Appendix A
Proof of Policy Invariance

The following proposition shows the one-step potential-based difference reward guarantees policy invariance and establishes the convergence of our method to a locally optimal policy theoretically.

Proposition 1. *By introducing the potential-based difference reward shaping in our framework, the policy invariant still holds and doesn't influence the convergence.*

Proof. For Agent 1, let $Q_1(s_1, a_1, a_2, a_3) = Q_1(s_1, \boldsymbol{a})$ be the original Q-function, and $\tilde{Q}_1(s_1, a_1, a_2, a_3) = \tilde{Q}_1(s_1, \boldsymbol{a})$ be the modified Q-function with the reward shaping method. We have:

$$Q_1(s_1, \boldsymbol{a}) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} \gamma^t R_1^t\right], \tag{31}$$

$$\begin{aligned}\tilde{Q}_1(s_1, \boldsymbol{a}) =& \mathbb{E}_{\boldsymbol{\pi}}\Bigg[\sum_{t=0}^{T} \gamma^t \Big(R_1^t + \gamma \mathbb{E}_{\tilde{a}_1^{t+1}}\left[R_1(s_1^{t+1}, a_2^{t+1}, a_3^{t+1}, \tilde{a}_1^{t+1})\right] \\ & - \mathbb{E}_{\tilde{a}_1^t}\left[R_1(s_1^t, a_2^t, a_3^t, \tilde{a}_1^t)\right]\Big)\Bigg].\end{aligned} \tag{32}$$

where the expectation $\mathbb{E}_{\boldsymbol{\pi}}$ is with respect to the state-action distribution induced by the joint policy $\{\pi_1, \pi_2, \pi_3\}$ Then, $\tilde{Q}_1(s_1, \boldsymbol{a})$ can be further:

$$\begin{aligned}\tilde{Q}_1 =& \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} \gamma^t R_1^t\right] + \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=1}^{T} \gamma^t \mathbb{E}_{\tilde{a}_1^t \sim \pi_1}\left[R_1(s_1^t, a_2^t, a_3^t, \tilde{a}_1^t)\right]\right] \\ & - \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} \gamma^t \mathbb{E}_{\tilde{a}_1^t \sim \pi_1}\left[R_1(s_1^t, a_2^t, a_3^t, \tilde{a}_1^t)\right]\right],\end{aligned} \tag{33}$$

$$= \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{t=0}^{T} \gamma^t R_1^t\right] - \mathbb{E}_{\boldsymbol{\pi}}\left[\gamma^0 \mathbb{E}_{\tilde{a}_1^0}\left[R_1(s_1^0, a_2^0, a_3^0, \tilde{a}_1^0)\right]\right], \tag{34}$$

$$= Q_1 - \mathbb{E}_{\tilde{a}_1^0}\left[R_1(s_1^0, a_2^0, a_3^0, \tilde{a}_1^0)\right]. \tag{35}$$

Note that the expectation $\mathbb{E}_{\boldsymbol{\pi}}$ is $\mathbb{E}_{(s_1^t, s_2^t, a_1^t, a_2^t) \sim \boldsymbol{\pi}}$, considering the specific initial state s^0, this expectation becomes deterministic, eliminating variability in trajectories induced by $\boldsymbol{\pi}$. Therefore, the outer expectation can be omitted. Similar arguments apply for $Q_2, \tilde{Q}_2$ and $Q_3, \tilde{Q}_3$. For brevity in the ensuing discussion, let's denote $\mathcal{E}_1(s_1^0, a_2^0, a_3^0) = \mathbb{E}_{\tilde{a}_1^0}[R_1(s_1^0, a_2^0, a_3^0, \tilde{a}_1^0)]$, and similar to $\mathcal{E}_2(s_2^0, a_1^0, a_3^0)$, and $\mathcal{E}_3(s_3^0, a_1^0, a_2^0)$.

