Listening to the Data:
Visual Learning from the Bottom Up

Yutong Bai
UC Berkeley

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Self Introduction

- I am currently a Postdoc Researcher at UC Berkeley, advised by Alyosha Efros, Jitendra Malik and Trevor Darrell. I obtained PhD degree at Johns Hopkins University advised by Alan Yuille.
- Research is representation learning, self-supervised learning, and generative modeling.

Sequential Modeling Enables Scalable Learning for Large Vision Models







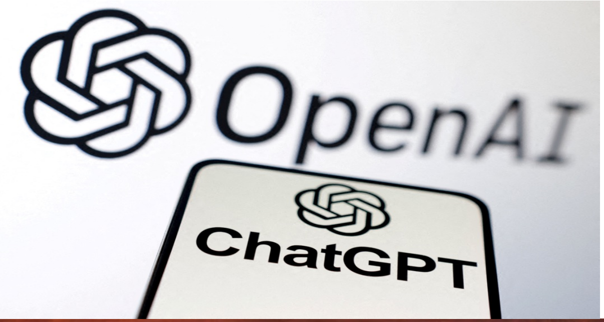

Yutong Bai*, Xinyang Geng*, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, Alexei A Efros

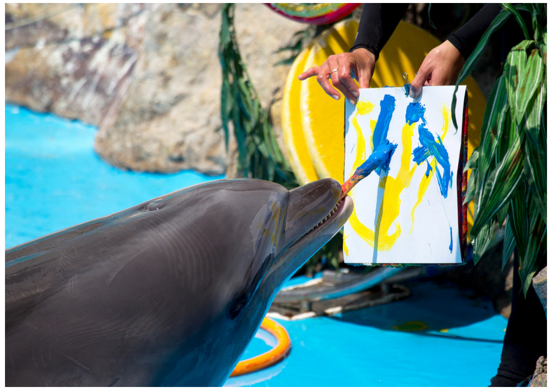
LVM: Why LLM without Language?

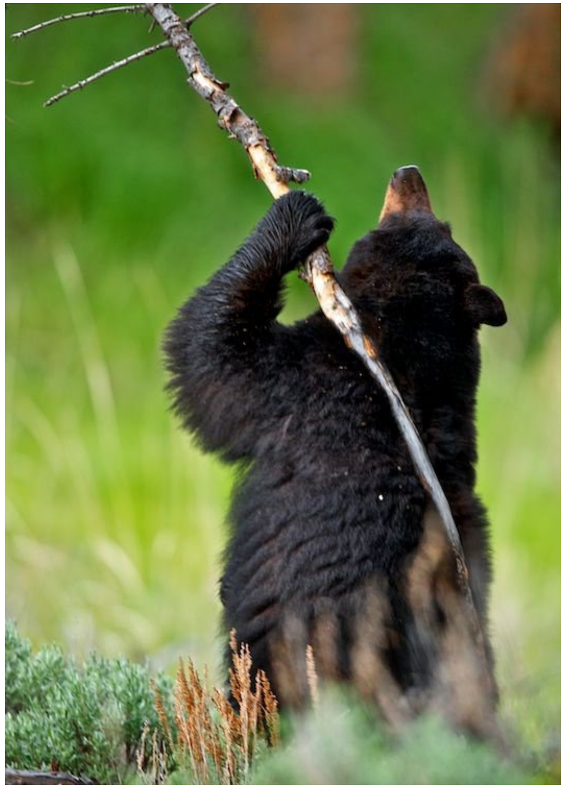
LVM: Why LLM without Language?

- Philosophical
- Practical

LLMs ->Intelligence?







LLMs ->Intelligence?

Scientific Question:
How far can we go from pixels **alone**?

LVM: Why LLM without Language?

- Philosophical
- Practical

Self-Supervised Learning

- **AKA: How to ‘torture’ both the model* and yourself***

Self-Supervised Learning

- **AKA: How to 'torture' both the model and yourself**

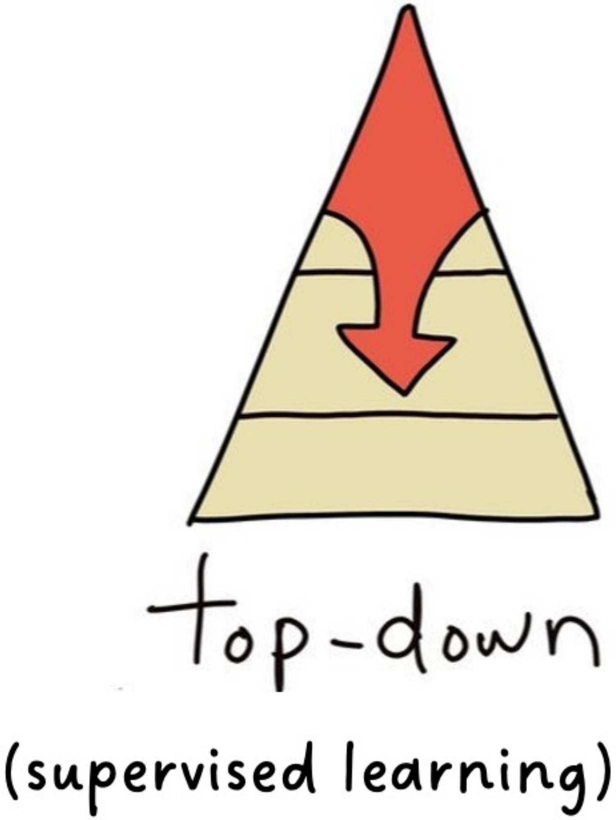
People who listen to my talk. (I wish)

Self-Supervised Learning

- **AKA: How to ‘torture’ both the model and yourself**

Language,
Semantics,
Concepts

Pixels
(raw sensory data



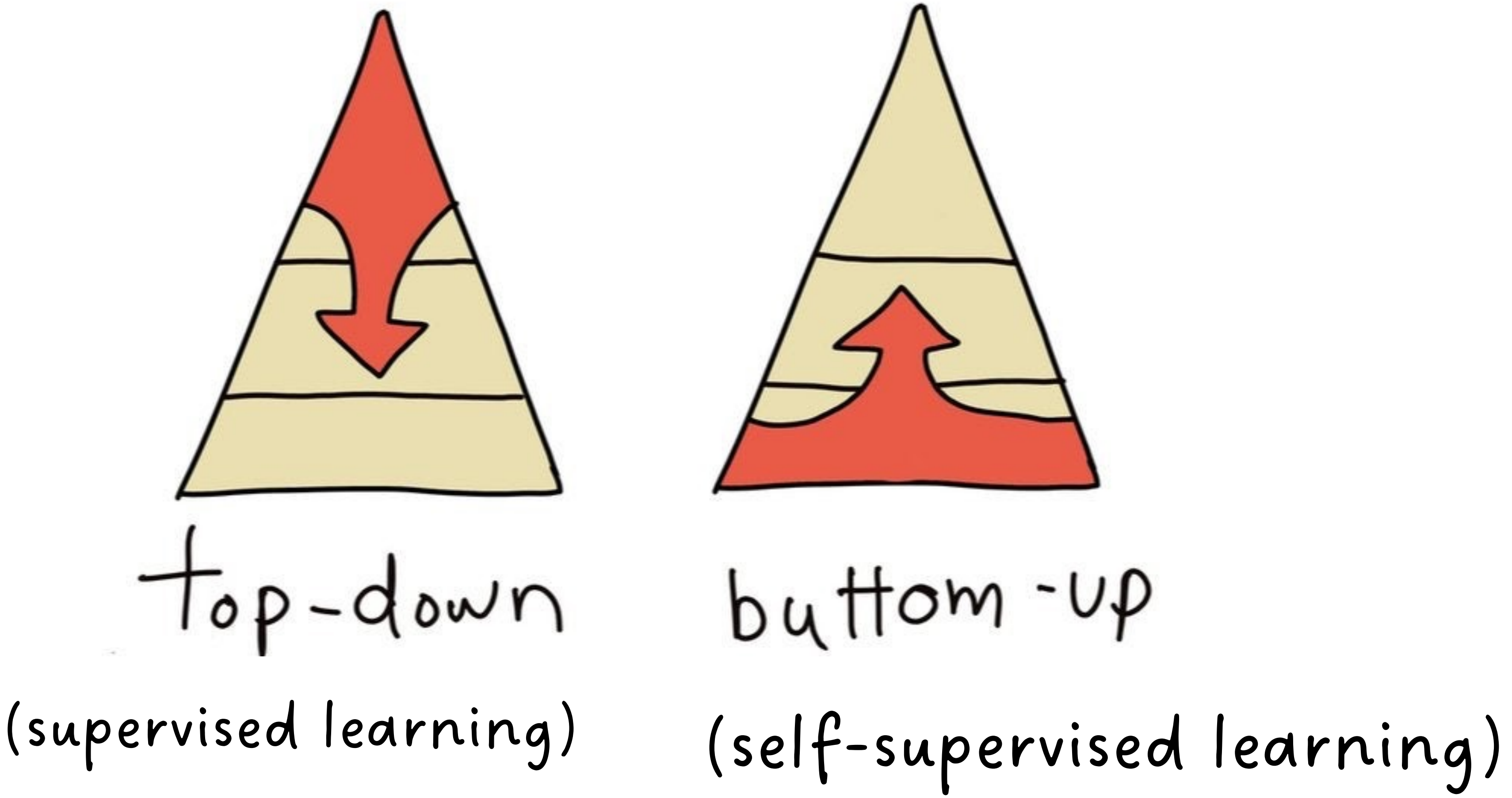
People who listen to my talk. (I wish)

Self-Supervised Learning

- **AKA: How to ‘torture’ both the model and yourself**

Language,
Semantics,
Concepts

Pixels
(raw sensory data

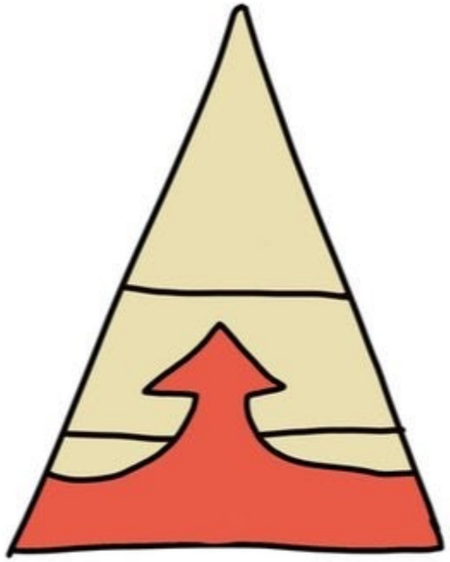


Self-Supervised Learning

- **AKA: How to 'torture' both the model and yourself**

A Difficult task!

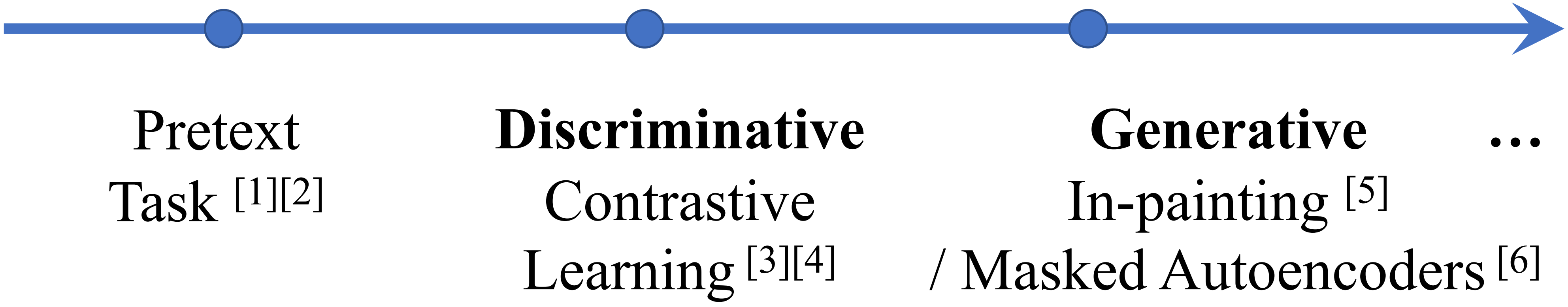
- **Non-trivial.**
- **Absorb in large amount of data.**



buttom-up

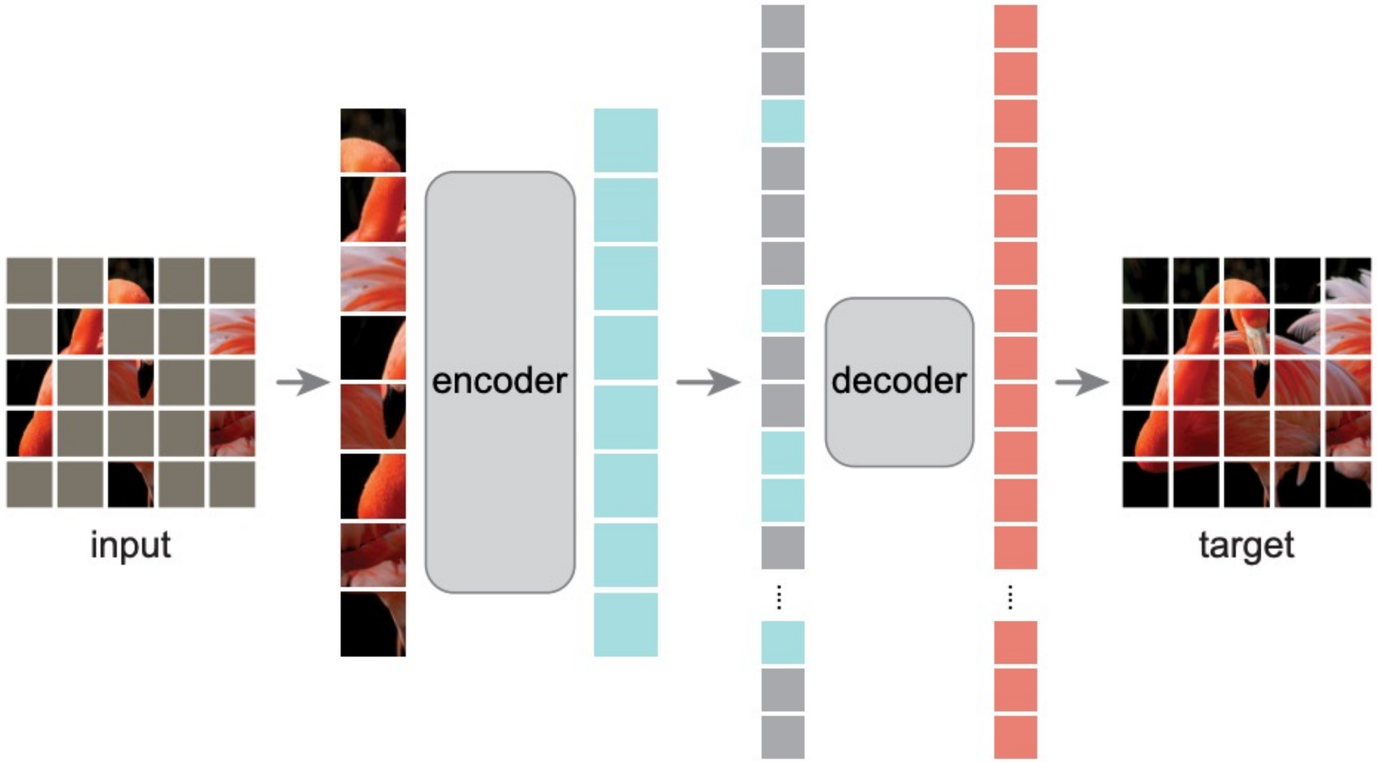
(self-supervised learning)

