SmartBrush: Text and Shape Guided Object Inpainting with Diffusion Model

Shaoan Xie [1,*], Zhifei Zhang[2], Zhe Lin[2], Tobias Hinz[2], Kun Zhang[1,3]
[1]Carnegie Mellon University
[2]Adobe Research
[3]Mohamed bin Zayed University of Artificial Intelligence
shaoan@cmu.edu, {zzhang, zlin, thinz}@adobe.com, kunz1@cmu.edu

CVPR 2023

Presenter: Hao Wang

Advisor: Prof. Chia-Wen Lin

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

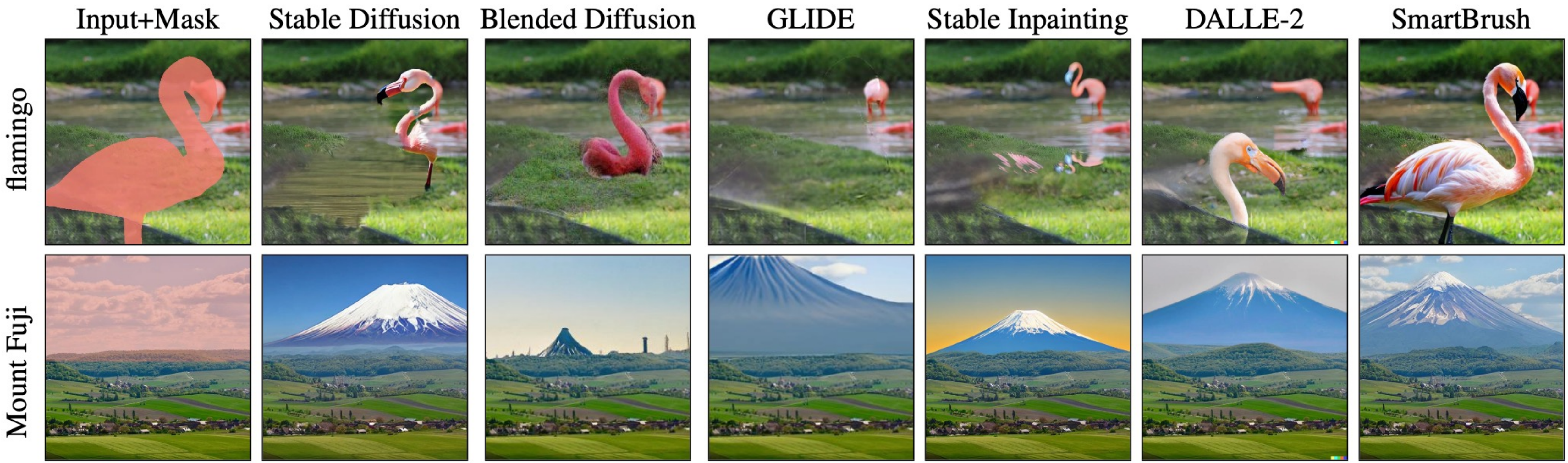
Introduction

- Introduce a text and shape guided object inpainting diffusion model, which is conditioned on **object masks of different precision**, achieving a new level of control for object inpainting.

- To **preserve the image background** with coarse input masks, the model is trained to **predict a foreground object mask** during inpainting for preserving original background surrounding the synthesized object.

- Propose a multi-task training strategy by jointly training object inpainting with text-to-image generation to leverage more training data.

