Instruction-aware Visual Feature Extraction for Multimodal Large Language Model

Jianhong Tu Erdong Chen Shuhan Zhang

Abstract

We present TA-LLaVA, an instruction-tuned multimodal large language model (MLLM) designed to be efficient and scalable for general vision-language tasks. TA-LLaVA introduces a novel cross-attention adapter design that effectively reduces the number of visual prefix tokens from 576 to 32, significantly reducing inference costs by over 50% compared to LLaVA-1.5, while maintaining strong task performance. Our key innovation lies in instruction-aware visual feature pooling, where visual information extraction is conditioned on the provided instructions, enabling the model to keep relevant visual features efficiently. Despite using a smaller language model and training dataset, TA-LLaVA achieves competitive results, outperforming InstructBLIP on tasks like MME and Science QA. However, we observe limitations such as hallucinations and reduced accuracy on benchmarks requiring precise perception (e.g., POPE), which we attribute to the limited prefix token capacity and insufficient training data. Our future direction includes adding support for multi-image an video inputs and integrating it with more powerful casual LLMs. This work demonstrate a promising step toward efficient and instruction-aware multimodal LLMs. Our code is available at `https://github.com/ToviTu/TA-LLaVA`.

1. Introduction

Recent advancements in large language models (LLMs) and vision-language models (VLMs) have enabled the development of self-supervised methods to learn robust joint semantic spaces for text and vision. These advancements have led to versatile multimodal large language models (MLLMs) that excel in various vision-language tasks, such as visual question answering and visual reasoning [2, 19, 23, 26]. The core objective of this line of research is to extend LLMs to process visual inputs and generate textual responses effectively. Recently, there has been growing interest in developing practical multimodal assistants through a technique known as visual instruction tuning [8,9,20,21]. This method extends the language-only supervised paradigm [24,32] by incorporating multimodal inputs. A dual-phase training paradigm, consisting of multimodal pre-training and supervised fine-tuning has proven to be a simple yet effective way to enhance zero-shot question answering performance in MLLMs using natural instructions.

In prefix multimodal LLM, such as LLava [21], image embeddings, treated as soft prompts, are prepended to standard text embeddings, enabling the LLM backbone to process multimodal inputs. Unlike end-to-end training of multimodal LLMs from scratch, adapting pretrained checkpoints substantially reduces pretraining costs while facilitating efficient knowledge transfer to vision-language tasks. However, a significant challenge remains: high computational cost when processing images. Specifically, the number of visual tokens increases drastically with high-resolution images, exacerbated by the quadratic time complexity of the attention mechanism. Moreover, recent studies highlight that image resolution critically impacts visual performance, suggesting an inevitable increase in computational burden for future MLLMs [14, 20]. A common approach to address this issue is the use of bottleneck mechanisms to down-sample visual signals. While instruction-aware compression mitigates information loss, it often requires additional modules trained separately [9].

In this work, we aim to enhance the training and inference efficiency of multimodal LLMs by introducing an instruction-aware prefix without extensive instruction-image pretraining. We propose a novel architecture and training method that leverages publicly available datasets containing approximately 1 million samples. Our design enables fast zero-shot generation using a relatively small language model.

Unlike previous designs that primarily focus on visual properties, such as preserving semantic locality [4], our approach incorporates language-aware factors into the visual feature pooling process at minimal additional cost. This strategy filters relevant visual information before passing it to the language model. Our method is inspired by the findings of [9], which demonstrates that not all image information is essential for answering visual questions; therefore, selectively discarding certain visual inputs is a viable optimization. The key advantage of our approach lies in significantly reducing computational cost by limiting the size of visual inputs attended to by the language model while preserving sufficient information for accurate responses.

We introduce TA-LLaVA, a prefix multimodal LLM similar to the LLaVA family, which uses a visual prefix and achieves image-conditioned text generation. TA-LLaVA is trained on a relatively small vision-language dataset but distinguishes itself through instruction-aware (TA) visual feature pooling. Specifically, we reduce the number of prefix tokens by applying a modified cross-attention mechanism that alternates between extracting textual and visual features. Additionally, we employ a curriculum learning technique to gradually increase the difficulty of training tasks, enabling steady model improvement.

To validate our approach, we follow the standard zero-shot evaluation protocol and benchmark TA-LLaVA on a suite of public vision-language datasets unseen during training. Empirical results demonstrate that TA-LLaVA achieves strong performance relative to models requiring significantly more computational resources, memory, and time during generation.

