A Expanding GMM from Conditional Moment Restriction

In order to show what connectivity exists between GMM and moment restriction, we start from Equation (1) of conditional moment restriction in our manuscript as:

$$\mathbb{E}_T[\psi_T(h) \mid Z] = \int_{t\in T} \psi_T(h)\mathrm{d}\mathbb{P}(T = t \mid Z) = \mathbf{0}, \tag{1}$$

To generate infinite moments by the test function g regarding numerous case of instrumental variables, GMM selects a moment with the hypothesis model and test function, which can be written as:

$$m(h, g) = \mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)]. \tag{2}$$

Here, this moment can be expressed with conditional moment restriction, then it satisfies zero as follows:

$$\begin{aligned}
\mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)] &= \int_{z\in Z, t\in T} \psi_t(h) \cdot g(z)\mathrm{d}\mathbb{P}(Z = z, T = t), \\
&= \int_{z\in Z, t\in T} \psi_t(h)\mathrm{d}\mathbb{P}(T = t \mid Z = z) \cdot g(z)\mathrm{d}\mathbb{P}(Z = z) \\
&= \int_{z\in Z} \mathbb{E}_T[\psi_T(h) \mid Z = z] \cdot g(z)\mathrm{d}\mathbb{P}(Z = z) \\
&= \mathbb{E}_Z[\underbrace{\mathbb{E}_T[\psi_T(h) \mid Z]}_{\text{conditional moment}} \cdot g(Z)] \\
&= \mathbb{E}_Z[\mathbf{0} \cdot g(Z)] \qquad (\because \text{Eq. (1)}) \\
&= 0.
\end{aligned} \tag{3}$$

From this proof, we can infer that once GMM achieves a reduction of moment magnitude, then it successfully expands conditional moment restriction to perform infinite moment restriction, where the intractable infinite number of moments generated by g is replaced with (infinite-dimensional) non-parametric test function g such as DNNs.

B Realizing Generalization Gap

We employ *Taylor Expansion* and *Identity mapping* to realize the generalization gap $\phi(g^*, g)$ as described in Equation (8) in our manuscript. By using them, we elaborate the equation to feasibly calculate the generalization gap.

B.1 Taylor Expansion

When we make use of AMR-GMM for adversarial instrumental variable regression, there happens generalization gap between ideal $m(h^*, g^*)$ and empirical moments $m(h^*, g)$ for test functions due to the absence of regularizing the direction for learning a test function on maximum moment restriction. Here, the generalization gap can be written as follows:

$$\phi(g^*, g) = m(h^*, g^*) - m(h^*, g), \tag{4}$$

where we suppose the empirical moment has sufficiently converged generalized residual function $\psi^{\Omega}_{T'|Z}(h^*)$ to a small constant value from the best estimator h^*, which can be written as:

$$m(h^*, g) = \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot (\Omega \circ g)(Z)]. \tag{5}$$

Note that, the generalized residual function $\psi^{\Omega}_{T'|Z}(h^*)$ of the ideal moments $m(h^*, g^*)$ is either a small constant value. From their assumption of moments, we can indicate the generalization gap of Eq. (5) with simple subtraction terms with inner product on the small constant of the generalized residual function, which can be written as follows:

$$\phi(g^*, g) = \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{(\Omega \circ g^*)(Z) - (\Omega \circ g)(Z)\}]. \tag{6}$$

In this spot, we unpack the log-likelihood function Ω by using *Taylor Expansion* that it satisfies:

$$\Omega(\omega + \Delta\omega) \approx \Omega(\omega) + \Omega'(\omega) \otimes \Delta\omega, \tag{7}$$

with a vector-valued function $\Omega : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^K$ (class number K) and its derivation function $\Omega' : \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{K \times HWC}$. In addition, the operator $\otimes$ denotes dimension squeeze (*i.e,* vectorize) and multiplication due to its tensor dimension of $\Delta\omega \in \mathbb{R}^{H \times W \times C}$ such that it satisfies $a \otimes b := a \times \text{Vec}(b)$. Then, Taylor Expansion of the log-likelihood function Ω in Eq. (7) can be also applied to a simple setup $\omega = \mathbf{0}$ for the following equation:

$$\Omega(\mathbf{0} + \Delta\omega) \approx \Omega(\omega = \mathbf{0}) + \Omega'(\omega = \mathbf{0}) \otimes \Delta\omega. \tag{8}$$