Self-supervised Learning



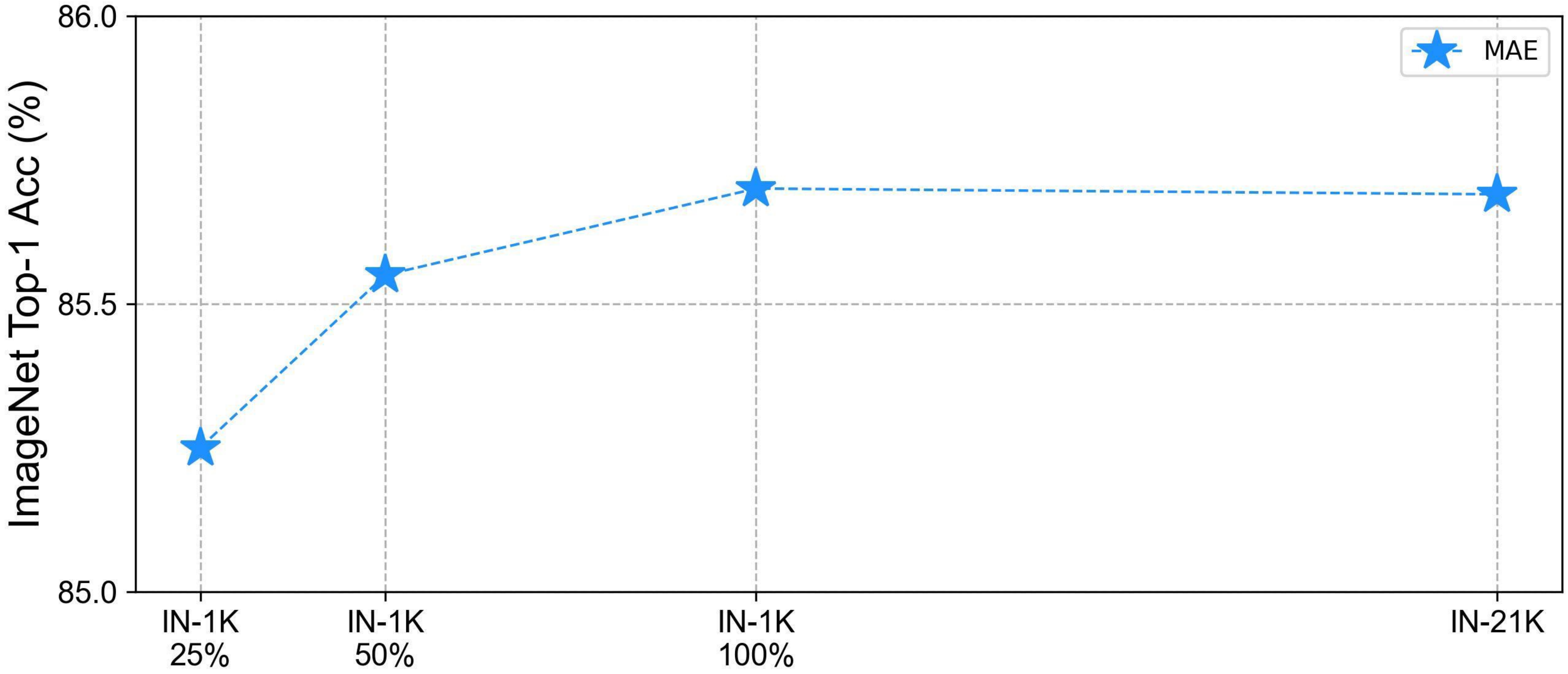
[1] Zhang, Isola, and Efros. "Colorful image colorization." ECCV 2016.
[2] Doersch, Gupta, and Efros. "Unsupervised visual representation learning by context prediction." ICCV 2015.
[3] Wu, Xiong, Yu and Lin. "Unsupervised feature learning via non-parametric instance discrimination. " CVPR 2018.
[4] He, Fan, Wu, Xie and Girshick. "Momentum contrast for unsupervised visual representation learning. " CVPR 2020.
[5] Pathak, Krahenbuhl, Donahue, Darrell and Efros. "Context encoders: Feature learning by inpainting. " CVPR 2016.
[6] He, Chen, Xie, Li, Dollár and Girshick. "Masked autoencoders are scalable vision learners." CVPR 2022.

Masked Autoencoder (MAE) for Transformer



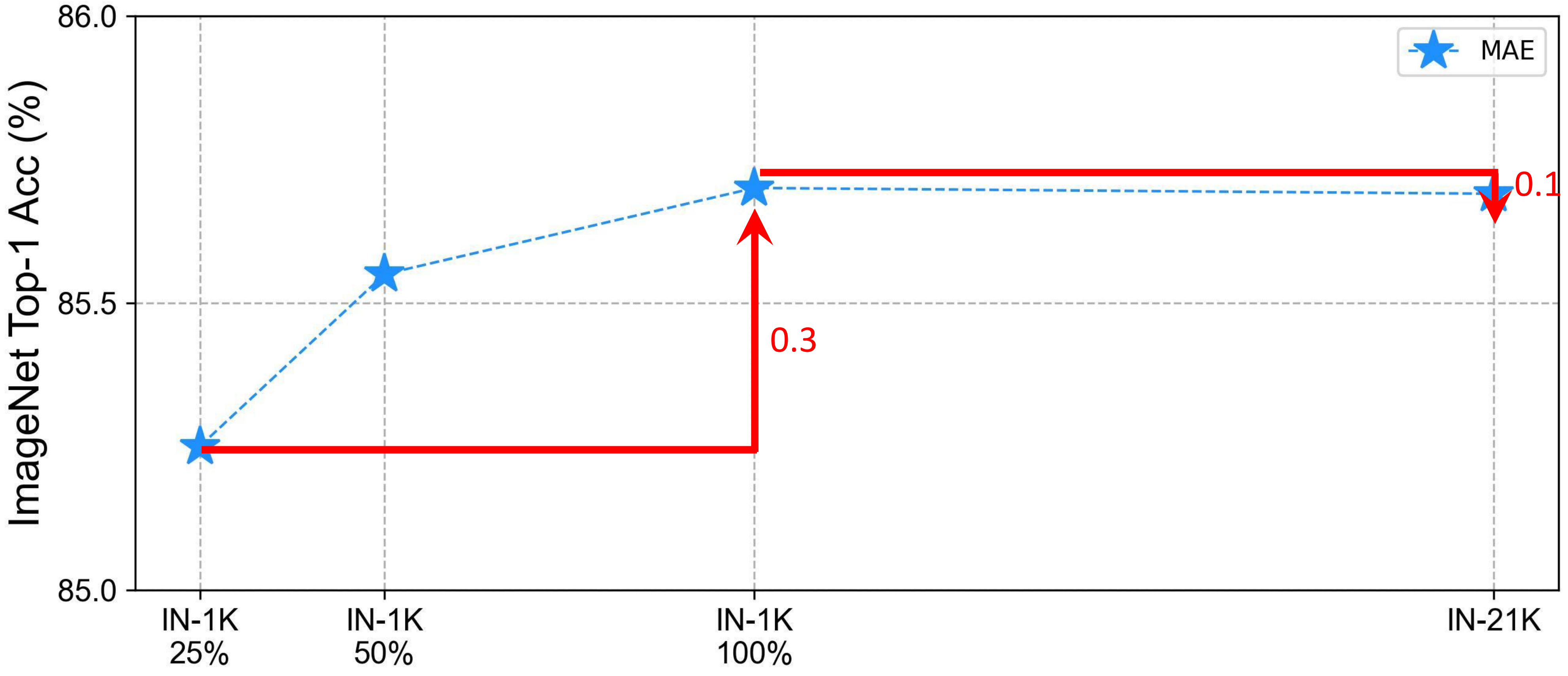
Pathak, Krahenbuhl, Donahue, Darrell and Efros. "Context encoders: Feature learning by inpainting. " CVPR 2016.
He, Chen, Xie, Li, Dollár and Girshick. "Masked autoencoders are scalable vision learners." CVPR 2022.

Scaling Behaviors of MAE on Data



Bai et al. Scalable Visual Pretraining Needs Better Targets

Scaling Behaviors of MAE on **Data**



Bai et al. Scalable Visual Pretraining Needs Better Targets

Rethink the Paradigm of MAE

Data: ImageNet , 1600 ep.

Architecture: Masked Autoencoders

Loss function: L2 regression loss

Task Specification: Finetune

Rethink the Paradigm of MAE

Data: ~~ImageNet , 1600 ep.~~ 1.68B of images, 420B tokens, 50 Datasets, 1 ep, no aug, deterministic training.

Architecture: Masked Autoencoders

Loss function: L2 regression loss

Task Specification: Finetune

Rethink the Paradigm of MAE

Data: ~~ImageNet , 1600 ep.~~ 1.68B of images, 420B tokens, 50 Datasets, 1 ep, no aug, deterministic training.

Architecture: ~~Masked Autoencoders~~ Autoregressive Model

Loss function: L2 regression loss

Task Specification: Finetune

Rethink the Paradigm of MAE

Data: ~~ImageNet , 1600 ep.~~ 1.68B of images, 420B tokens, 50 Datasets, 1 ep, no aug, deterministic training.

Architecture: ~~Masked Autoencoders~~ Autoregressive Model

Loss function: ~~L2 regression loss~~

Task Specification: Finetune

Rethink the Paradigm of MAE

Data: ~~ImageNet , 1600 ep.~~ 1.68B of images, 420B tokens, 50 Datasets, 1 ep, no aug, deterministic training.

Architecture: ~~Masked Autoencoders~~ Autoregressive Model

Loss function: ~~L2 regression loss~~ Cross Entropy for next token

Task Specification: Finetune

Rethink the Paradigm of MAE

Data: ~~ImageNet , 1600 ep.~~ 1.68B of images, 420B tokens, 50 Datasets, 1 ep, no aug, deterministic training.

Architecture: ~~Masked Autoencoders~~ Autoregressive Model

Loss function: ~~L2 regression loss~~ Cross Entropy for next token

Task Specification: ~~Finetune~~ prompting

across:
images,
videos,
supervised / unsupervised
synthetic /real,
all kinds of tasks
2D / 3D / 4D data etc.

Dataset	Tokens (Millions)	Annotation Type	Annotation Source
Unpaired Image Data			
LAION 5B [71] (1.5B images subset)	380690	-	-
Images with Annotations			
ImageNet 1K [25]	1317.40	Image Classification	Ground Truth
COCO [54]	363	Object Detection	MMDetection [16]
ADE 20K [100], Cityscapes [22]	66.88	Semantic Segmentation	Ground Truth
COCO [54], ImageNet 1K [25]	2078.06	Semantic Segmentation	Mask2Former [19]
COCO [54], lvmhp [51], mpii [4], Unite [49]	950.79	Human Pose	MMPose[21]
COCO [54], ImageNet 1K [25]	1623.85	Depth Map Image	DPT [67]
Subset of InstructPix2Pix [34]	415.46	Style Transfer	InstructPix2Pix [34]
COCO[54], ImageNet 1K[25]	1623.85	Surface Normal Image	NLL-AngMF [7]
COCO [54], ImageNet 1K [25]	1623.85	Edge Detection	DexiNed [79]
DID-MDN [98]	35.06	Rainy and Clean Image Pairs	Ground Truth
SIDD [3]	245.76	Denoised Image	Ground Truth
LOL[89]	0.458	Light Enhanced Image	Ground Truth
ImageNet 1K [25]	1321.07	Grayscale and Colorized Image Pairs	Ground Truth
ImageNet 1K [25]	1321.07	Inpainting	Ground Truth
Kitti [34]	9.21	Stereo	Ground Truth
Videos			
UCF101 [78]	109.11	-	-
DAVIS [65]	0.36	-	-
HMDB [48]	55.41	-	-
ActivityNet [13]	380.63	-	-
Moments in Time [59]	2979.00	-	-
Multi-moments in Time [60]	4124.04	-	-
Co3D [69]	228.75	-	-
Charades v1 [76]	241.53	-	-
Something-something v2 [37]	904.57	-	-
YouCook [23]	3.14	-	-
Kinetics 700 [14]	7092.04	-	-
MSR-VTT [92]	57.34	-	-
Youtube VOS [93]	63.70	-	-
jester [57]	606.47	-	-
diving48 [52]	150.73	-	-
MultiSports [53]	78.44	-	-
CharadesEgo [77]	193.06	-	-
AVA [61]	117.96	-	-
Ego4D [38]	1152.12	-	-
Videos with Annotations			
VIPSeg [58]	64.47	Video Panoptic Segmentation	Ground Truth
Hand14K [32]	1.96	Hand Segmentation	Ground Truth
AVA [61]	122.88	Video Detection	Ground Truth
JHMDB [43]	19.00	Optical Flow	Ground Truth
JHMDB [43]	37.92	Video Human Pose	Ground Truth
Synthetic 3D Views			
Objaverse [24] Rendered Multiviews	217.85	-	-

across:

images,

videos,

supervised / unsupervised

synthetic /real,

all kinds of tasks

2D / 3D / 4D data etc.