2. Related Work

Multimodal Large Language Models. The dominant approach to constructing MLLMs integrates a visual encoder with a pretrained large language model (LLM). Since mainstream LLMs adopt the Transformer architecture, the CLIP-series models [26], which employ Vision Transformer (ViT) layers, are particularly well-suited for this integration. CLIP models represent image inputs as flattened sequences of patch tokens, naturally aligning with the Transformer structure. Additionally, their unsupervised learning paradigm demonstrates robust cross-domain generalization. Prior studies have shown that freezing the visual encoder during MLLM training is often sufficient to achieve competitive performance on downstream tasks. Various adapter modules have been explored to project the activations from the visual encoder into the LLM embedding space. VILA [19], Palm-E [11], and LLaVA [21] choose a simple linear layer or MLP, whereas Blip-2 [16] and Flamingo [2] use cross-attention-based module modified to learn better vision-language representation. Recent studies on pre-training MLLMs highlight that image resolution plays a critical role in downstream performance, often surpassing the impact of model size [14, 20, 23]. However, the performance gain is at the cost of inference speed. For instance, increasing the image resolution from CLIP-ViT-L/14@224 to CLIP-ViT-L/14@336 effectively doubles the number of visual tokens, requiring the base LLM to process significantly more tokens. Given the quadratic time complexity of self-attention with respect to the token count, several bottleneck mechanisms have been introduced to condense visual representations and control inference costs [14, 16]. However, these methods face a trade-off between efficiency and performance, as compressing visual signals may result in information loss. Determining an optimal bottleneck size requires extensive empirical experimentation. To address this challenge, we propose introducing textual signals into the visual projection module, guiding the extraction of relevant visual information. By aligning the visual feature pooling process with textual instructions, we aim to preserve only the most critical visual inputs for processing, balancing efficiency and effectiveness.

Visual Instruction Tuning. Motivated the success of InstructGPT [25], and FLAN [32] in improving the zero-shot generalization of LLMs through supervised fine-tuning of conversation data, where the model's response is conditioned on human instructions, visual instruction tuning extends this learning paradigm to MLLMs by composing image-centric fine-tuning datasets. While LLaVA [21] constructs such datasets by prompting language-only GPT-4V, PoliteFlamingo [5] trains a rewriter model to annotate public vision-language datasets with human-preferred responses. Some MLLMs, such as Kosmos-1 [15] and Gemini [27], are inherently built as multimodal models from scratch using in-house datasets. However, the more widely adopted approach is to adapt a pretrained language-only model into a multimodal one. This typically involves aligning the text and image input spaces through multimodal pre-training, followed by end-to-end visual instruction tuning. Notably, the LLaVA series [20] achieves strong zero-shot performance with impressive training efficiency, requiring only 1 million training samples. Another line of work focuses on improving the inferencing efficiency to allow prolonged visual inputs (e.g. video) or lower generation costs. InstructBLIP [9] and LLaMA-VID [18] inject instruction information into visual feature pooling to make the extraction process instruction-aware. Their empirical results demonstrate that this approach significantly reduces the required visual prefix tokens. However, both works involve an additional module and a separate training phase to align text instructions to visual information, and neither fails to reuse the text embeddings outputted by the base LLM. In contrast, our proposed method introduces a compact design that merges instruction-aware feature pooling directly into the base model's forward pass. Inspired by the architecture of Flamingo [2], our design eliminates the need for separate alignment phases, achieving improved efficiency without compromising performance. [2].

3. Method

In this section, we detail our proposed modeling and training approach. For completeness, we begin with a brief review of the prefix multimodal language modeling method, which is prominently used in LLaVA [21].

3.1. Preliminaries

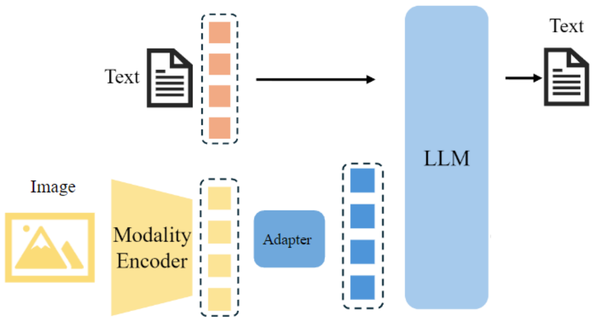

Figure 1. An overview of our multimodal LLM's architecture. The components of the model can be summarized into a vision encoder, a multimodal adapter, and a LLM backbone. Our work is designing a novel adapter that can freely attend to the image and the instruction. This model only allows single-image, in addition to text, inputs, and text-only outputs.

