哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

ICML
International Conference
On Machine Learning

SJD-SV: Speculative Jacobi Decoding with Semantics Verification for Autoregressive Image Generation

Baoquan Zhang
HIT

Bingqi Shan
HIT & PCL

Shihao Fang
HIT

Kenghong Lin
HIT

Xutao Li
HIT

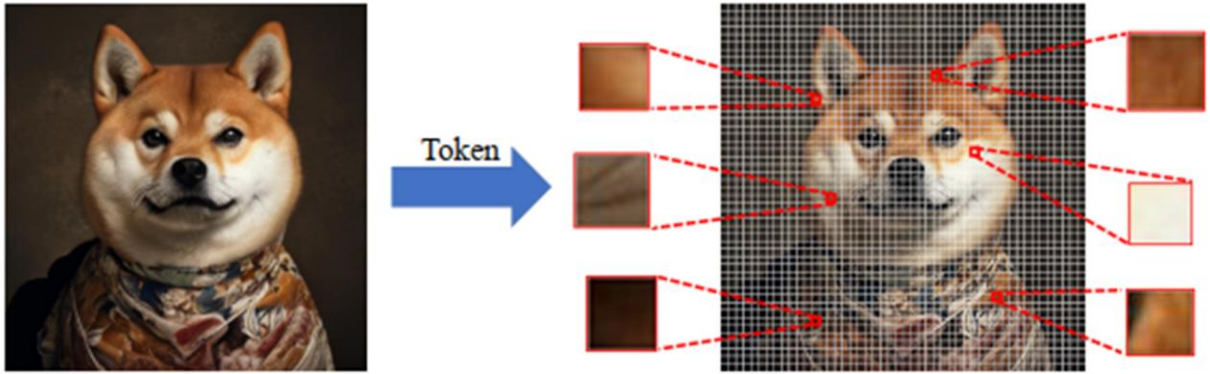
Yunming Ye
HIT

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

Content

PART 01

Problem & Idea



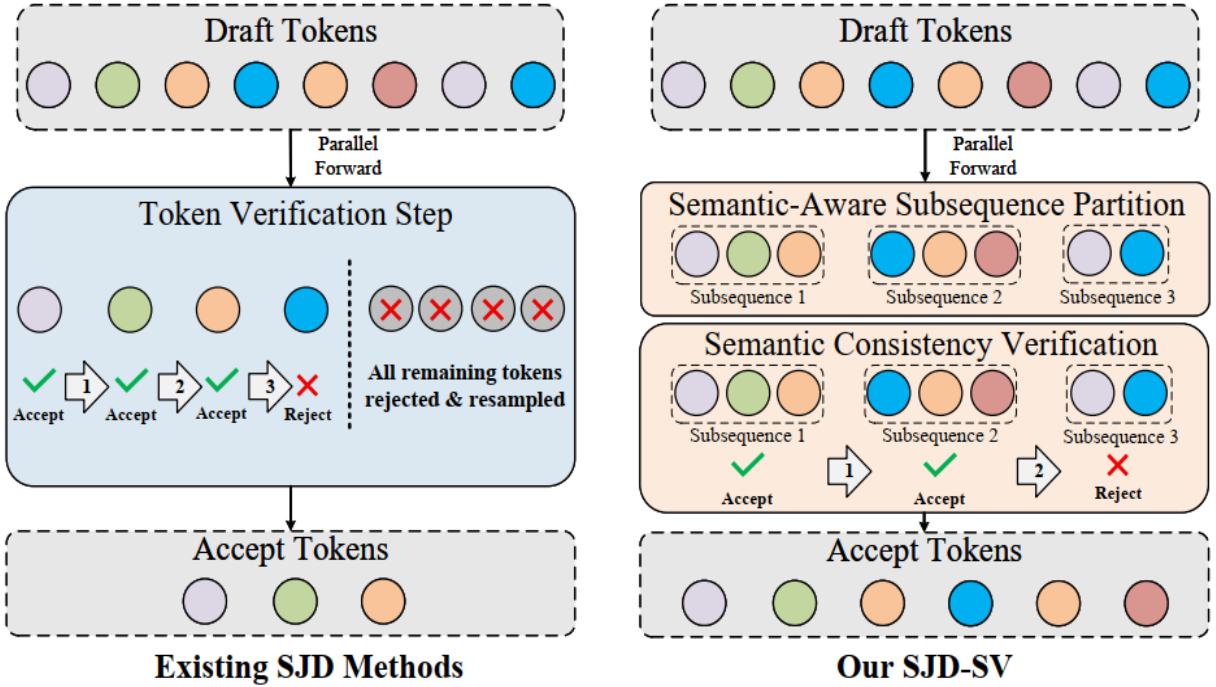
Root of the Token Ambiguity (**Problem**):

- Vision tokens are too **small and blurry** to convey clear semantic meaning
- **Lack of context information** makes multiple different candidates

Verify **Token by Token** → **Idea**: Split Draft to **Semantic** Subequence & Verify **Semantic** by Semantic

PART 02

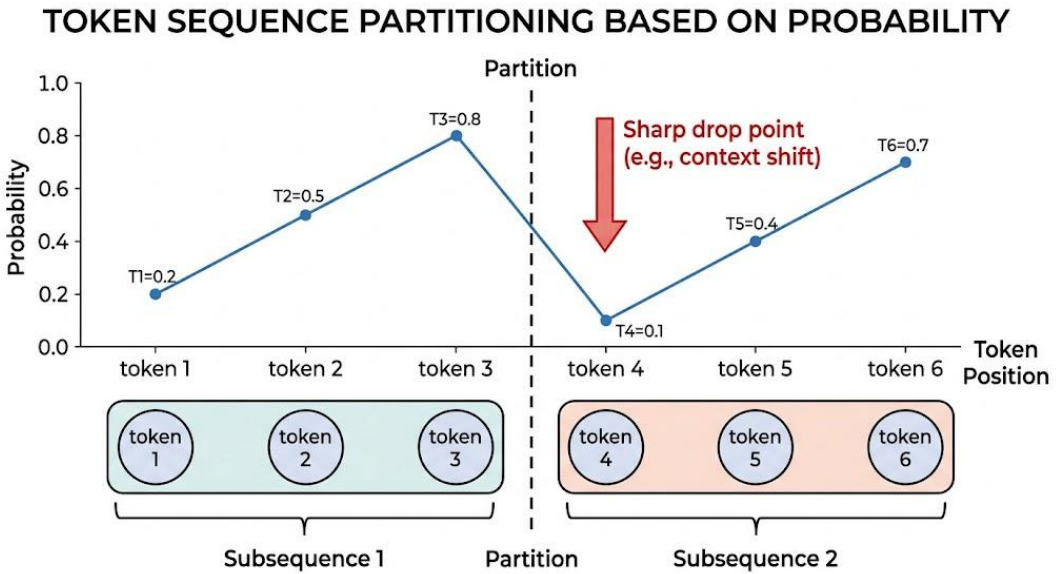
Method



Traditional SJD: **individually** verify daft tokens

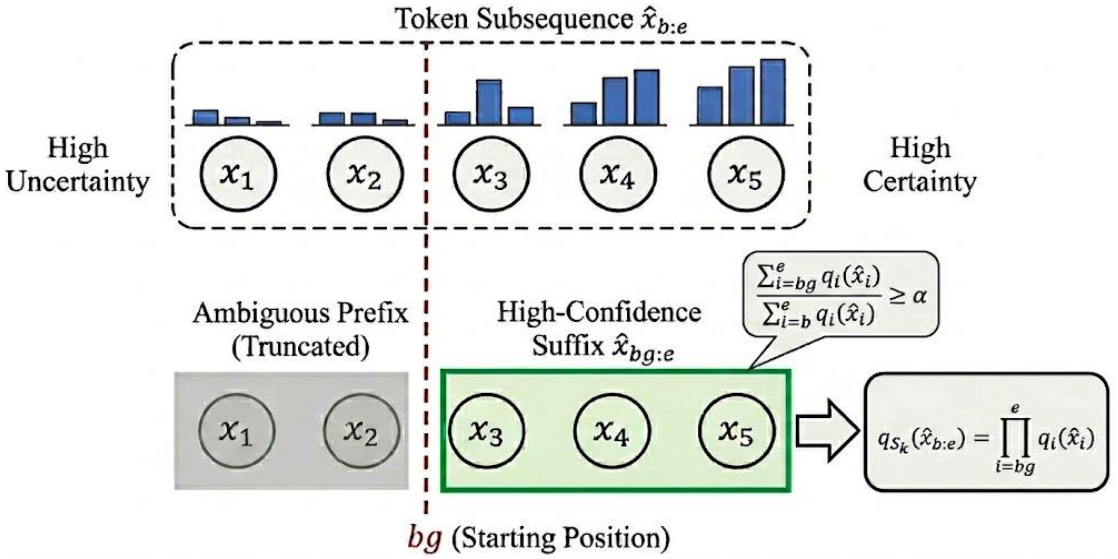
Our SJD-SV: split draft into **semantic sequences** and verify sequences

Subsequence Partition



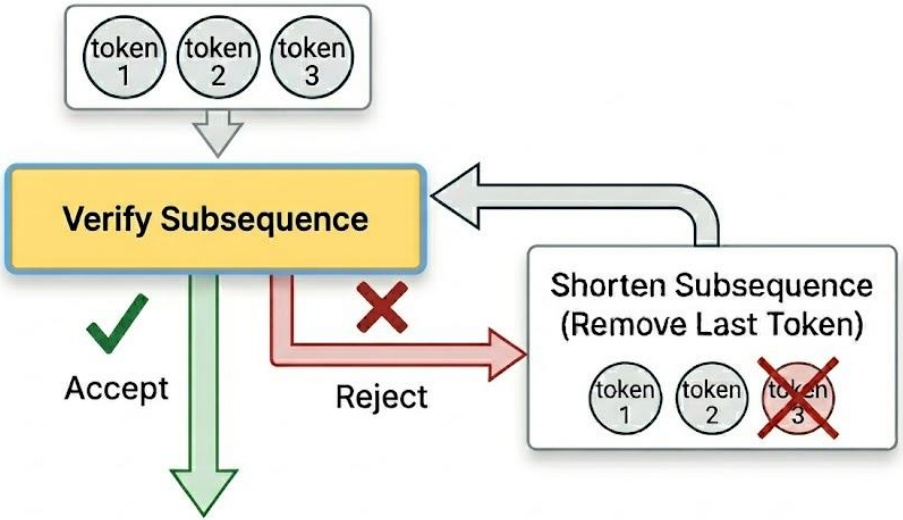
- Identify semantic subsequences based on **probability growth**.
- Grouping Rule: A token is grouped with the previous one only if its **probability increases**.
- Example: In *New York City*, the probability of *City* rises as context becomes certain, grouping them together.

Joint Verification



- Adaptive Suffix Selection: Verify only **high-confidence suffixes** where tokens dominate the total probability mass, filtering out ambiguous prefixes that introduce noise。
- Joint Probability Verification: Validate the selected suffix using its **joint distribution probability**.

Robust Fallback



- Reject & Retry: If verification fails, iteratively remove the last token and re-verify to find the longest valid prefix.
- Accept **the longest correct prefix** to balance accuracy and throughput.

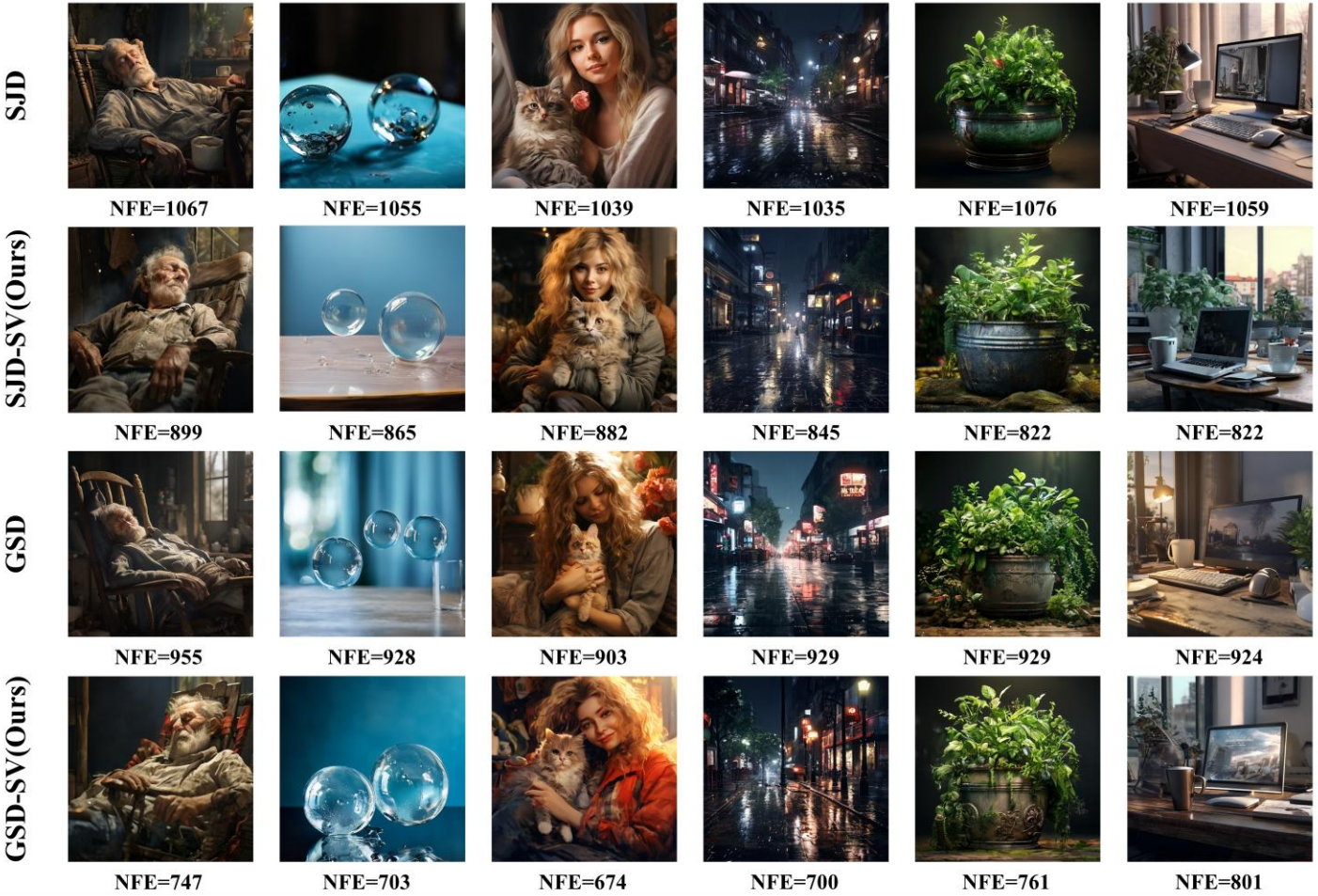
PART 03

Experiment Result

Quantitative Results

Configuration	Latency (↓)	NFE (↓)	Acceleration (↑)		FID (↓)	CLIP-Score (↑)
			Latency	NFE		
Parti-prompt						
Lumina-mGPT (Liu et al., 2024a)	79.37s	2392	1.00×	1.00×	–	32.09
Jacobi Decoding (Song et al., 2021)	82.21s	2300.0	0.97×	1.04×	–	32.09
SJD (Teng et al., 2024a)	36.07s	1035.3	2.20×	2.31×	–	32.09
SJD-SV	**30.48s**	**826.32**	**2.60×**	**2.89×**	–	**32.13**
GSD (So et al., 2025)	33.36s	898.97	2.38×	2.66×	–	32.11
GSD-SV	**30.77s**	**721.47**	**2.58×**	**3.32×**	–	**32.12**
LANTERN(Jang et al., 2024b)	31.40s	636.66	2.52×	3.76×	–	32.10
LANTERN-SV	**27.72s**	**552.93**	**2.86×**	**4.33×**	–	**32.11**
MS-COCO 2017						
Lumina-mGPT (Liu et al., 2024a)	86.55s	2379	1.00×	1.00×	30.79	31.31
Jacobi Decoding (Song et al., 2021)	85.64s	2312	1.01×	1.03×	30.78	31.31
SJD (Teng et al., 2024a)	40.10s	1058.6	2.16×	2.25×	30.78	31.31
SJD-SV	**32.92s**	**866.80**	**2.63×**	**2.74×**	**30.76**	**31.33**
GSD (So et al., 2025)	34.12s	925.89	2.54×	2.56×	31.50	31.33
GSD-SV	**29.96s**	**745.88**	**2.89×**	**3.19×**	**31.47**	**31.34**
LANTERN(Jang et al., 2024b)	33.80s	655.37	2.56×	3.63×	31.72	31.32
LANTERN-SV	**23.11s**	**568.30**	**3.75×**	**4.19×**	**31.70**	**31.34**

Qualitative Results



PART 04

Conclusion

Our Contribution:

•**Identified the Root Cause of Token Ambiguity:** We revealed that **Token Ambiguity** in visual AR models stems from the lack of semantics in individual fine-grained pixel patches.

•**Proposed SJD-SV:** A novel framework that shifts verification process from isolated tokens to **semantic-aware subsequences**.

•**Experimental Proven:** Extensive experiments prove SJD-SV is a **highly effective, plug-and-play** module, achieving significant acceleration with zero quality loss.

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

Thank You for Your Time !

Welcome to Connect & Collaborate

baoquanzhang@yeah.net