Supplementary Material
Continual SFT Matches Multimodal RLHF with Negative Supervision

Ke Zhu[1,2*] Yu Wang[2*] Yanpeng Sun[2,3] Qiang Chen[2] Jiangjiang Liu[2]
Gang Zhang[2] Jingdong Wang[2†]
[1]Nanjing University [2]Baidu
[3]Nanjing University of Science and Technology
zhuk@lamda.nju.edu.cn, {wangyu106,wangjingdong}@baidu.com

A. Theoretical derivation

A.1. Relations between DPO and SFT.

In this section, we want to analyze the relations between DPO and SFT, from the *gradient perspective*. We first define the logit of DPO loss function with and without the reference model as p_{dpo} and p'_{dpo}, respectively:

$$p_{\text{dpo}} = \log \frac{\pi_{\theta'}(\boldsymbol{y}_c|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_c|\boldsymbol{x})} - \log \frac{\pi_{\theta'}(\boldsymbol{y}_r|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_r|\boldsymbol{x})} \tag{16}$$

$$\begin{aligned} p'_{\text{dpo}} &= \log \pi_{\theta'}(\boldsymbol{y}_c|\boldsymbol{x}) - \log \pi_{\theta'}(\boldsymbol{y}_r|\boldsymbol{x}) && (17)\\ &= -(\mathcal{L}_{\text{sft}}(\boldsymbol{y}_c) - \mathcal{L}_{\text{sft}}(\boldsymbol{y}_r)) && (18)\end{aligned}$$

Then the standard DPO loss function changes to:

$$\mathcal{L}_d = -\log \sigma(\beta p_{\text{dpo}})\,. \tag{19}$$

If we take the partial derivation of the DPO loss function to the LLM parameter θ', we will obtain DPO gradient as:

$$\begin{aligned} \frac{\partial \mathcal{L}_d}{\partial \theta'} &= -\frac{1}{\beta p_{\text{dpo}}} \frac{\partial(\beta p_{\text{dpo}})}{\partial \theta'} && (20)\\ &= -\frac{1}{p_{\text{dpo}}} \frac{\partial p_{\text{dpo}}}{\partial \theta'} && (21)\\ &= -\frac{1}{p_{\text{dpo}}} \frac{\partial p'_{\text{dpo}}}{\partial \theta'} && (22)\\ &= \frac{1}{p_{\text{dpo}}} \left[\frac{\partial \mathcal{L}_{\text{sft}}(\boldsymbol{y}_c)}{\partial \theta'} - \frac{\partial \mathcal{L}_{\text{sft}}(\boldsymbol{y}_r)}{\partial \theta'} \right]. && (23)\end{aligned}$$

Note that during derivation, $\frac{\partial p_{\text{dpo}}}{\partial \theta'} = \frac{\partial p'_{\text{dpo}}}{\partial \theta'}$ since the reference model *do not* receive gradient. In the samewhile, the gradient of common SFT loss to the LLM parameter θ is represented as (we denote the parameter of SFT during continual learning as θ'):

$$\frac{\partial \mathcal{L}_{\text{sft}}(\boldsymbol{y})}{\partial \theta}\,. \tag{24}$$

That is, the DPO gradient is just a linear combination of two SFT gradient (positive response $\boldsymbol{y}_c$ and negative response $\boldsymbol{y}_r$), respectively, with just a dynamic scaling factor $\frac{1}{p'_{\text{dpo}}}$. This makes their optimization process similar, and can explain the inferior performance brought by the lack of negative supervision in SFT loss.

A.2. Gradient analysis

In this subsection, we analyze the gradient direction of DPO loss function towards the chosen response and reject responses, respectively. From Sec. 3, we know that:

$$t_1 = \frac{\pi_{\theta'}(\boldsymbol{y}_c|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_c|\boldsymbol{x})}, \quad t_2 = \frac{\pi_{\theta'}(\boldsymbol{y}_r|\boldsymbol{x})}{\pi_{\text{ref}}(\boldsymbol{y}_r|\boldsymbol{x})}\,, \tag{25}$$

$$\left| \frac{\partial \mathcal{L}}{\partial t_1} \Big/ \frac{\partial \mathcal{L}}{\partial t_2} \right| = t_2/t_1\,. \tag{26}$$

During the optimization process, t_1 tends to increase and t_2 tends to decrease (in order to optimize the final DPO loss), which make the division factor t_2/t_1 less than 1 [1]. As a result, the gradient will be biased towards t_2, the negative supervision in DPO loss function.

