

Listening to the Data: Visual Learning from the Bottom Up

Yutong Bai
UC Berkeley

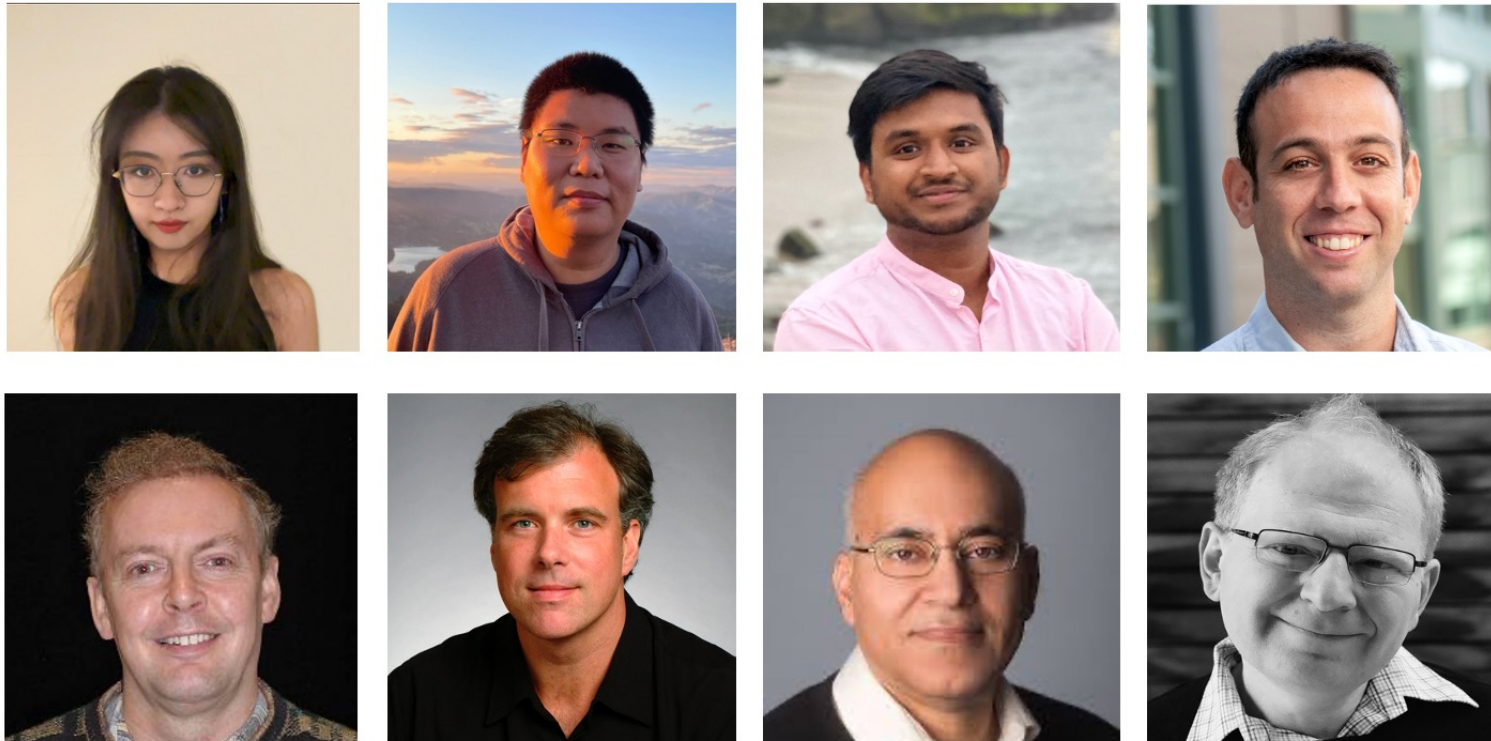


Self Introduction

- 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 **L**arge **V**ision **M**odels



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

LVM: Why LLM without Language?

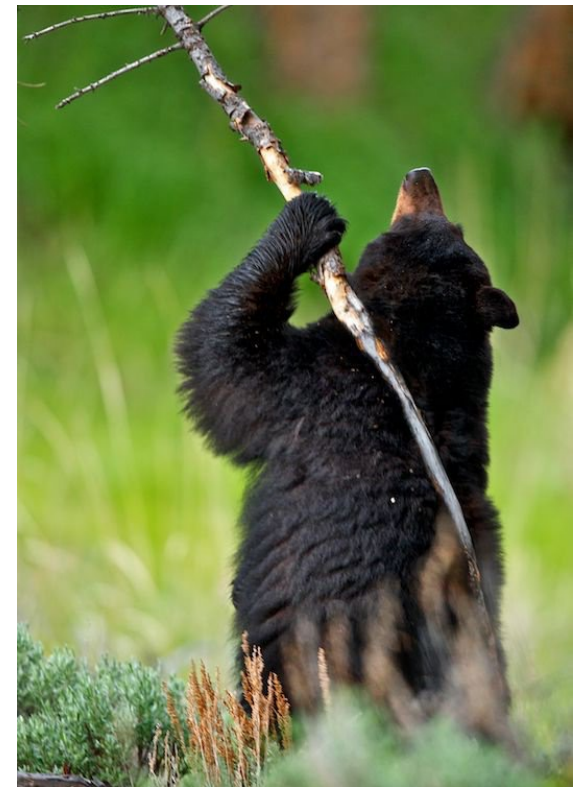
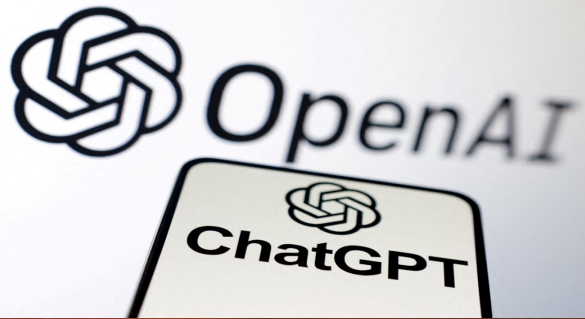


LVM: Why LLM without Language?

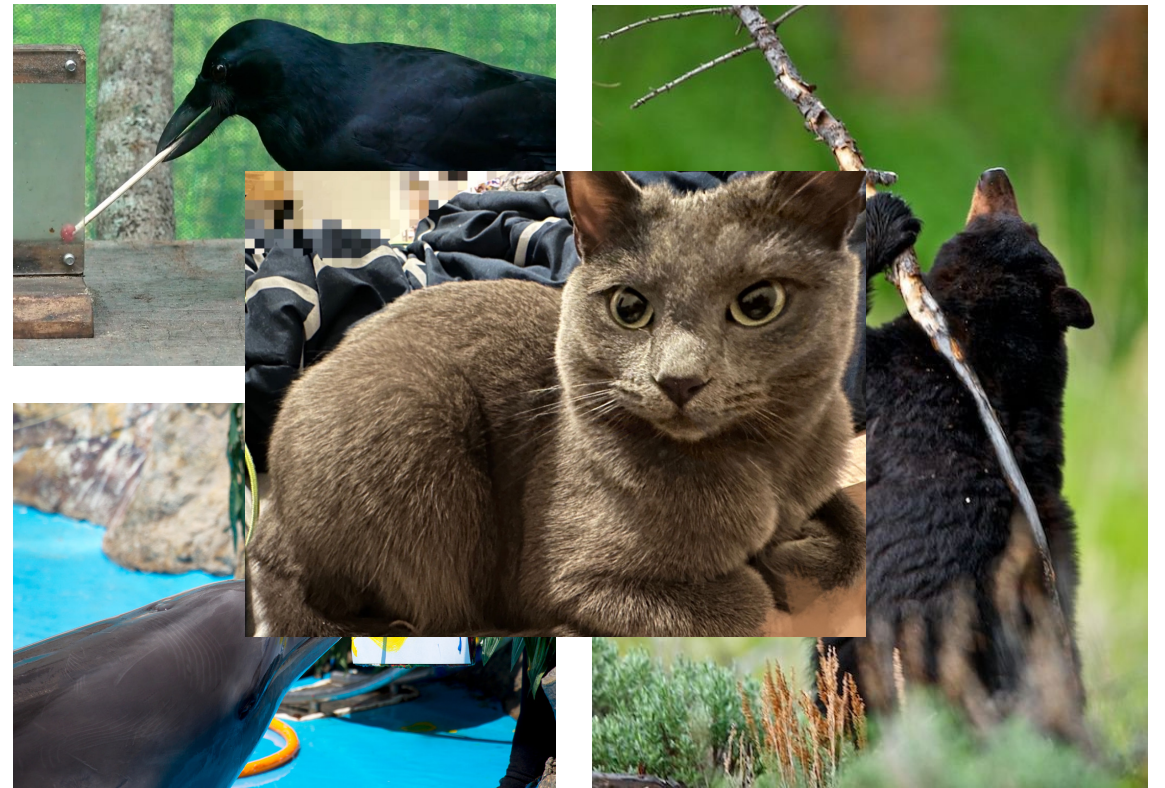


- Philosophical
- Practical

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)

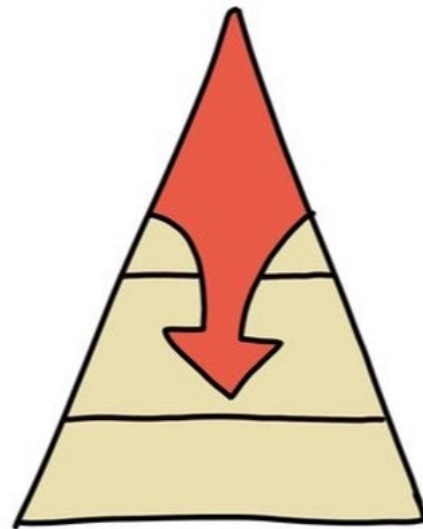


Self-Supervised Learning

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

Language,
Semantics,
Concepts

Pixels
(raw sensory data)



top-down

(supervised learning)

People who listen to my talk. (I wish)

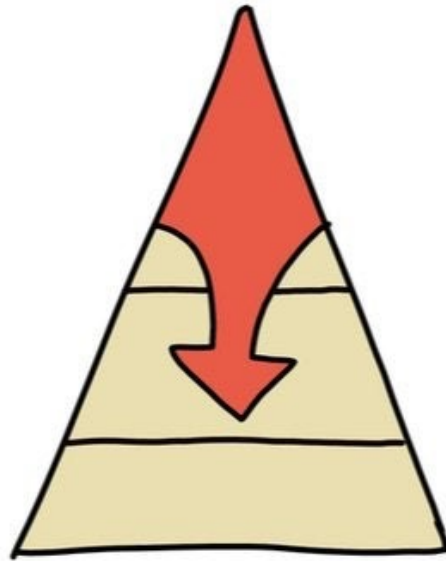


Self-Supervised Learning

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

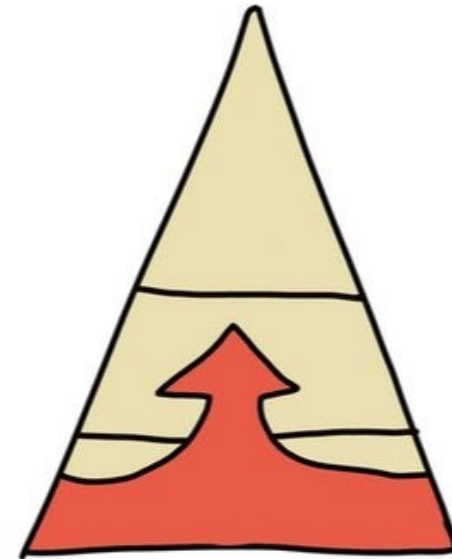
Language,
Semantics,
Concepts

Pixels
(raw sensory data)



top-down

(supervised learning)



bottom-up

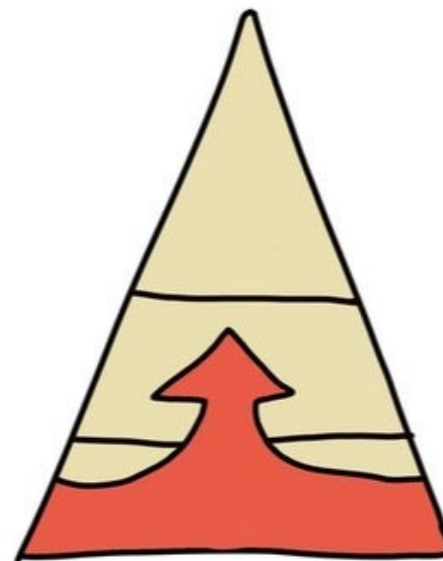
(self-supervised learning)

Self-Supervised Learning

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

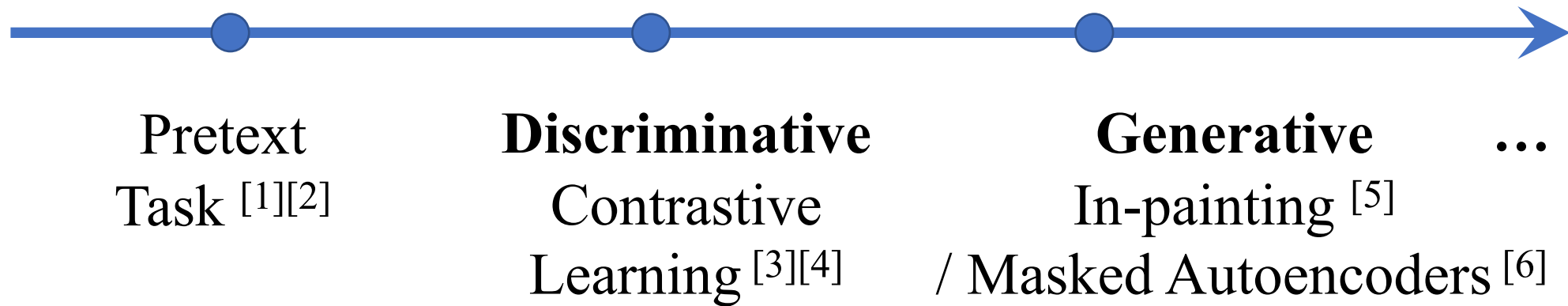
A Difficult task!

