On Speculative Decoding for Multimodal Large Language Models

Mukul Gagrani* Raghavv Goel* Wonseok Jeon Junyoung Park Mingu Lee Christopher Lott
Qualcomm AI Research

Abstract

Inference with Multimodal Large Language Models (MLLMs) is slow due to their large-language-model backbone which suffers from memory bandwidth bottleneck and generates tokens auto-regressively. In this paper, we explore the application of speculative decoding to enhance the inference efficiency of MLLMs, specifically the LLaVA 7B model. We show that a language-only model can serve as a good draft model for speculative decoding with LLaVA 7B, bypassing the need for image tokens and their associated processing components from the draft model. Our experiments across three different tasks show that speculative decoding can achieve a memory-bound speedup of up to 2.37× using a 115M parameter language model that we trained from scratch. Additionally, we introduce a compact LLaVA draft model incorporating an image adapter, which shows marginal performance gains in image captioning while maintaining comparable results in other tasks.

1. Introduction

Large Language Models (LLMs) have become ubiquitous across various domains due to their impressive performance. However, LLMs only take text queries as input but real-world data comes in the form of multiple modalities including visual data. Multi-modal Large Language Models (MLLMs) [1, 13, 21, 22] provides the LLMs with image understanding abilities, and the fusion of visual and textual tokens enhances the model's interaction with users, leading to more informed responses. MLLMs comprise of an image encoder to process the image information and an adapter which transforms the image encodings to the language model embedding space. In addition, MLLMs have a language-model backbone in the form of a LLM and thus inherit the auto-regressive generation and memory-bandwidth bottleneck which lead to slow inference [19].

Speculative decoding [3, 7, 9, 15, 20] has been proposed as a solution to accelerate the LLM inference without loss in accuracy, where a smaller draft model predicts multiple future tokens which are verified in a single call of the LLM. Given that MLLMs have a LLM backbone, speculative decoding can be used to make inference with MLLMs more efficient. Many recent works have studied the application of speculative decoding and its variants [2, 5, 7, 8, 18, 20] for LLMs, but no such work exists in the context of MLLMs to the best of our knowledge.

In this paper, we apply speculative decoding to LLaVA 7B model (with LLaMA 7B model as language-model backbone) to make inference more efficient, block diagram shown in Figure 1. Due to the lack of publicly available models of LLaVA and LLaMA families smaller than 7B parameters, we train a language model of size 115M from scratch for speculative decoding. We show that language-only model which does not consider the image tokens (and hence does not require the image encoder and adapter) can serve as a good draft model for LLaVA 7B. We conduct experiments on three different tasks including image QA on LLaVA Instruct 150K dataset [13], image captioning on Coco dataset [11] and ScienceQA dataset [14], using draft model candidates which have gone through different stages of training and fine-tuning. Our results show that we can achieve memory-bound speedup of upto 2.37× using only a language model as draft model. We also create a small LLaVA draft model which consists of an image adapter along with our trained language model and show that it improves the performance slightly on COCO captioning task and ScienceQA task while performing similar to language-model-only draft models on the other tasks.

2. Background

2.1. Speculative Decoding

SPeculative Decoding (SPD) [3, 9] involves a smaller draft model generating multiple tokens which are verified in parallel by the target LLM. Given an input con-

*Equal contribution.

Correspondence to {mgagrani,raghgoel,mingul}@qti.qualcomm.com.
Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

text $X_{1:n} := [X_1, \ldots, X_n]$, the draft model generates a sequence of tokens $\hat{X}_{n+1:n+L}$ in an auto-regressive fashion, $\hat{X}_{n+j} \sim p(\cdot|X_{1:n}, \hat{X}_{n+1:n+j-1})$. The draft tokens are then verified via a single call of the target LLM (q) using rejection sampling criteria that guarantees the same output token distribution as that of the target LLM. Specifically, token $\hat{X}_{n+j}$ is accepted with probability

$$\min\left\{1, \frac{q(\hat{X}_j|X_{1:n}, \hat{X}_{n+1:n+j-1})}{p(\hat{X}_j|X_{1:n}, \hat{X}_{n+1:n+j-1})}\right\}.$$

If a draft token $\hat{X}_{n+j}$ is rejected, then a new token is sampled from the residual distribution defined as $p_{res}(x) = \max(0, q(x) - p(x))$.

2.2. Multimodal Large Language Models

An image-based Multimodal Large Language Model (MLLM) consists of *1) a vision encoder* to encode the input image, *2) an adapter* to convert the image encodings to language model embeddings, and *3) a language-model backbone*. We describe the framework of the LLaVA model in more detail as follows; given an input image I and the text query Q, the image I is converted into a sequence $H_1, H_2, \ldots, H_m$ of m image encodings, and the text query is converted to a sequence of token embeddings $X_1, X_2, \ldots X_n$. The image encodings are further transformed via an adapter g_θ (a small multi-layer perceptron) to get image embeddings, $V_i = g_\theta(H_i)$. This is done to convert the encodings H_i to the language model embedding space. Tokens are then generated by the language model conditioning on the image embeddings and the token embeddings as follows:

$$X_{n+1} \sim q(\cdot|V_{1:m}, X_{1:n}) \quad (1)$$

3. SPD for MLLMs

To achieve higher gain with speculative decoding, we need a draft model significantly smaller than and well-aligned with our target model (LLaVA-7B). The most common choice for draft models in prior works on LLMs is to use a small pre-trained model from the same family of models as the target model or train a smaller model which has the same architecture as the target model [15]. Since there is no publicly available smaller model in the LLaVA family, we need to train a draft model from scratch. A natural choice for draft model architecture is to follow LLaVA's architecture where the draft model comprises an adapter and a language-model backbone with smaller number of parameters than the LLaVA 7B. In our approach, we use both, *1) a smaller LLaVA draft model* which consists of a smaller image adapter and a draft language model, and *2) the language-only draft model* which generates draft tokens by conditioning only on the input text tokens. Given an input image with image embeddings $V_{1:m}$, token embeddings $X_{1:n}$ the draft model generates the draft tokens $\hat{X}_{n+1:n+L}$ where the draft token

$$\hat{X}_{n+j} \sim p(\cdot|X_{1:n}, \hat{X}_{n+1:n+j-1})$$

is generated by conditioning only on the text tokens. The target LLaVA model verifies the draft tokens by computing the target distribution which is conditioned on both the image embeddings $V_{1:m}$ and the text token embeddings $X_{1:n}$, i.e., draft token $\hat{X}_{n+j}$ is accepted with probability

$$\min\left\{1, \frac{q(\hat{X}_{n+j}|V_{1:m}, X_{1:n}, \hat{X}_{n+1:n+j-1})}{p(\hat{X}_{n+j}|X_{1:n}, \hat{X}_{n+1:n+j-1})}\right\}.$$

Using the language-model-only draft model is more efficient than a draft model with LLaVA architecture since *1) it does not need an additional adapter* as it does not condition on the image embeddings for generating draft tokens, and *2) it does not require the training of the adapter*. Figure 1 shows SPD with MLLM consisting of the smaller draft language model doing autoregressive generation followed by the large target model evaluating the draft model predicted tokens in parallel while using the image.

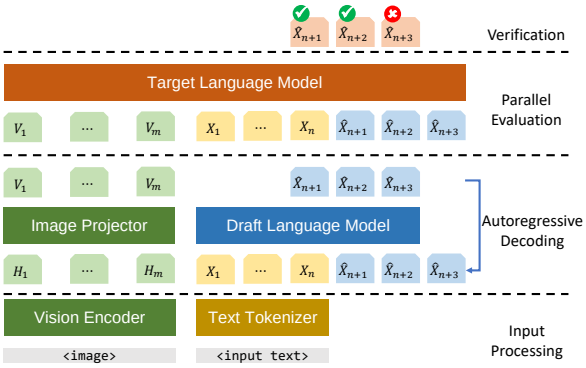

Figure 1. SPD with a MLLM as target having three components: vision encoder, image projector, and target language model, and the smaller language model as draft. The small draft model generates draft tokens autoregressively for block-size number of iterations followed by parallel evaluation by the target language model which also uses image features.

4. Experiments

We run experiments on three visual instruction tasks using SPD with LLaVA-7B [12] as our target model which uses the LLaMA-7B model as the language-model backbone. We employ draft models that underwent different stages of training with the size of the language part of each draft model fixed to 115M.

Draft Model Candidates. We train draft model of size $115M$ which follow the LLaMA-2 architecture. We follow the training pipeline of [6] to pre-train a draft model from scratch and fine-tune the draft model on instruction finetuning datasets using TVD++ loss [6]. We further fine-tune our draft model on a subset of LLaVA Instruct 150K dataset [13]. For our experiments, we consider the following four draft models after each stage of training and finetuning: *1) base-LLaMA*, a draft LLaMA model after pre-training using next-token-prediction loss on 600B English tokens, *2) chat-LLaMA*, an instruction fine-tuned draft LLaMA model following [6] initialized with base-LLaMA draft model, and *3) fine-tuned-LLaVA* (ft-llava), a fine-tuned LLaVA draft model where the image adapter was initialized using subcloning [17] of LLaVA-7B image adapter and the language model was initialized from the chat-LLaMA draft model (the model is then fine-tuned on LLaVA dataset). We also include another draft model *4) fine-tuned-LLaVA-text* (ft-llava-text), which simply uses the language model part of *3)*. Note that only the fine-tuned-LLaVA draft model uses image information while all other draft models only consume the text part of the input prompt; when the draft model uses image information, the vision encoder (CLIP-based [16]) is shared with the target model to avoid re-computation of image embeddings. The detailed parameters are given in Appendix A.1

Evaluation Tasks. We focus on open-ended text generation and multiple choice question-answering with reasoning to encourage a higher number of token generation, which is beneficial when using SPD. To this end, we evaluate on 1) **LLaVA Instruct 150K dataset** [13], 2) Image captioning task on images from **COCO dataset** [11], and 3) **Science QA (SQA)** with chain-of-thought (CoT) reasoning [14]. The system prompts settings for all the tasks are described in Appendix A.2

Metrics. The efficacy of SPD is evaluated with the following metrics; 1) **block efficiency** (τ), the average number of tokens generated per block (or target model run), for a block of size γ and input x, the maximum value of $\tau(x)$ can be $\gamma + 1$, block-size (γ) is also known as draft length (DL) in some works; 2) memory-bound speedup (**MBSU**), the hypothetical speedup achieved by SPD for a given block efficiency $\tau(x)$ and a relative latency c defined as ratio between number of parameters of draft to target model, i.e., $\text{MBSU}(x) = \frac{c\tau(x)}{c\gamma+1}$; 3) **token rate**, the total number of tokens generated divided by the total time of generation, giving an estimate of tokens generated for per second. We measure these metrics on various tasks using different block size γ in $\{3, 5\}$

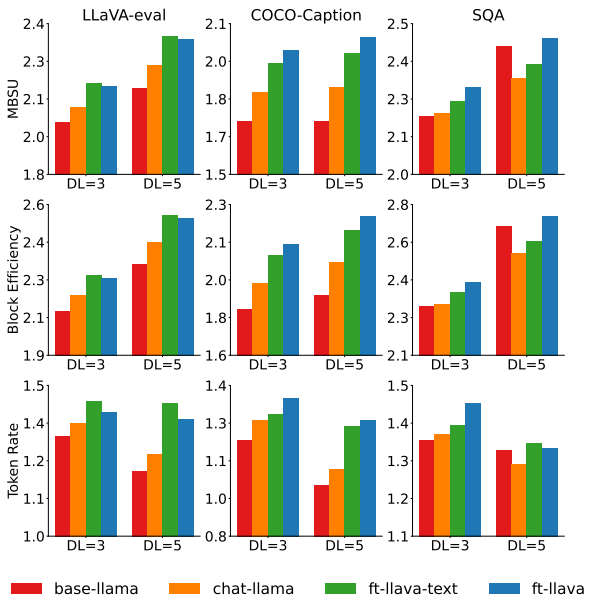

Figure 2. MBSU, block efficiency and token rate (relative to auto-regressive generation) for SPD are depicted; We consider LLaVA-eval, COCO-Caption and SQA datasets for evaluation; For draft models, base-LLaMA, chat-LLaMA, ft-LLaVA-text, ft-LLaVA are considered, we consider three text-only draft models and a single text and image draft model; For draft length (DL) (or block size of SPD), we consider either 3 or 5.

Decoding. We use greedy decoding for all experiments so that the SPD generation is identical to the target model's autoregressive generation. We leave it as future work to explore sampling-based decoding (varying temperature, varying top-p, top-k) in the context of SPD for MLLMs.

Results. Our results show that using SPD with LLaVA 7B target model gives considerable speedup in output generation, and we emphasize that when using a draft model without any image information, SPD can still give considerable and competitive speedup to that of a draft model using image information.

From Figure 2 (top and middle plots), we observe that using SPD gives more than $2\times$ gains in terms of block efficiency and MBSU. The performance trend when increasing the block size from 3 to 5 for each task is similar with the exception for SQA task where base-llama draft model performs better than other text-only draft models for block size $=$ 5. For LLaVA-eval task on both block sizes (either 3 or 5), the ft-llava-text draft model performs the best closely followed by ft-llava. For COCO-caption task, ft-llava performs best, followed by ft-llava-text for both block sizes. Lastly, for the SQA task, for block size=3, ft-llava draft model performs the best followed by ft-llava-text while for block-size=5, ft-

Figure 3. Example 1

Input prompt: ⟨s⟩A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: ⟨image⟩ Provide a detailed description of the given image ASSISTANT:
Output generation: The image features a white plate filled with a variety of vegetables, including several carrots, green beans, and broccoli. The carrots are scattered across the plate, with some placed closer to the center and others towards the edges. The green beans are also spread out, with some near the center and others closer to the edges. The broccoli is located towards the top left corner of the plate. The arrangement of the vegetables creates a colorful and appetizing display.

Figure 4. Example 2

Input prompt: ⟨s⟩A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: ⟨image⟩ Provide a detailed description of the given image ASSISTANT:
Output generation: The image depicts a kitchen counter filled with various food items and kitchen utensils. There are several bowls placed on the counter, with one large bowl towards the left side and two smaller bowls closer to the center. A few cups can be seen as well, with one near the left side of the counter and another towards the right side.
In addition to the bowls and cups, there are multiple bottles scattered across the counter, with some placed near the center and others towards the right side. A spoon is also visible on the counter, located near the center.
The kitchen counter is surrounded by various appliances, including an oven on the right side, a sink in the background, and a refrigerator on the left side. A microwave can be seen above the counter, and a knife is placed near the right edge of the counter.

Figure 5. SPD examples on COCO-caption task

llava draft model performs best followed by base-llama. In addition, all our draft models show the improved token rate compared to auto-regressive generation in Figure 2 (bottom) with block size 3 giving better token rate than block size 5, thus, SPD generates more token tokens per second than autoregressive decoding. The token rate shown corresponds to the ratio of the token rate of SPD using a particular draft model to the token rate of autoregressive generation using target model.

We further provide qualitative results on the COCO-captioning task to show the tokens accepted during a generation process when using the fine-tune-LLaVA-text draft model so no image information used by draft model in Figure 5. Based on the output generations in the figure, where tokens in blue and underlined are the accepted tokens, we observe that the draft model can predict common words and propositions, along with halves of words. For example, the draft model can predict "tables" given "vege". Similarly in the second example, given the context and additional token "app", the draft model was able to predict "liances". We believe in general open-ended text generation has several tokens comprising of common words, propositions, and word completions which do not require knowledge of image tokens, thus, even a draft model without using image information gives competitive performance. Moreover, draft model can also predict the repetition of certain tokens once they have been generated. For example, in the second image the word "counter" and "bowls" can be predicted by the draft model multiple times once it has been generated by the target model. Lastly, performing more rigorous training on a small multi-modal language model is left as our future work.

5. Conclusion

In this paper, we present the first effort towards using speculative decoding for accelerating inference when using multi-modal large language models, specifically for image-text domain. We show that using the text-only draft model achieves performance competitive to using a draft model utilizing image features. We perform various experiments on different visual question-answering tasks focusing on generating higher number output tokens: open-ended text generation and text generation with reasoning using different draft models (text-only and image-text). We achieved significant speedup of upto 2.37× for text-only draft model and marginal better speedup for image-text draft model, empirically showing the potential of using SPD for MLLMs.

Our work opens several future avenues owing to the general framework presented. Our work can be extended to other target models such as BLIP-2 [10], MiniGPT-4 [22] and OpenFlamingo [1], and other modalities such as audio [4] which are also bottlenecked by auto-regressive generation. Furthermore, recent advancement in SPD algorithm to tree-based decoding can also be used following [2, 7, 15, 20] to further increase generation speed.

References

[1] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1, 4

[2] Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, and Tri Dao. Medusa: Simple framework for accelerating llm generation with multiple decoding heads. https://github.com/FasterDecoding/Medusa, 2023. 1, 4

[3] Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023. 1

[4] Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023. 4

[5] Yichao Fu, Peter Bailis, Ion Stoica, and Hao Zhang. Breaking the sequential dependency of LLM inference using lookahead decoding, 2023. 1

[6] Raghavv Goel, Mukul Gagrani, Wonseok Jeon, Junyoung Park, Mingu Lee, and Christopher Lott. Direct alignment of draft model for speculative decoding with chat-fine-tuned llms. *arXiv preprint arXiv:2403.00858*, 2024. 3

[7] Wonseok Jeon, Mukul Gagrani, Raghavv Goel, Junyoung Park, Mingu Lee, and Christopher Lott. Recursive speculative decoding: Accelerating llm inference via sampling without replacement. *arXiv preprint arXiv:2402.14160*, 2024. 1, 4

[8] Sehoon Kim, Karttikeya Mangalam, Jitendra Malik, Michael W Mahoney, Amir Gholami, and Kurt Keutzer. Big little transformer decoder. *arXiv preprint arXiv:2302.07863*, 2023. 1

[9] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. 1

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3

[12] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2

[13] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1, 3, 6

[14] 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. 1, 3, 6

[15] Xupeng Miao, Gabriele Oliaro, Zhihao Zhang, Xinhao Cheng, Zeyu Wang, Rae Ying Yee Wong, Zhuoming Chen, Daiyaan Arfeen, Reyna Abhyankar, and Zhihao Jia. SpecInfer: Accelerating generative LLM serving with speculative inference and token tree verification. *arXiv preprint arXiv:2305.09781*, 2023. 1, 2, 4

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 6

[17] Mohammad Samragh, Mehrdad Farajtabar, Sachin Mehta, Raviteja Vemulapalli, Fartash Faghri, Devang Naik, Oncel Tuzel, and Mohammad Rastegari. Weight subcloning: direct initialization of transformers using larger pretrained ones. *arXiv preprint arXiv:2312.09299*, 2023. 3

[18] Andrea Santilli, Silvio Severino, Emilian Postolache, Valentino Maiorca, Michele Mancusi, Riccardo Marin, and Emanuele Rodolà. Accelerating transformer inference for translation via parallel decoding. *arXiv preprint arXiv:2305.10427*, 2023. 1

[19] Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019. 1

[20] Ziteng Sun, Ananda Theertha Suresh, Jae Hun Ro, Ahmad Beirami, Himanshu Jain, and Felix Yu. SpecTr: Fast speculative decoding via optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 4

[21] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 1

[22] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 4