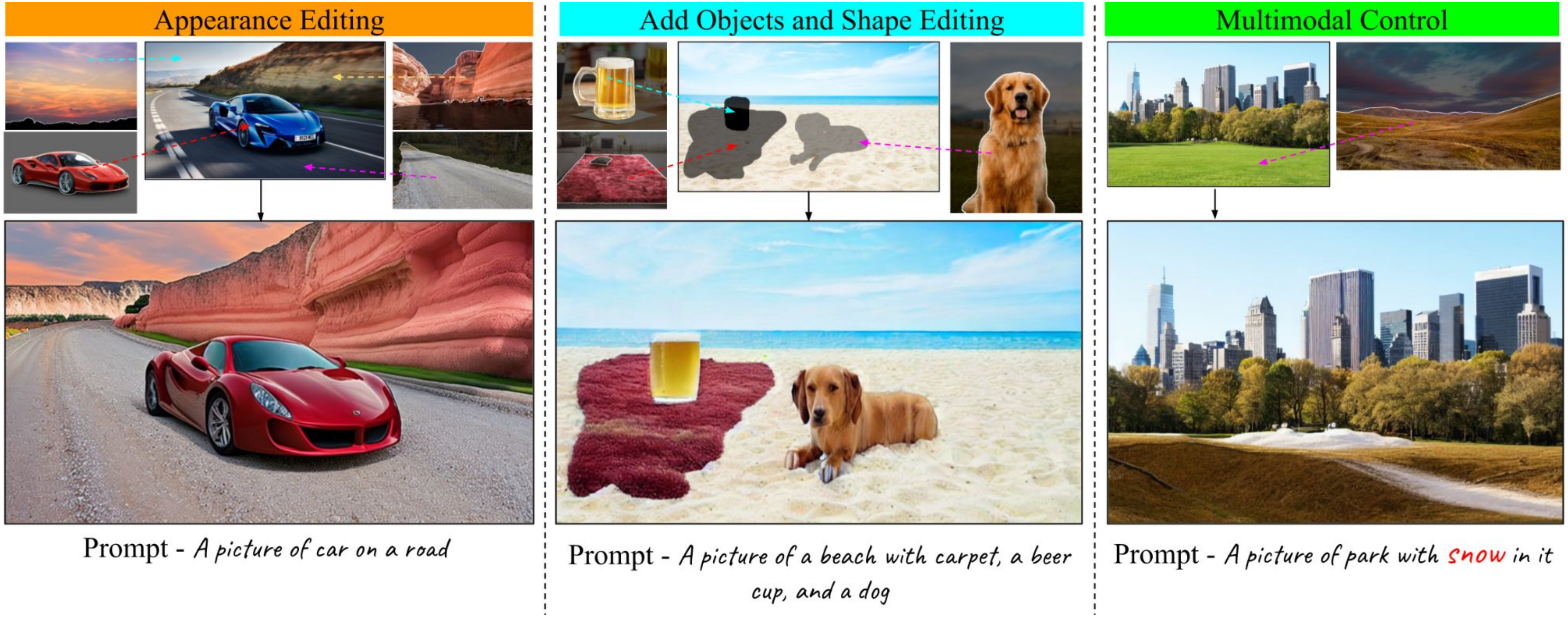
PAIR Diffusion: A Comprehensive Multimodal Object-Level Image Editor

Vidit Goel[1,2,*] Elia Peruzzo[3,*] Yifan Jiang[4] Dejia Xu[4] Xingqian Xu[1,2]
Nicu Sebe[3] Trevor Darrell[5] Zhangyang Wang[1,4] Humphrey Shi[1,2]

[1]Picsart AI Research (PAIR) [2]SHI Labs @ Georgia Tech & UIUC [3]University of Trento [4]UT Austin [5]UC Berkeley

Contributions



Appearance Editing

Prompt - *A picture of car on a road*

Add Objects and Shape Editing

Prompt - *A picture of a beach with carpet, a beer cup, and a dog*

Multimodal Control

Prompt - *A picture of park with snow in it*

- **General** framework to enable precise object-level editing in Diffusion Models.
- **Comprehensive** editing capabilities under a single model trained once.
- **Multimodal Classifier-free Guidance** to control all editing operations using both text and images

Introduction

- Each image can be seen as a composition of various objects.

- Each object has various properties such as shape, category, appearance, lighting.

- We aim to edit images at object level i.e. independent control over properties of the object. Further, we identify the **appearance** and **structure** to be some of the most intuitive and useful properties for editing

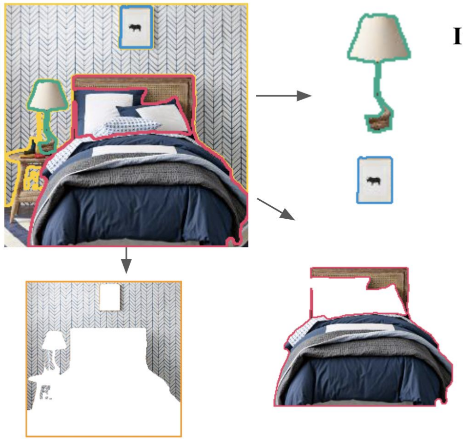
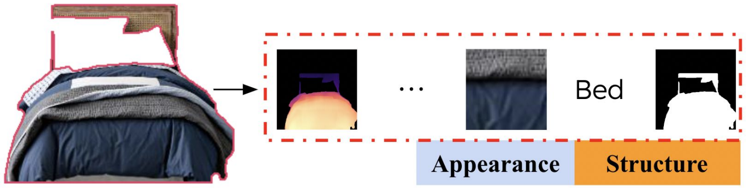


Image as Composition of Objects

- PAIR Diffusion allows localized free-form shape editing, appearance editing, editing shape and appearance simultaneously, adding objects in a controlled manner, and object-level image variation..