Eventually, the generalization gap can be possibly approximated by the following equation:

$$\begin{aligned}
\phi(g^*, g) &= \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{(\Omega \circ g^*)(Z) - (\Omega \circ g)(Z)\}] \\
&= \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{\underbrace{\Omega(\omega = \mathbf{0}) + \Omega'(\omega = \mathbf{0}) \otimes g^*(Z)}_{(\Omega \circ g^*)(Z)} - \underbrace{(\Omega(\omega = \mathbf{0}) + \Omega'(\omega = \mathbf{0}) \otimes g(Z))}_{(\Omega \circ g)(Z)}\}] \\
&= \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{\Omega'(\omega = \mathbf{0}) \otimes (g^*(Z) - g(Z))\}].
\end{aligned} \tag{9}$$

B.2 Identity Mapping

However, once we directly compute this equation, we will take a striking computational burden from the repeated procedure of tensor derivation Ω' and its dimension squeeze and multiplication $\otimes$. To be specific, computing the generalization gap in Eq. (9) naively induces a computational complexity $\mathcal{O}(K^2H^2W^2C^2)$, at least, per one iteration.

Therefore, we should practically compute the generalization gap and get its fast convergence. Here, *localized Rademacher* enables the operator $\otimes$ and the two weighted factors (ψ^{Ω}, Ω') for $g^*(Z) - g(Z)$ to be ignored in computing the generalization gap, and it allows the generalization gap to be uniform bound with the convergence rate λ, such that

$$|\phi(g^*, g)| \approx \sqrt{\lambda}|\mathbb{E}_Z[g^*(Z) - g(Z)]|, \tag{10}$$

where its complexity is even $\mathcal{O}(1)$ to our satisfaction. Then, we use an elementary algebraic trick with identity mapping $\mathcal{I}$ to approximate tight upper bound of the generalization gap by triangle inequality for its feasible computation within reach as follows:

$$\begin{aligned} |\phi(g^*, g)| = |m(h^*, g^*) - m(h^*, g)| &= |\underbrace{m(h^*, g^*) - m(h^*, \mathcal{I})}_{\phi(g^*,\mathcal{I})} + \underbrace{m(h^*, \mathcal{I}) - m(h^*, g)}_{\phi(\mathcal{I},g)}| \\ &\leq |\phi(g^*, \mathcal{I})| + |\phi(\mathcal{I}, g)| \approx \sqrt{\lambda}|\mathbb{E}_Z[g^*(Z) - Z]| + \sqrt{\lambda}|\mathbb{E}_Z[Z - g(Z)]|, \end{aligned} \tag{11}$$

where $|\phi(g^*, \mathcal{I})|$ in the upper bound is constant value with respect to g. Once we subtract $|\phi(g^*, \mathcal{I})|$ to the above inequality, we can get the supremum value of $|\phi(g^*, g)| - |\phi(g^*, \mathcal{I})|$, as follows:

$$\sup_{g\in\mathcal{G}} |\phi(g^*, g)| - |\phi(g^*, \mathcal{I})| \leq \sup_{g\in\mathcal{G}} |\phi(\mathcal{I}, g)|. \tag{12}$$

Here, we suppose that the absence of regularizing test function forges a significant difference between a feature variation (*i.e,* instrument) Z and its counterfactuals (*i.e,* test function) $g(Z)$. This postulation implies that the output of test function strays from the possible feature bound and the infimum value $\inf_{g\in\mathcal{G}} |\phi(\mathcal{I}, g)| \approx \sqrt{\lambda}|\mathbb{E}_Z[Z - g(Z)]|$ becomes large enough, thus we can realign Eq. (12) with total range of $|\phi(\mathcal{I}, g)|$, which can be written as follows:

$$\sup_{g\in\mathcal{G}} |\phi(g^*, g)| - |\phi(g^*, \mathcal{I})| \leq \inf_{g\in\mathcal{G}} |\phi(\mathcal{I}, g)| \leq |\phi(\mathcal{I}, g)| \leq \sup_{g\in\mathcal{G}} |\phi(\mathcal{I}, g)|. \tag{13}$$