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



bottom-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.

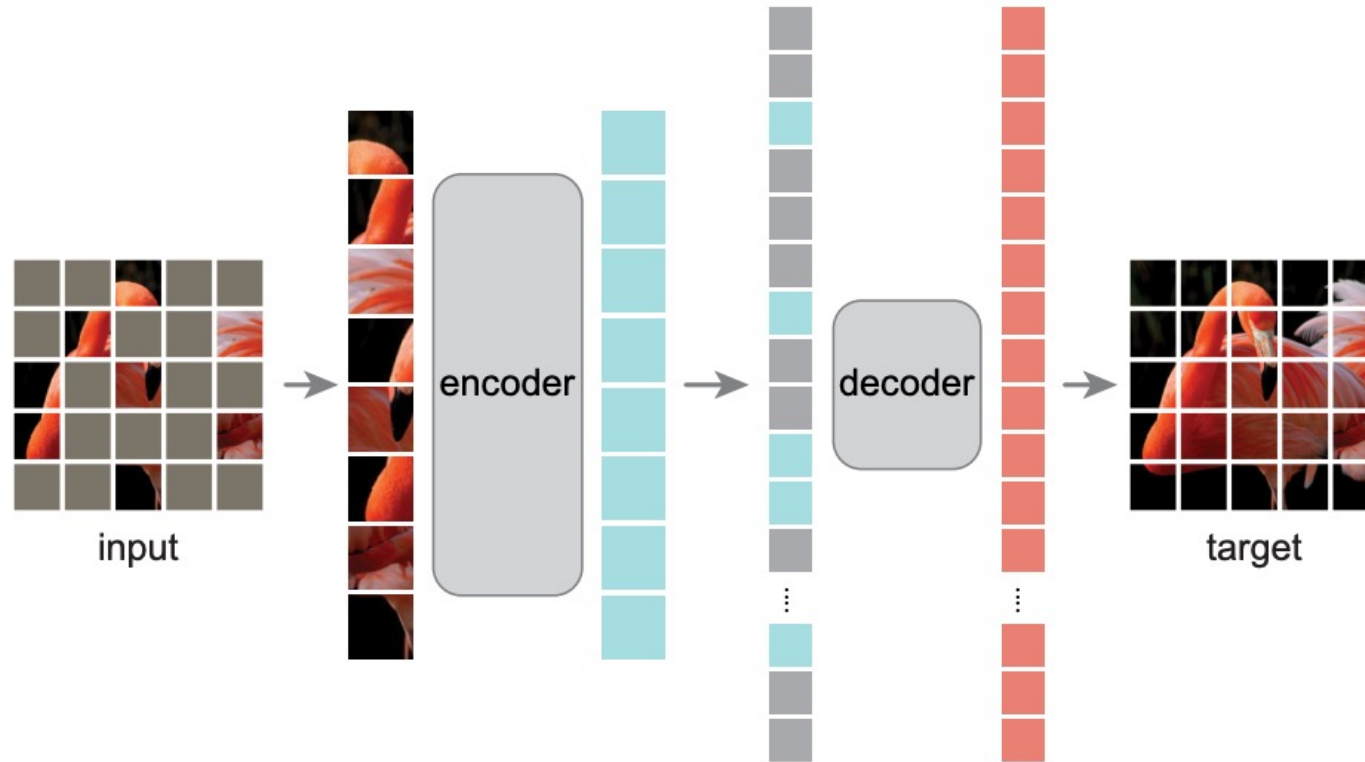
[3] Wu, Xiong, Yu and Lin. "Unsupervised feature learning via non-parametric instance discrimination." CVPR 2018.

[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.

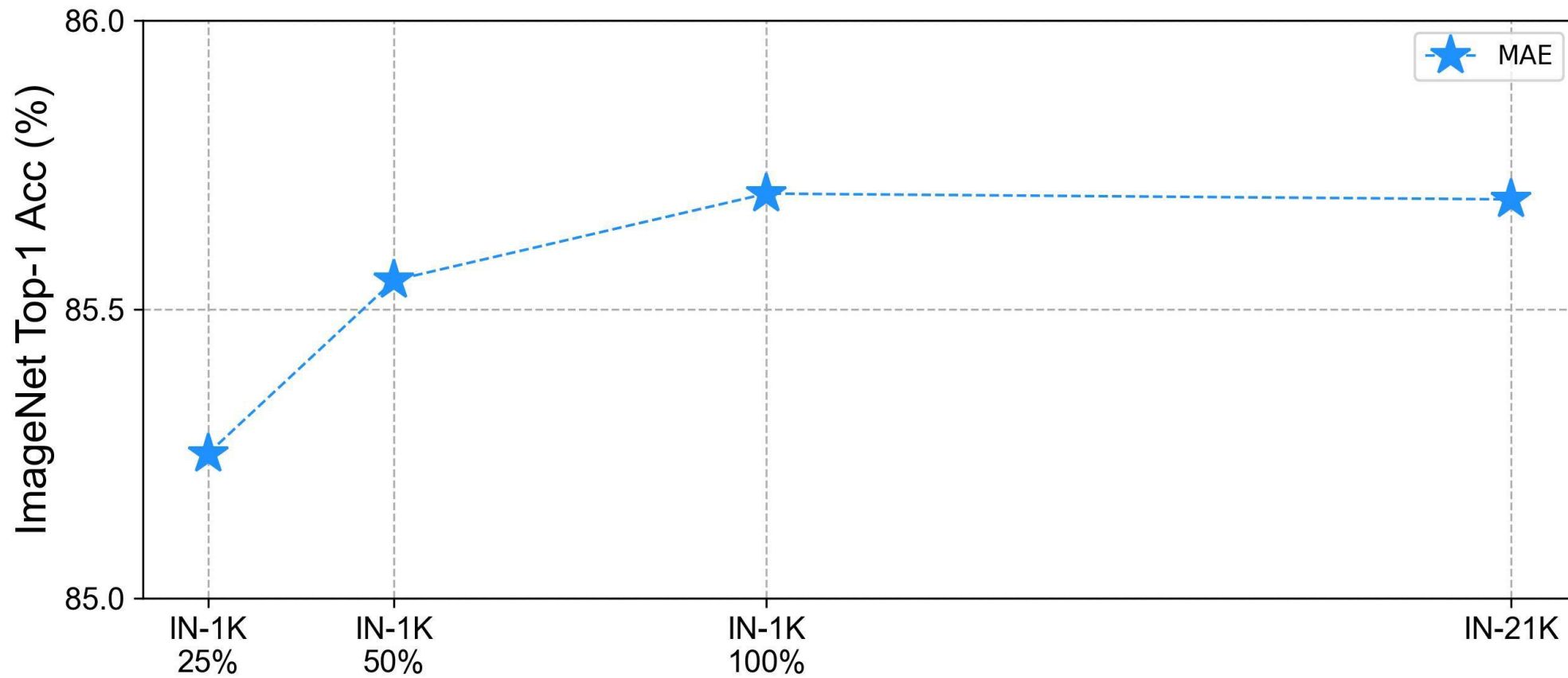
Masked Autoencoder (MAE) for Transformer



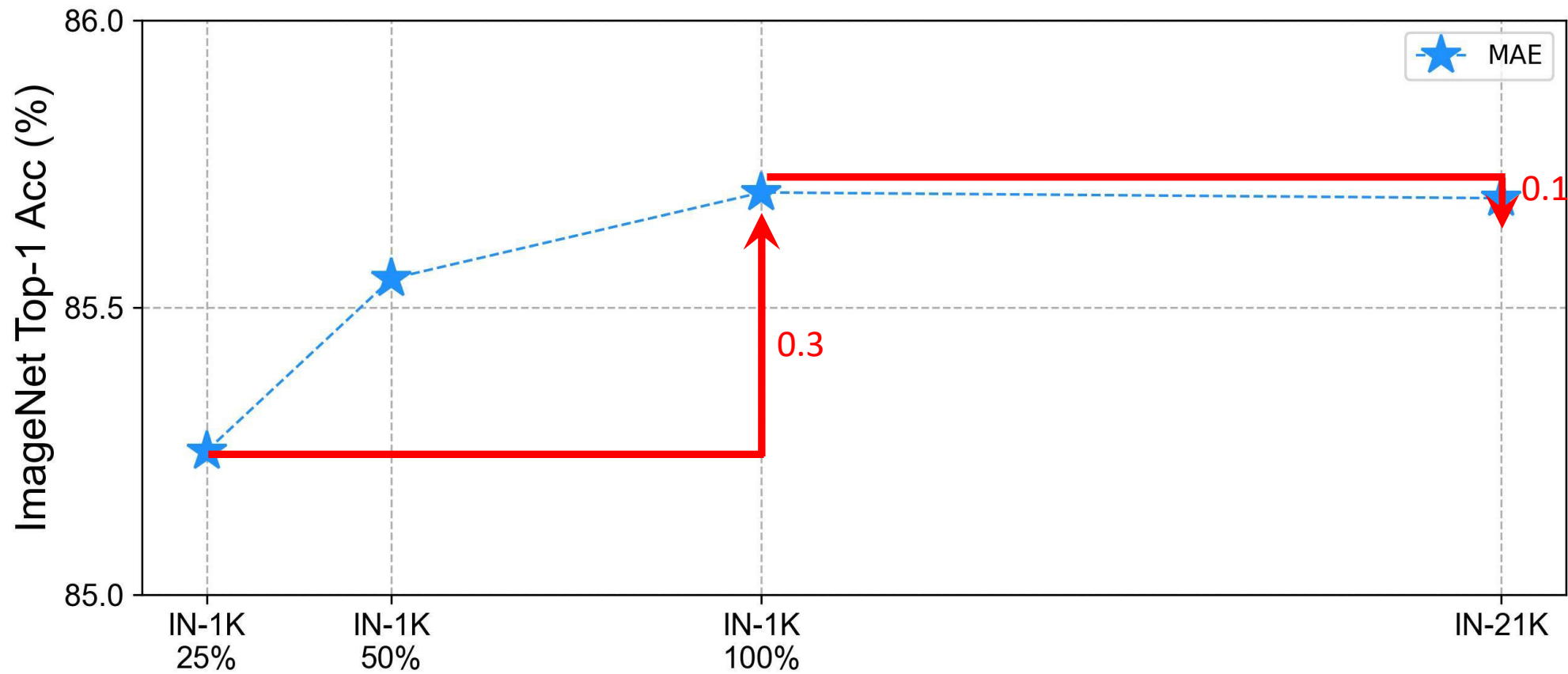
Pathak, Krahenbuhl, Donahue, Darrell and Efros. "Context encoders: Feature learning by inpainting." CVPR 2016.

He, Chen, Xie, Li, Dollár and Girshick. "Masked autoencoders are scalable vision learners." CVPR 2022.

