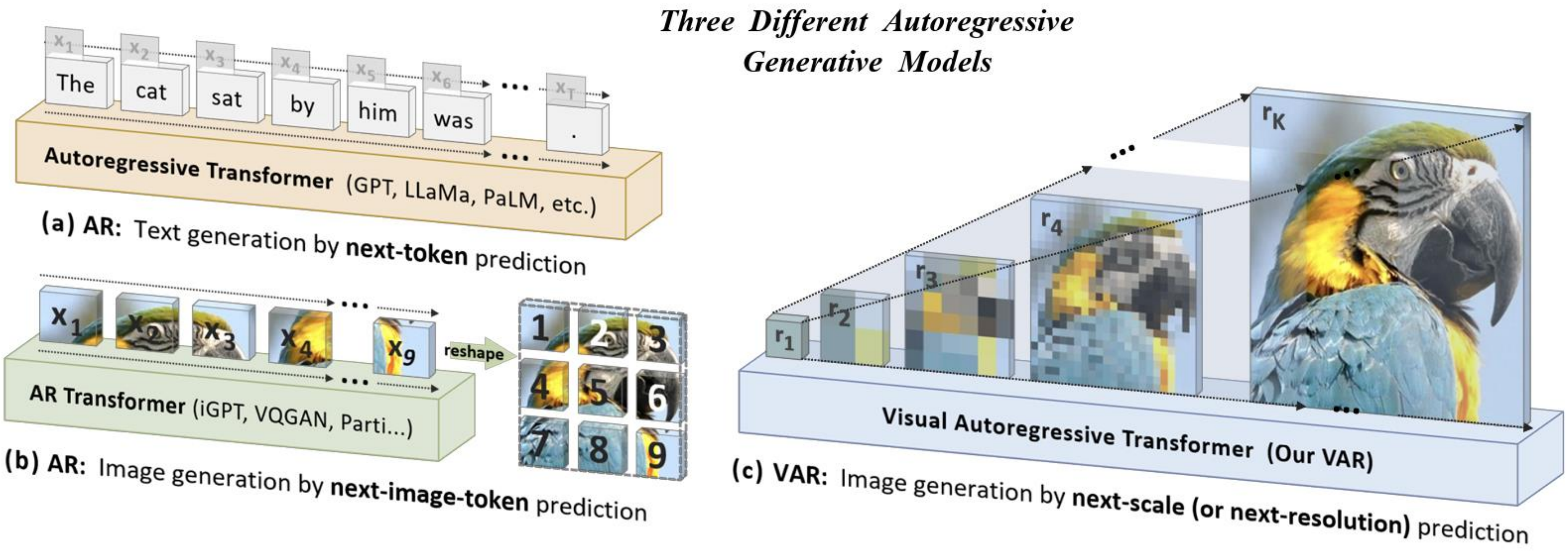
Frequency-Aware Autoregressive Modeling for Efficient High-Resolution Image Synthesis

Zhuokun Chen[1 2] Jugang Fan[1 2] Zhuowei Yu[3] Bohan Zhuang[4†] Mingkui Tan[1 2†]

[1] South China University of Technology [2] Pazhou Lab [3] University of California, Davis [4] Zhejiang University

Background & Motivation



- Text-to-image generation is widely used
- Autoregressive (AR) + next-scale prediction → strong quality & scalability
- Challenge: high-resolution stages involve **thousands of tokens** → expensive computation