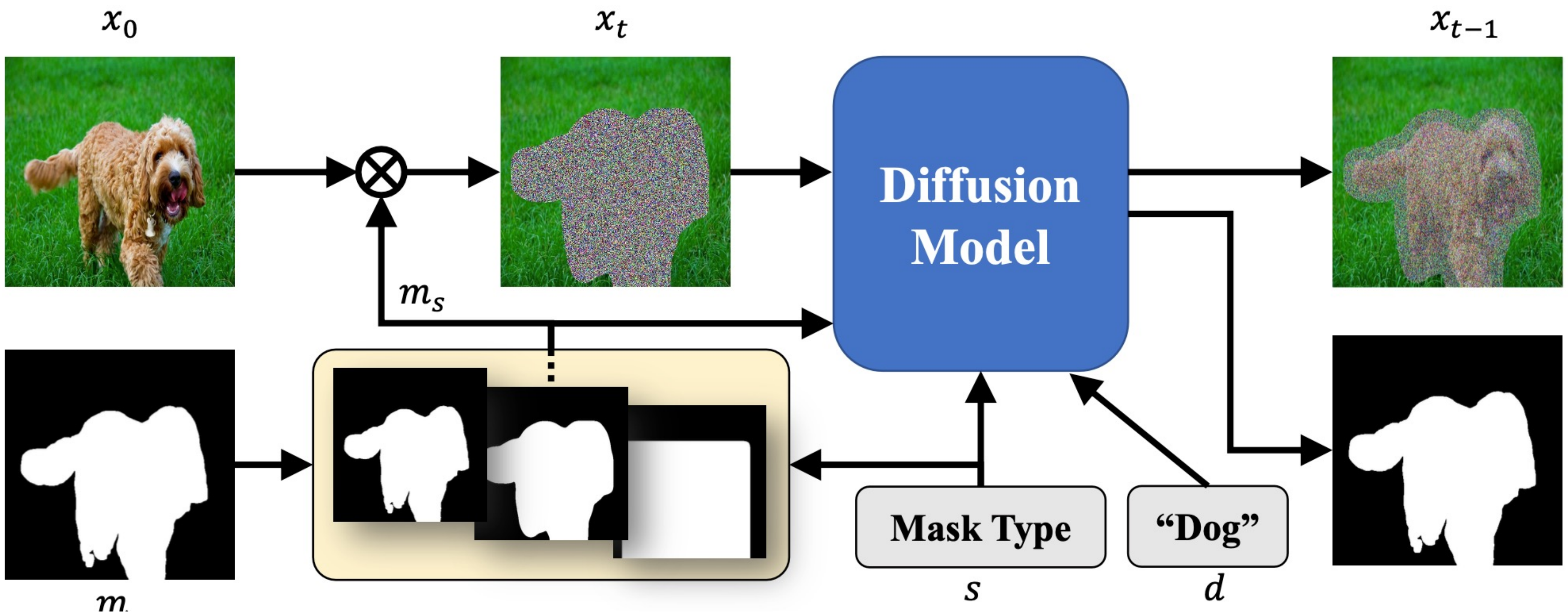
Introduction



Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Framework

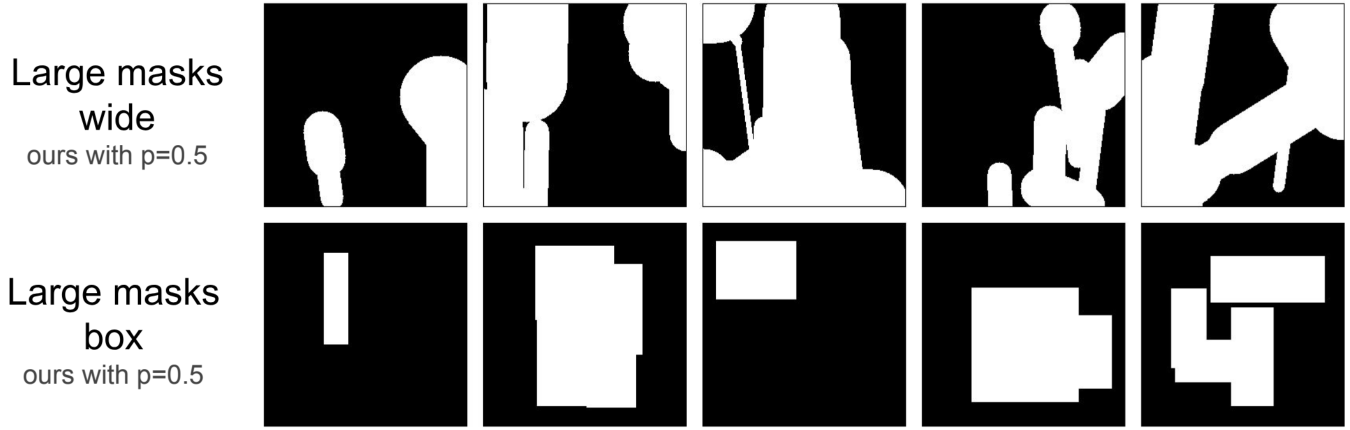


Outline

- Preliminary
- Framework
- Method
- Experiment
- Conclusion

Preliminary- SD Inpainting Training dataset

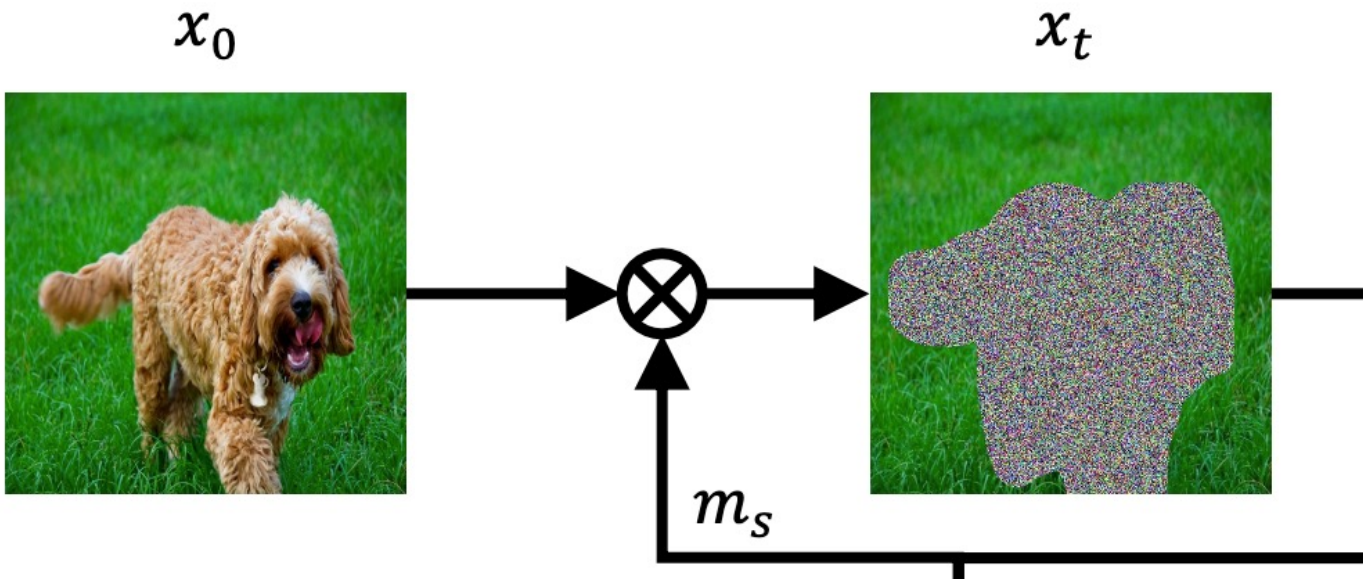
- Existing inpainting models randomly erase part of the images.



