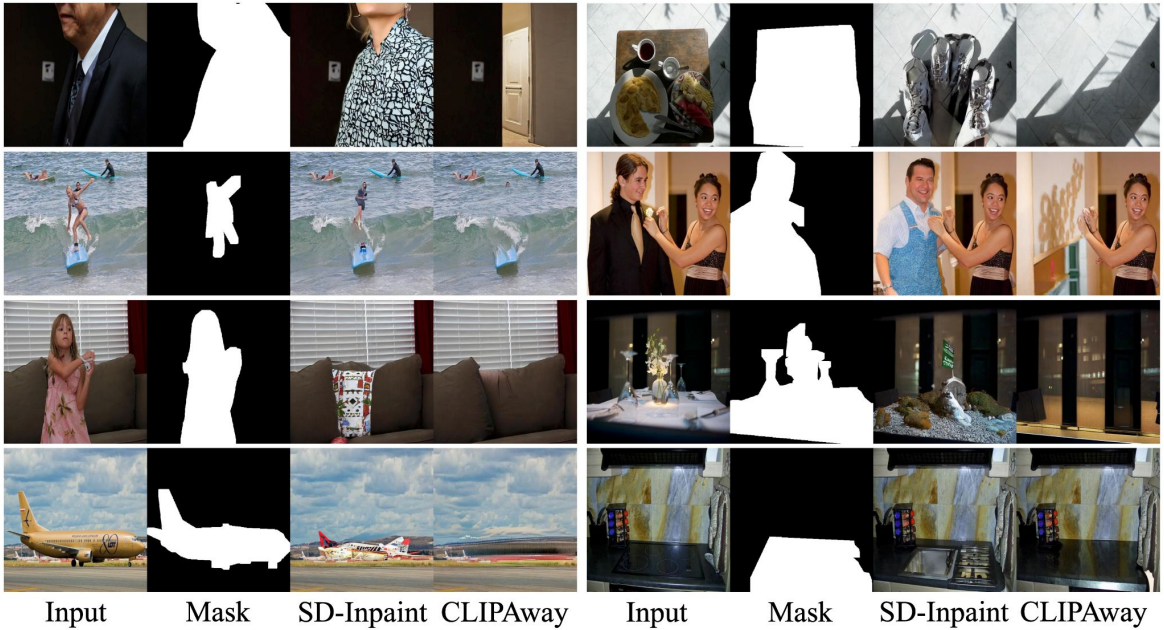
CLIPAway: Harmonizing Focused Embeddings for Removing Objects via Diffusion Models

Yiğit Ekin, Ahmet Burak Yildirim, Erdem Eren Caglar,
Aykut Erdem, Erkut Erdem, Aysegul Dundar

Problem Definition



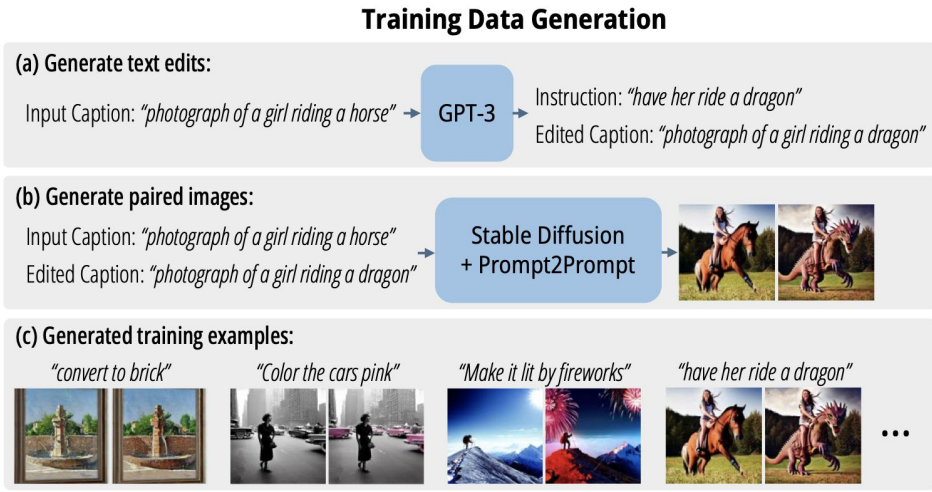
- Due to their training objective, SD-Inpaint [1] struggles with object removal
- Hallucinates rather than removing the object