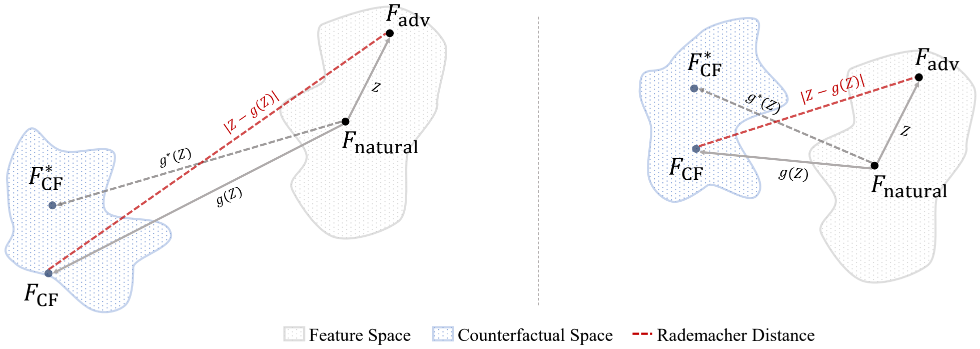
From this inequality, we can show the existence of the triangle inequality $\sup_{g\in\mathcal{G}} |\phi(g^*, g)| \leq |\phi(g^*, \mathcal{I})| + \inf_{g\in\mathcal{G}} |\phi(\mathcal{I}, g)|$ described in our manuscript. In addition, as our manuscript has already explained the connection between the generalization gap and Rademacher complexity, such that $\sup_{g\in\mathcal{G}} |\phi(g^*, g)| = 2b\mathcal{R}(\mathcal{G})$, we eventually get an indirect method to reduce Rademacher complexity once minimizing $|\phi(\mathcal{I}, g)|$ effortlessly. Then, we practically optimize the squared $|\phi(\mathcal{I}, g)|^2$, namely *localized Rademacher regularizer*, together with the main objective of AMR-GMM to maintain a low generalization gap for getting rich test function, which can be written as follows:

$$\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h) \cdot (\Omega \circ g)(Z)] - \lambda|\mathbb{E}_Z[Z - g(Z)]|^2. \tag{14}$$

This ensures the successful achievement of AMR-GMM where the output of test function does not deviate appreciably from the feature variation Z, so that it enables to find out the worst-case counterfactuals within adversarial feature bound. Appendix B.3 describes the triangle inequality clearly with figure and delineates how sufficiently rich test function works, through the lens of empirical evidence by conducting AMR-GMM without the localized Rademacher regularizer.

B.3 Rich Test Function by Rademacher Complexity

So far, we have verified the realization of the generalization gap on the triangle inequality $\sup_{g\in\mathcal{G}} |\phi(g^*, g)| \leq |\phi(g^*, \mathcal{I})| + \inf_{g\in\mathcal{G}} |\phi(\mathcal{I}, g)|$. To clearly understand it, we then transform representation domain of the triangle inequality to feature and counterfactual space as below figure.



(a) AMR-GMM w/o Regularizer

(b) AMR-GMM w/ Regularizer

Figure 1: Representing feature space, counterfactual space, and their space interval (*Rademacher Distance*) according to whether localized Rademacher regularizer is applied in AMR-GMM.

From Fig. 1, we can draw three factors $|\phi(g^*, g)|$, $|\phi(g^*, \mathcal{I})|$, $|\phi(\mathcal{I}, g)|$ of the triangle inequality to $\sqrt{\lambda}|\mathbb{E}_Z[F^*_{\text{CF}} - F_{\text{CF}}]|$, $\sqrt{\lambda}|\mathbb{E}_Z[F^*_{\text{CF}} - F_{\text{adv}}]|$, $\sqrt{\lambda}|\mathbb{E}_Z[F_{\text{adv}} - F_{\text{CF}}]|$ and then the inequality for the given instrument can be obviously shown to: $\sup_{F_{\text{CF}}|Z} |F^*_{\text{CF}} - F_{\text{CF}}| \leq |F^*_{\text{CF}} - F_{\text{adv}}| + \inf_{F_{\text{CF}}|Z} |F_{\text{adv}} - F_{\text{CF}}|$. Therefore, it becomes a more intuitively understandable formulation to explain their relationship.