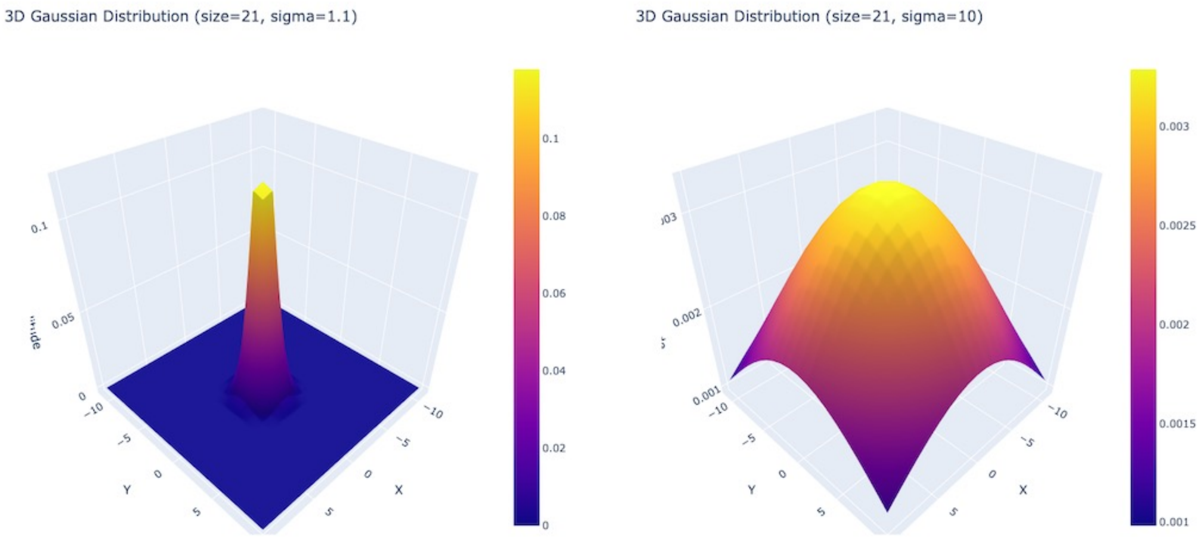
Text and Shape Guided Diffusion

- utilize the text and shape information from existing instance or panoptic segmentation datasets.

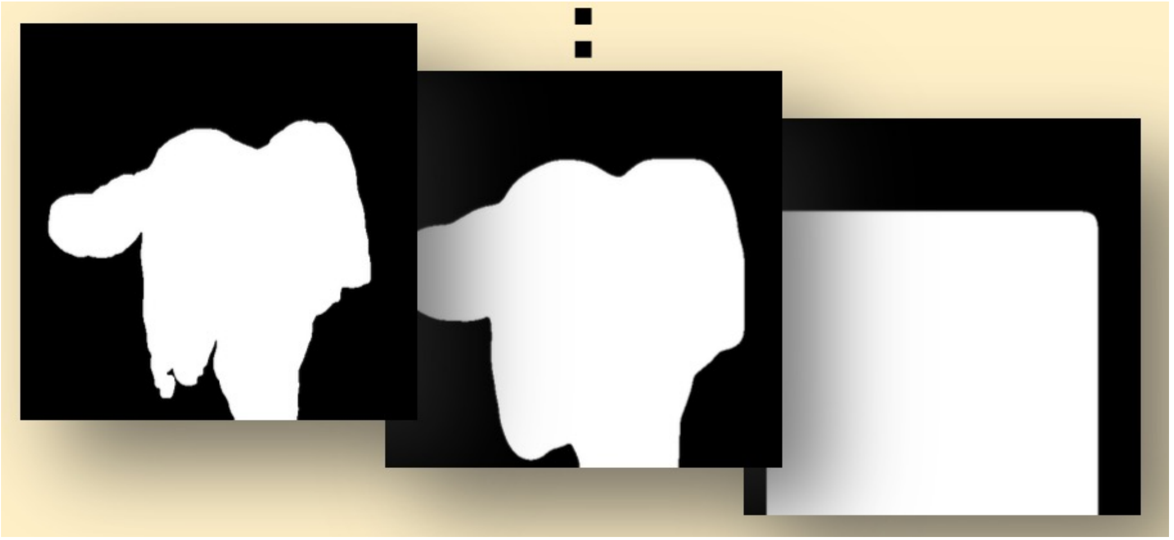
$$\tilde{x}_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$$
$$x_t = \tilde{x}_t \odot m + x_0 \odot (1 - m),$$



