A Appendix

Overview

A.1 List of prompts

We provide the list of prompts that are used to guide the diffusion model to generate diverse background changes, encompassing different distribution shifts with respect to the original data distribution.

Table 6: Prompts used to create background alterations

Background	**Prompts**
Class label	*"This is a picture of a class name"*
BLIP-2 Caption	Captions generated from BLIP-2 image to caption
$\text{Color}_{\text{prompt-1}}$	*"This is a picture of a vivid red background"*
$\text{Color}_{\text{prompt-2}}$	*"This is a picture of a vivid green background"*
$\text{Color}_{\text{prompt-3}}$	*"This is a picture of a vivid blue background"*
$\text{Color}_{\text{prompt-4}}$	*"This is a picture of a vivid colorful background"*
$\text{Texture}_{\text{prompt-1}}$	*"This is a picture of textures in the background"*
$\text{Texture}_{\text{prompt-2}}$	*"This is a picture of intricately textured background"*
$\text{Texture}_{\text{prompt-3}}$	*"This is a picture of colorful textured background"*
$\text{Texture}_{\text{prompt-4}}$	*"This is a photo of distorted textures in the background"*
Adversarial	Captions generated from BLIP-2 image to caption.

A.2 Algorithm

We provide the algorithm (Algo. 1) for our approach of generating adversarial backgrounds by optimizing the textual and visual conditioning of the diffusion model. We also tried to optimize only the conditional embeddings or the latent embeddings, but achieve better attack success rate by optimizing both. Note that for crafting adversarial examples on `COCO-DC` we use ImageNet trained ResNet-50 classifiers and our adversarial objective is to maximize the feature representation distance between clean and adversarial samples. Furthermore, for introducing desired non-adversarial background changes using the textual description $\mathcal{T}'$, the optimization of the latent and embedding is not needed.

Algorithm 1 Background Generation

Require: Conditioning module $\mathcal{C}$, Diffusion model ϵ_θ, Autoencoder $\mathcal{V}$, CLIP text encoder $\psi_{\texttt{CLIP}}$, image $\mathcal{I}$, class label $\boldsymbol{y}$, classifier $\mathcal{F}_\phi$, denoising steps T, guidance scale λ, attack iterations N, and learning rate β for AdamW optimizer $\mathcal{A}$.

Get the textual and visual conditioning from the image $\mathcal{I}$

$$\mathcal{C}(\mathcal{I}, \boldsymbol{y}) = \mathcal{T}_{\mathbf{B}}, \mathcal{M}$$

Modify $\mathcal{T}_{\mathbf{B}}$ to $\mathcal{T}$ for desired background change.
Map the mask $\mathcal{M}$ and image $\mathcal{I}$ to latent space: $i, m \leftarrow \mathcal{V}_{\texttt{ENC}}(\mathcal{I}, \mathcal{M})$
Get the embedding of the textual discription $\mathcal{T}$: $e_{\mathcal{T}} \leftarrow \psi_{\texttt{CLIP}}(\mathcal{T})$
Randomly initialize the latent z_T
Get the denoised latent z_t at time step t.
for $n \in [1, 2, \ldots N]$ **do**
    for $t \in [t, t+1, \ldots T]$ **do**
        $\hat{\epsilon}^t_\theta(z_t, e_{\mathcal{T}}, i, m) = \epsilon^t_\theta(z_t, i, m) + \lambda\left(\epsilon^t_\theta(z_t, e_{\mathcal{T}}, i, m) - \epsilon^t_\theta(z_t, i, m)\right)$
        From noise estimate $\hat{\epsilon}_\theta$ get z_{t-1}.
    end for
    Project the latents to pixel space: $\mathcal{I}_{adv} \leftarrow \mathcal{V}_{\texttt{DEC}}(z_0)$
    Compute Adversarial Loss:

$$\mathcal{L}_{adv} = \mathcal{L}_{CE}(\mathcal{F}_\phi(\mathcal{I}_{adv}), \boldsymbol{y})$$

    Update z_t and $e_{\mathcal{T}}$ using $\mathcal{A}$ to maximize $\mathcal{L}_{adv}$:

$$z_t, e_{\mathcal{T}} \leftarrow \mathcal{A}\left(\nabla_{z_t}\mathcal{L}_{adv}, \nabla_{e_{\mathcal{T}}}\mathcal{L}_{adv}\right)$$

end for

Generate Adversarial image $\mathcal{I}_{adv}$ using updated z_t and $e_{\mathcal{T}}$.

A.3 Object Distortion in LANCE

In [44], LANCE method is proposed, which is closely relevant to our approach. LANCE leverages the capabilities of language models to create textual prompts, facilitating diverse image alterations using the prompt-to-prompt image editing method [22] and null-text inversion [41] for real image editing. However, this reliance on prompt-to-prompt editing imposes constraints, particularly limiting its ability to modify only specific words in the prompt. Such a limitation restricts the range of possible image transformations. Additionally, the global nature of their editing process poses challenges in preserving object semantics during these transformations. In contrast, our method employs both visual and textual conditioning, effectively preserving object semantics while generating varied background changes. This approach aligns well with our goal of studying object-to-background context. We use open-sourced code from LANCE to compare it against our approach both quantitatively and qualitatively. We use a subset of 1000 images, named $\texttt{ImageNet-B}_{1000}$, for comparison. We observe that our natural object-to-background changes including color and texture perform

favorably against LANCE, while our adversarial object-to-background changes perform significantly better as shown in the Table 1. Since LANCE relies on global-level image editing, it tends to alter the object semantics and distort the original object shape in contrast to our approach which naturally preserves the original object and alters the object-to-background composition only. This can be observed in qualitative examples provided in Figures 9 and 10. We further validate this effect by masking the background of original and LANCE-generated counterfactual images. As reported in Table 7, when the background is masked in LANCE-generated counterfactual images, overall accuracy drops from 84.35% to 71.57%. This drop in accuracy compared to original images with masked background, shows that the LANCE framework has distorted the original object semantics during optimization. In contrast to this, our proposed approach allows us to study the correlation of object-to-background compositional changes without distorting the object semantics.

We calculate the FID score by comparing the background changes applied on our `ImageNet-B` dataset with the original ImageNet val. set (Tab. 8). Our background modifications such as *Class Label*, *BLIP Caption*, & *Color* achieve FID scores close to the original images, while our more complex background changes (*Texture*, *Adversarial*) show significant improvement over related works [44, 61].

Table 7: Performance evaluation and comparison on $\texttt{ImageNet-B}_{1000}$ dataset. The drop in accuracy of LANCE dataset when the background is masked clearly highlight the image manipulation being done on the object of interest.

Dataset	**Masked Background**							
	ViT-T	ViT-S	Swin-T	Swin-S	Res-50	Res-152	Dense-161	Average
Original	70.5	86.1	84.2	87.6	87.2	91.2	83.7	84.35
LANCE	59.5	72.5	72.3	75.3	71.9	77.5	72.0	71.57

Table 8: FID comparison *(lower is better)*.

LANCE [44]	BG [61]	Material [61]	Texture [61]	Ours:	Class	BLIP	Color	Texture	Adv
88.51	68.99	120.18	132.28		35.05	30.98	31.65	45.11	67.57

A.4 Qualitative Comparison with Related Works

Fig. 9: Background Compositional changes on $\texttt{ImageNet-B}_{1000}$ dataset using LANCE and our method. LANCE fails to preserve object semantics, while our method exclusively edits the background.

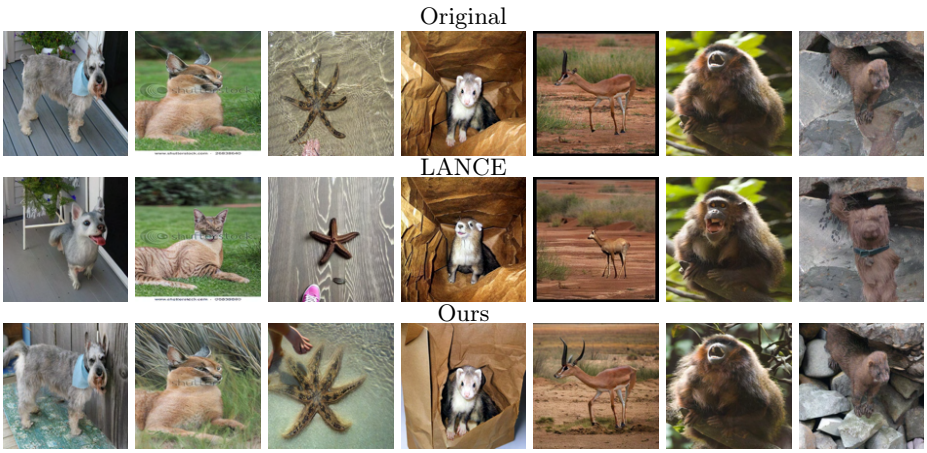

Fig. 10: Background Compositional changes on $\texttt{ImageNet-B}_{1000}$ dataset using LANCE and our method. LANCE fails to preserve object semantics, while our method exclusively edits the background.

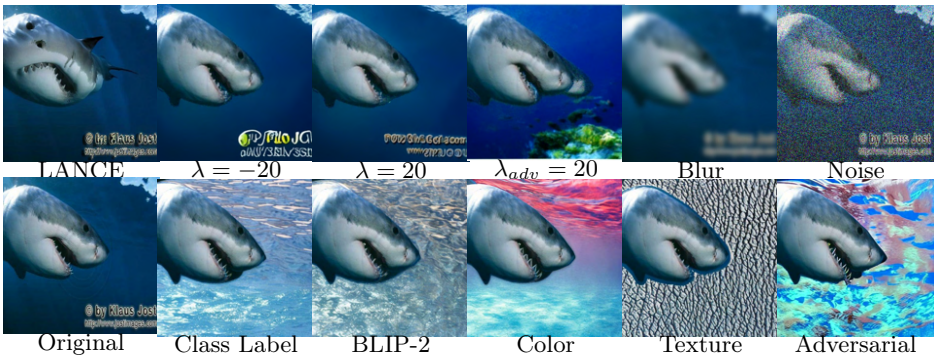

Fig. 11: Qualitative comparison of our background changes *(bottom row)* with previous related work *(top row)*. Our method enables diversity and controlled background edits.

A.5 Ablation on Background Changes

In this section, we report results on `ImageNet-B` and `COCO-DC` for uni-modal classifiers in Table 9 and Table 10 reports zero-shot classification results on `ImageNet-B`. Furthermore, ablations across diverse color and texture prompts is provided in Table 11 and 12.

Table 9: Resilience of Transformer and CNN models trained on ImageNet and COCO training sets against our proposed object-to-background context variations. We report top-1 (%) accuracy. We observe that CNN-based models are relatively more robust than Transformers.

Datasets	Background	Transformers				CNN			
		ViT-T	ViT-S	Swin-T	Swin-S	Res-50	Res-152	Dense-161	Average
`ImageNet-B`	Original	96.04	98.18	98.65	98.84	98.65	99.27	98.09	98.25
	Class label	92.82	94.75	96.18	96.55	97.24	97.56	95.8	95.84 (-2.41)
	BLIP-2 Caption	86.77	90.41	92.71	93.60	94.46	95.35	91.62	92.13 (-6.12)
	Color	70.64	84.52	86.84	88.84	89.44	92.89	83.19	85.19 (-13.06)
	Texture	68.24	79.73	81.09	84.41	83.21	87.66	77.29	80.23 (-18.02)
$\texttt{ImageNet-B}_{1000}$	Original	95.01	97.50	97.90	98.30	98.50	99.10	97.20	97.64
	Adversarial	18.40	32.10	25.00	31.70	2.00	28.00	14.40	21.65 (-75.99)
`COCO-DC`	Original	82.96	86.24	88.55	90.23	88.55	89.08	86.77	87.21
	BLIP-2 Caption	82.69	84.73	86.24	86.95	88.46	86.69	85.01	85.67 (-1.54)
	Color	55.54	61.04	70.09	72.13	74.97	75.10	66.19	66.66 (-20.55)
	Texture	52.52	58.82	68.05	70.09	70.71	74.77	63.79	63.99 (-23.22)
	Adversarial	49.68	55.72	61.93	69.12	55.45	61.13	57.76	58.68 (-28.52)

Table 10: Comparative Evaluation of Zero-shot CLIP and Eva CLIP Vision-Language Models on `ImageNet-B` and $\texttt{ImageNet-B}_{1000}$. Top-1(%) accuracy is reported. We find that Eva CLIP models showed more robustness in all object-to-background variations.

