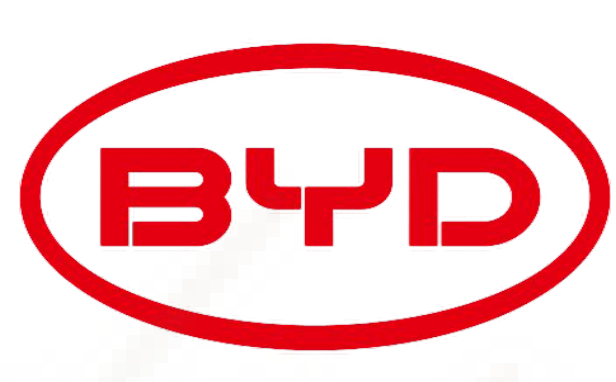




Length-Adaptive Decoding for Masked Diffusion Machine Translation

Yan Zhan¹ Mengkai Hou¹ Wanting Zhang² Zhijun Gao¹

¹Peking University ²BYD Company Limited

Contact: gaozhijun@pku.edu.cn

Code: github.com/Entropy-Valley/Entropy-Valley

Dataset & Model: huggingface.co/collections/YanZhanPKU/entropy-valley



arXiv
2608.22274



GitHub
Code



Hugging Face
Data & Models



About Me
Personal Website



WeChat
yalezhan



WhatsApp
yalezhan

Problem

Fixed-canvas masked diffusion MT must choose the target length before denoising. A single corpus ratio cannot adapt to sentence-specific compression and expansion needs.

Failure modes

Case 1: Too short canvas drops the action			
Source (EN)	Fixed Ratio (selected length=5)	Entropy-Valley (selected length=6)	
Tap Reset Now.	 Output (ZH) 现在。 COMET 0.47	 Output (ZH) 点击重置。 COMET 0.91	
Reference (ZH)	Missing Action Verb	Preserve Action	
现在点击重置。	 One extra canvas slot preserves the main action.		

Case 2: Too short canvas loses placeholder				
Source (EN) Under #PRS_ORG#, tap Sign out.	Fixed Ratio (selected length=10)	Entropy-Valley (selected length=12)		
Reference (ZH) 在#PRS_ORG#，点击退出。	 Output (ZH) 在私人组织下按退出。	 COMET 0.43	 Output (ZH) 在#PRS_ORG#，按下退出。	 COMET 0.87
	Missing #PRS_ORG#		Preserve Placeholder	
	 Enough canvas length helps preserve literal source tokens.			

Case 3: EV fails when the best length is outside its range			
Source (EN) Please give me a moment.	Fixed Ratio (selected length=6; matches reference)	Entropy-Valley (selected length=7)	
	 Output (ZH) 请等一下。  0.93	 Output (ZH) 请给我一会儿。  0.77	
Reference (ZH) 请等我一下。	Concise and Correct	Grid Overshoots True Ratio: EV drop range	
 EV is limited by its candidate length range.			
			

Findings

Too short: content disappears.
Too long: redundancy and dilution.

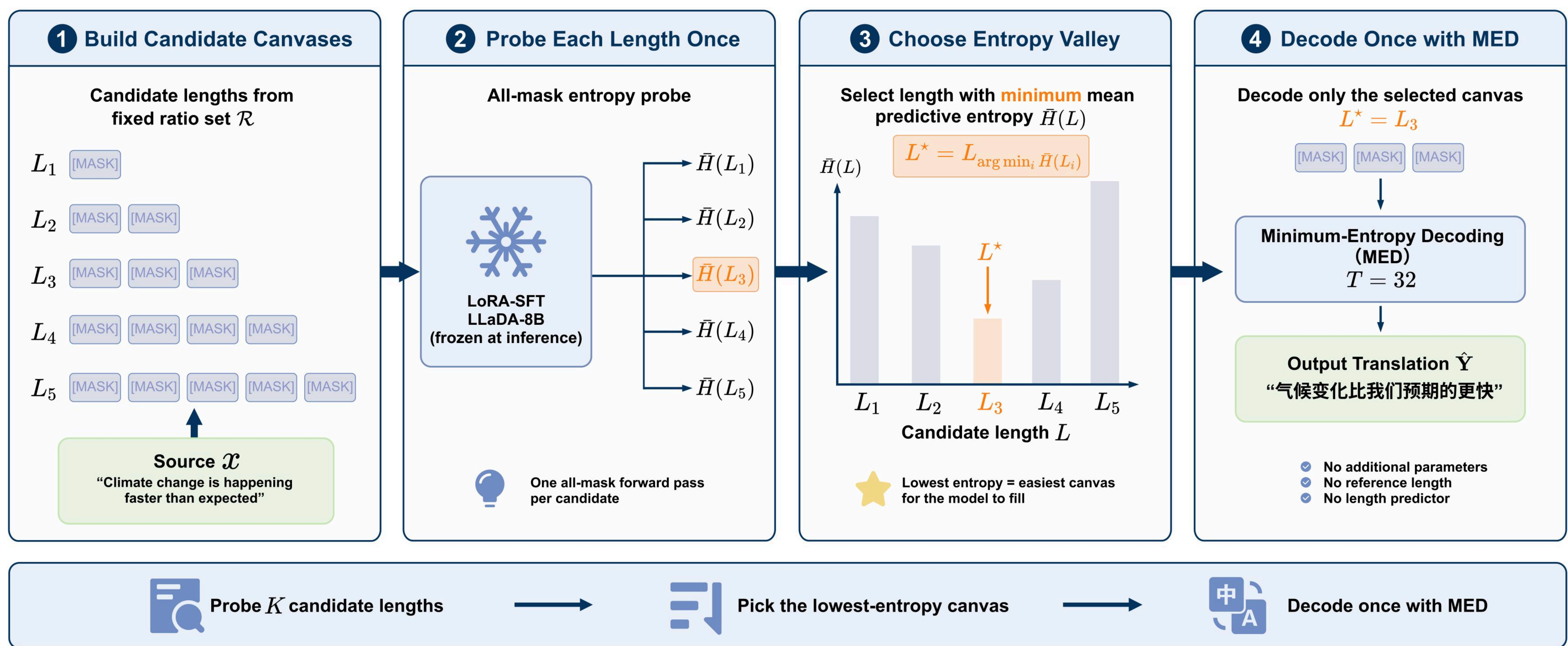
Motivation

- 1 The backbone already exposes a length signal before decoding: predictive uncertainty changes when the canvas too short or too long.
- 2 The desired canvas need not equal the human reference length; it should be a length the frozen model can fill reliably.
- 3 Mean all-mask entropy turns this signal into a test-time criterion, without a new length head or a change to denoising.

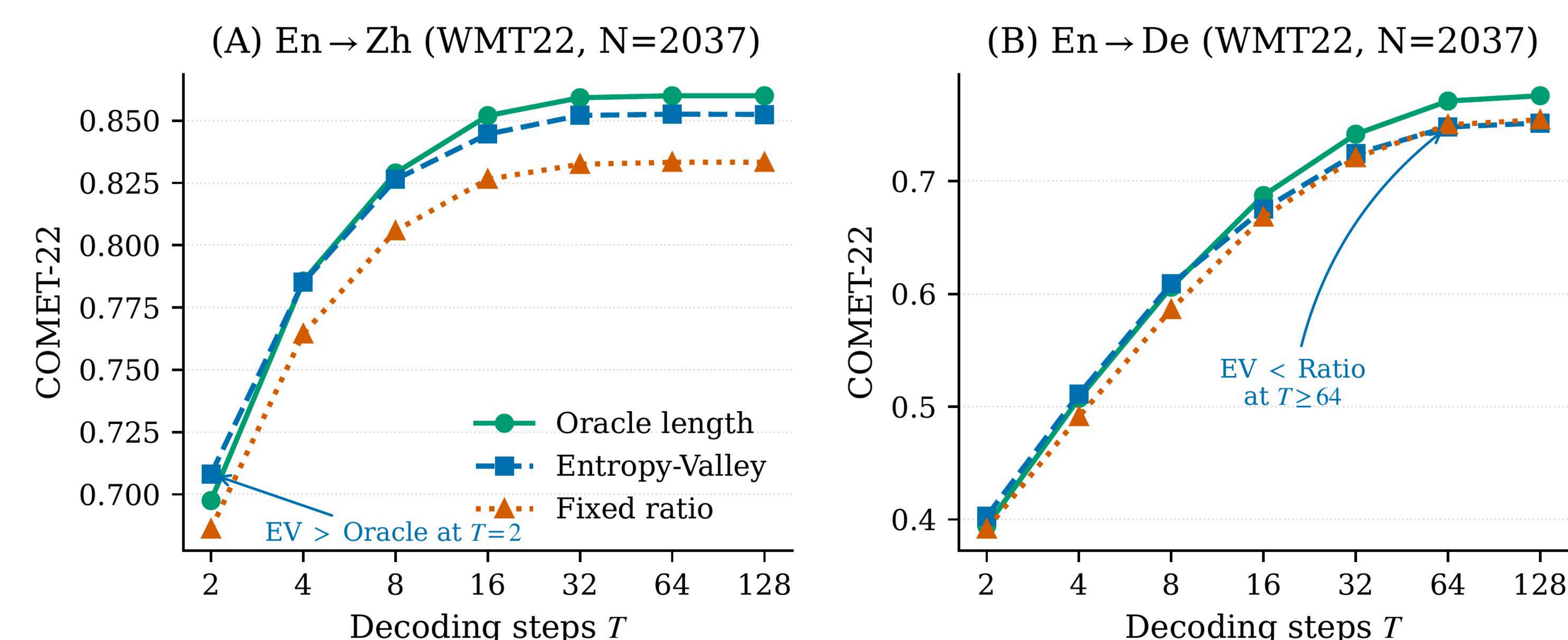
A target canvas is the number of target slots allocated before denoising.
Too few slots can omit content, while too many slots dilute predictions.

Reference length is an evaluation-only oracle.
A learned predictor would add training and state.
Just choose a reliable canvas before any token is revealed.

Method



Ablation Study



Matched reference length separates reveal order from canvas choice.

SO WHAT →

With the same target length, tested reveal schedules span 0.0081 COMET;
with the same MED schedule, length choices span 0.0265 COMET.

Headline Numbers

64.9%

COMET gap closure
En → Zh

65.3%

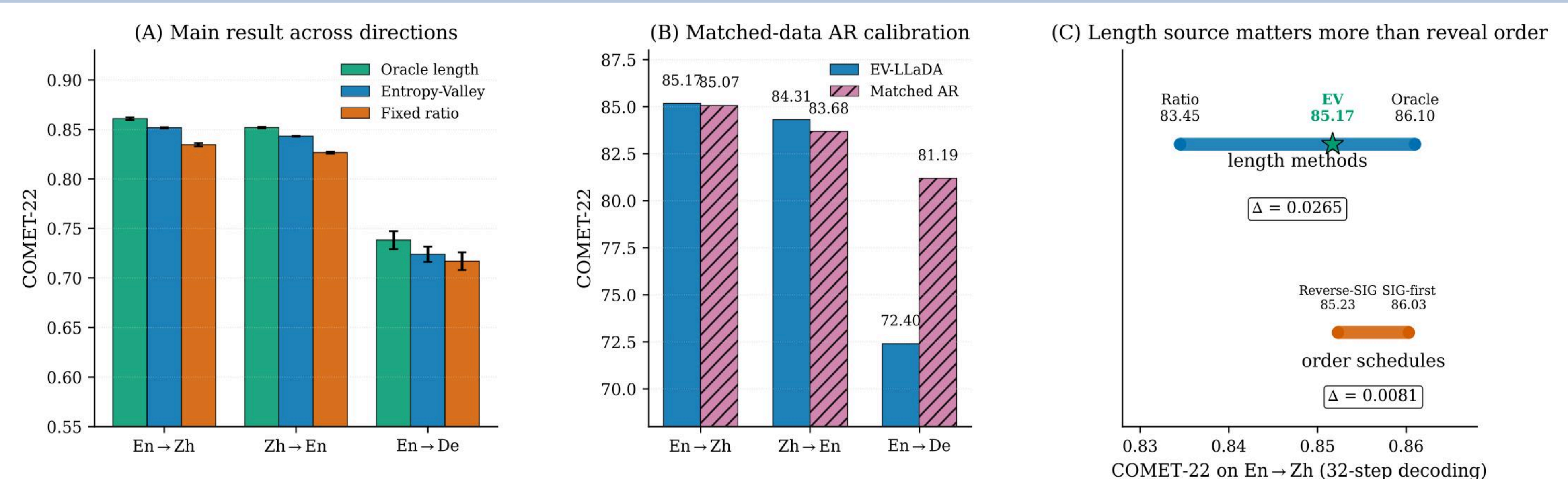
Zh → En
gap closure

33.0%

En → De
gap closure

Gap closure is measured relative to the fixed-ratio baseline and the reference-length oracle.

Key Results



SO WHAT →

Sentence-level canvas selection closes the largest gaps in the Chinese directions;
English → German is a weaker boundary case.