LIMA: Less Is More for Alignment

Chunting Zhou[μ*] Pengfei Liu[π*] Puxin Xu[μ] Srini Iyer[μ] Jiao Sun[λ]
Yuning Mao[μ] Xuezhe Ma[λ] Avia Efrat[τ] Ping Yu[μ] Lili Yu[μ] Susan Zhang[μ]
Gargi Ghosh[μ] Mike Lewis[μ] Luke Zettlemoyer[μ] Omer Levy[μ]

[μ] Meta AI
[π] Carnegie Mellon University
[λ] University of Southern California
[τ] Tel Aviv University

LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention

Renrui Zhang[*1,2], Jiaming Han[*1], Chris Liu[*1], Peng Gao[*†‡1], Aojun Zhou[2]
Xiangfei Hu[1], Shilin Yan[1], Lu Pan[3], Hongsheng Li[†2], Yu Qiao[†1]

[1]Shanghai Artificial Intelligence Laboratory [2]CUHK MMLab
[3]University of California, Los Angeles
{zhangrenrui, hanjiaming, gaopeng, qiaoyu}@pjlab.org.cn

LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model

Peng Gao[*‡†1], Jiaming Han[*1], Renrui Zhang[*1,2], Ziyi Lin[*2], Shijie Geng[3], Aojun Zhou[2]
Wei Zhang[1], Pan Lu, Conghui He[1], Xiangyu Yue[2], Hongsheng Li[†2], Yu Qiao[†1]

[1]Shanghai Artificial Intelligence Laboratory [2]CUHK MMLab
[3]Rutgers University
{gaopeng, hanjiaming, zhangrenrui, qiaoyu}@pjlab.org.cn

2023.08.24
ghdxheh123@naver.com
홍성태

LIMA: Less Is More for Alignment

Chunting Zhou[μ*] **Pengfei Liu**[π*] **Puxin Xu**[μ] **Srini Iyer**[μ] **Jiao Sun**[λ]

Yuning Mao[μ] **Xuezhe Ma**[λ] **Avia Efrat**[τ] **Ping Yu**[μ] **Lili Yu**[μ] **Susan Zhang**[μ]

Gargi Ghosh[μ] **Mike Lewis**[μ] **Luke Zettlemoyer**[μ] **Omer Levy**[μ]

[μ] Meta AI
[π] Carnegie Mellon University
[λ] University of Southern California
[τ] Tel Aviv University

Introduction

- Existing alignment methods require significant amounts of compute and specialized data to achieve ChatGPT-level performance

- Given a strong pre-trained language model, remarkably strong performance can be achieved by simply fine-tuning on **1,000 carefully curated training examples**.

- We hypothesize that alignment can be a simple process where the model **learns the style or format** for interacting with users, to expose the knowledge and capabilities that were already acquired during pretraining.

Alignment Data

- Community Questions & Answers

Source	#Examples	Avg Input Len.	Avg Output Len.
Training			
Stack Exchange (STEM)	200	117	523
Stack Exchange (Other)	200	119	530
wikiHow	200	12	1,811
Pushshift r/WritingPrompts	150	34	274
Natural Instructions	50	236	92
Paper Authors (Group A)	200	40	334
Dev			
Paper Authors (Group A)	50	36	N/A
Test			
Pushshift r/AskReddit	70	30	N/A
Paper Authors (Group B)	230	31	N/A

Table 1: Sources of training prompts (inputs) and responses (outputs), and test prompts. The total amount of training data is roughly 750,000 tokens, split over exactly 1,000 sequences.

- Superficial Alignment Hypothesis

  A model's knowledge and capabilities are learnt almost entirely during pretraining, while alignment teaches it which subdistribution of formats should be used when interacting with users

  ➔ Superficial Alignment Hypothesis is that one could sufficiently tune a pretrained language model with a rather small set of examples

- 1,000 prompts and responses, where the outputs (responses) are stylistically aligned with each other, but the inputs (prompts) are diverse.

- We also collect a test set of 300 prompts and a development set of 50.

- Three community Q&A websites: Stack Exchange, wikiHow, and the Pushshift Reddit Dataset

Alignment Data
- Community Questions & Answers

- Stack Exchange
  - 75 STEM exchanges (including programming, math, physics, etc.) and 99 other (English, cooking, travel, and more)
  - highest score that are self-contained in the title (no body). then select the top answer for each question
  - Filter answers that are too short (less than 1200 char), too long (more than 4096 char), written in the first person, or reference other answers
  - Remove links, images, and other HTML tags from the response, retaining only code blocks and lists

- wikiHow
  - Variety and high - quality
  - Title as prompt, Article's body as the response
  - preprocessing heuristics to prune links, images, and certain sections of the text.

- Reddit
  - Manually select examples from within the most upvoted posts in each r/AskReddit and r/WritingPrompts,
  - r/AskReddit, use for the test set
  - r/Writing, use for Training set

Alignment Data

- Manually Authored Examples

- To further diversify our data
    1. Group A and Group B, to create 250 prompts each, inspired by their own interests or those of their friends
    2. 200 for training and 50 for development set from Group A
    3. 250 data from Group B filtered to 230 and used it test set

- Also include 13 training prompts with some degree of toxicity or malevolence. (30 prompts in test set)
    ➔ carefully write responses that partially or fully reject the command, and explain why the assistant will not comply.

- Sample 50 training examples from Super-Natural Instructions such as summarization, paraphrasing, and style transfer
    - Edit some of the examples to conform with the style of our 200 manual examples.

Experiments

- Human Eval & GPT-4

- **LIMA: LLaMa 65B, fine-tune on our 1,000-example alignment training set**

- **Baseline**
  - Alpaca 65B (LLaMa 65B, tuned 52,000 dataset)
  - DaVinci003 (RLHF)
  - Bard (Based on PaLM)
  - Claude (52B parameter model trained with RL)
  - GPT-4 (RLHF)

  Methodology : with a single prompt and two possible responses, Asking annotators to label which response they preferred

- **Results**
  - Despite training on 52 times more data, Alpaca 65B tends to produce less preferable outputs than LIMA, Same is true for Davinci003
  - There is a non-trivial amount of cases where LIMA does actually produce better responses, compared to GPT-4 and Claude
  - Perhaps ironically, even GPT-4 prefers LIMA outputs over its own 19% of the time.

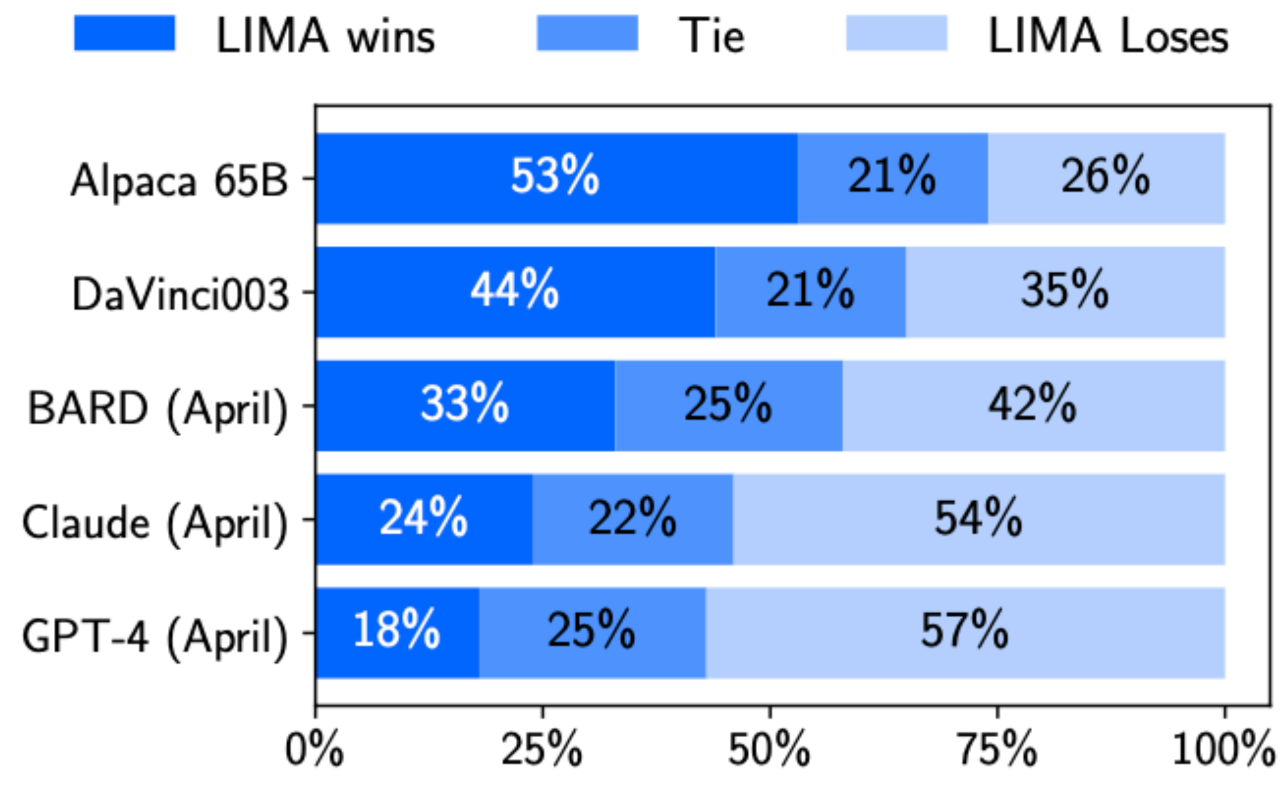


Figure 1: Human preference evaluation, comparing LIMA to 5 different baselines across 300 test prompts.

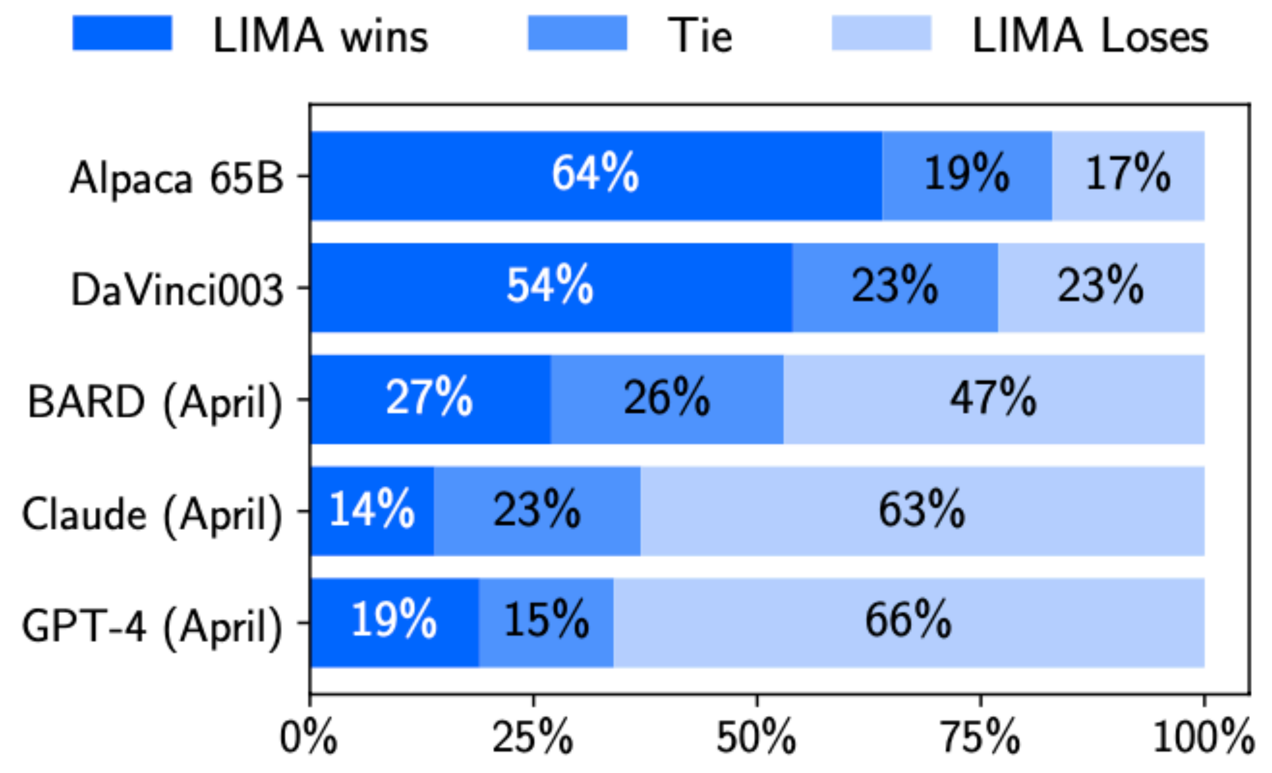


Figure 2: Preference evaluation using GPT-4 as the annotator, given the same instructions provided to humans.

Experiments

- Analysis

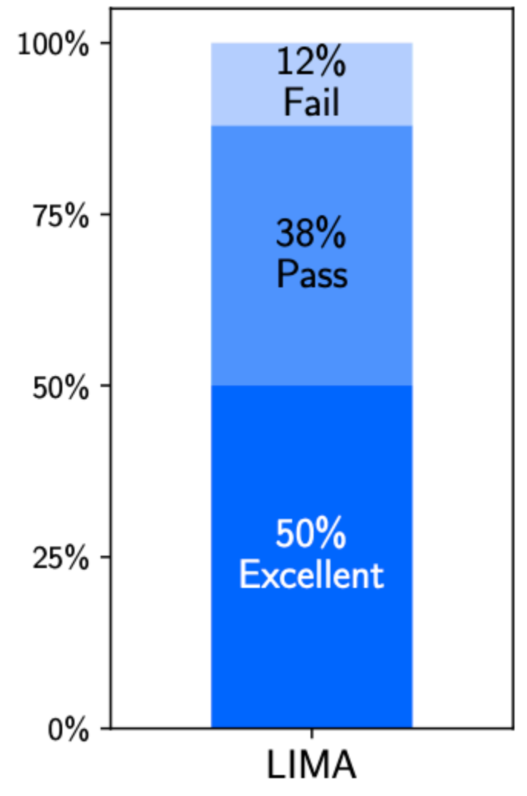


Figure 3: Analysis of LIMA over 50 test prompts.

- Result
  - Some of these baselines are actually highly-tuned products that may have been exposed to millions of real user prompts during training
    ➔ We thus provide an *absolute* assessment by manually analyzing 50 random examples
  - We label each example into one of three categories: **Fail**, **Pass**, **Excellent**

- Out of Distribution
  - 43 have a training example that is somewhat related in terms of format (e.g. question answering, advice, letter writing, etc)
  - Analyze 13 additional out-of-distribution examples (20 in total), 20% of responses fail, 35% pass, and 45% are excellent

  ➔ LIMA achieves similar absolute performance statistics outside of its training distribution, suggesting that it is able to generalize well.

- Safety
  - LIMA's response to 30 potentially sensitive prompts from the test set, and find that LIMA responds safely to 80% of them
  - when the malicious intent is implicit, LIMA is more likely to provide unsafe responses

Experiments

- Why is Less More? Ablations on Data Diversity, Quality, and Quantity

•Investigate the effects of training data diversity, quality, and quantity through ablation experiments

Setup: sample 5 responses, evaluate by ChatGPT, 1-6 likert scale(helpfulness)

Diversity

: To test the effects of prompt diversity

- 2,000 training examples from each source
- More diverse Stack Exchange data yields significantly higher performance

Quality

: To test the effects of response quality

- 2,000 examples from Stack Exchange *without* any quality or stylistic filters
- 0.5 point difference between models

Quantity

- Exponentially increasing amounts of data, sampled from (quality-filtered) Stack Exchange
- Scaling laws of alignment are not necessarily subject to *quantity* alone

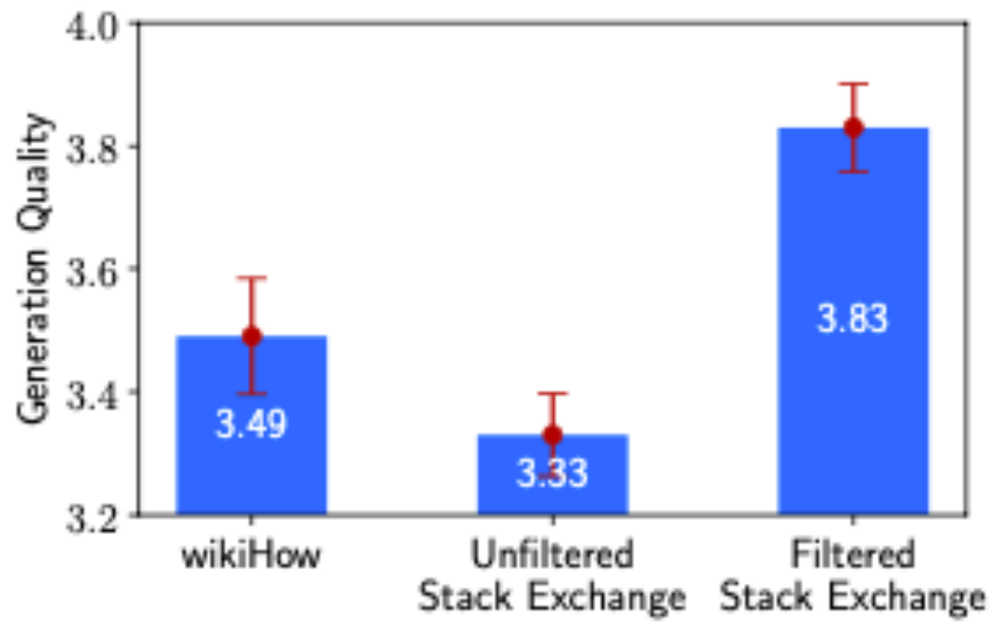


Figure 5: Performance of 7B models trained with 2,000 examples from different sources. **Filtered Stack Exchange** contains diverse prompts and high quality responses; **Unfiltered Stack Exchange** is diverse, but does not have any quality filters; **wikiHow** has high quality responses, but all of its prompts are "how to" questions.

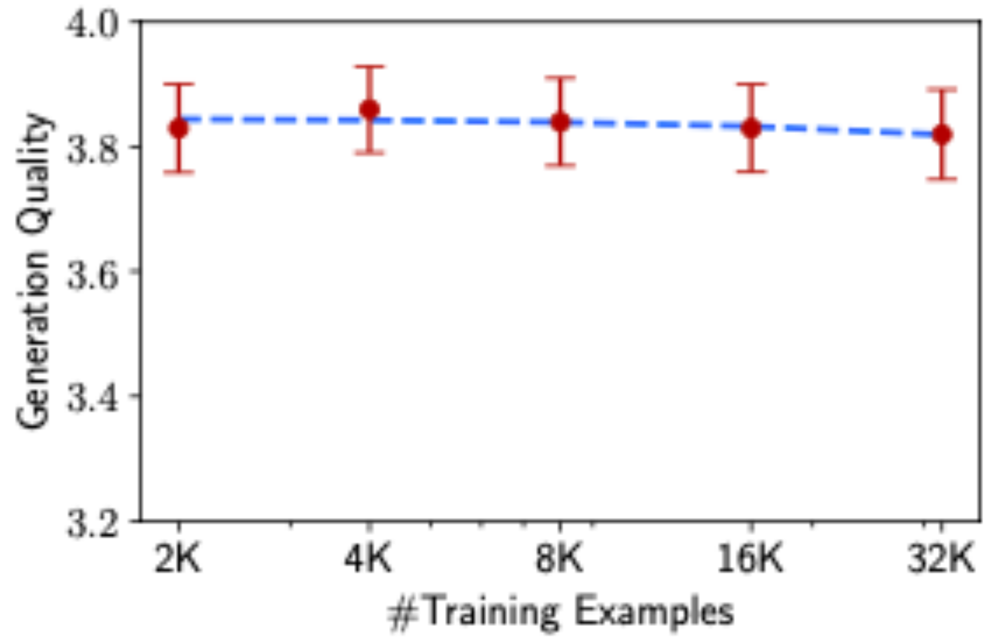


Figure 6: Performance of 7B models trained with exponentially increasing amounts of data, sampled from (quality-filtered) Stack Exchange. Despite an up to 16-fold increase in data size, performance as measured by ChatGPT plateaus.

Multi-turn Dialogue

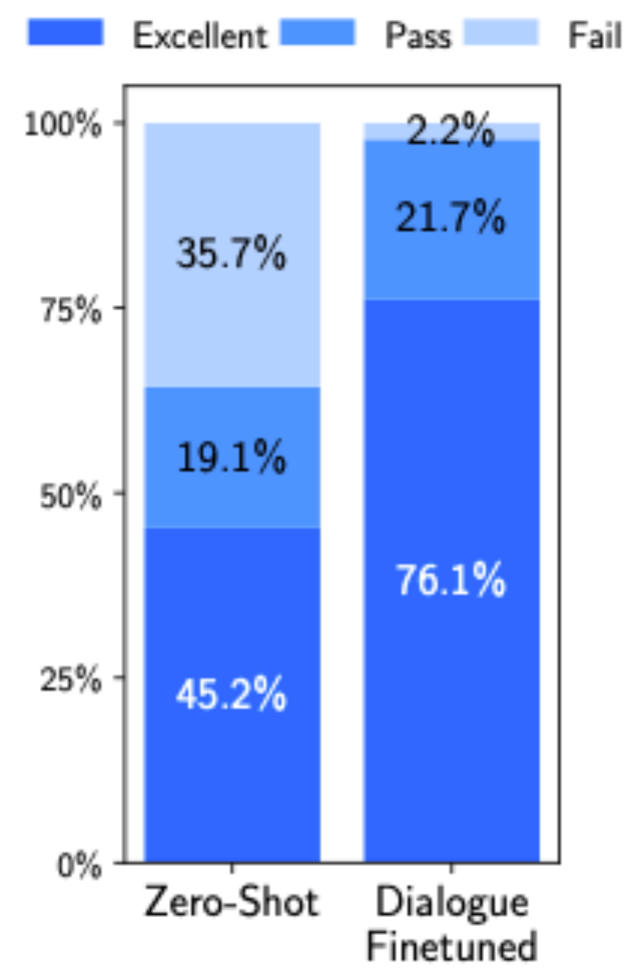


Figure 7: Analysis of dialogue turns, averaged over 10 test chats.

:Test LIMA across 10 live conversations (Excellent/Pass/Fail)

:Response Coherent for a zero-shot chatbot, referencing information from previous steps in the dialogue.

:In 6 out of 10 conversations, LIMA fails to follow the prompt within 3 interactions.

•**To improve its ability to converse, we gather 30 multi-turn dialogue chains.**

- Finetune with 1,030 examples ➔ Test with previous set
- Adding conversations substantially improves generation quality, raising the proportion of excel- lent responses from 45.2% to 76.1%.
- significantly better in 7 out of 10 conversations, and tied with the zero-shot model in 3.

