Pretraining is All You Need for Image-to-Image Translation

Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang,
Dong Chen, Qifeng Chen, Fang Wen

NIPS'22 Under Review

2022.08.08 윤주열

Image-to-Image Translation

Changing the domain of an image.

e.g., SYNTHIA-to-Cityscapes, Label-to-Image, Sketch-to-Image, …

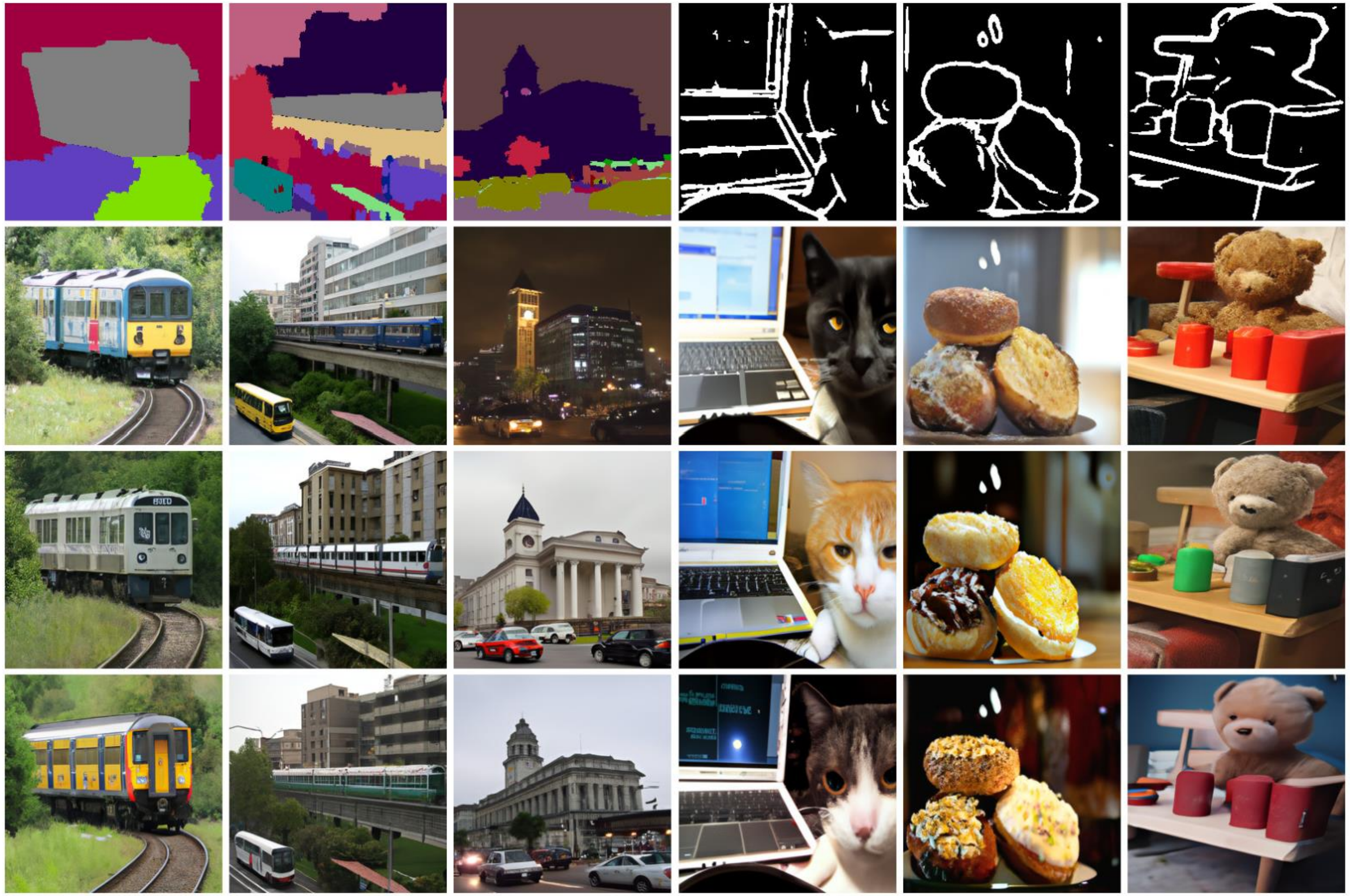


Figure 1: Diverse images sampled by our method given semantic layouts or sketches.

