Controllable image generation

Marlene Careil

Image generation 2014-2023

Image editing

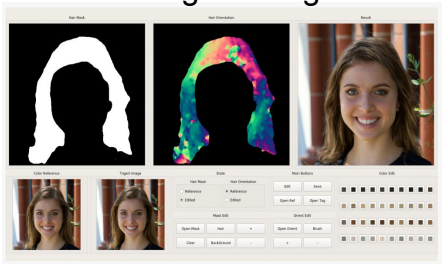
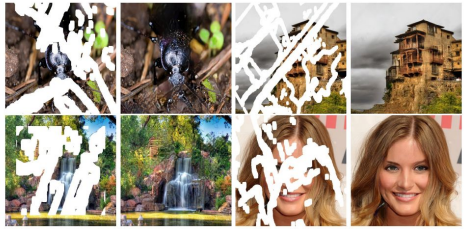


Image In/out-painting



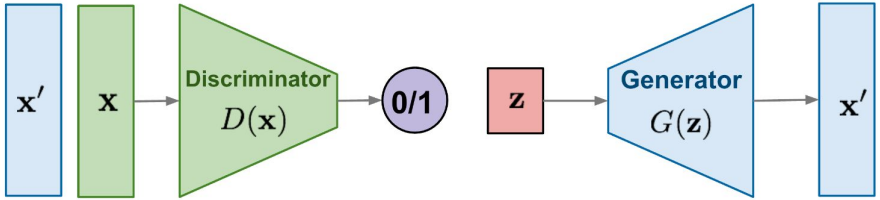
Virtual try-on

Text-to-image

Images from: Image Inpainting for Irregular Holes Using Partial Convolutions, Liu et al., ECCV 2018
Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization, Huang and Belongie, ICCV 2017
VITON: An Image-based Virtual Try-on Network, Han et al., CVPR 2018
You only need adversarial supervision for semantic image synthesis, Sushko et al., ICLR 2021
High-Resolution Image Synthesis With Latent Diffusion Models, Rombach et al, CVPR 2022

Generative models

GAN: Adversarial training



Diffusion models: Gradually add Gaussian noise and then reverse

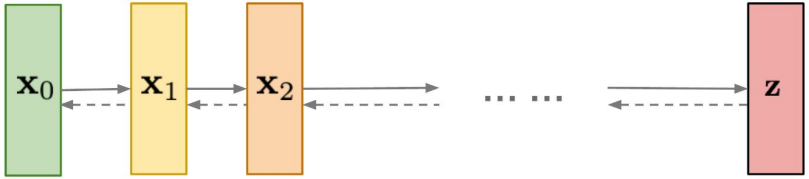
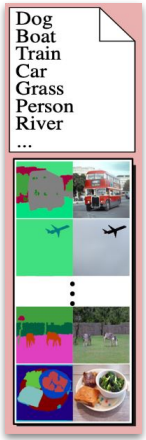


Figure from https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

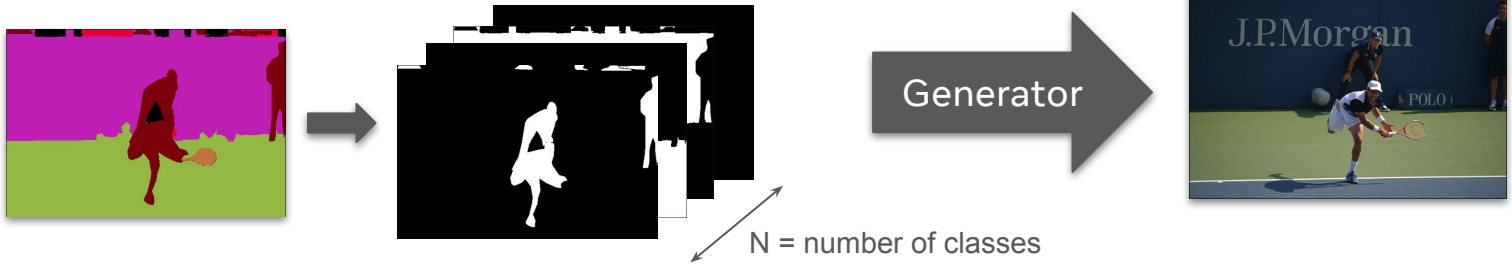
Task: Semantic Image Synthesis

Training dataset



Segmentation map

One-hot encoding

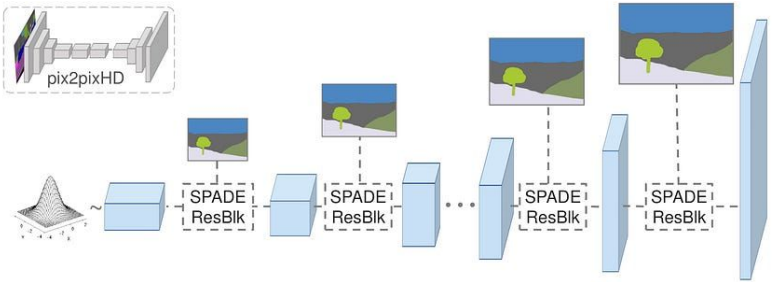


- Each pixel is described by a class.
- Trained on an annotated dataset containing a finite set of classes.

GANs and Diffusion models for semantic image synthesis

GANs

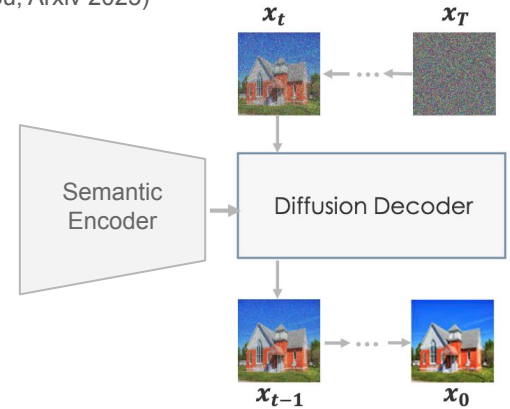
- **Pix2pix** (P.Isola et al., CVPR 2017), **SPADE** (T. Park, CVPR 2019), **OASIS** (E. Schonfeld, ICLR 2021)
- **DP-SIMS** (T. Berrada, CVPR 2024)



SPADE

Diffusion models

- **PITI** (T. Wang, Arxiv 2022), **SDM** (W. Wang, Arxiv 2022)
- **ControlNet** (L. Zhang, ICCV 2023), **T2I-Adapter** (C. Mou, Arxiv 2023)



PITI

Trained on large-scale annotated datasets such as COCO (TY Lin, 2014) containing 110k images or ADE20K (20k images).

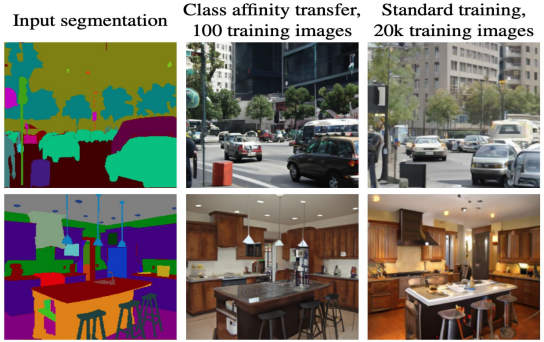
Motivations for few-shot learning

- Very costly to annotate images with segmentation maps
- At the time of this project, little work on few-shot transfer for class-conditional generation

Objective
Develop a few-shot transfer method for semantic image synthesis

Motivations for few-shot learning

- Very **costly to annotate** images with segmentation maps
- **Related works:** Many works on few shot learning for classification, and unconditional or class-level condition generation but none for semantic synthesis



Our diffusion-based transfer results using training set of 100 ADE20K images (2nd col.) compared to the same model trained from scratch on full dataset (20k images, 3rd col.)

Main idea: Start from a pre-trained semantic synthesis model.

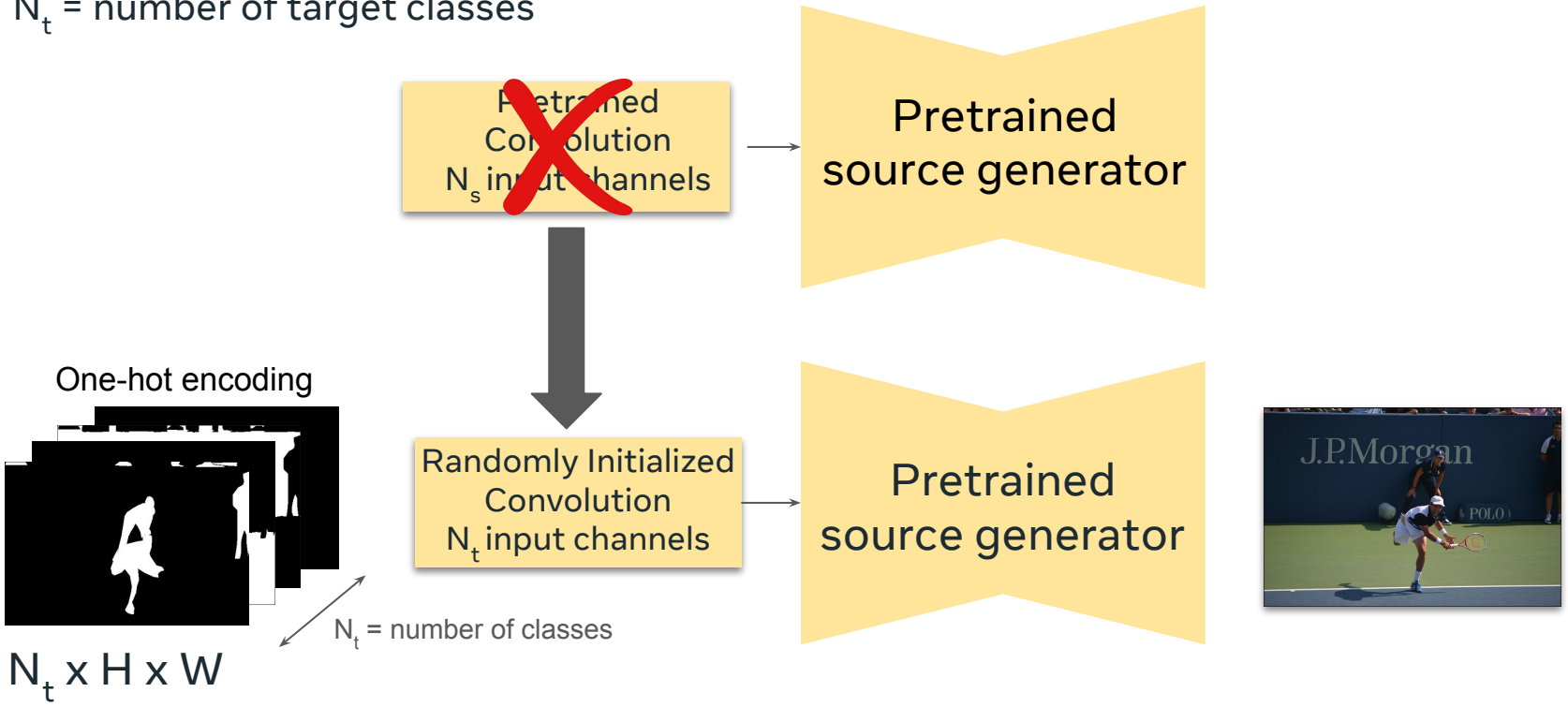
Few-shot Semantic Image Synthesis with Class Affinity Transfer

Marlène Careil[1,2] Jakob Verbeek[2] Stéphane Lathuilière[1]
[1]LTCI, Télécom Paris, IP Paris [2]Meta AI

Figure 1. **Can we train a semantic image synthesis model from only 100 images?** Our diffusion-based transfer results using training set of 100 ADE20K images (2nd col.) compared to the same model trained from scratch on full dataset (20k images, 3rd col.).