Datasets	Background	CLIP							
		ViT-B/32	ViT-B/16	ViT-L/14	Res50	Res101	Res50x4	Res50x16	Average
`ImageNet-B`	Original	75.56	81.56	88.61	73.06	73.95	77.87	83.25	79.12
	Class label	80.83	84.41	89.41	78.87	79.33	81.94	85.67	82.92 (+3.80)
	BLIP-2 Captions	69.33	73.66	79.07	67.44	68.70	71.55	75.78	72.22 (-6.90)
	Color	53.02	63.08	71.42	53.53	55.87	60.05	71.28	61.18 (-17.94)
	Texture	51.01	62.25	69.08	51.35	53.46	61.10	70.33	59.79 (-19.33)
$\texttt{ImageNet-B}_{1000}$	Original	73.90	79.40	87.79	70.69	71.80	76.29	82.19	77.43
	Adversarial	25.5	34.89	48.19	18.29	24.40	30.29	48.49	32.87 (-46.25)
Datasets	Background	EVA-CLIP							
		g/14	g/14+	B/16	L/14	L/14+	E/14	E/14+	Average
`ImageNet-B`	Original	90.80	93.71	90.24	93.71	93.69	95.38	95.84	93.34
	Class label	90.48	93.53	90.20	93.47	93.49	94.78	95.18	93.02 (-0.32)
	BLIP-2 Caption	80.56	85.23	81.88	85.28	86.24	88.13	88.68	85.14 (-8.20)
	Color	77.25	83.96	76.24	83.63	85.79	88.70	88.33	83.41 (-9.93)
	Texture	75.93	82.76	74.44	82.56	86.35	87.84	88.44	82.62 (-10.72)
$\texttt{ImageNet-B}_{1000}$	Original	88.80	92.69	89.19	91.10	91.99	93.80	94.60	91.74
	Adversarial	55.59	62.49	48.70	65.39	73.59	70.29	73.29	64.19 (-27.55)

Table 11: Performance evaluation of naturally trained classifiers and zero-shot CLIP models on `ImageNet-B`. The text prompts used for color and texture changes are provided in Table 6.

Background	**Naturally Trained Models**							
	ViT				**CNN**			
	ViT-T	ViT-S	Swin-T	Swin-S	ResNet50	ResNet152	DenseNet	Average
Clean	96.04	98.18	98.65	98.84	98.65	99.27	98.09	98.25
$\text{Color}_{\text{prompt-1}}$	76.58	86.43	88.92	91.23	91.08	93.79	86.05	87.72
$\text{Color}_{\text{prompt-2}}$	77.09	87.57	89.33	90.99	90.62	93.40	86.61	87.94
$\text{Color}_{\text{prompt-3}}$	76.80	86.97	88.74	90.99	90.62	93.18	87.41	87.82
$\text{Color}_{\text{prompt-4}}$	70.64	84.52	86.84	88.84	89.44	92.89	83.19	85.19
$\text{Texture}_{\text{prompt-1}}$	79.07	87.92	90.17	91.68	91.18	94.42	88.28	88.96
$\text{Texture}_{\text{prompt-2}}$	75.29	85.84	87.74	90.32	89.01	93.04	84.77	86.57
$\text{Texture}_{\text{prompt-3}}$	67.97	82.54	86.17	87.99	87.99	91.28	82.99	83.85
$\text{Texture}_{\text{prompt-4}}$	68.24	79.73	81.09	84.41	83.21	87.66	77.29	80.23

Background	**CLIP Models**							
	ViT			**CNN**				
	ViT-B/32	ViT-B/16	ViT-L/14	ResNet50	ResNet101	ResNet50x4	ResNet50x16	Average
Clean	75.56	81.56	88.61	73.06	73.95	77.87	83.25	79.12
$\text{Color}_{\text{prompt-1}}$	58.32	65.54	72.75	57.43	60.92	65.97	73.04	64.49
$\text{Color}_{\text{prompt-2}}$	57.91	67.28	74.44	58.67	60.12	65.9	74.13	65.49
$\text{Color}_{\text{prompt-3}}$	57.27	66.77	74.07	57.89	59.03	66.10	74.06	65.03
$\text{Color}_{\text{prompt-4}}$	53.02	63.08	71.42	53.53	55.87	60.05	71.28	61.18
$\text{Texture}_{\text{prompt-1}}$	59.05	68.50	75.67	60.38	61.78	66.99	74.31	66.67
$\text{Texture}_{\text{prompt-2}}$	58.60	68.01	74.40	58.29	59.56	66.34	74.67	65.69
$\text{Texture}_{\text{prompt-3}}$	52.89	64.30	68.70	53.29	55.35	61.58	69.35	60.78
$\text{Texture}_{\text{prompt-4}}$	51.01	62.25	69.08	51.35	53.46	61.10	70.33	59.79

Table 12: Performance evaluation of naturally trained classifiers on `COCO-DC` dataset. The text prompts used for color and texture changes are provided in Table 6.

Background	**ViT**				**CNN**		
	ViT-T	ViT-S	Swin-T	Swin-S	ResNet50	Dense-161	Average
Clean	82.96	86.24	88.55	90.23	88.55	86.77	87.21
$\text{Color}_{\text{prompt-1}}$	61.66	65.92	73.38	73.73	75.86	71.6	70.35
$\text{Color}_{\text{prompt-2}}$	64.86	70.01	76.84	77.10	77.81	75.06	73.61
$\text{Color}_{\text{prompt-3}}$	62.64	67.52	73.29	74.09	77.28	73.64	71.41
$\text{Color}_{\text{prompt-4}}$	55.54	61.04	70.09	72.13	74.97	66.19	66.66
$\text{Texture}_{\text{prompt-1}}$	67.96	70.36	75.42	78.70	79.94	73.55	74.32
$\text{Texture}_{\text{prompt-2}}$	63.97	69.56	74.62	77.72	78.97	75.15	73.33
$\text{Texture}_{\text{prompt-3}}$	52.52	58.82	68.05	70.09	70.71	63.79	63.99
$\text{Texture}_{\text{prompt-4}}$	56.16	61.57	66.72	70.18	69.56	67.25	65.24

A.6 Evaluation on Adversarially Trained models

In this section, we evaluate adversarially trained Res-18, Res-50, and Wide-Res-50 models across background changes induced by our methods and baseline methods (See Table 13, 14, and 15).

Table 13: Performance evaluation and comparison of our dataset with state of the art methods on adversarially trained Res-18 models. The images are generated on $\texttt{ImageNet-B}_{1000}$ dataset. We report top-1 average accuracy of models trained on various adversarial budget.

Datasets	ℓ_∞				ℓ_2			
	ε=0.5	ε=2.0	ε=4.0	ε=8.0	ε=0.5	ε=2.0	ε=4.0	ε=8.0
Original	88.00	78.30	69.70	54.40	84.60	81.40	68.80	57.50
ImageNet-E (λ=-20)	84.50	77.01	69.10	54.80	83.40	80.00	68.80	59.00
ImageNet-E (λ=20)	81.41	74.94	66.36	52.52	81.61	75.75	65.65	55.05
ImageNet-E (λ_{adv} = 20)	75.15	66.16	56.36	45.35	72.82	66.66	55.75	45.05
LANCE	76.37	66.78	59.13	45.17	74.99	73.19	61.60	48.68
Class label	87.10	79.90	69.40	57.30	85.00	79.90	70.90	57.80
BLIP-2 Caption	80.90	73.10	67.10	51.00	79.50	75.30	63.10	53.40
Color	56.90	45.80	35.40	25.70	53.20	46.80	32.80	23.40
Texture	59.20	47.10	38.60	28.70	54.30	48.10	35.50	26.20
Adversarial	12.10	19.80	24.60	26.80	10.90	12.40	16.90	17.20

Table 14: Performance evaluation and comparison of our dataset with state of the art methods on adversarially trained Res-50 models. The images are generated on $\texttt{ImageNet-B}_{1000}$ dataset. We report top-1 average accuracy of models trained on various adversarial budget.

Datasets	ℓ_∞				ℓ_2			
	ε=0.5	ε=2.0	ε=4.0	ε=8.0	ε=0.5	ε=2.0	ε=4.0	ε=8.0
Original	95.20	89.30	83.20	72.40	94.30	91.10	80.90	70.80
ImageNet-E (λ=-20)	93.10	89.20	82.00	70.70	91.70	88.50	79.60	69.10
ImageNet-E (λ=20)	92.52	86.36	80.60	67.97	90.40	86.96	76.26	68.08
ImageNet-E (λ_{adv} = 20)	84.44	78.78	71.71	58.58	80.50	76.76	65.75	56.86
LANCE	81.94	78.96	70.11	59.72	83.52	80.46	69.83	61.32
Class label	92.40	88.50	82.70	72.50	90.70	88.60	80.20	70.50
BLIP-2 Caption	87.90	83.70	79.00	67.90	86.60	84.60	73.70	65.70
Color	70.80	60.30	53.20	39.50	67.20	58.50	44.40	34.20
Texture	69.70	61.00	54.60	43.40	64.90	59.70	48.00	37.70
Adversarial	10.80	17.10	18.10	16.60	10.70	11.90	14.70	13.40

Table 15: Performance evaluation and comparison of our dataset with state of the art methods on adversarially trained Wide Res-50 models. The images are generated on $\texttt{ImageNet-B}_{1000}$ dataset. We report top-1 average accuracy of models trained on various adversarial budget.

Datasets	ℓ_∞				ℓ_2			
	ε=0.5	ε=2.0	ε=4.0	ε=8.0	ε=0.5	ε=2.0	ε=4.0	ε=8.0
Original	96.20	92.60	89.10	78.70	95.60	94.00	87.00	78.30
ImageNet-E (λ=-20)	94.10	91.10	86.60	76.40	93.20	91.60	84.20	76.90
ImageNet-E (λ=20)	92.82	89.29	83.53	75.15	91.21	88.68	81.11	73.63
ImageNet-E (λ_{adv} = 20)	87.17	76.56	66.16	82.92	82.22	82.00	71.41	62.42
LANCE	84.18	81.34	77.07	64.51	83.48	83.09	77.40	66.09
Class label	93.50	90.60	87.30	78.80	92.10	90.80	83.70	75.50
BLIP-2 Caption	90.20	86.20	82.80	74.30	88.90	86.70	80.00	69.20
Color	72.2	66.60	60.70	51.20	68.10	65.10	51.40	40.90
Texture	73.80	66.70	61.70	53.60	67.20	64.60	52.00	44.20
Adversarial	13.90	22.90	28.00	32.00	12.40	15.10	19.50	20.60

A.7 Evaluation on Recent Vision Models

We have conducted experiments on recent transformer CNN based models like DeiT [57] and ConvNeXt [37], and their results are presented in Table 16. We observe a consistent trend in model performance on our dataset, revealing that even the modern vision models are vulnerable to background changes.

Table 16: Performance evaluation on naturally trained classifiers on `ImageNet-B` and $\texttt{ImageNet-B}_{1000}$ dataset. All models exhibit a marked decrease in accuracy when the background is modified, highlighting their sensitivity to changes in the environment. The decline in performance is minimal with class label backgrounds but more pronounced with texture and color alterations. The most significant accuracy drop occurs under adversarial conditions, underscoring the substantial challenge posed by such backgrounds to the classifiers.