This leap in capability from a mere 30 examples reinforces the hypothesis that such capabilities are learned during pretraining, and can be invoked through limited supervision

Discussion

- **1,000 carefully curated examples can produce remarkable, competitive results on a wide range of prompts**

- **LIMA is not as robust as product-grade models**
    ➔ while LIMA typically generates good responses, an unlucky sample during decoding or an adversarial prompt can often lead to a weak response

- **This work demonstrates the potential of tackling the complex issues of alignment with a simple approach.**

LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention

Renrui Zhang[*1,2], **Jiaming Han**[*1], **Chris Liu**[*1], **Peng Gao**[*†‡1], **Aojun Zhou**[2]
Xiangfei Hu[1], **Shilin Yan**[1], **Lu Pan**[3], **Hongsheng Li**[†2], **Yu Qiao**[†1]

[1]Shanghai Artificial Intelligence Laboratory [2]CUHK MMLab
[3]University of California, Los Angeles
{zhangrenrui, hanjiaming, gaopeng, qiaoyu}@pjlab.org.cn

Introduction

- Massive corpuse and advanced hardware의 영향으로 LLM은 놀라운 이해와 생성 능력을 보여주며 언어 작업을 더 높은 수준으로 이끌어냄

- 최근 instruction-following models은 상당한 성능을 보임. (e.g ChatGPT 와 GPT-3.5 (text-davinci-003))

- Prevalence of instruction models is largely impeded by the closed-source restriction and high development costs

- Complete fine-tuning of LLaMA is still time-consuming, computation-intensive, multi-modality unsupported

- LLaMA-Adapter, an efficient fine-tuning method that adapts LLaMA into a well-performed instruction-following model

Introduction

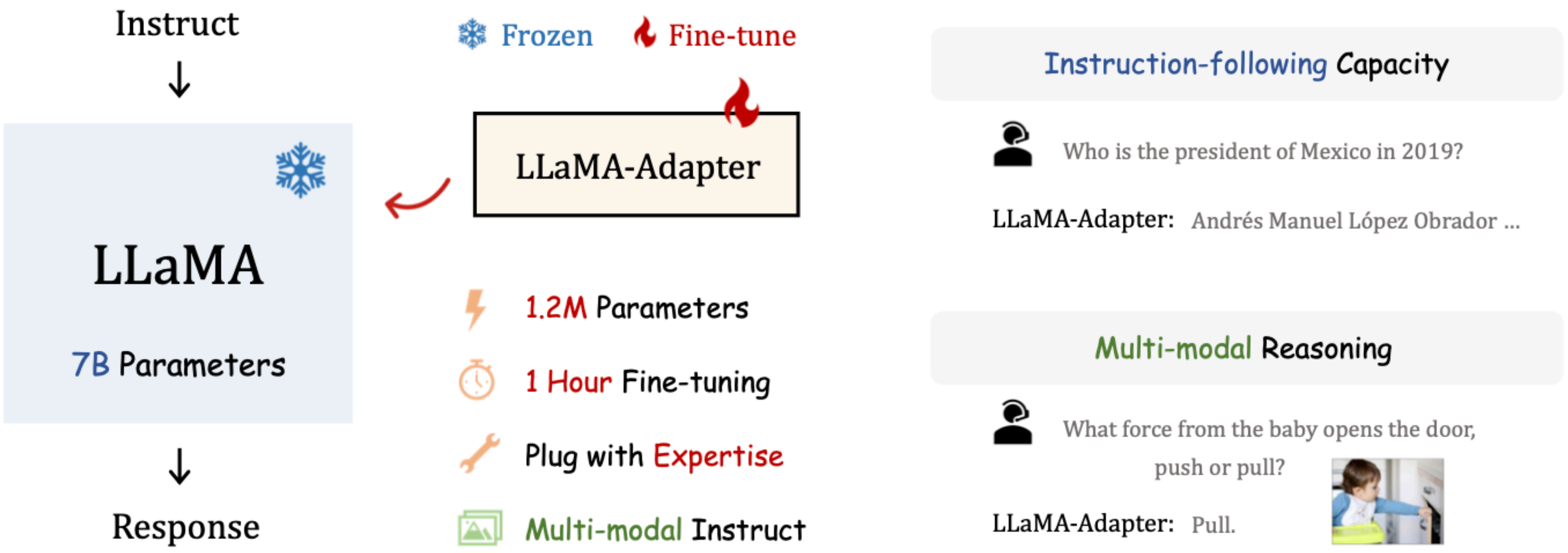


Figure 1: **Characteristics of LLaMA-Adapter.** Our lightweight adaption method efficiently fine-tunes LLaMA [61] 7B model with only 1.2M learnable parameters within one hour. After training, LLaMA-Adapter exhibits superior instruction-following and multi-modal reasoning capacity.

- **1.2M Parameters** : Freeze the pre-trained LLaMA and only learn the adaption prompts with 1.2M parameters on top.
- **One-hour Fine-tuning**: the training convergence of LLaMA-Adapter costs less than one hour on 8 A100 GPUs, which are three times faster than Alpaca.
- **Plug with Expertise**: For different scenarios, it is flexible to insert their respective adapters and endow LLaMA with different expert knowledge.
- **Multi-modal Instruction**: Besides textual instruction, our approach can also take images as input for multi-modal reasoning.

Introduction

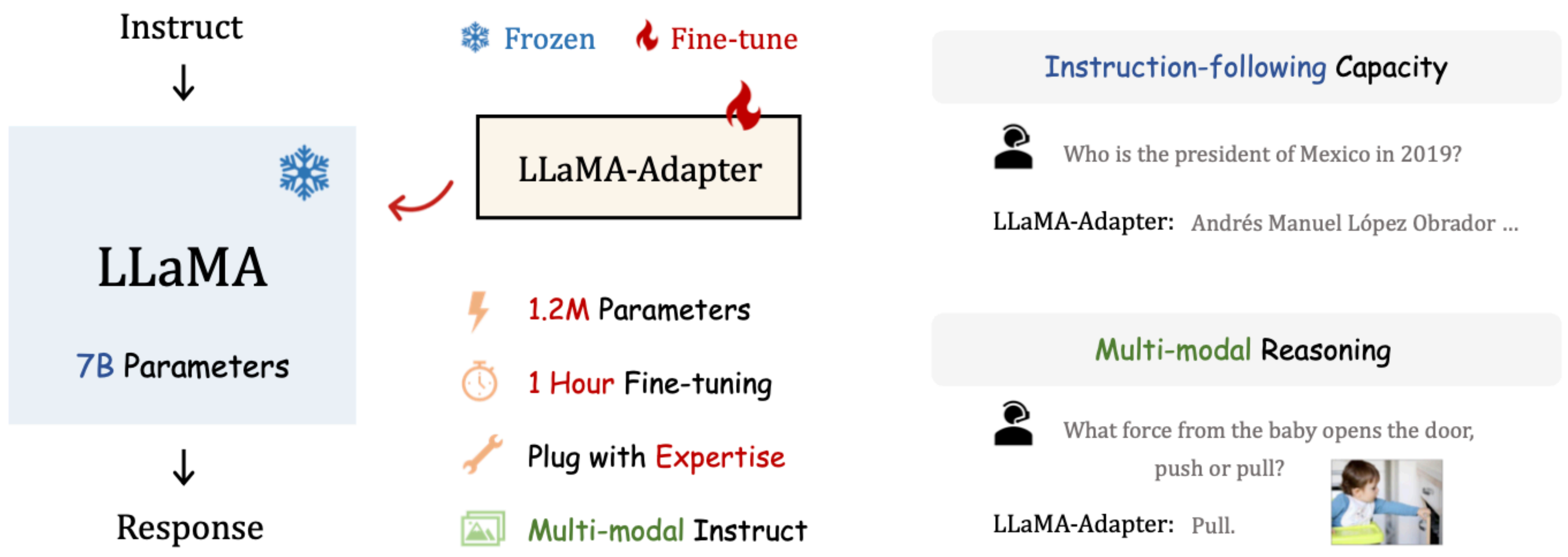


Figure 1: **Characteristics of LLaMA-Adapter.** Our lightweight adaption method efficiently fine-tunes LLaMA [61] 7B model with only 1.2M learnable parameters within one hour. After training, LLaMA-Adapter exhibits superior instruction-following and multi-modal reasoning capacity.

- Utilize the 52K instruction-output data, Freeze the entire LLaMA mode

- Append a set of learnable **adaption prompts** as prefix to the input instruction tokens in higher layer

➔ These prompts learn to adaptively inject new instructions (conditions) into the frozen LLaMA

- To avoid noise from adaption prompts at the early stage, modify the vanilla attention at inserted layers to be zero-initialized attention, with a learnable gating factor.

- Initialized by zero vectors, the gating can firstly preserve the original knowledge in LLaMA, and progressively incorporate instructional signals during training.

LLaMA-Adapter

- Learnable Adaption prompts

- N-layer Transformer
- Prompts for L transformer layers as $\{P_l\}_{l=1}^{L}$, where $P_l \in R^{(K \times C)}$

  K: prompt length for each layer,

  C: feature dimension of LLaMA's transformer.

- Note that we insert the prompts into the topmost L layers of the transformer ($L \leq N$)

  ➔ This can better tune the language representations with higher-level semantics.

- l-th inserted layer as an example ($l \leq L$), M-length word tokens as $T_l \in R^{(M \times C)}$, which represent the input instruction and the already generated response

- The learnable adaption prompt is concatenated with T_l along the token dimension as prefix, formulated as. $[P_l; T_l] \in R^{(K+M) \times C}$

LLaMA-Adapter

- Zero-initialized Attention

$$Q_l = \mathrm{Linear_q}(\ t_l\ ); \qquad t_l \in R^{(1 \times C)}$$
$$K_l = \mathrm{Linear_k}(\ [P_l;\ T_l;\ t_l]\ );$$
$$V_l = \mathrm{Linear_v}(\ [P_l;\ T_l;\ t_l]\ ).$$

Then, the attention scores of Q_l and K_l before the softmax function are calculated as

$$S_l = Q_l K_l^T / \sqrt{C} \ \in \mathbb{R}^{1\times(K+M+1)},$$

Which records the feature similarities between the new word t_l and all K + M + 1 tokens.

