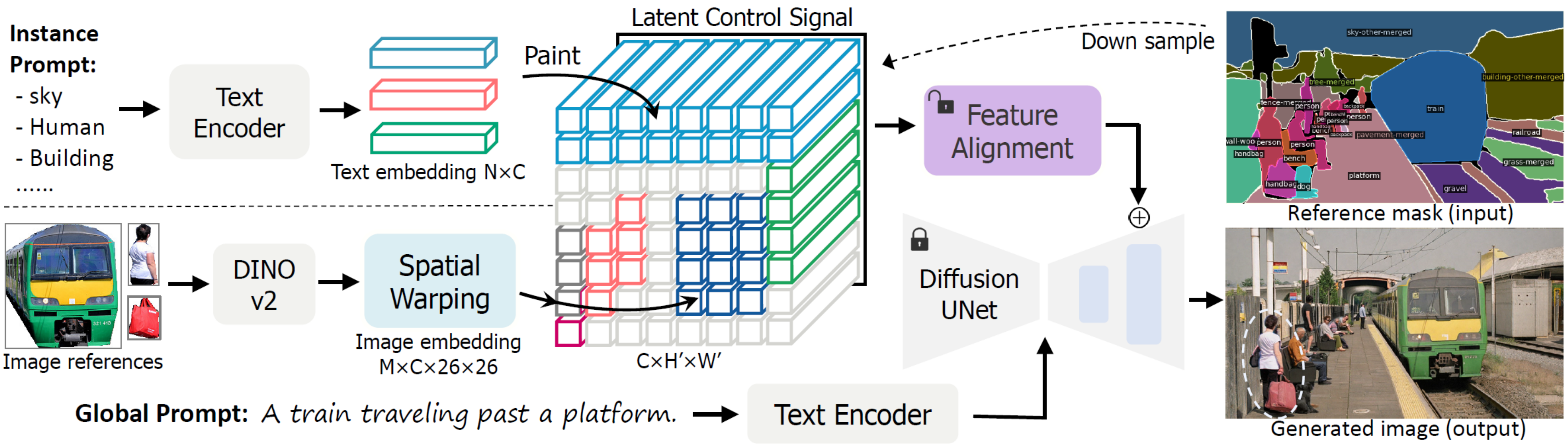
OmniBooth: Learning Latent Control for Image Synthesis with Multi-modal Instruction

Leheng Li[1], Weichao Qiu[3], Xu Yan[3], Jing He[1], Kaiqiang Zhou[3], Yingjie Cai[3], Qing Lian[2], Bingbing Liu[3], Ying-Cong Chen[1,2]

[1]HKUST(GZ), [2]HKUST, [3]HUAWEI Noah's Ark Lab

Open vocabulary image generation

- Input: per point embedding + mask guidance
- Control image: N*C*H*W
- The embedding can be obtained from text of image



Extend RGB condition into latent condition

- ControlNet: 3*H*W
- The condition can be semantic mask, depth map, 3d box map

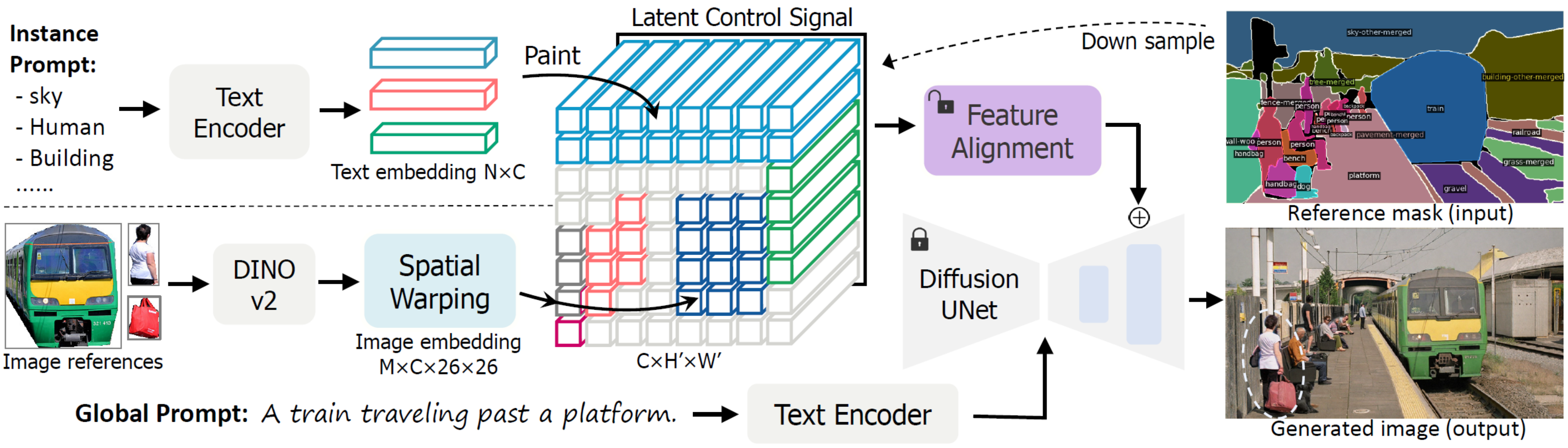
- SyntheOcc: D*H*W
- D=256: number of MPI

- OmniBooth: C*H*W
- C=1024 is the dimension of latent feature

- The latent condition thus contain meaningful input instruction

Open vocabulary image generation

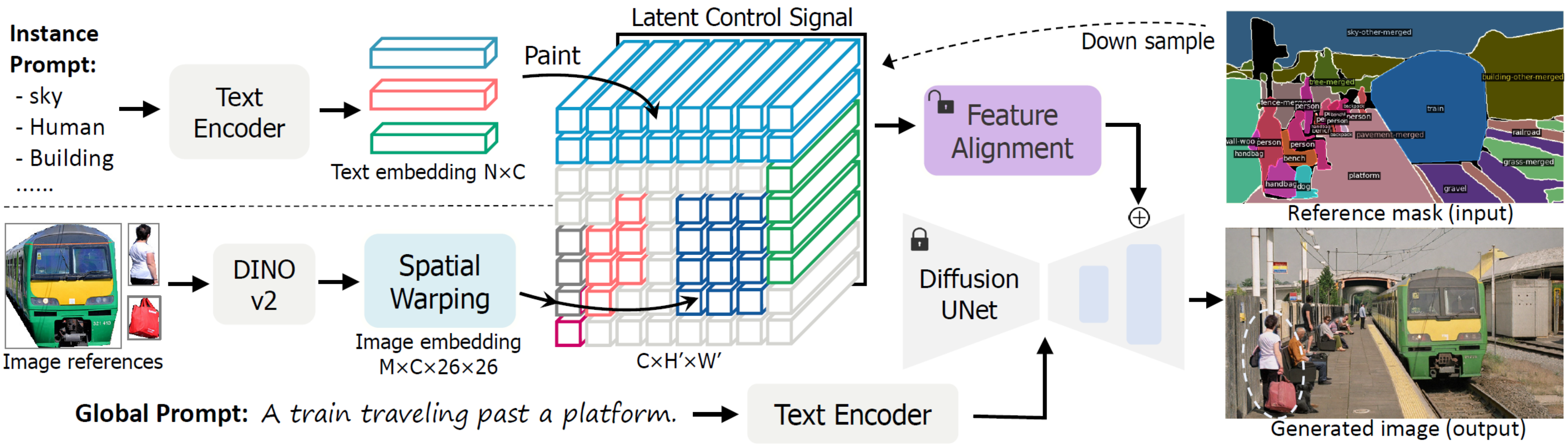
- Input: per point embedding + mask guidance
- Control image: N*C*H*W
- The embedding can be obtained from text of image