Datasets	Background	**Transformers**				**CNN**			
		DeiT-T	DeiT-S	DeiT-B	Average	ConvNeXt-T	ConvNeXt-B	ConvNeXt-L	Average
`ImageNet-B`	Original	96.36	99.27	99.41	98.34	99.07	99.21	99.40	99.22
	Class label	94.18	96.85	97.74	96.25	97.60	97.51	97.51	97.54
	BLIP-2 Caption	89.33	94.29	95.07	92.89	94.64	94.82	95.47	94.97
	Color	80.96	89.48	91.11	87.13	92.11	93.58	93.58	93.09
	Texture	74.15	84.01	86.75	81.63	88.50	89.50	91.13	89.71
$\texttt{ImageNet-B}_{1000}$	Original	95.44	99.10	99.10	97.88	99.00	99.00	92.92	96.97
	Adversarial	20.40	29.62	34.81	28.27	32.88	42.52	48.60	41.33

A.8 Evaluation on DINOv2 models

Our findings underscore the necessity of training vision models to prioritize discriminative and salient features, thereby diminishing their dependence on background cues. Recent advancements, such as the approaches by [53] employing a segmentation backbone for classification to improve adversarial robustness and by [9] using additional learnable tokens known as *registers* for interpretable attention maps, resonate with this perspective. Our preliminary experiments with the DINOv2 models [43], as presented in Table 17, corroborate this hypothesis. Across all the experiments, models with registers *(learnable tokens)* provide more robustness to background changes, with significant improvement seen in the adversarial background changes.

Table 17: Performance comparison of classifiers that are trained different on `ImageNet-B` dataset. The DINOv2 model with registers generally shows higher robustness to background changes, particularly in the presence of color, texture and adversarial backgrounds. This suggests that the incorporation of registers in DINOv2 enhances its ability to maintain performance despite challenging background alterations.

Dataset	Background	**Dinov2**				**Dinov2$_{\textbf{registers}}$**			
		ViT-S	ViT-B	ViT-L	Average	ViT-S	ViT-B	ViT-L	Average
`ImageNet-B`	Original	96.78	97.18	98.58	97.51	97.71	98.05	99.14	98.30
	Class label	94.62	96.02	97.18	95.94	95.55	96.44	97.94	96.64
	BLIP-2 Caption	89.22	91.73	94.33	91.76	90.86	92.10	95.02	92.66
	Color	83.85	89.68	93.31	88.94	85.88	91.15	94.64	90.55
	Texture	83.63	89.08	92.44	88.38	84.98	91.03	93.97	89.99
`ImageNet-B`$_{1000}$	Original	95.12	96.50	98.10	96.57	97.91	97.80	99.00	98.23
	Adversarial	54.31	71.62	80.87	68.93	58.30	76.21	84.50	73.00

A.9 Vision Language model for Image Captioning

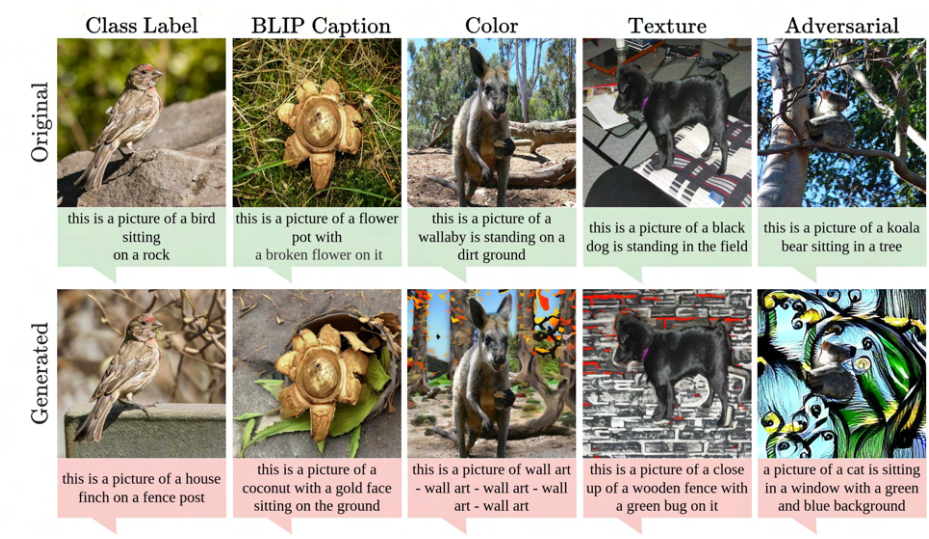

Fig. 12: A visual comparison of BLIP-2 captions on clean and generated datasets. The top row shows captions on clean images, while the bottom row displays captions on generated images. As background complexity increases, BLIP-2 fails to accurately represent the true class in the image.

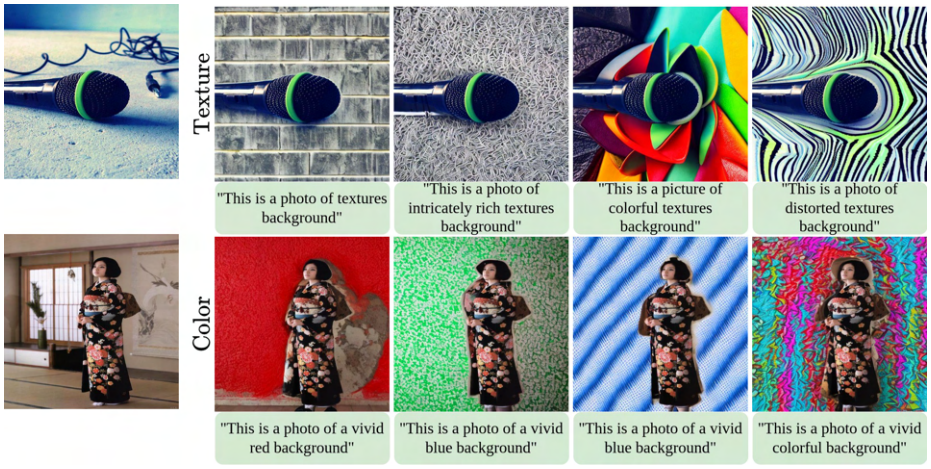

Fig. 13: The figure illustrates the introduction of background variations achieved through a diverse set of texture and color text prompts

A.10 Qualitative Results on Detection

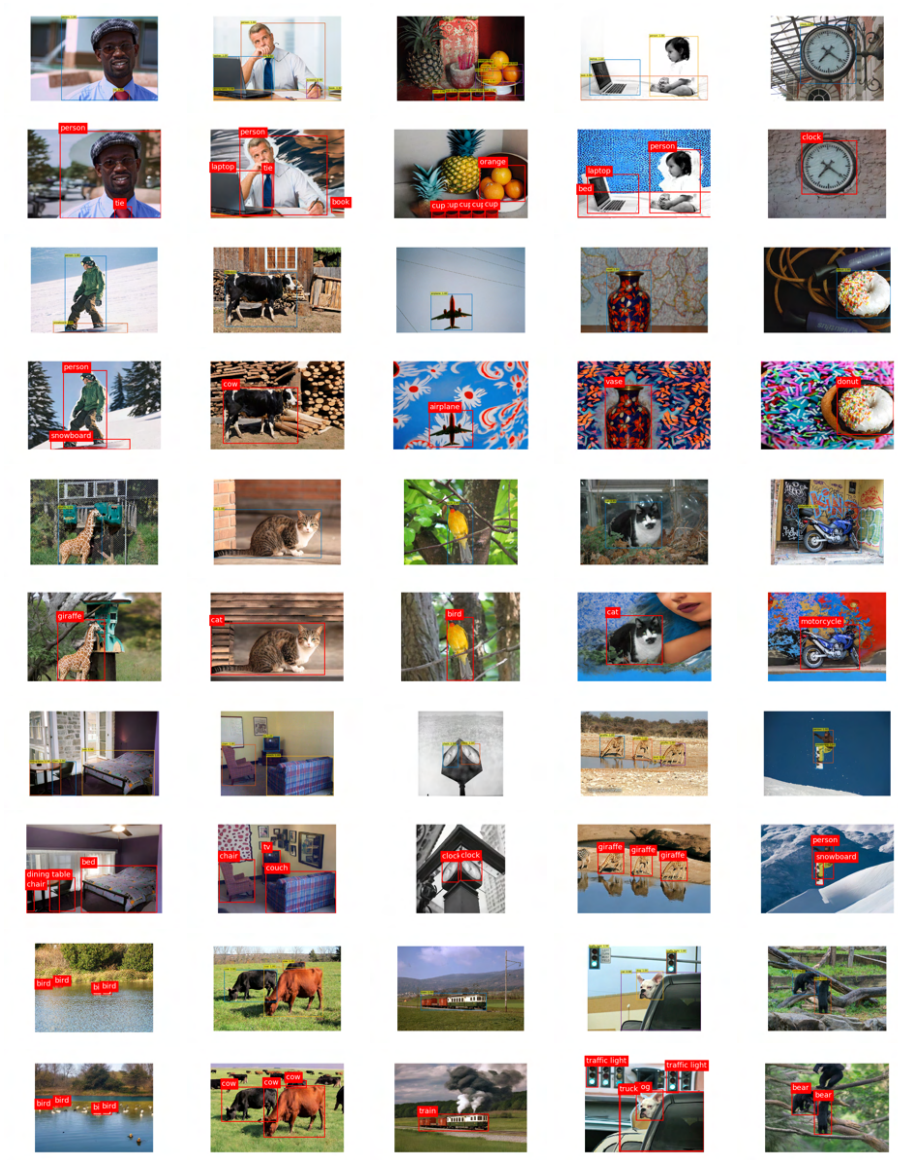

Fig. 14: We use diverse prompts to capture the diverse background shifts on samples from COCO-DC. The figure illustrate a comparison of prediction of Mask-RCNN on both clean and generated samples on COCO-DC. Each two adjacent rows represents the prediction of Mask-RCNN on clean *(top)* and generated images *(bottom)*.

A.11 Effect of Background Change on Segmentation Models

Figure 15, 16, and 17 provide failure cases of FastSAM to correctly segment the object in the images where background has been changed in terms of color, texture, and adversarial, respectively. Since we obtain the object masks for `ImageNet-B` using FastSAM, we compare those masks using IoU with the ones obtained by FastSAM on the generated dataset (see Table 18).

Table 18: IoU distribution of FastSAM. Percentage of images within an IoU range is reported.

Background	0.0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
Class Label	8.10	5.93	8.02	13.03	64.92
BLIP-2 Caption	5.70	4.81	6.92	13.01	69.56
Color	1.65	1.39	2.31	4.99	89.65
Texture	2.11	1.02	1.78	4.07	91.02
Adversarial	4.87	2.91	4.32	10.63	77.27

Table 19: DETR Object detection evaluation on `COCO-DC`

Background	Box AP	Recall AR
Original	0.65	0.81
BLIP-2 Caption	0.53	0.76
Color	0.52	0.73
Texture	0.52	0.71
Adversarial	0.42	0.62

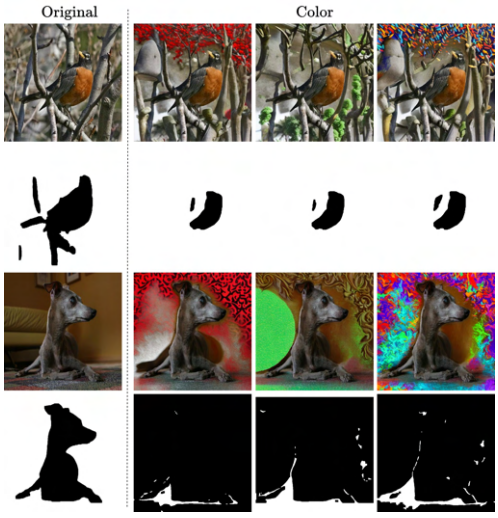

Fig. 15: Instances illustrating FastSAM model's failure to accurately segment masks for the background color changes on `ImageNet-B` samples..

