CO-SPY: Combining Semantic and Pixel Features to Detect Synthetic Images by AI

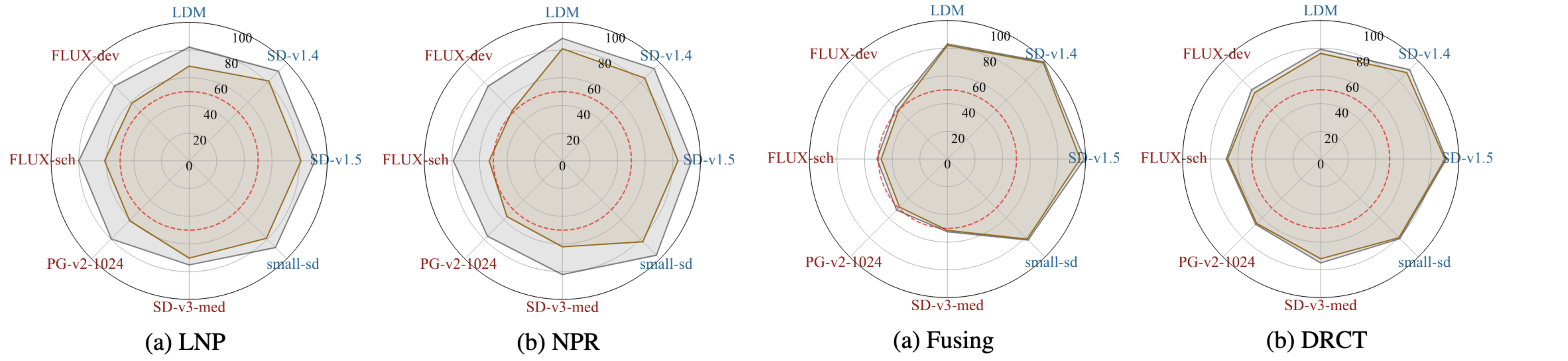
Siyuan Cheng*, Lingjuan Lyu†, Zhenting Wang, Xiangyu Zhang, Vikash Sehwag

* Work done during the internship at Sony AI † Corresponding Author

Background & Limitation of Existing Methods

AI-generated (Synthetic) images raise significant concerns regarding misuse. Existing methods can be largely classified into two categories: (1) detectors based on ***semantic*** features, and (2) detectors based on texture-level ***artifacts***.

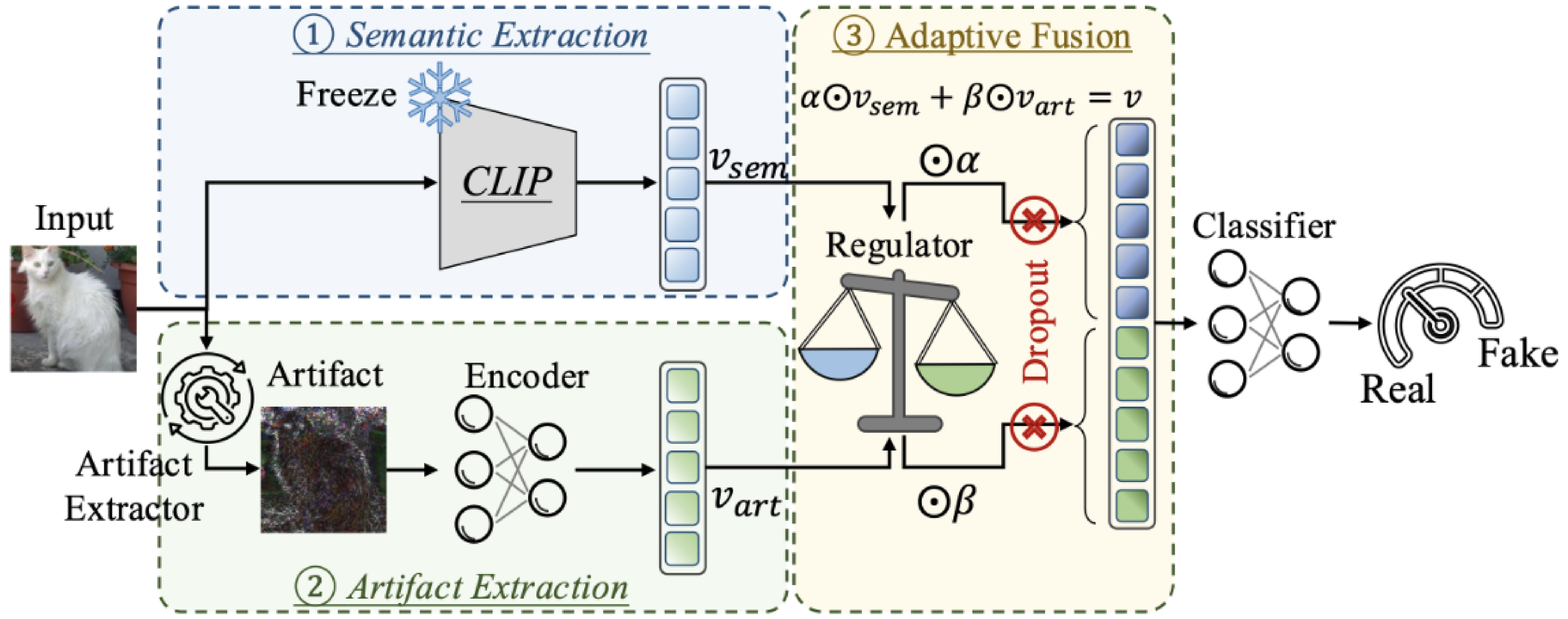
Generalization	Artifact		Semantic		Fusion	
	Existing	Enhanced	Existing	Enhanced	Simple	Enhanced
Diverse Models	●	●	○	◐	◐	●
Lossy Formats	○	◐	●	●	◐	●
Unseen Objects	●	●	○	◐	◐	●



