MaskDiME: Adaptive Masked Diffusion for Precise and Efficient Visual Counterfactual Explanations

Supplementary Material

A. Additional Dataset and Setup Details

A.1. More Dataset Details

Following the evaluation protocol of ACE [8], we evaluate the proposed MaskDiME on five visual datasets covering facial attributes, driving scenes, and natural image classification. Detailed dataset descriptions and experimental settings are provided below, and an overview is in Tab. 6.

CelebA [10]: CelebA is a large-scale facial attribute dataset containing over 200,000 aligned images annotated with 40 attributes. In this work, we focus on two binary attributes, *Smiling* and *Young*. Following ACE and DiME [7], we use the validation split (19,867 images) and generate one counterfactual per image for each attribute. All images are resized to 128×128. We adopt the unconditional DDPM [4] and the DenseNet121 [5] classifier provided by DiME, using their pretrained weights.

CelebA-HQ [9]: A high-quality extension of CelebA with 30,000 high-resolution face images. Following ACE, we use 2,824 test images resized to 256×256 and generate one counterfactual per image per attribute. We adopt the CelebA-HQ DDPM from ACE and a DenseNet121 [5] classifier with STEEX [6] pretrained weights.

BDD100K [14] / BDD-OIA [13]: BDD100K is a large-scale autonomous driving dataset, and BDD-OIA is a task-specific extension with additional object-induced action annotations. We use both datasets for the same binary driving decision task: *Forward* vs. *Slow Down*. Images are resized to 512×256. Following ACE, we use 10,000 and 2,259 validation images from BDD100K and BDD-OIA, respectively, and generate one counterfactual per image. Both datasets use the ACE driving-scene DDPM and a DenseNet121 classifier with STEEX pretrained weights.

ImageNet [1]: Following ACE, we evaluate MaskDiME on more complex natural images using an ImageNet subset of 7,800 resized 256×256 images from three category pairs with strong visual contrasts: *Sorrel* vs. *Zebra*, *Persian* vs. *Egyptian Cat*, and *Cougar* vs. *Cheetah*. These pairs exhibit pronounced differences in texture, color, and semantics, providing a more challenging setting for evaluating counterfactual explanation. We adopt the ImageNet pretrained diffusion model from Guided Diffusion [2] and a ResNet50 [3] classifier with PyTorch pretrained weights.

A.2. Additional Implementation Details

Across all datasets, we adopt a unified masking strategy with only a small set of dataset-specific hyperparameters. For facial datasets (CelebA/CelebA-HQ), the noisy-level top-k is determined by the typical spatial extent of each attribute: 5% for *Smiling* and 10% for *Age*. For the clean-level mask top-ρk, CelebA uses $\rho = 0.5$, while CelebA-HQ adopts a smaller $\rho = 0.25$, with gradient scales of 8 and 10, respectively. We use a smaller ρ and larger gradient scale for CelebA-HQ as attribute editing is more challenging; a smaller ρ localizes updates and yields more informative gradients, while a larger scale amplifies them. For non-facial datasets (BDD100K, BDD-OIA, and ImageNet), we fix $k = 10\%$ and $\rho = 0.5$, and adjust the gradient scale s to account for differences in gradient magnitudes ($s = 14$ for BDD100K/BDD-OIA and $s = 6.5$ for ImageNet). The batch size is 50 for CelebA and 25 for all other datasets.

We further validate these parameter choices through additional ablation studies on CelebA and CelebA-HQ, as shown in Tab. 5. The results indicate that a larger k leads to worse FID but higher COUT. Across both datasets, $k = 0.05$ achieves a good balance between COUT and FID. We also verify the necessity of the shrinkage parameter ρ. When $\rho = 1$, the same mask is applied to both the noisy and clean images. Although this setting achieves comparable FID, it results in significantly lower COUT, indicating less precise and less confident attribute manipulation.

Table 5. Additional ablation study of MaskDiME on *smile* of CelebA ($\rho = 0.5$) and CelebA-HQ ($\rho = 0.25$).