PAIR Diffusion

We represent images as amalgamation of objects $\mathbb{O} = \{o_1, o_2, \dots o_n\}$

Each object is further represented as (s_i , f_i) where s_i represents structure and f_i represents appearance

$$p(x|\mathbb{O}, y) = p(x|\{(s_1\, f_1), \dots, (s_n,\, f_n)\}, y)$$

Structure: It is represented by category (c_i) and shape (m_i). We use panoptic segmentation to extract c_i and m_i for each object.

Appearance: Appearance is calculated using features in initial layers of a pre-trained image encoder E_G .

$$\boldsymbol{g}_i^l = \frac{\sum_{j,k} E_G^l(x) \odot m_i}{\sum_{j,k} m_i}$$

We use a combination of VGG and DINOv2 as E_G.

Final appearance vectors are represented as $f_i = (\boldsymbol{g}_i^{Vl_1}, \boldsymbol{g}_i^{Dl_2}, \boldsymbol{g}_i^{Dl_3})$

PAIR Diffusion

Image Editing Formulation

Appearance Editing $(s_i, f_i) \rightarrow (s_i, f_i')$: We can simply manipulate f_i using $f_i' = a_0 f_i + a_1 f_i^R$

Shape Editing $(s_i, f_i) \rightarrow (s_i', f_i)$: User can directly edit m_i

Object Addition $\mathbb{O} \rightarrow \mathbb{O} \cup \{o_{n+1}\}$: User can define new object using app. from reference image. Structure can come from ref or defined by user

Object Appearance Variation : Owing to pooling operation in calculating f_i and stochastic nature of generative model we can variations at object level.

Our representations are general can be used with any diffusion model.

PAIR Diffusion

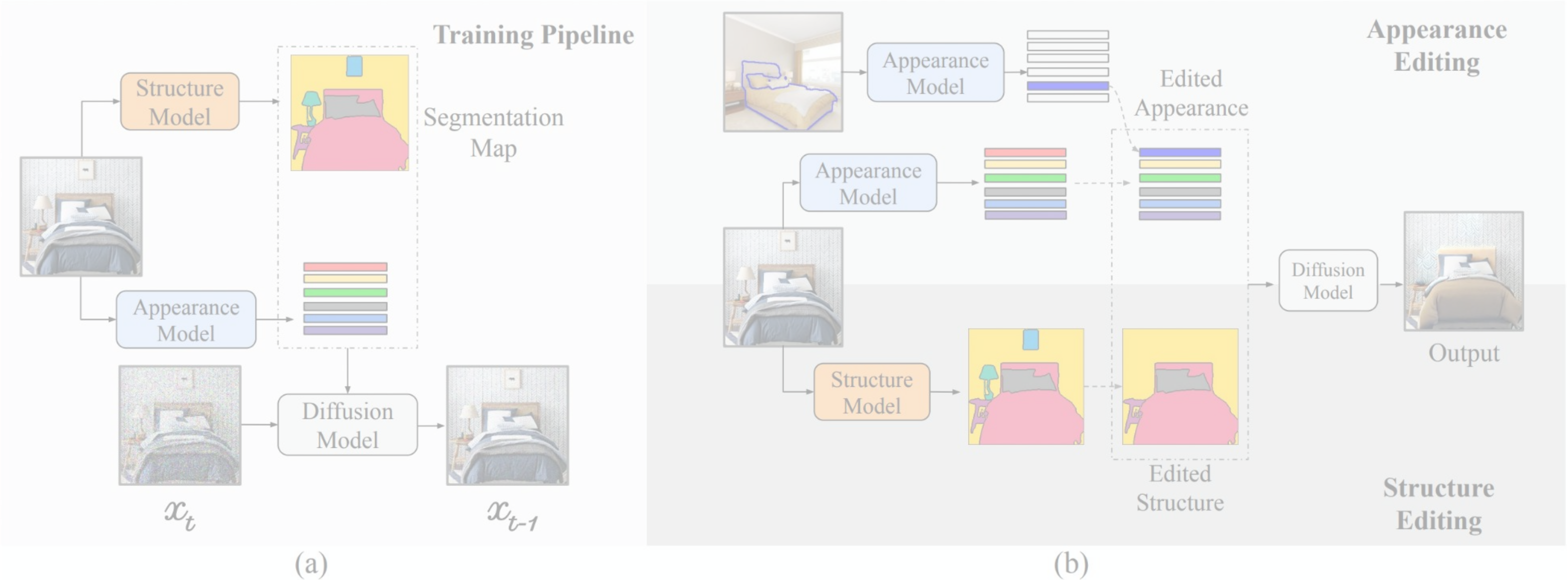


Figure 3: Overview of PAIR diffusion. An image can be seen as a composition of objects each defined by different properties like structure (shape and category), appearance, depth, etc. We focus on controlling structure and appearance. (a) During training, we extract structure and appearance information and train a diffusion model in a conditional manner. (b) At inference, we can perform multiple editing operations, independently controlling the structure and appearance of any real image at the object level.