Prefix language modeling was initially introduced as an efficient fine-tuning method for causal language models (CLMs). The key idea is to condition the model on a set of continuous prefix tokens that influence its behavior without participating in next-token prediction or contributing to the loss computation during training.

LLaVA [21] extends this learning paradigm by using tokenized image inputs V as a prefix concatenated with the regular text tokens X. The language model is then trained to process the combined input and attend to the visual signal appropriately. For instruction tuning, the input text sequence X is typically divided into two parts: instruction and desired response $X = [X_{instruct}, X_{response}]$. Since the instruction is provided during inference, both visual inputs and the instruction are used as the prefix, and the model is trained only to generate the response. Mathematically, for a response of length T, the probability of the answer is computed as:

$$p(X_{resp}|V, X_{inst}) = \prod_{t=1}^{T} p_\theta(x_t|V, X_{inst}, X_{resp,0:t-1})$$

where $X_{resp,0:t}$ denote the previously generated response token up until time i.

During inference, the model generates the response tokens by repeatedly predicting a probability distribution over the language model's vocabulary. In our evaluation, we employ greedy decoding, where at each time step, the token with the highest probability is selected:

$$x_t = \arg\max_{x \in V} p_\theta(x|V, X_{inst}, X_{resp,0:t-1})$$

where V is the model's vocabulary. This simple yet effective strategy ensures deterministic generation and is widely adopted in language model evaluations.

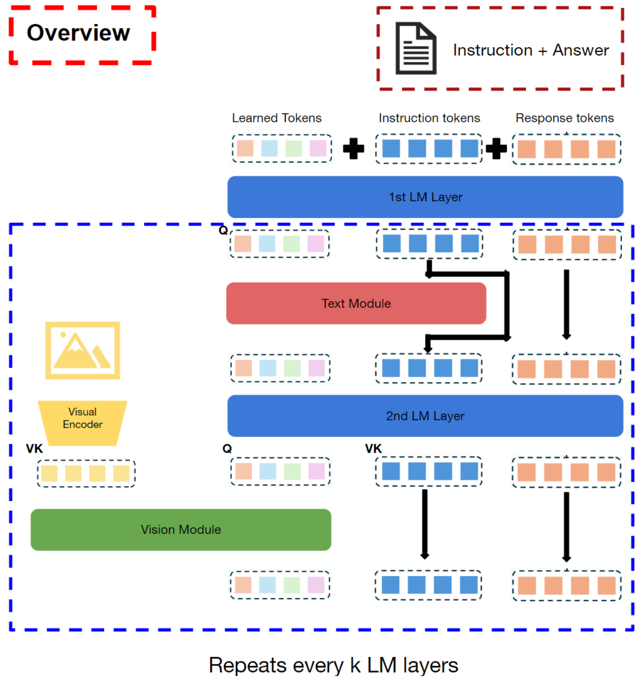

Figure 2. Detailed diagram about the proposed multimodal LLM. Learnable tokens are used as the memory by the adapter to hold relevant instruction and visual information. They are then concatenated to the text prompt for causal text generation. The adapter alternates between pooling image and instruction features.

3.2. Instruction-Aware Visual Feature

Algorithm 1 Our Proposed Adapter

```
procedure CONTEXTAWAREATTENTION(V, X, Y)
    [Y, X] ← [Y, X] + SelfAttention([Y, X])
    [Y, X] ← [Y, X] + FFN([Y, X])
    Initialize [X_instruct, X_response] ← X
    if current layer index % k == 0 then
        Y ← Y + XAttn_text(q = Y, v = X_instruct)
        Y ← Y + FFN(Y)
    else if current layer index % k == 1 then
        Y ← XAttn_vision(q = Y, v = V)
        Y ← Y + FFN(Y)
    end if
    return Y, X
end procedure
```

Our proposed model adheres to the established MLLM paradigm, where visual features are attached as prefix tokens to the text embeddings. However, instead of aligning the visual features directly to the input layer of the LLM, we allow the base LLM to interact with them at certain LM layers. To enable instruction-aware visual feature extraction, previous approaches often employ an independent cross-attention model, trained separately with a distinct objective and often on a different dataset. In contrast, we reuse the language model backbone as the text encoder, alternating between extracting information from the language tokens and the visual tokens. This is achieved using a novel adapter design based on the cross-attention mechanism. The learnable prefix tokens serve as a "memory," storing summarized relevant features from both the instruction and image inputs. To facilitate this process, we initialize two modality-specific adapters: one for the vision modality and one for the text modality. To guide visual feature extraction using instruction features, the instruction feature pooling step occurs before the image feature pooling. The adapters are inserted between the original layers of the language model, enabling them to update the learnable prefix tokens iteratively. The architecture design resembles that of Flamingo [2]. However, a crucial difference is that the text embeddings in the language model can only attend to the prefix tokens but not the visual tokens.

