DataComp
Creating the largest public multimodal Dataset

Institute for Found
MACHINE L

Alex Dimakis
UT Austin

Datacomp is a collaboration with Ludwig Schmidt and the DataComp team

DataComp: In search of the next generation of multimodal datasets

Samir Yitzhak Gadre*[2] Gabriel Ilharco*[1] Alex Fang*[1] Jonathan Hayase[1] Georgios Smyrnis[5]
Thao Nguyen[1] Ryan Marten[7,10] Mitchell Wortsman[1] Dhruba Ghosh[1] Jieyu Zhang[1] Eyal Orgad[3]
Rahim Entezari[11] Giannis Daras[5] Sarah Pratt[1] Vivek Ramanujan[1] Yonatan Bitton[12]
Kalyani Marathe[1] Stephen Mussmann[1] Richard Vencu[6] Mehdi Cherti[6,8,9] Ranjay Krishna[1]
Pang Wei Koh[1] Olga Saukh[11] Alexander Ratner[1] Shuran Song[2] Hannaneh Hajishirzi[1,7]
Ali Farhadi[1] Romain Beaumont[6] Sewoong Oh[1] Alexandros G. Dimakis[5] Jenia Jitsev[6,8,9]
Yair Carmon[3] Vaishaal Shankar[4] Ludwig Schmidt[1,6,7]

Abstract

Large multimodal datasets have been instrumental in multiple breakthroughs like CLIP, DALL-E, Stable Diffusion, Flamingo and GPT-4, yet datasets rarely receive the same research attention as model architectures or training algorithms. To address this shortcoming in the machine learning ecosystem, we introduce DataComp, a participatory benchmark where the training code is fixed and researchers innovate by proposing new training sets. Concretely, we provide an experimental testbed centered around a new

What are foundation models

- First answer: ChatGPT
- Also StyleGAN2, Dalle2, Stable Diffusion, CLIP, Whisper, SAM, Dino v2
- Some are generative (StyleGAN, GPT, Dalle), some are classifiers (CLIP)
- Trained at **billion** scale datasets. (vs **million** scale for Imagenet).
- Most FMs are closed (OpenAI...). This is a key problem for a competitive AI ecosystem.
- We need large open multimodal datasets and models.
- In IFML we are
  1. Creating new Foundation models
  2. Teaching Foundation models new tricks.
  (Solve new problems, Teach new concepts, Improve Consistency, Accelerate)

Datasets are the foundation of progress in AI

Example: **language**

GPT-1 (2018): ≈ 3 billion tokens

GPT-2 (2019): ≈ 30 billion tokens

GPT-3 (2020): ≈ 300 billion tokens

GPT-4 (2023): ≈ 3,000 billion tokens (?) → **1,000x growth in 5 years**

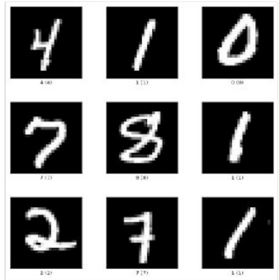
Example: **images**

ImageNet (2009): ≈ 1 million images

LAION-5B (2022): ≈ 5 billion images → **5,000x growth in 13 years**

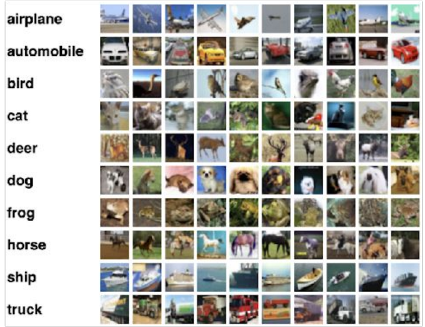
X

Discoveries in ML enabled by datasets



MNIST (1994)

Convolutional neural networks



CIFAR-10 (2009)

Training on GPUs

ImageNet (2012)

Deep learning resurgence, ResNets, transfer learning, etc.

OpenAI

WebImageText (2021)

Zero-shot classification (CLIP), text-guided image generation (DALL-E) x

The future is multimodal

The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision)

Zhengyuan Yang*, **Linjie Li***, **Kevin Lin***, **Jianfeng Wang***, **Chung-Ching Lin***, **Zicheng Liu**, **Lijuan Wang***♠
Microsoft Corporation

* Core Contributor ♠ Project Lead

Abstract

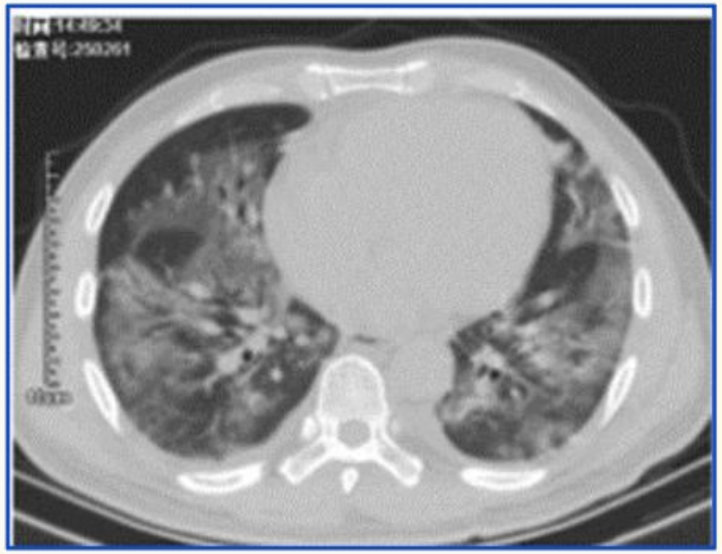
Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision)[1], to deepen the understanding of LMMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and genericity of

GPT4V= GPT4+ Images as input

- LMMs (Large MultiModal Models) integrate image and text capabilities.

Prompt:

Look at the CT scan of lung in the image below, tell me what's wrong.



GPT-4V:

There appears to be multiple areas of consolidation and ground-glass opacities in both lungs, which could indicate a lung infection or inflammation. There is also a possible mass or nodule in the right upper lobe. Further evaluation and clinical correlation is needed to determine the exact diagnosis.