Meanwhile, S_l can be reformulated by two components as

$$S_l = [S_l^K;\ S_l^{M+1}]^T, \quad \text{where } S_l^K \in \mathbb{R}^{K\times 1} \text{ and } S_l^{M+1} \in \mathbb{R}^{(M+1)\times 1}$$

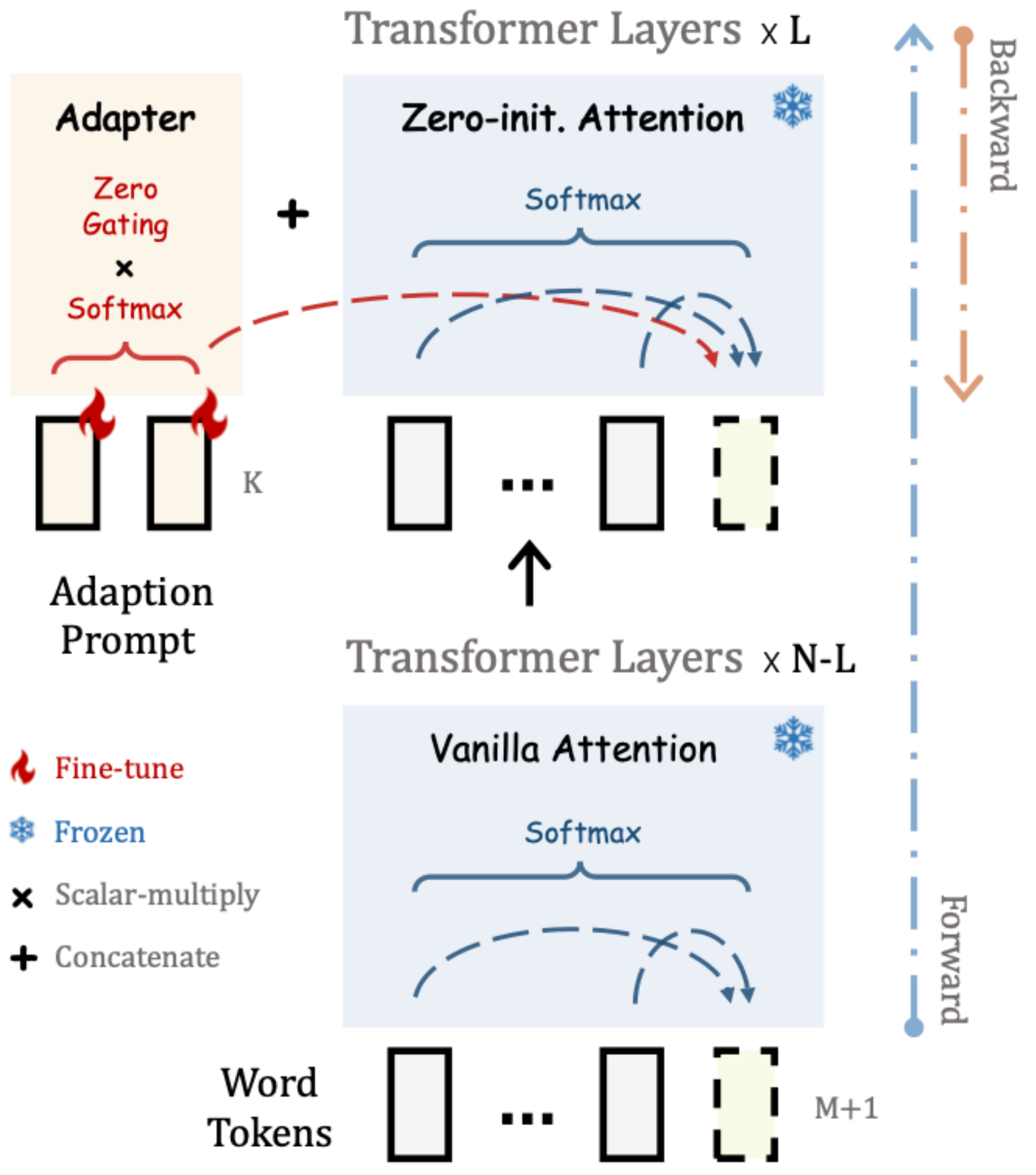


Figure 2: **Details of LLaMA-Adapter.** We insert lightweight adapters with learnable prompts into L out of N transformer layers of LLaMA. To progressively learn the instructional knowledge, we adopt zero-initialized attention with gating mechanisms for stable training in early stages.

LLaMA-Adapter

- Zero-initialized Attention

Apply the softmax functions to the two components previous Equation , and multiply the first term by gl, formulated as

$$S_l^g = [\mathrm{softmax}(S_l^K) \cdot g_l;\ \ \mathrm{softmax}(S_l^{M+1})]^T.$$

Calculate the output of the l-th attention layer with a linear projection layer as

$$t_l^o = \mathrm{Linear}_o(S_l^g V_l) \in \mathbb{R}^{1\times C}$$

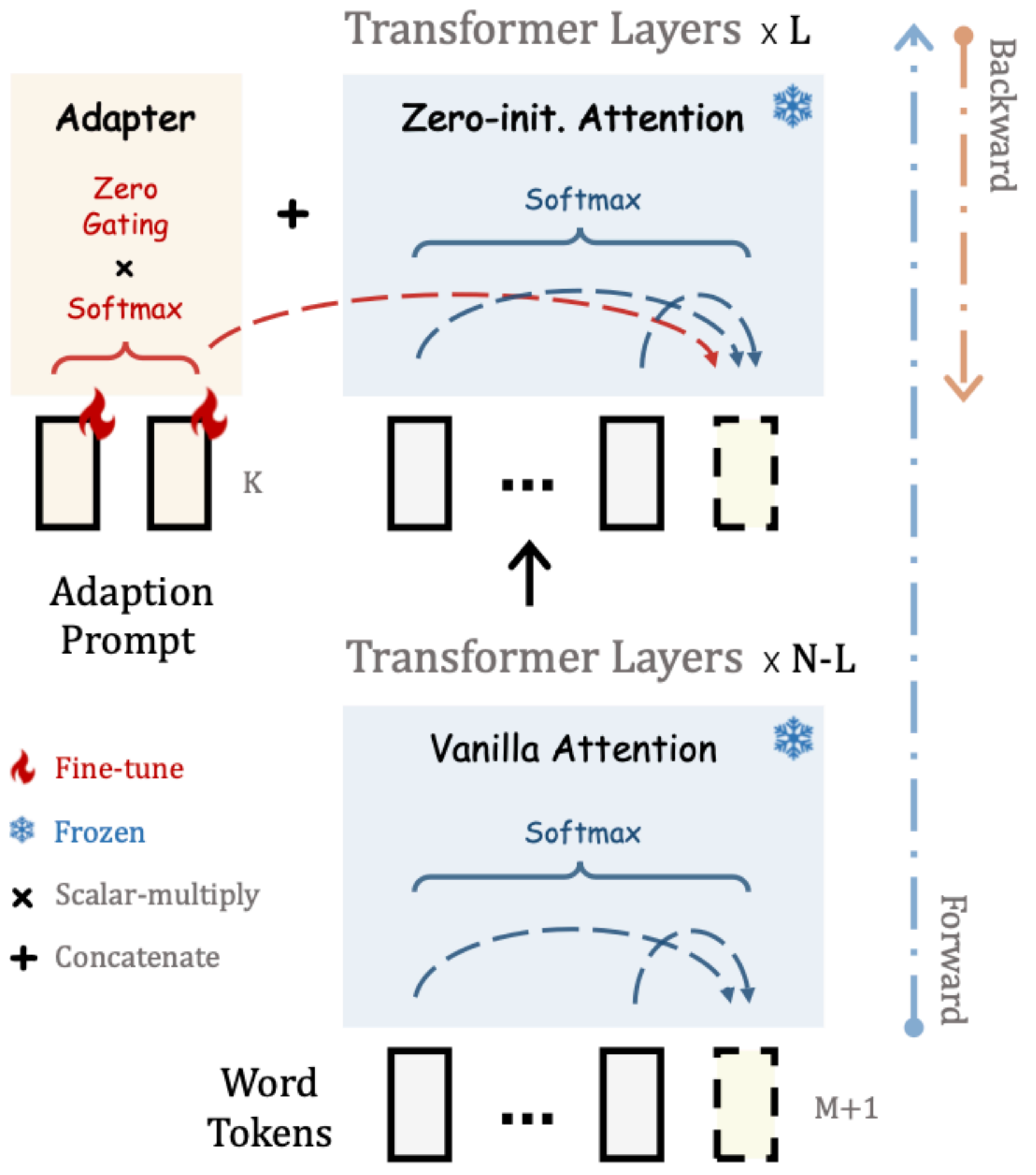


Figure 2: **Details of LLaMA-Adapter.** We insert lightweight adapters with learnable prompts into L out of N transformer layers of LLaMA. To progressively learn the instructional knowledge, we adopt zero-initialized attention with gating mechanisms for stable training in early stages.

LLaMA-Adapter

- Multi-modal Reasoning

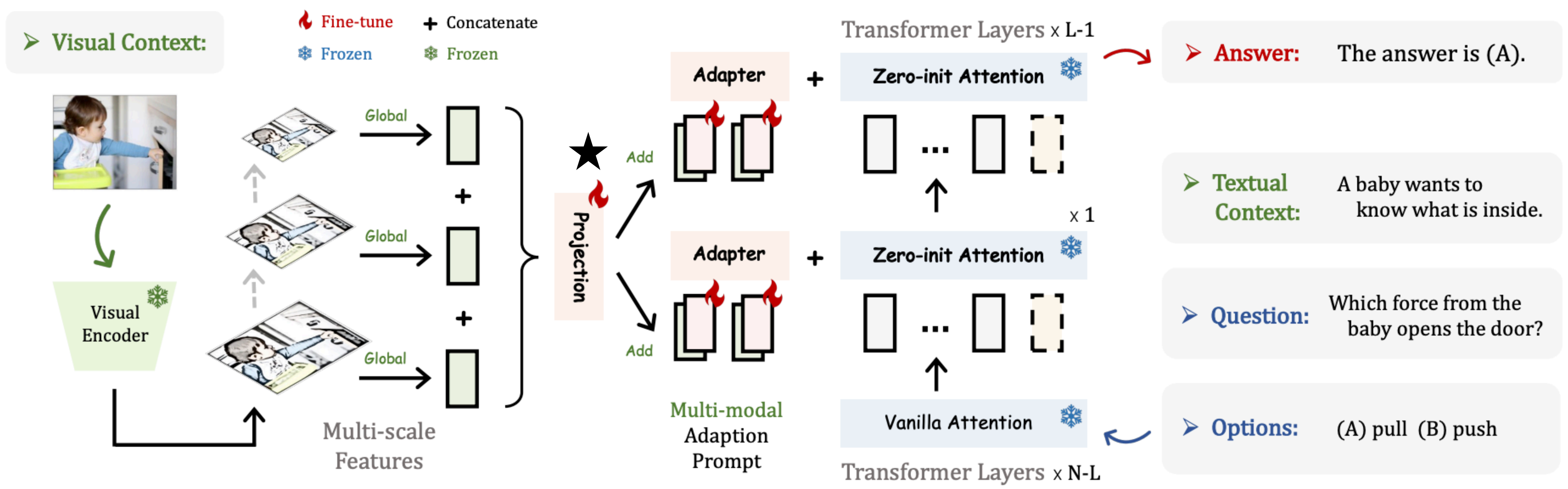


Figure 3: **Multi-modal Reasoning of LLaMA-Adapter.** On ScienceQA benchmark [41], LLaMA-Adapter is extended to a multi-modal variant for image-conditioned question answering. Given an image as the visual context, we acquire the global image token by multi-scale aggregation, and element-wisely add it onto the adaption prompts for visual instruction following.

$$\bigstar I_p = \text{Projection}\left(\text{Concat}\left(\{I_m\}_{m=1}^{M}\right)\right), \text{ where } I_p \in \mathbb{R}^{1\times C}$$

$$P_l^v = P_l + \text{Repeat}(I_p) \in \mathbb{R}^{K\times C},$$

- Using CLIP, to extract its multi-scale global features, denoted as $\{I_m\}_{m=1}^{M}$
- I_P is overall image token with the same feature dimension as our adaption prompts
- P_l^v denotes the adaption prompt incorporating visual information from the given image context

Experiments

- Instruction-following Evaluation

Setting

- Utilize

  Data: 52K instruction-following data (Alpaca)

  Model: LLaMA model with 7B parameters and N = 32 transformer layers

- prompt length K = 10 and insert the adaption prompts into the last L = 30 layers

- For quantitative evaluation , we ask GPT-4 to assess the response quality of instruction-following models on 80 questions

Experiments

- Instruction-following Evaluation

Figure 6: **Quantitative Comparison** between LLaMA-Adapter, Alpaca [60] and Alpaca-LoRA [1], evaluated by GPT-4 [45].