Adapter In practice, the visual encoder's hidden dimension often misaligns with that of the language model. Therefore, a projector, either an MLP or a linear layer, is used to make the visual embeddings compatible. Mathematically, let the output activation of the visual encoder upscaled by the projector to be $V \in \mathbb{R}^{M \times d_{model}}$ and the text embeddings be $X \in \mathbb{R}^{N \times d_{model}}$. Inspired by the iterative attention mechanism, we initialized a group of fixed-size learnable tokens $Y \in \mathbb{R}^{L \times d_{model}}$ as the prefix such that $L << M$. The prefix tokens are updated iteratively using cross-attention adapters $f_{xattn,text}$ and $f_{xatten,vision}$, which process text and visual information, respectively. We assume the input text prompts follow the structure $X = [X_{instruct}, X_{response}]$. To make the prefix instruction-aware, we first summarize the instruction part of the input prompt by updating the prefix Y as follows:

$$Y_{t+1} \leftarrow f_{xattn,text}(Y_t, X_{instruct})$$

where Y_t serves as the query, and X_t provides the key and value. Here, t corresponds to the layer index in the stacked transformer-based language model. Next, we incorporate the visual information into the prefix tokens by applying the vision adapter:

$$Y_{t+2} \leftarrow f_{xattn,vision}(Y_{t+1}, V)$$

where V represents the precomputed visual tokens stored to save computation. In each LM layer, the prefix Y occupies the beginning of the input sequence to the LLM so that all following tokens can freely attend to Y as in the regular causal language model. We insert the adapters for every k LM layer, and feature pooling repeats until the final layer of the language model. Again, the prefix Y and the instruction $X_{instruct}$ do not participate in the text generation and loss calculation. A linear layer, known as the LM head, finally predicts a probability distribution over all possible words for each token in the output sequence.

Conditional Auto-regressive Generation The KV cache is a widely used and efficient technique for LLM inference. Since a causal language model generates one token at a time, generating a complete response may require hundreds of forward passes. Without optimization, the key-value (KV) pairs for past tokens would be recomputed repeatedly at each generation step. The KV cache improves efficiency by storing the precomputed keys and values from previous steps, enabling the newly generated tokens (queries) to attend directly to these cached keys and values. *Restricting the text-aware feature pooling to user instruction is detrimental to efficiency because it avoids repeated updating of the visual prefix and allows easy causal generation readily compatible with most LLM serving frameworks.* During inference, only the user instruction is available as input, and the model iteratively predicts the response conditioned on the image and the instruction. To ensure that the instruction-only pooling is maintained, we use an attention mask to prevent the prefix tokens Y from attending to the response tokens $x_{response}$. During response generation, the visual prefix Y remains fixed and is not updated with newly generated tokens. Consequently, the prefix tokens Y along with the instruction embeddings can be cached once using standard KV cache implementations. This design seamlessly integrates with existing causal generation techniques, ensuring efficient auto-regressive decoding without requiring modifications to standard LLM inference pipelines.

4. Experiment

Our design of the adapter may be flexibly applied to most causal LM. However, due to limited time and computing resources, we mainly experiment with the approach on one small language model and compared it against other MLLMs for visual instruction tuning. We adopt the zero-shot evaluation protocol described in [21] and report scores on standard vision-language benchmarks.