PAIR Diffusion

Multimodal Classifier-free Guidance

Diffusion models learn the score function which in our case is $\nabla_{z_t} p(z_t|\mathbb{O}, y) = \nabla_{z_t} p(z_t|\mathbf{S}, F, y)$,

We take the case of Stable Diffusion where y is text prompt and we want to edit images using both ref images and prompt

$$\nabla_{z_t} \log p(z_t|\mathbf{S}, F, y) = \nabla_{z_t} \log p(z_t|\mathbf{S}, F) + \nabla_{z_t} \log p(z_t|y) - \nabla_{z_t} \log p(z_t)$$

$$\begin{aligned} \tilde{\epsilon}_\theta(z_t, \mathbf{S}, F, y) &= \epsilon_\theta(z_t, \phi, \phi, \phi) + s_S \cdot (\epsilon_\theta(z_t, \mathbf{S}, \phi, \phi) - \epsilon_\theta(z_t, \phi, \phi, \phi)) \\ &+ s_F \cdot (\epsilon_\theta(z_t, \mathbf{S}, F, \phi) - \epsilon_\theta(z_t, \mathbf{S}, \phi, \phi)) + s_y \cdot (\epsilon_\theta(z_t, \phi, \phi, y) - \epsilon_\theta(z_t, \phi, \phi, \phi)) \end{aligned}$$