Here, we newly define $|F_{\text{adv}} - F_{\text{CF}}| = |Z - g(Z)|$ as Rademacher Distance (red dotted lines) measuring space interval between feature space and its counterfactual space. These red dotted lines are highly related to the localized Rademacher regularizer $|\phi(\mathcal{I}, g)|^2 \approx \lambda|\mathbb{E}_Z[Z - g(Z)]|^2$ as explained in Appendix B.2. Consequently, using this regularizer makes their space interval close compared to not using it, thereby pushing the conterfactual space towards possible feature space.

C Statistical Distance for Causal Inversion

Table 1: Measuring distance metric (unit: m) of KL divergence $\mathcal{D}_{\text{KL}}$, between the prediction of causal features and *causal inversion*, *natural input*, and *adversarial example* of which perturbation budget varies on small and large dataset. The elements below δ_{causal} denote maximum magnitude of causal perturbation budget chosen by a heuristic search to estimate causal features.

Network	CIFAR-10				SVHN				Tiny-ImageNet			
	δ_{causal}	Inversion	Natural	Adversary	δ_{causal}	Inversion	Natural	Adversary	δ_{causal}	Inversion	Natural	Adversary
VGG	8/255	**6.3**	55.5	586.0	4/255	**4.7**	25.6	1011.3	1/255	**35.8**	83.4	800.4
ResNet	4/255	**2.2**	16.5	549.5	1/255	**2.1**	11.6	768.4	.5/255	**35.1**	80.8	762.5
WRN	2/255	**1.2**	7.2	671.8	1/255	**1.6**	6.0	937.9	.5/255	**33.4**	57.5	1062.1

Table 1 shows the statistical distance away from confidence score for model prediction of causal features, compared with that of causal inversion, natural input, and adversarial examples. It implies how well the generated causal inversion represents causal features on feasible bound so that networks themselves enable to exhibit the causal features. In other words, it does not harm causal prediction much according to the smaller statistical distance in Table 1, thus we employ it to effectively inject causal features into the defense networks without direct aid of hypothesis model.

D Algorithm Detail of AMR-GMM

Algorithm 1 Adversarial Moment Restriction based Generalized Method of Moments (AMR-GMM)

```
Require: Data Samples $\mathcal{S}$, Pre-trained Network $f$, Log-likelihood Function $\Omega$
Initialize parameters $\theta_h$ and $\theta_g$ of hypothesis model $h$ and test function $g$
for $(X, Y) \sim \mathcal{S}$ do
    $X_\epsilon \leftarrow \text{Attack}(X, Y)$    ▷ PGD Attack
    $F_{\text{adv}} \leftarrow f_l(X_\epsilon), F_{\text{natural}} \leftarrow f_l(X)$    ▷ Adversarial/Natural Features
    $Z \leftarrow F_{\text{adv}} - F_{\text{natural}}$    ▷ Instrumental Variables
    $T' \leftarrow g(Z), Y' \leftarrow \log Y$    ▷ Counterfactual Treatment and Pseudo Target Label
    $\psi^{\Omega}_{T'|Z}(h) \leftarrow Y' - (\Omega \circ h)(T')$    ▷ Generalized Residual Function for AMR
    $\mathcal{L}_{\text{AMR-GMM}}(\theta_h, \theta_g) \leftarrow \psi^{\Omega}_{T'|Z}(h) \cdot (\Omega \circ g)(Z)$    ▷ Main objective of AMR-GMM
    $\mathcal{L}_{\text{Reg}}(\theta_g) \leftarrow \lambda |Z - g(Z)|^2$    ▷ Localized Rademacher Regularizer
    $\theta_g \leftarrow \theta_g + \alpha \frac{\partial}{\partial \theta_g}(\mathcal{L}_{\text{AMR-GMM}} - \mathcal{L}_{\text{Reg}})$    ▷ Update $\theta_g$ ($\alpha$: lr) for Maximizing AMR-GMM Loss
    $\theta_h \leftarrow \theta_h - \alpha \frac{\partial}{\partial \theta_h} \mathcal{L}_{\text{AMR-GMM}}$    ▷ Update $\theta_h$ ($\alpha$: lr) for Minimizing AMR-GMM Loss
end for
```

Both hypothesis model and test function comprise a bundle of the convolutional layers as a simple CNN structure, trained on AdamW with a learning rate of $\alpha = 10^{-4}$ in 10 epochs, where we set the convergence rate $\lambda = 1$. For ImageNet, we train it with perturbation budget $2/255$ and its 2.5 times step size using fast adversarial training based on FGSM. More details are described in our code at supplementary material.