Scaling Behaviors of MAE on Data



Scaling Behaviors of MAE on **Data**



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

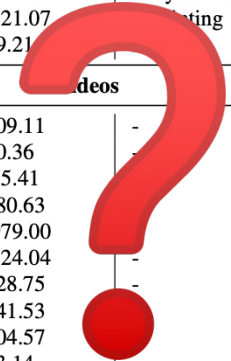
Task Specification: ~~Finetune~~ prompting

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	Image Denoising	Ground Truth
Kitti [34]	9.21	Ground Truth	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

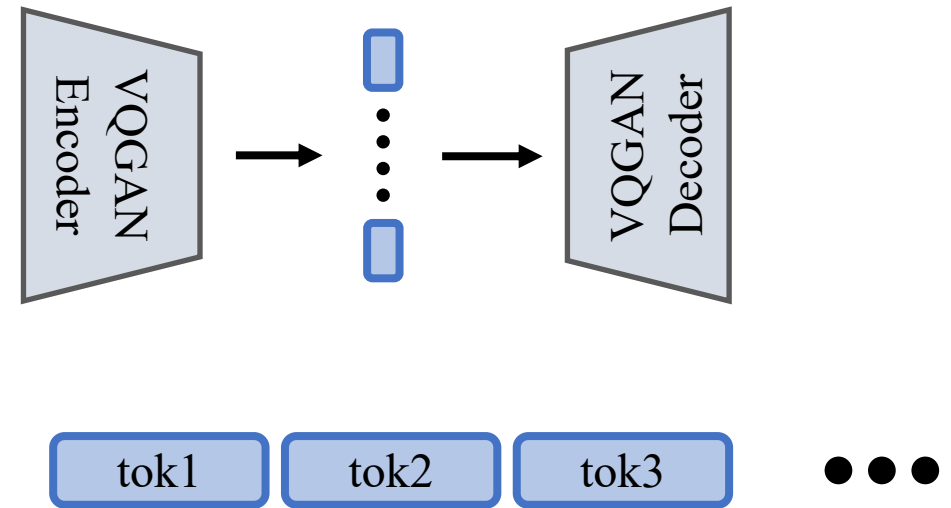


Image sequences

<BOS>



●●● <EOS>

Image sequences



Image sequences

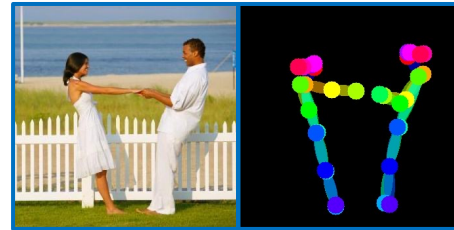
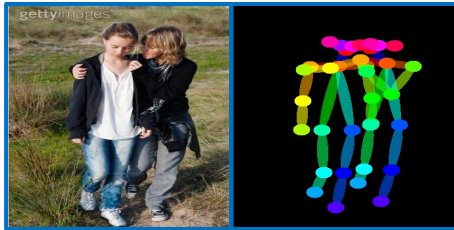
<BOS>



... <EOS>

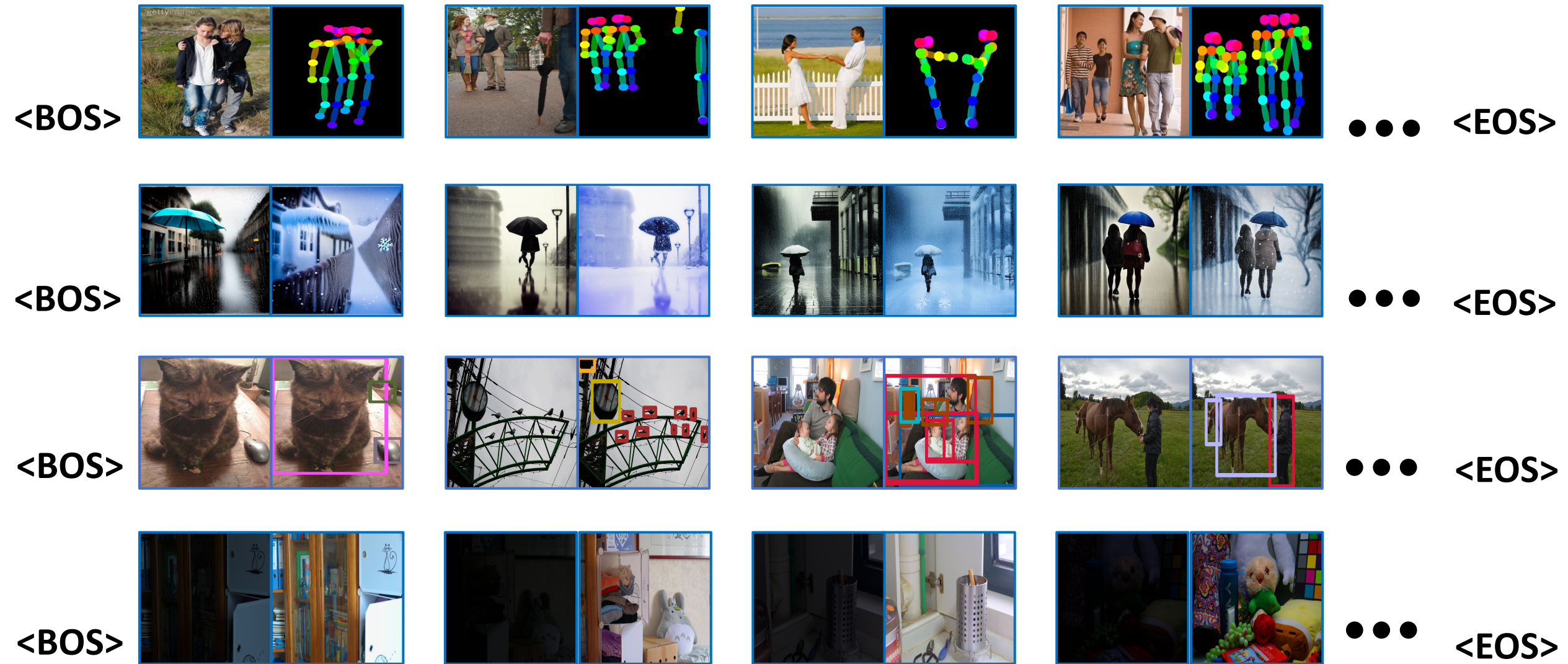
Images with annotation

<BOS>

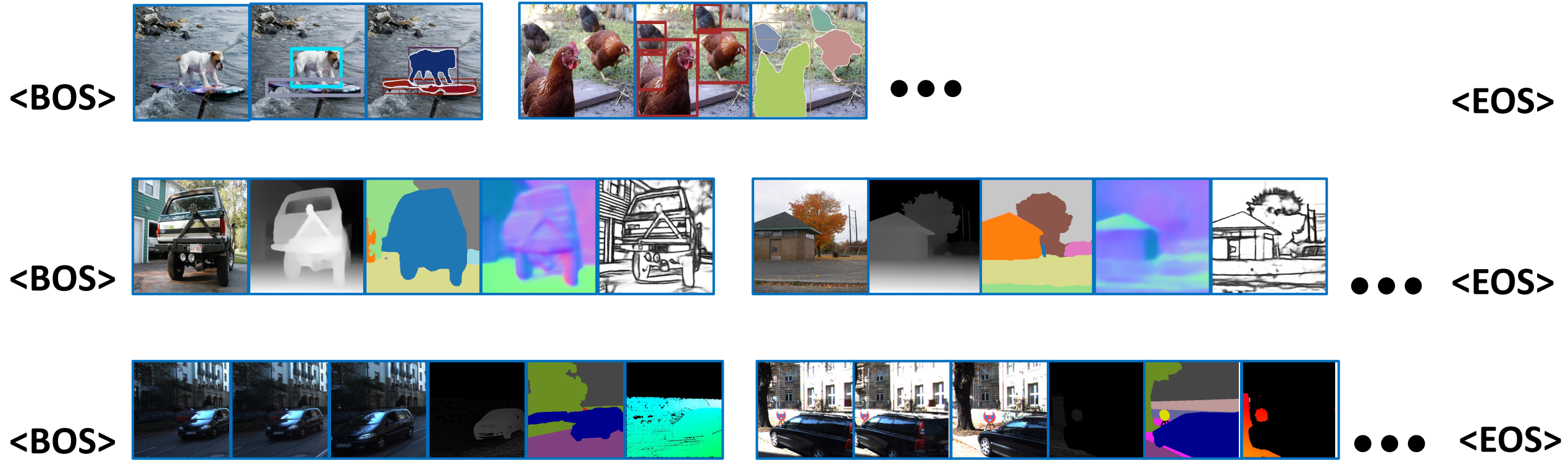


● ● ● <EOS>

Images with annotation

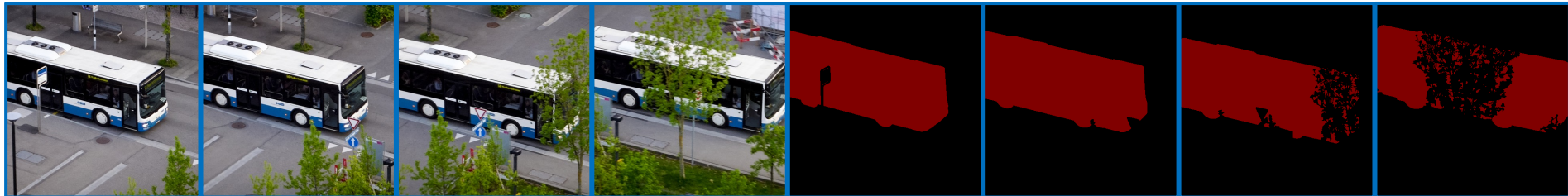


Images with free form annotation



Videos with annotation

<BOS>



<EOS>

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

Visual Sentences

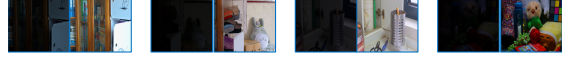
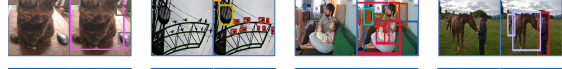


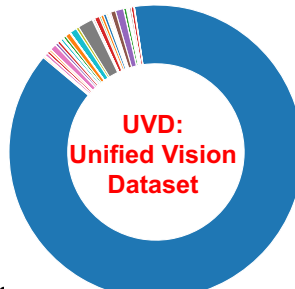
Image sequences, e.g. videos, 3D rotations, synthetic viewpoints

Images with annotation, e.g. style transfer, object detection, low light enhancement

Images with free form annotation, e.g. object detection + instance segmentation etc

Videos with annotation, e.g. video segmentation

Dataset Names	
Baion (380,000 tokens)	coco_generated_edge (0,3028 tokens)
multiSports (0,0748 tokens)	kinetics700_ab (2,3640 tokens)
denoise (0,2458 tokens)	ik_seq (1,3218 tokens)
ik_seq (1,3218 tokens)	kiti (0,0028 tokens)
mpii (0,0638 tokens)	ade20k (0,0517 tokens)
coco_generated_mixed (0,7570 tokens)	aug_detection (0,1228 tokens)
inpainting_coco (0,3028 tokens)	coco_generated_normal (0,3028 tokens)
cityscapes (0,0132 tokens)	hand24k (0,0020 tokens)
coco_pose (0,3048 tokens)	dic_mom_heavy (0,0088 tokens)
ucf101 (0,1091 tokens)	eg4d (1,1521 tokens)
charades_v1 (0,2415 tokens)	sisu (0,1108 tokens)
DIC_MDR_light (0,0081 tokens)	light_enhance (0,0005 tokens)
charades_ego (0,1931 tokens)	mpii_back_2d (0,0638 tokens)
diving48 (0,1078 tokens)	activity1m (0,3806 tokens)
ik_category_edge (1,1395 tokens)	youtuve_vis_annotation (0,0711 tokens)
hmdb_optical_flow (0,0190 tokens)	ik_length (1,3218 tokens)
kinetics (3,8476 tokens)	ik_caterpillar (1,3218 tokens)
kinetics700_12 (2,3640 tokens)	mat_vt1 (0,0573 tokens)
ik_category_2s (0,6557 tokens)	momentum_time (2,9790 tokens)
obaverse (0,1741 tokens)	hmdb51 (0,0548 tokens)
coco_pose_generated (0,2123 tokens)	hmdb_pose (0,0190 tokens)
coco_cat (0,3628 tokens)	ycb_cook (0,0001 tokens)
ik_category_1s (0,6568 tokens)	3d_pose (0,0438 tokens)
coco_pose_generated_back_2d (0,2123 tokens)	kin8mp (0,0394 tokens)
ik_category_depth (1,3150 tokens)	airair (0,0318 tokens)
ik_painting (1,3218 tokens)	shv2 (0,9048 tokens)
coco_generated_img_200 (0,7570 tokens)	kin8mp_back_2d (0,0394 tokens)
kinetics700_24 (2,3640 tokens)	ik_category_normal (1,1395 tokens)
refortization (0,0768 tokens)	hmdb_pose (0,0190 tokens)
ik_category_4s (0,6588 tokens)	sisu (0,0004 tokens)
instruc_xvizpix (0,4155 tokens)	ik_normal (1,3218 tokens)
inpainting_w (7,2699 tokens)	youtuve_vis_kfns (0,0637 tokens)
ik_category_img (1,1395 tokens)	coco_generated_depth (0,3028 tokens)
ik_cat (1,3027 tokens)	vis_seq (0,0645 tokens)
mpii_cat (0,0500 tokens)	co3d_seq (0,2288 tokens)
DIC_MDR_medium (0,0081 tokens)	poster (0,0008 tokens)



420B tokens,
60s Datasets.