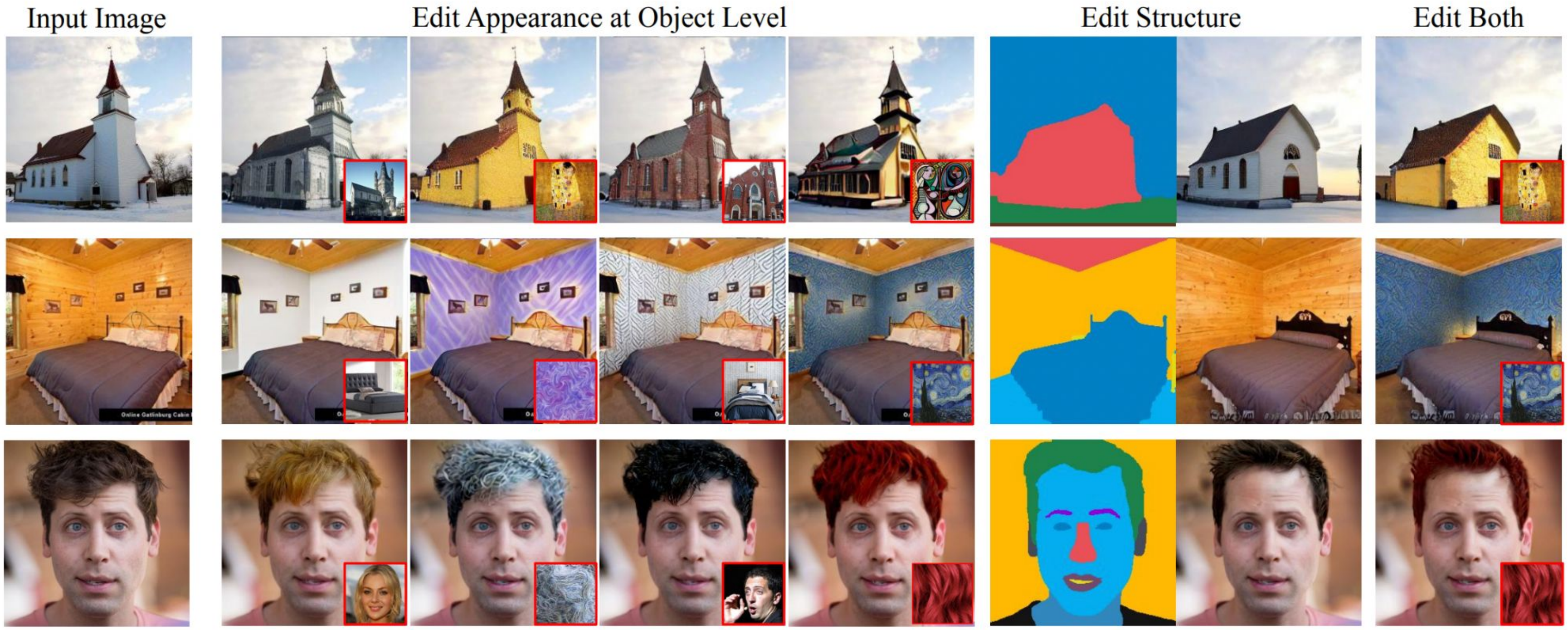
reference image control

text control

Results: Unconditional Models

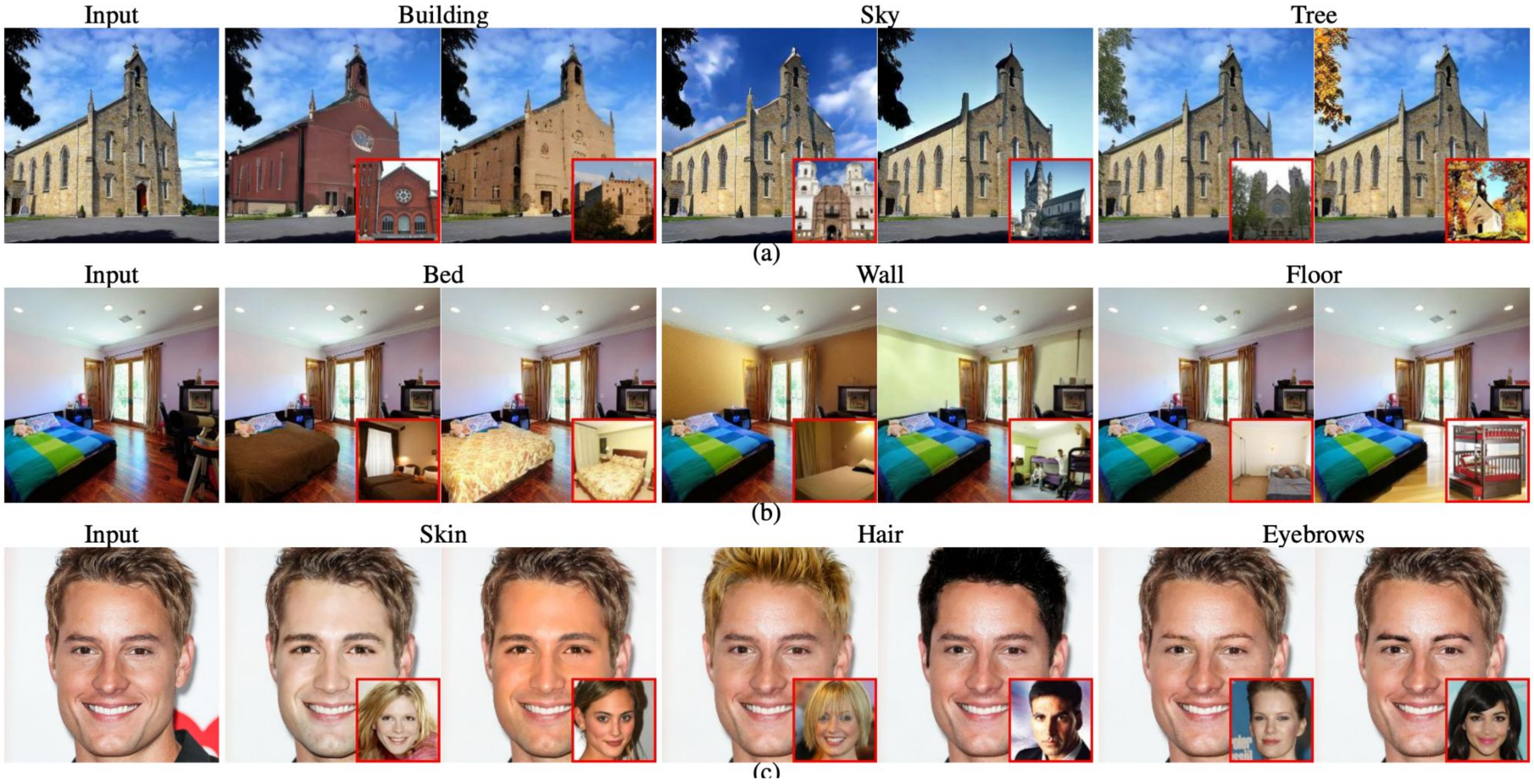
We experimented with both unconditional image generation models and foundation text-2-image model i.e. Stable Diffusion.

