GRIFFIN: Effective Token Alignment for Faster Speculative Decoding

Shijing Hu[1], Jingyang Li[2], Xingyu Xie[2], Zhihui Lu[1], Kim-Chuan Toh[2], Pan Zhou[3]
[1]Fudan University, [2]National University of Singapore, [3]Singapore Management University

Overview

- We propose **GRIFFIN**, a novel speculative decoding framework that addresses **token misalignment** between training and decoding phases.
- We design a **token-alignable training strategy** with dynamic loss masking that excludes highly misaligned tokens during training.
- We introduce a **token-alignable draft model** with Token-Guided Fusion (TGF) module to correct feature inconsistencies.
- GRIFFIN achieves **8% higher acceptance length** and **7% speedup ratio** compared to state-of-the-art methods on diverse LLMs.

Motivation

Background: Speculative decoding employs a lightweight draft model $\mathcal{M}$ to generate multiple tokens through multiple forward passes, then a target model $\mathcal{T}$ verifies them in a single pass.

The Problem: EAGLE-style speculative decoding methods operate at the feature level predicting hidden states $\mathbf{F}_t$ from $\mathcal{T}$'s final layer. This creates a critical train-decode mismatch. This mismatch causes two fundamental issues:

1. **Feature Misalignment:** Using $\mathbf{F}_t$ instead of $\bar{\mathbf{F}}_t$ for prediction.
2. **Token Misalignment:** Draft tokens $\bar{\mathbf{x}}_t$ differ from ground-truth $\mathbf{x}_t$ used in training.

Token misalignment is particularly severe:

- HASS still suffers **37% misalignment token rate**.
- Error **accumulates across forward passes** early mistakes compound in subsequent steps.
- Acceptance length **plateaus after 3 forward passes** despite continued training.

Why naive solutions fail:

- Replacing $\bar{\mathbf{x}}_t$ with $\mathbf{x}_t$ during training creates inconsistent input-feature pairs $(\mathbf{x}_t, \mathbf{F}_t)$.
- EAGLE/HASS precompute and store $\mathbf{F}_t$ for all $\mathbf{x}_t$ before training since regenerating data is computationally prohibitive.
- Naive substitution significantly degrades acceptance length.

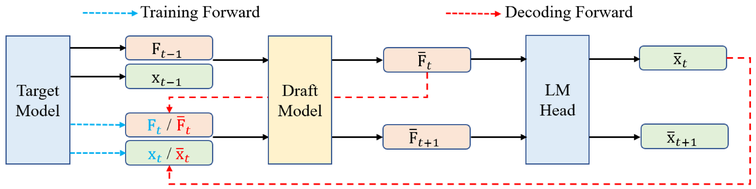

Figure: Token and feature misalignment in EAGLE.

GRIFFIN: A Token-Alignable Framework

GRIFFIN addresses token misalignment through two synergistic components: 1) token-alignable training strategy and 2) token-alignable draft model.

Token-Alignable Training Strategy

Core Idea: Progressive training that mirrors decoding behaviorâgradually shifting from ground-truth to self-generated tokens across n forward passes.

***n*-th Forward Pass ($n \geq 2$):** Introduce token alignment masks to handle misalignment:

- **Top-k Predictable Mask:** $\tilde{\mathbf{m}}_t = 1$ if $\bar{\mathbf{x}}_t \in \text{Top-}k(\mathbf{x}_t)$, else 0
- **Cumulative Alignment Mask:** $\mathbf{m}_t = \prod_{i=t-n+1}^{t-1} \tilde{\mathbf{m}}_i$
- **Masked Loss:** Only backpropagate through aligned tokens:

$$\mathcal{L}_{\mathcal{M}}^{(n)} = \frac{1}{\sum_{t=1}^{l} \mathbf{m}_t} \sum_{t=1}^{l} \mathbf{m}_t \ell(\bar{\mathbf{x}}_t, \mathbf{x}_t, \bar{\mathbf{F}}_t, \mathbf{F}_t)$$

Token-Alignable Draft Model

Two key modules enhance draft token accuracy:

1. Token-Guided Fusion (TGF): Addresses feature inconsistency by prioritizing token embeddings.

- *Feature Expansion:* Normalize and expand to $4d$: Higher dimension enables disentangling complex token-feature relationships