Let $\boldsymbol{\theta}$ be the parameters of joint actor policies $\theta_1, \theta_2, \theta_3$, the multi-agent policy gradient with the reward shaping is [16]:

$$\begin{aligned}g = \mathbb{E}_{\boldsymbol{\pi}}[&\nabla_{\boldsymbol{\theta}} \log \pi_1(a_1^t|s_1^t)\tilde{Q}_1(s_1, \boldsymbol{a}) \\ &+ \nabla_{\boldsymbol{\theta}} \log \pi_2(a_2^t|s_2^t)\tilde{Q}_2(s_2, \boldsymbol{a}) \\ &+ \nabla_{\boldsymbol{\theta}} \log \pi_3(a_3^t|s_3^t)\tilde{Q}_3(s_3, \boldsymbol{a})],\end{aligned} \tag{36}$$

$$\begin{aligned}= \mathbb{E}_{\boldsymbol{\pi}}[&\nabla_{\boldsymbol{\theta}} \log \pi_1(a_1^t|s_1^t)(Q_1(s_1, \boldsymbol{a}) - \mathcal{E}_1(s_1^0, a_2^0, a_3^0)) \\ &+ \nabla_{\boldsymbol{\theta}} \log \pi_2(a_2^t|s_2^t)(Q_2(s_2, \boldsymbol{a}) - \mathcal{E}_2(s_2^0, a_1^0, a_3^0)) \\ &+ \nabla_{\boldsymbol{\theta}} \log \pi_3(a_3^t|s_3^t)(Q_3(s_3, \boldsymbol{a}) - \mathcal{E}_3(s_3^0, a_1^0, a_2^0))].\end{aligned} \tag{37}$$

Let's focus on the second term, and let $d^{\boldsymbol{\pi}}(s_i)$ be the stationary state distribution [11], and $-i$ be the other agent indicator:

$$\begin{aligned}g_p = - \mathbb{E}_{\boldsymbol{\pi}}[&\nabla_{\boldsymbol{\theta}} \log \pi_1(a_1^t|s_1^t)\mathcal{E}_1(s_1^0, a_2^0, a_3^0) \\ &+ \nabla_{\boldsymbol{\theta}} \log \pi_2(a_2^t|s_2^t)\mathcal{E}_2(s_2^0, a_1^0, a_3^0) \\ &+ \nabla_{\boldsymbol{\theta}} \log \pi_3(a_3^t|s_3^t)\mathcal{E}_3(s_3^0, a_1^0, a_2^0)]\end{aligned} \tag{38}$$

$$= - \mathbb{E}_{\boldsymbol{\pi}}[\sum_{i=1}^{3} \nabla_{\boldsymbol{\theta}} \log \pi_i(a_i^t|s_i^t)\mathcal{E}_i(s_i^0, a_{-i}^0)], \tag{39}$$

$$\begin{aligned}= - \sum_{i=1}^{3}\sum_{t=0}^{T}\sum_{s_i} d^{\boldsymbol{\pi}}(s_i^t) \sum_{a_{-i}} \pi_{-i}(a_{-i}^t|s_{-i}^t) \cdot \\ \sum_{a_i} \nabla_{\boldsymbol{\theta}} \pi_i(a_i^t|s_i^t)\mathcal{E}_i(s_i^0, a_{-i}^0),\end{aligned} \tag{40}$$

$$\begin{aligned}&= - \sum_{i=1}^{3}\sum_{t=0}^{T}\sum_{s_i} d^{\boldsymbol{\pi}}(s_i^t) \sum_{a_{-i}} \pi_{-i}(a_{-i}^t|s_{-i}^t)\mathcal{E}_i(s_i^0, a_{-i}^0)\nabla_{\boldsymbol{\theta}} 1 \\ &= 0.\end{aligned} \tag{41}$$

Therefore, we have:

$$\mathbb{E}_{\boldsymbol{\pi}}[\sum_{i=1}^{3} \nabla_{\boldsymbol{\theta}} \log \pi_i(a_i|s_i)\tilde{Q}_i(s_i, a_i, a_{-i})] \tag{42}$$

$$= \mathbb{E}_{\boldsymbol{\pi}}[\sum_{i=1}^{3} \nabla_{\boldsymbol{\theta}} \log \pi_i(a_i|s_i)Q_i(s_i, a_i, a_{-i})]. \tag{43}$$

This demonstrates this reward-shaping technique in our framework doesn't inherently change the expected gradient. Moreover, this proof remains valid for continuous actions. By employing Gaussian policies, we can treat the action as a Gaussian distribution, which only changes the calculation of $\nabla_\theta \log_\pi(a|s)$, and in the proof, we can replace $\sum_a$ with $\int_a$.

