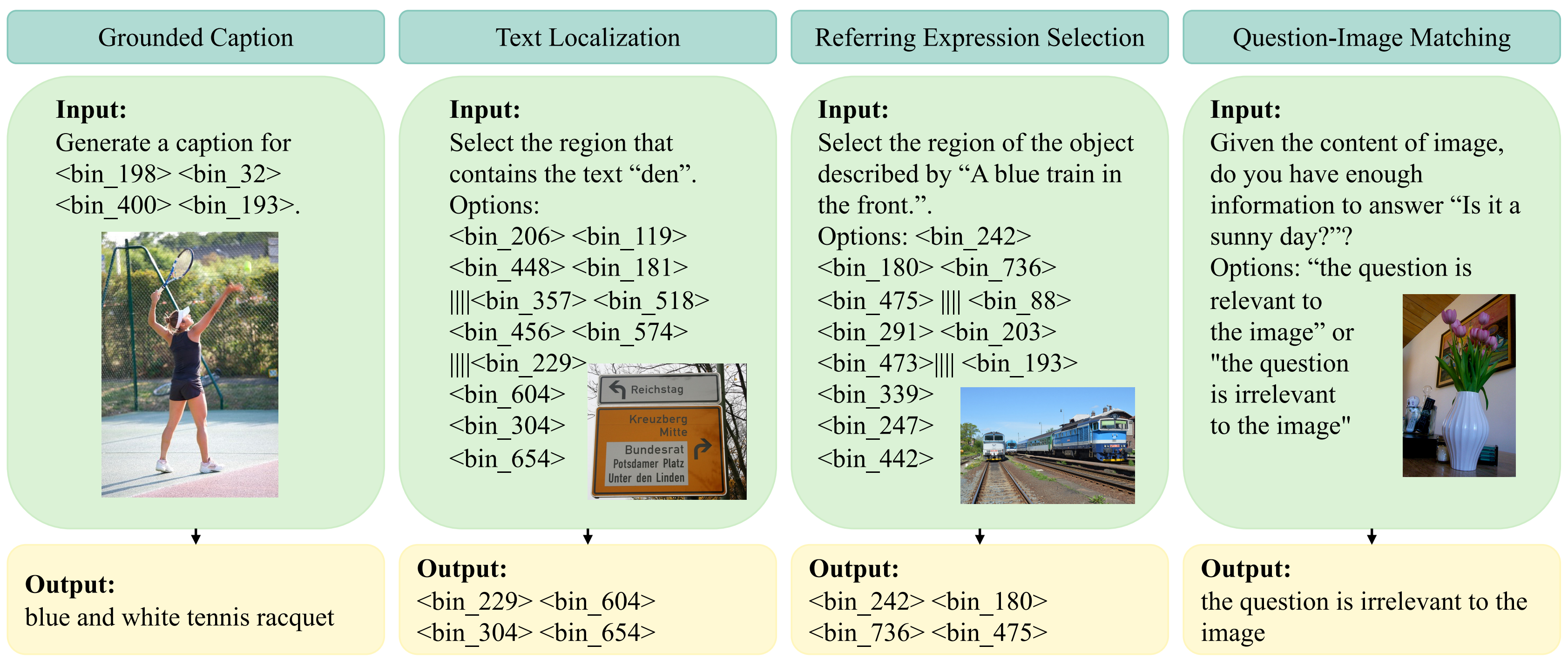
MultiInstruct: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning

Zhiyang Xu*, Ying Shen*, Lifu Huang

Computer Science Department
Virginia Tech

Motivation

Instruction tuning has shown promising zero-shot performance on various natural language processing tasks. However, it has yet to be explored for vision and multimodal tasks. In this work, we introduce MultiInstruct, the first multimodal instruction tuning benchmark dataset.



Grounded Caption

Input:
Generate a caption for <bin_198> <bin_32> <bin_400> <bin_193>.

Output:
blue and white tennis racquet

Text Localization

Input:
Select the region that contains the text "den".
Options:
<bin_206> <bin_119>
<bin_448> <bin_181>
||||<bin_357> <bin_518>
<bin_456> <bin_574>
||||<bin_229>
<bin_604>
<bin_304>
<bin_654>

Output:
<bin_229> <bin_604>
<bin_304> <bin_654>

Referring Expression Selection

Input:
Select the region of the object described by "A blue train in the front.".
Options: <bin_242>
<bin_180> <bin_736>
<bin_475> |||| <bin_88>
<bin_291> <bin_203>
<bin_473>|||| <bin_193>
<bin_339>
<bin_247>
<bin_442>