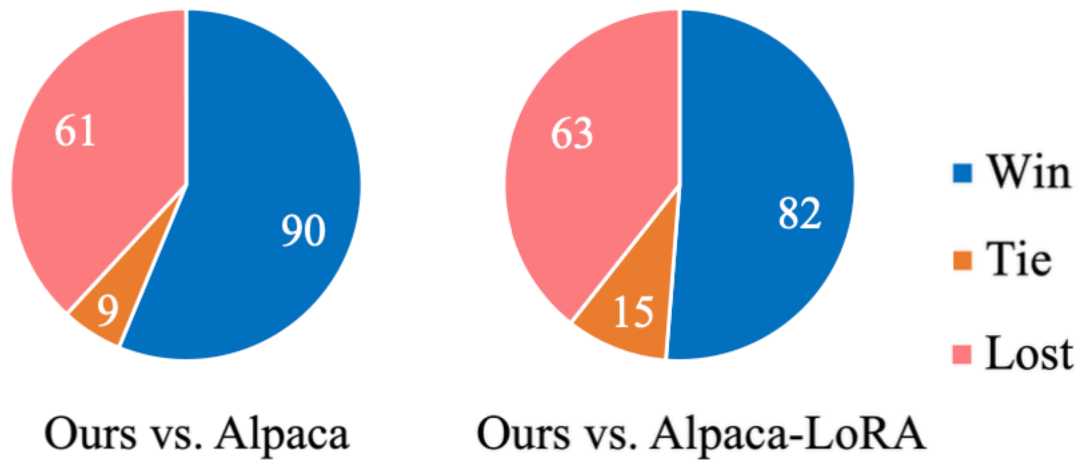


Table 1: **Efficiency Comparison** of different instruction-following methods. The training time is tested on 8 A100 GPUs.

Model	Tuned Params	Storage Space	Training Time
Alpaca [60]	7B	13G	3 hours
Alpaca-LoRA [1]	4.2M	16.8M	1.5 hours
LLaMA-Adapter	**1.2M**	**4.7M**	**1 hour**

Performance

- Reasonable responses comparable to the fully fine-tuned Alpaca, including question answering, language translation, and code generation

- LLaMA-Adapter obtains more 'win' compared to Alpaca and Alpaca-LoRA
  ➔ This fully demonstrates the effectiveness of our adapters with zero-initialized attention mechanisms.

Efficiency

- As a lightweight plug-and-play module, LLaMA-Adapter enjoys superior training efficiency with only 1.2M parameters, 4.9M storage, and one-hour training

Experiments

- Instruction-following Evaluation

Instruction: Tell me about alpacas.

Alpaca: Alpacas are small, fluffy animals related to camels and llamas. They are native to Peru and Bolivia, and were first domesticated around 5,000 years ago. They are kept mainly for their fine, soft fleece, which is used to make knitwear and other garments. Alpacas are herd animals and live in small family groups. They are highly social animals and can form strong bonds with their owners.

LLaMA-Adapter: Alpacas are members of the camelid family and are native to the Andes Mountains of South America. They are typically found in herds of 10-20 animals and are known for their gentle nature and their ability to thrive in harsh climates. Alpacas are typically white or grey in color and have long, silky coats. They are also known for their ability to produce high-quality wool, which is used in the production of clothing and other products.

Instruction: Translate the sentence 'I have no mouth but I must scream' into Spanish.

Alpaca: No tengo boca, pero debo gritar.

LLaMA-Adapter: No tengo boca pero tengo que gritar.

Instruction: Write a Python program that prints the first 10 Fibonacci numbers.

Alpaca:

```python
numbers = [0, 1]
for i in range(2, 11):
    numbers.append(numbers[i-2] + numbers[i-1])
print(numbers)
```

LLaMA-Adapter:

```python
def fibonacci():
    a, b = 0, 1
    for i in range(10):
        a, b = b, a + b
        print(a, b)
    print("Fibonacci numbers: ", a, b)
fibonacci()
```

Experiments

- Multi-modal Evaluation

Table 2: **Question Answering Accuracy (%) on ScienceQA's [41] test set.** We report GPT-3 [4], ChatGPT [2], and GPT-4 [45] for zero-shot inference. CoT denotes to utilize additional chain of thought for question answering. T denotes the single-modal model with text-only input.

Model	Tuned Params	Avg	NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12
Random Choicc [41]	-	39.83	40.28	46.13	29.25	47.45	40.08	33.66	39.35	40.67
Human [41]	-	88.40	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42
MCAN [65]	95M	54.54	56.08	46.23	58.09	59.43	51.17	55.40	51.65	59.72
VisualBERT [33, 34]	111M	61.87	59.33	69.18	61.18	62.71	62.17	58.54	62.96	59.92
UnifiedQA [27]	223M	70.12	68.16	69.18	74.91	63.78	61.38	77.84	72.98	65.00
UnifiedQA$_{CoT}$	223M	74.11	71.00	76.04	78.91	66.42	66.53	81.81	77.06	68.82
GPT-3 [4]	0M	74.04	75.04	66.59	78.00	74.24	65.74	79.58	76.36	69.87
GPT-3$_{CoT}$	0M	75.17	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68
ChatGPT$_{CoT}$ [2]	0M	78.31	78.82	70.98	83.18	77.37	67.92	86.13	80.72	74.03
GPT-4$_{CoT}$ [45]	0M	83.99	85.48	72.44	90.27	82.65	71.49	92.89	86.66	79.04
MM-COT$_T$ [74]	223M	70.53	71.09	70.75	69.18	71.16	65.84	71.57	71.00	69.68
MM-COT	223M	84.91	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37
LLaMA-Adapter$_T$	**1.2M**	**78.31**	**79.00**	**73.79**	**80.55**	**78.30**	**70.35**	**83.14**	**79.77**	**75.68**
LLaMA-Adapter	**1.8M**	**85.19**	**84.37**	**88.30**	**84.36**	**83.72**	**80.32**	**86.90**	**85.83**	**84.05**

LLaMA-Adapter demonstrates superior parameter efficiency while achieving competitive question answering capacity

Experiments

- Ablation Study

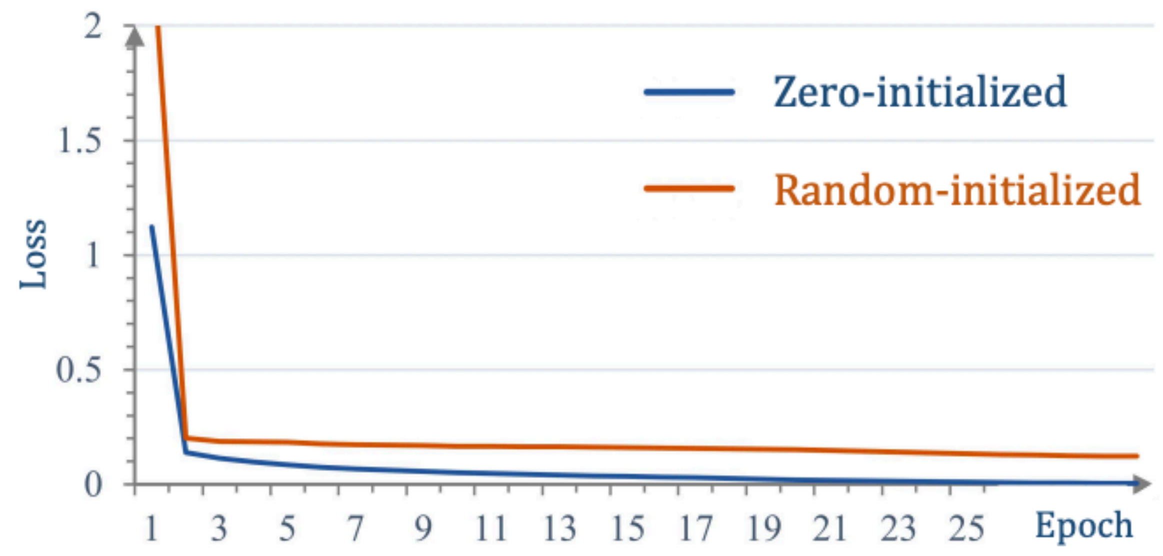
Table 3: **Ablation on Inserted Layers** of LLaMA's transformer.

Layers	Params	Val Acc (%)
10	0.97	55.95
20	1.37	73.36
30	1.79	**83.85**

Table 4: **Ablation on Zero-initialized Attention.** Blue highlights the gain.

Setting	Val Acc (%)
Rand-Init Attention	40.77
Zero-Init Attention	**83.85**
Gain	+43.08

Figure 7: **Loss Curves** with (blue) and without (orange) zero-initialized attention.



Insertion Layers

: To investigate the number of transformer layers to be inserted in LLaMA-Adapter

- Increasing the layer numbers introduces more parameters, but leads to a large improvement in the accuracy of ScienceQA's validation set
- It indicates that more adaption prompts at different layers can provide stronger task-specific guidance to the pre-trained LLaMA

Zero-initialized Attention

- Significant +43.08% performance gain on the validation set.
- This comparison demonstrates the decisive role of zero-initialized attention in our approach
- zero-init attention' converges faster and reaches lower loss bounds than 'rand-init attention

Conclusion

- 1.2M parameters and one-hour training, our approach effectively fine-tunes LLaMA with superior efficiency compared to the 7B-parameter Alpaca

- Introduce zero-initialized attention with gating mechanism, which adaptively incorporates instructional signals, while preserving the pre-trained knowledge in LLaMA.

- LLaMA-Adapter can be generalized to image conditions for multi-modal reasoning, achieving competitive results on ScienceQA and COCO Caption benchmarks

LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model

Peng Gao[*‡†1], Jiaming Han[*1], Renrui Zhang[*1,2], Ziyi Lin[*2], Shijie Geng[3], Aojun Zhou[2]
Wei Zhang[1], Pan Lu, Conghui He[1], Xiangyu Yue[2], Hongsheng Li[†2], Yu Qiao[†1]

[1]Shanghai Artificial Intelligence Laboratory [2]CUHK MMLab
[3]Rutgers University