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)

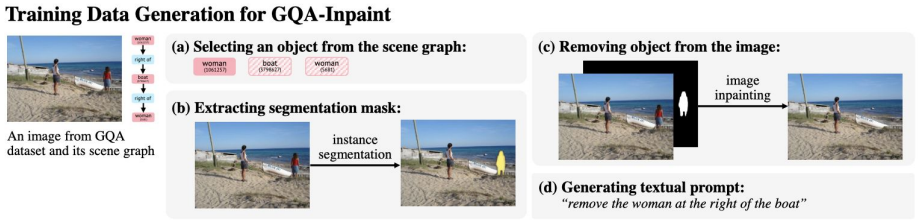
Previous Work

- Existing approaches rely on fine-tuning SD with synthetic datasets
- Alters the model's generation capability

SD-Inpaint has built-in capacity for object removal, no need for additional training.



Instructpix2pix [2]

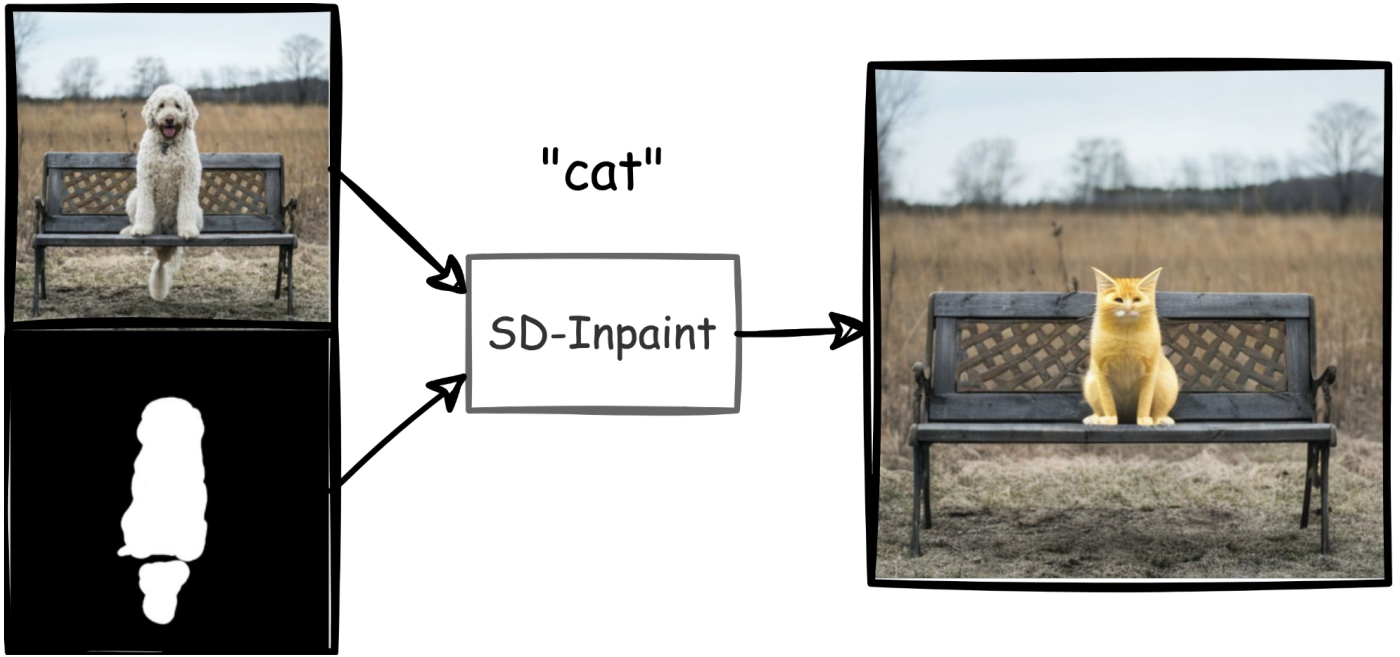


Inst-Inpaint [3]