Then the expected policy gradient is:

$$\begin{aligned}g' =& \mathbb{E}_{\boldsymbol{\pi}}[\sum_{i=1}^{3} \nabla_{\boldsymbol{\theta}} \log \pi_i(a_i^t|s_i^t)Q_i(s_i, a_i, a_{-i})] \\ =& \mathbb{E}_{\boldsymbol{\pi}}[\nabla_{\boldsymbol{\theta}} \log \prod_{i=1}^{3} \pi_i(a_i^t|s_i^t)Q_i(s_i, a_i, a_{-i})] \\ =& \mathbb{E}_{\boldsymbol{\pi}}[\nabla_{\boldsymbol{\theta}} \log \boldsymbol{\pi}(\boldsymbol{a}^t|\boldsymbol{s}^t)Q_i(s_i, a_i, a_{-i})]\end{aligned} \tag{44}$$

where $\boldsymbol{\pi}(\boldsymbol{a}^t|\boldsymbol{s}^t) = \pi_1(a_1^t|s_1^t) \cdot \pi_2(a_2^t|s_2^t) \cdot \pi_3(a_3^t|s_3^t)$. Konda and Tsitsiklis [33] showed that a single-agent actor-critic with a given gradient converges to a local maximum of expected return under specific conditions, and the most important one is the policy is differentiable. Given that our policy gradient (i.e., the joint learner is broken down into separate, independent actors) in equation (44) is differentiable and the policy parameterization doesn't hinder convergence, our actor-critic model's convergence remains intact and assured. □

Appendix B
Numerical Setting

Under a single EVS, we support 5 to 9 users with varying settings. The total transmission power of the EVS is denoted as P Watts, and the EVS operates at a frequency of 10 GHz. We set the cycles per bit to a constant value of 150 [18]. Each pixel is represented using 16 bits [34], and the compression rate is randomly sampled from a uniform distribution ranging between 800 and 1000. We consider four different resolution settings: 720p (1280 × 720 pixels), 1080p (1920 × 1080 pixels), 2k (2560 × 1440 pixels), and 3k (3840 × 1920 pixels). The number of total frames per second T is fixed at 100, and the available bandwidth per channel is 10^6 Hz. We model the channel gain as $h_n^t = \sqrt{\beta_n^t} g_n^t$. For small-scale fading, we use Rician fading, where $g_n^t = \sqrt{\frac{K}{K+1}}\bar{g}_n^t + \sqrt{\frac{1}{K+1}}\tilde{g}_n^t$. Here, $\bar{g}_n^t$ represents the Line-Of-Sight (LOS) component, and $\tilde{g}_n^t$ characterizes the Non-LOS (NLOS) component, both following a standard complex normal distribution $\mathcal{CN}(0,1)$. Large-scale fading is modeled as $\beta_n^t = \beta_0(L_n)^{-\alpha}$, where L_n denotes the distance between the nth VU and the server. Here, β_0 represents the channel gain at the reference distance of $L_0 = 1$ m, and α is the path-loss exponent. For our simulations, we set α to 2 and the Rician factor K to 3. All experiments utilize the multiple same global random seeds, and we include error margins in our results.

References

[1] S. Kum, S. Oh, J. Yeom, and J. Moon, "Optimization of edge resources for deep learning application with batch and model management," *Sensors*, vol. 22, no. 17, p. 6717, 2022.

[2] Y. Li, C. Dou, Y. Wu, W. Jia, and R. Lu, "Noma assisted two-tier vr content transmission: A tile-based approach for qoe optimization," *IEEE Transactions on Mobile Computing*, 2023.

[3] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for ai-enabled wireless networks: A tutorial," *IEEE Communications Surveys & Tutorials*, 2021.

[4] B. W. Wah, X. Su, and D. Lin, "A survey of error-concealment schemes for real-time audio and video transmissions over the internet," in *Proceedings International Symposium on Multimedia Software Engineering*. IEEE, 2000, pp. 17–24.

