IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation

Yizhi Song[1], Zhifei Zhang[2], Zhe Lin[2], Scott Cohen[2], Brian Price[2], Jianming Zhang[2], Soo Ye Kim[2], He Zhang[2], Wei Xiong[2], Daniel Aliaga[1]
[1]Purdue University, [2]Adobe Research

Poster: WED-PM-7884
Project page: https://song630.github.io/IMPRINT-Project-Page/

Adobe

PURDUE UNIVERSITY

1. Task definition

- Generative object composition: first defined in *ObjectStitch*
- Given the location + scale, insert the object into a background image

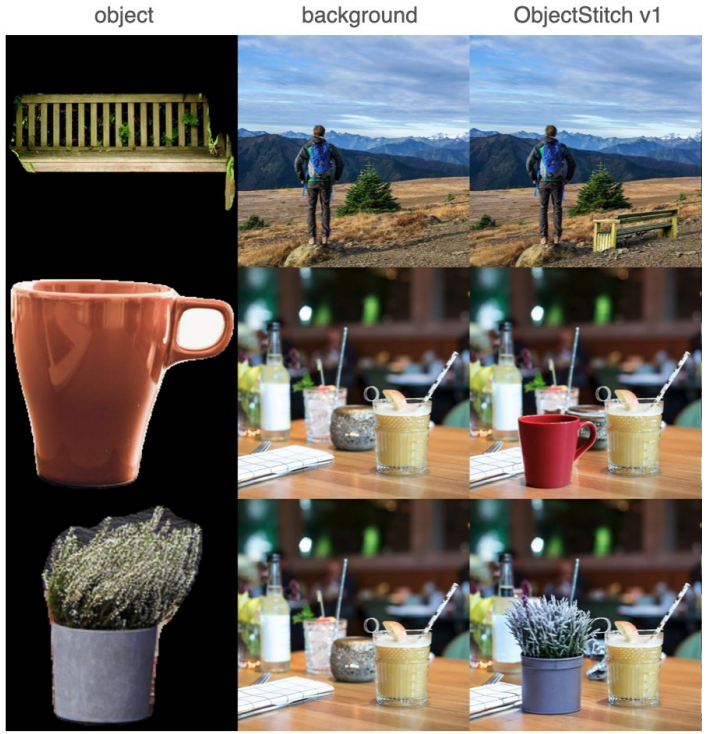
Requirements:

- Harmonization
- Shadow synthesis
- View synthesis
- Identity preservation

Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y. and Aliaga, D., 2023. Objectstitch: Object compositing with diffusion model. CVPR 2023.