Setting	CelebA		CelebA-HQ	
	FID↓	COUT↑	FID↓	COUT↑
$k = 0.025$	0.63	0.67	1.95	0.59
$k = 0.1$	0.82	0.93	4.54	0.87
$k = 0.05$, $\rho = 1$	0.71	0.81	2.36	0.60
MaskDiME ($k = 0.05$)	0.71	0.87	2.51	0.69

B. Quantitative efficiency

In Tab. 7 we report the computational efficiency of DiME [7], FastDiME [12], ACE [8], RCSB [11], and MaskDiME on the CelebA 1000-image subset, including their complexity, runtime, and peak GPU memory usage.

C. Additional qualitative results

In Figs. 7 to 11 we present additional qualitative comparisons with prior methods. For all datasets, we show the input image, the counterfactuals produced by different methods, and their corresponding difference maps. Note that we include a brief discussion of the results in the captions.

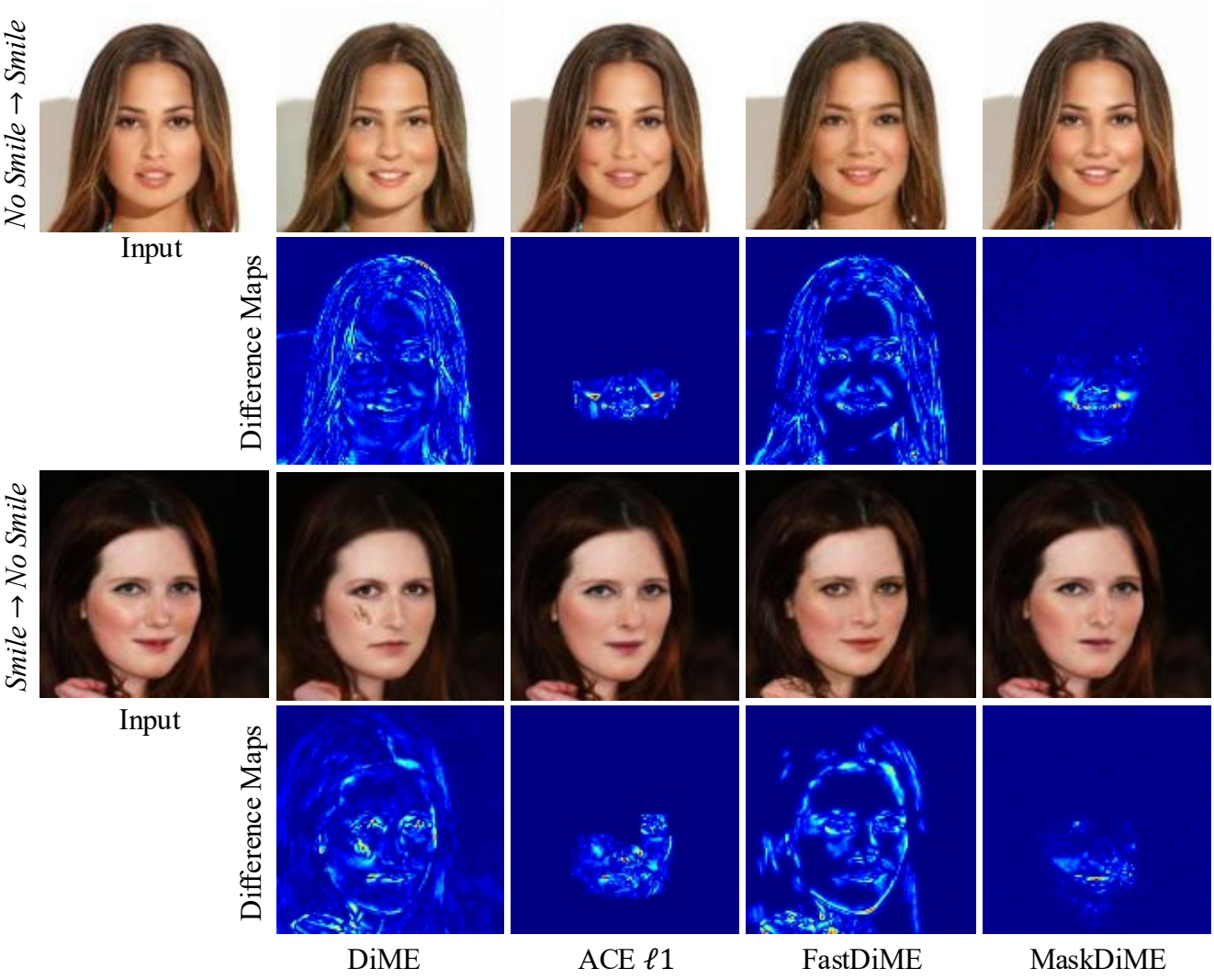

Figure 7. **Additional CelebA qualitative results for the *Smiling* attribute.