VLMDiff: Leveraging Vision-Language Models for Multi-Class Anomaly Detection with Diffusion

Supplementary Material

6. Overview

In this supplementary material, we first share the quantitative results on MVTec-AD [3] and VISA [67] datasets and comparisons with the same methods presented in the main text. Then, we present per-class results on the Real-IAD [46], and per-split results on the COCO-AD [59] dataset, only for the diffusion-based methods to compare. Finally, we show more visual results of our method and diffusion methods on each class of the Real-IAD dataset, and on each split of the COCO-AD dataset.

Table 8. Details of MVTec-AD and VISA datasets.

Dataset	Categories		Images		
	Train	Test	Train	Test	
			Normal	Anomaly	Normal
MVTec AD [3]	15	15	3,629	1,258	467
VisA [67]	12	12	8,659	962	1,200

6.1. MVTec and VISA results

Dataset statistics for MVTec-AD [3] and VISA [67] are presented in Table 8. We trained the best-performing diffusion methods for 100 and 300 epochs on the MVTec-AD [3] and VISA [67] datasets to compare their performance on the same epoch training regime. We present the results in Table 9 and Table 10 for MVTec-AD and VISA, respectively. On MVTec-AD, our method achieved the best ROC_P and PRO scores among the diffusion-based approaches, which show the exceptional localization performance of VLMDiff. VISA dataset results show similar patterns and our method achieved the best ROC_P score by improving more than 5 points.

	Method	ROC_I	ROC_P	PRO
Aug.	DRAEM [55]	54.5/55.2	47.6/48.7	14.3/15.8
	SimpleNet [29]	95.4/79.2	96.8/82.4	86.9/62.0
	RealNet [64]	84.8/82.9	72.6/69.8	56.8/51.2
Emb.	CFA [24]	57.6/55.8	54.8/43.9	25.3/19.3
	PatchCore [41]	98.8/ -	98.3/ -	94.2 / -
	CFLOW-AD [18]	91.6/92.7	95.7/95.8	88.3/89.0
	PyramidalFlow [25]	70.2/66.2	80.0/74.2	47.5/40.0
Hyb.	UniAD † [53]	92.5/96.8	95.8/96.8	89.3/91.0
	RD++ [45]	97.9/95.8	97.3/97.3	93.2/92.9
	DesTSeg [63]	96.4/96.3	92.0/92.6	83.4/82.6
Rec.	RD [13]	93.6/90.5	95.8/95.9	91.2/91.2
	ViTAD † [60]	98.3/98.4	97.6/97.5	92.0/91.7
	MambaAD [19]	97.8/98.5	97.4/97.6	93.4/93.6
Dif.	DiffAD [62]	80.7/91.8	79.7/88.4	65.1/78.4
	TransFusion [15]	90.4/**95.3**	80.9/90.6	72.4/83.5
	DiAD † [20]	88.9/92.0	89.3/89.3	63.9/64.4
	VLMDiff †	86.9/90.6	94.9/**95.9**	86.7/**89.4**

Table 9. Results on the MVTec AD dataset [3] for 100/300 epochs training. †: multi-class setting.

	Method	ROC_I	ROC_P	PRO
Aug.	DRAEM [55]	55.1/56.2	37.5/45.0	10.0/16.0
	SimpleNet [29]	86.4/80.7	96.6/94.4	79.2/74.2
	RealNet [64]	71.4/79.2	61.0/65.4	27.4/33.9
Emb.	CFA [24]	66.3/67.1	81.3/83.0	50.8/48.7
	CFLOW-AD [18]	86.5/87.2	97.7/97.8	86.8/87.3
	PyramidalFlow [25]	58.2/69.0	77.0/79.1	42.8/52.6
Hyb.	UniAD † [53]	89.0/91.4	98.3/98.5	86.5/89.0
	RD++ [45]	93.9/93.1	98.4/98.4	91.9/91.4
	DesTSeg [63]	89.9/89.0	86.7/84.8	61.1/57.5
Rec.	RD [13]	90.6/93.9	98.0/98.1	91.9/91.9
	ViTAD [60]	90.4/90.3	98.2/98.2	85.7/85.8
	MambaAD [19]	94.5/93.6	98.4/98.2	92.1/90.5
Dif.	DiffAD [62]	78.6/89.2	82.9/85.5	65.7/76.7
	TransFusion [15]	87.4/**92.5**	82.1/90.3	55.4/64.7
	DiAD † [20]	84.8/90.5	82.5/83.4	44.5/44.3
	VLMDiff †	79.0/80.9	96.0/**97.0**	77.0/**81.0**

Table 10. Results on the VISA AD dataset [67] for 100/300 epochs training. †: multi-class setting.

We conducted extended ablation studies to thoroughly evaluate our method. These experiments focused on three key aspects: 1) the choice of VLMs for extracting image descriptions during training, 2) the impact of including a specific prompt during the inference stage, and 3) the selection of the feature extractor for inference.