- Unconditional Models on LSUN churches, bedroom and faces



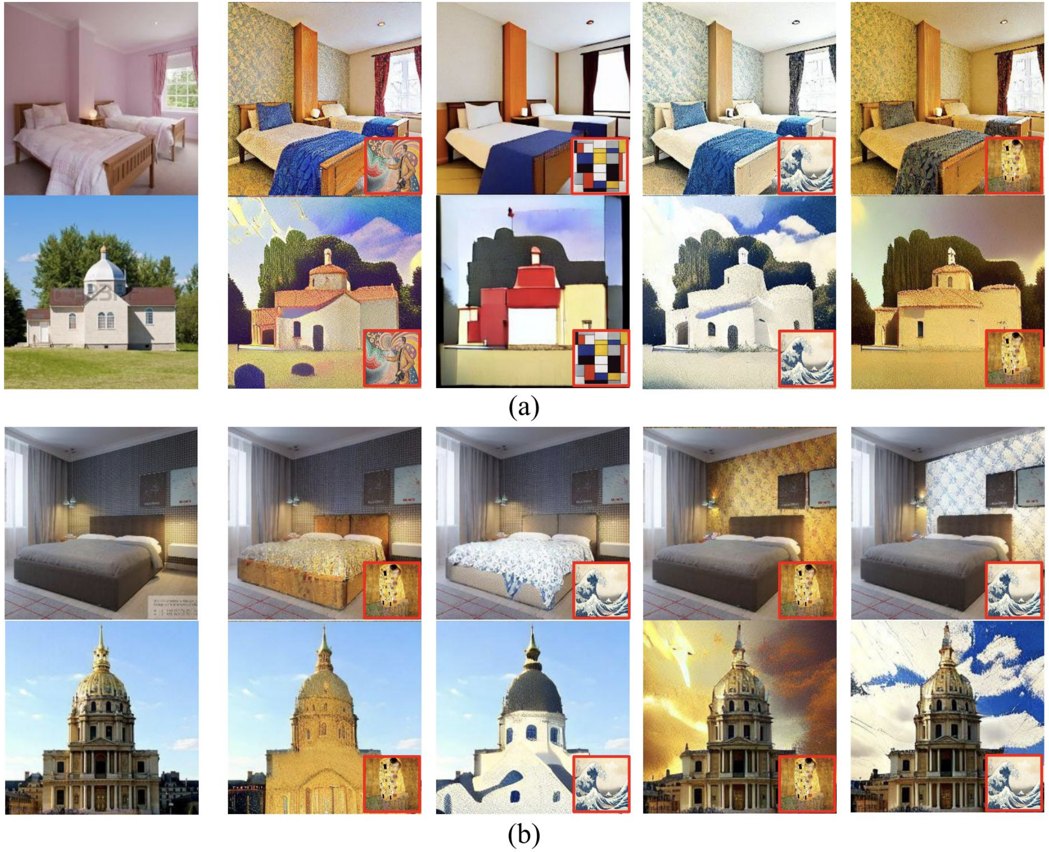
Results: Unconditional Models

Appearance Editing



Results: Unconditional Models

Appearance Editing

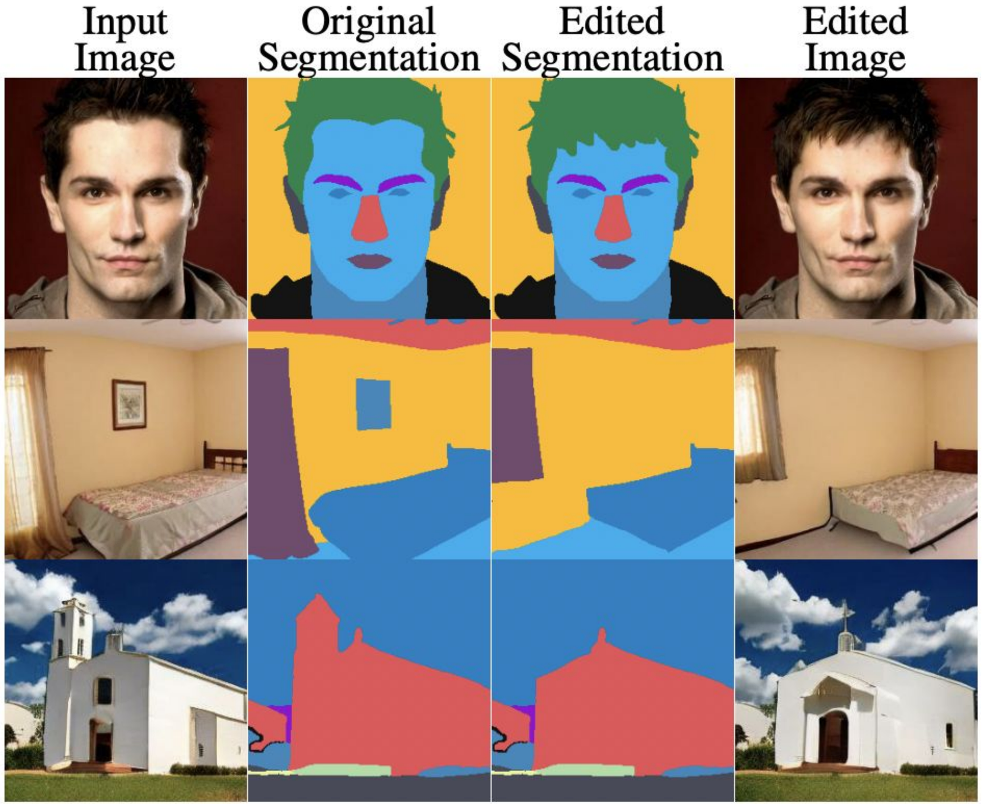


(a)

(b)

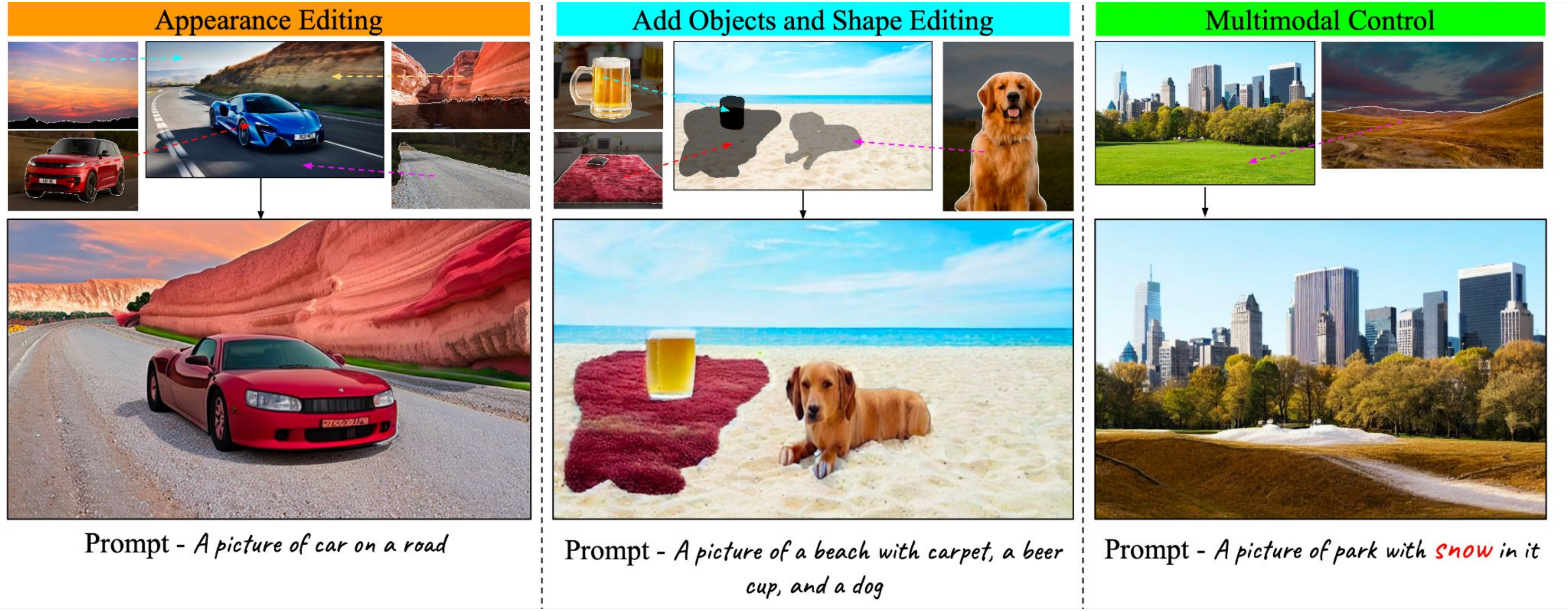
Results: Unconditional Models

Structure Editing



Results: Stable Diffusion

- We used ControlNet to condition the SD because it is data efficient
- We use 118k training images from COCO Stuff dataset to train the model



