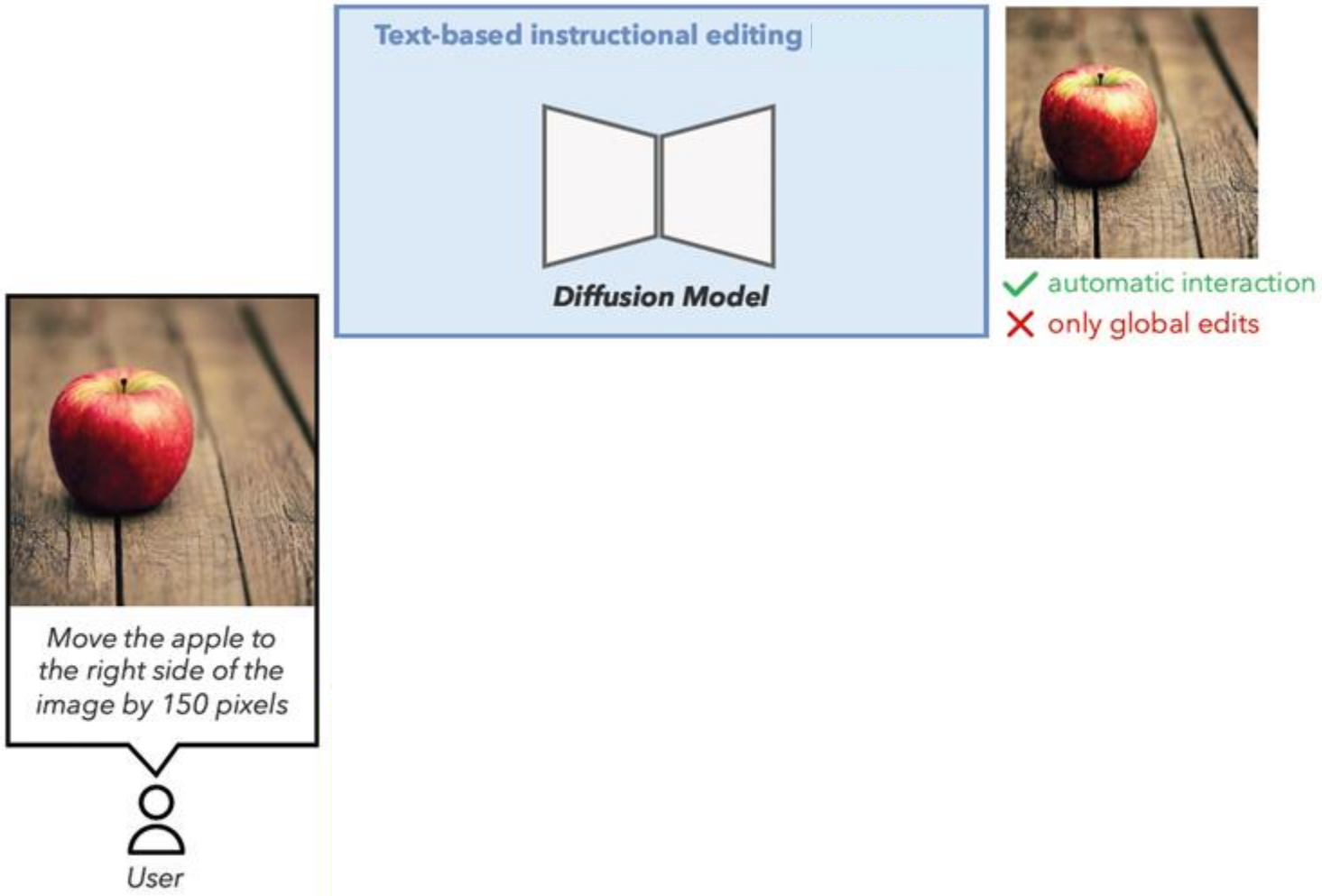
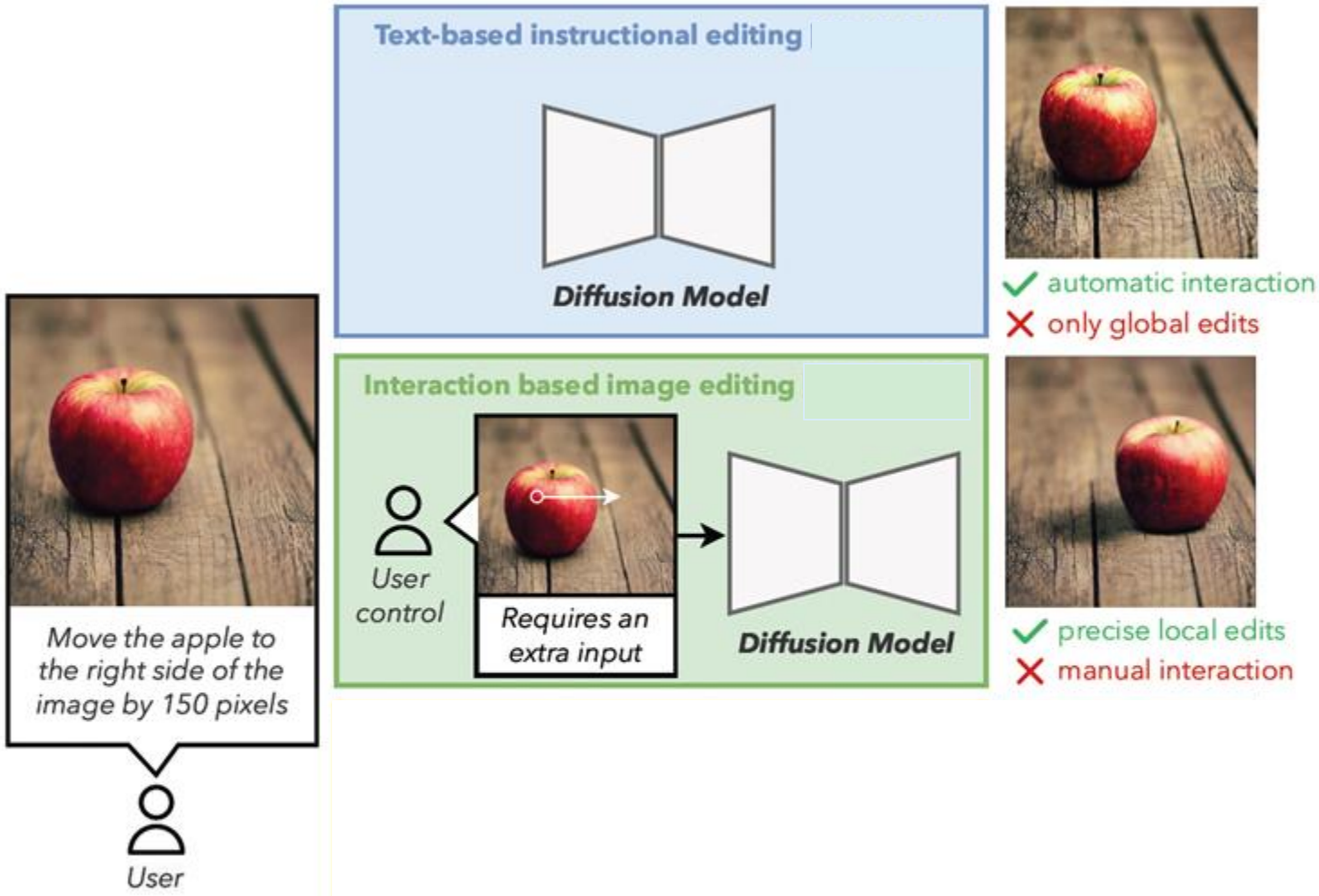
POEM: Motivation