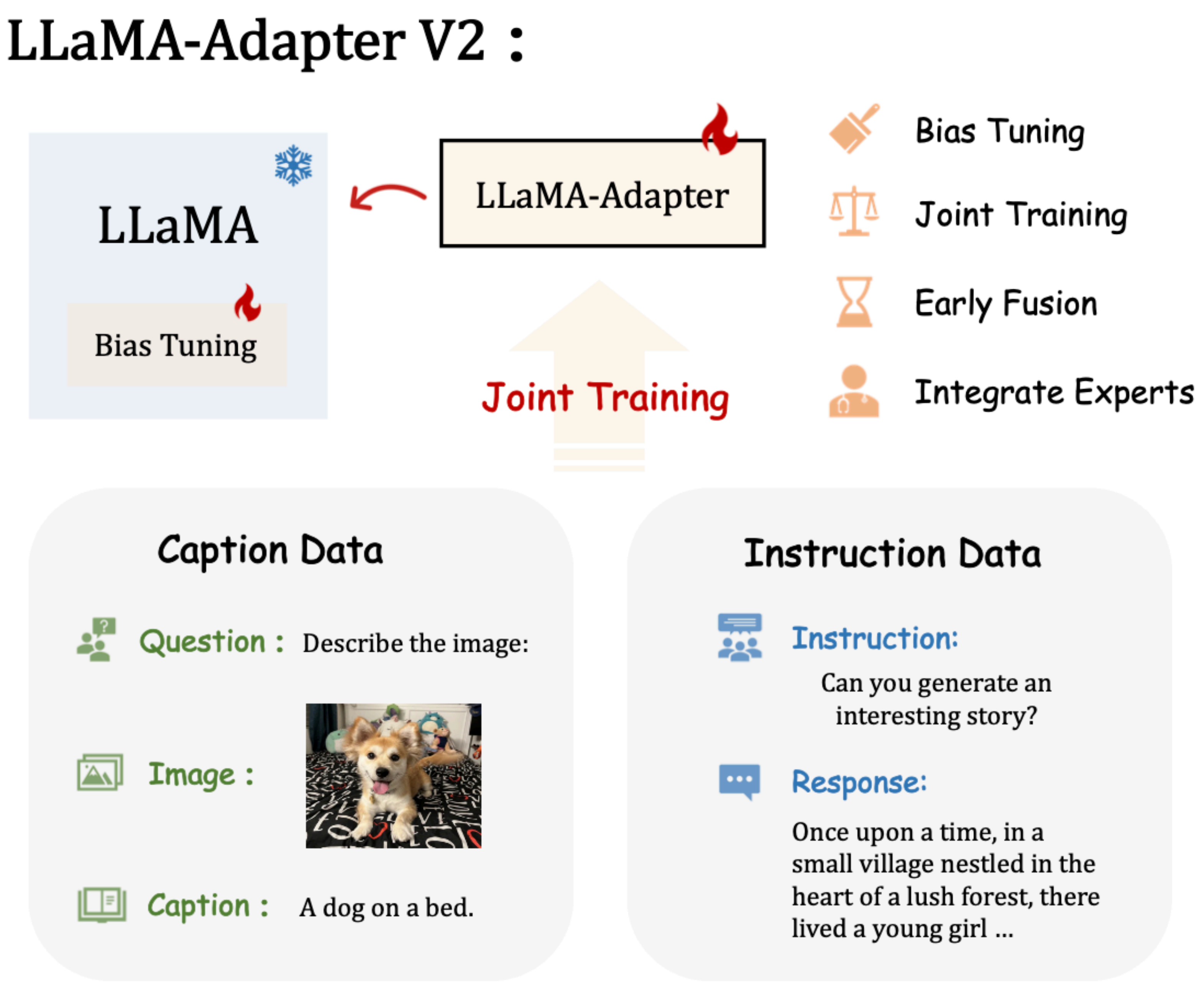
{gaopeng, hanjiaming, zhangrenrui, qiaoyu}@pjlab.org.cn

Introduction

- 최근 MiniGPT-4와 LLaVA 같은 연구에서 language-only instruction models을 multi-modal로 확장하는 연구 다수
- MiniGPT-4는 frozen visual encoder와 LLM을 1억 3400만 개의 이미지-텍스트 쌍으로 사전 훈련 후 aligned된 이미지-텍스트 데이터셋을 통해 추가 튜닝
- LLaVA도 image-text pair를 활용, GPT-4에서 생성한 15만 개의 high-quality multi-modal instruction 데이터로 전체 LLM을 추가 튜닝
- 상기 방법들은 인상적인 다중 모달 이해 능력을 보이지만, 수십억 개의 모델 매개변수를 업데이트하고 대량의 다중 모달 훈련 데이터를 정교하게 수집해야함

Model	Language Instruction Data			Image-Text Data		Visual Instruction Data		Tuning Parameters
	Source	Type	Size	Source	Size	Source	Size	Size
MiniGPT-4 [78]	ShareGPT [1]	Conversation	70K	CC-VG-SBU-L400*	134M	CC+ChatGPT	5K	13B
LLaVA [38]	ShareGPT [1]	Conversation	70K	CC*	595K	COCO+GPT4	158K	13B
LLaMA-Adapter V2	GPT-4-LLM [50]	Single-turn	**52K**	COCO [6]	**567K**	-	**0**	**14M**

Introduction



- **Stronger Language Instruction Model**: LLaMA-Adapter V2 surpasses its predecessor LLaMA-Adapter in terms of language instruction-following performance.

- **Balanced Visual Instruction Tuning:** Strategy to solve the interference between image-text alignment and instruction- following learning targets

- **Integration of Expert Systems :** Different expert models can be integrated into our framework

LLaMA-Adapter V2

- Bias Tuning of Linear Layers

내부 parameter는 수정되지 않기 때문에 fine-tuning 능력에 제한

In LLaMA-Adapter V2

1. Unfreeze all the normalization layers in LLaMA & each linear layer, add a bias and a scale factor as two learnable parameters

2. bias, scale factor 를 각각 0과 1로 초기화하여 초기 단계의 훈련 과정을 안정화

$$\mathbf{y} = \mathbf{W} \cdot \mathbf{x} \quad \rightarrow \quad \mathbf{y} = s \cdot (\mathbf{W} \cdot \mathbf{x} + b),$$
$$\text{where} \quad b = \text{Init}(0), \quad s = \text{Init}(1).$$

LLaMA-Adapter V2

- Joint Training with Disjoint Parameters

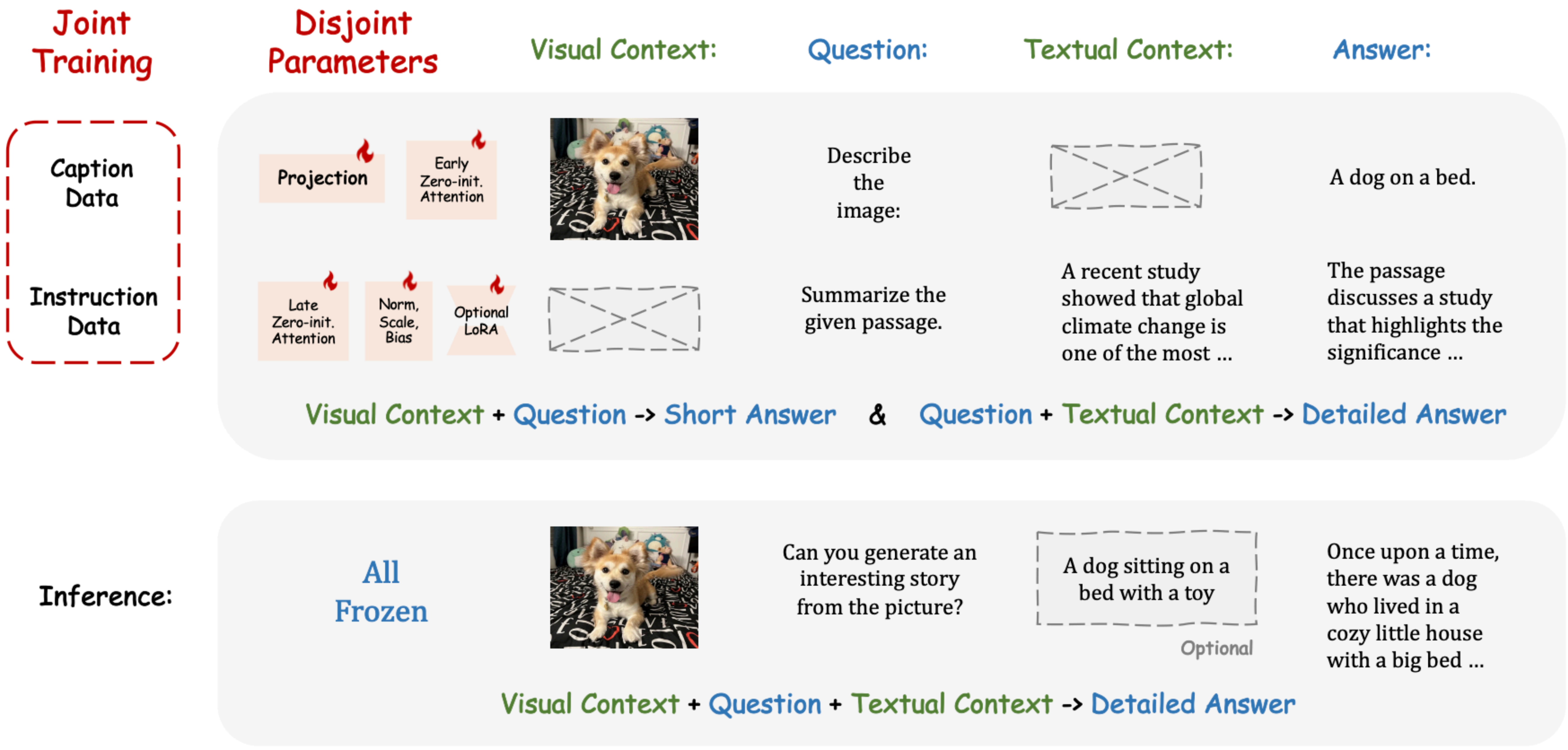


Figure 2. **Joint Training Paradigm in LLaMA-Adapter V2.** We utilize both image-text caption and language-only instruction data to jointly train LLaMA-Adapter V2, optimizing disjoint groups of learnable parameters.

500K 개의 image-text caption data와 50K개의 instruction data 간의 크기 차이로 인해 단순 학습 시 instruction-following 기능 손상 가능성

1. Visual projection layers and early zero-initialized attention with gating are trained for image-text captioning data

2. Adaptation prompts together with zero gating, the unfrozen norm, newly added bias and scale factors are utilized for learning from the instruction-following data