Pretrained Generative Models

Similar to the pretraining-finetuning paradigm of discriminative tasks (classification, semantic segmentation, and object detection),

Generative tasks often leverage a pretrained StyleGAN2 for downstream tasks (editing and synthesis).

We can utilize even more powerful generative models based on diffusion.

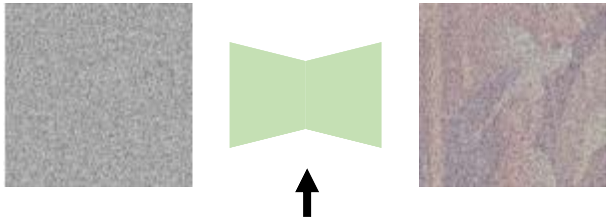
Since we are targeting a conditional generation task, we utilize a pretrained text-to-image model (GLIDE).

GLIDE

Text-to-Image Generation model by openAI (trained on the DALL-E dataset).

Consists of a low-resolution diffusion model, and a upsampling diffusion model.

<Base Model>



“A portrait of Lena”

<Upsample Model>

“A portrait of Lena”

“a hedgehog using a calculator” “a corgi wearing a red bowtie and a purple party hat” “robots meditating in a vipassana retreat” “a fall landscape with a small cottage next to a lake”

“a surrealist dream-like oil painting by salvador dalí of a cat playing checkers” “a professional photo of a sunset behind the grand canyon” “a high-quality oil painting of a psychedelic hamster dragon” “an illustration of albert einstein wearing a superhero costume”