The future is multimodal

The Dawn of LMMs:
Preliminary Explorations with GPT-4V(ision)

We need large open multimodal datasets

Abstract

Large multimodal models (LMMs) extend large language models (LLMs) with multi-sensory skills, such as visual understanding, to achieve stronger generic intelligence. In this paper, we analyze the latest model, GPT-4V(ision)[1], to deepen the understanding of LMMs. The analysis focuses on the intriguing tasks that GPT-4V can perform, containing test samples to probe the quality and genericity of

The LAION dataset

- 5B image-caption pairs from the internet.
- Start with Common Crawl which is a crawl of the internet (images + alt text). Filter image-caption pairs that are good fit.
- Dataset creation is a data filtration problem.
- LAION was used to train Stable Diffusion, OpenCLIP and numerous other world-changing open projects.
- (NeurIPS 2022 best paper award).

DataComp= LAION 2.0

- The largest public multimodal dataset (13B images+captions)
- The first data filtration competition.
- Given a pool of billions of image-caption pairs,
- Select a subset of high quality.
- Run fixed training code on fixed model to get a CLIP model.
- Evaluated in Imagenet zero-shot and several other tasks.
- We also detected and blurred faces which makes LAION usable by many companies.

DataComp

- Key idea: Lets re-do LAION-5B but in a flexible and re-usable way
- Publish not only the dataset but the entire **Tooling** around it
- Make it a **Benchmark** (Datacomp= Dataset Competition).
- Goal: Enable **collaborative data-centric research**,
  similar to model-centric research

DataComp: Shift to Data-Centric AI

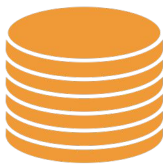
Traditional ML Benchmark: Dataset fixed (e.g. Imagenet), improve the model (Model Centric)

DataComp: Model training code fixed.

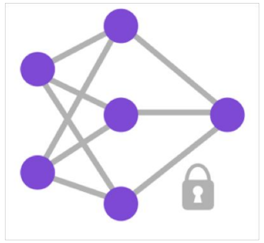
Improve the dataset. (Data Centric)

DataComp

- DataComp Provides: Large Pool of candidate image-text pairs from common crawl (12M, 128M, 1.2B, 12.8B).

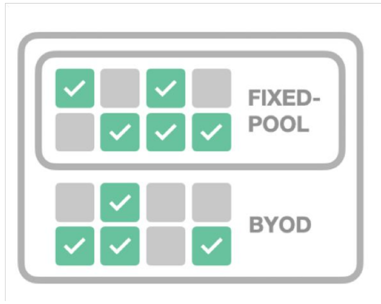


- Safety: toxic content filtering, face bluring, data poisoning protection
- Standardized training code for CLIP models.



- Evaluation testbed on 38 downstream datasets.

- Participants submit either:
- A subset of of the pool + filtering method
- A new data source (bring your own data track)



DataComp impact

CLIP model size	Training dataset	ImageNet accuracy	ObjectNet accuracy
Large (L/14)	OpenAI	75.6%	69.0%
Large (L/14)	LAION	72.8% (-2.8)	59.9% (-9.1)
Large (L/14)	**DataComp**	**79.2% (+3.6)**	**74.2% (+5.2)**

First public training set that is better than OpenAI
Already in use at Anthropic, Apple, CMU, Meta, many others
Used for OpenCLIP, the most used open source CLIP implementation
50k downloads (git clones) per day
Datacomp used in Stable Diffusion

Datacomp.ai

datacomp.ai/leaderboard.html

DataComp | Home | Participate | Tracks | FAQs | Workshop | Team | Leaderboard

Select the track and scale

Filtering track	BYOD track

small	medium	large	xlarge

Leaderboard

Rank	Created	Submission	ImageNet acc.	Average perf.	Dataset size	Authors
1	4-28-2023	Baseline: Image-based ∩ CLIP score (L/14 30%)	0.792	0.653	1.4e+9	DataComp team
2	4-28-2023	Baseline: CLIP score (L/14 30%)	0.764	0.641	3.8e+9	DataComp team
3	4-28-2023	Baseline: LAION-2B filtering	0.755	0.627	1.3e+9	DataComp team
4	4-28-2023	Baseline: No filtering	0.723	0.611	1.3e+10	DataComp team

Conclusions

- Making dataset curation a first-class citizen in AI research
- Develop novel models to do filtration and weighting
- Creating benchmarking pipelines to evaluate zero-shot performance, fairness and robustness
- Industry is not releasing datasets, so good for academia to do
- Create a standardized benchmark to evaluate synthetic datasets

fin

Previous datasets

Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

Talk overview

- ~~Part 1. Creating large multimodal datasets (LAION and Datacomp)~~

- Part 2.Teaching existing Foundation models new tricks

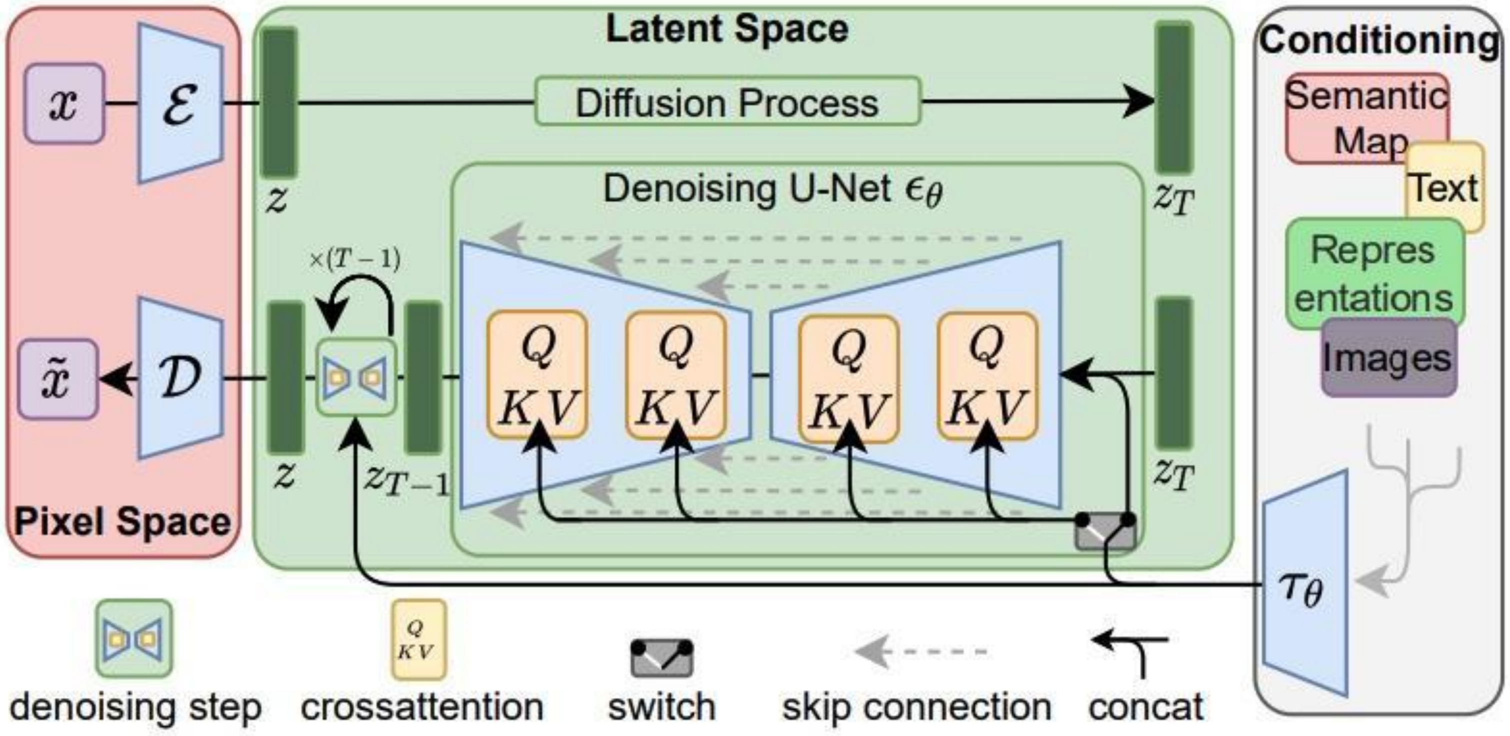
-Inpainting using Stable Diffusion
-Enforcing Consistency
-Finetuning with corrupted data: Ambient Diffusion
-Teaching them custom concepts.