Here we want to clarify that this conclusion also trivially holds in our derivation where the reference term is omitted (e.g., the DPO loss function changes to $\mathcal{L}'_d$). In such cases, we can simply let:

$$t'_1 = \pi_{\theta'}(\boldsymbol{y}_c|\boldsymbol{x}), \quad t'_2 = \pi_{\theta'}(\boldsymbol{y}_r|\boldsymbol{x})\,, \tag{27}$$

and derive the same conclusion where the loss function $\mathcal{L}'_d$ will be biased towards optimizing t'_2. This will exactly match the basic form of our derivation in Sec. A.1 (how DPO loss is related to SFT loss without a reference term).

*Equal Contributions
†Corresponding Author

B. Experiment details

B.1. How to construct nSFT?

We now describe the constructed nSFT data in detail. Note that in OCRVQA, the newly constructed conversation length is 2 (manually constructed). In TextCaps and LLaVA-150k, the new constructed conversation length is 5 (GPT-4 is adopted).

OCRVQA. These dataset are set of book title pages (usually be viewed as object-centric images). As described in our experiment sections. We construct doubled Q-A (question-answer) pairs for each mistake that the model made, and appended these constructed pairs into the tail of the original GT conversation. During our implementation, we found that without the original GT conversation (only the negative constructed pairs are used for training), the training process can be unstable and the performance is unsatisfactory. However, this phenomenon does not apply to TextCaps and LLaVA-150k dataset, where the role of GT information is *not* necessary, as shown in Table 6 in main paper. We conjecture that the constructed conversation in OCRVQA contains few information, as it derives from the wrong answers that only contain one or few tokens that cannot fully describe the whole images.

TextCaps & LLaVA-150k. In these two dataset, we adopted GPT-4 to identify the error content in the rejected responses: *the model's original answer without temperature sampling*. The GT annotations are adopted as reference information of the image the guide the identification process. During experiment, we constructed 5 conversations per image, as we found more conversations will increase the overlapping probability with previous constructed data.

B.2. Experimental settings

Here we will primarily focus on two evaluation metrics.

CHAIR. This benchmark refers to the evaluation metrics adopted in [6]. It contains two core metrics:

$$\text{CHAIR}^i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|} \tag{28}$$

$$\text{CHAIR}^s = \frac{|\{\text{sentence with hallucinated objects}\}|}{|\{\text{all sentence}\}|} \tag{29}$$

In experiment, we randomly sample 1,000 COCO validation images, and pair each image with 5 questions of producing detailed captions utilized in LLaVA [3]. We then use the tools adopted in CHAIR evaluation process to match the object token in MS-COCO [2] and calculate the final results. Finally, we simply average the score obtained by CHAIR^i and CHAIR^s:

$$\text{CHAIR} = (\text{CHAIR}^i + \text{CHAIR}^s)/2 \tag{30}$$

ChatGPT For In-Domain Evaluation

```
# Prompt start
You are an AI visual assistant that analyzes a single image based on its ground-truth DETAILED IMAGE CAPTION.
You are also provided with a DESCRIPTION generated by a caption model.

# Task specific prompt
Evaluate the accuracy of the provided answer in relation to the image description and the question asked. Please address the following evaluation criteria specifically and provide a score for each criterion on a scale from 0 (poor) to 10 (excellent):

Instruction Following (0-10):
Assess whether the answer directly addresses the question asked and relates appropriately to the image description.

Detail Accuracy (0-10):
Check whether the details mentioned in the answer correctly reflect those described in the image, considering both visual elements and any text mentioned.

# Final Output
Your OUTPUT should be:
Instruction Following score: [score from 1-10, DO NOT ADD EXPLANATION]
Detail Accuracy score: [score from 1-10, DO NOT ADD EXPLANATION]

Your Task:
[task-specific output format]
```

Figure 7. Visualizations of prompt to evaluate in-domain results

GPT version	Alignment methods			
	baseline	GT-DPO	SeVa	nSFT
ChatGPT3.5	2.80	2.83	2.90	3.01
GPT-4	2.04	2.05	2.12	2.20

Table 8. The evaluation consistency introduced by GPT-3.5/4.0

MMHal. MMhal evaluation is proposed by [7], which aims to evaluate the hallucination ratios from range 0 to 6. Due to the quota limit in the company and heavy evaluation in our experiment, we adopted ChatGPT-3.5 (instead of GPT-4) to evaluate the generated responses. As a result, the number evaluated by ChatGPT-3.5 in Table 1 (in the main paper) and and Table 8 (in the appendix) is slightly higher than the results in GPT-4. To verify whether the version changes will influence the final comparison, we conduct a short experiment, by involving GPT's both version. As shown in Table 8, the evaluation score by GPT-3.5/4.0 is different (e.g., the GPT-3.5's is generally higher). However, we have observed a same growing trend with regard to the 4 methods, showing that a replacement of 3.5 version is valid.

