

Evaluating Text-to-Image Models

Shobhita Sundaram

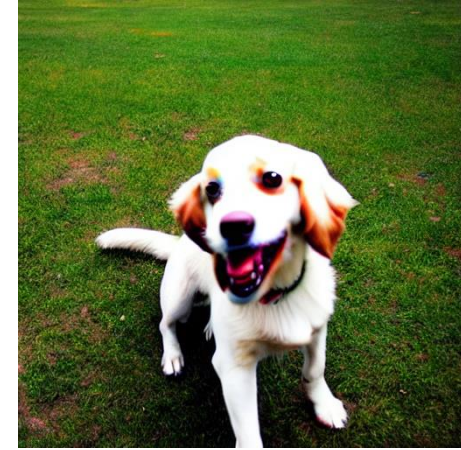
"Generate a photo of a dog playing outside"



- ✓ Shows a dog
- ✗ Not a photo



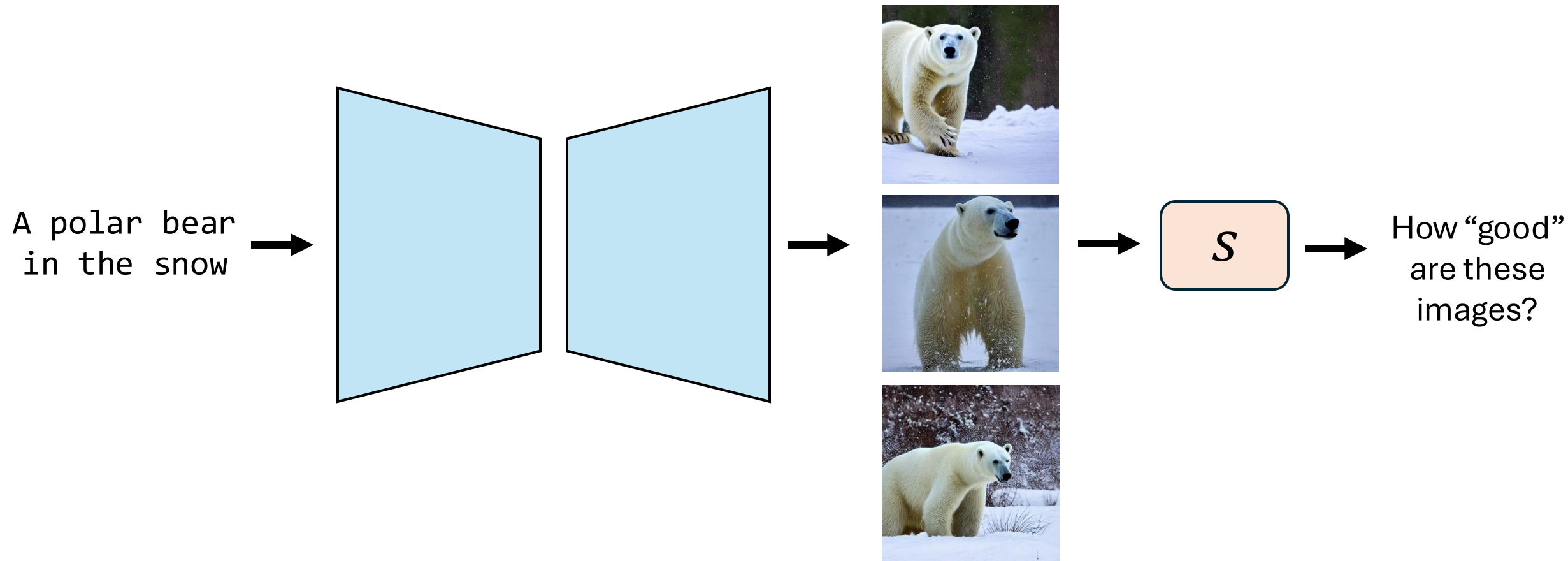
- ✓ Dog is playing
- ✓ Aesthetically pleasing



- ✓ Is a photo
- ✗ Strange lighting/artifacts

How do we evaluate generative models and their outputs?

Evaluating T2I models



Agenda

- What are the current image evaluation metrics?
- What are the best/most popular metrics for T2I models?
- How do you design a good metric that reflects human preferences?

Agenda

- **What are the current image evaluation metrics?**
- What are the best/most popular metrics for T2I models?
- How do you design a good metric that reflects human preferences?

What are the tools for image evaluation?

	Low-Level	High-Level
Holistic $s(x)$	Blurriness, No-Reference IQA	PickScore, ImageReward
Similarity $s(x, x_{ref})$	PSNR, SSIM, LPIPS, DISTS	DreamSim
Distribution $s(p(x), p_{ref})$	FID, InceptionScore, CMMD	
Text-Alignment $s(x, y_{ref})$	SOA, CLIPScore	

	Low-Level	High-Level
Holistic $s(x)$	Blurriness, No-Reference IQA	PickScore, ImageReward

“A cat on a propaganda poster”



“A demon exiting through a portal...”









Agenda

- What are the current image evaluation metrics?
- **What are the best/most popular metrics for T2I models?**
- How do you design a good metric that reflects human preferences?

	Low-Level	High-Level
Holistic $s(x)$	Blurriness, No-Reference IQA	PickScore, ImageReward
Similarity $s(x, x_{ref})$	PSNR, SSIM, LPIPS, DISTS	DreamSim
Distribution $s(p(x), p_{ref})$	FID, InceptionScore, CMMD	
Text-Alignment $s(x, y_{ref})$	SOA, CLIPScore	

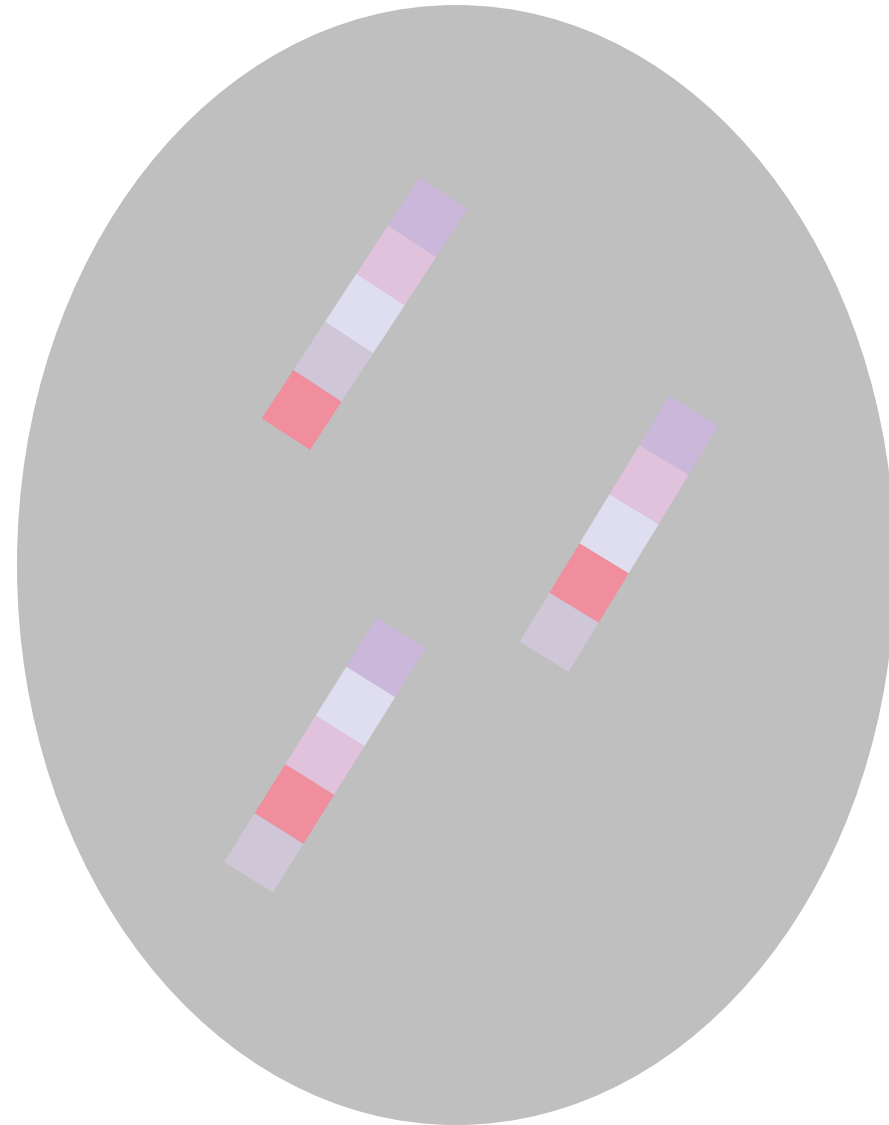
Why compare image distributions?

Caption	Generated Image	Real Image
<p>A shoe rack with some shoes and a dog sleeping on them.</p>	 A generated image showing a golden retriever dog lying on top of a wooden shoe rack. The rack is filled with various styles of shoes, including sneakers and dress shoes.	 A real image showing a fluffy brown dog sitting inside a wire cage. There are some items in the cage, including a red container and a black bag with the word "RUCANOR" on it.
<p>Bunk bed with a narrow shelf sitting underneath it</p>	 A generated image of a modern, light-colored metal bunk bed in a room. A narrow shelf is visible underneath the lower bunk.	 A real image of a wooden bunk bed in a room. A white chest of drawers is visible underneath the lower bunk.
<p>A table full of food such as peas and carrots, bread, salad and gravy</p>	 A generated image of a table set with a variety of dishes, including a roasted chicken, bread, peas, carrots, and gravy.	 A real image of a table set with a variety of dishes, including a large flatbread, salad, and other food items.

How do we compare image distributions?



How do we compare image distributions?



FID & CMMD slides

Agenda

- What are the current image evaluation metrics?
- What are the best/most popular metrics for T2I models?
- **How do you design a good metric that reflects human preferences?**

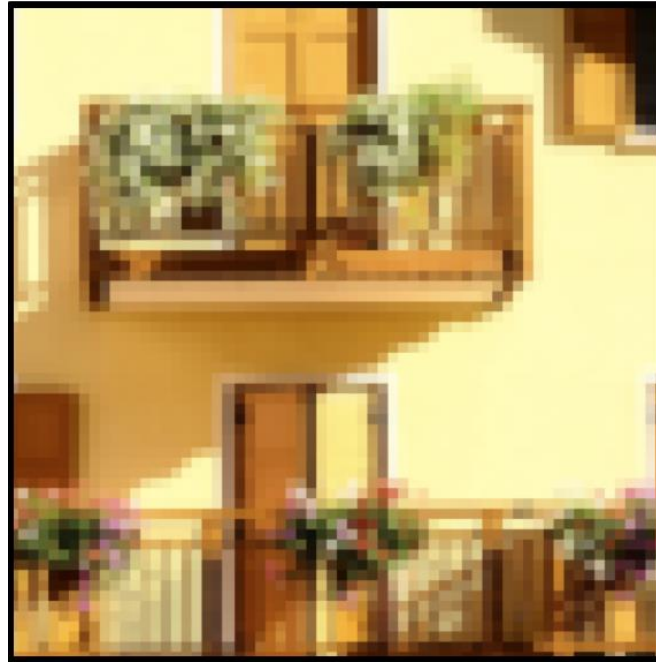
	Low-Level	High-Level
Holistic $s(x)$	Blurriness, No-Reference IQA	PickScore, ImageReward
Similarity $s(x, x_{ref})$	PSNR, SSIM, LPIPS, DISTS	DreamSim
Distribution $s(p(x), p_{ref})$	FID, InceptionScore, CMMD	
Text-Alignment $s(x, y_{ref})$	SOA, CLIPScore	

D ( , )

Which patch is more similar to the middle?



< Clap >



Humans

L2/PSNR

SSIM/FSIMc

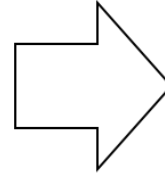
Deep Networks?



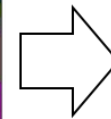
< C ✓ >

“Perceptual Losses”

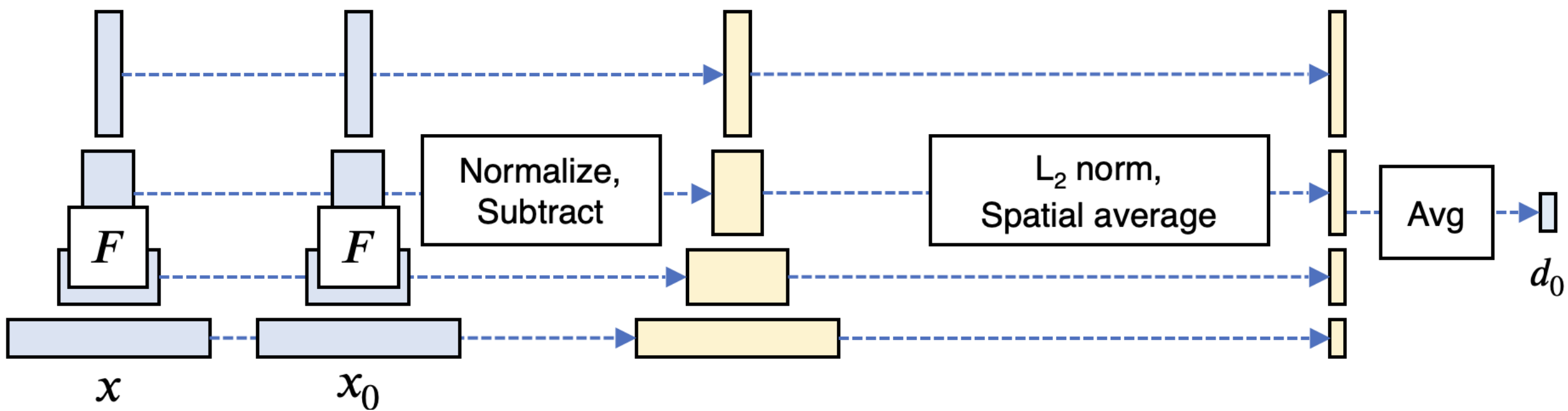
Gatys et al. In CVPR, 2016.
Johnson et al. In ECCV, 2016.
Dosovitskiy and Brox. In NIPS, 2016.



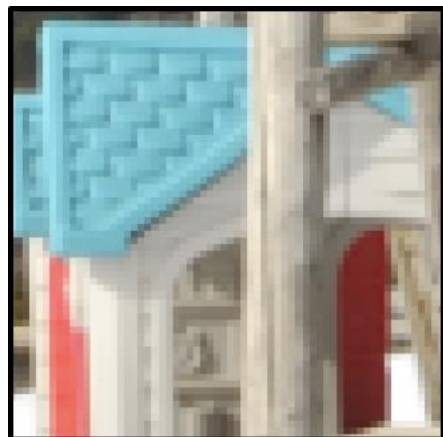
Chen and Koltun. In ICCV, 2017.



Deep Networks as a Perceptual Metric

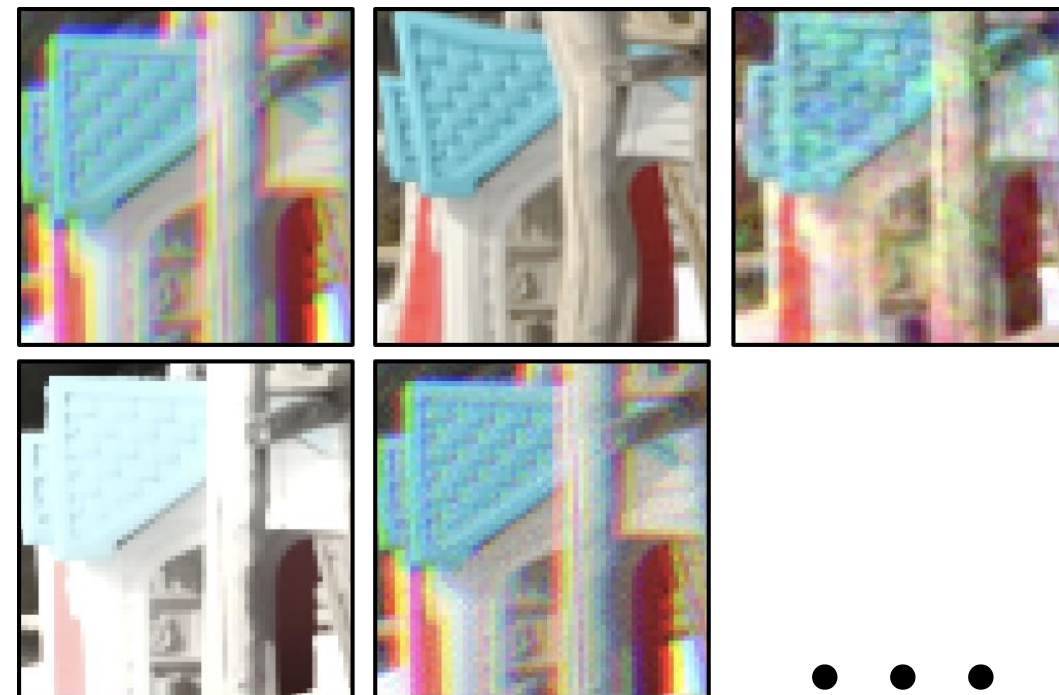


Distortions



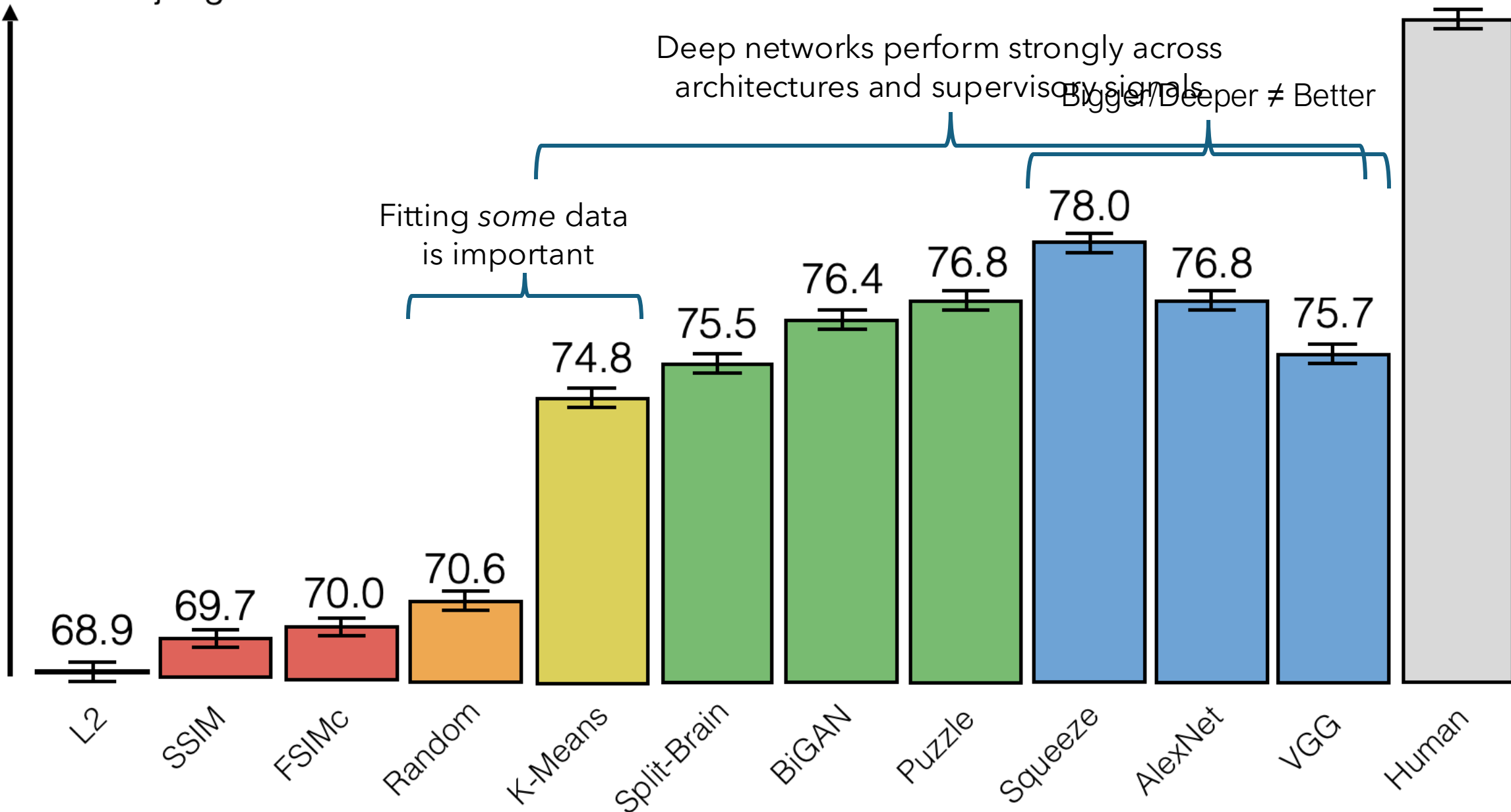
Original Patch

Noise
Photometric
Spatial warps
Compression
Blur



Distorted Patches

% agreement with human judges



How different are *these* images?

D (





,



)

DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data

 <https://dreamsim-nights.github.io/> 



Stephanie Fu^{*1}



Netanel Y. Tamir^{*2}



Shobhita Sundaram^{*1}



Lucy Chai¹



Richard Zhang³



Tali Dekel²



Phillip Isola¹

^{*}Equal contribution, order decided by random seed



Which image, A or B, is more similar to the reference?

A



Reference



B



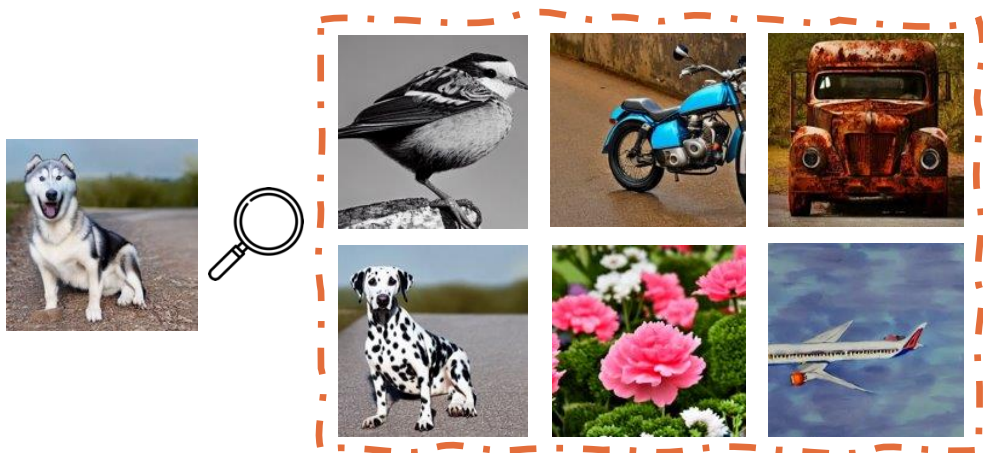
 LPIPS  DINO  CLIP

 Humans

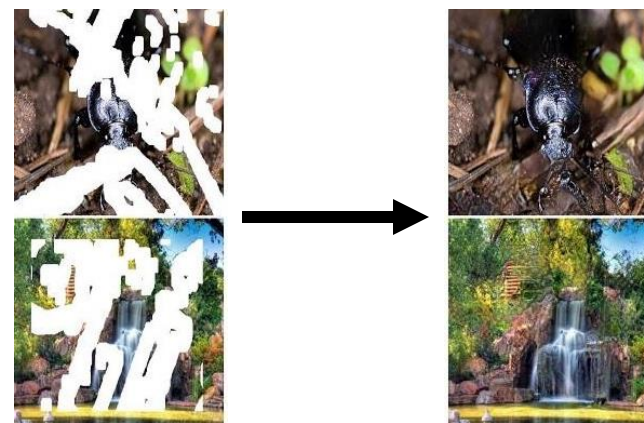
 DreamSim

$$f\left(\text{img}_1, \text{img}_2\right) = d$$

Image retrieval

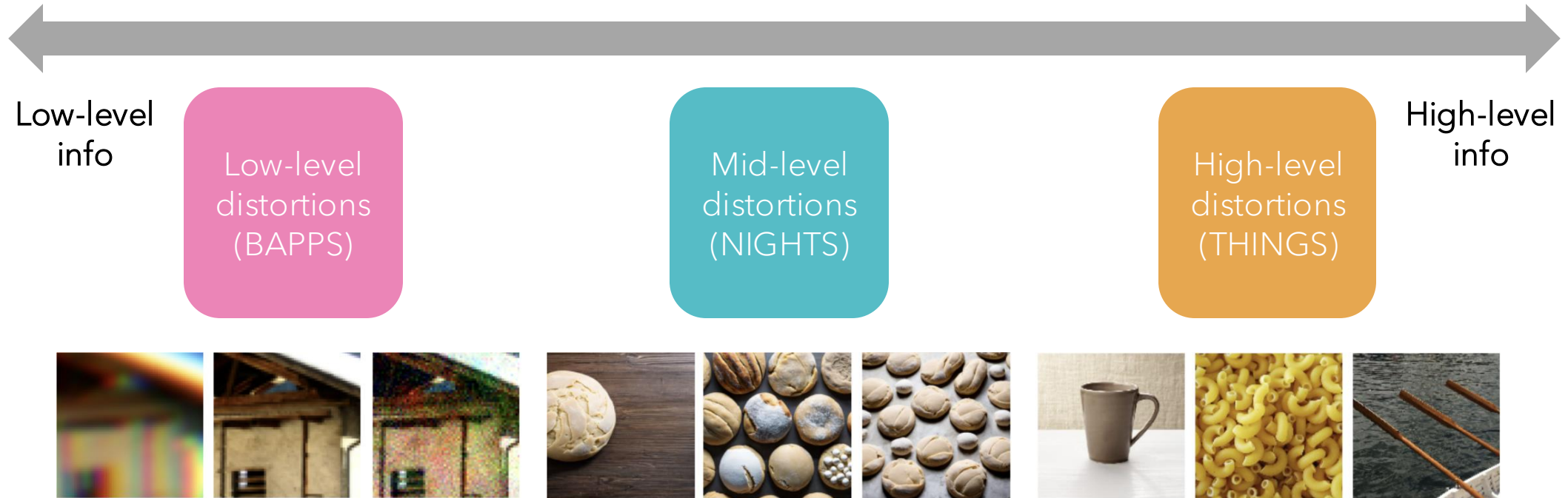


Loss function



Liu et al, Image Inpainting for Irregular Holes Using Partial Convolutions, *ECCV 2018*

Perceptual similarity datasets





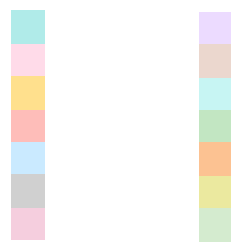
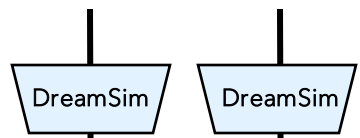
Low-level
info

Low-level
distortions
(BAPPS)

Mid-level
distortions
(NIGHTS)

High-level
distortions
(THINGS)

High-level
info



$D (\text{bread image} , \text{bread image})$

NIGHTS – Novel Image Generations with Human-Tested Similarity

Goal: create a dataset of triplets which exhibit changes in **mid-level** information

“An image of
a **ski lodge**”



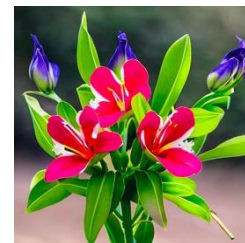
Stable
Diffusion



3 seeds

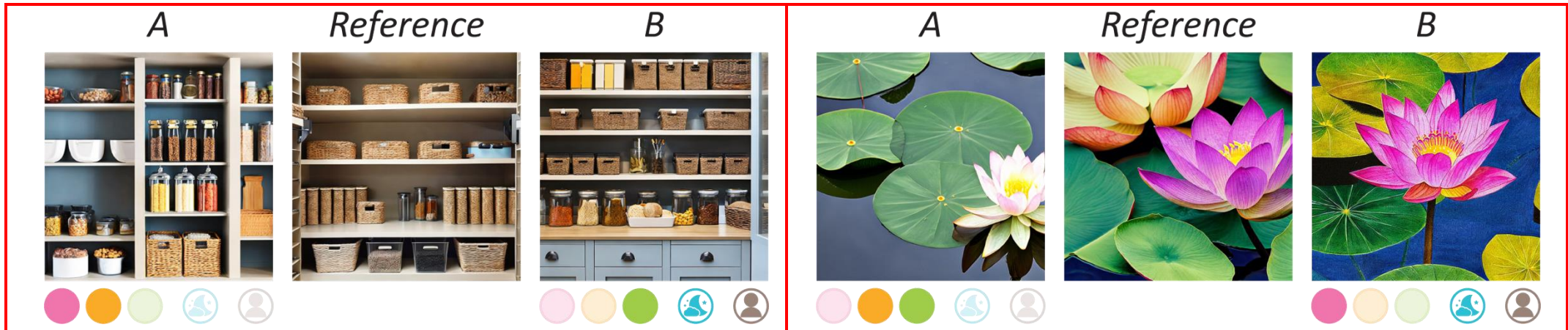
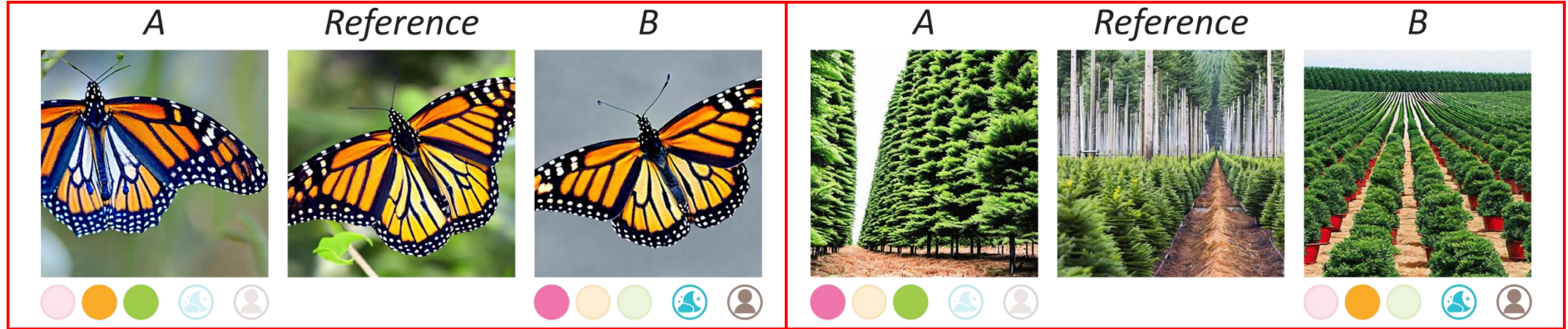
Two-alternative forced choice (2AFC) test

Which image, A or B, is more similar to the
reference?



- ~20k **synthetic** image triplets with unanimous human votes
- Average of 7 votes per triplet
- Classes taken from ImageNet, Food-101, SUN397, etc.

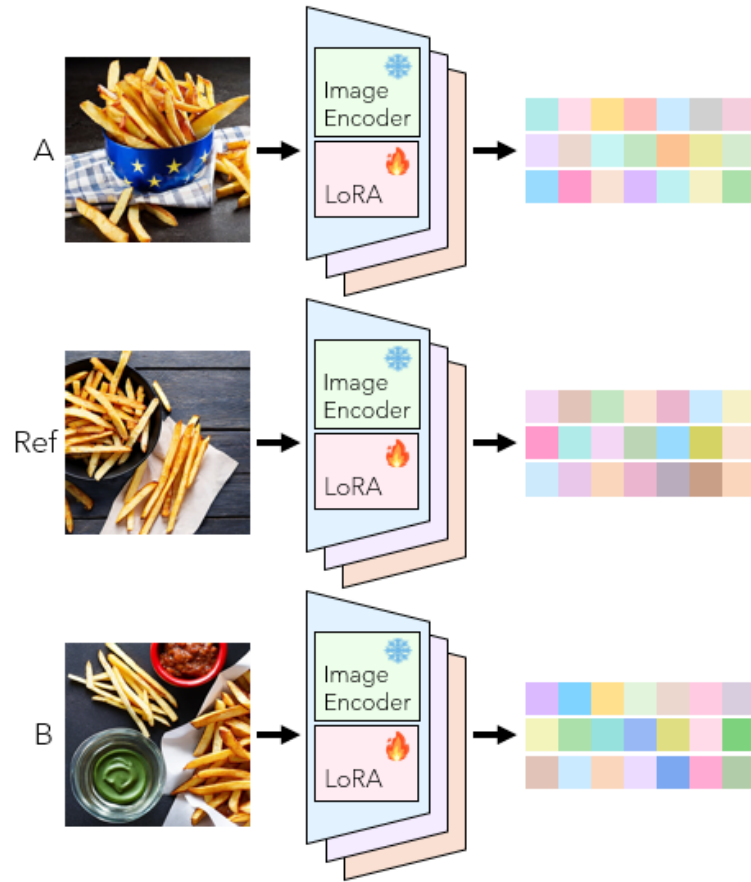
Examples of NIGHTS triplets



 LPIPS  DINO  CLIP  DreamSim  Humans

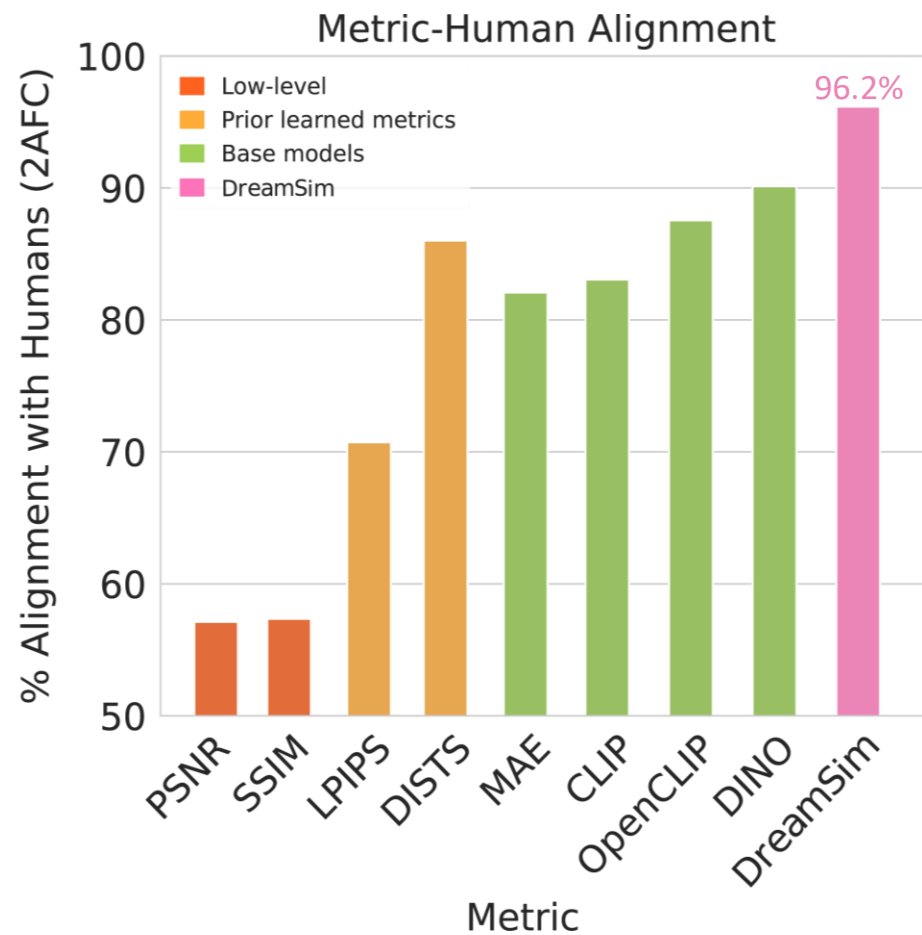
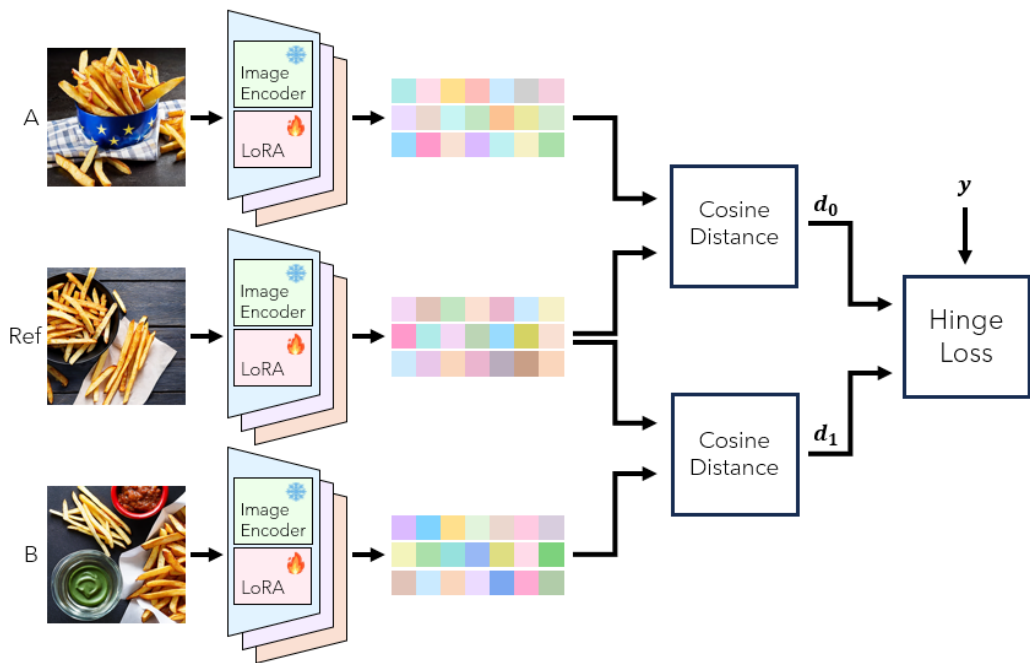
Training & Inference

Training: use hinge loss between distances (= triplet loss between embeddings)



Inference: cosine distance between embeddings of two images

Training & Inference



Nearest Neighbors

Input



Nearest Neighbors

LPIPS



DISTS



OpenCLIP



DINO



Ours



Generation

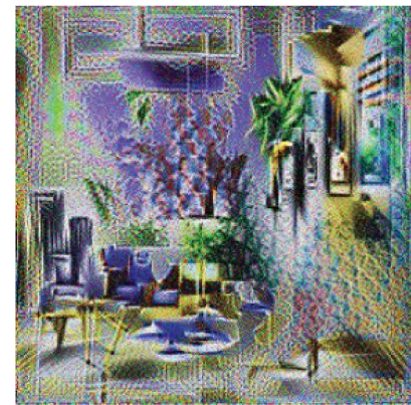
Target



OpenCLIP



DINO



Ours



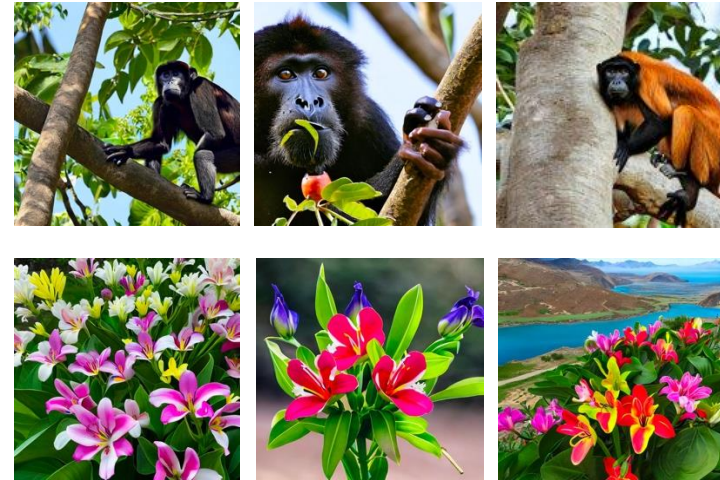
Guided Diffusion Optimization



Summary

NIGHTS dataset

- Diverse synthetic image triplets
 - Focus on mid-level distortions
- Labeled with human judgments



DreamSim

- Tuned on NIGHTS
- Applications in image retrieval
 - Performance generalizes to real images



Paper, Code & Dataset
dreamsim-nights.github.io