Open vocabulary image generation

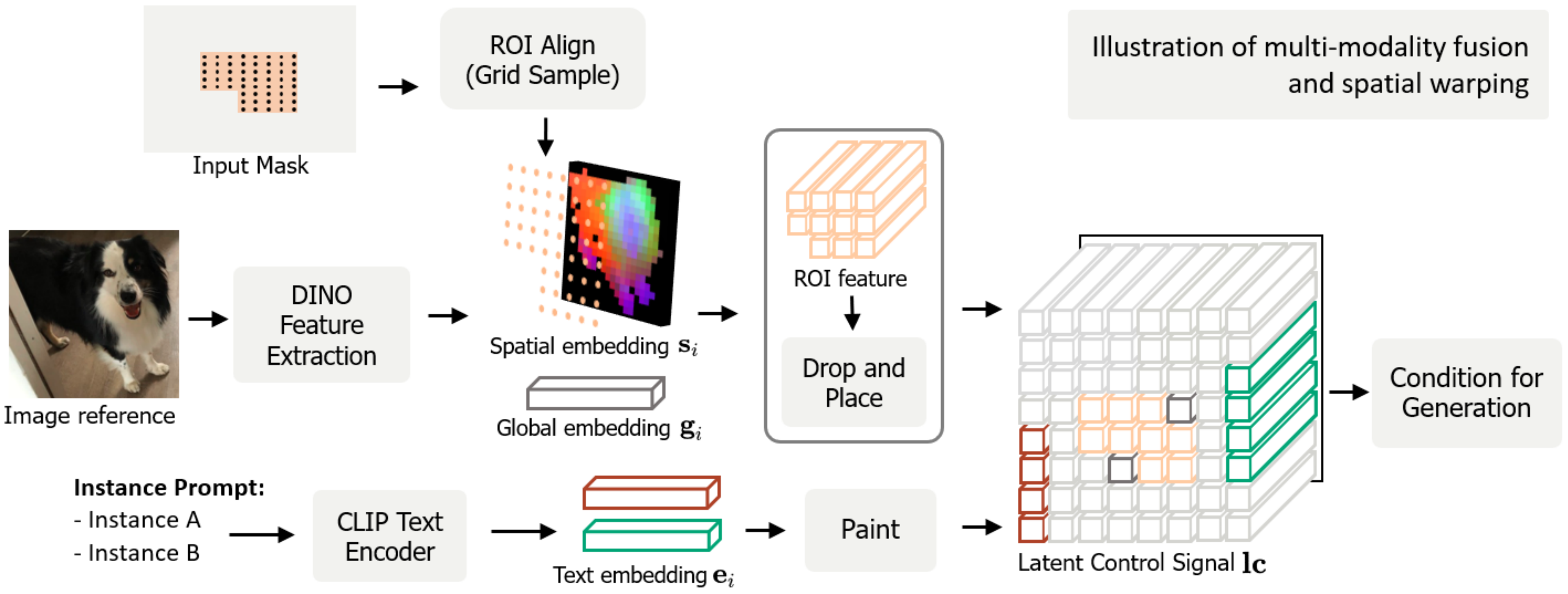
- Draw inspiration from SyntheOcc:
- change the depth dimension of 3D MPI to latent dimension
- Two branch: text condition and image condition
- Control image: N*C*H*W

$$
\begin{aligned}
\text{Instruction:}\quad & \mathbf{s} = (\mathbf{P}, \mathbb{M}, \mathbb{D}), \quad \text{with} \\
\text{Instance masks:}\quad & \mathbb{M} = [\mathbf{M}_1, \cdots, \mathbf{M}_N], \\
\text{Descriptions:}\quad & \mathbb{D} = [(\mathbf{T}_1 \textit{ or } \mathbf{I}_1), \cdots, (\mathbf{T}_N \textit{ or } \mathbf{I}_N)],
\end{aligned}
$$



Spatial warping

- Motivation: inject 2D spatial feature for condition, rather than 1D embedding
- First DINOv2 extract spatial feature, then warping it to latent control signal



Results

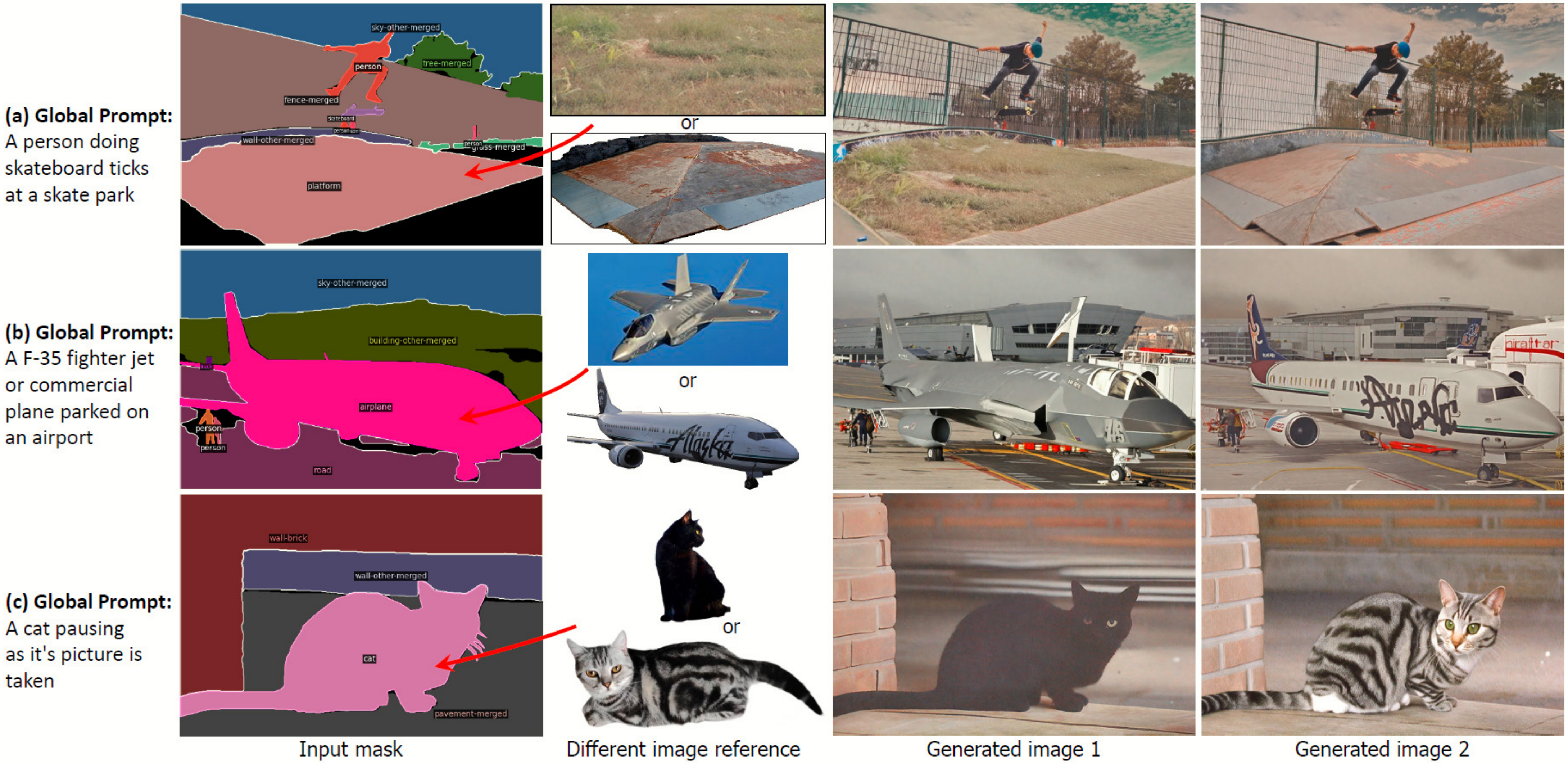
(a) Global Prompt:
A young man doing a flip on a skateboard in a busy street.
Instance Prompt:
- a skateboard
- a person
......

(b) Global Prompt:
A plate topped with a piece of cake.
Instance Prompt:
- a silver fork
- a piece of cake with frosting

(c) Global Prompt:
A woman stands in the dining area at the table.
Instance Prompt:
- a wooden floor
- a dining table
- a red vase
......