2. Token-Enhanced Head (TEH): Dual-head design decouples conflicting objectives.

- Predict feature $\bar{\mathbf{F}}^{\mathbf{P}}_{t+1}$ for token prediction
- Regress feature $\bar{\mathbf{F}}^{\mathbf{R}}_{t+1}$ for subsequent passes

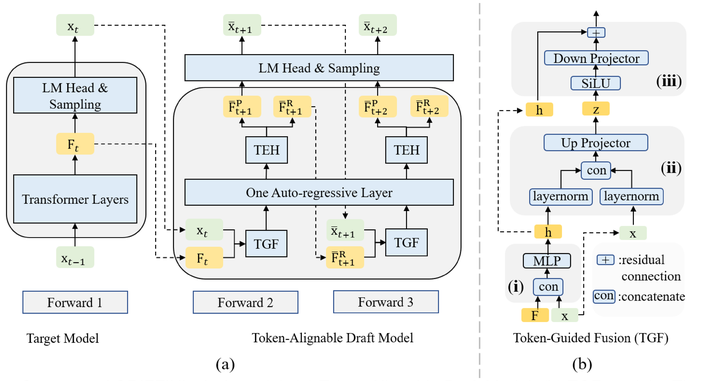

Figure: Structure of GRIFFIN's darft model. (a) Token-Alignable Draft Model. (b) TGF module. The diagram depicts the shared architecture used in both training and decoding phasesâarrows indicating token flow correspond to valid data dependencies in both regimes.

Experimental Results

Table: Comparison of different speculative decoding methods. This table presents evaluation results on standard LLM benchmarks with temperature $T \in \{0, 1\}$, including speedup ratio SR and acceptance lengths τ. Higher values indicate better performance.

Model	Method	Temperature = 0								Temperature = 1							
		MT-bench		HumanEval		GSM8K		Average		MT-bench		HumanEval		GSM8K		Average	
		SR	τ	SR	τ	SR	τ	SR	τ	SR	τ	SR	τ	SR	τ	SR	τ
LLaMA2 Chat 7B	PLD	1.41	1.46	1.51	1.57	1.34	1.39	1.42	1.47	N/A, since the acceptance conditions are relaxed							
	Lookahead	1.64	1.71	1.75	1.81	1.57	1.63	1.65	1.72								
	EAGLE-2	2.69	4.50	3.22	5.24	2.77	4.72	2.89	4.82	2.41	4.29	3.00	5.01	2.63	4.66	2.68	4.65
	FSPAD	2.89	4.82	3.38	5.62	2.95	4.99	3.07	5.14	2.61	4.53	3.14	5.35	2.84	4.88	2.86	4.92
	HASS	2.97	4.97	3.46	5.69	3.06	5.12	3.17	5.26	2.72	4.64	3.18	5.22	2.83	5.08	2.91	4.98
	GRIFFIN	**3.12**	**5.11**	**3.61**	**5.93**	**3.10**	**5.27**	**3.28**	**5.44**	**2.81**	**4.81**	**3.33**	**5.63**	**3.06**	**5.26**	**3.07**	**5.23**
LLaMA3 Instruct 8B	EAGLE	1.32	2.96	2.07	3.76	1.88	3.61	1.76	3.44	1.28	2.71	1.42	3.36	1.66	3.31	1.45	3.13
	EAGLE-2	2.56	4.18	3.36	5.05	2.53	4.41	2.82	4.54	2.26	3.75	2.63	4.77	2.46	4.30	2.45	4.27
	EAGLE-3	2.93	4.71	3.59	5.72	3.17	5.01	3.23	5.15	2.51	4.18	3.27	5.47	2.90	4.85	2.89	4.83
	FSPAD	2.72	4.52	3.40	5.39	2.95	4.77	3.02	4.89	2.43	4.09	3.04	5.18	2.75	4.60	2.74	4.62
	HASS	2.75	4.63	3.51	5.70	3.09	5.06	3.12	5.13	2.41	4.15	3.09	5.41	2.92	4.90	2.81	4.82
	GRIFFIN	**3.09**	**4.85**	**3.65**	**5.97**	**3.30**	**5.31**	**3.35**	**5.38**	**2.62**	**4.35**	**3.31**	**5.62**	**3.07**	**5.08**	**3.00**	**5.02**
Vicuna1.5 7B	SPS	1.81	2.34	2.04	2.68	1.73	2.28	1.86	2.43	1.49	1.85	1.57	1.99	1.52	1.80	1.53	1.88
	Medusa	1.97	2.60	2.07	2.75	1.93	2.65	1.99	2.67	N/A, since the acceptance conditions are relaxed							
	EAGLE-2	3.56	4.74	3.92	5.30	3.69	5.03	3.72	5.02	3.15	4.20	3.30	4.62	3.41	4.65	3.29	4.49
	FSPAD	3.73	5.16	4.12	5.74	3.85	5.37	3.90	5.42	3.27	4.53	3.45	5.11	3.56	4.98	3.42	4.87
	HASS	3.91	5.15	4.22	5.86	3.97	5.41	4.03	5.47	3.34	4.52	3.62	5.16	3.70	5.03	3.55	4.90
	GRIFFIN	**4.02**	**5.36**	**4.53**	**6.29**	**4.14**	**5.63**	**4.23**	**5.76**	**3.38**	**4.64**	**4.12**	**5.68**	**3.88**	**5.29**	**3.79**	**5.20**
Qwen2 Instruct 7B	EAGLE-2	2.32	3.80	2.90	4.73	2.70	4.32	2.64	4.28	2.00	3.01	2.71	4.18	2.60	3.98	2.43	3.72
	HASS	2.59	4.23	3.18	5.46	2.91	4.86	2.89	4.85	2.17	3.23	2.83	4.52	2.79	4.38	2.59	4.04
	GRIFFIN	**2.76**	**4.67**	**3.34**	**5.75**	**3.02**	**5.13**	**3.04**	**5.18**	**2.27**	**3.36**	**3.04**	**4.82**	**2.96**	**4.71**	**2.76**	**4.30**
LLaMA2 Chat 13B	EAGLE-2	2.97	4.68	3.61	5.59	3.05	4.97	3.21	5.08	2.77	4.45	3.41	5.45	2.97	4.83	3.05	4.91
	FSPAD	3.09	5.05	3.91	5.98	3.32	5.35	3.44	5.46	3.03	4.85	3.51	5.71	3.21	5.25	3.25	5.27
	HASS	3.11	5.05	4.16	6.05	3.38	5.33	3.55	5.47	3.05	4.90	3.66	5.85	3.22	5.30	3.31	5.35
	GRIFFIN	**3.33**	**5.27**	**4.29**	**6.26**	**3.61**	**5.56**	**3.74**	**5.70**	**3.36**	**5.07**	**3.94**	**6.13**	**3.61**	**5.49**	**3.64**	**5.56**
LLaMA3 Instruct 70B	EAGLE-2	2.96	4.13	4.03	5.08	3.21	4.42	3.40	4.54	3.04	4.05	3.65	5.01	3.20	4.32	3.34	4.46
	HASS	3.36	4.59	4.61	5.73	4.01	5.21	3.99	5.17	3.35	4.48	4.23	5.65	3.84	5.17	3.80	5.10
	GRIFFIN	**3.52**	**4.66**	**4.71**	**6.03**	**4.09**	**5.39**	**4.11**	**5.36**	**3.49**	**4.54**	**4.33**	**5.94**	**3.90**	**5.30**	**3.91**	**5.26**
Mixtral-v0.1 Instruct 8x7B	EAGLE-2	1.96	3.39	2.34	4.13	2.19	3.79	2.16	3.77	1.93	3.32	2.28	3.98	2.09	3.71	2.10	3.67
	HASS	2.17	3.67	2.63	4.76	2.39	4.58	2.39	4.33	2.09	3.61	2.53	4.58	2.24	4.46	2.28	4.21
	GRIFFIN	**2.29**	**3.97**	**2.82**	**5.25**	**2.51**	**4.86**	**2.54**	**4.69**	**2.22**	**3.89**	**2.71**	**5.08**	**2.39**	**4.72**	**2.44**	**4.56**

Code & Paper

For More Details:

- **Code:** `https://github.com/hsj576/GRIFFIN`
- **Paper:** `https://arxiv.org/abs/2502.11018`
- **Contact:** sjhu24@m.fudan.edu.cn