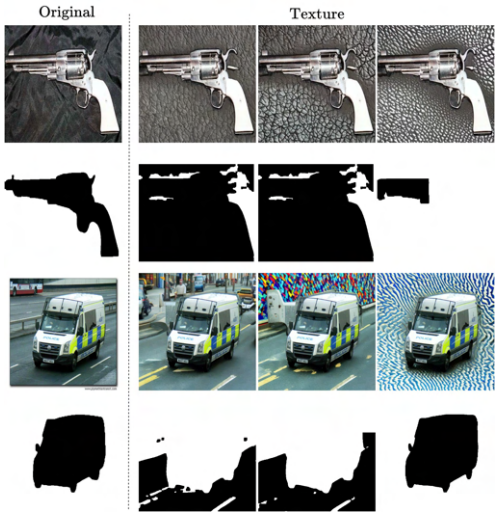

Fig. 16: Instances illustrating FastSAM model's failure to accurately segment masks for the background texture changes on `ImageNet-B` samples.

Fig. 17: Instances illustrating FastSAM model's failure to accurately segment masks for the adversarial background changes on $\texttt{ImageNet-B}_{1000}$ samples.

A.12 Exploring Feature Space of Vision Models

In Figure 18 and 19, we explore the visual feature space of vision and vision language model using t-SNE visualizations. We observe that as the background changes deviate further from the original background, a noticeable shift occurs in the feature space. The distinct separation or clustering of features belonging to the same class appears to decrease. This observation suggests a significant correlation between the model's decision-making process and the alterations in the background. Furthermore, we also show the GradCAM [50] on generated background changes. We observe that diverse background changes significantly shift the attention of the model as can be seen from Figure 20 and 21.

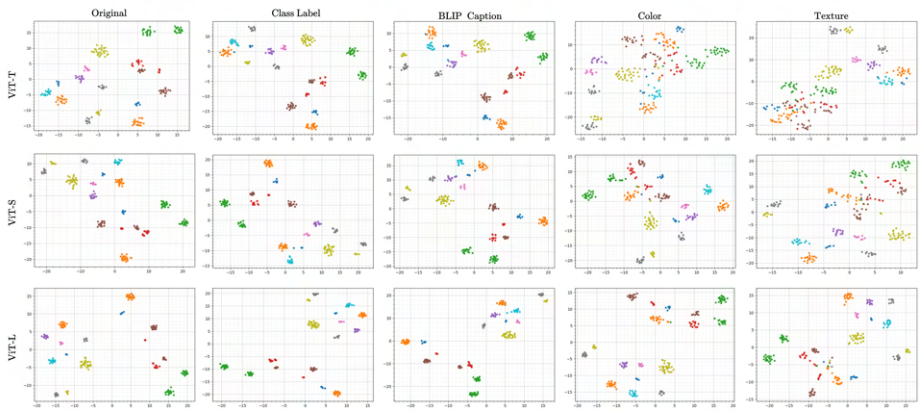

Fig. 18: t-SNE visualization of classifier models on `ImageNet-B` dataset.

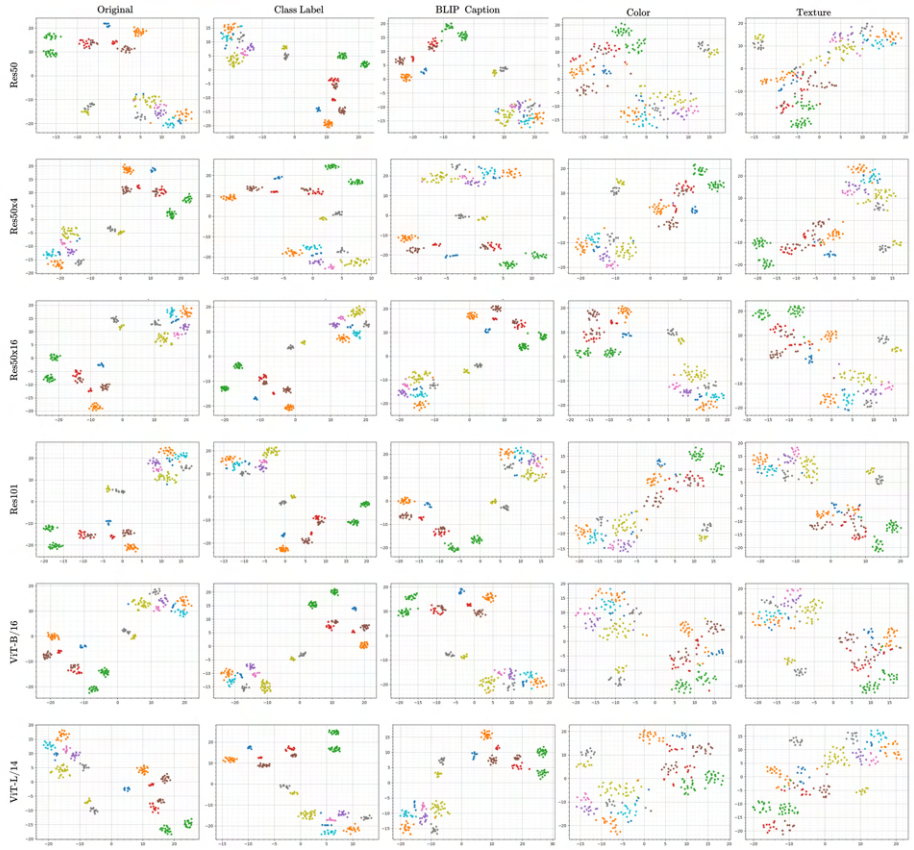

Fig. 19: t-SNE visualization of CLIP Vision Encoder features on `ImageNet-B` dataset.

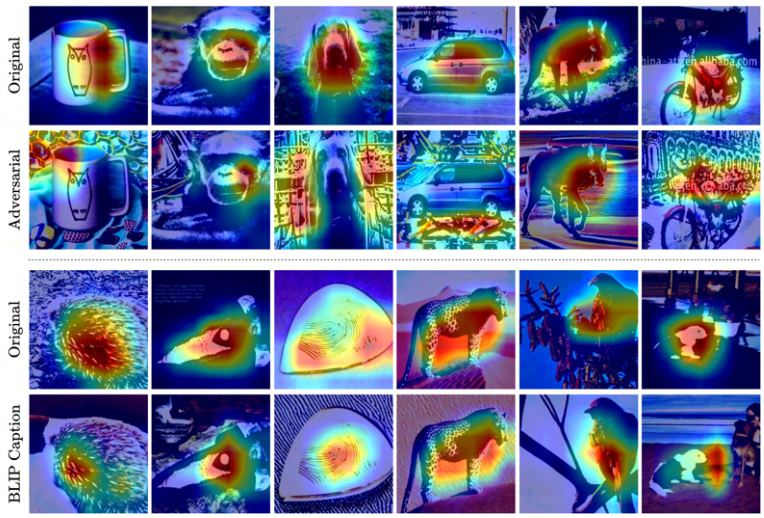

Fig. 20: GradCAM [50] visualization of adversarial and BLIP-2 background examples. The activation maps were generated on ImageNet pre-trained Res-50 model.

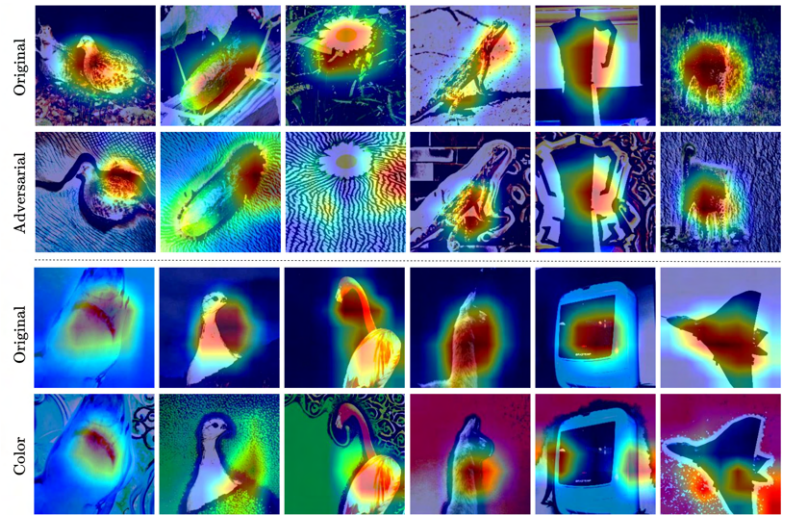

Fig. 21: GradCAM [50] visualization of texture and color background changes. The activation maps were generated on ImageNet pre-trained Res-50 model.

A.13 Diversity and Diffusion parameter ablation

In this section, we qualitatively analyze the diversity in visual results of the diffusion model. In Figure 22, we show that keeping textual and visual guidance fixed, the diffusion model is still able to generate diverse changes with similar background semantics at different seeds for the noise z_T. Furthermore, we explore the diversity in generating realistic background changes across an original image by using diverse class agnostic textual prompts, capturing different realistic backgrounds. Figure 23 and 24 show some of the qualitative results obtained on `ImageNet-B` samples using prompts generated from ChatGPT). Furthermore, we show the visual examples of color, texture, and adversarial attack on `ImageNet-B` dataset in Figure 25, 26, and 27. We also provide a visualization in Figure 28 showing the effect of changing diffusion model parameters.

Fig. 22: In this figure, examples are generated using BLIP-2 captions by altering the seed from left to right in the row. This highlights the high diversity achievable with the diffusion model when employing different starting noise latents.

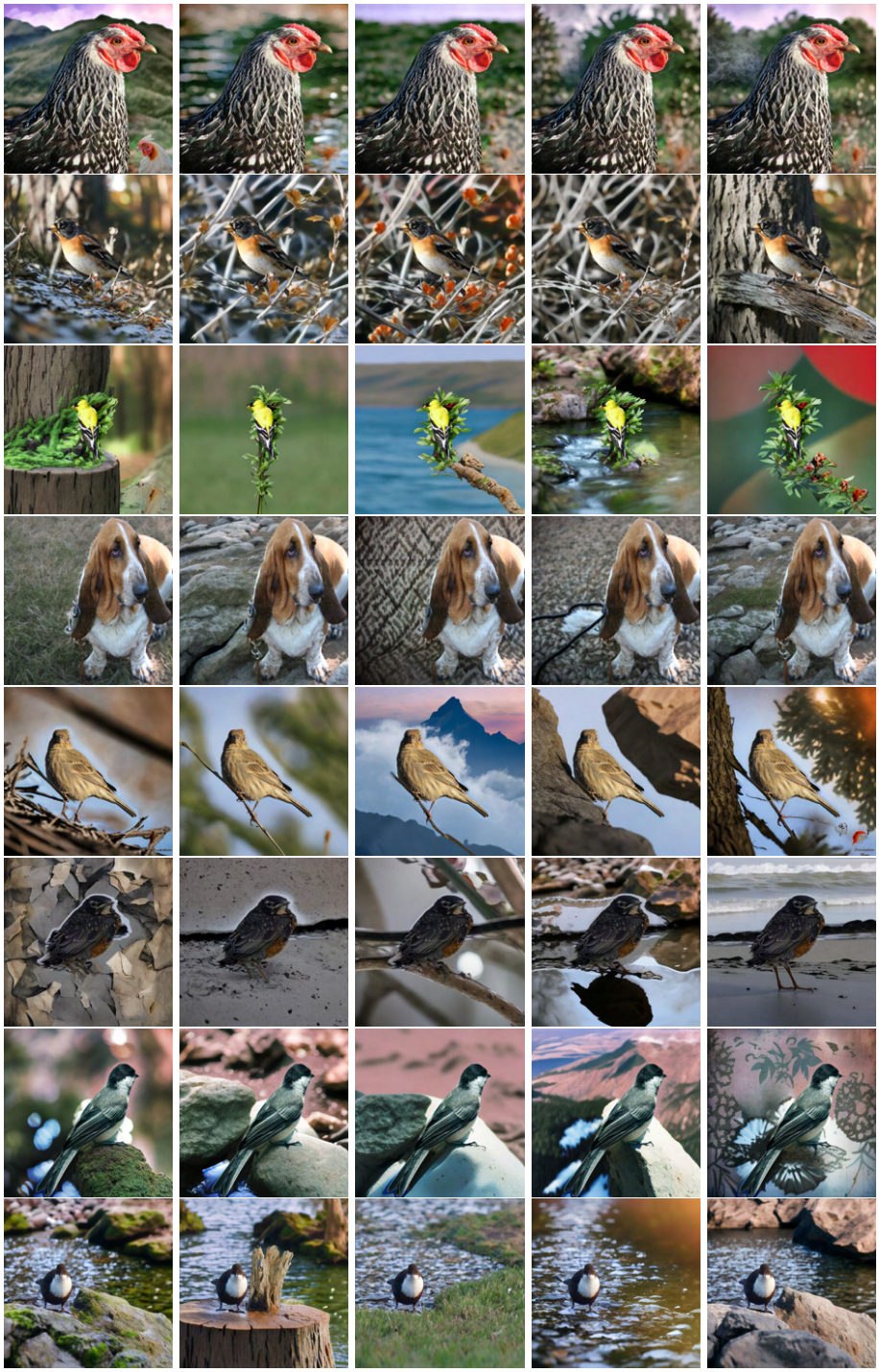

Fig. 23: Using diverse prompts to capture for diverse background shifts on samples from `ImageNet-B`.