Our comparison of VLMs for image description extraction (Table 11) revealed that InternVL-2 consistently achieved the best overall performance across both datasets. Further investigation into inference-time prompting (Table 12) with InternVL-2 showed that employing prompt $\mathcal{P}_{\mathcal{D}}$ led to a noticeable performance drop on both datasets. Lastly, our analysis of different feature extractors during inference (Table 13) indicated that DINO ViT-S with a patch

Variants	MVTec-AD			VISA		
	ROC_I	ROC_P	PRO	ROC_I	ROC_P	PRO
Blip2	89.8	95.7	88.6	82.3	97.0	80.7
DeepSeekv3-1.3B	90.7	95.5	89.0	81.8	96.8	81.0
InternVL-2-8B	90.6	95.9	89.4	80.9	97.0	81.0

Table 11. Ablation experiments using different VLMs to extract anomaly descriptions on MVTec-AD and VISA datasets.

Dataset	ROC_I	ROC_P	PRO
MVTec-AD	-2.2	-2.4	-6.5
VISA	-8.1	-2.9	-9.9

Table 12. Relative change when we use $\mathcal{P}_{\mathcal{D}}$ query during inference. Using text description from InternVL-2 for inference has a negative impact on all metrics.

size of 8 delivered the strongest overall results.

6.2. COCO-AD per split results

Per-split results on COCO-AD [59] are shown in Table 14. VLMDiff shows a noticeable improvement compared to the baselines, especially in the first split where there are significantly fewer normal images.

6.3. Real-IAD per class results

Tables 15 and 16 present per-class results on the Real-IAD [46] dataset for diffusion-based methods. Except for a few cases, VLMDiff achieves the best ROC_P and PRO on all classes. A detailed overview of the performance of other methods can be found in [58].

6.4. More visuals for Real-IAD

We present more visual comparisons in Figures 6-11 on Real-IAD. Specifically, we show two results per object category in the dataset. VLMDiff shows superior localization capability compared to strong baselines. Moreover, in some cases, we observe that it even finds unmarked potential defective pixels. For instance, in Figure 6 first *bottle cap* image has a small blue dot which is marked as an anomaly by our method. Similarly, both *sim card* objects have small defective pixels which are again detected by VLMDiff.

6.5. More visuals for COCO-AD

Figures 12 and 13 show three example results per split, and in each split, we pick different anomaly classes to show the performance across various objects. As a real-world dataset, COCO-AD is more complex and challenging compared to previous industrial domain datasets. Nevertheless, VLMDiffachieves significantly better anomaly localization across multiple classes.

Model	Metrics		
	ROC_I	ROC_P	PRO
ImageNet R50	88.8	92.6	81.0
DINO-v2 ViTS/14	64.6	61.3	16.8
DINO R50	86.8	90.3	71.0
DINO ViTB/16	89.8	92.8	79.3
DINO ViTS/16	85.1	90.2	74.3
DINO ViTB/8	87.5	93.6	82.6
DINO ViTS/8	**90.6**	**95.9**	**89.4**

Table 13. Ablation experiments using variants of DINO and an ImageNet pretrained Resnet-50 for anomaly segmentation on MVTec-AD dataset with InternVL-2 descriptions.

	Method	ROC_I	ROC_P	PRO
Split-0	DiAD †	57.5	67.0	28.8
	TransFusion	56.1	54.8	12.8
	VLMDiff †	**62.6**	**74.3**	**43.8**
Split-1	DiAD †	54.4	71.3	28.8
	TransFusion	**57.1**	62.2	6.6
	VLMDiff †	52.7	**71.5**	**37.5**
Split-2	DiAD †	**63.8**	68.0	33.2
	TransFusion	61.4	58.4	2.7
	VLMDiff †	62.9	**69.3**	**40.7**
Split-3	DiAD †	**60.1**	**65.9**	32.3
	TransFusion	59.0	56.1	5.0
	VLMDiff †	58.2	60.8	**33.2**
Avg	DiAD †	59.0	68.1	30.8
	TransFusion	58.4	57.8	6.8
	VLMDiff †	**59.1**	**69.0**	**38.8**

Table 14. Per split results on the COCO-AD dataset [59] for 100 epochs training. †: multi-class setting.