** DiME introduces noticeable artifacts when removing the smile attribute, while FastDiME exhibits similar global modifications that may alter irrelevant regions. ACE ℓ_1 tends to produce concentrated high-contrast change areas that appear less natural. In contrast, MaskDiME achieves smooth and localized smile manipulations in both directions, resulting in more natural and visually coherent edits.

Table 6. Overview of datasets and pretrained models used for evaluating **MaskDiME**.

Datasets	**Domain**	**Resolution**	**#Images**	**Attributes / Classes**	**DDPM Weights**	**Classifier**	**Weights**
CelebA [10]	Faces	128×128	19,867	Smile, Age	DiME [7]	DenseNet121 [5]	DiME [7]
CelebA-HQ [9]	Faces	256×256	2,824	Smile, Age	ACE [8]	DenseNet121 [5]	STEEX [6]
BDD100K [14]	Autonomous driving	512×256	10,000	Forward / Slow Down	ACE [8]	DenseNet121 [5]	STEEX [6]
BDD-OIA [13]	Autonomous driving	512×256	2,259	Forward / Slow Down	ACE [8]	DenseNet121 [5]	STEEX [6]
ImageNet (subset) [1]	Image classification	256×256	7,800	Sorrel–Zebra, Persian–Egyptian cat, Cougar–Cheetah	Guided-diffusion [2]	ResNet50 [3]	PyTorch