[2] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
[3] Yildirim, A.B., Baday, V., Erdem, E., Erdem, A., Dundar, A.: Inst-inpaint: Instructing to remove objects with diffusion models. arXiv preprint arXiv:2304.03246 (2023)

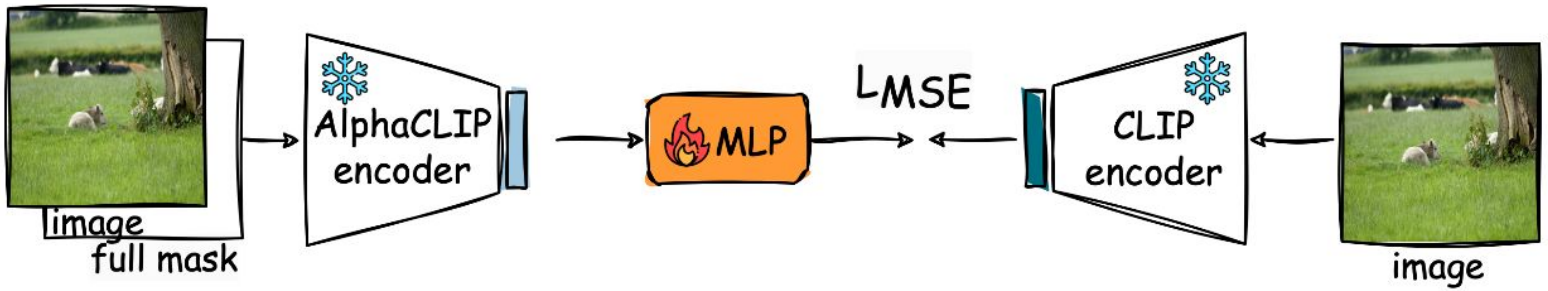
Motivation

- SD-Inpaint correctly fill the intended region with the given prompt
- **What if the prompt is the background?**



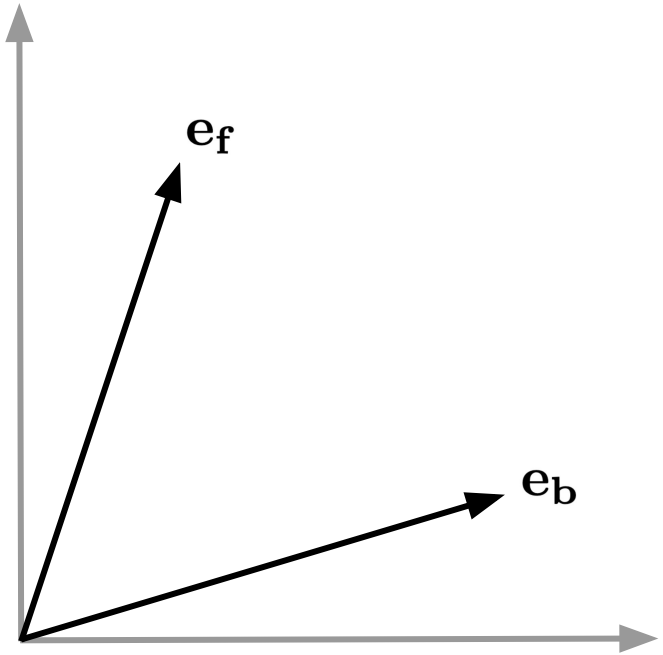
Training

- IP-Adapter [4] and AlphaCLIP [5] have different CLIP [6] spaces (Vit-H and Vit-L14)
- An MLP is trained to bridge such domain shift