Goal: image editing pipeline, with localized and precise object edit



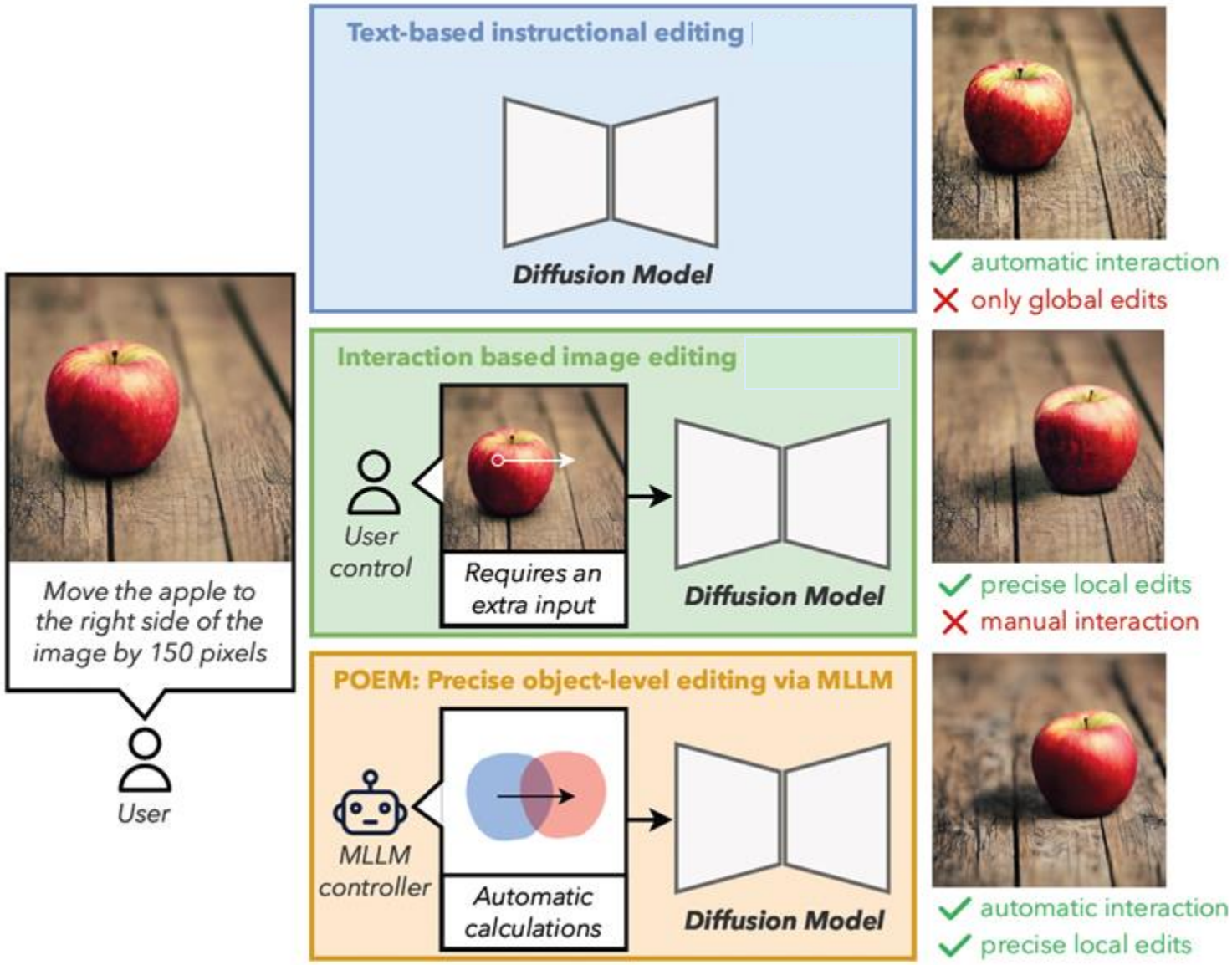
POEM: Motivation

Goal: image editing pipeline, with localized and precise object edit

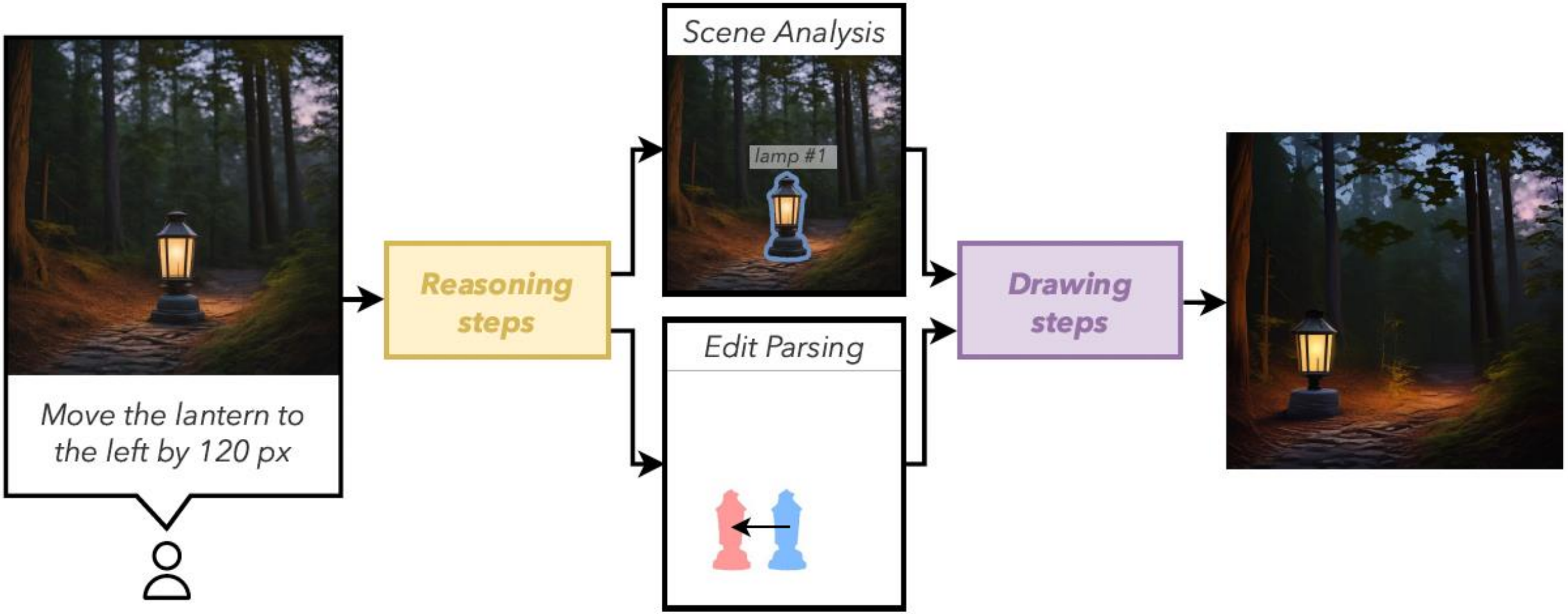


POEM: Motivation

Goal: image editing pipeline, with localized and precise object edit

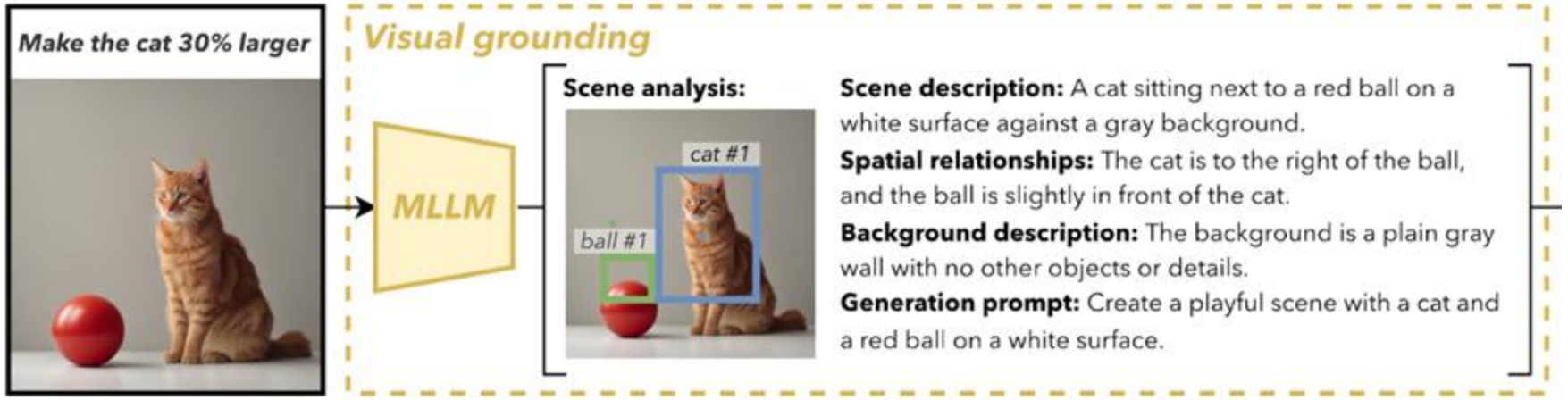


POEM: Pipeline



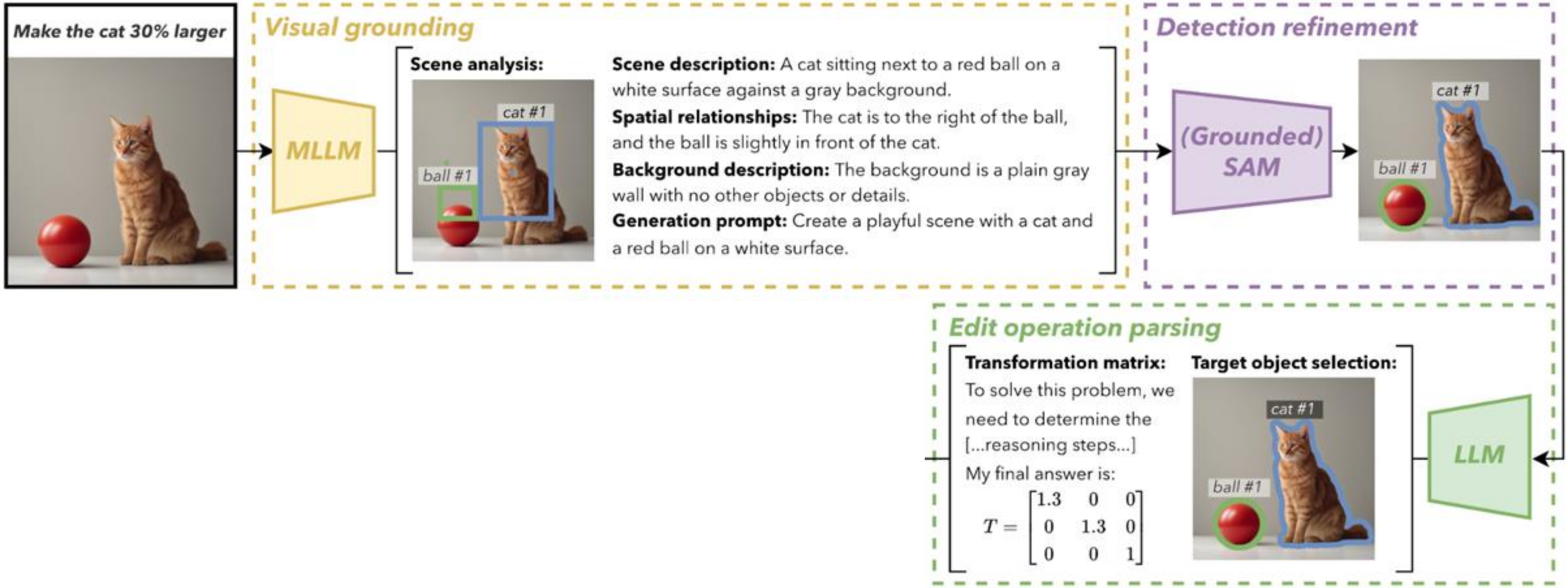
POEM: Detailed Pipeline

POEM: Detailed Pipeline

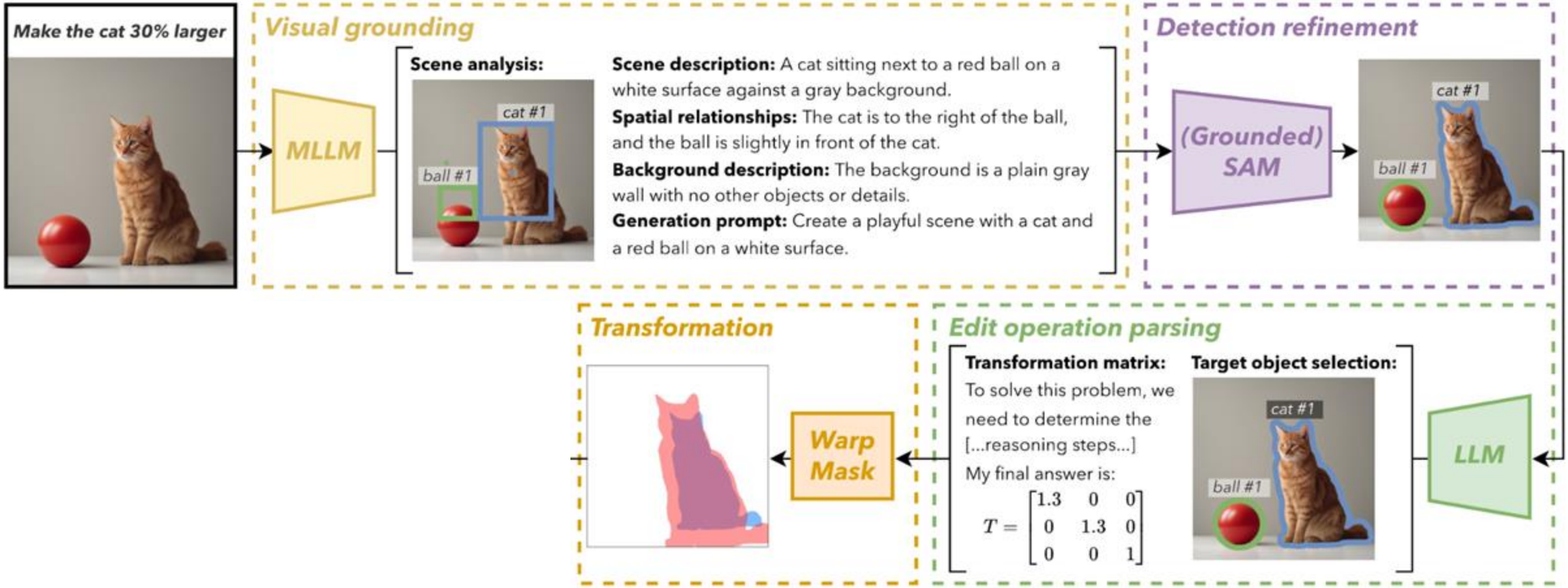


POEM: Detailed Pipeline