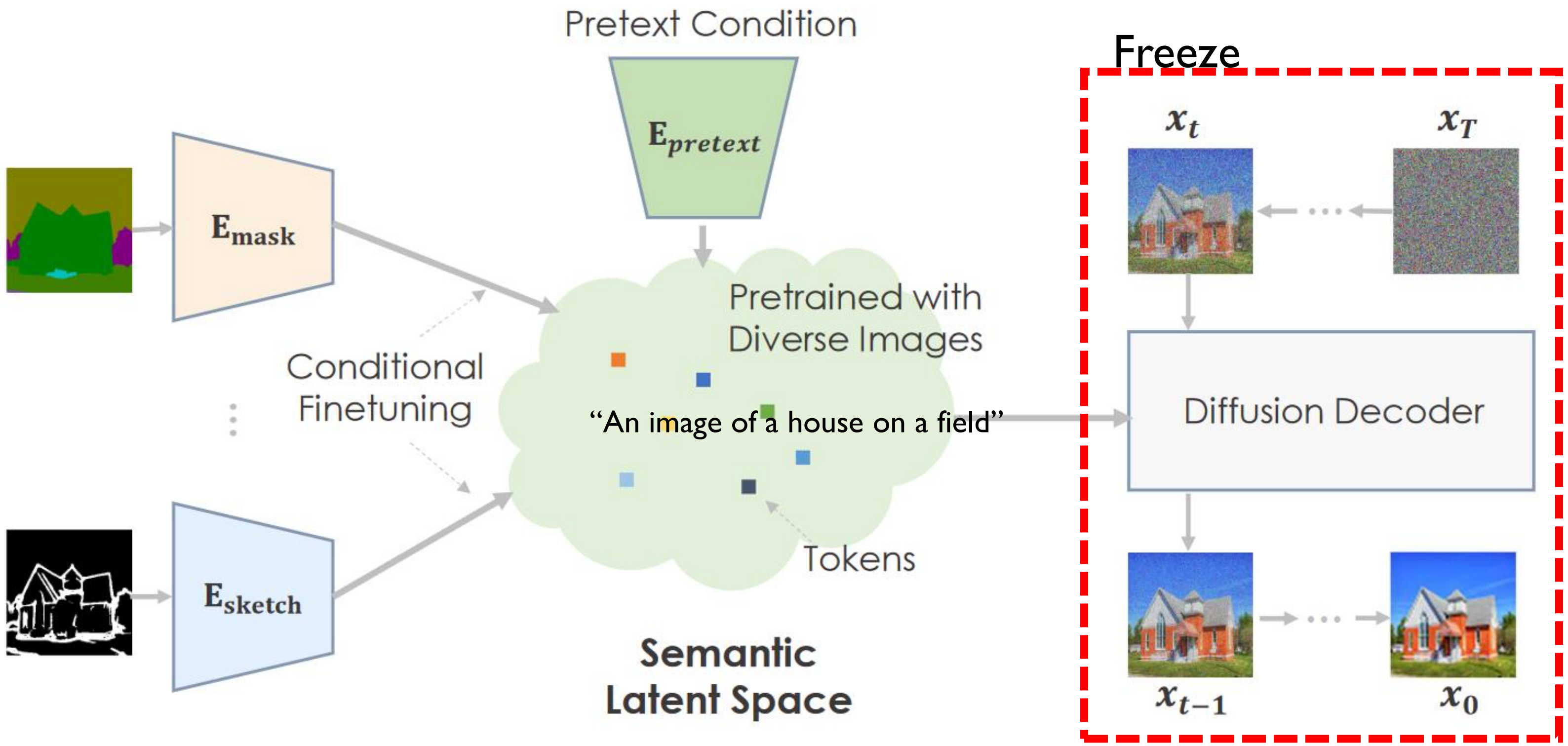
PITI

Replace the original text condition with a new condition by introducing a task-specific encoder.

Adopts a two-stage finetuning scheme.

First Stage: Freeze base model of GLIDE, and train condition encoder.

Second Stage: Jointly finetune the entire pipeline.



PITI

Finetune the **upsampler** with additional loss functions and severe augmentation to prevent the model from producing overly smooth images.

Add various noise and blur functions to the input (e.g., $\{\mathbf{B}_{\mathtt{iso}}, \mathbf{B}_{\mathtt{aniso}}, \mathbf{N}_{\mathrm{G}}, \mathbf{N}_{\mathrm{JPEG}}, \mathbf{N}_{\mathrm{S}}\}$).

Add a perceptual loss and an adversarial loss function.

$$\mathcal{L}_{\text{perc}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}} \|\boldsymbol{\psi}_m(\hat{\boldsymbol{x}}_0^t) - \boldsymbol{\psi}_m(\boldsymbol{x}_0)\|, \quad (4)$$

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\epsilon}} \left[\log D_\theta(\hat{\boldsymbol{x}}_0^t)\right] + \mathbb{E}_{\boldsymbol{x}_0} \left[\log(1 - D_\theta(\boldsymbol{x}_0))\right], \quad (5)$$

Normalized Classifier-Free Guidance (during sampling)

Classifier-Free Guidance leads to mean an variance shifts and causes overly saturated and overly smooth images.

Normalize the output mean and variance.

$$\hat{\boldsymbol{\epsilon}}_\theta\left(\boldsymbol{x}_t|\boldsymbol{y}\right) = \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t|\boldsymbol{y}\right) + w \cdot \left(\boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t|\boldsymbol{y}\right) - \boldsymbol{\epsilon}_\theta\left(\boldsymbol{x}_t|\boldsymbol{\emptyset}\right)\right)$$

$$\tilde{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t|\boldsymbol{y}) = \frac{\sigma}{\hat{\sigma}}(\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t|\boldsymbol{y}) - \hat{\mu}) + \mu.$$

Results

Quantitative results

Table 5: Quantitative comparison on diverse image translation tasks.

	ADE20K		COCO (Mask)		Flickr (Mask)		COCO (Sketch)		Flickr (Sketch)		DIODE	
Method	FID-I	FID-C	FID-I	FID-C	FID-I	FID-C	FID-I	FID-C	FID-I	FID-C	FID-I	FID-C
Pix2PixHD	61.8	35.3	67.7	37.5	41.5	26.1	38.7	27.1	26.9	16.8	66.0	18.2
SPADE	33.9	18.9	22.6	15.0	27.7	17.4	89.2	48.9	43.6	29.5	61.2	17.0
OASIS	28.3	14.8	17.0	8.8	24.4	10.5	-	-	-	-	-	-
Ours (from scratch)	35.7	16.3	25.1	13.0	26.9	10.6	33.6	13.0	24.8	9.4	70.2	13.9
Ours	**27.3**	**8.9**	**15.8**	**5.2**	**21.2**	**6.1**	**21.4**	**8.8**	**20.3**	**6.0**	**59.6**	**11.5**

Table 3: Ablation study of the proposed PITI on ADE20K dataset.