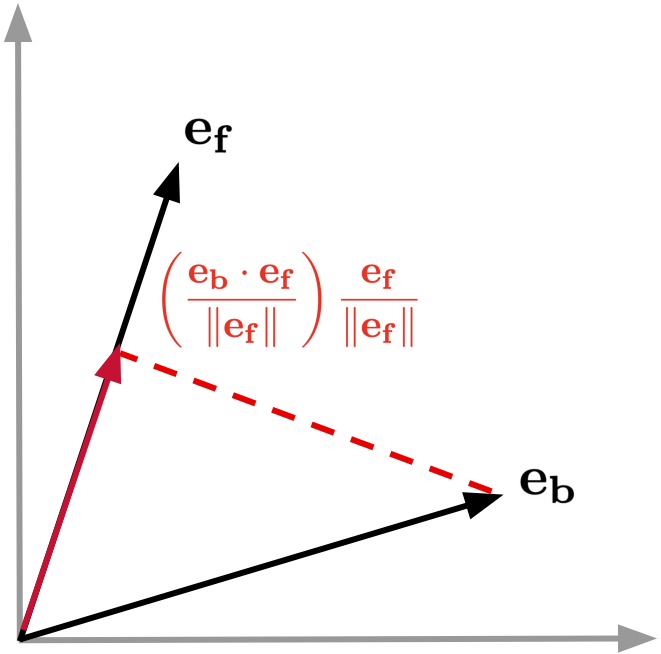
[4] Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
[5] Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D., Wang, J.: Alpha-clip: A clip model focusing on wherever you want. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
[6] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021)

Projection Block



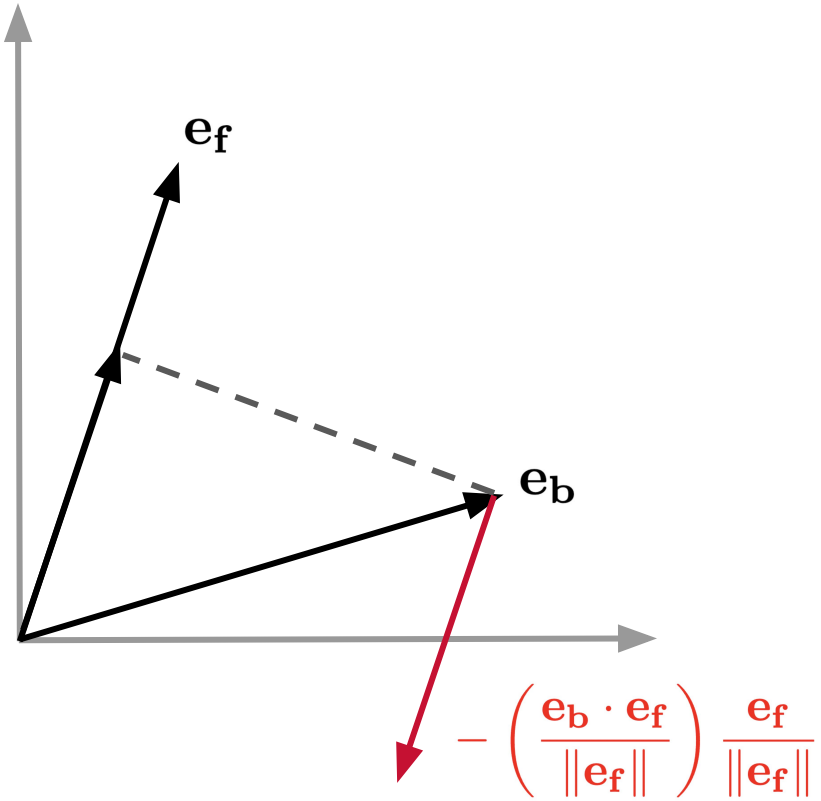
$$\mathbf{e}_{\text{final}} = \mathbf{e_b} - \left(\frac{\mathbf{e_b} \cdot \mathbf{e_f}}{\|\mathbf{e_f}\|} \right) \frac{\mathbf{e_f}}{\|\mathbf{e_f}\|}$$