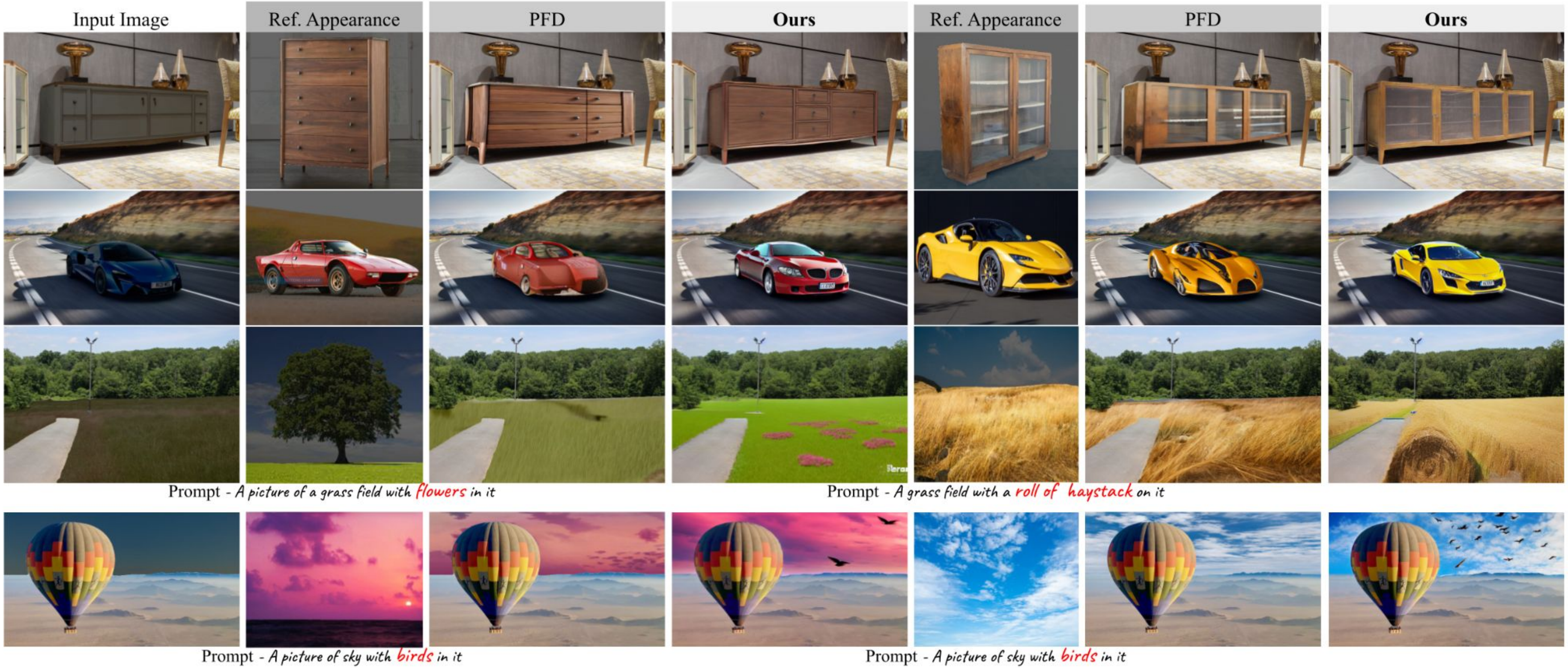
Prompt - *A picture of car on a road*

Prompt - *A picture of a beach with carpet, a beer cup, and a dog*

Prompt - *A picture of park with snow in it*

Results: Stable Diffusion

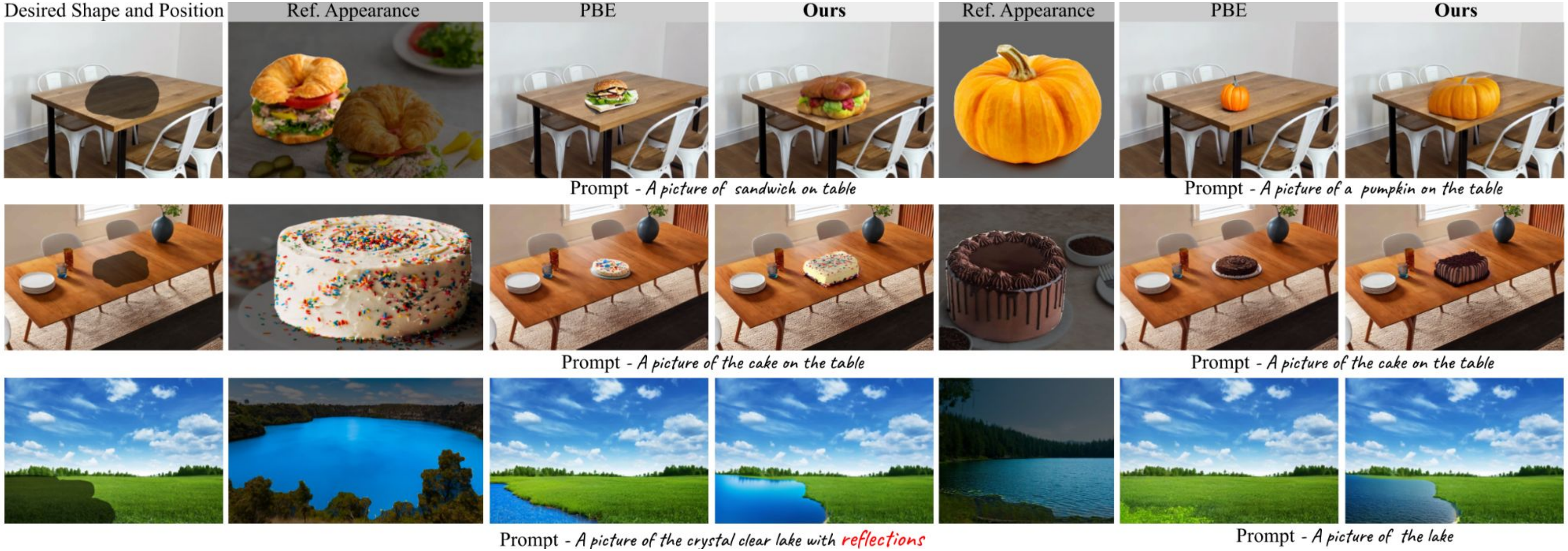
Appearance Editing


Input Image
Ref. Appearance
PFD
Ours
Ref. Appearance
PFD
Ours
Prompt - A picture of a grass field with flowers in it
Prompt - A grass field with a roll of haystack on it
Prompt - A picture of sky with birds in it
Prompt - A picture of sky with birds in it

