Generation by Search:
Scaling Test-Time Compute for Autoregressive Image Generation

Zhitong Gao[1], Parham Rezaei[1], Ali Cy[1], Nataša Jovanović[1], Mingqiao Ye[1], Jesse Allardice[2], Afshin Dehghan[2], Roman Bachmann[1*], Oğuzhan Fatih Kar[2*], Amir Zamir[1*]

[1]EPFL, [2]Apple

*Equal Technical Advising

***Presenter*: Zhitong Gao**

***Date*: 10.10.2025**

The Bitter Lesson

*"The biggest lesson that can be read from 70 years of AI research is that **general methods that leverage computation are ultimately the most effective**, and by a large margin…*

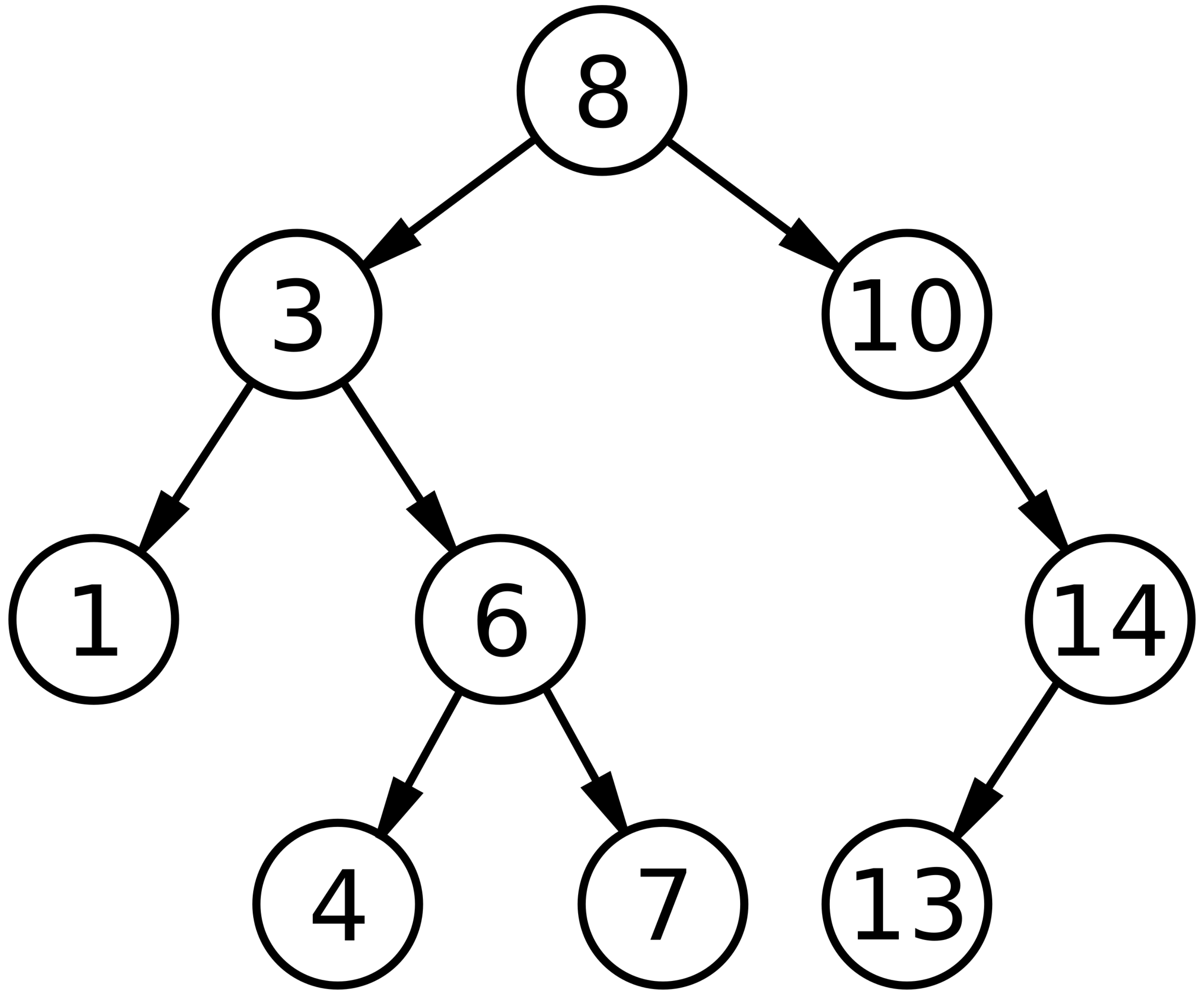
*The two methods that seem to scale arbitrarily in this way are **search** and **learning**."*

— Rich Sutton (2019)

[https://alberta-wealth.com/speakers/dr-richard-s-sutton/]

What is search?

“Search" means exploring a sequence of actions to achieve a goal.



[Image from https://en.wikipedia.org/wiki/Search_tree]

Search in Game-Playing Agents

DeepBlue (1997)

[Photo: © Yvonne Hemsey / Getty Images]

AlphaGo (2016)

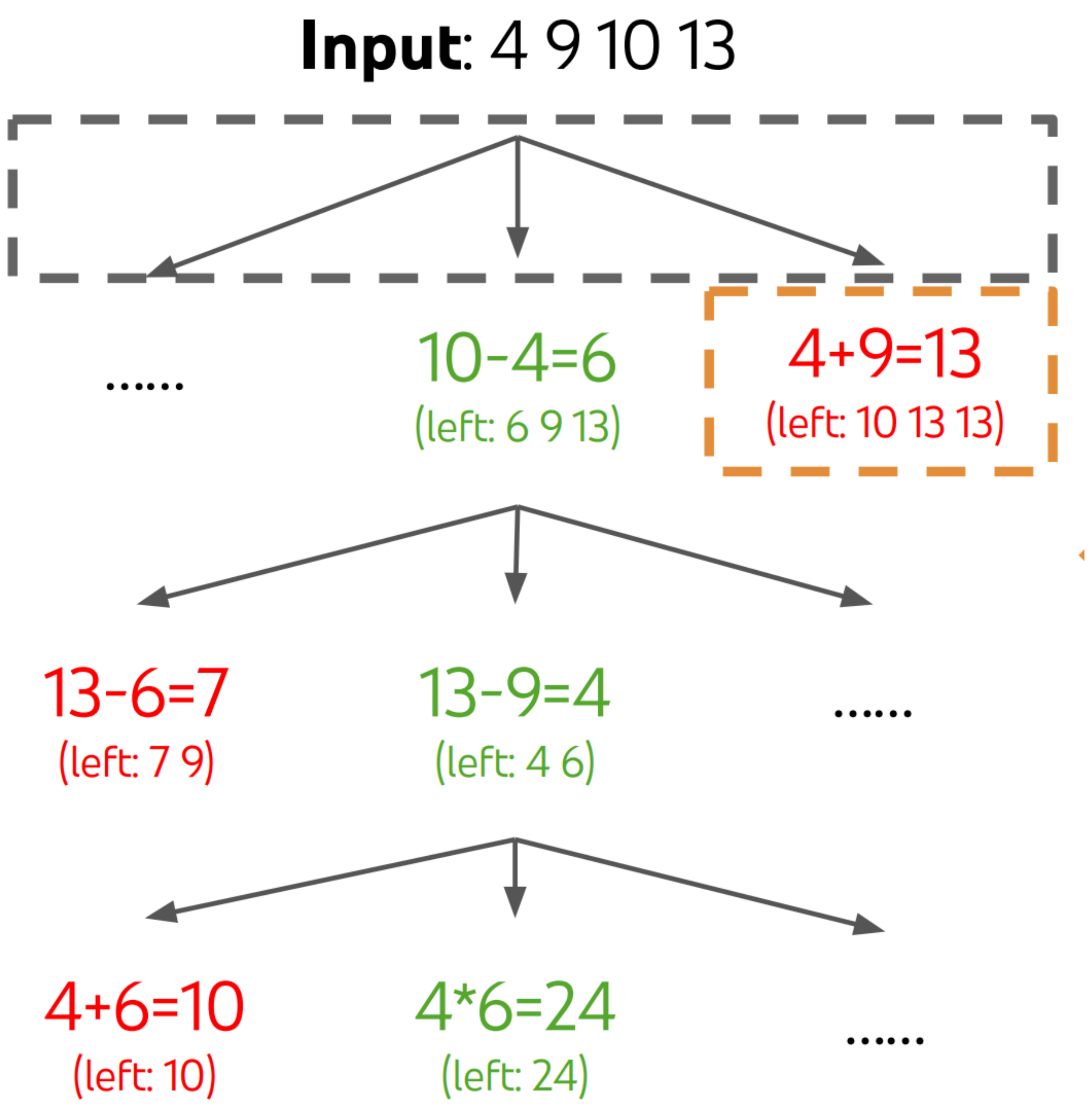
[DeepMind, Mastering the game of Go with deep neural networks and tree search, 2016]

MuZero (2020)

[Schrittwieser et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model, 2019]

Search in LLMs

Task: Use 4 numbers and +-*/ to obtain 24.



[Yao et al. Tree of thoughts: Deliberate problem solving with large language models, 2023]

Let's analyze each option.

Option A: "because appetite regulation is a field of staggering complexity."

Is that a good explanation? Hmm.

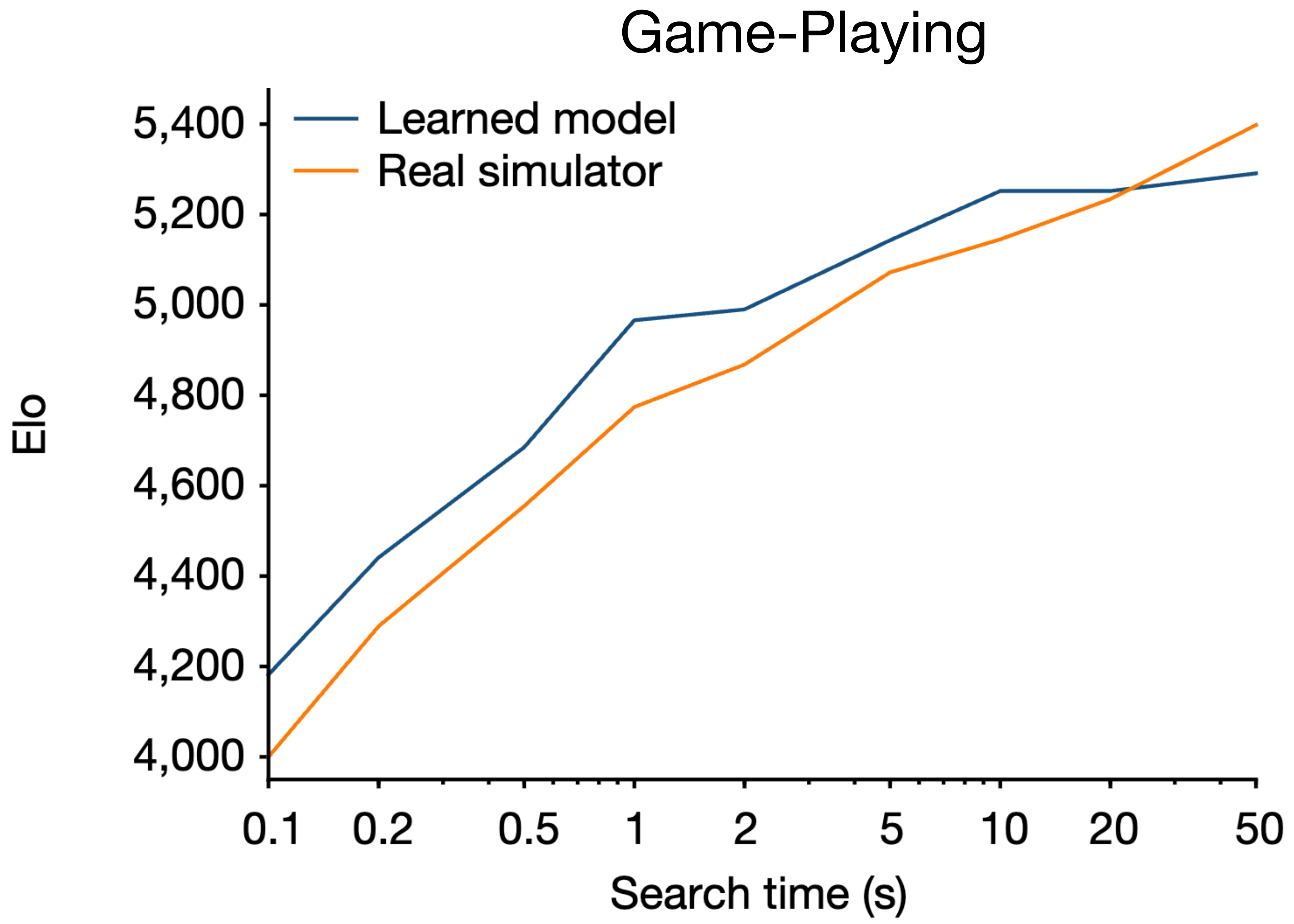
Option B: "because researchers seldom ask the right questions."

Does this make sense with the main clause?

[OpenAI o1, 2024]

Search as Test-Time Scaling

Search scales with more compute.

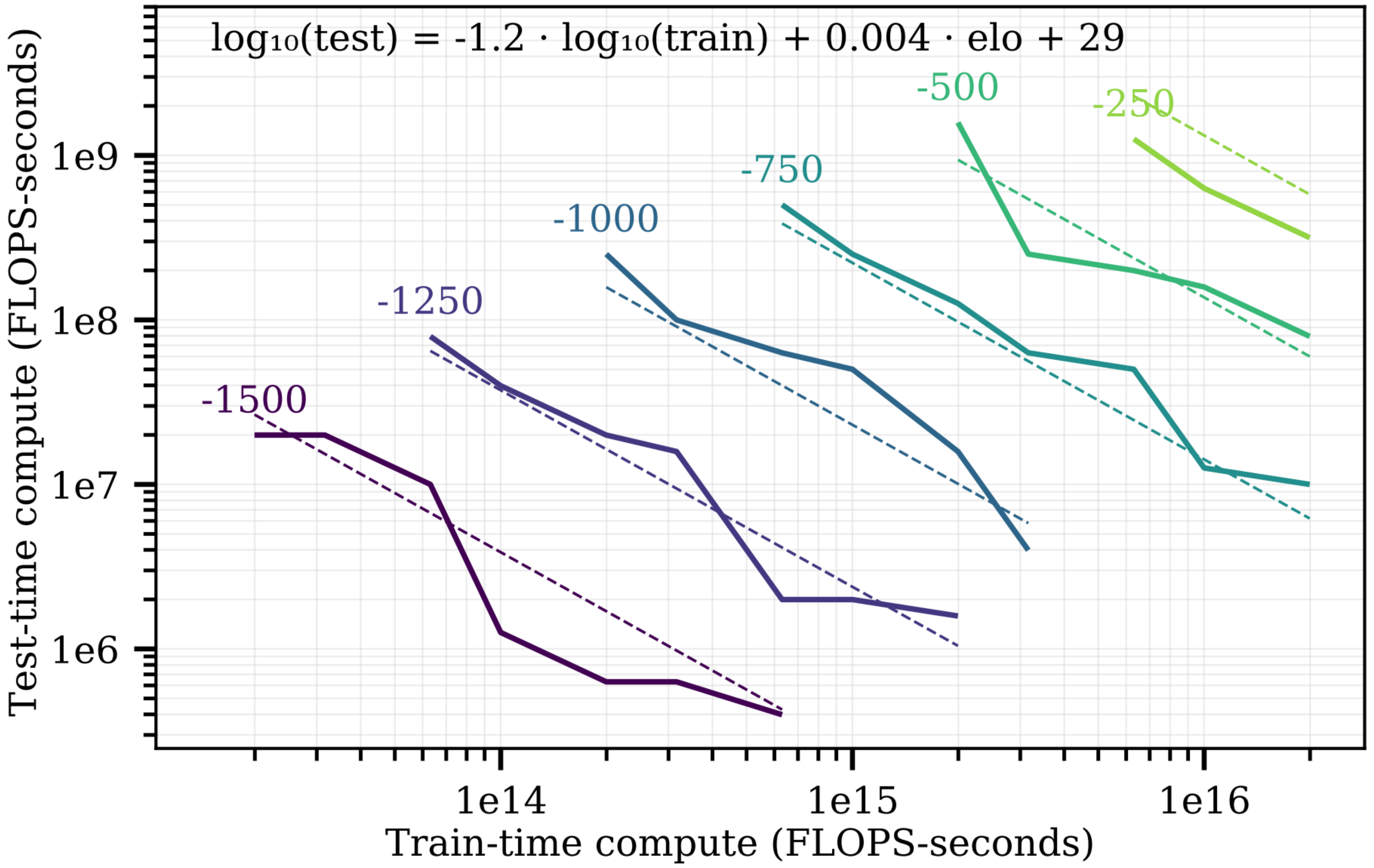


[Schrittwieser et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model, 2019]

[OpenAI, Learning to reason with LLMs, 2024]

Search as Test-Time Scaling

Compensate for training-time compute.

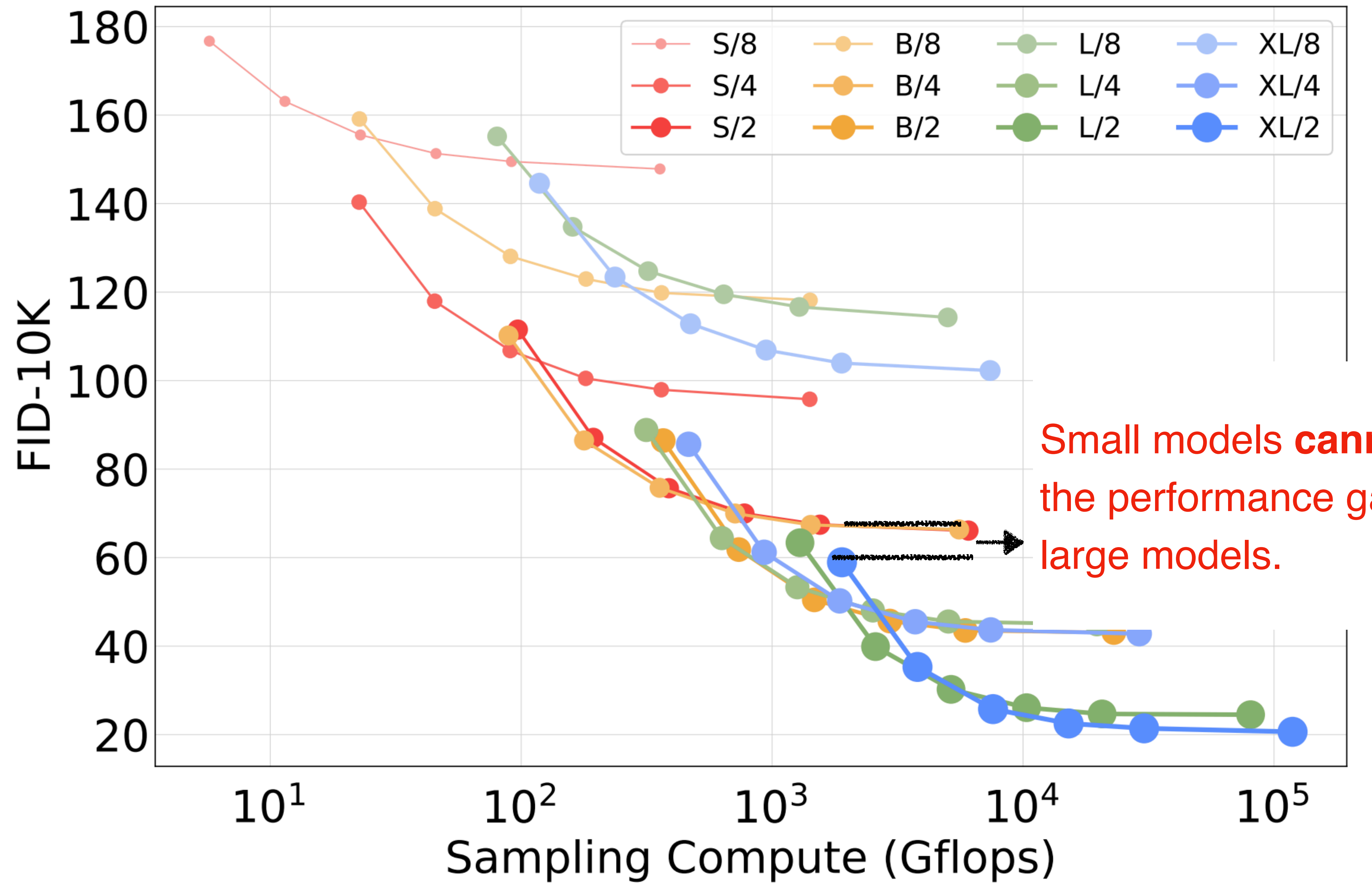


[Jones, Scaling Scaling Laws with Board Games, 2021]

[Wu et al. Inference Scaling Laws: An Empirical Analysis of Compute-Optimal Inference for Problem-Solving with Language Models, 2024]

Test-Time Scaling in Image Generation Models

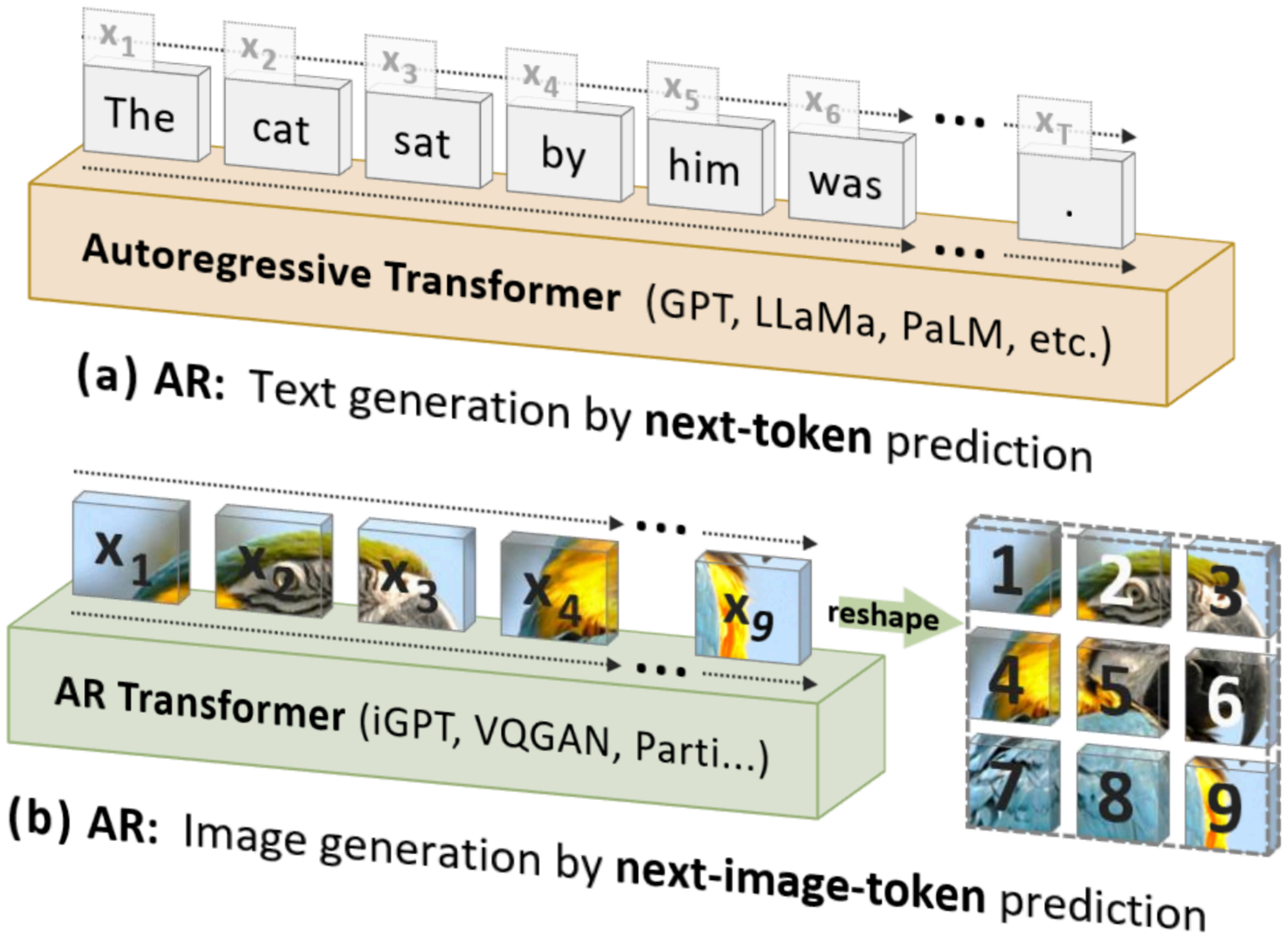
Diffusion models



Small models **cannot** close the performance gap with large models.