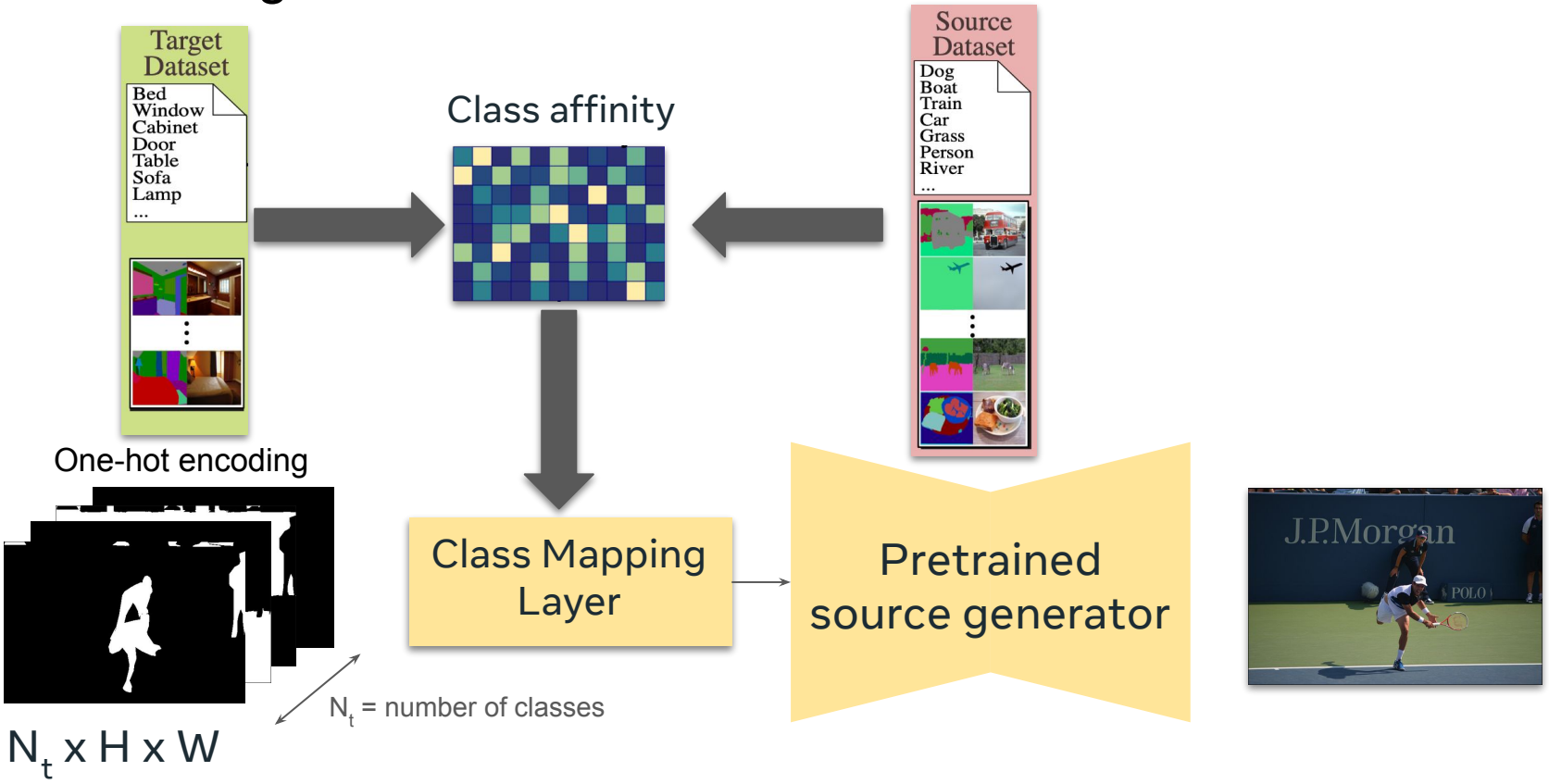
Transfer learning for Semantic Image Synthesis

N_s = number of source classes
N_t = number of target classes

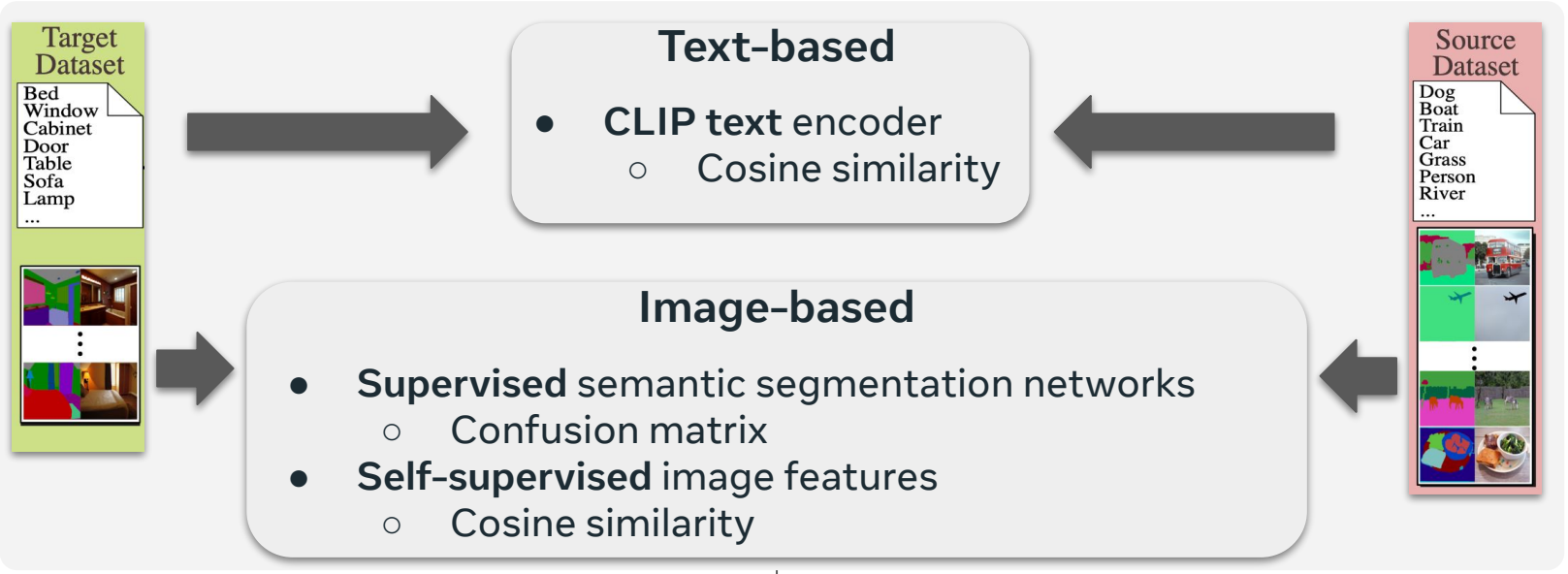


Can we do better than random initialization?

