A. Proof of Eq. 9

$$
\begin{aligned}
&\frac{1}{L}\sum_{v\in V}\phi^m(v)\\
=&\frac{1}{L}\sum_{v\in V}E_{S\subseteq V\setminus\{v\},|S|=m}[f(S\cup\{v\})-f(S)]\\
=&\frac{1}{L}\sum_{v\in V}\frac{m!(L-1-m)!}{(L-1)!}\sum_{S\subseteq V\setminus\{v\},|S|=m}[f(S\cup\{v\})-f(S)]\\
=&\frac{1}{L}\frac{m!(L-1-m)!}{(L-1)!}\sum_{S\subseteq V,|S|=m+1}[f(S)-\sum_{v\in S}f(S\setminus\{v\})]\\
=&\frac{(m+1)!(L-1-m)!}{L!}\sum_{S\subseteq V,|S|=m+1}[\frac{1}{m+1}f(S)-\frac{1}{m+1}\sum_{v\in S}f(S\setminus\{v\})]\\
=&\frac{1}{m+1}E_{S\subseteq V,|S|=m+1}[f(S)-\sum_{v\in S}f(S\setminus\{v\})]
\end{aligned}
\tag{10}
$$

B. Discussion about the State-of-the-art

To further validate our findings, we provided analyses about the SOTA method in [14] *w.r.t.* our proposed hypotheses. Following the methods in the main paper, results in Tab. 3 show that when aligned with the same backbone, *CADDM* [14] with better generalization abilities tended to encode low-order interactions with fewer negative contributions (*i.e.*, larger values of D^m, which were averaged among different images.) and less strength (*i.e.*, smaller values of ρ^m.), which were consistent with our previous analyses [1].

Backbone	Model	Train	Test	Manipulation									
		FF++	Celeb-DF (v2)	DeepFakes		FaceShifter		Face2Face		FaceSwap		NeuralTextures	
		V-AUC	V-AUC	$D^m\uparrow$	$\rho^m\downarrow$	$D^m\uparrow$	$\rho^m\downarrow$	$D^m\uparrow$	$\rho^m\downarrow$	$D^m\uparrow$	$\rho^m\downarrow$	$D^m\uparrow$	$\rho^m\downarrow$
ResNet-18	*Base*	0.998	0.658	-0.023	0.049	-0.022	0.048	-0.022	0.048	-0.021	0.047	-0.022	0.049
	Base+DA	0.998	0.776	-0.011	0.044	-0.013	0.045	-0.012	0.045	-0.011	0.045	-0.012	0.045
	CADDM [14]	0.998	0.890	**-0.008**	**0.028**	**-0.006**	**0.029**	**-0.008**	**0.028**	**-0.008**	**0.028**	**-0.008**	**0.028**
ResNet-34	*Base*	0.999	0.641	-0.050	0.102	-0.049	0.102	-0.049	0.102	-0.050	0.103	-0.049	0.102
	Base+DA	0.998	0.801	-0.018	0.068	-0.021	0.068	-0.021	0.068	-0.023	0.069	-0.021	0.069
	CADDM [14]	0.997	0.912	**-0.011**	**0.067**	**-0.011**	**0.066**	**-0.011**	**0.067**	**-0.012**	**0.067**	**-0.011**	**0.067**
Xception	*Base*	0.998	0.585	-0.039	0.095	-0.039	0.095	-0.038	0.095	-0.038	0.095	-0.037	0.094
	Base+DA	0.998	0.815	-0.010	0.056	-0.018	0.057	-0.012	0.058	-0.016	0.059	-0.010	0.057
	CADDM [14]	0.999	0.867	**-0.009**	**0.039**	**-0.012**	**0.040**	**-0.011**	**0.039**	**-0.012**	**0.049**	**-0.009**	**0.039**

Table 3. Verifying hypotheses on the SOTA method *CADDM* [14]. Here *DA* denotes data augmentations. *Base* denotes the model trained without data augmentations. Results show that *CADDM* exhibited better generalization abilities, when encoding low-order interactions with less negative contributions and less strength, consistent with our proposed hypotheses.

[1] To gain better comparisons, we calculated the proposed metrics D^m and ρ^m in a smaller range of lower-order interactions (*i.e.*, $m < 0.1n$) for the backbone of Xception here, which may cause values of metrics to be different from the main paper for *Base* and *Base+DA*.