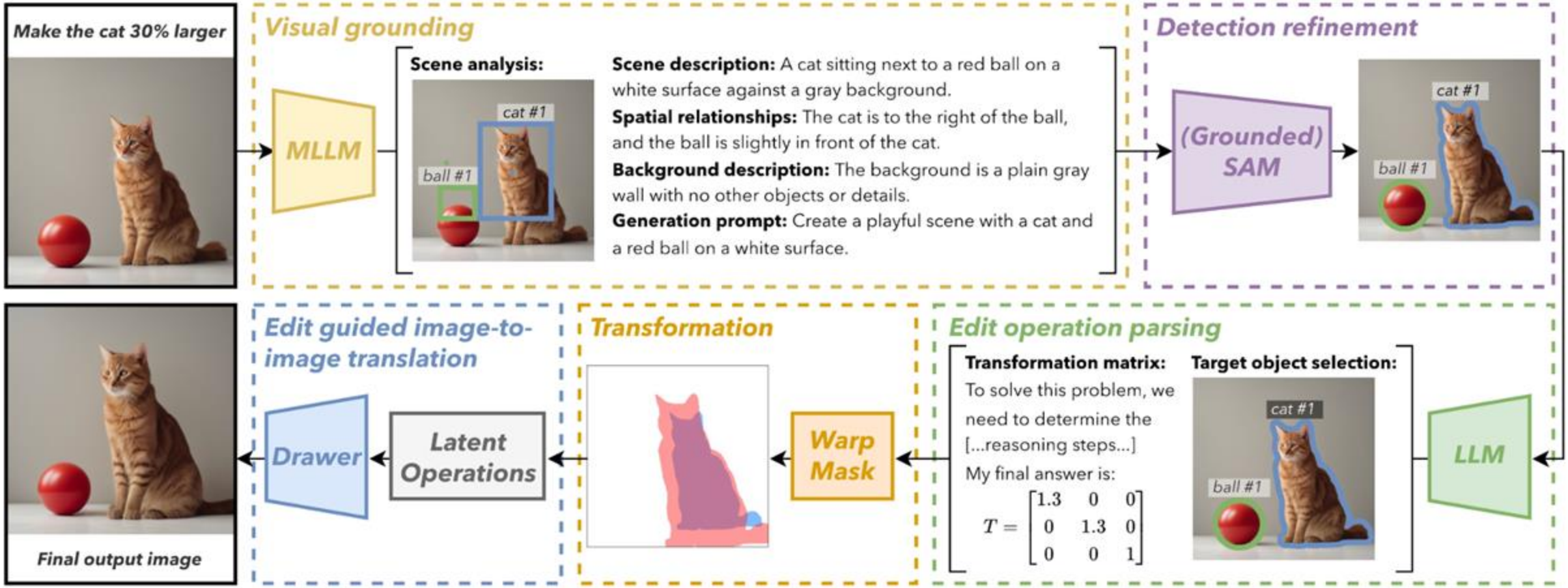
POEM: Detailed Pipeline



POEM: Detailed Pipeline



POEM: Detailed Pipeline



POEM: qualitative results

Scale the bus by 0.56

LEDITS++ [Brack, CVPR 2024]

Ours

POEM: qualitative results

Move the pear left by 150px, and make it red

IP2P [Brooks, CVPR 2023]

Ours

Scaling Test-Time Compute

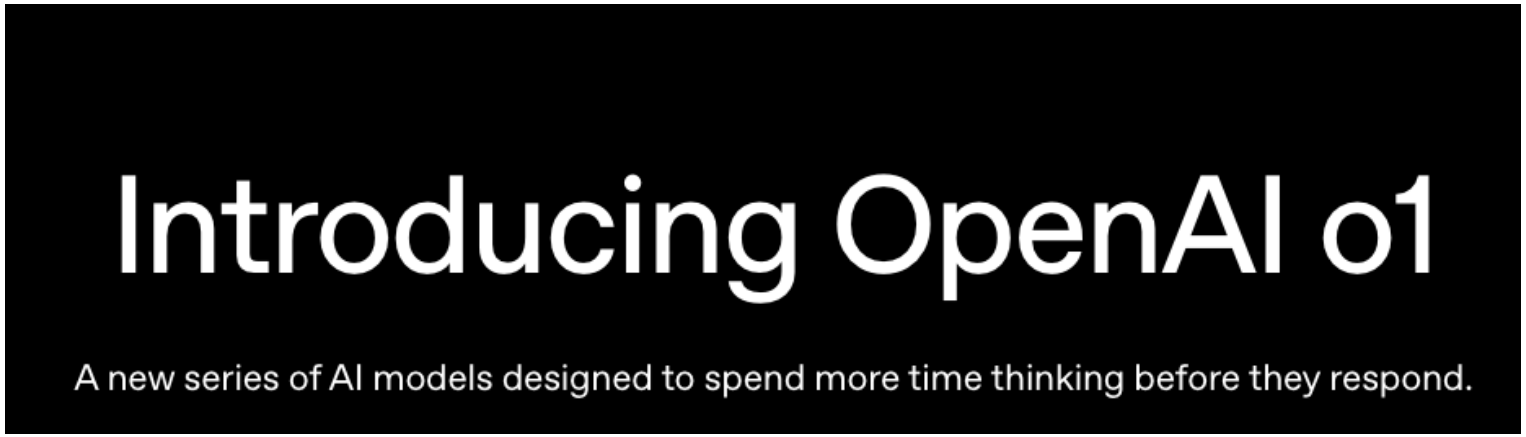
Test Time Scaling for Image Generation

Erik Wold Riise, Mehmet Onurcan Kaya, **Dim P. Papadopoulos**

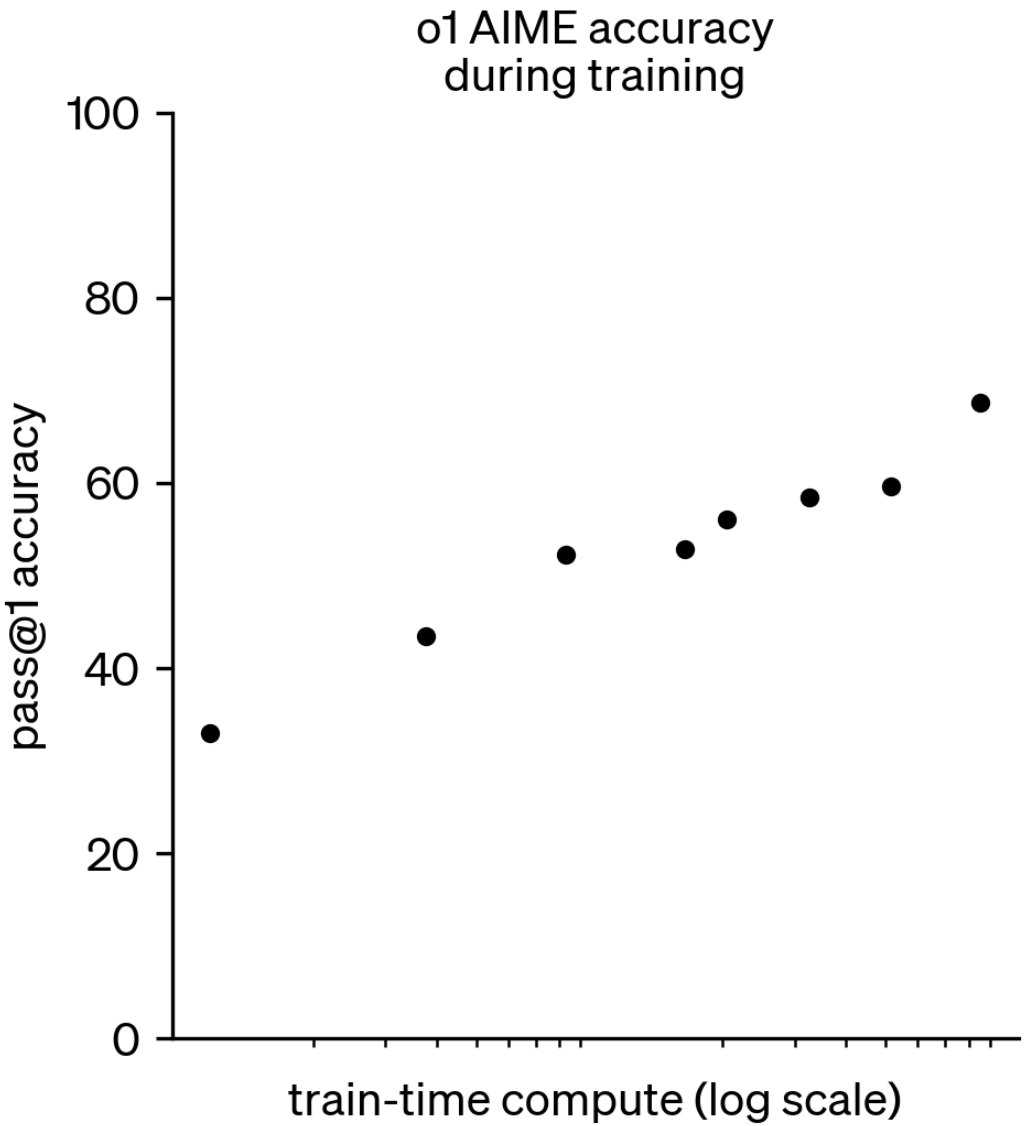
[Work in progress]