Shape Precision Control

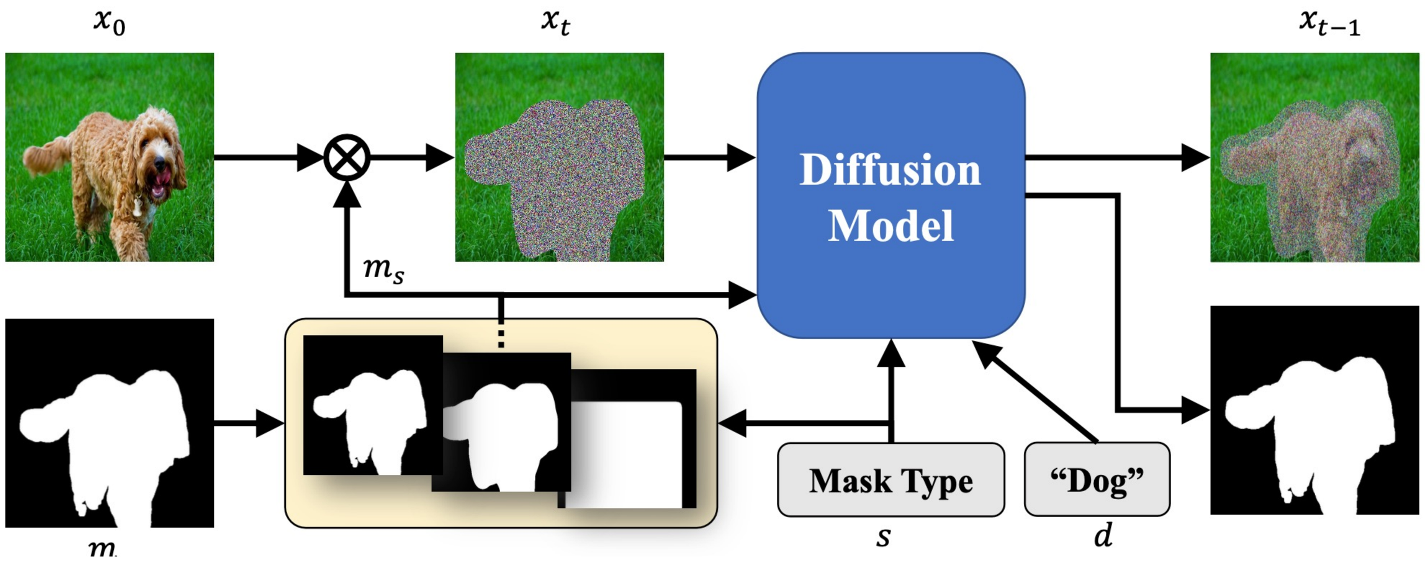


$$m_s = \text{GaussianBlur}(m, k_s, \sigma_s), \tag{7}$$



Shape Precision Control

- we can control whether the generated object should align with the input mask by specifying different mask precision indicators *s*

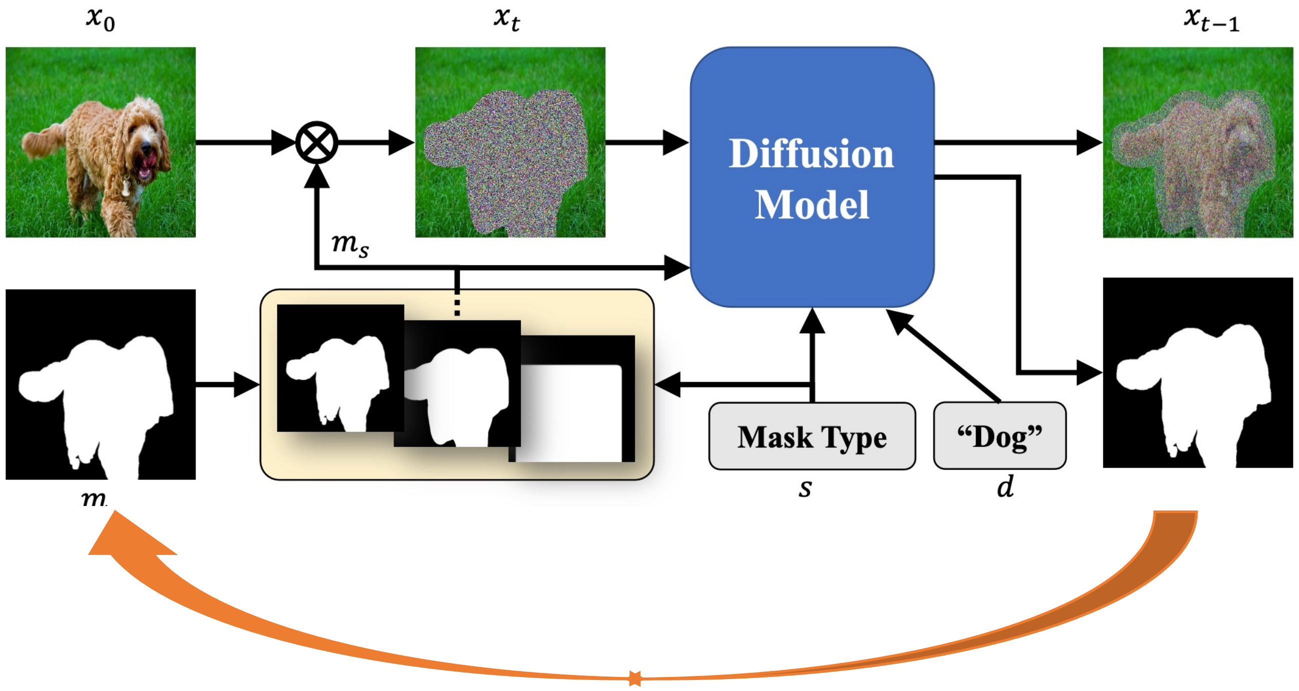


$$\mathcal{L}_{\text{seg-DM}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(x_t, t, m_s, c, s)\|_2^2 \right]. \quad (8)$$

Background Preservation

- Background in the masked region will be changed if the input masks are coarse

- Also predict an accurate instance mask m from the coarse input version m_s



$$\mathcal{L}_{\text{prediction}} = H(\epsilon_\theta(m_s), m), \qquad H(X,Y) = 1 - \frac{2|X \cap Y|}{|X|+|Y|}$$

Training Strategy

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seg-DM}} + \lambda \mathcal{L}_{\text{prediction}}, \qquad (10)$$

- $\lambda = 0.01$
- Jointly training our main task and input **mask to cover the entire image**.
- Pair the segmentation label or **BLIP caption** to the corresponding mask.
- Model can be built based on pre-trained generation models.