E Algorithm Detail of CAFE

Algorithm 2 CAusal FEatures (CAFE)

```
Require: Data Samples $\mathcal{S}$, Pre-trained Network $f$ and Hypothesis Model $h$, Defense Loss $\mathcal{L}_{\text{defense}}$
1:  for $(X, Y) \sim \mathcal{S}$ do
2:      $X_\epsilon \leftarrow \text{Attack}(X, Y)$    ▷ PGD Attack
3:      $F_{\text{adv}} \leftarrow f_l(X_\epsilon), F_{\text{natural}} \leftarrow f_l(X)$    ▷ Adversarial/Natural Features
4:      $Z \leftarrow F_{\text{adv}} - F_{\text{natural}}$    ▷ Instrumental Variables
5:      $F_{\text{AC}} \leftarrow F_{\text{natural}} + h(Z)$    ▷ Calculating Causal Features
6:      $\delta_{\text{causal}} = \arg\min_{\|\delta\|_\infty \leq \gamma} \mathcal{D}_{\text{KL}} \left( f_{l+}(F_{\text{AC}}) \mid\mid f(X_\delta) \right)$    ▷ Causal Perturbation
7:      $X_{\text{causal}} \leftarrow X + \delta_{\text{causal}}$    ▷ Causal Inversion
8:      $\hat{F}_{\text{AC}} \leftarrow f_l(X_{\text{causal}})$    ▷ Estimated Causal Features
9:      $\mathcal{L}_{\text{CAFE}}(\theta_f) \leftarrow \mathcal{L}_{\text{Defense}} + \mathcal{D}_{\text{KL}}(f_{l+}(\hat{F}_{\text{AC}}) \mid\mid f_{l+}(F_{\text{adv}}))$    ▷ CAFE Loss with parameter $\theta_f$ of $f$
10:     $\theta_f \leftarrow \theta_f - \alpha \frac{\partial}{\partial \theta_f} \mathcal{L}_{\text{CAFE}}$    ▷ Update $\theta_f$ ($\alpha$: lr) for Minimizing CAFE Loss
11: end for
```

As described in line 9, we readily add a causal regularizer $\mathcal{D}_{\text{KL}}$ to pre-defined defense loss $\mathcal{L}_{\text{Defense}}$ and train all networks from scratch to show the true effectiveness of CAFE. Note that, the number of steps for causal inversion is each 10 for CIFAR-10, SVHN and 3 (regarding speed) for Tiny-ImageNet. More details are either described in our code at supplementary material.

F Efficacy of CAusal FEatures (CAFE)

F.1 CAFE without Causal Inversion (CAFE†)

We experiment ablation study of CAFE without causal inversion to show the effectiveness of the causal inversion for CAFE. In Algorithm 2, we first remove the procedures of getting the causal inversion and the estimated causal features in line 6-8, and we name it *CAusal FEatures without causal inversion (CAFE†)* of which algorithm is explained in the following Algorithm 3.

Algorithm 3 CAusal FEatures without Causal Inversion (CAFE†)

```
Require: Data Samples $\mathcal{S}$, Pre-trained Network $f$ and Hypothesis Model $h$, Defense Loss $\mathcal{L}_{\text{defense}}$
for $(X, Y) \sim \mathcal{S}$ do
    $X_\epsilon \leftarrow \text{Attack}(X, Y)$                                          ▷ PGD Attack
    $F_{\text{adv}} \leftarrow f_l(X_\epsilon)$, $F_{\text{natural}} \leftarrow f_l(X)$                  ▷ Adversarial/Natural Features
    $Z \leftarrow F_{\text{adv}} - F_{\text{natural}}$                                     ▷ Instrumental Variables
    $F_{\text{AC}} \leftarrow F_{\text{natural}} + h(Z)$                                   ▷ Calculating Causal Features
    $\mathcal{L}_{\text{CAFE}^\dagger} \leftarrow \mathcal{L}_{\text{Defense}} + \mathcal{D}_{\text{KL}}(f_{l+}(F_{\text{AC}}) \mid\mid f_{l+}(F_{\text{adv}}))$      ▷ CAFE$^\dagger$ Loss
    $\theta_f \leftarrow \theta_f - \alpha \frac{\partial}{\partial \theta_f} \mathcal{L}_{\text{CAFE}^\dagger}$                    ▷ Update $\theta_f$ ($\alpha$: lr)
end for
```