Better weight initialization

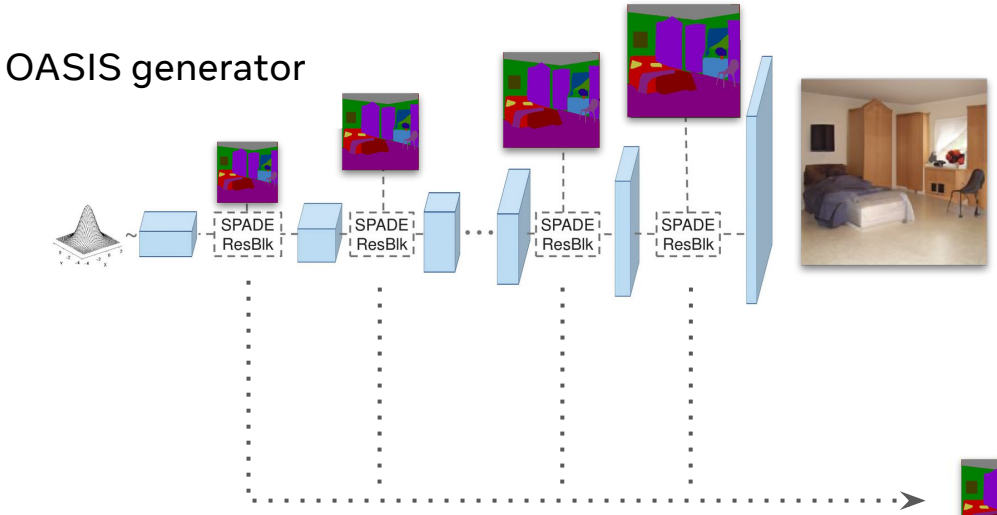


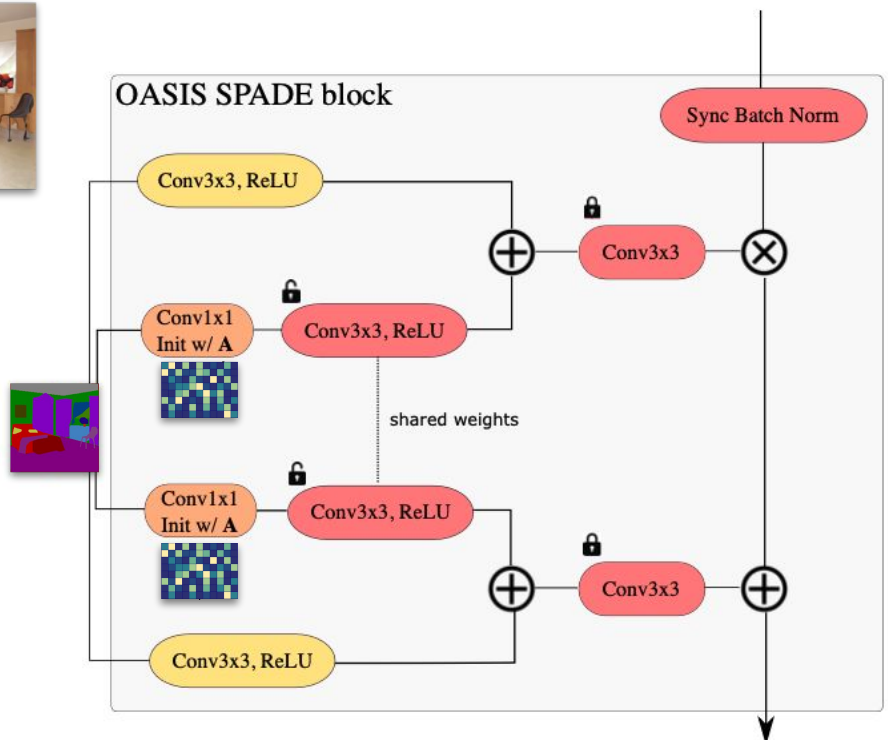
Class Affinity Estimation



Combination via majority voting : aggregate all the three methods

Class Affinity Transfer (CAT) with GANs

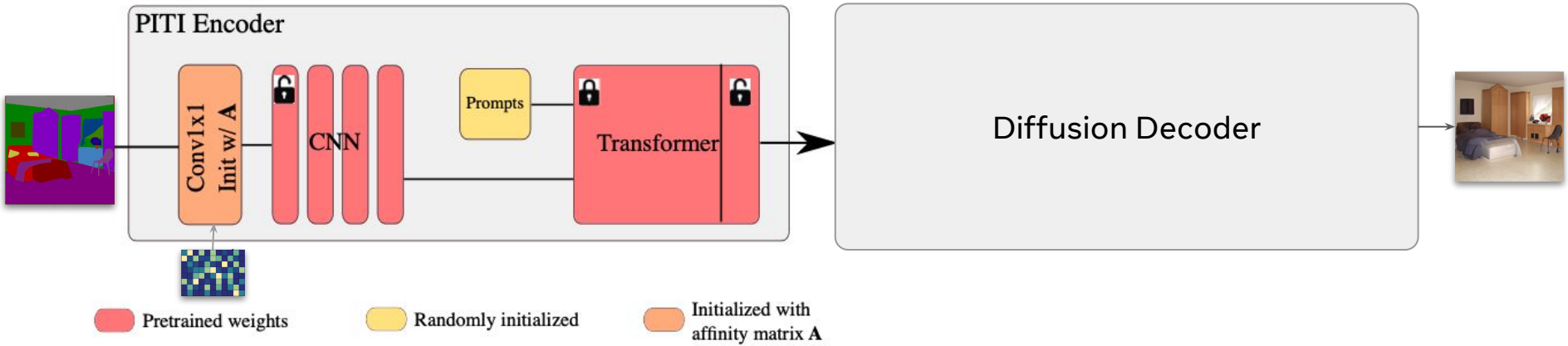




- Based on GAN-based OASIS model

OASIS: You only need adversarial supervision for semantic image synthesis. Edgar Schönfeld et al., ICLR 2021

Class Affinity Transfer (CAT) with Diffusion models



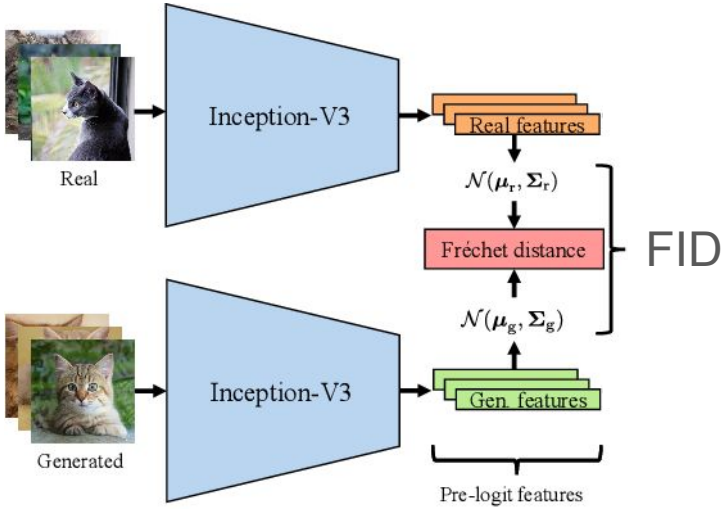
- Based on the PITI diffusion model
- Modification of the PITI semantic encoder to incorporate affinity matrix in the weights.

PITI: Pretraining is All You Need for Image-to-Image Translation, Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, Fang Wen

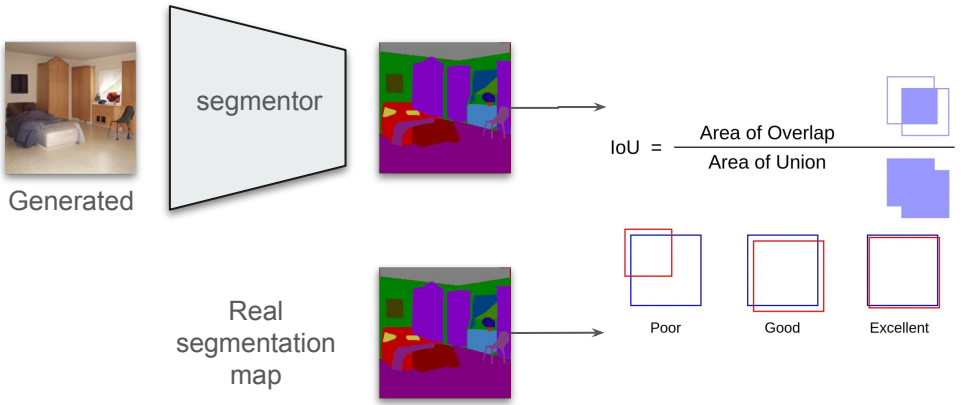
Evaluation setup

- Source dataset:
  - COCO (117k images)
  - ADE (20k images)
- Target dataset:
  - subsets of COCO, ADE, Cityscapes
  - from 25 to 400 images
- Metrics:
  - FID: image realism
  - mIoU: segmentation consistency

Evaluation setup



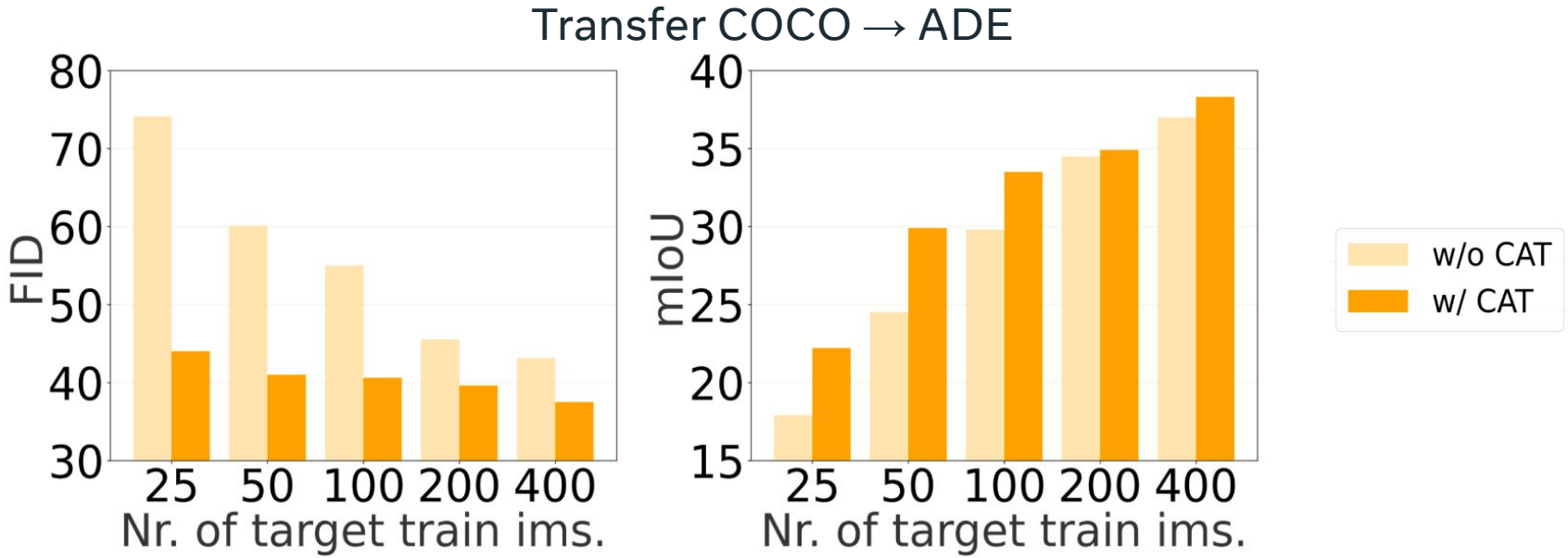
FID



mIoU

Figure taken from the paper : The Role of ImageNet Classes in Fréchet Inception Distance, T Kynkäänniemi et al., 2022

Different target dataset sizes



- Gain is larger with small target dataset size.
- **Better image quality** (FID) and **segmentation consistency** (mIoU) as the dataset size increases.

Comparison of class affinity methods

Affinity matrix initialization	Training-free		After Finetuning	
	FID	mIoU	FID	mIoU
Random	216.0	0.5	54.0	30.0
Text-based	44.6	23.9	41.1	30.4
Segmentation	47.0	22.8	42.0	30.8
Self-supervised	45.5	22.7	41.3	29.8
Combination	**43.1**	**25.1**	**40.9**	**31.4**

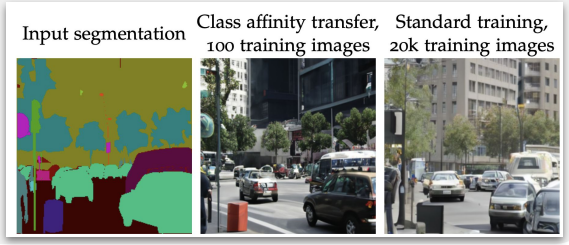
- From COCO -> ADE on OASIS, 100 training images
- Best results with majority voting (last row)

Qualitative results : ADE -> COCO

Target dataset size : 100 images

Take-aways of CAT

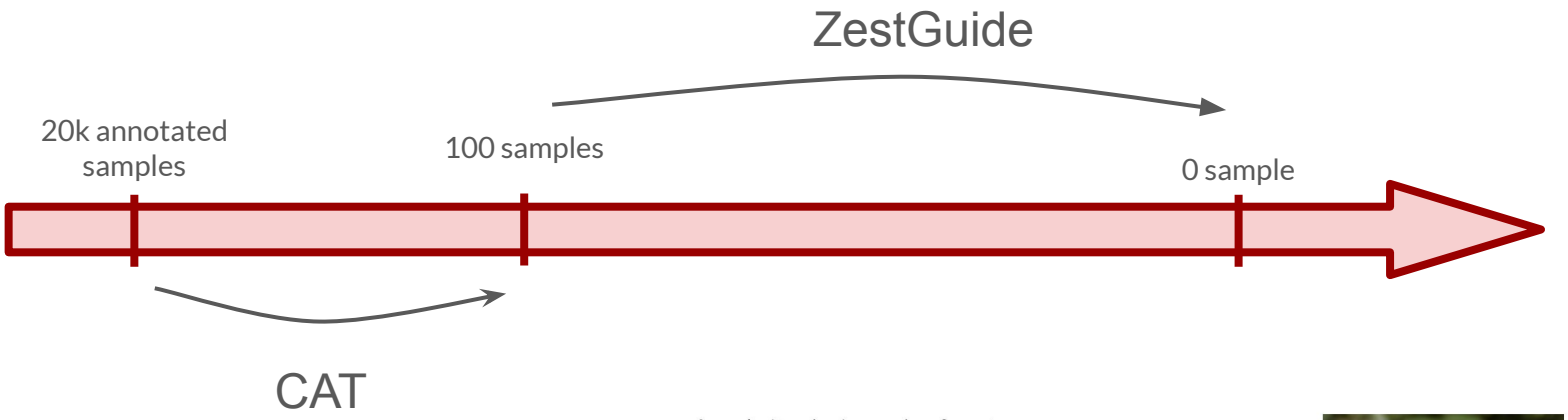
- ➔ Leverage a pretrained semantic image synthesis model.
- ➔ Class affinity between source and target classes.
- ➔ Different ways to estimate class affinities

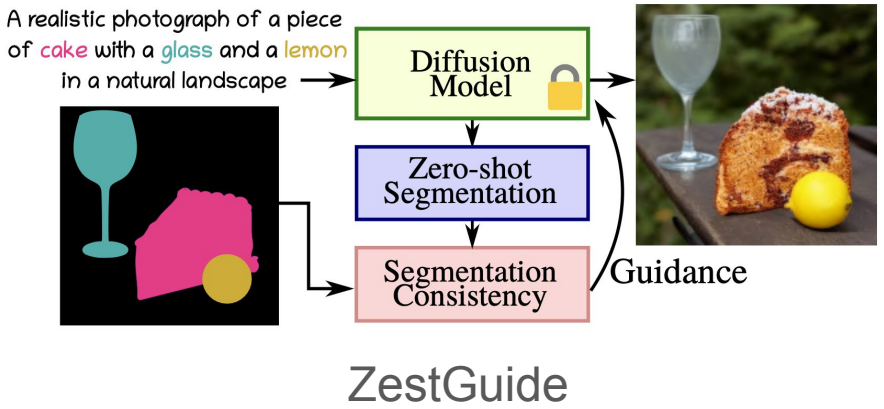


- ➔ Efficiently transfer to very small target datasets
- ➔ Works on GAN-based and diffusion-based models

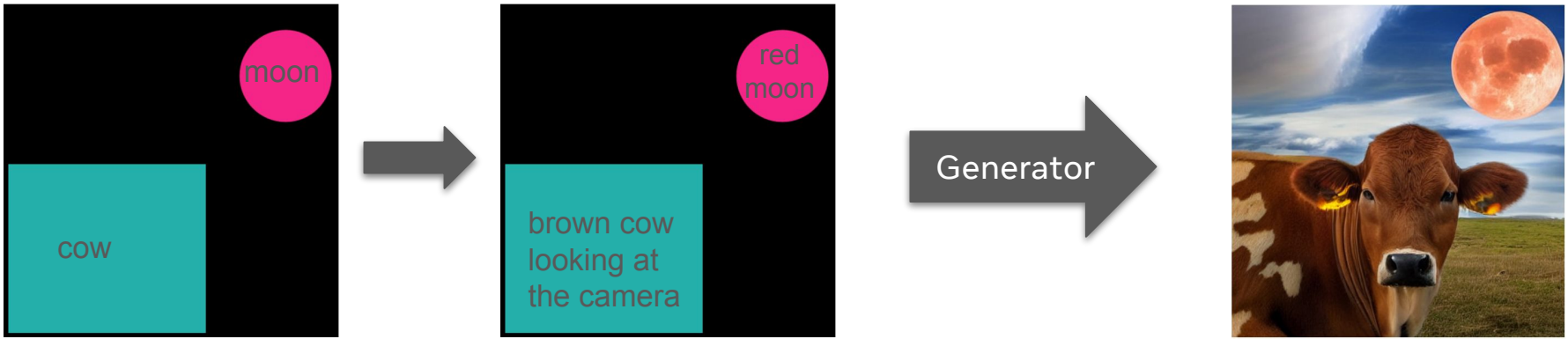
- ➔ Transfers poorly when little class semantic overlap
- ➔ Limitations of closed-vocabulary paradigm
- ➔ Requires few annotated data

Training-free method





Lack of flexibility of semantic image synthesis



- Open real-world cannot be represented by a finite number of classes.
- Introduce free-form textual description.

Can we leverage a pretrained text-to-image diffusion models?

- Generation process conditioned on a **global prompt**
- High-quality image synthesis

Generation




Inputs

Freshly made hot floral tea in glass kettle on the table, angled shot, midday warm

A master jedi cat in star wars holding a lightsaber, wearing a jedi cloak hood, dramatic, cinematic lighting.

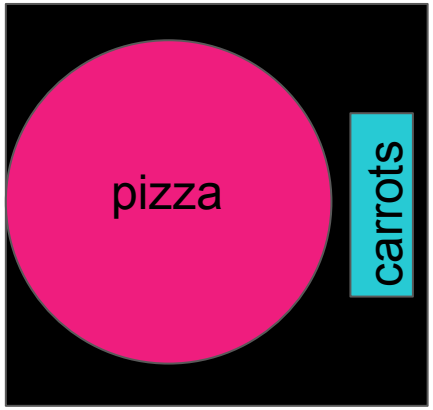
Abandoned city with ruined buildings, long deserted streets, cars aged by time, trees, flowers, scattered leaves, empty streets, vibrant colors.

A living room, bright modern Scandinavian style house, large windows, magazine photoshoot.

Objective

Can we leverage a pretrained text-to-image diffusion model with local constraints ?

Spatial constraints with text-to-image models



A realistic photograph of carrots to the right of pizza.

Generator

Stable Diffusion XL

- High-quality image synthesis with latest text-to-image diffusion models
- Inaccurate and cumbersome to indicate spatial conditioning with text-to-image generative models

Best of both world: semantic maps + text

Recap on Diffusion models

$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$

$q(x_t \mid x_{t-1})$

p_θ estimates the reverse of the diffusion process q. It is parameterized by a U-Net.

Diffusion process computationally more efficient in the latent space of a VQ-GAN .

Recap on Diffusion models

$p_\theta(x_{t-1} \mid x_t) = N(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t))$

Data

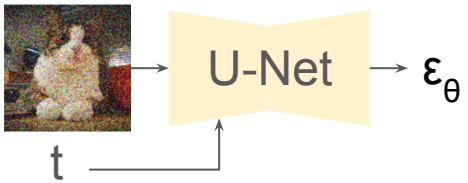
Noise

$q(x_t \mid x_{t-1})$

Re-parametrization: We predict the noise ε_θ instead of the image

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\right)$$

Noise estimate



How to condition a diffusion model ?

