

Bi-level Implicit Adversarial (BIA)

Best-Response Gradient

$$\widetilde{\mathcal{L}}_{total} \quad G_{BR} \quad \text{Best-Response Gradient}$$

$$\begin{array}{c|c}
\widetilde{\mathcal{L}}_{total} & G_{BR} \\
\hline
\text{GMP} \\
\text{IDR} \\
\hline
\text{Hyper}. \quad \nabla_{\varsigma}\widetilde{\mathcal{L}}_{total} + \nabla_{\omega}\widetilde{\mathcal{L}}_{total} \nabla_{\varsigma}\omega
\end{array}$$

$$\begin{array}{c|c}
\widetilde{\mathcal{L}}_{total} \\
\hline
\text{GMP} \\
\hline
\text{IDR} \\
\hline
\text{Hyper}. \quad \omega(\varrho) \\
\hline
\omega(\varsigma) \\
\hline
\omega_{k+1} \quad \nabla_{\omega}\mathcal{L}_{adv}$$

$$\begin{array}{c|c}
\overline{\mathcal{L}}_{adv} \\
\hline
\end{array}$$

$$\begin{array}{c|c}
\overline{\mathcal{L}_{adv} \\
\hline
\end{array}$$

$$\begin{array}{c|c}
\overline{\mathcal{L}}_{adv} \\
\hline
\end{array}$$