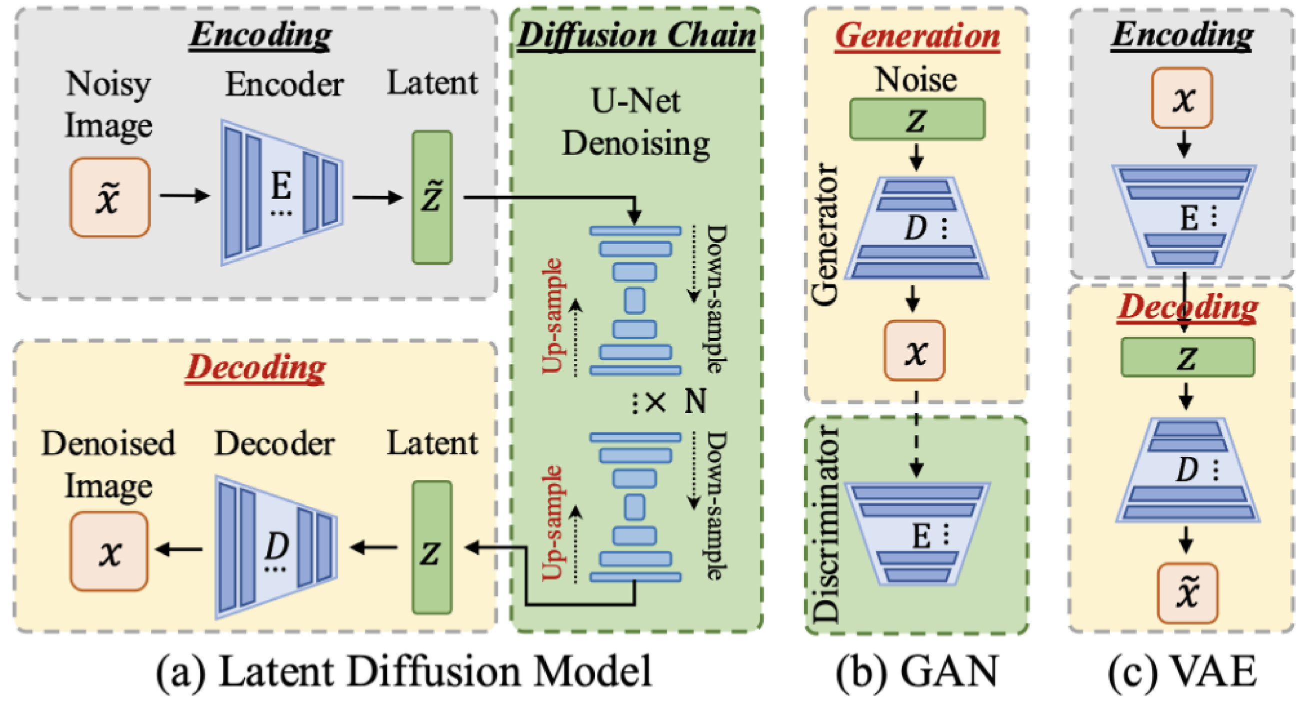
① Artifact Detectors Do Not Support Lossy Formats