(a) Finetune strategy.

Finetune strategy	FID
Fixed decoder	12.6
One-stage finetune	13.3
Two-stage finetune	8.9

(b) Upsampling strategy.

Degradation	$\mathcal{L}_{\text{perceptual}}$	$\mathcal{L}_{\text{adversarial}}$	FID
			14.5
✓			12.1
✓	✓		9.8
✓	✓	✓	8.9

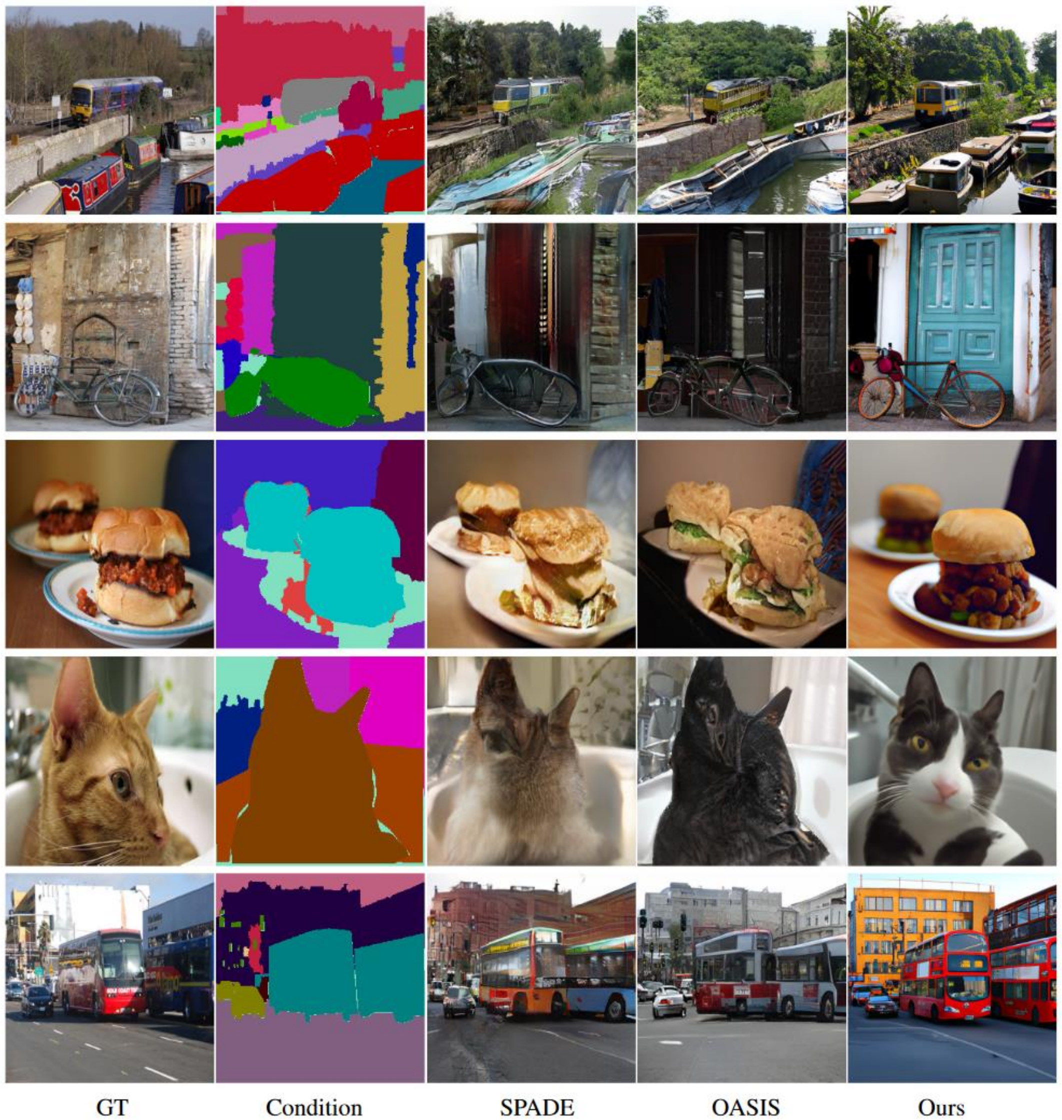
Results

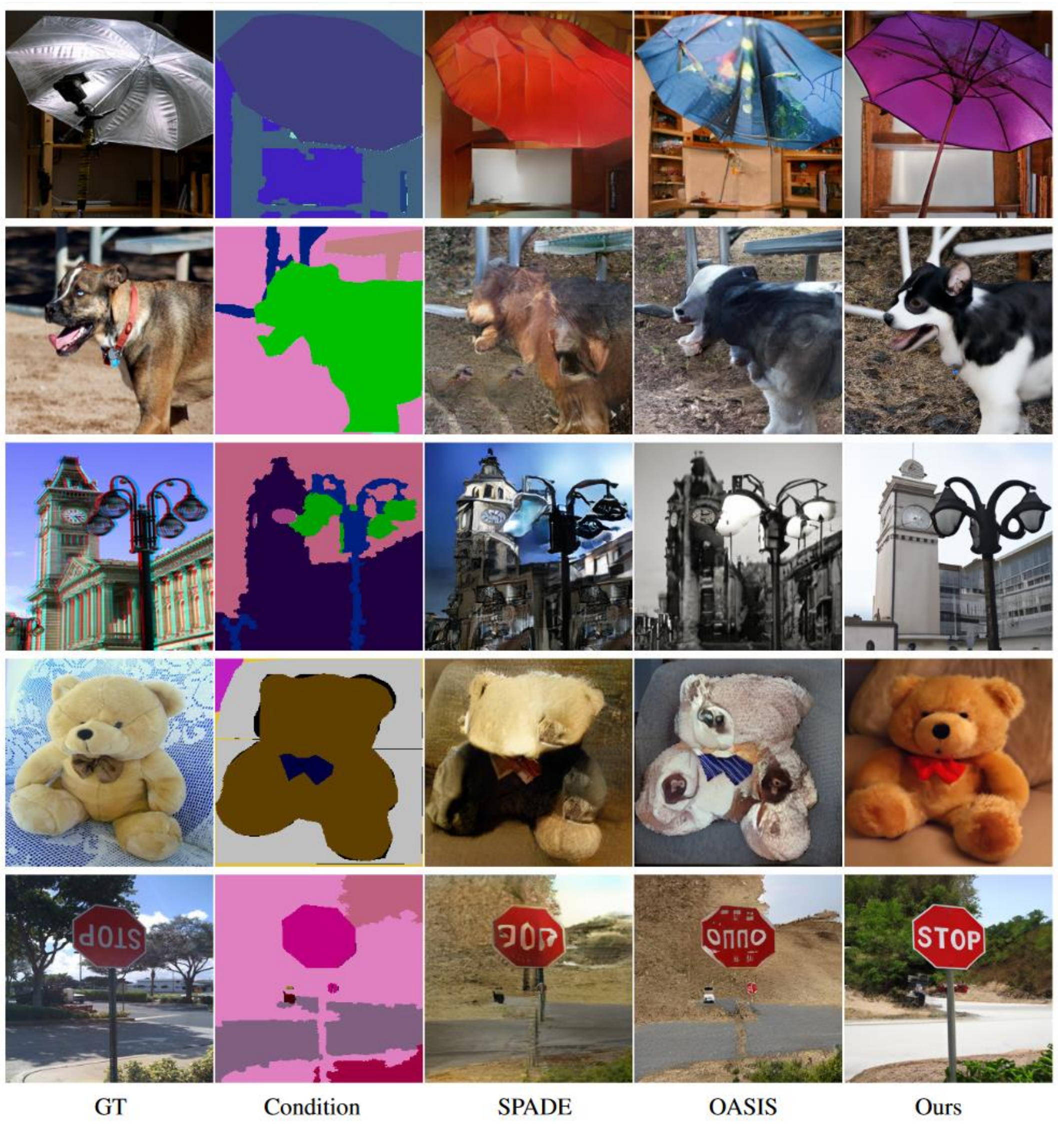
Qualitative results

GT Condition SPADE OASIS Ours (Scratch) Ours

Results

Qualitative results





Ablation Study

Effect of two-stage training

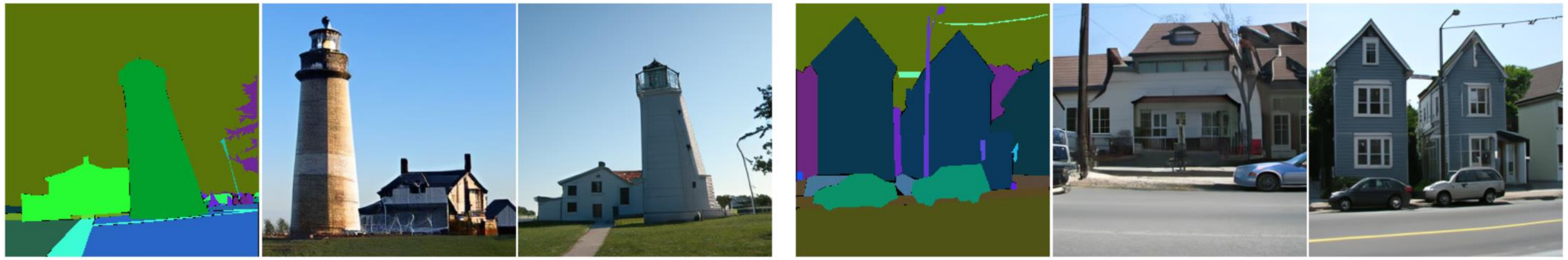
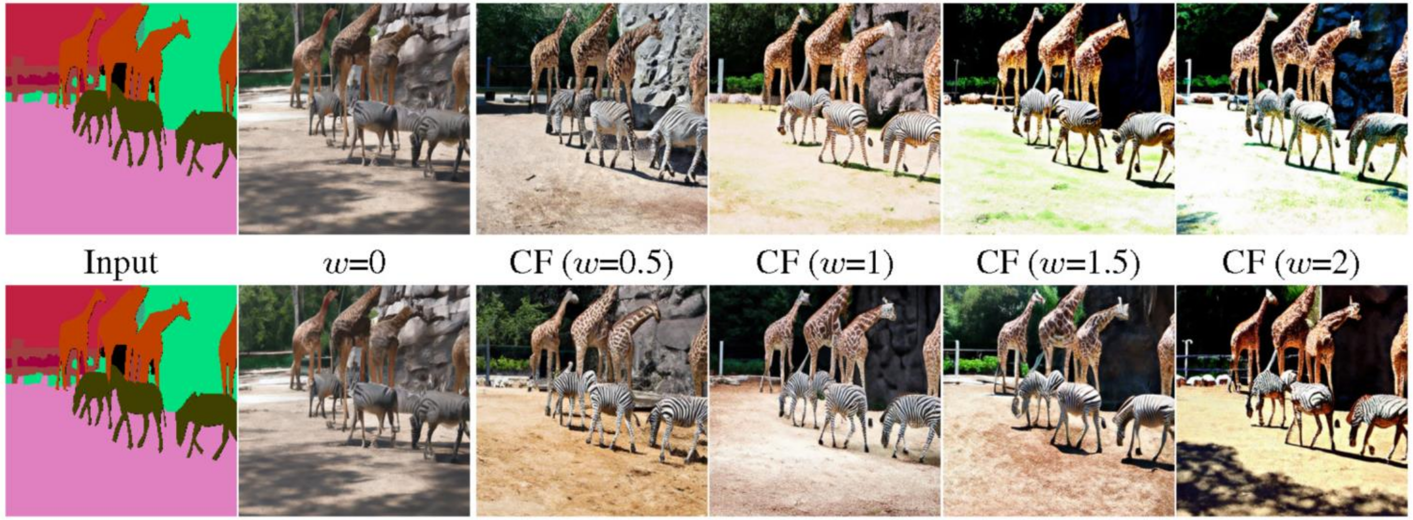


Figure 5: Fixing the decoder generates high-quality images but fails to align with the condition.

Normalized Classifier-Free Guidance



Limitations

Intra-image correlation and misalignment in small regions.