	Method	ROC_I	ROC_P	PRO
audiojack	DiAD	76.5	91.6	63.3
	Transfusion	**80.3**	85.9	51.2
	VLMDiff	77.5	**97.8**	**87.1**
bottle cap	DiAD	**91.6**	94.6	73.0
	Transfusion	65.4	70.9	43.3
	VLMDiff	77.3	**98.4**	**92.3**
button battery	DiAD	80.5	84.1	66.9
	Transfusion	**88.1**	94.5	**76.8**
	VLMDiff	72.8	**96.9**	74.5
end cap	DiAD	**85.1**	81.3	38.2
	Transfusion	64.3	56.6	32.8
	VLMDiff	71.6	**96.4**	**85.9**
eraser	DiAD	**80.0**	91.1	67.5
	Transfusion	74.3	75.8	55.1
	VLMDiff	78.9	**98.2**	**89.6**
fire hood	DiAD	**83.3**	91.8	66.7
	Transfusion	72.0	84.9	57.7
	VLMDiff	73.2	**98.3**	**89.4**
mint	DiAD	**76.7**	91.1	64.2
	Transfusion	60.8	68.8	30.8
	VLMDiff	64.3	**93.8**	**73.0**
mounts	DiAD	75.3	84.3	48.8
	Transfusion	**81.5**	86.1	73.2
	VLMDiff	78.7	**98.8**	**93.2**
pcb	DiAD	**86.0**	92.0	66.5
	Transfusion	77.7	94.9	64.6
	VLMDiff	82.5	**98.3**	**88.9**
phone battery	DiAD	**82.3**	**96.8**	85.4
	Transfusion	77.0	88.0	66.6
	VLMDiff	80.8	91.1	**90.2**
plastic nut	DiAD	71.9	81.1	38.6
	Transfusion	75.4	90.7	59.5
	VLMDiff	**80.6**	**98.7**	**93.4**
plastic plug	DiAD	**88.7**	92.9	66.1
	Transfusion	82.2	91.9	76.6
	VLMDiff	70.5	**97.1**	**86.0**
porcelain doll	DiAD	**72.6**	93.1	70.4
	Transfusion	70.2	85.8	64.2
	VLMDiff	**72.6**	**97.7**	**88.7**
regulator	DiAD	72.1	84.2	44.4
	Transfusion	**74.3**	81.1	49.0
	VLMDiff	61.7	**97.7**	**88.0**
rolled strip	DiAD	68.4	87.7	63.4
	Transfusion	**98.0**	87.1	81.0
	VLMDiff	86.6	**99.6**	**98.3**

Table 15. Per class results on Real-IAD dataset for diffusion models, part 1.

	Method	ROC_I	ROC_P	PRO
sim card	DiAD	72.6	89.9	60.4
	Transfusion	91.8	96.6	82.5
	VLMDiff	**92.9**	**98.3**	**90.0**
switch	DiAD	73.4	90.5	64.2
	Transfusion	82.0	86.6	59.4
	VLMDiff	**83.9**	**96.8**	**90.7**
tape	DiAD	73.9	81.7	47.3
	Transfusion	**91.9**	94.6	83.7
	VLMDiff	89.5	**99.3**	**96.9**
terminalblock	DiAD	62.1	75.5	38.5
	Transfusion	70.6	85.6	70.3
	VLMDiff	**82.1**	**99.4**	**96.2**
toothbrush	DiAD	**91.2**	82.0	54.5
	Transfusion	88.5	87.5	66.1
	VLMDiff	80.3	**95.1**	**84.8**
toy	DiAD	66.2	82.1	50.3
	Transfusion	**81.0**	74.8	56.0
	VLMDiff	68.4	**90.8**	**78.8**
toy brick	DiAD	68.4	93.5	66.4
	Transfusion	65.1	76.2	47.0
	VLMDiff	**72.8**	**96.1**	**85.6**
transistor1	DiAD	73.1	88.6	58.1
	Transfusion	**86.9**	85.0	56.9
	VLMDiff	82.4	**96.7**	**85.5**
u block	DiAD	75.2	88.8	54.2
	Transfusion	78.9	91.0	65.6
	VLMDiff	**79.8**	**98.5**	**90.3**
usb	DiAD	58.9	78.0	28.0
	Transfusion	80.8	87.1	68.3
	VLMDiff	**88.7**	**99.4**	**96.3**
usb adaptor	DiAD	**76.9**	94.0	75.5
	Transfusion	69.9	87.2	57.8
	VLMDiff	71.6	**95.6**	**78.5**
vcpill	DiAD	64.1	90.2	60.8
	Transfusion	72.8	76.1	45.1
	VLMDiff	**83.5**	**97.3**	**85.2**
wooden beads	DiAD	62.1	85.0	45.6
	Transfusion	**79.3**	76.0	53.8
	VLMDiff	73.5	**97.2**	**85.8**
woodstick	DiAD	74.1	90.9	60.7
	Transfusion	**77.5**	91.2	67.5
	VLMDiff	69.2	**95.6**	**79.2**
zipper	DiAD	86.0	90.2	53.5
	Transfusion	**98.3**	87.4	85.4
	VLMDiff	91.8	**97.5**	**87.7**
Avg	DiAD	75.6	88.0	58.1
	Transfusion	**78.6**	84.2	61.6
	VLMDiff	78.0	**97.1**	**87.7**

Table 16. Per class results on Real-IAD dataset for diffusion models, part 2.