Teaching Foundation Models new tricks: Our Recent work

- 1. **PSLD**: Solving Linear Inverse Problems Provably via Posterior Sampling with Latent Diffusion Models
- Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G. Dimakis, Sanjay Shakkottai
  https://arxiv.org/pdf/2307.00619.pdf Neurips 2023

Stable Diffusion

- A Latent diffusion model uses an Auto-encoder to go into a latent space and performs a diffusion there.
- Much faster and memory efficient compared to pixel-space diffusion models.
- Stable Diffusion is the state-of-the art foundation model for generating images (trained on LAION 5B).
- Can generate great images but we want to solve inverse problems, like inpainting.

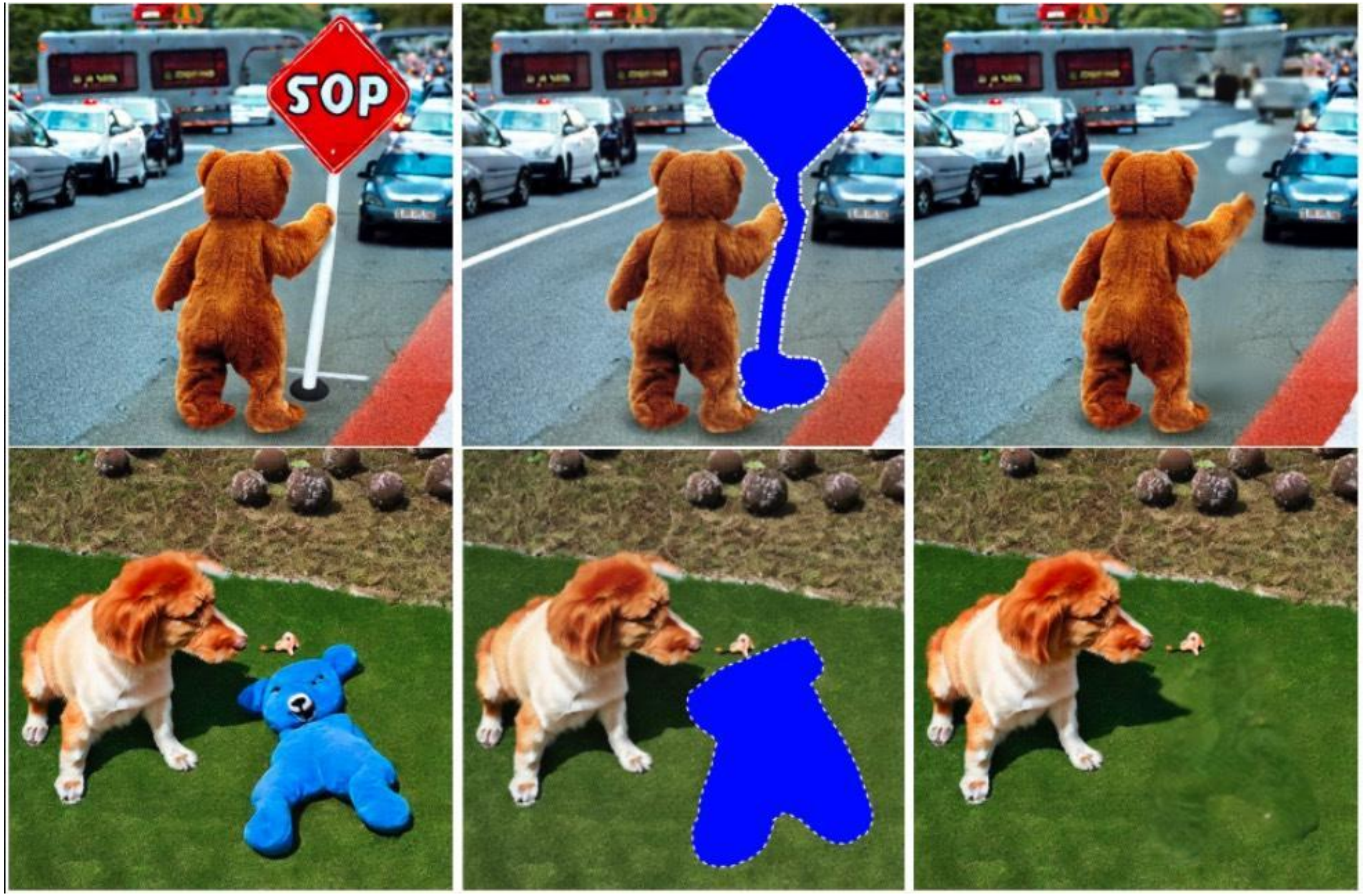


Inpainting

- Given an image with missing pixels,

Complete the pixels to preserve the statistics of the image.