4.1. Implementation

Dataset. This work focuses mainly on validating the efficacy of our proposed design. Therefore, we utilize the publicly released datasets from LLava-1.5 [20], which consists of a pre-training and a fine-tuning dataset. The former, LLaVA-CC3M-Pretrain-558k, contains sampled instances from a large-scale dataset, LAION [29], annotated by BLIP-2 [16]. Although not intended for instruction tuning, the dataset complies with the instruction-following format with a user request and a model response. The sequences feature straightforward and concise descriptions of

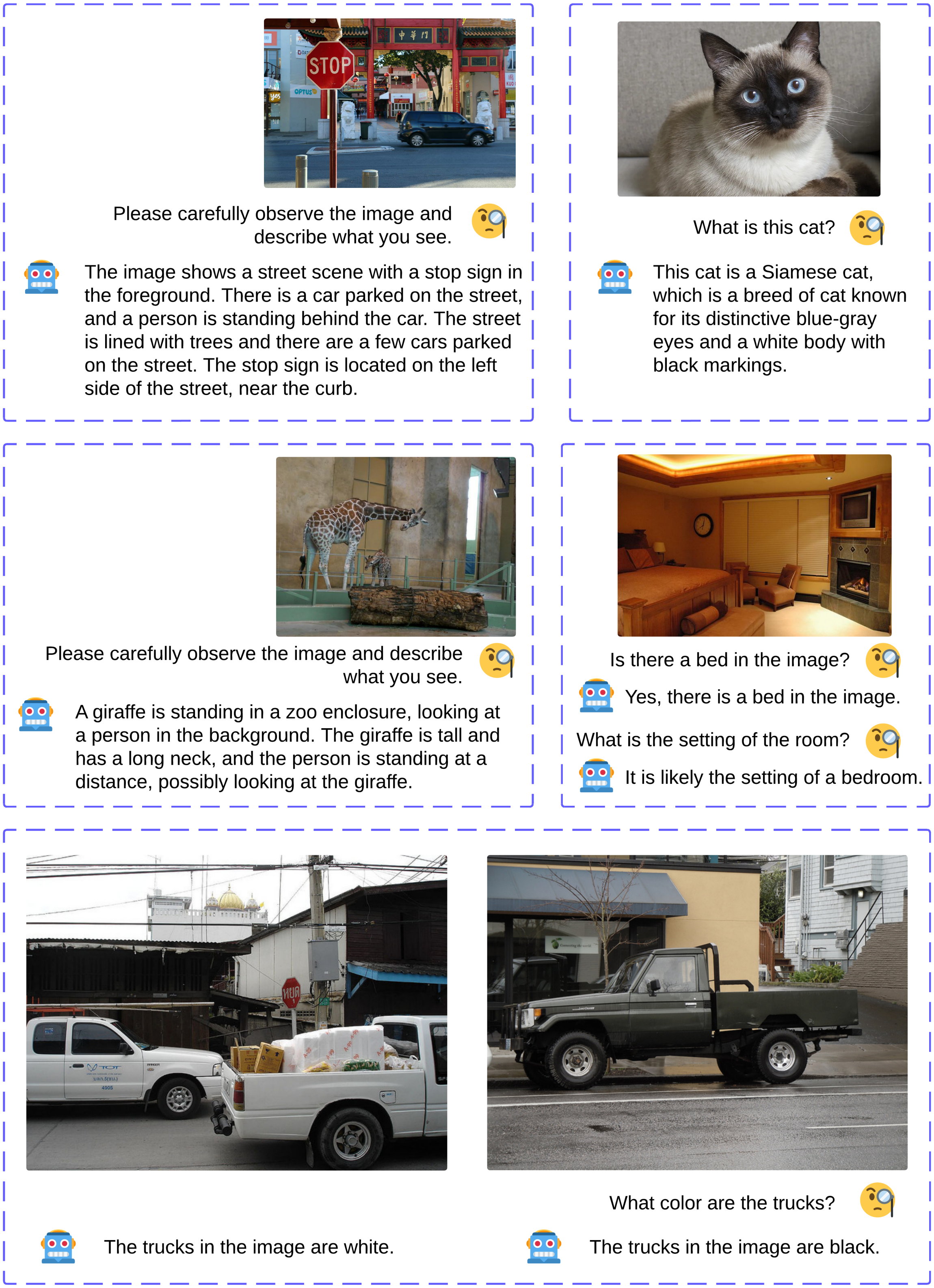

Figure 3. A few qualitative examples generated by our TA-LLaVA model. The model is instructed to answer vision-language questions about the scene or to describe the scene in detail. The model demonstrates strong instruction-following abilities but suffers from hallucination.

the image. The other dataset, LLaVA-v1.5-mix665k, contains training samples from three sources: GPT-generated LLaVA-instruct-150k [21], public vision-language datasets, and ShareGPT [6]. It should be noted that the ShareGPT dataset is text-only, while others have both images and text paragraphs. Although past studies reveal that stronger language ability can positively benefit vision-language performance [8, 19, 20], we choose to remove the language-only instances as their verbosity causes out-of-memory issues during training. For the same practical reason, we discard samples that are longer than 1024 tokens to make training feasible. The filtered instruction-tuning dataset contains roughly 90% of the training instances. No modification is done to the pre-training dataset. For all training instances, we adopt a consistent, prompt template to structure the input: "user:<question> model: <response>." In particular, the token corresponding to the word "model" serves as a special token that specifies the end of the instruction. Though theoretically possible, our model is not trained on multi-image or video datasets. We leave support for multi-image inputs as a feature work.