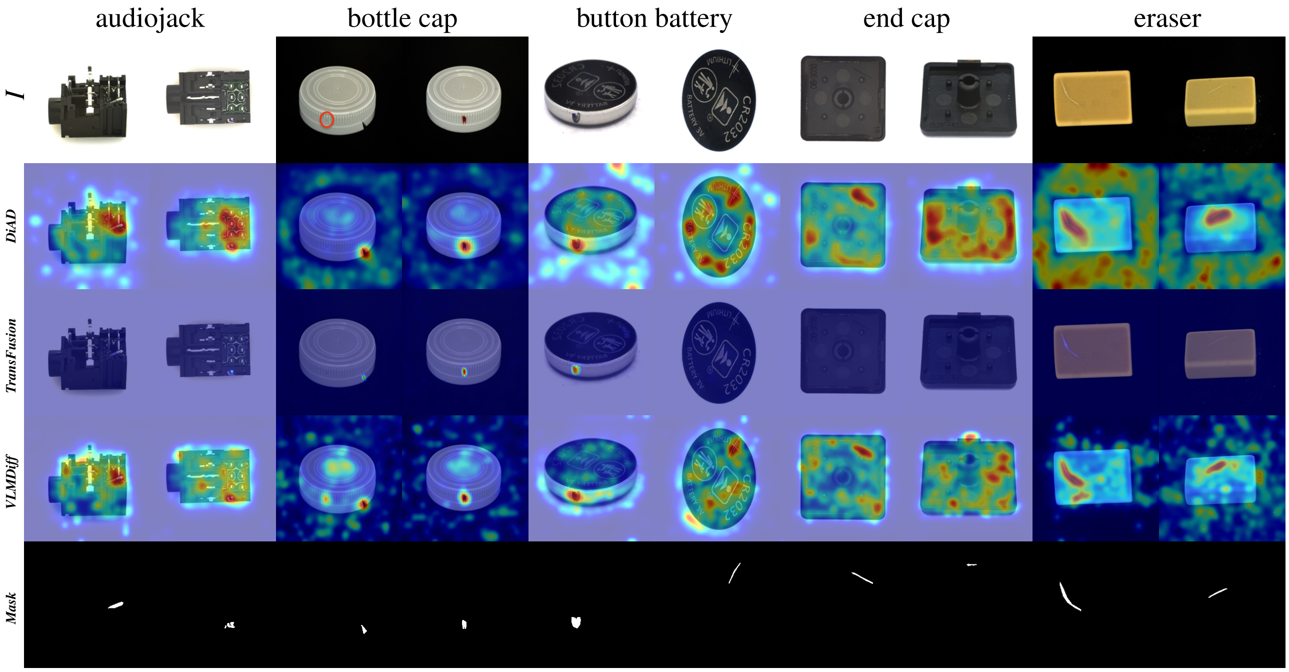

Figure 6. Visual comparison of diffusion-based methods on Real-IAD dataset.

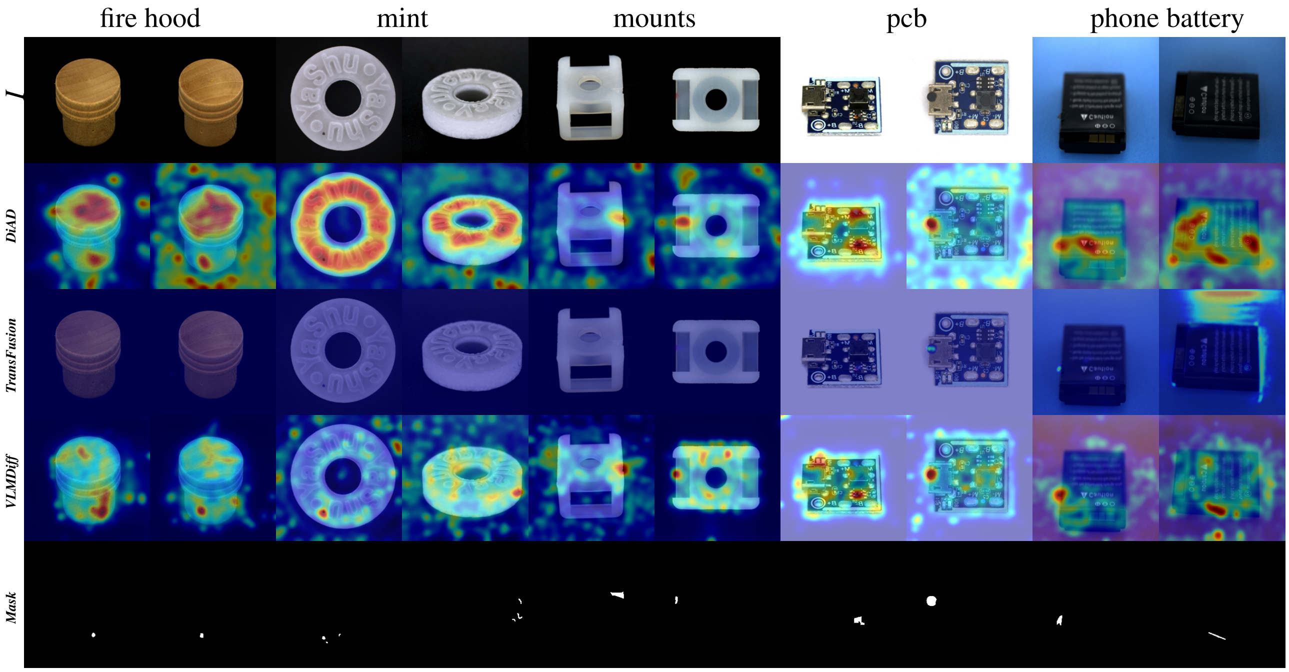

Figure 7. Visual comparison of diffusion-based methods on Real-IAD dataset.