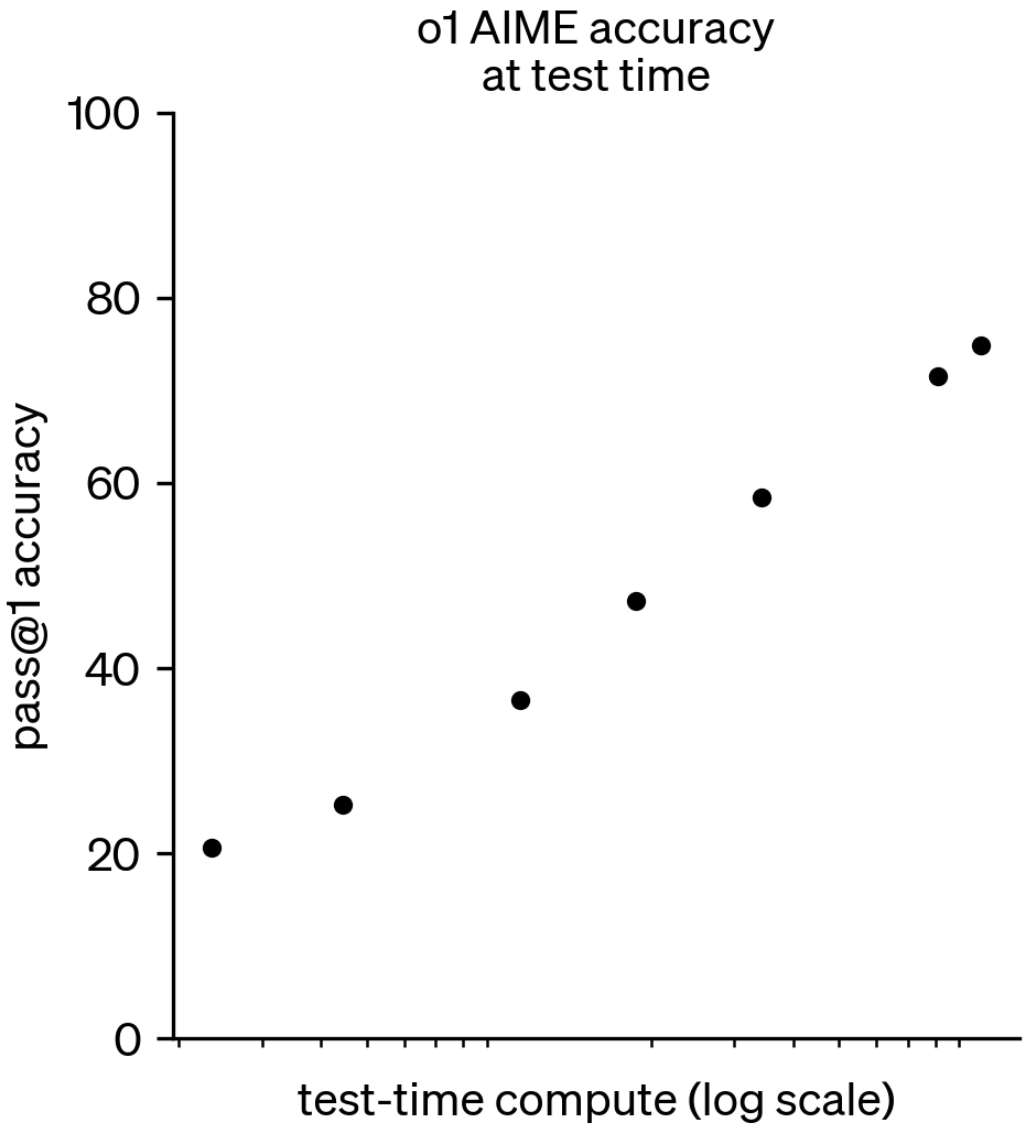
Test-time Scaling

Rather than relying on ever-larger pretraining budgets, test-time methods use dynamic inference strategies that allow models to "think longer" on harder problems

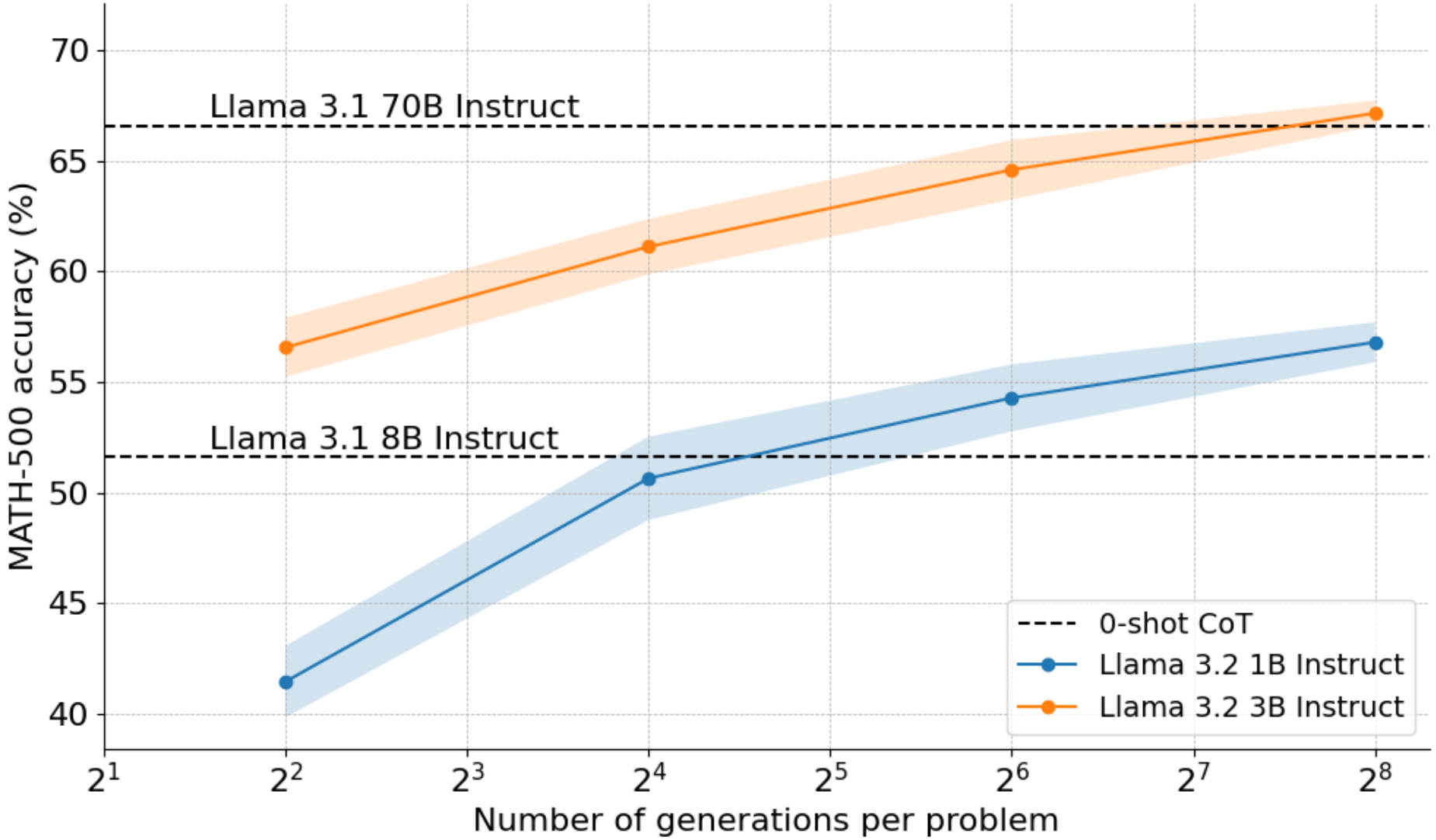


Test-time Scaling (LLMs)