- Information
- Diversity

LVM: Large Vision Model

Visual Sentences

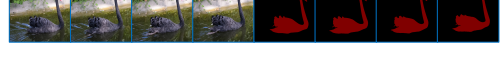
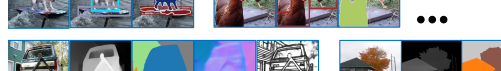
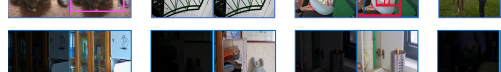
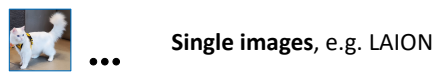


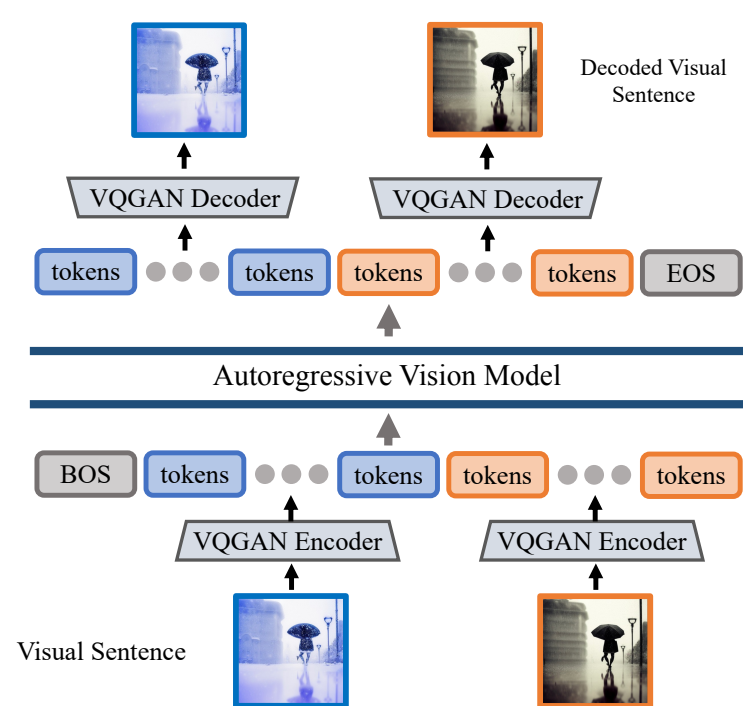
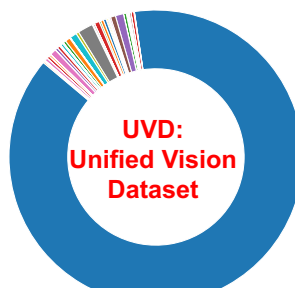
Image sequences, e.g. videos, 3D rotations, synthetic viewpoints

Images with annotation, e.g. style transfer, object detection, low light enhancement

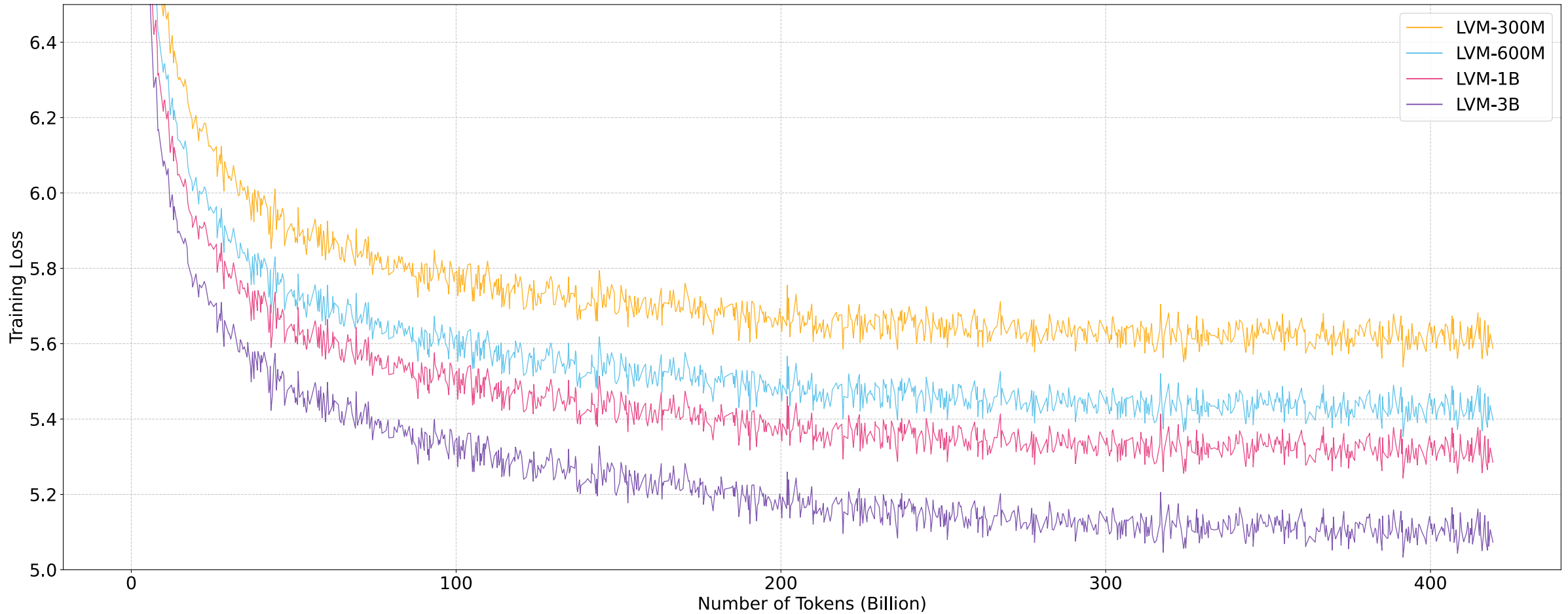
Images with free form annotation, e.g. object detection + instance segmentation etc

Videos with annotation, e.g. video segmentation

Dataset Names	
Baion (380,000 tokens)	coco_generated_edge (0,30288 tokens)
multiSports (0,07848 tokens)	kinetics700_s8 (2,36408 tokens)
denoise (0,24588 tokens)	ik_seq (1,32118 tokens)
ik_seq (1,32118 tokens)	Mini (0,00928 tokens)
mpii (0,06398 tokens)	ade20k (0,05178 tokens)
coco_generated_mixed (0,75708 tokens)	swg_detection (0,12298 tokens)
inspaining_coco (0,30288 tokens)	coco_generated_normal (0,30288 tokens)
cityscape (0,01328 tokens)	hand24k (0,00208 tokens)
coco_pose (0,38148 tokens)	DIU_NW_heavy (0,00818 tokens)
ucf101 (0,10918 tokens)	egqa4 (1,15218 tokens)
charades_v1 (0,21458 tokens)	sisu (0,11808 tokens)
ICD_MDR_light (0,00818 tokens)	light_enhance (0,00058 tokens)
charades_ego (0,19318 tokens)	mpii_back_29 (0,06398 tokens)
diving48 (0,10718 tokens)	activity_net (0,38068 tokens)
ik_category_edge (1,31958 tokens)	youtuve_vis_annotation (0,07118 tokens)
jhmdb_optical_flow (0,01908 tokens)	ik_length (1,32118 tokens)
kinetics700_s12 (2,36408 tokens)	ik_categorization (1,32118 tokens)
kinetics (3,84768 tokens)	msc_vt (0,05738 tokens)
kinetics700_s12 (2,36408 tokens)	moments_in_time (2,97908 tokens)
ik_category_2s (0,65878 tokens)	hmdb51 (0,00548 tokens)
obaverse (0,17418 tokens)	jhmdb_pose (0,01908 tokens)
coco_pose_generated (0,21238 tokens)	coco_cot (0,06238 tokens)
ik_category_1s (0,65688 tokens)	3d_pose (0,04388 tokens)
coco_pose_generated_bow_29 (0,21238 tokens)	kin8mp (0,03948 tokens)
ik_category_depth (1,31508 tokens)	aircraft (0,03118 tokens)
ik_painting (1,32118 tokens)	shv2 (0,90468 tokens)
coco_generated_img_2900 (0,70708 tokens)	kin8mp_bow_29 (0,03948 tokens)
kinetics700_s24 (2,36418 tokens)	ik_category_normal (1,31958 tokens)
refortization (0,07688 tokens)	jhmdb_bow_pose (0,01908 tokens)
ik_category_4s (0,65988 tokens)	divis (0,00048 tokens)
instruc_xc2pix (0,41558 tokens)	ik_normal (1,32118 tokens)
inspaining_v1 (7,26998 tokens)	youtuve_vis_fips (0,06378 tokens)
ik_category_img (1,31958 tokens)	coco_generated_depth (0,30288 tokens)
ik_cot (3,30278 tokens)	vis_seq (0,06458 tokens)
mpii_cot (0,05008 tokens)	co3d_seq (0,22888 tokens)
ICD_MDR_medium (0,00818 tokens)	poster (0,00858 tokens)



Training Loss (1 epoch) ~ Validation Loss



Larger Model, More Data, Better Downstreams.

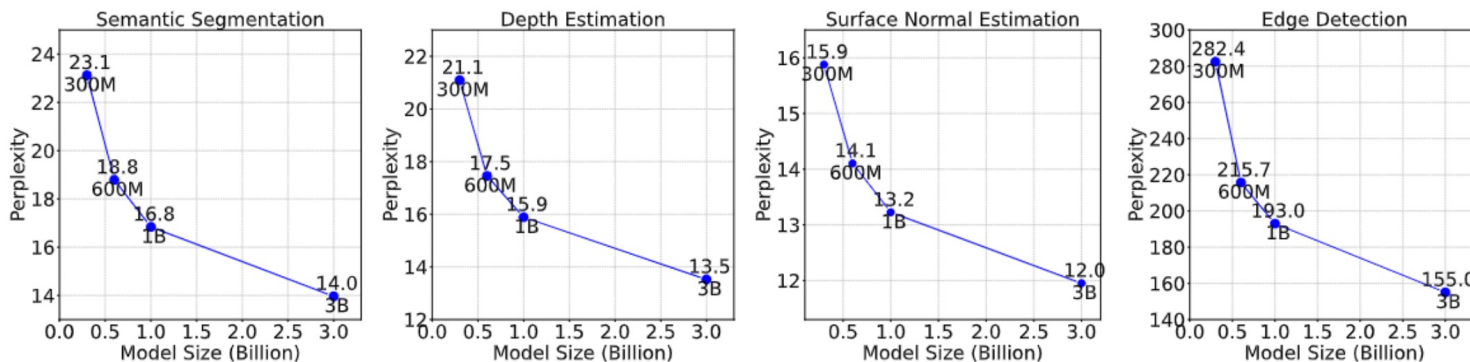


Figure 4. **Larger LVMs perform better on downstream tasks.** We evaluate LVM models of varying sizes on 4 different downstream tasks, following the 5 shot setting on the ImageNet validation set and report the perplexity. We find that perplexity decreases with larger models across all tasks, indicating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

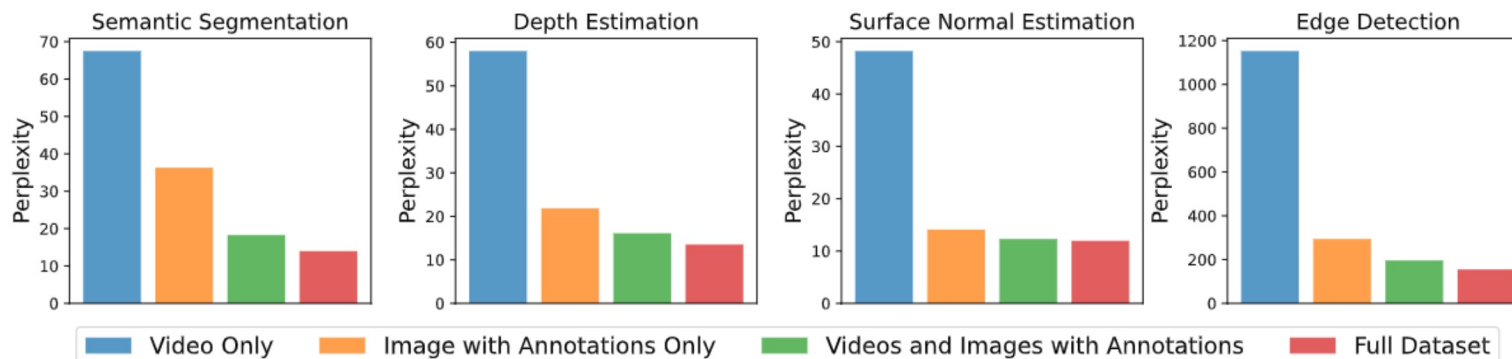
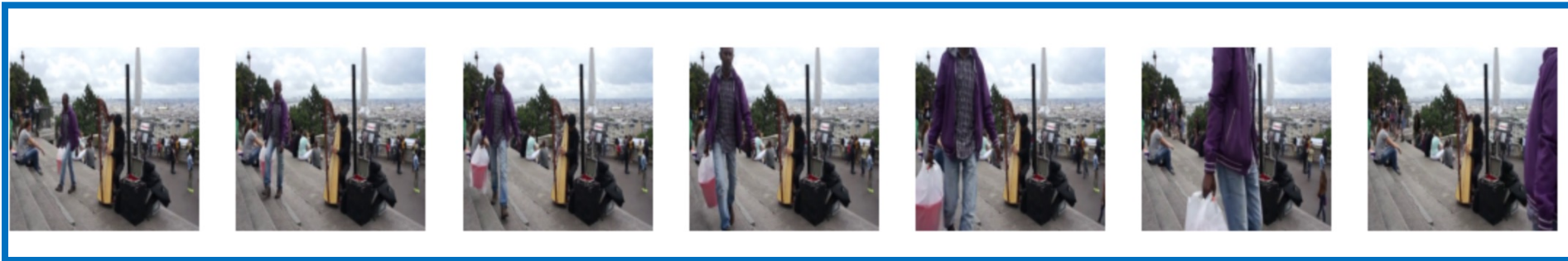