In-domain evaluation. Here we explain our in-domain evaluation results in detail, which was done in Table 3 in our main paper. The motivation of this experiment is to ablate the affect of multimodal RLHF as well as our SFT method.

Since common multimodal evaluation benchmarks [4, 9] usually possess a large domain shift between training and evaluation dataset, which can induce potential noise that the

core conclusion might not be emerged. For example, the LLaVA-1.5 continually trained on OCRVQA could probably behave well on books title page recognition, but not specialized for benchmarks that relies on image reasoning (e.g., MMVet). In this paper, we first train LLaVA-1.5-7B on 5k subset of OCRVQA, TextCaps and LLaVA-150k, respectively. Then we evaluate each model on its own data source using 500 instances in a *held-out* manner. The evaluation metrics was the instruction following ability (whether the answers clearly resolves the questions asked) and detailed accuracy (whether the answer clearly reflect the content in the image). As shown in Table 3 of our main paper, all methods jointly improve the LLaVA-1.5-7B results, showing that this in-domain dataset design is *indeed* valid. Interestingly, we found that the worst accuracy score is relatively higher in RLHF paradigm (e.g., GT-DPO and SeVa obtains the best score in ACC_{10}^{w}), while the metric ACC_{10}^{b} lean towards our nSFT approach. We conjecture this can be attributed to different training paradigm.

In SFT, although we integrate the negative supervision into the final nSFT loss, the overall optimization objective can be *positive*. However, during DPO optimization process, the negative response is deeply integrated into DPO's 'logit'. As a result, nSFT implicitly avoid making unpreferred answers, while DPO behave in an explicit way.

B.3. More data scale comparison

Please refer to Table 9 for a whole demonstration. Here we list the results by applying GT-DPO, SeVa [10], SIMA [8], Cont. SFT and nSFT to 3 different databases: OCRVQA [5], TextCaps and LLaVA-150k. The dataset scale was chosen at 5k and 10k for each specific data source. Although more data scale might further improve the VLM's comprehension ability, we found that this effect becomes weaker when the data scale are beyond 15k-20k. We guess this is due to the data diversity issue. That is, the data for alignment has partially been seen in LLaVA-1.5 SFT stage, which calls for more diverse pretraining or SFT databases. A similar phenomenon can be observed in previous preference alignment articles, where they mostly adopt less than 20k datasource during preference alignment.

To study how data scale affect alignment is interesting. However, this is beyond the scope of this paper, and we will leave this as future work.

C. Visualizations

C.1. Prompt of in-domain evaluation

We provide the prompt to evaluate the in-domain results in Fig. 7. This prompt is sent to GPT-4 to evaluate both the instruction following score, as well as the detailed accuracy score, and the results are shown in Table 3 in main paper.

C.2. Prompt of nSFT

Please refer to Fig. 9 for the total prompt to construct the negative supervision. This prompt also contains the element of our vision error codebook, which are highlighted with bold phase. Fig. 9 is best viewed in color.

C.3. Wordcloud visualization

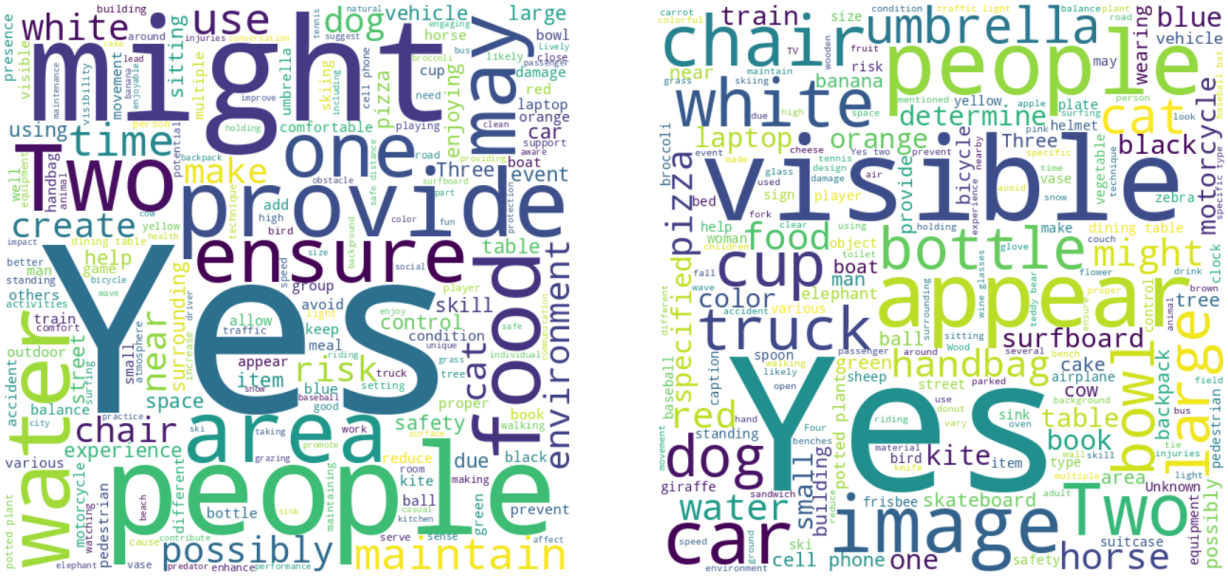
Please refer to Fig. 8a-8b for the word visualization of our nSFT without and with our vision error codebook. The conversation data are sourced from a 5k subset of LLaVA-150k. As shown in Fig. 8, when the conversation is *not* guided by the numerated vision error, it will mostly hinges on non-object phrase, such as 'might', 'provide', 'ensure'. When we force LLM to focus on specific object type, more object related word emerged, such as 'truck', 'cup', 'bottle' and 'chair'. This has emphasized the importance of the vision error codebook.

C.4. OCRVQA, TextCaps and LLaVA

Please refer to Fig. 10 to see alignment dataset. Note that the OCRVQA dataset has the shortest response token length, where the responses contains only single word or phrase. The annotations of TextCaps is longer, similar to the length of MS-COCO [2] captions. The captions of LLaVA-150k is much longer, which is constructed by GPT-4 with the original annotations in MS-COCO.

C.5. Negative supervision of nSFT

Please refer to Fig. 11 for more examples of nSFT. As shown in these figure, models are usually tend to make wrong existence of object (also called image hallucinations). Our nSFT re-inforce these false information by asking the model about the image content. In practical, we balance the 'Yes' and 'No' ratio in the conversation by randomly erase some 'No' answers in the constructed conversations. Empirically this could lead to a more robust result.



(a) w/o vision error codebook

(b) w/ vision error codebook

Figure 8. Visualization of the constructed conversation w/o and w/ our vision error codebook, derived from a 5k subset of LLaVA-150k.

ChatGPT For Negative Construction

Prompt start
You are an AI visual assistant that analyzes a single image based on its ground-truth DETAILED IMAGE CAPTION.
You are also provided with a DESCRIPTION generated by a caption model.

Task specific prompt
Task: Comparing the generated DESCRIPTION and the ground-truth DETAILED IMAGE CAPTION, indentify if there are hallucinations in the generated DESCRIPTION. If hallucinations exists, then Design one conversation between you (as "GPT") and a person (as "User") who asks about the image. Each conversation must include five question-answer pairs. The questions and answers should be ***logically connected***. The aim of the conversation should mostly focus on hallucinations correction.

 # The Vision-Error Codebook ----Instance Level
Consider the following hallucinations when inspecting the generated sentences:
First, in the instance level:
1. Object Identity: whether if the object class or category is correctly classified
2. Object attribute: whether if the object shape, color or counting is correctly classified
3. Object actions: whether if the that action the object is doing is correct

 # The Vision-Error Codebook ----Image Level
Second, in the image level
1. Object locations: whether if the object is correctly located in the image as described
2. Relative positions between objects: whether the position location between two described object is correct
3. Background knowledge of objects: whether the background knowledge or surroundings are correct
4. Events in the image: identify if the whole image's atmosphere or fully description of the scene is correct
5. Event planning based on objects: identify whether the prediction of the objects movement is correct.
6. Reasoning: identify whether the image-level reasoning is mostly consistent with the GT description.

More general guidelines for generating topics:
Conversation Guidelines:
1. For each topic, Ask a variety of questions and provide corresponding answers.
2. Only include questions that have definite answers.
3. Answer as if you are directly looking at the image.
4. Responses should be concise, within 20 tokens.
5. Do not mention that the information comes from a description.
6. Do not aways ask question that starts with "Is" or "Are".

If there are no hallucinations, just end this chat.

Final output:
OUTPUT Format (If hallucinations exists):
Hallucinations:
[hallucination-specific output format]

Constructed Topic1: [Your chosen topic]
[topic-specific output format]

Figure 9. The prompt template used for error identification and conversation reconstruction.

Align. Data	Method	Align. tax				Comprehension				Hallucinations			
		SQA	GQA	VQA^T	total	MMVet	MME	MMB	total	POPE	$CHAIR^*$	MMHal	total
OCRVQA-5k	baseline	66.8	62.0	58.0	(+0.0)	30.5	1510	64.3	(+0.0)	85.9	32.0	2.80	(+0.0)
	GT-DPO	67.6	61.5	58.2	(+0.5)	32.3	1339	63.9	(+1.3)	84.4	32.0	2.91	(+0.3)
	SeVa	67.8	61.9	57.9	(+0.8)	32.2	1511	64.6	(+2.0)	86.5	28.5	2.92	(+6.1)
	SIMA	67.7	61.8	58.2	(+0.9)	32.8	1460	64.5	(+2.5)	85.8	30.5	2.90	(+3.1)
	Cont. SFT	67.7	61.8	57.3	(+0.0)	30.8	1453	63.9	(-0.1)	86.1	32.0	2.83	(+0.7)
	nSFT	**68.0**	**62.0**	**58.1**	(+1.3)	**32.6**	**1512**	**64.8**	(+2.6)	**86.7**	**28.2**	**2.93**	(+6.8)
OCRVQA-10k	baseline	66.8	62.0	58.0	(+0.0)	30.5	1510	64.3	(+0.0)	85.9	32.0	2.80	(+0.0)
	GT-DPO	67.8	61.4	57.7	(+0.1)	32.5	1412	63.9	(+1.6)	84.3	31.5	2.90	(+0.6)
	SeVa	67.6	62.0	57.5	(+0.3)	32.5	1502	64.9	(+2.6)	86.6	27.3	3.00	(+8.7)
	SIMA	68.0	61.9	58.2	(+1.3)	32.5	1486	64.8	(+2.5)	86.2	29.4	2.93	(+5.1)
	Cont. SFT	67.9	61.7	56.9	(-0.3)	33.3	1490	64.5	(+3.0)	87.0	34.0	2.76	(-1.6)
	nSFT	**68.1**	**62.0**	**58.1**	(+1.4)	**34.0**	**1515**	**64.9**	(+4.1)	**87.1**	**26.5**	**2.93**	(+8.9)
TextCaps-5K	baseline	66.8	62.0	58.0	(+0.0)	30.5	1510	64.3	(+0.0)	85.9	32.0	2.80	(+0.0)
	GT-DPO	67.7	61.9	57.8	(+0.6)	33.0	1503	64.0	(+2.2)	86.3	29.6	2.83	(+3.3)
	SeVa	67.7	61.8	57.8	(+0.5)	32.8	1498	65.0	(+3.0)	86.0	27.8	2.90	(+6.0)
	SIMA	68.0	62.0	58.1	(+1.3)	32.8	1477	65.0	(+3.0)	85.9	29.3	2.90	(+4.4)
	Cont. SFT	67.3	61.5	56.6	(-1.4)	32.5	1518	63.6	(+1.3)	86.3	31.5	2.91	(+2.7)
	nSFT	**68.1**	**62.0**	**58.1**	(+1.4)	**33.0**	**1515**	**64.8**	(+3.0)	**86.8**	**27.5**	**2.91**	(+7.2)
TextCaps-10K	baseline	66.8	62.0	58.0	(+0.0)	30.5	1510	64.3	(+0.0)	85.9	32.0	2.80	(+0.0)
	GT-DPO	68.0	61.7	57.5	(+0.4)	34.2	1500	64.2	(+3.6)	86.5	29.2	2.83	(+3.9)
	SeVa	68.1	61.7	57.8	(+0.8)	34.6	1480	65.0	(+4.8)	86.3	26.3	2.90	(+7.8)
	SIMA	68.0	62.1	58.0	(+1.3)	32.2	1473	64.9	(+2.3)	85.9	27.6	2.87	(+5.6)
	Cont. SFT	66.9	61.3	56.6	(-2.0)	31.0	1520	64.4	(+0.6)	86.3	30.5	2.83	(+2.4)
	nSFT	**68.4**	**62.3**	**58.2**	(+2.1)	**33.7**	**1521**	**65.3**	(+4.2)	**87.2**	**26.2**	**2.97**	(+9.9)
LLaVAData-5k	baseline	66.8	62.0	58.0	(+0.0)	30.5	1510	64.3	(+0.0)	85.9	32.0	2.80	(+0.0)
	GT-DPO	67.9	61.9	58.0	(+1.0)	32.6	1512	64.1	(+1.9)	86.1	28.5	2.94	(+6.0)
	SeVa	67.6	61.7	58.2	(+0.7)	32.6	1505	64.7	(+2.5)	86.0	28.3	2.93	(+6.0)
	SIMA	67.9	**62.1**	58.2	(+1.4)	32.1	1505	64.5	(+1.8)	86.3	28.1	2.95	(+6.8)
	Cont. SFT	67.5	61.0	56.7	(-1.6)	31.0	1475	63.5	(-0.3)	85.6	29.5	2.82	(+2.5)
	nSFT	**68.2**	61.9	**58.2**	(+1.5)	**33.0**	**1533**	**65.0**	(+3.2)	**86.5**	**27.2**	**2.99**	(+8.6)
LLaVAData-10k	baseline	66.8	62.0	58.0	(+0.0)	30.5	1510	64.3	(+0.0)	85.9	32.0	2.80	(+0.0)
	GT-DPO	68.1	61.6	57.6	(+0.5)	33.9	1497	63.9	(+3.0)	85.9	30.7	2.80	(+1.3)
	SeVa	67.5	61.4	58.0	(+0.1)	32.5	1490	64.7	(+2.4)	85.6	28.2	2.94	(+5.8)
	SIMA	67.9	62.2	58.2	(+1.5)	32.1	1511	64.9	(+2.2)	86.9	26.2	2.97	(+9.6)
	Cont. SFT	67.1	60.9	57.0	(-1.8)	31.2	1480	64.0	(+0.4)	86.3	29.1	2.91	(+5.1)
	nSFT	**68.4**	**62.3**	**58.4**	(+2.3)	**34.2**	**1550**	**65.2**	(+4.6)	**87.4**	**25.4**	**3.02**	(+11.8)