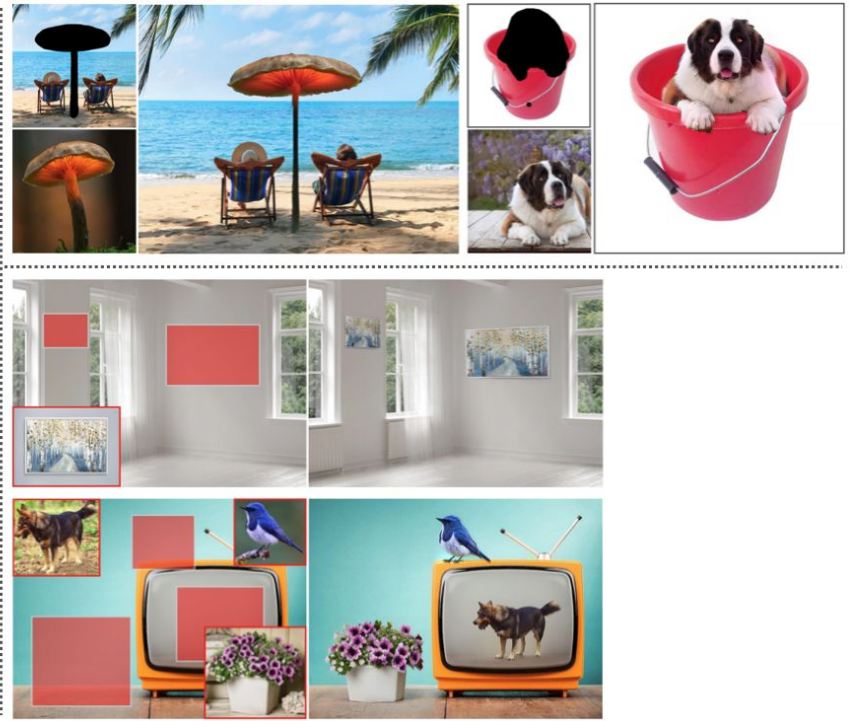
2. Limitations

- Cannot well preserve the identity
- No control over the generation



3. *IMPRINT*: overview

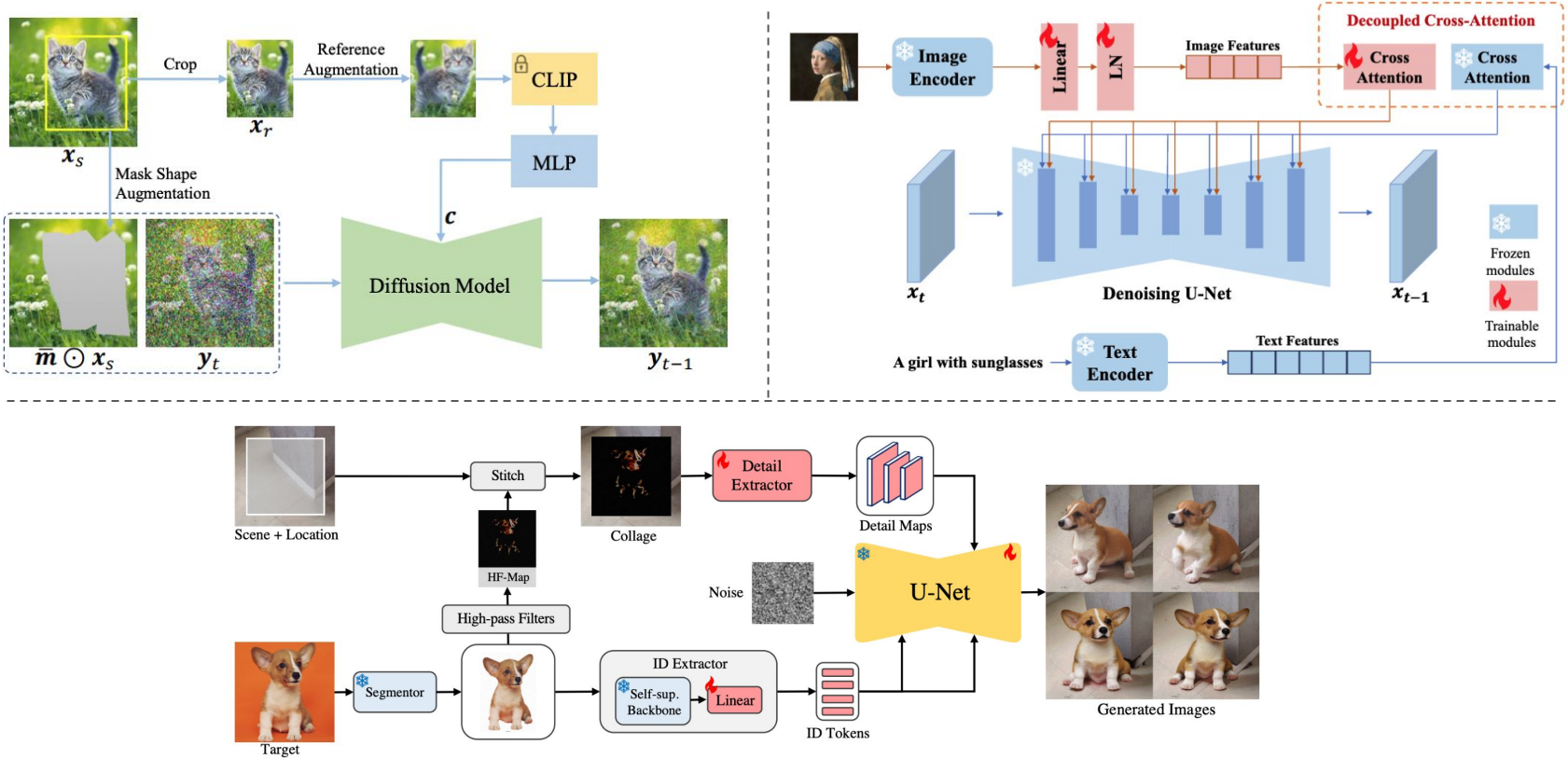
- Identity-preservation is improved
- Mask-control is supported to edit object poses



Song, Y., Zhang, Z., Lin, Z., Cohen, S., Price, B., Zhang, J., Kim, S.Y., Zhang, H., Xiong, W. and Aliaga, D., 2024. IMPRINT: Generative Object Compositing by Learning Identity-Preserving Representation. CVPR 2024.