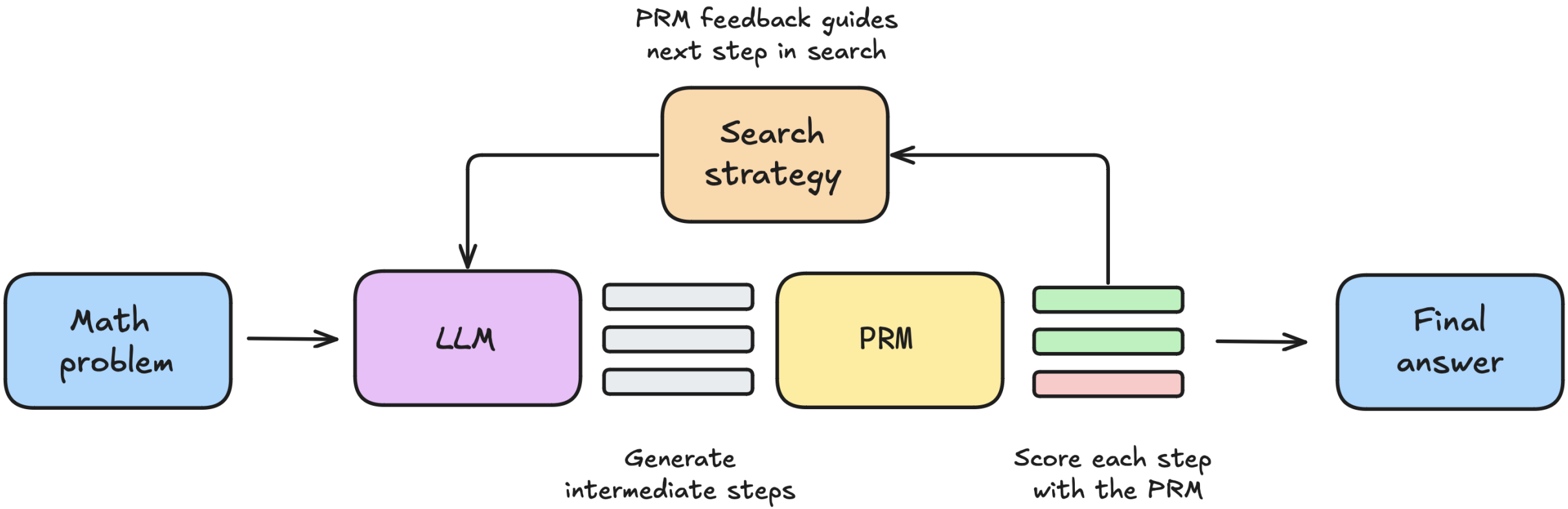




Test-time Scaling (LLMs)



Test-time Scaling (LLMs)



Requires an extrernal verifier!!

Test-time Scaling (image generation)

Scaling Inference Time Compute for Diffusion Models

Nanye Ma[1,3,*] Shangyuan Tong[2,3,*] Haolin Jia[3] Hexiang Hu[3] Yu-Chuan Su[3]
Mingda Zhang[3] Xuan Yang[3] Yandong Li[3] Tommi Jaakkola[2] Xuhui Jia[3] Saining Xie[1,3]
[1]New York University [2]MIT [3]Google

Abstract

Generative models have made significant impacts across various domains, largely due to their ability to scale during training by increasing data, computational resources, and model size, a phenomenon characterized by the scaling laws. Recent research has begun to explore inference-time scaling behavior in Large Language Models (LLMs), revealing how performance can further improve with additional computation during inference. Unlike LLMs, diffusion models inherently possess the flexibility to adjust inference-time computation via the number of denoising steps, although the performance gains typically flatten after a few dozen. In this work, we explore the inference-time scaling behavior of diffusion models beyond increasing denoising steps and investigate how the generation performance can further improve with increased computation. Specifically, we consider a search problem aimed at identifying better noises for the diffusion sampling process. We structure the design space along two axes: the verifiers used to provide feedback, and the algorithms used to find better

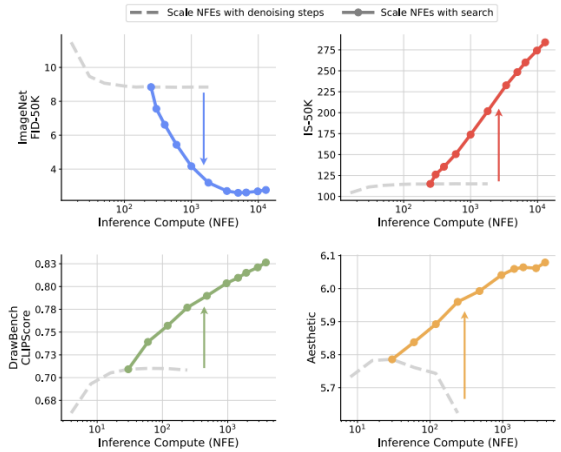


Figure 1. **Inference scaling beyond increasing denoising steps.** We demonstrate the performance with respect to FID ↓, IS ↑ on ImageNet, and CLIPScore ↑, Aesthetic Score ↑ on DrawBench. Our search framework exhibits substantial improvements in all settings over purely scaling NFEs with increasing denoising steps.

FLUX-1.dev [38]

"A laptop on top of a teddy bear."

"Photo of an athlete cat explaining it's latest scandal at a press conference to journalists."

[Deepmind, CVPR 2025]

Test-time Scaling (image generation)

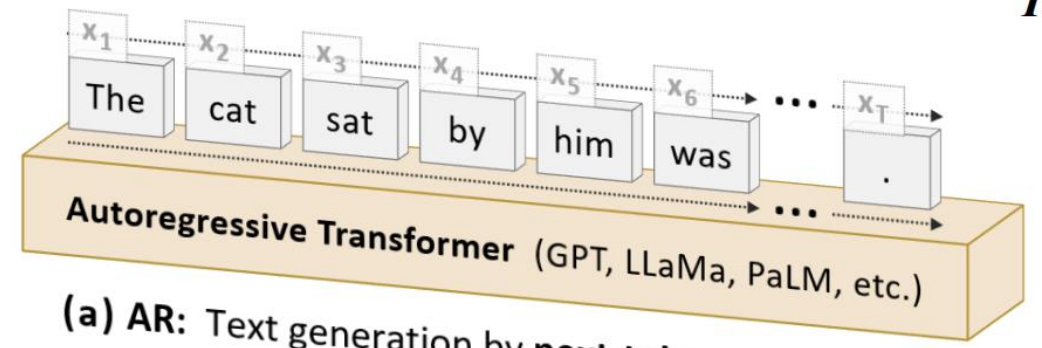
Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction

Keyu Tian[1,2], Yi Jiang[2,†], Zehuan Yuan[2,*], Bingyue Peng[2], Liwei Wang[1,3,*]
[1]Center for Data Science, Peking University [2]Bytedance Inc.
[3]State Key Lab of General Artificial Intelligence, School of Intelligence Science and Technology, Peking University
keyutian@stu.pku.edu.cn, jiangyi.enjoy@bytedance.com, yuanzehuan@bytedance.com, bingyue.peng@bytedance.com, wanglw@pku.edu.cn

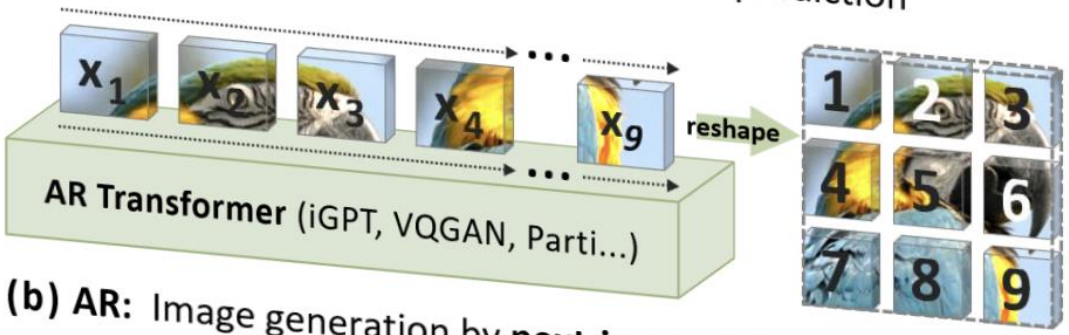
Try and explore our online demo at: https://var.vision
Codes and models: https://github.com/FoundationVision/VAR

Figure 1: **Generated samples from Visual AutoRegressive (VAR) transformers trained on ImageNet**. We show 512×512 samples (top), 256×256 samples (middle), and zero-shot image editing results (bottom).

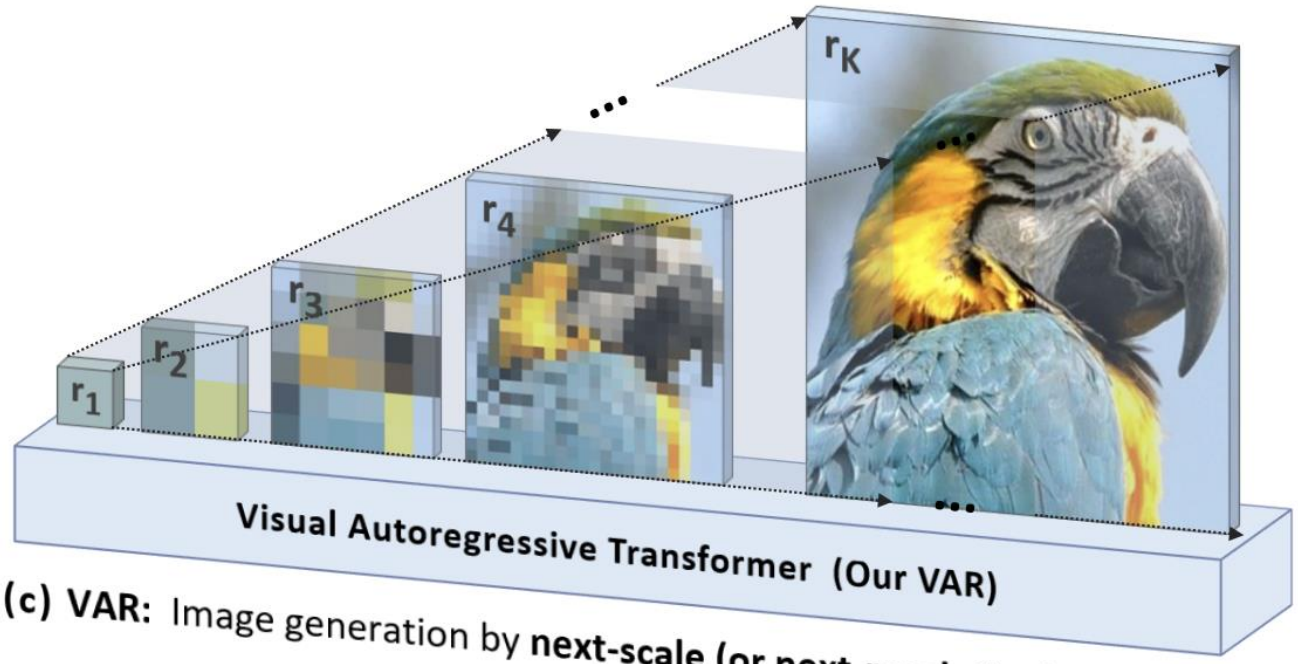
Three Different Autoregressive Generative Models



