Text-Conditioned Sampling Framework for Text-to-Image Generation with Masked Generative Models

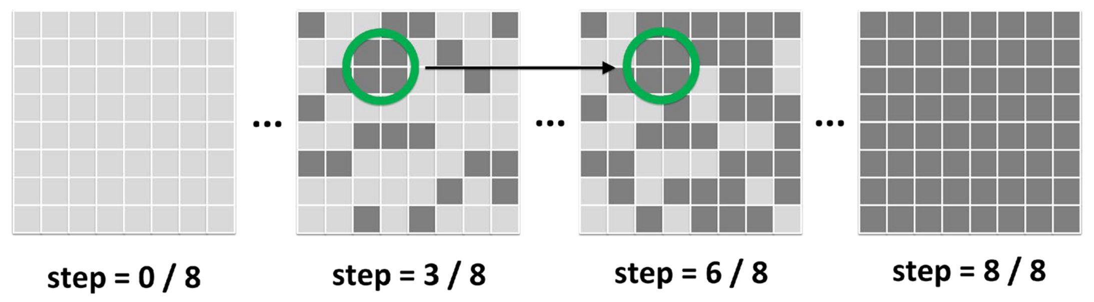
Jaewoong Lee[1], Sangwon Jang[1], Jaehyeong Jo[1], Jaehong Yoon[1], Yunji Kim[2], Jinhwa Kim[2], Jungwoo Ha[2], Sungju Hwang[1]

[1]KAIST, [2]NAVER AI Lab

KAIST · NAVER AI · ICCV23 PARIS

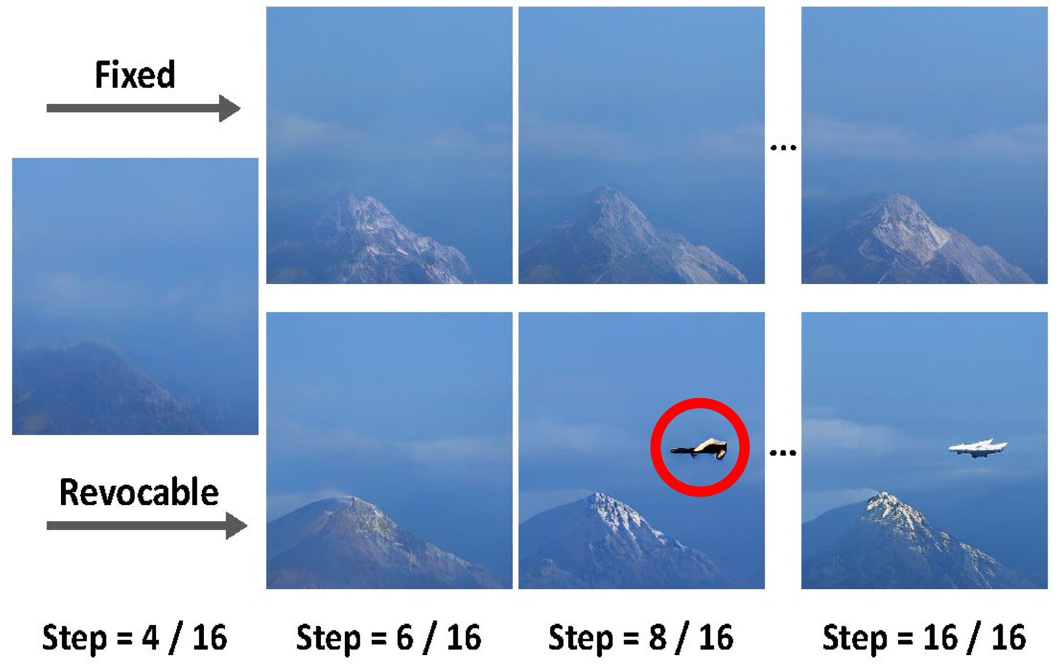
Introduction

Fixed Sampling in Masked Generative Models



- In fixed sampling, tokens that are once sampled cannot be revised afterward negatively affecting the text alignment.