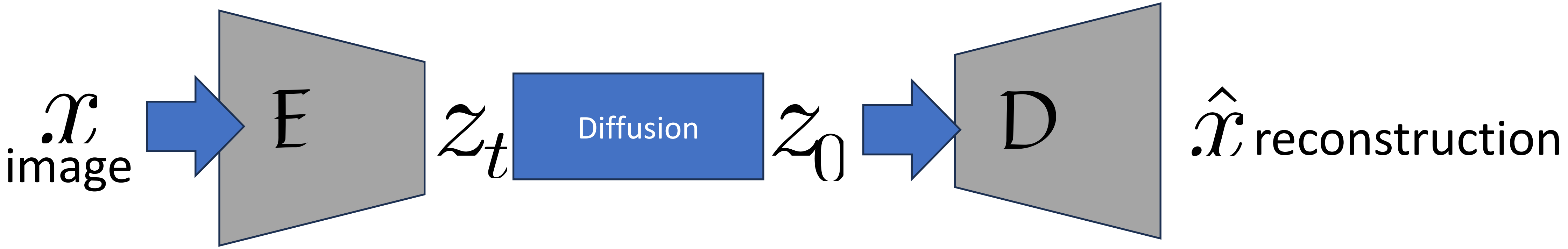
Missing data problem for pixels.



Prior work

- A lot of significant work has studied inpainting using generative models. We are building on, and comparing to
- DPS: H. Chung, J. Kim, M. Mccann, M. Klasky, and J.C. Ye. "Diffusion Posterior Sampling for General Noisy Inverse Problems" ICLR 23
- DDRM: B. Kawar, M. Elad, S. Ermon, and J. Song. "Denoising Diffusion Restoration Models". NeurIPS 22.
- PnP ADMM: S.Chan, X. Wang, O. Elgendy. "Plug-and-play ADMM for image restoration: Fixed-point convergence and applications". 2016
- DeepRED:G. Mataev, P. Milanfar, and M. Elad. "DeepRED: Deep image prior powered by RED". ICCV 2019

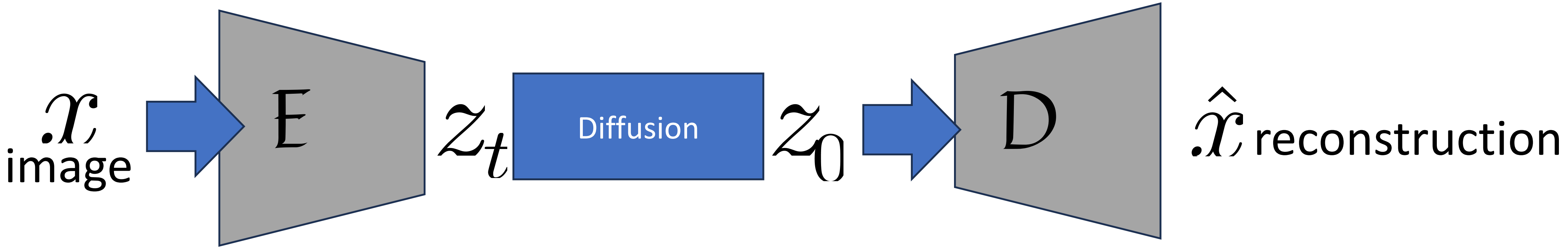
Foundational Idea



The encoder is a many to one mapping: There are multiple images x that map to the same z.

Diffusion is happening in the latent space z_t

Foundational Idea



The encoder is a many to one mapping: There are multiple images x that map to the same z.

$$y = \mathcal{A}x_0 + \sigma_y \eta$$

During diffusion, we use the Conditional reverse SDE

$$dx = \big(f(x,t) - g^2(t)\left(\nabla_{x_t} \log p_t(x_t) + \nabla_{x_t} \log p(y|x_t)\right)\big)\,\mathrm{d}t + g(t)\mathrm{d}w.$$

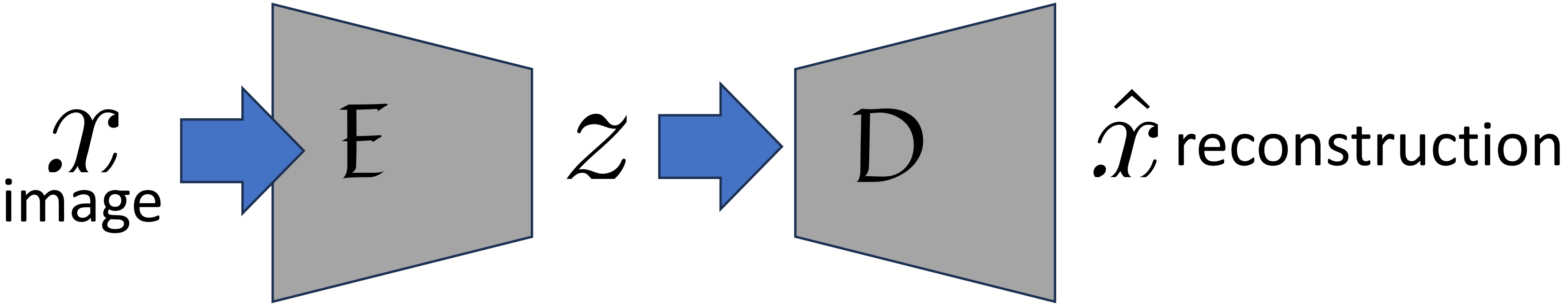
We discovered that this leads us to bad **z** latent vectors.

