KLASS :

KL-Guided Fast Inference in Masked Diffusion Models

Seo Hyun Kim*, Sunwoo Hong*, Hojung Jung, Youngrok Park, Se-Young Yun (* Equal contribution)

Paper

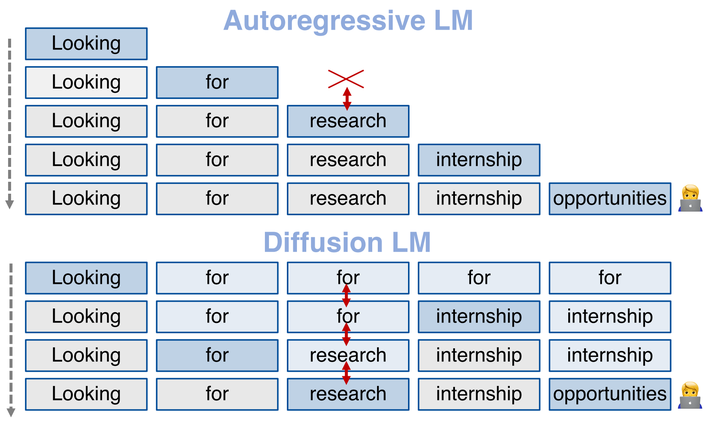
Code

NEURAL INFORMATION PROCESSING SYSTEMS

KAIST AI
Kim Jaechul Graduate School

- **Plug-and-play SOTA sampler with up to 2.78x inference speedup**
- **Unmasks stable tokens in parallel using token-level KL divergence**

Why KL?

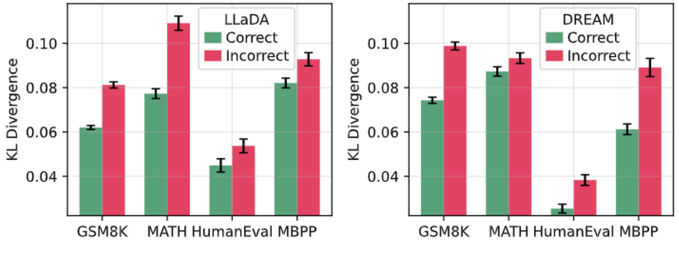


- **Iteratively predicts** x0 at each step
  → enables **KL** measurement
- **Parallel generation** → potential speedup

Existing Problem

Standard samplers use **fixed** samplers.
Too slow!

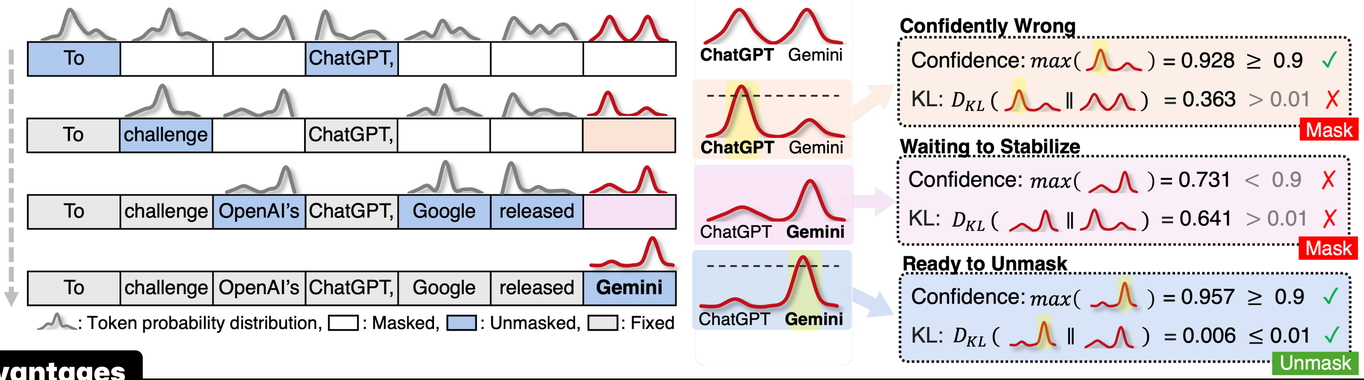
Auxiliary planners add significant computational **overhead** and **latency**.
Too heavy!



→ **Low KL** serves as an indicator of **correctness**

How does KLASS Work?

KL-Adaptive Stability Sampling: Unmask multiple tokens in parallel when they are both **confident** and **stable**



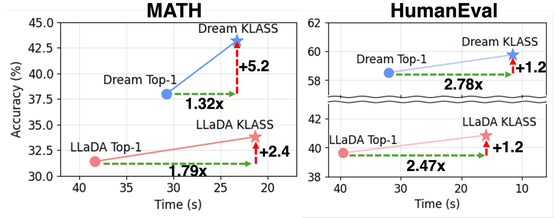
Key Advantages

✓ **Adaptive Parallelism**: Unmasking count is **not fixed**, speeding up on **stable** tokens & slowing down to refine **unstable** ones
✓ **Plug-and-Play Efficiency**: A strictly **training-free** method using only existing logits
+ **negligible overhead** (<1.6% memory, <0.2% latency) and **no external planners**

What are the Results?

✗ **Top-k Confidence**
… Therefore, the number of cars that drove through in the first 15 minutes is:\n \\[\n 25 - 20 = 10\n \\]
Therefore, the number of cars that drove through the traffic jam in the first 15 minutes is \\(\\boxed{ 10 }\\).
conf: 0.9241, kl: 0.4517

✓ **KLASS**
… Therefore, the number of cars that drove through in the first 15 minutes is:\n \\[\n 25 - 20 = 5\n \\]
Therefore, the number of cars that drove through the traffic jam in the first 15 minutes is \\(\\boxed{ 5 }\\).
conf: 0.7587, kl: 0.0193



Method	Parallel	MATH Acc↑	MATH Steps↓	GSM8K Acc↑	GSM8K Steps↓	HumanEval Acc↑	HumanEval Steps↓	MBPP Acc↑	MBPP Steps↓
LLaDA									
Top-1	✗	31.4	256	75.13	256	39.63	256	46.69	256
Random	✗	26.2	256	67.10	256	20.21	256	29.18	256
Top-2	✓	29.6	128	72.40	128	33.54	128	37.74	128
Confidence	✓	31.6	96.46	75.21	74.35	37.80	54.41	47.08	85.20
KL divergence	✓	32.6	172.21	74.52	155.88	40.24	111.93	45.53	150.47
KLASS (ours)	✓	**33.8**	128.62	**76.50**	98.57	**40.85**	91.98	**47.86**	119.59
Dream									
Top-1	✗	37.97	256	**79.55**	256	58.53	256	63.81	256
Random	✗	18.73	256	37.35	256	18.09	256	28.14	256
Top-2	✓	33.60	128	71.69	128	42.88	128	47.08	128
Confidence	✓	41.80	95.10	73.67	74.81	50.00	52.47	57.59	72.49
KL divergence	✓	41.27	162.49	76.70	150.02	**59.35**	73.94	62.65	108.15
KLASS (ours)	✓	**43.20**	149.72	79.43	155.67	**59.35**	74.88	**64.59**	111.24

✓ **Accurate**: Achieves SOTA performance, boosting accuracy by up to +5.23%
✓ **Fast**: Cuts sampling steps by 40-70%, accelerates up to 2.78x wall-clock speedup

Text Generation (Unconditional)		
Method	MAUVE ↑	PPL (LLaMA3) ↓
AR	0.880	14.9
SEDD	0.022	110.2
MDLM	0.115	54.3
KLASS (Ours)	**0.179**	**49.0**

Image Generation (Step 16)		
Method	FID ↓	IS ↑
Confidence	34.48	75.72
KLASS (Ours)	**30.48**	**93.07**

Molecular Generation (QED)		
Method	Reward ↑	NFEs ↓
MDLM	0.526	32.0
KLASS (Ours)	**0.546**	**18.8**

✓ **Universal**: Proven effective across Text, Image, and Molecular generation tasks