Collaborative Training of Tiny-Large Vision Language Models

Anonymous Authors

1 EXPERIMENTS

1.1 Effect of the proportion of Distillation

The proportion of distillation loss significantly impacts model performance during the knowledge distillation process. Properly calibrating this loss proportion is crucial for achieving the optimal balance between knowledge transfer and model training efficiency. In this experiment, we explored the impact of the distillation ratio on the performance of our model. As shown in Table1, we observed When $\lambda = 0.7$ tiny LLMs get low performance in vision-language tasks. High distillation loss weight may lead to overfitting on the large model's specific behaviors, potentially ignoring the intrinsic learning capabilities of the tiny model. Due to their mutual influence, the large model ends up with suboptimal features, leading both models into a vicious cycle where their performance simultaneously deteriorates. When $\lambda = 0.1$, tiny LLMs might not sufficiently capture the nuanced knowledge or expertise from the Large model, resulting in suboptimal performance. Through experimentationwe use $\lambda = 0.3$ during our training process.

2 DETAILED TRAINING SETTINGS

2.1 Settings of Stage I

As shown in Table2, in this stage, the image encoder is initialized using CLIP-VIT-L/14 336, the tiny language model is initialized using OPT-125m, and the large language model is initialized using Vicuna-7B. All parameters except the image encoder and large language model are fully trainable. We employ the AdamW optimizer, weight decay at 0.1 and cosine learning rate schedule starting at 1e-3 for tiny language model. The training involves a total batch size of 256 across 8 A800 GPUs, extending over 300K iterations to process about 93 million samples. The input resolution is 336 ×336.

2.2 Settings of Stage II

In this stage, the Tiny language model inherits weights from the first stage. All parameters except the image encoder are fully trainable. The input images are processed at a resolution of 336 ×336.

For optimization, the AdamW optimizer is employed with weight decay seen at 0.05, and a total batch size of 64. The training extends over 40K steps across 8 A800 GPUs, inclusive of 2K warm-up steps. More detailed training settings are listed in Table 2.

Table 1: Ablation study of the proportion of KL-divergence Distillation Loss, we compare both Tiny and Large models' performance in vision-language tasks with different λ

	λ	VQAv2	GQA	Text VQA
Tiny LMs	0.1	59.8	43.5	38.7
	0.3	**68.3**	**49.3**	**45.2**
	0.7	40.3	25.6	23.4
Large LMs	0.1	78.9	62.7	60.5
	0.3	**79.9**	**63.8**	**62.3**
	0.7	63.8	54.9	53.1

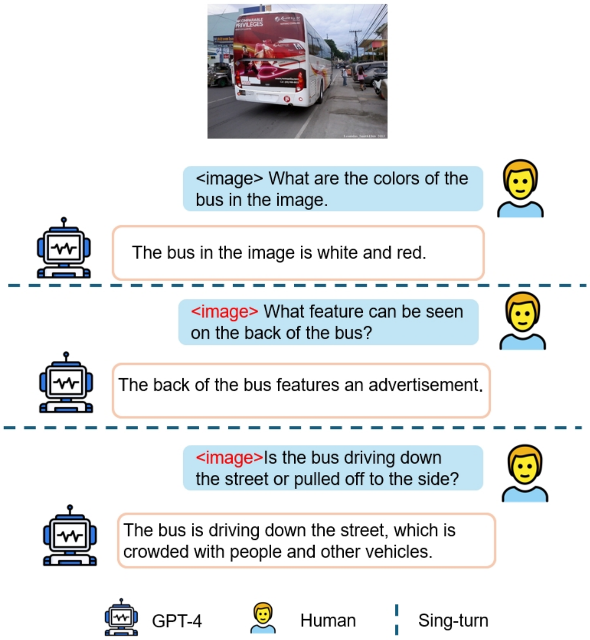
Table 2: Training setting of CTVLMs' stage 1 and stage 2.

config	Stage 1	Stage 2
iamge enc. Weight init	from clip	from clip
tiny text enc. Weight init	from opt125m	from stage 1
large text enc. Weithg init	from vicuna 7B	from stage 1
image enc. Peak learning rate	-	-
tiny text enc. Peak learning rate	1e-3	1e-4
large text enc. Peak learning rate	frozen	1e-5
cross attn peak learning rate	5e-5	-
learning rate schedule	cosine decay	cosine decay
optimizer	AdamW	AdamW
weight decay	0.1	0.05
input resolution	336	336
patch size	32	8
total batch size	256	64
warm-up iterations	5000	2000
samples seen	93M	1.5M
drop path rate	0	0
λ	0.3	0.3
numerical precision	DeepSpeed bf16	DeepSpeed bf16
trainable parameters	150M	7.1B
GPUs for training	$8\times A800(80G)$	$8\times A800(80G)$