$p_\theta(x_{t-1} \mid x_t, y)$ p_θ is also conditioned on condition y

$q(x_t \mid x_{t-1})$

1 Training :U-Net is conditioned on y

How to condition a diffusion model

$p_\theta(x_{t-1} \mid x_t, y)$ p_θ is also conditioned on condition y

$q(x_t \mid x_{t-1})$

1 Training :U-Net is conditioned on y

Classifier-free guidance: $\varepsilon'_\theta(x_t, y) = (1+\omega)\, \varepsilon_\theta(x_t, y) - \omega\, \varepsilon_\theta(x_t)$

Classifier Guidance

$p_\theta(x_{t-1} \mid x_t, y) = p_\theta(x_{t-1} \mid x_t) p(y \mid x_t)$

Data

Noise

Guidance $\nabla_x p(y \mid x_t)$

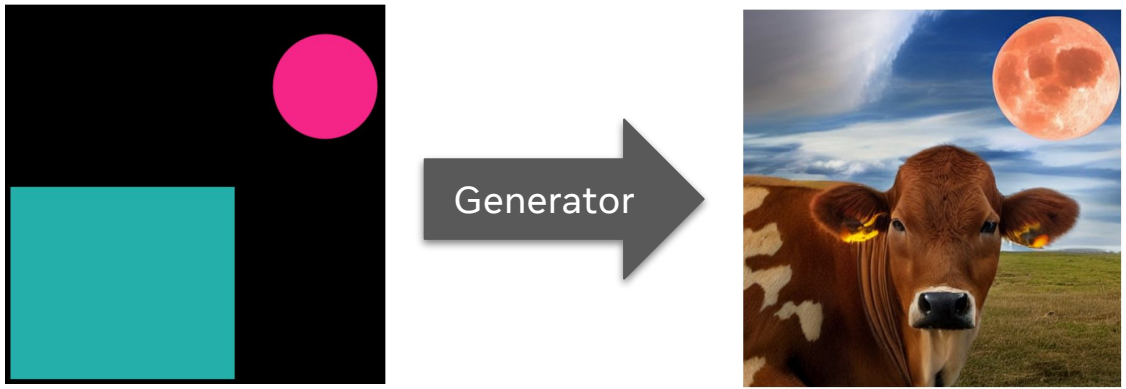
classifier

2 Inference : do classifier guidance for class-conditional generation

$$\hat{\epsilon}(x_t) := \epsilon_\theta(x_t) - \sqrt{1-\bar{\alpha}_t}\,\nabla_{x_t} \log p_\phi(y|x_t)$$

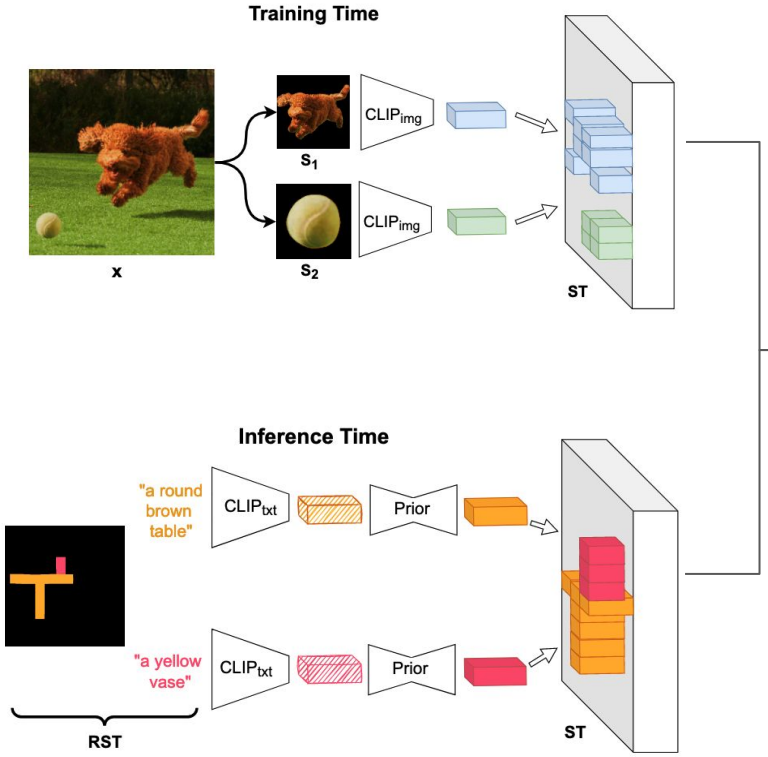
Task: free-form local textual conditioning

'A brown cow in a field, cloudy sky, red full moon.'



Free-form textual description for each segment

SpaText baseline



Conditional diffusion model

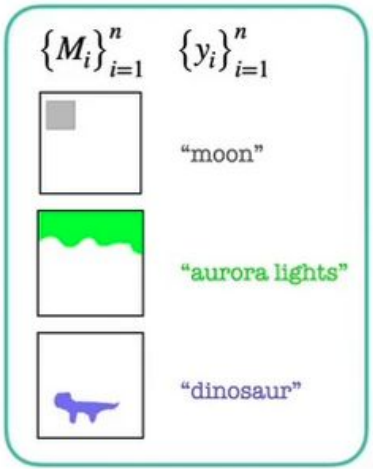
- **Finetune** a text-to-image diffusion model with **additional inputs** coming from CLIP outputs
- **No need** of annotated dataset
- Leverage **shared image and text space** of CLIP model

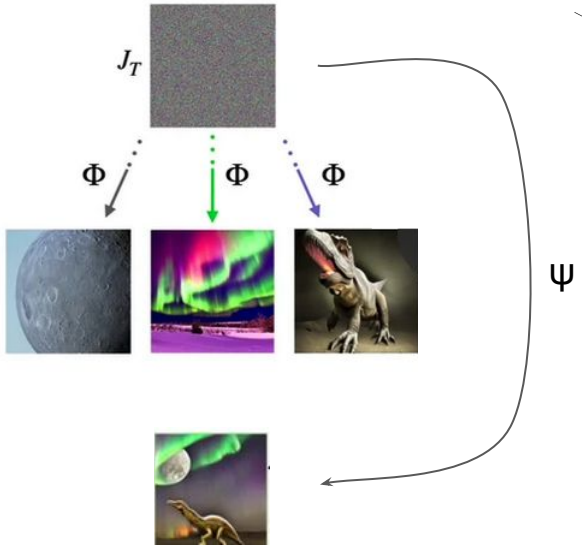
- Need to finetune a pretrained diffusion model on a large datasets

Can we leverage a pretrained text-to-image diffusion model with a **training-free method** ?

SpaText: Spatio-Textual Representation for Controllable Image Generation. Omri Avrahami et al., CVPR 2023

MultiDiffusion baseline





Approximate ψ with Φ

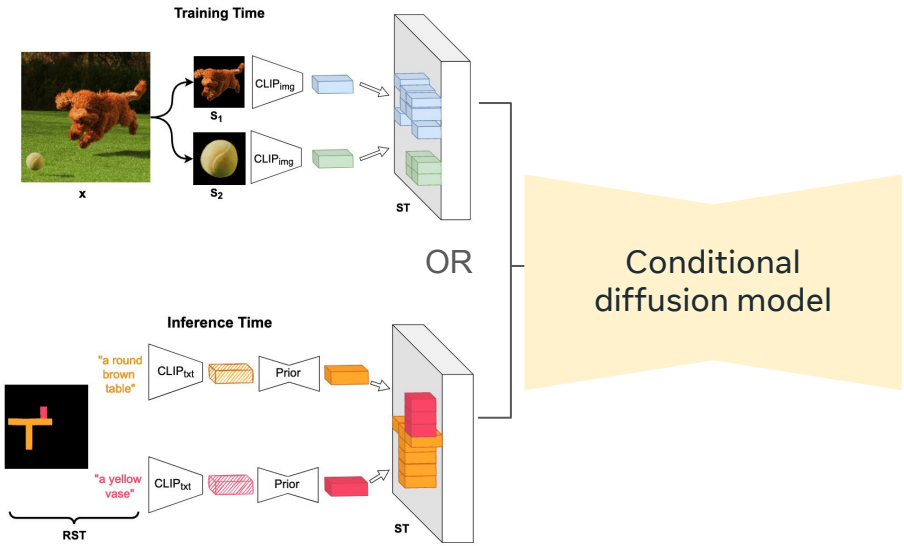
- Frozen pretrained text-to-image diffusion model
- Generate each object separately with a masking strategy at each denoising step
- Merge of the different object noise estimates

Significant increase of compute (proportional to number of objects in the image)

MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. Omer Bar-Tal et al., ICML 2023

Concurrent works

SpaText (w/ finetuning)



MultiDiffusion (w/o finetuning)

- ➔ Defines a new diffusion process
- ➔ At each step, each object generated **separately** with a **masking strategy** + fusion

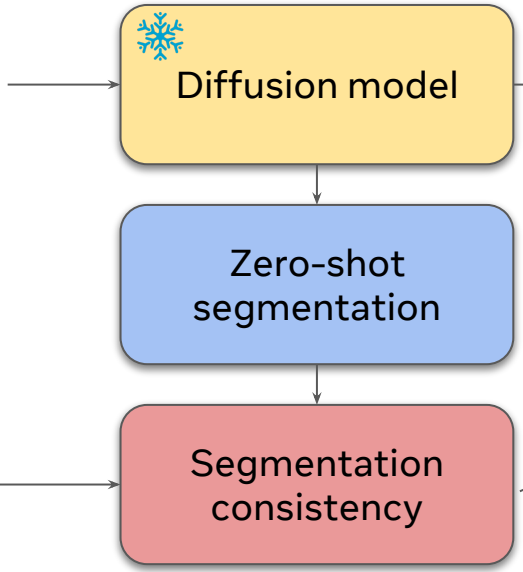
Paint-with-words (w/o finetuning)

- ➔ Exploits cross-attention maps in a pretrained text-to-image diffusion model

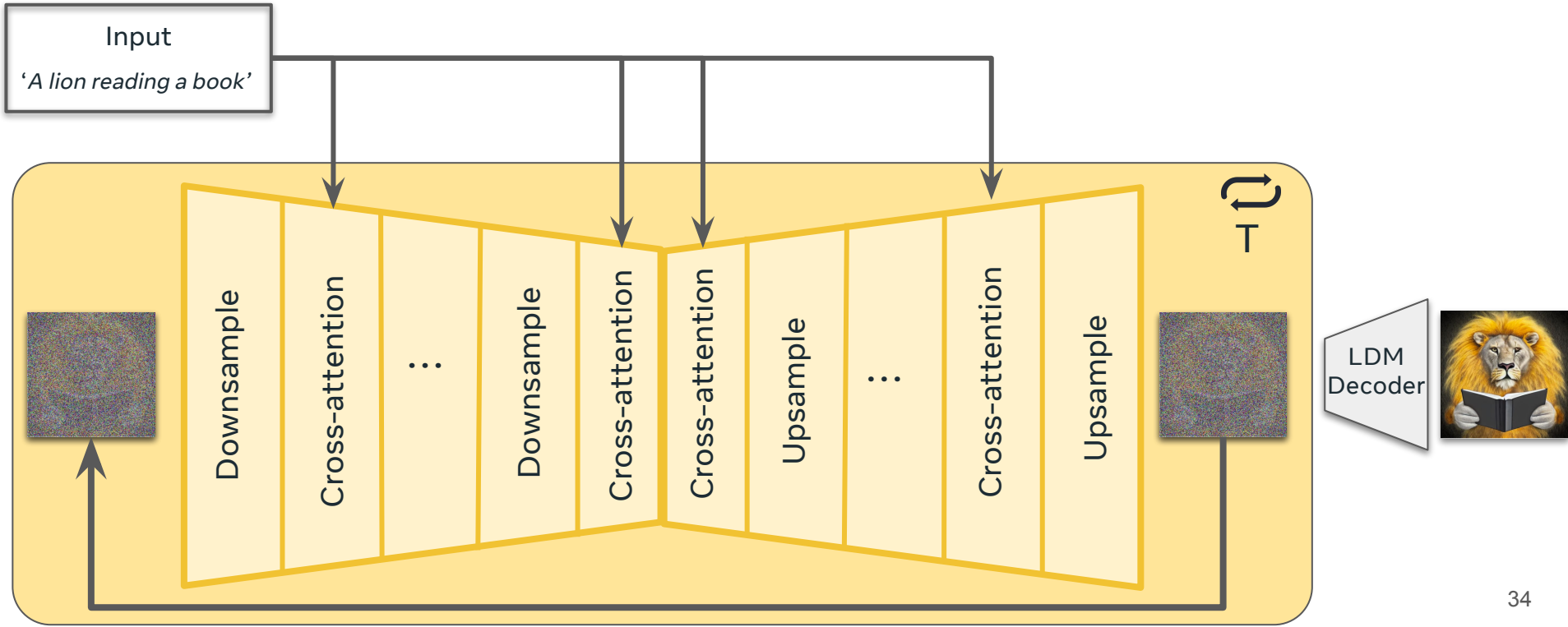
MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. Omer Bar-Tal et al., ICML 2023
SpaText: Spatio-Textual Representation for Controllable Image Generation. Omri Avrahami et al., CVPR 2023
eDiff-I, Paint-with-Words: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers, Yogesh Balaji et al., ArXiv 2022

ZestGuide

'A realistic photograph of a piece of cake with a glass and a lemon in a natural landscape'

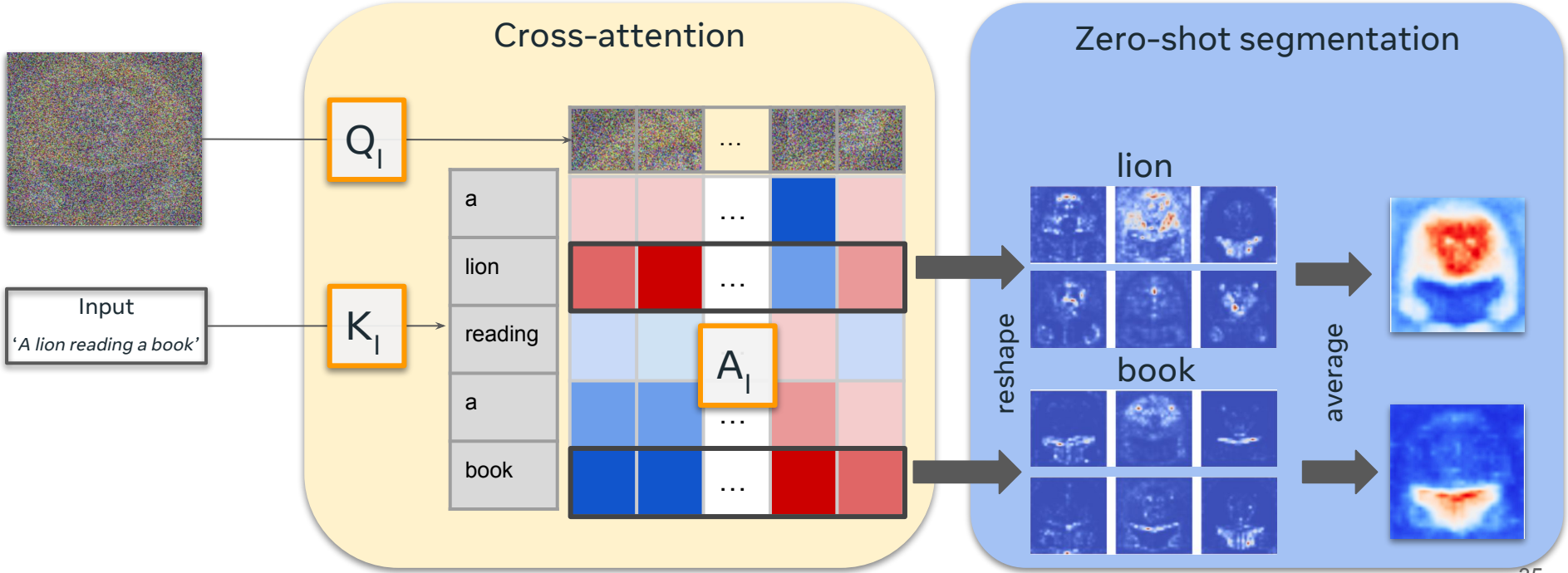


Text conditioning in diffusion models

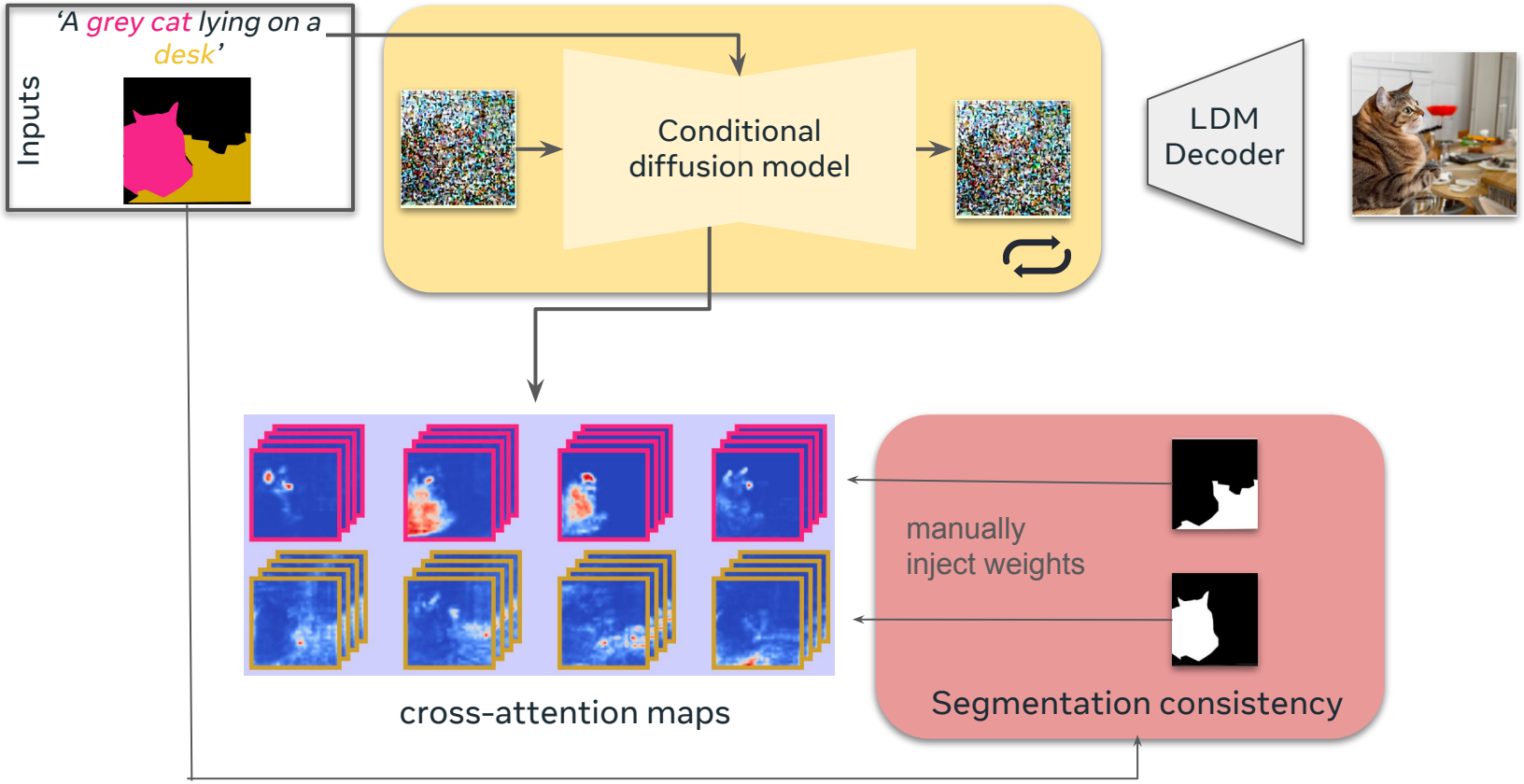


Spatio-textual interaction in diffusion models

$$\mathbf{A}_l = \text{Softmax}\left(\frac{\mathbf{Q}_l\mathbf{K}_l^T}{\sqrt{d}}\right)$$