Projection Block



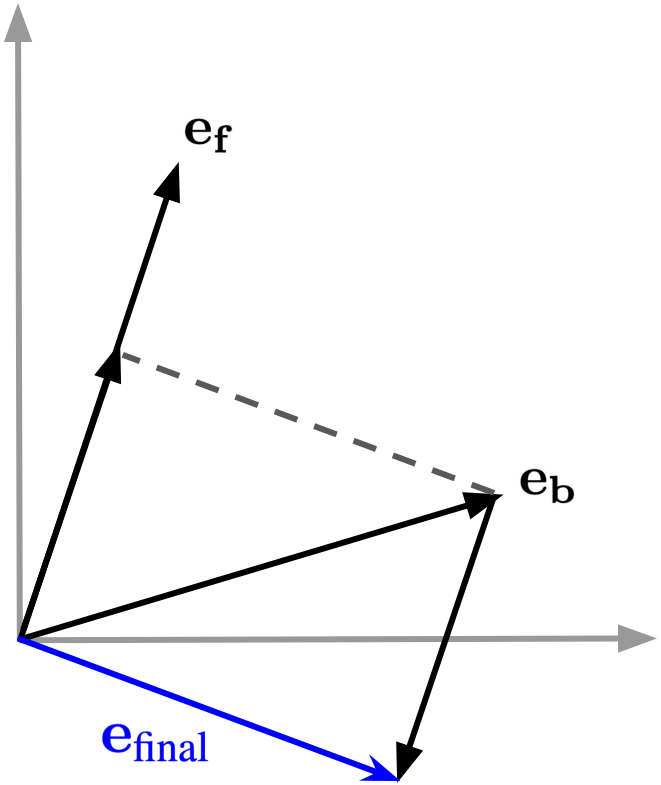
$$\mathbf{e}_{\text{final}} = \mathbf{e_b} - \left(\frac{\mathbf{e_b} \cdot \mathbf{e_f}}{\|\mathbf{e_f}\|}\right) \frac{\mathbf{e_f}}{\|\mathbf{e_f}\|}$$

Projection Block



$$\mathbf{e}_{\text{final}} = \mathbf{e_b} - \left(\frac{\mathbf{e_b}\cdot\mathbf{e_f}}{\|\mathbf{e_f}\|}\right)\frac{\mathbf{e_f}}{\|\mathbf{e_f}\|}$$

Projection Block



$$\mathbf{e}_{\text{final}} = \mathbf{e_b} - \left(\frac{\mathbf{e_b} \cdot \mathbf{e_f}}{\|\mathbf{e_f}\|} \right) \frac{\mathbf{e_f}}{\|\mathbf{e_f}\|}$$

