Efficient Brains That Imagine

Vicky Kalogeiton

50th Pattern Recognition and Computer Vision Colloquium

Czech Technical University, Prague

09 October 2025

Scale is religion?

Vicky Kalogeiton

50th Pattern Recognition and Computer Vision Colloquium

Czech Technical University, Prague

09 October 2025

My academic story

Lagrange

Poincare

- *Professor*, 2020 –
  - VISTA Group, Ecole Polytechnique, France
  - Assistant Professor 2020-2025
- *Post-doc,* 2018 – 2020
  - Visual Geometry Group, University of Oxford, UK
  - Andrew Zisserman
- *PhD*, 2013 – 2018
  - University of Edinburgh, UK, INRIA, Grenoble, France
  - Vittorio Ferrari, Cordelia Schmid

Monge

Deslandres

Poisson

From perception to imagination

Seeing the world
- Visual recognition, mapping, detection

Imagining futures
- Multiple outcomes, belief modeling, inference

Acting with intent
- Goal driven generation, semantic control, motion

We need **efficient generative brains** models that learn and adapt from very little, imagine, infer, and act → only from a chip in our living room

Image generated with ChatGPT 2025

Challenges

efficient generative brains

Training data
- Collecting, filtering
- Privacy

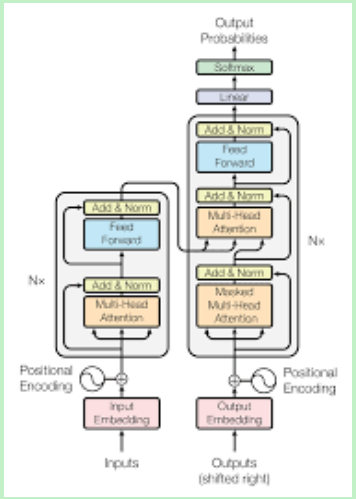
Model
- Large model size
- Scaling resolution

Training & conditioning
- How to condition?
- Long training times

Inference & post-training
- Multiple denoising steps
- Apply RL?

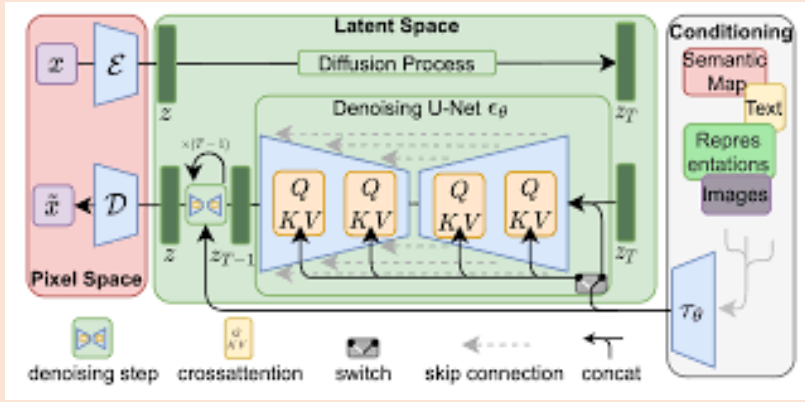
Popular Generative AI architectures

Transformers



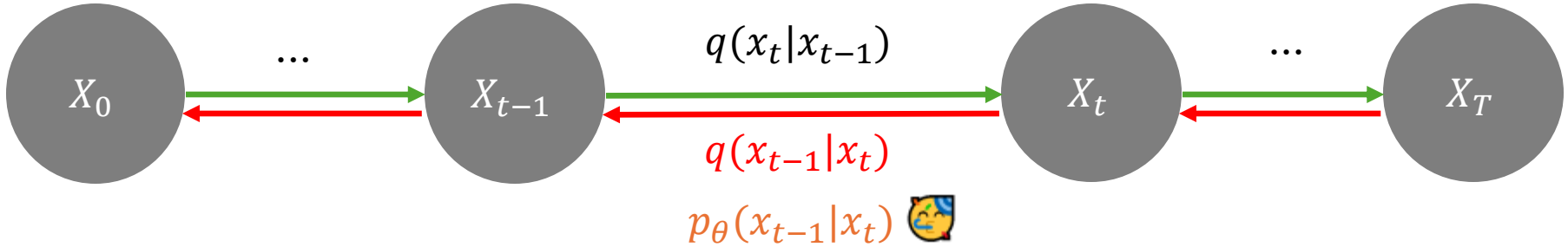
Use: Text, code, tokens, sequential data
Examples: GPT series, LLaMA, Claude

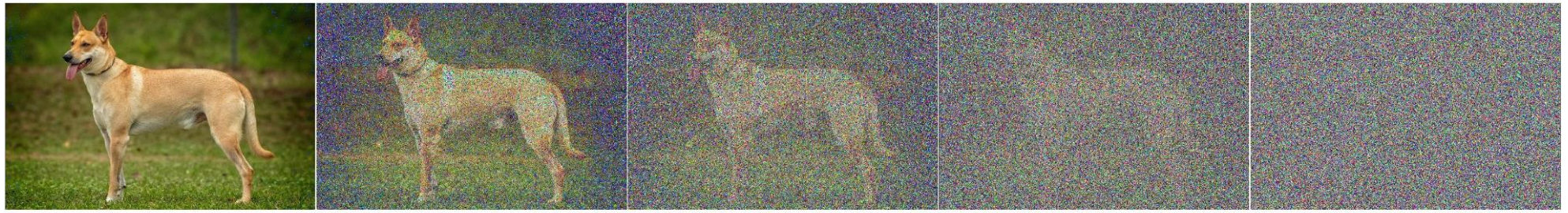
Diffusion Models



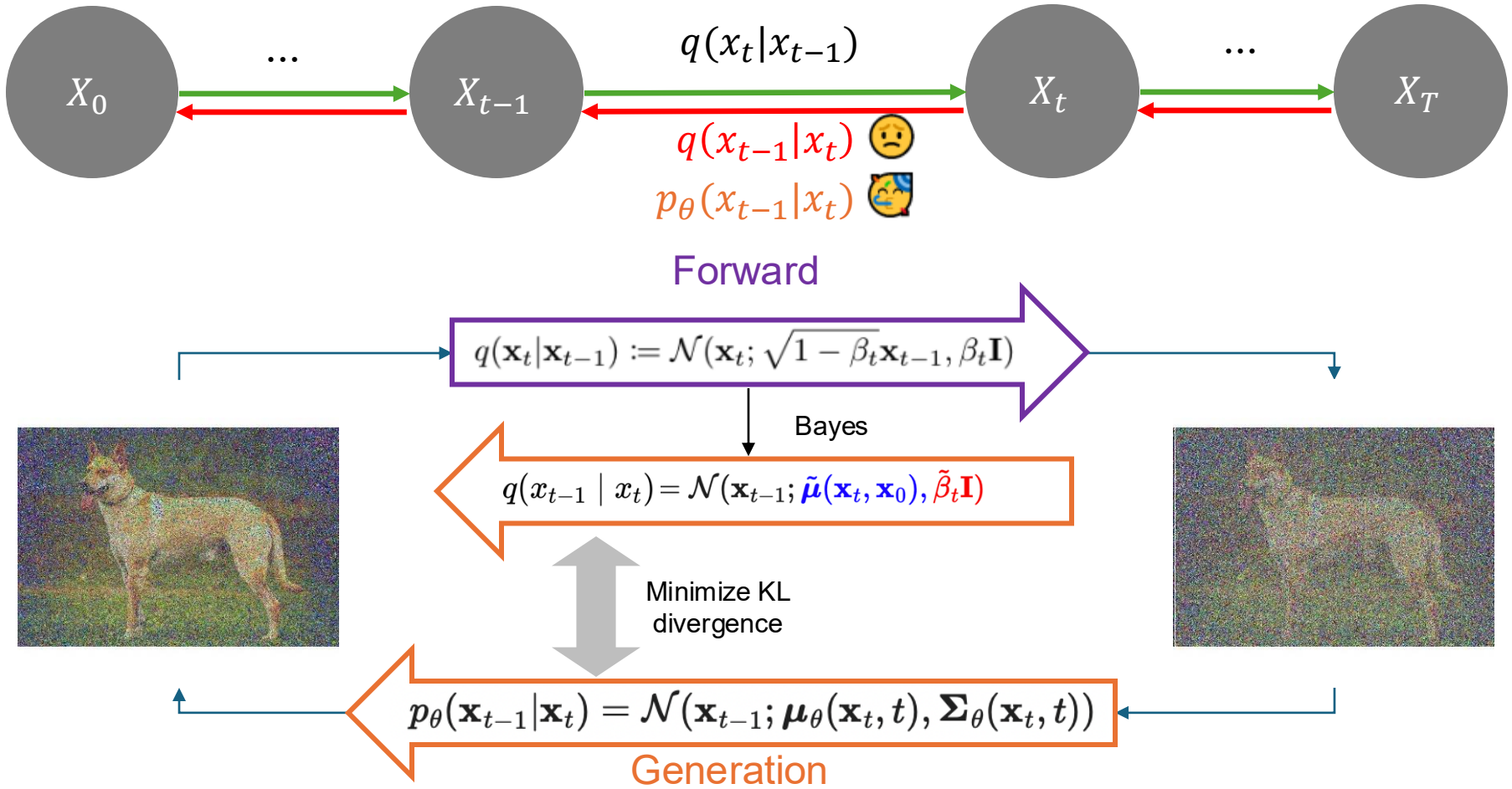
Use: Image, audio, continuous data
Examples: Stable Diffusion, DALL-E

Diffusion Models





Diffusion Models



Dalle-2 (Text-to-Image)

A bowl of soup as a planet in the universe

An astronaut riding a horse in a photorealistic style

Teddy bears mixing sparkling chemicals as mad scientists

Diffusion Models

OpenAI: DALL-E3

Midjourney

music, audio, animation, video, physical etc....

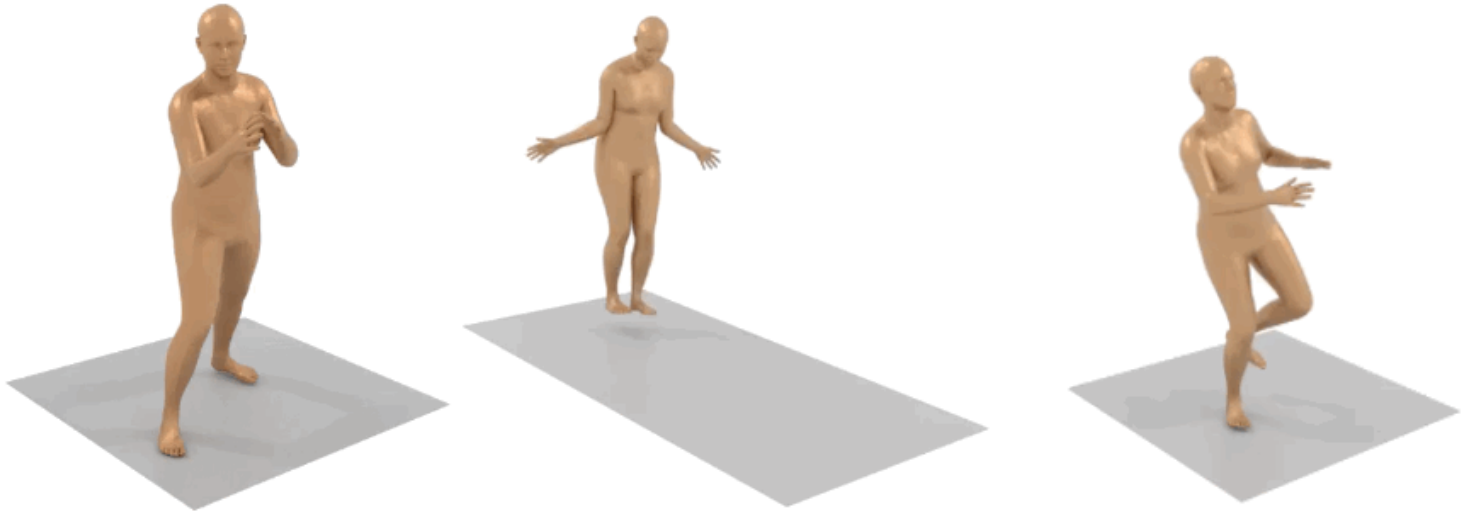
Stable Diffusion 3

February 2024

SORA (Text-to-Video)

February 2024

Human Motion Diffusion (Text-to-Motion)



"A person punches in a manner consistent with martial arts."

"A person is skipping rope."

"a man kicks with something or someone with his left leg."

Efficiency: Challenges

Training data
- Collecting, filtering
- Privacy

Model
- Large model size
- Scaling resolution

Training & conditioning
- How to condition?
- Long training times

Inference & post-training
- Multiple denoising steps
- Apply RL?

Efficiency

Model
- Large model size
- Scaling resolution

Inference & post-training
- Multiple denoising steps
- Apply RL?

Training & conditioning
- How to condition?
- Long training times

Training data
- Collecting, filtering
- Privacy

Don't drop your samples! Coherence-aware training benefits Conditional diffusion

Nicolas Dufour, Victor Besnier, David Picard, Vicky Kalogeiton

CVPR 2024 Highlight

Code: https://github.com/nicolas-dufour/CAD

Website with weights and demo: https://nicolas-dufour.github.io/cad

Motivation: Datasets are noisy by nature

- Diffusion models are easy to condition
- But, aligned datasets are rare and usually contain annotation noise (e.g. webscrapped text/image datasets)
- Noise in annotations makes training harder

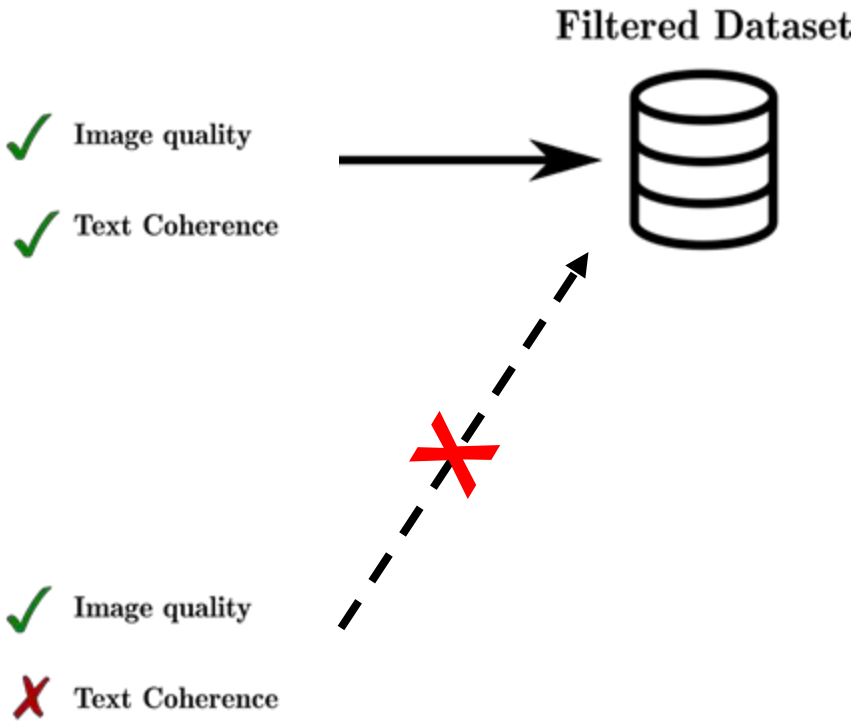
Vancouver could delay plastic straw and foam food container ban until 2020

17 best Backyard images on Pinterest

Related work: Collect billions of data, filter and discard

Two Impala Rams squaring off.

The food on Jeju Island was fishy, in the best possible way. Jeju Island, Sea Urchin, Noodle Soup, Noodles, Spaghetti, Eat, Ethnic Recipes, Food, Macaroni Pasta



Filtering discards useful data!

- Datasets: filtered on a coherence score:
  - measures how coherent the label is with the data
- → Discard too noisy annotations

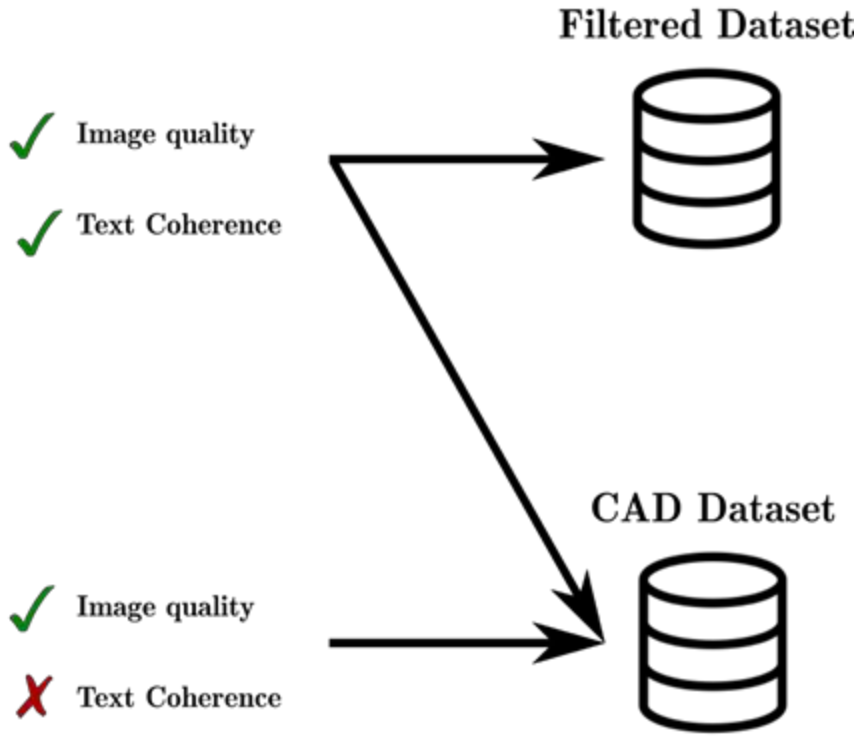
Stable diffusion, Imagen, E-Diffi

Our approach: Coherence-Aware Diffusion (CAD)

Our finding: Providing the coherence score to the model, it can learn what to do with the conditioning in presence of low coherence scores

Two Impala Rams squaring off.