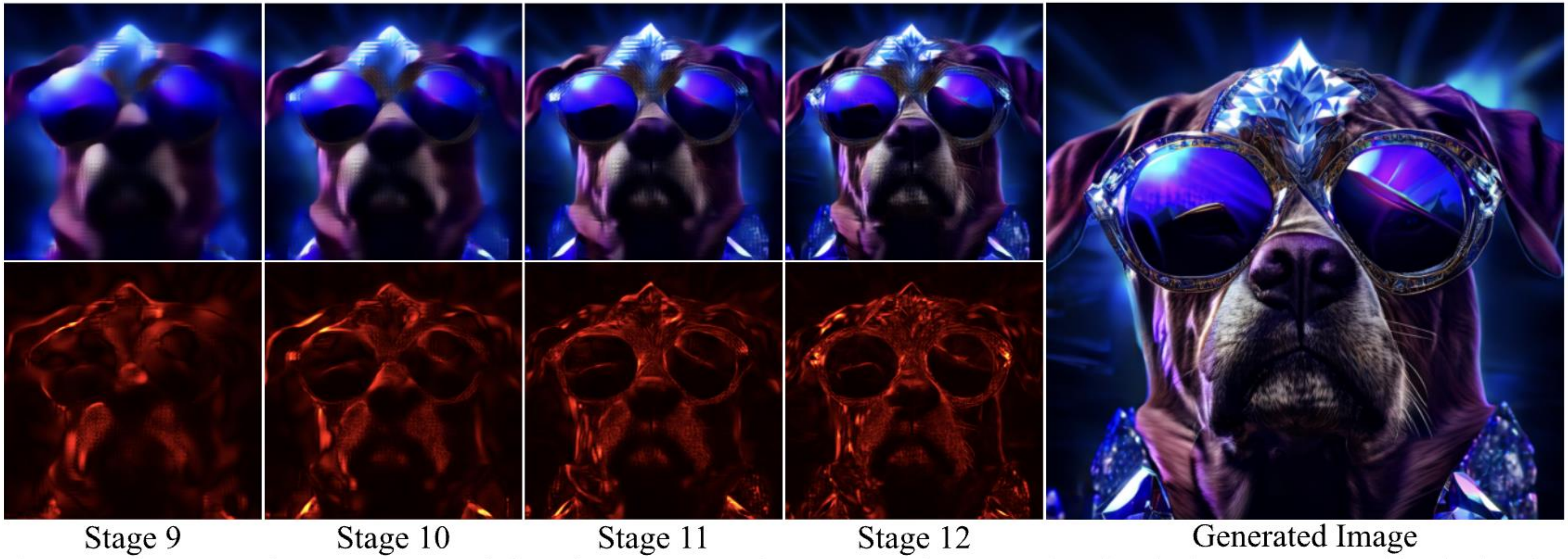
Related work of token reduction

- **Token Selection: keep only the most important tokens**
  - rank tokens by attention/saliency, drop the rest
  - **Limitation:** unsuitable for generative models → tokens are highly interdependent

- **Token Merging: reduce redundancy by combining similar tokens**
  - cluster tokens or merge by similarity
  - **Limitation:** clustering & similarity search very costly at high resolution → impractical for generation

Emprical Insights

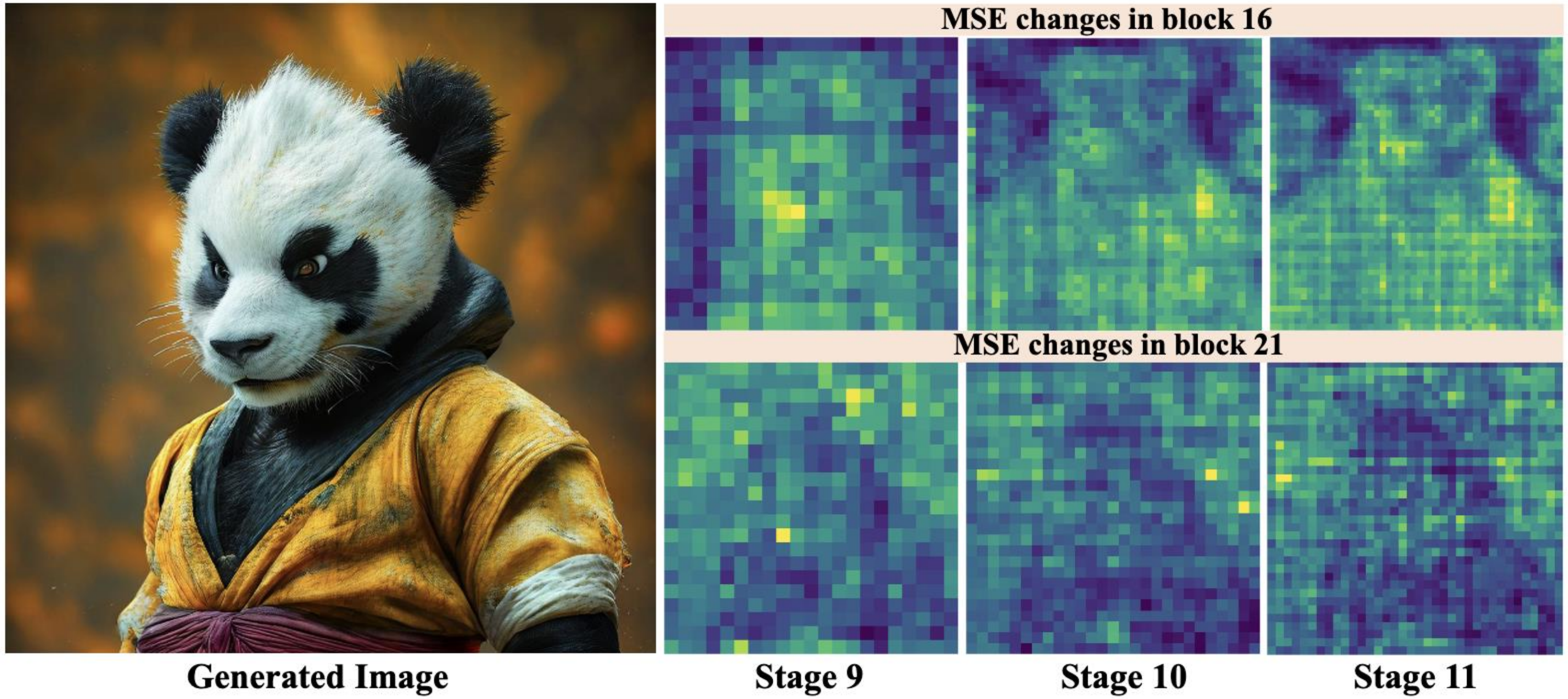
Observation 1: Residuals at high-resolution stages have minimal impact on low-freq regions