4. Related works

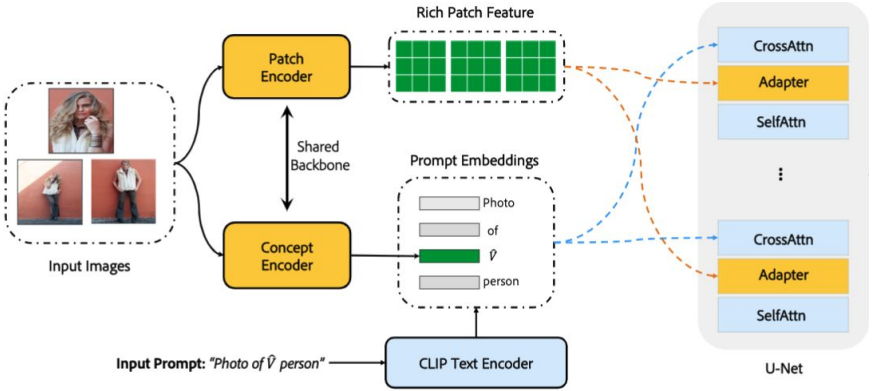
- Encoder-based methods: *Paint-by-Example*, *AnyDoor*, *IP-Adapter*

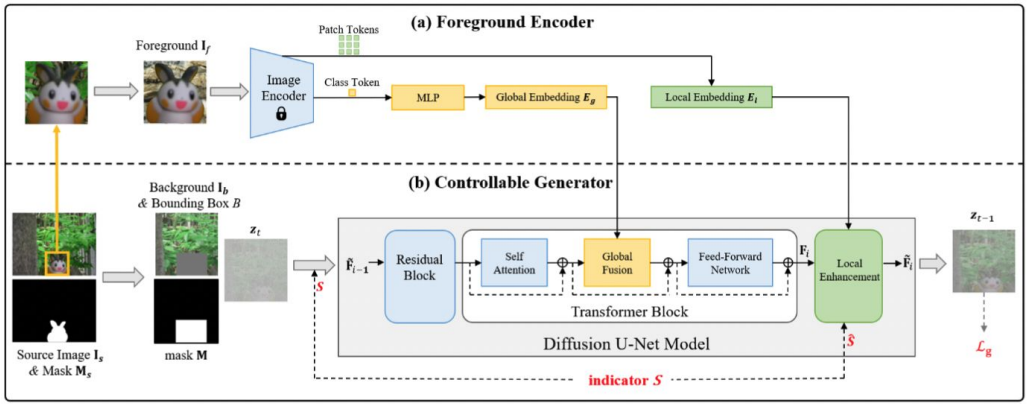


Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D. and Wen, F., 2023. Paint by example: Exemplar-based image editing with diffusion models. CVPR 2023.
Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D. and Zhao, H., 2023. Anydoor: Zero-shot object-level image customization. CVPR 2024.
Ye, H., Zhang, J., Liu, S., Han, X. and Yang, W., 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721.