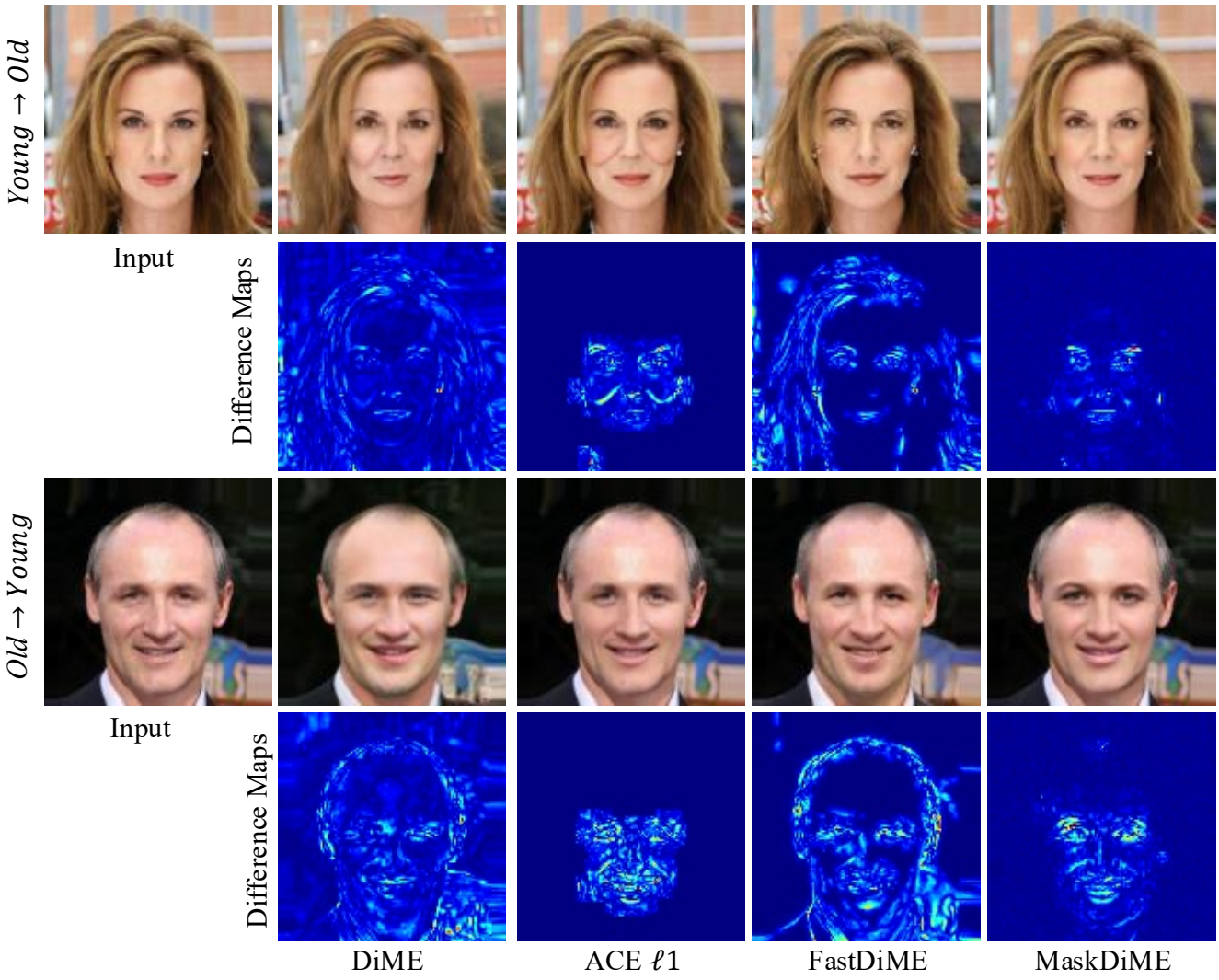

Figure 8. **Additional CelebA qualitative results for the *Age* attribute.** DiME and FastDiME tend to apply more global modifications, whereas ACE ℓ_1 and MaskDiME restrict edits to localized facial regions. Among region-constrained methods, MaskDiME produces more natural age transitions with fewer unintended changes, as revealed by the difference maps.

Table 7. Efficiency comparison in terms of total runtime (s) and peak GPU memory (MB). T denotes the number of diffusion steps, and N is the number of adversarial update steps in ACE. MaskDiME (w/o s) corresponds to fixing $s = 1$, while MaskDiME (w/o mask) refers to the variant without applying any masking. All results are obtained on the CelebA *Smiling* 1000-image subset with a batch size of 5. Peak GPU memory is measured using `torch.cuda.max_memory_allocated()` by recording the maximum value across all sampling steps.

Method	Complexity	Total Time	Peak GPU
DiME	$\mathcal{O}(T^2)$	35,034	**962**
ACE ℓ_1	$\mathcal{O}(NT)$	7,413	12,543
ACE ℓ_2	$\mathcal{O}(NT)$	7,375	12,543
FastDiME	$\mathcal{O}(T)$	*2,767*	963
FastDiME-2	$\mathcal{O}(T)$	4,094	963
FastDiME-2+	$\mathcal{O}(T)$	5,289	963
RCSB	$\mathcal{O}(T)$	5,381	11,147
MaskDiME (w/o s)	$\mathcal{O}(T)$	2,359	965
MaskDiME (w/o mask)	$\mathcal{O}(T)$	1,301	964
MaskDiME	$\mathcal{O}(T)$	**1,147**	965

