Table 1. Soundness of Morpho-MNIST image counterfactuals. Effectiveness for digit class (d) is measured using accuracy (Acc) from a pre-trained classifier and mean absolute percentage error (MAPE) for slant (s), thickness (t) and intensity (i). VAE, HVAE and VCI denote the mechanisms from Pawlowski et al. (2020); De Sousa Ribeiro et al. (2023); Wu et al. (2024) resp. in our evaluation setup.

	Slant Intervention ($do(s)$)					Class Intervention ($do(d)$)					Null
	Effectiveness				Rev.	Effectiveness				Rev.	Comp.
Mechanism	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑	$L_1^{\downarrow}$	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑	L_1 ↓	L_1 ↓
VAE	4.63e-2	6.98e-2	2.91e-1	97.27	4.28e-2	5.90e-2	8.03e-2	2.10e-1	94.92	4.88e-2	3.62e-2
HVAE	3.39e-2	4.93e-3	3.88e-1	95.02	1.23e-2	4.39e-2	5.08e-3	1.23e-1	95.31	3.25e-2	1.60e-4
VCI	3.08e-2	6.52e-3	9.07e-2	90.04	3.19e-2	2.63e-2	6.32e-3	9.12e-2	94.62	1.98e-2	1.31e-2
Spatial:	2.78e-2	5.52e-3	2.45e-1	96.62	5.47e-2	3.11e-2	5.92e-3	2.13e-1	96.29	7.56e-2	1.11e-3
$\{\omega$=1.5, $p_{\varnothing}$=0.1$\}$	2.17e-2	5.91e-3	1.53e-1	99.02	7.03e-2	2.26e-2	5.46e-3	5.01e-2	99.51	9.00e-2	2.02e-2
$\{\omega$=3, $p_{\varnothing}$=0.1$\}$	1.85e-2	2.91e-3	8.85e-2	99.84	1.05e-1	1.87e-2	2.92e-3	8.55e-2	99.90	1.14e-1	6.57e-2
$\{\omega$=4.5, $p_{\varnothing}$=0.1$\}$	1.87e-2	3.06e-3	6.67e-2	99.90	1.19e-1	1.85e-2	3.19e-2	5.11e-2	99.98	1.26e-1	7.42e-2
$\{\omega$=1.5, $p_{\varnothing}$=0.5$\}$	2.88e-2	7.39e-3	2.43e-1	97.75	5.87e-2	3.44e-2	8.02e-3	1.29e-1	93.95	8.61e-2	1.32e-2
$\{\omega$=3, $p_{\varnothing}$=0.5$\}$	2.01e-2	3.88e-3	9.84e-2	99.64	8.35e-2	2.55e-2	4.45e-3	9.32e-2	97.63	9.86e-2	3.43e-2
$\{\omega$=4.5, $p_{\varnothing}$=0.5$\}$	2.02e-2	3.93e-3	1.84e-1	99.71	8.98e-2	2.48e-2	4.60e-3	7.83e-2	98.14	1.06e-1	3.54e-2
Semantic:	4.98e-2	9.81e-3	4.79e-1	90.33	7.87e-2	7.43e-2	1.63e-2	7.15e-1	83.01	1.12e-1	2.77e-3
$\{\omega$=1.5, $p_{\varnothing}$=0.1$\}$	2.83e-2	1.20e-2	1.51e-1	98.44	5.96e-2	3.88e-2	1.13e-2	1.43e-1	97.66	8.66e-2	1.88e-2
$\{\omega$=3, $p_{\varnothing}$=0.1$\}$	2.14e-2	1.00e-2	1.48e-1	99.80	8.81e-2	2.15e-2	9.41e-3	6.99e-2	99.80	1.07e-1	4.34e-2
$\{\omega$=4.5, $p_{\varnothing}$=0.1$\}$	2.13e-2	7.96e-3	1.78e-1	99.80	1.07e-1	7.67e-2	9.41e-3	7.62e-2	99.93	1.23e-1	5.74e-2

Table 4. Soundness of Morpho-MNIST image counterfactuals. Effectiveness for digit class (d) is measured using the accuracies (Acc) from a pre-trained classifier and mean absolute percentage error (MAPE) for slant (s), thickness (t) and intensity (i). VAE, HVAE and VCI denote the mechanisms from Pawlowski et al. (2020); De Sousa Ribeiro et al. (2023); Wu et al. (2024) resp. in our evaluation setup.