Dataset	Tokens (Millions)	Annotation Type	Annotation Source
Unpaired Image Data			
LAION 5B [71] (1.5B images subset)	380690	-	-
Images with Annotations			
ImageNet 1K [25]	1317.40	Image Classification	Ground Truth
COCO [54]	363	Object Detection	MMDetection [16]
ADE 20K [100], Cityscapes [22]	66.88	Semantic Segmentation	Ground Truth
COCO [54], ImageNet 1K [25]	2078.06	Semantic Segmentation	Mask2Former [19]
COCO [54], lvmhp [51], mpii [4], Unite [49]	950.79	Human Pose	MMPose[21]
COCO [54], ImageNet 1K [25]	1623.85	Depth Map Image	DPT [67]
Subset of InstructPix2Pix [34]	415.46	Style Transfer	InstructPix2Pix [34]
COCO[54], ImageNet 1K[25]	1623.85	Surface Normal Image	NLL-AngMF [7]
COCO [54], ImageNet 1K [25]	1623.85	Edge Detection	DexiNed [79]
DID-MDN [98]	35.06	Rainy and Clean Image Pairs	Ground Truth
SIDD [3]	245.76	Denoised Image	Ground Truth
LOL[89]	0.458	Light Enhanced Image	Ground Truth
ImageNet 1K [25]	1321.07	Grayscale and Colorized Image Pairs	Ground Truth
ImageNet 1K [25]	1321.07	[illegible]ting	Ground Truth
Kitti [34]	9.21	[illegible]	Ground Truth
Videos			
UCF101 [78]	109.11	-	-
DAVIS [65]	0.36	-	-
HMDB [48]	55.41	-	-
ActivityNet [13]	380.63	-	-
Moments in Time [59]	2979.00	-	-
Multi-moments in Time [60]	4124.04	-	-
Co3D [69]	228.75	-	-
Charades v1 [76]	241.53	-	-
Something-something v2 [37]	904.57	-	-
YouCook [23]	3.14	-	-
Kinetics 700 [14]	7092.04	-	-
MSR-VTT [92]	57.34	-	-
Youtube VOS [93]	63.70	-	-
jester [57]	606.47	-	-
diving48 [52]	150.73	-	-
MultiSports [53]	78.44	-	-
CharadesEgo [77]	193.06	-	-
AVA [61]	117.96	-	-
Ego4D [38]	1152.12	-	-
Videos with Annotations			
VIPSeg [58]	64.47	Video Panoptic Segmentation	Ground Truth
Hand14K [32]	1.96	Hand Segmentation	Ground Truth
AVA [61]	122.88	Video Detection	Ground Truth
JHMDB [43]	19.00	Optical Flow	Ground Truth
JHMDB [43]	37.92	Video Human Pose	Ground Truth
Synthetic 3D Views			
Objaverse [24] Rendered Multiviews	217.85	-	-