Table 2: Measuring adversarial robustness of CAFE† not using causal inversion (Algorithm 3) and comparing the robustness with original CAFE (Algorithm 2) on five defense baselines: ADV, TRADES, MART, AWP, HELP, trained with VGG-16 for CIFAR-10, SVHN, Tiny-ImageNet under six attack modes: FGSM, PGD, CW$_\infty$, AP, DLR, AA.

Method	CIFAR-10							SVHN							Tiny-ImageNet						
	Natural	FGSM	PGD	CW$_\infty$	AP	DLR	AA	Natural	FGSM	PGD	CW$_\infty$	AP	DLR	AA	Natural	FGSM	PGD	CW$_\infty$	AP	DLR	AA
ADV	78.5	49.8	44.8	42.6	43.2	42.9	40.7	91.9	64.8	52.1	48.9	48.0	48.5	45.2	53.2	25.3	21.5	21.0	20.2	20.8	19.6
ADV$_{\text{CAFE}^\dagger}$	**79.5**	50.6	45.0	43.8	43.5	44.1	41.7	**92.0**	64.9	52.0	49.5	47.6	48.7	45.4	**53.7**	25.4	21.8	21.2	20.6	21.1	20.0
ADV$_{\text{CAFE}}$	78.4	**52.2**	**47.9**	**44.1**	**46.4**	**44.5**	**42.7**	91.5	**67.0**	**55.3**	**50.0**	**51.3**	**49.6**	**46.1**	52.6	**26.0**	**22.8**	**22.1**	**21.8**	**22.0**	**21.0**
$\Delta_{\text{CAFE}^\dagger}$(%)	***1.3***	*1.5*	*0.4*	*3.0*	*0.8*	*2.6*	*2.4*	***0.2***	*0.2*	*-0.2*	*1.2*	*-0.8*	*0.5*	*0.4*	***0.9***	*0.6*	*1.7*	*1.2*	*2.2*	*1.5*	*1.8*
Δ_{CAFE}(%)	*-0.1*	***4.8***	***7.1***	***3.7***	***7.5***	***3.8***	***4.9***	*-0.4*	***3.4***	***6.1***	***2.2***	***6.8***	***2.3***	***1.9***	*-1.2*	***3.0***	***6.4***	***5.2***	***7.8***	***5.6***	***6.9***
TRADES	**79.5**	50.4	45.7	43.2	44.4	42.9	41.8	**91.9**	66.4	53.6	49.1	49.1	47.7	45.2	52.8	25.9	22.5	21.9	21.5	21.8	20.7
TRADES$_{\text{CAFE}^\dagger}$	78.2	50.0	45.1	43.5	43.9	43.6	41.9	90.6	64.1	52.8	49.5	48.5	48.8	45.9	**53.5**	25.5	22.1	21.5	20.9	21.5	20.3
TRADES$_{\text{CAFE}}$	77.0	**51.6**	**47.9**	**44.0**	**47.0**	**43.9**	**42.7**	90.3	**67.8**	**56.1**	**50.0**	**53.6**	**49.1**	**47.5**	52.1	**26.5**	**23.6**	**22.6**	**22.5**	**22.6**	**21.6**
$\Delta_{\text{CAFE}^\dagger}$(%)	***-1.7***	*-0.8*	*-1.4*	*0.5*	*-1.0*	*1.6*	*0.4*	***-1.4***	*-3.4*	*-1.5*	*1.0*	*-1.1*	*2.2*	*1.5*	***1.3***	*-1.9*	*-1.6*	*-1.4*	*-3.0*	*-1.4*	*-2.0*
Δ_{CAFE}(%)	*-3.1*	***2.2***	***4.8***	***1.8***	***5.8***	***2.3***	***2.3***	*-1.8*	***2.1***	***4.6***	***1.9***	***9.3***	***2.9***	***5.0***	*-1.3*	***2.2***	***5.2***	***3.6***	***4.6***	***3.7***	***4.2***
MART	**79.7**	52.4	47.2	43.4	45.5	43.8	42.0	**92.6**	66.6	54.2	47.9	49.6	47.1	44.4	53.1	25.0	21.5	21.2	20.4	21.0	19.9