- Residuals concentrate on **high-frequency regions** (edges, textures)
- **Low-frequency regions** remain largely unchanged → redundancy

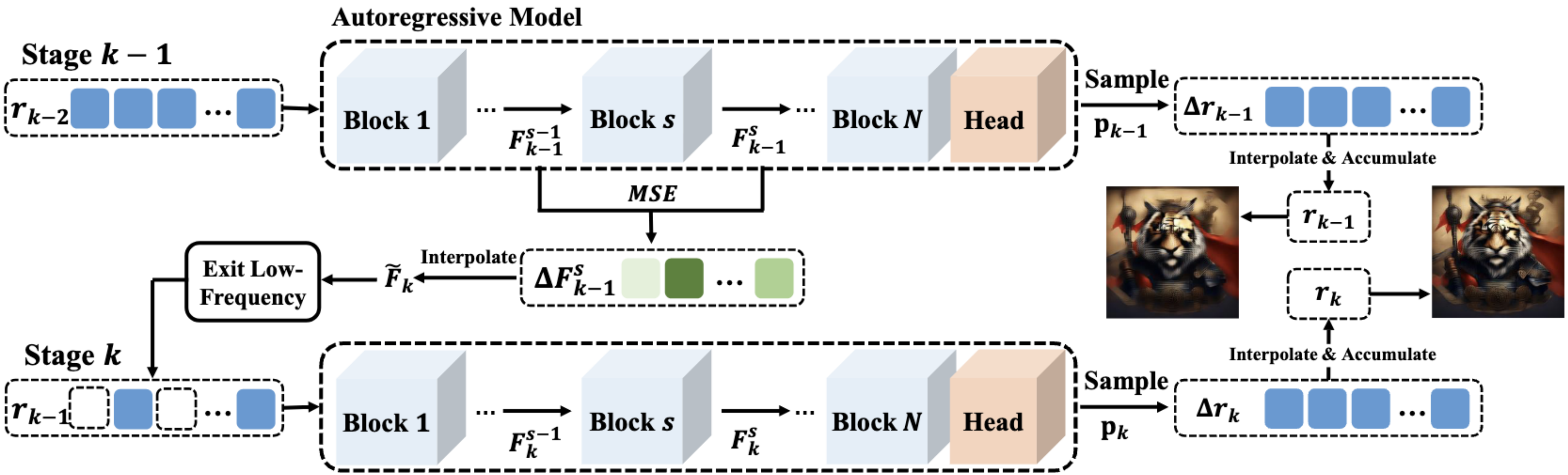
Emprical Insights

Observation 2: Different blocks in next-scale prediction models focus on distinct regions