Inference

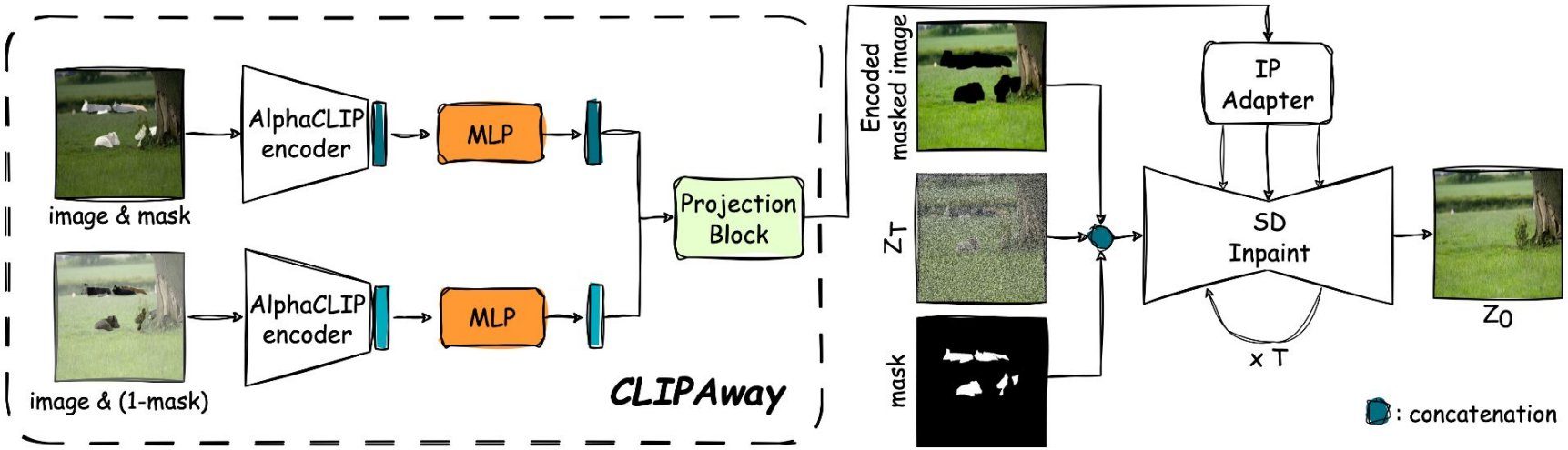
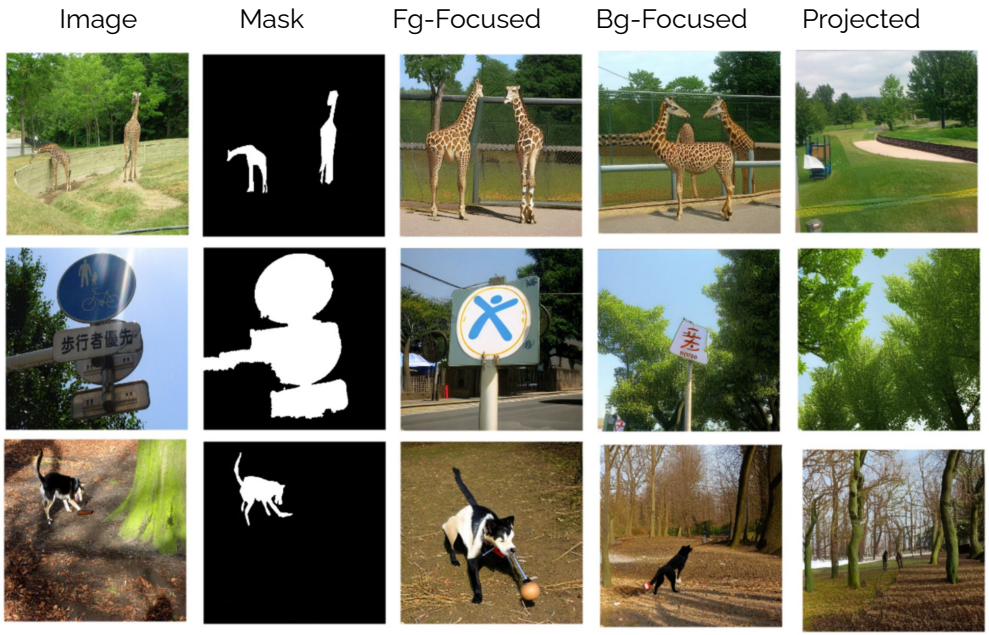
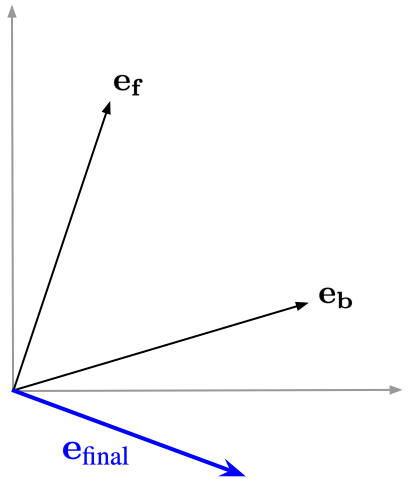


image & mask
AlphaCLIP encoder
MLP
image & (1-mask)
AlphaCLIP encoder
MLP
Projection Block
CLIPAway
Encoded masked image
z_T
mask
IP Adapter
SD Inpaint
x T
z_0
: concatenation

Effect of Projection Block





Results

- Evaluated on COCO2017 [7] validation split
- FID [8], KID [9], CMMD [10] measures photorealism
- CLIP Dist, CLIP@1, CLIP@3, CLIP@5 measure object removal ability