4. Related works

- Attention manipulation: *InstantBooth*, *ControlCom*, *TF-ICON*

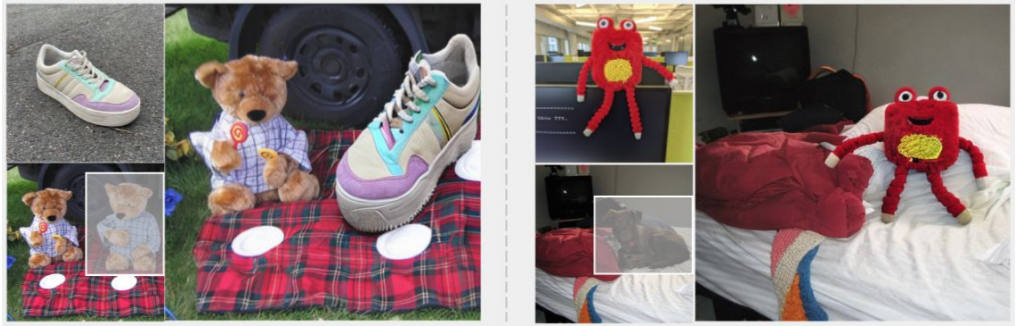




Shi, J., Xiong, W., Lin, Z. and Jung, H.J., 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. CVPR 2024.
Zhang, B., Duan, Y., Lan, J., Hong, Y., Zhu, H., Wang, W. and Niu, L., 2023. Controlcom: Controllable image composition using diffusion model. arXiv preprint arXiv:2308.10040.
Lu, S., Liu, Y. and Kong, A.W.K., 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. ICCV 2023.

4. Related works - common limitations

- Trade-off between identity and diversity
- Their capacity for geometric correction / 3D rotation is significantly limited

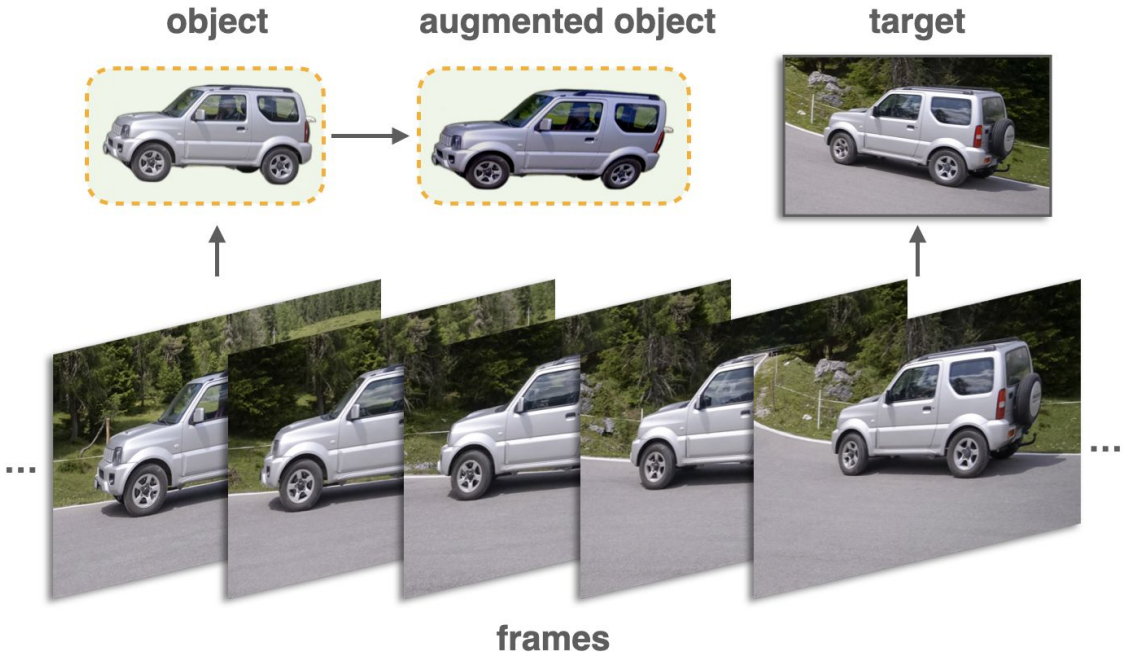


‘ a professional photograph of a teddy bear, ultra realistic’

‘ a professional photograph of a mailbox on the grass, ultra realistic’