PFD - Xu, Xingqian, et al. "Prompt-Free Diffusion: Taking" Text" out of Text-to-Image Diffusion Models." CVPR 2024

Results: Stable Diffusion

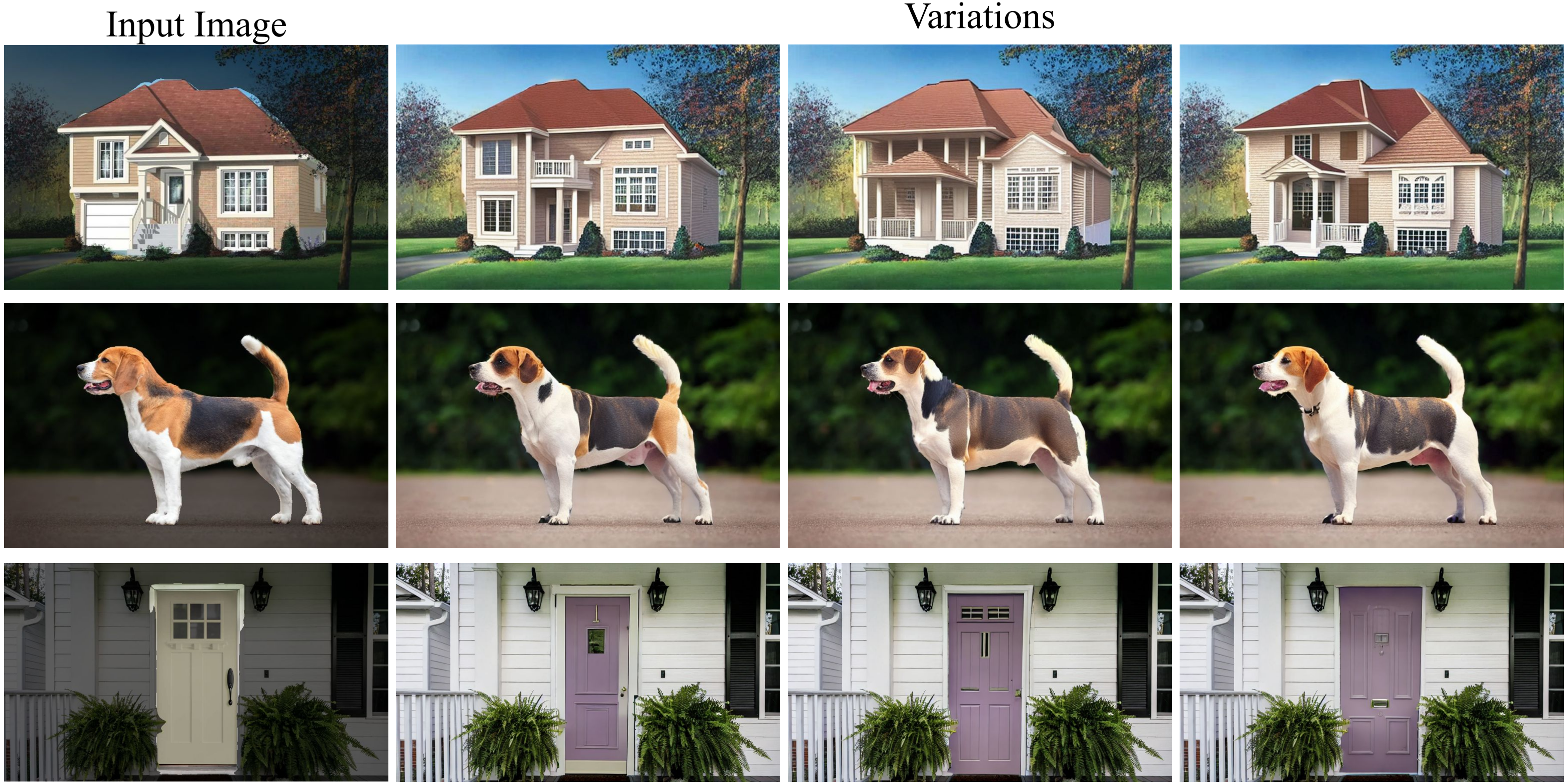
Add Objects and Shape Editing


Desired Shape and Position
Ref. Appearance
PBE
Ours
Ref. Appearance
PBE
Ours
Prompt - A picture of sandwich on table
Prompt - A picture of a pumpkin on the table
Prompt - A picture of the cake on the table
Prompt - A picture of the cake on the table
Prompt - A picture of the crystal clear lake with reflections
Prompt - A picture of the lake

PBE - Yang, Binxin, et al. "Paint by example: Exemplar-based image editing with diffusion models." CVPR 2023.