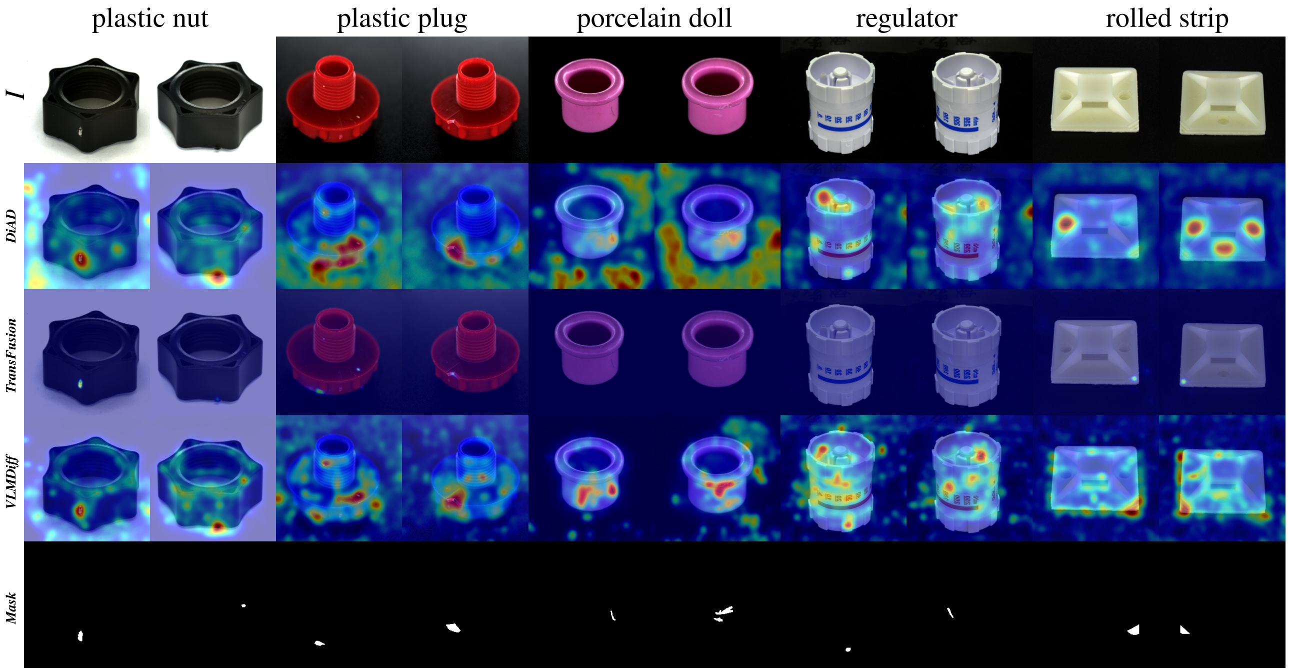

Figure 8. Visual comparison of diffusion-based methods on Real-IAD dataset.