5. Approaches - dataset

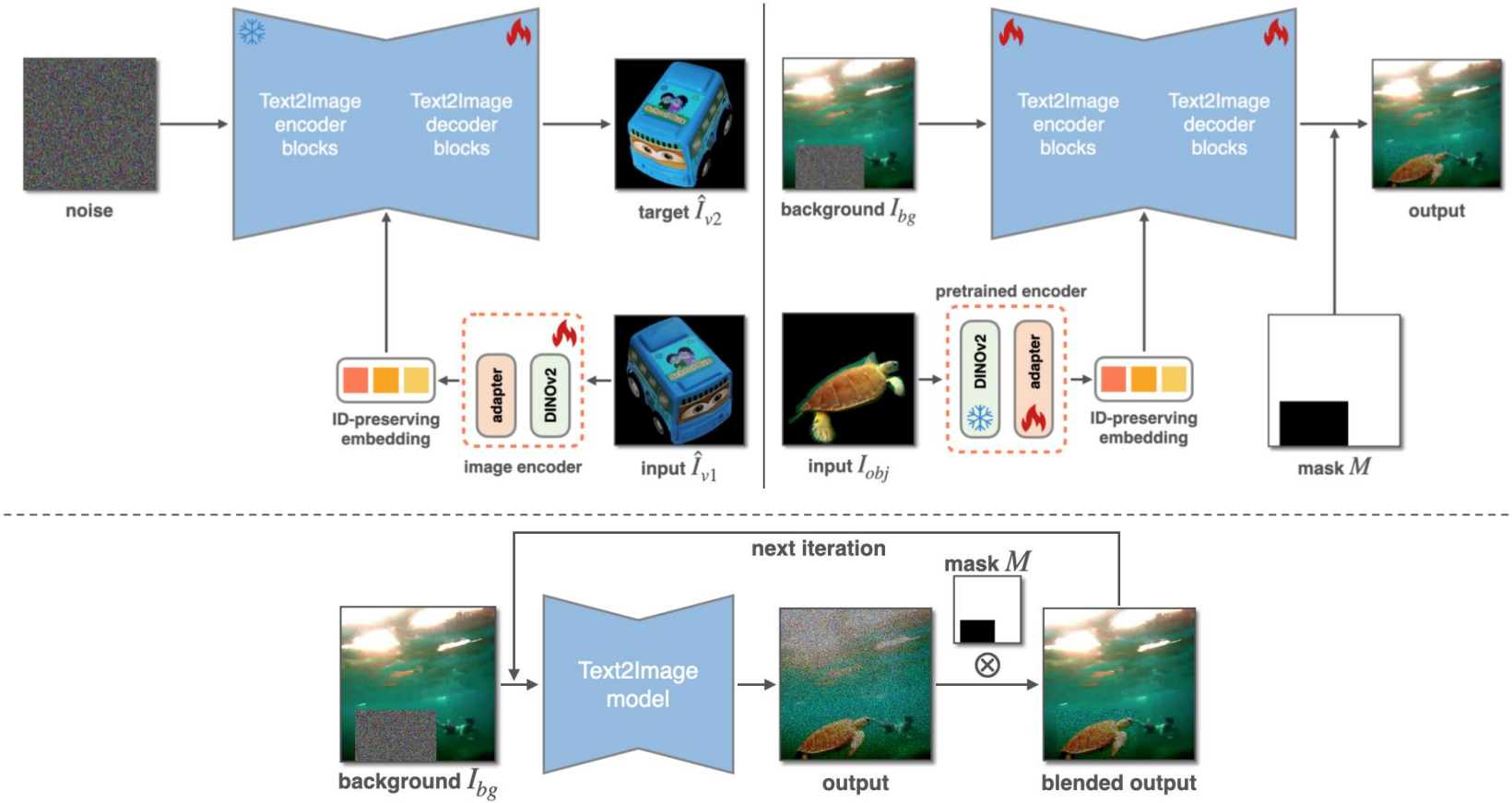
- Paired data from video datasets (VIPSeg, YoutubeVOS, PPR10K, MVImgNet)
- Training pair generation: randomly select frame a as input and frame b as target



Datasets	Pixabay	VIPSeg	YoutubeVOS	PPR10K
Training	116,820	51,743	42,868	6,020
Validation	6,490	5,487	3,690	102