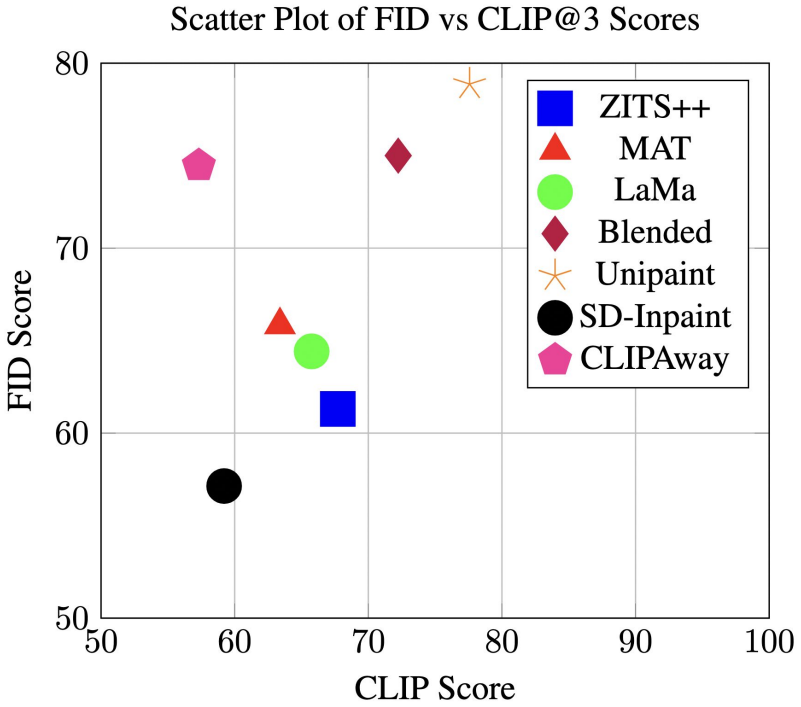
Models	FID ↓	KID ↓	CMMD ↓	CLIP Dist.↑	CLIP@1 ↑	CLIP@3 ↑	CLIP@5 ↑
ZITS++	67.72	0.0208	0.74	0.66	76.15	61.31	52.91
MAT	63.39	0.0278	0.93	0.76	80.62	65.85	60.10
LaMa	65.76	0.0195	0.81	0.66	78.34	64.42	56.85
SD-Inpaint + LaMa	51.33	0.0117	0.45	0.75	72.29	57.61	50.01
Blended Diff.	72.24	0.0362	0.89	0.85	85.69	75.01	69.34
+ CLIPAway	61.66 (-10.58)	0.0194 (-0.0168)	0.78 (-0.11)	0.83 (-0.02)	87.28 (+1.59)	78.87 (+3.86)	73.20 (+3.86)
Unipaint	77.58	0.0360	0.98	0.78	85.38	74.48	67.44
+ CLIPAway	62.18 (-15.40)	0.0199 (-0.0161)	0.79 (-0.19)	0.84 (+0.06)	88.26 (+2.88)	78.65 (+4.17)	73.05 (+5.61)
SD-Inpaint	59.21	0.0145	0.54	0.75	70.45	57.14	49.88
+ CLIPAway	57.32 (-1.89)	0.0108 (-0.0037)	0.53 (-0.01)	0.81 (+0.06)	84.82 (+14.37)	74.42 (+17.28)	67.76 (+17.88)

Table 1: **Evaluation results.** Improvements of CLIPAway over the base models are given in parenthesis for each metric.