Table 9. Nine benchmark results by applying 5 continual learning methods. We list the outcomes using alignment data of 5k and 10k. In our main paper, we already conduct such an experiment (*cf*. experimental settings in main paper) with data scale of 5k and 10k for each datasource (OCRVQA, TextCaps and LLaVA-150k). However, due the constraints of format and space, we only show 10k results in Table 1 in our main paper, and provide all these results here to help the readers get a better understanding.

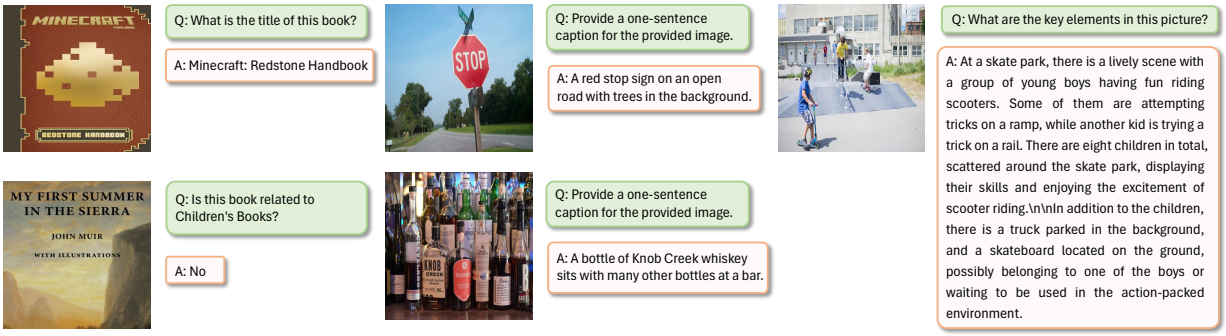


Figure 10. Visualization of groundtruth (GT) captions in OCRVQA, TextCaps and LLaVA-150k dataset. The GT length from left to right has seen a steady growth. For clarity, we omit the suffix of TextCaps question (e.g., 'Reference OCR token').