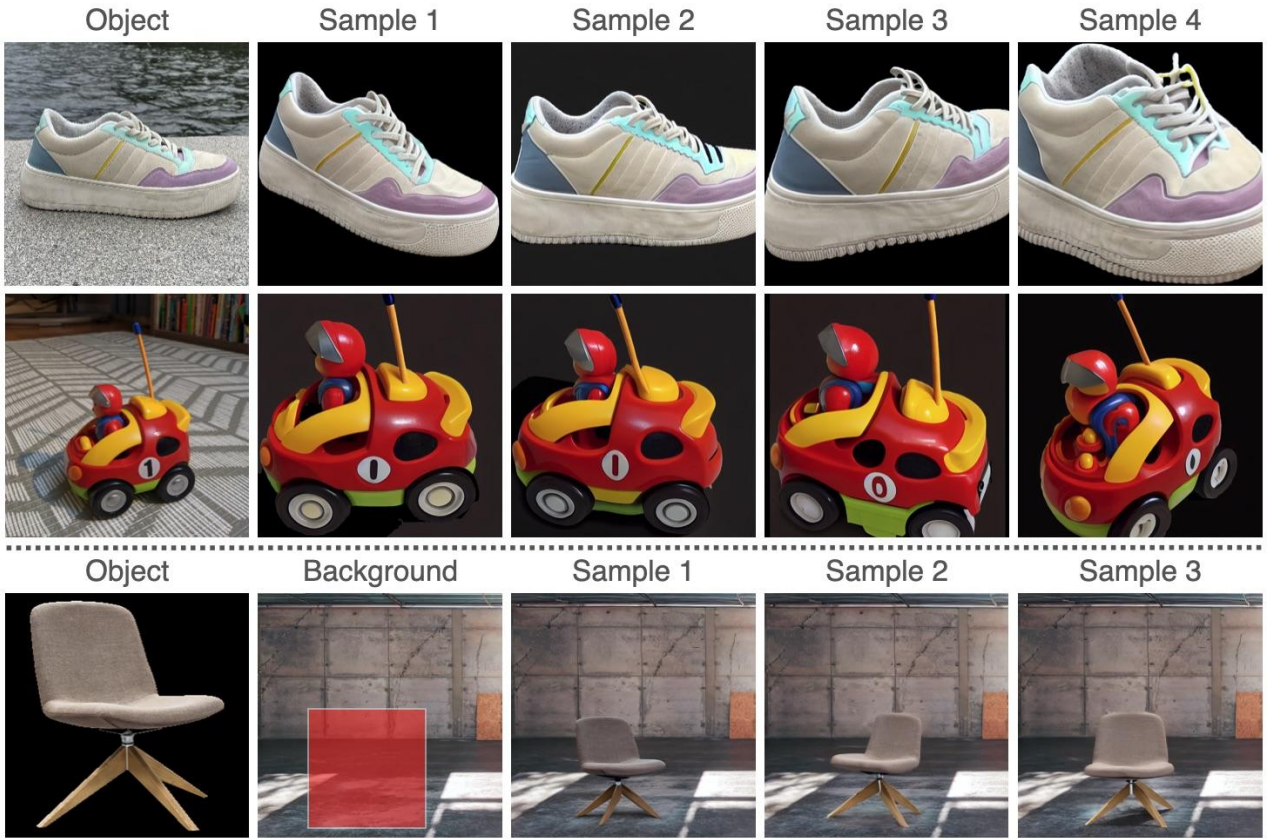
5. Approaches - pipeline

- Decouple the compositing task into two stages: identity preserving and background alignment
- A new image encoder to learn *id-preserving embedding*