(a) AR: Text generation by **next-token** prediction



(b) AR: Image generation by **next-image-token** prediction



(c) VAR: Image generation by **next-scale (or next-resolution)** prediction

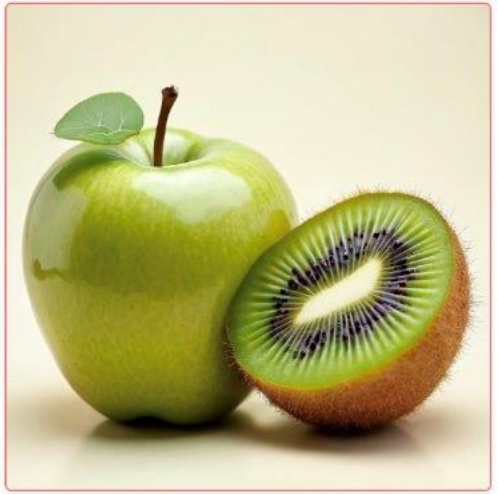
[Deepmind, CVPR 2025]

Qualitative Results

"bird next to refrigerator"

"giraffe right of wallet"

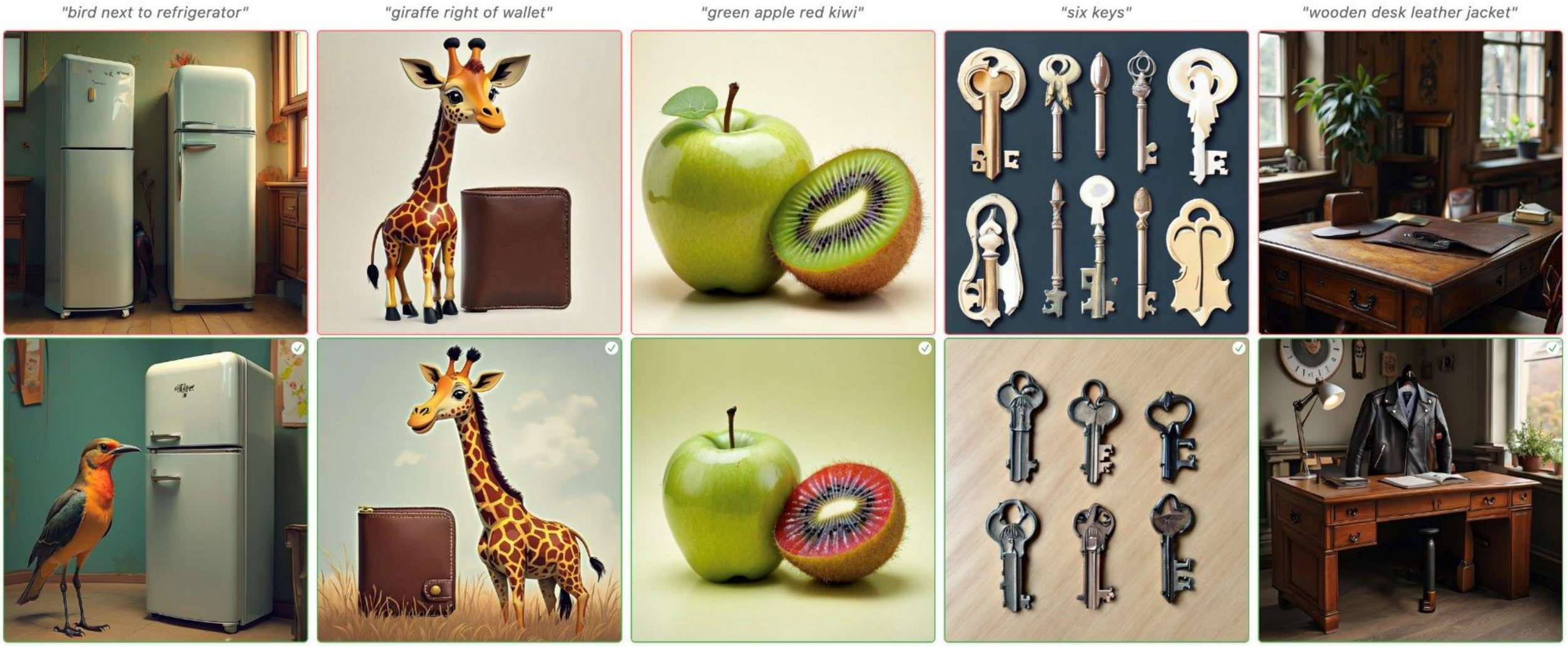
"green apple red kiwi"



"six keys"

"wooden desk leather jacket"

Qualitative Results



Quantitative Results (DrawBench)

Method	Params	NFEs	Images	Aesthetic	CLIPScore	ImageReward
FLUX.1-dev Baseline	12B	30	1	5.79	0.71	0.97
FLUX.1-dev + Search (Best)	12B	2880	96	6.38	0.82	1.58
Ours (2B) Baseline	2B	13	1	6.06	0.71	0.94
Ours (2B) + Search (Best)	2B	1365	195	**7.38**	**0.83**	**1.59**

The Infinty-2B model surpasses the 12B model's best result across all metrics, while using less than half the computational search budget (1365 vs 2880 NFEs).

Quantitative Results (Compositional tasks)

Model & Method	Color	Shape	Texture	Spatial	Numeracy	Complex
FLUX.1-dev (12B) + Search	0.8204 (+5.12)	0.5959 (+7.72)	0.7197 (+9.10)	0.3043 (+6.14)	0.6623 (+4.56)	0.3754 (+1.54)
Ours (2B) + Search	**0.8327** (+8.21)	**0.6389** (+17.38)	**0.7603** (+14.57)	**0.3601** (+10.45)	**0.6727** (+13.41)	**0.4245** (+3.85)

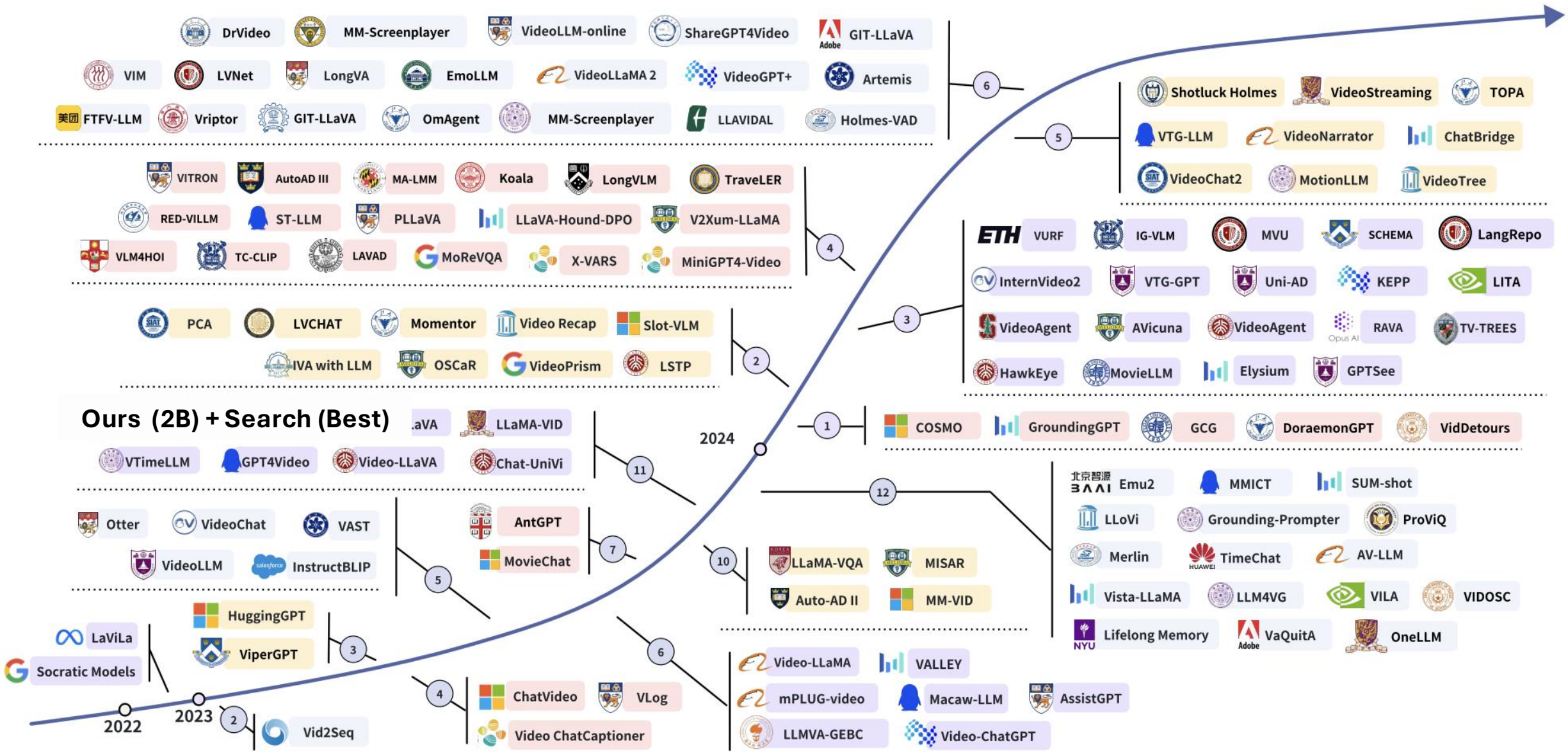
Despite its 6x smaller size, my autoregressive approach achieves **2x larger improvements** on average and a **higher absolute score in every single category**.