Figure 5. We evaluate the perplexity of 4 models trained on different sub-components of our datasets on tasks using the ImageNet validation set. All models are 3B parameters and all evaluations are conducted in the 5-shot setting. We can see that the model benefits from each of single images, videos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

Sequential Prompting

Prompts



Sequential Prompting

Prompts

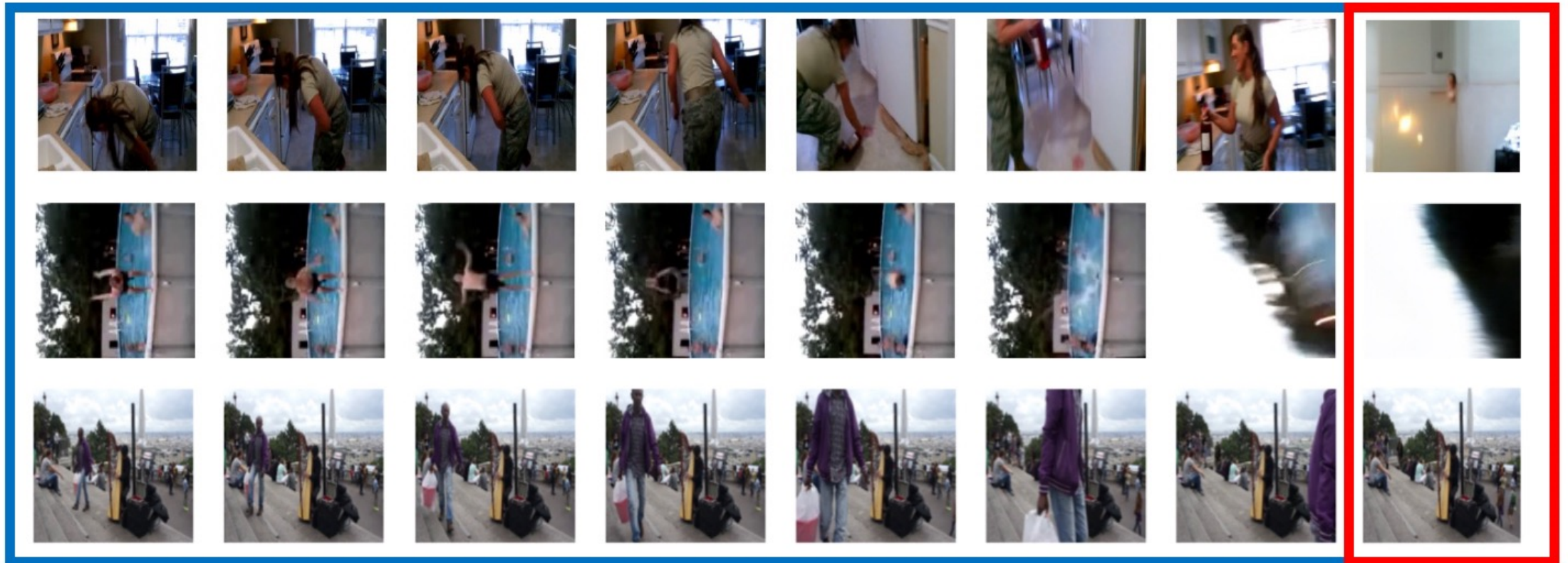
Generated



Sequential Prompting

Prompts

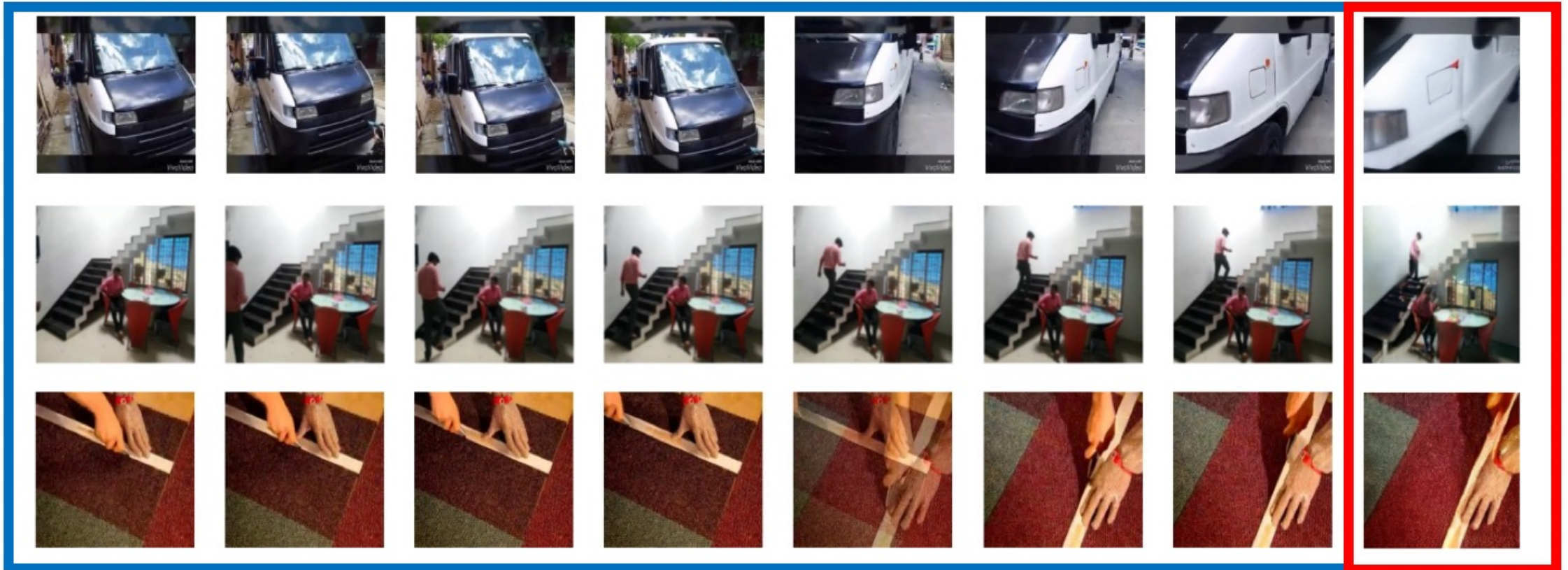
Generated



Sequential Prompting

Prompts

Generated



Longer Contexts



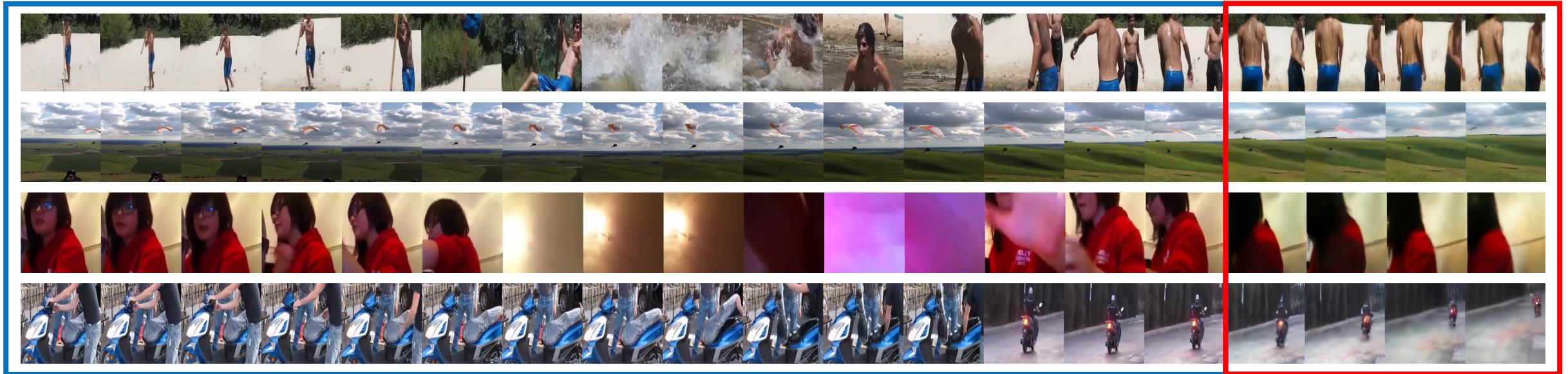
Generated



Longer Contexts

Prompts

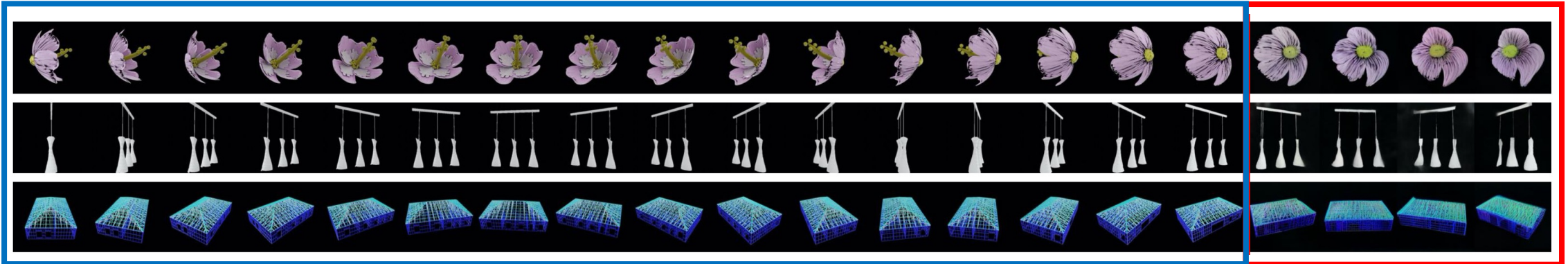
Generated



Sequential Prompting

Prompts

Generated



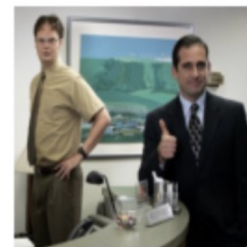
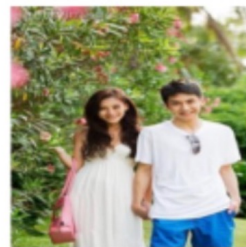
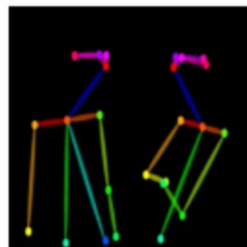
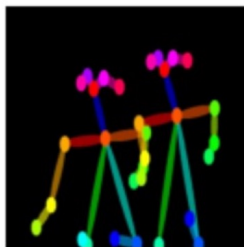
Sequential Prompting

Prompts

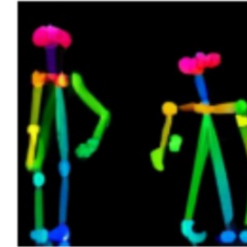
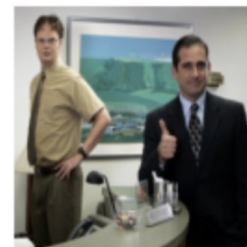
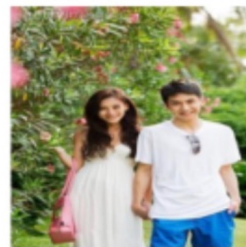
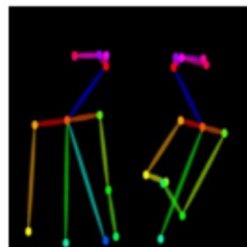
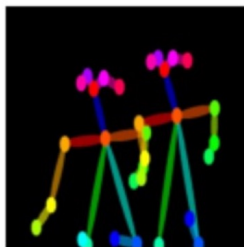
Generated



Analogy Prompting



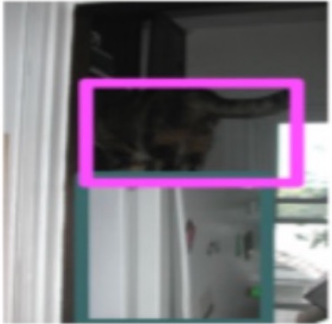
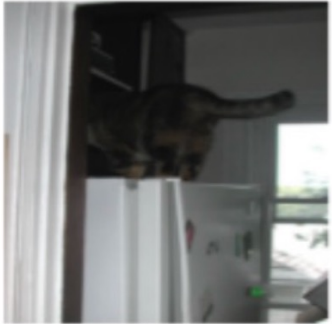
Analogy Prompting