[Peebles et al. Scalable Diffusion Models with Transformers, 2022]

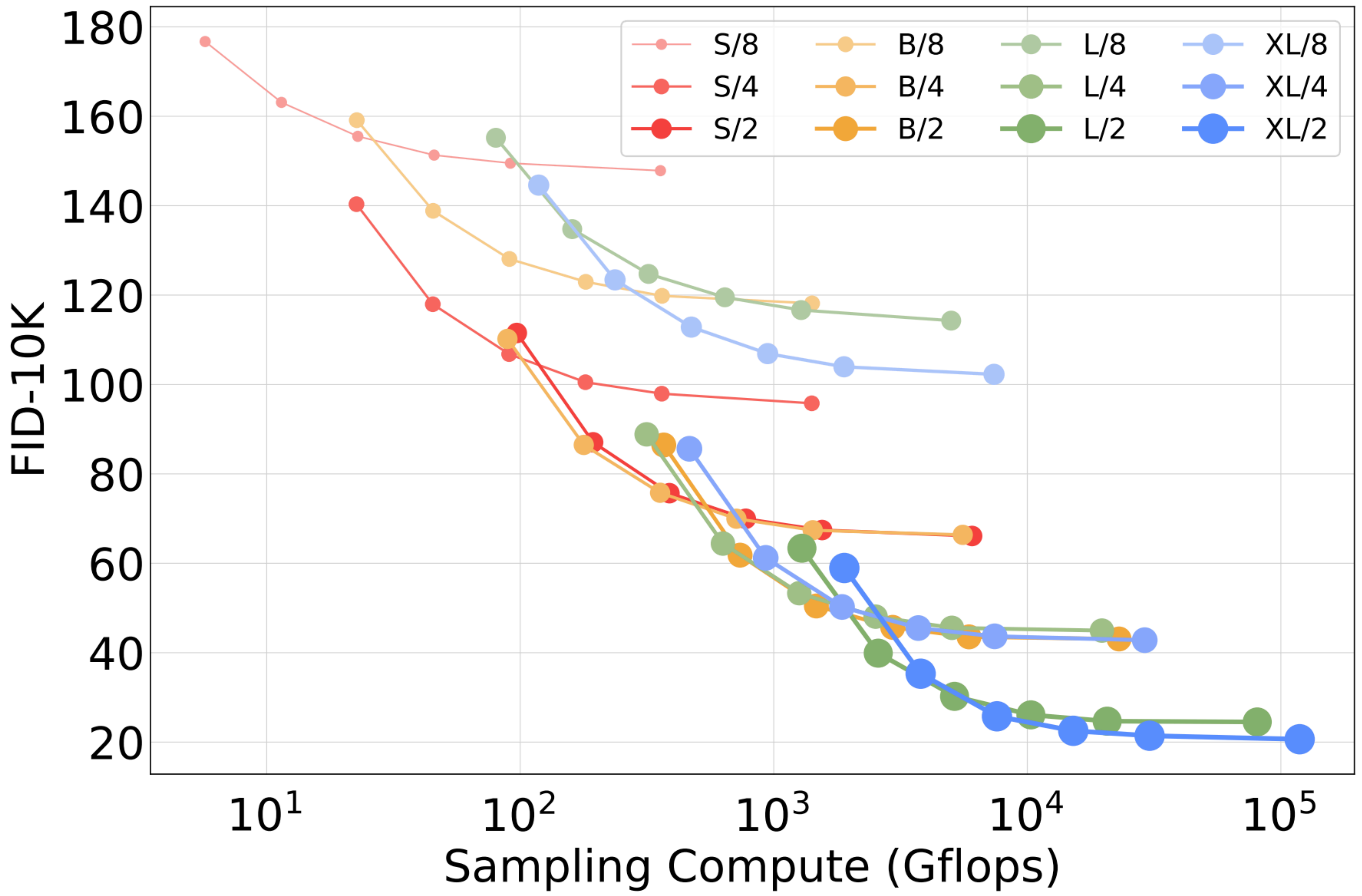
Autoregressive (AR) models.



[Tian et al. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction, 2024.]

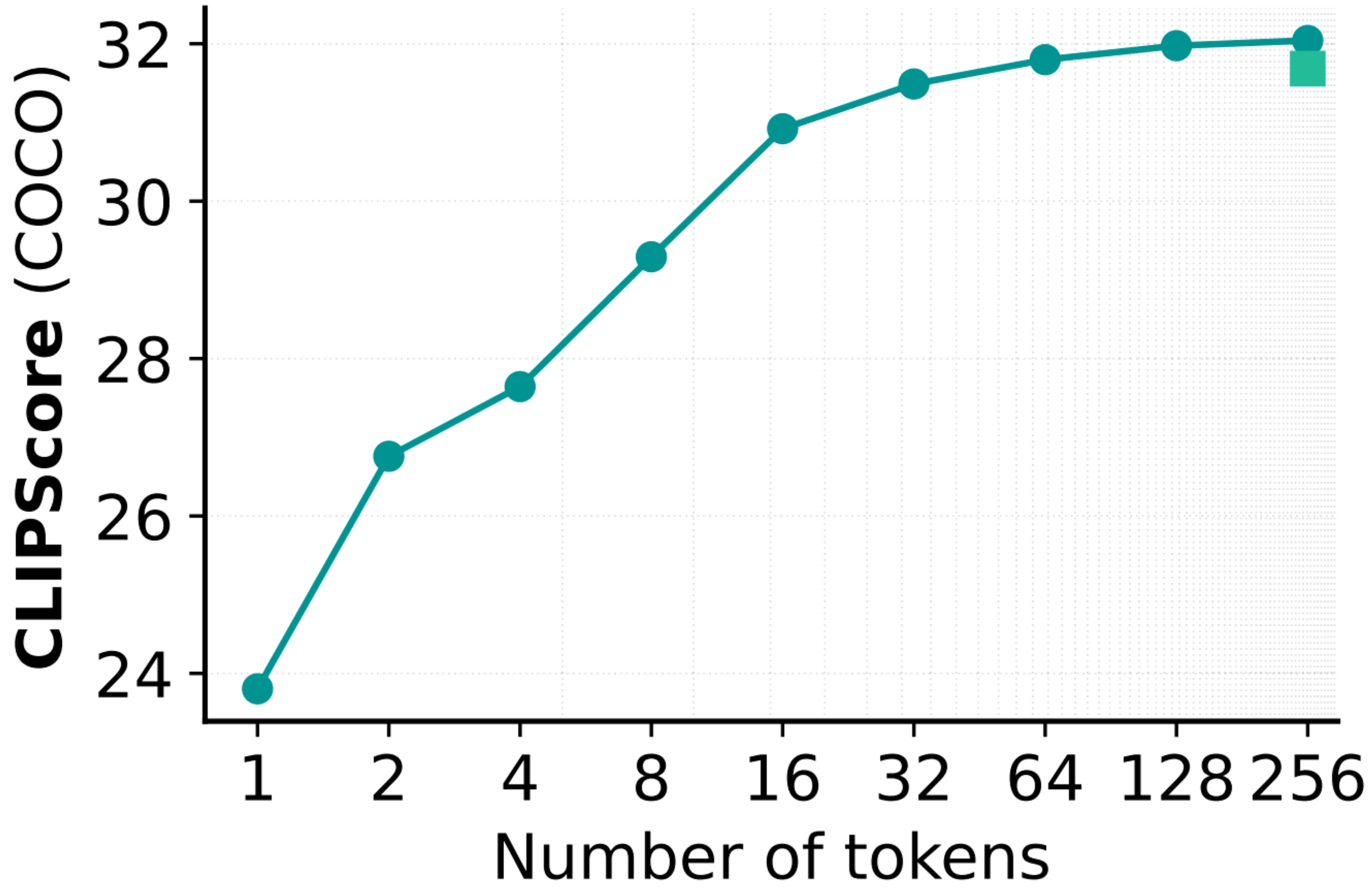
Test-Time Scaling in Image Generation Models

Diffusion models



[Peebles et al. Scalable Diffusion Models with Transformers, 2022]

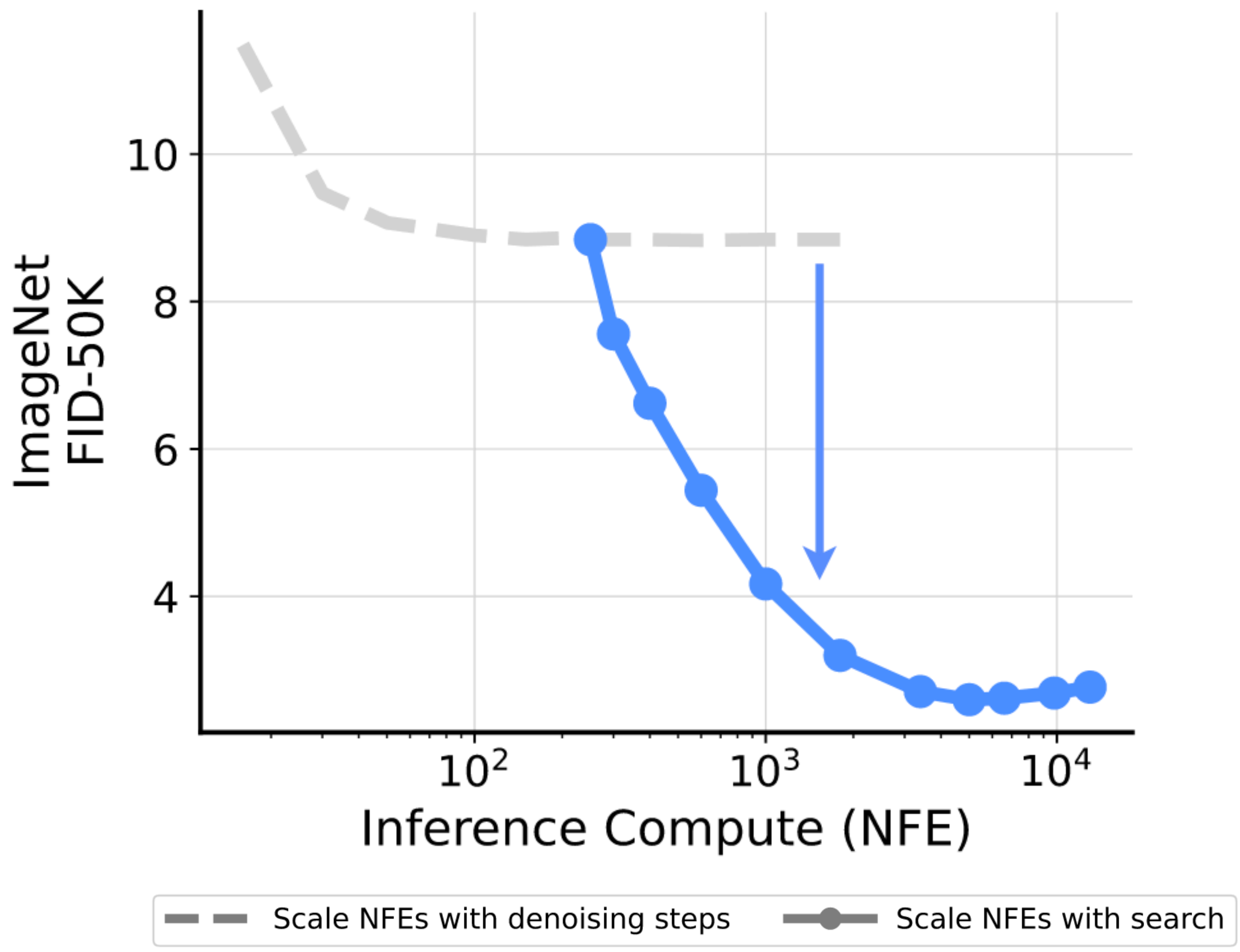
Autoregressive (AR) models.



[Bachmann et al. FlexTok: Resampling Images into 1D Token Sequences of Flexible Length, 2022]

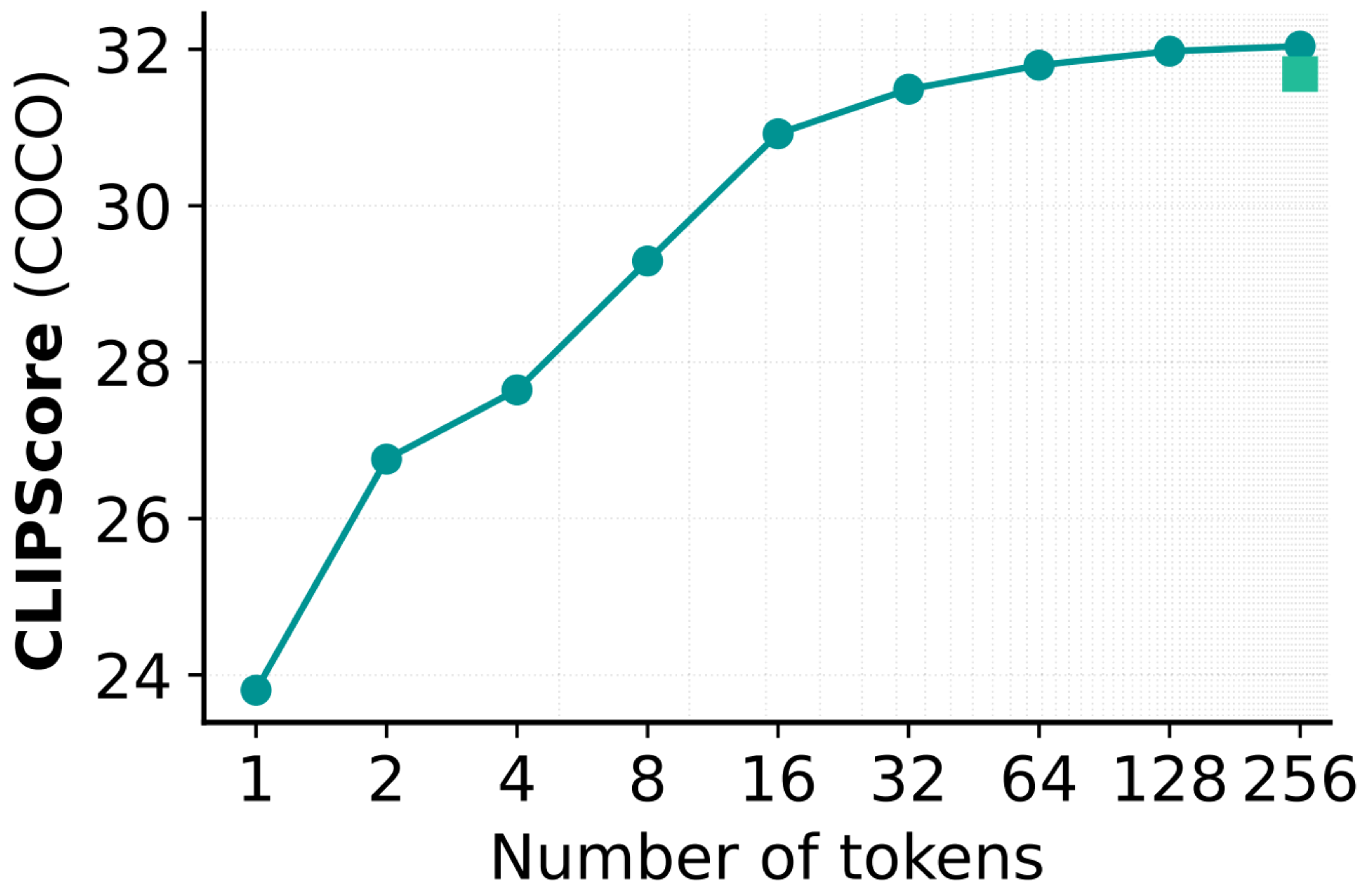
Test-Time Scaling in Image Generation Models