Stable diffusion for Inpainting has a problem

Input

Previous SOTA
(DPS in SD)

Foundational Idea (PSLD)



$$\mathcal{D}[\mathcal{E}(x)] = \hat{x}$$ This is how the AutoEncoder is trained

$$\mathcal{E}[\mathcal{D}(z)] = z$$ We want to keep the diffusion to stay within such z vectors.

$$||z - \mathcal{E}[\mathcal{D}(z)]||^2$$ By adding this term in the diffusion

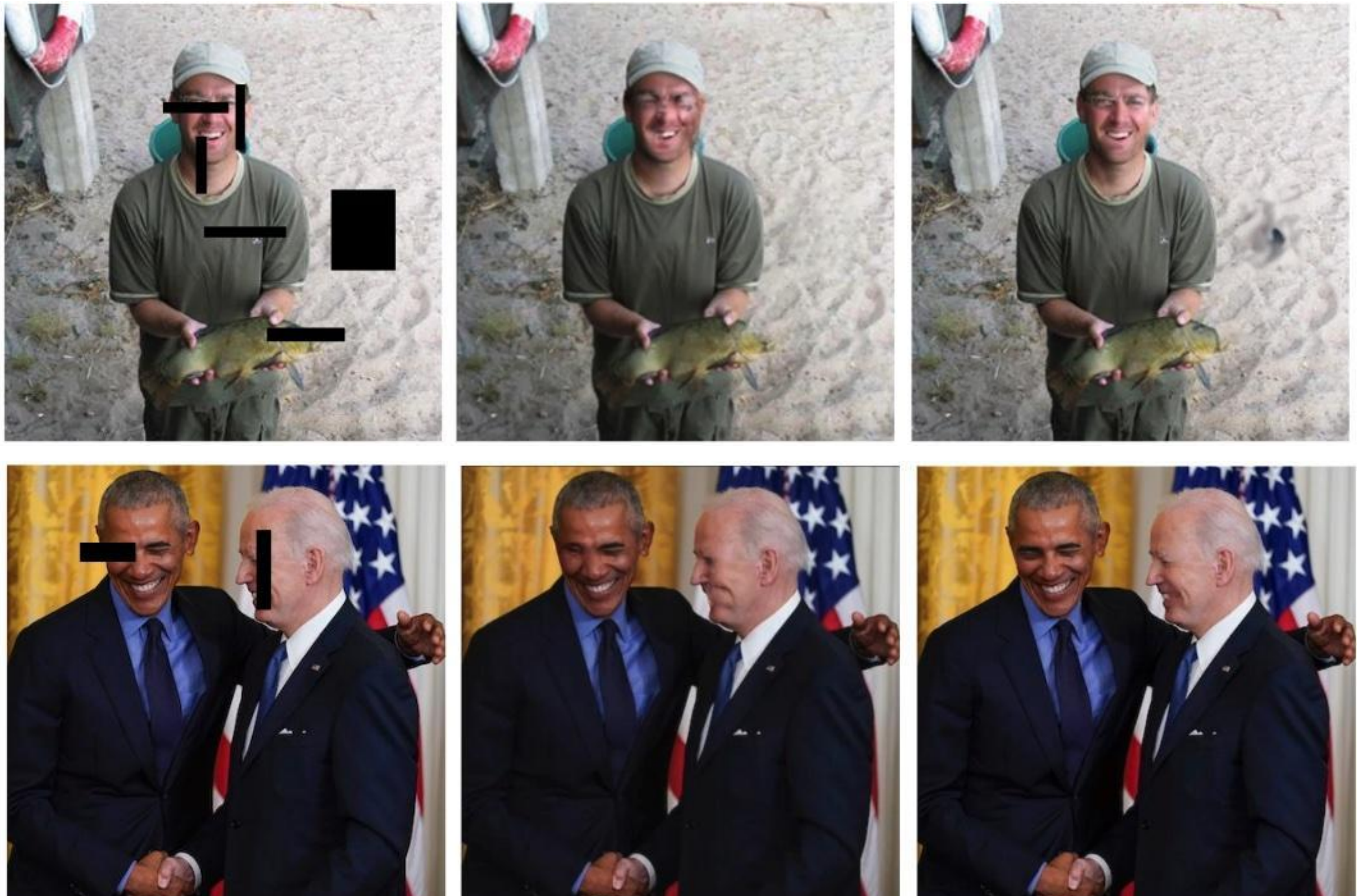
Gluing Loss: Known piece of D(z*)=y. Unknown piece is y=AD(z). Add them together to get an image with boundary effect x_{Glued}= Concat[AD(z) , y]
We want z to be equal to E(x_{Glued})
New loss term is || z- E(x_{Glued}) || this is the gluing loss objective. Leads to PSLD algorithm

Foundational Idea-> New algorithm (PSLD)

Input

Previous SOTA
(DPS in SD)

PSLD (Ours)



(a) Input (b) Groundtruth (c) DPS [11] (d) PSLD (Ours)

Figure 13: Gaussian deblur results on ImageNet 256 [17] (out-of-distribution).

Quantitative Comparisons

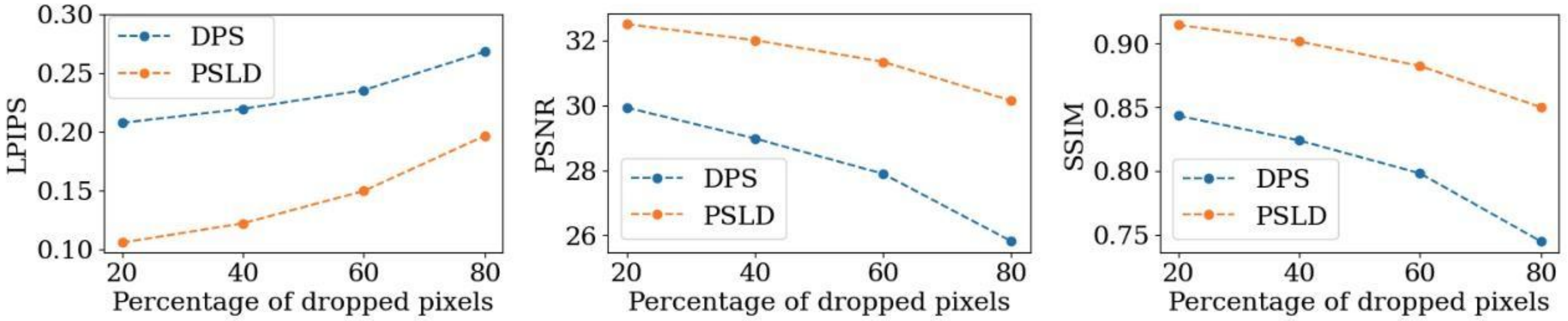
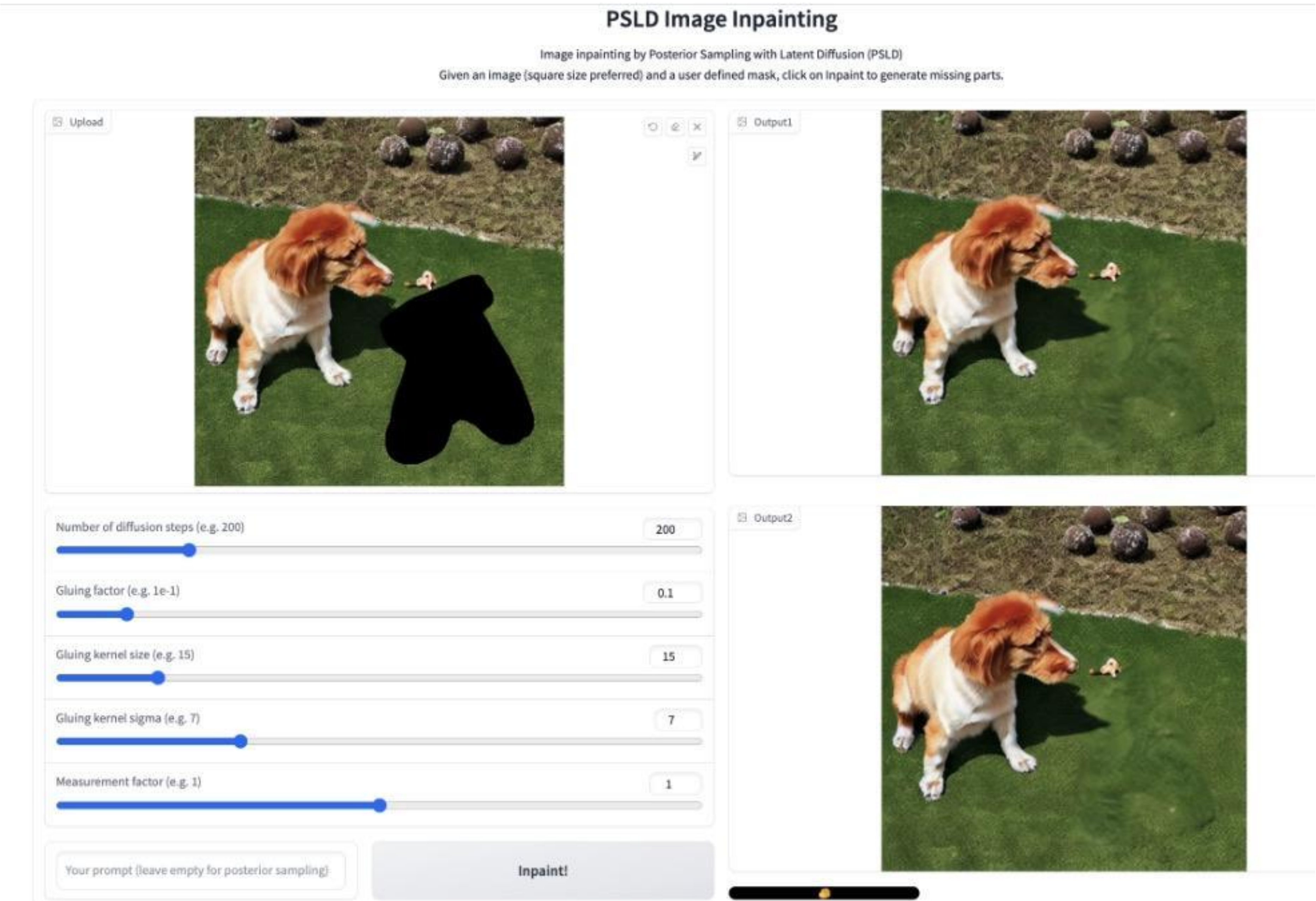


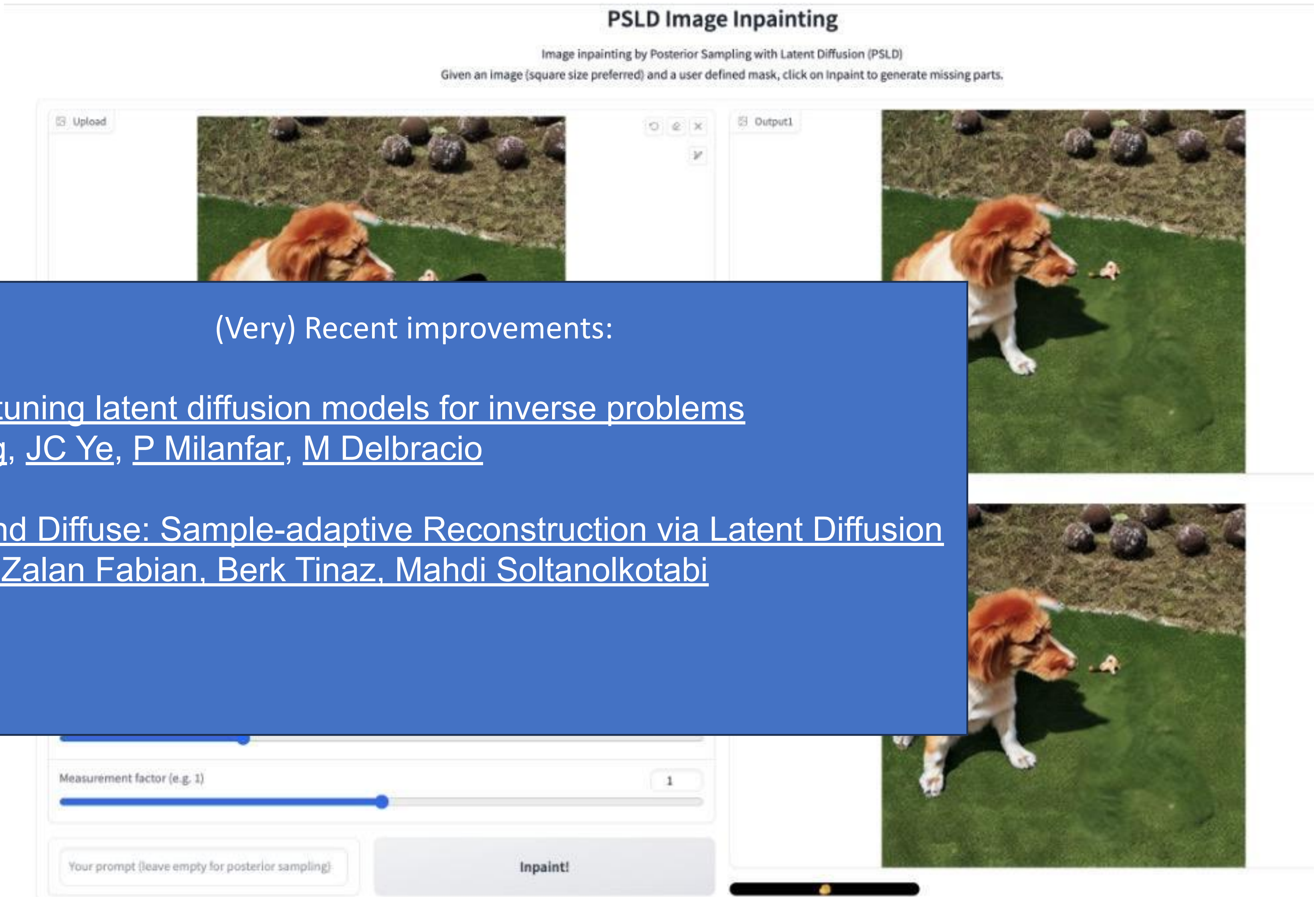
Figure 5: Comparing DPS and PSLD performance in random inpainting on FFHQ 256 [25, 11], as the percentage of masked pixels increases. PSLD with Stable Diffusion outperforms DPS.