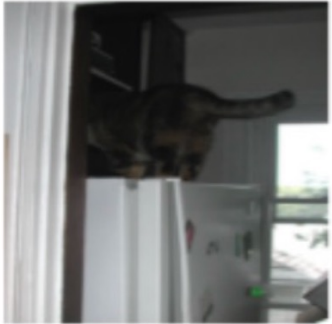
Analogy Prompting

	Inpaint MSE	Color MSE	Depth MSE	Surf MSE	Seg mIOU	KP Det PCKh	3DRot MSE	Denoise PSNR	Derain PSNR	LOL PSNR
<i>Bar et al</i> [6]	0.32	0.67	0.72	0.85	27.17	32.81	0.73	49.25	39.21	25.74
<i>Wang et al</i> [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

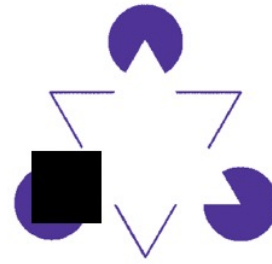
Analogy Prompting – Out of Domain Data



Analogy Prompting – Out of Domain Data

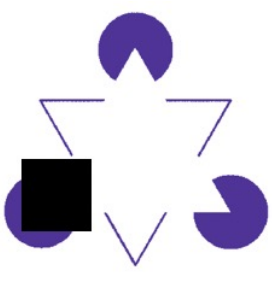


Analogy Prompting – Out of Domain Data



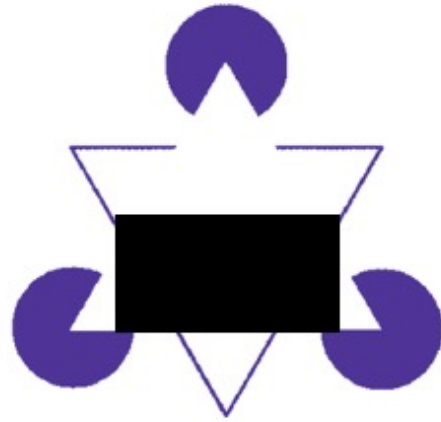
Analogy Prompting – Out of Domain Data

- corners

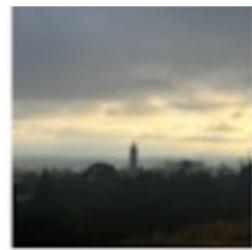
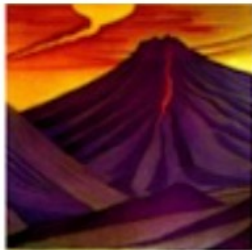


Analogy Prompting – Out of Domain Data

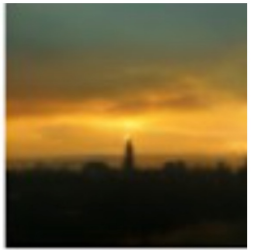
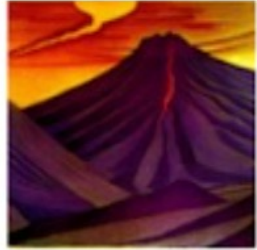
- edges



Analogy Prompting – Out of Domain Data



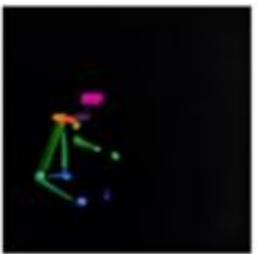
Analogy Prompting – Out of Domain Data



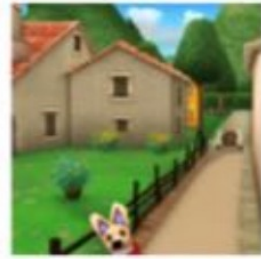
Analogy Prompting – Out of Domain Data



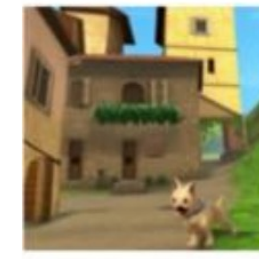
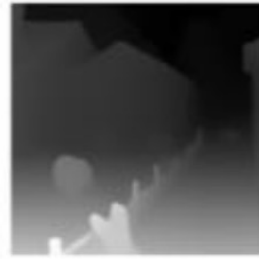
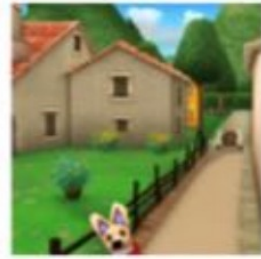
Analogy Prompting – Out of Domain Data



Analogy Prompting – Out of Domain Data



Analogy Prompting – Out of Domain Data



Compositional Prompts



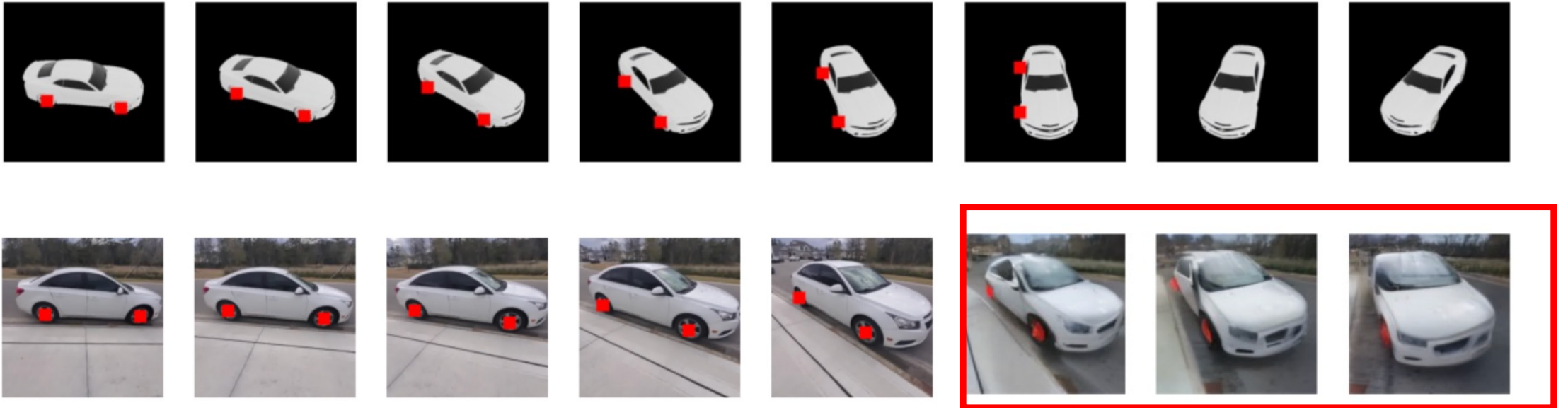
More complicated



More complicated



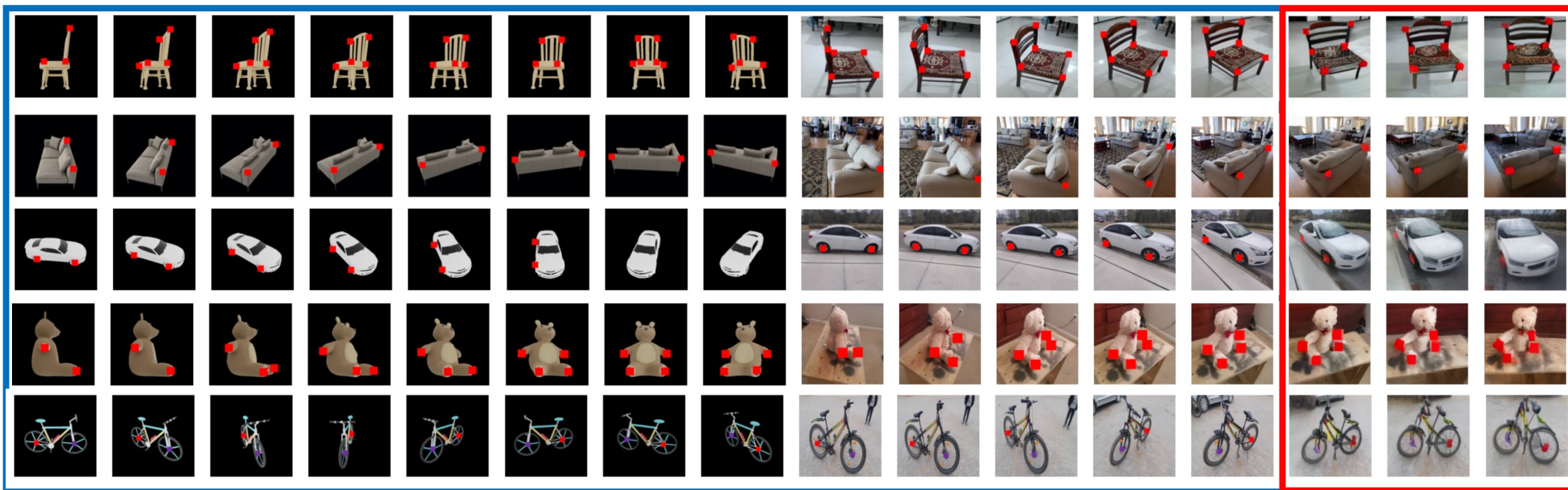
More complicated



Compositional Prompts

Prompts

Generated



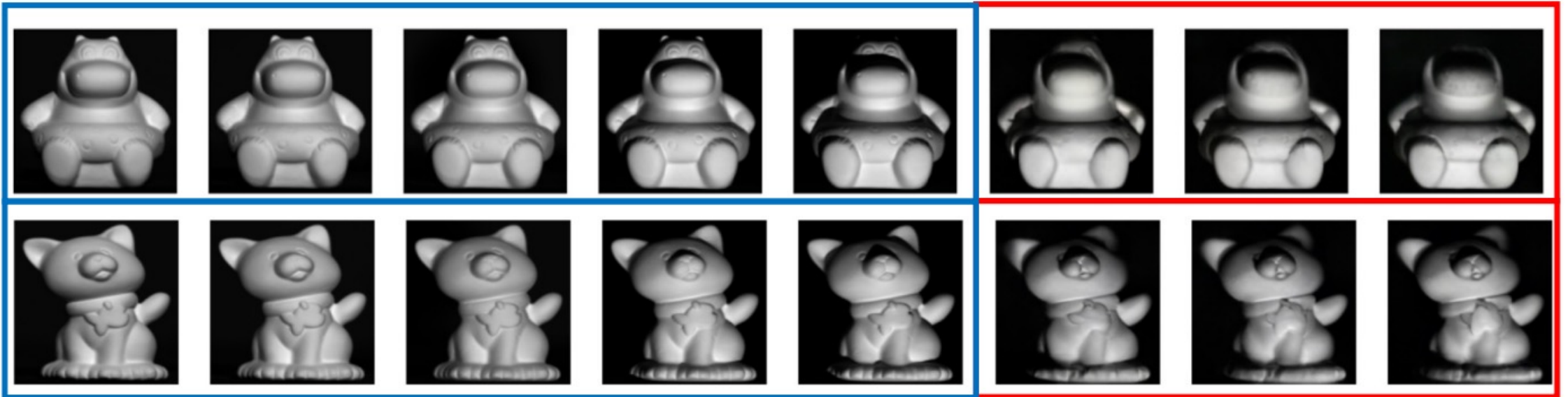
Unseen tasks

Prompts

Generated



Unseen tasks



Unseen tasks



Not easily describable



Not easily describable



Not easily describable



Not easily describable



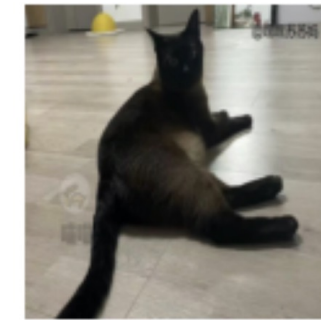
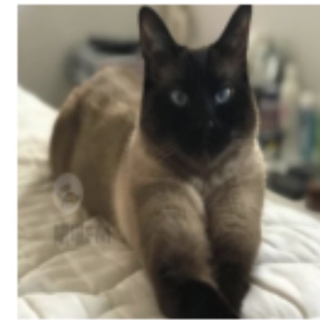
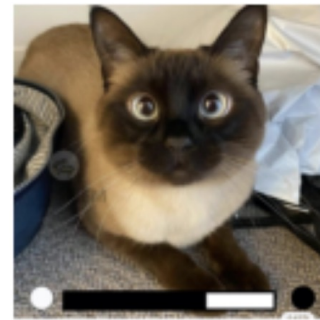
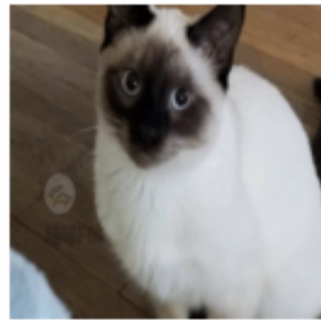
Not easily describable



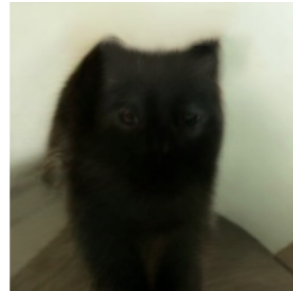
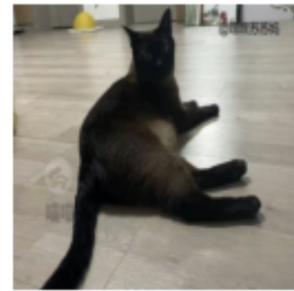
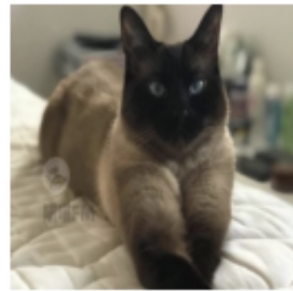
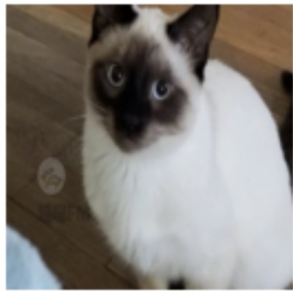
Not easily describable

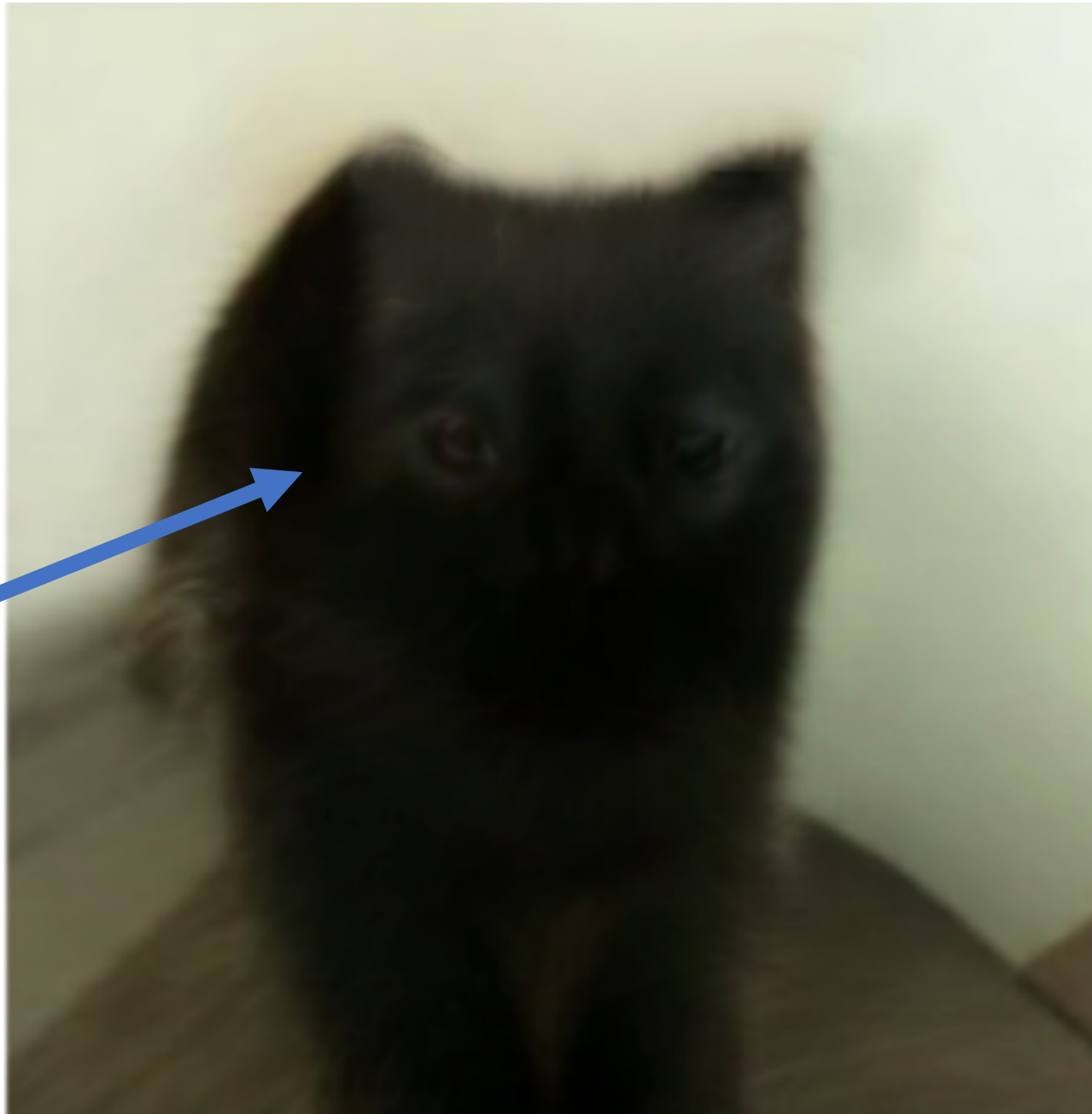
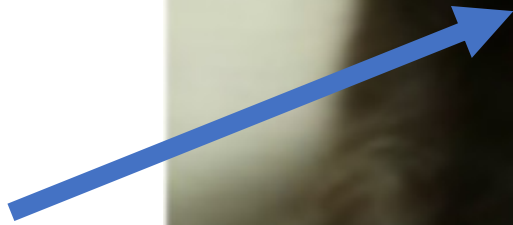


Not easily describable

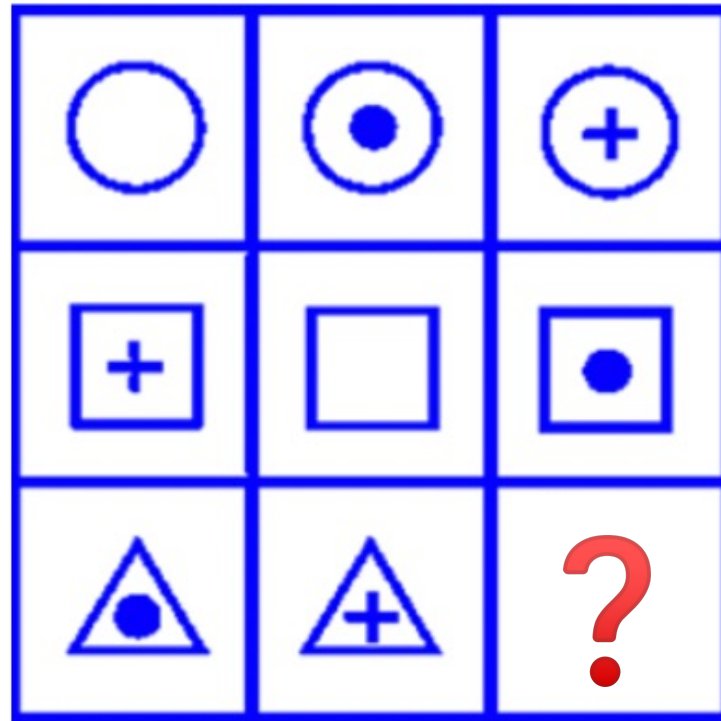


Not easily describable



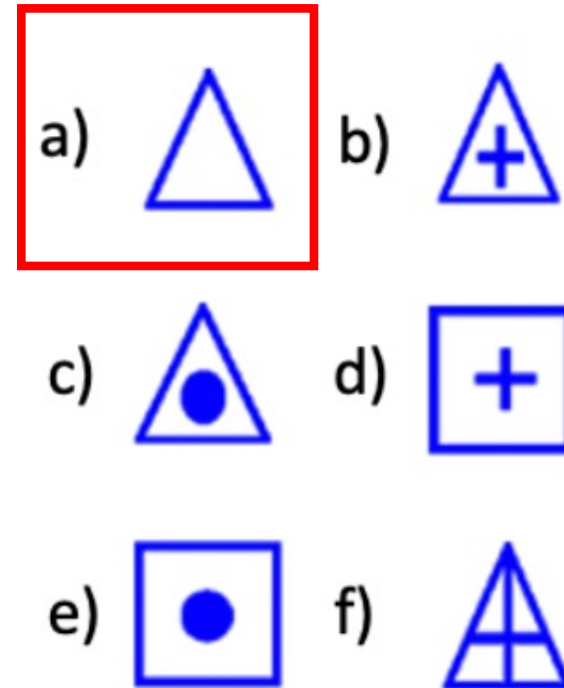
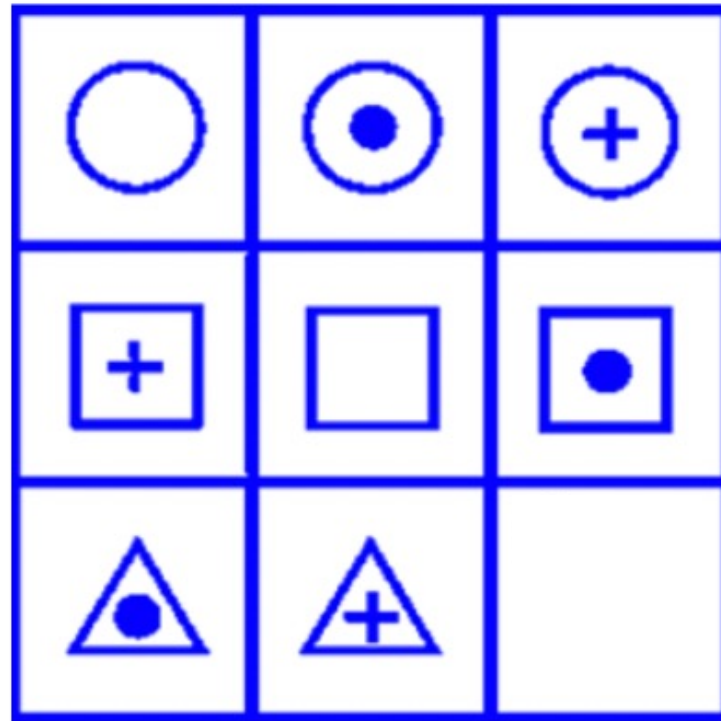


Raven's Progressive Test (Non-verbal IQ test)

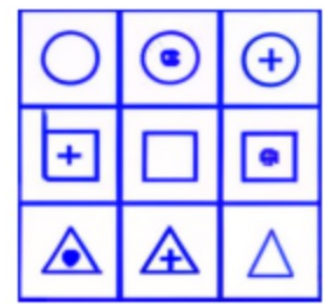
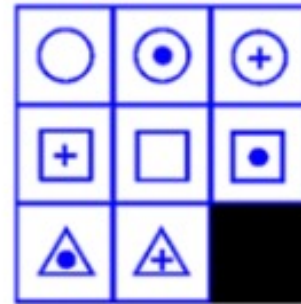
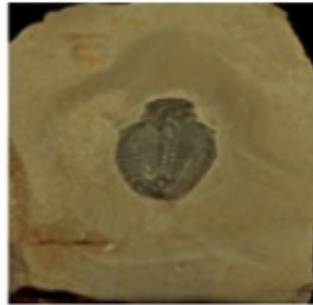
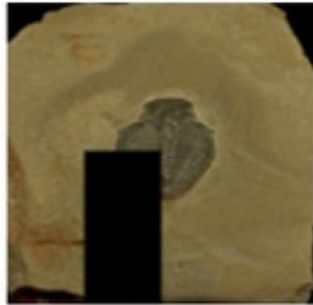


- a)
- b)
- c)
- d)
- e)
- f)

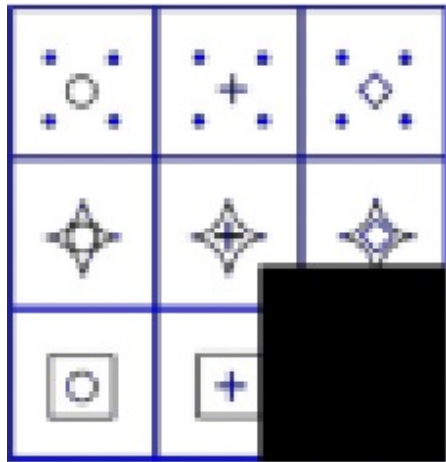
Raven's Progressive Test (Non-verbal IQ test)



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

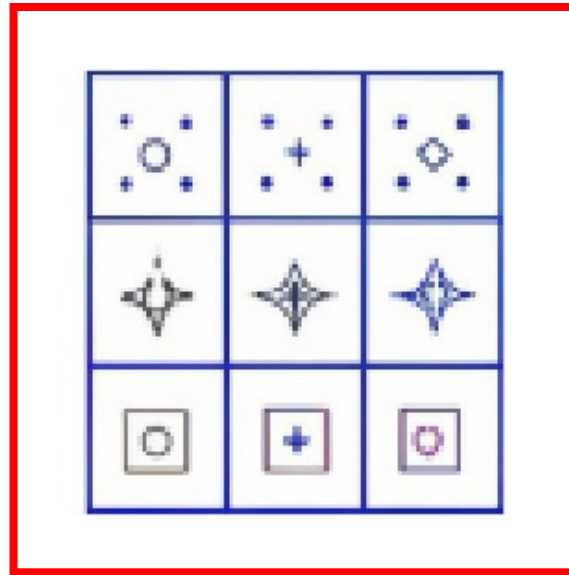
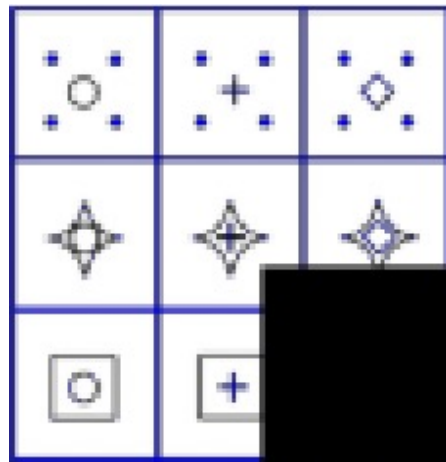


More Difficult?



