Spec-LLaVA: Accelerating Vision-Language Models with Dynamic Tree-Based Speculative Decoding

Mingxiao Huo [1 *] **Jiayi Zhang** [2 *] **Hewei Wang** [1 *] **Jinfeng Xu** [3] **Zheyu Chen** [4] **Huilin Tai** [5] **Yijun Chen** [6]

Abstract

Vision-Language Models (VLMs) enable powerful multimodal reasoning but suffer from slow autoregressive inference, limiting their deployment in real-time applications. We introduce Spec-LLaVA, a system that applies speculative decoding to accelerate VLMs without sacrificing output quality. Spec-LLaVA pairs a lightweight draft VLM with a large target model: the draft speculates future tokens, which the target verifies in parallel, allowing multiple tokens to be generated per step. To maximize efficiency, we design a dynamic tree-based verification algorithm that adaptively expands and prunes speculative branches using draft model confidence. On MS COCO out-of-domain images, Spec-LLaVA achieves up to 3.28× faster decoding on LLaVA-1.5 (7B, 13B) with no loss in generation quality. This work presents a lossless acceleration framework for VLMs using dynamic tree-structured speculative decoding, opening a path toward practical real-time multimodal assistants. Importantly, the lightweight draft model design makes the framework amenable to resource-constrained or on-device deployment settings. Project page: https://zhangjiayi24.github.io/Spec-LLaVA/.

1. Introduction

Large vision-language models (VLMs), such as LLaVA (Liu et al., 2023), combine image understanding with language generation to enable rich multimodal interactions. However, their autoregressive decoding and large parameter sizes make inference slow. Generating a single response may require hundreds of forward passes through a 7B or 13B model, resulting in high latency that hinders real-time deployment. Existing acceleration methods—such as quantization (Frantar et al., 2022), early exit (Schuster et al., 2021), or distillation (Hinton et al., 2015)—offer limited speedups (up to ∼3.5×) and often degrade output quality or require extensive tuning.

Recently, *speculative decoding* (Leviathan et al., 2023) has emerged as a promising approach for accelerating language model inference without altering outputs. A small draft model predicts several tokens ahead, which the target model then verifies in parallel. If predictions match, multiple tokens are accepted in a single step, significantly reducing compute. This yields lossless acceleration—output identical to the baseline, but faster. While previous works such as SpecInfer (Miao et al., 2023), EAGLE-2 (Li et al., 2024b), OPT-Tree (Wang et al., 2024), and Sequoia (Chen et al., 2024) have applied this technique to LLMs, they focus on static sequences or predefined trees. For instance, Wen et al. (Wen et al., 2024) proposed a CTC-based drafting method to improve acceptance in text-only decoding. However, speculative decoding remains unexplored for multimodal models, where greater output variability demands more flexible strategies.

We present Spec-LLaVA, a system that extends speculative decoding to VLMs. It combines a small, distilled draft model (68M or 160M parameters) with a full-scale LLaVA-1.5 target. Both draft and target take image and text inputs, enabling speculative token trees guided by visual grounding. This grounding imposes semantic constraints that improve alignment between draft and target distributions. The compact draft model also enables low-latency inference in resource-constrained settings such as mobile or edge devices (Xu et al., 2024), where full VLMs are impractical.

To maximize accepted tokens, we introduce a dynamic tree-based verification algorithm inspired by OPT-Tree (Wang et al., 2024) and adapted for uncertainty-aware decoding. When confident, the draft expands a narrow tree; when uncertain, it explores multiple branches. A leaf-to-root verification strategy ensures exact match with the target model, enabling lossless acceleration. This architecture supports a hybrid inference setup, where speculative generation occurs locally, with periodic verification deferred to a larger model

[*]Equal contribution [1]Carnegie Mellon University [2]University of Nottingham [3]The University of Hong Kong [4]The Hong Kong Polytechnic University [5]Columbia University [6]University of California, Berkeley.

Published at TTODLer Workshop at ICML 2025, Vancouver, Canada.

in the cloud or server. Our contributions are as follows:

- We propose Spec-LLaVA, a speculative decoding method for vision-language models, achieving lossless acceleration without compromising output quality.
- We develop a dynamic tree-based verification algorithm for Spec-LLaVA that adapts structure via draft confidence, beating static or fixed-width methods.
- We construct small draft VLMs trained with the same data and loss as the target, improving acceptance length and reducing KL divergence via distillation.
- Experiments on MS COCO and out-of-domain images show up to 3.28× speedup on LLaVA-1.5 (7B/13B), with analysis of alignment, efficiency, and scalability.