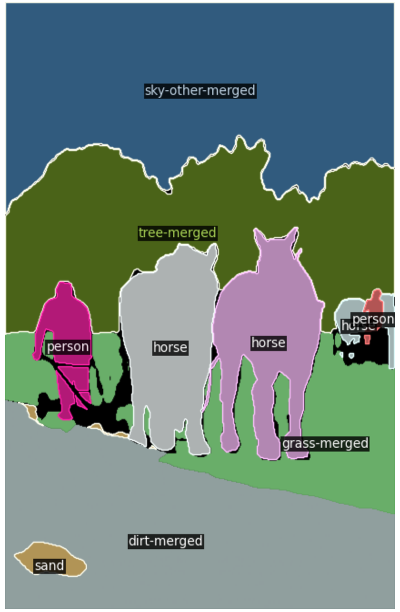
Language instruction | Input mask | InstanceDiffusion | Ours | Ground Truth

Results



(a) Global Prompt:
A person doing skateboard ticks at a skate park
sky-other-merged
tree-merged
person
fence-merged
skateboard
wall-other-merged
grass-merged
platform
or
(b) Global Prompt:
A F-35 fighter jet or commercial plane parked on an airport
sky-other-merged
building-other-merged
airplane
person
road
or
(c) Global Prompt:
A cat pausing as it's picture is taken
wall-brick
wall-other-merged
cat
pavement-merged
or
Input mask
Different image reference
Generated image 1
Generated image 2