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Text and Shape Guided Inpainting

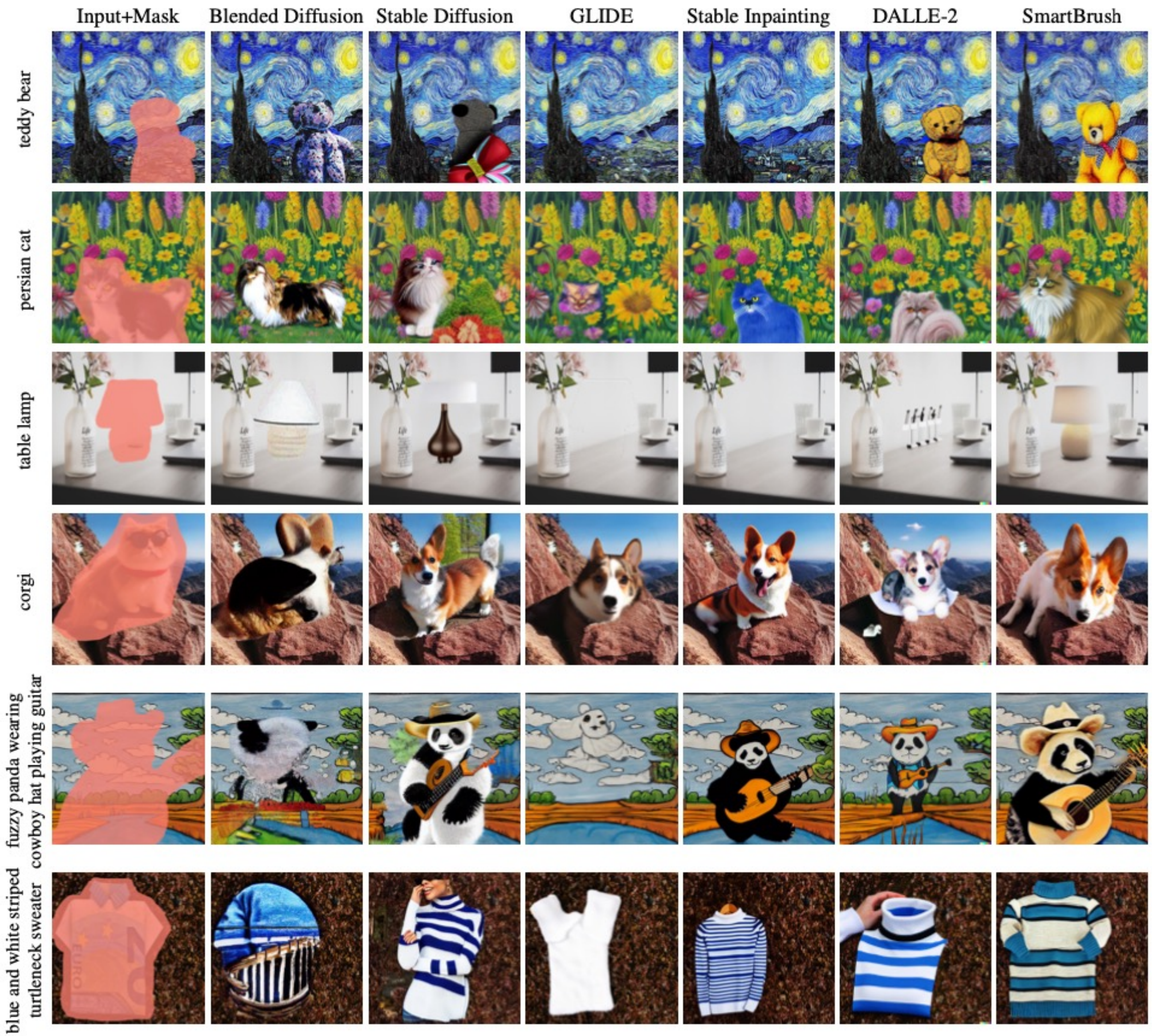
Table 1. Text-guided object inpainting with bounding box mask.

	OpenImages			MSCOCO		
	Local FID ↓	CLIP Score ↑	FID ↓	Local FID ↓	CLIP Score ↑	FID ↓
Blended Diffusion [2]	29.16	0.265	11.05	41.43	0.251	12.68
GLIDE [16]	22.45	0.252	9.70	30.72	0.241	9.32
Stable Diffusion [20]	15.28	0.265	9.10	25.61	0.250	12.29
Stable Inpainting [20]	12.57	0.264	7.07	18.13	0.246	8.50
SmartBrush (Ours)	**9.71**	**0.266**	**6.00**	**13.22**	**0.252**	**8.05**