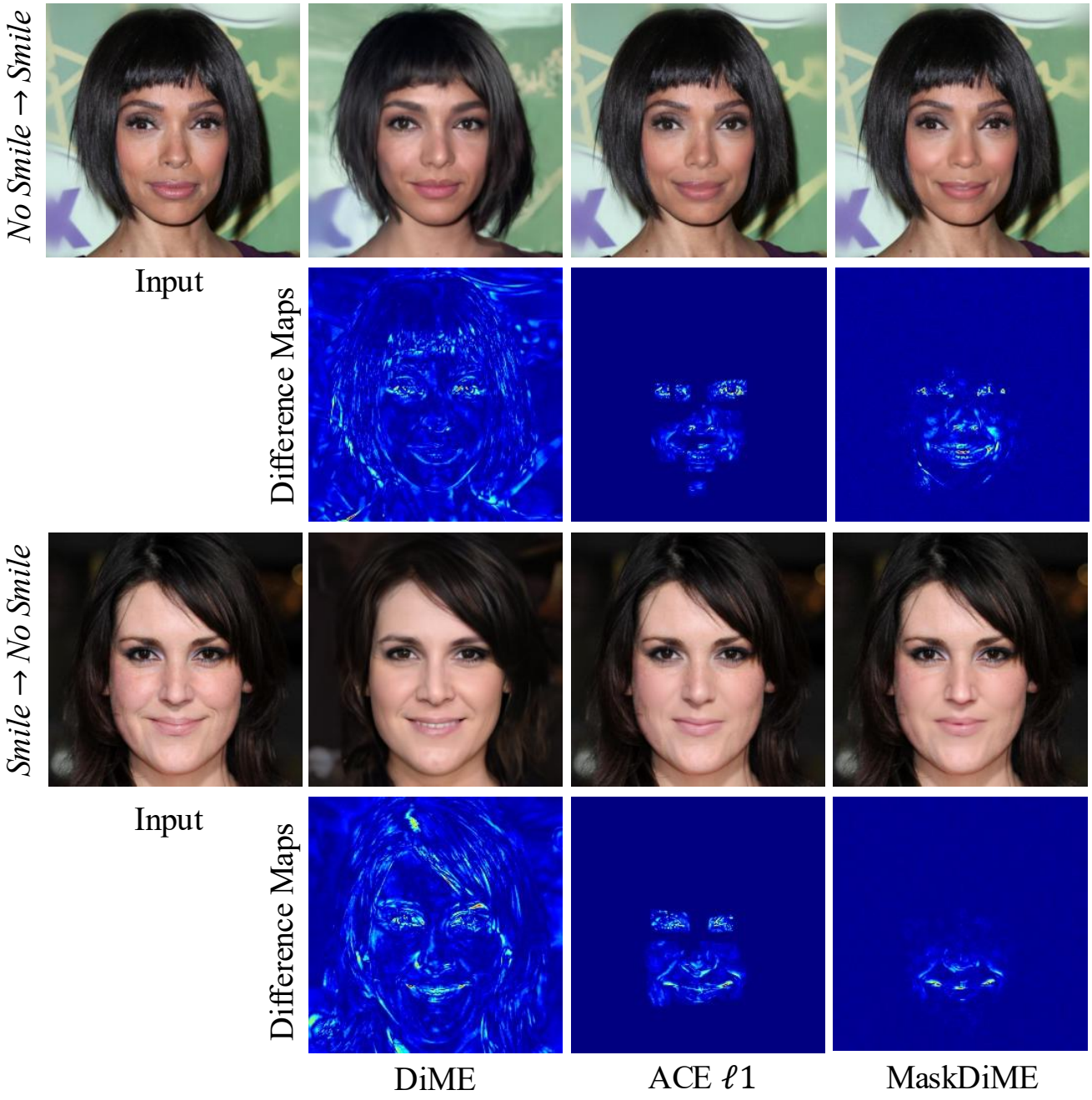

Figure 9. **Additional CelebA-HQ qualitative results for the *Smiling* attribute.** In the *Smile* → *No Smile* setting, DiME fails to produce the expected changes, does not achieve an effective counterfactual transformation, and additionally alters many irrelevant regions. Compared with ACE ℓ_1, MaskDiME preserves local edits while exhibiting clearer and more natural attribute changes.