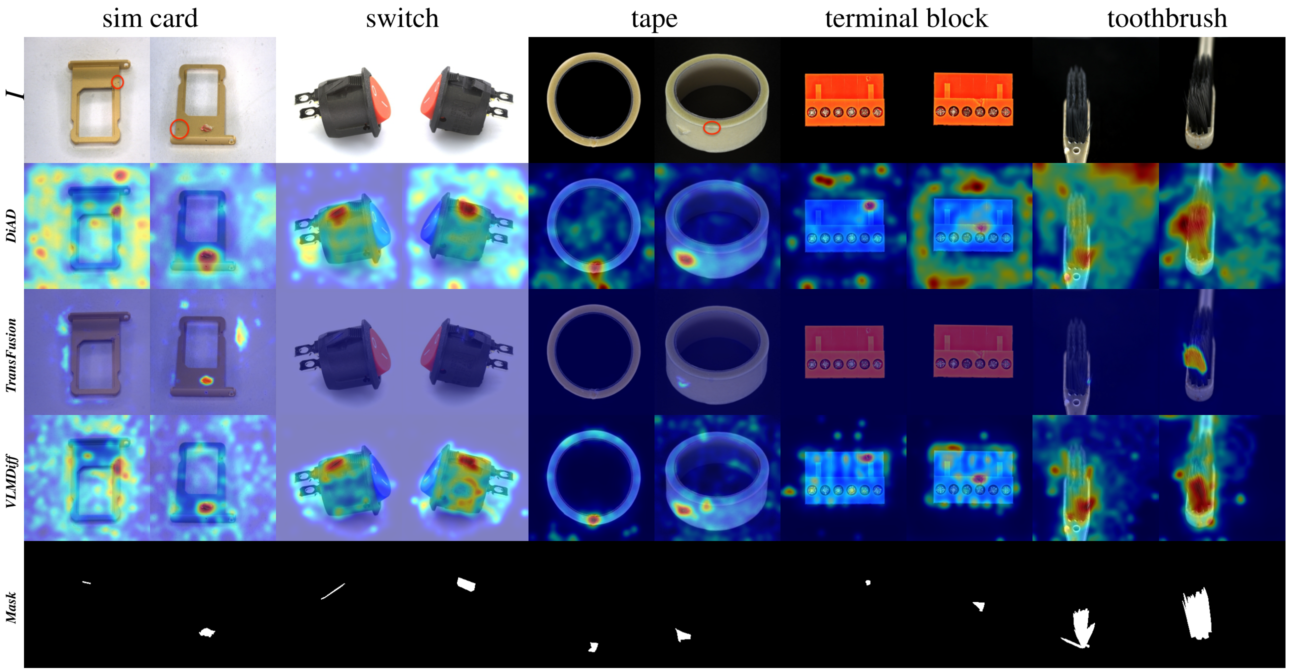

Figure 9. Visual comparison of diffusion-based methods on Real-IAD dataset.

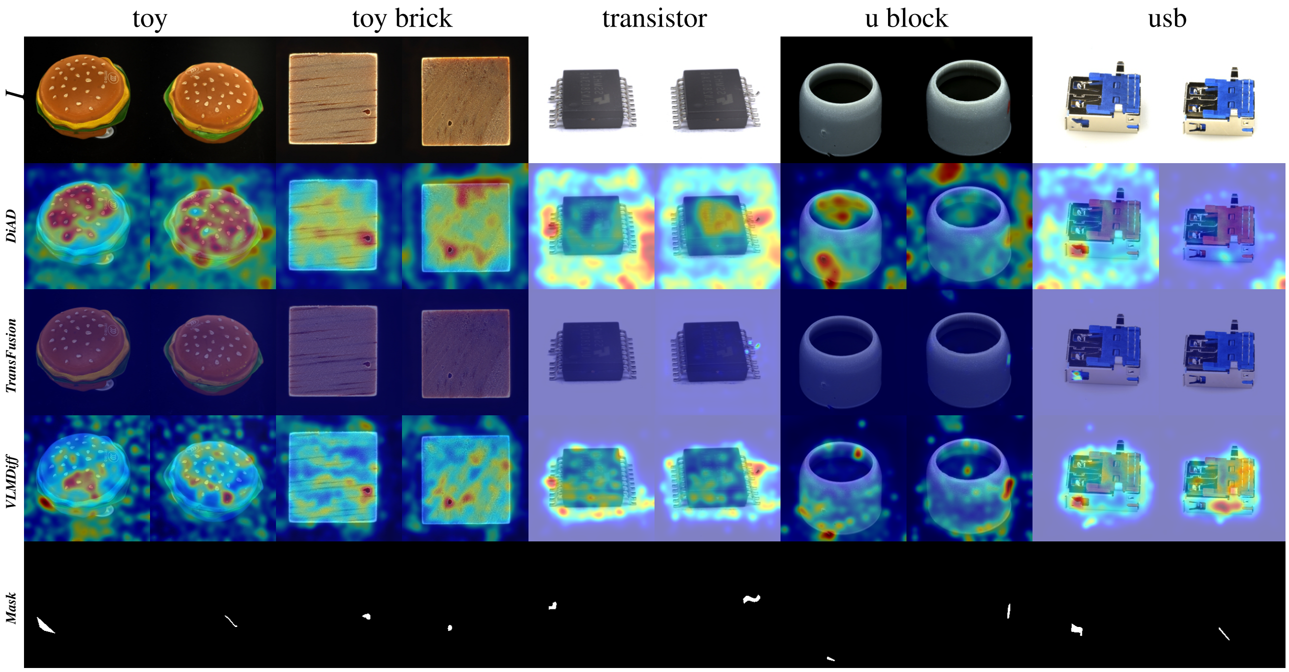

Figure 10. Visual comparison of diffusion-based methods on Real-IAD dataset.