Sentence -> Visual Sentence

Single images

<BOS>  <EOS>

Tokenizer

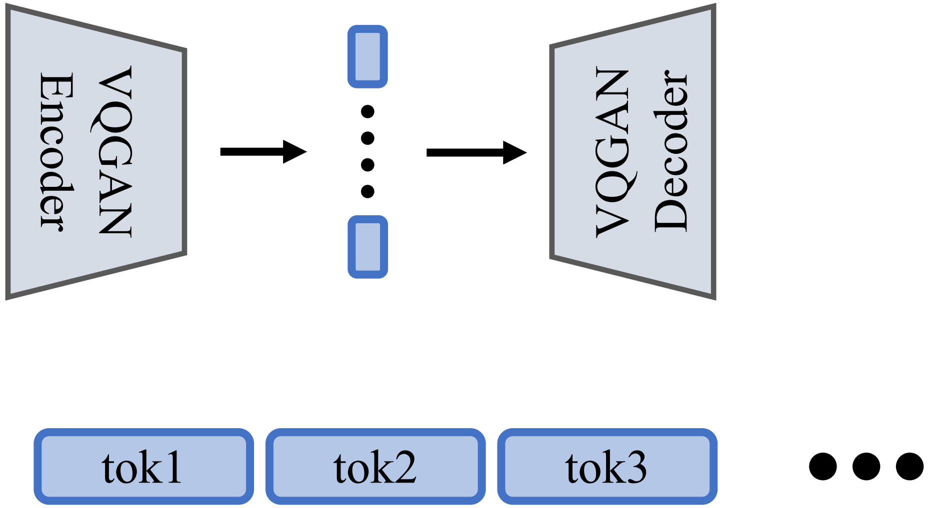


Image sequences

<BOS>

••• <EOS>

Image sequences

<BOS>  ••• <EOS>

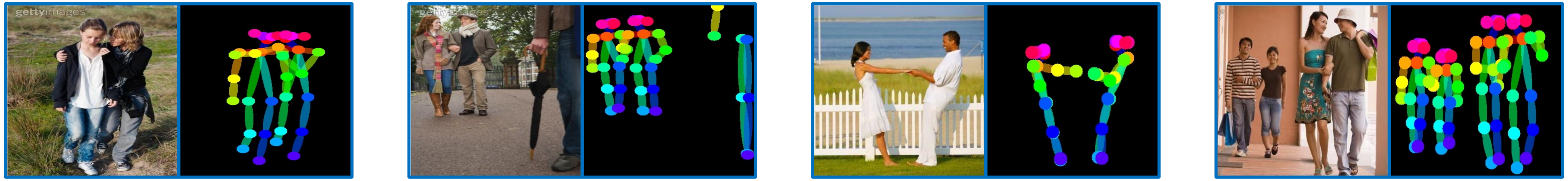
Image sequences

<BOS>

••• <EOS>

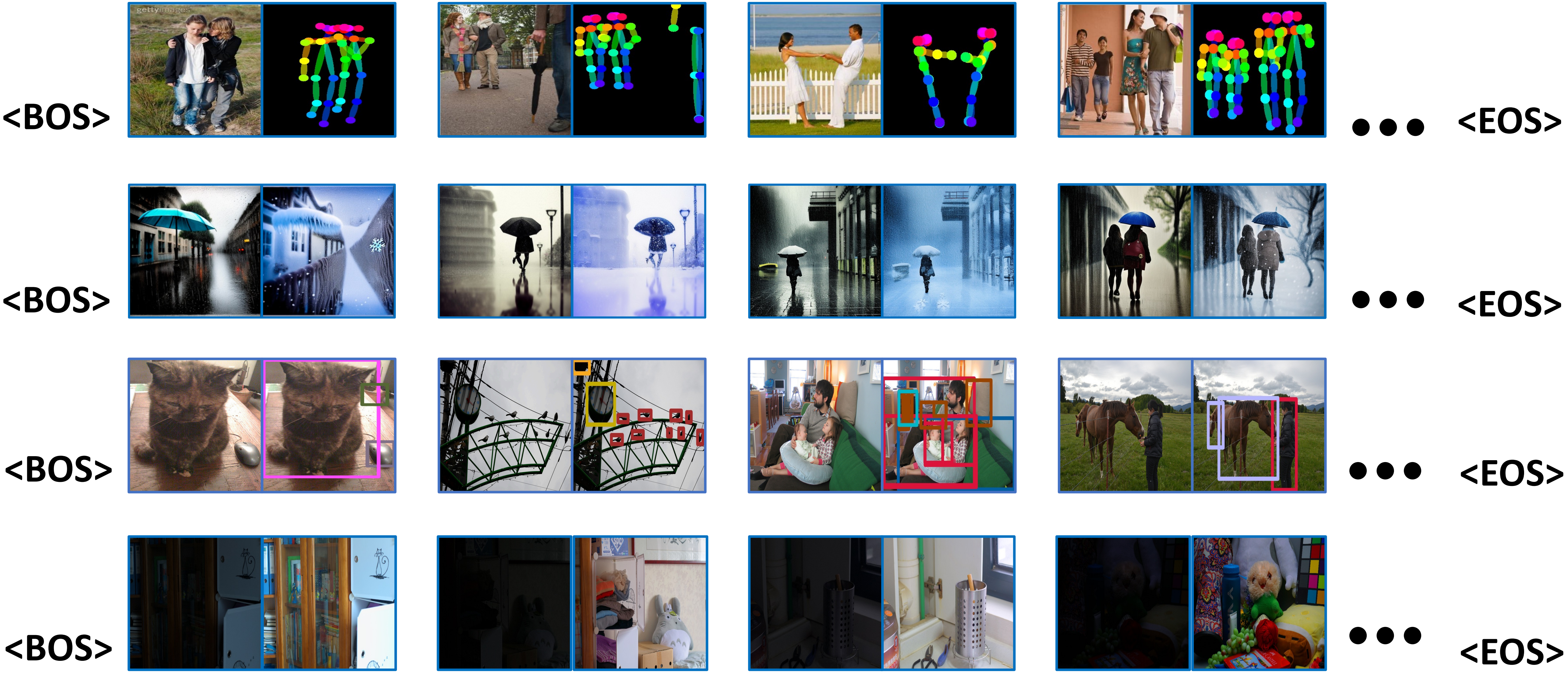
Images with annotation

<BOS>

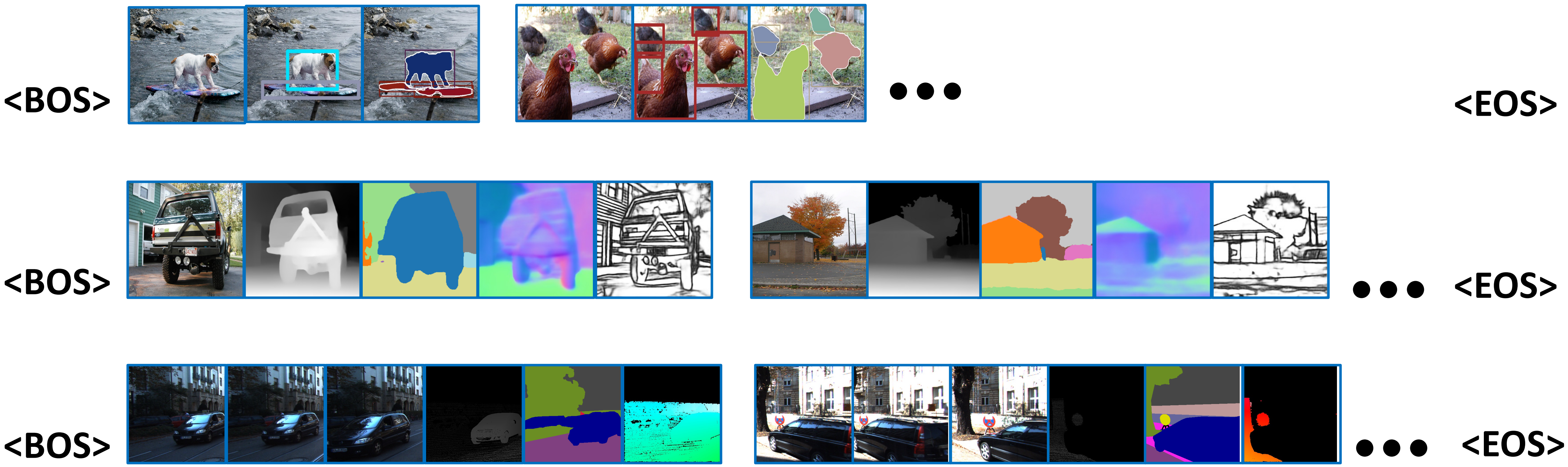


••• <EOS>

Images with annotation

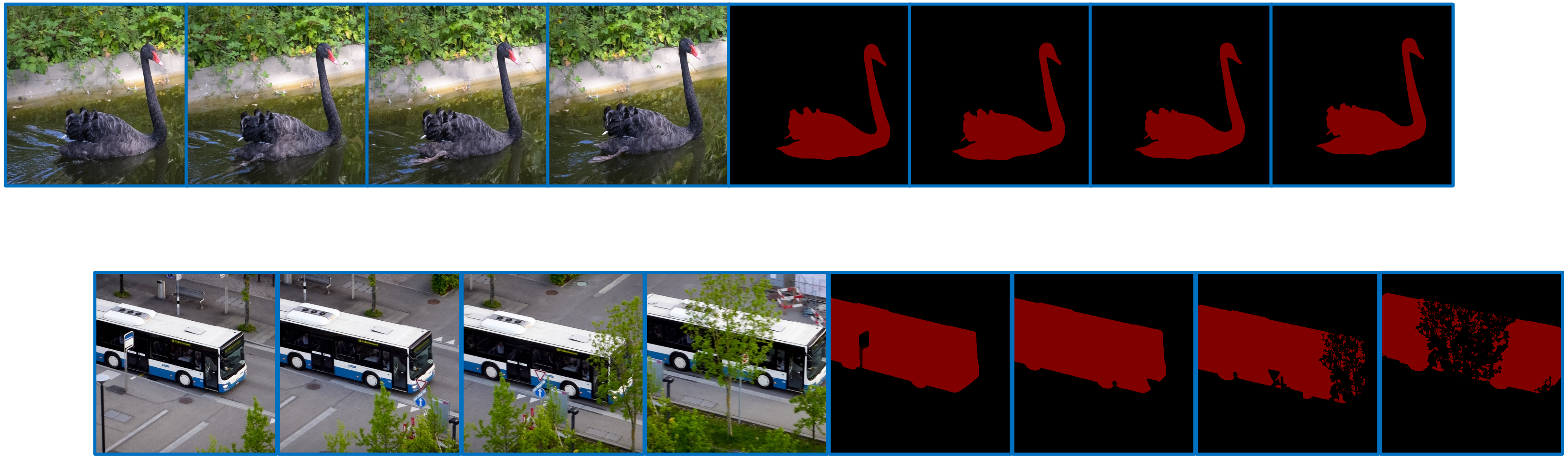


Images with free form annotation



Videos with annotation

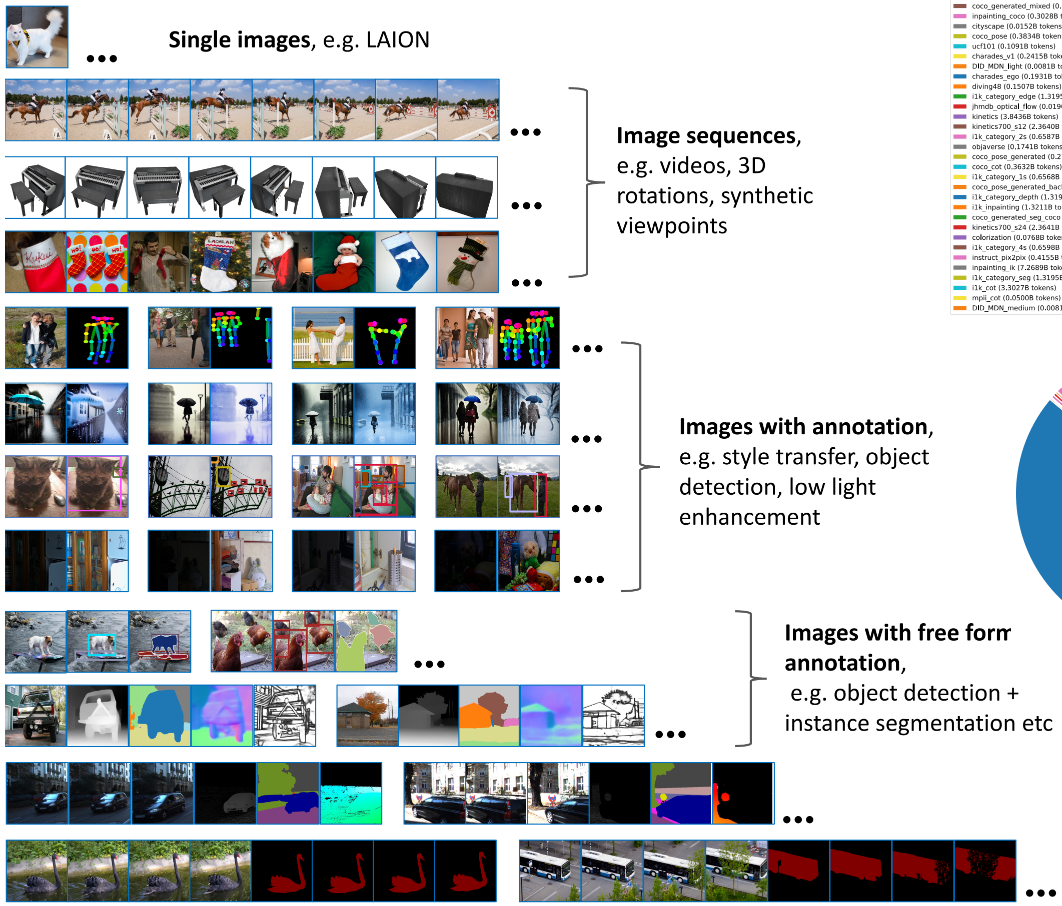
<BOS>

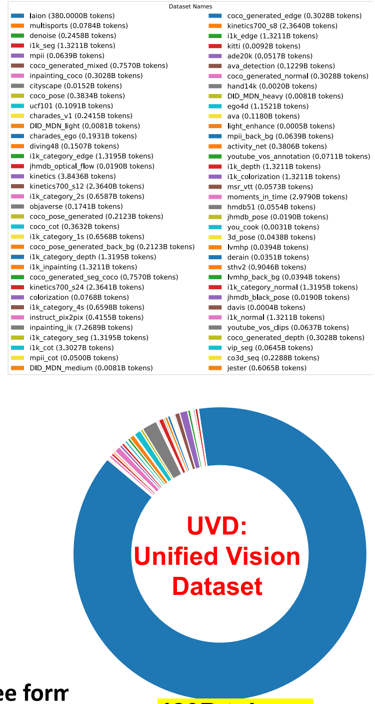


<EOS>

“Data! Data! Data! I can’t make bricks without clay!”
-- SHERLOCK HOLMES

Visual Sentences

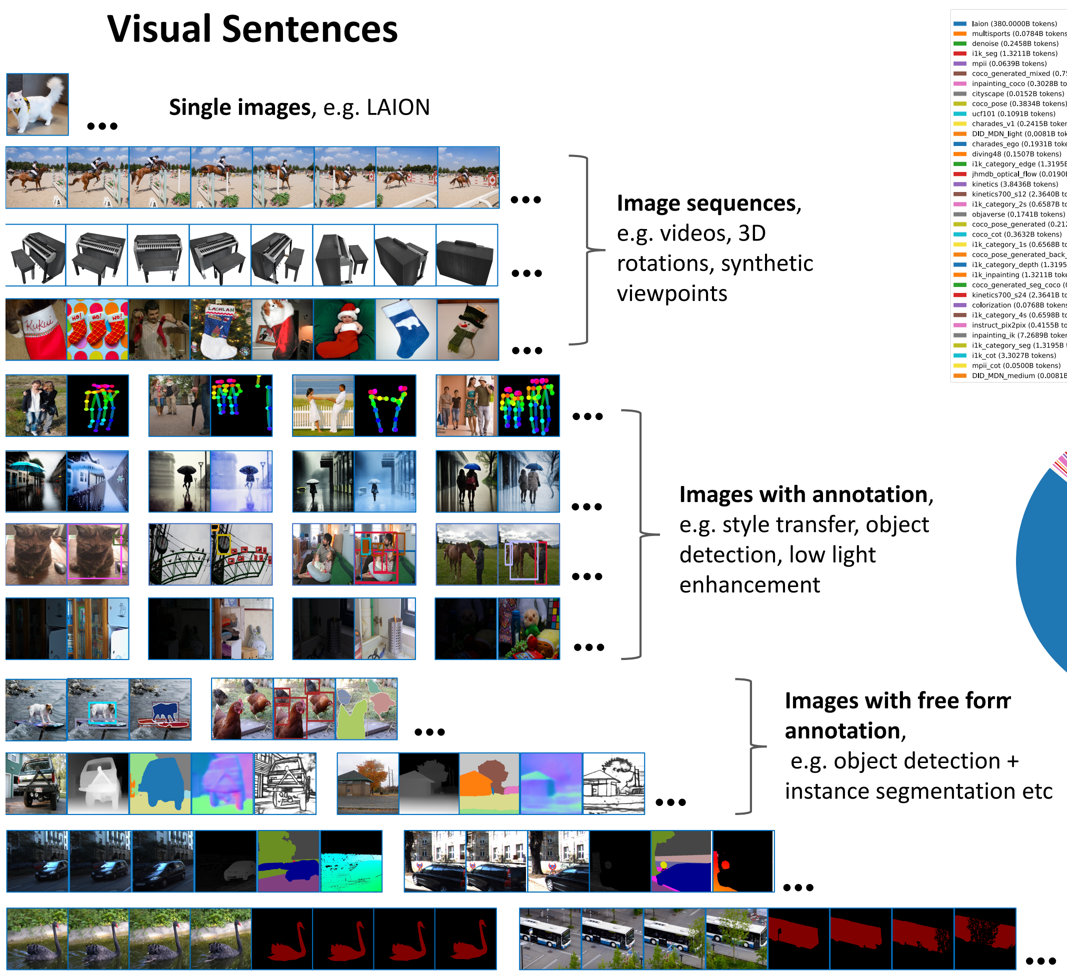


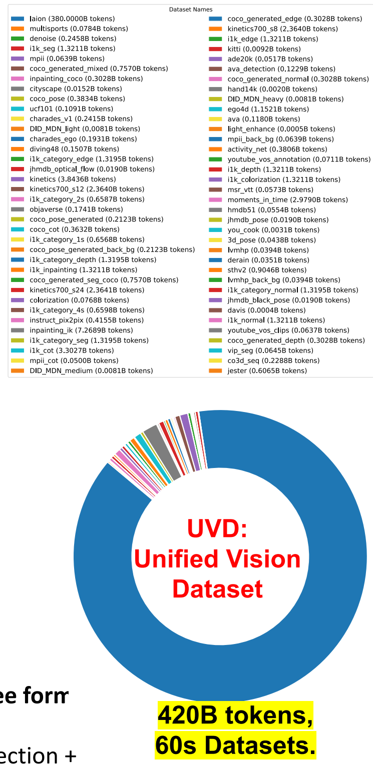


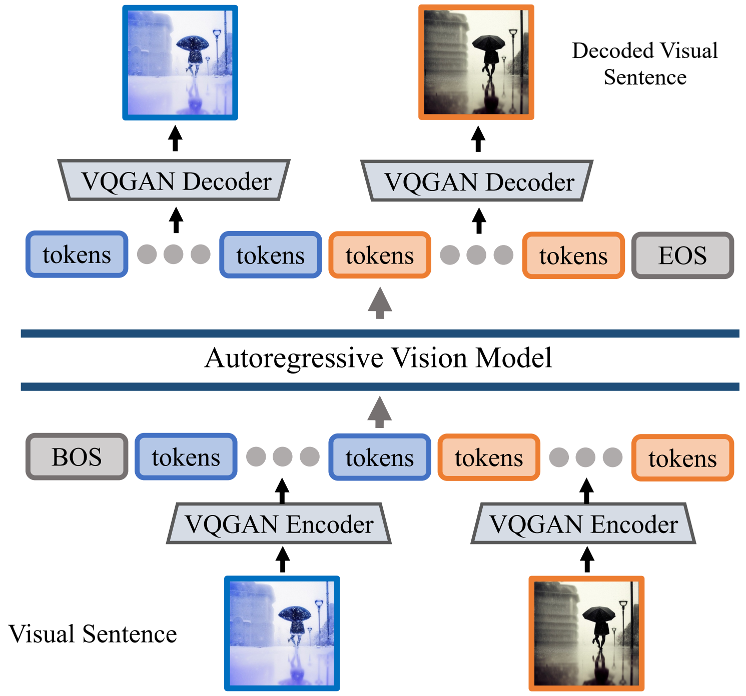
420B tokens,
60s Datasets.

- **Information**
- **Diversity**

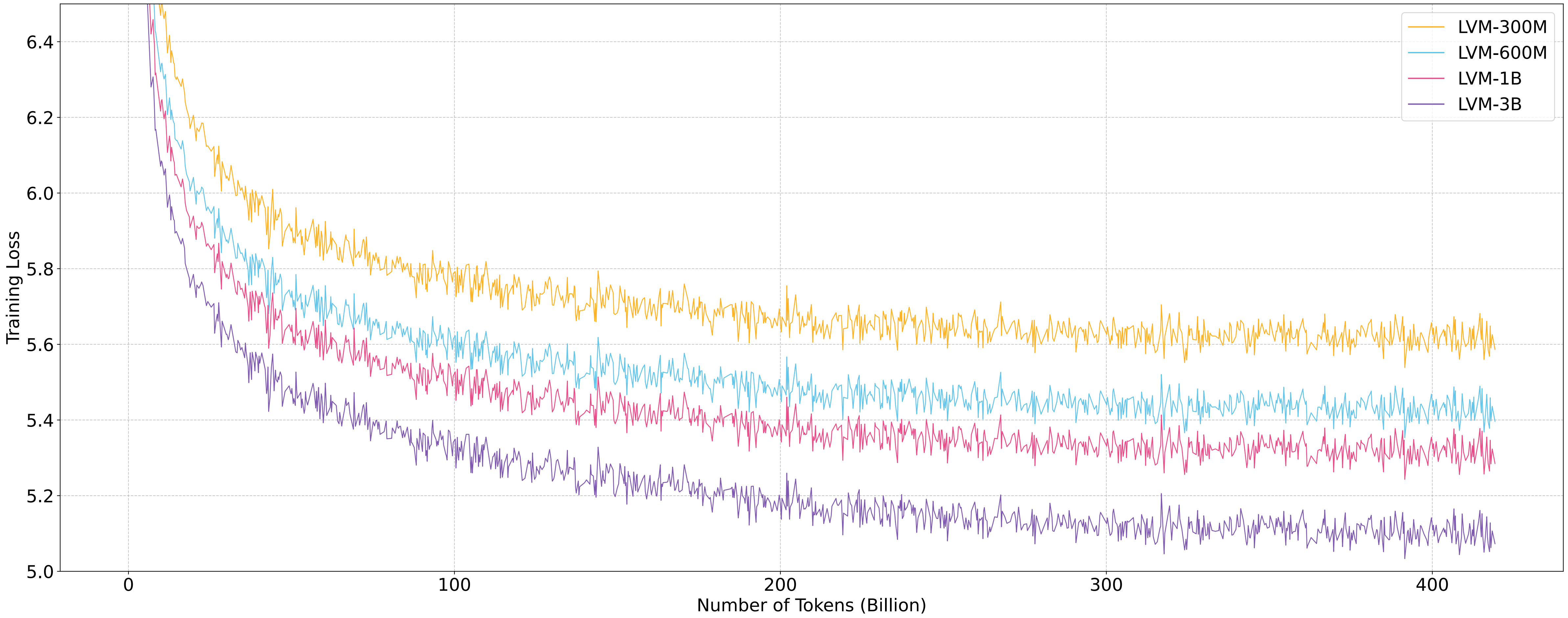
LVM: Large Vision Model







Training Loss (1 epoch) ~ Validation Loss



Larger Model, More Data, Better Downstreams.

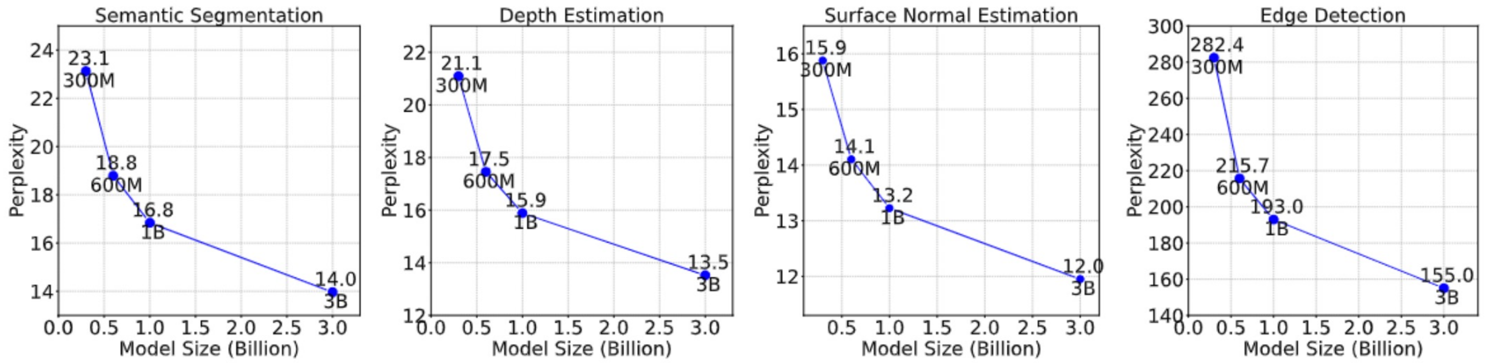


Figure 4. **Larger LVMs perform better on downstream tasks.** We evaluate LVM models of varying sizes on 4 different downstream tasks, following the 5 shot setting on the ImageNet validation set and report the perplexity. We find that perplexity decreases with larger models across all tasks, indicating the strong scalability of LVM.

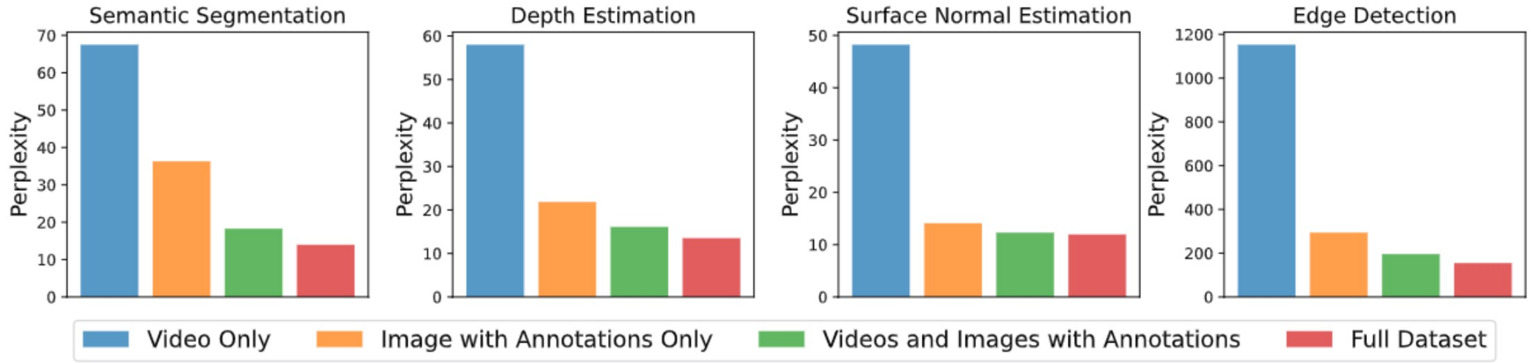
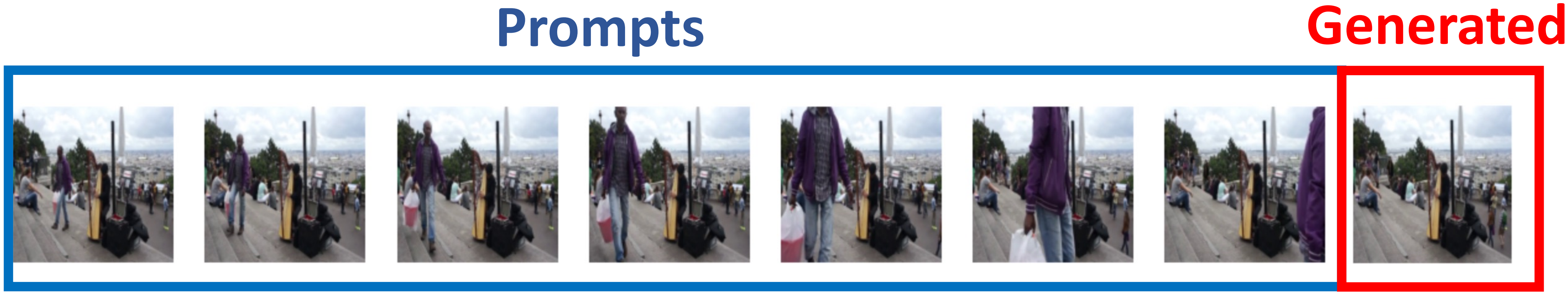


Figure 5. We evaluate the perplexity of 4 models trained on different sub-components of our datasets on tasks using the ImageNet validation set. All models are 3B parameters and all evaluations are conducted in the 5-shot setting. We can see that the model benefits from each of single images, videos and annotations, demonstrating the importance of our training dataset diversity.