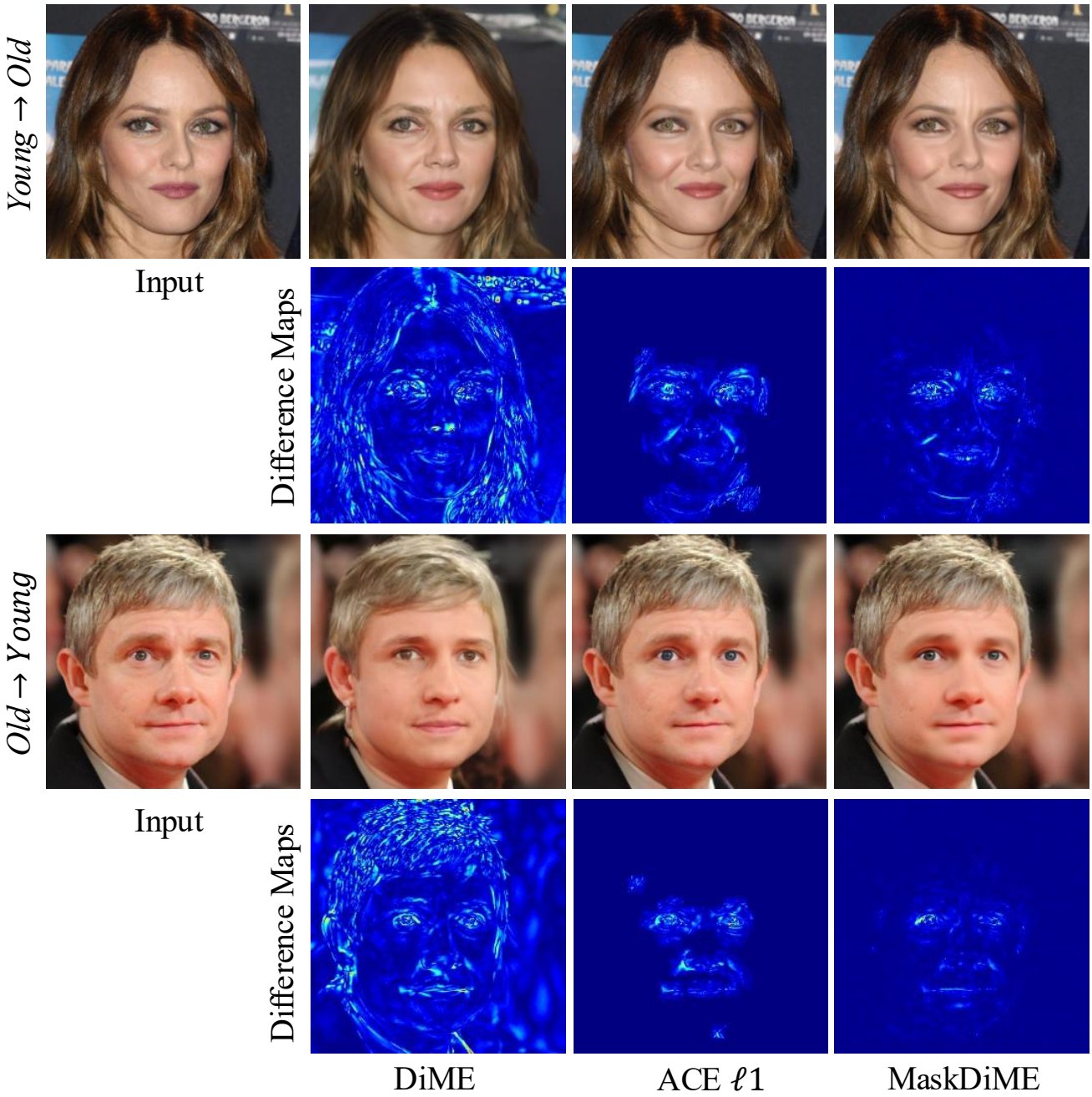

Figure 10. **Additional CelebA-HQ qualitative results for the *Age* attribute.** In the Age editing results, DiME successfully performs the counterfactual transformation but alters the entire face, resulting in a noticeable identity shift. Both ACE ℓ_1 and MaskDiME achieve localized edits; however, MaskDiME produces more natural and perceptually meaningful changes, such as clearer wrinkle formation on the forehead when aging and smoother facial appearance when rejuvenating.