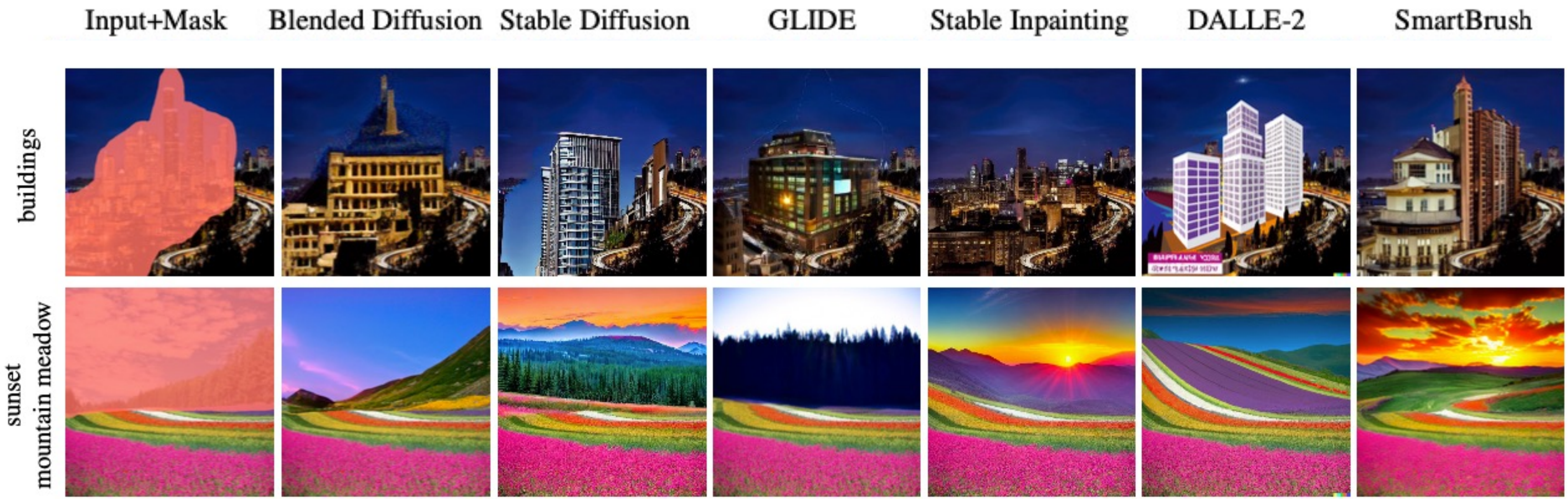
Table 2. Text-guided object inpainting with object layout mask.

	OpenImages			MSCOCO		
	Local FID ↓	CLIP Score ↑	FID ↓	Local FID ↓	CLIP Score ↑	FID ↓
Blended Diffusion [2]	21.93	0.261	9.72	26.25	0.244	8.16
GLIDE [16]	21.09	0.250	9.03	24.25	0.235	6.98
Stable Diffusion [20]	12.27	**0.263**	6.90	17.16	0.246	7.78
Stable Inpainting [20]	10.98	0.261	5.84	15.16	0.243	6.54
SmartBrush (Ours)	**7.82**	**0.263**	**4.70**	**9.80**	**0.249**	**5.76**