	Thickness Intervention ($do(t)$)					Intensity Intervention ($do(i)$)				
	Effectiveness				Rev.	Effectiveness				Rev.
Mechanism	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑	$L_1^{\downarrow}$	MAPE (t) ↓	MAPE (i) ↓	MAPE (s) ↓	Acc (d) ↑	L_1 ↓
VAE	4.48e-2	6.76e-2	3.88e-1	97.75	4.26e-2	8.06e-2	7.55e-2	4.83e-1	97.85	4.31e-2
HVAE	3.05e-2	6.75e-3	1.82e-1	95.61	6.78e-3	3.92e-2	4.71e-3	1.60e-1	94.92	5.80e-3
VCI	3.08e-2	6.26e-3	9.13e-2	92.97	2.10e-2	6.99e-2	2.61e-2	3.97e-1	82.52	3.75e-2
Spatial	2.99e-2	5.06e-3	1.75e-1	96.55	2.67e-2	4.33e-2	7.35e-3	3.51e-1	92.39	2.69e-2
$\{\omega$=1.5, $p_{\varnothing}$=0.1$\}$	1.96e-2	3.55e-3	1.44e-1	99.61	4.70e-2	3.46e-2	4.96e-3	3.07e-1	97.17	4.55e-2
$\{\omega$=3, $p_{\varnothing}$=0.1$\}$	1.95e-2	2.96e-3	1.16e-1	99.89	9.56e-2	3.04e-2	6.29e-3	2.10e-1	95.84	9.75e-2
$\{\omega$=4.5, $p_{\varnothing}$=0.1$\}$	1.98e-2	3.27e-3	1.26e-1	99.98	1.14e-1	2.43e-2	9.36e-3	1.91e-1	98.63	1.14e-1
$\{\omega$=1.5, $p_{\varnothing}$=0.5$\}$	2.29e-2	4.78e-3	2.78e-1	98.93	2.95e-2	2.91e-2	6.37e-3	2.92e-1	97.75	2.13e-2
$\{\omega$=3, $p_{\varnothing}$=0.5$\}$	2.21e-2	3.89e-3	1.02e-1	99.71	6.21e-2	2.93e-2	1.13e-2	1.34e-1	98.35	6.89e-2
$\{\omega$=1.5, $p_{\varnothing}$=0.5$\}$	2.03e-2	3.82e-3	2.15e-1	99.90	6.74e-2	2.56e-2	1.02e-2	1.59e-1	99.12	9.56e-2
Semantic	4.17e-2	7.18e-3	3.18e-1	94.43	3.44e-2	5.90e-2	1.62e-2	3.00e-1	87.50	4.46e-2
$\{\omega$=1.5, $p_{\varnothing}$=0.1$\}$	2.36e-2	1.29e-2	2.30e-1	98.54	3.72e-2	3.38e-2	1.40e-2	2.12e-1	96.39	3.39e-2
$\{\omega$=3, $p_{\varnothing}$=0.1$\}$	2.16e-2	1.01e-2	1.56e-1	99.51	7.04e-2	3.41e-2	1.45e-2	2.11e-1	97.75	6.54e-2
$\{\omega$=4.5, $p_{\varnothing}$=0.1$\}$	2.24e-2	7.57e-3	2.73e-1	99.61	9.43e-2	4.21e-2	1.55e-2	2.56e-1	98.24	8.85e-2