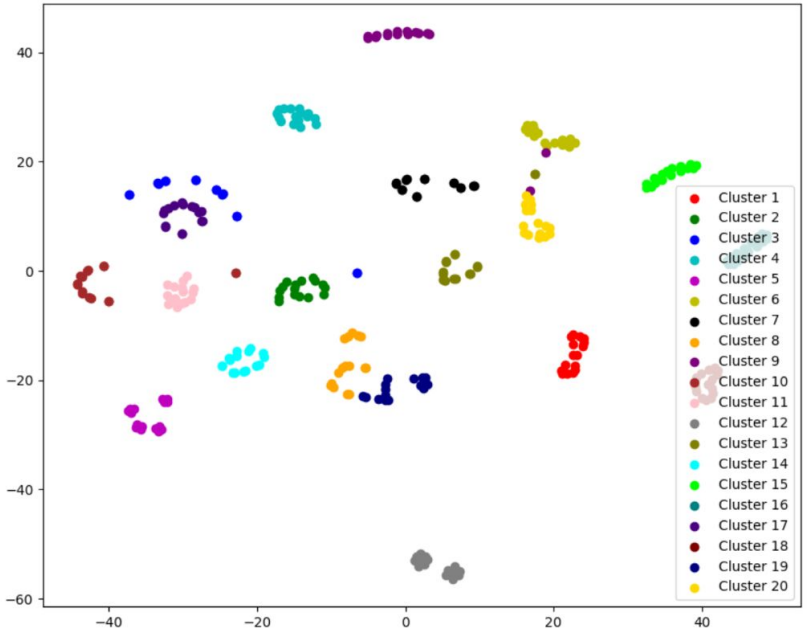
5. Approaches - pipeline

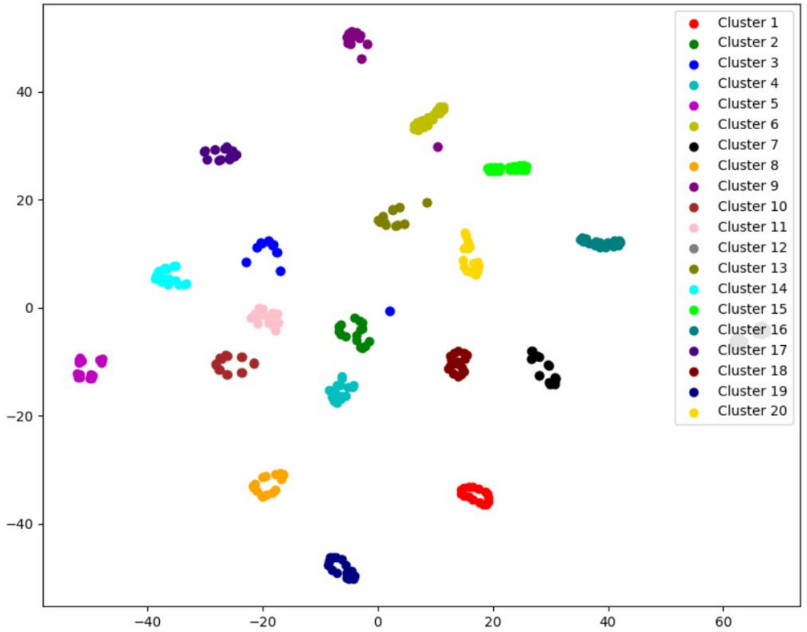
- A new image encoder to learn *id-preserving embedding*



5. Approaches - pipeline

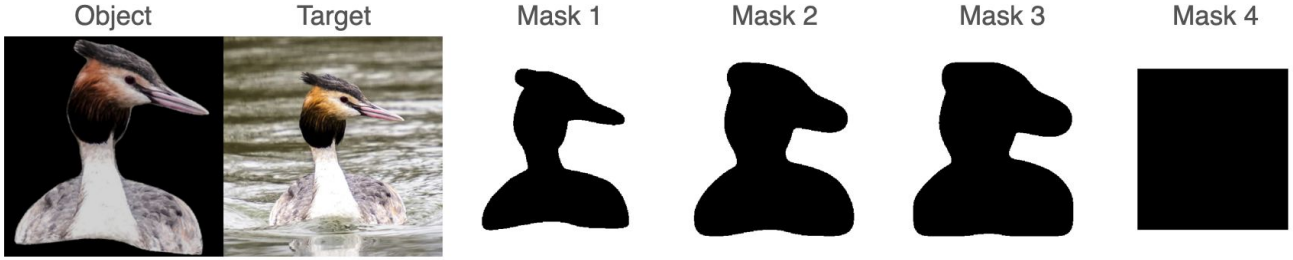
- use DINOv2 (before and after 1st stage) to predict embeddings of different views of 20 Objaverse objects; the embeddings are then clustered

