

# Supplementary Materials: Learning by Correction: Efficient Tuning Task for Zero-Shot Generative Vision-Language Reasoning

Type	Content
Task Instruction	Check the caption: “{}” Check the caption according to the image: “{}” Based on the image, please correct the caption: “{}”
replace Answer	“{}” should be “{}” “{}” could be “{}” “{}” is “{}” “{}” actually is “{}”
swap Answer	“{}” and “{}” are swapped “{}” and “{}” need to switch “{}” and “{}” should exchange positions “{}” and “{}” need to be swapped

Table 1. **The instruction prompts and answer templates for the ICCC task.** The “{}” in each task instruction is a placeholder for augmented caption; the “{}” in each answer template is the mismatched concept and the correct one, respectively.

We present a comprehensive breakdown of the task instruction and ground truth answer templates for the ICCC task, the essential components in constructing the caption correction task for model training. Additionally, we offer an extended set of data samples, showcasing the ICCC task through its construction process.

## A. Templates for Task Construction

The comprehensive template for the task construction (Section 4.3) is illustrated in Table 1. To mitigate potential overfitting to the language bias inherent in the correction task, we employ various prompts and answer formats during data construction, as informed by empirical experiments. Specifically, for the BLIP-2 VLM models [2], the shortest instruction, “check the caption,” is utilized. In the case of instructBLIP [1] coupled with the Vicuna Large Language Model (LLM), multiple instructions are employed during tuning to prevent the degradation of generalization. With the different choice of these templates, it shows considerable effectiveness in training empirically.

## B. Example of Generated Samples

To vividly illustrate the data samples generated by our ICCC framework, we present examples of mismatched captions in Fig. 1. Specifically, we showcase five examples, highlighting all concept types and the augmentation methods applied.

As illustrated in Fig. 1, we parse the original caption (Ori. Cap) into a dependency tree (Dep. Tree), enabling the extraction of linguistic units associated with each concept type. Subsequently, specific types of linguistic units undergo replacement (replace), where affected elements are highlighted in **red**, and swapping (swap) operations, with modified components emphasized in **orange**. The operating results are the mismatched captions (Mismatched Cap.) with perturbed linguistic units, serving as input for VLMs to correct in the ICCC task.

Given that each caption involves multiple linguistic units for augmentation, the task constructor adeptly produces a variety of captions with mismatched concepts. Consequently, this ICCC task significantly enhances the image-conditioned text generation of VLMs, focusing on a general semantic concept, thereby effectively improving zero-shot capabilities across various downstream Vision-Language tasks.

## C. More Experiment Results

### C.1. Performance on Other VLMs

Our choice of VLMs is representative of the latest generative VLM architectures, which share similar structures by incorporating an adaptor to bridge the visual encoder and large language model. Notably, at the time of submission, InstructBLIP performed as the SoTA among the published generative VLMs, surpassing the zero-shot performance of Flamingo. Therefore, we selected InstructBLIP as a representative SoTA model for our experiments in the paper. To demonstrate the efficacy of our approach with the latest and different models, we conducted experiments on the latest accepted SoTA method, LLaVA, published on NeurIPS 2023. We follow the evaluation setup in the latest version, LLaVA-1.5, but exclude datasets used during instruction


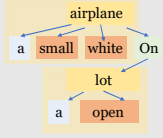

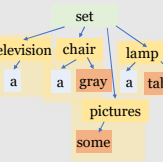


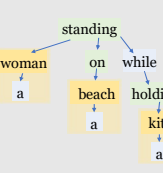
Image	Ori. Cap.	Dep. Tree	Type	Mismatched Cap.
	A small white airplane on a open lot.		entity phrase	- <b>An infant</b> on a open lot. - <b>A open lot</b> on <b>a small white airplane</b> .
	A lamp shines on the nightstand beside the bed.		predicate phrase	- A lamp <b>stuck on</b> the nightstand beside the bed. - A lamp <b>beside</b> the nightstand <b>shines on</b> the bed.
	A television set a gray chair, a table lamp and some pictures.		attribute phrase	- A television set a gray chair, <b>a shelving</b> lamp and some picture - A television set <b>some</b> chair, a table lamp and <b>a gray</b> pictures.
	Food stands are set up on the platform of a train stop.		noun word	- Food stands are set up on the platform of a <b>Mexican</b> stop. - <b>Train</b> stands are set up on the platform of a <b>food</b> stop. - Food stands are set up on the <b>train</b> of a <b>platform</b> stop.
	A woman standing on a beach while holding a kite.		verb word	- A woman <b>perform</b> on a beach while holding a kite. - A woman <b>holding</b> on a beach while <b>standing</b> a kite.

Figure 1. **Examples of constructed mismatched captions categorized by modification concept type.** The linguistic units operated by replace are highlighted in **red**, while the language units operated by swap are highlighted in **orange**.

tuning, such as GQA and VQAv2, to maintain a zero-shot setting. As shown in Tab. 2, our method outperforms the baseline model on two evaluation benchmarks by keeping consistency with the method setup outlined in the main text, even without adjusting hyperparameters. This underscores the complementarity of our approach with existing instruction tuning methods.

LLaVA	ScienceQA-IMG	MM-VET
Vicuna-7B	66.8	30.8
Vicuna-7B w/ ICC	<b>67.7</b>	<b>31.7</b>

Table 2. The zero-shot evaluation results on LLaVA-7B with our tuning method.

## C.2. Hallucinations and VSR

We evaluate the effect of the method on hallucinations by CHAIR score [3] on the COCO dataset. The results indicate that our approach proves beneficial in addressing caption hallucinations. On the VSR dataset, our method also leads to further improvements on BLIP-2. The hallucinations of VLMs are a very interesting research question, and we will conduct further exploration in future work.

## References

- [1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1
- [2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip:

BLIP-2	COCO		VSR
	CHAIRs	CHAIRi	
OPT 2.7B	3.2	2.3	<b>48.28</b>
OPT 2.7B w/ ICC	<b>3.0</b>	<b>2.1</b>	47.79
OPT 6.7B	3.2	2.2	48.53
OPT 6.7B w/ ICC	<b>3.0</b>	<b>2.1</b>	<b>51.55</b>
FlanT5XL	2.3	1.6	63.42
FlanT5XL w/ ICC	<b>2.1</b>	<b>1.5</b>	<b>64.24</b>

Table 3. The zero-shot evaluation results on BLIP-2 on CHAIR metrics for image caption hallucinations and performance of VSR.

Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. [1](#)

- [3] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. [2](#)