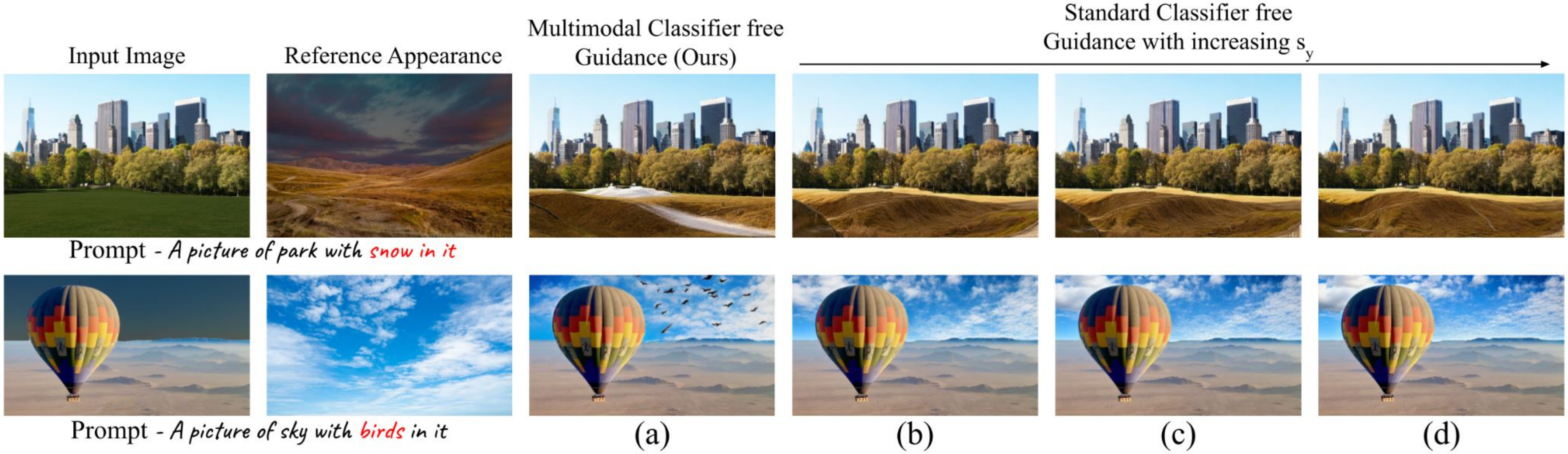
Results: Stable Diffusion

Variations



Prompt - *A picture of a purple door*

Ablation


Input Image
Reference Appearance
Multimodal Classifier free Guidance (Ours)
Standard Classifier free Guidance with increasing s_y
Prompt - A picture of park with snow in it
Prompt - A picture of sky with birds in it
(a)
(b)
(c)
(d)

Standard CFG

$$\tilde{\epsilon_\theta}(z_t, \mathbf{S}, F, y) = \epsilon_\theta(z_t, \phi, \phi, \phi) + s_S \cdot \left(\epsilon_\theta\left(z_t, \mathbf{S}, \phi, \phi\right) - \epsilon_\theta\left(z_t, \phi, \phi, \phi\right)\right) \\ + s_F \cdot \left(\epsilon_\theta\left(z_t, \mathbf{S}, F, \phi\right) - \epsilon_\theta\left(z_t, \mathbf{S}, \phi, \phi\right)\right) + s_y \cdot \left(\epsilon_\theta\left(z_t, \mathbf{S}, F, y\right) - \epsilon_\theta(z_t, \mathbf{S}, F, \phi)\right)$$

Thank you

Paper: https://arxiv.org/pdf/2303.17546
Code: https://github.com/Picsart-AI-Research/PAIR-Diffusion
Project Page: https://vidit98.github.io/publication/conference-paper/pair_diff.html

Related Works

- A line of work uses text to edit images such as Prompt2Prompt[1], InstructPix2Pix [2], DiffEdit [3].
  - Using text we cannot have control over shape
  - Describing appearance using text can be ambiguous
  - No control over each object in the image

- Many works use extra conditioning such binary masks[4], sketches[5], bounding box[6]
  - They do have control over structure
  - However, good appearance control and object level understanding is still missing

- Recently, works such as Composer[7] and T2I[8] adapter can give multiple conditioning at once
  - They do can control structure and appearance of overall image
  - However, object level understanding is missing

[1] Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." *arXiv preprint arXiv:2208.01626* (2022).
[2] Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).
[3] Couairon, Guillaume, et al. "DiffEdit: Diffusion-based semantic image editing with mask guidance." arXiv preprint arXiv:2210.11427 (2022).
[4] Park, Dong Huk, et al. "Shape-Guided Diffusion with Inside-Outside Attention." arXiv preprint arXiv:2212.00210 (2022).
[5] Voynov, Andrey, Kfir Aberman, and Daniel Cohen-Or. "Sketch-Guided Text-to-Image Diffusion Models." arXiv preprint arXiv:2211.13752 (2022).
[6] Yang, Zhengyuan, et al. "ReCo: Region-Controlled Text-to-Image Generation." arXiv preprint arXiv:2211.15518 (2022).
[7] Huang, Lianghua, et al. "Composer: Creative and controllable image synthesis with composable conditions." arXiv preprint arXiv:2302.09778 (2023).
[8] Mou, Chong, et al. "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models." arXiv preprint arXiv:2302.08453 (2023).

Conclusion and Future Work

Conclusion

- In this work we propose PAIR diffusion an object level image editing method.
- We show that object level design can lead to comprehensive editing capabilities.
- We propose spatial classifier free guidance to further improve the independent control over structure and appearance.

Future Work

- The level of details that we want from reference image can be improved.
- Having control over appearance can be applied to 3D or videos where we need to have consistent appearance across different views or time.
- We can have control over more variables such lighting in future.
- Different ways of extracting the appearance information needs to explored.
- Current framework support precise structure editing, adding a coarse structure editing will lead to more comprehensive and complete image editor

Q/A