How can one guess the purpose of this group photo?

To guess the purpose of this group photo, one can observe that it consists of several people standing in front of a wall. Each person is positioned in front of a different picture, which is either a large photograph or a collage of smaller photos. These pictures may represent various themes, such as family, friends, or colleagues. Given the diversity of the individuals and the presence of the photographs or collage, it is likely that this group photo captures a gathering, event, or celebration where people have come together to share memories, show support, or appreciate their connections. The photo could be used to commemorate a special occasion, reunion, or milestone, and the presence of the yacht in the photo might hint at a more luxurious or significant event.

User: What is the actual background in this photo?
GPT: The background is a ***tennis court with a net****.*

User: Are there any pictures or collages on the wall?
GPT: ***No****, there are* ***no pictures or collages*** *on a wall.*

User: What are the people in the photo holding?
GPT: They are holding ***tennis racquets.***

User: Is there a yacht in this photo?
GPT: ***No, there is no yacht in the photo.***

User: What is the likely purpose of this group photo?
GPT: It seems to ***capture camaraderie*** *among* ***tennis players or teammates.***

Describe the following image.

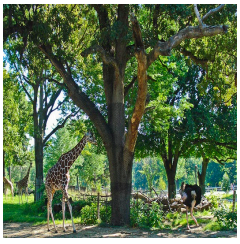
The image depicts a large group of people gathered together, forming a crowd in a small, confined, and busy space. The crowd in the presented image is composed of a very diverse mix of people, with individuals located in various positions within the image. Some people are standing closer to the edges of the image frame, while two people are more centrally located, trying to walk across the road, The whole image creating a busy and lively atmosphere. (missing information about other vehicles in the image, like cars and truck, etc)

User: How many people are in the image?
GPT: There are ***three people*** *visible in the center,* ***two in the*** *left edge of the figure*

User: Is these three people trying to walk across the street?
GPT: ***No, they are standing, perhaps talking about traffic.***

User: What about the atmosphere? Is it busy and lively?
GPT: The atmosphere ***appears to be ordinary, not very busy or lively.***

User: Are there any other notable details in the image?
GPT: Yes, there ***are several vehicles, traffic lights, and a cell phone*** *visible.*



What factors contribute to the unique interaction between the ostrich and the giraffes in the image?

In the image, the ostrich and the giraffes are seen in a shared environment, likely at a zoo. The presence of multiple trees and a wooden fence provides a natural setting for the animals. The unique interaction between the ostrich and the giraffes in this setting is attributed to the fact that ostriches and giraffes are not typically found together in their natural habitat. Giraffes are typically native to the savannahs of Africa, while ostriches are native to the African grasslands and savannahs, but not necessarily in the same locations. Additionally, ostriches and giraffes are different species with distinct behaviors, which could result in different social interactions between the two animals.

User: Are the ostrich and giraffes in this image located at a zoo?
GPT: ***No, the image shows*** *them in their* ***natural habitat, not a zoo.***

User: Do ostriches and giraffes typically coexist in the same natural environment?
GPT: ***Yes, both species are native*** *to Africa and may coexist in savannahs, forests, or grasslands.*

User: Are there any indications in the image that suggest they are in a captivity or enclosed space?
GPT: ***No, the presence of large trees and open grasslands indicates a natural, non-captive environment.***

User: So, the interaction between the ostrich and giraffes is not rare or unusual?
GPT: ***Correct, their interaction is a result of their shared habitat*** *and coexistence in the same natural environment.*

Figure 11. Random sampled cases of our negative constructed supervision. The middle part are the model's original responses to the question, and the right part shows the reconstructed conversations (our nSFT sample). Error content are red color coded.

References

[1] Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024. 1

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[3] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[4] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 2

[5] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 3

[6] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 2

[7] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2

[8] Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Furong Huang, et al. Enhancing visual-language modality alignment in large vision language models via self-improvement. *arXiv preprint arXiv:2405.15973*, 2024. 3

[9] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 2

[10] Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. Self-supervised visual preference alignment. *arXiv preprint arXiv:2404.10501*, 2024. 3