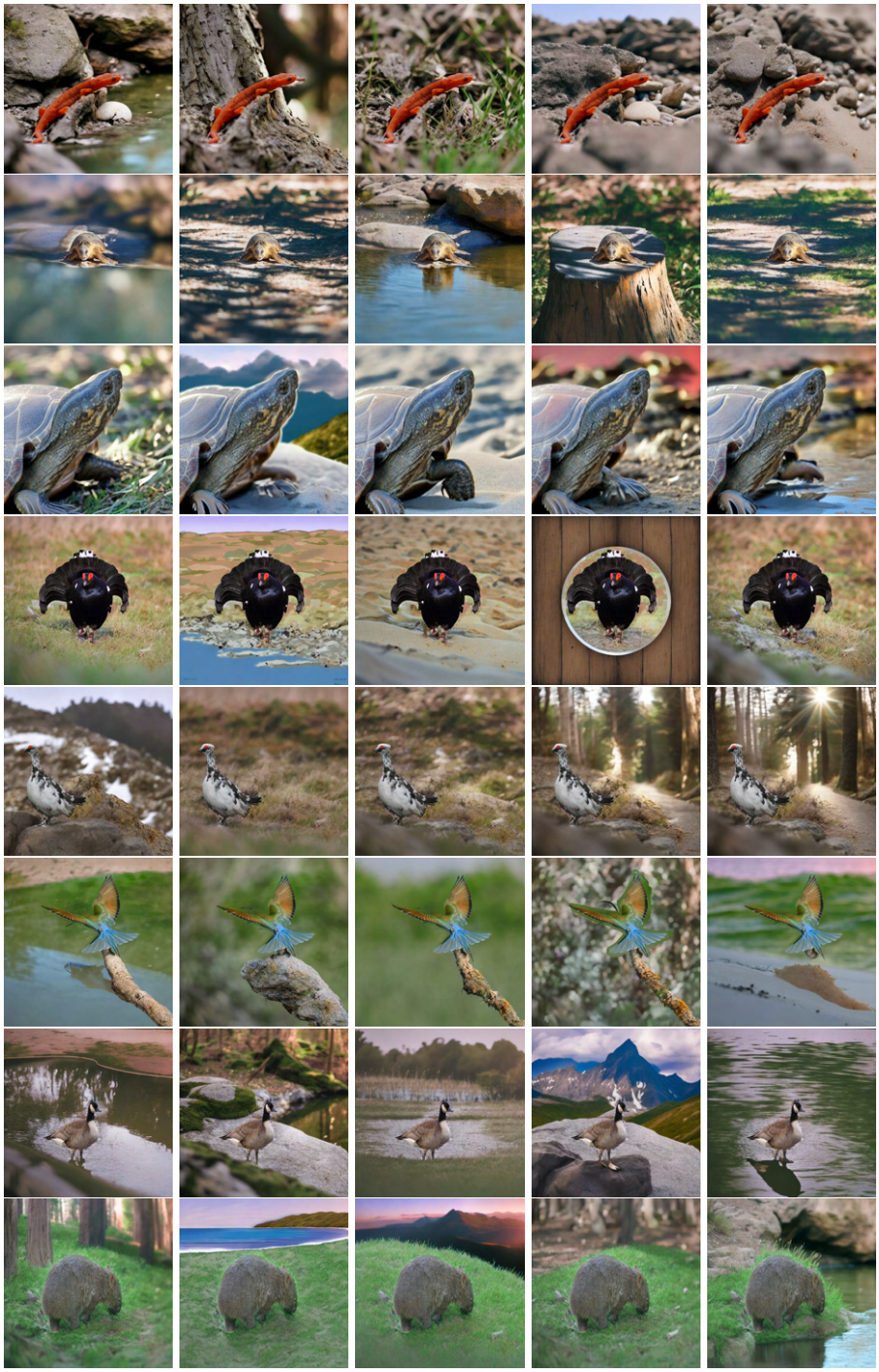

Fig. 24: Using diverse prompts to capture for diverse background shifts on samples from `ImageNet-B`.

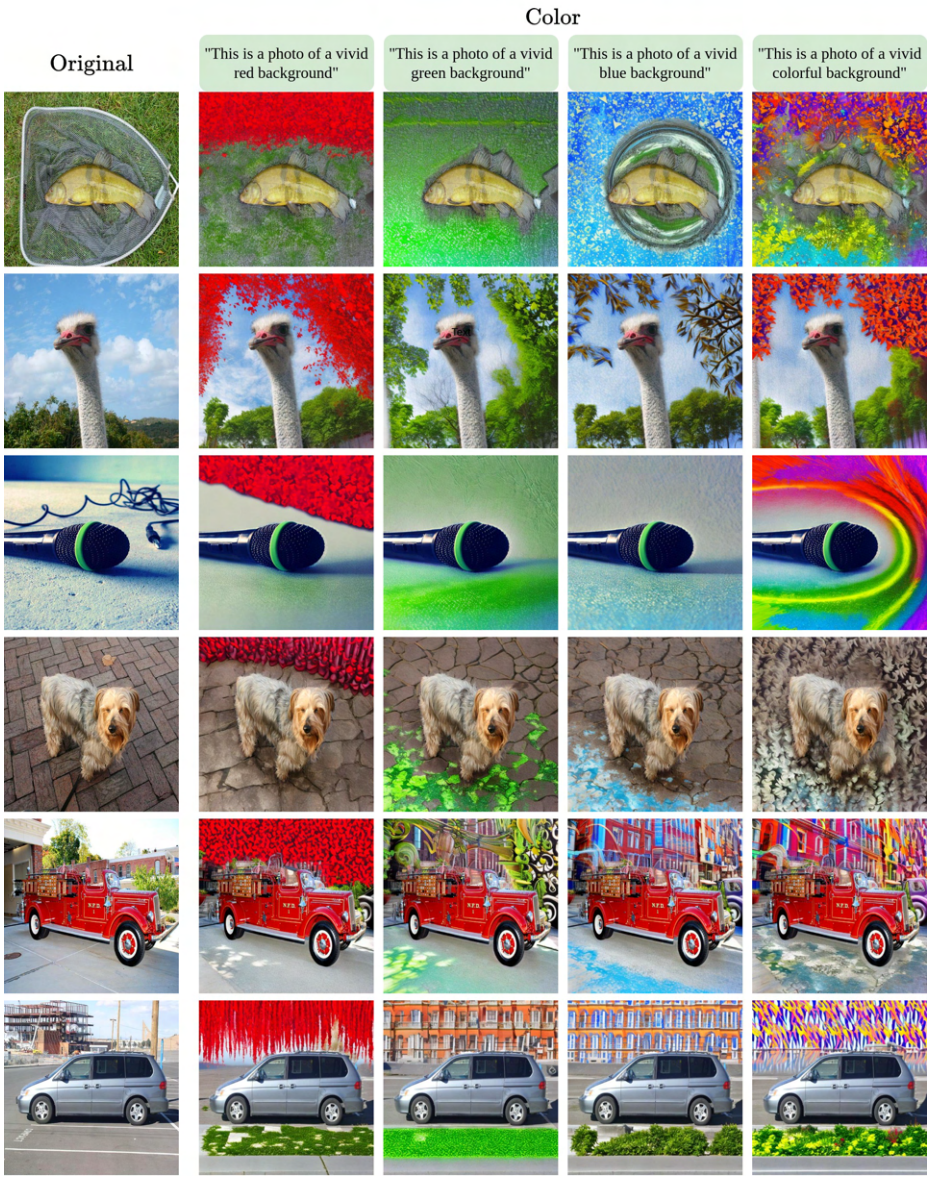

Fig. 25: Images generated through diverse color prompts on `ImageNet-B`.

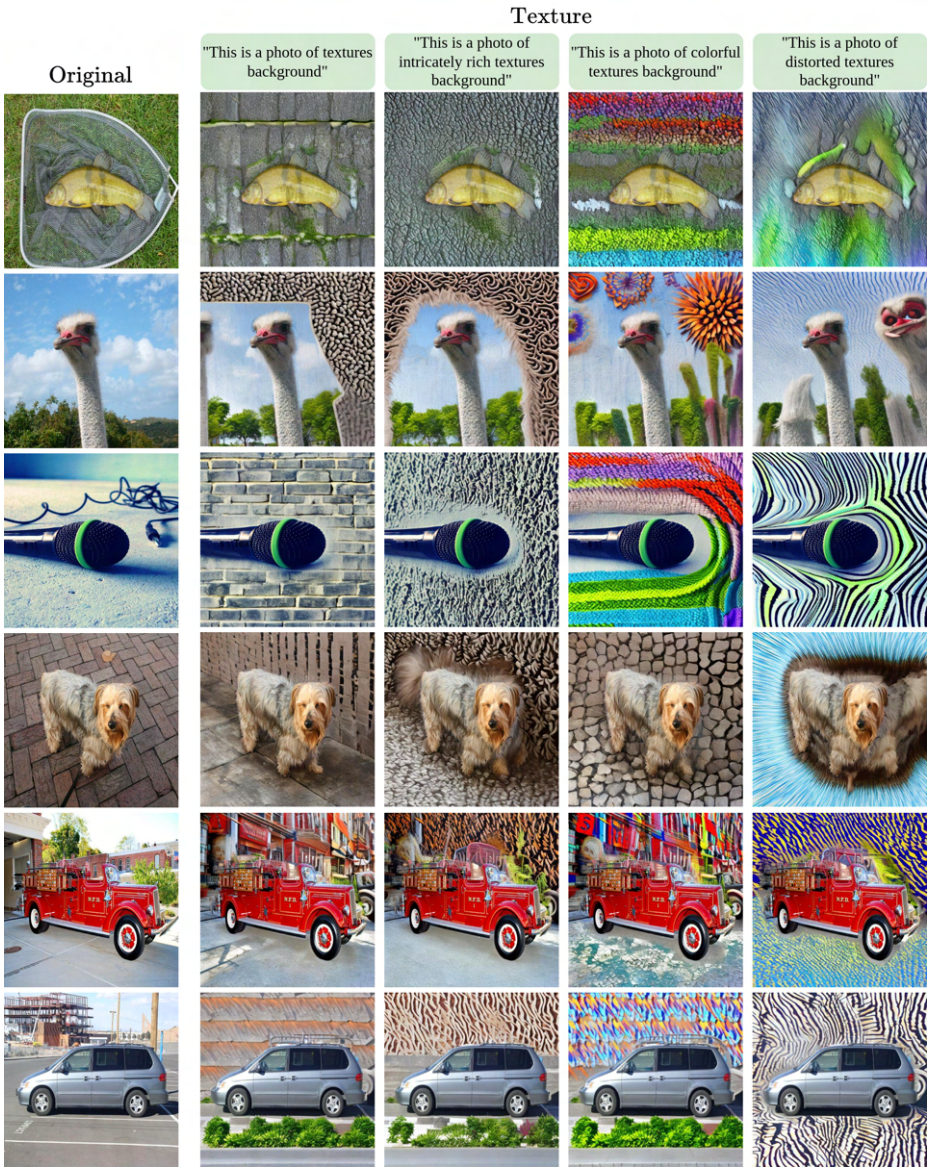

Fig. 26: Images generated through diverse texture prompts on `ImageNet-B`.