Test Time Augmentation for MLLMs

Mehmet Onurcan Kaya, Desmond Elliott, **Dim P. Papadopoulos**

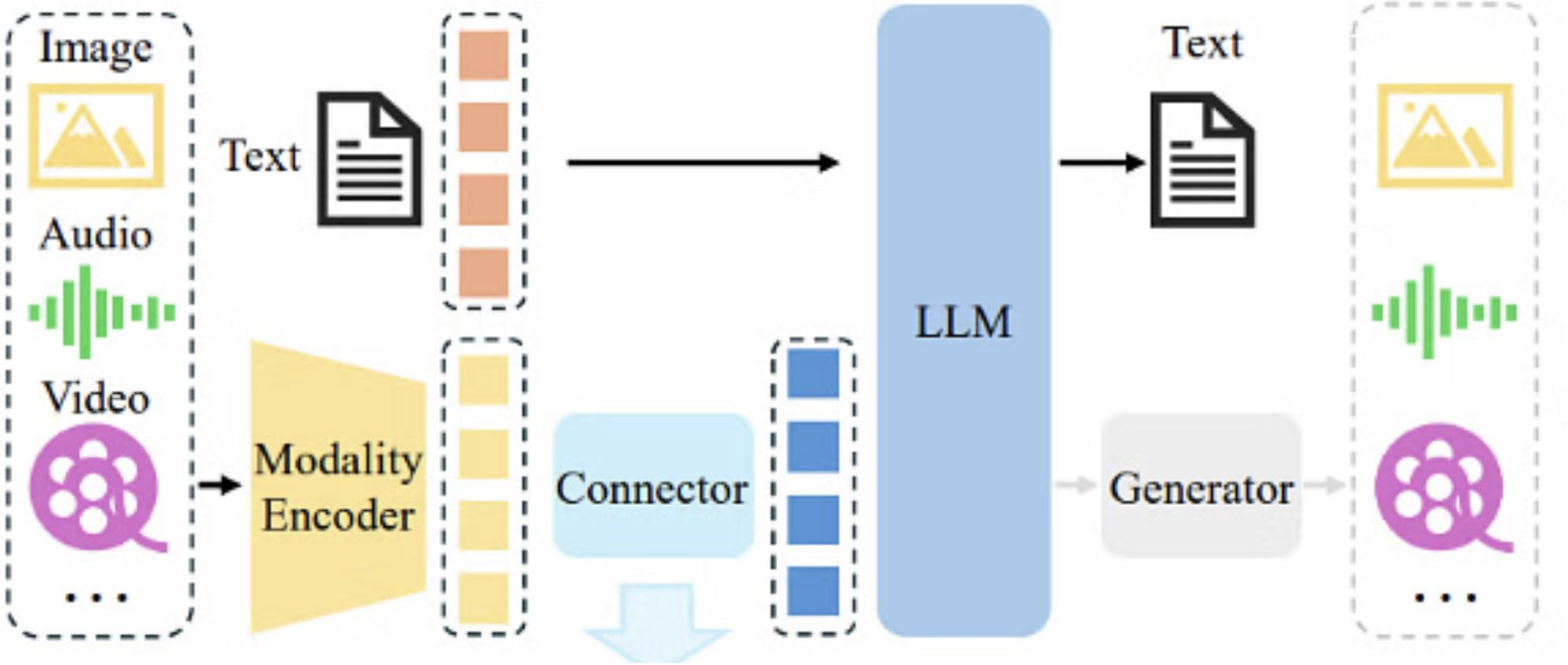
[Work in progress]

MLLMs (VLMs)

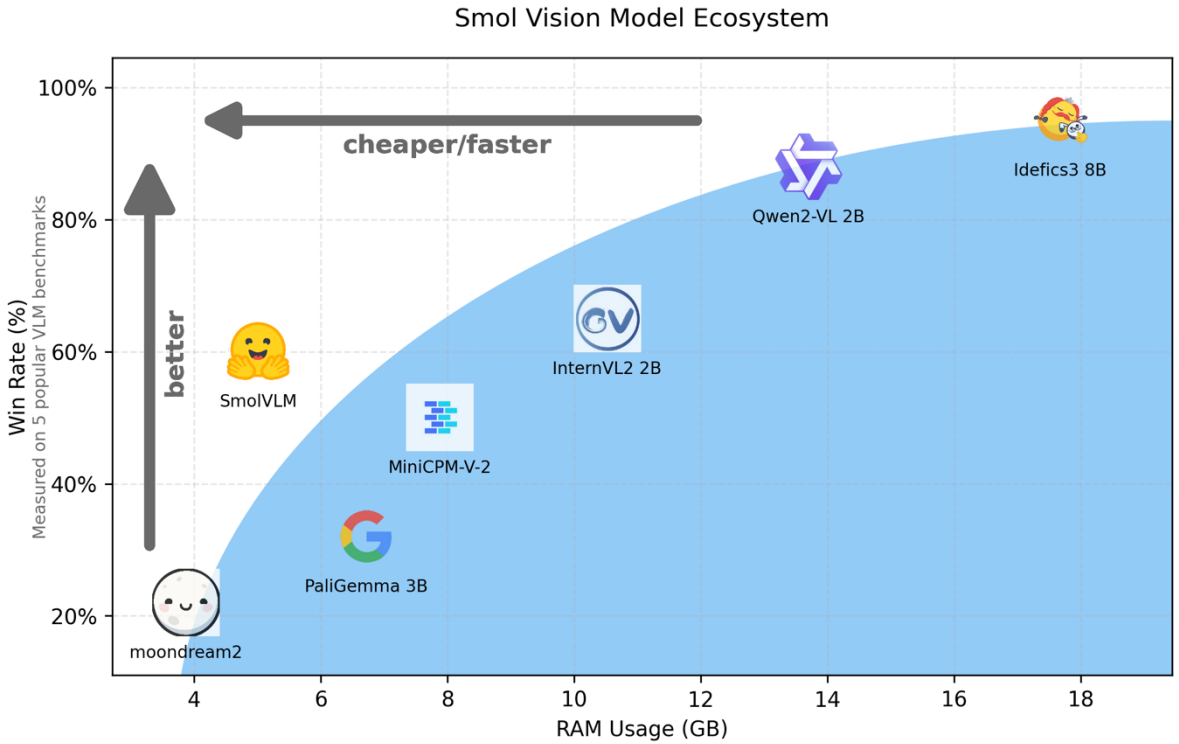


Survey paper: https://arxiv.org/pdf/2312.17432

Leveraging LLMs

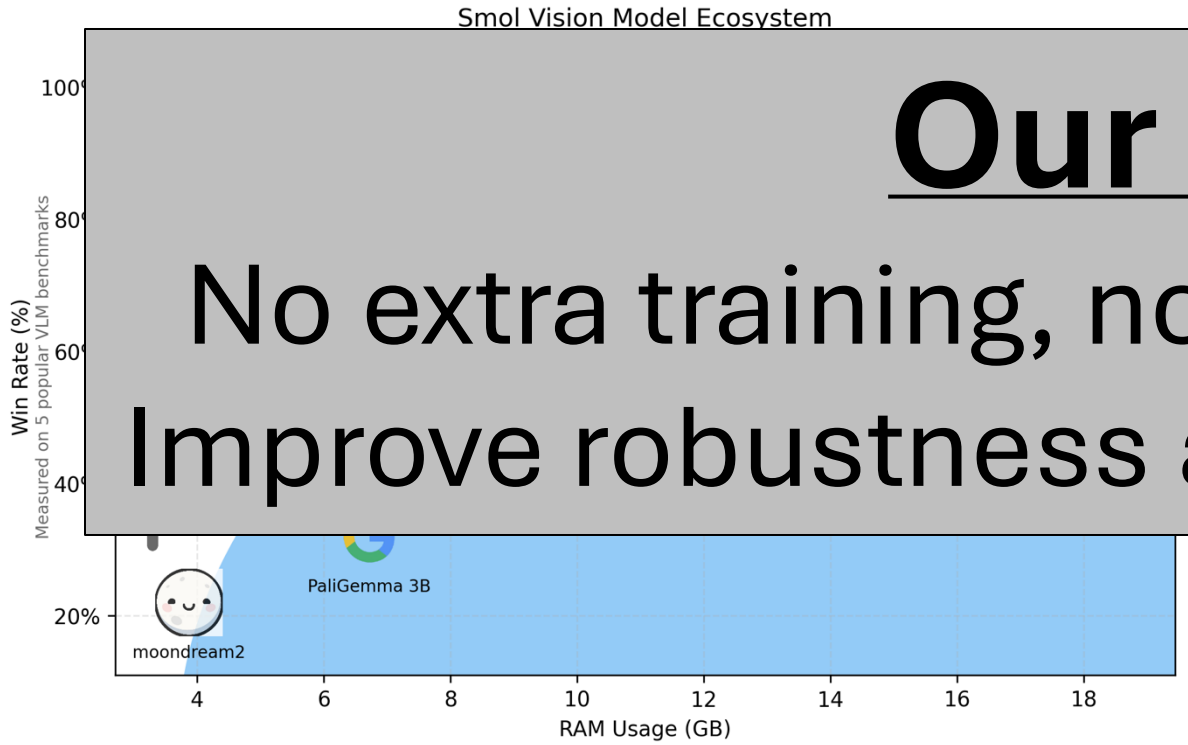


MSLMs (Multimodal Small Language Models)



- Efficient but performance **drops under domain shift**
- Existing test-time compute **require extra models**
- **Test-Time Augmentation** is underexplored in multimodal settings!!

MSLMs (Multimodal Small Language Models)



- Efficient but performance **drops under domain shift**

Our Goal :

No extra training, no additional models → Improve robustness and accuracy using TTA

- **Test-Time Augmentation** is underexplored in multimodal settings!!

