Re-Imagen

Retrieval-grounded text-to-image generation

Presenter: Wenhu Chen

Collaborators

Hexiang Hu

Chitwan sahariac

William Cohen

Acknowledgement

William Chan

Jason Baldridge

Agenda

Existing Text-to-Image Models

Motivation

Model Design

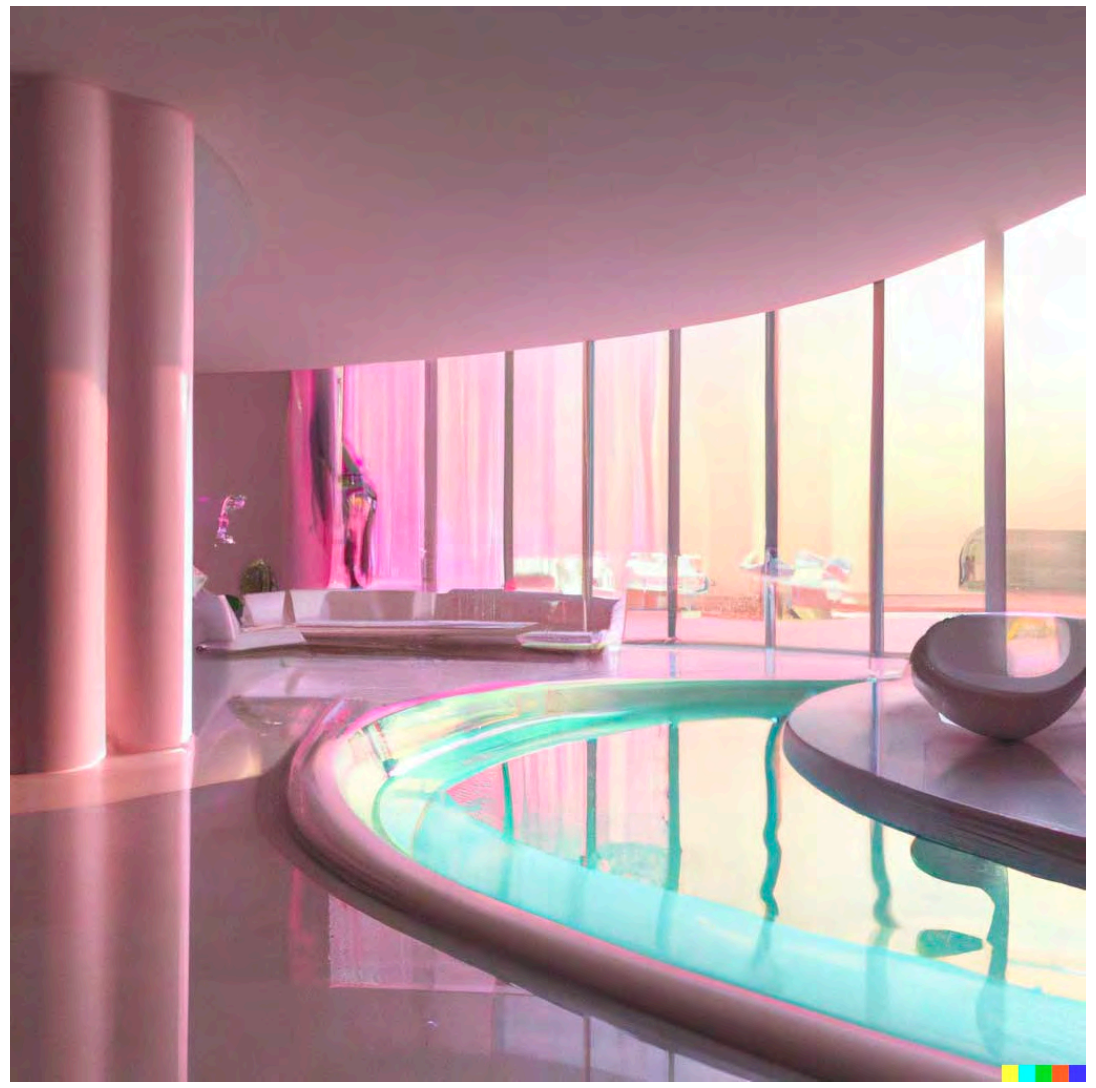
Experimental Results

Limitations and future directions

Existing Text-to-Image Models

Recent Progress in text-to-image generation

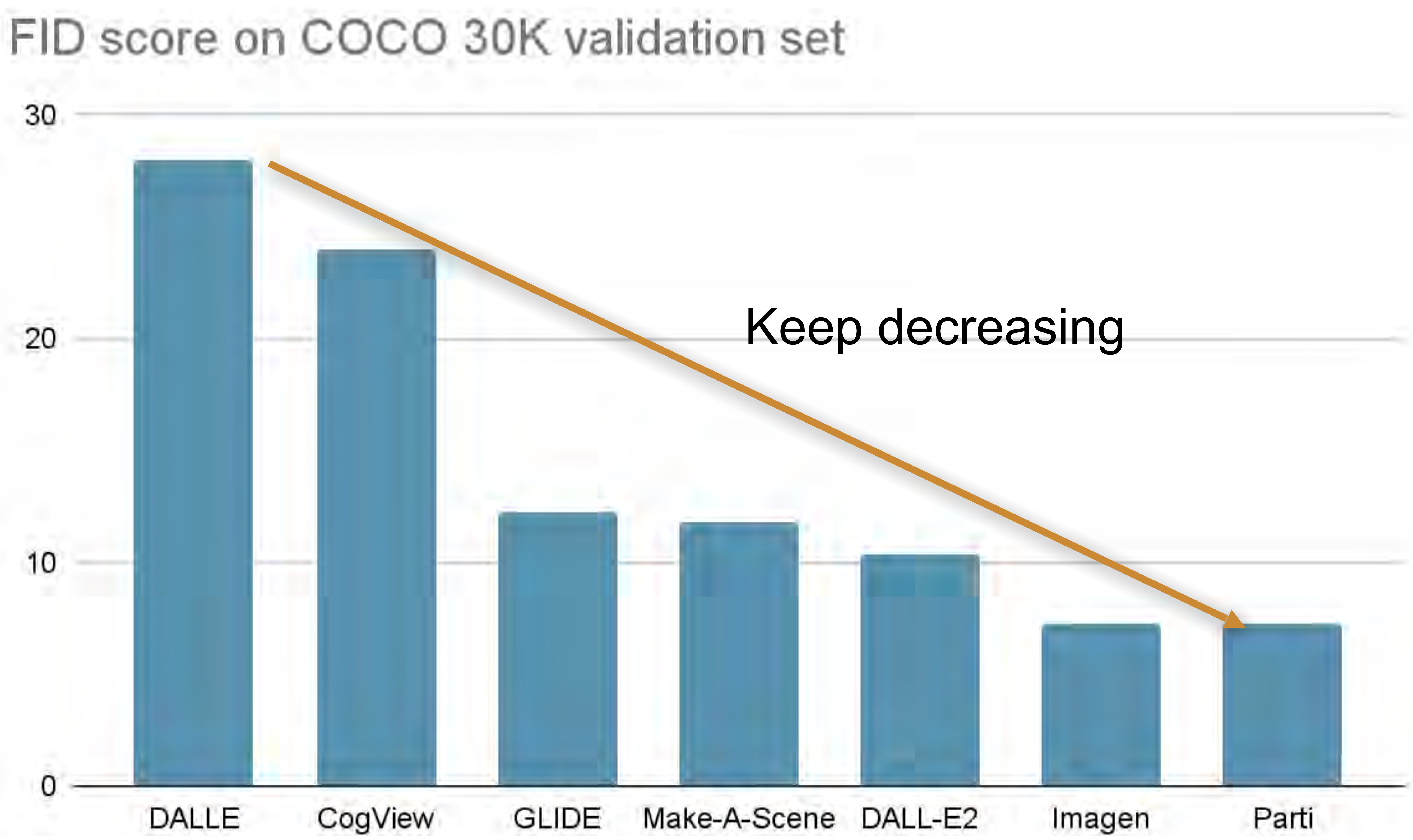
Imagen



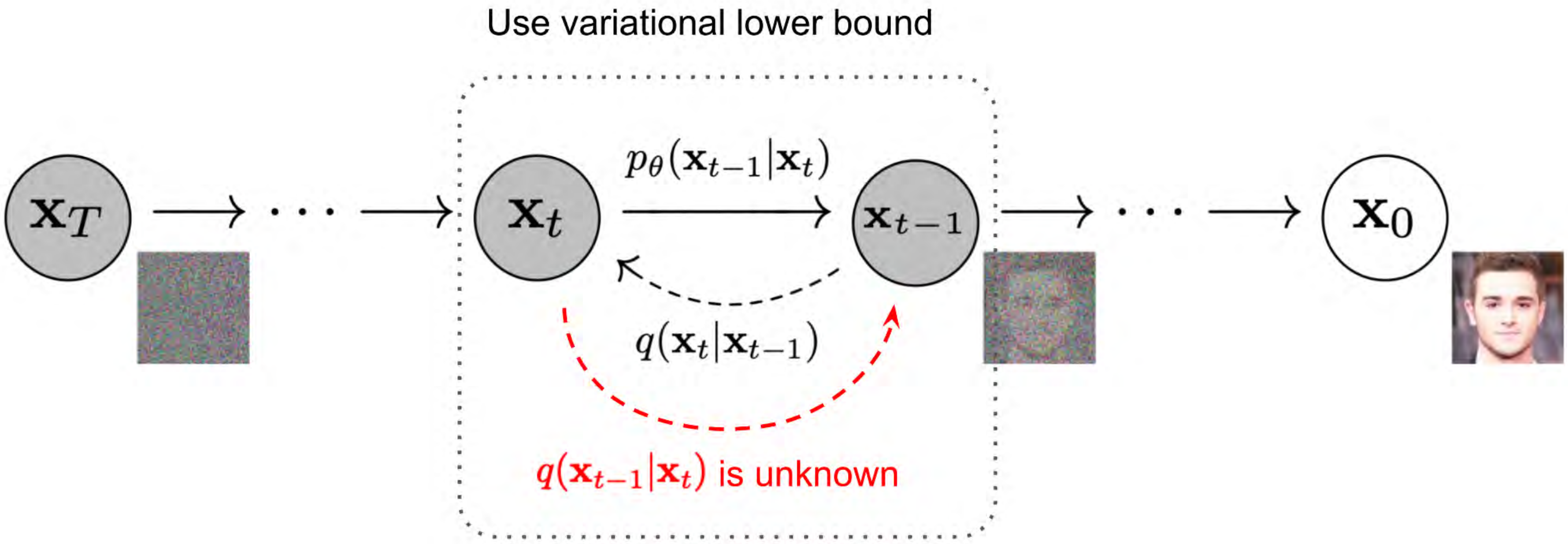
Dall-E2

Stable Diffusion

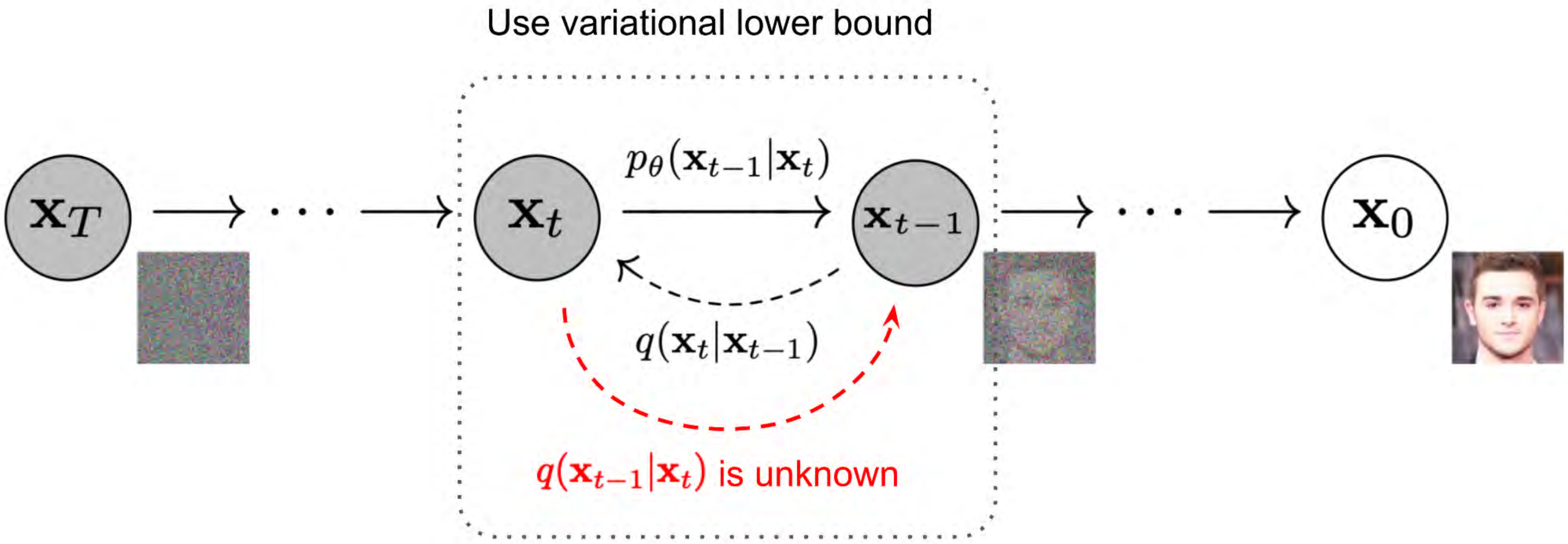
Recent Progress in text-to-image generation



Diffusion Model Training (Ho et al. 2020)

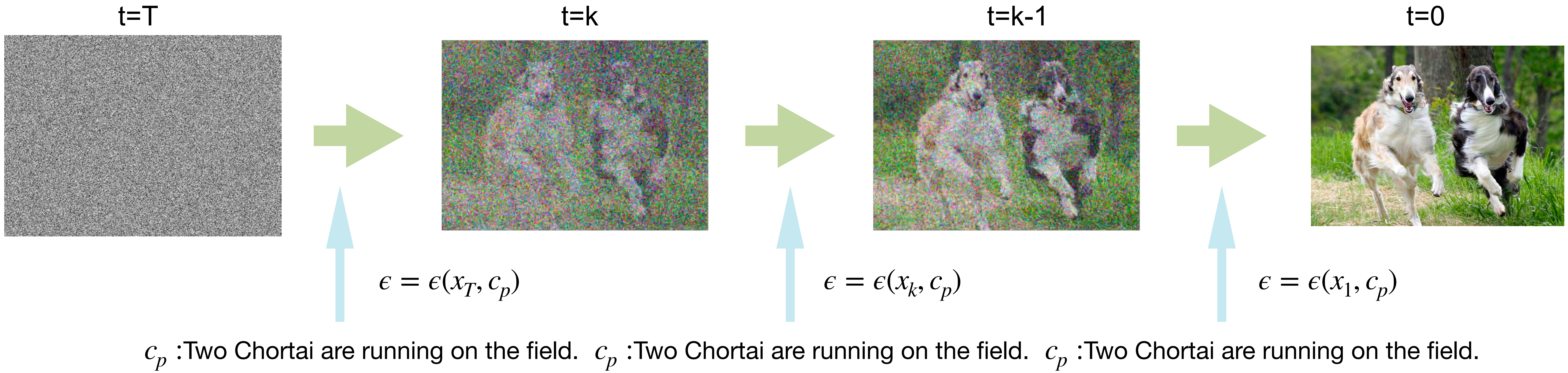


Diffusion Model Training (Ho et al. 2020)



$$E_q\Big[\sum_{t>1} KL(q(x_{t-1}|x_t, x_0)\,||\,p_\theta(x_{t-1}|x_t))\Big] \xrightarrow{\text{Reparameterization}} E_{x_0,\epsilon}[w_t\,||\,\epsilon - \epsilon_\theta(x_t(x_0,\epsilon), t)\,||^2]$$

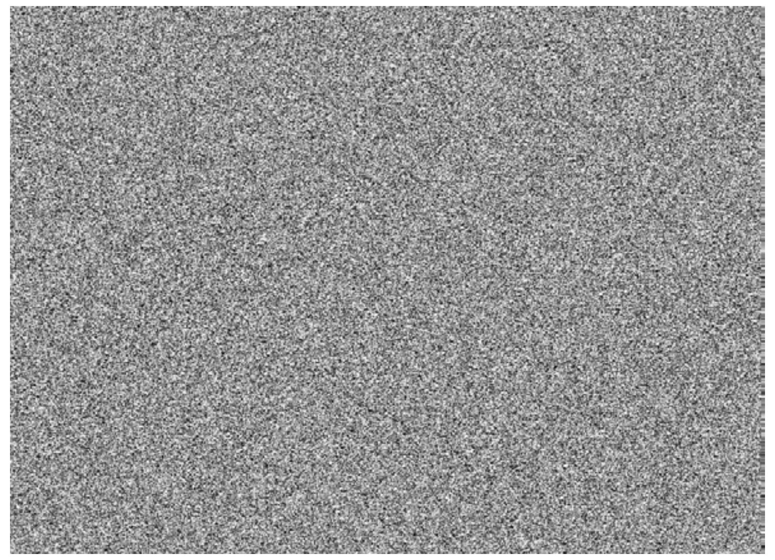
Diffusion Model Inference (Ho et al. 2020)



Classifier-free Guidance (Ho et al. 2022)

$$\epsilon = \gamma\epsilon(x_t, c_t) - (\gamma - 1)\epsilon(x_t)$$

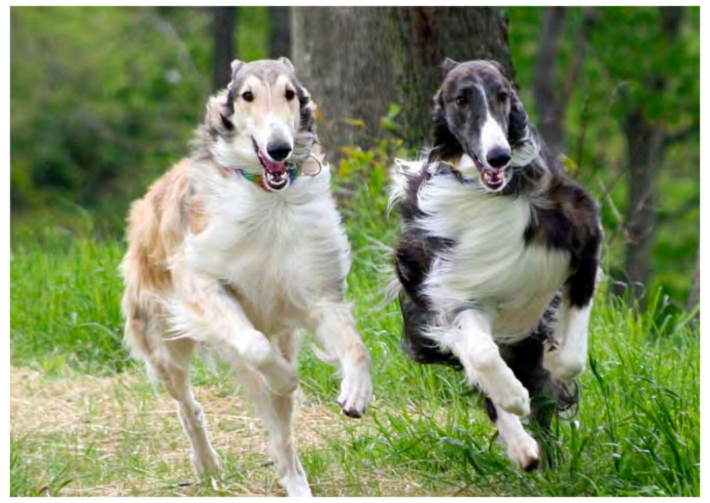
t=T



t=k

t=k-1

t=0



$\epsilon = \gamma\epsilon(x_T, c_p) - (\gamma - 1)\epsilon(x_T)$

c_p :Two Chortai are running on the field.

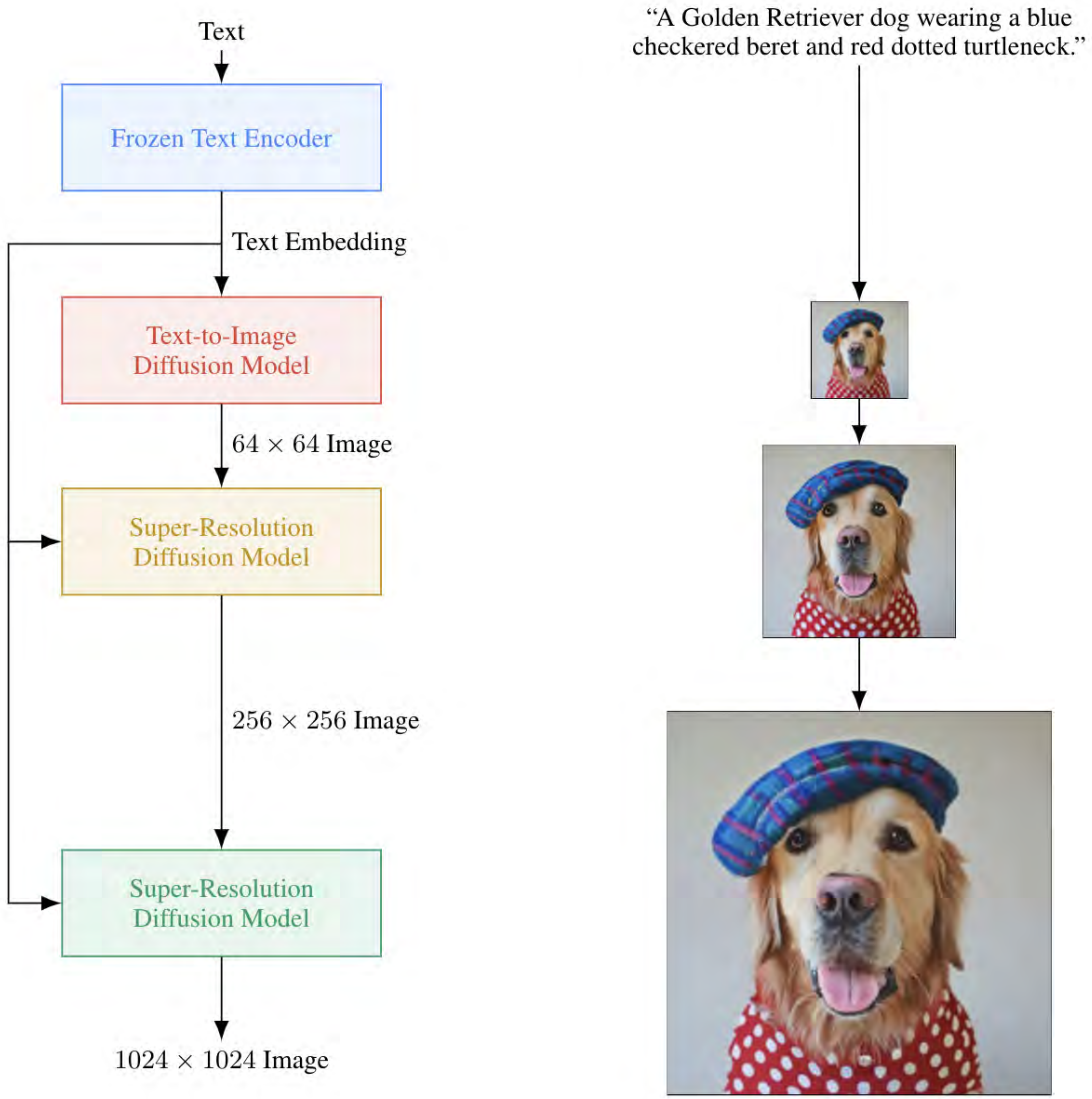
$\epsilon = \gamma\epsilon(x_k, c_p) - (\gamma - 1)\epsilon(x_k)$

c_p :Two Chortai are running on the field.

$\epsilon = \gamma\epsilon(x_1, c_p) - (\gamma - 1)\epsilon(x_1)$

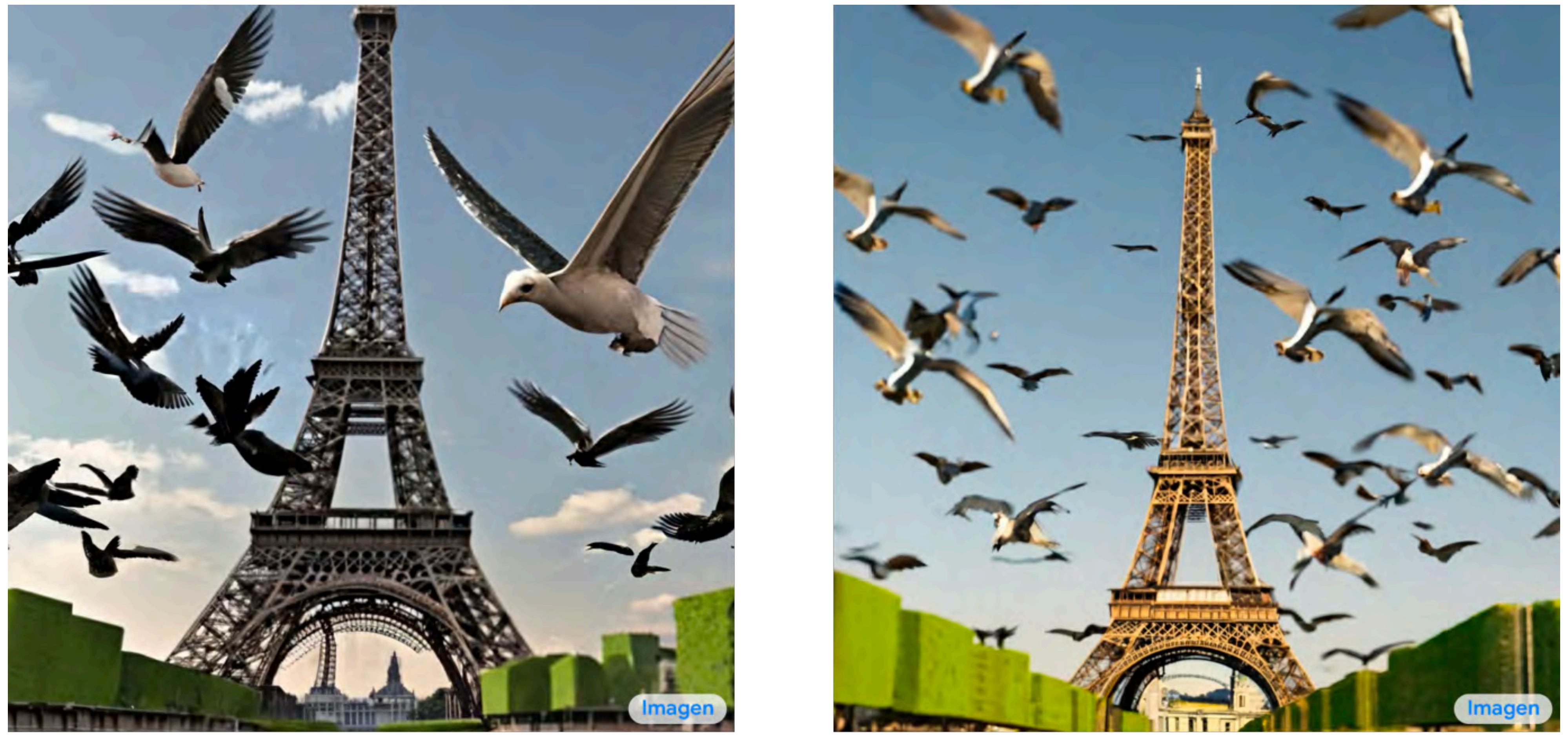
c_p :Two Chortai are running on the field.

Cascaded Diffusion Model (Saharia et al. 2022)



The models are really good at frequent entities/objects

A Golden Retriever is standing by the river.



Birds flying around Eiffel Tower.

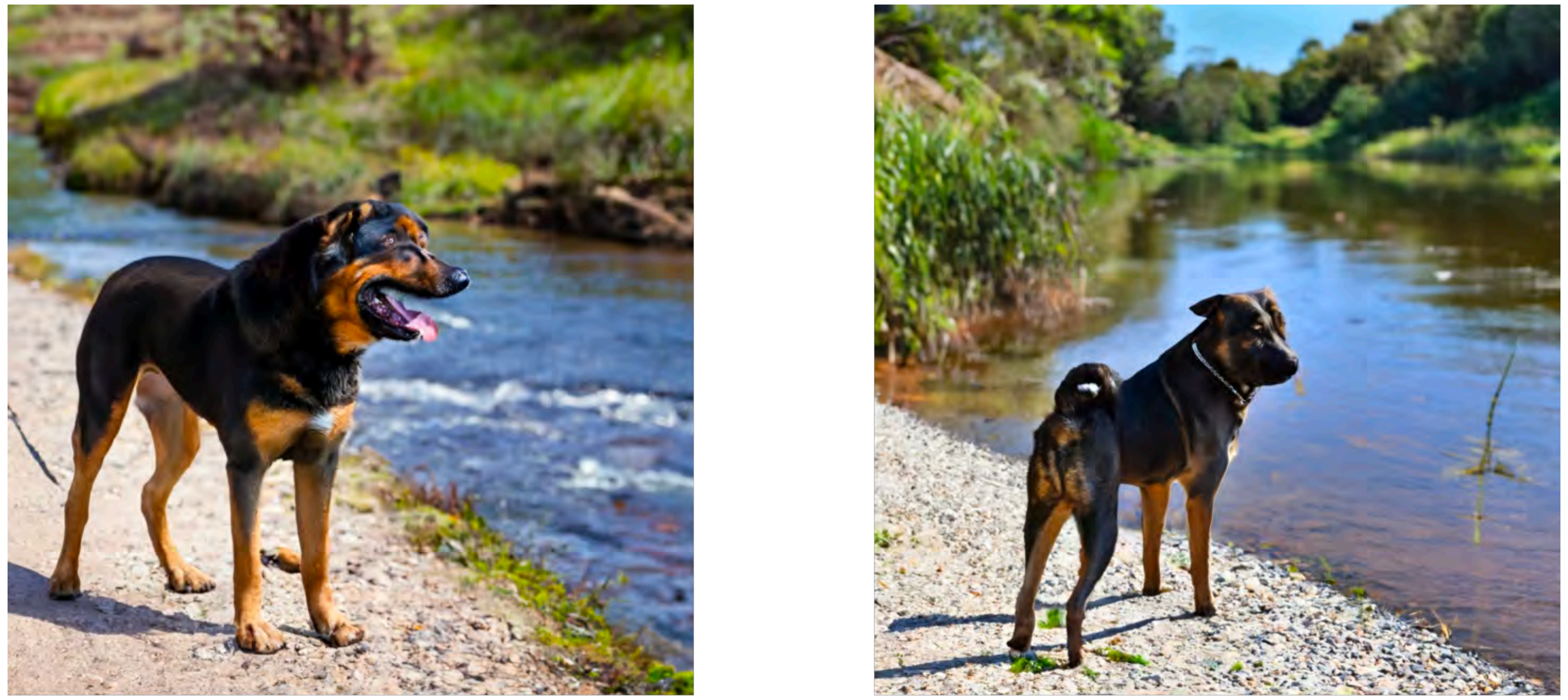
Peperoni Pizza is served with wine.

… not so good with infrequent entities/objects

Hawaiian Pizza is served with wine.

A Barbado da Terceira is standing by a river.

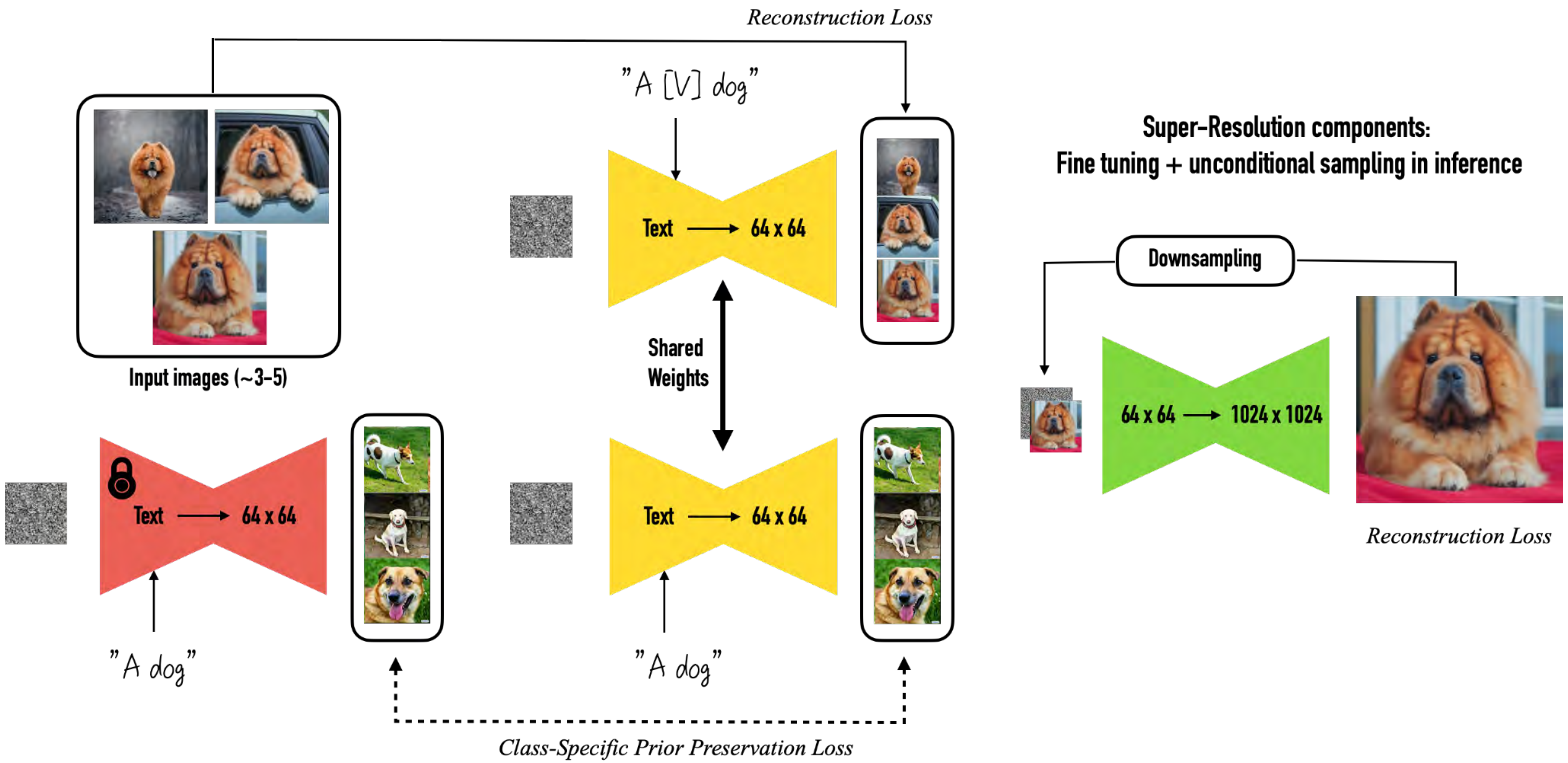
Barbado da Terceira



A Barbado da Terceira (dog) is standing by a river.

Potential Ways to address this? Fine-tune the model!

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. (Nataniel et al. 2022)

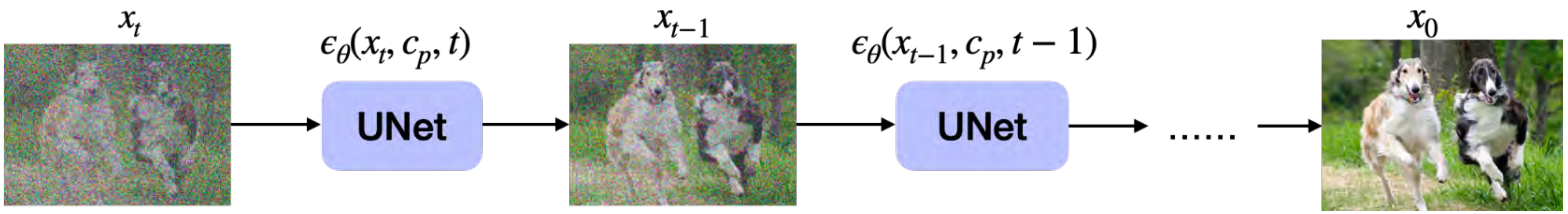


1. Expensive, requires 15 minutes fine-tuning for each new entity.
2. Require 3-5 images about the same entity.
3. Requires additional entity images of the same category to optimize prior preservation loss.

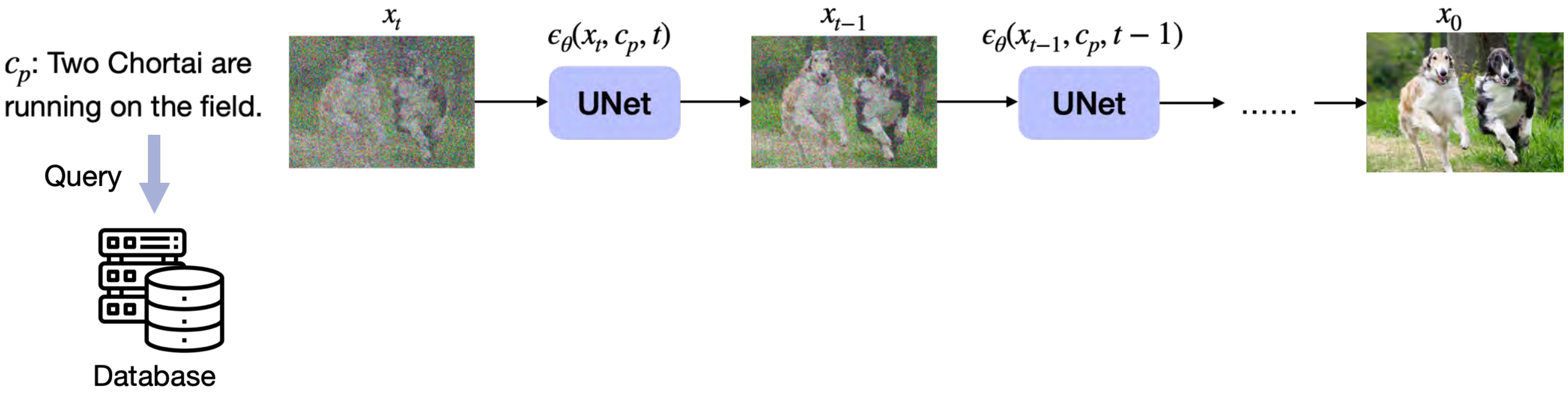
Re-Imagen: Retrieval Augmentation

Our approach: Retrieval-augmented Model

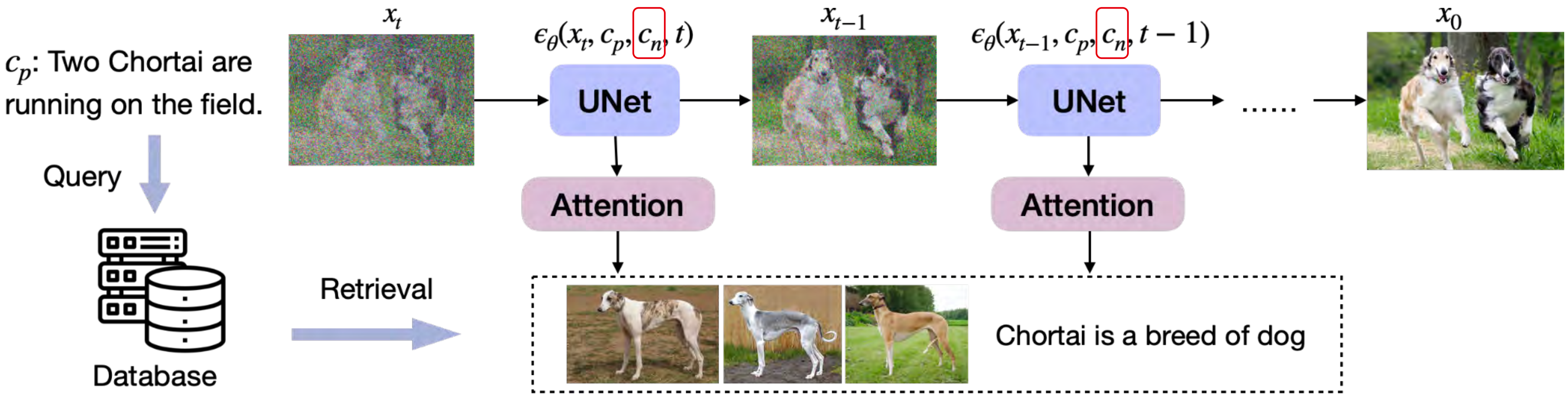
c_p: Two Chortai are running on the field.



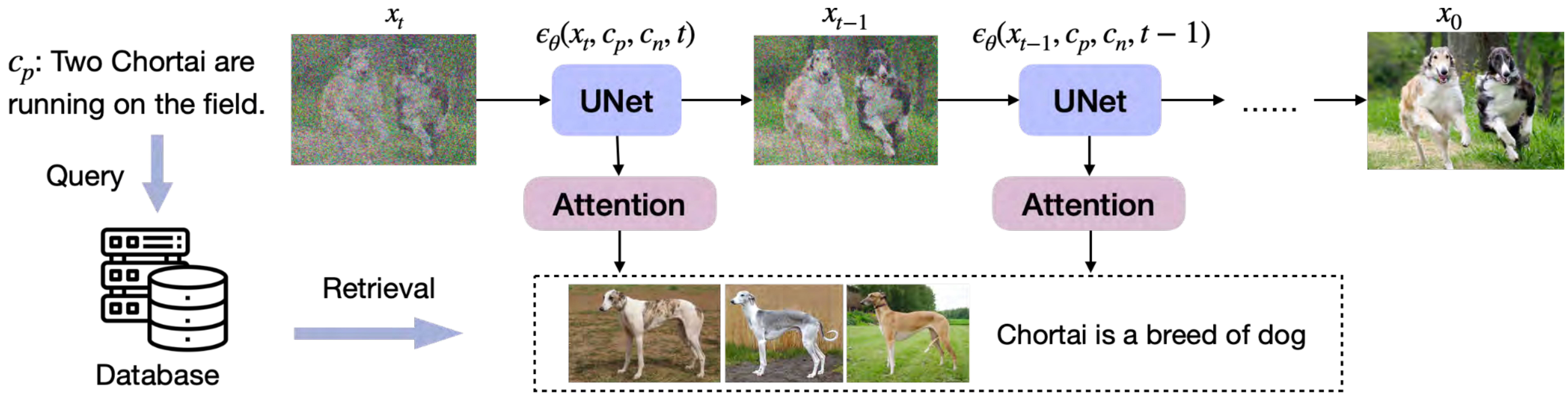
Our approach: Retrieval-augmented Model



Our approach: Retrieval-augmented Model



Advantage over Optimization-based Model



Train a retrieval-augmented model to ground on reference image-text pairs

1. No more fine-tuning during inference, only 30 seconds for inference
2. Only need one reference image, no other assumption.
3. No need for additional image of the same category.

Imagen Architecture

UNet Downstack $f(x_t, c_p)$: a feature map

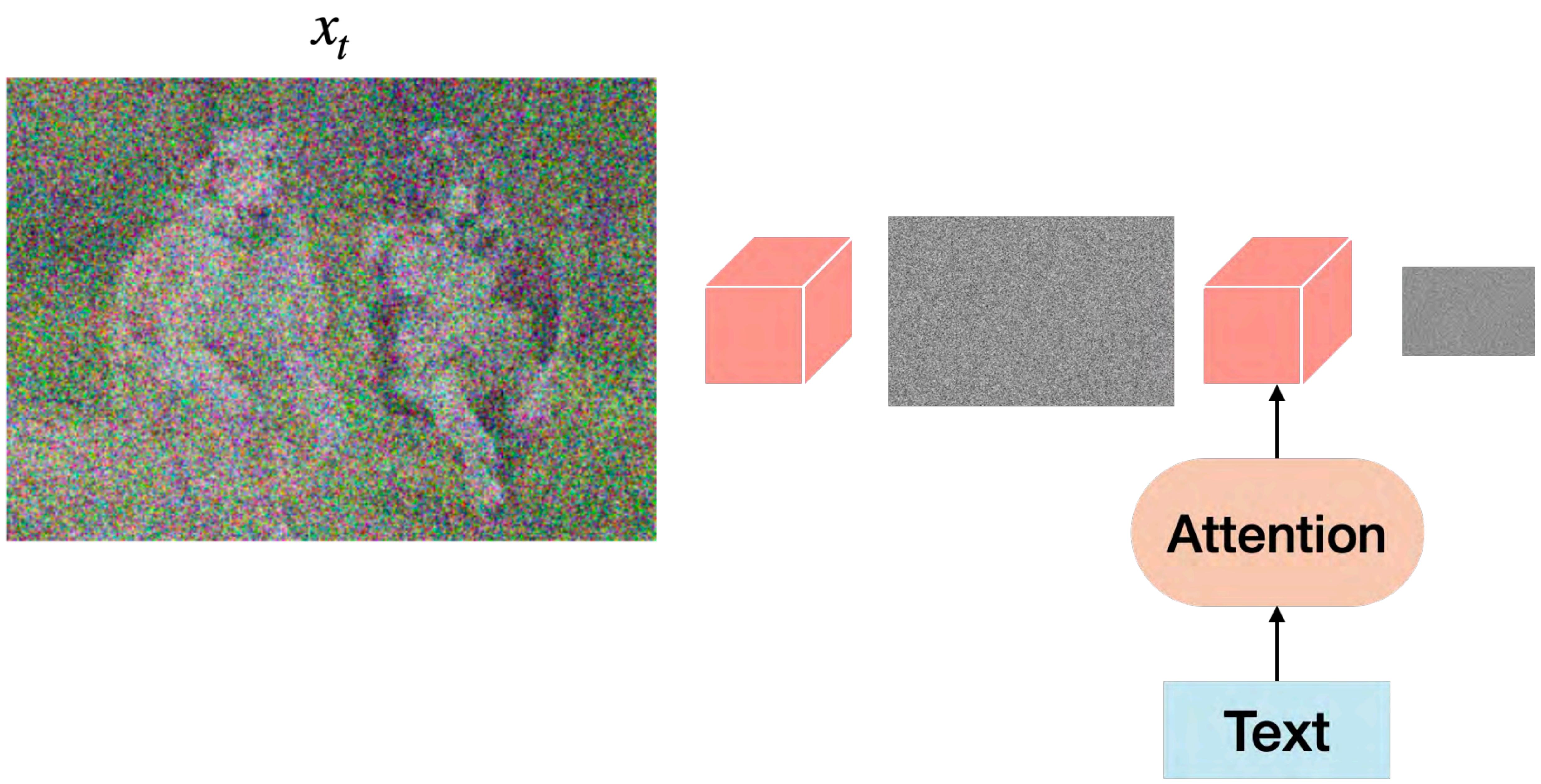
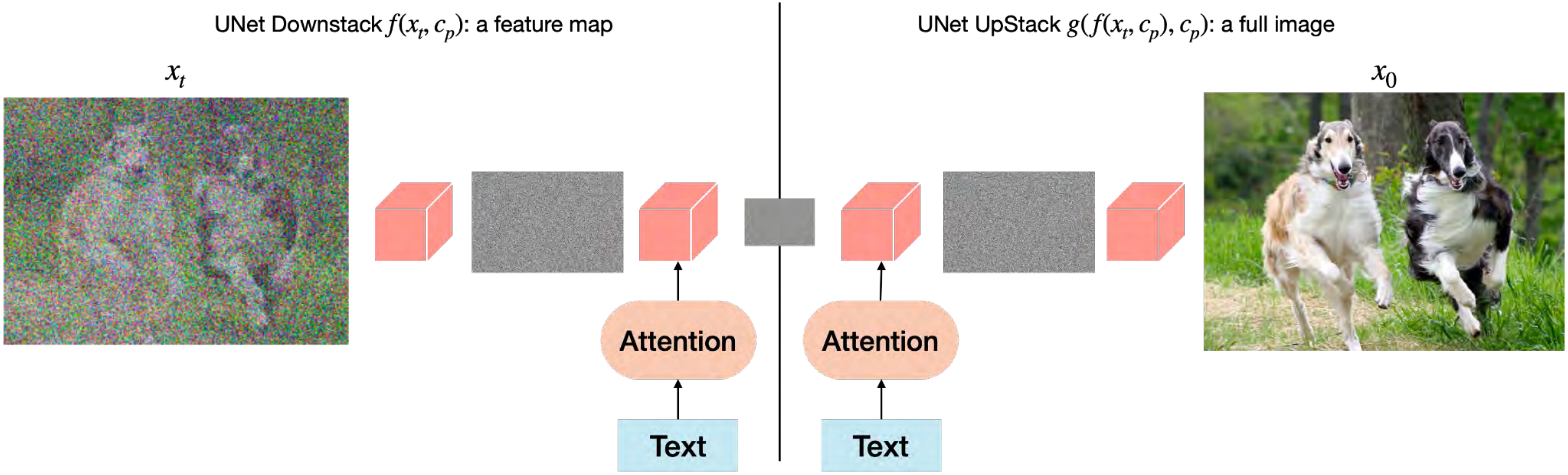
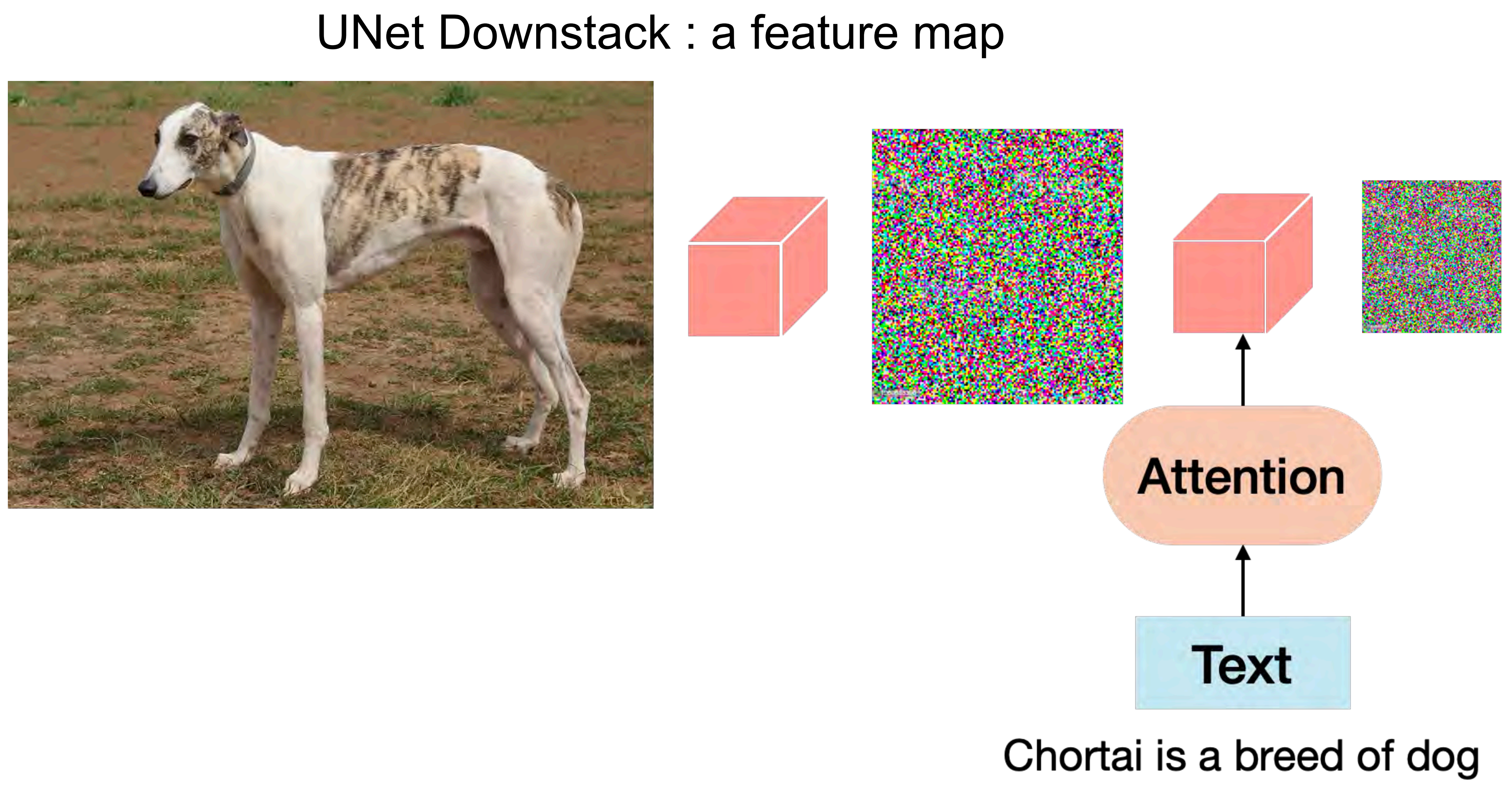


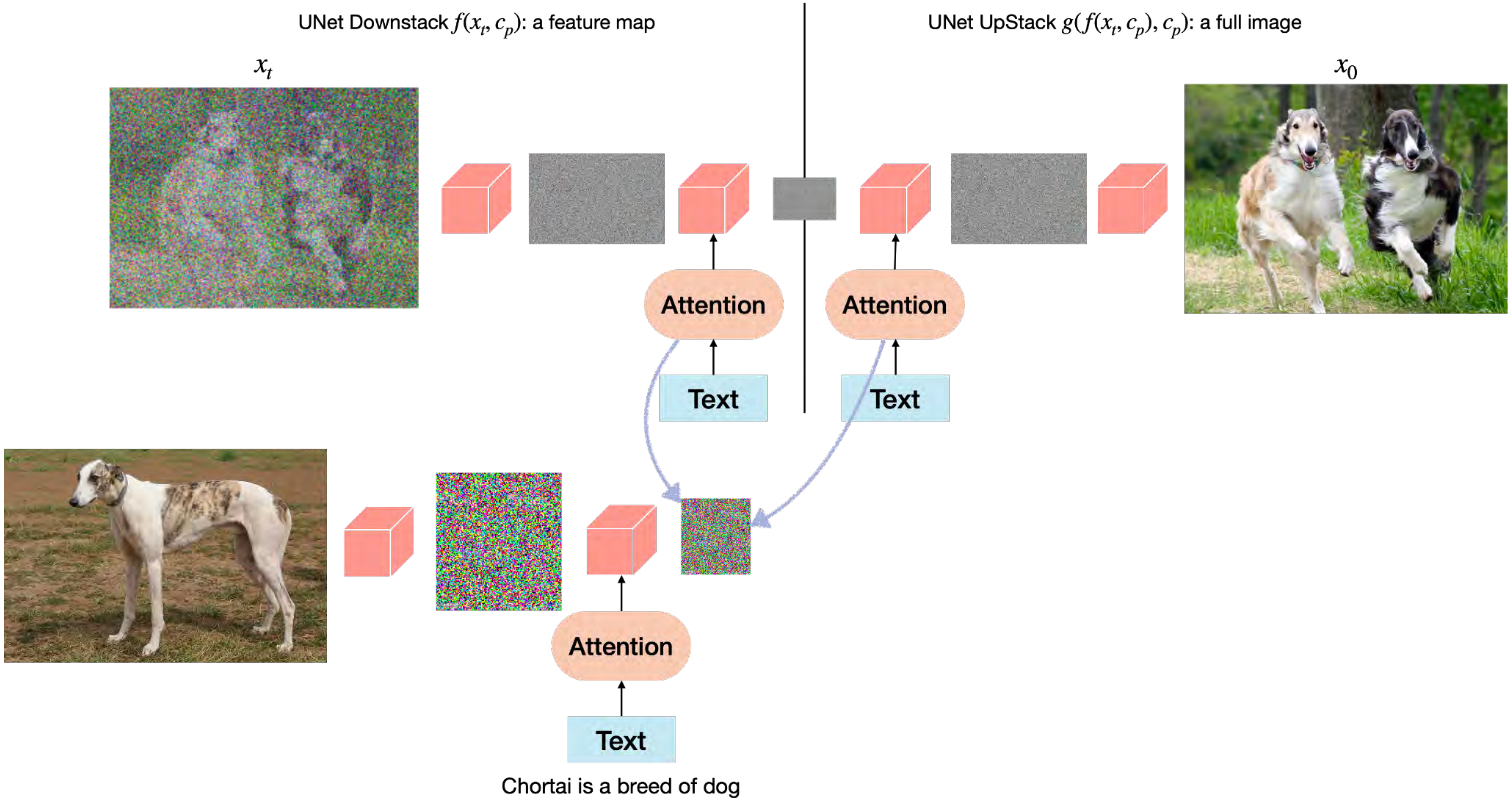
Imagen Architecture



Re-Imagen Architecture



Re-Imagen Architecture



Training Dataset (40M Internal Dataset)

For each (image, text) pair, we search over itself to find similar (image, text) pair with BM25 score.

Top-2 Neighbors

Palm Leaf Placemats |
The Inkabilly Emporium

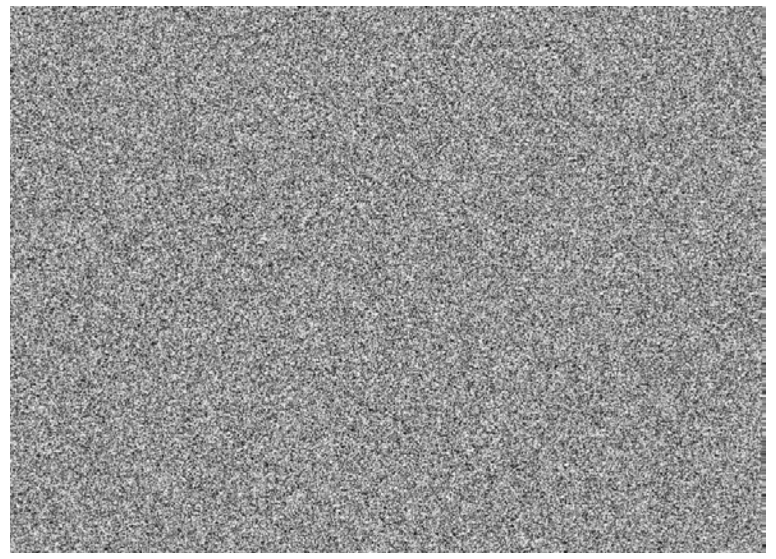
Palm Leaf Placemat Set, with
bamboo | The Inkabilly Emporium

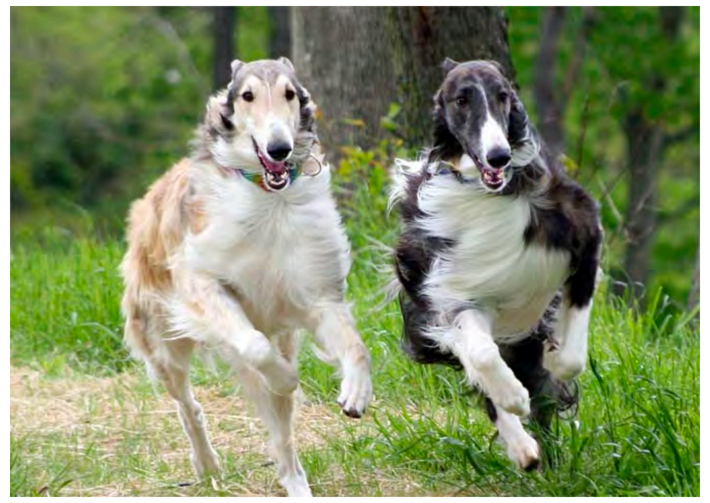
Target

Palm Leaf Print Placemats |
The Inkabilly Emporium

Standard Classifier-free Guidance (Ho et al. 2022)

condition-enhanced: $\epsilon(c_p) = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t, c_n)$





$\epsilon = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t)$

$\epsilon = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t)$

$\epsilon = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t)$

Two Chortai are running on the field.

Two Chortai are running on the field.

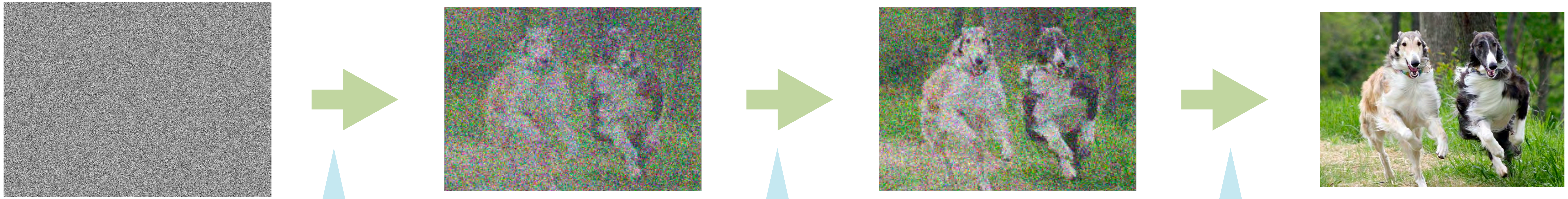
Two Chortai are running on the field.

Entangled Condition Form: the generation is easily dominated by one of the condition

Interleaved Classifier-free Guidance

text-enhanced: $\epsilon(c_p) = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t, c_n)$

neighbor-enhanced: $\epsilon(c_n) = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t, c_p)$



$\epsilon(c_p) = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t, c_n)$ $\epsilon(c_n) = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t, c_p)$ $\epsilon(c_p) = \gamma\epsilon(x_t, c_n, c_p) - (\gamma - 1)\epsilon(x_t, c_n)$

Two Chortai are running on the field.

Two Chortai are running on the field.

Two Chortai are running on the field.

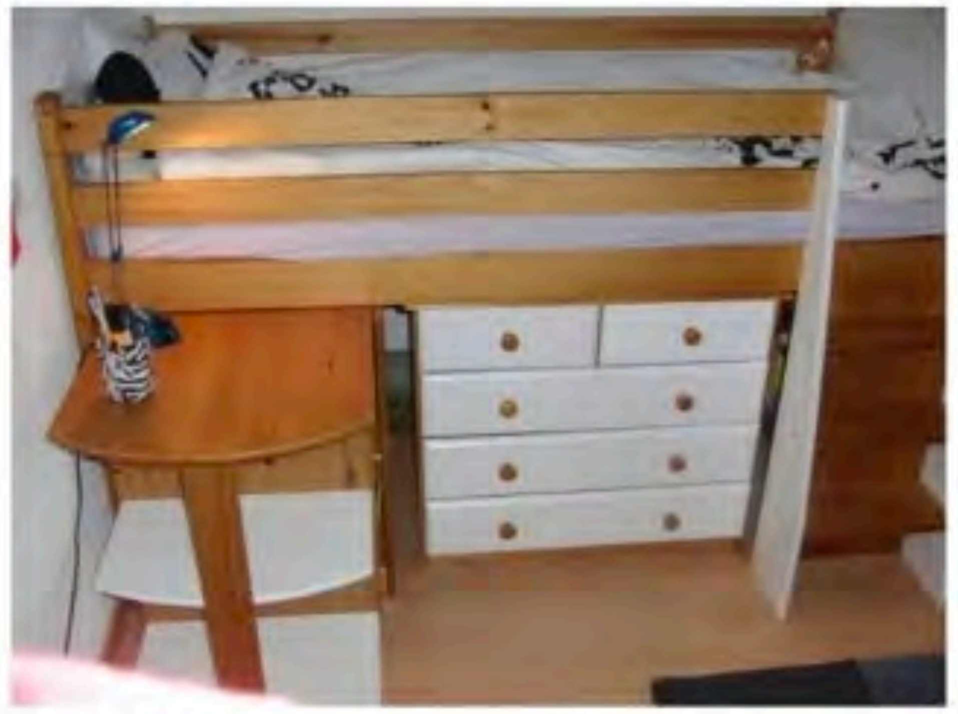
We can adjust the ratio of two guidance by setting η

Evaluation (Quantitative)

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.



Bunk bed with a narrow shelf sitting underneath it.

MSCOCO-30K (Validation Set)

a full length photographic portrait of the photographer Charles Jones

Red tulips in a private garden in Bonfeld, Bad Rappenau, Germany.

WikiCommons Images 20K (Validation Set)

MSCOCO

FID results on MSCOCO-30K (Validation Set)

Model	# of Params	FID-30K	Zero-shot FID-30K
GLIDE (Nichol et al., 2021)	5B	-	12.24
DALL-E 2 (Ramesh et al., 2022)	$\sim$5B	-	10.39
VQ-Diffusion (Gu et al., 2022)	0.4B	-	19.75
KNN-Diffusion (Ashual et al., 2022)	0.8B	-	16.66
Stable-Diffusion (Rombach et al., 2022)	1B	-	12.63
Imagen (Saharia et al., 2022)	3B	-	7.27
Make-A-Scene (Gafni et al., 2022)	4B	7.55	11.84
Parti (Yu et al., 2022)	20B	**3.22**	7.23
Re-Imagen (γ=BM25; $\mathcal{B}$=COCO; k=2)	3.6B	**5.25**†	-
Re-Imagen (γ=CLIP; $\mathcal{B}$=COCO; k=2)	3.6B	5.29†	-
Re-Imagen (γ=BM25; $\mathcal{B}$=ImageText; k=2)	3.6B	-	7.02
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=2)	3.6B	-	6.88

Database: COCO-Train, Internal-40M, LAION-400M

MSCOCO

FID results on MSCOCO-30K (Validation Set)

Model	# of Params	FID-30K	Zero-shot FID-30K
GLIDE (Nichol et al., 2021)	5B	-	12.24
DALL-E 2 (Ramesh et al., 2022)	$\sim$5B	-	10.39
VQ-Diffusion (Gu et al., 2022)	0.4B	-	19.75
KNN-Diffusion (Ashual et al., 2022)	0.8B	-	16.66
Stable-Diffusion (Rombach et al., 2022)	1B	-	12.63
Imagen (Saharia et al., 2022)			7.27
Make-A-Scene (Gafni et al., 2022)			11.84
Parti (Yu et al., 2022)			7.23
Re-Imagen (γ=BM25; $\mathcal{B}$=COCO; k=2)	3.6B	**5.25**†	-
Re-Imagen (γ=CLIP; $\mathcal{B}$=COCO; k=2)	3.6B	5.29†	-
Re-Imagen (γ=BM25; $\mathcal{B}$=ImageText; k=2)	3.6B	-	7.02
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=2)	3.6B	-	6.88

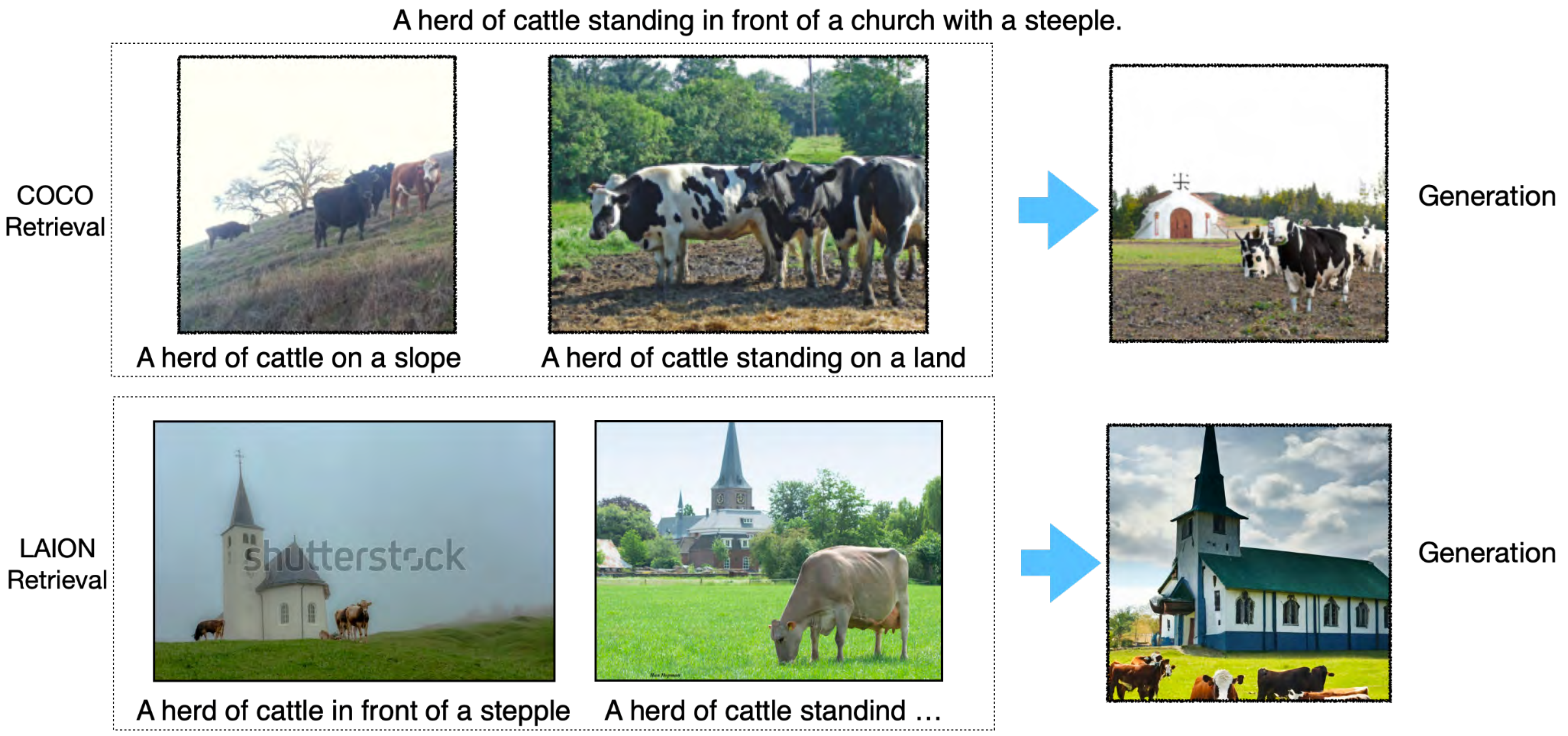
2% improvement using train-set retrieval

only 0.4% improvement with out-of-domain retrieval

Database: COCO-Train, Internal-40M, LAION-400M

MSCOCO Analysis

- MSCOCO dataset does not contain entities, thus the “entity appearance” grounding does not help much.
- Retrieving from in-domain training set can help the model know the “style” of COCO images, thus improving FID significantly.



WikiImages

FID results on WikiCommons-20K (Validation Set)

Model	# of Params	FID-30K	Zero-shot FID-20K
Stable-Diffusion (Rombach et al., 2022)	1B	-	7.50
Imagen (Saharia et al., 2022)	3B	-	6.44
Re-Imagen (γ=BM25; $\mathcal{B}$=WikiImages; k=2)	3.6B	5.88	-
Re-Imagen (γ=CLIP; $\mathcal{B}$=WikiImages; k=2)	3.6B	5.85	-
Re-Imagen (γ=BM25; $\mathcal{B}$=ImageText; k=2)	3.6B	-	6.04
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=1)	3.6B	-	5.94
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=2)	3.6B	-	5.82
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=3)	3.6B	-	**5.80**

WikiImages

FID results on WikiCommons-20K (Validation Set)

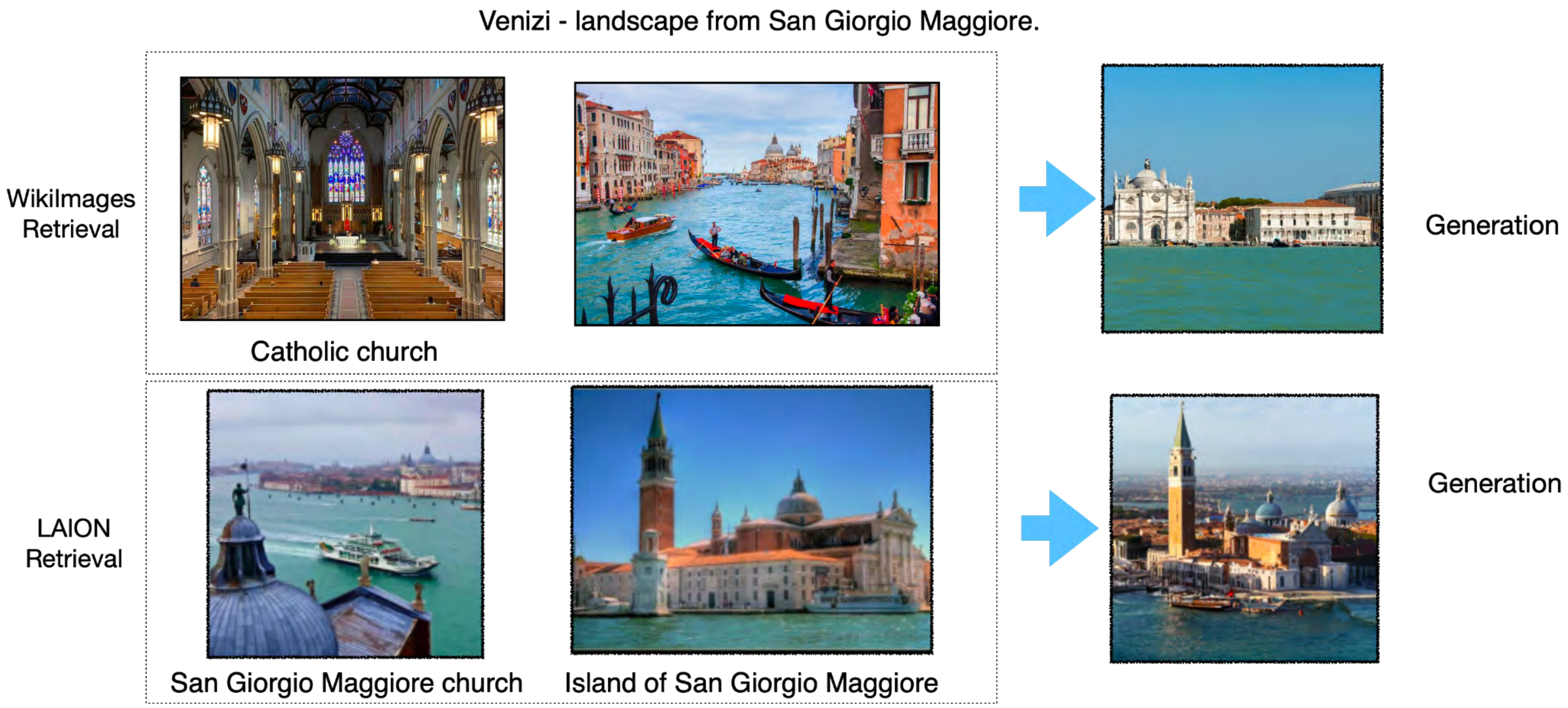
Model	[illegible]	FID-20K	Zero-shot FID-20K
Stable-Diffusion (Rombach et al., 2022)			7.50
Imagen (Saharia et al., 2022)			6.44
Re-Imagen (γ=BM25; $\mathcal{B}$=WikiImages; k=2)	3.6B	5.88	-
Re-Imagen (γ=CLIP; $\mathcal{B}$=WikiImages; k=2)	3.6B	5.85	-
Re-Imagen (γ=BM25; $\mathcal{B}$=ImageText; k=2)	3.6B	-	6.04
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=1)			5.94
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=2)			5.82
Re-Imagen (γ=BM25; $\mathcal{B}$=LAION; k=3)			**5.80**

0.6% improvement using train-set retrieval

0.6% improvement with out-of-domain retrieval

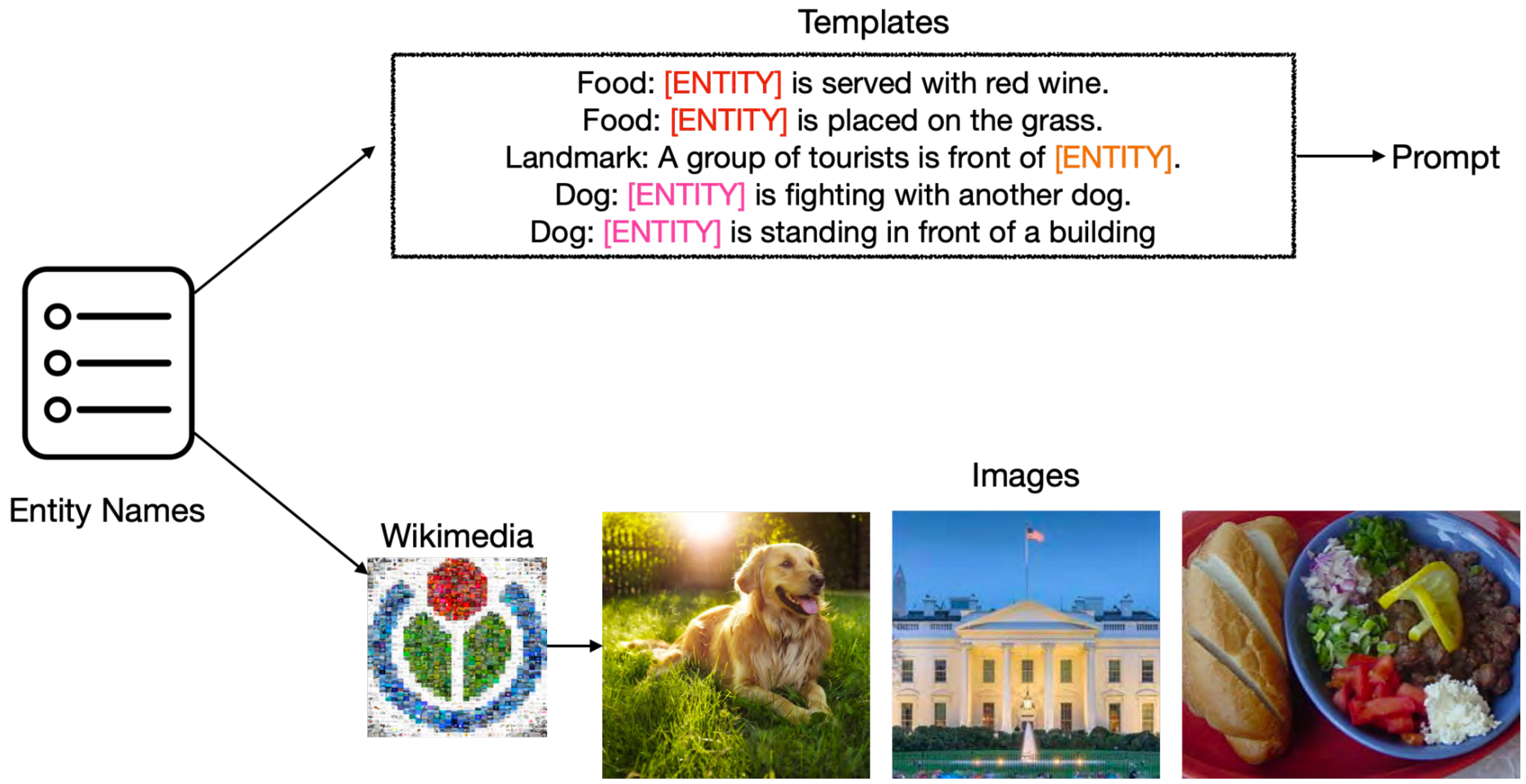
WikiImages Analysis

- WikiImages contains mostly entity-focused images, having “entity appearance” becomes more helpful.
- LAION-400M has much higher coverage for entities, thus providing the same amount of gains as in-domain database.



Evaluation (Qualitative)

Metric: Human evaluation -> Faithfulness and Photorealism



150 <Prompt, (Image, Text)> pairs

Evaluation (Qualitative)

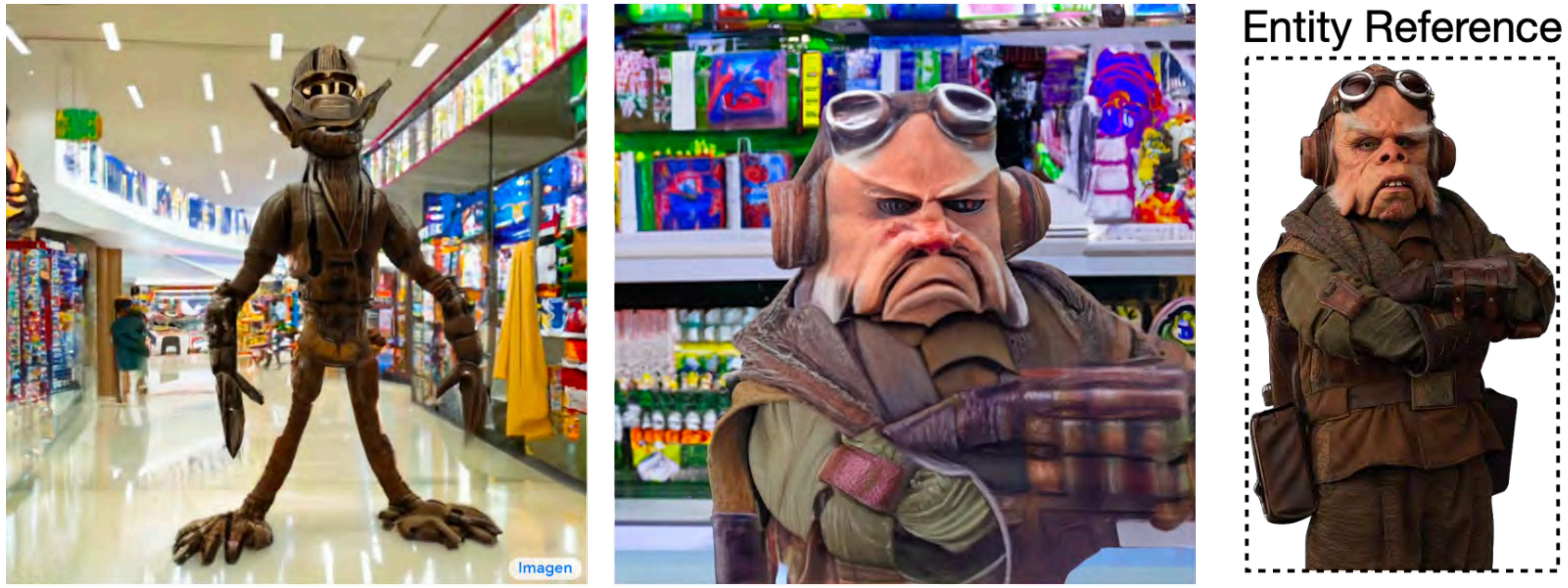
Model	Faithfulness				Photorealism
	Dogs	Foods	Landmarks	All	All
Imagen	0.28 ± 0.02	0.26 ± 0.02	0.27 ± 0.02	0.27	**0.98**
DALL-E 2	0.60 ± 0.02	0.47 ± 0.02	0.36 ± 0.04	0.48	**0.98**
Stable-Diffusion	0.16 ± 0.02	0.24 ± 0.04	0.12 ± 0.06	0.17	0.92
Re-Imagen	**0.68** ± 0.04	**0.70** ± 0.02	**0.74** ± 0.04	**0.71**	0.97

Examples (StarWars)

Imagen | Re-Imagen | Reference

StarWars character Weequay is drinking beer.

Entity Reference



The StarWars character Ugnaught is in a shopping mall.

Examples (Dogs)

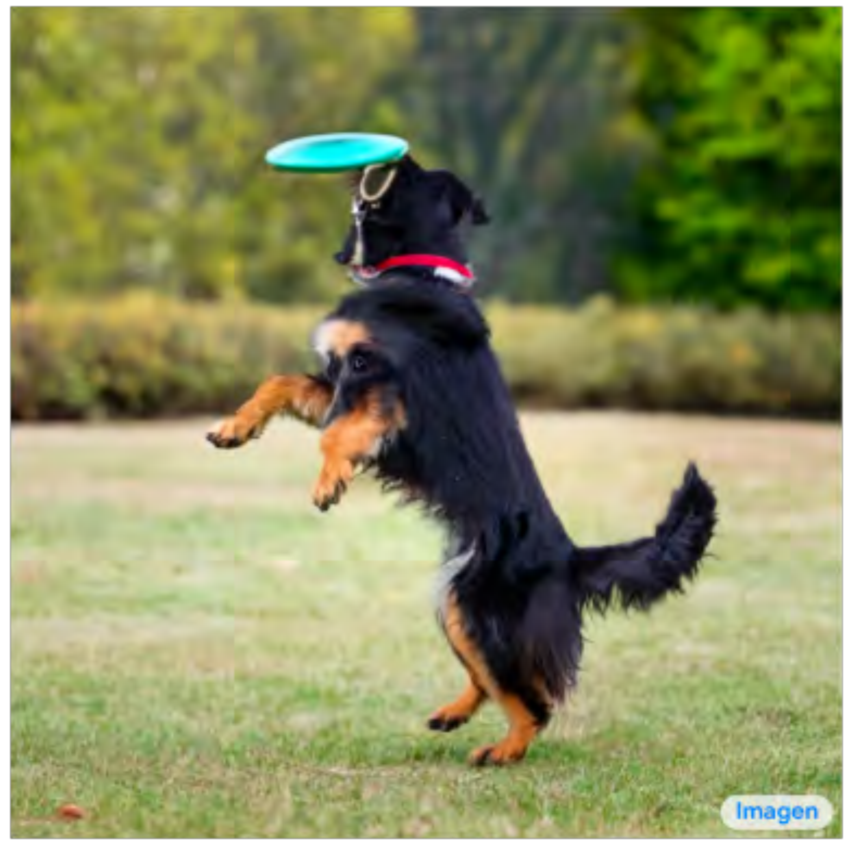
Re-Imagen | Imagen | DALLE-2 | Stable-Diffusion | Entity Refenrence





Tri-colour Armant

A Tri-colour Armant is taking a shower.


Bergamasco shepherd

A Bergamasco shepherd dog is catching a frisbee.

Examples (Food)

Re-Imagen | Imagen | DALLE-2 | Stable-Diffusion | Entity Refenrence




Chilaquiles

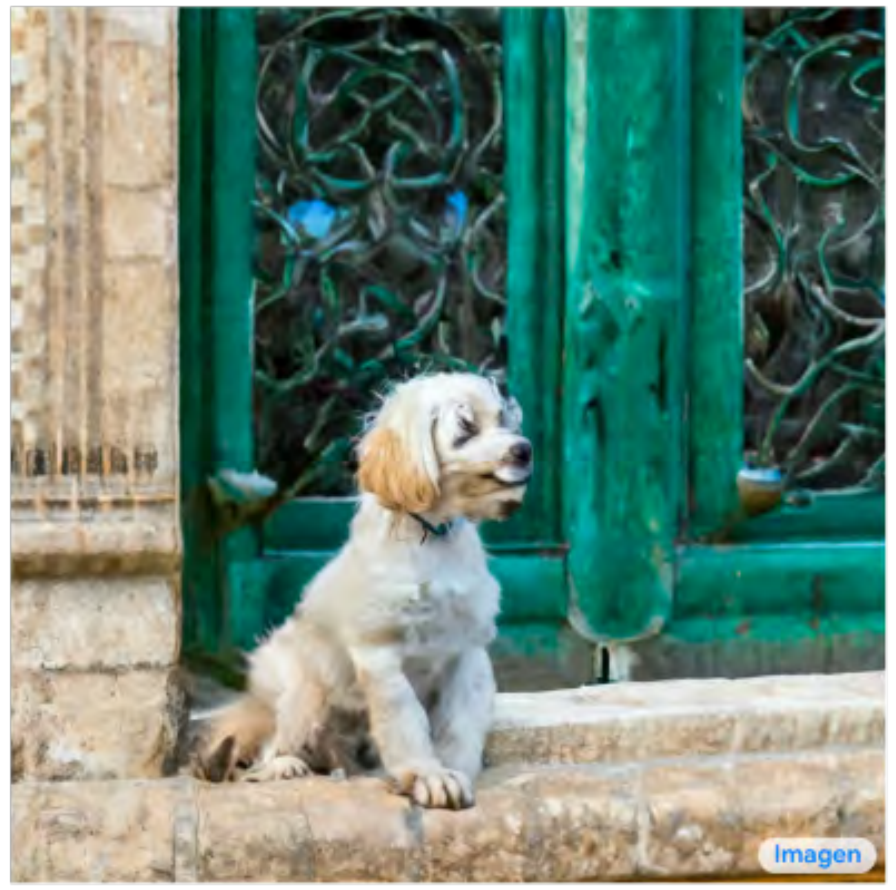
Chilaquiles with popcorns on the side.




Tomato bredie

Tomato bredie is served with wine

Examples (Landmarks)

Re-Imagen	Imagen	DALLE-2	Stable-Diffusion	Entity Refenrence
				

A dog is sitting in front of Palau Güell.

Palau Güell.

A flock of birds fly around Visoki Dečani church.

Visoki Dečani

Ablation Studies of Re-Imagen

Impact of interleaved ratio η (text: all)

Neighbor overwhelming

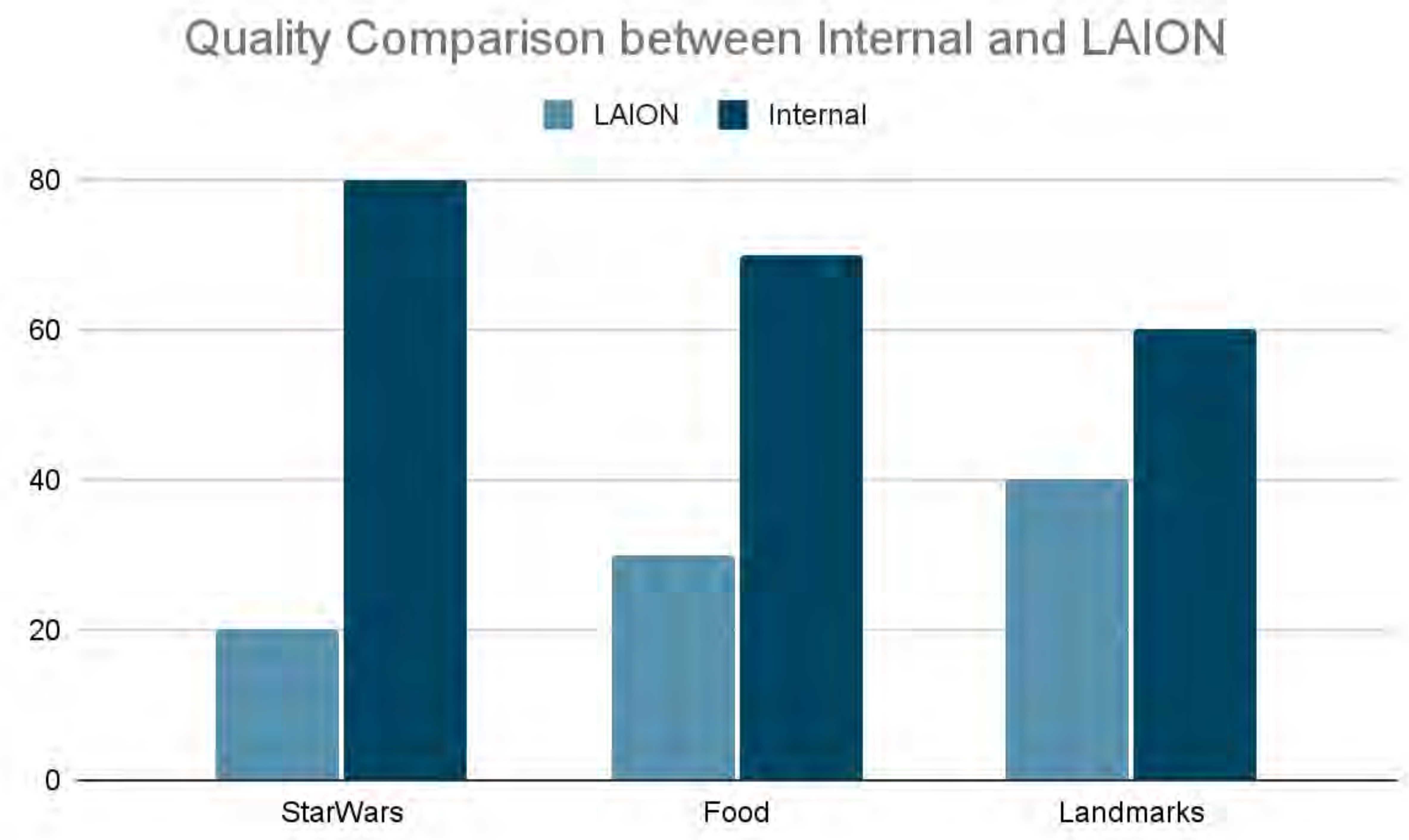
Neighbor overwhelming

A Cretan Hound is running on the moon.

Reference

$\eta = 0.1$ $\eta = 0.4$ $\eta = 0.50$ $\eta = 0.60$ $\eta = 1.0$

Impact of the training dataset



Limitations of Re-Imagen

What are the failure cases [Text Grounding]

Retrievals

Bergen op Zoom.

Escudella

Austrian Pinscher

Generations

A dog is sitting in front of Bergen op Zoom.

Escudella is placed on the grass.

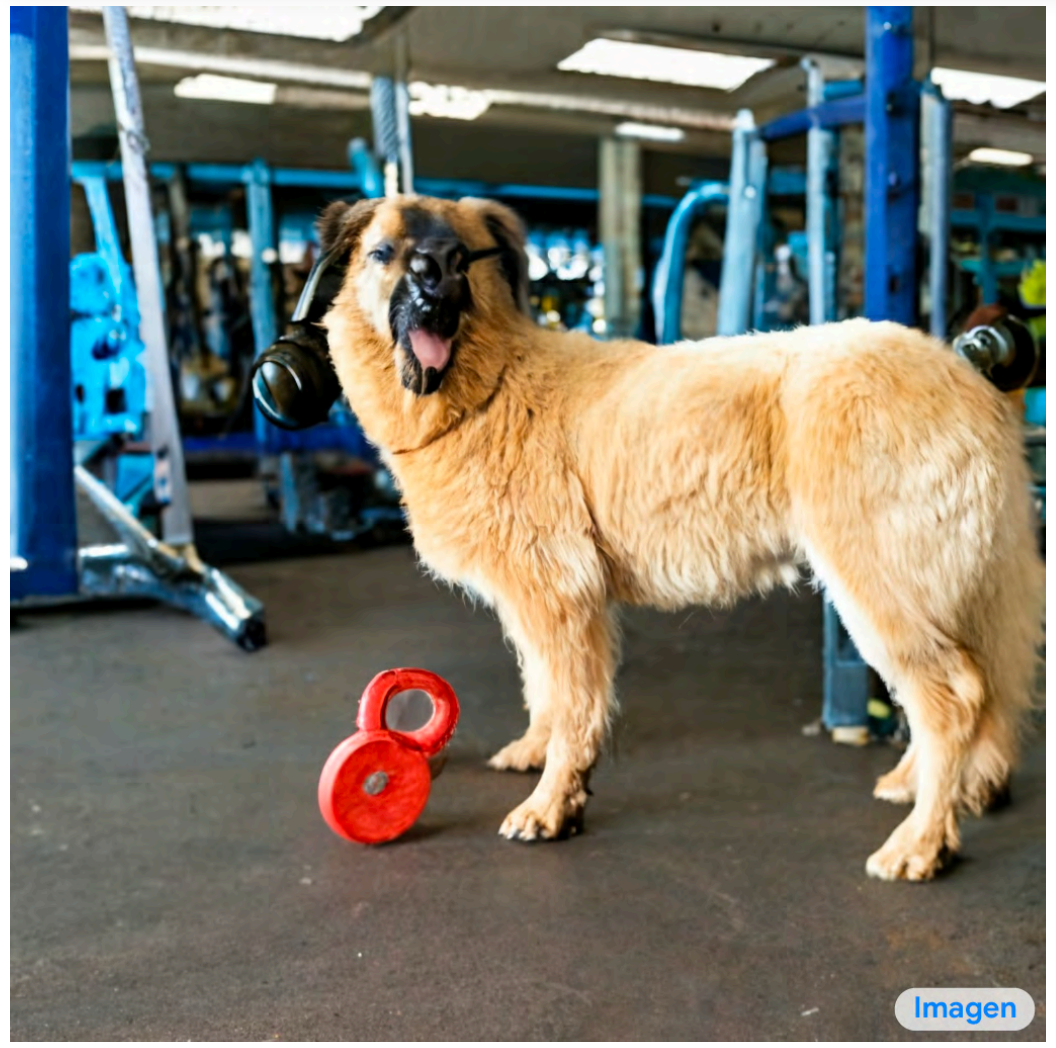
An Austrian Pinscher is chasing a rabbit.

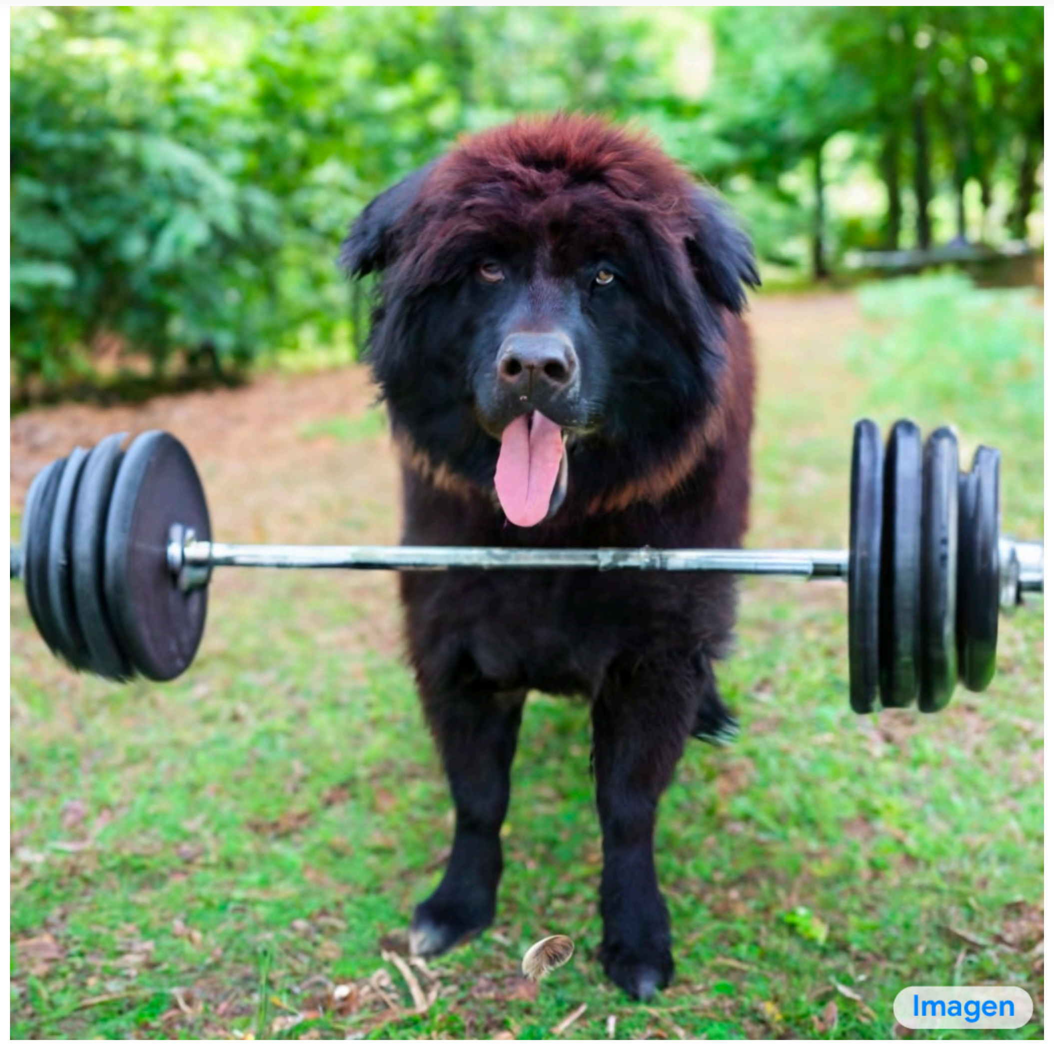
What are the failure cases [Complex Prompts]

Bergamasco shepherd

Re-Imagen

a Bergamasco shepherd is lifting heavy weights.





a Bergamasco shepherd is lifting heavy weights.

The current training dataset is weakly supervised

Cardboard Boxes in Warehouse

Cardboard boxes in warehouse

Not similar

Modern warehouse full of cardboard boxes. 3d illustration

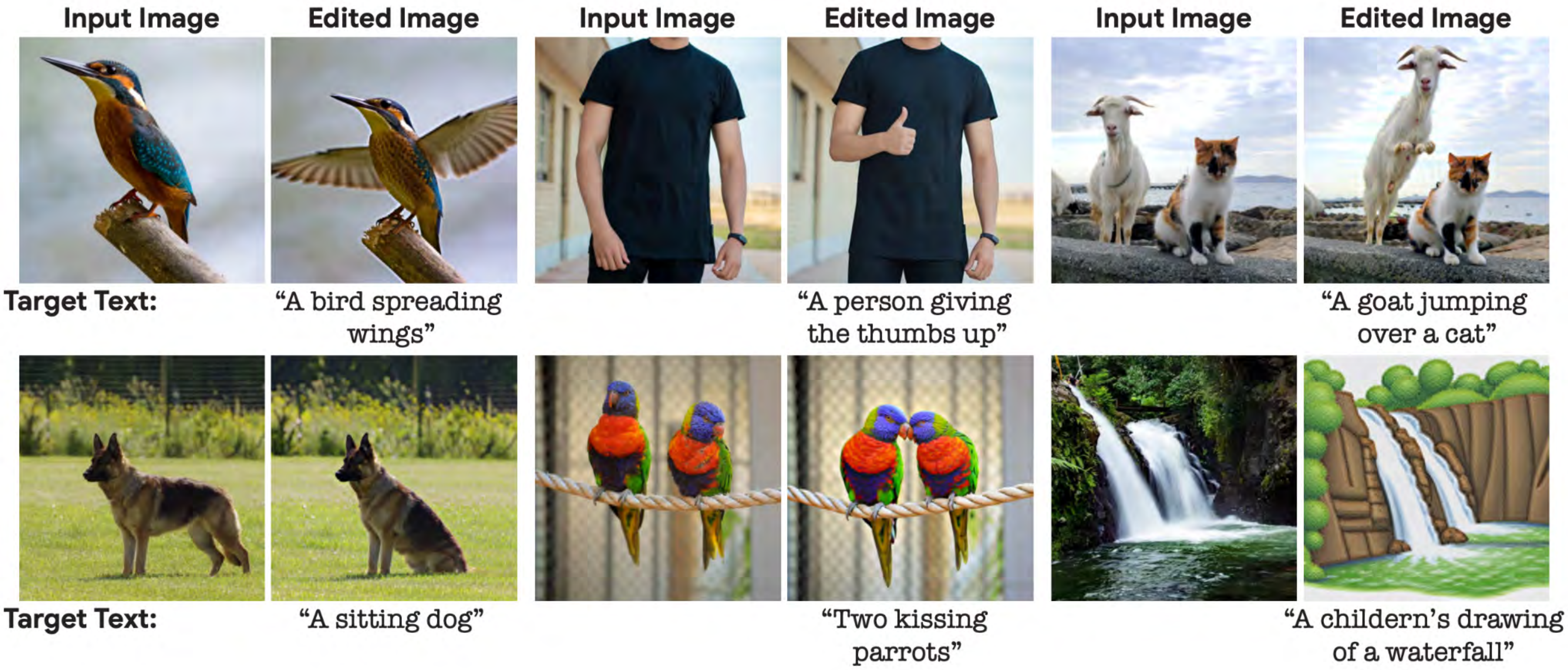
Plaza de los fusilados, Barcelona

Apartados

Almost same

Plaza de los fusilados by Francisco Franco in Barcelona

We need Training Dataset like this!



How to construct better training dataset

Conclusion

Pros:

1. Re-Imagen shows strong capability to ground on retrievals to generate images.
2. Re-Imagen works really well on long-tail entities, which the model cannot capture.
3. Re-Imagen can also be use to perform fast domain adaptation without fine-tuning.

Cons:

1. Re-Imagen still grounds on wrong concepts.
2. Re-Imagen is not good at generating complex prompts about entities.
3. Re-Imagen cannot handle compositional cases well.