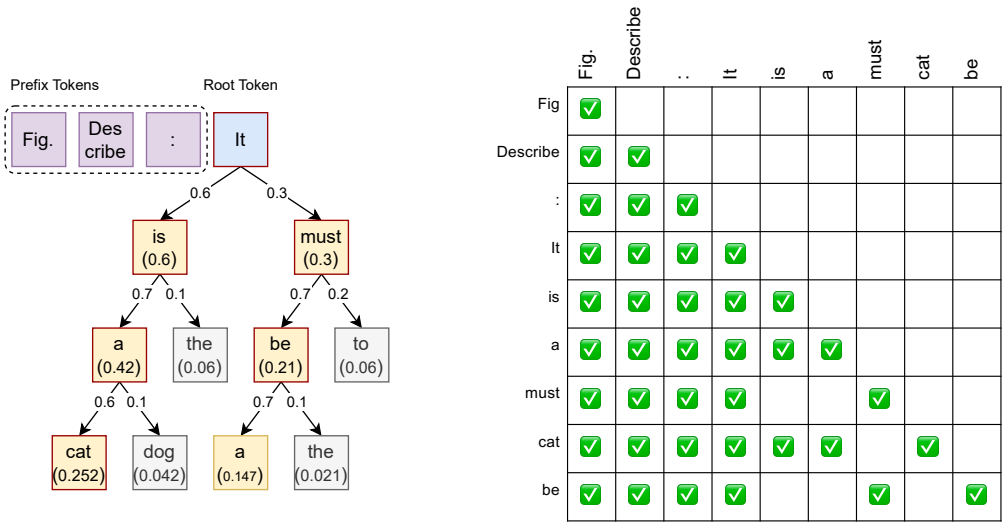
Supplementary Materials

A1: Tree Based Generation Process

To better illustrate the tree-based generation process, we present a representative example. As shown in Figure 1a, we begin from a prefix sequence "Fig. Describe :" and a token "It," which serves as the root node of the draft tree and corresponds to the most recently verified token provided by the target model. Here, the special prefix token "Fig." denotes the visual tokens representation of the input image. From the root, the draft model expands a tree of candidate tokens across multiple depth layers.

We expand the draft tree by sampling candidate tokens at each depth in parallel. To maintain autoregressive integrity, we apply a tree-based attention mask. Each node in the draft tree only attends to its ancestral path, thereby preventing information leakage across branches.



(a) Tree-based expansion process.

(b) Flattened sequence and its tree-based attention mask.

Figure 1: Tree-based generation and attention mechanism. (a) The draft model generates a tree of candidate tokens starting from a root token ("It") following the prefix ("Fig. Describe :"). Tokens are expanded layer-by-layer with associated global acceptance scores. Colors denote different roles: purple for prefix tokens, blue for the root, yellow for retained tokens, and gray for pruned branches. (b) The tree-based attention mask ensures that each token attends only to its ancestors in the tree, thereby preserving autoregressive structure and preventing cross-branch leakage.

Top-k Expansion To prevent exponential expansion, we selectively expand only the top-k tokens with the highest estimated global acceptance scores at each depth layer of the draft tree:

$$C_i = \prod_{d_j \in \text{Path}(\text{root}, d_i)} c_j$$

where c_j denotes the draft model's confidence score for token d_j, and $\text{Path}(\text{root}, d_i)$ is the ancestral path from the root node to token d_i. These global scores estimate the likelihood of a token passing

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

verification and guide the selection of which branches to expand. As illustrated in Figure 1a, when top-$k = 2$, yellow nodes indicate the tokens selected for further expansion.

Reranking Strategy To mitigate the verification cost associated with a large draft tree, we use the reranking step that prunes low-confidence branches. Since tokens at deeper levels tend to have lower acceptance probabilities, we rerank all candidate tokens across all depths based on their global scores and select the top-m tokens for verification.

To preserve structural consistency, we retain all ancestors of selected tokens and resolve ties in favor of shallower nodes. The selected subtree is then flattened into a linear sequence for verification. During this phase, we continue to apply the tree-based attention mask, ensuring that each token can only attend to its ancestral tokens, thus preserving autoregressive constraints. In Figure 1a, red-bordered tokens represent those selected during the reranking process with top-$m = 6$, including the root node. Figure 1b visualizes the flattened sequence and its corresponding attention mask.

A4: Statistical Significance Analysis

We evaluate the statistical significance of DREAM's improvements over six baseline methods on two key metrics—speedup ratio (S) and average accepted token length (τ)—across six datasets (MMT-Bench [7], SEED-Bench-2 [1], ScienceQA [5], OCRBench [3] , ChartQA [6], and MathVista [4]), using LLaVA-v1.6-Vicuna-13B [2] at temperature $T = 0$.

DREAM consistently achieves statistically significant improvements in speedup ratio (S) across all baselines, with large margins and strong effect sizes. For average accepted token length (τ), DREAM also shows significant improvements in nearly all comparisons, demonstrating its ability to not only accelerate inference but also maintain high output quality.

These results highlight DREAM's robustness across diverse datasets and baselines, offering both faster decoding and better verification efficiency under strong backbone models.

Table 1: Significance test of DREAM vs. baselines, with Bonferroni-corrected t-tests ($\alpha = 0.0083$)

Method	ΔS			$\Delta \tau$		
	Mean	95% CI	p	Mean	95% CI	p
SPD	2.09	[1.66, 2.52]	0.0001	3.36	[3.01, 3.72]	0.0000
Kangaroo	1.70	[1.17, 2.23]	0.0004	3.15	[2.46, 3.84]	0.0001
Medusa	1.32	[1.02, 1.63]	0.0001	2.10	[1.41, 2.78]	0.0005
Hydra	1.19	[0.91, 1.47]	0.0001	1.88	[1.23, 2.53]	0.0007
EAGLE	0.96	[0.68, 1.25]	0.0003	1.46	[0.95, 1.97]	0.0007
EAGLE-2	0.79	[0.56, 1.02]	0.0005	0.42	[0.05, 0.79]	0.0280

References

[1] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models. *arXiv preprint arXiv:2311.17092*, 2023.

[2] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[3] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024.

[4] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

[5] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, Jiayi Lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi, 2024.