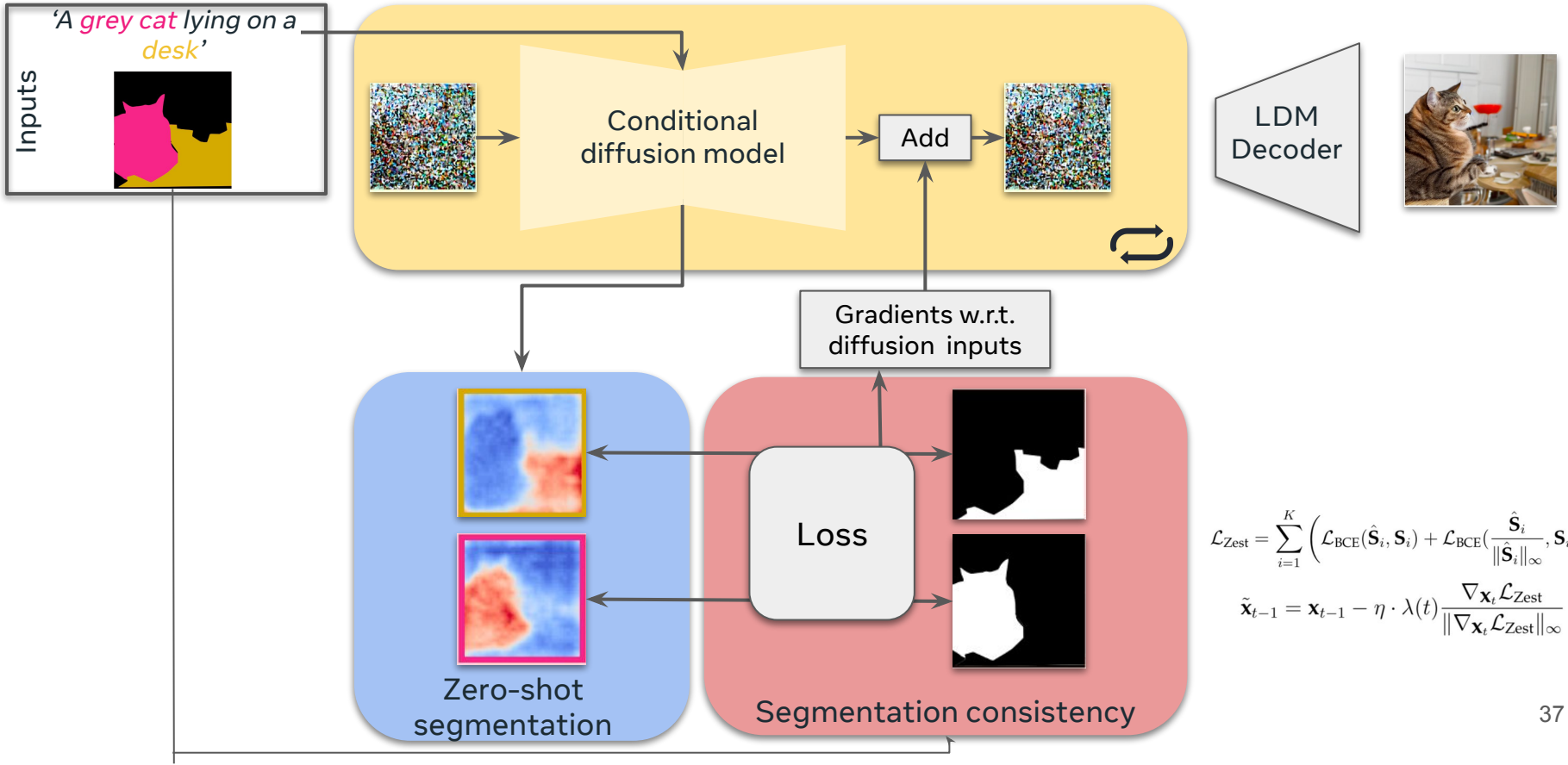


Paint-with-Words baseline



eDiff-I: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers, *Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, Ming-Yu Liu*, Arxiv 2022

ZestGuide



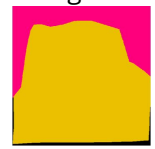
$$\mathcal{L}_{\text{Zest}} = \sum_{i=1}^{K} \left(\mathcal{L}_{\text{BCE}}(\hat{\mathbf{S}}_i, \mathbf{S}_i) + \mathcal{L}_{\text{BCE}}(\frac{\hat{\mathbf{S}}_i}{\|\hat{\mathbf{S}}_i\|_\infty}, \mathbf{S}_i) \right)$$

$$\tilde{\mathbf{x}}_{t-1} = \mathbf{x}_{t-1} - \eta \cdot \lambda(t) \frac{\nabla_{\mathbf{x}_t} \mathcal{L}_{\text{Zest}}}{\|\nabla_{\mathbf{x}_t} \mathcal{L}_{\text{Zest}}\|_\infty}$$

Evolution of attention maps with timesteps

Inputs

A big burly grizzly bear is shown with grass in the background



Showing the first 12 denoising steps out of 50 steps

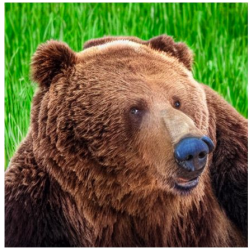
timesteps

w/o guidance

Bear

Grass

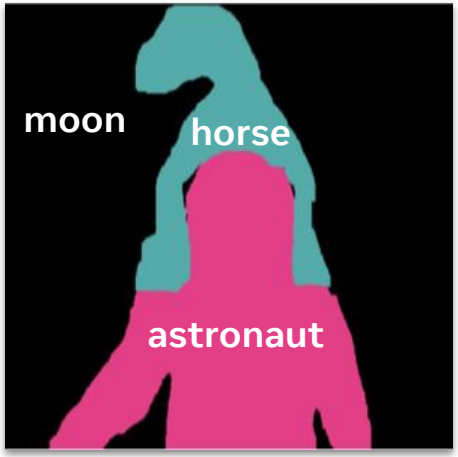
w/ guidance



Bear

Grass

Hyperparameters tuning

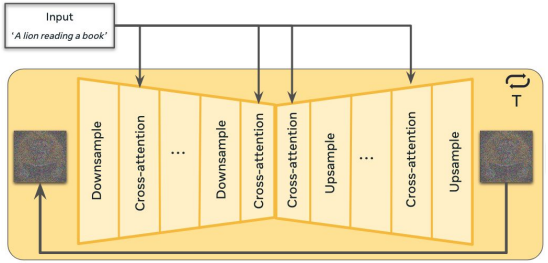


increasing number of updates (from 10% to 50%)

increasing learning rate (from 0.1 to 2)

$$\tilde{\mathbf{x}}_{t-1} = \mathbf{x}_{t-1} - \eta \cdot \lambda(t) \frac{\nabla_{\mathbf{x}_t} \mathcal{L}_{\text{Zest}}}{\|\nabla_{\mathbf{x}_t} \mathcal{L}_{\text{Zest}}\|_\infty}$$

Ablation on cross-attention layers



Diffusion U-Net

Layers used	↓FID	↑mIoU	↑CLIP
All layers	33.74	40.17	30.19
Only decoder layers	33.81	40.02	30.05
Only encoder layers	30.98	38.24	30.67
Only res32 layers	**29.35**	39.49	**30.75**
Only res16 layers	33.59	40.27	30.23
res16 and res32 layers (ours)	31.53	**43.34**	30.44

Ablation on which cross-attention layers to use to compute the average estimated mask S_i

$$\mathcal{L}_{\text{Zest}} = \sum_{i=1}^{K} \left(\mathcal{L}_{\text{BCE}}(\hat{\mathbf{S}}_i, \mathbf{S}_i) + \mathcal{L}_{\text{BCE}}(\frac{\hat{\mathbf{S}}_i}{\|\hat{\mathbf{S}}_i\|_\infty}, \mathbf{S}_i) \right)$$

Ablation on loss type

Components	↓FID	↑mIoU	↑CLIP
Loss for each attention head	33.6	32.1	29.9
Loss for each layer	31.6	42.7	30.5
Loss for global average (ours)	31.5	43.3	30.4

Ablation on how we compute the loss -> global average is the best

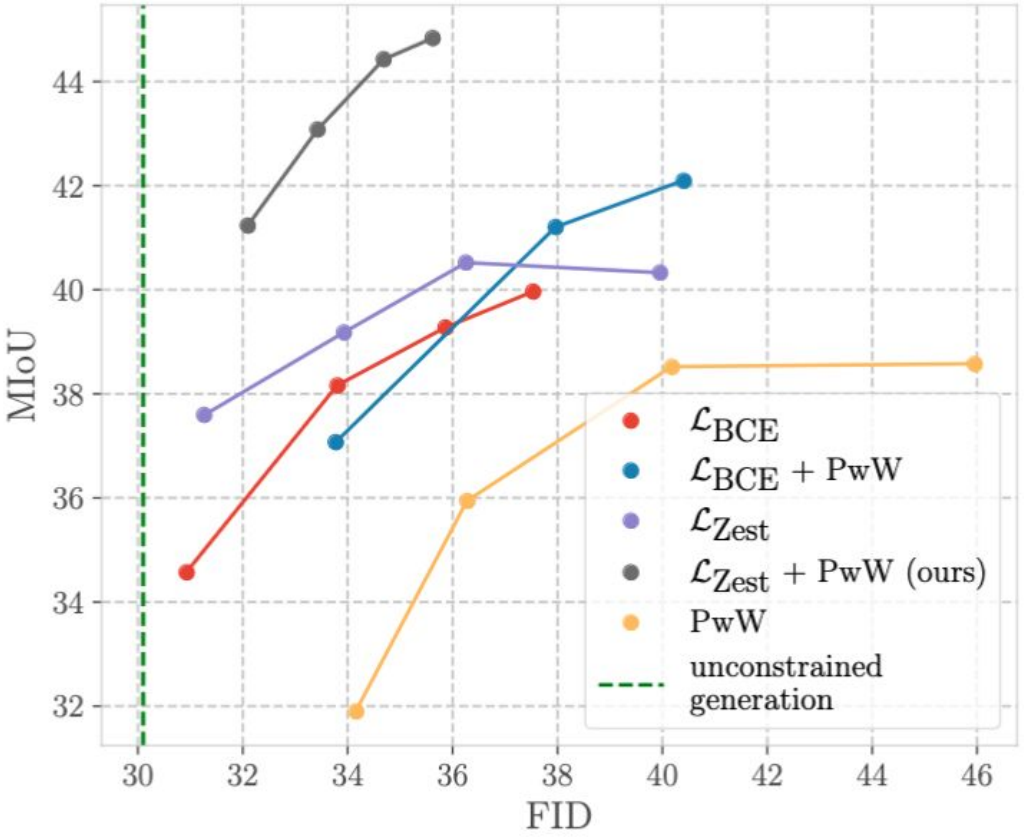
$$\mathcal{L}_{\text{Zest}} = \sum_{i=1}^{K} \left(\mathcal{L}_{\text{BCE}}(\hat{\mathbf{S}}_i, \mathbf{S}_i) + \mathcal{L}_{\text{BCE}}(\frac{\hat{\mathbf{S}}_i}{\|\hat{\mathbf{S}}_i\|_\infty}, \mathbf{S}_i) \right)$$

Tradeoff between FID and mIoU

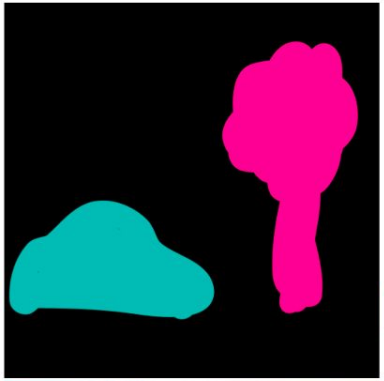
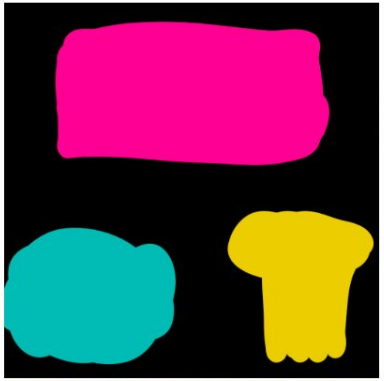
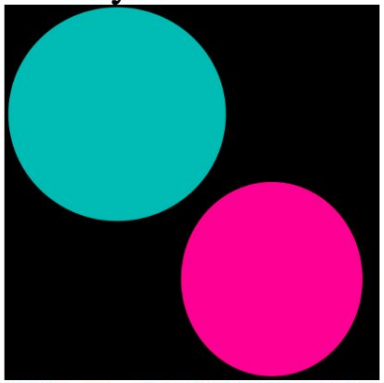
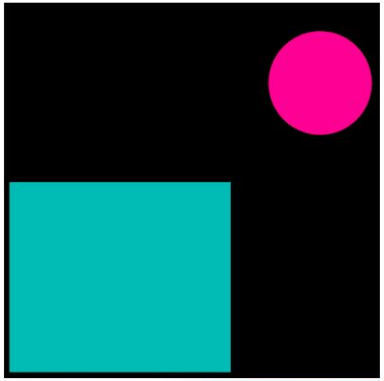
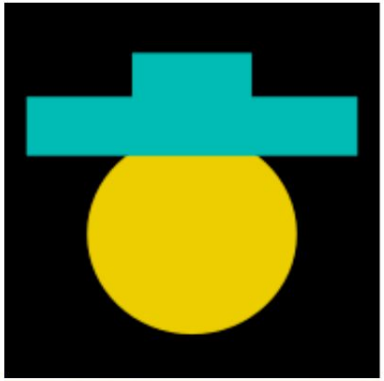
$$\mathcal{L}_{\text{Zest}} = \sum_{i=1}^{K} \left(\mathcal{L}_{\text{BCE}}(\hat{\mathbf{S}}_i, \mathbf{S}_i) + \mathcal{L}_{\text{BCE}}(\frac{\hat{\mathbf{S}}_i}{\|\hat{\mathbf{S}}_i\|_\infty}, \mathbf{S}_i) \right)$$

$$\tilde{\mathbf{x}}_{t-1} = \mathbf{x}_{t-1} - \eta \cdot \lambda(t) \frac{\nabla_{\mathbf{x}_t} \mathcal{L}_{\text{Zest}}}{\|\nabla_{\mathbf{x}_t} \mathcal{L}_{\text{Zest}}\|_\infty}$$

ZestGuide can be further improved by using Paint-with-Words



Qualitatives

“A car and a tree, at the beach.”	“ A mirror, sink and flowers in a bathroom.”	“Plate with cookies and cup of coffee, fancy tablecloth ”	“A brown cow in a field, cloudy sky, red full moon”	“A mouse wearing a hat in the desert.”
				
				

Qualitatives

Quantitative evaluation

Evaluation metrics:

- **FID**: image realism
- **mIoU**: consistency with segmentation map
- **CLIP**: consistency with caption

Evaluation setting:

- Dataset : COCO-5K
- selecting objects covering more than **5%** of the conditioning segmentation mask
- choosing **$1 \leq K \leq 3$ objects per mask**

Method	free-form mask texts	Zero-shot	FID↓	mIoU↑	CLIP↑
OASIS [1]	✗	✗	46.8	41.4	N/A
SpaText [3]	✓	✗	**16.2**	23.8	30.2
MultiDiffusion [2]	✓	✓	21.1	19.6	29.0
Paint with Words [4]	✓	✓	20.3	36.3	**31.2**
ZestGuide (ours)	✓	✓	21.0	**46.9**	30.3

[1] You only need adversarial supervision for semantic image synthesis. Edgar Schönfeld et al., ICLR 2021
[2] MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. Omer Bar-Tal et al., ICML 2023
[3] SpaText: Spatio-Textual Representation for Controllable Image Generation. Omri Avrahami et al., CVPR 2023
[4] eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers, Yogesh Balaji et al., ArXiv 2022

Quantitative evaluation

Evaluation metrics:

- **FID**: image realism
- **mIoU**: consistency with segmentation map
- **CLIP**: consistency with caption

Evaluation setting:

- Dataset : COCO-5K
- Selecting objects covering more than **5%** of the conditioning segmentation mask

Method	free-form mask texts	Zero-shot	FID↓	mIoU↑	CLIP↑
OASIS [1]	✗	✗	18.2	53.7	N/A
SpaText [3]	✓	✗	28.9	19.2	30.1
Paint with Words [4]	✓	✓	23.4	31.8	31.4
ZestGuide (ours)	✓	✓	23.1	43.3	31.3

[1] You only need adversarial supervision for semantic image synthesis. Edgar Schönfeld et al., ICLR 2021
[2] MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. Omer Bar-Tal et al., ICML 2023
[3] SpaText: Spatio-Textual Representation for Controllable Image Generation. Omri Avrahami et al., CVPR 2023
[4] eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers, Yogesh Balaji et al., ArXiv 2022

Quantitative evaluation

Evaluation metrics:

- **FID**: image realism
- **mIoU**: consistency with segmentation map
- **CLIP**: consistency with caption

Evaluation setting:

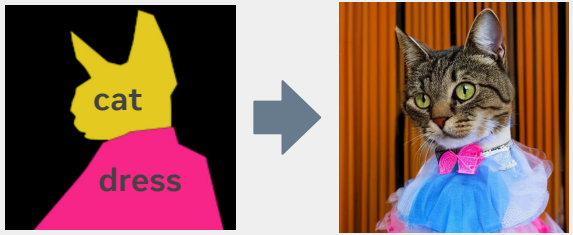
- Dataset : COCO-5K
- **All objects** in the masks

Method	free-form mask texts	Zero-shot	FID↓	mIoU↑	CLIP↑
OASIS [1]	✗	✗	15.0	52.1	N/A
SpaText [3]	✓	✗	19.8	16.8	30.0
Paint with Words [4]	✓	✓	22.9	27.9	31.5
ZestGuide (ours)	✓	✓	22.8	33.1	31.9

[1] You only need adversarial supervision for semantic image synthesis. Edgar Schönfeld et al., ICLR 2021
[2] MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. Omer Bar-Tal et al., ICML 2023
[3] SpaText: Spatio-Textual Representation for Controllable Image Generation. Omri Avrahami et al., CVPR 2023
[4] eDiff-I: Text-to-Image Diffusion Models with Ensemble of Expert Denoisers, Yogesh Balaji et al., ArXiv 2022

Take-aways of ZestGuide

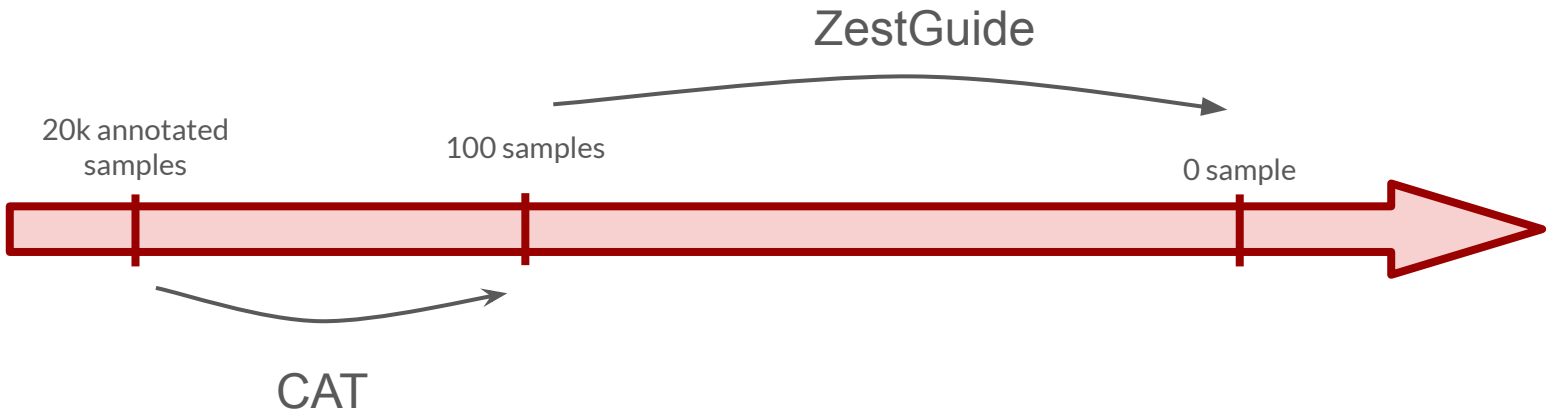
- ➔ Method for free-form mask texts conditioning of diffusion models.
- ➔ Averaged cross-attention maps to estimate positions of objects in the prompt
- ➔ Classifier guidance to enforce segmentation consistency



- ➔ Training-free method
- ➔ Flexible conditioning with both texts and segments
- ➔ Works well with coarse masks

- ➔ Object boundaries not always well respected
- ➔ Sometimes, small object are ignored

Conclusion



Different types of conditioning :

- CAT [1]: Per-pixel class name
- ZestGuide [2] : Per-pixel textual description

  …
- PerCo [3] : low-level and semantic information conditioning with texts and learned spatial quantized maps

[1] Few-shot Semantic Image Synthesis with Class Affinity Transfer, *Marlène Careil, Jakob Verbeek, Stéphane Lathuilière,* CVPR 2023
[2] Zero-shot spatial layout conditioning for text-to-image diffusion models, *Guillaume Couairon*, Marlène Careil*, Matthieu Cord, Stéphane Lathuilière, Jakob Verbeek,* ICCV 2023
[3] Towards image compression with perfect realism at ultra-low bitrates, *Marlène Careil, Matthew J. Muckley, Jakob Verbeek, Stéphane Lathuilière,* ICLR 2024

Overview Practical

Jupyter notebook to play with the different ZestGuide hyper-parameters:

- ➔ Number of guidance steps
- ➔ Learning rate
- ➔ Which attention layers to use
- ➔ Seed