Diffusion models



[Ma et al. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps, 2025]

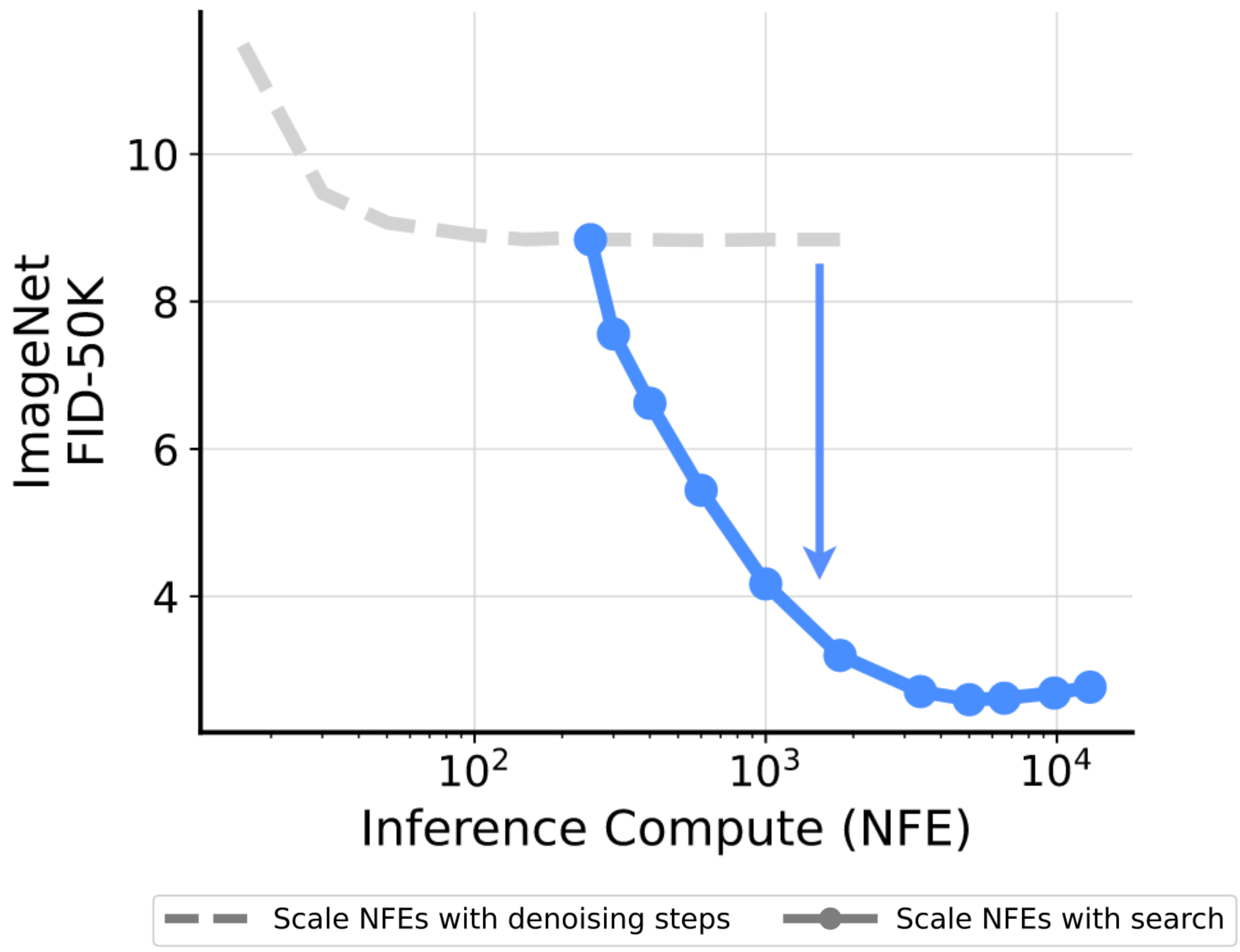
Autoregressive (AR) models.



[Bachmann et al. FlexTok: Resampling Images into 1D Token Sequences of Flexible Length, 2022]

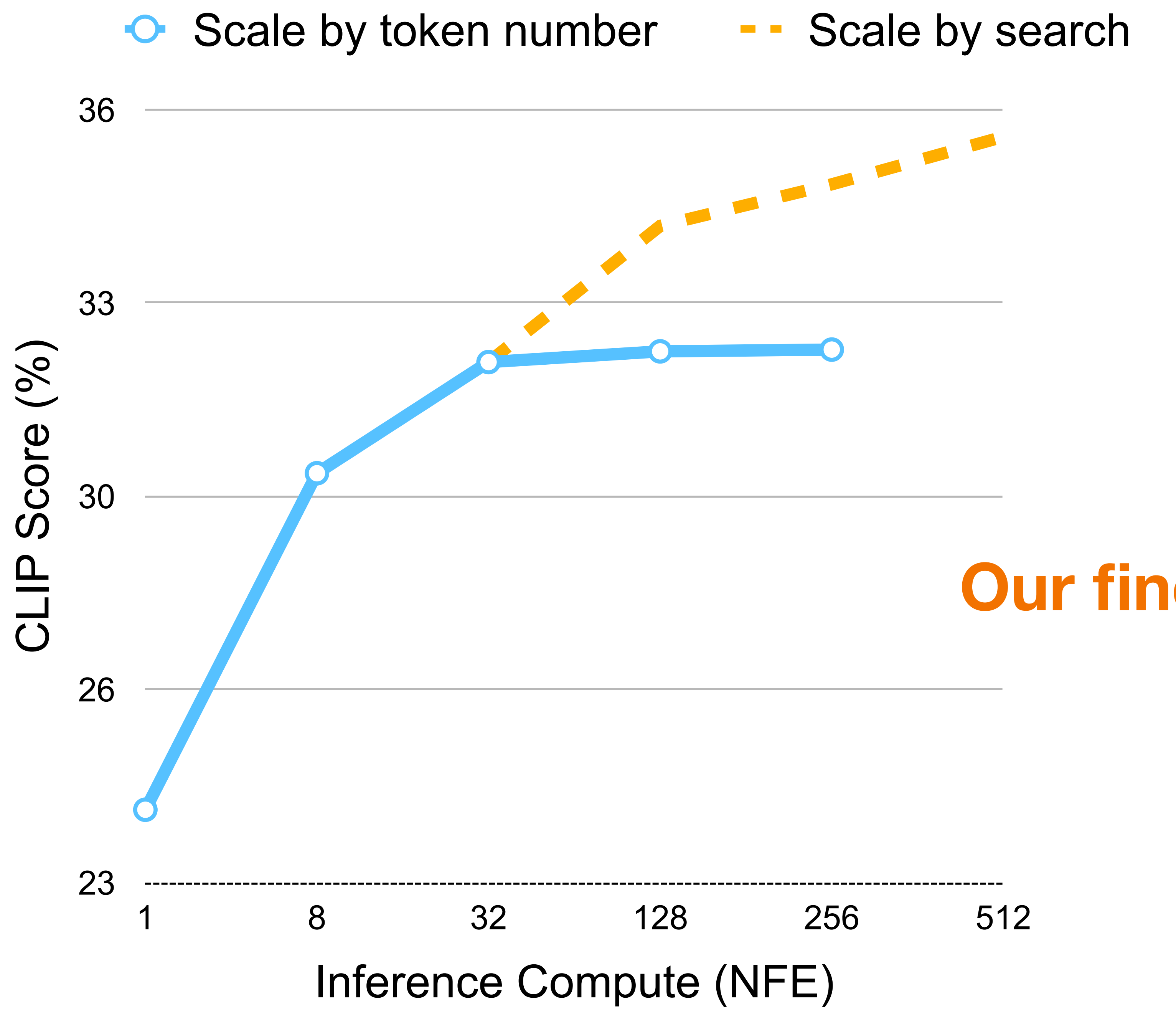
Test-Time Scaling in Image Generation Models

Diffusion models



[Ma et al. Inference-Time Scaling for Diffusion Models beyond Scaling Denoising Steps, 2025]

Autoregressive (AR) models.



Our finding

[Bachmann et al. FlexTok: Resampling Images into 1D Token Sequences of Flexible Length, 2022]

Image generation as a search problem

Image generation as a search problem

Image generation as a search problem

➡ Action

Image generation as a search problem

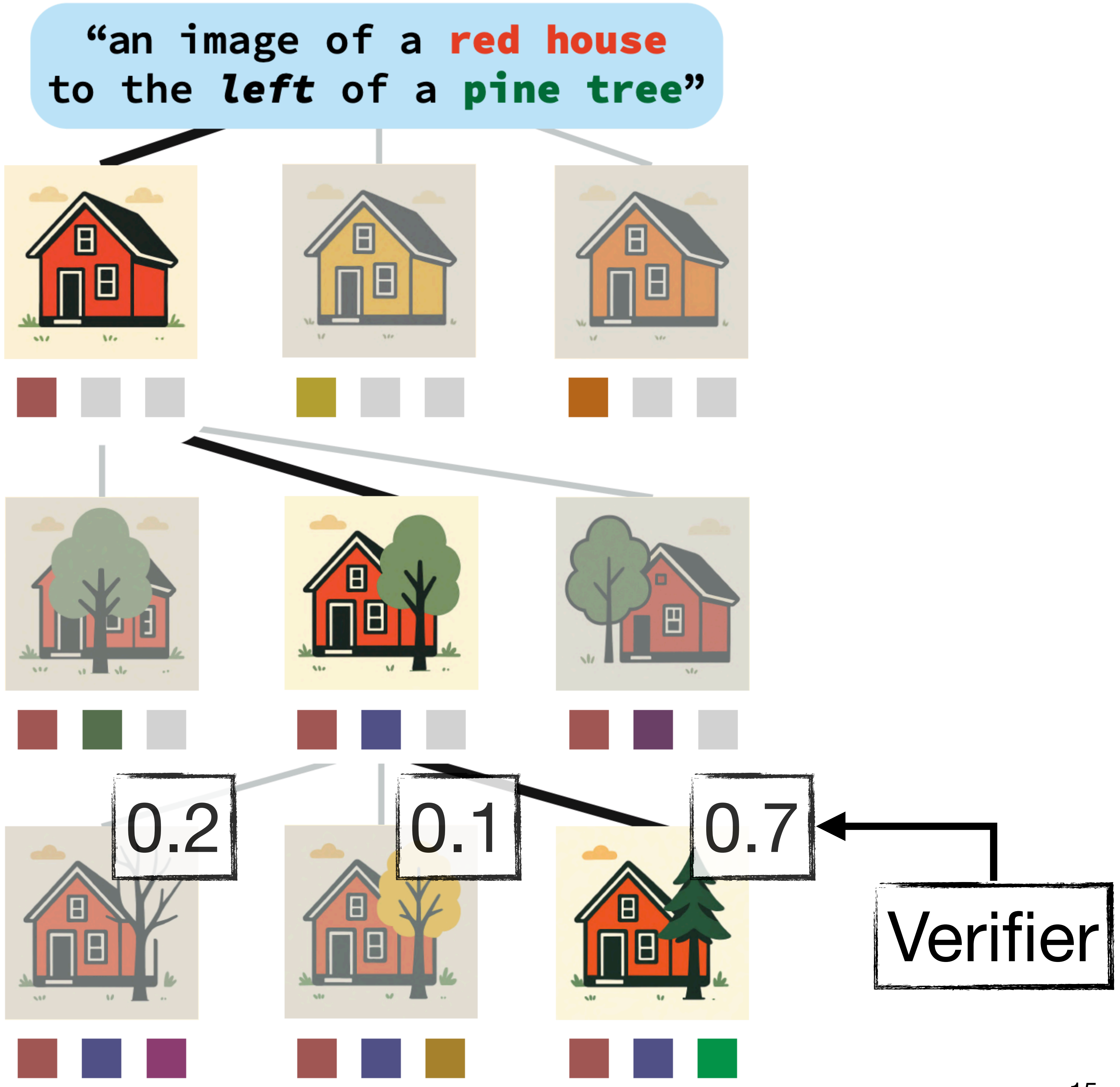
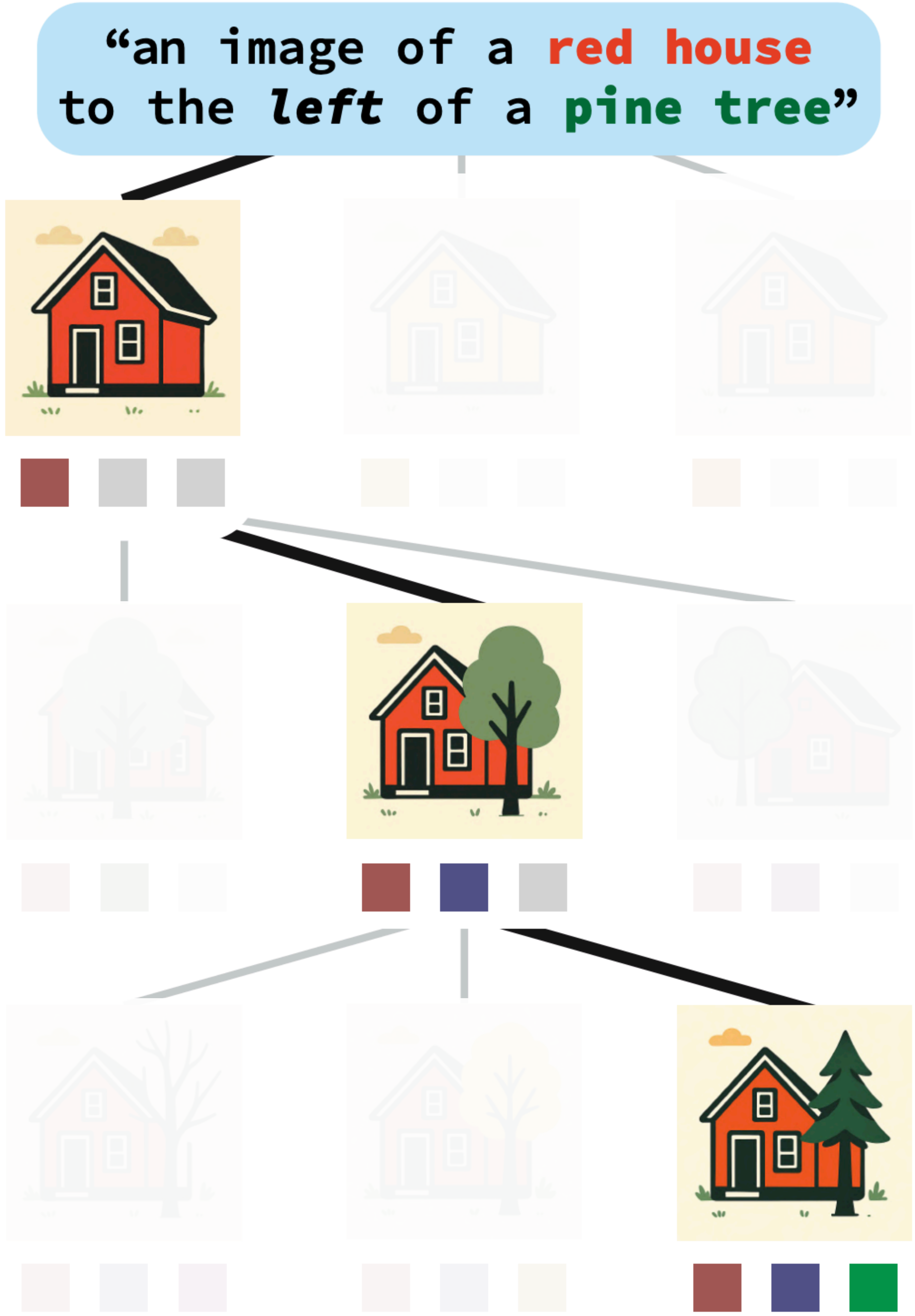


Image generation as a search problem



Standard AR generation: greedy search with likelihood as verifier.

- Greedy decoding may **miss globally optimal** sequences.
- High-likelihood **doesn't** always match **desired image quality or alignment.**

Image generation as a search problem

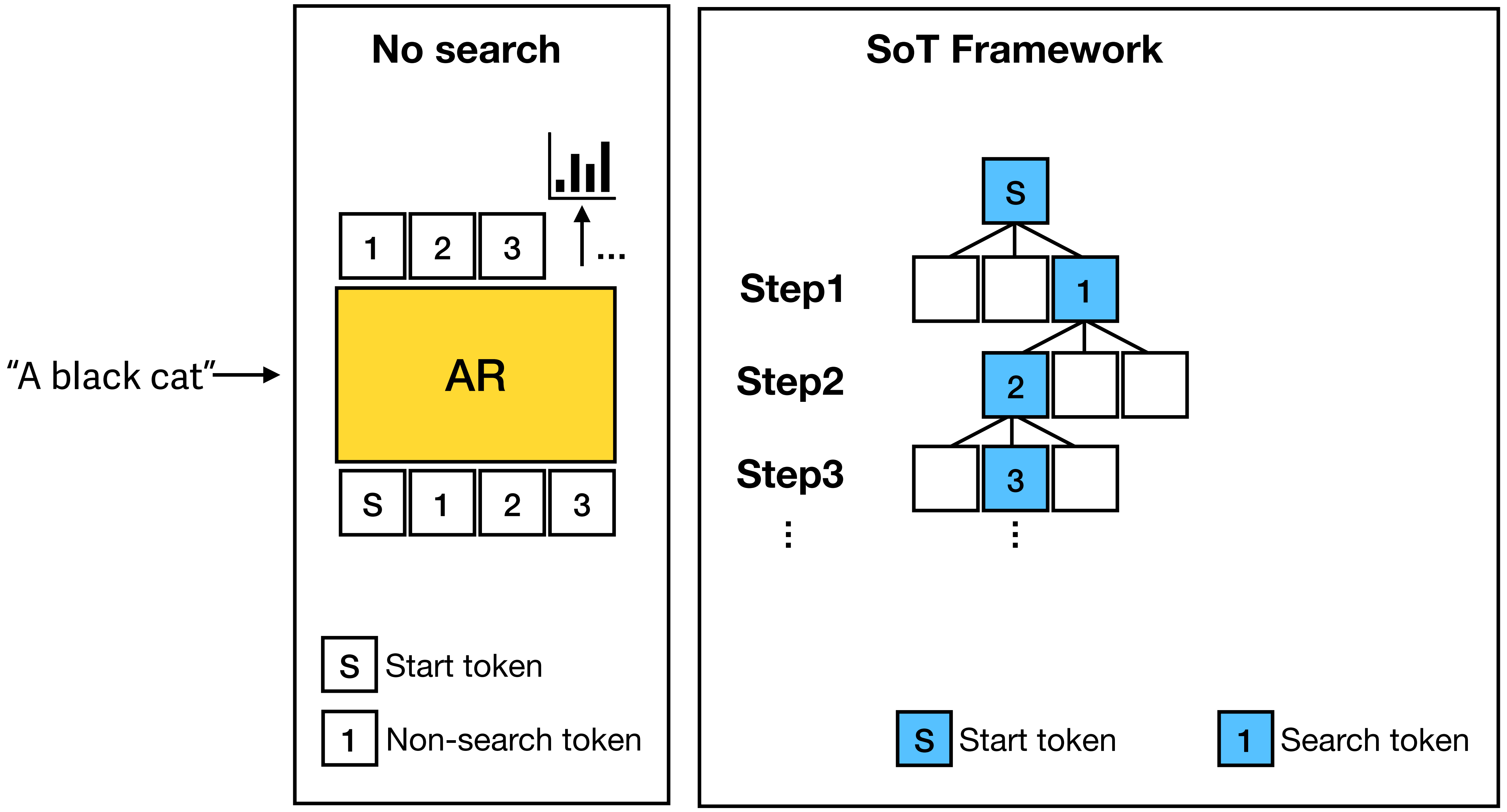
Standard AR generation: greedy search with likelihood as verifier.

- Greedy decoding may **miss globally optimal** sequences.
- High-likelihood **doesn't** always match **desired image quality or alignment.**

Search over tokens (SoT)

- Explore more sophisticated search algorithms.
- Use verifiers to directly measure expected task utility.

Search over Token (SoT) Framework

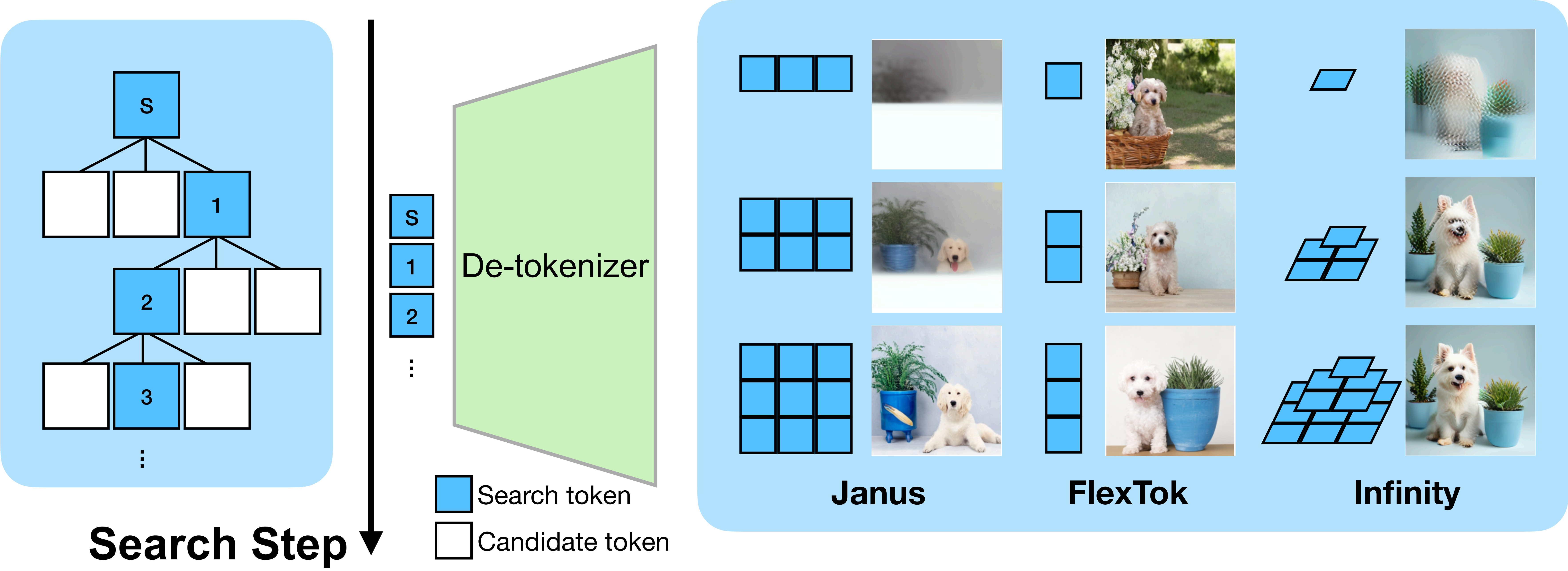


Search over Token (SoT) Framework

Search over Token (SoT) Framework

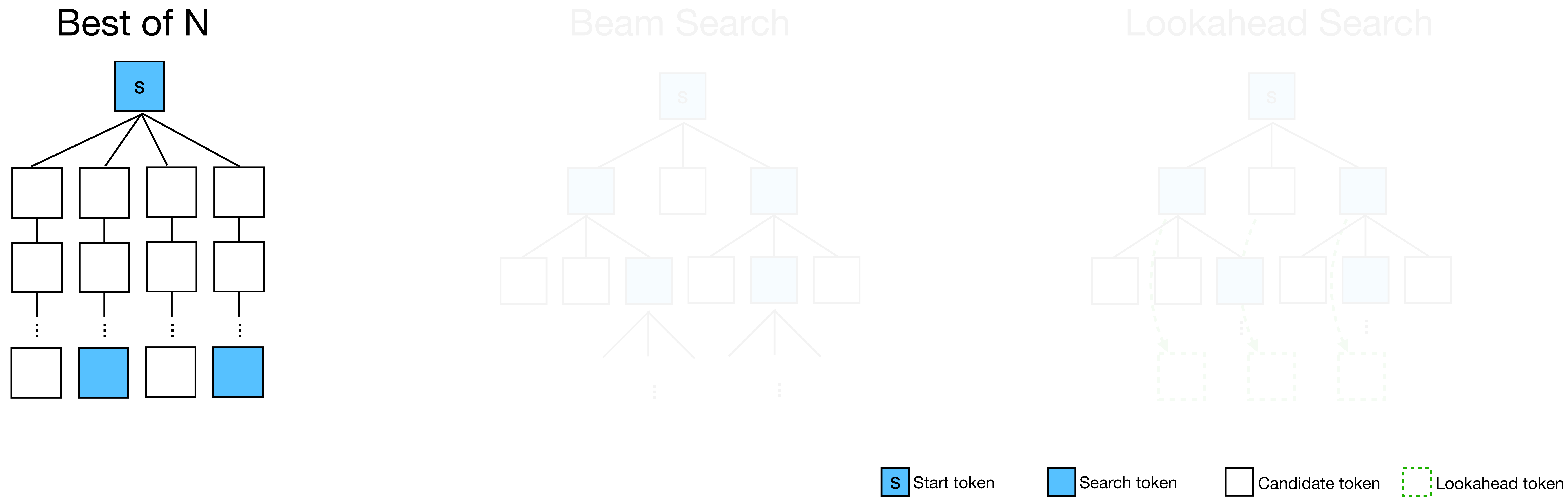
Four Components: Token structure Search Algorithm Verifier AR Prior