MART$_{\text{CAFE}^\dagger}$	79.4	51.7	45.8	43.3	44.1	43.7	41.6	92.0	65.8	53.1	49.1	48.2	48.2	44.9	**53.5**	25.4	21.8	21.3	20.7	21.3	20.2
MART$_{\text{CAFE}}$	78.3	**54.2**	**49.7**	**43.9**	**48.1**	**44.5**	**42.7**	91.3	**67.6**	**57.3**	**49.5**	**54.2**	**48.3**	**46.4**	53.0	**25.6**	**22.3**	**21.6**	**21.3**	**21.5**	**20.5**
$\Delta_{\text{CAFE}^\dagger}$(%)	***-0.5***	*-1.3*	*-3.0*	*-0.2*	*-3.2*	*-0.3*	*-0.8*	***-0.6***	*-1.2*	*-2.0*	*2.3*	*-2.8*	*2.4*	*1.1*	***0.7***	*1.5*	*1.7*	*0.9*	*1.7*	*1.5*	*1.5*
Δ_{CAFE}(%)	*-1.8*	***3.4***	***5.1***	***1.2***	***5.6***	***1.6***	***1.9***	*-1.4*	***1.4***	***5.9***	***3.3***	***9.2***	***2.7***	***4.6***	*-0.2*	***2.4***	***4.0***	***1.8***	***4.3***	***2.5***	***3.1***
AWP	**78.0**	51.7	48.2	43.5	47.2	43.4	42.6	90.8	65.5	56.6	50.4	54.0	49.7	48.6	52.6	28.0	25.7	23.6	24.8	23.5	22.8
AWP$_{\text{CAFE}^\dagger}$	76.3	50.9	47.0	43.8	45.9	44.0	42.4	83.2	58.0	51.8	49.0	49.8	48.7	47.0	52.5	26.5	23.4	22.6	22.4	22.5	21.6
AWP$_{\text{CAFE}}$	77.4	**54.8**	**51.4**	**44.2**	**50.2**	**44.9**	**43.5**	**91.9**	**67.9**	**58.6**	**51.2**	**55.9**	**51.1**	**49.7**	**52.9**	**28.8**	**26.4**	**24.2**	**25.6**	**24.1**	**23.4**
$\Delta_{\text{CAFE}^\dagger}$(%)	*-2.2*	*-1.6*	*-2.4*	*0.6*	*-2.7*	*1.5*	*-0.4*	*-8.3*	*-11.4*	*-8.6*	*-2.9*	*-7.9*	*-2.0*	*-3.3*	*-0.2*	*-5.4*	*-8.7*	*-4.0*	*-9.6*	*-4.3*	*-5.3*
Δ_{CAFE}(%)	***-0.8***	***5.8***	***6.8***	***1.7***	***6.4***	***3.6***	***2.2***	***1.2***	***3.8***	***3.4***	***1.6***	***3.6***	***2.7***	***2.3***	***0.6***	***3.0***	***2.7***	***2.7***	***3.3***	***2.5***	***2.9***
HELP	**77.4**	51.8	48.3	43.9	47.3	43.9	42.9	91.2	65.8	56.6	50.9	53.9	50.2	48.8	**53.0**	28.3	25.9	23.9	25.1	23.8	23.1
HELP$_{\text{CAFE}^\dagger}$	76.2	51.0	47.2	43.9	46.1	44.2	42.7	87.6	61.7	53.7	49.6	51.3	49.2	47.3	52.9	27.0	24.1	23.0	23.2	23.0	22.1
HELP$_{\text{CAFE}}$	75.6	**54.4**	**51.4**	**44.6**	**50.4**	**44.8**	**43.7**	**91.5**	**67.3**	**58.5**	**51.6**	**56.2**	**51.4**	**50.0**	52.6	**29.4**	**27.1**	**24.7**	**26.4**	**24.4**	**23.9**
$\Delta_{\text{CAFE}^\dagger}$(%)	***-1.6***	*-1.6*	*-2.2*	*0.0*	*-2.4*	*0.9*	*-0.4*	*-4.0*	*-6.3*	*-5.1*	*-2.4*	*-5.0*	*-2.0*	*-3.0*	***-0.1***	*-4.4*	*-7.0*	*-3.8*	*-7.8*	*-3.2*	*-4.2*
Δ_{CAFE}(%)	*-2.3*	***5.0***	***6.4***	***1.5***	***6.6***	***2.2***	***1.8***	***0.3***	***2.3***	***3.3***	***1.4***	***4.2***	***2.4***	***2.5***	*-0.8*	***3.9***	***4.7***	***3.1***	***5.0***	***2.4***	***3.5***

