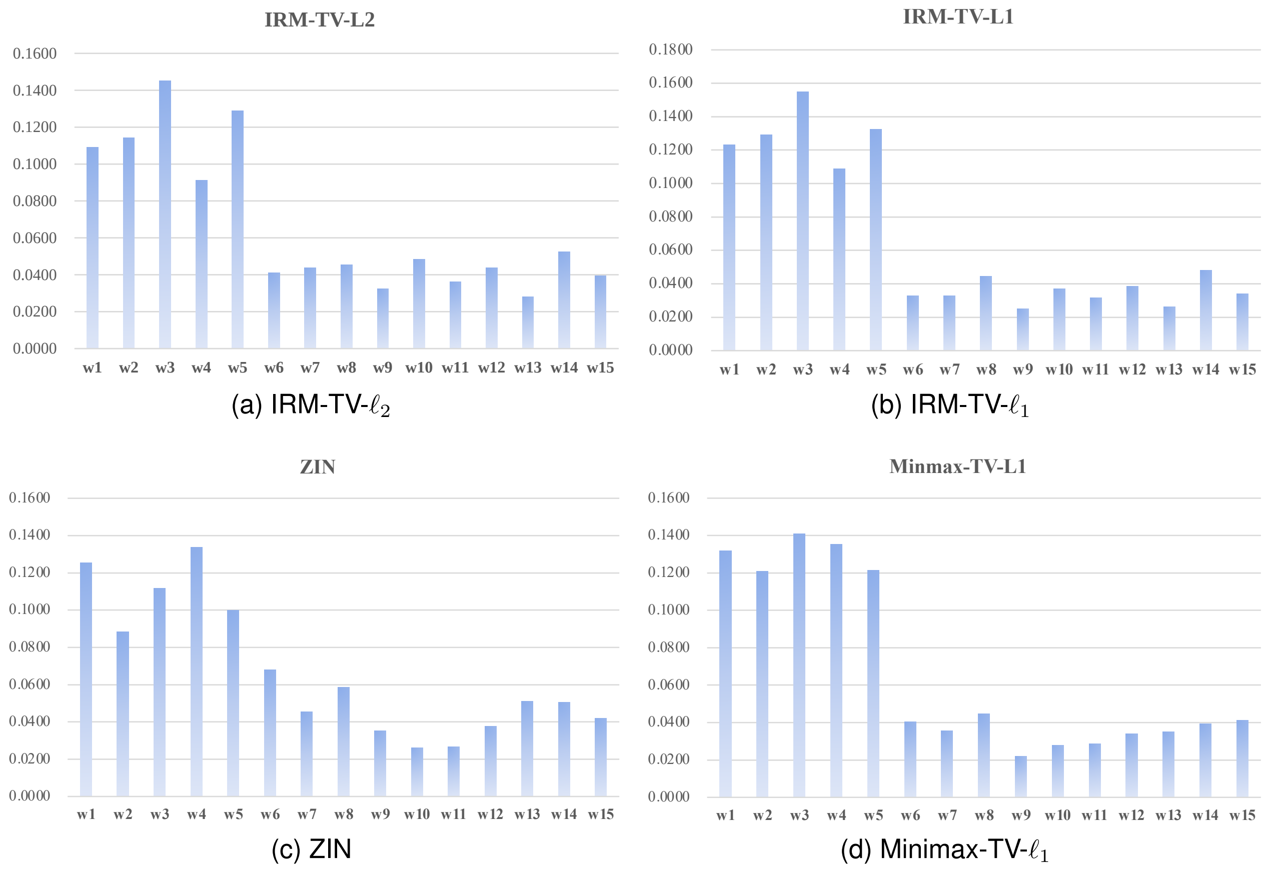
A. Invariant Feature Identifiability

We take the simulation study with $(p_s^-, p_s^+, p_v(t)) = (0.999, 0.9, 0.8)$ as an example to show how IRM-TV-ℓ_1 improves identifiability of invariant features over IRM-TV-ℓ_2. There are 15 features in this experiment, where **No. 1∼5** are **invariant** features and **No. 6∼15** are **spurious** features. We run 10 times of this experiment with different simulated samples and compute the normalized average absolute values of the feature weights, shown in Table 1 and Figure 1. The invariant features of IRM-TV-ℓ_1 and Minimax-TV-ℓ_1 take up 64.88% and 65.08% of the total feature weights, respectively. In contrast, the invariant features of IRM-TV-ℓ_2 and ZIN take up only 58.88% and 55.90% of the total feature weights, respectively. Hence the TV-ℓ_1 models extract more invariant features than the TV-ℓ_2 models. Moreover, the gap between the invariant and spurious feature weights in Figures 1(b) or 1(d) is larger than that in Figures 1(a) or 1(c), respectively. For example, the invariant feature weight w2$= 0.0884$ and the spurious feature weights w6$= 0.0679$ and w8$= 0.0586$ for ZIN, while w2$= 0.1210$, w6$= 0.0403$, and w8$= 0.0446$ for Minimax-TV-ℓ_1, respectively. It indicates that Minimax-TV-ℓ_1 enhances the invariant feature w2 while suppresses the spurious features w6 and w8.

Table 1: Normalized absolute values of feature weights for different methods in simulation study with $(p_s^-, p_s^+, p_v(t)) = (0.999, 0.9, 0.8)$. **w1∼w5** correspond to **invariant** features and **w6∼w15** correspond to **spurious** features.

METHODS	w1	w2	w3	w4	w5	w6	w7	w8	w9	w10	w11	w12	w13	w14	w15
IRM-TV-ℓ_2	0.1091	0.1143	0.1452	0.0912	0.1290	0.0412	0.0439	0.0454	0.0323	0.0483	0.0361	0.0438	0.0282	0.0526	0.0395
IRM-TV-ℓ_1	0.1232	0.1292	0.1548	0.1090	0.1326	0.0329	0.0329	0.0446	0.0249	0.0371	0.0317	0.0386	0.0262	0.0482	0.0341
ZIN	0.1253	0.0884	0.1118	0.1336	0.0999	0.0679	0.0454	0.0586	0.0353	0.0261	0.0267	0.0376	0.0510	0.0505	0.0419
MINIMAX-TV-ℓ_1	0.1320	0.1210	0.1410	0.1353	0.1215	0.0403	0.0356	0.0446	0.0221	0.0279	0.0285	0.0341	0.0352	0.0395	0.0412



(a) IRM-TV-ℓ_2

(b) IRM-TV-ℓ_1

(c) ZIN

(d) Minimax-TV-ℓ_1

Figure 1: Normalized absolute values of feature weights for different methods in simulation study with $(p_s^-, p_s^+, p_v(t)) = (0.999, 0.9, 0.8)$. **w1∼w5** correspond to **invariant** features and **w6∼w15** correspond to **spurious** features.