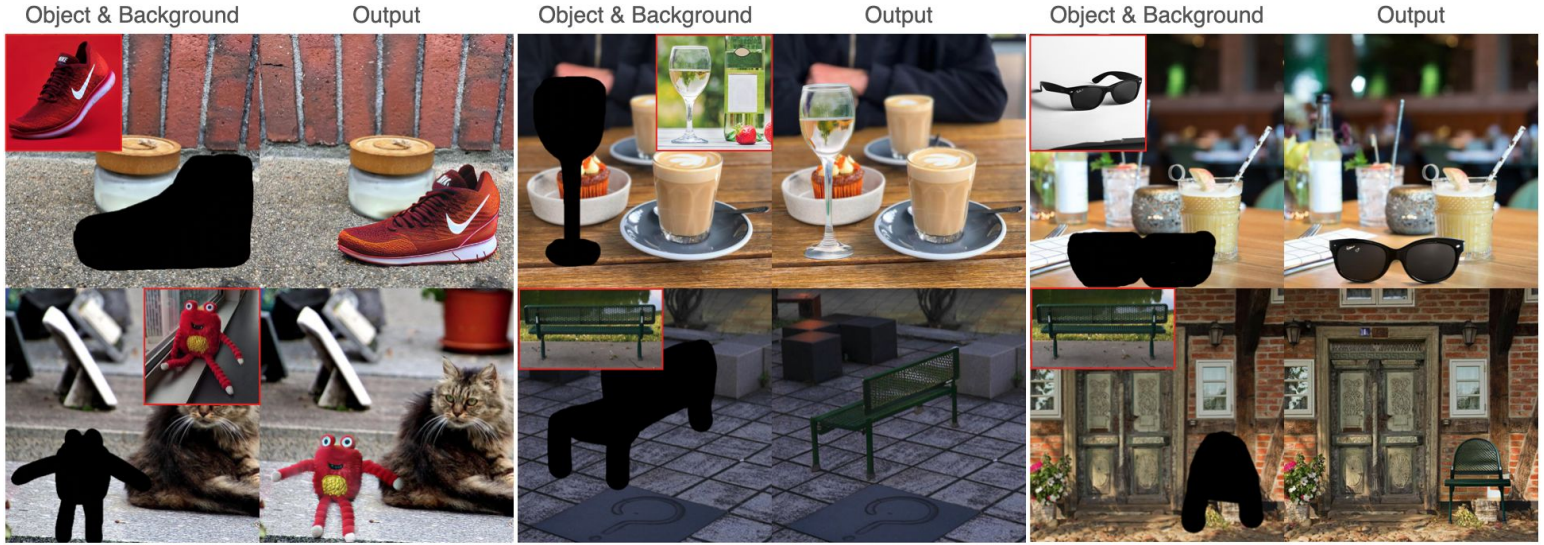




5. Approaches - mask control

- Train with 4 levels of coarse masks
- Depending on the shape of the coarse mask, it can operate different types of editing, including changing the view of an object, and applying non-rigid transformation





6. Evaluation

- Quantitative evaluation: tested on DreamBooth objects
- User study: user preference on pairwise comparisons

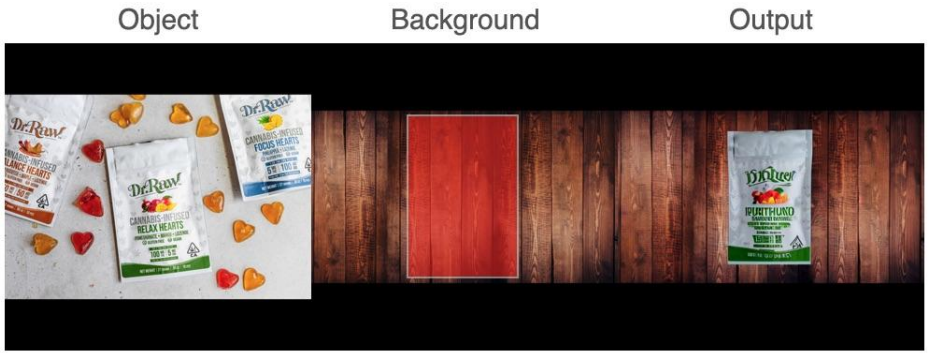
Method	FID ↓	CLIP-score↑	DINO-score↑	DreamSim ↓
PbE	-	71.5000	31.3765	0.4954
OS	-	73.6250	32.9739	0.4297
T-I	-	75.1250	39.2863	0.3661
Ours	-	**77.0625**	**43.4463**	**0.2898**
PbE	23.2663	93.6250	85.2260	0.1907
OS	22.4934	94.9375	90.3853	0.1422
T-I	63.9730	88.3125	73.2155	0.3219
Ours	**16.4487**	**96.1875**	**94.705**	**0.0831**

	Ours	OS	Ours	PbE	Ours	T-I
Realism	**50.68**	49.32	**62.84**	37.16	**53.38**	46.62
Fidelity	**80.41**	19.59	**86.49**	13.51	**73.65**	26.35

(a) Object (b) Background (c) PbE (d) OS (e) TF-ICON (f) Ours

7. Future work

- Degradation in large 3D transformations
- Artifacts in texts / human faces



Thank you!