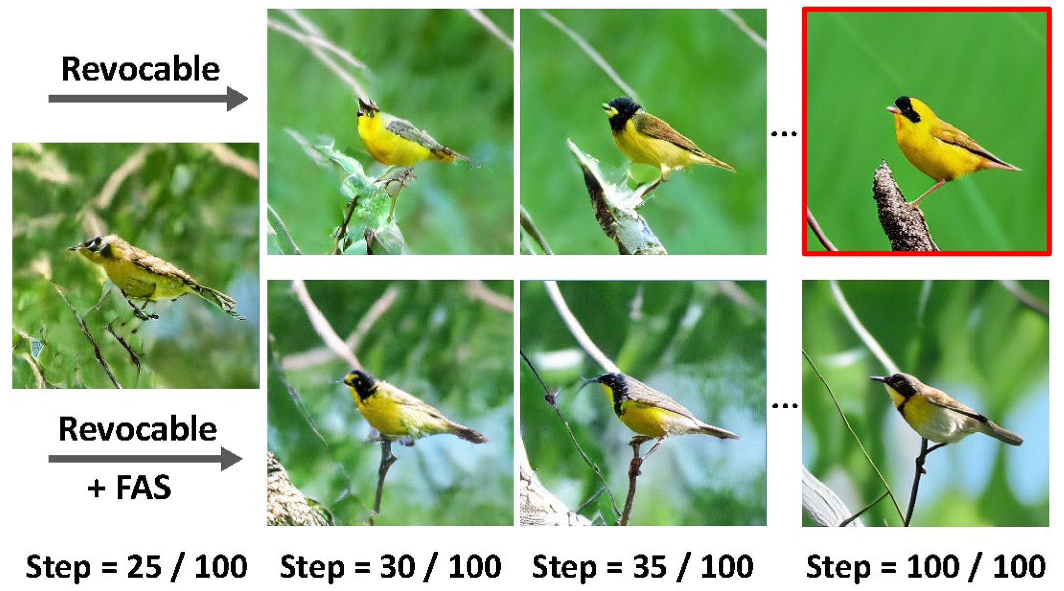
Revocable Sampling



"A view of the end of an ***airplane*** *in the sky over mountains"*

- Compared to revocable methods, fixed method fixes the misaligned tokens causing joint distribution issue.

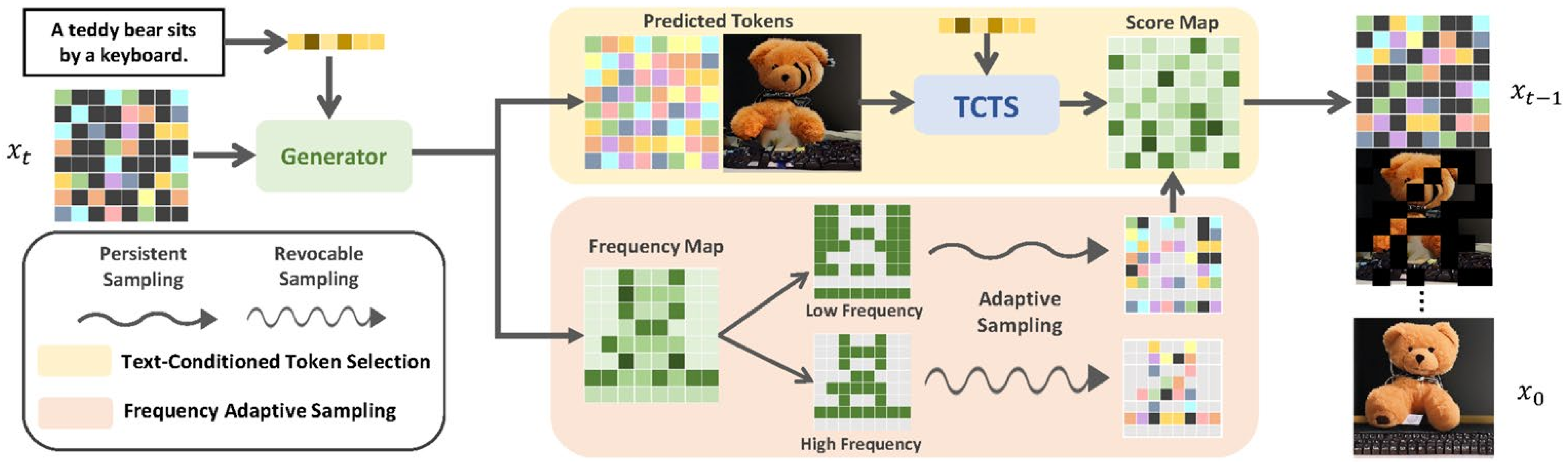
Over-simplification with Revocable Sampling



- In longer step generation, revocable methods cause over-simplification in the low-frequency areas because of excessive resampling.