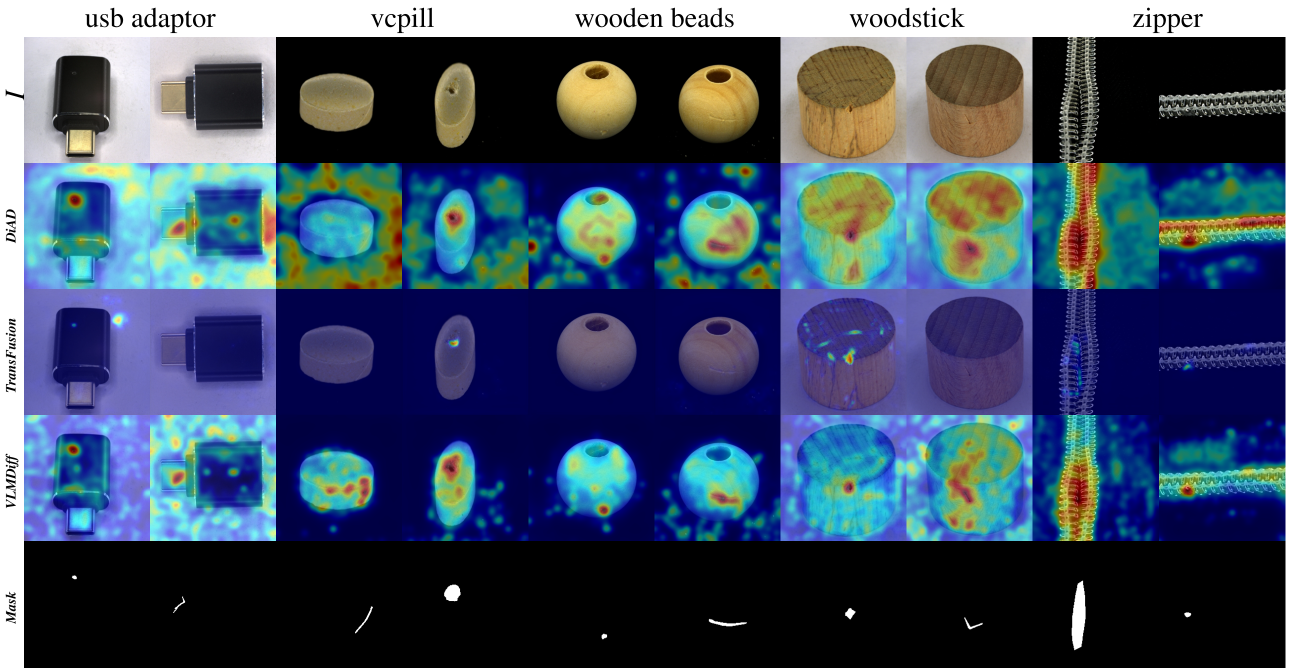

Figure 11. Visual comparison of diffusion-based methods on Real-IAD dataset.

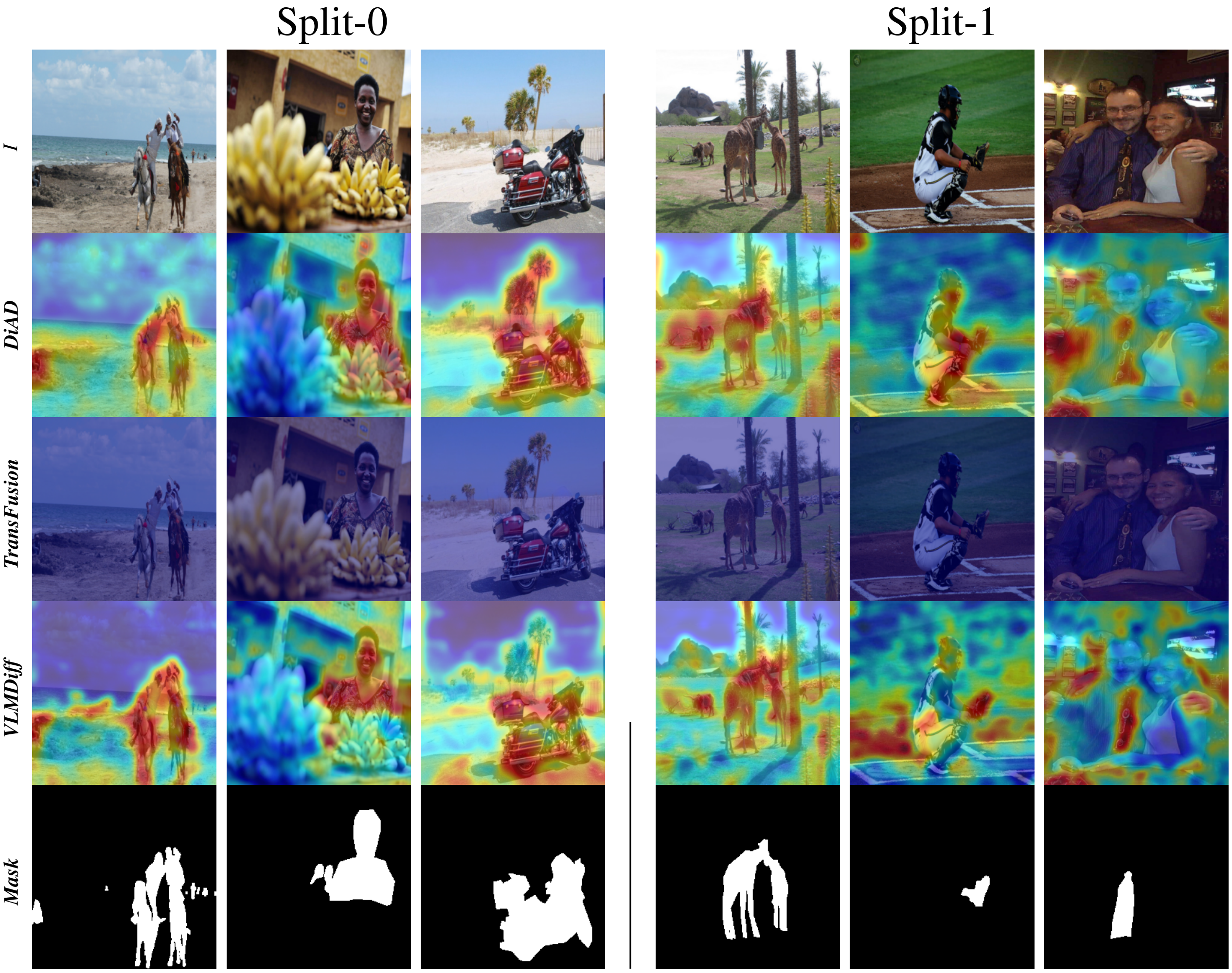

Figure 12. Visual comparison of diffusion-based methods on COCO-AD dataset on Split-0 and Split-1.

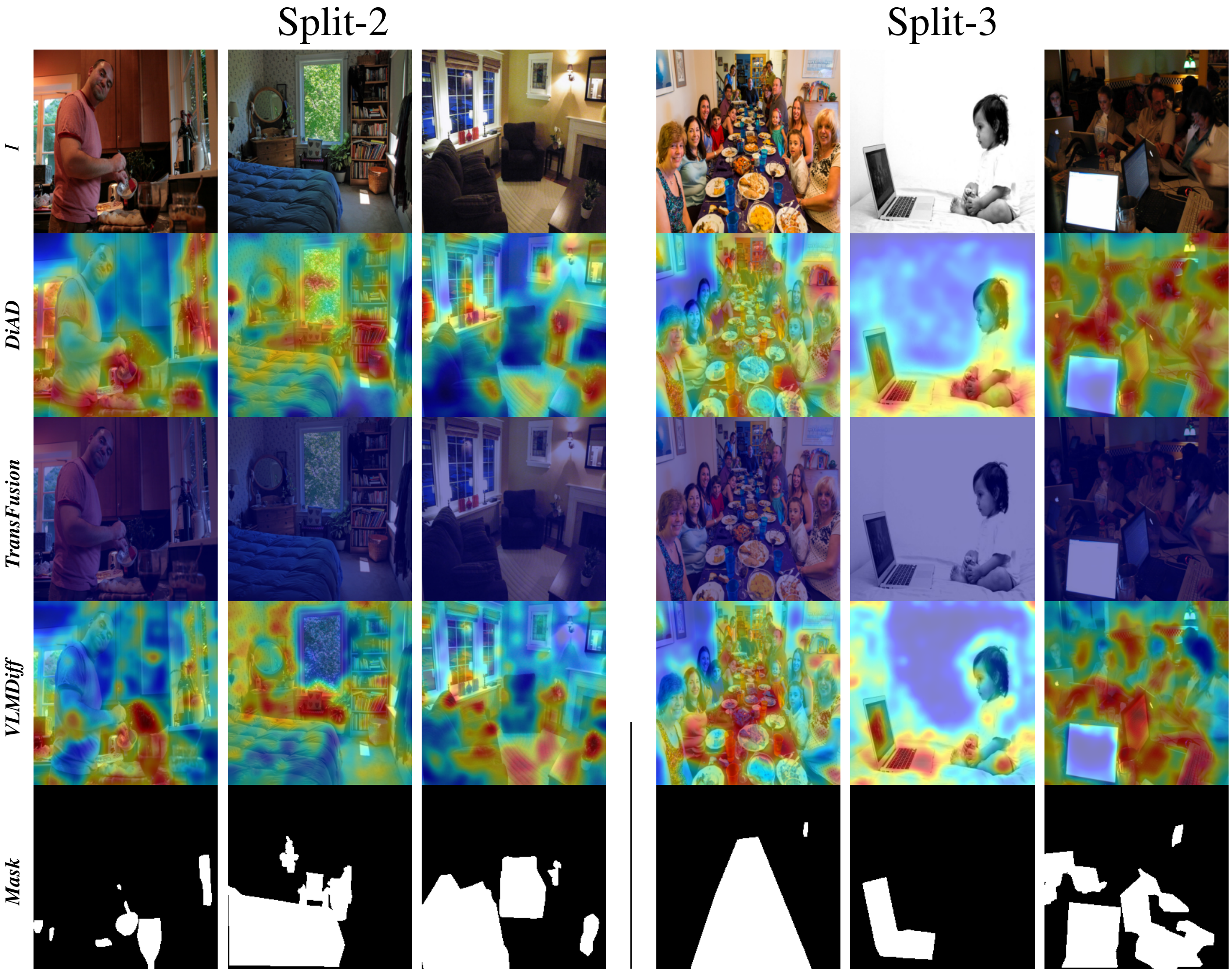

Figure 13. Visual comparison of diffusion-based methods on COCO-AD dataset on Split-2 and Split-3.