Gumbel Distillation for Parallel Text Generation

ICLR

TEXAS
The University of Texas at Austin

Chi Zhang*[1], Xixi Hu*[1], Bo Liu[1], Qiang Liu[1]
[1] The University of Texas at Austin,
* Equal Contribution

Paper Code Personal

Thanks for stopping by!

Let's chat:
chizhang@cs.utexas.edu

Motivation

> Autogressive LMs are high-quality but only decode one token at a time during inference

> Parallel decoders are fast but assume *conditional independence*, causing repetition and incoherrence and thus lower quality

> **Goal:** make parallel decoders better capture the *joint* distribution from an AR teacher

Key Idea: Gumbel as a blueprint

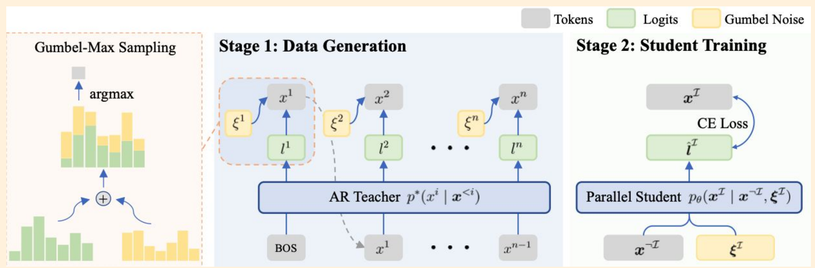
> Gumbel-Max Trick turns AR sampling into a *deterministic* function of gumbel noise ξ

$$x^i = \arg\max_k \left(l_k + \xi_k\right)$$

> Train the student to reconstruct tokens given the paired gumbel noise from the AR teacher

$$\mathcal{L} = -\mathbb{E}\left[\log p_\theta(\boldsymbol{x}^{\mathcal{I}} \mid \boldsymbol{x}^{\neg\mathcal{I}}, \boldsymbol{\xi}^{\mathcal{I}})\right]$$

Distributon Matching => Supervised Learning



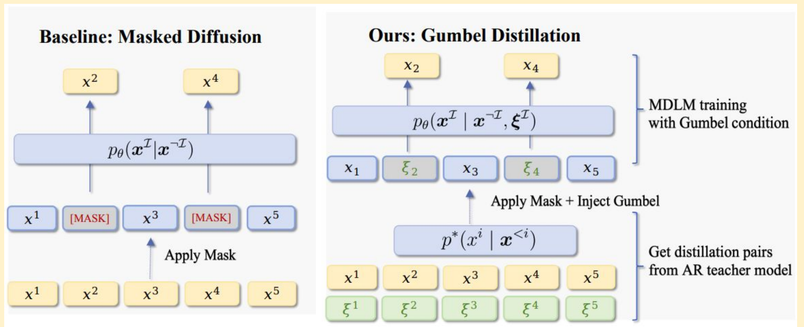
Stage I: Get (gumbel, token) pairs from teacher

Stage II: Train student conditional on gumbel

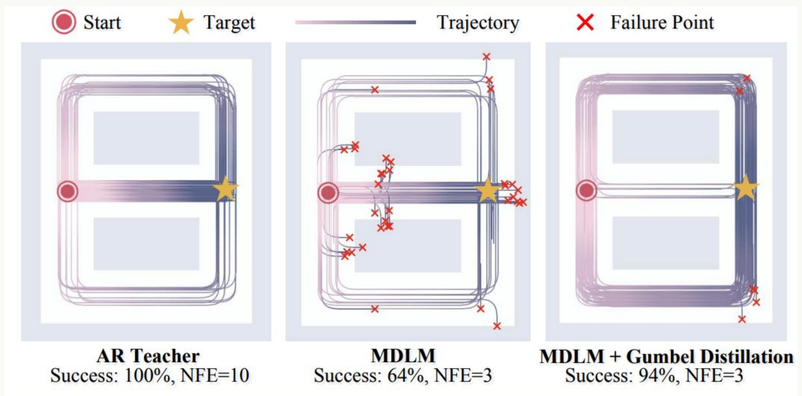
Plug-and-play Integration

> **Masked Diffusion LMs** (MDLM, BD3-LM):
- Modify <MASK> with Gumbel embeddings
- Same training procedure + objective

> **Multi-token Prediction** (Medusa):
- Feed Gumbel signal into each prediction head
- Same training procedure + objective

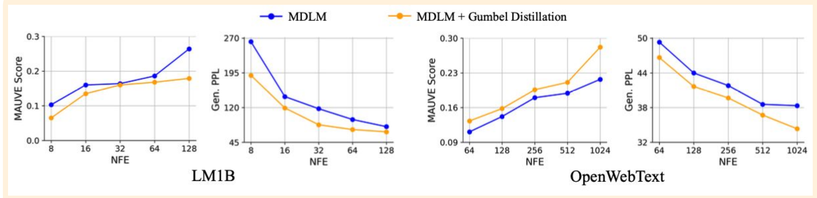


Toy Case: Maze Navigation



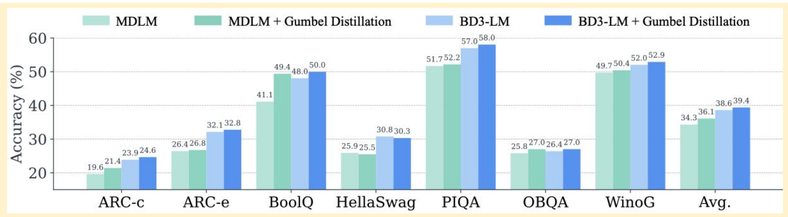
> Baseline MDLM fails: conditional independence
> +Gumbel Distillation matches AR at low NFE

Generation Quality



Model	Gen-PPL ↓	MAUVE ↑
MDLM	38.34	0.217
+ Gumbel	**34.33**	**0.282**
BD3-LM ($L'=4$)	26.40	0.251
+ Gumbel	**24.37**	**0.304**

Zero-shot Benchmarks



MDLM avg. acc.: 34.3→36.1%; BD3-LM: 38.6→39.4%

Multi-token Prediction

Backbone	Head 1	Head 2	Head 3
GPT-2-Small	+4.5%	+11.0%	**+22.0%**
Vicuna-7B	+8.9%	+28.1%	**+37.6%**

> Farther positions benefit more from conditioning
> Scales from 120M to 7B w/o architectural changes