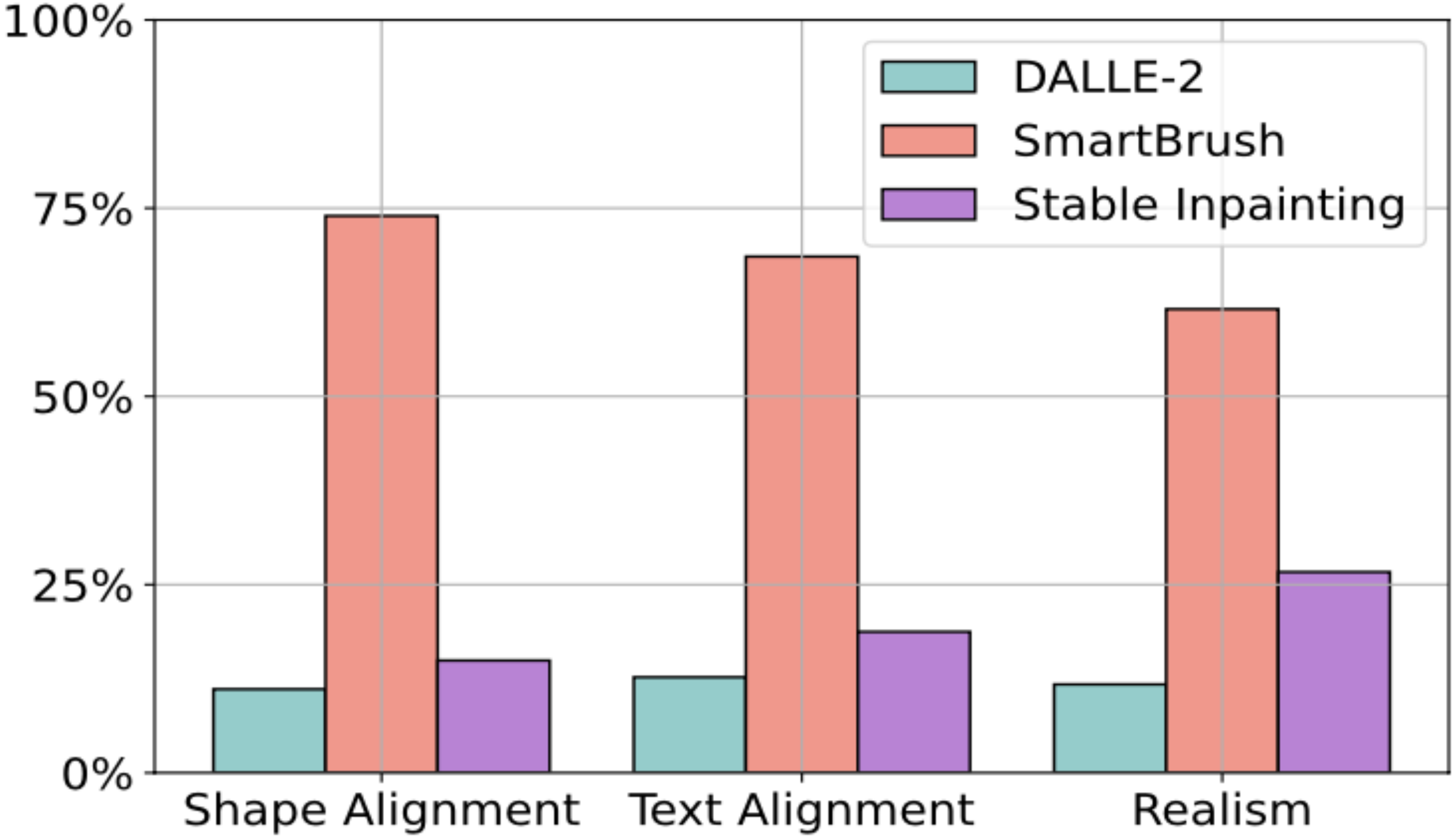
Accurate object masks



Bounding box masks



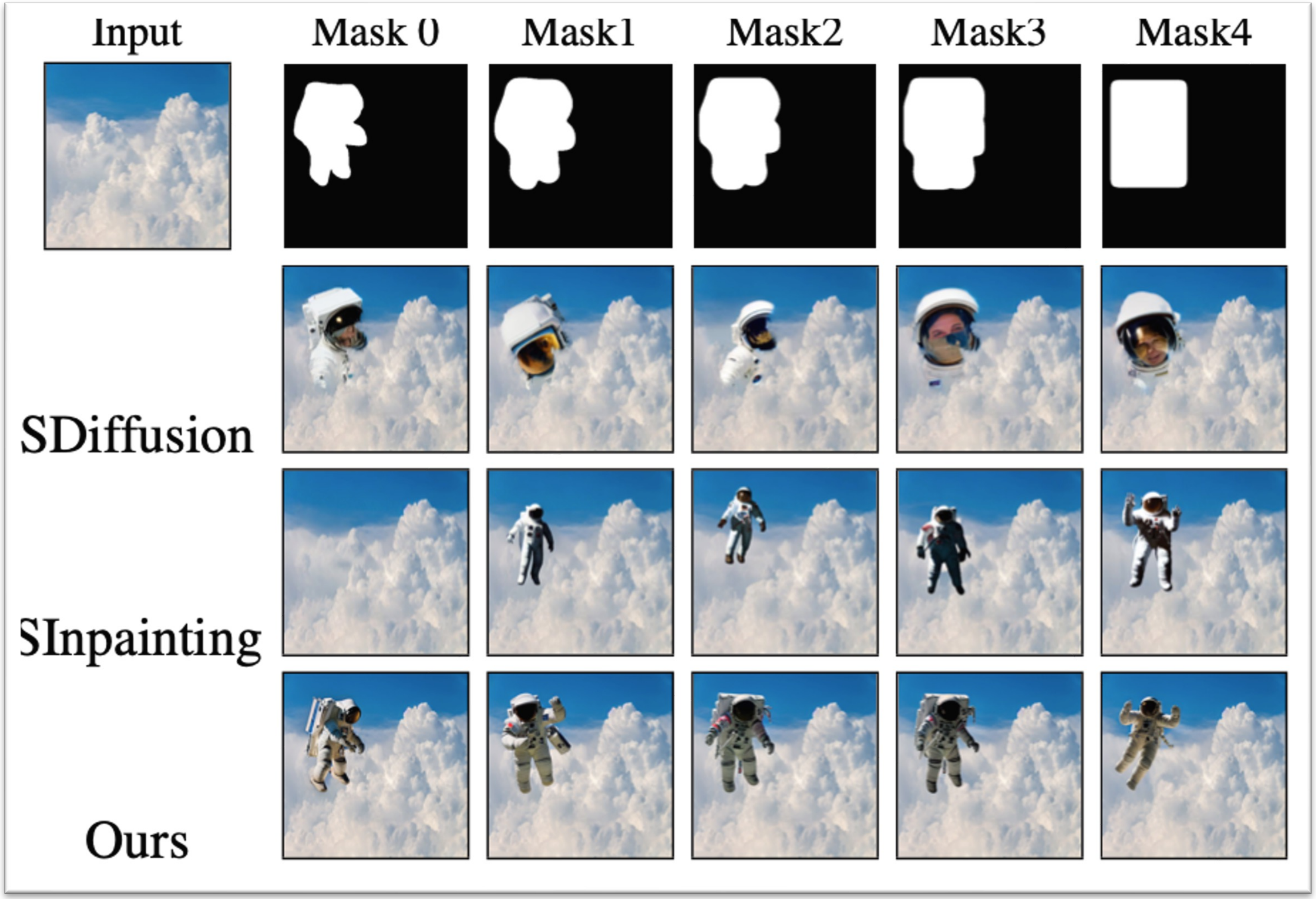
User study



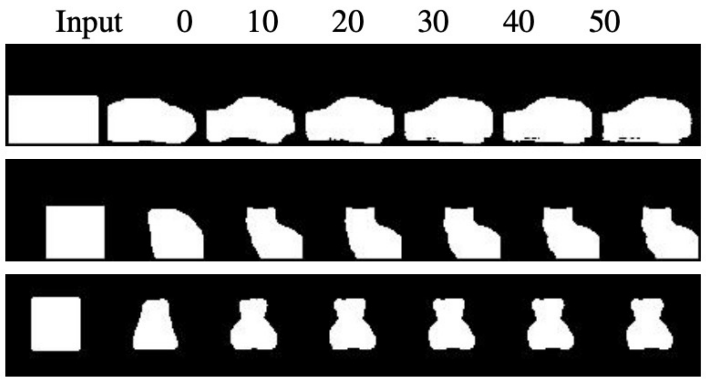
100%
75%
50%
25%
0%
DALLE-2
SmartBrush
Stable Inpainting
Shape Alignment
Text Alignment
Realism