Table 2 shows that CAFE without causal inversion (CAFE†) cannot further enhance adversarial robustness of networks compared with that of original CAFE with causal inversion, and even CAFE† has mostly worse robustness than its corresponding baselines. Due to its deviated prediction, we introduce a causal inversion that helps to estimate causal features and fit their prediction. We can then enlighten causal inversion has a remarkable effect to elevate robustness in all of the defense networks and conclude the effectiveness of the CAFE comes from the causal inversion.

F.2 Power of Test Function

As described in Section 3.1 (Revisiting Non-parametric IV regression), the test function is responsible for generating infinite moment restrictions. In practical, however, it is impossible to handle the such infinite moments. To deal with the analogue limitation, the adversarial learning [7, 41] utilizes test function that provokes the upper-bound of the moment m by finding the extreme part of IV, then minimizes m with hypothesis model to obtain more robustified one from counterfactuals. It is highly aligned with our problem setup, because our main goal is either to acquire hypothesis model extracting causal features, despite given the possible worst-case variation of IV. From the power of test function, a lot of works: AGMM [41], DeepGMM [7], AGMM+RKHS [20], MMR-IV [44] have employed min-max optimization for performing IV regression, and it has been empirically verified that they greatly outperform the constraint optimization that does not fulfill min-max optimization with a test function: RandomForest, DirectNN, GMM [28], 2SLS [71], DeepIV [29], KernelIV [61]. Beyond the roles of our hypothesis model and test function, our proposed method (AMR-GMM) also follows recent IV regression methods utilizing the min-max optimization.

	CIFAR-10		Tiny-ImageNet	
Method	Natural	AA	Natural	AA
ADV	84.3	45.6	**60.9**	23.9
$\text{ADV}_{\text{CAFE}\setminus g}$	84.6	47.1	60.3	23.7
ADV_{CAFE}	**85.7**	**49.5**	60.6	**25.4**

Table 3: Ablation study of test function g

For further verification, we train hypothesis model without test function and inoculate causal features acquired from the hypothesis model to DNNs where we use minimizing optimization instead of min-max one: $\text{ADV}_{\text{CAFE}\setminus g}$ to identify its regression effect. Then, we observe the experimental results are significantly degraded in ADV_{CAFE} for CIFAR-10 and Tiny-ImageNet each with WRN-34-10. From these results, we can say that min-max optimization fits in our problem setup.

F.3 Stability of CAFE

In Appendix E, we noted that our hypothesis model and test function were trained for 10 epochs on CIFAR-10 and SVHN and 3 epochs on Tiny-ImageNet and it was deemed sufficient convergence, as observed through empirical evidence.

	CIFAR-10		SVHN		Tiny-ImageNet	
Networks	Natural	AA	Natural	AA	Natural	AA
VGG-16	0.027	0.084	0.005	0.027	0.058	0.030
ResNet-18	0.015	0.028	0.003	0.021	0.039	0.028
WRN-34-10	0.007	0.021	0.001	0.019	0.018	0.024

Table 4: Standard deviations of CAFE for Stability

To verify the stability of CAFE, we experiment standard deviation results of CAFE for the accuracy (%) of adversarial robustness under 20 repetitions on the all datasets with all networks for ADV_{CAFE}. This table implies that CAFE has either consistent stability of the performances, aligned with that of acquiring causal estimator h in the concept of Lewis *et al.* [41] with *Set Identification* [14] and *Lipschitz* [6] to finding ϵ-equilibrium of the zero-sum game.