PSLD: Solving Linear Inverse Problems Provably via Posterior Sampling with Latent Diffusion Models
Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G. Dimakis, Sanjay Shakkottai
https://arxiv.org/pdf/2307.00619.pdf

Web-demo of our method



Web-demo of our method



(Very) Recent improvements:

Prompt-tuning latent diffusion models for inverse problems
H Chung, JC Ye, P Milanfar, M Delbracio

Adapt and Diffuse: Sample-adaptive Reconstruction via Latent Diffusion Models, Zalan Fabian, Berk Tinaz, Mahdi Soltanolkotabi

Talk overview

- ~~Part 1. Creating large multimodal datasets (LAION and Datacomp)~~

- ~~Part 2.Teaching existing Foundation models new tricks~~

~~-Inpainting using Stable Diffusion~~
~~-Finetuning with corrupted data~~
-Teaching them custom concepts.

Multiresolution Textual inversion: Teaching new concepts to Stable Diffusion.

A pre-trained Stable Diffusion can produce ‘a painting of a dog.’

But what if you a painting of **your** dog.

Learning concepts from a few sample images by fine-tuning stable diffusion.

Related to DreamBooth and Textual inversion but with different pseudowords trained at each level of the diffusion.

https://arxiv.org/abs/2211.17115
Multiresolution Textual Inversion, Daras and Dimakis, Neurips Workshop 2022

Multiresolution Textual inversion

‘An image of <alex>’

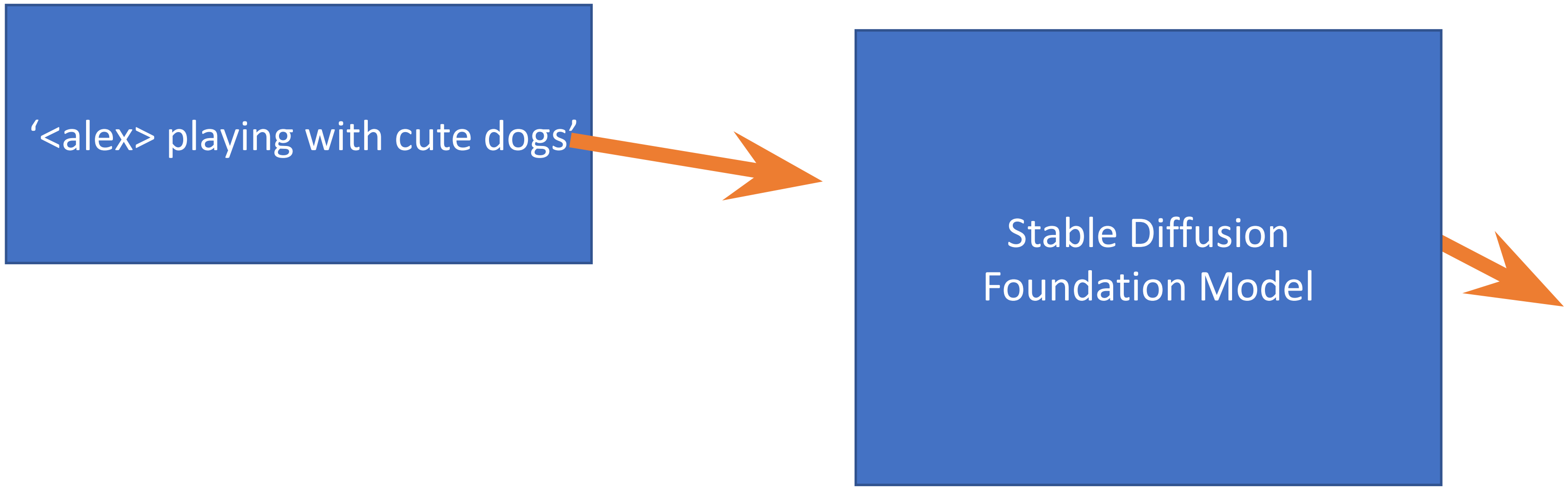
Stable Diffusion
Foundation Model

Loss

Perform CSGM:
Train xxx so that min|| G(‘<alex>’)- target||

https://arxiv.org/abs/2211.17115
Multiresolution Textual Inversion, Daras and Dimakis, Neurips Workshop 2022

Use learned pseudoword xxx in sentence



https://arxiv.org/abs/2211.17115
Multiresolution Textual Inversion, Daras and Dimakis, Neurips Workshop 2022

Use learned pseudoword xxx in sentence

‘<alex> playing with cute dogs’

Stable Diffusion Foundation Model

Use learned pseudoword xxx in sentence

‘<alex> as agent Smith in the Matrix movie’

Stable Diffusion Foundation Model

https://arxiv.org
Multiresolution

Fine Tuning on Ten Photographs of Jay Hartzell

jayH as a spartan gladiator in american football court. closeup, detailed face.

jayH as a spartan gladiator closeup, detailed face.

4:43 PM

Multiresolution textual inversion

Concept we want to learn:
<Jane>

Given a few images, we learn pseudo-words
that represent a concept at different resolutions.

Then we want to Generate with the prompt:
A painting of a dog in the style of <jane>

https://arxiv.org/abs/2211.17115
Multiresolution Textual Inversion, Daras and Dimakis, Neurips Workshop 2022

Multiresolution textual inversion

Given a few images, we learn pseudo-words
that represent a concept at different resolutions.