Model. In this study, we build our proposed MLLM from a small yet performant language model, Gemma-2-2B [28], released by Google research. Specifically, we use the version that has been instruction-tuned with reinforcement learning from human feedback [24]. The model is trained with the standard causal language modeling objective and contains 24 transformer layers. For the visual encoder, we use the CLIP-vit-l/14@336 [26] to encode images of resolution 336^2 into 576 visual tokens by extracting the activations from the last-second transformer layer. A linear projector layer is added after the visual encoder to upscale the visual tokens to 2048 dimensions, thus making them compatible with the language model. 32 learnable prefix tokens are initialized and subsequently trained to hold instruction and image information. Most importantly, our cross-attention implementation is adopted from the improved self-attention module in Gemma-2 [28]. Novel techniques, including attention soft capping [28], grouped query attention [1], and rotary positional embedding [31], are applied to stabilize training and increase expressivity. In every forward pass, we first look for the special token "model" to identify the end of the user instruction and create an attention mask accordingly. The response part of the text prompt is masked out to ensure that only instruction information is used in visual feature extraction. One cross-attention adapter is initialized for each modality and inserted after every four LM layers. Although the adapters are used repeatedly, the same adapters of each modality instead of multiple adapters of different weights are used to reduce the number of parameters.

Training. We propose a three-phase curriculum learning schedule to progressively expose the model to tasks of increasing difficulty. In phase one, we focus on training the model to learn an alignment between the visual and the textual modality. With the vision encoder and the LLM frozen, we pre-train the adapter and the projector on LLaVA-Pretrain-558k. Captioning is relatively easy as the semantic relationship between the image and the response is straightforward, and instruction-aware feature extraction is not important since an understanding of the global context is required. In phase two, we train the model to utilize complex instruction information to extract only useful visual features. The dataset used is LLaVA-v1.5-mix665k, which contains complex tasks such as question answering, reasoning, and conversation. Again, both the vision encoder and the LLM remain frozen. This stage is considered more challenging as high-level skills are required to solve the mentioned tasks, and the model is forced to summarize only the useful information in the restricted 32 prefix tokens. Lastly, phase three focuses on fine-tuning the language model to fully adapt it to vision-language tasks. Therefore, only the vision encoder is frozen, but the rest of the model is fine-tuned on the same dataset as in phase two, namely LLaVA-v1.5-mix665k. In three training stages, we steadily train the model to gain more and more complex skills that are useful in tackling downstream vision-language tasks.

Hyperparameters. For all training sessions, we use the Adam optimizer without weight decay and $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To stabilize training, we apply linear warm-up, gradually increasing the learning rate during the initial steps. The specific hyperparameters are provided in Table 1. All models are trained utilizing 8 Nvidia A6000 GPUs and completed within 1 day. Techniques, including gradient checkpointing and gradient accumulation, are adopted to improve memory efficiency. Further acceleration is possible with FlashAttention [10]. However, we fail to implement it due to an incompatibility issue with the cross-attention mechanism associated with the Transformers package.

4.2. Quantitative Results

We empirically evaluate the proposed architecture on five standard benchmarks: POPE [17], VQAv2 [13], MME [12], Science QA [22], and 2017 COCO Caption [7]. The task types span image question answering, reasoning, OCR, captioning, and domain knowledge testing. We compare our method against other architectures, including LLaVA [21], LLaVA-1.5 [20], InstructBLIP [9], and Qwen-VL [3]. We follow the zero-shot evaluation protocol: during inference, no demonstration examples are provided, but the model receives task-specific instructions. The model generates responses by greedily decoding the next token based on the highest probability.

We summarize the performance and the size of the training dataset in Table 2. Despite the small sizes of the base LLM and the training dataset, our final model, TA-LLaVA,

Phase	Vsual Encoder	Adapter	LLM	LR	Warm-Up	Batch size	# Epoch
Phase 1	Frozen	Trained	Frozen	1×10^{-3}	0.03%	256	1
Phase 2	Frozen	Trained	Frozen	5×10^{-5}	0.03%	128	1
Phase 3	Frozen	Trained	Trained	2×10^{-5}	0.03%	128	1

Table 1. Hyperparameters for each training phase.