Sequential Prompting

Prompts

Sequential Prompting

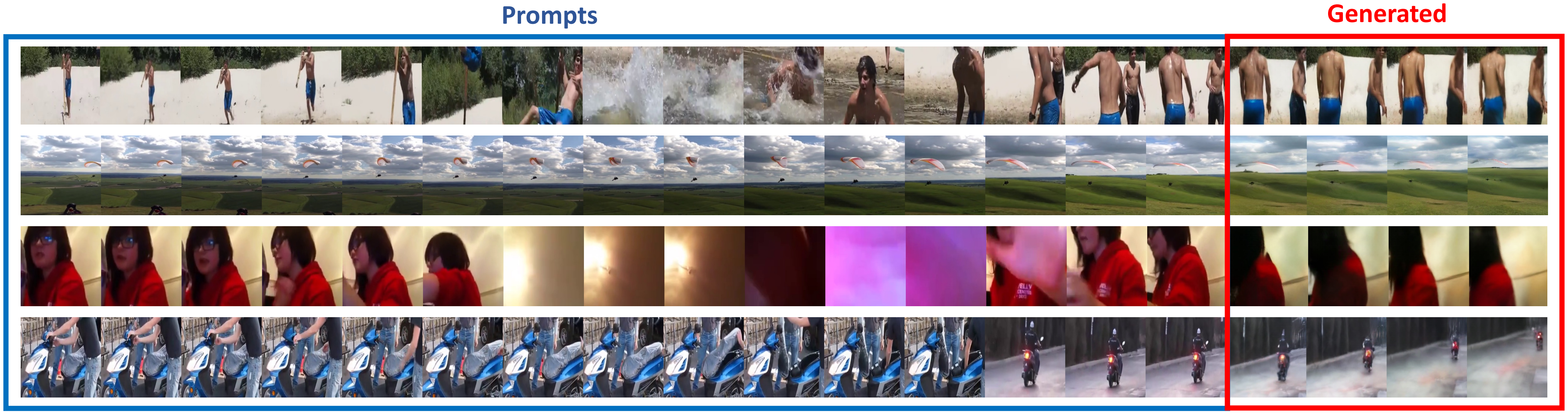


Sequential Prompting

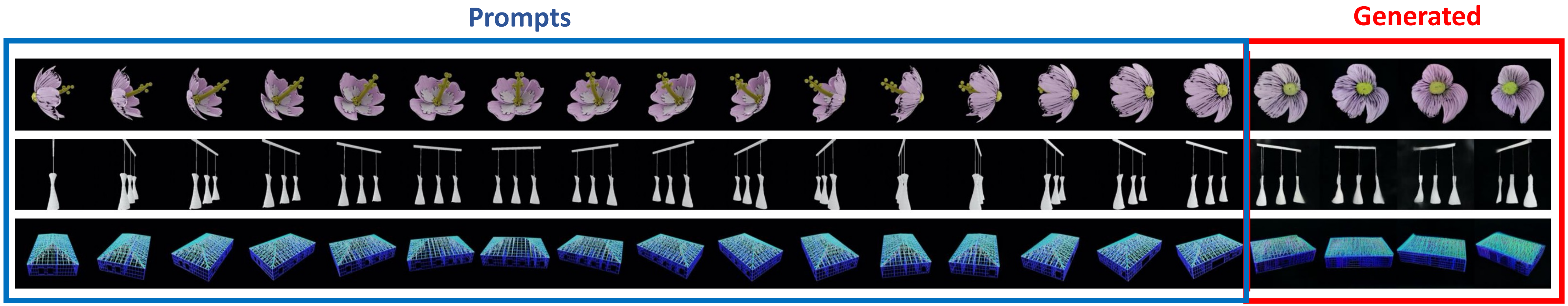
Sequential Prompting

Longer Contexts

Longer Contexts



Sequential Prompting

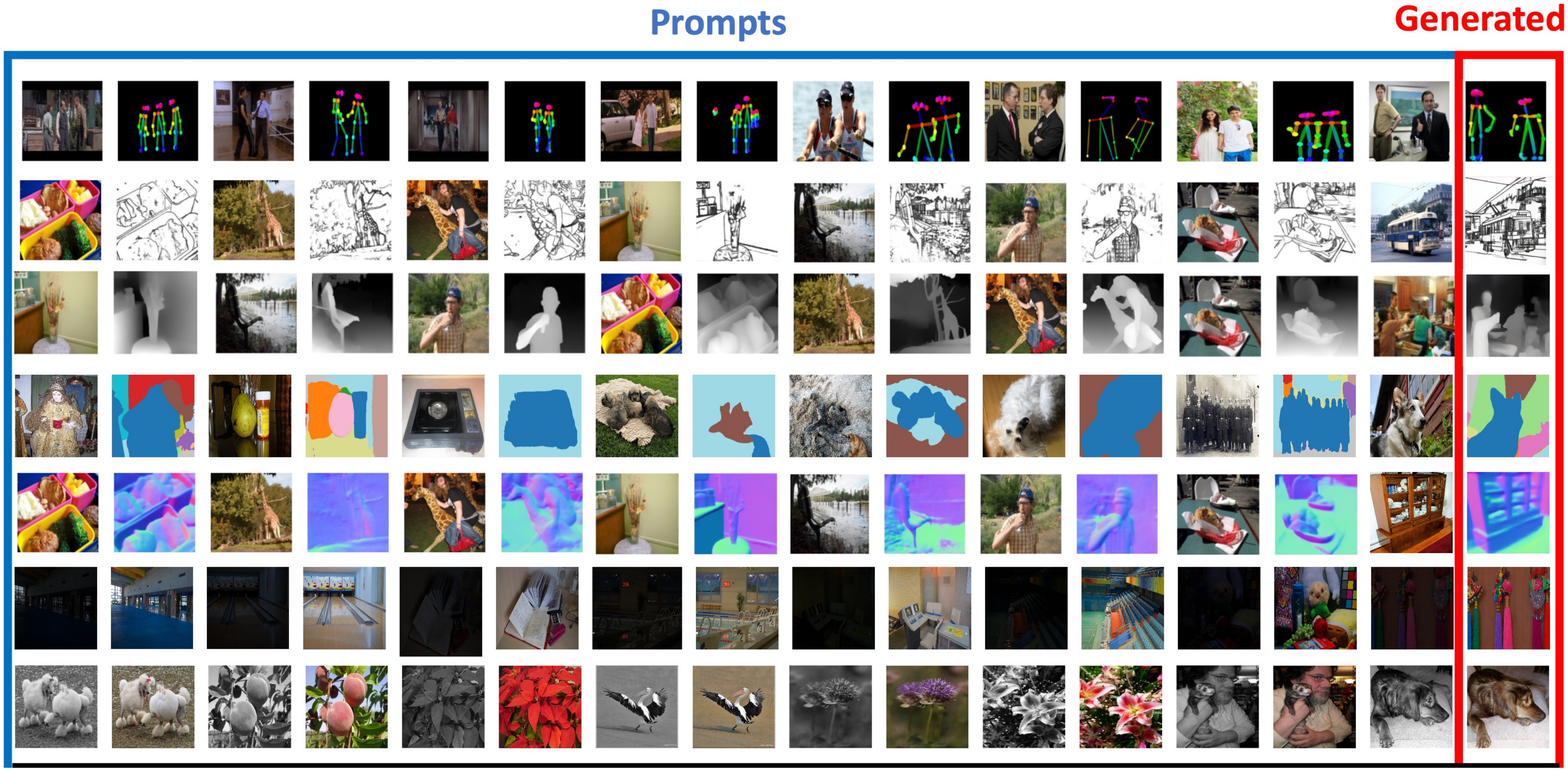


Sequential Prompting

Analogy Prompting

Analogy Prompting

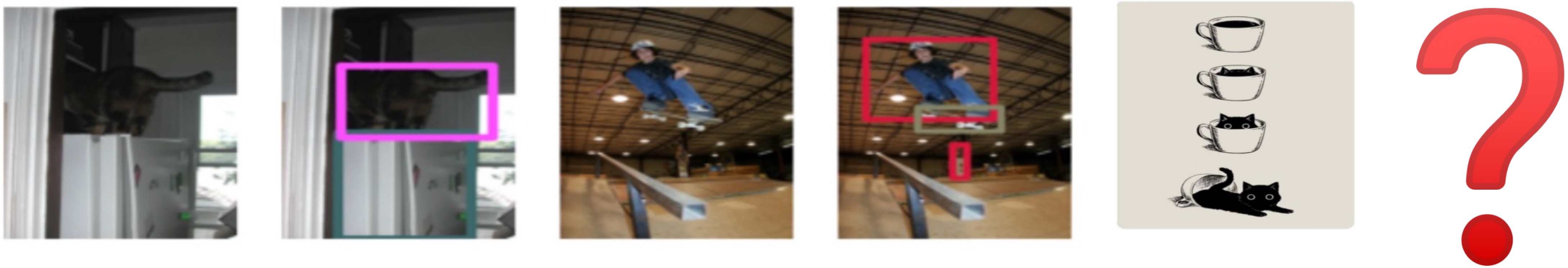
Analogy Prompting



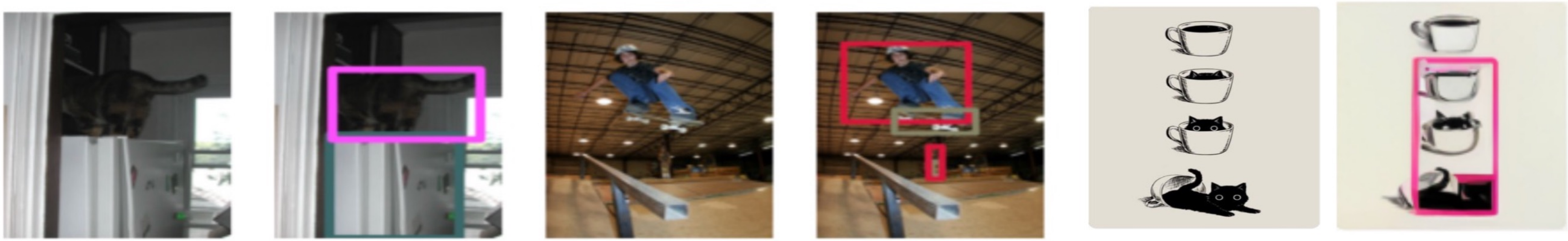
Analogy Prompting

	Inpaint MSE	Color MSE	Depth MSE	Surf MSE	Seg mIOU	KP Det PCKh	3DRot MSE	Denoise PSNR	Derain PSNR	LOL PSNR
Bar et al [6]	0.32	0.67	0.72	0.85	27.17	32.81	0.73	49.25	39.21	25.74
Wang et al [75]	1.27	1.50	0.75	1.37	13.76	78.67	1.79	**38.88**	29.49	22.40
Ours	**0.11**	**0.51**	**0.18**	**0.25**	**49.68**	**81.34**	**0.13**	35.50	**30.15**	**23.21**