- Block 16 → focuses more on high-frequency regions (e.g., contours, edges)
- Block 21 → emphasizes low-frequency regions (e.g., background)
- Insight: Block choice determines which regions are emphasized → enables dynamic distinction of high- vs. low-frequency regions

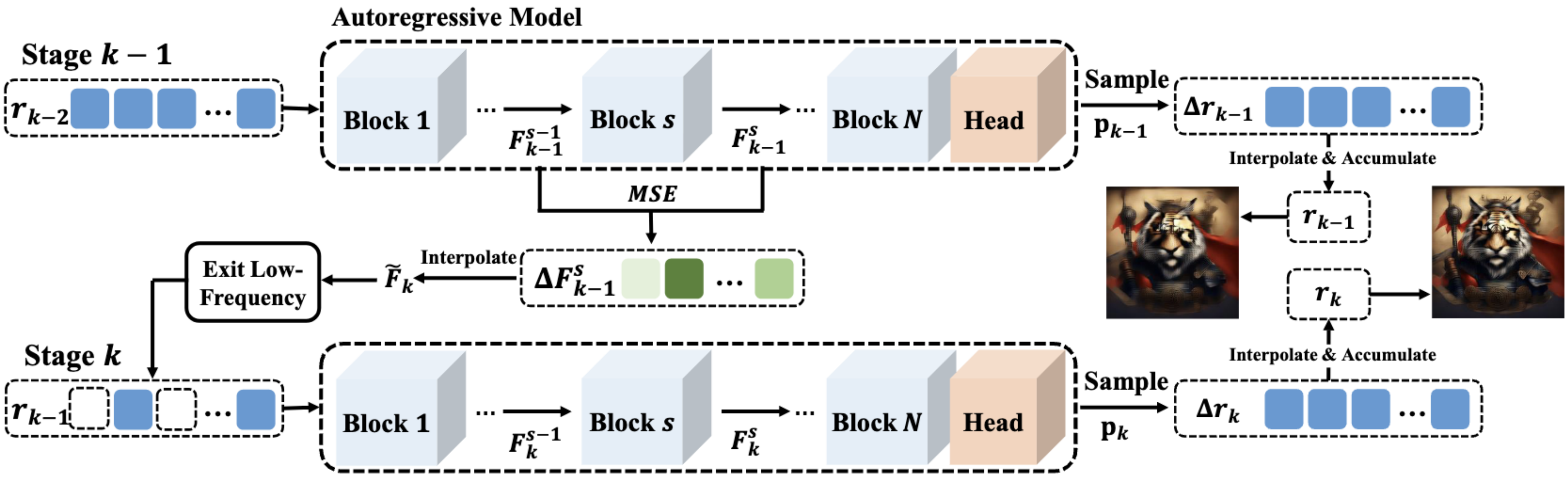
Method Overview



SparseVAR:

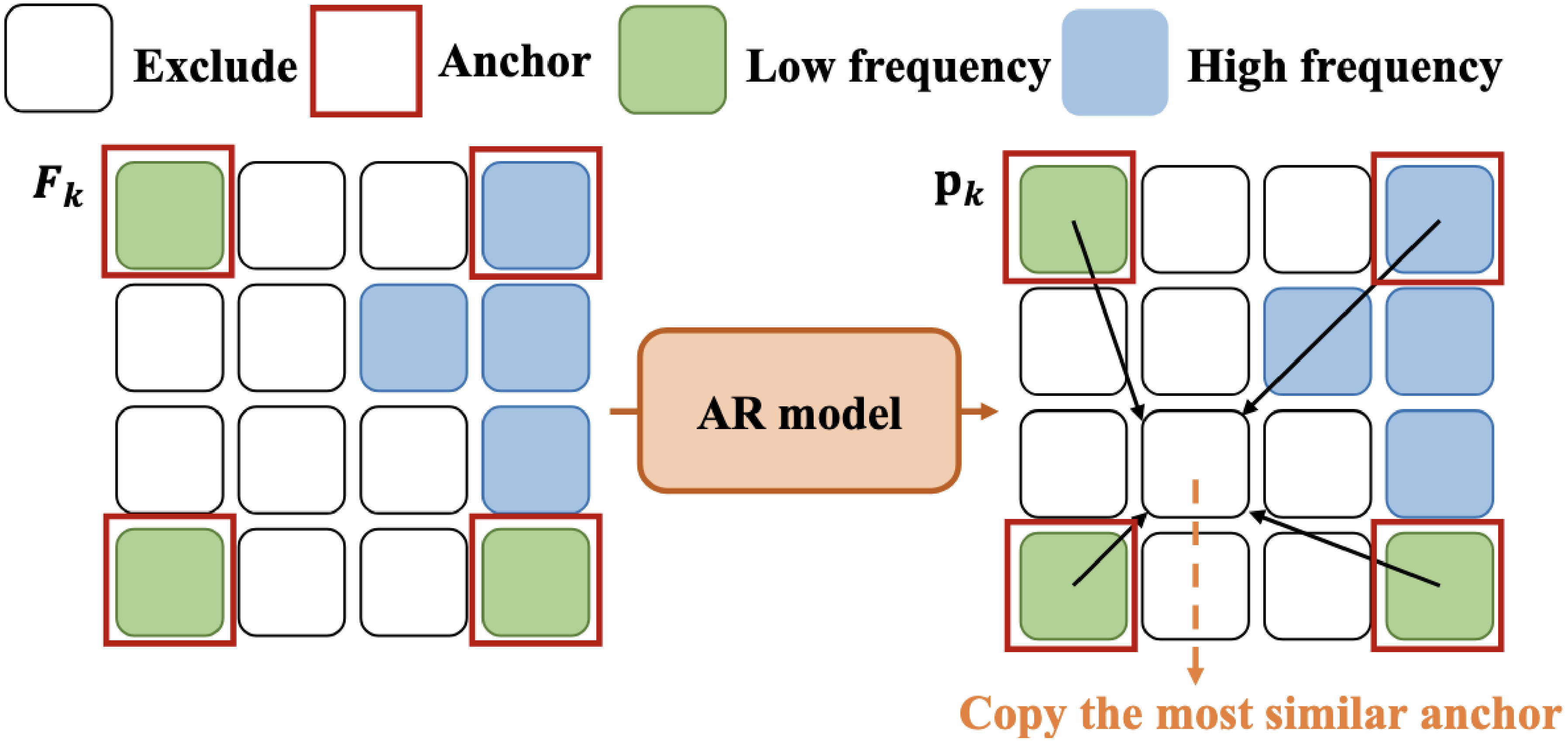
- Dynamic exclusion of low-frequency tokens
- Retaining anchor tokens for quality consistency
- Plug-and-play: no retraining required

Dynamic Exclusion



- Start from stage P
- Compute MSE change at a specific block
- Tokens below threshold τ → excluded

Keep anchor tokens



- One anchor per α×α grid
- Excluded tokens copy predictions from nearest anchor
- Maintain global fidelity

Performance on GenEval and DPG

Model	τ	GenEval↑							Latency (s)↓
		Two Obj.	Position	Color Attri.	Counting	Colors	Sin Obj.	Overall	
Infinity-2B	-	0.8586	0.4175	0.5525	0.6844	0.8431	1.0000	0.7260	2.78
+ SparseVAR	0.4	0.8485	0.4250	0.5625	0.7000	0.8457	1.0000	0.7303	2.64
	0.5	0.8359	0.4250	0.5600	0.6781	0.8351	1.0000	0.7224	1.87
	0.6	0.8409	0.4125	0.5475	0.6812	0.8404	1.0000	0.7204	1.47
	0.7	0.8460	0.4225	0.5475	0.6719	0.8378	1.0000	0.7209	1.36
HART-0.7B	-	0.6919	0.1625	0.2825	0.3688	0.8617	0.9938	0.5602	1.32
+ SparseVAR	0.4	0.7071	0.1450	0.2650	0.3938	0.8777	0.9906	0.5632	1.25
	0.5	0.7045	0.1600	0.2575	0.3969	0.8644	0.9906	0.5623	1.18
	0.6	0.7071	0.1600	0.2825	0.3562	0.8670	0.9906	0.5606	0.99
	0.7	0.6035	0.1200	0.2125	0.3344	0.8351	0.9656	0.5119	0.81