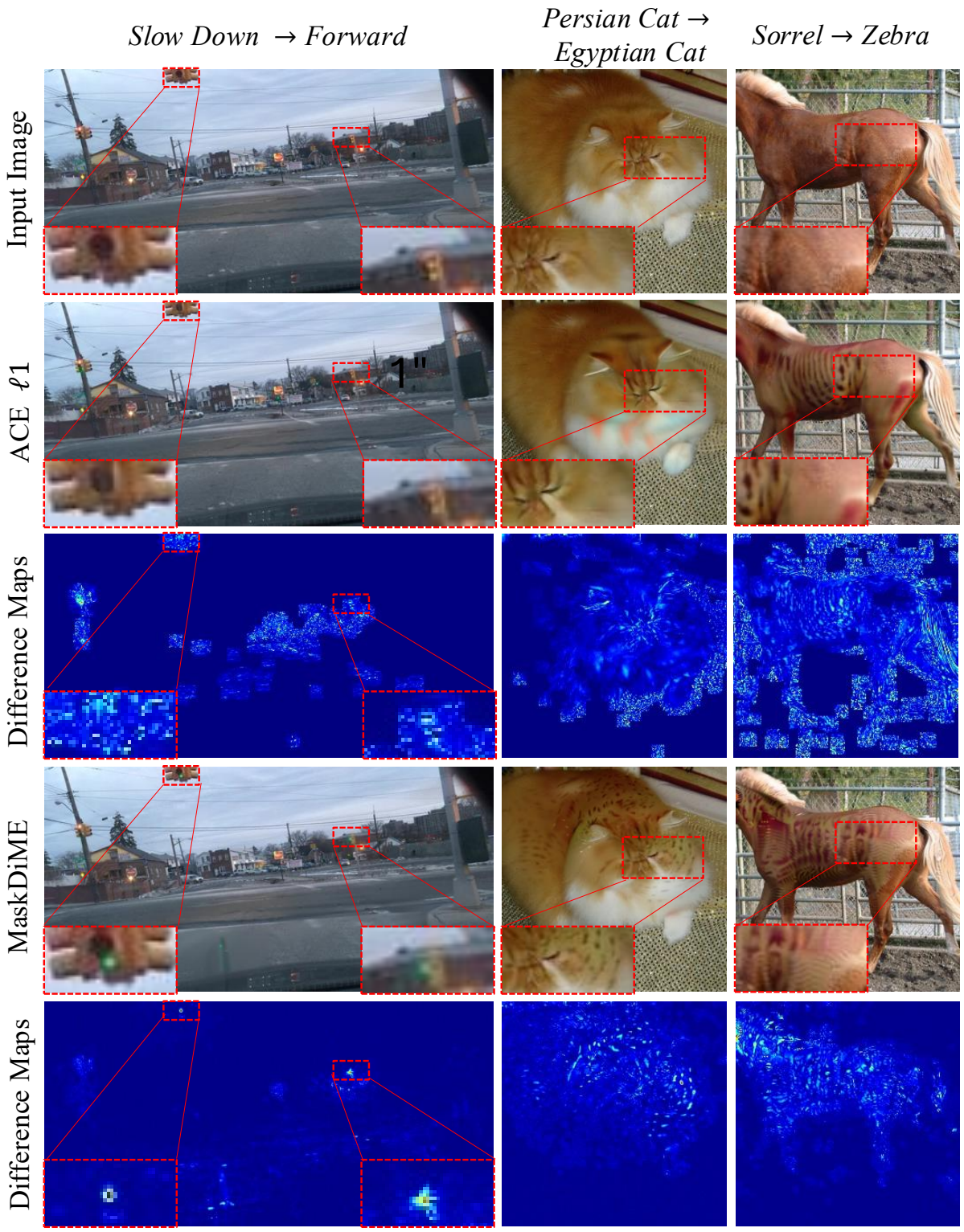

Figure 11. **Additional qualitative results on BDD and ImageNet.** Zoomed-in crops (red boxes) highlight the modified regions. On BDD, ACE ℓ_1 yields scattered and poorly localized changes and barely alters the semantic evidence for the target action, whereas MaskDiME focuses its edits around the traffic light and changes its color, producing a more interpretable counterfactual. A similar pattern is observed on ImageNet: ACE ℓ_1 modifies large parts of the object in an over-smoothed, less class-specific way, while MaskDiME applies compact, class-relevant changes (e.g., breed-specific fur texture and zebra stripes), leading to semantically meaningful counterfactual transformations.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1, 2

[6] Paul Jacob, Éloi Zablocki, Hedi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. Steex: steering counterfactual explanations with semantics. In *European Conference on Computer Vision*, pages 387–403. Springer, 2022. 1, 2

[7] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian conference on computer vision*, pages 858–876, 2022. 1, 2

[8] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16425–16435, 2023. 1, 2

[9] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5549–5558, 2020. 1, 2

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 1, 2

[11] Bartlomiej Sobieski, Jakub Grzywaczewski, Bartłomiej Sadlej, Matthew Tivnan, and Przemyslaw Biecek. Rethinking visual counterfactual explanations through region constraint. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[12] Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, pages 338–357. Springer, 2024. 1

[13] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9523–9532, 2020. 1, 2

[14] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 1, 2