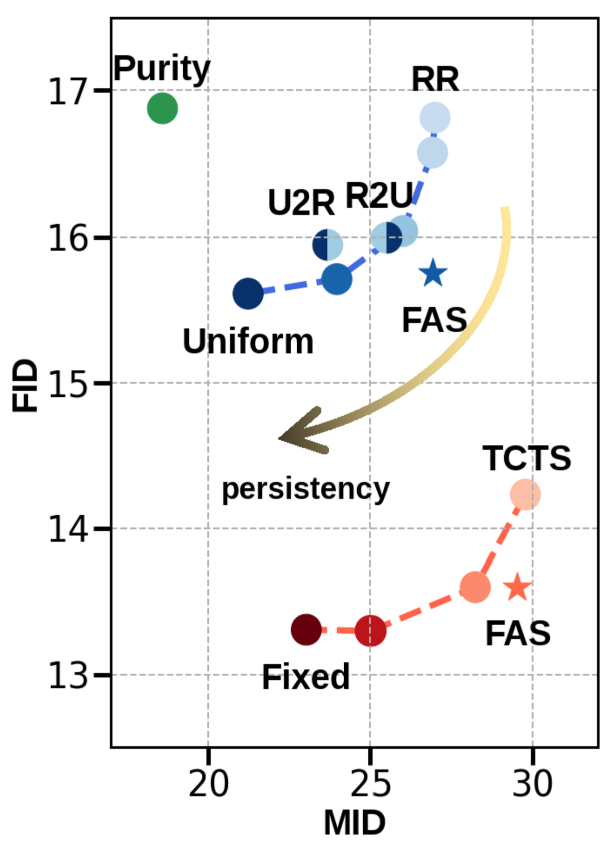
Proposed Method

Text-Conditioned Token Selection with Frequency Adaptive Sampling



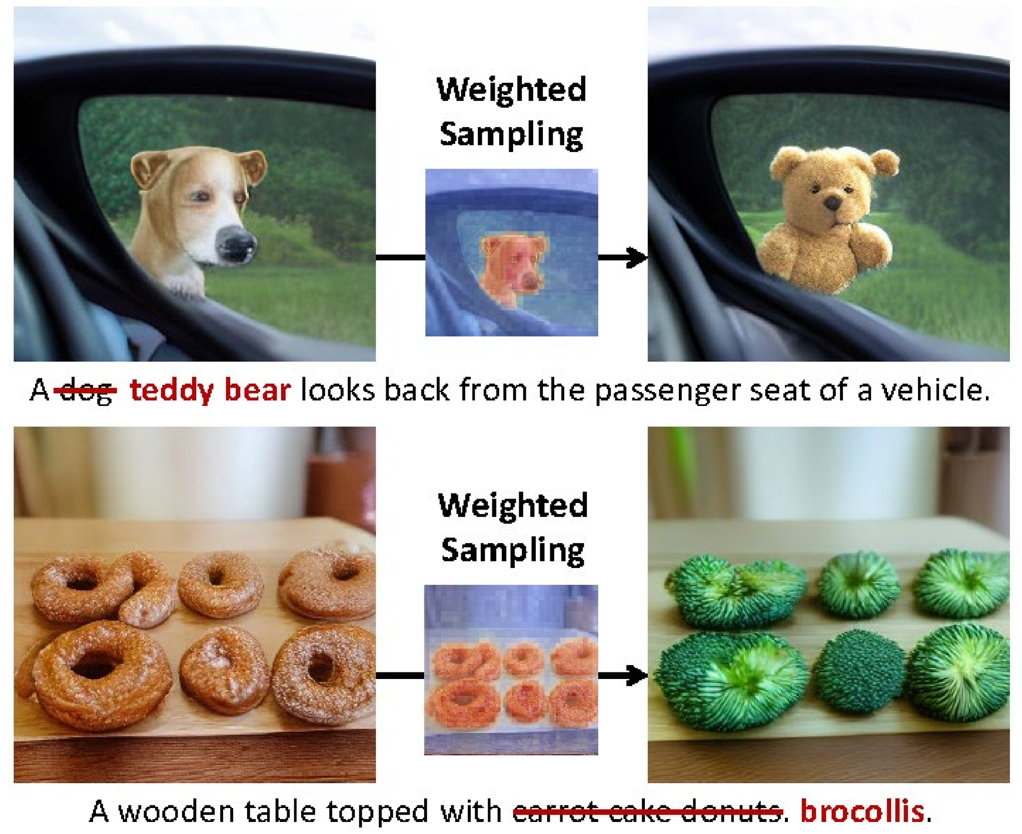
- Our model consists of two main components, **Text-Conditioned Token Selection (TCTS)** and **Frequency Adaptive Sampling (FAS)**.
- TCTS exploits the text condition to detect misaligned tokens and alleviate the error accumulation.
- FAS utilizes the generator's **self-attention map** to limit resampling only in the low-frequency areas preventing over-simplification.

Trade-off between Text Alignment (CLIPs, MID) and Image Quality (FID)



- A trade-off can be seen that as the sampling strategy gets closer to the RR sampling, text alignment gets better, while the image quality gets worse.
- TCTS can generate high-quality images with improved text-alignment in even fewer sampling steps enhancing the trade-off compared to the naive generative model.
- Our model (TCTS + FAS) successfully generates high-quality images which contain a clear semantic connection to the given text captions.

Mask-free Object Editing with Cross-Attention Map



- We can leverage a cross-attention map corresponding to the word of the object instead of self-attention.
- It can perform mask-free editing that better preserves the original content with fewer steps.