② Semantic Detectors Do Not Generalize to Unseen Models and Unseen Contents

CO-SPY Overview



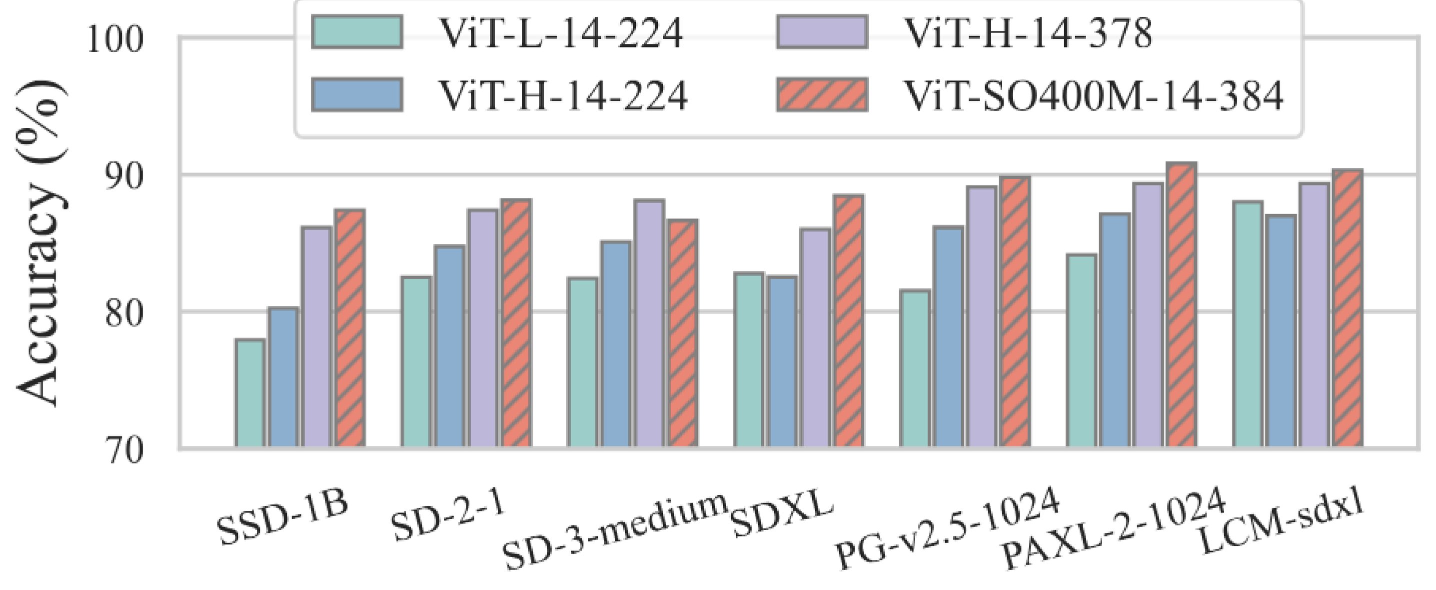
Step ① Enhanced Artifact Detectors



Leverage a pre-trained Variational Autoencoder (VAE) to extract artifacts.

$$\underbrace{\mu, \sigma = E(x)}_{\text{VAE Encoding}} \qquad \underbrace{x' = E(\mu)}_{\text{VAE Decoding}} \qquad \underbrace{\Delta = |x' - x|}_{\text{Artifact Extraction}}$$

Step ② Enhanced Semantic Detectors



Leverage the latest CLIP encoder to extract the semantic features.

