith: a single image can jailbreak one million multimodal llm agents exponentially fast," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[38] Z. Chen, Z. Xiang, C. Xiao, D. Song, and B. Li, "Agentpoison: Red-teaming LLM agents via poisoning memory or knowledge bases," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

[39] X. Ma, Y. Wang, Y. Yao, T. Yuan, A. Zhang, Z. Zhang, and H. Zhao, "Caution for the environment: Multimodal LLM agents are susceptible to environmental distractions," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., 2025.

[40] M. Russinovich, A. Salem, and R. Eldan, "Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack," in *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2025.

[41] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, ""do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

[42] T. Xie, D. Zhang, J. Chen, X. Li, S. Zhao, R. Cao, T. J. Hua, Z. Cheng, D. Shin, F. Lei, Y. Liu, Y. Xu, S. Zhou, S. Savarese, C. Xiong, V. Zhong, and T. Yu, "Osworld: benchmarking multimodal agents for open-ended tasks in real computer environments," in *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2024.

[43] W. Yu, K. Hu, T. Pang, C. Du, M. Lin, and M. Fredrikson, "Infecting LLM agents via generalizable adversarial attack," in *Proceedings of the Conference on Neural Information Processing Systems Workshop (NeurIPS Workshop)*, 2025.

[44] A. Amayuelas, X. Yang, A. Antoniades, W. Hua, L. Pan, and W. Y. Wang, "Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate," in *Proceedings of the Findings of the Association for Computational Linguistics (EMNLP)*, 2024.

[45] T. Ju, Y. Wang, X. Ma, P. Cheng, H. Zhao, Y. Wang, L. Liu, J. Xie, Z. Zhang, and G. Liu, "Flooding spread of manipulated knowledge in llm-based multi-agent communities," *arXiv preprint arXiv:2407.07791*, 2024.

[46] OpenAI, "Gpt-4v(ision)," 2025. [Online]. Available: https://openai.com/index/gpt-4v-system-card/

[47] Aliyun, "Qwen-vl-max," 2025. [Online]. Available: https://modelscope.cn/studios/qwen/Qwen-VL-Max

[48] Volcengine, "Doubao-vision-pro," 2025. [Online]. Available: https://www.volcengine.com/product/doubao

[49] A. Fourney, G. Bansal, H. Mozannar, C. Tan, E. Salinas, Erkang, Zhu, F. Niedtner, G. Proebsting, G. Bassman, J. Gerrits, J. Alber, P. Chang, R. Loynd, R. West, V. Dibia, A. Awadallah, E. Kamar, R. Hosn, and S. Amershi, "Magentic-one: A generalist multi-agent system for solving complex tasks," *arXiv preprint arXiv:2411.04468*, 2024.

[50] camel ai.org, "Owl: Optimized workforce learning for general multi-agent assistance in real-world task automation," 2025. [Online]. Available: https://github.com/camel-ai/owl

[51] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. YU, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, and O. Levy, "LIMA: Less is more for alignment," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[52] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[53] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[54] Y. Cao, N. Gu, X. Shen, D. Yang, and X. Zhang, "Defending large language models against jailbreak attacks through chain of thought prompting," in *Proceedings of the International Conference on Networking and Network Applications (NaNA)*, 2024.

[55] J. Mao, F. Meng, Y. Duan, M. Yu, X. Jia, J. Fang, Y. Liang, K. Wang, and Q. Wen, "Agentsafe: Safeguarding large language model-based multi-agent systems via hierarchical data management," *arXiv preprint arXiv:2503.04392*, 2025.

[56] B. Hui, H. Yuan, N. Gong, P. Burlina, and Y. Cao, "Pleak: Prompt leaking attacks against large language model applications," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024.

[57] C. McCauley, K. Yeung, J. Martin, and K. Schulz, "Novel universal bypass for all major llms," 2025. [Online]. Available: https://hiddenlayer.com/innovation-hub/novel-universal-bypass-for-all-major-llms/

[58] Z. Zhang, P. Zhao, D. Ye, and H. Wang, "Enhancing jailbreak attacks on llms via persona prompts," *arXiv preprint arXiv:2507.22171*, 2025.

[59] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[60] H. Xiao, Y. Sun, Z. Duan, Y. Huo, J. Liu, M. Luo, Y. Li, and Y. Zhang, "A study of model iterations of fitts' law and its application to human–computer interactions," *Applied Sciences*, vol. 14, 2024.

[61] OpenAI, "Gpt-4o," 2025. [Online]. Available: https://platform.openai.com/docs/models/gpt-4o

[62] Anthropic, "claude-3-7-sonnet," 2025. [Online]. Available: https://www.anthropic.com/news/claude-3-7-sonnet