The food on Jeju Island was fishy, in the best possible way. Jeju Island, Sea Urchin, Noodle Soup, Noodles, Spaghetti, Eat, Ethnic Recipes, Food, Macaroni Pasta



Estimate alignment with coherence score

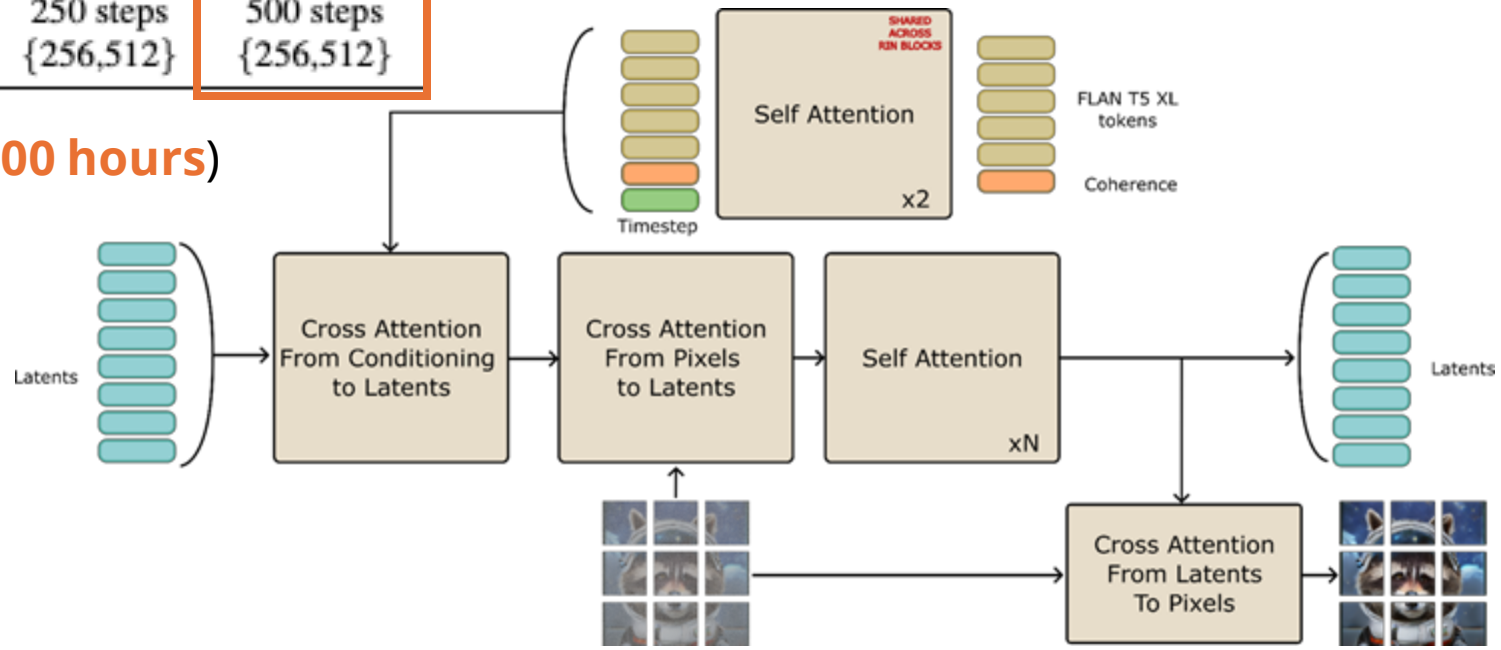
- annotator confidence
- annotator agreement
- expert network (e.g. CLIPScore)

Condition the model by coherence score

CAD: Relatively *small* diffusion model

Features	FLUX	SDXL	SD1.5	CAD (Ours)
Coherence Awareness	-	-	-	✓
Training set size	>1B	>1B	>1B	18M
Training set release & protocol	closed	closed	closed	open
Parameters	12+B	2.6+B	1.45+B	332M
Inference time	4 steps	4 steps	250 steps	500 steps
Resolution	up to 2048	{512,1024}	{256,512}	{256,512}

Trained for 700k steps (~**3k A100 hours**)



Results on text-to-image generation

"a raccoon wearing an astronaut suit. The racoon is looking out of the window at a starry night; unreal engine, detailed, digital painting,cinematic,character"

"An armchair in the shape of an avocado"

Qualitative samples

portrait photo of a asia old warrior chief tribal panther make up blue on red side profile looking away serious eyes 50mm portrait photography hard rim lighting photography

Pirate ship trapped in a cosmic maelstrom nebula rendered in cosmic beach whirlpool engine volumetric lighting spectacular ambient lights light pollution cinematic atmosphere art nouveau style illustration art artwork by SenseiJaye intricate detail.

an oil painting of rain at a traditional Chinese town

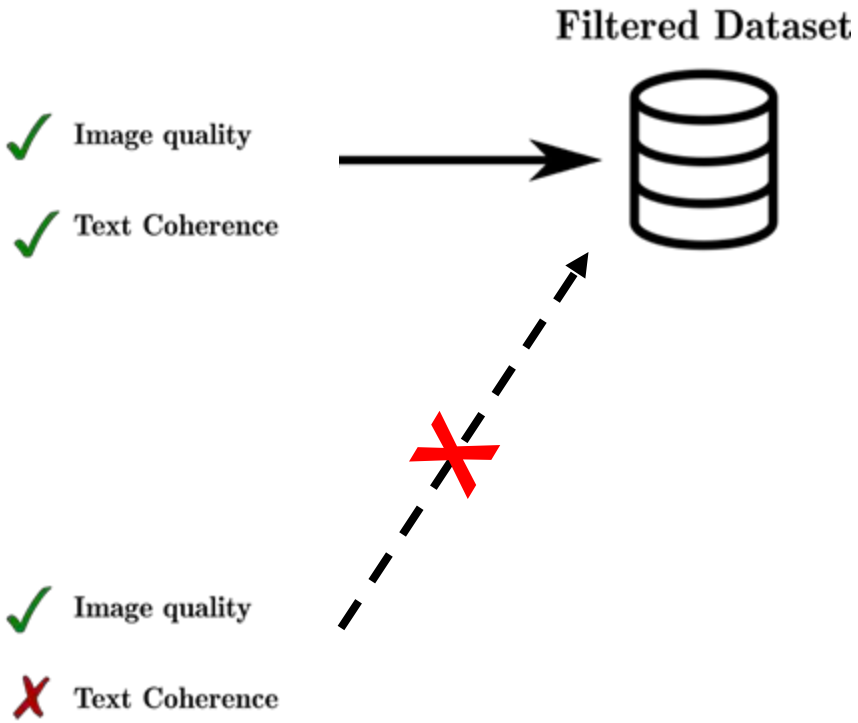
Conclusion

- Don't drop your samples!
- Evaluate the conditioning coherence instead and condition your network with it
- Resulting model: conditional + unconditional, allowing classifier-free guidance without dropping the conditioning
- Bonus: academia can also **train** a text-to-image diffusion model ☺

Related work: After filtering Training→ Billions of data

Two Impala Rams squaring off.

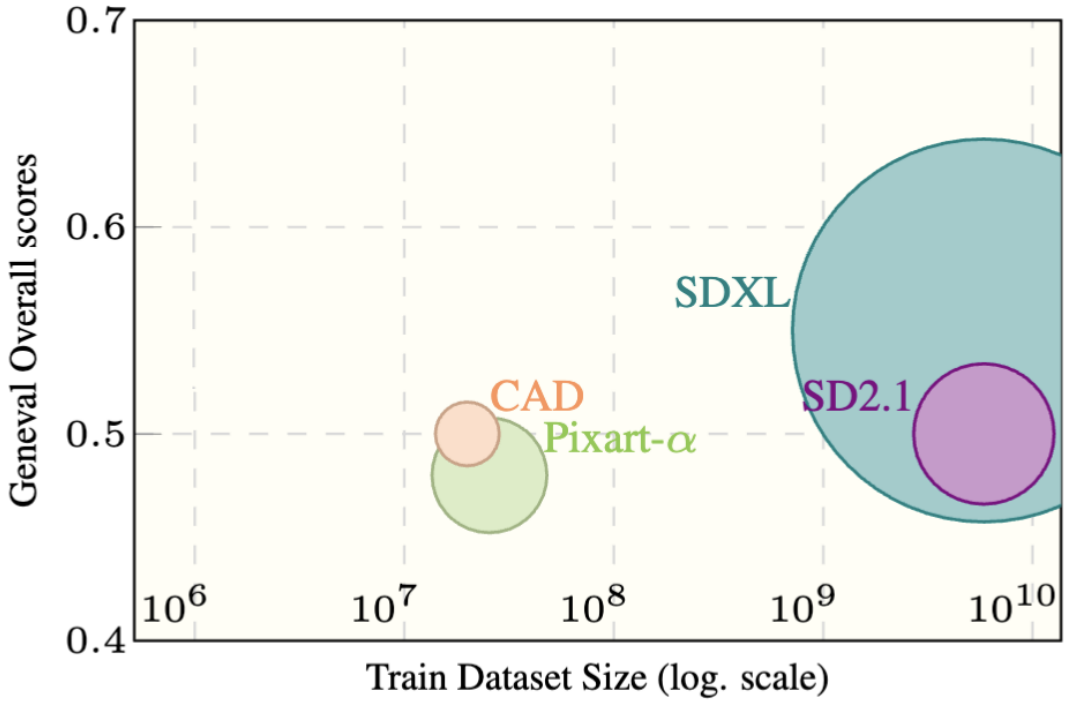
The food on Jeju Island was fishy, in the best possible way. Jeju Island, Sea Urchin, Noodle Soup, Noodles, Spaghetti, Eat, Ethnic Recipes, Food, Macaroni Pasta

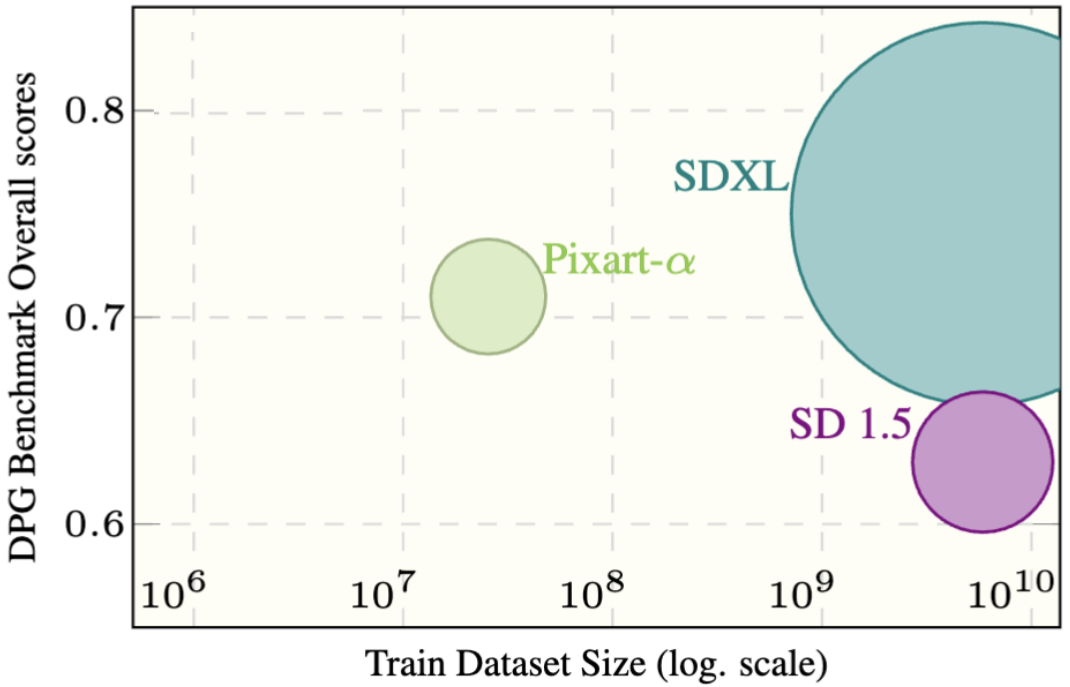


Stable diffusion, Imagen, E-Diffi

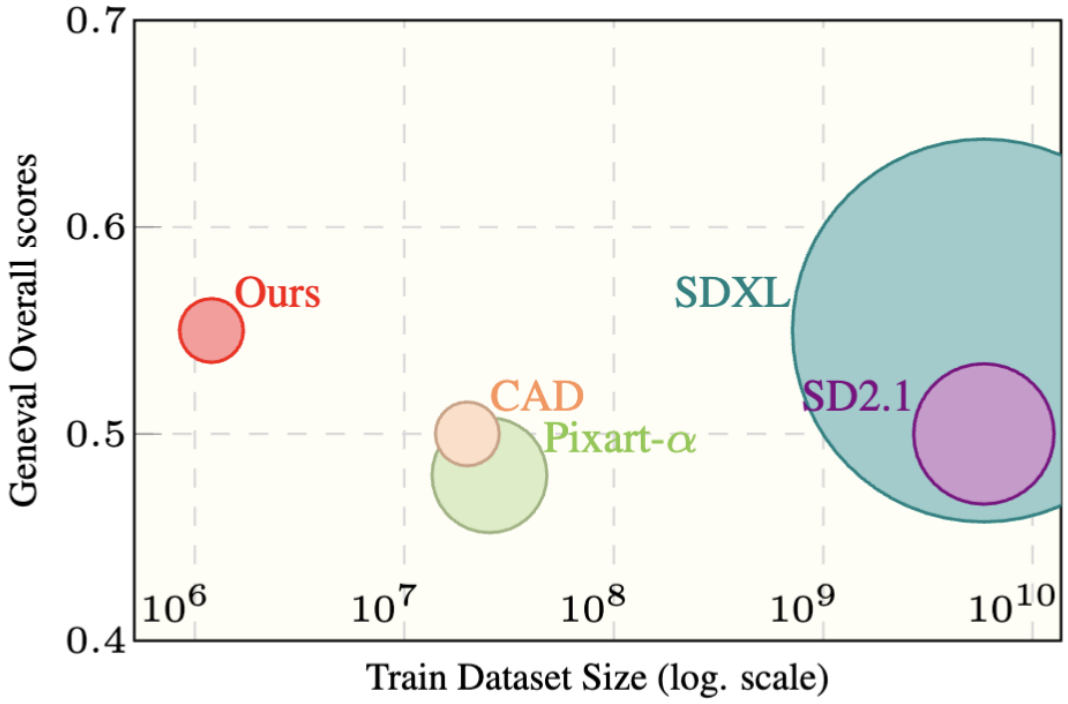
Motivation: Train datasets: billions of data

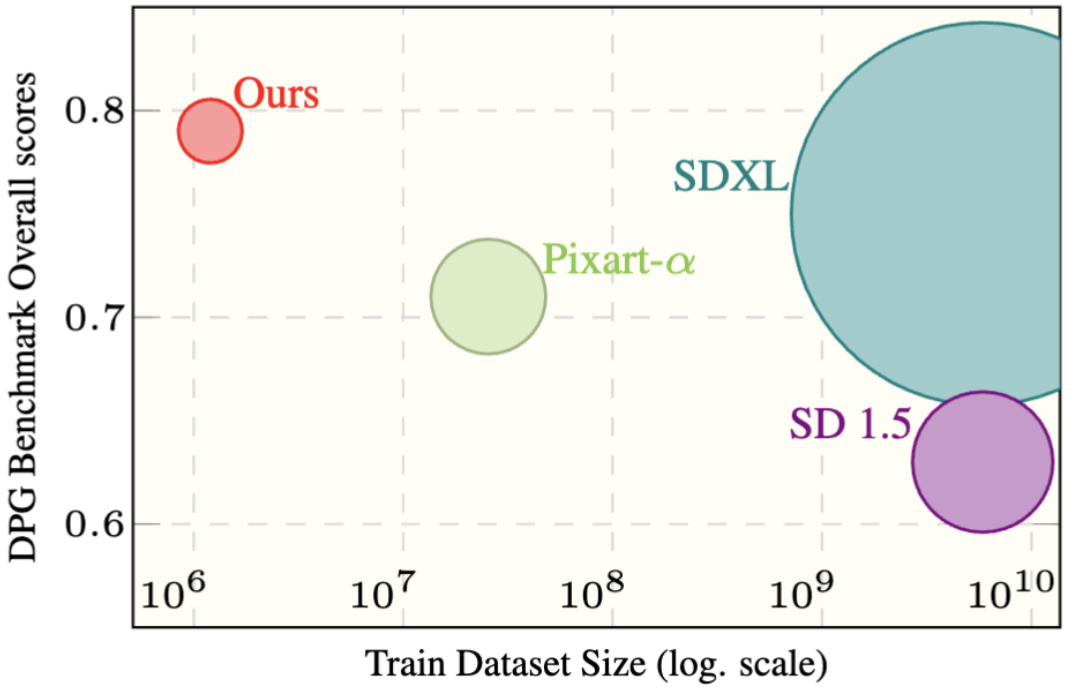
- The larger the dataset size, the higher the performance
- But, do we really exploit training data?





How far can we go with ImageNet for text-to-image generation?





How far can we go with ImageNet for text-to-image generation?




Lucas Degeorge*, Arijit Ghosh*, Nicolas Dufour, David Picard, Vicky Kalogeiton
arXiv 2025

Efficiency: Challenges

Training data
- Collecting, filtering
- Privacy

Model
- Large model size
- Scaling resolution

Training & conditioning
- How to condition?
- Long training times

Inference & post-training
- Multiple denoising steps
- Apply RL?

Efficiency

Training data
- Collecting, filtering
- Privacy

Model
- Large model size
- Scaling resolution

Training & conditioning
- How to condition?
- Long training times

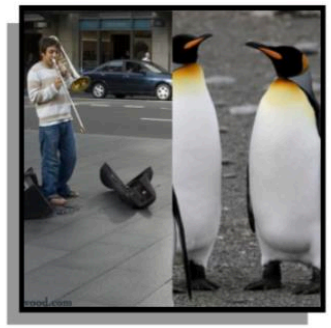
Inference & post-training
- Multiple denoising steps
- Apply RL?

Augmentations

an image of (AOI) + a cabbage Butterfly

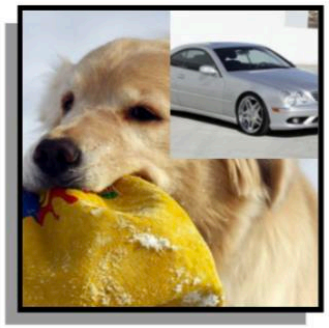
A delicate white butterfly… The flower, a stunning shade of purple .... The butterfly, positioned slightly to the left of the flower's center, …

a trombone and a king penguin



On the left side, a person is playing the trumpet on a street. .... On the right side of the image, there are two penguins standing ....

a golden retriever and a car wheel



... a silver sports car in the background. ... a Golden Retriever, is on the left side of the frame .... The sports car, positioned on the right, ...

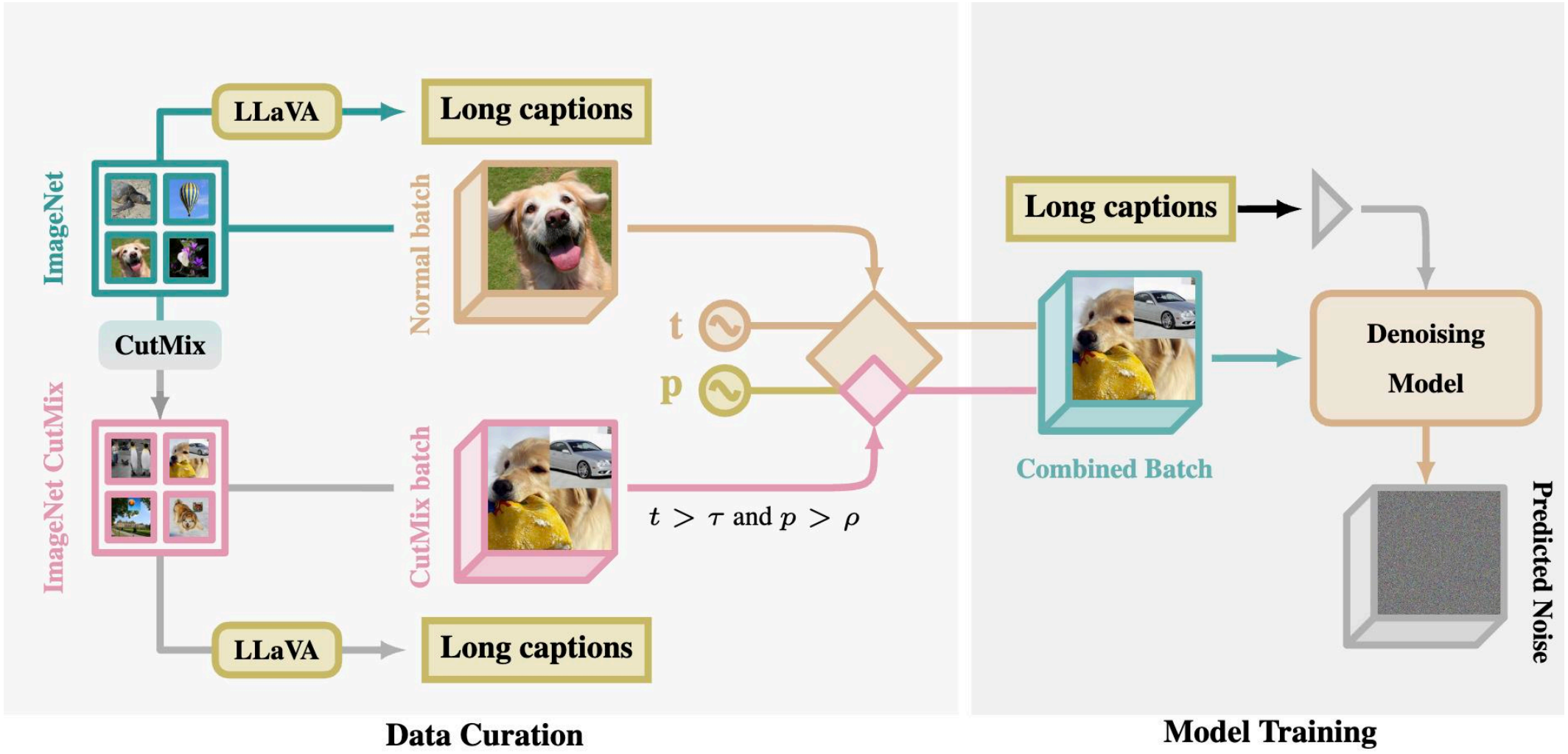
a palace and a balloon

... a palace or manor house, ... In front of the building is a well-maintained garden ... In the sky, there is a single hot air balloon ....

a malamute and a red wine

... a husky dog resting in the snow. ... Next to the dog's side, there is a wine glass with red wine and a few purple flowers...

How far can we go with ImageNet for text-to-image generation?



Quantitative Analysis: Augmentations

Model	IA	Overall↑	One obj.↑	Two obj.↑	Count.↑	Col.↑	Pos.↑	Col. attr.↑
DiT-I	✗	0.55	0.95	0.61	0.36	0.80	0.28	0.33
	Crop	0.54	0.96	0.56	0.38	0.79	0.22	0.33
	CutMix	0.58	0.95	0.67	0.43	0.80	0.30	0.35
CAD-I	✗	0.55	0.97	0.60	0.42	0.74	0.26	0.35
	Crop	0.54	0.96	0.61	0.40	0.71	0.23	0.33
	CutMix	0.57	0.94	0.68	0.40	0.70	0.35	0.36

Table 2: **GenEval scores** of TA and TA + IA models. All models are trained with long captions. A Prompt Extender was used before generating images. Models are evaluated at 256^2 resolution.

[Thanks to Alyosha for the crop-only augmentation idea!]

Qualitative Analysis: Augmentations

« An image of »	« An image of » + TA	« An image of » + TA + IA
	A teddy bear driving a motorbike	
	A teapot and cookies on a table	
	A goat on a mountain top	

TA: Text Augmentation
IA: Image Augmentation

Qualitative results

Ours SDXL Ours SDXL

A corgi wearing a red bowtie and a purple party hat.

A mountain

An old man with a long grey beard and green eyes

A bird and its reflection in a fountain.

What about resolution?

Resolution 512^2: Quantitative results

- Initialization: 250k steps DiT checkpoint
- Fine-tuning: 50k steps on the same data
  → adjusting the image tokenization to handle larger input size

Resolution	Overall↑	One obj.↑	Two obj.↑	Count.↑	Col.↑	Pos.↑	Col. attr.↑
DiT-I 256^2	0.58	0.95	0.67	0.43	0.80	0.30	0.35
DiT-I 512^2	0.61	0.98	0.73	0.43	0.76	0.34	0.40

Table 4: **GenEval scores** of models with different resolution. The 512^2 is finetuned from the 256^2.

→ Is it all about initialization?

Resolution 512^2 : Qualitative results

→ Scale is not required!

Model	#params	#train data	Overall↑	One obj.↑	Two obj.↑	Count.↑	Col.↑	Pos.↑	Col. attr.↑
SD v1.5	0.9B	5B+	0.43	0.97	0.38	0.35	0.76	0.04	0.06
PixArt-α	0.6B	25M	0.48	<u>0.98</u>	0.50	0.44	0.80	0.08	0.07
PixArt-Σ (512)	0.6B	35M+	0.52	<u>0.98</u>	0.59	0.50	0.80	0.10	0.15
SD v2.1	0.9B	5B+	0.50	<u>0.98</u>	0.51	0.44	<u>0.85</u>	0.07	0.17
SDXL	3.5B	5B+	0.55	<u>0.98</u>	<u>0.74</u>	0.39	<u>0.85</u>	0.15	0.23
SD3 M (512)	2B	1B+	0.62	<u>0.98</u>	<u>0.74</u>	<u>0.63</u>	0.67	**0.34**	0.36
SANA-0.6	0.6B	⊘	<u>0.64</u>	**0.99**	0.71	<u>0.63</u>	**0.91**	0.16	<u>0.42</u>
FLUX-dev	12B	⊘	**0.67**	**0.99**	**0.81**	**0.79**	0.74	<u>0.20</u>	**0.47**
Ours (512^2)	0.4B	1.2M	0.61	<u>0.98</u>	0.73	0.43	0.76	**0.34**	0.40

Table 5: **Results on GenEval**. Results are reported from their papers. **Bold** indicates best, <u>underline</u> second best.

What about aesthetics?

- Initialization: 300k steps DiT 512^2 checkpoint
- Upscale to 1024^2
- Fine-tuning: on LAION-POP (400K images) for high aesthetics targets
  → adjusting the image tokenization to handle larger input size

Quantitative results: Finetuning on high-aesthetic dataset

Model	#params	#train data	Aes. Score↑	PickScore↑	HPSv2.1↑	ImageReward↑
SD v1.5	0.9B	5B+	5.68	21.3	0.25	0.24
SD v2.1	0.9B	5B+	5.81	21.5	0.26	0.38
PixArt-α	0.6B	25M	<u>6.47</u>	22.6	<u>0.29</u>	0.97
PixArt-Σ	0.6B	35M+	6.44	22.5	<u>0.29</u>	1.02
CAD	0.35B	-	5.56	21.4	0.26	0.69
Sana-0.6B	0.6B	-	6.31	<u>22.8</u>	**0.30**	**1.23**
Sana-1.6B	1.6B	-	6.36	<u>22.8</u>	**0.30**	**1.23**
SDXL	2.6B	5B+	5.94	22.0	0.25	0.46
SD3-Medium	2B	1B+	6.18	22.5	0.30	1.15
FLUX-dev	12B	-	**6.56**	**22.9**	**0.30**	<u>1.19</u>
Ours (ft. Laion-POP 1024^2)	0.4B	1.5M	6.28	21.6	<u>0.29</u>	0.64

Table 11: **Results on Reward Metrics.** Results are computed using the PartiPrompts Yu et al. (2022). SOTA scores are computing using HuggingFace checkpoints at their native resolution. **Bold** indicates best, <u>underline</u> second best.

Finetuning on high-aesthetic dataset

Ours | **SDXL** | **Pixart-a** | **SD3-M**

A fox with glasses and dressed in white shirt and khaki cargo

Finetuning on high-aesthetic dataset

SDXL **Pixart-a** **SD3-M** **Ours**

Plants, flowers, trees being mixed in a bowl

Finetuning on high-aesthetic dataset

A harsh winter landscape with mountains, a river, and forest, where a lone man walks through deep snow beneath birds flying

A cat with bunny ears

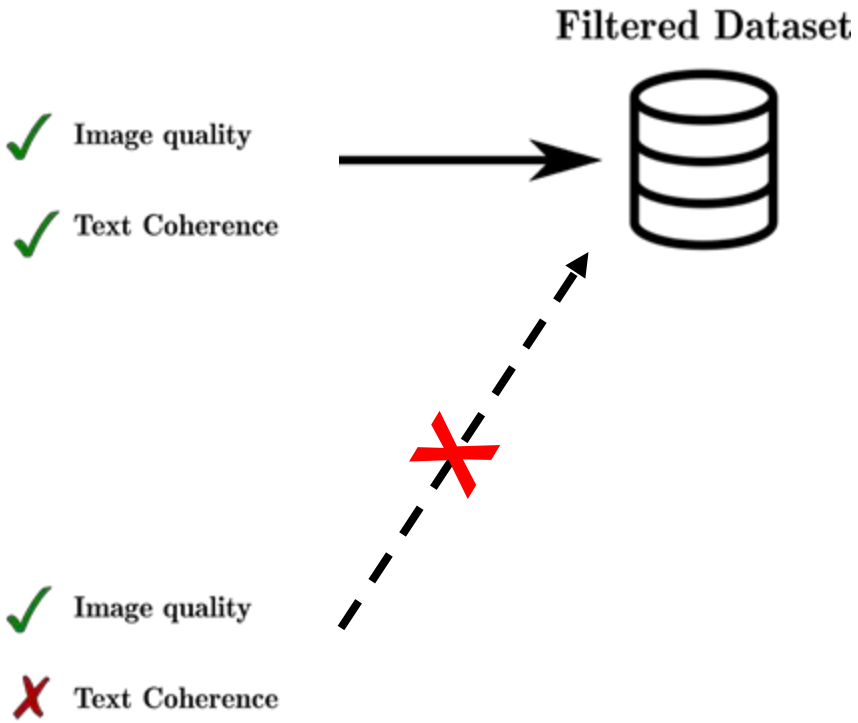
Conclusions

- Text-to-image generation: data is ~~not~~ enough
- From billion-scale datasets → 1M image datasets with **augmentations**
- Performance on par to modern models, while trained with
  - x10 fewer #params and
  - x10^2-10^3 fewer #data

Related work
Collect billions of data, filter and discard

Two Impala Rams squaring off.

The food on Jeju Island was fishy, in the best possible way. Jeju Island, Sea Urchin, Noodle Soup, Noodles, Spaghetti, Eat, Ethnic Recipes, Food, Macaroni Pasta



Filtering discards useful data!

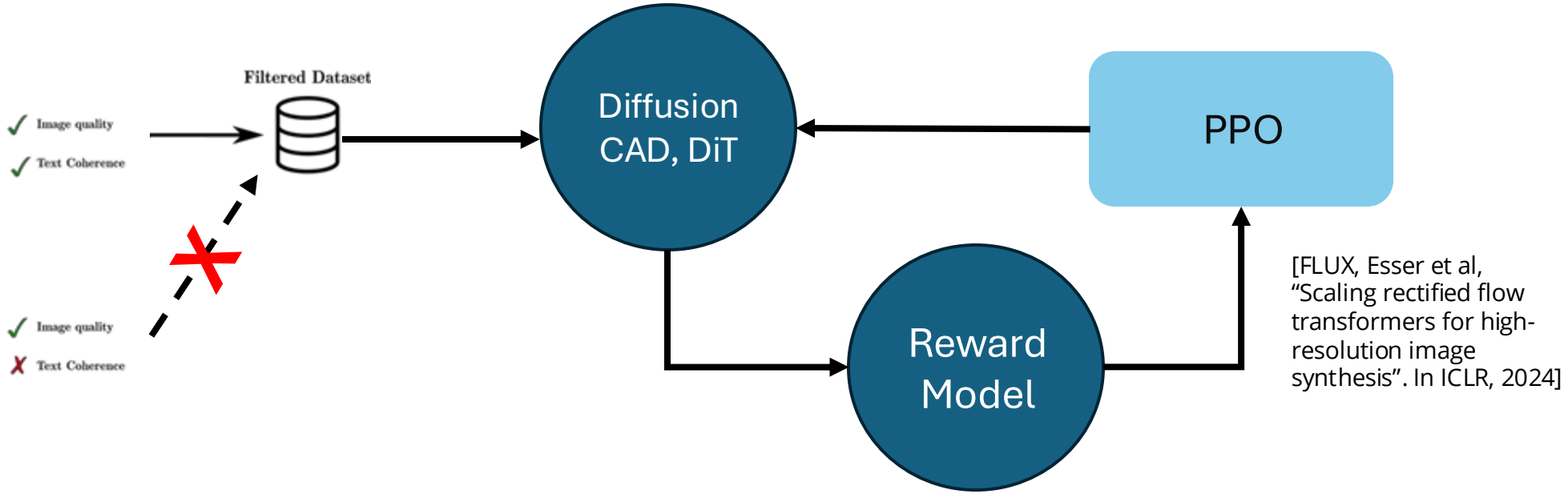
- Datasets: filtered on a coherence score:
  - measures how coherent the label is with the data
- → Discard too noisy annotations

Stable diffusion, Imagen, E-Diffi

Related work
Aligning with human preferences

Two Impala Rams squaring off.

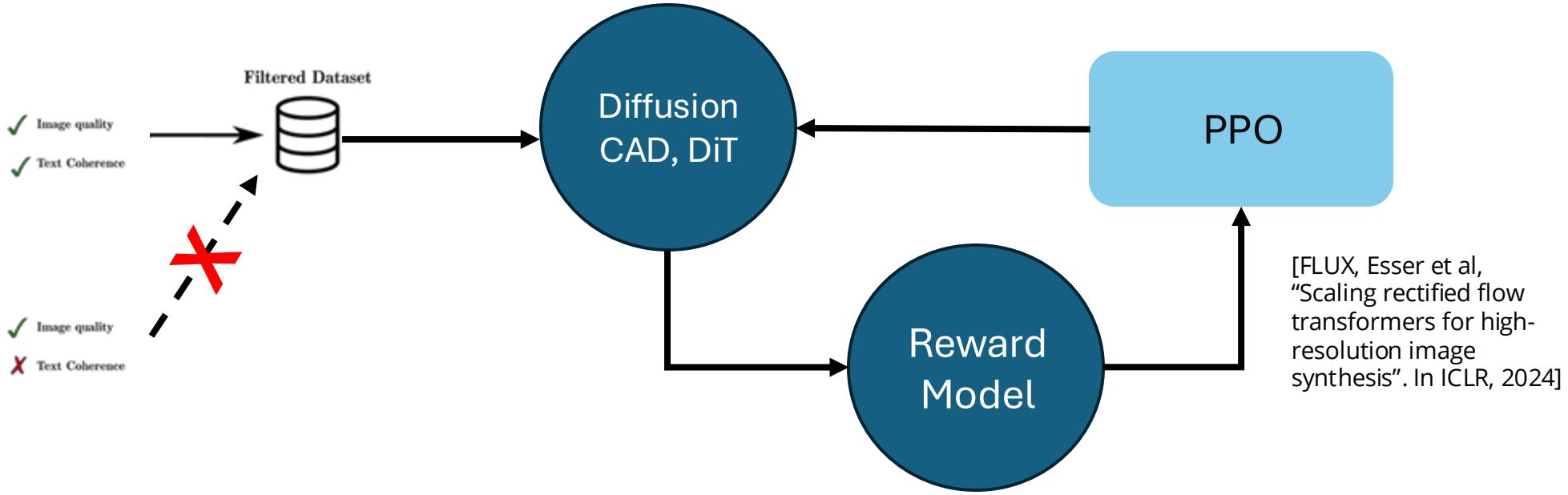
The food on Jeju Island was fishy, in the best possible way. Jeju Island, Sea Urchin, Noodle Soup, Noodles, Spaghetti, Eat, Ethnic Recipes, Food, Macaroni Pasta



[FLUX, Esser et al, "Scaling rectified flow transformers for high-resolution image synthesis". In ICLR, 2024]

❌ Discards informative "low-quality" data
❌ Complicates training with additional optimization step
❌ Overfit to a single reward →
❌ harming diversity (mode collapse) or semantic fidelity and efficiency

Related work
Aligning with human preferences



[FLUX, Esser et al, "Scaling rectified flow transformers for high-resolution image synthesis". In ICLR, 2024]

Rather than correcting a pre-trained text-to-image model

Can we teach model how to trade off multiple rewards *from the beginning*?

Efficiency: Challenges

Training data
- Collecting, filtering
- Privacy

Model
- Large model size
- Scaling resolution

Training & conditioning
- How to condition?
- Long training times

Inference & post-training
- Multiple denoising steps
- Apply RL?

Efficiency

Training & conditioning
- How to condition?
- Long training times

Model
- Large model size
- Scaling resolution

Inference & post-training
- Multiple denoising steps
- Apply RL?

Training data
- Collecting, filtering
- Privacy

MIRO: Multi-Reward Conditioned Retraining improves text-to-image quality and efficiency




Nicolas Dufour, Lucas Degeorge*, Arijit Ghosh*, David Picard, Vicky Kalogeiton
arXiv 2025

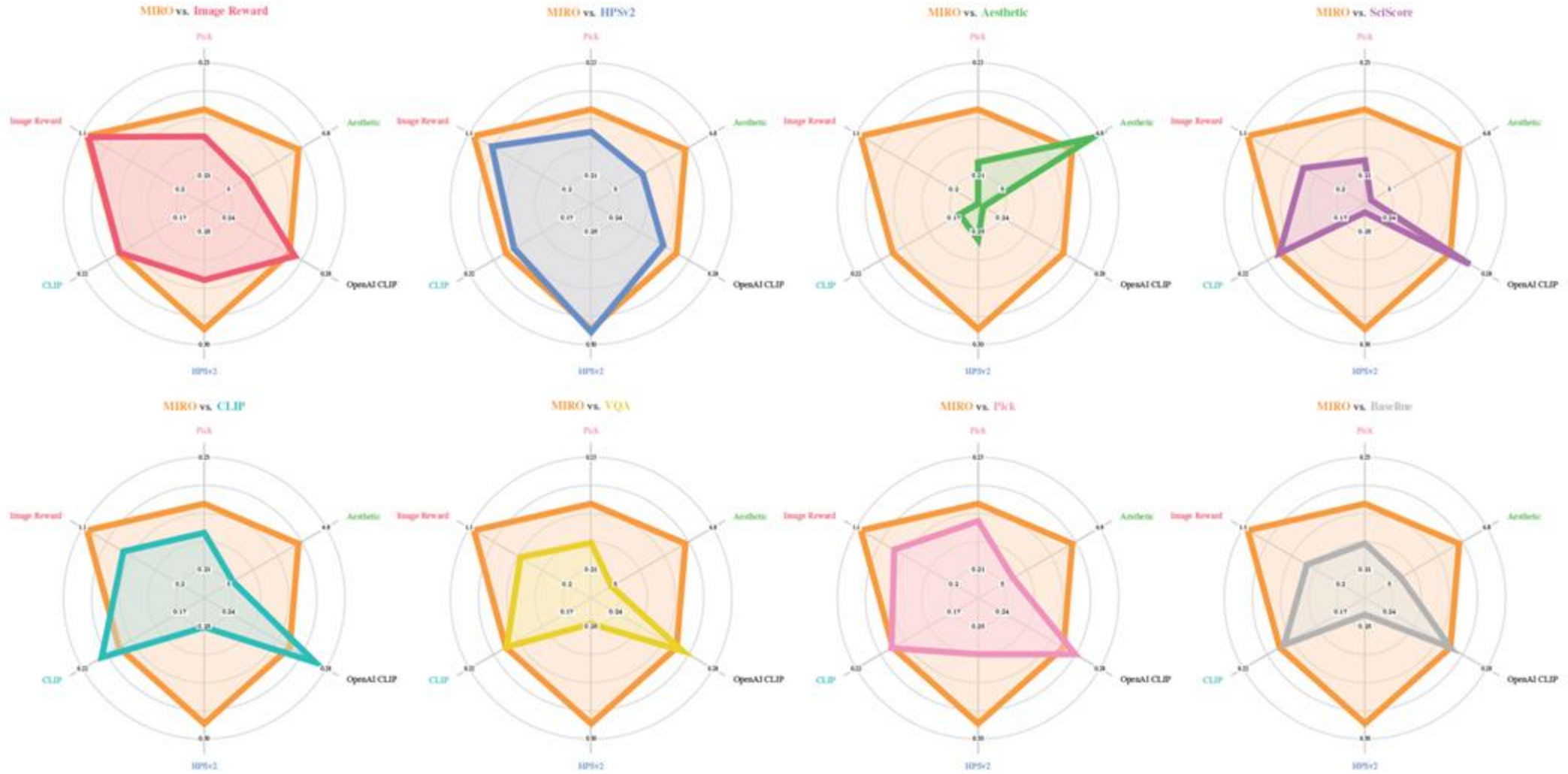
Qualitative results

MIRO: improves model quality

1. Outperforms single-reward approaches across all metrics
2. Mitigates reward hacking
3. Accelerates training convergence
4. Outperforms synthetic captioning alone
5. MIRO + synthetic captions: strongest performance
6. Enhances compositional understanding
7. Single-reward models exhibit varying alignment capabilities

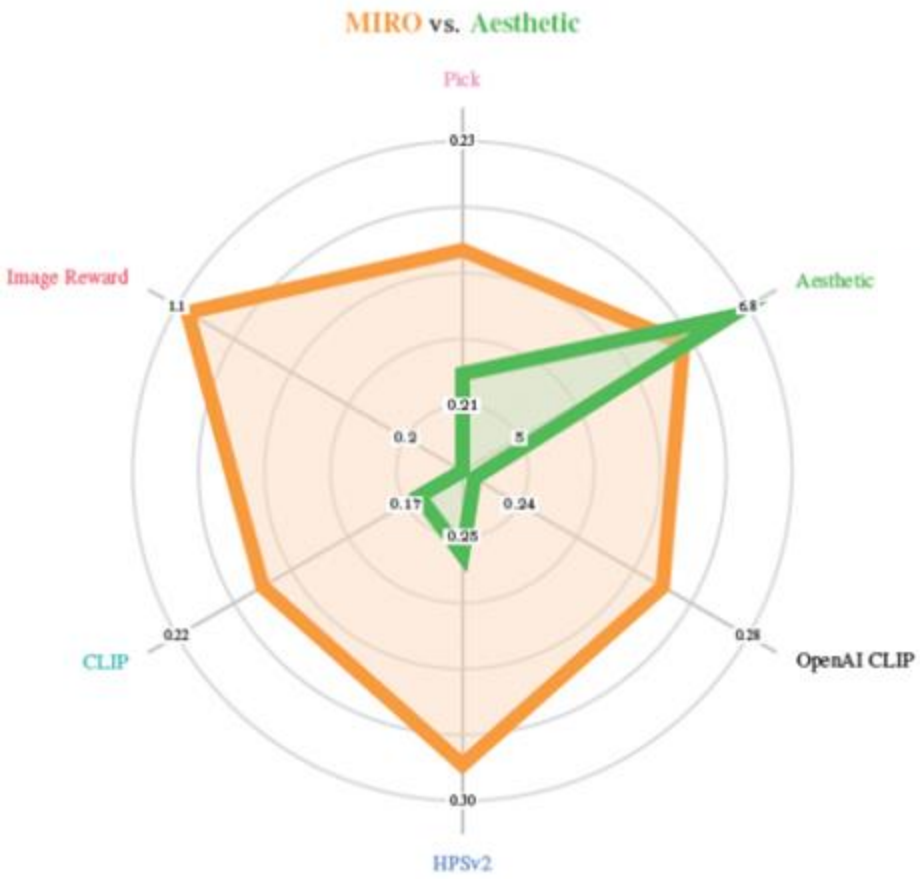
Leveraging multiple rewards

1. Outperforms single-reward approaches across all metrics



Leveraging multiple rewards

2. Mitigates reward hacking



Leveraging multiple rewards

3. Accelerates training convergence

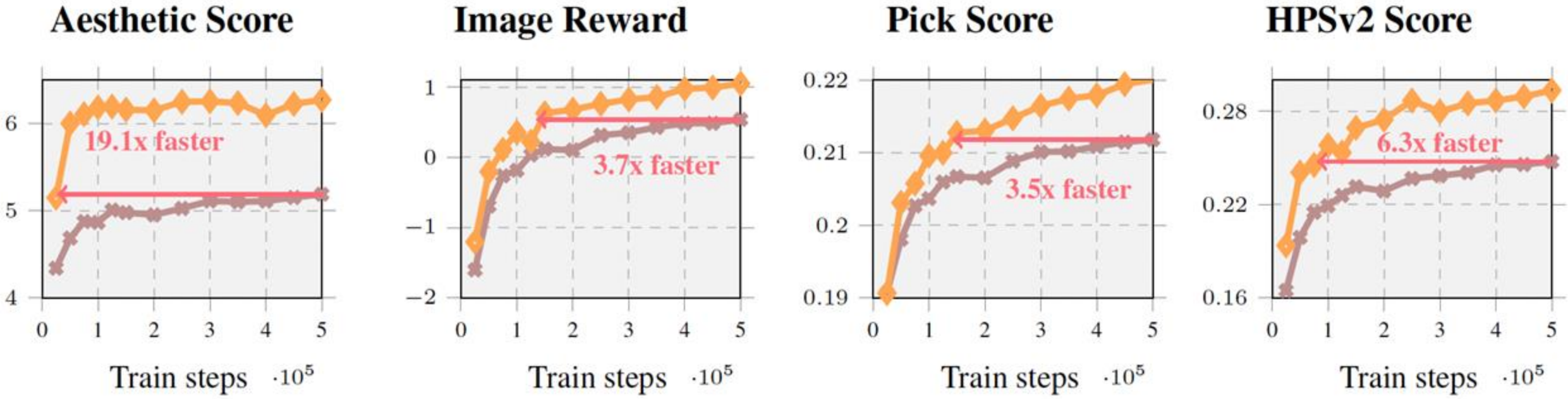
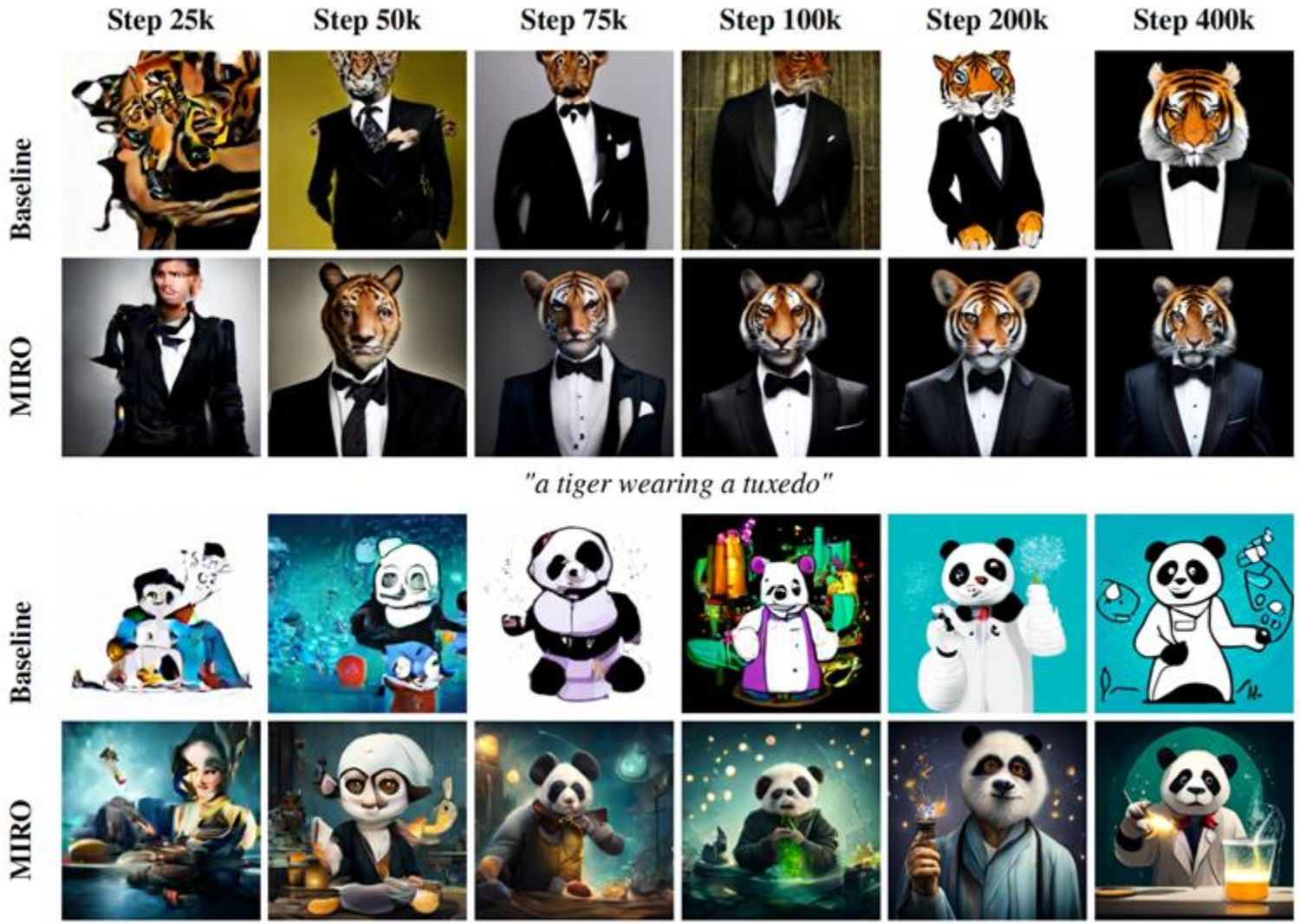


Figure 3: Training curves showing reward evolution during training. × Baseline, ◇ MIRO.

Leveraging multiple rewards

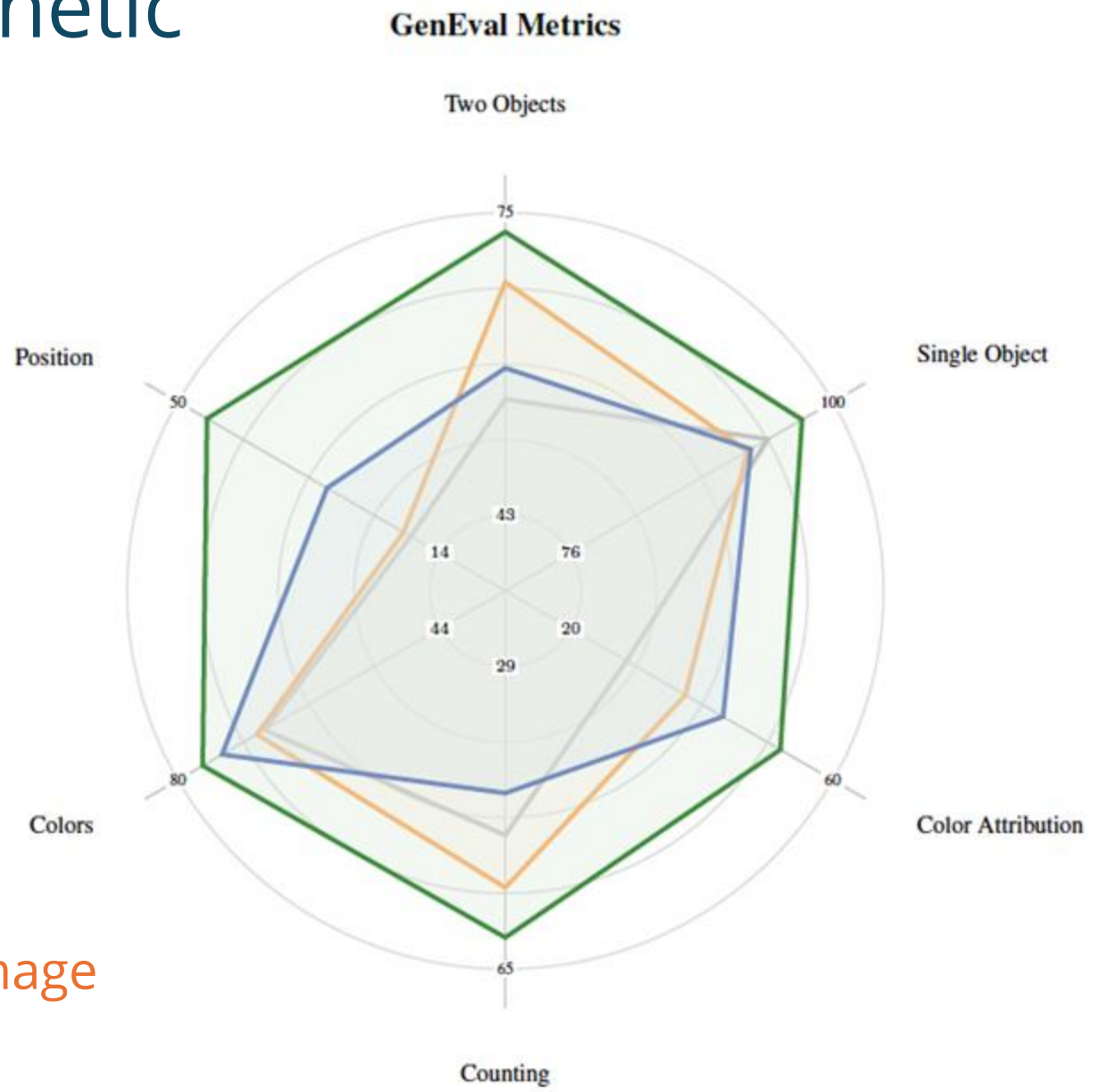
3. Accelerates training convergence



4. MIRO outperforms synthetic captioning alone

Legend (GenEval Overall):

- Baseline - **52.2** (Real only)
- MIRO - **56.8** (Real only)
- Synth Baseline - **56.9** (50% Real + 50% Synth)
- Synth MIRO - **67.7** (50% Real + 50% Synth)



More effective approach to improving text-image alignment than synthetic captioning alone

Model	Params (B)	Inference TFLOPs	GenEval							PartiPrompts			
			Overall	Single Obj.	Two Obj.	Position	Counting	Colors	Color Attr.	Aesthetic	Image	HPSv2	PickAScore
SOTA Baselines													
SD v1.5	0.9	-	43	97	38	4	35	76	6	5.68	0.24	0.25	0.213
SD v2.1	0.9	-	50	98	51	7	44	85	17	5.81	0.38	0.26	0.215
PixArt-α	0.6	-	48	98	50	8	44	80	7	6.47	0.97	0.29	0.226
PixArt-Σ	0.6	-	52	98	59	10	50	80	15	6.44	1.02	0.29	0.225
CAD	0.35	20.8	50	95	56	11	40	76	22	5.56	0.69	0.26	0.214
Sana-0.6B	0.6	-	64	**99**	71	16	63	**91**	42	6.31	1.23	0.30	0.228
Sana-1.6B	1.6	-	66	**99**	79	18	63	88	47	6.36	1.23	0.30	0.228
SDXL	2.6	-	55	98	74	15	39	85	23	5.94	0.46	0.25	0.220
SD3-medium	2.0	-	62	98	74	34	63	67	36	6.18	1.15	0.30	0.225
FLUX-dev	12.0	1540	67	**99**	81	20	**79**	74	47	6.56	1.19	0.30	**0.229**
CAD-like Models (our models)													
Image Reward	0.36	4.16	57	97	59	21	56	76	33	5.31	1.04	0.27	0.214
HPSv2	0.36	4.16	56	95	63	15	52	78	31	5.47	0.90	0.29	0.215
Aesthetic	0.36	4.16	33	74	37	6	24	42	15	6.65	0.00	0.26	0.209
SciScore	0.36	4.16	58	94	62	24	61	72	35	4.62	0.56	0.24	0.209
CLIP	0.36	4.16	57	97	63	24	57	70	32	5.04	0.73	0.25	0.214
VQA	0.36	4.16	57	97	58	20	57	76	37	4.88	0.64	0.25	0.212
Pick	0.36	4.16	57	93	62	17	58	75	34	5.16	0.76	0.26	0.216
Real Caption Models (our models)													
Baseline	0.36	4.16	52	94	55	18	49	68	29	5.18	0.52	0.25	0.212
MIRO	0.36	4.16	57	92	68	19	55	69	38	6.28	1.06	0.29	0.220
Synthetic Caption Models (50% Real + 50% Synth) (our models)													
Baseline	0.36	4.16	57	93	59	30	44	74	43	4.96	0.52	0.24	0.211
MIRO	0.36	4.16	68	97	73	46	61	77	52	6.28	1.11	0.29	0.220
Inference Scaled + Synthetic Caption Models (MIRO + 128 samples inference scaled) (our models)													
Aesthetic Scaled MIRO	0.36	532	63	97	68	40	57	75	45	**6.81**	1.04	0.29	0.219
Image Reward Scaled MIRO	0.36	532	**75**	98	**84**	**52**	69	82	**65**	6.28	**1.61**	0.30	0.223
HPSv2 Scaled MIRO	0.36	532	74	98	83	47	74	80	**65**	6.28	1.35	**0.32**	0.225
PickAScore Scaled MIRO	0.36	532	74	98	83	44	76	81	59	6.27	1.32	0.31	**0.229**

Model	Params (B)	Inference TFLOPs	GenEval							PartiPrompts			
			Overall	Single Obj.	Two Obj.	Position	Counting	Colors	Color Attr.	Aesthetic	Image	HPSv2	PickAScore
SOTA Baselines													
SD v1.5	0.9	-	43	97	38	4	35	76	6	5.68	0.24	0.25	0.213
SD v2.1	0.9	-	50	98	51	7	44	85	17	5.81	0.38	0.26	0.215
PixArt-α	0.6	-	48	98	50	8	44	80	7	6.47	0.97	0.29	0.226
PixArt-Σ	0.6	-	52	98	59	10	50	80	15	6.44	1.02	0.29	0.225
CAD	0.35	20.8	50	95	56	11	40	76	22	5.56	0.69	0.26	0.214
Sana-0.6B	0.6	-	64	**99**	71	16	63	**91**	42	6.31	1.23	0.30	0.228
Sana-1.6B	1.6	-	66	**99**	79	18	63	88	47	6.36	1.23	0.30	0.228
SDXL	2.6	-	55	98	74	15	39	85	23	5.94	0.46	0.25	0.220
SD3-medium	2.0	-	62	98	74	34	63	67	36	6.18	1.15	0.30	0.225
FLUX-dev	12.0	1540	67	**99**	81	20	**79**	74	47	6.56	1.19	0.30	**0.229**
CAD-like Models (our models)													
Image Reward	0.36	4.16	57	97	59	21	56	76	33	5.31	1.04	0.27	0.214
HPSv2	0.36	4.16	56	95	63	15	52	78	31	5.47	0.90	0.29	0.215
Aesthetic	0.36	4.16	33	74	37	6	24	42	15	6.65	0.00	0.26	0.209
SciScore	0.36	4.16	58	94	62	24	61	72	35	4.62	0.56	0.24	0.209
CLIP	0.36	4.16	57	97	63	24	57	70	32	5.04	0.73	0.25	0.214
VQA	0.36	4.16	57	97	58	20	57	76	37	4.88	0.64	0.25	0.212
Pick	0.36	4.16	57	93	62	17	58	75	34	5.16	0.76	0.26	0.216
Real Caption Models (our models)													
Baseline	0.36	4.16	52	94	55	18	49	68	29	5.18	0.52	0.25	0.212
MIRO	0.36	4.16	57	92	68	19	55	69	38	6.28	1.06	0.29	0.220
Synthetic Caption Models (50% Real + 50% Synth) (our models)													
Baseline													0.211
MIRO													0.220
Inference Scaled + Synthetic Caption Models (MIRO + 128 samples inference scaled) (our models)													
Aesthetic Scaled MIRO	0.36	532	63	97	68	40	57	75	45	**6.81**	1.04	0.29	0.219
Image Reward Scaled MIRO	0.36	532	**75**	98	**84**	**52**	69	82	**65**	6.28	**1.61**	0.30	0.223
HPSv2 Scaled MIRO	0.36	532	74	98	83	47	74	80	**65**	6.28	1.35	**0.32**	0.225
PickAScore Scaled MIRO	0.36	532	74	98	83	44	76	81	59	6.27	1.32	0.31	**0.229**

5. MIRO + synthetic captions: strongest performance

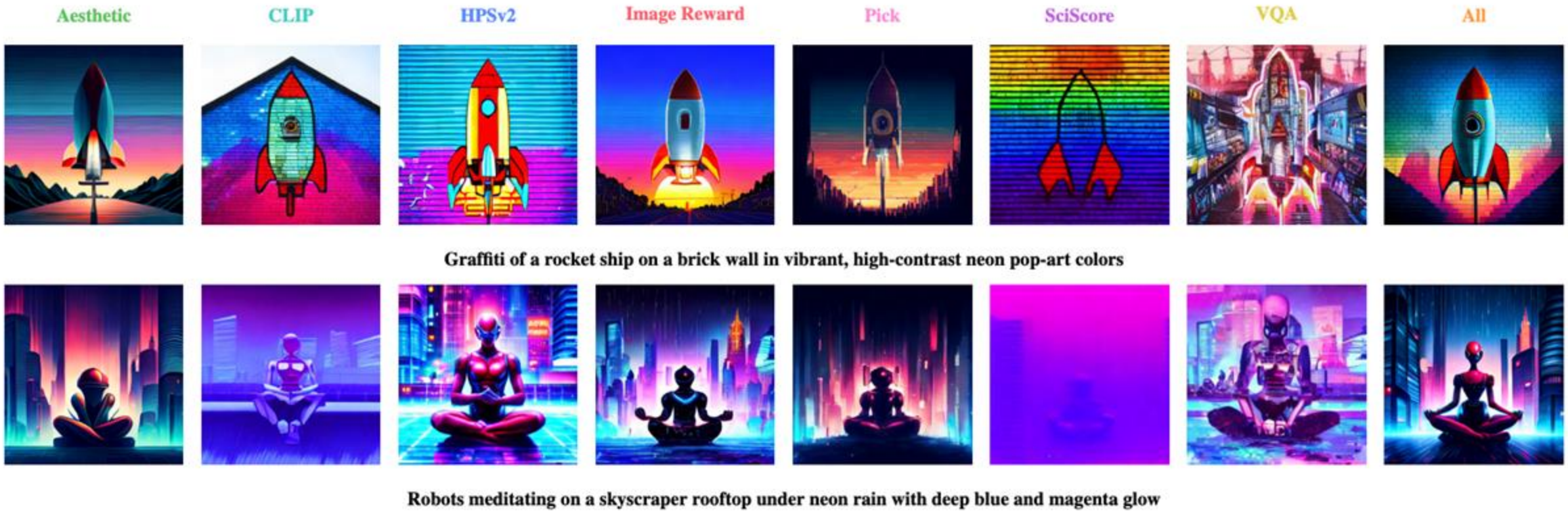
Model	Params (B)	Inference TFLOPs	GenEval							PartiPrompts			
			Overall	Single Obj.	Two Obj.	Position	Counting	Colors	Color Attr.	Aesthetic	Image	HPSv2	PickAScore
SOTA Baselines													
SD v1.5	0.9	-	43	97	38	4	35	76	6	5.68	0.24	0.25	0.213
SD v2.1	0.9	-	50	98	51	7	44	85	17	5.81	0.38	0.26	0.215
PixArt-α	0.6	-	48	98	50	8	44	80	7	6.47	0.97	0.29	0.226
PixArt-Σ	0.6	-	52	98	59	10	50	80	15	6.44	1.02	0.29	0.225
CAD	0.35	20.8	50	95	56	11	40	76	22	5.56	0.69	0.26	0.214
Sana-0.6B	0.6	-	64	**99**	71	16	63	**91**	42	6.31	1.23	0.30	0.228
Sana-1.6B	1.6	-	66	**99**	79	18	63	88	47	6.36	1.23	0.30	0.228
SDXL	2.6	-	55	98	74	15	39	85	23	5.94	0.46	0.25	0.220
SD3-medium	2.0	-	62	98	74	34	63	67	36	6.18	1.15	0.30	0.225
FLUX-dev	12.0	1540	67	**99**	81	20	**79**	74	47	6.56	1.19	0.30	**0.229**
CAD-like Models (our models)													
Image Reward	0.36	4.16	57	97	59	21	56	76	33	5.31	1.04	0.27	0.214
HPSv2	0.36	4.16	56	95	63	15	52	78	31	5.47	0.90	0.29	0.215
Aesthetic	0.36	4.16	33	74	37	6	24	42	15	6.65	0.00	0.26	0.209
SciScore	0.36	4.16	58	94	62	24	61	72	35	4.62	0.56	0.24	0.209
CLIP													0.214
VQA													0.212
Pick													0.216
Real Caption Models (our models)													
Baseline	0.36	4.16	52	94	55	18	49	68	29	5.18	0.52	0.25	0.212
MIRO	0.36	4.16	57	92	68	19	55	69	38	6.28	1.06	0.29	0.220
Synthetic Caption Models (50% Real + 50% Synth) (our models)													
Baseline	0.36	4.16	57	93	59	30	44	74	43	4.96	0.52	0.24	0.211
MIRO	0.36	4.16	68	97	73	46	61	77	52	6.28	1.11	0.29	0.220
Inference Scaled + Synthetic Caption Models (MIRO + 128 samples inference scaled) (our models)													
Aesthetic Scaled MIRO	0.36	532	63	97	68	40	57	75	45	**6.81**	1.04	0.29	0.219
Image Reward Scaled MIRO	0.36	532	**75**	98	**84**	**52**	69	82	**65**	6.28	**1.61**	0.30	0.223
HPSv2 Scaled MIRO	0.36	532	74	98	83	47	74	80	**65**	6.28	1.35	**0.32**	0.225
PickAScore Scaled MIRO	0.36	532	74	98	83	44	76	81	59	6.27	1.32	0.31	**0.229**

6. Enhances compositional understanding

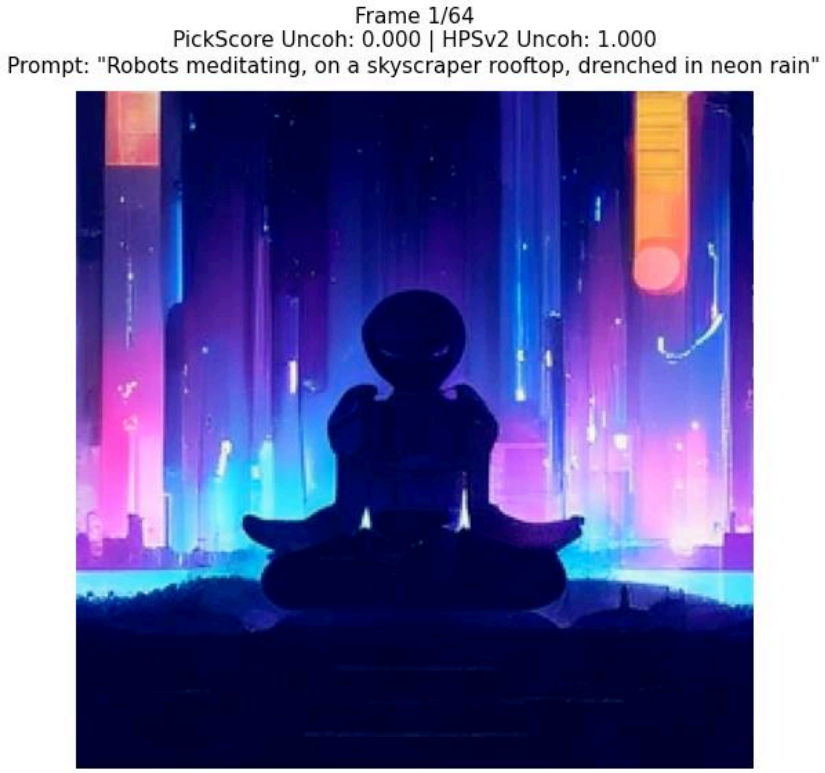
Model	Params (B)	Inference TFLOPs	GenEval Overall	Single Obj.	Two Obj.	Position	Counting	Colors	Color Attr.	PartiPrompts Aesthetic	Image	HPSv2	PickAScore
SOTA Baselines													
SD v1.5	0.9	-	43	97	38	4	35	76	6	5.68	0.24	0.25	0.213
SD v2.1	0.9	-	50	98	51	7	44	85	17	5.81	0.38	0.26	0.215
PixArt-α	0.6	-	48	98	50	8	44	80	7	6.47	0.97	0.29	0.226
PixArt-Σ	0.6	-	52	98	59	10	50	80	15	6.44	1.02	0.29	0.225
CAD	0.35	20.8	50	95	56	11	40	76	22	5.56	0.69	0.26	0.214
Sana-0.6B	0.6	-	64	**99**	71	16	63	**91**	42	6.31	1.23	0.30	0.228
Sana-1.6B	1.6	-	66	**99**	79	18	63	88	47	6.36	1.23	0.30	0.228
SDXL													
SD3-medium													
FLUX-dev													
CAD-like Models (our models)													
Image Reward	0.36	4.16	57	97	59	21	56	76	33	5.31	1.04	0.27	0.214
HPSv2	0.36	4.16	56	95	63	15	52	78	31	5.47	0.90	0.29	0.215
Aesthetic	0.36	4.16	33	74	37	6	24	42	15	6.65	0.00	0.26	0.209
SciScore	0.36	4.16	58	94	62	24	61	72	35	4.62	0.56	0.24	0.209
CLIP	0.36	4.16	57	97	63	24	57	70	32	5.04	0.73	0.25	0.214
VQA	0.36	4.16	57	97	58	20	57	76	37	4.88	0.64	0.25	0.212
Pick	0.36	4.16	57	93	62	17	58	75	34	5.16	0.76	0.26	0.216
Real Caption Models (our models)													
Baseline	0.36	4.16	52	94	55	18	49	68	29	5.18	0.52	0.25	0.212
MIRO	0.36	4.16	57	92	68	19	55	69	38	6.28	1.06	0.29	0.220
Synthetic Caption Models (50% Real + 50% Synth) (our models)													
Baseline	0.36	4.16	57	93	59	30	44	74	43	4.96	0.52	0.24	0.211
MIRO	0.36	4.16	68	97	73	46	61	77	52	6.28	1.11	0.29	0.220
Inference Scaled + Synthetic Caption Models (MIRO + 128 samples inference scaled) (our models)													
Aesthetic Scaled MIRO	0.36	532	63	97	68	40	57	75	45	**6.81**	1.04	0.29	0.219
Image Reward Scaled MIRO	0.36	532	**75**	98	**84**	**52**	69	82	**65**	6.28	**1.61**	0.30	0.223
HPSv2 Scaled MIRO	0.36	532	74	98	83	47	74	80	**65**	6.28	1.35	**0.32**	0.225
PickAScore Scaled MIRO	0.36	532	74	98	83	44	76	81	59	6.27	1.32	0.31	**0.229**

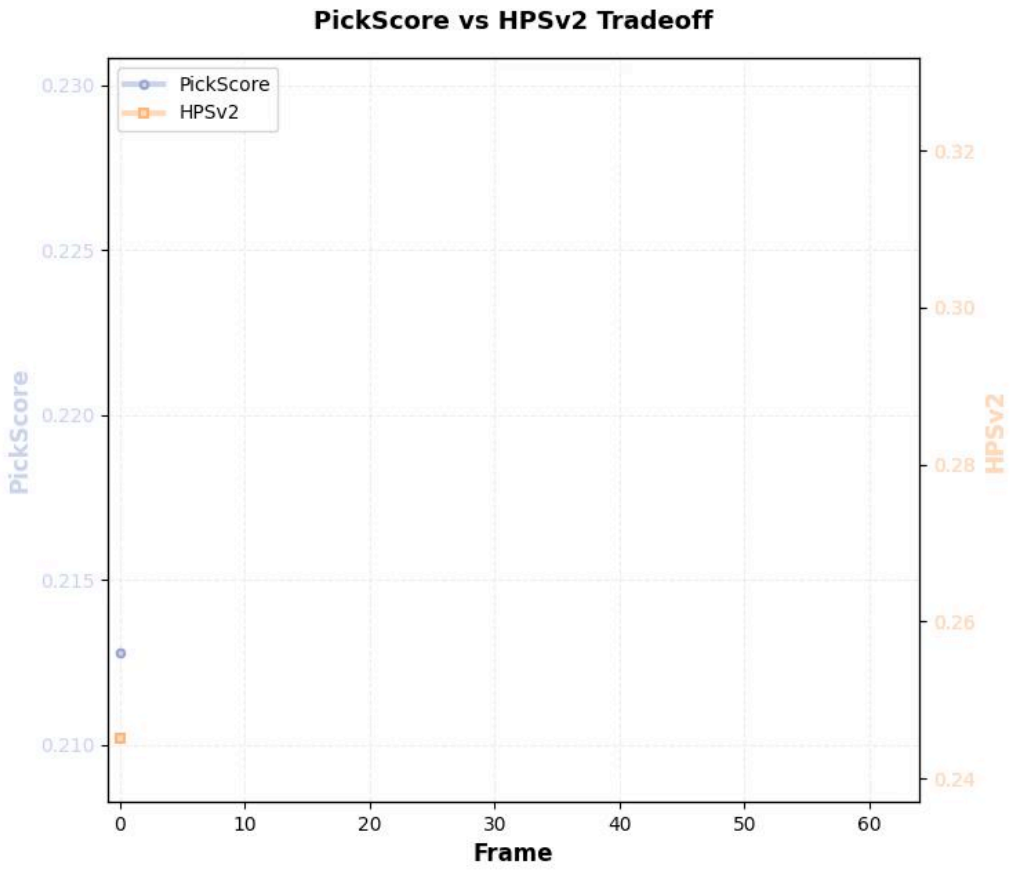
7. Single-reward models exhibit varying alignment capabilities

Reward controllability

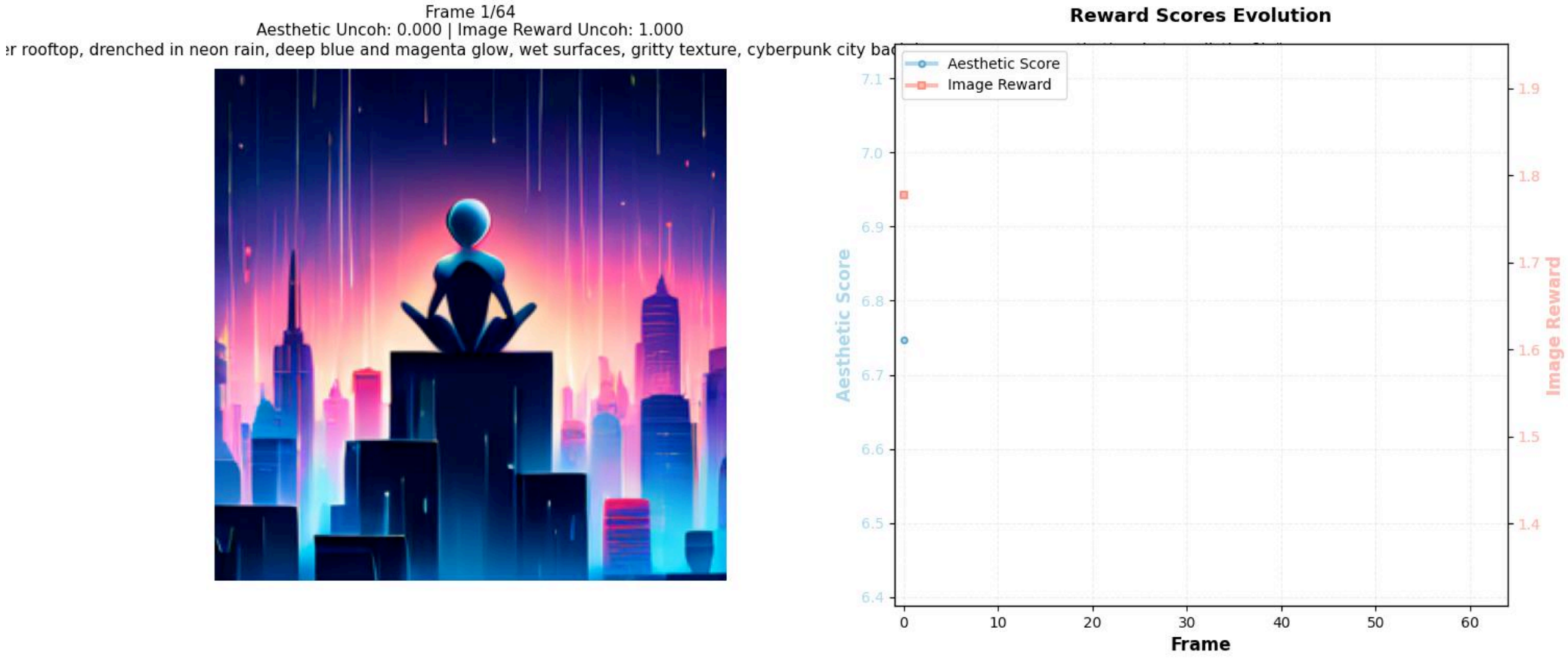


Trade-off animation between aesthetic-reward scores

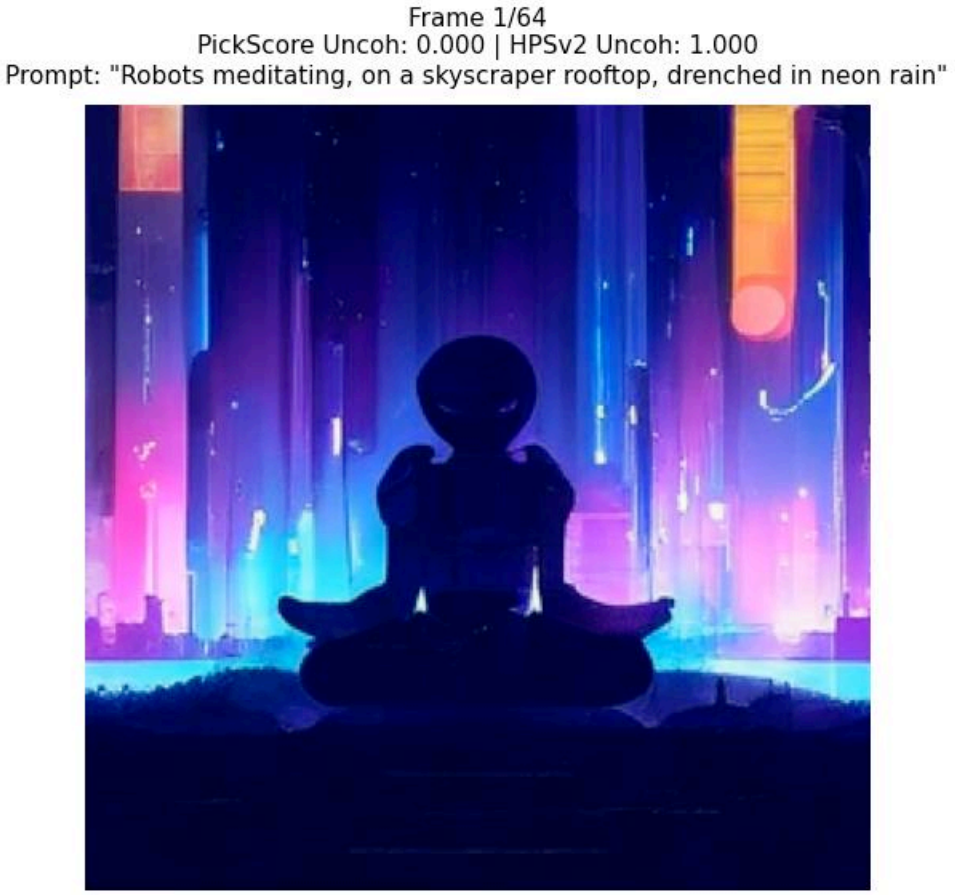


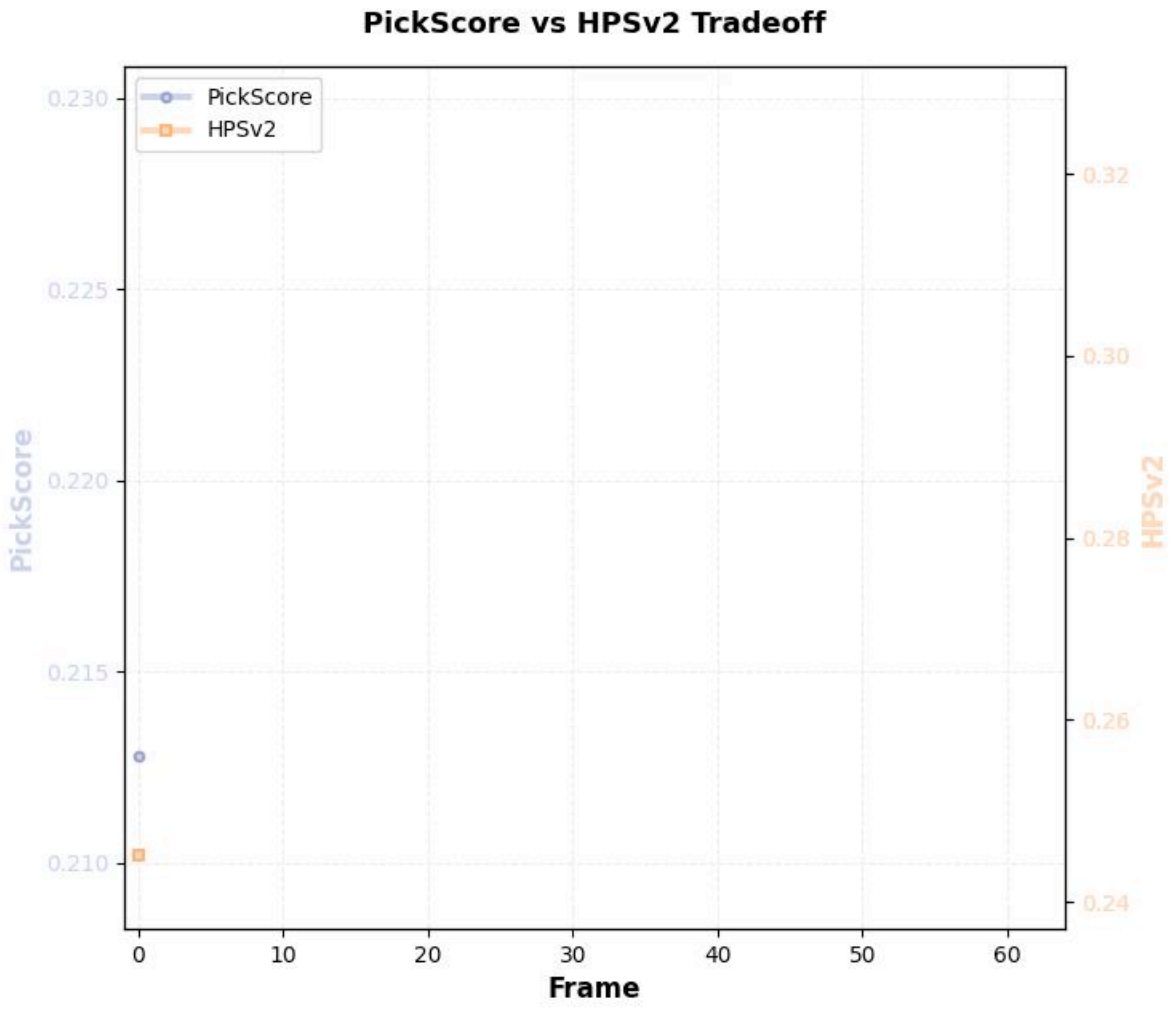


Trade-off animation between aesthetic-reward scores

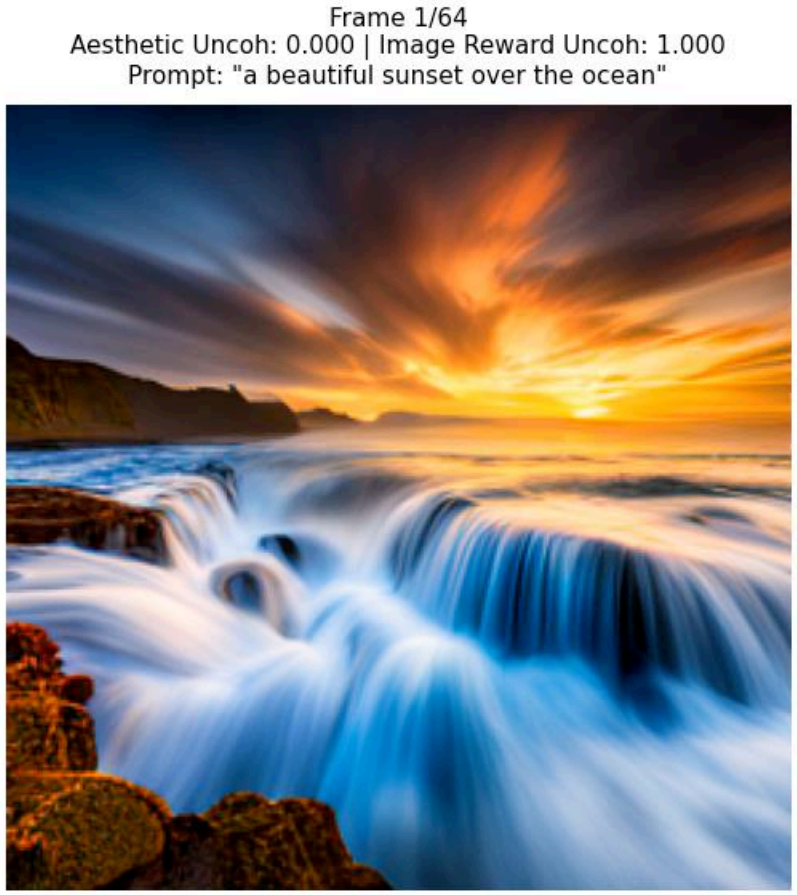


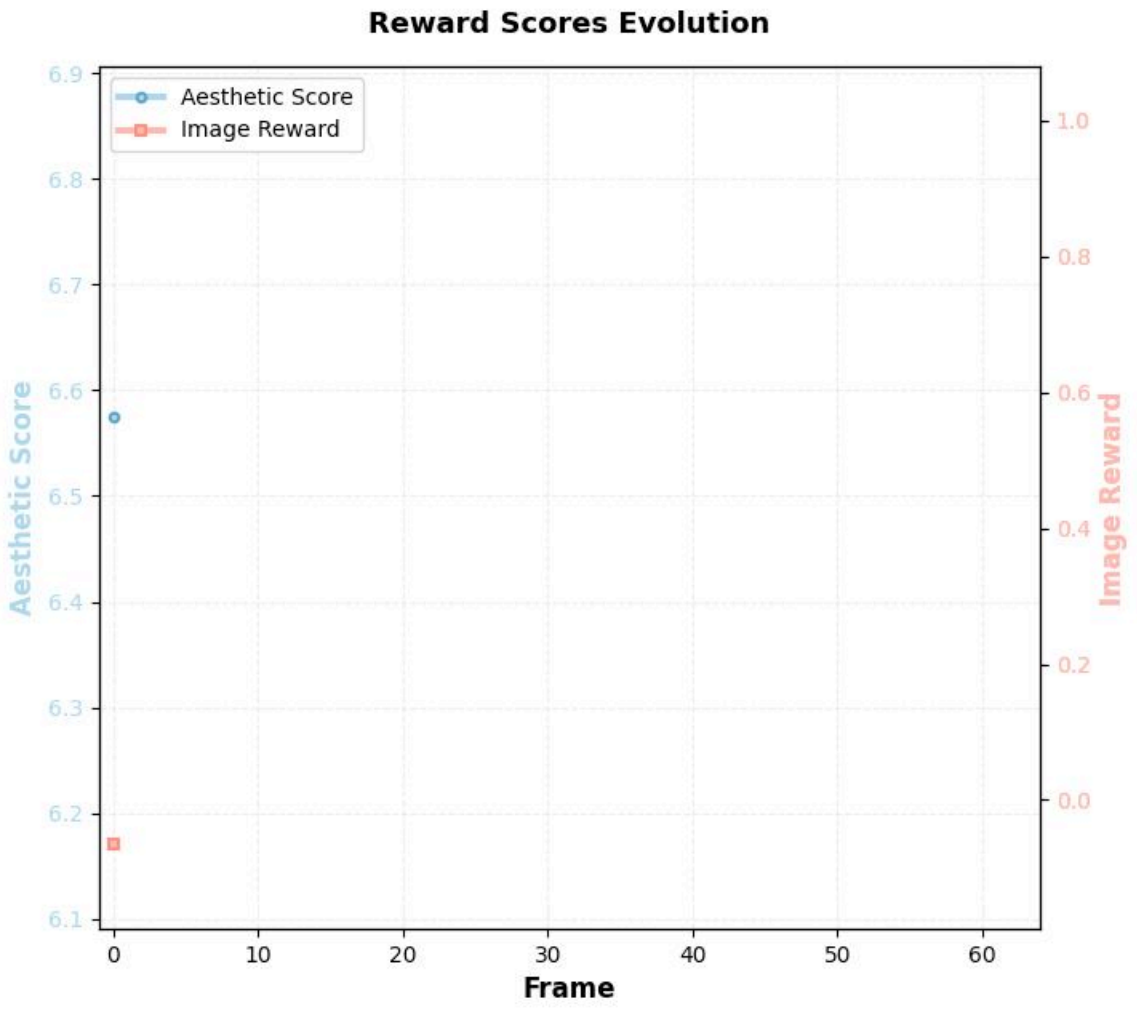
Trade-off animation between aesthetic-reward scores





Trade-off animation between aesthetic-reward scores





Test-time scaling

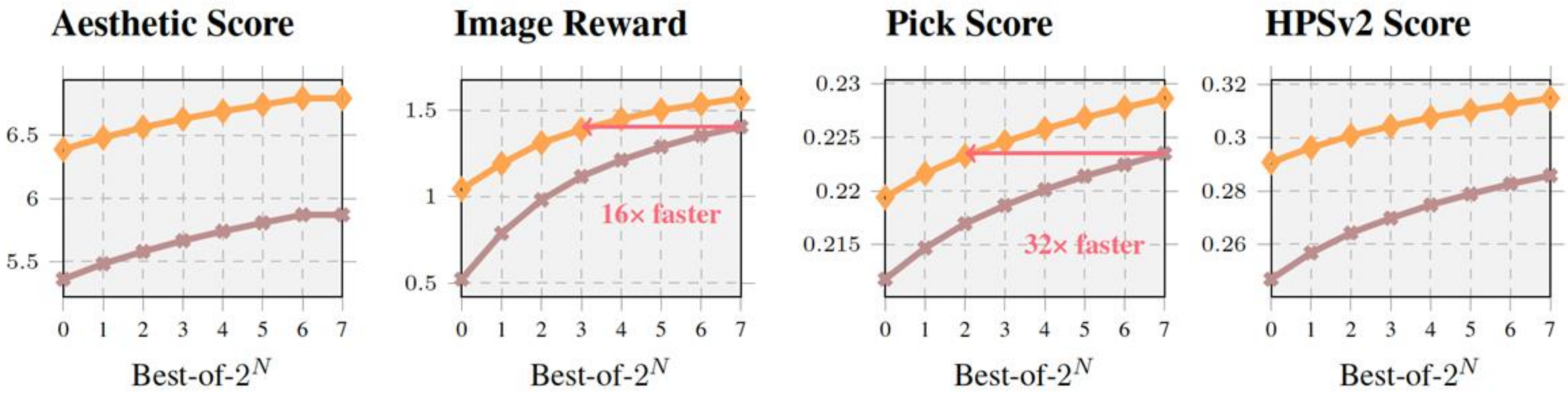
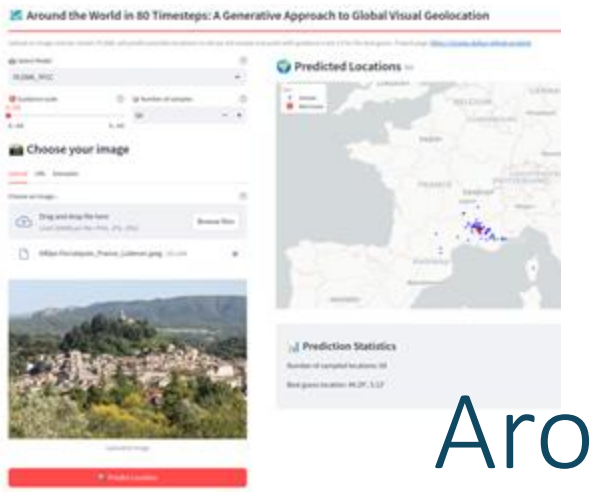


Figure 6: Test-time scaling showing performance vs. Best-of-2^N sampling. × Baseline, ◇ MIRO.

Conclusions

- MIRO:
  - Condition on a vector of reward scores to integrate alignment into training vs post-hoc
- Goodies:
  - outperforms no-conditioning and single-reward baselines
  - converges substantially faster
  - mitigates reward hacking
  - strengthens compositional alignment
- State-of-the-art results
  - on PartiPrompts w/ inference-time scaling, while more compute-efficient
  - 🏆 Outperforms FLUX-dev on GenEval and PartiPrompts at a fraction of the compute
- Future:
  - MIRO trained on ImageNet
  - MIRO + DiT
  - Personalized rewards? Can we easily find your style as a function of other rewards and give you the style you prefer

Bonus
(For Jiří: added 30 minutes before presentation!)



Around the World in 80 Time Steps
A Generative Approach to Global Geolocation

https://nicolas-dufour.github.io/plonk

Nicolas Dufour, David Picard, Vicky Kalogeiton, Loic Landrieu
CVPR 2025

🎈 Geolocation as Coordinates/Distribution generation

Challenge: ambiguity

- Canada? Norway? Elsewhere?

- Regression predicts the mean of two modes, one per region

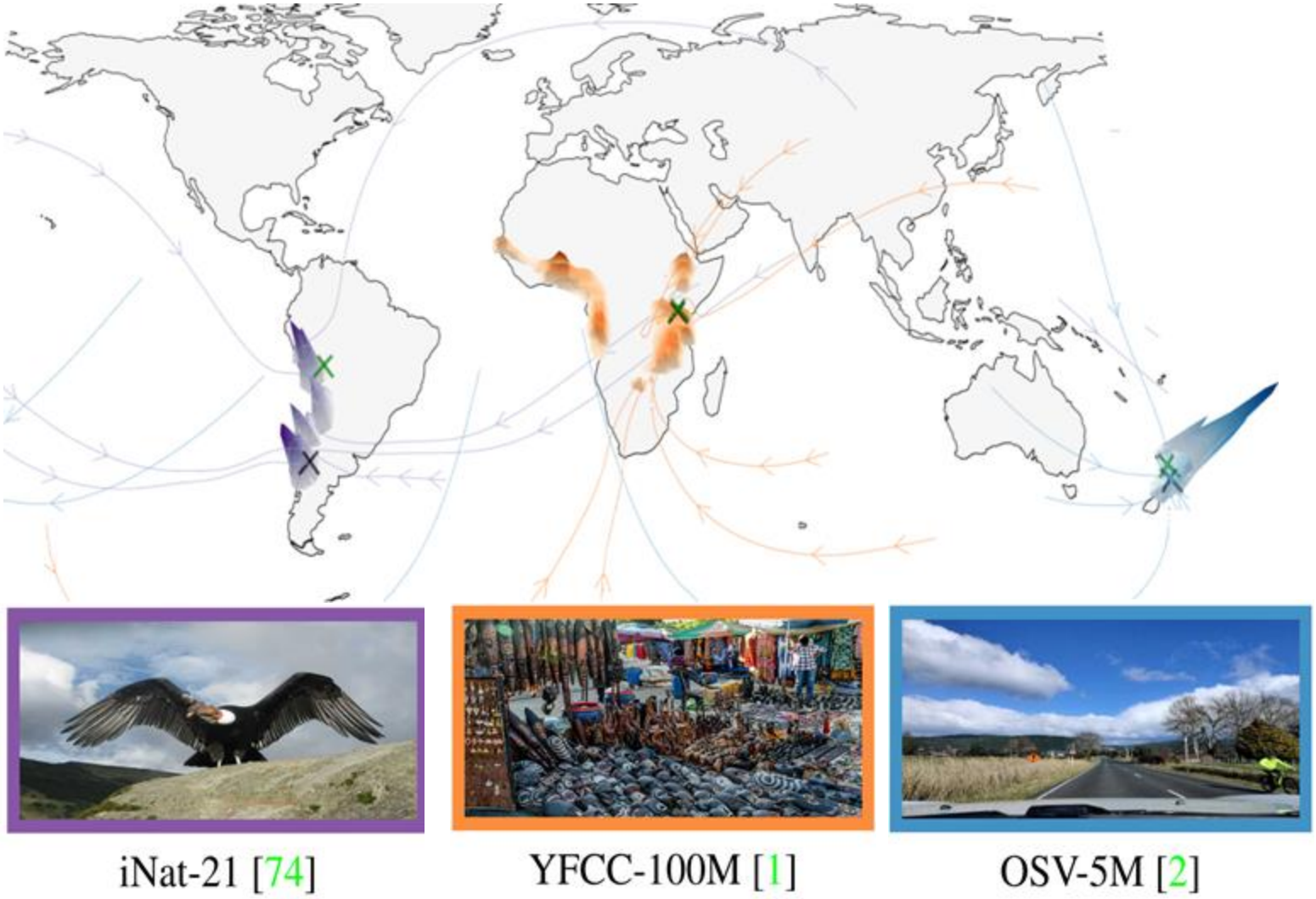
Middle of the Atlantic
Wrong prediction!

Plonk: Generative Geolocalization

- **Solution:** predict (conditional) distributions instead of “hard” locations

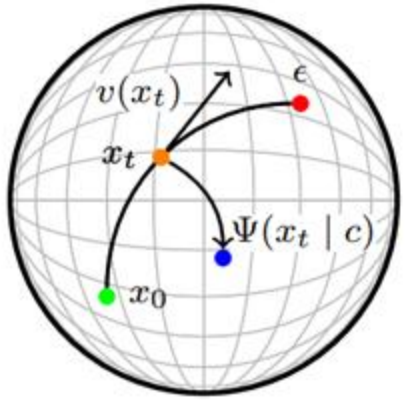
- Sample multiple locations to access the distribution of possible locations

- More interpretable



Geolocation as a Generative Process

- **Method**: we use Diffusion / Flow Matching techniques
- **In Practice:** we learn to “correct” noisy coordinates, given an image

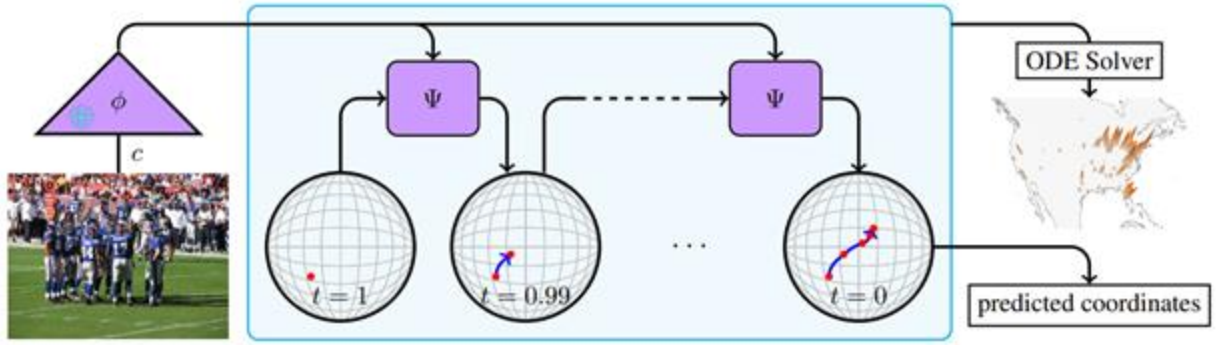


- x_0: true location
- ϵ: sampled noise
- x_t: noisy location
- $\psi(x_t \mid c)$: prediction
- $\rightarrow v(x_t)$: velocity field

Diffusion
$x_t = \sqrt{1-\kappa(t)}x_0 + \sqrt{\kappa(t)}\epsilon$ $\mathcal{L}_{\text{D}} = \|\psi(x_t \mid c) - \epsilon\|^2$
Flow Matching
$x_t = (1-\kappa(t))x_0 + \kappa(t)\epsilon$ $\mathcal{L}_{\text{FM}} = \|\psi(x_t \mid c) - v(x_t)\|^2$
Riemannian Flow Matching
$x_t = \exp_{x_0}\left(\kappa(t)\log_{x_0}(\epsilon)\right)$ $\mathcal{L}_{\text{RFM}} = \|\psi(x_t \mid c) - v(x_t)\|^2_{x_t}$

$\kappa(t)$: noise scheduler

Trick: Riemannian flow matching to take into account Earth’s geometry



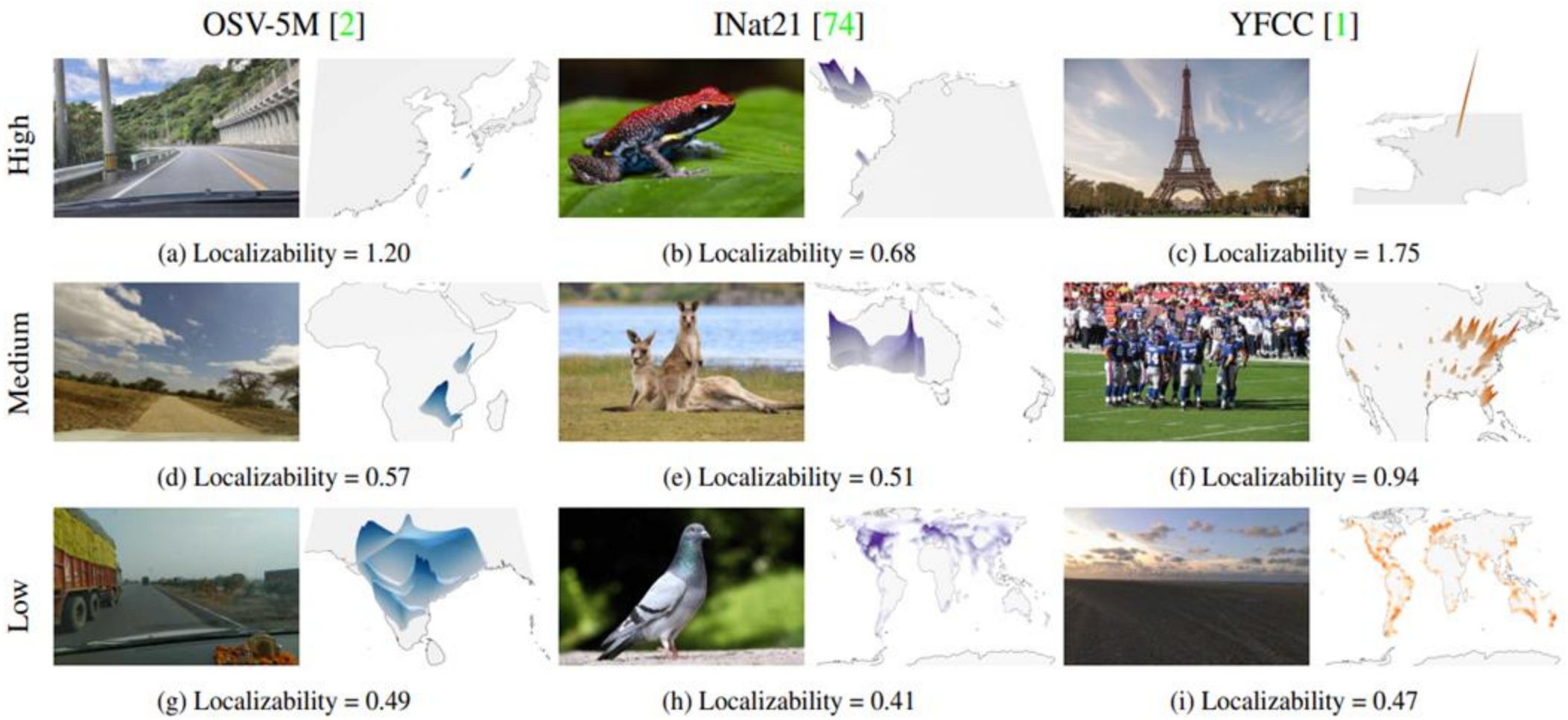
Results: SOTA for geolocation

		OSV-5M [2]					iNat21 [74]
		geos. ↑ /5000	dist ↓ (km)	accuracy ↑ (in %) country	region	city	dist ↓ (km)
deterministic	SC 0-shot [25]	2273	2854	38.4	20.8	14.8	
	Regression [2]	3028	1481	56.5	16.3	0.7	
	ISNs [52]	3331	2308	66.8	39.4	4.2	
	Hybrid [2]	3361	1814	68.0	39.4	5.9	
	SC Retrieval [25]	3597	1386	73.4	**45.8**	**19.9**	
generative	Uniform	131	10052	2.4	0.1	0.0	10,010
	vMF	2776	2439	52.7	17.2	0.6	6270
	vMFMix [36]	1746	5662	34.2	11.1	0.3	4701
	Diff $\mathbb{R}^3$ (ours)	3762	1123	75.9	40.9	3.6	3057
	FM $\mathbb{R}^3$ (ours)	3688	1149	74.9	40.0	4.2	2942
	RFM $\mathcal{S}_2$ (ours)	**3767**	**1069**	**76.2**	**44.2**	5.4	**2500**

		YFCC-4k [1, 76]					
		geos. ↑ /5000	dist ↓ (km)	accuracy ↑ (in %) 25km	200km	750km	2500km
deterministic	PlaNet [77]			14.3	22.2	36.4	55.8
	CPlaNet [66]			14.8	21.9	36.4	55.5
	ISNs [52]			16.5	24.2	37.5	54.9
	Translocator [63]			18.6	27.0	41.1	60.4
	GeoDecoder [11]			24.4	33.9	50.0	68.7
	PIGEON [26]			24.4	40.6	**62.2**	**77.7**
generative	Uniform	131.2	10052	0.0	0.0	0.3	3.8
	vMF	1847	3563	4.8	15.0	30.9	53.4
	vMFMix [36]	1356	4394	0.4	8.8	20.9	41.0
	Diff $\mathbb{R}^3$ (ours)	2845	2461	11.1	37.7	54.7	71.9
	FM $\mathbb{R}^3$ (ours)	2838	2514	22.1	35.0	53.2	73.1
	RFM $\mathcal{S}_2$ (ours)	2889	2461	23.7	36.4	54.5	73.6
	RFM$_{10M}$ $\mathcal{S}_2$ (ours)	**3210**	**2058**	**33.5**	**45.3**	61.1	**77.7**

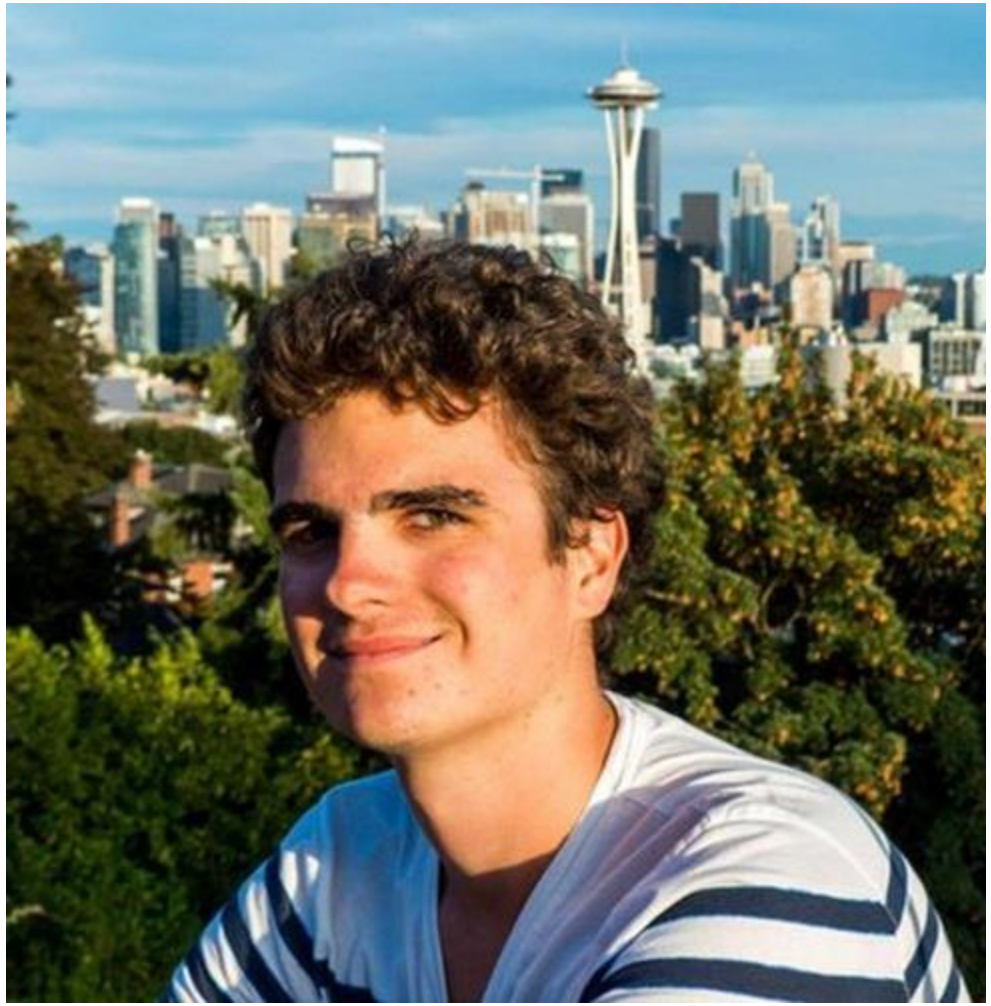
- OSV-5M (street-view)
- iNaturalist21 (animals)
- YFCC100M (uploaded on Flickr)

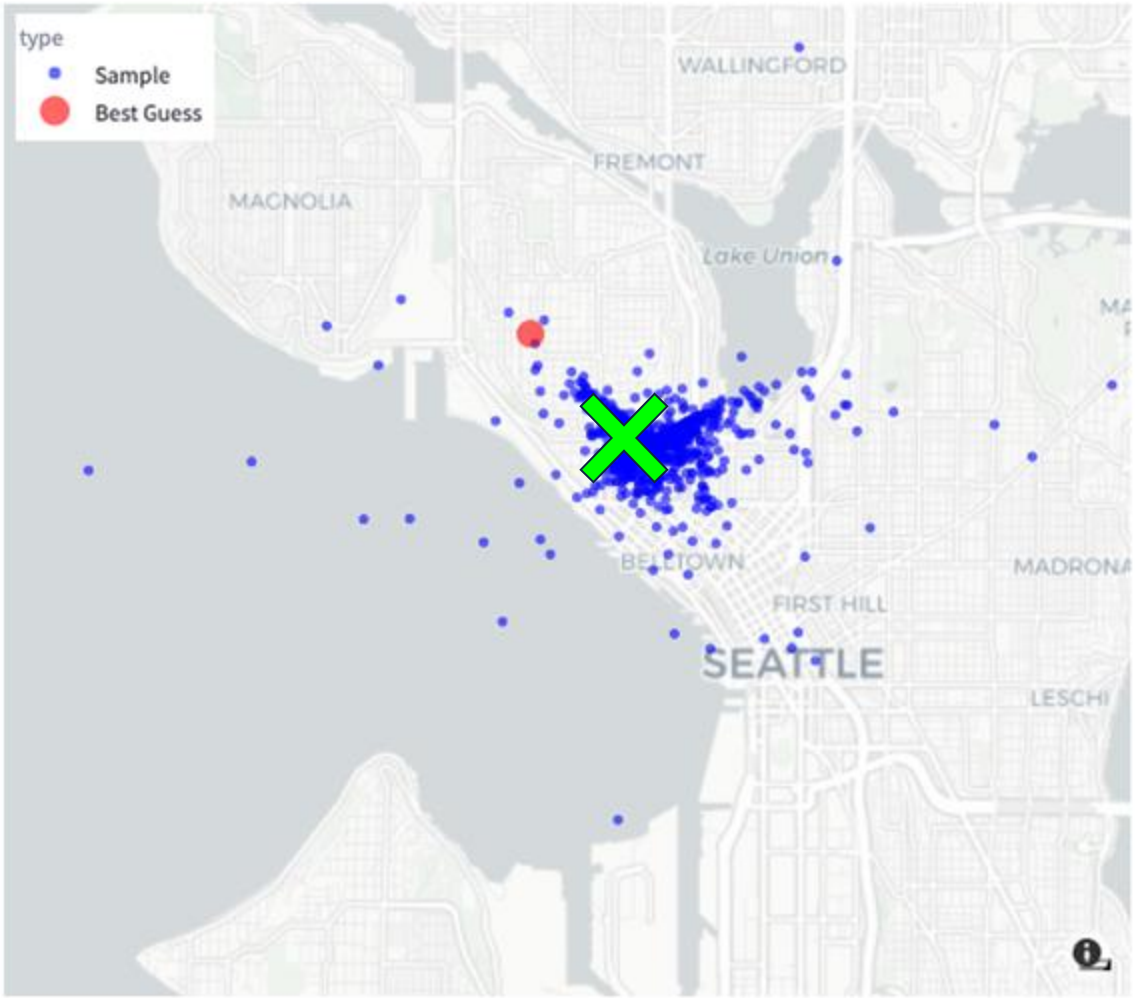
Results: Probabilistic geolocation



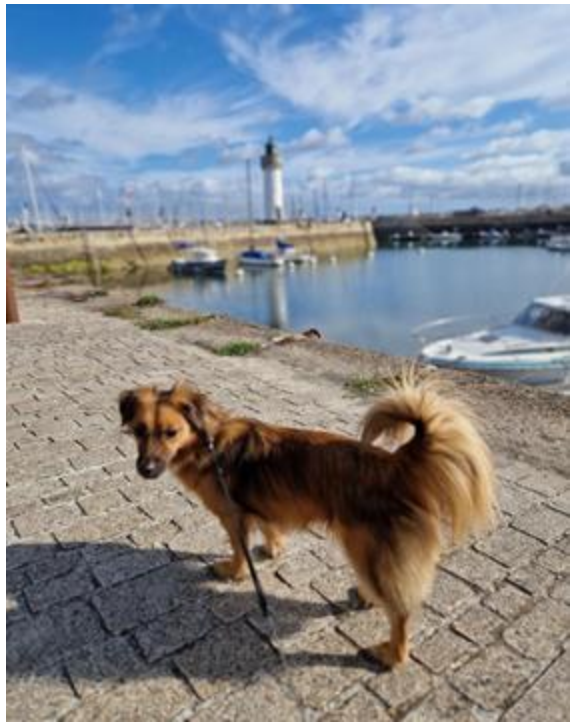
Localizability: correlates with our intuition

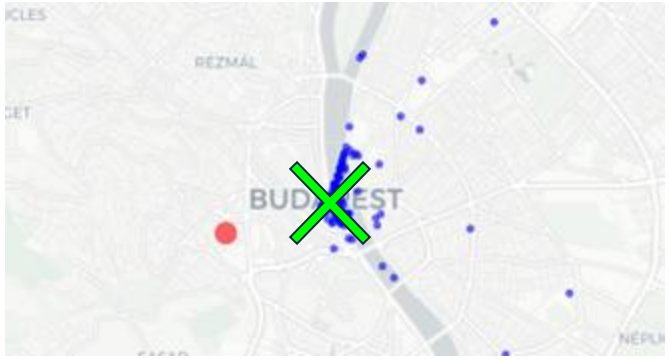
Some fun samples: Localizing profile pictures





Some fun samples: Localizing profile pictures





Can you cheat in Geoguessr?

GT

Plonk

Efficiency: Challenges

Training data
- Collecting, filtering
- Privacy

Model
- Large model size
- Scaling resolution

Training & conditioning
- How to condition?
- Long training times

Inference & post-training
- Multiple denoising steps
- Apply RL?

+ Bonus!

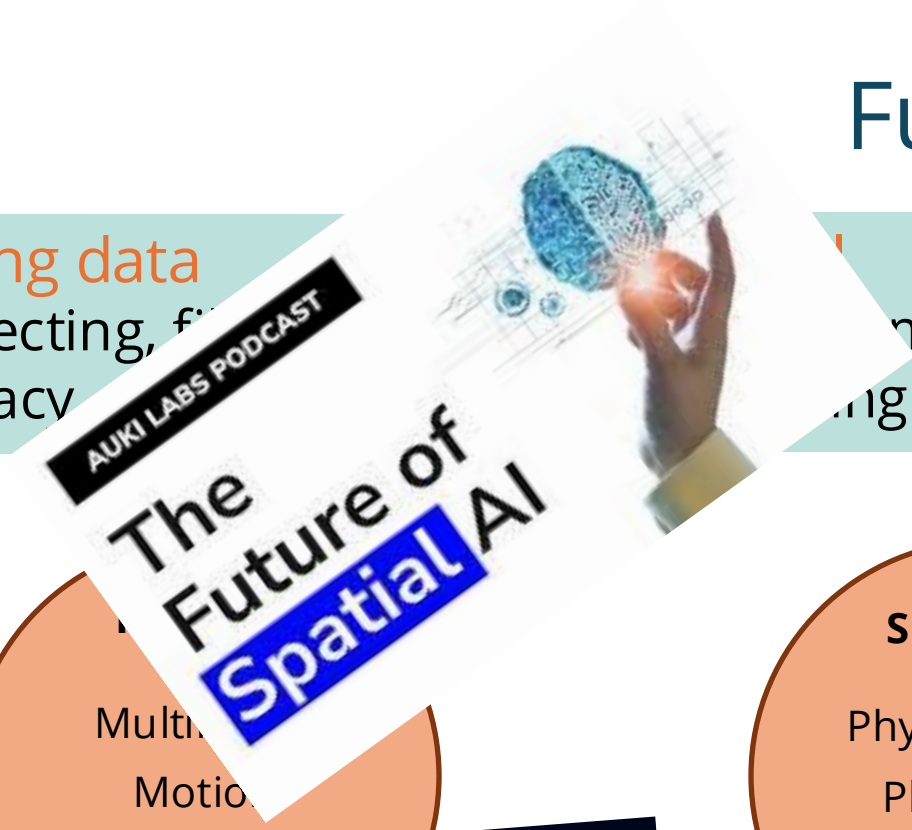
Futu

Training data

- Collecting, fi
- Privacy

nditioning

- model s
- ng resolution



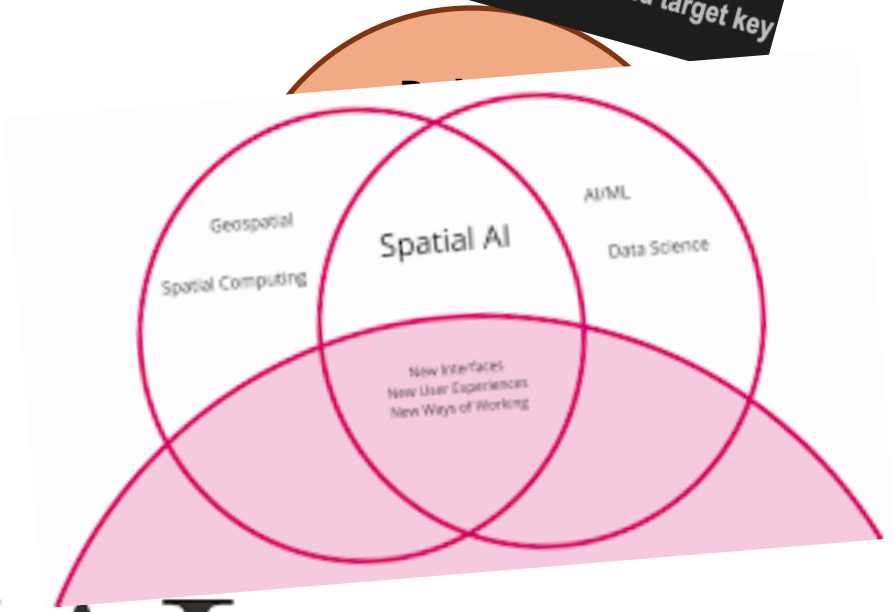
Multi
Motio

Spatial AI

Physics-aware

Physical AI

Cybernetics



SPATIAL

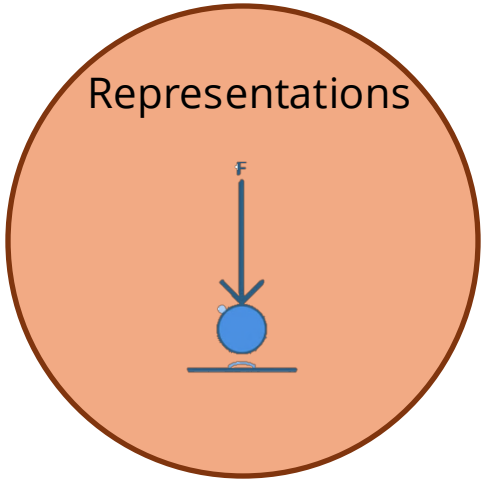
Screenshots taken 9 Sep 2025

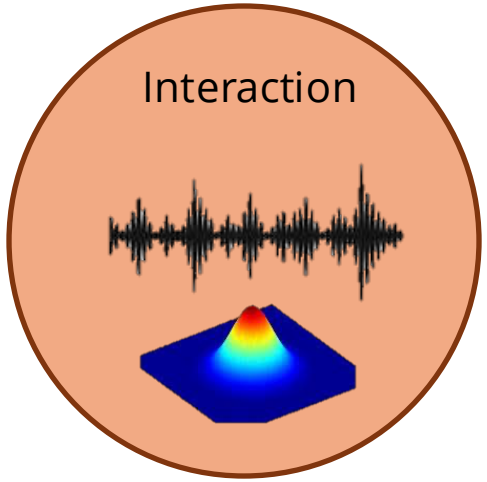
Future

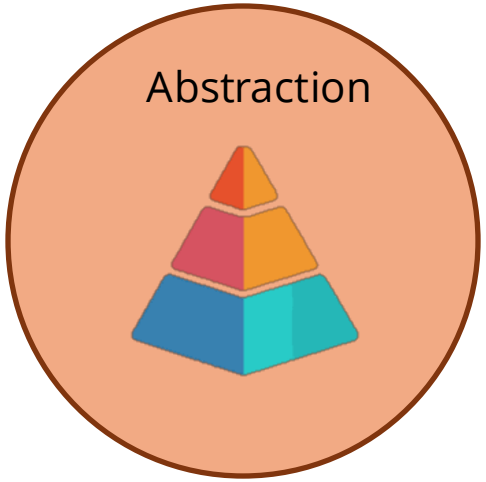
Training data
- Collecting, filtering
- Privacy

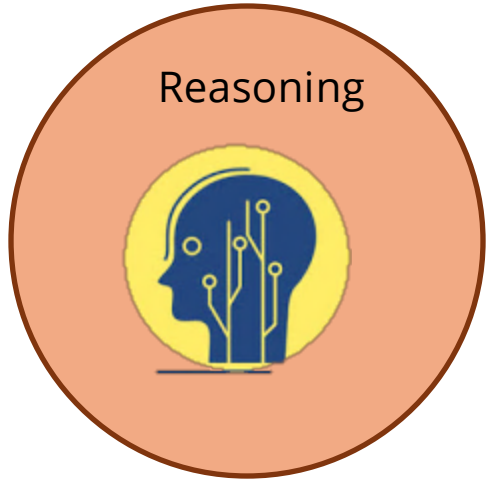
Model
- Large model size
- Scaling resolution

Training & conditioning
- How to condition?
- Long training times









Efficient human-robot collaboration

Logos generated with Claude, ChatGPT 2024, 2025

From perception to imagination to collaboration

e.ffi.ci.en.tly

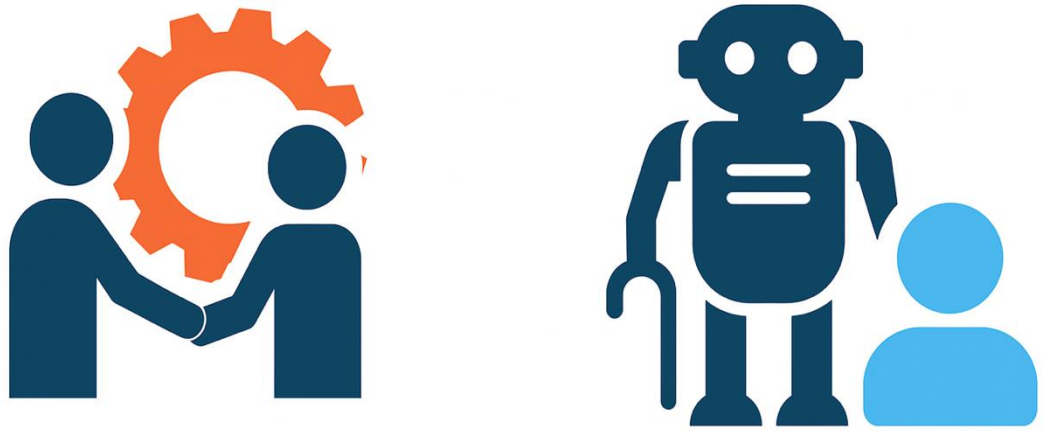
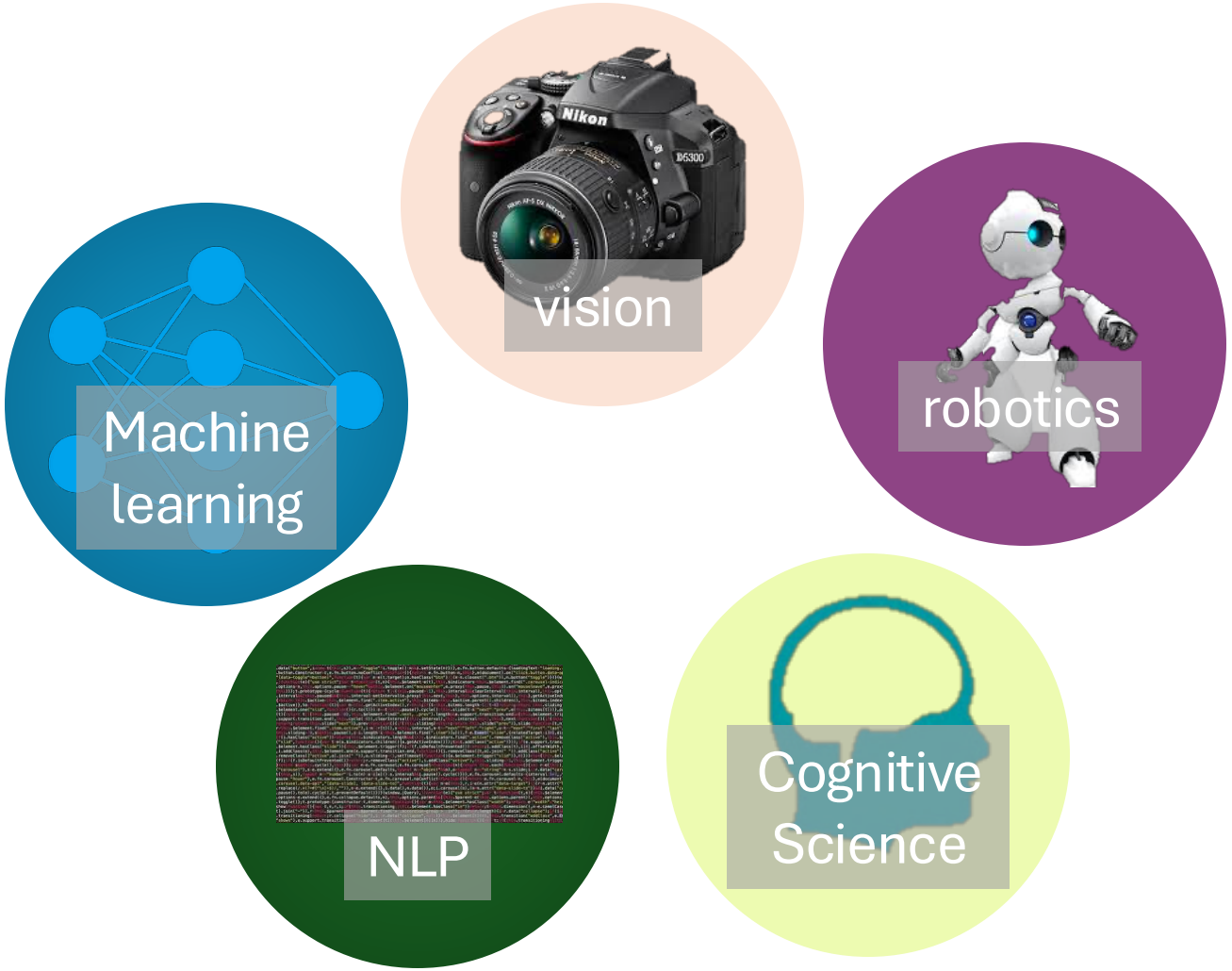


Image generated with ChatGPT 2025

Future





🤘👏

Děkuju!