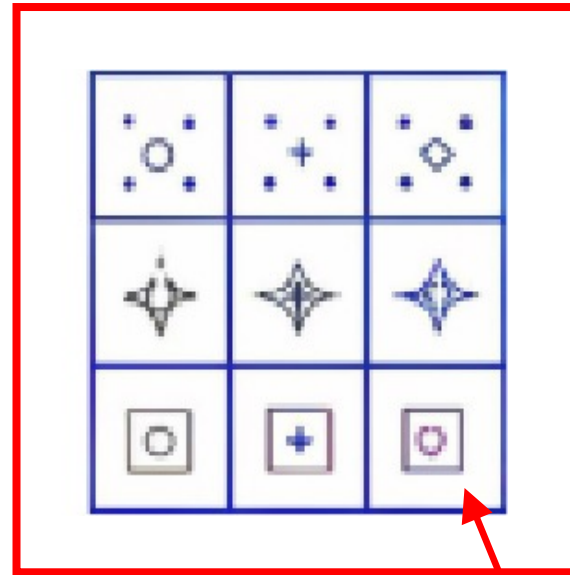
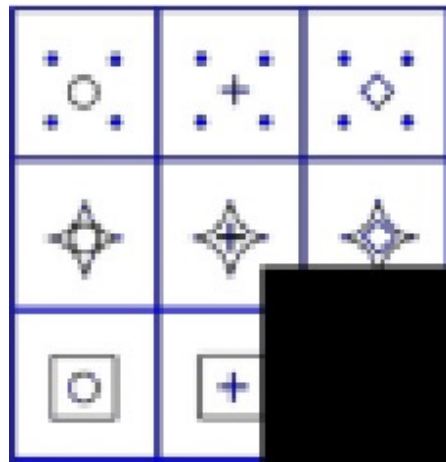
More Difficult?

Generated



More Difficult?

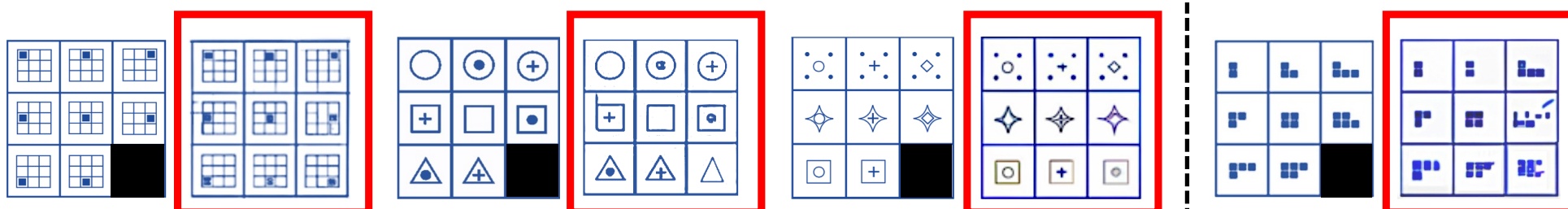
Generated



Hard to tell if correct or not 🤔

Perplexity

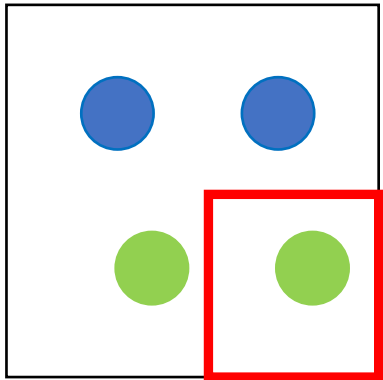
- 10 Questions:
 - performed perplexity analysis on classic Raven 5-way multiple-choice Matrices, choosing the answer with lowest perplexity.



	Raven's Progressive Matrices
Chance	20%
Ours	30%

Synthetic Reasoning

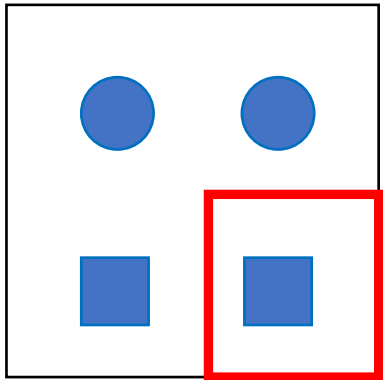
- **Color Change:** choose from 3 random generated colors.



	color
Chance	33%
Ours	42%

Synthetic Reasoning

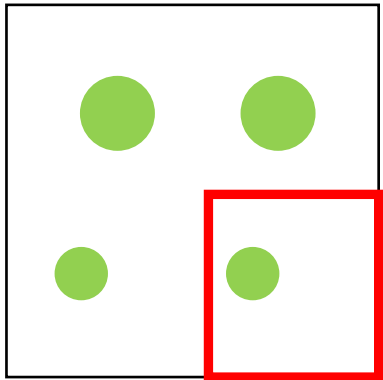
- **Shape Change:** choose from 3 random generated shapes.



	shape
Chance	33%
Ours	45%

Synthetic Reasoning

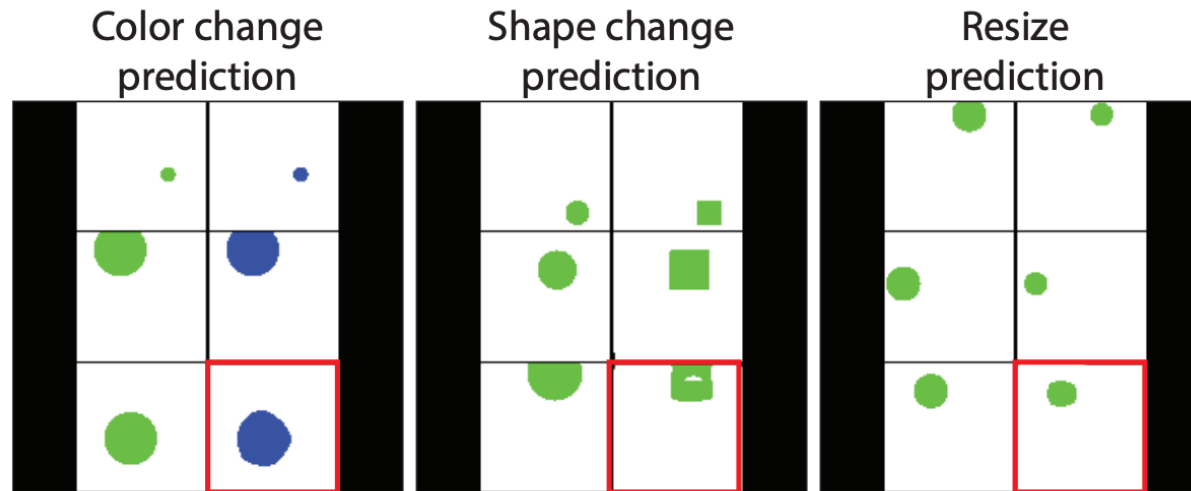
- **Size Change:** choose from 2 random generated sizes. (resolution)



	size
Chance	50%
Ours	94%

Synthetic Reasoning

- In total 900 experiments

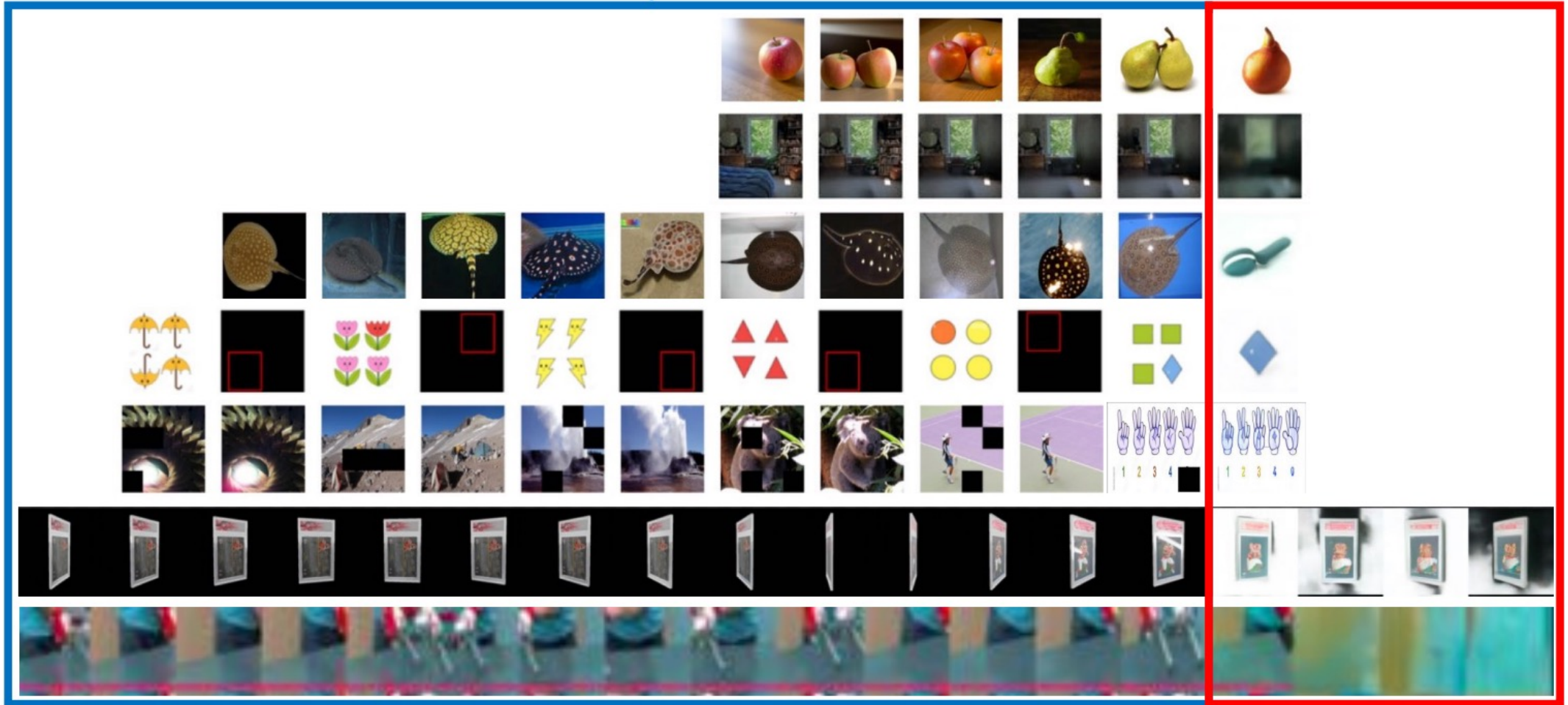


	color	shape	size
Chance	33%	33%	50%
Ours	42%	45%	94%

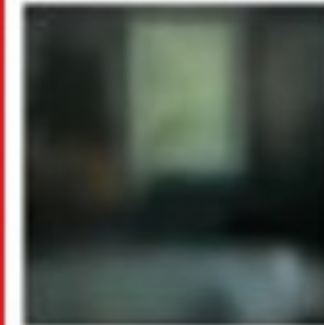
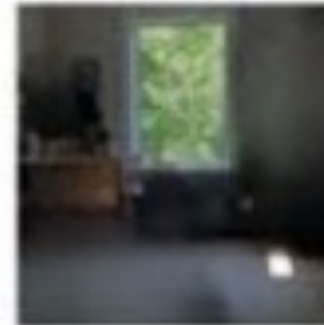
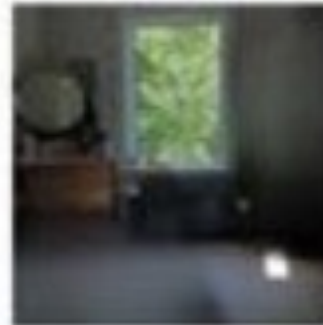
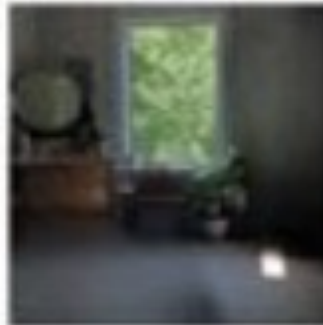
Failure Case

Prompts

Generated



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**

