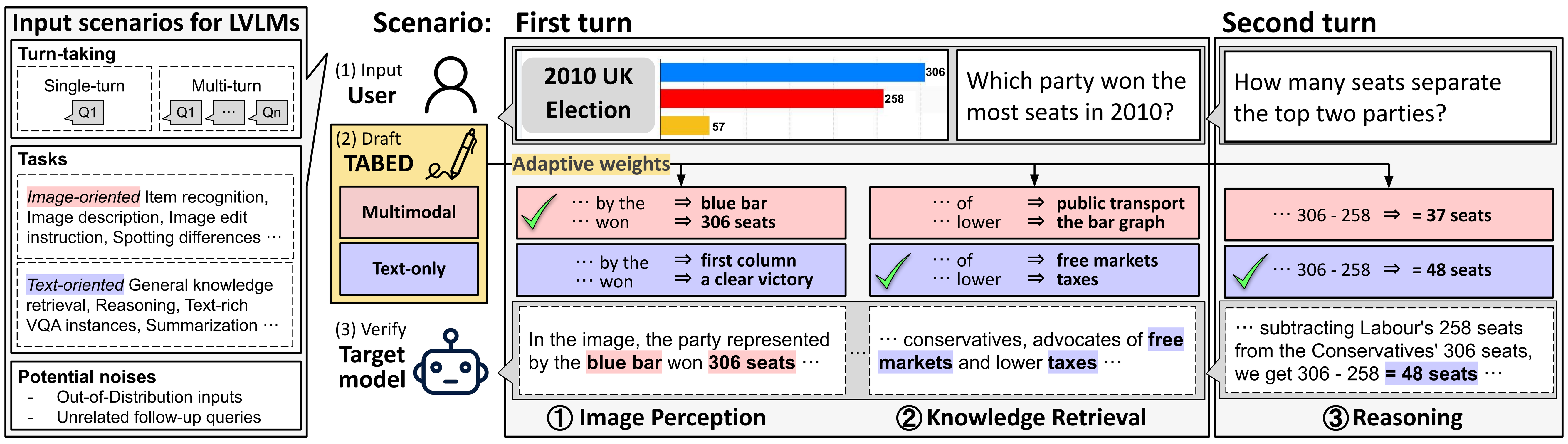
TABED: Test-Time Adaptive Ensemble Drafting for Robust Speculative Decoding in LVLMs

Minjae Lee*, Wonjun Kang*, Byeongkeun Ahn, Christian Classen,
Kevin Galim, Seunghyuk Oh, Minghao Yan, Hyung Il Koo, Kangwook Lee
FuriosaAI, University of Wisconsin-Madison, KRAFTON

EACL 2026
RABAT · MOROCCO
Mars • March 24-29, 2026 • مارس

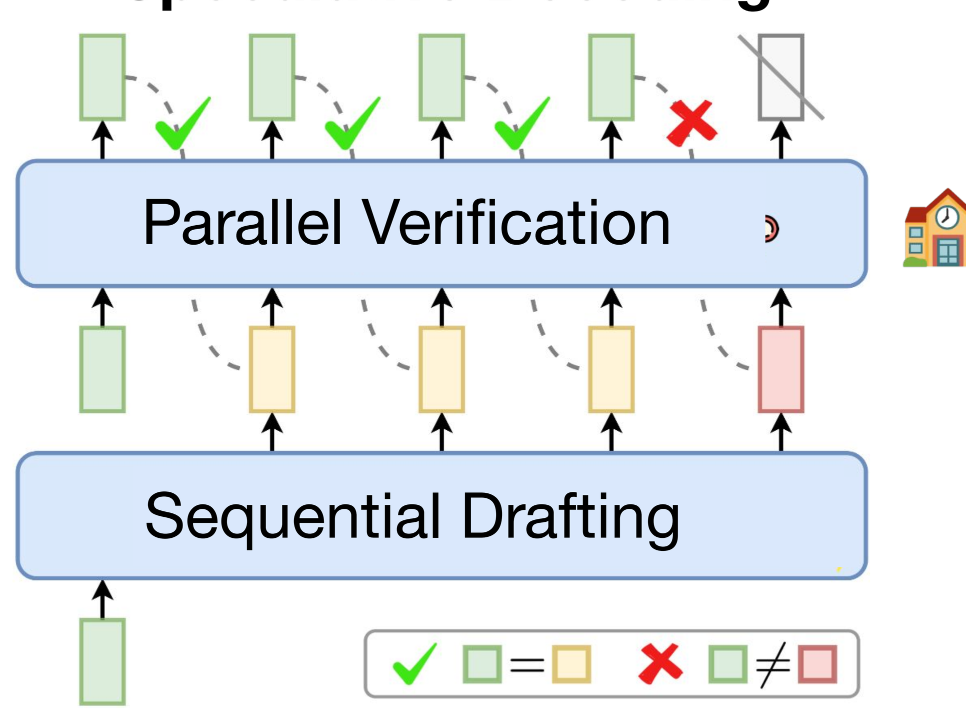
Motivation



- **Large Vision–Language Models (LVLMs)** must handle **diverse input scenarios.**
- To effectively accelerate LVLMs with **Speculative Decoding (SD)**, different drafting strategies are required for intra-response and inter-response settings.
- While SD has proven effective for LLMs, it remains **underexplored for LVLMs.**

Preliminary

Speculative Decoding[1]

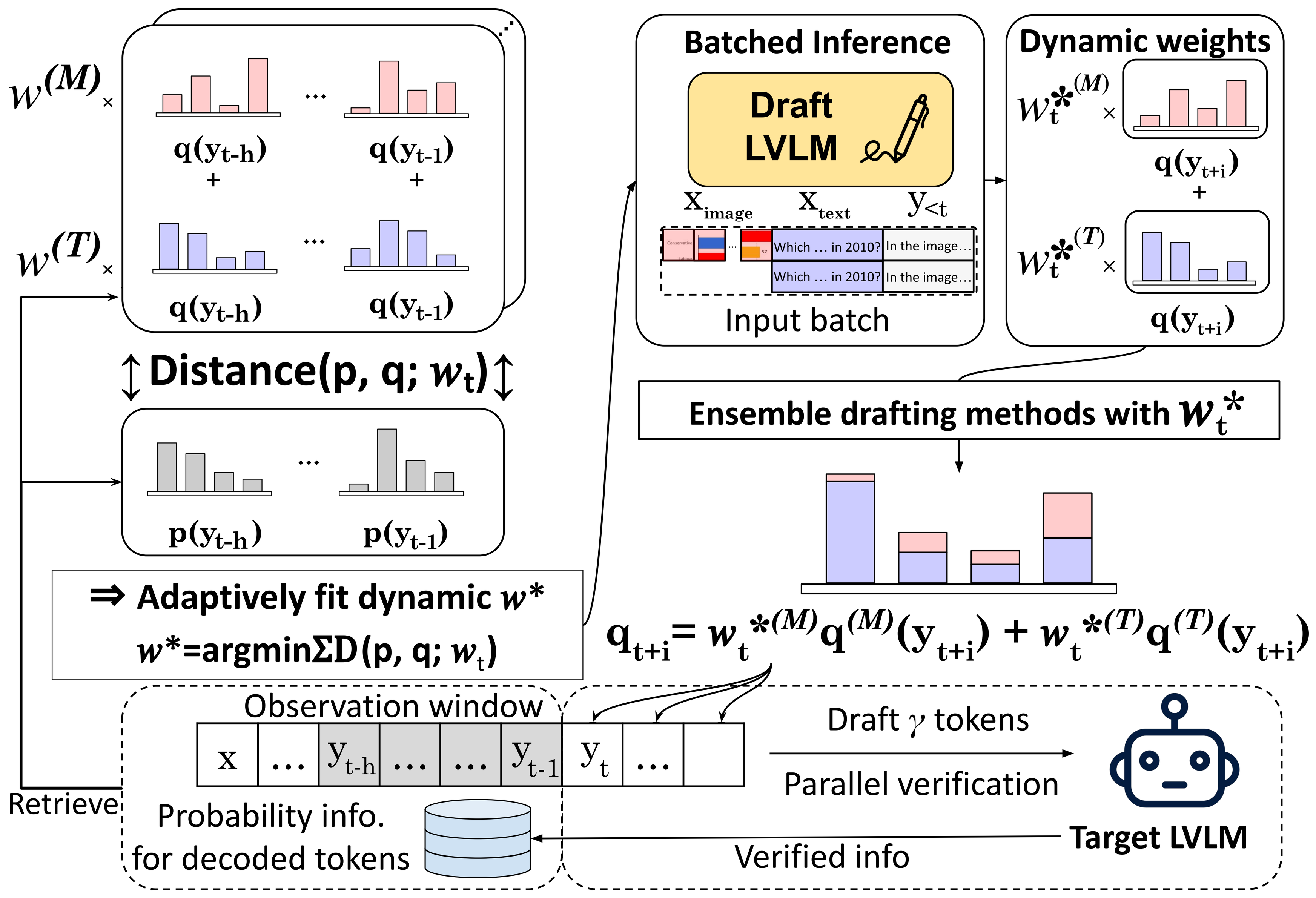


1. A small **draft model** speculates a specified number of draft tokens.
2. The larger **target model** verifies these proposed tokens in parallel.

Contribution

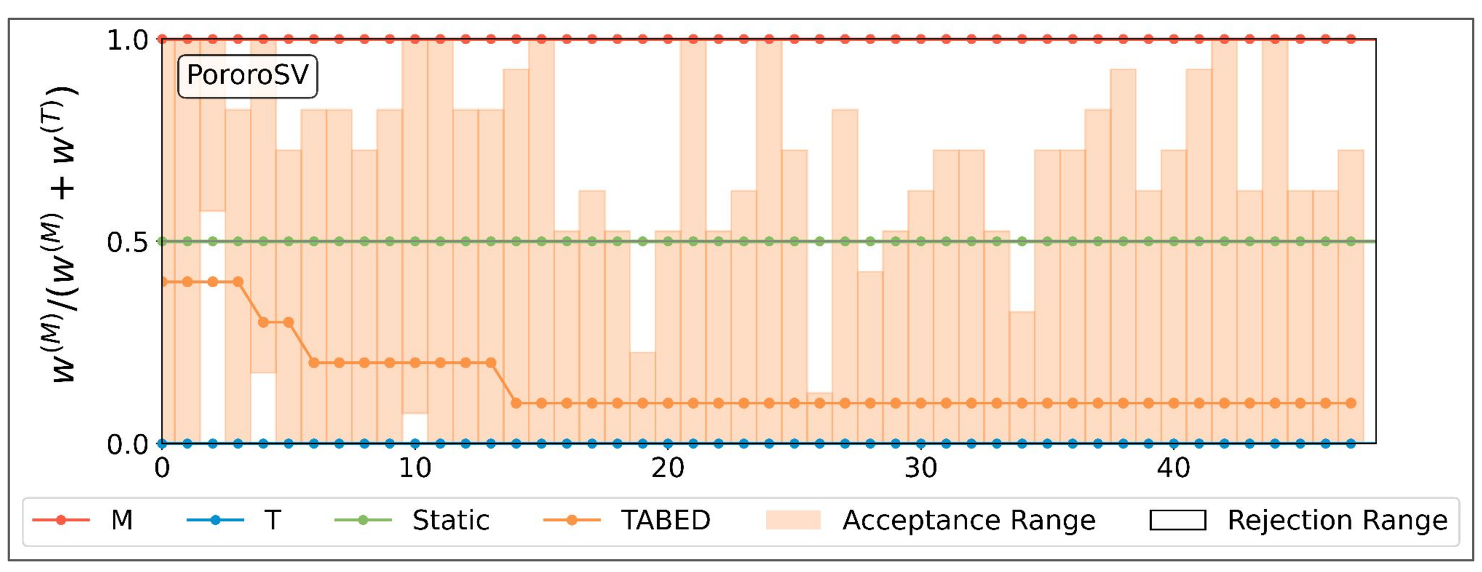
- We **benchmark** existing drafting methods and find that their performance fluctuates across diverse LVLM input scenarios.
- We propose **Test-time Adaptive Batched Ensemble Drafting (TABED)**, which achieves **superior and robust** performance.
- TABED is further enhanced via its **plug-and-play integration** with advanced verification and alternative drafting methods.

Method



$$q_{t+i} = w_t^{*(M)} q^{(M)}(y_{t+i}) + w_t^{*(T)} q^{(T)}(y_{t+i})$$

1. **Batched Inference**: TABED generates multiple drafts via batched inference with shared model parameters.
2. **Test-time Adaptive Ensemble Weights**: TABED dynamically ensembles drafts by leveraging deviations from past ground truths available in the SD setting.



Adaptation Behavior of Dynamic Weights:

- TABED effectively stay within the shaded acceptance range while avoiding the unshaded rejection range

Experimental Results

Drafting		Benchmark Datasets (First Turn)								OOD Datasets	
Type	Method	LLaVA-W	DocVQA	POPE	MMVet	IEdit	MB	Spot	Avg.	PSV	VIST
Single	M	**2.28**	<u>2.15</u>	**2.56**	**2.21**	2.19	1.96	<u>2.34</u>	<u>2.24</u>	1.19	1.16
	T [16]	2.19	2.08	2.31	2.16	<u>2.23</u>	<u>2.34</u>	2.27	2.23	**2.05**	**2.05**
Ensemble	**TABED^{MT}**	<u>2.26</u>	**2.16**	<u>2.52</u>	**2.21**	<u>2.29</u>	**2.39**	**2.36**	**2.31**	<u>2.02</u>	<u>2.04</u>
Drafting		**Benchmark Datasets (Second Turn)**								**NLP Datasets**	
Type	**Method**	**LLaVA-W**	**DocVQA**	**POPE**	**MMVet**	**IEdit**	**MB**	**Spot**	**Avg.**	**NQ**	**GSM8K**
Single	M	2.10	1.96	2.78	2.18	1.61	1.53	1.83	2.00	1.98	2.25
	T [16]	**2.32**	**2.23**	<u>2.91</u>	**2.56**	**1.87**	**2.01**	**2.08**	**2.28**	**2.03**	**2.30**
Ensemble	**TABED^{MT}**	<u>2.29</u>	**2.23**	**2.93**	**2.56**	<u>1.85</u>	<u>1.99</u>	<u>2.05</u>	<u>2.27</u>	**2.03**	<u>2.29</u>

Benchmarking Results

- TABED consistently achieves either the **best** or <u>second-best</u> performance across diverse input scenarios.
- TABED achieves an average **robust walltime speedup** of 1.74× over autoregressive decoding and a 5% improvement over single drafting methods (Multimodal and Text-only).
- TABED supports **plug-and-play** compatibility with **no further training,** and maintains **negligible ensembling costs.**

Verification	Drafting		Block Efficiency	
Tree width	Type	Method	Benchmark	OOD
$d = 2$	Sgl.	M	<u>2.89</u>	1.30
		T [16]	2.85	**2.72**
	Ens.	**TABED^{MT}**	**2.99**	<u>2.64</u>
$d = 3$	Sgl.	M	<u>3.28</u>	1.38
		T [16]	3.24	**3.19**
	Ens.	**TABED^{MT}**	**3.39**	<u>3.09</u>

Drafting		Benchmark		OOD
Type	Method	First	Second	
Single	M	2.24	2.00	1.18
	T [16]	2.23	2.28	2.05
	C	<u>2.29</u>	<u>2.30</u>	<u>2.09</u>
	P	2.23	2.25	2.08
Ensemble	**TABED^{MTCP}**	**2.32**	**2.32**	**2.13**

Plug-and-Play Extensions

- (Top) TABED integrated with **token-tree verification** using tree width d.
- (Bottom) TABED integrated with **alternative drafting methods** (caption-based and pooled-multimodal).

Scan & Connect 😄

[1] Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding