

LLaVA-Gemma: Accelerating Multimodal Foundation Models with a Compact Language Model

Musashi Hinck* Matthew L. Olson* David Cobbley Shao-Yen Tseng
Vasudev Lal
Cognitive AI, Intel Labs
Santa Clara, CA USA

{musashi.hinck, matthew.lyle.olson, david.j.cobbley, shao-yen.tseng, vasudev.lal}@intel.com

Abstract

We train a suite of multimodal foundation models (MMFM) using the popular LLaVA framework with the recently released Gemma family of large language models (LLMs). Of particular interest is the 2B parameter Gemma model, which provides opportunities to construct capable small-scale MMFMs. In line with findings from other papers in this space, we test the effect of ablating three design features: pretraining the connector, utilizing a more powerful image backbone, and increasing the size of the language backbone. The resulting models, which we call LLaVA-Gemma, exhibit moderate performance on an array of evaluations, but fail to improve past the current comparably-sized SOTA models. Closer analysis of performance shows mixed effects; skipping pretraining tends to reduce performance, larger vision models sometimes improve performance, and increasing language model size has inconsistent effects. We publicly release training recipes, code and weights for our models for the LLaVA-Gemma models¹.

1. Introduction

In this paper, we introduce LLaVA-Gemma, a suite of vision-language assistants trained from the Gemma Large Language Model (LLM) variants, Gemma-2B and Gemma-7B [17]. Our work is inspired by the rapid progress in small but capable visual language models (VLMs), such as LLaVA-Phi [23], which have demonstrated remarkable efficiency and effectiveness in various language understanding tasks. LLaVA-Gemma distinguishes itself among small VLMs due to the public release of similarly trained, different-sized LLMs Gemma-2B and Gemma-7B.

The unique release of the Gemma models offers an opportunity to contrast model performance in relation to pa-

rameter size and visual encoding capabilities. By possessing two variants with different parameter sizes, LLaVA-Gemma allows researchers to investigate the trade-offs between computational efficiency and the richness of visual and linguistic understanding. With these two variants, we perform a deeper exploration of how varying levels of model complexity influence the effectiveness of visual encoding, providing valuable insights into the optimization of small VLMs for diverse tasks and environments. Furthermore, the use of significantly more unique tokens, at 256k, offers an opportunity to investigate how a massively increased token set effects multi-modal performance.

Recent advancements in (LLMs) [20] and multimodal foundation models (MMFMs) [7] have propelled the interest and development of Large Multimodal Models (LMMs). Notable models like GPT-4 [1], LLaVA [9, 10], and their derivatives have demonstrated significant performance in vision-language tasks such as Visual Question Answering (VQA) and image captioning [5]. However, the computational demands of deploying these models have led to the exploration of small-scale LMMs. Our work aims to provide a unified analysis of small-scale LMMs, examining how model selections, training recipes, and data contribute to performance, which is distinct from existing works such as LLaVA-Phi.

Our contributions are as follows:

1. We introduce LLaVA-Gemma, a MMFM that leverages the compact yet powerful Gemma language models for efficient multimodal interactions.
2. We extensively evaluate Gemma-2B and Gemma-7B model variants provides valuable insights into the trade-offs between computational efficiency and the richness of visual and linguistic understanding in LLMs.
3. We present a deep exploration into alternate design choices and visualize attention with relevancy maps to enhance our understanding of the model’s performance and attention.

*Equal Contributions, order decided by LLaVA-Gemma 2b

¹<https://huggingface.co/intel/llava-gemma-2b/>,
<https://huggingface.co/intel/llava-gemma-7b/>

Language Backbone	Vision Backbone	Pretrain Connector	MME			MM-Vet	POPE		VQAv2	MMVP	ScienceQA Image
			GQA	Cog.	Per.		Acc.	F1			
gemma-2b-it	CLIP	Yes	0.531	236	1130	17.7	0.850	0.839	70.7	0.287	0.564
gemma-2b-it	CLIP	No	0.481	249	935	13.1	0.784	0.762	61.7	0.180	0.549
gemma-2b-it	DinoV2	Yes	0.587	307	1133	19.1	0.853	0.838	71.4	0.227	0.555
gemma-2b-it	DinoV2	No	0.501	309	959	14.5	0.793	0.772	61.7	0.180	0.568
gemma-7b-it	CLIP	Yes	0.472	254	895	18.2	0.848	0.829	68.7	0.327	0.625
gemma-7b-it	CLIP	No	0.472	278	857	19.1	0.782	0.734	65.1	0.240	0.636
gemma-7b-it	DinoV2	Yes	0.519	257	1021	14.3	0.794	0.762	65.2	0.327	0.628
gemma-7b-it	DinoV2	No	0.459	226	771	12.2	0.693	0.567	57.4	0.267	0.598
Phi-2b	CLIP	Yes	-	-	1335	28.9	-	0.850	71.4	-	0.684
Llama-2-7b	CLIP	Yes	0.620	348	1511	30.6	0.850	0.859	78.5	46.1	0.704

Table 1. Performance of LLaVA-Gemma models across seven benchmarks. Highlighted box indicates strongest performance amongst LLaVA-Gemma models. Bottom two rows show self-reported performance of Llava Phi-2 and LLaVA-v1.5 respectively.

2. Methods

We follow the LLaVA framework [9] with a few design modifications. This framework combines a pretrained vision encoder (such as CLIP [14]) and pretrained language model (such as Llama-2 [19]) into a multimodal model using a MLP connector and a two-stage training procedure.

The first stage pretrains the MLP connector by freezing the vision and language models and training on custom dataset of 595k samples filtered from CC3M [15]. The second stage jointly finetunes the language model and connector using a custom mixture 665k multimodal instruction tuning examples. This dataset includes synthetic data generated [10], as well as examples from established vision-language training sets such as GQA [5] and TextCaps [16].

We deviate from the original recipe in three ways: the language model, the vision encoder and the pretraining stage. For the language backbone, we use the recently released Gemma models [17]. Two aspects of Gemma make it an interesting candidate for our experiments. Whereas LLaVA uses the 7 and 13-billion parameter vicuña language models [22], Gemma offers 2 and 7-billion parameter versions. Next, Gemma uses a significantly larger token set than any other LLM, with 256k unique tokens (compared to a standard 50k), which offers a unique opportunity to see the effects of a massively more diverse embeddings space. Other papers exploring the design space of Vision Language Models (VLMs) find the vision encoder is important for achieving strong performance [12]. Correspondingly, we explore the use of the larger 1-billion parameter DINOv2 image encoder [13] as the vision tower. Related work on VLMs [6] finds that skipping the initial pretraining stage improves downstream performance. For all designs, we train a version with and without the initial pretraining step.

3. Results

We evaluate the LLaVA-Gemma models on a similar collection of benchmarks to other LMM works: **GQA** [5]; **MME** [3]; **MM-Vet** [21]; **POPE** (accuracy and F1) [8]; **VQAv2**

[4]; **MMVP** [18]; the image subset of **ScienceQA** [11]. Our experiments provide insights into the efficacy of various design choices within the LLaVA framework. As shown in table 1, the performance of LLaVA-Gemma models across seven benchmarks reveals interesting patterns, particularly concerning the choice of vision encoder and the impact of pretraining the connector.

One item of note is that for the ScienceQA dataset, the larger models consistently perform better than smaller due to the datasets task requiring diverse general knowledge captured better by the larger models.

3.1. Influence of Vision Encoder on Performance

For the 2B backbone, exchanging the CLIP vision encoder for DinoV2 appears to generally improve performance, with DinoV2 variants outperforming CLIP variants on all benchmarks except POPE-F1 and MMVP. When using a 7B backbone, the picture is murkier; although we see improvements for GQA and MME, we see a decline in performance on MM-Vet, POPE, VQA and ScienceQA. This may suggest an interaction between the capability of the language model and the richness of the representation provided by the vision encoder, or to the possibility that the 7b-Dino combination is undertrained.

3.2. Effects of Pretraining

We find that skipping the initial connector pretraining almost always reduces model performance. With the exceptions of 2B-Dino on MME Cognition and 7B-CLIP on MME Cognition, MM-Vet and ScienceQA, the variant with a pretrained connector outperforms its counterpart that skipped pretraining. These results do not support the hypothesis posited in Karamcheti et al. [6].

3.3. Comparison to Baselines

Contrasting the results of LLaVA-Gemma with the self-reported performances of Phi-2b and Llama-2-7b models provides additional context. The LLaVA-Gemma models

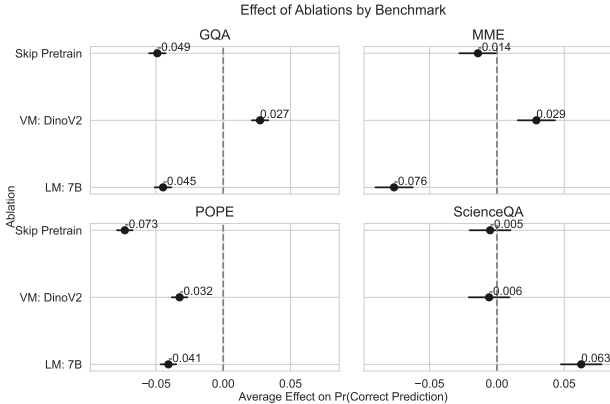


Figure 1. *Effect of design choices differs between evaluations.* Point indicates average change in probability of correct answer versus baseline design.

only reach parity on comparably-sized baselines for the VQA benchmark between 2B models. Given the absence of strong *a priori* reasons to expect Gemma-based LLaVA models to perform worse, understanding this “poor” performance is a direction of future interest.

3.4. Speed of Training and Inference

We compare the training and eval speed for the two models sizes. In our experiments, the training time for the Gemma-2B model on 8 Intel Gaudi 2[®] AI accelerators was 4 hours, while the larger Gemma-7B model required 16 hours to train under the same conditions. This indicates that the Gemma-7B model, with its increased parameter count, takes approximately four times longer to train compared to the Gemma-2B model. The relative speed of the Gemma-7B model is thus 0.25x compared to the Gemma-2B model. We find a similar speed ratio during inference. These results highlight the trade-off between model size and training efficiency, with larger models requiring significantly more computational resources.

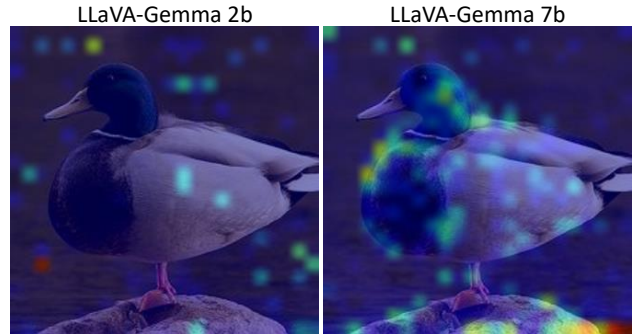
4. Analysis

4.1. Impact of Alternate Design Choices

Table 1 suggests that the `gemma-2b-dino` recipe generally provides stronger evaluation results, but these results are mixed. To better assess the effect of the design choices, we fit a collection of linear models to measure the average associated change in the probability of a correct prediction as a function of each of the three ablations: skipping pre-training, changing the vision backbone, and increasing the size of the LM backbone from 2B to 7B. We study these effects separately for each benchmark.

Figure 1 shows the average effects of design choices for four benchmarks where we have observation-level errors. Skipping pretraining appears to either have a strong negative (GQA, POPE) or weak/insignificant effect (MME, ScienceQA). Changing the vision encoder to DinoV2 improves performance on GQA and MME, but slightly worsens per-

Question: Is the duck floating? (a) Yes (b) No



The duck is **floating** on the water.

(b) No

Figure 2. Relevancy map comparison between LLaVA-Gemma 2b (Left) and LLaVA-Gemma 7b (Right) with gradients on the first relevant output token. For the question “Is the duck floating? (a) Yes (b) No”, despite using the identical CLIP vision encoder, the smaller model does not attend to the visual input.

formance on POPE and has no significant effect on the probability of correct predictions on ScienceQA. Notably, in our experiments increasing the LM backbone to the 7B parameter variant had a strong negative effect on GQA, MME and POPE, but strong positive effect on ScienceQA. Taken together, these heterogeneous results underscore the need for more granular analysis of errors and design choices.

4.2. Visualizing Attention with Relevancy Maps

To better understand the differences between our the LLaVA-Gemma models, we use relevancy maps [2] to visualize where the model focuses its attention. These relevancy maps provide a token-wise understanding of the model’s attention by highlighting the most relevant parts of the input and is specially designed to maintain the total relevancy across layers for transformer based models.

We apply an qualitative example of these relevancy maps from the Eyes-wide-shut (MMVP) dataset. This dataset is of particular interest as it is designed to find image-caption pairs that a CLIP model finds to be similar, but are distinct. As the traditional LLaVA recipe uses CLIP, we compare our CLIP backbone models to find a case where the Gemma 2b model fails, but Gemma 7b is successful.

Figure 2 shows an example of the differences in attention to the visual aspects of the scene between the LLaVA-Gemma 2b and LLaVA-Gemma 7b models. The relevancy maps for the LLaVA-Gemma 2b model show a dispersed and unfocused pattern of attention, which correlates with its failure to accurately interpret the scene. In contrast, the LLaVA-Gemma 7b model exhibits a more concentrated and relevant pattern of attention, particularly focusing border between objects: the duck, the water, and the rock being stood on. This visualization not only highlights the superior performance of the LLaVA-Gemma 7b model, but also illuminates an interesting case where leveraging a more powerful LLM ensures improved visual token attention.

5. Discussion

In this paper, we introduced LLaVA-Gemma, a compact vision-language model leveraging the Gemma Large Language Model in two variants, Gemma-2B and Gemma-7B. Our work provides a unique opportunity for researchers to explore the trade-offs between computational efficiency and multimodal understanding in small-scale models. The availability of both variants allows for a comparative analysis that sheds light on how model size impacts performance in various tasks. Our evaluations demonstrate the versatility and effectiveness of LLaVA-Gemma across a range of datasets, highlighting its potential as a benchmark for future research in small-scale vision-language models. With these models, future practitioners can optimize the performance of small-scale multimodal models more directly.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *Int. Conf. Comput. Vis.*, pages 397–406, 2021. 3
- [3] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2
- [4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6904–6913, 2017. 2
- [5] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6700–6709, 2019. 1, 2
- [6] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*, 2024. 2
- [7] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *arXiv preprint arXiv:2309.10020*, 1(2):2, 2023. 1
- [8] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2
- [9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 2
- [10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Adv. Neural Inform. Process. Syst.*, 36, 2024. 1, 2
- [11] Pan Lu, Swaroop Mishra, Tanglin 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. *Adv. Neural Inform. Process. Syst.*, 35:2507–2521, 2022. 2
- [12] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xi-anzhi Du, Futang Peng, Floris Weers, et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*, 2024. 2
- [13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2
- [14] 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. 2
- [15] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 2
- [16] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Eur. Conf. Comput. Vis.*, pages 742–758. Springer, 2020. 2
- [17] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 1, 2
- [18] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. 2
- [19] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. 2
- [20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017. 1
- [21] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2
- [22] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, et al. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 2
- [23] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava-phi: Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024. 1