Search over Token (SoT) Framework



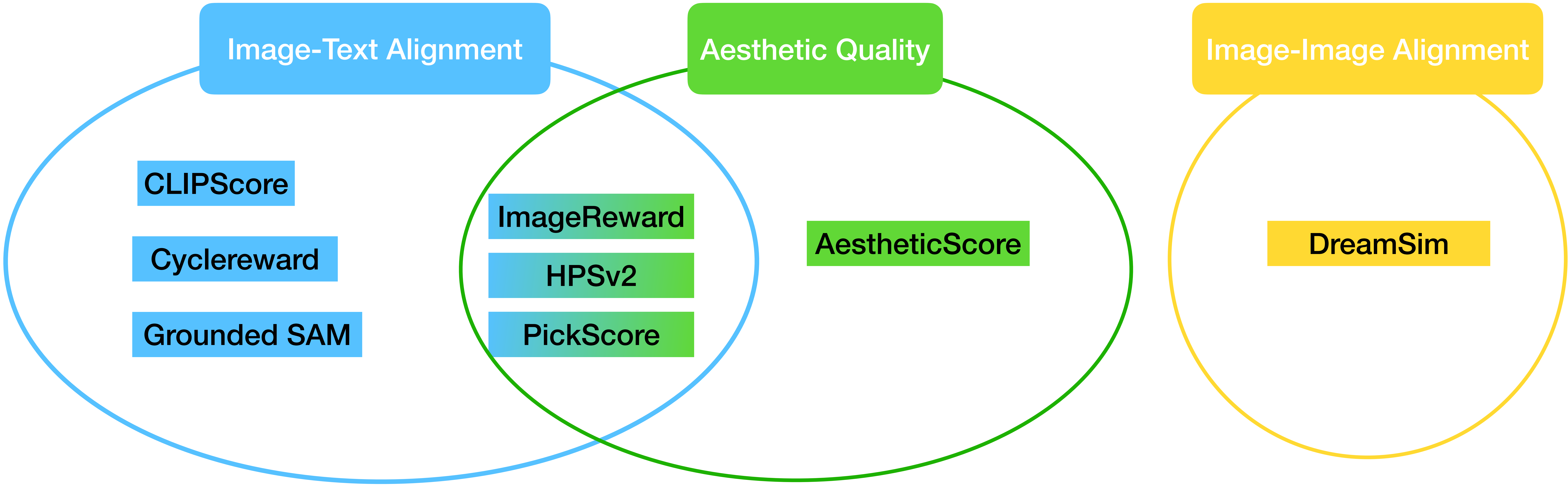
Four Components: Token structure Search Algorithm Verifier AR Prior

Search over Token (SoT) Framework



Four Components: Token structure | Search Algorithm | Verifier | AR Prior

Search over Token (SoT) Framework



Four Components: Token structure Search Algorithm Verifier AR Prior

Search over Token (SoT) Framework

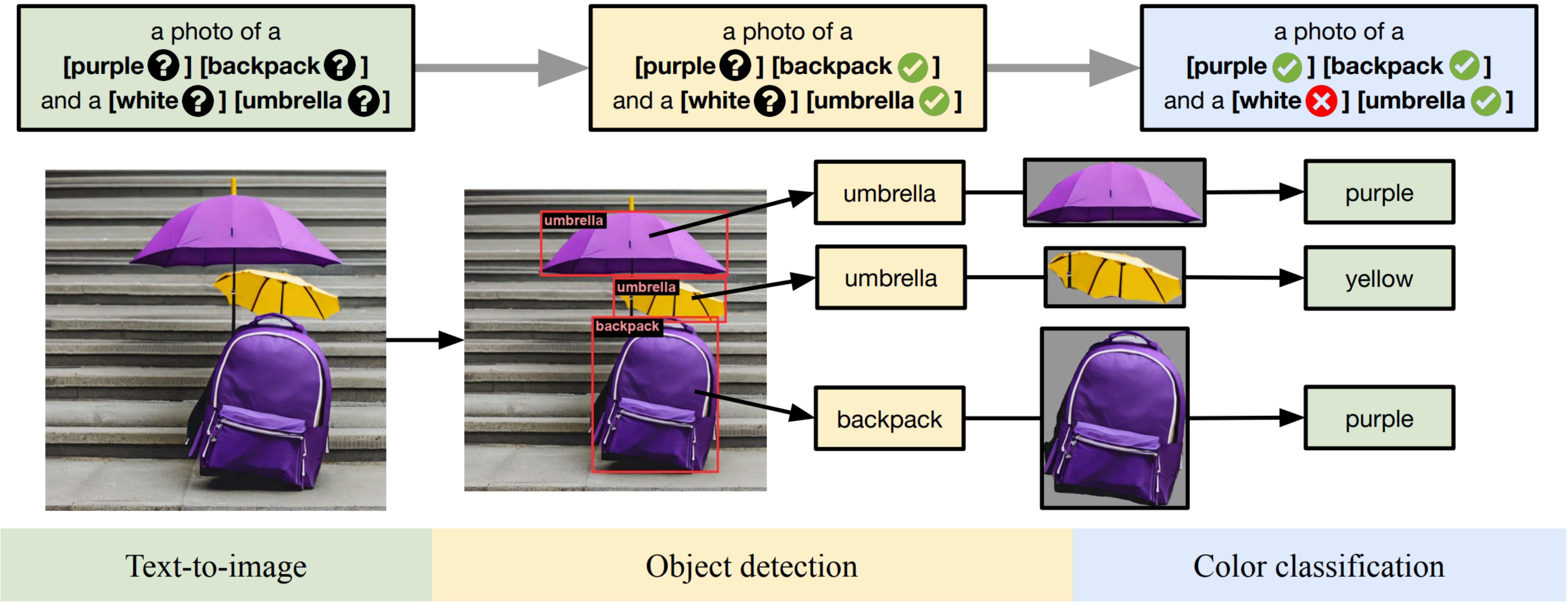
Can we generate images just by searching, without using an AR model?

Four Components: Token structure | Search Algorithm | Verifier | AR Prior

Main Results

1. Search improves condition alignment across AR models.

- GenEval Benchmark



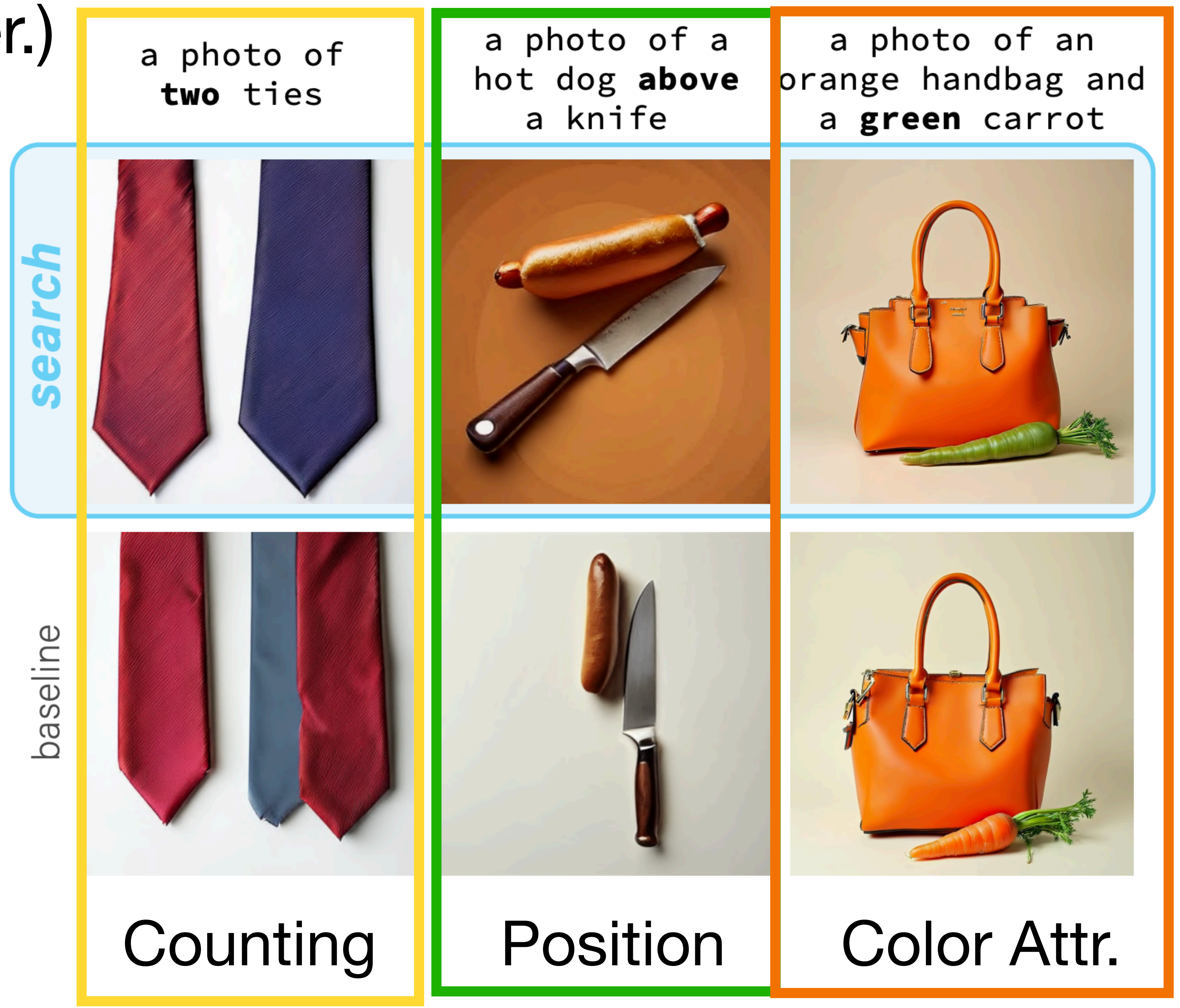
[Ghosh et al. GENEVAL: An Object-Focused Framework for Evaluating Text-to-Image Alignment, 2023.]

Main Results

1. Search improves condition alignment across AR models.

- Results on GenEval Benchmark. (Imagereward as verifier.)

Model	Search	Single obj.	Two obj.	Counting	Colors	Position	Color attri.	Overall ↑
FlexTok (Bachmann et al., 2025)	–	95	59	56	80	16	35	57
	BoN	100 +5	84 +25	69 +13	90 +10	24 +8	57 +22	68 +11
	Beam	100 +5	88 +29	69 +13	91 +11	23 +7	53 +18	70 +13
Infinity (Han et al., 2025)	–	98	82	65	83	27	64	70
	BoN	100 +2	93 +11	71 +6	83 +0	30 +3	67 +3	74 +4
	LA	100 +2	93 +11	69 +4	91 +8	36 +9	74 +10	77 +7
Janus (Wu et al., 2024)	–	96	60	38	85	43	44	61
	BoN	96 +0	91 +31	51 +13	90 +5	65 +22	55 +11	75 +14
	LA	100 +4	94 +34	58 +20	90 +5	70 +27	79 +35	82 +21
Janus-Pro (Chen et al., 2025a)	–	100	86	60	91	76	60	79
	BoN	97 -3	91 +5	74 +14	90 -1	77 +1	78 +18	85 +6
	LA	100 +0	95 +9	76 +16	94 +3	81 +5	79 +19	87 +8



Main Results

1. Search improves condition alignment across AR models.

- Results on GenEval Benchmark. (Imagereward as verifier.)

Model	Single obj.	Two obj.	Counting	Colors	Position	Color attri.	Overall ↑
Models without test-time search							
CLIP Retrieval Beaumont (2022)	89	22	37	62	3	0	35
SD-XL (Podell et al., 2023)	98	74	39	85	15	23	55
LlamaGen (Sun et al., 2024)	75	26	20	55	42	32	31
LlamaGen-GRPO (Yuan et al., 2025)	79	26	23	59	40	30	32
Emu3-Gen Wang et al. (2024)	98	71	34	81	17	21	54
FlexTok† (Bachmann et al., 2025)	95	59	56	80	16	35	57
Janus† (Wu et al., 2024)	96	60	38	85	43	44	61
Show-o (Xie et al., 2024)	98	80	66	84	31	50	68
Infinity† (Han et al., 2025)	98	82	65	83	27	64	70
Janus-Pro† (Chen et al., 2025a)	100	86	60	91	76	60	79
GPT-4o-Image (Yan et al., 2025)	99	92	**85**	89	74	71	84
Models with test-time search							
TTS-VAR (Chen et al., 2025b)	–	95	74	–	–	68	75
SoT-Janus-Pro (Ours)	**100**	**95**	76	**94**	**81**	**79**	**87**

Best result on the GenEval benchmark.