Method Metrics	Sample Size	Base LLM	POPE F1	VQAv2 Acc	MME Acc	SciQA Acc	COCO-Cap CiDER
LLaVA-7B	0.71M	LLaVA-7B	-	76.3	809.6	-	-
LLaVA-1.5-7B	1.22M	Vicuna-1.5-7B	87.3	78.5	1510.7	67.1	-
InstructBLIP	130M	Vicuna-1.5-13B	87.7	-	1212.8	63.1	-
Qwen-VL-Chat	1.4B	Qwen-7B	-	78.2	1487.5	68.2	-
TA-LLaVA-Phase2	1.15M	Gemma-2-2b-it	47.2	50.3	870.6	46.3	69.1
TA-LLaVA	1.15M	Gemma-2-2b-it	78.9	60.5	1251.8	63.1	79.4

Table 2. TA-LLaVA's zero-shot performance on unseen vision-language benchmarks compared with the SoTA models. Our model attains strong performance comparable to InstructBLIP on complex tasks while using significantly less data. The scores for the other models are reported by [20].

attains a strong performance compared with the other SoTA MLLMs. Particularly, it scores 1251.8 and 63.1 on MME and Science QA, outperforming InstructBLIP. However, we acknowledge a significant gap between TA-LLaVA and the SoTA methods, especially in POPE and VQAv2, where LLaVA-1.5-7B beats our method by 8.4 and 18 points. This gap suggests that TA-LLaVA may still lag behind in fundamental vision capabilities.

To investigate the effect of training further, we also evaluate the intermediate model after Phase Two training and compare it against the final model. We notice that the final model significantly improves the scores by training on the same data but with LLM unfrozen. Therefore, it seems full fine-tuning is detrimental to both modality alignment and knowledge transfer as both scores on simpler (POPE and VQAv2) and more complex (MME, Science QA, and COCO Caption) increase by a large margin. However, the most fair comparison may be between InstructBLIP and TA-LLaVA-Phase2, as both models have instruction-aware feature extraction, and the base LLM is not fine-tuned explicitly. The observation that InstructBLIP has much better performance suggests that there is still a lot of potential for tuning the adapter, which may be undertrained. Yet, a confounding factor remains: the InstructBLIP has a much (6x) bigger LLM backbone, which may partially account for the performance boost. These findings suggest that a large-scale ablation study compares the methods in a fairer setting, where all methods should use the same base LLM.

Thanks to the small size and the efficient prefix design, our MLLM has a significantly lower inference cost. Following the inference cost analysis as in [30], assuming a sequence of 40 text tokens and one input image, one forward call to the LLaVA-1.5-7B model has an estimated computation cost of 9.3 TeraFLOPS, while our TA-LLaVA only requires 3.56 TeraFLOPS. Our method effectively cuts the inference cost by more than 50%.

4.3. Qualitative Results

In addition to the standard benchmarks, we also qualitatively examine the outputs of our model. We present a few inference samples in Figure 3. As illustrated, our model demonstrates decent capabilities in solving a wide range of vision-language questions. It is able to understand the user's instructions and recognize objects of interest as requested. Furthermore, the model can supply additional details by accessing its internal knowledge about the world. However, a notable problem with TA-LLaVA is hallucination. In the giraffe example, the model points out that there is a person in the background looking at the giraffe, but it is apparent that no one is present in the scene other than the two giraffes. Additionally, in the stop sign example, the model claims there are "a few cars" while there is only one car. These observations suggest that the visual prefix fails to keep all details in the image, although the model captures the global context.

5. Conclusion

In this project, we present TA-LLaVA for instruction-tuned multimodal LLM, a scalable and efficient model to solve general vision-language tasks. The key contribution is that our novel adapter design 1) effectively reduces the number of visual prefix tokens from 576 to 32, and

2) condition visual feature extraction on the provided instruction. Compared to LLaVA-1.5, TA-LLaVA reduces inference costs by more than 50% while maintaining strong performance on complex tasks. Remarkably, our model achieves performance on par with InstructBLIP, which is trained on datasets 100 times larger. Qualitative evaluations further demonstrate that TA-LLaVA possesses strong instruction-following abilities, comprehensive scene understanding, and broad world knowledge.