Data Augmentation


Original


Horizontal Flip


Vertical Flip


Horizontal + Vertical


Color Profile 1


Color Profile 2


Color Profile 3


Color Profile 4


Rotate Left


Rotate Right


Noise 1

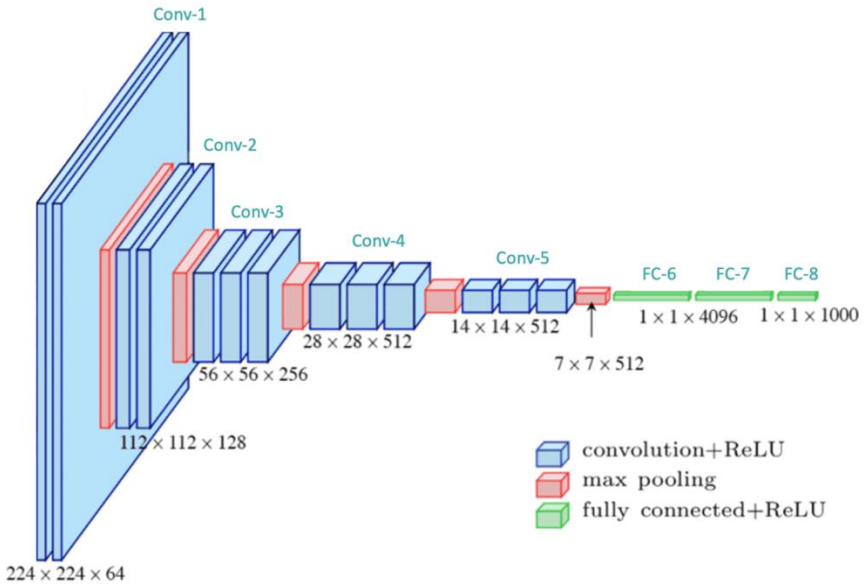

Noise 2


Crop 1


Crop 2


Resize 1


Resize 2



dog

VQA

What is the color of the car of the left?

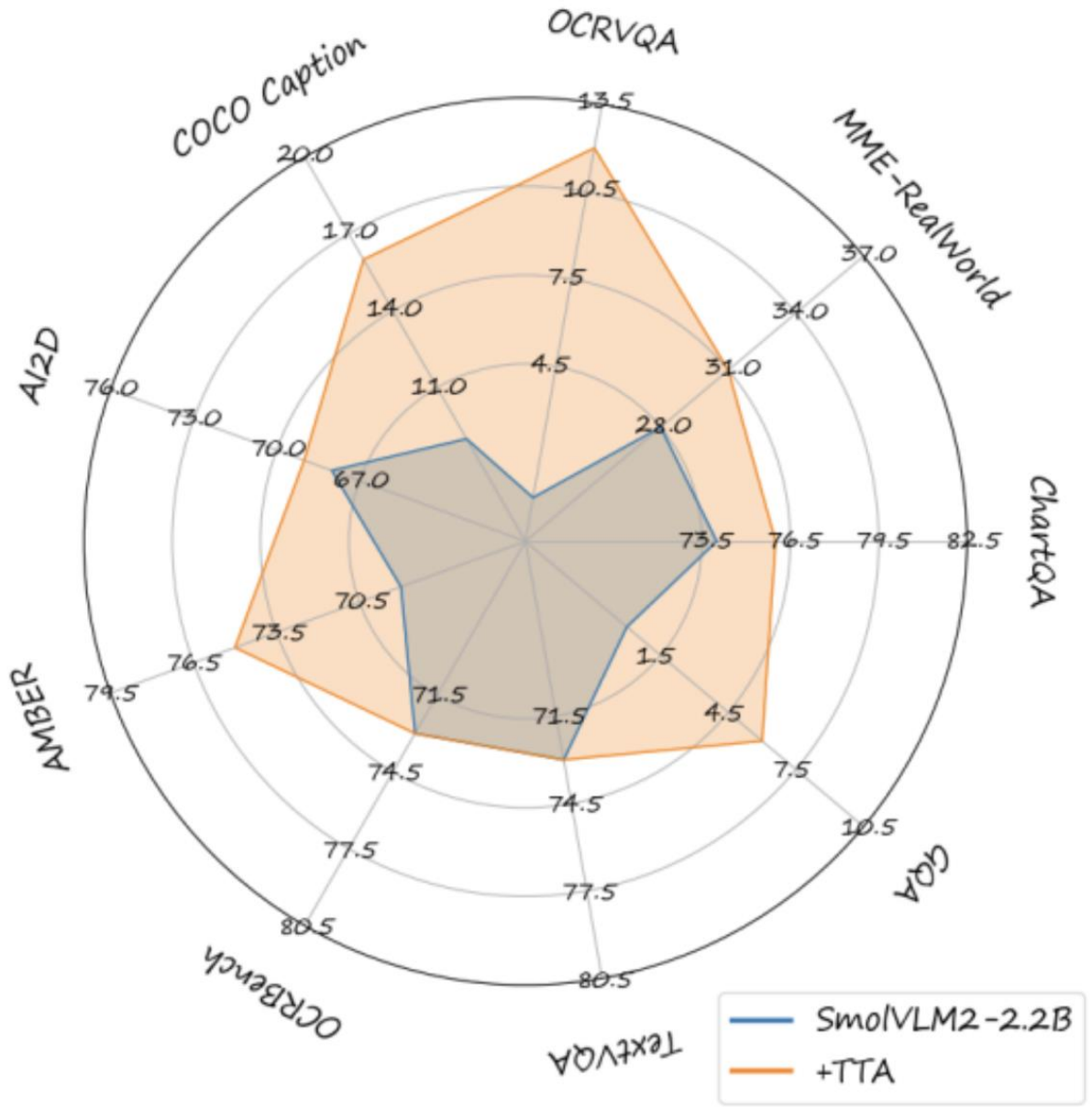
How many cars appear in the image?

What color is the car in the middle?

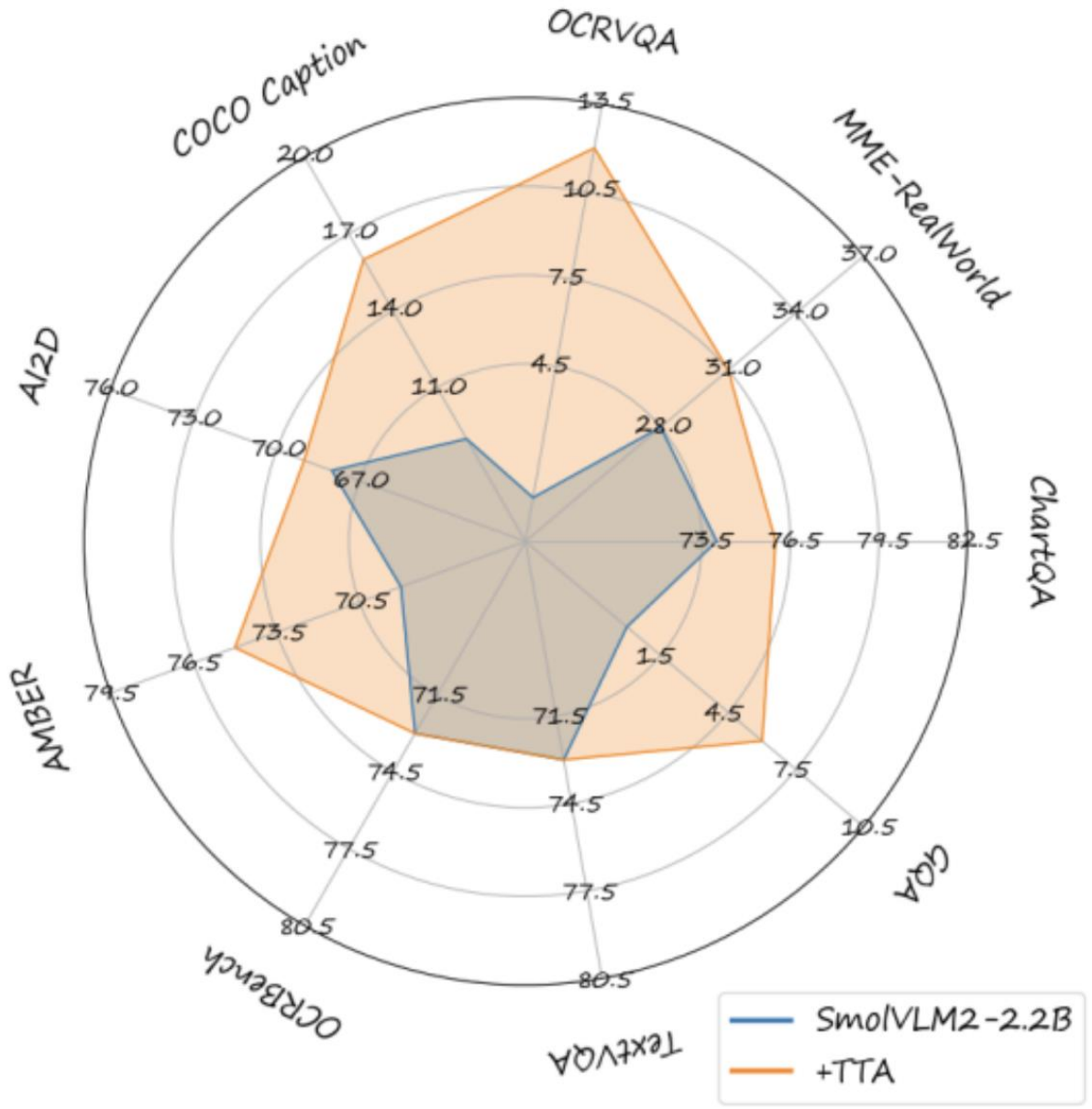
Test time Augmentation

- *how to augment?*
- *how to aggregate?*
- *where to aggregate?*

Test time Augmentation



Test time Augmentation



Computational Overhead

Number of Augmentations	Peak GPU Memory Overhead (GB)		Runtime Overhead (s)	
	Parallel	Sequential	Parallel	Sequential
2	+0.31	+0.04	+0.71	+1.41
4	+0.83	+0.16	+1.02	+4.23
8	+1.86	+0.27	+1.56	+9.81
16	+4.15	+0.38	+3.34	+21.17
32	+8.29	+0.67	+7.16	+43.72

Our approach yields improvement with minimal compute and memory overhead, making it practical even on consumer GPUs.


PIONEER
CENTRE
FOR AI
Thank you!!!
DTU