Experiments

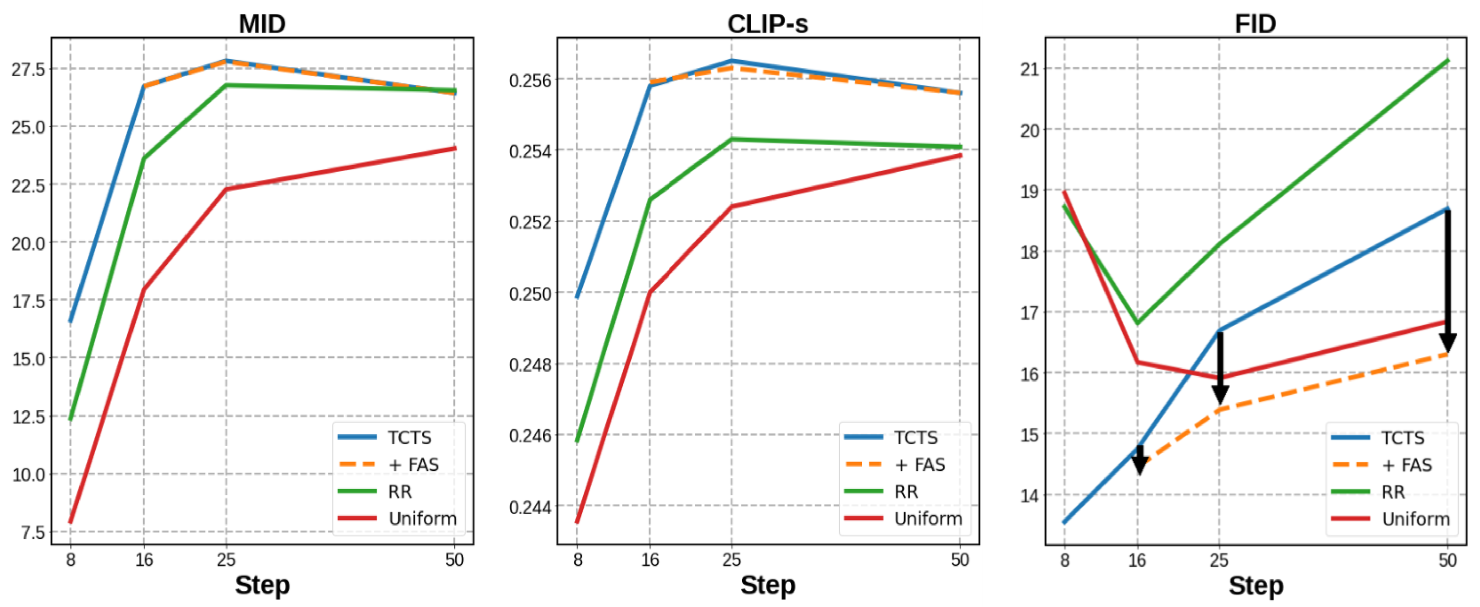
Quantitative Evaluation of Sampling Methods

Step	Method	MID-L ↑	SOA-I ↑	CLIP-S ↑	FID-30K ↓
16	Purity	11.02	72.38	0.2474	19.20
	Uniform	17.94	74.80	0.2500	16.17
	RR	23.60	78.79	0.2526	17.10
	TCTS + FAS	**26.72**	**79.52**	**0.2559**	**14.45**
25	Purity	16.84	75.21	0.2487	18.39
	Uniform	22.27	77.08	0.2524	15.91
	RR	26.77	**81.10**	0.2543	18.43
	TCTS + FAS	**27.79**	80.87	**0.2563**	**15.39**

Step	Method	MID-L ↑	CLIP-S ↑	FID ↓
16	Purity	-24.21	0.2410	15.21
	Uniform	-25.60	0.2404	16.57
	RR	-25.03	0.2371	17.38
	TCTS + FAS	**-19.88**	**0.246**	**12.35**
25	Purity	-21.26	0.2384	**12.60**
	Uniform	-23.04	0.2396	13.02
	RR	-23.29	0.2364	14.53
	TCTS + FAS	**-18.31**	**0.2409**	13.67

- We can observe the trade-off between CLIP score (MID) and FID, and ours outperforms other baselines in most of the metrics.

Performance Comparison with Different Steps



- TCTS outperforms other baseline methods in terms of MID and CLIP scores, while slightly compromising the FID score in longer step generation.
- FAS significantly enhances the FID score of TCTS without compromising the alignment between the image and text.

Conclusion

- We empirically find that the revocable sampling significantly improves the text alignment yet degrades the quality of the generated images
- We propose a simple token sampling strategy TCTS with guidance sampling training, pushing the boundary of the trade-off between image quality and text alignment.
- We find that collaborative sampling in a persistent and revocable manner (FAS) surprisingly alleviates over-simplification issues in the generated backgrounds.