Model	τ	DPG-Bench↑			Latency (s)↓
		Global.	Relation	Overall	
Infinity-2B	-	0.8419	0.9283	0.8289	2.55
+ SparseVAR	0.4	0.8541	0.9246	0.8282	2.34
	0.5	0.8480	0.9242	0.8254	1.69
	0.6	0.8632	0.9237	0.8260	1.35
	0.7	0.8511	0.9270	0.8256	1.20
HART-0.7B	-	0.8710	0.9295	0.8099	1.31
+ SparseVAR	0.4	0.8571	0.9233	0.8092	1.24
	0.5	0.8602	0.9233	0.8082	1.19
	0.6	0.8602	0.9246	0.8069	1.00
	0.7	0.8663	0.9254	0.8072	0.83

- Infinity-2B: up to 2× faster, minimal quality drop
- HART-0.7B: ~25% faster, negligible degradation

Performance on Human Preference

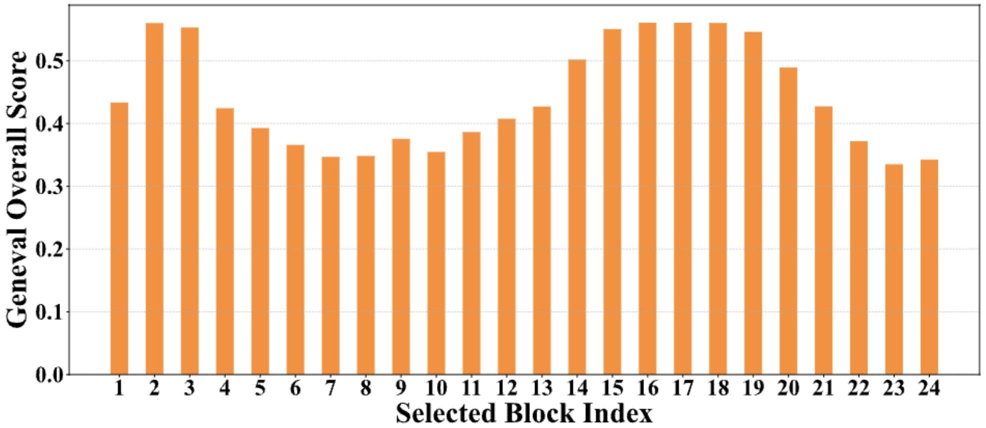
Model	τ	ImageReward		HPSv2.1					
		Score↑	Latency(s)↓	Anime	Concept-Art	Paintings	Photo	Overall↑	Latency(s)↓
Infinity-2B	-	0.9212	2.64	31.63	30.26	30.28	29.27	30.36	2.61
+ SparseVAR	0.4	0.9147	2.37	31.58	30.13	30.16	29.22	30.27	2.35
	0.5	0.8969	1.77	31.40	29.95	29.96	29.05	30.09	1.79
	0.6	0.8943	1.42	31.29	29.82	29.77	28.94	29.95	1.40
	0.7	0.8946	1.33	31.21	29.75	29.71	28.88	29.89	1.32
HART-0.7B	-	0.8656	1.32	31.22	29.61	29.10	28.21	29.53	1.30
+ SparseVAR	0.4	0.8818	1.25	31.19	29.58	29.08	28.19	29.51	1.26
	0.5	0.8818	1.20	31.06	29.47	28.96	28.09	29.40	1.19
	0.6	0.8121	1.01	30.25	28.68	28.13	27.51	28.64	1.02
	0.7	0.4333	0.82	27.18	25.60	25.13	24.93	25.71	0.81