3 DATA PREPARATION

To fully utilize web-scale image-text data, we adopted different data filtering strategies in stage 1 and stage 2. In Stage 1, we first applied only minor data filtering, thus retaining the vast majority of the data. We considered six factors: CLIP similarity, watermark probability, unsafe probability, aesthetic score, image resolution, and caption length, to remove extreme data points and avoid disrupting training stability. Then we deleted most of the low-quality data based on the captions, mainly considering the length, completeness, readability, and whether they were gibberish or boilerplate (like menus, error messages, or duplicate text), contained offensive language, placeholder text, or source code. We retained only 93 million entries.

In stage 2, we adopted a data augmentation method that deconstructs multi-turn conversations into several single-turn dialogues in LLaVA-mix-665k dataset. For instance, as shown in Figure**??**, we prepend a special image token to the question for the purely textual conversations without image inputs, allowing the model to input image features during training. This method not only maintains the multimodal nature of the training data but also ensures that the model remains adept at handling visual contexts alongside textual information. By training the model in this enriched environment, we aim to improve its ability to understand and generate responses that are contextually aligned with both text and imagery, thereby enhancing its overall multimodal processing capabilities.



<image> What are the colors of the bus in the image.

The bus in the image is white and red.

<image> What feature can be seen on the back of the bus?

The back of the bus features an advertisement.

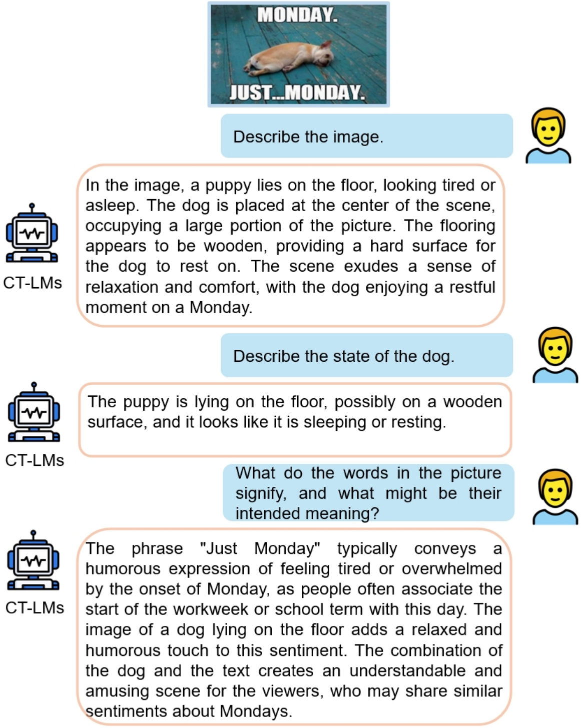
<image>Is the bus driving down the street or pulled off to the side?

The bus is driving down the street, which is crowded with people and other vehicles.

Figure 1: Sample for our data augmentation method. We have deconstructed the multi-turn dialogue in the figure into three separate single-turn dialogues, and for each question, we have inserted an image token.

4 VISUALIZATION

In this section, we visualized an example to demonstrate the conversational capabilities of our model. As shown in figure2, our model can understand images and the text contained within them, and it can provide corresponding descriptions or reasoning.



Describe the image.

In the image, a puppy lies on the floor, looking tired or asleep. The dog is placed at the center of the scene, occupying a large portion of the picture. The flooring appears to be wooden, providing a hard surface for the dog to rest on. The scene exudes a sense of relaxation and comfort, with the dog enjoying a restful moment on a Monday.

Describe the state of the dog.

The puppy is lying on the floor, possibly on a wooden surface, and it looks like it is sleeping or resting.

What do the words in the picture signify, and what might be their intended meaning?

The phrase "Just Monday" typically conveys a humorous expression of feeling tired or overwhelmed by the onset of Monday, as people often associate the start of the workweek or school term with this day. The image of a dog lying on the floor adds a relaxed and humorous touch to this sentiment. The combination of the dog and the text creates an understandable and amusing scene for the viewers, who may share similar sentiments about Mondays.

Figure 2: An example showcasing the capabilities of our CTLMs in image-text dialogues.