[7] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
[8] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017)
[9] Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans (2021), https://arxiv.org/abs/1801.01401
[10] Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. arXiv preprint arXiv:2401.09603 (2023)

Results

- FID only measures photorealism
- CLIP score only measures object removal ability
- A good model should perform good in both



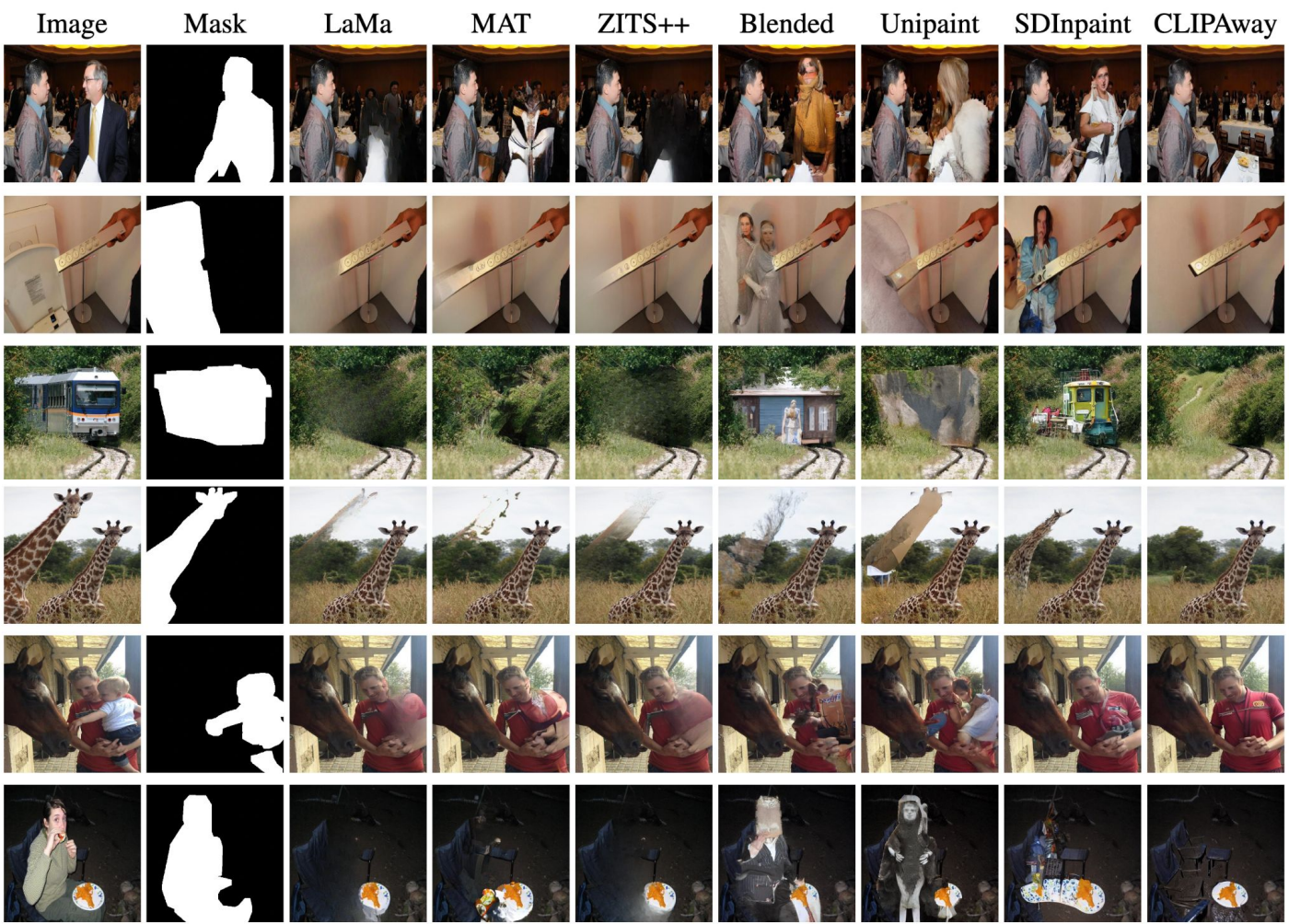


Image
Mask
LaMa
MAT
ZITS++
Blended
Unipaint
SDInpaint
CLIPAway

Final Remarks

- Using a simple arithmetic in CLIP's image embedding space, CLIPAway removes objects better than their competitors
- Stable Diffusion possesses a built-in capacity for object removal, and there is no need for additional training