Step ③ Combining Two Features

$$\underbrace{\alpha = R_{sem}(v_{sem})}_{\text{Regulate Semantic Features}}, \quad \underbrace{\beta = R_{art}(v_{art})}_{\text{Regulate Artifact Features}}, \quad \underbrace{v = \alpha * v_{sem} \oplus \beta * v_{art}}_{\text{Adaptive Fusion}},$$

Evaluation

CO-SPY-Bench is a high-quality and diverse benchmark for synthetic image detection. It (1) comprises over one million images, (2) includes real images sourced from five established databases, and (3) covers synthetic images produced by 22 state-of-the-art text-to-image diffusion models.

Detector	CNNDet		FreqFD		Fusing		LNP		UnivFD		DIRE		FreqNet		NPR		DRCT		CO-SPY	
	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.
LDM	90.57	77.56	74.88	54.17	98.18	83.03	96.04	84.87	85.76	79.07	86.46	66.25	92.37	74.92	93.45	84.34	88.28	79.25	98.91	95.04
SD-v1.4	97.00	89.95	92.40	62.91	99.95	99.16	99.12	95.92	88.35	80.87	96.51	83.55	90.55	69.20	96.88	90.90	91.61	81.18	97.80	91.95
SD-v1.5	97.03	89.75	92.30	62.56	99.96	99.12	99.23	96.21	88.57	80.88	96.72	83.77	90.33	68.86	97.05	91.30	91.08	81.06	98.02	91.31
SSD-1B	87.35	66.55	48.90	49.72	84.65	53.96	93.81	79.12	86.46	76.47	74.59	56.63	50.28	49.18	52.84	47.87	82.04	75.83	95.40	83.20
tiny-sd	87.44	66.37	80.01	52.19	98.01	77.12	95.03	81.48	84.58	76.96	87.83	63.65	88.03	63.56	95.67	88.42	88.06	79.99	95.99	84.80
SegMoE-SD	91.12	74.41	80.21	51.74	97.16	73.58	96.36	86.62	89.59	83.07	88.95	65.98	88.55	64.40	97.21	93.79	79.29	75.12	97.39	89.49
small-sd	89.78	70.15	81.81	52.57	99.06	82.65	94.81	80.75	85.67	77.45	91.31	68.38	89.44	65.42	95.77	89.14	90.08	81.20	96.22	85.80
SD-2-1	86.78	68.14	52.95	49.93	92.64	59.32	81.26	57.19	89.00	81.74	88.11	65.25	64.40	51.62	71.62	51.31	81.60	76.12	96.89	88.53
SD-3-medium	79.00	60.68	57.99	49.98	81.86	52.47	75.08	53.69	87.62	78.42	76.64	56.75	57.19	49.64	71.36	50.00	79.95	74.95	95.04	82.91
SDXL-turbo	96.42	88.67	92.98	61.34	95.19	59.69	95.07	83.47	90.58	84.31	90.97	72.97	87.04	66.62	94.63	83.57	90.46	80.36	99.17	95.39
SD-2	85.93	65.73	50.79	49.84	89.08	55.92	76.79	54.56	83.24	73.78	83.69	60.07	59.28	50.64	72.97	51.19	80.13	75.14	94.94	83.67
SDXL	83.39	61.79	43.93	49.70	76.81	51.01	94.00	80.33	72.48	63.64	64.48	51.95	47.90	48.84	46.99	47.75	80.62	75.18	91.68	74.12
PG-v2.5-1024	65.10	53.65	47.54	49.70	75.22	50.41	93.46	79.07	82.98	78.23	61.97	52.32	55.09	48.67	50.95	47.71	79.16	71.33	96.45	88.65
PG-v2-1024	83.85	63.48	48.93	49.70	85.62	52.08	74.18	52.48	83.77	78.55	76.95	56.63	53.95	48.79	63.62	48.40	70.06	66.62	96.72	89.14
PG-v2-512	77.73	57.94	55.59	49.87	74.90	51.58	59.60	49.40	69.21	58.90	71.35	53.63	45.21	49.09	65.15	49.09	83.55	77.61	85.02	64.86
PG-v2-256	81.40	63.30	57.88	50.19	75.17	51.10	72.63	54.77	72.40	62.99	81.13	60.32	49.26	49.47	60.43	49.74	78.19	73.55	90.22	72.92
PAXL-2-1024	71.18	56.29	54.28	49.76	86.44	53.60	74.24	53.41	84.97	80.08	70.61	54.77	64.81	49.40	72.84	51.16	76.22	71.80	97.94	93.94
PAXL-2-512	83.05	65.44	80.53	52.20	95.97	68.77	91.30	74.25	85.36	80.32	81.29	62.18	82.72	57.23	94.24	81.25	79.95	75.53	98.63	94.96
LCM-sdxl	93.11	81.55	81.10	52.44	97.74	70.75	95.96	85.46	81.52	78.04	89.14	68.57	85.17	62.29	70.29	50.87	91.96	81.05	98.72	96.20
LCM-sdv1-5	97.67	92.29	94.70	68.17	98.76	81.22	97.87	90.87	83.87	79.67	93.58	79.02	93.14	76.20	98.37	93.71	87.27	79.59	99.63	97.14
FLUX.1-sch	71.72	56.04	56.01	50.02	76.74	51.03	74.75	54.38	80.14	72.89	73.67	56.27	64.35	50.39	77.91	53.14	74.38	68.39	95.52	85.24
FLUX.1-dev	70.75	57.44	53.60	50.04	82.30	52.92	73.25	54.13	82.50	75.95	73.74	56.32	53.83	49.06	71.09	50.74	77.88	70.70	96.16	86.10
Average	84.88	69.42	67.24	53.12	89.15	65.02	86.54	71.93	83.57	76.47	81.80	63.42	70.59	57.43	77.79	65.70	82.81	75.98	96.02	87.06

Table 1: CO-SPY outperforms existing baselines on CO-SPY-Bench.

Detector	CNNDet		FreqFD		Fusing		LNP		UnivFD		DIRE		FreqNet		NPR		DRCT		CO-SPY	
	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.	AP	Acc.
Civitai	88.08	76.85	82.61	55.95	97.22	80.40	95.83	89.85	66.05	61.25	98.34	88.60	89.13	83.80	89.34	86.20	40.94	43.30	95.82	88.95
DALL-E 3	53.83	51.10	50.23	49.68	71.33	51.50	74.61	65.18	71.14	62.98	70.33	54.33	35.70	45.52	37.12	42.98	70.33	62.62	83.94	77.03
instavibe.ai	59.01	53.60	42.76	50.12	65.02	50.90	31.00	44.42	77.97	64.35	52.89	50.82	34.47	45.67	31.59	42.15	68.27	59.80	75.21	67.73
Lexica	59.45	53.08	40.72	49.43	67.43	51.38	33.59	45.00	79.58	64.71	73.24	53.52	37.21	45.95	42.92	43.38	86.70	65.65	85.62	78.42
Midjourney-v6	44.73	48.62	39.77	49.58	55.94	49.98	37.43	45.80	67.13	60.93	61.69	51.90	35.03	45.42	35.26	42.23	60.34	57.53	82.65	74.92
Average	61.02	56.65	51.22	50.95	71.39	56.83	54.49	58.05	72.37	62.84	71.30	59.84	46.31	53.27	47.24	51.39	65.32	57.78	84.65	77.41

Table 2: CO-SPY outperforms existing baselines for in-the-wild testing.

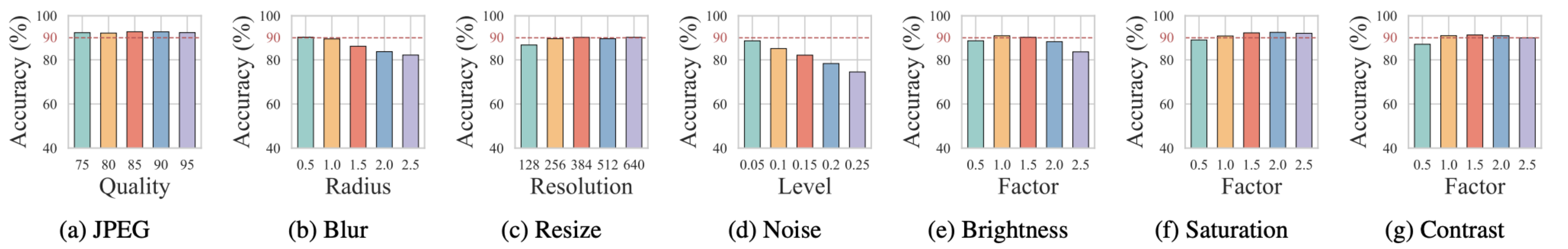


Figure 1: CO-SPY is resilient against a range of post-transformations.