6. Limitation

However, a notable limitation of our model is hallucination, where TA-LLaVA struggles to accurately recognize elements within an image. This issue is particularly evident in the POPE benchmark, where the model exhibits significantly lower accuracy. Additionally, when tasked with image descriptions, which require both holistic and precise perception, the model is prone to generating erroneous answers. We hypothesize that the limited number of prefix tokens (32) may cause information loss, especially in tasks demanding richer visual details.

Furthermore, while TA-LLaVA demonstrates strong performance, it still lags behind state-of-the-art models on standard benchmarks. Addressing this gap will require further architectural and training improvements. In future work, we plan to extend TA-LLaVA in three key directions. 1) We aim to implement the adapter design in other causal LLMs, such as Qwen [3] and Vicuna [33]. This will enable a more fair comparison against existing methods and validate the adapter's compatibility with mainstream LLM backbones. 2) An exciting extension involves enabling the model to process multi-image or even video inputs by concatenating sequences of visual embeddings. This requires the collection of dedicated multi-image datasets for both pre-training and fine-tuning stages. 3) To further improve scalability and speed, we plan to integrate advanced techniques such as FlashAttention [10], which can optimize the attention mechanism for better memory and computational efficiency.

7. Statement of Individual Contribution

7.1. Jianhong Tu

Jianhong Tu is primarily responsible for designing the architecture and implementing the code. He developed and deployed a training framework onto computation nodes for large-scale training. He also contributed to the method and related work section of the report and the presentation, thanks to his familiarity with the field. Lastly, he plans the empirical experimentation and specifies the procedure for quantitative evaluation.

7.2. Erdong Chen

Erdong is responsible for both quantitative and qualitative evaluation. He contributed by preparing a codebase for automatic evaluation on five vision-language benchmarks. He also manually tests the model's performance using many examples. Erdong also assisted in the model architecture and training, as well as literature reviews and presentations.

7.3. Shuhan Zhang

Shuhan enhanced the dataset by organizing it into subsections, generating example prompts, and performing usage analysis. Shuhan also assisted in performing evaluations on the final model and creating slides for demonstration.

8. External Resources Used

The base LLM that we use as the foundation for our multimodal LLM is accessed through the HuggingFace platform at `https://huggingface.co/google/gemma-2-2b-it`. The final model is majorly implemented using PyTorch `https://pytorch.org/`, and both training and inferencing functionality rely on API offered by the Transformers package `https://github.com/huggingface/transformers`. For efficient modeling training, we use model sharding with the DeepSpeed framework `https://github.com/microsoft/DeepSpeed`. Finally, empirical evaluation is performed on the LMMs-Eval platform `https://github.com/EvolvingLMMs-Lab/lmms-eval`.

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4895–4901. Association for Computational Linguistics, 2023. 6

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New*

Orleans, LA, USA, November 28 - December 9, 2022, 2022. 1, 2, 4

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 6, 8

[4] Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal LLM. *CoRR*, abs/2312.06742, 2023. 1

[5] Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17745–17753. AAAI Press, 2024. 2

[6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. *CoRR*, abs/2406.04325, 2024. 6

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 6

[8] Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NVLM: open frontier-class multimodal llms. *CoRR*, abs/2409.11402, 2024. 1, 6

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 2, 6

[10] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 6, 8

[11] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 8469–8488. PMLR, 2023. 2

[12] 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, 2024. 6

[13] 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. *CoRR*, abs/1612.00837, 2016. 6

[14] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: An LMM perceiving any aspect ratio and high-resolution images. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIII*, volume 15141 of *Lecture Notes in Computer Science*, pages 390–406. Springer, 2024. 1, 2

[15] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023. 2, 4

[17] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. 6

[18] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVI*, volume 15104 of *Lecture Notes in Computer Science*, pages 323–340. Springer, 2024. 2

[19] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: on pre-training for visual language models. *CoRR*, abs/2312.07533, 2023. 1, 2, 6

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744, 2023. 1, 2, 4, 6, 7

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1, 2, 3, 4, 6

[22] 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. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 6

[23] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. MM1: methods, analysis and insights from multimodal LLM pre-training. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXIX*, volume 15087 of *Lecture Notes in Computer Science*, pages 304–323. Springer, 2024. 1, 2

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 1, 6

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[27] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530, 2024. 2

[28] Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Mil-

lican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024. 6

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 4

[30] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *CoRR*, abs/2403.15388, 2024. 7

[31] Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 6

[32] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 1, 2

[33] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 8