A painting of a dog in the style of <jane(0)>", <jane(0.5)> <jane(0.7)>
gives different levels of freedom to match the <jane> style based on the number index.

https://arxiv.org/abs/2211.17115
Multiresolution Textual Inversion, Daras and Dimakis, Neurips Workshop 2022

Learning different concepts at different resolutions

(a) Inputs learned as <jane> and <cat>.

Use prompt:

`Generate an image of a` **`<cat:Low Res>`**
`made with` **`<Jane, Details only>`**

Learning different concepts at different resolutions

(a) Inputs learned as <jane> and <cat>. (c) A photo of <cat(5)> made with <jane|0.1|>.

Our method allows combining different objects at different resolutions.

For example, given a painting made of buttons and a toy cat as inputs, we can generate an image that has the shape of the cat and the texture of the buttons.

Conclusions

- Part 1.Making better multimodal datasets: DataComp

Part 2: Teaching Foundation Models new tricks:

- 1. PSLD: Solving Linear Inverse Problems Provably via Posterior Sampling with Latent Diffusion Models

- 2. MultiResolution Textual Inversion: Customizing Foundation models to new datasets.

- 3. Ambient Diffusion: Learning Clean Distributions from Corrupted Data.

- Future directions: Uncertainty Quantification, Reducing training set memorization, Personalizing FMs: Multiresolution Textual Inversion, DreamBooth, etc. Improving Diversity, Reducing memorization.
- Training Foundation Models for MRI, Proteins and other domains.

Fin

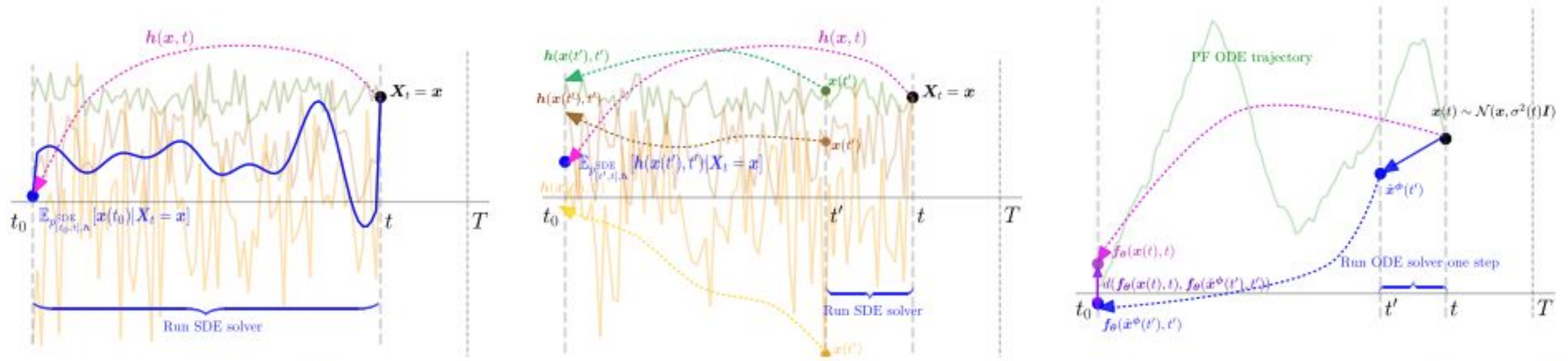
Enforcing Consistency in Diffusion Models

News: There are 3 types of Consistency in Diffusions that appeared in the recent literature:

On the Equivalence of Consistency-Type Models

Table 1. Comparison of existing consistency-type models.

Models	Purpose	Trajectory	Object of Eq.	Approach
CDM (Daras et al., 2023)	Sample quality	Backward SDE	Samples	DSM + Martingale regularizer
CM (Song et al., 2023)	Sampling speed	PF ODE	Samples	Specific NN structure + New training scheme
FP-Diffusion (Lai et al., 2022)	Likelihood	Score FPE (a PDE)	Scores	DSM + Score FPE-regularizer



(a) Illustration of Def. 3.1. A consistent SDE-denoiser indicates that the SDE-denoiser prediction $\boldsymbol{h}(\boldsymbol{x}, t)$ (endpoint of the magenta arrow) aligns with the average of SDE predictions $\mathbb{E}_{p^{\mathrm{SDE}}_{[t_0,t],\boldsymbol{h}}}[\boldsymbol{x}(t_0)|\boldsymbol{X}_t = \boldsymbol{x}]$ (blue dot).

(b) Illustration of Prop. 3.2. The SDE-denoiser prediction $\boldsymbol{h}(\boldsymbol{x}, t)$ (endpoint of the magenta arrow) aligns with the average prediction of intermediate points obtained by first applying an SDE solver and subsequently applying the SDE-denoiser $\mathbb{E}_{p^{\mathrm{SDE}}_{[t',t],\boldsymbol{h}}}[\boldsymbol{h}(\boldsymbol{x}(t'), t')|\boldsymbol{X}_t = \boldsymbol{x}]$ (blue dot).

(c) Illustration of Alg. 2 in (Song et al., 2023). The objective of CM is to align the prediction of the direct denoiser (endpoint of the magenta arrow) with the prediction obtained by first applying a one-step ODE solver and subsequently applying the denoiser.

Figure from: “On the Equivalence of Consistency-Type Models: Consistency Models, Consistent Diffusion Models, and Fokker-Planck Regularization”
CH Lai, Y Takida, T Uesaka, N Murata et al.

DataComp Baselines

Scale	Filtering strategy	Training dataset size	ImageNet	ImageNet dist. shifts	Average over 38 datasets
small	No filtering	13M	0.025	0.033	0.132
	Text-based	3M	0.046	0.052	0.156
	Image-based	5M	0.035	0.045	0.146
	CLIP score (L/14 30%)	4M	0.051	0.055	0.172
medium	No filtering	128M	0.176	0.152	0.254
	Text-based	31M	0.255	0.215	0.301
	Image-based	45M	0.238	0.198	0.292
	CLIP score (L/14 30%)	38M	0.273	0.230	0.323
large	No filtering	1B	0.459	0.378	0.428
	Text-based	317M	0.561	0.465	0.466
	Image-based	449M	0.527	0.433	0.461
	CLIP score (L/14 30%)	384M	0.578	0.474	0.520
xlarge	CLIP score (L/14 30%)	1.8B	**0.769**	**0.656**	**0.634**

Scaling with dataset size

Performance of filtering methods consistent across tracks: