

Supplementary Materials: SemGIR: Semantic-Guided Image Regeneration based method for AI-generated Image Detection and Attribution

Anonymous Authors

1 OVERVIEW

Our supplementary materials consist of three parts: the attribution experiment for VQDM, the precision metric of detection performance, and a visual comparison between the regeneration process of SemGIR and DIRE.

2 ATTRIBUTION EXPERIMENT FOR VQDM

In addition to attempting attribution for SDv1.5 and SDv2.1, we also attempt attribution for VQDM. The results are shown in Table 1.

From the table, it is evident that attribution for VQDM is more challenging compared to attribution for SDv1.5 and SDv2.1. However, we still achieved SOTA performance among numerous methods with the highest accuracy of 85.33%.

Table 1: The attribution accuracy comparison between SemGIR-A and baselines. Among all detectors, the best result and the second-best result are denoted in boldface and underlined, respectively.

Attribution Method	VQDM	SDV1.5	SDV2.1
CNNSpot	0.7530	0.7599	0.8021
FreDect	0.7918	0.7563	0.8332
Fusing	0.7534	0.8674	<u>0.9718</u>
GramNet	0.7776	0.8148	0.8054
LGrad	0.7229	0.8133	0.8433
LNP	<u>0.8484</u>	0.7387	0.9701
DIRE	0.8288	0.9114	0.9020
UnivFD	0.8000	0.7748	0.8040
SemGIR-A	0.8533	<u>0.9016</u>	0.9801

3 PRECISION METRIC OF DETECTION

In the main text, we mentioned using accuracy and precision metrics to evaluate the performance of various detection methods. The reason for using these two metrics is that accuracy can measure the overall performance of the detection method, while precision can measure the accuracy of the model when predicting the positive class. Due to page limitations, we only present the accuracy metric in the main text, and we will list the precision metric in this section.

We list the performance evaluation of various detection methods on the original images in Table 2. Additionally, we also present the detection performance after applying JPEG compression with quality factors of 70, 80, and 90 to the images in Table 3, Table 4, and Table 5, respectively.

From the tables, it can be observed that our method not only achieves the highest accuracy and precision for detecting original images but also maintains remarkable robustness when faced with JPEG compression at different levels.

4 VISUALIZATION

Finally, in Figure 1, we present a visualization of the regeneration results of SemGIR and DIRE. Given an AI-generated image, SemGIR performs semantically guided image regeneration, which is a semantic-level image regeneration, while DIRE directly regenerates the image at the pixel level, thus failing to ensure semantic-level consistency. We obtain captions for the re-generated images through BLIP to support our point.

From Figure 1, it can be observed that our method ensures semantic consistency between the re-generated image and the original AI-generated image, while the image reconstructed using DIRE fails to guarantee this semantic consistency. This also validates our viewpoint: as semantic consistency is ensured between the original and regenerated images, it effectively removes the influence of the text prompt, leading to a more thorough decoupling.

Table 2: The detection performance comparison between SemGIR and baselines. Among all detectors, the best result and the second-best result are denoted in boldface and underlined, respectively.

Generator	Metrics	Detection Methods								
		CNNSpot	FreDect	Fusing	GramNet	LGrad	LNP	DIRE	UnivFD	SemGIR
SDv1.5	Acc	0.5630	0.5250	<u>0.9960</u>	0.7950	0.8730	0.9980	0.7980	0.8840	0.9929
	Prec	0.9302	0.4649	<u>0.9999</u>	0.8652	0.9988	1.0000	0.9783	0.9371	1.0000
SDv2.1	Acc	0.7937	0.4850	<u>0.9405</u>	0.6680	0.8710	0.9985	<u>0.9710</u>	0.7960	0.9561
	Prec	0.6350	0.4508	<u>0.9983</u>	0.8090	0.9944	1.0000	0.9168	0.7414	0.9967
Midjourney	Acc	0.5763	0.5145	<u>0.7463</u>	0.8136	0.9845	1.0000	0.9522	0.8327	<u>0.9954</u>
	Prec	0.5497	0.4572	<u>0.9676</u>	0.7299	0.9639	0.9999	0.9280	0.9001	<u>0.9519</u>
ADM	Acc	0.6735	0.5250	<u>0.4990</u>	0.4995	<u>0.8205</u>	0.4880	0.5130	0.6380	0.9960
	Prec	0.7561	0.5388	<u>0.5755</u>	0.5277	0.9486	0.4874	<u>0.9578</u>	0.8626	0.9920
VQDM	Acc	0.4980	0.5005	<u>0.5220</u>	0.6455	<u>0.8725</u>	0.7665	0.6320	0.8285	0.9970
	Prec	0.6234	0.4832	<u>0.6450</u>	0.6875	0.9306	0.8961	0.9978	0.9872	<u>0.9950</u>
StyleGAN2	Acc	0.6260	0.5080	<u>0.9370</u>	0.7235	0.5250	0.6345	0.5147	0.8970	0.9546
	Prec	0.9575	0.4900	0.9939	0.8249	0.9323	0.9105	0.7601	<u>0.9599</u>	0.9509
Average	Acc	0.6217	0.5096	<u>0.7734</u>	0.6908	<u>0.8244</u>	0.8142	0.7301	0.8127	0.9820
	Prec	0.7790	0.5040	<u>0.7381</u>	0.6800	<u>0.9371</u>	0.7646	0.9052	0.9366	0.9793

Table 3: The detection performance under JPEG70 compression comparison between SemGIR and baselines. Among all detectors, the best result and the second-best result are denoted in boldface and underlined, respectively.

Generator	Metrics	Detection Methods								
		CNNSpot	FreDect	Fusing	GramNet	LGrad	LNP	DIRE	UnivFD	SemGIR
SDv1.5	Acc	0.5635	0.5250	0.9960	0.7970	0.7440	0.8695	0.5540	0.9160	0.9940
	Prec	0.9517	0.4548	1.0000	0.8777	0.9956	<u>0.9999</u>	0.8335	0.9769	1.0000
SDv2.1	Acc	0.5475	0.4850	<u>0.9650</u>	0.6695	0.8330	0.9975	0.8395	0.827	0.9501
	Prec	0.8157	0.4631	<u>0.9995</u>	0.7460	0.9931	1.0000	0.9000	0.7412	0.9969
Midjourney	Acc	0.6836	0.4981	<u>0.4736</u>	0.8590	<u>0.9725</u>	0.9985	0.9072	0.9018	0.8770
	Prec	0.6920	0.0584	<u>0.8752</u>	0.7156	0.9431	1.0000	0.4722	0.9510	<u>0.9706</u>
ADM	Acc	0.7020	0.5250	<u>0.4990</u>	0.4840	<u>0.8336</u>	0.4578	0.4766	0.6365	0.9970
	Prec	0.7598	0.5490	<u>0.5760</u>	0.5072	<u>0.9036</u>	0.4154	0.6983	0.8299	0.9960
VQDM	Acc	0.4850	0.5005	<u>0.5270</u>	0.6270	<u>0.8647</u>	0.5994	0.4766	0.8245	0.9980
	Prec	0.5409	0.4832	<u>0.6405</u>	0.6604	0.9270	0.6341	0.6466	<u>0.9777</u>	0.9970
StyleGAN2	Acc	0.6260	0.5095	<u>0.9370</u>	0.7235	0.5250	0.5050	0.5050	0.8970	0.9617
	Prec	0.9306	0.4902	0.9939	0.8292	0.8825	0.9468	0.7603	<u>0.9599</u>	0.9519
Average	Acc	0.6012	0.5071	<u>0.7329</u>	0.6933	0.7954	0.7379	0.6264	<u>0.8338</u>	0.9629
	prec	0.7817	0.4164	<u>0.8475</u>	0.7226	<u>0.9408</u>	0.8327	0.7184	0.9061	0.9854

Table 4: The detection performance under JPEG80 compression comparison between SemGIR and baselines. Among all detectors, the best result and the second-best result are denoted in boldface and underlined, respectively.

Generator	Metrics	Detection Methods								
		CNNSpot	FreDect	Fusing	GramNet	LGrad	LNP	DIRE	UnivFD	SemGIR
SDv1.5	Acc	0.5630	0.5250	0.9960	0.7940	0.7835	0.9575	0.5925	0.9105	<u>0.9940</u>
	Prec	0.9180	0.4615	1.0000	0.8733	0.9965	<u>0.9999</u>	0.8981	0.9712	1.0000
SDv2.1	Acc	0.5495	0.4850	0.9320	0.6680	0.8770	0.9985	0.8450	0.7820	<u>0.9753</u>
	Prec	0.7753	0.4484	<u>0.9986</u>	0.7386	0.9948	1.0000	0.9130	0.7486	0.9968
Midjourney	Acc	0.5145	0.5063	0.5263	0.7772	0.9825	1.0000	0.9102	0.8638	0.9430
	Prec	0.4720	0.0584	0.9152	0.7193	0.9576	1.0000	0.4657	0.9082	<u>0.9429</u>
ADM	Acc	0.6565	0.5250	0.4990	0.5025	<u>0.8063</u>	0.4610	0.4766	0.6330	0.9960
	Prec	0.7363	0.5326	0.5241	0.5305	<u>0.8899</u>	0.4487	0.6471	0.8128	0.9900
VQDM	Acc	0.5005	0.5005	0.5135	0.6530	<u>0.8631</u>	0.5642	0.4766	0.8255	0.9975
	Prec	0.6227	0.4784	0.6008	0.6859	0.9176	0.5720	0.5921	<u>0.9805</u>	0.9950
StyleGAN2	Acc	0.6260	0.5055	<u>0.9370</u>	0.7235	0.5250	0.5045	0.5095	0.8970	0.9551
	Prec	0.9354	0.4899	0.9723	0.8283	0.8878	0.9161	0.7686	<u>0.9599</u>	0.9570
Average	Acc	0.5683	0.5078	0.7339	0.6863	0.8062	0.7476	0.6350	<u>0.8186</u>	0.9768
	prec	0.7432	0.4115	0.8351	0.7293	<u>0.9407</u>	0.8227	0.7141	0.8968	0.9802

Table 5: The detection performance under JPEG90 compression comparison between SemGIR and baselines. Among all detectors, the best result and the second-best result are denoted in boldface and underlined, respectively.

Generator	Metrics	Detection Methods								
		CNNSpot	FreDect	Fusing	GramNet	LGrad	LNP	DIRE	UnivFD	SemGIR
SDv1.5	Acc	0.5630	0.5250	0.9960	0.7960	0.8545	0.9890	0.6775	0.9095	0.9940
	Prec	0.9314	0.4623	1.0000	0.8745	0.9985	<u>0.9999</u>	0.9506	0.9694	1.0000
SDv2.1	Acc	0.5520	0.4850	<u>0.9560</u>	0.6675	0.8750	0.9990	0.8475	0.7830	0.9516
	Prec	0.8000	0.4540	0.9988	0.7425	0.9946	1.0000	0.9148	0.7497	<u>0.9989</u>
Midjourney	Acc	0.5363	0.5018	0.5263	0.8136	<u>0.9785</u>	1.0000	0.9072	0.8809	0.9375
	Prec	0.5012	0.0583	0.9318	0.7212	<u>0.9601</u>	1.0000	0.4151	0.9119	0.9429
ADM	Acc	0.6460	0.5250	0.4990	0.4920	<u>0.8178</u>	0.4684	0.4766	0.6375	0.9965
	Prec	0.6978	0.5420	0.5429	0.5184	<u>0.9078</u>	0.4929	0.6122	0.8329	0.9910
VQDM	Acc	0.4955	0.5005	0.5245	0.6450	<u>0.8447</u>	0.6810	0.4766	0.8285	0.9970
	Prec	0.6039	0.4828	0.6375	0.6745	0.9025	0.7154	0.6024	<u>0.9827</u>	0.9950
StyleGAN2	Acc	0.6260	0.5085	<u>0.9370</u>	0.7235	0.5250	0.5045	0.5065	0.8970	0.9551
	Prec	0.9407	0.4900	0.9801	0.8292	0.9032	0.9451	0.7919	<u>0.9599</u>	0.9500
Average	Acc	0.5698	0.5076	0.7398	0.6896	0.8159	0.7736	0.6486	<u>0.8227</u>	0.9719
	prec	0.7458	0.4149	0.8485	0.7267	<u>0.9444</u>	0.8588	0.7145	0.9010	0.9796

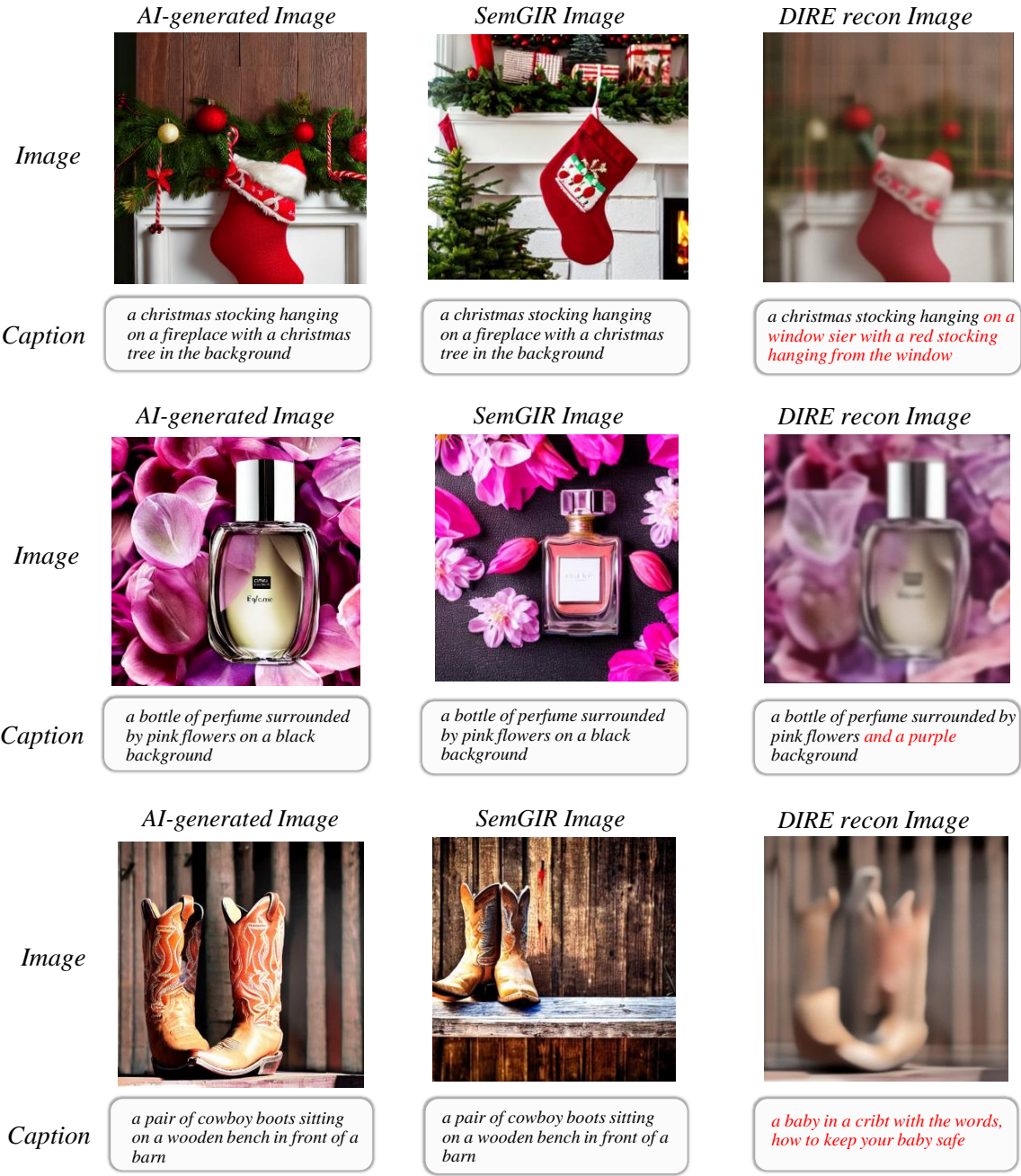


Figure 1: The visual comparison between the regeneration process of SemGIR and DIRE.