LLaMA-Adapter V2

- Early Fusion of Visual Knowledge

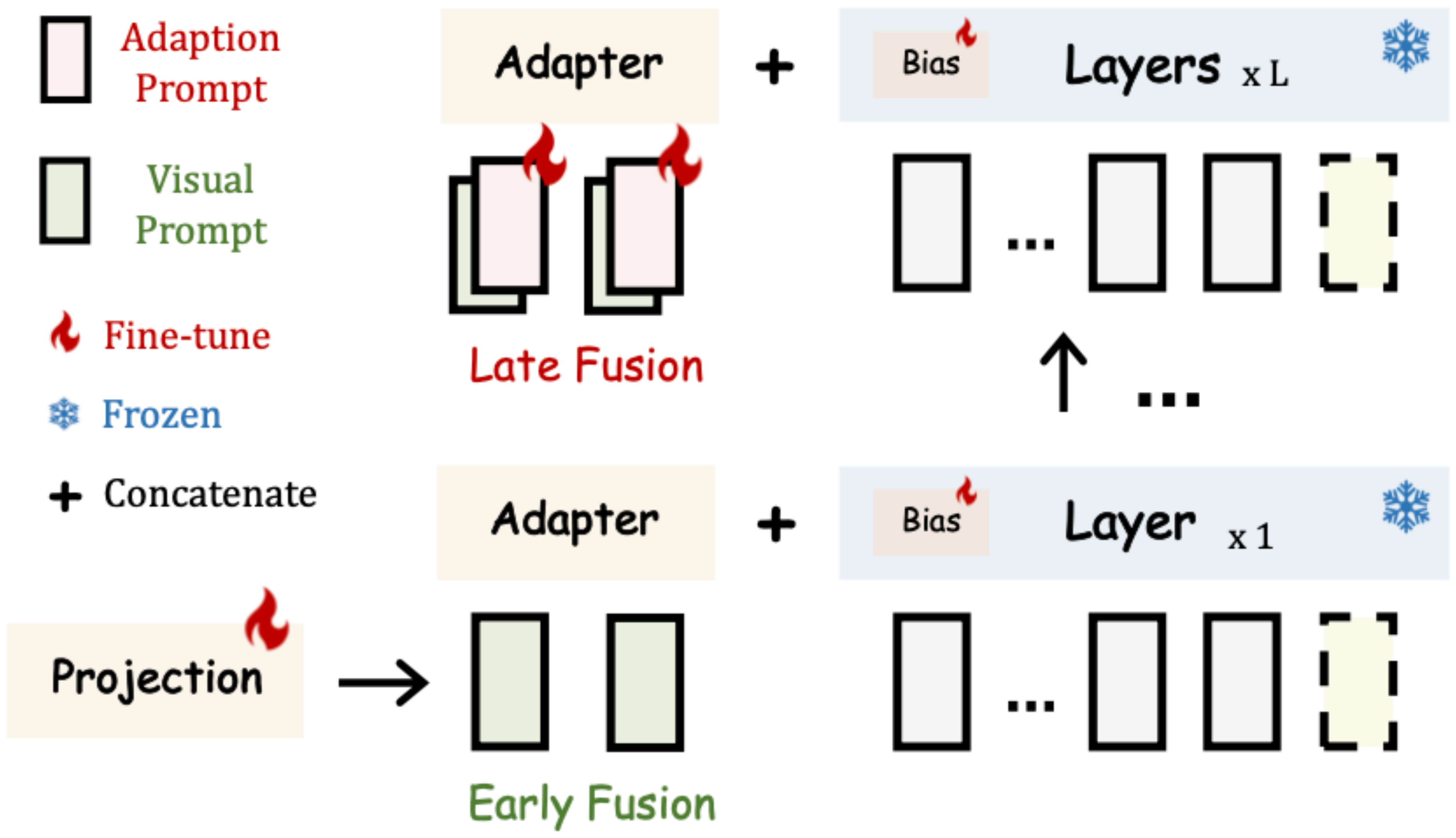


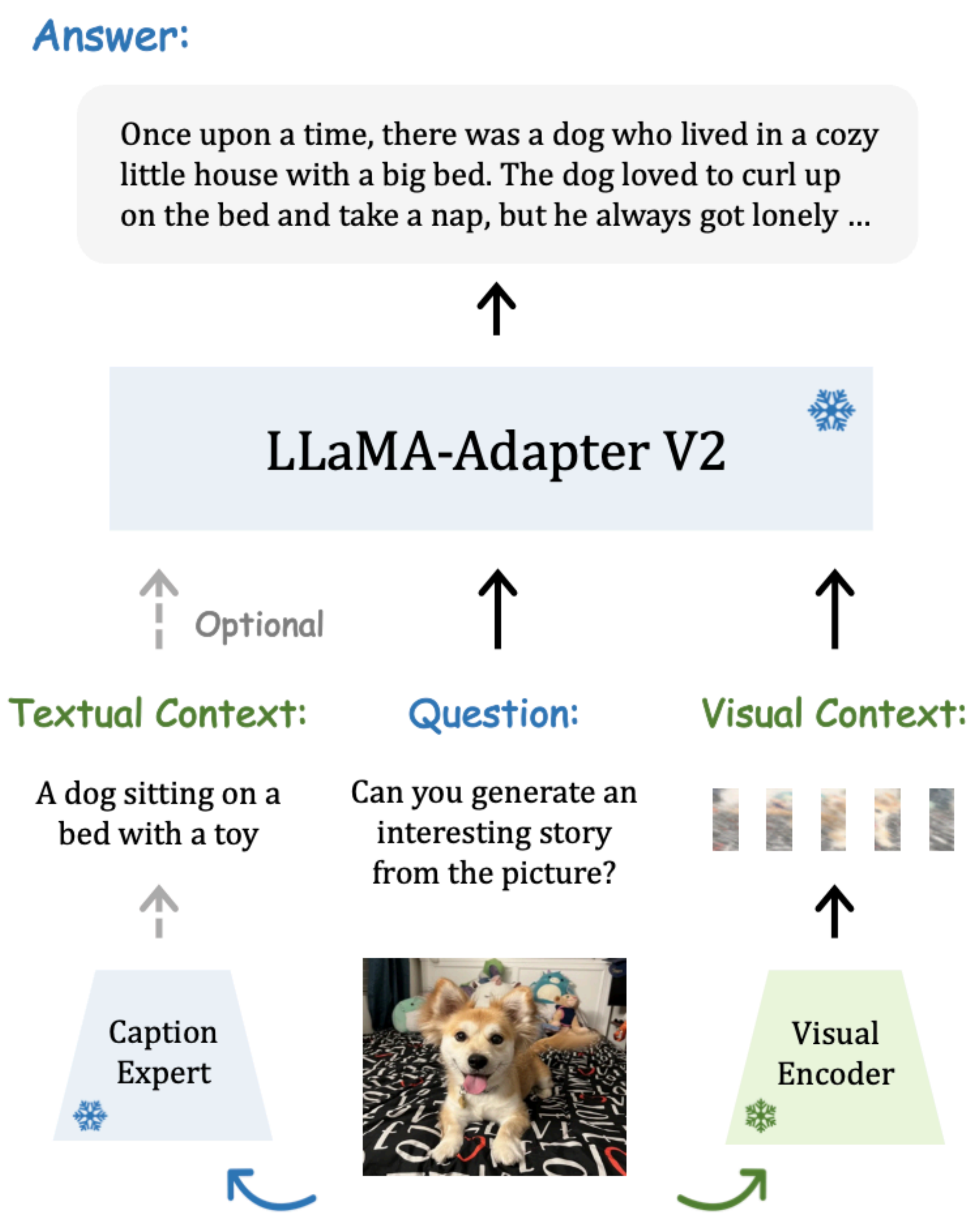
Figure 3. **Early Fusion of Visual Knowledge.** Following LLaMA-Adapter, we insert static adaptation prompts into the last L layers. For visual prompts, we insert them in the early stage of LLM, disjointing with adaptation prompts.

Visual prompt의 경우 zero-initialized attention과 함께 첫 번째 transformer layer의 단어 토큰에 직접 추가되고 후기 L 레이어에서 adaptation 프롬프트와 합쳐짐.

1. this simple early fusion strategy of visual tokens can effectively resolve the conflict between the two types of fine-tuning targets

LLaMA-Adapter V2

- Integration with Experts



LLaMA-Adapter V2 접근 방식의 이미지 이해 능력은 MiniGPT- 4, LLaVA 에 비해 상대적으로 약하며, 때로는 부정확하거나 관련 없는 응답을 유발

1. Rather than collecting more image-text data or adopting stronger multi-modal modules

   ➔ Integrating expert systems, such as captioning, OCR, and search engines

Experiments

- Stronger Language Instruction Model

Setting

- Utilize

  Data: 52K single-turn instruction data from GPT- 4-LLM and 567K captioning data from COCO Caption, 80K conversation data collected by ShareGPT

  Model: LLaMA model with 7B parameters and N = 32 transformer layers

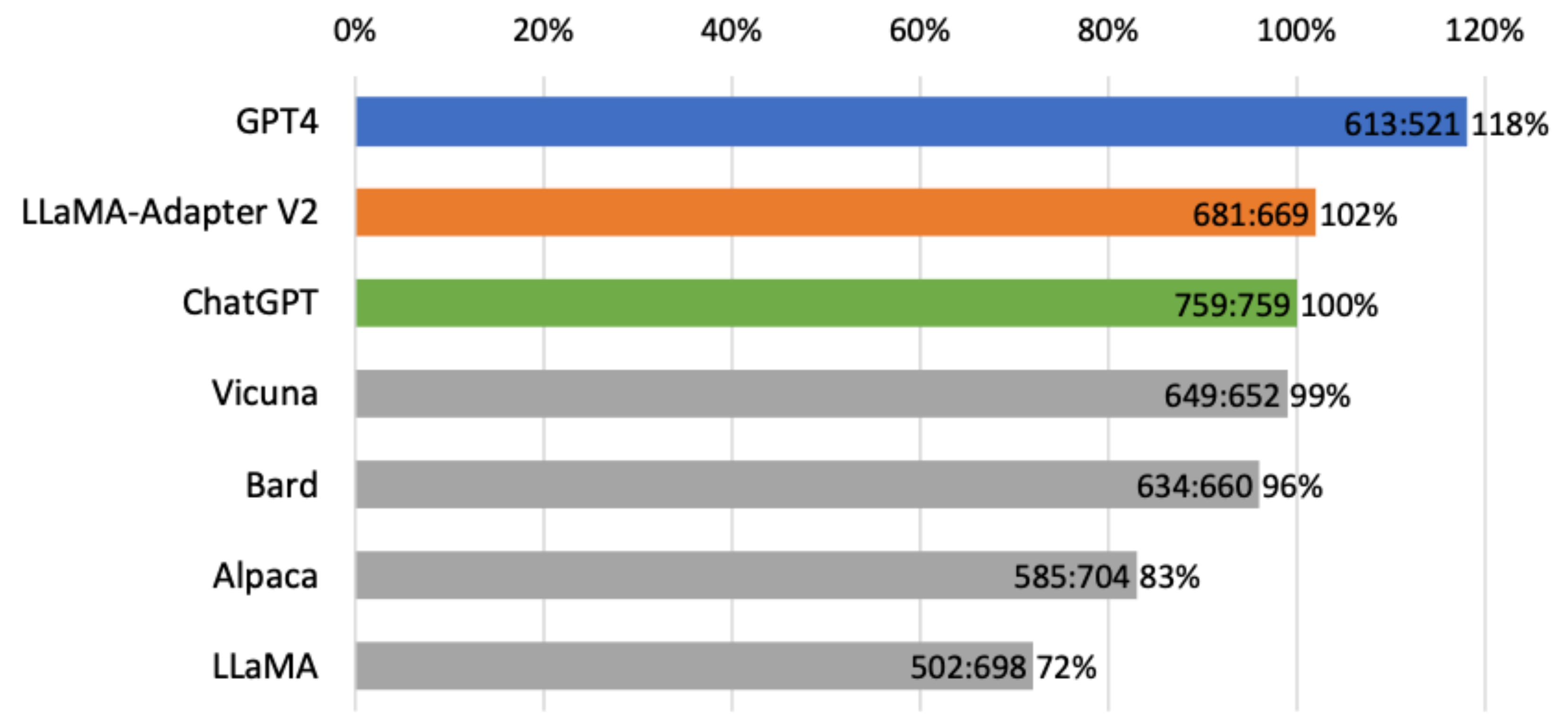
- Dynamic visual prompts to the first layer

- Prompt length K = 20 and insert the adaption prompts into the last L = 31 layers

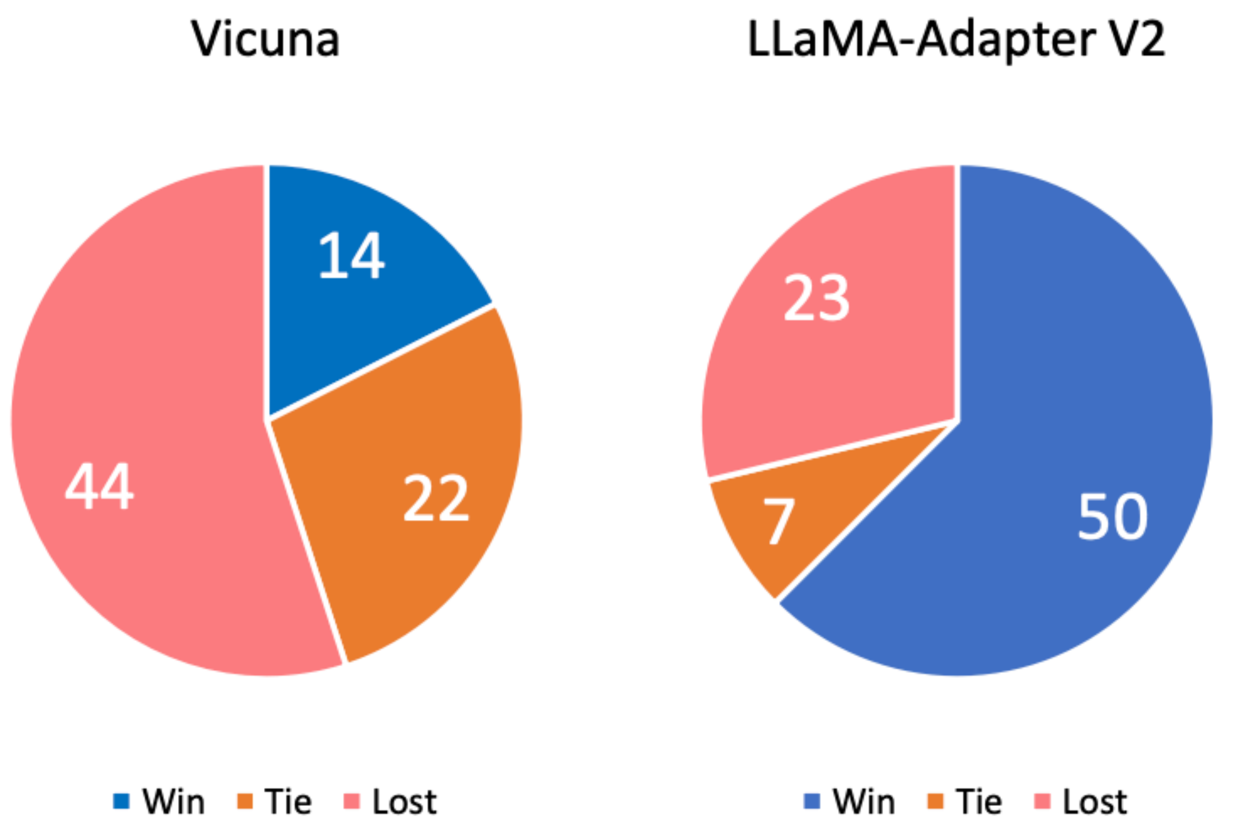
- All the parameters in normalization layers, linear layer bias and scale are set to be updated during training, while the remaining parameters in LLaMA are kept frozen.