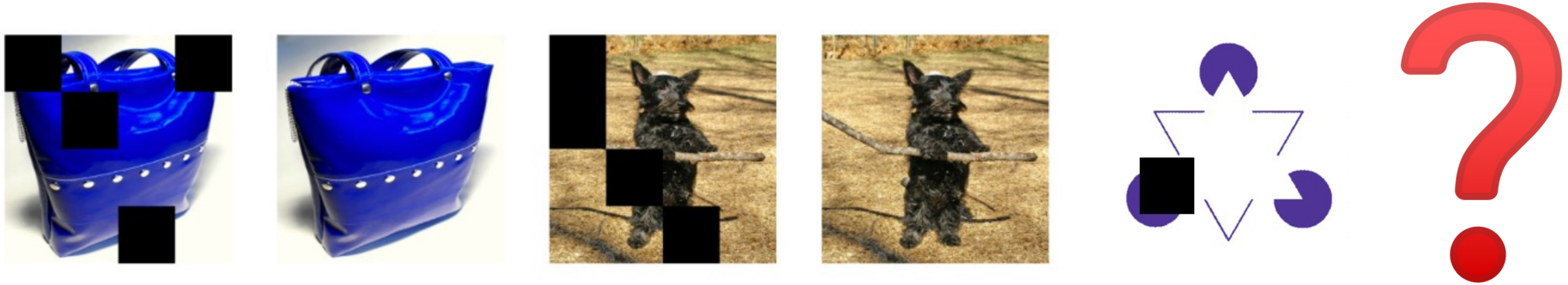
Analogy Prompting – Out of Domain Data



Analogy Prompting – Out of Domain Data

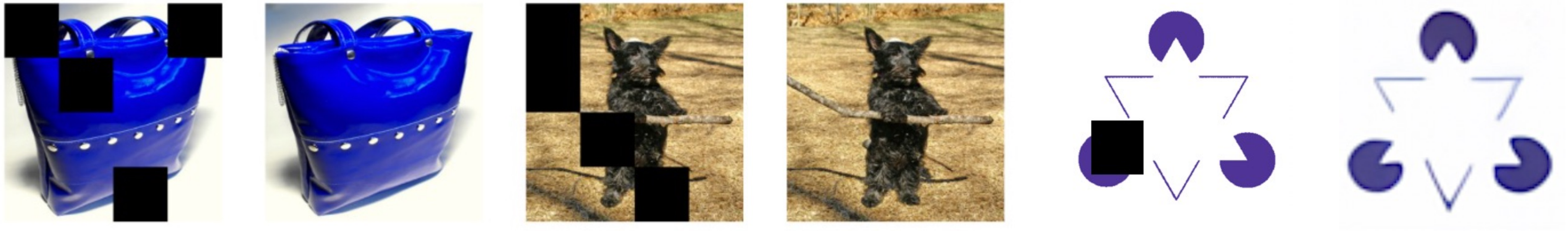


Analogy Prompting – Out of Domain Data



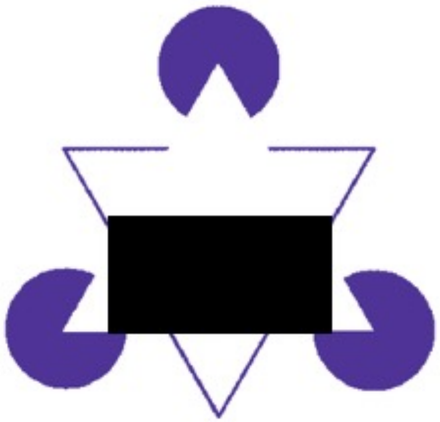
Analogy Prompting – Out of Domain Data

- corners

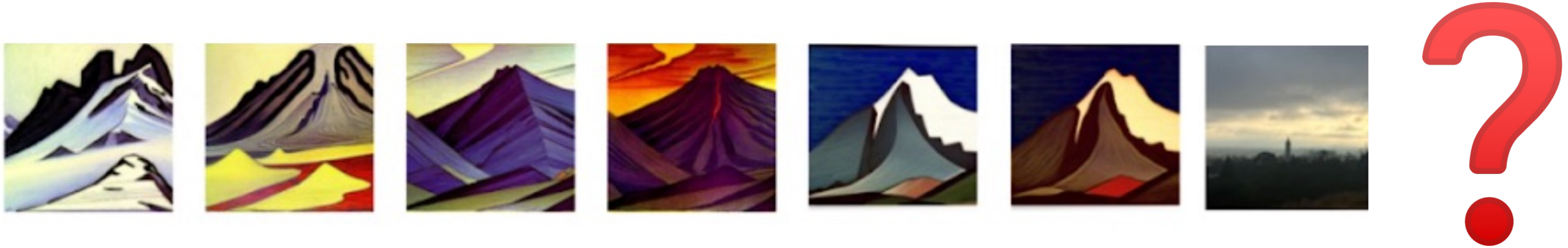


Analogy Prompting – Out of Domain Data

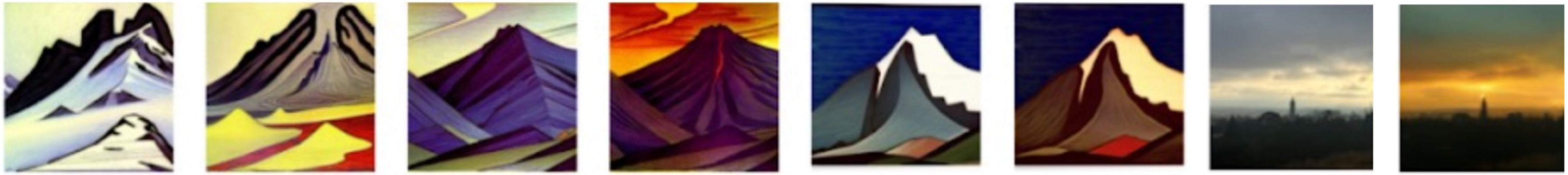
- edges



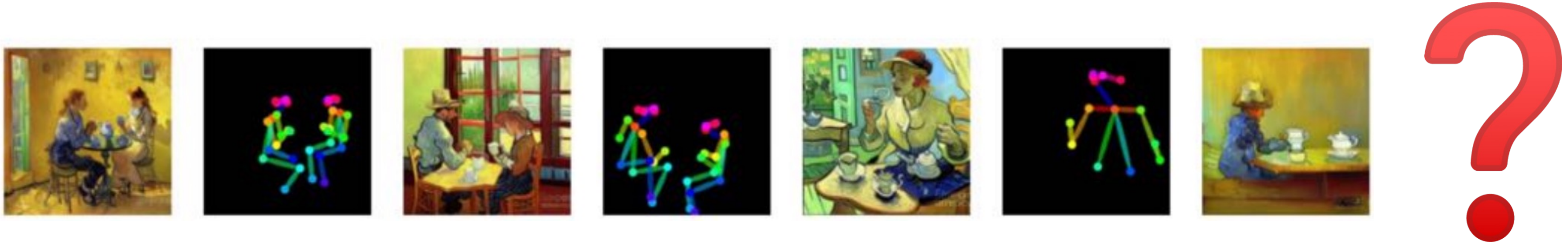
Analogy Prompting – Out of Domain Data



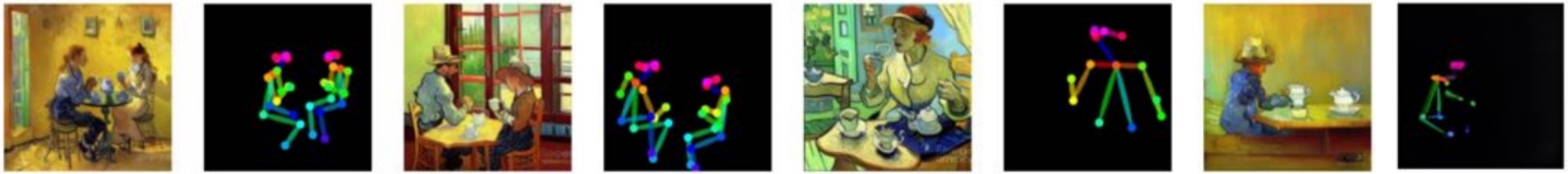
Analogy Prompting – Out of Domain Data



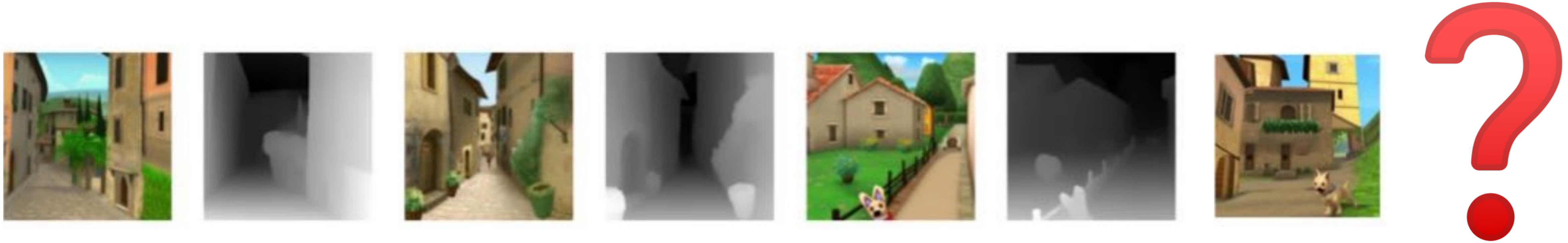
Analogy Prompting – Out of Domain Data



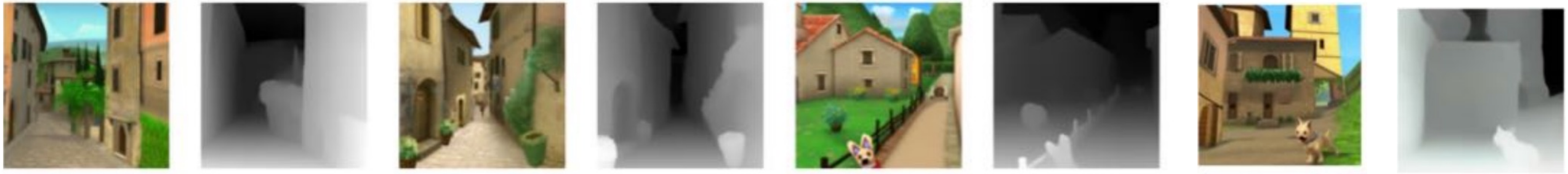
Analogy Prompting – Out of Domain Data



Analogy Prompting – Out of Domain Data



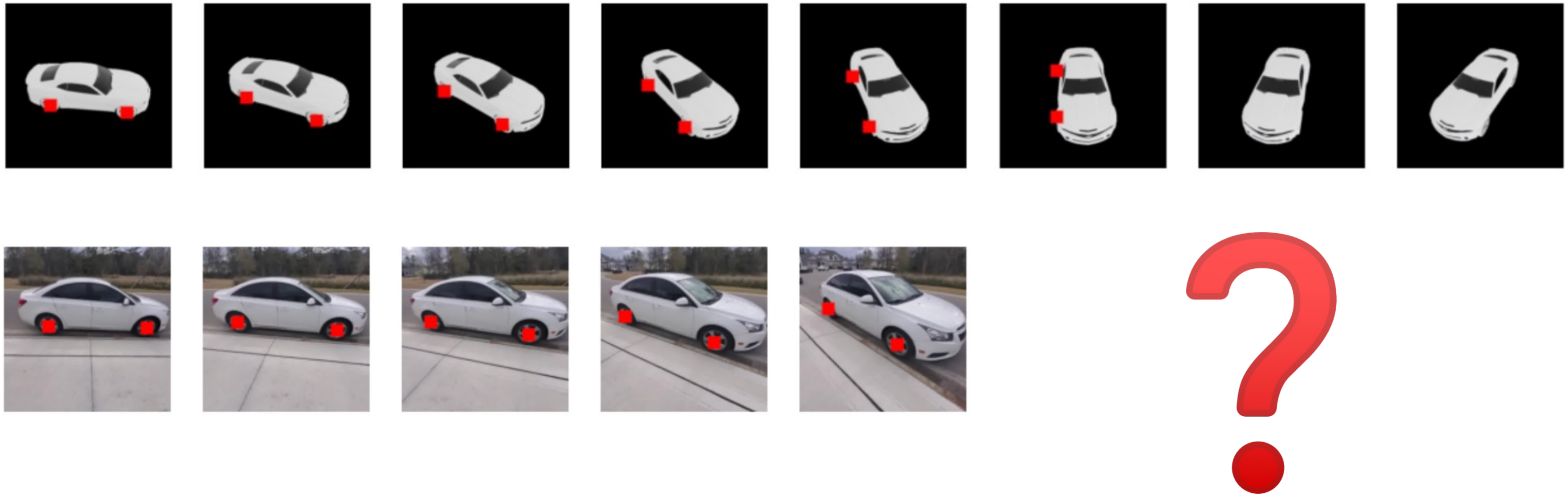
Analogy Prompting – Out of Domain Data



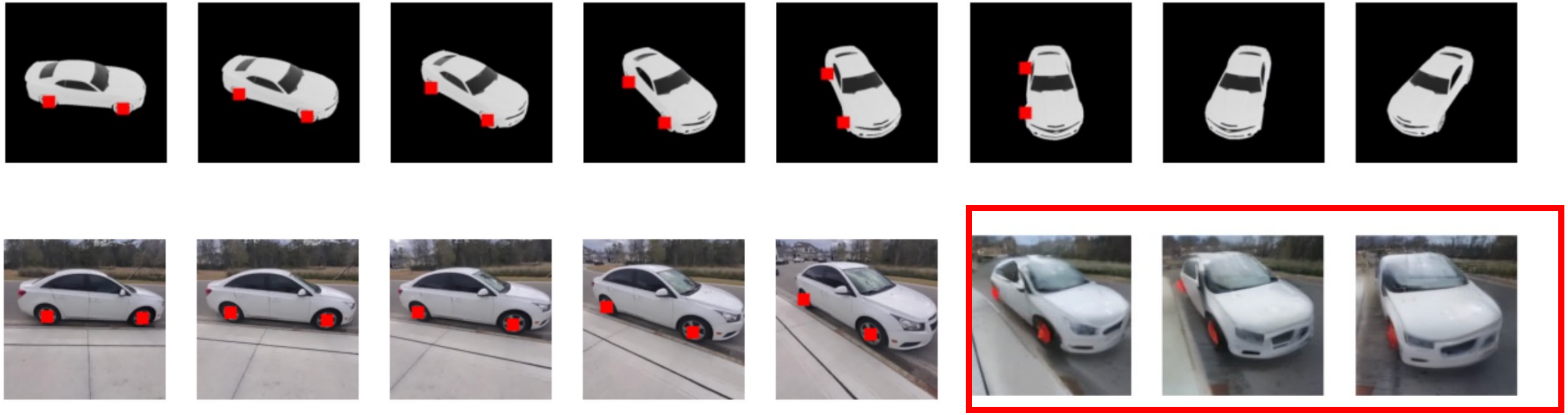
Compositional Prompts

More complicated

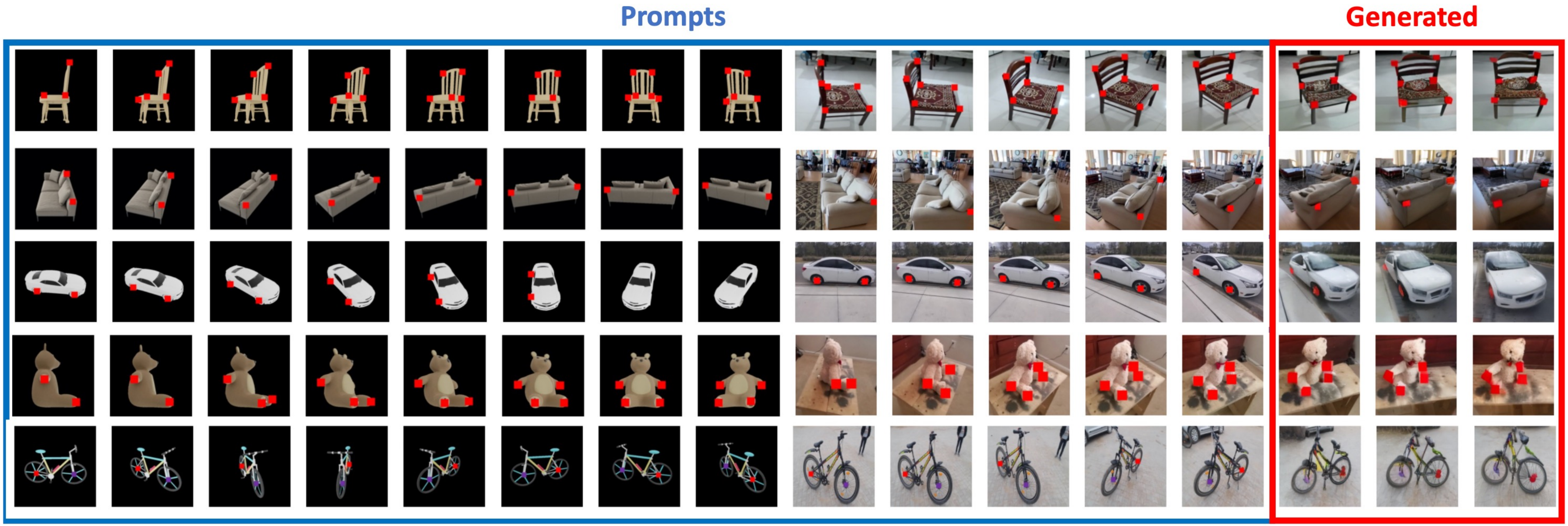
More complicated



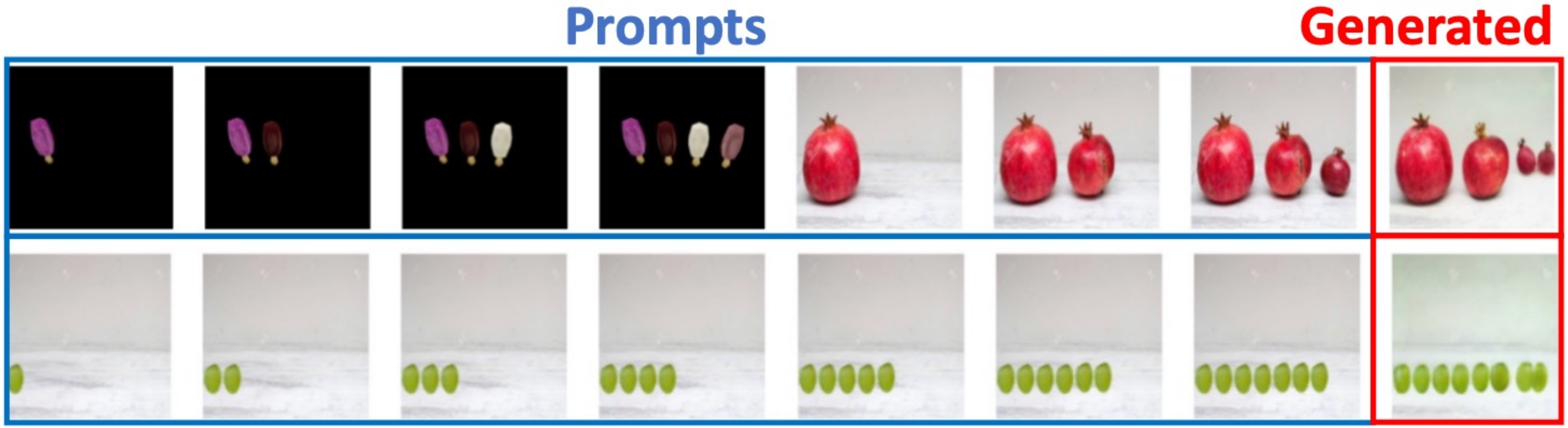
More complicated



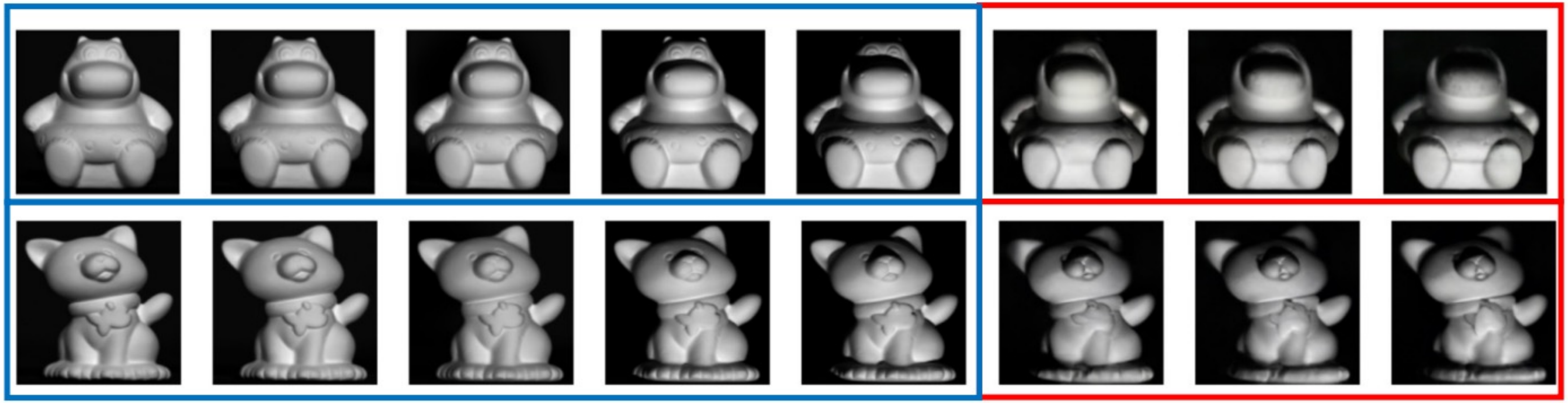
Compositional Prompts



Unseen tasks



Unseen tasks



Unseen tasks

Not easily describable

Not easily describable

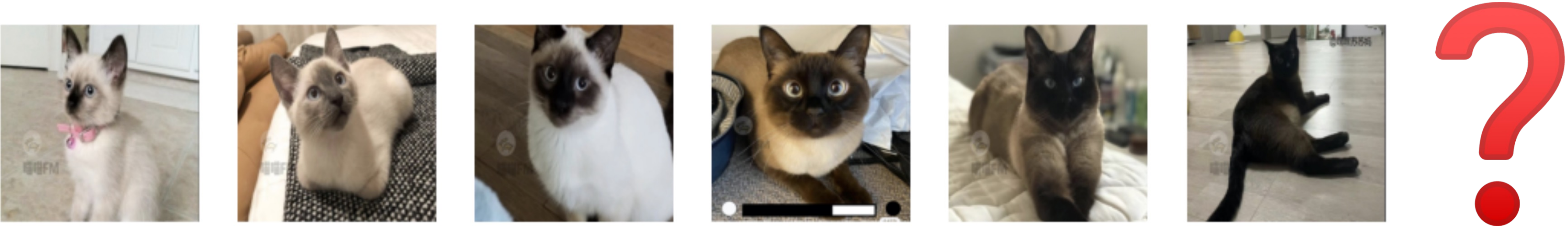
Not easily describable

Not easily describable

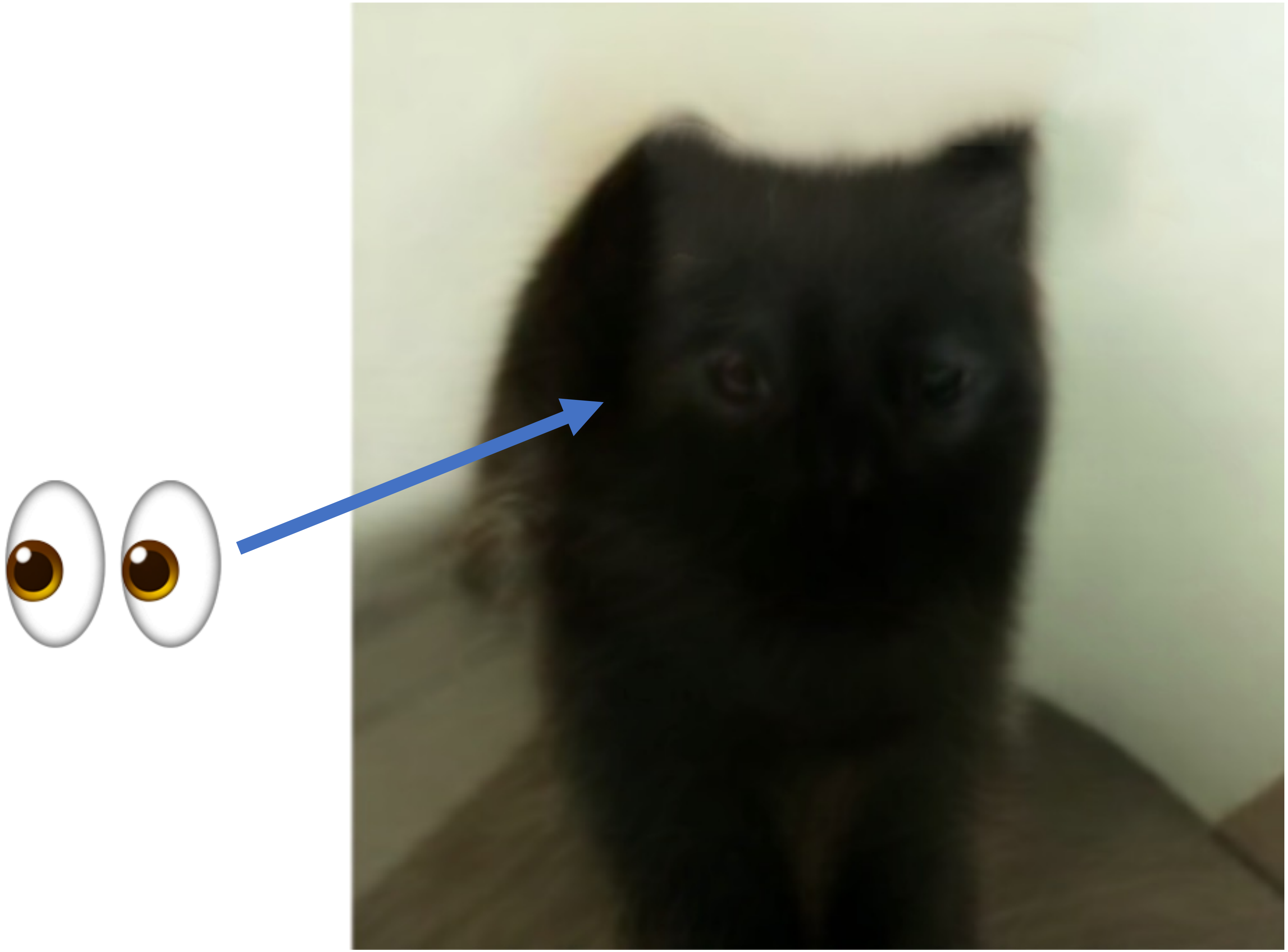
Not easily describable

Not easily describable

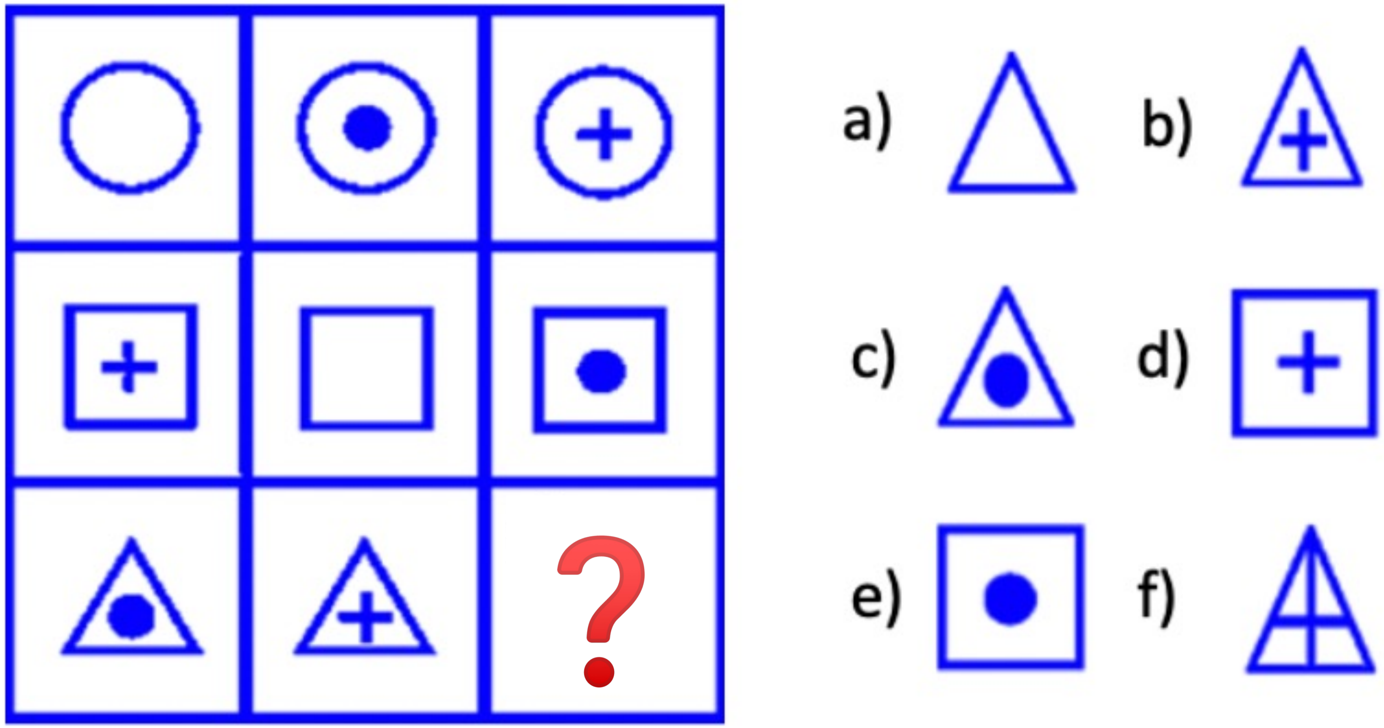
Not easily describable



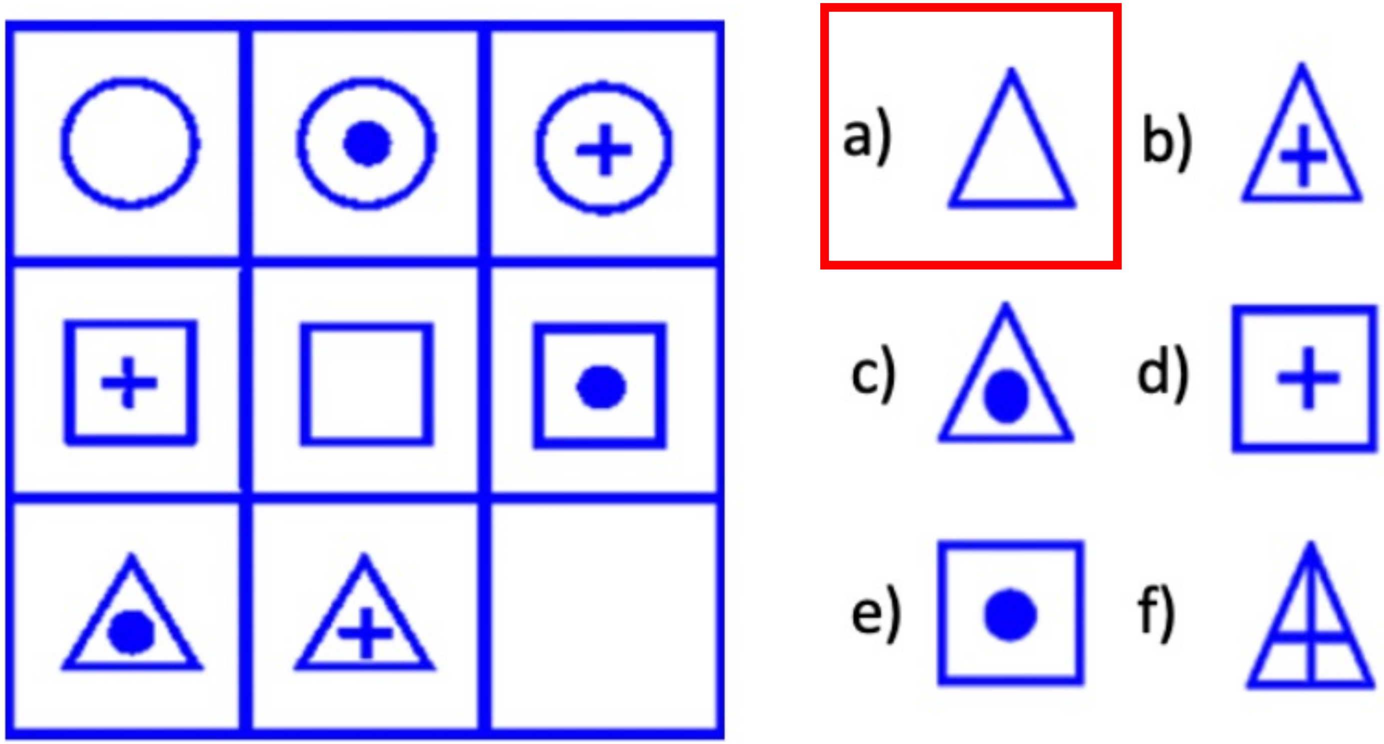
Not easily describable



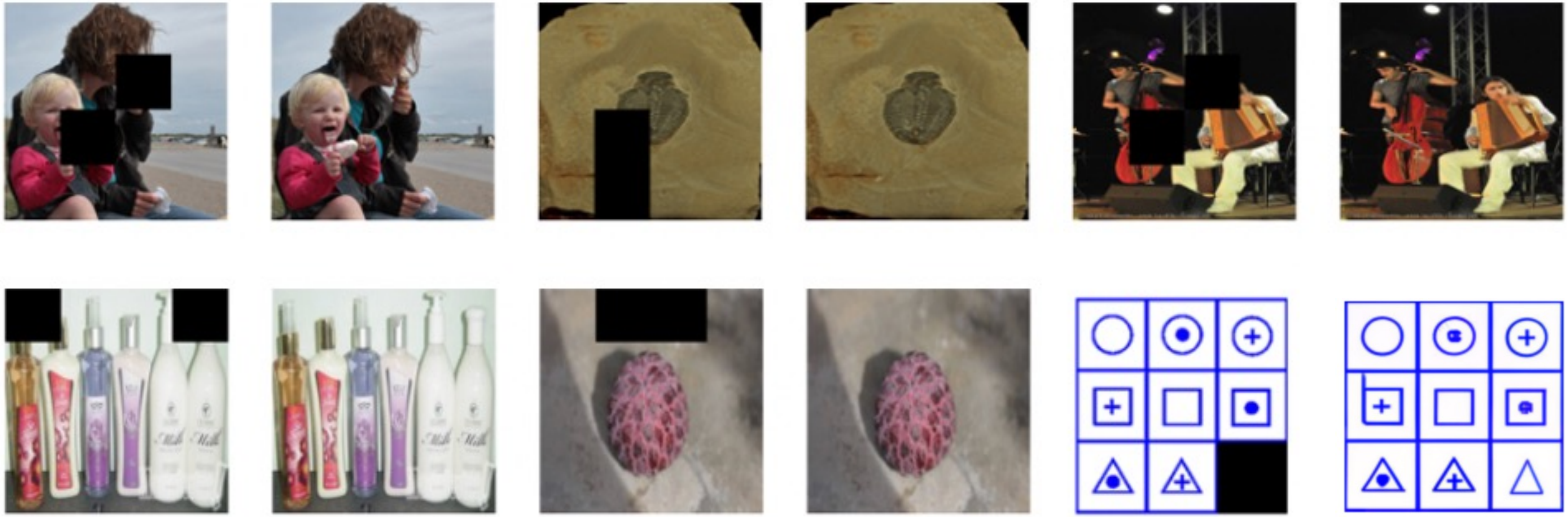
Raven’s Progressive Test (Non-verbal IQ test)



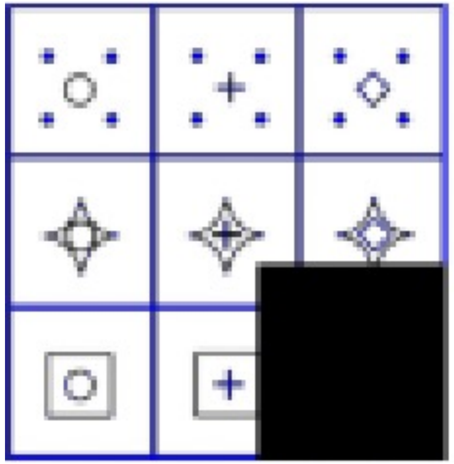
Raven’s Progressive Test (Non-verbal IQ test)



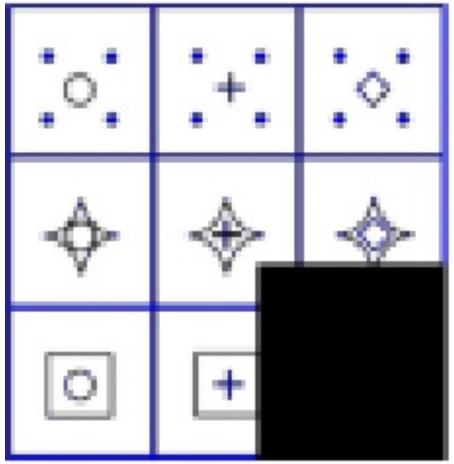
Raven’s Progressive Test (Non-verbal IQ test)



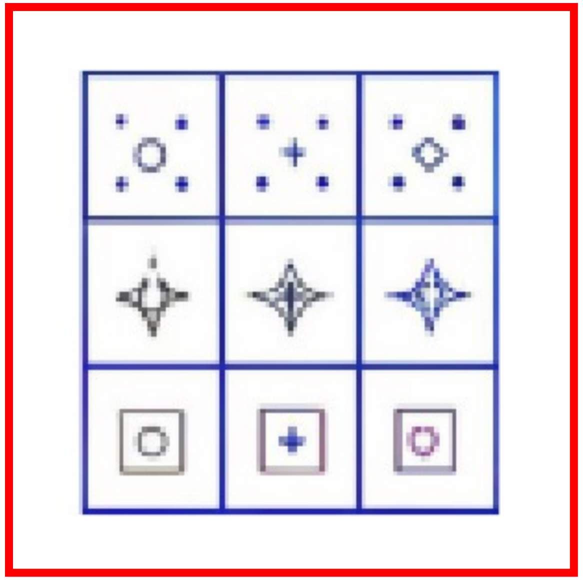
More Difficult?



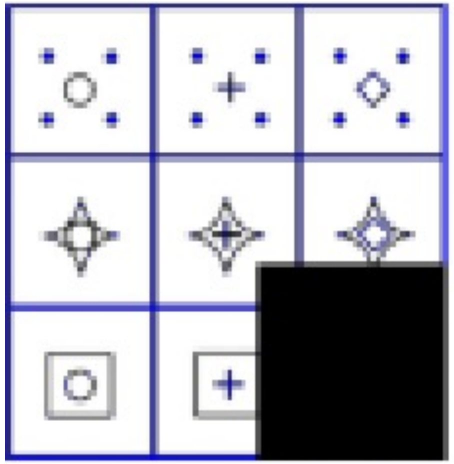
More Difficult?



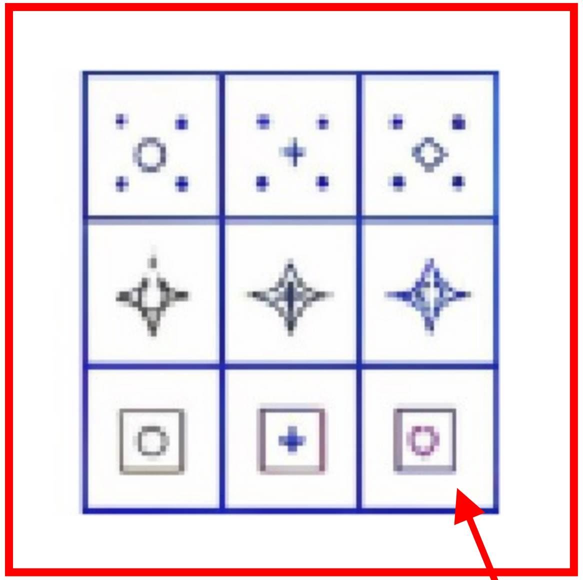
Generated



More Difficult?



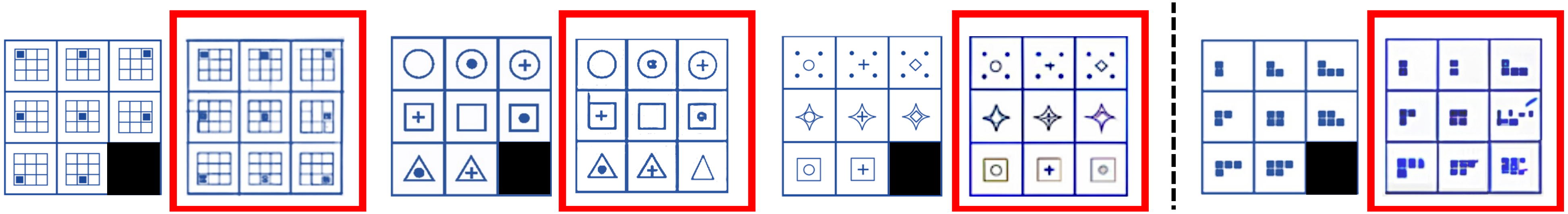
Generated



Hard to tell if correct or not 🤔

Perplexity

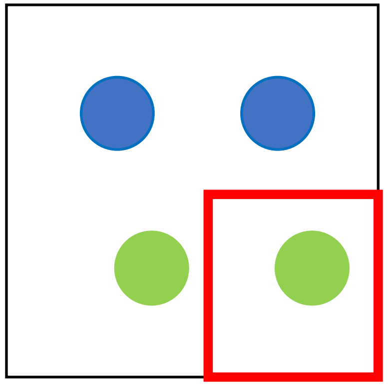
- 10 Questions:
  - performed perplexity analysis on classic Raven 5-way multiple-choice Matrices, choosing the answer with lowest perplexity.



	Raven’s Progressive Matrices
Chance	20%
Ours	30%

Synthetic Reasoning

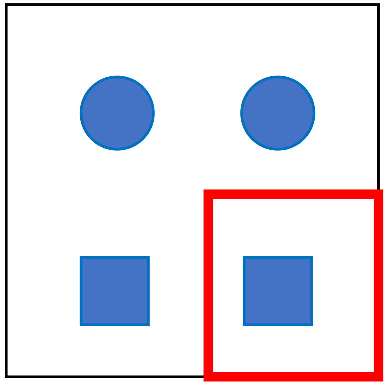
- **Color Change:** choose from 3 random generated colors.



	color
Chance	33%
Ours	**42%**

Synthetic Reasoning

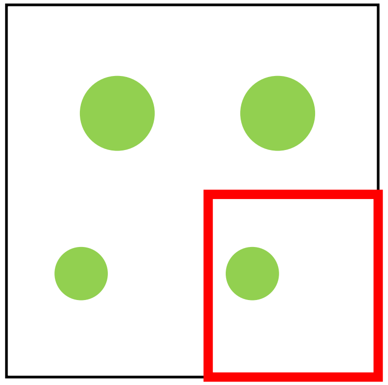
- **Shape Change:** choose from 3 random generated shapes.



	shape
Chance	33%
Ours	**45%**

Synthetic Reasoning

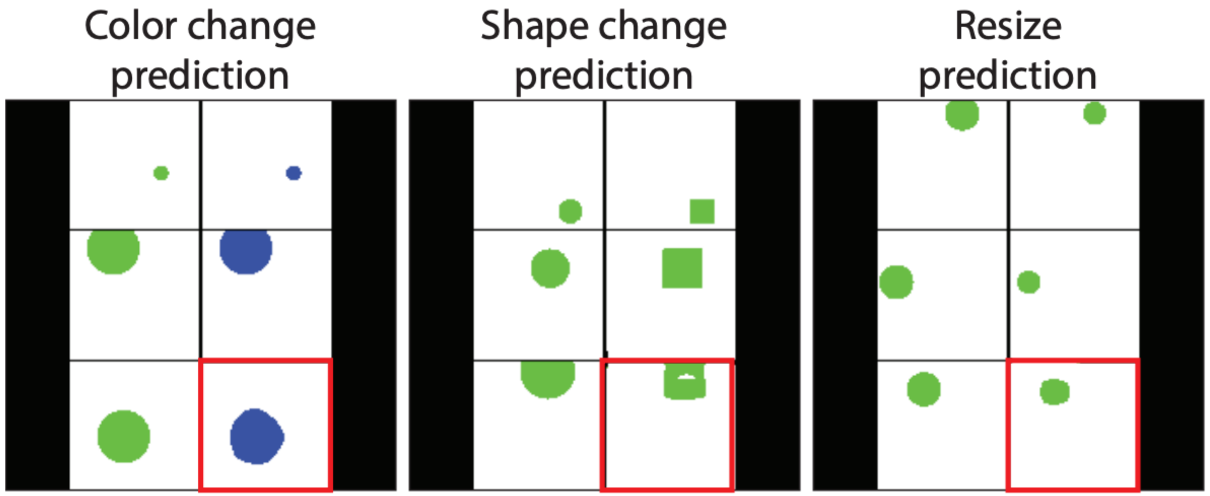
- **Size Change:** choose from 2 random generated sizes. (resolution)



	size
Chance	50%
Ours	**94%**

Synthetic Reasoning

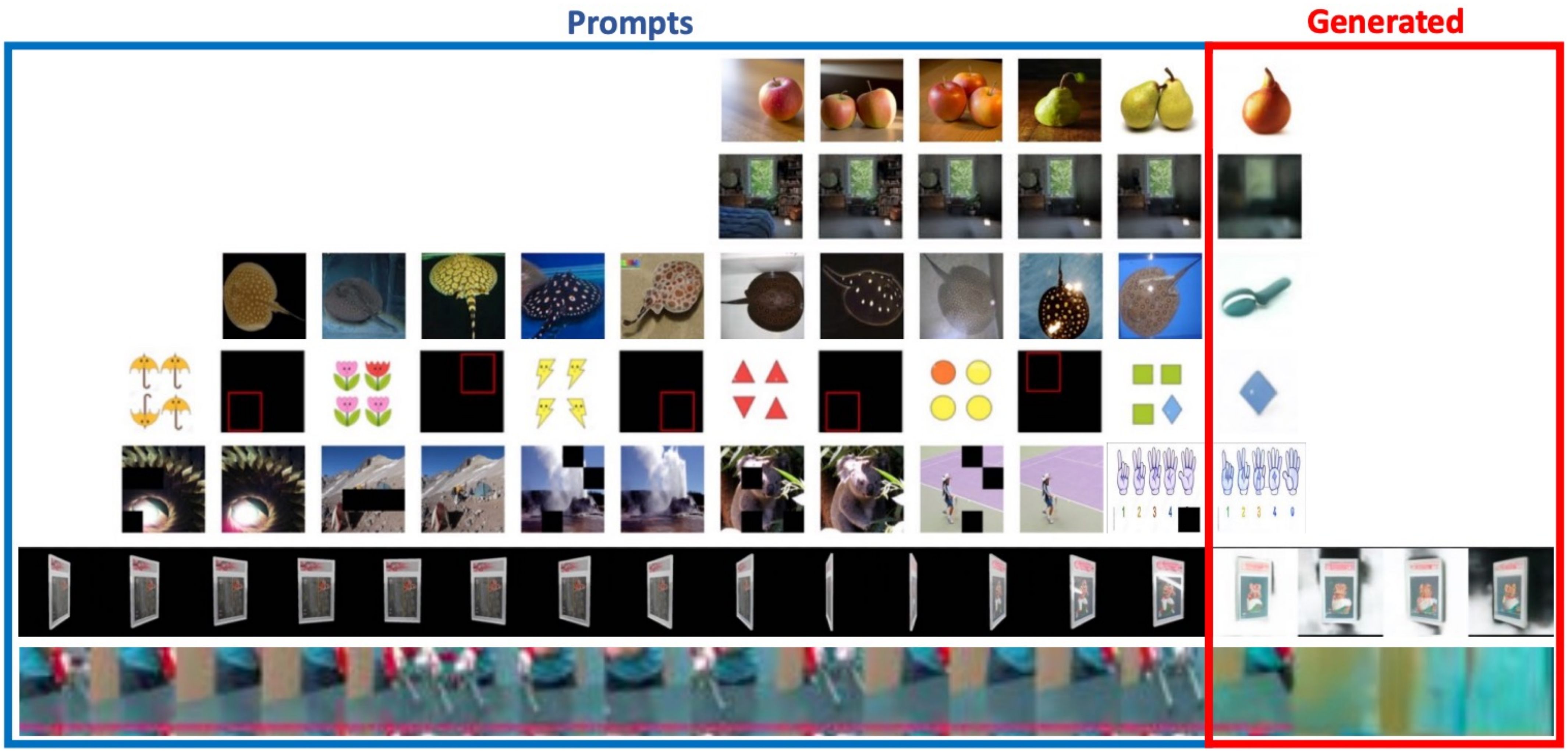
- In total 900 experiments



	color	shape	size
Chance	33%	33%	50%
Ours	**42%**	**45%**	**94%**

Visual Prompting via Image Inpainting, Bar et al.

Failure Case



Intrinsic Difference

Intrinsic Difference

People who listen to my talk. (I wish)

What’s not satisfying to me, yet

- Data:
  - Dataset distribution is so different from real life!
- Evaluations when things become more complicated.
  - Imagine you are driving in a dark night, rainy, and a person just walked passed your window…
  - Not a disentangled task.
- Training.
  - Is it hard enough for self-supervised learning yet?

Something to think about, maybe.

- ‘Supervised Training is an opium’.

Something to think about, maybe.

- ‘Supervised Training is an opium’.
- If ‘Supervised Training is an opium’, how about Language to Vision?

Something to think about, maybe.

- ‘Supervised Training is an opium’.
- If ‘Supervised Training is an opium’, how about Language to Vision?
- Do we bottom-up enough to fully unleash the power of visual data?

Thanks for listening

- Just a beginning.
- Despite this being one of the biggest vision models to date, it is still very small in comparison with modern Large Language Models

Code, Model, Demo courtesy of Hugging Face