Main Results

1. Search improves condition alignment across AR models.

- Results on long prompts.

A contented sloth, with a wide grin on its face, is decked out in an eclectic ensemble featuring a sleek black leather jacket and a brown cowboy hat atop its head. It's also sporting a traditional tartan kilt paired with a smart red bowtie around its neck. **In one claw**, the sloth firmly grips a wooden quarterstaff, while **the other** supports a large, thick book with a leather-bound cover.

Two ceramic cups filled with steaming coffee are placed on a wooden table with a natural grain finish. The cup on the left showcases intricate latte art spelling out **the word "LOVE" with a heart-shaped design**, while the cup on the right has the word "**PEACE**" beautifully crafted atop its frothy surface. Both cups have a glossy finish, and the warm lighting accentuates the creamy texture of the latte art.

Base model | Base model + SoT

Main Results

2. Search enables zero-shot multimodal control.

- Results on DreamBench++. (Dreamsim as verifier)

Text Prompts:

A photograph of a dog lazily sunbathing by a serene lake.

A photograph of a dog sitting attentively at a bustling street market.

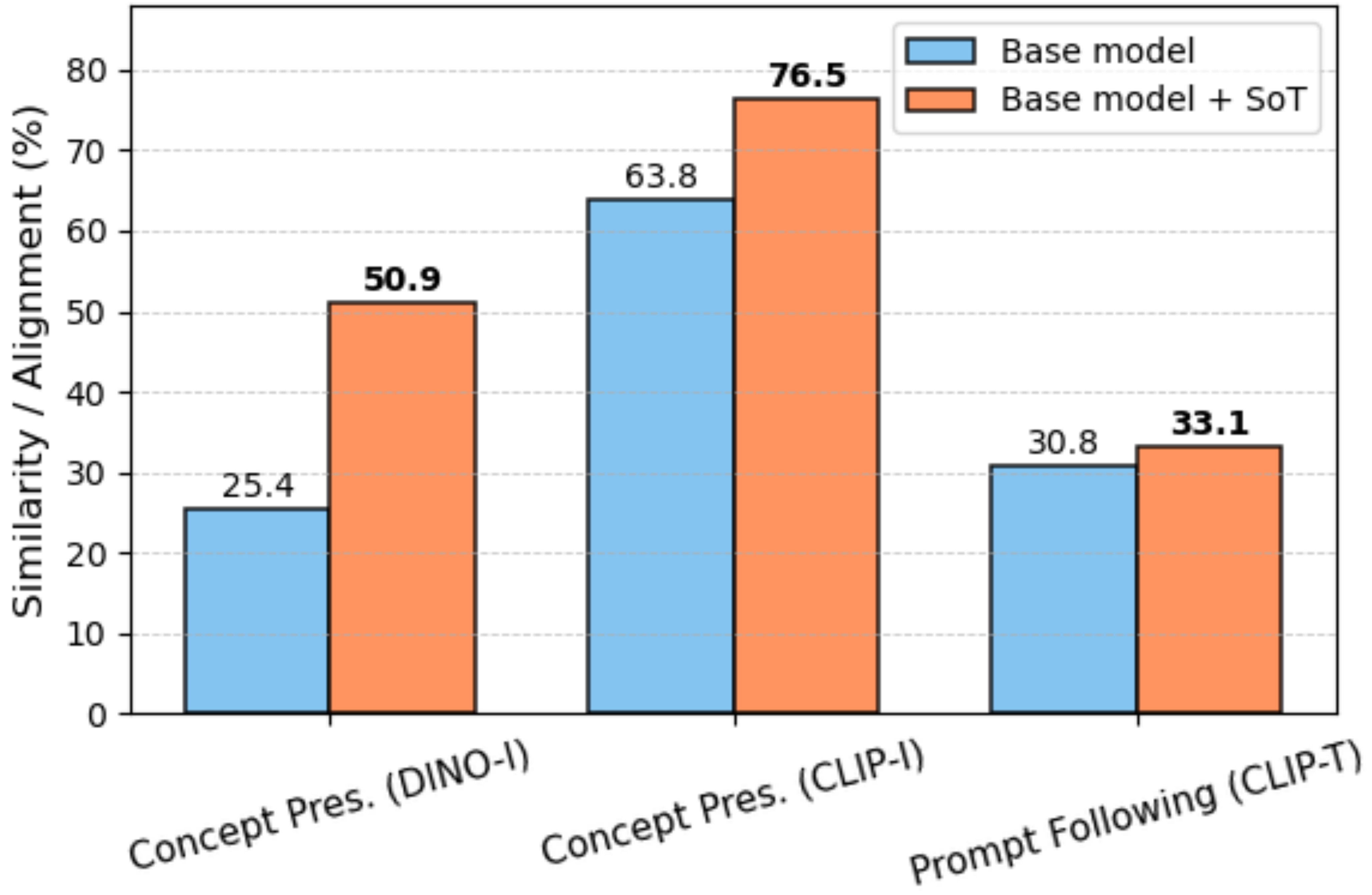
A photo of a dog curled up beside a crackling campfire under starlight.

Text Prompts:

A photo of a dog sitting patiently at a busy city crosswalk.

A retro-style painting of a dog sitting beside an old gramophone.

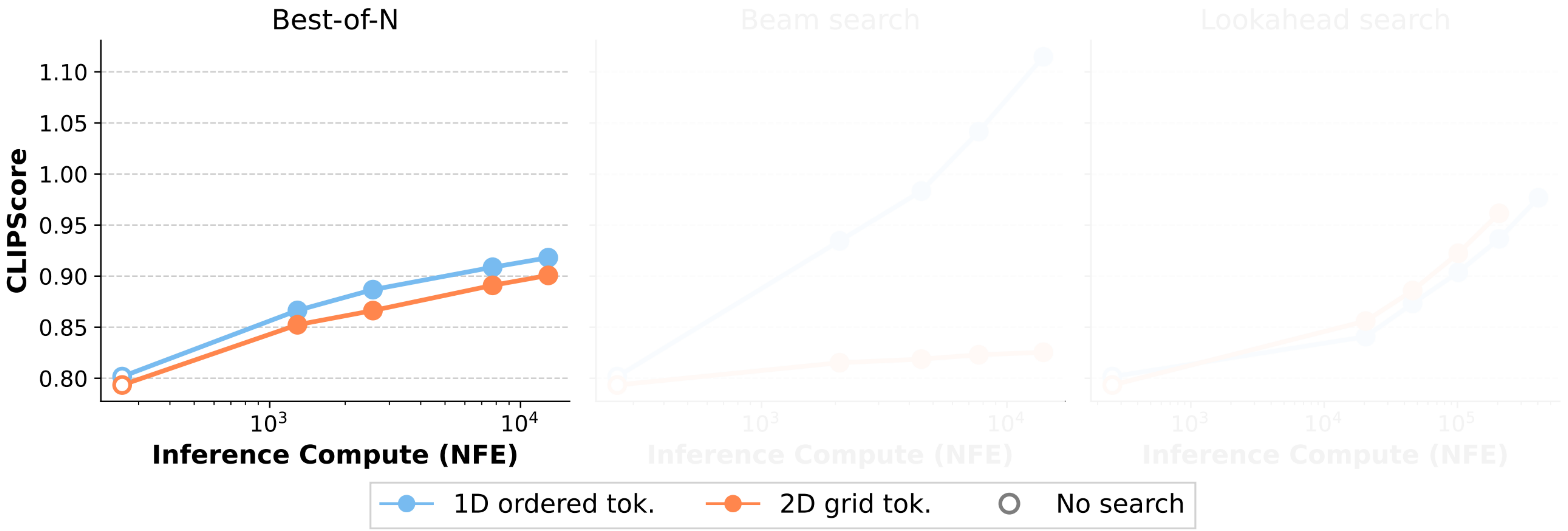
A photograph of a dog leaping through a field of tall grass.



Understanding SoT design Space

1. Token structure and search algorithms.

- Apples-to-apples comparison between 1D ordered token and 2D grid token.



Understanding SoT design Space

2. Comparison of different verifiers

- Leave-one-out evaluation on different verifiers.

Verifier \ Evaluation metric	CLIPScore	AestheticScore	CycleReward	HPSv2	ImageReward	Grounded SAM	PickScore	GenEval acc.	Avg. Rank
No search	0.800 [9]	4.617 [10]	0.636 [9]	0.249 [9]	0.543 [10]	0.660 [10]	0.063 [9]	0.570 [10]	9.50
CLIPScore	**0.891** [1]	4.677 [8]	0.949 [4]	0.255 [7]	0.817 [6]	0.692 [7]	0.079 [6]	0.631 [5]	5.50
AestheticScore	0.794 [10]	**5.430** [1]	0.610 [10]	0.259 [5]	0.654 [7]	0.680 [9]	0.078 [7]	0.587 [9]	7.25
CycleReward	0.826 [3]	4.691 [6]	**1.354** [1]	0.256 [6]	1.040 [3]	0.721 [3]	0.087 [5]	0.619 [6]	4.12
HPSv2	0.821 [4]	4.957 [3]	0.911 [5]	**0.289** [1]	0.983 [4]	0.701 [6]	0.136 [3]	0.663 [4]	3.75
ImageReward	0.818 [6]	4.778 [5]	1.098 [3]	0.264 [4]	**1.294** [1]	0.705 [5]	0.090 [4]	0.675 [2]	3.75
Grounded SAM	0.801 [8]	4.674 [9]	0.735 [7]	0.248 [10]	0.588 [9]	**0.942** [1]	0.060 [10]	0.603 [8]	7.75
PickScore	0.819 [5]	4.912 [4]	0.895 [6]	0.271 [3]	0.848 [5]	0.709 [4]	**0.198** [1]	0.612 [7]	4.38
GenEval Verifier	0.805 [7]	4.678 [7]	0.707 [8]	0.252 [8]	0.631 [8]	0.681 [8]	0.065 [8]	**0.761** [1]	6.88
Ensemble	0.841 [2]	5.036 [2]	1.123 [2]	0.277 [2]	1.121 [2]	0.775 [2]	0.144 [2]	0.674 [3]	**2.12**

Relative score per column: 0% 25% 50% 75% 100%

Understanding SoT design Space

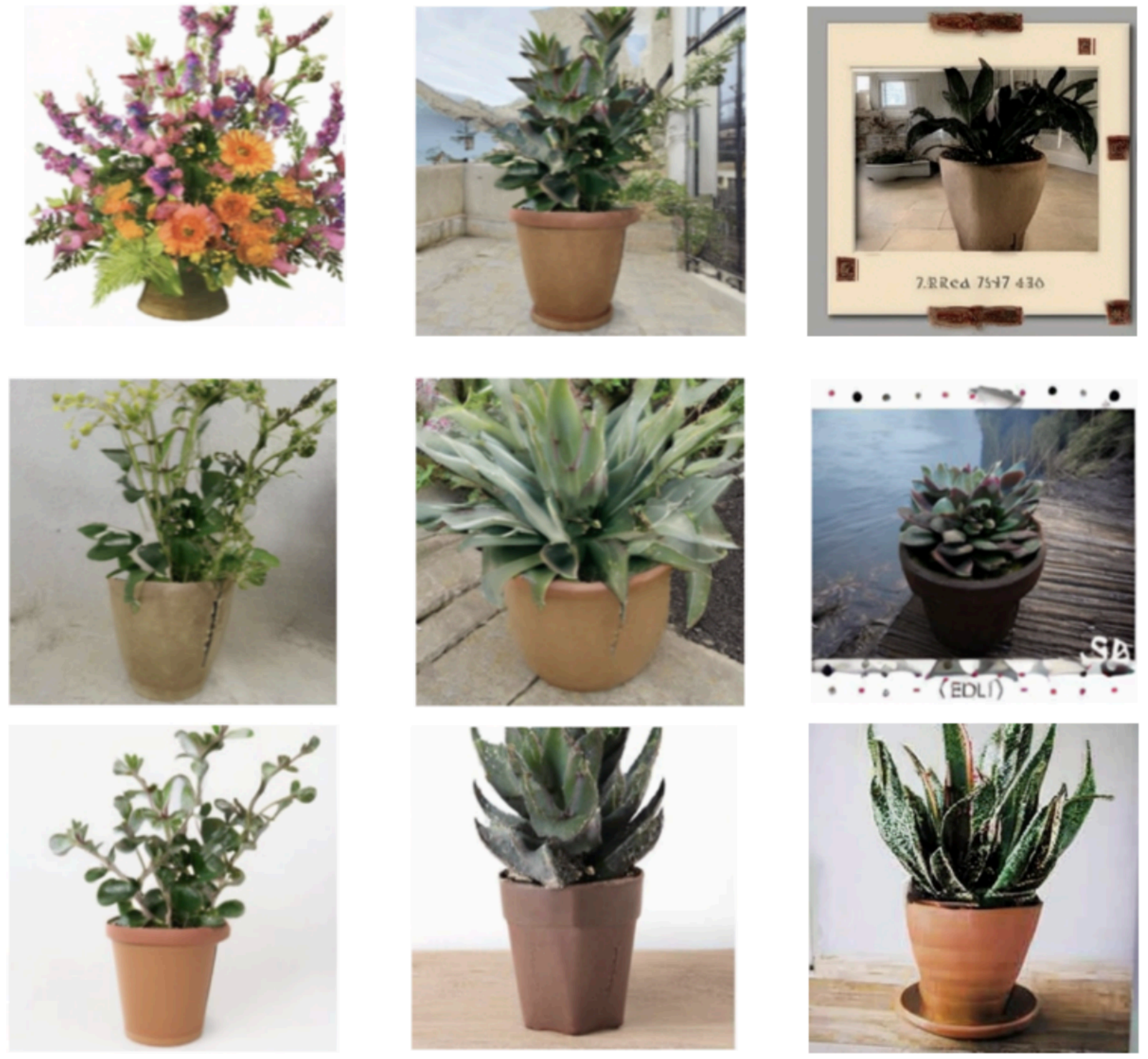
3. AR Priors

"A photo of a potted plant."

Uniform priors

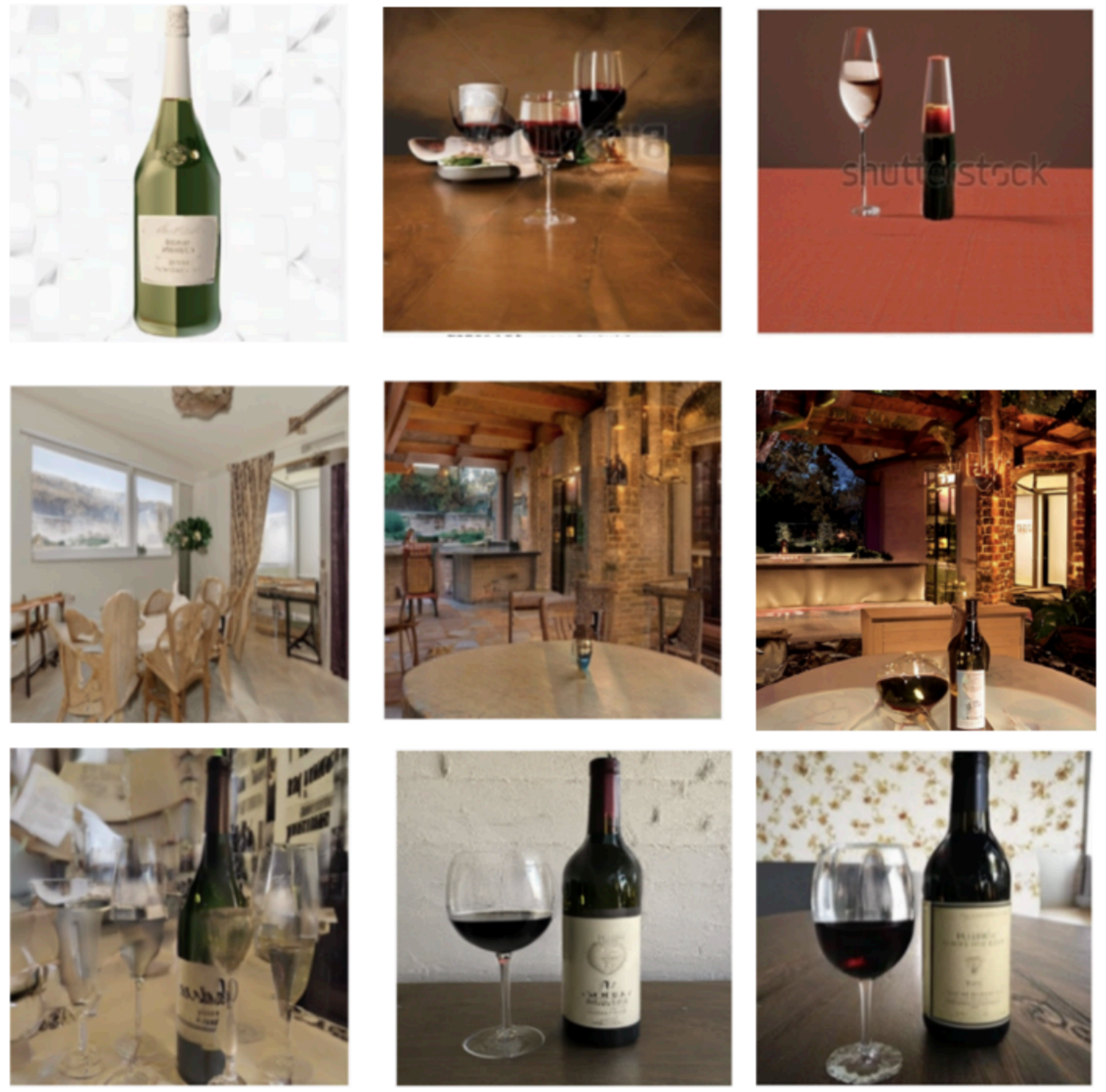
Unconditional priors

Conditional priors



Search over: Token 1 Token 8 Token 32

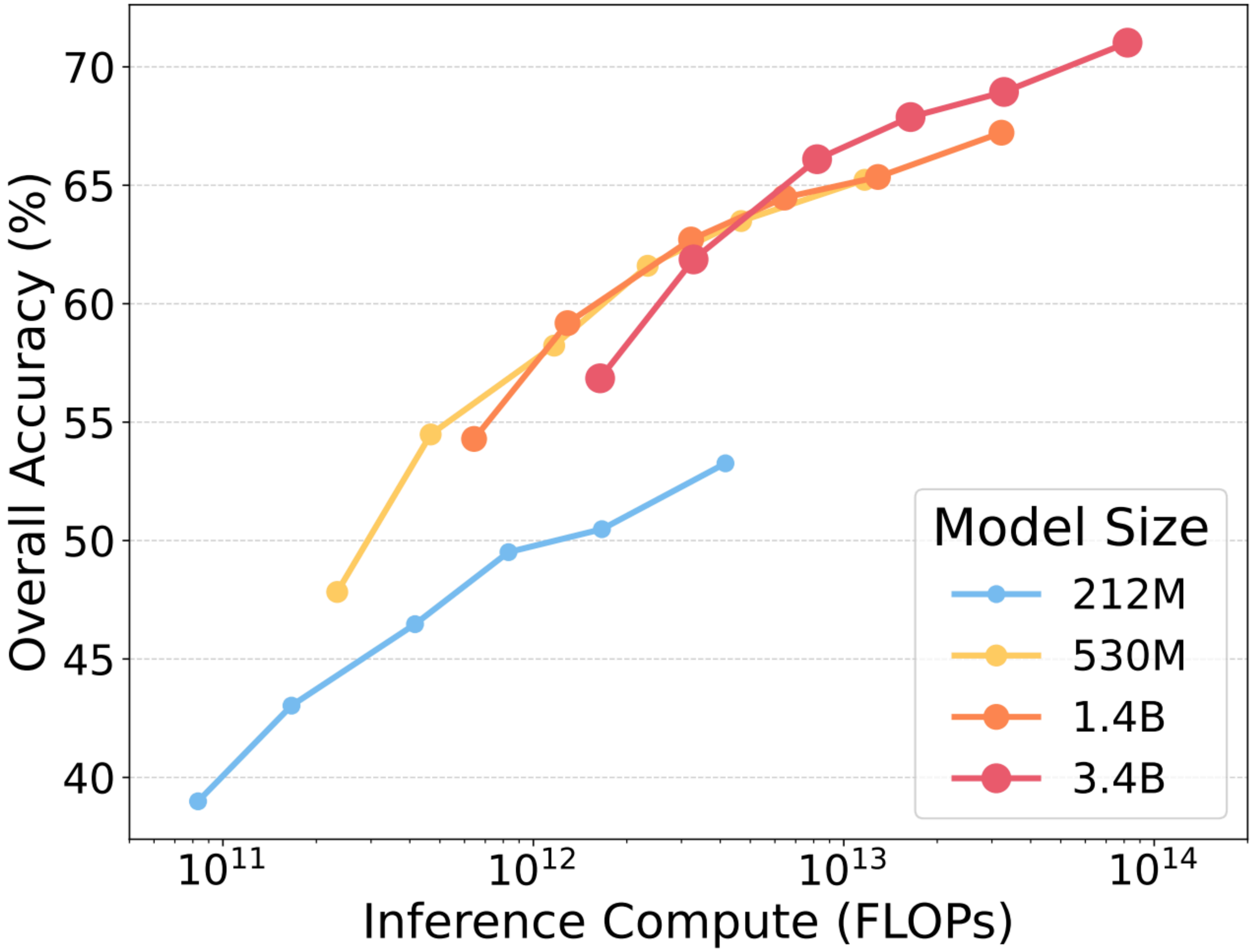
"A glass of wine is shown with a wine bottle right next to it on a table."

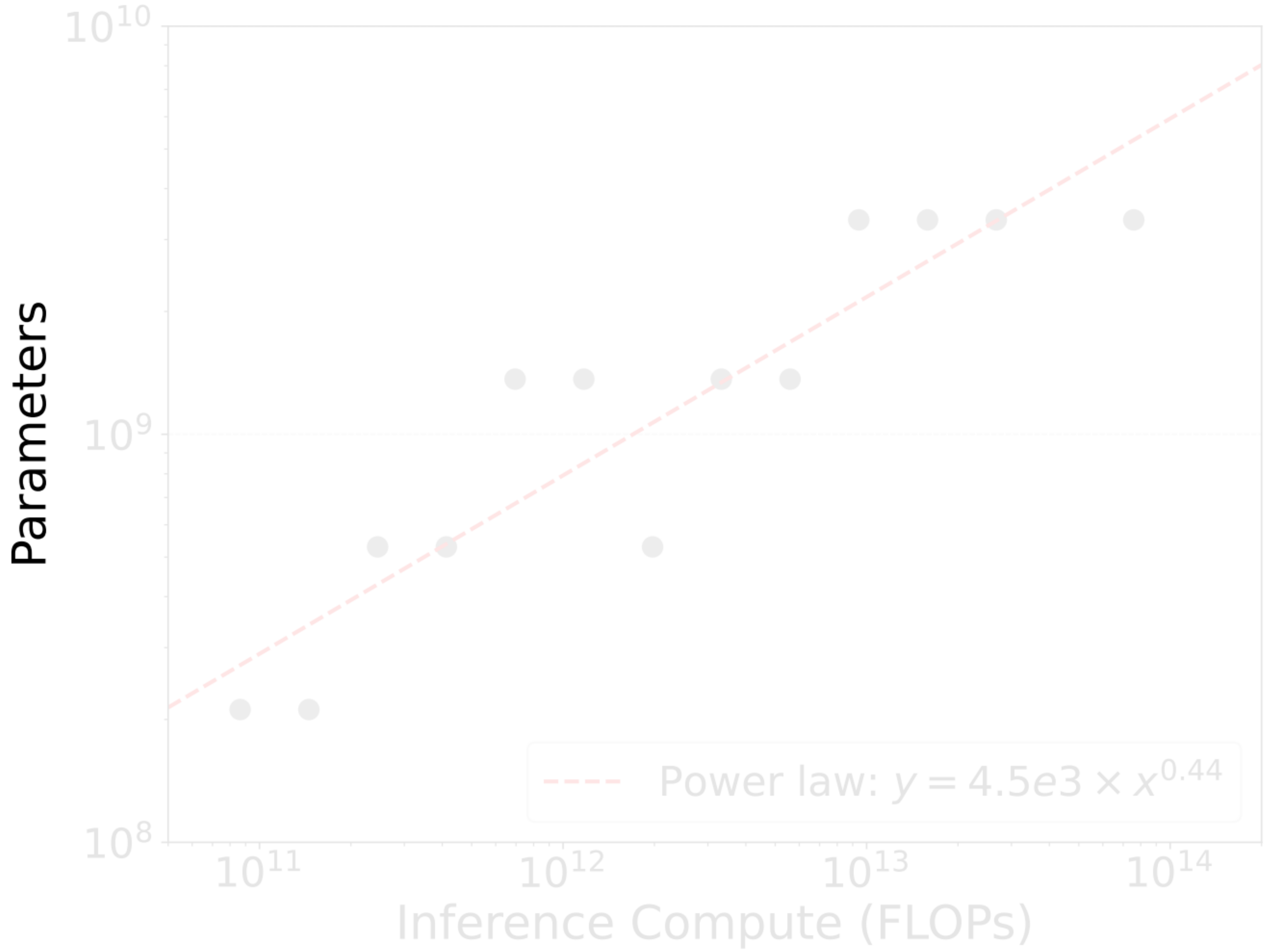


Token 1 Token 8 Token 32

Prior	Search	Single Object	Two Object
Uniform Prior	yes	79	32
Unconditional AR	yes	85	33
Conditional AR	yes	**100**	**81**
Conditional AR	no	97	48

Test-time scaling for different model sizes.





Conclusion

1. Image generation can be reformulated as a **search problem**. It is compatible with existing AR generation models.

2. Search **improves performance and steerability** for image generation, and even enables zero-shot multimodal controlling.

3. There are **four critical axes** of test-time search: token structure, search algorithm, verifier, and AR model, and we study their roles and interactions.

Future Work

1. **Verifier quality.**

2. **Adaptive Search.**

3. **Search over multiple images/frames/videos.**