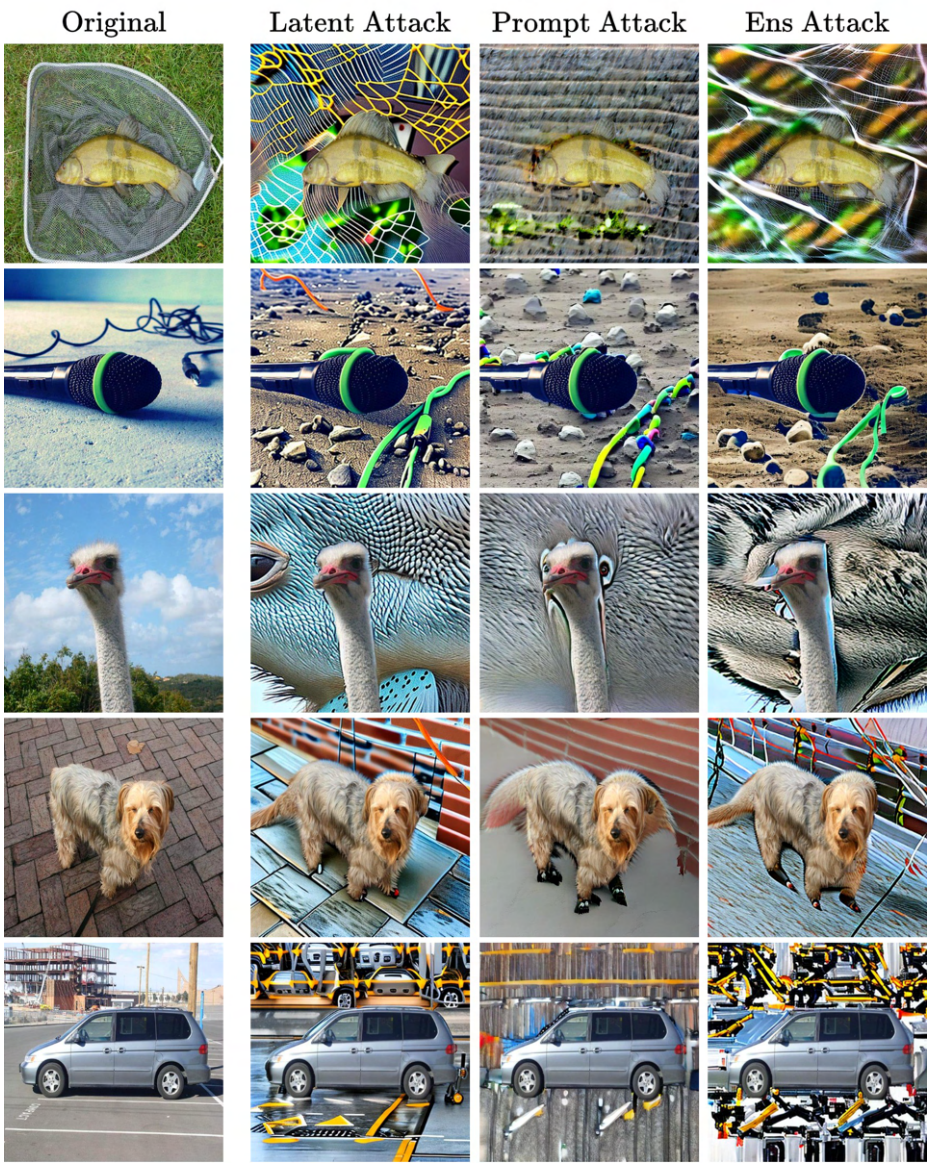

Fig. 27: Images generated under various attack scenarios on $\texttt{ImageNet-B}_{1000}$. Here we show the visualization for latent, prompt, and ensemble attack that are generated by optimizing latent, text prompt embeddings, and both respectively.

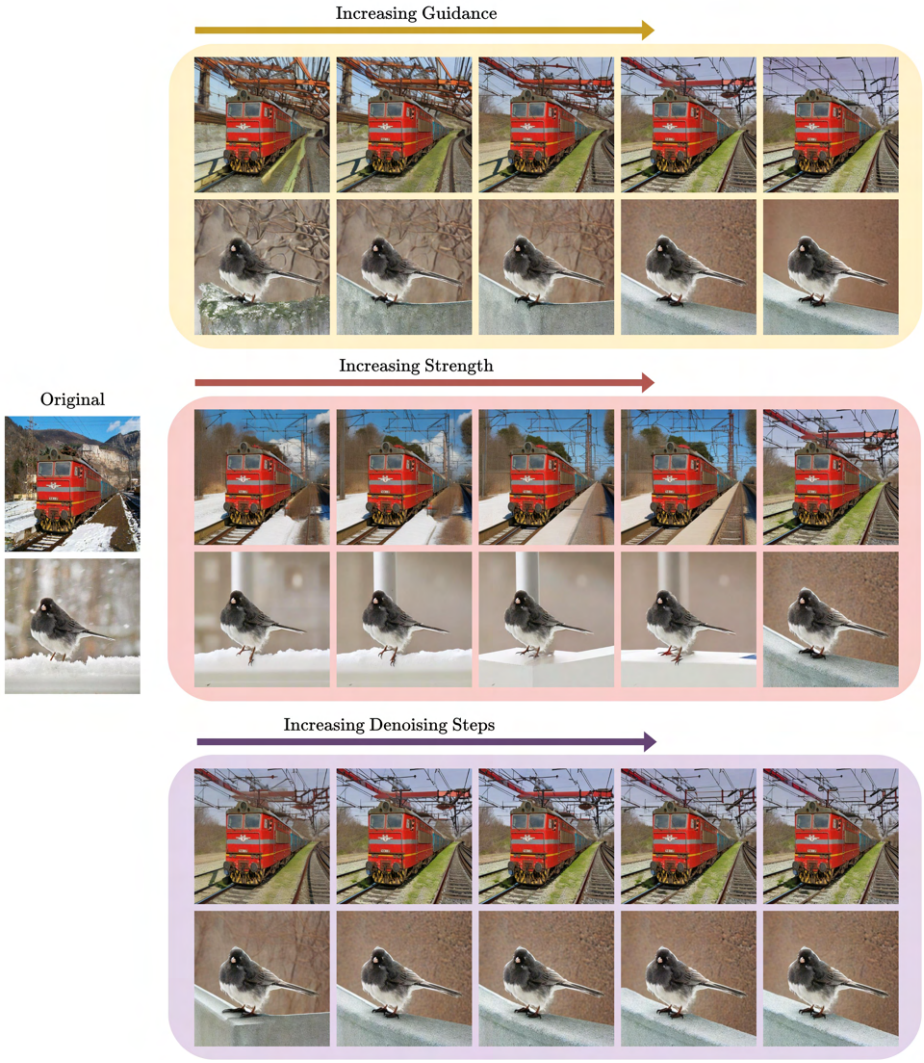

Fig. 28: Visualization on samples taken from `ImageNet-B`. Varying parameters like guidance, strength, and denoising steps while using BLIP-2 caption as the prompt. Increasing guidance leads to more fine-detailed background changes. Additionally, greater strength correlated with more pronounced alterations from the original background. And, augmenting diffusion steps improves image quality.

A.14 Misclassified Samples

We observe that there exist images which get misclassified *(by ResNet-50)* across several background alterations as can be seen from from Figure 29. In Figure 30 we show examples on which the highly robust *EVA-CLIP* ViT-E/14+ model fails to classify the correct class. After going through the misclassified samples, we visualize some of the *hard* examples in Figure 32. Furthermore, we also provide visualisation of images misclassified with adversarial background changes in Figure 31.

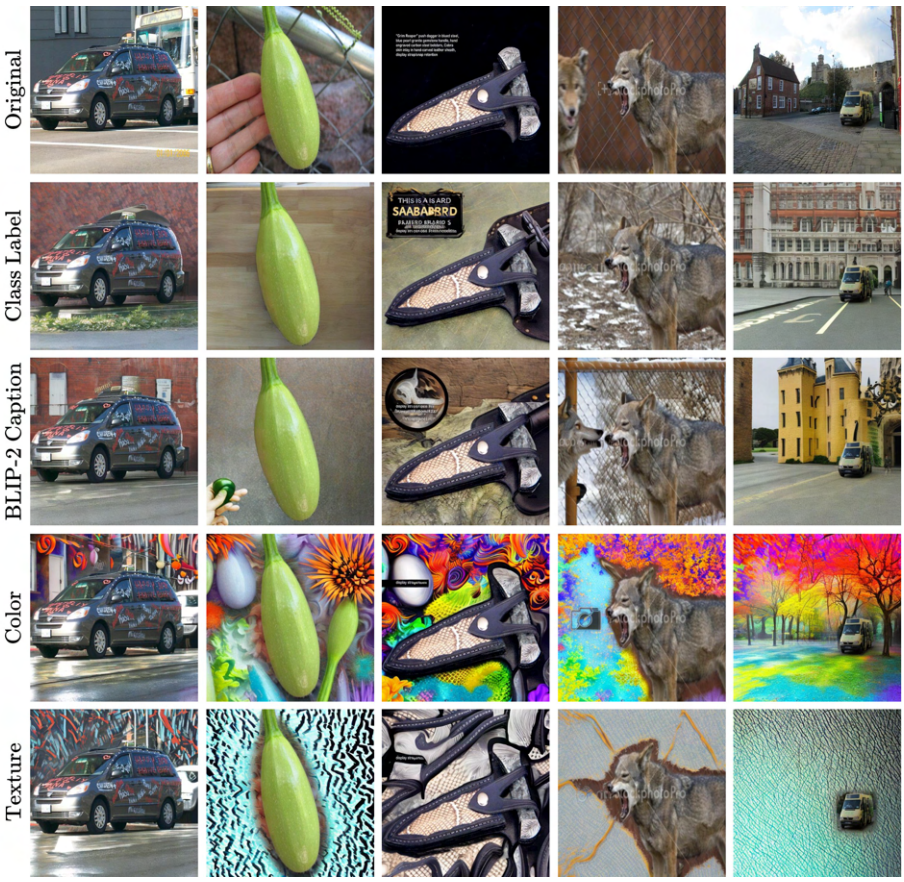

Fig. 29: Images misclassified by Res-50 across different background changes

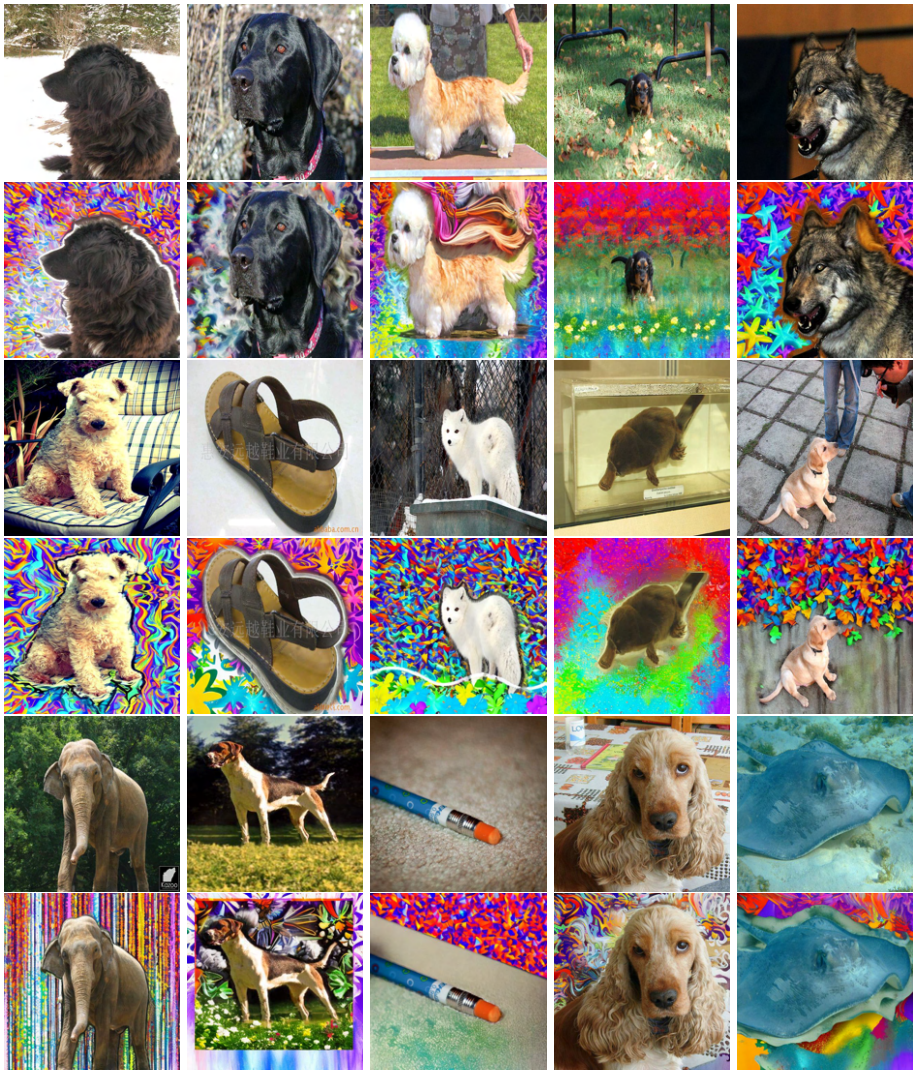

Fig. 30: Visual illustration of misclassified samples on color background and corresponding clean image samples. In two adjacent rows, *first row* represent the clean images and the *second row* represent the corresponding colorful background images