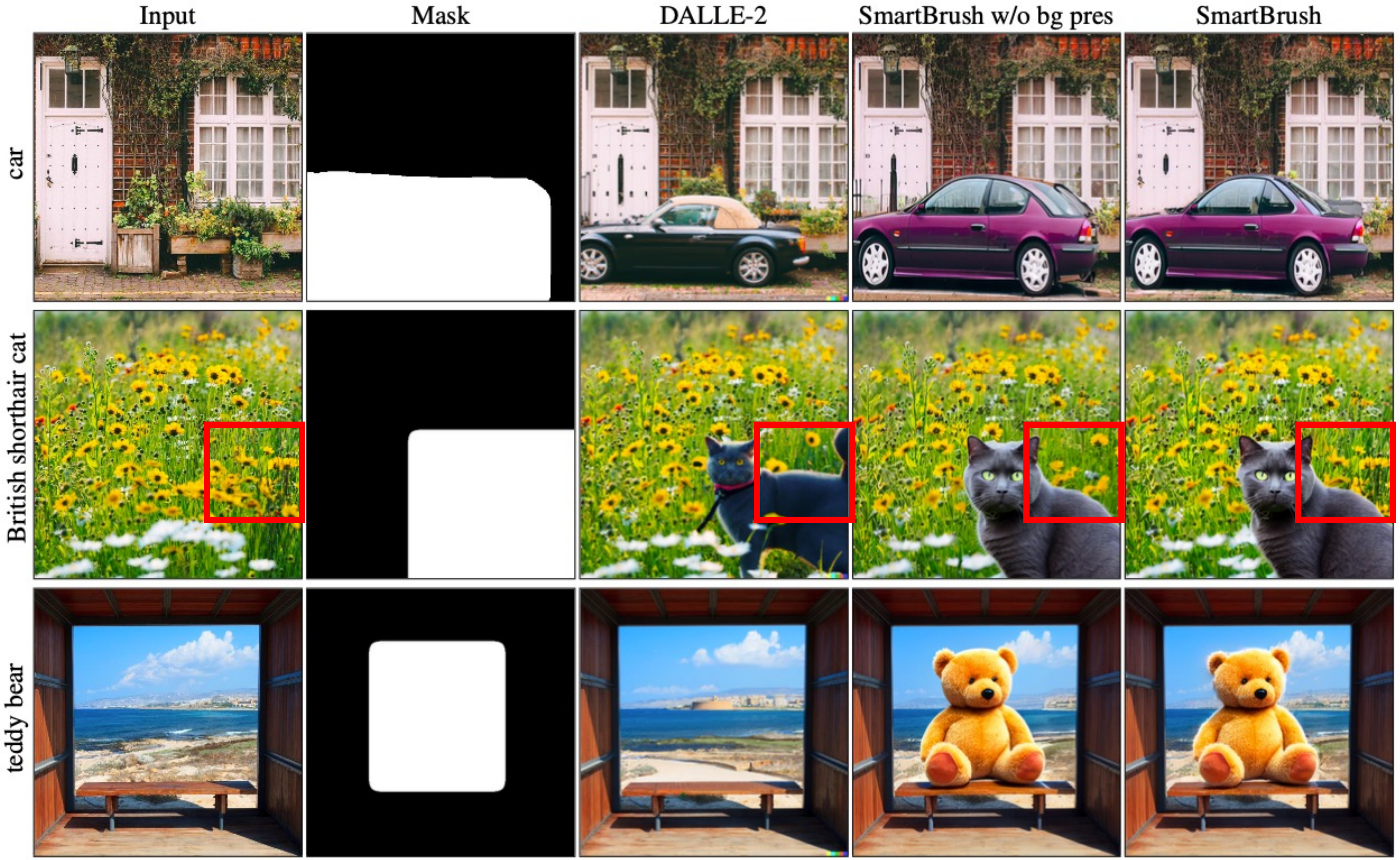
Mask Precision Control

- The Stable Diffusion
  - are not affected by mask
- Stable Inpainting
  - only change the object size with the mask size
  - not follow the mask shape
- SmartBrush
  - strictly follow the mask shape when providing a finer mask
  - roughly following the mask if given a coarser mask.



Background Preservation





Ablation Study

Method	**LFID**$\downarrow$	**CLIP**$\uparrow$	**FID** $\downarrow$
Ours	13.22	0.252	8.05
+ Background Preservation	12.26	0.251	7.19
- Mask Precision Cond	15.31	0.252	8.57
- BLIP Prompts	13.52	0.249	10.69
- Multi-Task	15.26	0.250	8.26
Stable Inpainting (SOTA)	18.13	0.246	8.50
+ Finetune on Our Dataset	18.34	0.245	8.38

Outline

- Introduction
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- In this paper, we propose a novel training method that **utilizes the text and shape guidance from the segmentation dataset** to address the text misalignment problem.
- Then we further propose to create **different levels of masks** to allow precision control of the generation.
- Encourage the model to make object predictions and utilize **the predicted mask** to avoid unnecessary changes inside the mask.