2. Related Work

Speculative decoding was initially proposed to accelerate large language models (LLMs) by using a lightweight draft model to generate candidate tokens, which are then verified in parallel by a larger target model (Leviathan et al., 2023; Chen et al., 2023). This enables lossless acceleration where outputs remain unchanged while latency is reduced. Early implementations such as Draft-and-Verify (Zhang et al., 2023) used simple linear verification, while later approaches like Medusa (Cai et al., 2024) and PASS (Monea et al., 2023) introduced multi-head decoding and parallel sampling to improve throughput.

Subsequent works explored tree-based speculative decoding. SpecInfer (Miao et al., 2023) and EAGLE-2 (Li et al., 2024b) used static trees with fixed-width branching, which are less effective under varying draft confidence. OPT-Tree (Wang et al., 2024) introduced adaptive branching with efficiency guarantees, while Sequoia (Chen et al., 2024) applied global dynamic programming for optimal tree construction. More recent methods such as BiTA (Lin et al., 2024) enabled lossless acceleration via bidirectional tuning and self-executed trees, and NEST (Li et al., 2024a) enhanced speculative decoding with nearest-neighbor retrieval. Hydra (Ankner et al., 2024) improved draft model quality through sequentially-dependent draft heads, highlighting the role of refinement.

For VLMs, prior acceleration strategies include distillation (Hinton et al., 2006; Zhou et al., 2023), quantization (Shoeybi et al., 2019), and model simplification like MoE (Rajbhandari et al., 2022), often trading off quality or requiring retraining. Spec-LLaVA is the first to apply speculative decoding to VLMs. Its dynamic tree-based inference with visual grounding enables lossless, efficient generation, and the lightweight draft model supports low-latency deployment in edge settings.

3. Intuition for VLM Speculation

Speculative decoding is particularly effective for vision-language models due to several factors. First, visual inputs often provide strong grounding that constrains the space of plausible textual outputs. For example, given an image of a cat and the prompt "What is the animal doing?", both small and large VLMs are likely to begin with similar responses such as "The cat is". This visual context reduces uncertainty, increasing the likelihood that the draft model's guesses align with the target model's outputs. The reduced entropy in early token distributions creates favorable conditions for multi-token acceptance.

Second, many VLM tasks are descriptive or factual in nature, such as captioning or visual question answering. These outputs require less linguistic variation or creativity than open-ended text generation, making them easier for a small model to predict accurately. As a result, the draft and target distributions tend to be well aligned over many steps.

Third, we apply the same training manner to train the draft model on outputs from the target VLM. This minimizes the divergence between the two models by explicitly teaching the draft to mimic the target's behavior, including stylistic preferences and phrasing. For example, if the target often begins answers with "Sure, here is ...", the draft will learn to replicate that prefix, improving acceptance. Such stylistic alignment improves not only local prefix matching but also global structural consistency.

Together, these factors contribute to long acceptance lengths during inference, even with relatively small draft models. Our empirical results confirm that VLMs are well suited to speculative decoding, achieving substantial speedup without compromising output fidelity. These properties also suggest that small, distilled draft models can serve as effective local inference agents for real-time speculative generation on resource-constrained or on-device platforms.

4. Method

4.1. Draft Model Construction

We use LLaVA-1.5 as the target vision-language model, which integrates a CLIP ViT-L/14 vision encoder with a LLaMA-based language decoder (7B or 13B parameters). To build a lightweight draft model, we construct two variants—LLaVA-68M and LLaVA-160M—sharing the same vision encoder to avoid redundant image encoding. The language decoders are significantly smaller, containing 68M and 160M parameters respectively.

The 68M model uses an 8-layer Transformer with hidden size 512 and 8 attention heads, while the 160M model employs 12 layers with hidden size 768 and 12 heads. CLIP-extracted image features are projected into the language

embedding space. Both models take image-prompt pairs as input and generate speculative continuations. These compact architectures support fast speculative generation under compute and memory constraints, making them particularly suitable for deployment in edge or embedded systems.

To align the draft model with the target distribution, we apply a same training manner procedure: the draft is trained using the same multimodal instruction data as LLaVA-1.5, minimizing KL divergence with respect to the target model's output distribution.

As shown in Fig. 1, our experiments on large language models reveal a clear correlation between KL divergence and acceptance length. Specifically, a smaller KL divergence consistently leads to improved acceptance length. Motivated by this observation, we design the draft model to match the target model in both training methodology and dataset. This alignment helps maintain a low KL divergence, thereby improving the overall quality and efficiency of the decoding process.

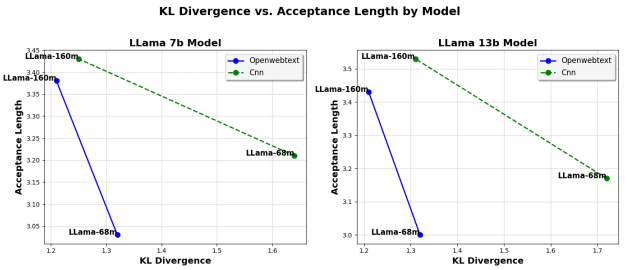

Figure 1: Lower draft-target KL divergence is associated with longer acceptance lengths, indicating better alignment.

4.2. Dynamic Tree-Based Verification

At inference time, the draft model generates a speculative token tree rooted at the current decoding context. The branching factor at each step is determined dynamically by the draft model's token-level confidence: if the distribution is peaked (low entropy), the top-1 token is used; if uncertain, multiple top-b tokens are expanded. During inference, the tree is pruned based on output logits, retaining only the tree structure and the top-n tokens for the whole tree.

Verification proceeds in a leaf-to-root manner. The target model traverses the draft tree, comparing its predicted token at each step with the candidates generated by the draft. If a match is found at the current depth, the token is accepted and the verification proceeds to the next step. Otherwise, the speculative block is truncated, and the target model resumes greedy generation from that point onward. This conservative strategy guarantees that the final output is identical to that of the target model running alone.

Compared to static tree-based decoding (e.g., SpecInfer, EAGLE) or global dynamic programming (e.g., Sequoia), our approach performs online, heuristic tree expansion using draft model logits. It requires no offline optimization, enabling seamless integration into VLM pipelines. Inspired by OPT-Tree, we further prune invalid branches early during traversal, reducing wasted computation and enabling longer acceptance spans. This method maximizes output entropy, which increases the likelihood of accepting tokens during speculative decoding.

5. Experiments

We evaluate Spec-LLaVA on vision-language generation tasks to investigate speculative decoding effectiveness in multimodal contexts. Specifically, we focus on: (1) practical speedup achieved, (2) output quality preservation, and (3) the influence of draft model size and alignment on acceptance length and acceleration.

Setup. We use LLaVA-1.5 (7B/13B) as target models with two lightweight drafts (68M and 160M). The evaluation includes 200 image-prompt pairs from MS COCO and a small out-of-domain set, covering descriptive captioning and visual question answering. All experiments run on a single NVIDIA L40 GPU, comparing Spec-LLaVA against baseline greedy decoding. We report wall-clock decoding times, average acceptance length (γ), and verify output exactness to baseline.

Table 1: Comparison of KL divergence and acceptance length for fine-tuned draft model and original model

Target	**Draft**	**KL** ($\downarrow$)	**Length** ($\uparrow$)
Llama2-7B	JF68M	1.32	3.03
Llama2-7B	FT-JF68M	**1.19**	**3.29**
Llama2-13B	JF68M	1.32	3.00
Llama2-13B	FT-JF68M	**1.19**	**3.27**

Acceptance Length and KL Divergence. We hypothesize that reducing the KL divergence between the draft and target models leads to improved speculative decoding performance. To test this, we fine-tune the draft model on the same dataset used to train the target model, encouraging the two models' output distributions to align more closely. As shown in Table 1, fine-tuning the draft model (FT-JF68M) consistently reduces the KL divergence and increases the acceptance length across both LLaMA2-7B and LLaMA2-13B target models. For instance, with LLaMA2-13B, fine-tuning reduces the KL divergence from 1.32 to 1.19 and improves the acceptance length from 3.00 to 3.27. These results validate our hypothesis and motivate our design choice: to construct the draft model for VLMs using the same training data and methodology as the target model, thereby mini-

Table 2: Acceptance length and speedup for Spec-LLaVA on LLaVA-1.5 (7B and 13B)

Target	Draft	γ (↑)	Speedup (↑)
LLaVA-7B	68M	2.5	2.41×
LLaVA-7B	160M	3.5	3.28×
LLaVA-13B	68M	2.1	2.12×
LLaVA-13B	160M	3.0	2.95×

mizing KL divergence and enhancing speculative decoding efficiency.

Speedup and Acceptance Length. Table 2 summarizes Spec-LLaVA's performance. The 160M draft model provides up to 3.28× speedup on LLaVA-7B and 2.95× on LLaVA-13B, outperforming the 68M model due to longer acceptance spans (γ of 3.5 vs. 2.5). Dynamic branching significantly boosts performance, particularly for smaller drafts. Allowing multiple speculative branches during uncertain predictions improves acceptance length by 15–20% compared to a single-path strategy. In contrast, the more confident 160M draft requires fewer branches, demonstrating the adaptive efficiency of dynamic speculative decoding.

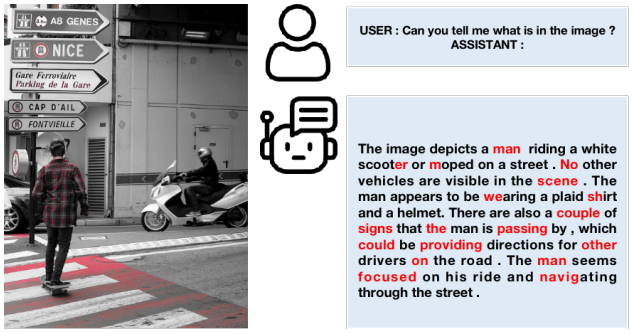

Figure 2: Speculative decoding in Spec-LLaVA (COCO example (Lin et al., 2014)). Draft tokens accepted are shown vs. those verified by the target, showing efficiency.

6. Benchmarking Methodology

To ensure reliable and reproducible speedup measurements, we benchmark all methods under controlled conditions. All models run on the same NVIDIA L40 GPU. To reduce variability from visual processing, image features are pre-extracted via the CLIP encoder. Each image-prompt pair is processed sequentially without batching to simulate realistic interactive use cases such as assistants or captioning tools. This setup also reflects low-latency scenarios typical of on-device or edge deployments.

We measure decoding latency from the first token to the final output. For speculative decoding, we log the number of verification cycles and accepted tokens per cycle. Speedup is computed as the ratio of baseline decoding time to speculative decoding time, with outputs verified as identical. All results use greedy decoding ($T = 0$), ensuring determinism and removing sampling variance in timing or quality.

7. Model Implementation Details

We construct draft models (68M and 160M) to explore the tradeoff between size and decoding efficiency. These are empirically selected: the 68M model offers lower resource cost and faster training, while the 160M variant better aligns with the target distribution. Both reuse the CLIP ViT-L/14 encoder from LLaVA-1.5 and accept image-text inputs via LLaVA's interface. The 68M draft is especially suited for resource-constrained or on-device deployment.

Training uses 600K image-prompt pairs with AdamW (learning rate 1e-4, linear decay) for three epochs on 8 A100 GPUs with mixed precision. Distillation minimizes a weighted sum of cross-entropy and KL divergence w.r.t. target outputs. The KL term improves alignment and increases acceptance length. Checkpoints are chosen by validation acceptance length. All drafts share the target's tokenizer; longer training improves alignment without overfitting.

8. Conclusion

We present Spec-LLaVA, a framework to apply speculative decoding to vision-language models in a lossless manner. By combining a compact draft model with a dynamic tree-based verification algorithm, Spec-LLaVA achieves up to 3.28× faster decoding without compromising output quality or altering outputs. Our results show that VLMs are well suited to speculative decoding due to grounded semantics, predictable output patterns, and strong alignment between visual and linguistic representations. The ability to offload draft inference to lightweight local models also makes the framework attractive for edge or on-device deployment.

This work opens several directions for future research. Combining speculative decoding with quantization or cascading draft models may yield further speedups. Extensions to multi-turn dialogues, long-form visual reasoning, and modalities like video or audio are promising. These directions could enable real-time generation for more complex multimodal systems, especially via hybrid pipelines combining local speculative generation with remote validation.

References

Ankner, Z., Parthasarathy, R., Nrusimha, A., Rinard, C., Ragan-Kelley, J., and Brandon, W. Hydra: Sequentially-dependent draft heads for medusa decoding. *arXiv*

preprint arXiv:2402.05109, 2024.

Cai, T., Li, Y., Geng, Z., Peng, H., Lee, J. D., Chen, D., and Dao, T. Medusa: Simple llm inference acceleration framework with multiple decoding heads. *arXiv preprint arXiv:2401.10774*, 2024.

Chen, C., Borgeaud, S., Irving, G., Lespiau, J.-B., Sifre, L., and Jumper, J. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Chen, Z., May, A., Svirschevski, R., Huang, Y., Ryabinin, M., Jia, Z., and Chen, B. Sequoia: Scalable, robust, and hardware-aware speculative decoding. *arXiv preprint arXiv:2402.12374*, 2024.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv e-prints*, pp. arXiv–1503, 2015.

Hinton, G. E., Osindero, S., and Teh, Y. W. A fast learning algorithm for deep belief nets. *Neural Computation*, 18: 1527–1554, 2006.

Leviathan, Y., Kalman, M., and Matias, Y. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pp. 19274–19286. PMLR, 2023.

Li, B., Zhang, B., Xu, Z., Liu, S., Shi, H., and Li, L. Nearest neighbor speculative decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2024a.

Li, Y., Wei, F., Zhang, C., and Zhang, H. Eagle-2: Faster inference of language models with dynamic draft trees. *arXiv preprint arXiv:2406.16858*, 2024b.

Lin, F., Yi, H., Li, H., Yang, Y., Yu, X., Lu, G., and Xiao, R. Bita: Bi-directional tuning for lossless acceleration in large language models. *arXiv preprint arXiv:2401.12522*, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Miao, X., Oliaro, G., Zhang, Z., Cheng, X., Wang, Z., Zhang, Z., Wong, R. Y. Y., Zhu, A., Yang, L., Shi, X., et al. Specinfer: Accelerating generative large language model serving with tree-based speculative inference and verification. *arXiv preprint arXiv:2305.09781*, 2023.

Monea, G., Joulin, A., and Grave, E. Pass: Parallel speculative sampling. *arXiv preprint arXiv:2311.13581*, 2023.

Rajbhandari, S. et al. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022. URL https://proceedings.mlr.press/v162/rajbhandari22a.html.

Schuster, T., Fisch, A., Jaakkola, T., and Barzilay, R. Consistent accelerated inference via confident adaptive transformers. In *Proceedings of EMNLP 2021*, pp. 4962–4979, 2021. doi: 10.18653/v1/2021.emnlp-main.406.

Shoeybi, M., Patwary, M. M. A., et al. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. URL https://arxiv.org/abs/1909.08053.

Wang, J., Su, Y., Li, J., Xia, Q., Ye, Z., Duan, X., Wang, Z., and Zhang, M. Opt-tree: Speculative decoding with adaptive draft tree structure. *arXiv preprint arXiv:2406.17276*, 2024.

Wen, Q., Chen, R., Yan, X., Zhao, W. X., and Wen, J.-R. Fast and lossless llm decoding via ctc-based drafting. In *Proceedings of NeurIPS 2024*, 2024.

Xu, J., Zhao, Y., Sun, J., Lin, J., and Han, S. On-device language models: A comprehensive review. *arXiv preprint arXiv:2403.05645*, 2024.

Zhang, J., Wang, J., Li, H., Shou, L., Chen, K., Chen, G., and Mehrotra, S. Draft & verify: Lossless large language model acceleration via self-speculative decoding. *arXiv preprint arXiv:2309.08168*, 2023.

Zhou, Y., Lyu, K., Rawat, A. S., Menon, A. K., Rostamizadeh, A., Kumar, S., Kagy, J.-F., and Agarwal, R. Distillspec: Improving speculative decoding via knowledge distillation. *arXiv preprint arXiv:2310.08461*, 2023.