- Human evaluations show consistent quality
- ~50% latency reduction

Ablation Studies

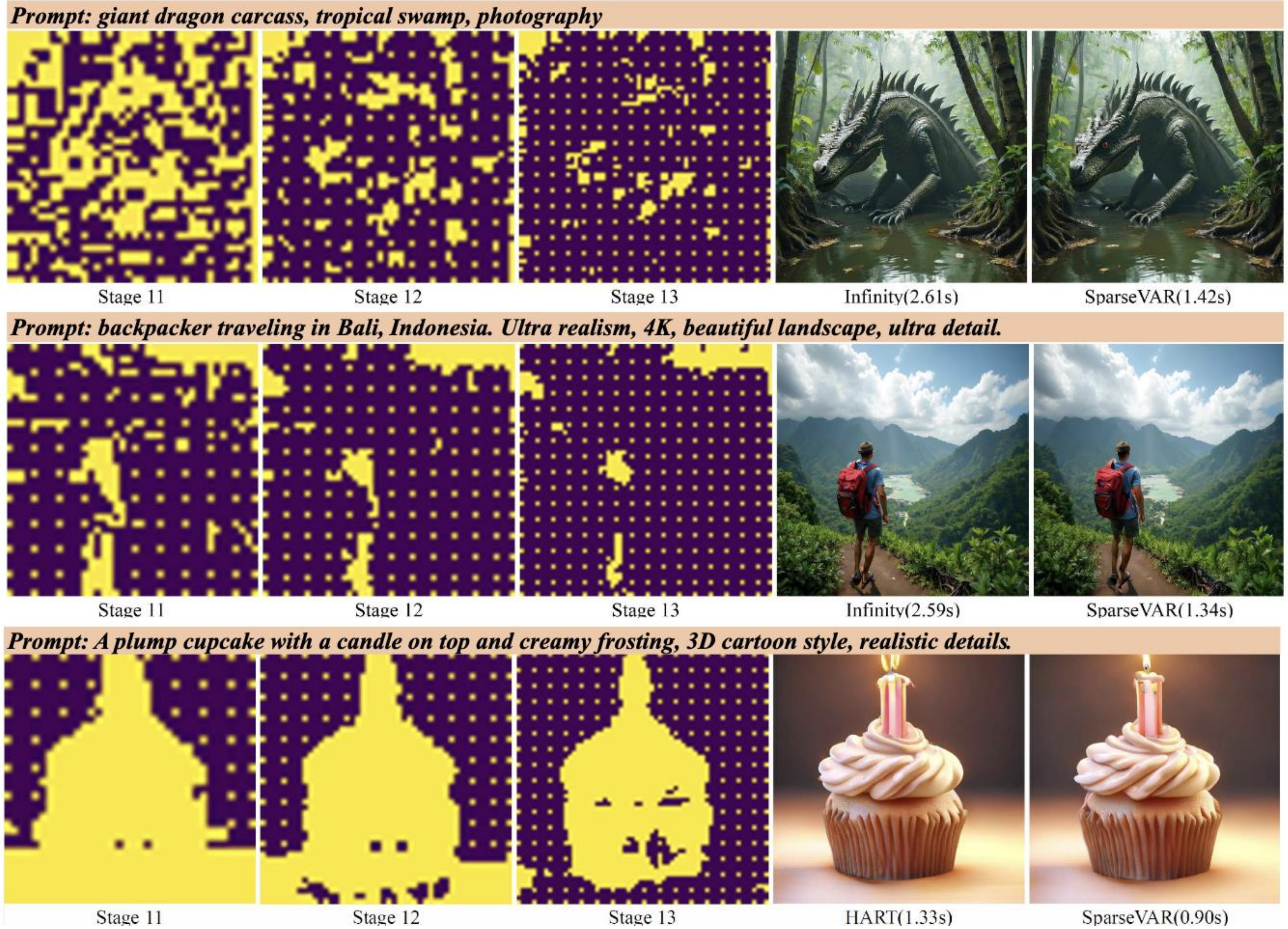
α	Infinity		HART	
	Score	Latency(s)	Score	Latency(s)
2	0.7235	1.76	0.5615	1.11
3	0.7210	1.54	0.5578	1.06
4	0.7204	1.47	0.5606	0.99
5	0.7200	1.46	0.5560	0.97
-	0.7190	1.38	0.5502	0.93

P	Infinity		HART	
	Score	Latency(s)	Score	Latency(s)
6	0.6805	1.29	0.5529	0.98
8	0.7085	1.35	0.5577	0.98
9	0.7126	1.39	0.5565	0.99
10	0.7204	1.47	0.5602	0.99
11	0.7261	1.76	0.5625	1.03
12	0.7274	2.21	0.5607	1.05



- Anchor tokens → essential for quality
- τ & P → control quality-speed trade-off
- Block selection matters (16th block best)

Visualizaitons



Conclusion

- SparseVAR: frequency-aware plug-and-play acceleration
- Efficient high-resolution autoregressive generation
- Future: broader autoregressive tasks, adaptive anchor strategy

Paper: https://arxiv.org/abs/2507.20454
Code: https://github.com/Caesarhhh/SparseVAR

Thanks for listening!