Instance-level manipulation

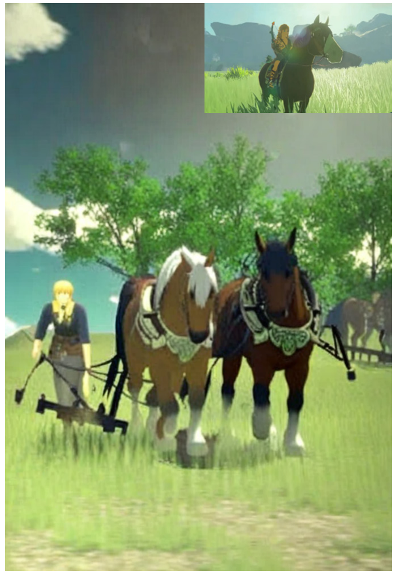


(a) Input mask

(b) Our generation

(c) Horse→Sheep

(d) Horse→Dog



(e) Style of Zelda using IP-Adapter

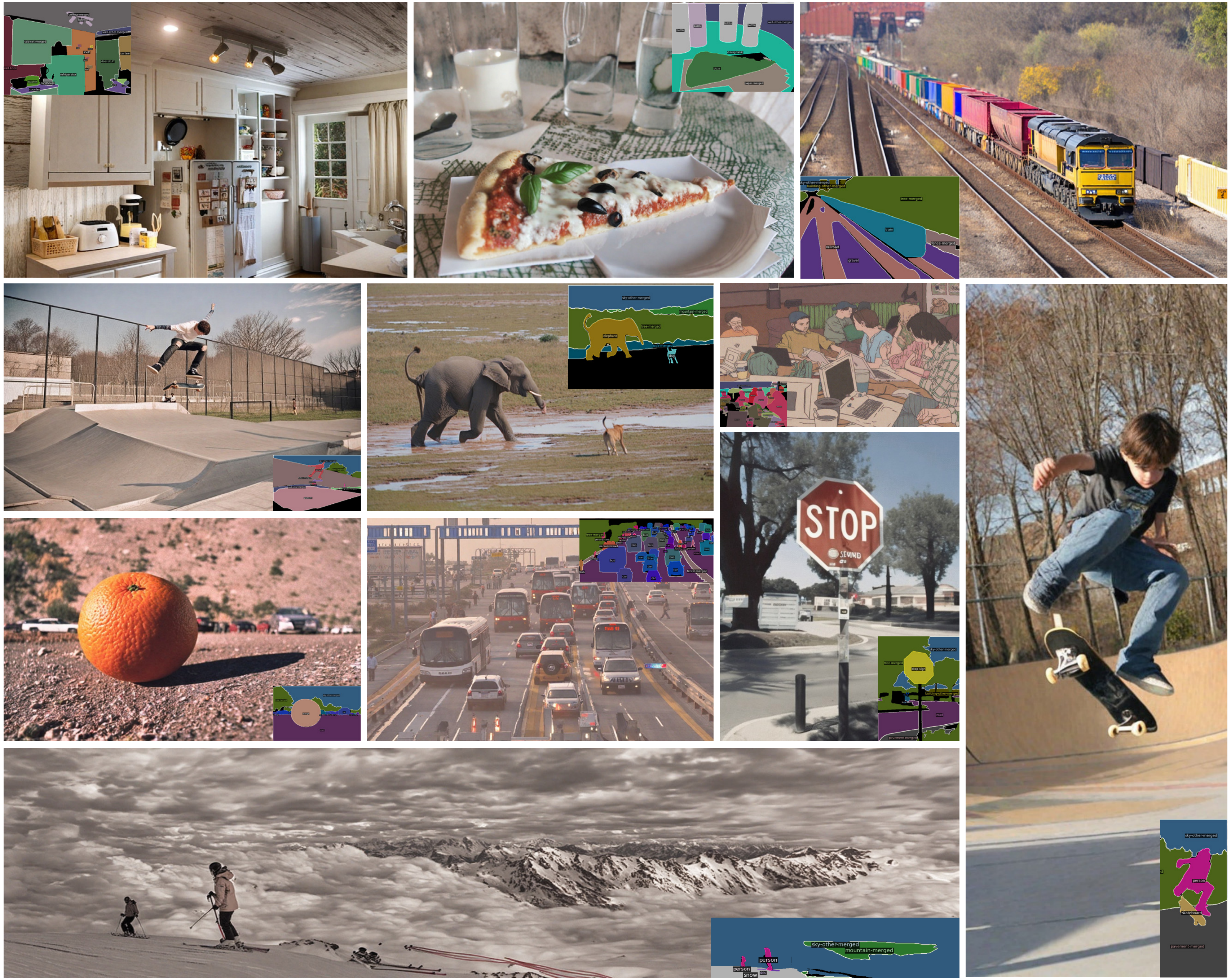
(f) Ground Truth

Global Prompt:

Two draft horses pulling plow, under cloudy skies with trees and other horses in background.

Instance Prompt:

- a horse with harness


STOP
sky-other-merged
mountain-merged
person
person
snow

Evaluation

Methods	Type	DINO	CLIP-I	CLIP-T
Real Images	-	0.774	0.885	-
Textual Inversion (Gal et al., 2022)	Fine-Tune	0.569	0.780	0.255
DreamBooth (Ruiz et al., 2023)	Fine-Tune	0.668	0.803	0.305
ELITE (Wei et al., 2023)	Zero-Shot	0.621	0.771	0.293
BLIP-Diffusion (Li et al., 2024a)	Zero-Shot	0.594	0.779	0.300
Subject-Diffusion (Ma et al., 2023)	Zero-Shot	0.711	**0.787**	0.293
OmniBooth	Zero-Shot	**0.736**	0.776	**0.310**

- Dataset: Instance segmentation in COCO dataset
- Generate images of val-set using its mask annotation, then use perception network to inference

Method	COCO Instance Segmentation							FID
	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}	AP^{mask}_{small}	AP^{mask}_{large}	AR^{mask}_{1}	AR^{mask}_{100}	
Oracle (YOLOv8)	40.8	63.5	43.6	21.9	58.2	32.9	56.0	-
SpaText (Avrahami et al., 2023)	5.3	12.1	5.8	3.1	11.2	10.7	14.2	23.1
ControlNet (Zhang et al., 2023)	6.5	13.8	6.1	3.6	12.5	12.9	15.1	20.3
InstanceDiff. (Wang et al., 2024b)	26.4	**48.4**	25.3	4.7	**47.0**	24.1	37.7	23.9
OmniBooth	**28.0**	46.7	**29.1**	**10.0**	46.7	**25.1**	**41.0**	**17.8**

Table 1: Downstream evaluation on the **MS COCO** val2017 set. We report YOLO score and FID to evaluate the alignment accuracy and image quality of our method.