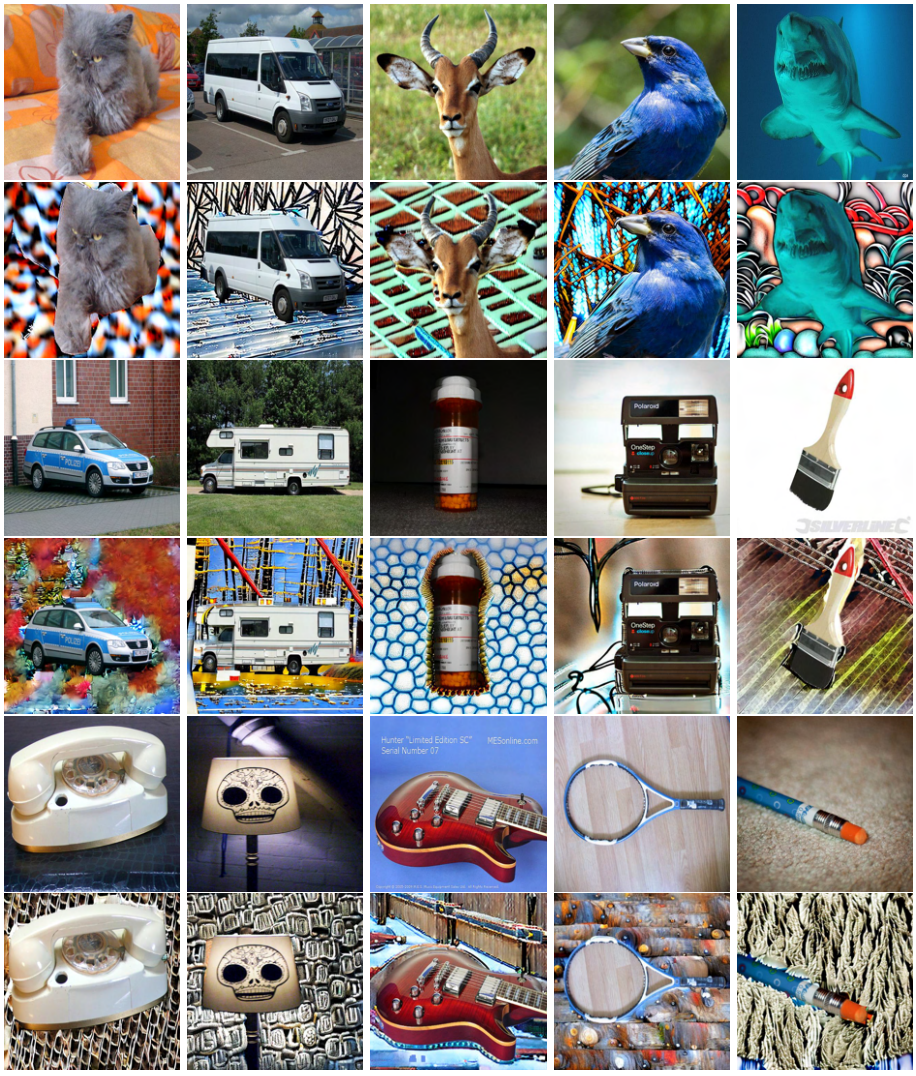

Fig. 31: Visual illustration of misclassified samples on adversarial background and corresponding clean image samples. In two adjacent rows, *first row* represents the clean images and the *second row* represents the corresponding adversarial images

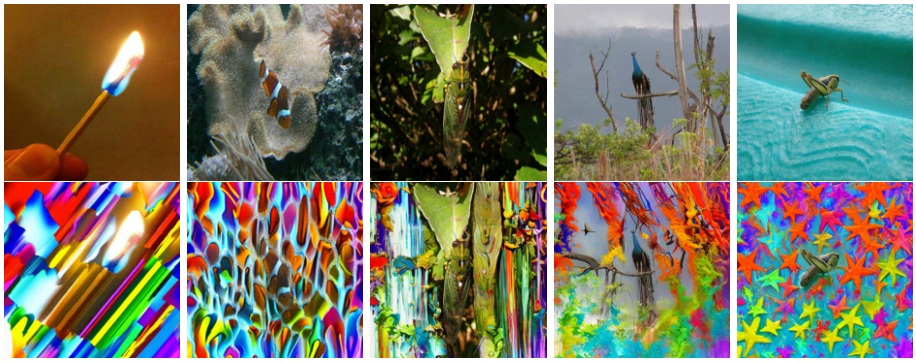

Fig. 32: Visual illustration of *hard* samples on color background

A.15 Potential External Factors

When composing object-to-background change with texture, color, or adversarial patterns, the target models can perceive those as some other class if that pattern or composition is dominant in that class during the training of the models. We discuss the potential external factors and how our proposed method minimizes the effect of those external factors fo object-to-background compositional changes.

Preserving Object Semantics: We preserve object semantics by using strong visual guidance via SAM for precise object delineation.

Possibility of extra objects in the Background: Kindly note that *a)* we use a pretrained diffusion model that is conditioned on a pretrained CLIP text encoder, this means that the generated output follows the latent space of the CLIP text encoder which is aligned with CLIP visual encoder. Therefore, we can measure the faithfulness of the generated sample w.r.t the textual prompt used to generate it. We can measure this by encoding the generated output and its corresponding text prompt within CLIP latent space. For a given sample, CLIP or EVA-CLIP performs zero-shot evaluation by measuring the similarity between embedding of class templates (e.g. 1000 templates of ImageNet class) with a given image. Thus, if we add the template for a textual prompt used to generate the object-to-background changes, then we can measure its alignment with the background changes. For instance, instead of using a "a photo of a fish" template for zero-shot classification, we add the relevant template that is with background change, such as "a photo of a fish in the vivid colorful background". In other words, the relevant template represents the object and background change we introduced. We validate this observation on the EVA-CLIP ViT-E/14+, a highly robust model. Using the class templates such as "a photo of a ", the model achieves 95.84% accuracy on the original images (`ImageNet-B` dataset), which decreases to 88.33% when our color background changes are applied (see Table 10 in Appendix A.5). However, when using the

relevant template, the performance improves to 92.95%, significantly reducing the gap between the performance on the original and color background changes from 7.51% to 2.89%. These results show that accuracy loss from background changes isn't due to unwanted background objects of other classes. Furthermore, we manually assess 2.89% of misclassified samples (very few samples, see Figure 30 and 32 in Appendix A.14). These can be considered the hardest examples in our dataset. We observe that even in such hard cases the model's confusion often stemmed from the complex background patterns instead of the addition of unwanted objects. We observe a similar trend in the case of adversarial patterns as well (see Figure 31 in Appendix A.14). b) Another empirical evidence of how our generated output closely follows the given textual prompts can be observed with BLIP-2 Caption of the original image. In this case, object-background change has similar results as compared to original images across different vision models (Table 2 in the main paper).

Extension of Objects: As already detailed in Appendix A.18, we encountered challenges when dealing with objects that occupy a small region in the image, sometimes leading to certain unwanted extensions to objects. To mitigate this, we filtered our dataset to focus on images where the object covers a significant area. Additionally, we slightly expand object masks computed using SAM to better define boundaries and prevent object shape distortion in the background.

The design choices discussed above, such as strong visual guidance and class-agnostic textual guidance, contribute to the well-calibrated results of our study. This indicates that our results using the conventional metrics such as classification accuracy are well calibrated as well in the context of our high quality of generated data as mentioned above. We note that these choices ensure that the models are primarily challenged by diverse changes in the background, rather than being misled by the presence of unwanted objects. This careful approach underlines the reliability of our findings and highlights the specific factors influencing model performance.

A.16 Dataset Distribution and Comparison

`ImageNet-B` dataset comprises a wide variety of objects belonging to different classes, as illustrated in Figure 33. Our dataset maintains a clear distinction between the background and objects, achieved through a rigorous filtering process applied to the ImageNet validation dataset. Additionally, we provide the list of prompts in Table 6 utilized for the experiments.

As shown in Tab. 20, our curated **ImageNet-B** dataset is the largest in terms of both the number of images *and* classes compared to closely related works [34, 44, 61]. In contrast to [34, 44, 61], we extend our analysis to object detection by introducing the **COCO-DC** dataset. Our proposed background changes on **ImageNet-B** & **COCO-DC**, enable us to evaluate on more than 70k samples for classification & 5k samples for detection. Our automated framework of delineating between foreground & background facilitates future dataset expansion.

Fig. 33: Our `ImageNet-B` dataset encompasses a diverse variety of images spanning 582 distinct classes. In this illustration, we showcase images distribution among all the classes. The figure is plotted in decreasing order of images present in each class.

Table 20: Dataset Comparison

	Dataset	#Classes	#Images	Classification	Detection
Baseline	LANCE(NeurIPS 2023) [44]	15	750	✓	×
	ImageNet-E(CVPR 2023) [34]	373	47872	✓	×
	ImageNet-D(CVPR 2024) [61]	113	4835	✓	×
Ours	`ImageNet-B`	582	77070	✓	×
	`COCO-DC`	66	5635	✓	✓

A.17 Evaluation on Background/Foreground Images

In this section, we systematically evaluate vision-based models by focusing on background and foreground elements in images. This evaluation involves masking the background of the original image, allowing us to assess the model's performance in recognizing and classifying the foreground without any cues from the background context. Conversely, we also mask the object or foreground from the image. This step is crucial to understand to what extent the models rely on background information for classifying the image into a specific class. This dual approach provides a comprehensive insight into the model's capabilities in image classification, highlighting its reliance on foreground and background elements.

Table 21: Evaluation of Zero-shot CLIP Models on `ImageNet-B` dataset while masking the object or the background of the image. Top-1(%) accuracy is reported. The accuracy drop is observe when we remove the object clues from the background such as in texture or color background

Background	Foreground							
	Res50	Res101	Res50x4	Res50x16	ViT-B/32	ViT-B/16	ViT-B/14	Average
Original	54.76	58.89	64.86	70.80	59.47	69.42	79.12	65.33
	Background							
Original	15.84	17.74	18.47	20.67	17.72	21.28	28.99	20.10
Class label	27.17	29.08	33.02	35.93	31.35	38.74	46.88	34.59
BLIP-2 Caption	19.05	21.39	23.37	24.57	22.39	27.21	34.42	24.62
Color	3.92	5.46	5.64	6.53	5.64	6.95	10.28	6.34
Texture	3.65	5.12	5.12	5.84	5.43	6.68	10.04	5.98

Table 22: DINOv2 model evaluation by masking either the object or the background within the `ImageNet-B` dataset. The integration of the additional token in the DINOv2 model proves beneficial, contributing to enhanced accuracy. However, our observations reveal that these models remain susceptible to background cues, particularly evident in class labels and the BLIP-2 Caption dataset. Interestingly, as we transition towards more generic texture or color backgrounds, a discernible drop in accuracy is observed.