[5] A. M. Seid, J. Lu, H. N. Abishu, and T. A. Ayall, "Blockchain-enabled task offloading with energy harvesting in multi-uav-assisted iot networks: A multi-agent drl approach," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 12, pp. 3517–3532, 2022.

[6] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in d2d networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 1363–1378, 2020.

[7] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for noma with deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2200–2210, 2019.

[8] D. Guo, L. Tang, X. Zhang, and Y.-C. Liang, "Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 124–13 138, 2020.

[9] W. Yu, T. J. Chua, and J. Zhao, "User-centric heterogeneous-action deep reinforcement learning for virtual reality in the metaverse over wireless networks," *IEEE Transactions on Wireless Communications*, 2023.

[10] T. Li, K. Zhu, N. C. Luong, D. Niyato, Q. Wu, Y. Zhang, and B. Chen, "Applications of multi-agent reinforcement learning in future internet: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 2, pp. 1240–1279, 2022.

[11] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[12] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, 2022.

[13] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[14] W. Yu, T. J. Chua, and J. Zhao, "Asynchronous hybrid reinforcement learning for latency and reliability optimization in the metaverse over wireless communications," *IEEE Journal on Selected Areas in Communications*, 2023.

[15] H. Van Seijen, M. Fatemi, J. Romoff, R. Laroche, T. Barnes, and J. Tsang, "Hybrid reward architecture for reinforcement learning," *Advances in Neural Information Processing Systems*, 2017.

[16] Y. Li, G. Xie, and Z. Lu, "Difference advantage estimation for multi-agent policy gradients," in *International Conference on Machine Learning*. PMLR, 2022, pp. 13 066–13 085.

[17] A. Hanyu, Y. Kawamoto, and N. Kato, "Adaptive channel selection and transmission timing control for simultaneous receiving and sending in relay-based uav network," *IEEE Transactions on Network Science and Engineering*, 2020.

[18] C. You, Y. Zeng, R. Zhang, and K. Huang, "Asynchronous mobile-edge computation offloading: Energy-efficient resource management," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7590–7605, 2018.

[19] M. Kazemi, M. Ghanbari, and S. Shirmohammadi, "A review of temporal video error concealment techniques and their suitability for hevc and vvc," *Multimedia Tools and Applications*, 2021.

[20] Y. Qi and M. Dai, "The effect of frame freezing and frame skipping on video quality," in *2006 international conference on intelligent information hiding and multimedia*. IEEE, 2006, pp. 423–426.

[21] H. T. Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang, "A subjective study on qoe of 360 video for vr communication," in *2017 IEEE 19th international workshop on multimedia signal processing (MMSP)*. IEEE, 2017.

[22] L. Liberti, "Undecidability and hardness in mixed-integer nonlinear programming," *RAIRO-Operations Research*, 2019.

[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT press, 2018.

[24] Y. Wang, B. Han, T. Wang, H. Dong, and C. Zhang, "Off-policy multi-agent decomposed policy gradients," 2020.

[25] P. W. Glynn and D. L. Iglehart, "Importance sampling for stochastic simulations," *Management Science*, 1989.

[26] V. Gullapalli and A. G. Barto, "Shaping as a method for accelerating reinforcement learning," in *Proceedings of the 1992 IEEE international symposium on intelligent control*. IEEE, 1992, pp. 554–559.

[27] S. Proper and K. Tumer, "Modeling difference rewards for multiagent learning." in *AAMAS*, 2012.

[28] F. Schweitzer, *Modeling complexity in economic and social systems*. World scientific, 2002.

[29] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *ICML*, 1999.

[30] E. Wiewiora, G. W. Cottrell, and C. Elkan, "Principled methods for advising reinforcement learning agents," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003.

[31] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[32] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, 2015.

[33] V. Konda and J. Tsitsiklis, "Actor-critic algorithms," *Advances in neural information processing systems*, 1999.

[34] M. Chen, W. Saad, and C. Yin, "Virtual reality over wireless networks: Quality-of-service model and learning-based resource management," *IEEE Transactions on Communications*, 2018.