Experiments

- Stronger Language Instruction Model



(a) The total quality scores compared with ChatGPT.



(b) Left: Vicuna *vs.* ChatGPT. Right: LLaMA-Adapter V2 *vs.* ChatGPT.

Figure 5. **Response Quality Comparisons** assessed by GPT4. The comparison baseline is ChatGPT. We use GPT4 to assess the response quality of different methods on 80 questions [7]. Note that our model is based on LLaMA-65B, while Vicuna is built on LLaMA-13B. But we only fine-tune **14M** parameters, while Vicuna requires fine-tuning the whole 13B parameters.

we also use GPT4 to assess the response quality, where LLaMA-Adapter V2 even beats ChatGPT on the total scores and 50/80 questions.

Experiments

- Visual Instruction Model (captioning)

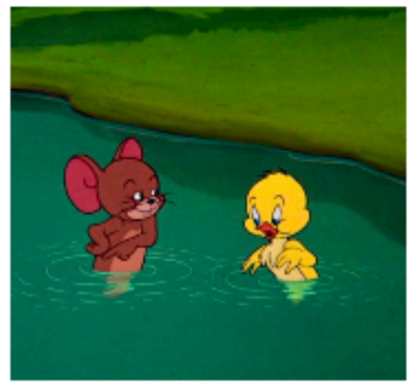
Success Case 1

V1: A bowl of cereal, a cup of coffee and a sandwich on a table.
V2: The image features a dining table set up with a variety of food items, including a bowl of fruit, a cup of coffee, and a plate of pastries. The table is set with a knife, fork, and a spoon, suggesting that the food is ready to be consumed. The dining table is situated in a kitchen, with a chair placed near the table. The chair is positioned next to a small table, which may be used for additional food items or drinks. The room is well-lit, with a lamp illuminating the area.

Success Case 2

V1: A group of people waiting to board a bus.
V2: In the image, a large bus is parked on the side of the road, with several people standing nearby. There are at least three people standing near the bus, with one person holding a backpack and another person holding a bag. A few more people can be seen in the background, walking towards the bus. The bus is painted white, and its front end is visible, with a few people standing near the front. A few cars are also visible in the background, adding to the busy street scene.

Failure Case



V1: A mouse and a duck in a body of water.
V2: In the image, two men are standing on a dock near a lake or river. One man is wearing a blue shirt and the other is wearing a white shirt. Both men are looking down and appear to be fishing. They are standing near a small rocky area and are in the water up to their knees. In the image, there is also a dog, a yellow labrador retriever, swimming in the water near the men. The dog appears to be swimming towards the men, perhaps attracted by the presence of the fishermen. The dog is wearing a red collar and appears to be enjoying its time in the water with the two men.

Figure 6. **Comparisons of Image Captioning Results** between LLaMA-Adapter and LLaMA-Adapter V2.

This failure is the lack of an image-text alignment stage, which models like MiniGPT-4 and LLaVA have incorporated.

➔ This motivates us to employ additional expert systems to enhance the image understanding ability.

Experiments

- Visual Instruction Model (Visual Understanding)

Instruction: Explain why this image is beautiful.
Answer: This image is beautiful because it captures the serene and peaceful atmosphere of the beach, with the sun setting in the background. The image features a beautiful, golden sunset over the ocean, creating a stunning backdrop for the solitary figure walking along the sandy beach. The combination of the warm, golden light of the sunset, the blue water of the ocean, and the white sandy beach creates a breathtaking and calming scene. The image also highlights the vastness of the ocean, with the waves crashing against the shore, creating a soothing sound that adds to the tranquility of the setting. The image conveys a sense of relaxation and tranquility, making it a perfect representation of a beautiful beach scene.

Instruction: What city might this picture be taken from? Can you give me some suggestions if I want to visit this city?
Answer: The image might be taken from Beijing, China. Beijing is the capital city of China and one of the most populous cities in the world. It is known for its rich history and culture, as well as its modern infrastructure and architectural landmarks. Some popular attractions in Beijing include the Forbidden City, Temple of Heaven, and the Great Wall of China, among others. If you want to visit Beijing, it is recommended to explore its historical and cultural heritage, as well as its modern developments.

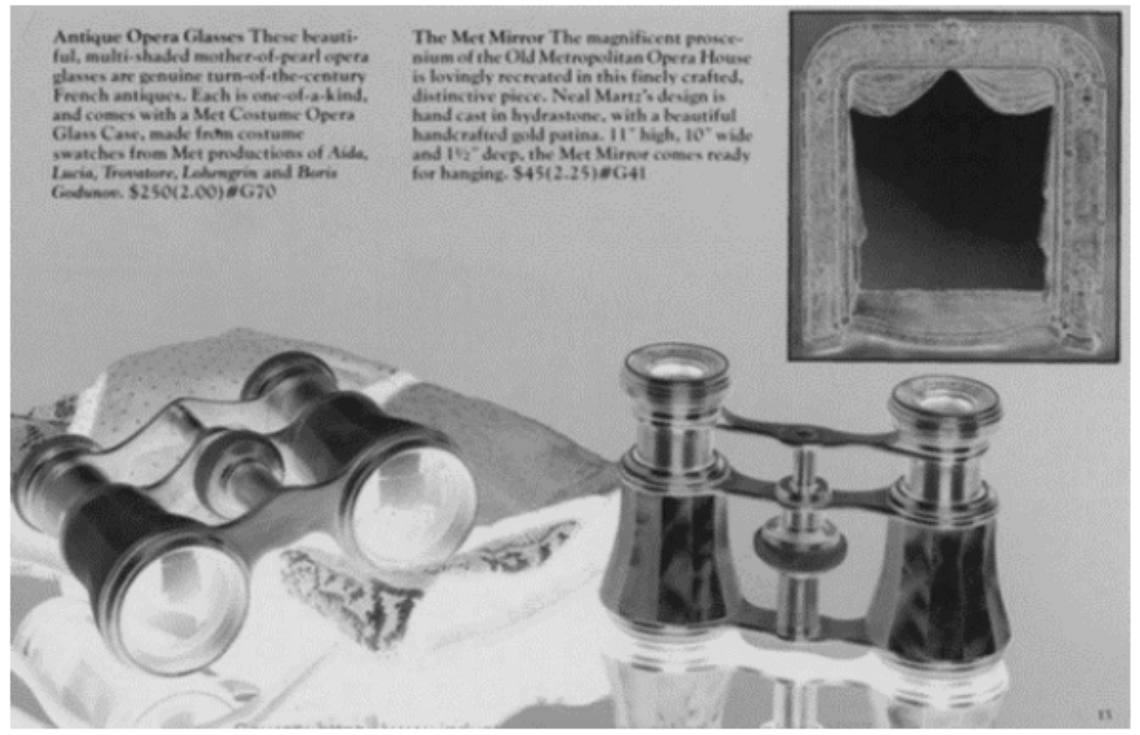
Instruction: What should the car do here?
Answer: The car should follow the pedestrian crossing the street, as indicated by the red light. The car should stop at the red light and wait for the pedestrian to cross the street safely before proceeding. This is a common practice to ensure the safety of pedestrians and drivers alike.

Figure 7. **The Visual Understanding Examples** of LLaMA-Adapter V2.

Joint training and early fusion techniques, LLaMA-Adapter V2 exhibits exceptional multi-modal understanding capabilities

Experiments

- Visual Instruction Model



Question: What is the cost of the Antique Opera glasses? Do you think it's worth buying the Antique Opera glasses and can you explain the reason in detail?
OCR Context: 1. Antique Opera Glasses These beauti- 2. ful, multi-shaded mother-of-pearl opera 3. The Met Mirror The magnificent prosce- 4. glasses are genuine turn-of-the-century 5. nium of the Old Metropolitan Opera House 6. French antiques. Each is one-of-a-kind, 7. is lovingly recreated in this finely crafted, 8. and comes with a Met Costume Opera 9. distinctive piece. Neal Martz's design is 10. hand cast in hydrastone, with a beautiful 11. Glass Case, made from costume 12. swatches from Met productions of Aida, 13. handcrafted gold patina. 11 high, 10 wide 14. Lucia, Trovatore, Lohengrin and Boris 15. and 11/2 deep, the Met Mirror comes ready 16. Godunov. $250(2.00) #G70 17. for hanging. $45(2.25) #G41 18. 13
Language Instruction Model: I'm sorry, but I am an AI language model and I don't have the ability to access images or prices. Could you please provide me with the price of the Antique Opera glasses so that I can provide you with an accurate response?
Visual Instruction Model w/o Context: The cost of the Antique Opera glasses is not mentioned in the image. However, the image suggests that the glasses are being sold in a store, which means that the cost may vary depending on the location, the vendor, and the specific glasses being sold. It is important to note that the value of the Antique Opera glasses is subjective and may depend on the individual's taste, preference, and budget. Some people may find the antique design and the history behind the glasses appealing and may be willing to pay a higher price for them.
Visual Instruction Model w/ Context: The cost of the Antique Opera glasses is $250 (2.00) and it is worth buying because they are genuine turn-of-the-century French antiques that are lovingly recreated in this finely crafted, hand cast in hydrastone, with a beautiful gold patina. The Met Mirror comes ready for hanging and is $45 (2.25).

Figure 9. **Visual Understanding with the help of OCR Experts.** The example and OCR context are from DocVQA [44].

Integration of experts thus largely extends the power of LLaMA-Adapter V2 and fully unleashes its multi-modal reasoning potential.

Conclusion

•사전 훈련된 LLMs를 제로샷 Visual instruction 모델로 변환할 수 있음

•Simple bias-tuning strategy and early-fusion of visual features

• Instruction-following 성능뿐만 아니라 챗봇과 같은 멀티턴 대화 능력도 갖춤

• 부정확한 이미지 설명 문제에 대해, OCR and image captioner과 같은 expert visual systems 결합 가능

• 향후 Visual instruction- following 능력 향상을 위해, Multi-model instruction dataset 이나 다른 PEFT method 결합하여 튜닝

감사합니다