Background	Foreground							
	ViT-S	ViT-B	ViT-L	Average	ViT-S_{reg}	ViT-B_{reg}	ViT-L_{reg}	Average
Original	88.73	93.86	94.89	92.49	96.34	89.95	97.25	94.51
	Background							
Original	27.72	37.78	51.44	38.98	30.10	42.08	55.18	42.45
Class label	42.70	54.73	66.68	54.70	46.88	58.81	68.97	58.22
BLIP-2 Caption	30.51	40.74	50.57	40.60	33.62	42.48	52.40	42.83
Color	2.96	5.03	8.39	5.46	3.68	5.75	9.50	6.31
Texture	2.83	4.92	7.88	5.21	3.45	5.57	9.28	6.10

A.18 Insights

Across Architectures. Our results (Tab. 1,2, 3, & 16) show that CNNs perform better than transformers across various background changes. We note that the mixing of background & foreground token features through the global attention mechanism may result in the reliance of transformer models prediction on outlier/background tokens. This is validated when we evaluate transformers which are trained to prioritize learning more salient features [9] (Tab. 17 & 22), resulting in improved performance under background changes.

Across Training Methods. Analysis of adversarially trained models (Fig. 6 & Tab. 13, 14, 15) reveal their robustness is confined to adversarial background changes, leaving them vulnerable to other background variations. Similar behavior is observed for models trained on stylized ImageNet dataset(Tab. 3). However, self-supervised training of uni-modal models on large extensively curated datasets shows performance gain across background changes(Fig. 7). Similarly, for multi-modal models, we observe that stabilizing training on large-scale

datasets (as in EVA-CLIP) leads to significant improvement in zero-shot performance across all background changes(Tab. 2,10, & 11).
Across Vision Tasks. The absence of object-to-background context during classification model training creates a significant vulnerability to background changes. In contrast, object detection & segmentation models (Tab. 5, Fig. 8 & 14), which explicitly incorporate object-to-background context during training, show notably better resilience to background variations. Based on the above insights, we discuss current limitations and future directions next.

Limitations. In Figure 34, we observe that for objects covering a small region in the image, relying solely on the class name to guide the diffusion model can result in alterations of the object shape, expanding the influence of the class name semantics to larger image regions. However, by supplementing with descriptive captions that encompass the object-to-background context, we partially mitigate this effect. Furthermore, the generated textured background can inadvertently camouflage the object. To address this concern, we slightly expand the object mask to clearly delineate the object boundaries.

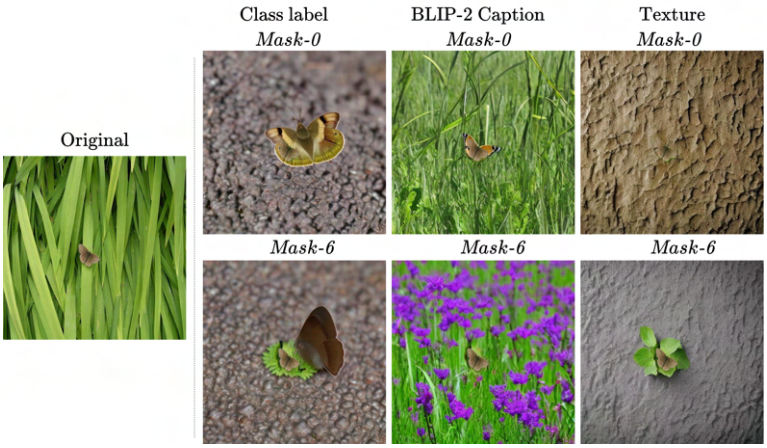

Fig. 34: Limitation: Background changes on small objects in the scene. Enlarging the mask (here by 6 pixels) helps in mitigating the issue to some effect.

Future Directions. Our current work represents one of the preliminary efforts in utilizing diffusion models to study the object-to-background context in vision-based systems. Based on our observations and analysis, the following are the interesting future directions.

– Since large capacity models in general show better robustness to object-to-background compositions, coming up with new approaches to effectively

distill knowledge from these large models could improve how small models cope with background changes. This can improve resilience in small models that can be deployed in edge devices.
- Another direction is to set up object-to-background priors during adversarial training to expand robustness beyond just adversarial changes. To some extent, successful examples are recent works [9,53] where models are trained to discern the salient features in the image foreground. This leads to better robustness.
- Our work can be extended to videos where preserving the semantics of the objects across the frames while introducing changes to the background temporally will help understand the robustness of video models.
- Additionally, the capabilities of diffusion models can be explored to craft complex changes in the object of interest while preserving the semantic integrity. For instance, in [59], diffusion models are employed to generate multiple viewpoints of the same object. Additionally, in [28], non-rigid motions of objects are created while preserving their semantics. By incorporating these with our approach, we can study how vision models maintain semantic consistency in dynamic scenarios.

A.19 Calibration Metrics

Model calibration refers to how well a model's predicted confidence levels correspond to its actual accuracy. Confidence represents the probability a model assigns to its predictions, while accuracy measures how often those predictions are correct. For example, if a model predicts with 70% confidence, a well-calibrated model should have an actual accuracy close to 70%. To quantify this, we use the Expected Calibration Error (ECE). ECE works by dividing the predictions into M bins based on confidence intervals (e.g., 60%-70%, 70%-80%). Within each bin, the average confidence and accuracy are calculated, and the ECE is the weighted average of the differences between these values. To visually evaluate model calibration, reliability diagrams are used. These diagrams plot predicted confidence against actual accuracy, allowing us to compare different models. A well-calibrated model will show points that lie close to the diagonal on these plots. In Figures 35 and 36, we plot the reliability diagrams for different convolutional and transformer-based models, respectively.

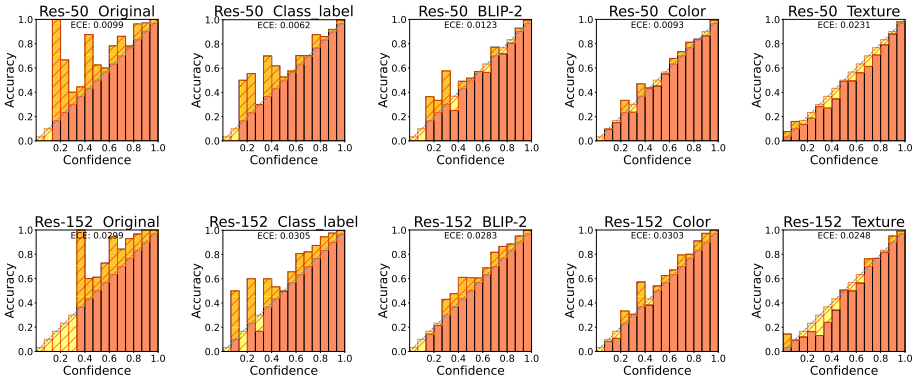

Fig. 35: Calibration results comparison of CNN-based models.

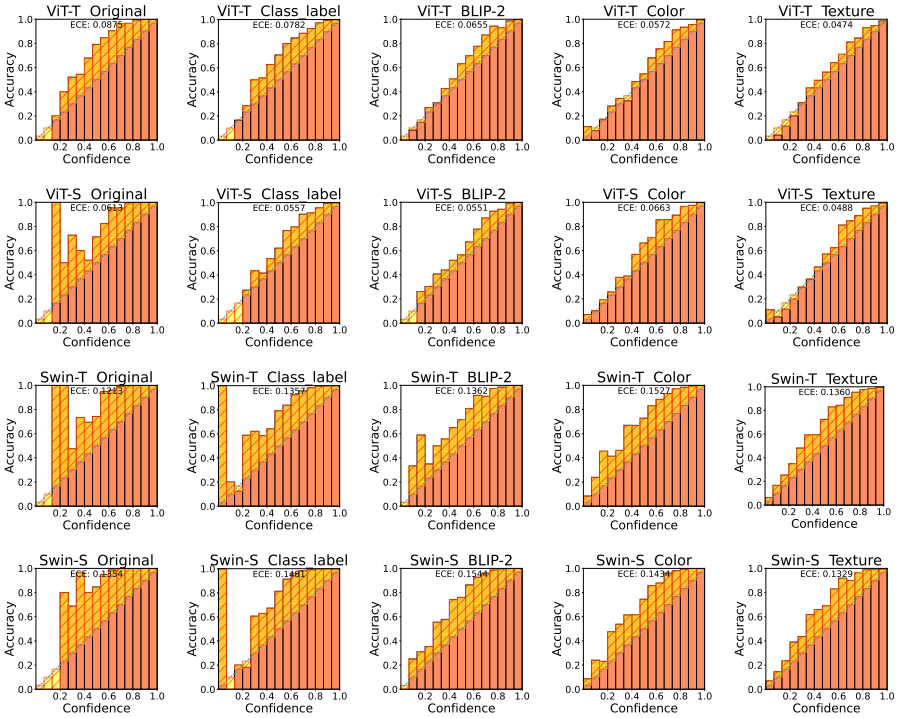

Fig. 36: Caliberation results comparison of ViT model

A.20 Ablation on Adversarial Loss

Tab.23 presents an ablation study on adversarial loss, showing optimizing both vision & text based emebedding, leads to the most performance drop.

Table 23: Ablation on different losses

	ViT-T	ViT-S	Swin-T	Swin-S	Res-50	Res-152	Dense-161	Average
Text	28.4	46.5	40.9	47.2	19.0	46.5	31.5	**37.1**
Latent	32.3	47.5	42.8	50.2	1.6	44.5	26.4	**35.1**
Combined	18.4	32.1	25.0	31.7	2.0	28	14.4	**21.6**

A.21 Reproducibility and Ethics Statement

Reproducibility Statement: Our method uses already available pre-trained models and the codebase is based on several open source implementations. We highlight the main components used in our framework for reproducing the results presented in our paper, `a) Diffusion Inpainting Implementation:` We use the open-source implementation of Stable-Diffusion-Inpainting method (`https://github.com/huggingface/diffusers/blob/main/src/diffusers/`) with available pretrained weights *(Stable-Diffusion-v-1-2)* for background generation. `b) Image-to-Segment Implementation:` We use the official open-source implementation of FastSAM (`https://github.com/CASIA-IVA-Lab/FastSAM`) to get the segmentation masks of filtered ImageNet dataset. `c) Image-to-Text Implementation:` We use the official open-source implementation of BLIP-2(`https://github.com/salesforce/LAVIS/tree/main/projects/blip2`) to get the captions for each image. We will also provide captions for each image in our dataset. `d) Adversarial Attack:` We intent to open-source our codebase and release the script for crafting adversarial examples. `e) Dataset:` In the paper, we describe the procedure of filtering the images from ImageNet and COCO val. set. Furthermore, we will provide the filtered datasets, object masks as well as prompts used to generate different backgrounds.

Ethics Statement: Our work focuses on evaluating resilience of current vision and language models against natural and adversarial background changes in real images. This work can be utilized by an attacker to generate malicious backgrounds on real images as well as generate adversarial backgrounds which can fool the deployed computer-vision systems. Nevertheless, we believe that our research will pave the way for improved evaluation protocols to assess the resilience of existing models. This, in turn, is likely to drive the development of enhanced techniques for bolstering the resilience of deployed systems. Since we are benchmarking vision and vision-language models using a subset of images from publicly available ImageNet and COCO datasets, it's relevant to mention that these datasets are known to have images of people which poses a privacy risk and further it is known to have biases which can encourage social stereotypes.

In the future, we intend to benchmark our models on a less biased dataset to mitigate these concerns and ensure a fair evaluation.