Output:
<bin_242> <bin_180>
<bin_736> <bin_475>

Question-Image Matching

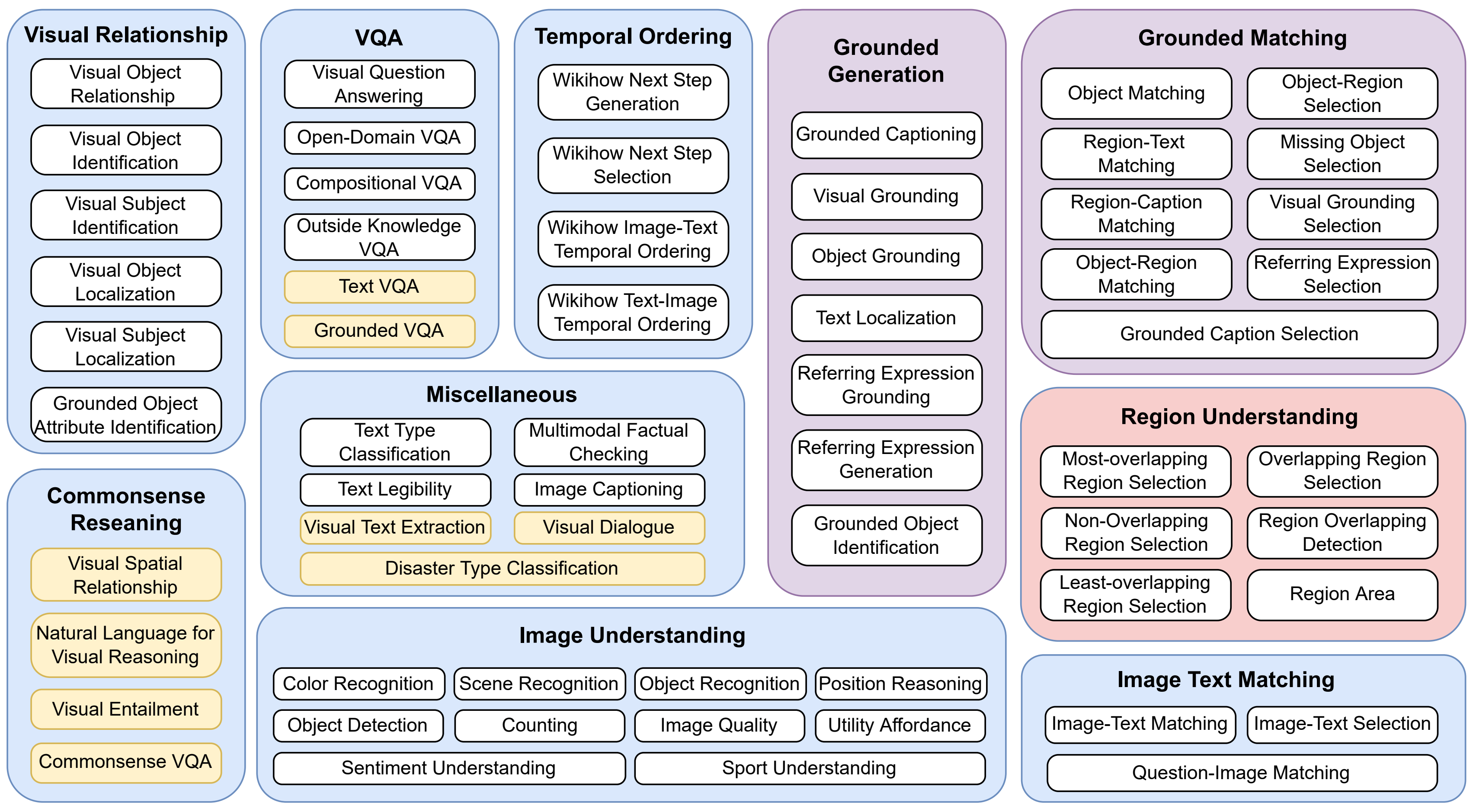
Input:
Given the content of image, do you have enough information to answer "Is it a sunny day?"?
Options: "the question is relevant to the image" or "the question is irrelevant to the image"

Output:
the question is irrelevant to the image

Figure 1. **Example Instances from MultiInstruct for Four Tasks.**

MultiInstruct

MultiInstruct, the first multimodal instruction tuning benchmark dataset, **consists of 62 diverse multi-modal tasks** in a unified seq-to-seq format covering 10 broad categories. The tasks are derived from 21 existing open-source datasets and **each task is equipped with 5 expert-written instructions**.



Visual Relationship: Visual Object Relationship; Visual Object Identification; Visual Subject Identification; Visual Object Localization; Visual Subject Localization; Grounded Object Attribute Identification

VQA: Visual Question Answering; Open-Domain VQA; Compositional VQA; Outside Knowledge VQA; Text VQA; Grounded VQA

Temporal Ordering: Wikihow Next Step Generation; Wikihow Next Step Selection; Wikihow Image-Text Temporal Ordering; Wikihow Text-Image Temporal Ordering

Grounded Generation: Grounded Captioning; Visual Grounding; Object Grounding; Text Localization; Referring Expression Grounding; Referring Expression Generation; Grounded Object Identification

Grounded Matching: Object Matching; Object-Region Selection; Region-Text Matching; Missing Object Selection; Region-Caption Matching; Visual Grounding Selection; Object-Region Matching; Referring Expression Selection; Grounded Caption Selection

Commonsense Reseaning: Visual Spatial Relationship; Natural Language for Visual Reasoning; Visual Entailment; Commonsense VQA

Miscellaneous: Text Type Classification; Multimodal Factual Checking; Text Legibility; Image Captioning; Visual Text Extraction; Visual Dialogue; Disaster Type Classification

Region Understanding: Most-overlapping Region Selection; Overlapping Region Selection; Non-Overlapping Region Selection; Region Overlapping Detection; Least-overlapping Region Selection; Region Area

Image Understanding: Color Recognition; Scene Recognition; Object Recognition; Position Reasoning; Object Detection; Counting; Image Quality; Utility Affordance; Sentiment Understanding; Sport Understanding

Image Text Matching: Image-Text Matching; Image-Text Selection; Question-Image Matching

Figure 2. **Task Groups Included in MultiInstruct.** The yellow boxes represent tasks used for evaluation, while the white boxes indicate tasks used for training.

Multimodal Instruction Tuning

- We finetune OFA on 53 tasks from the MultiInstruct dataset and evaluate it on 9 unseen tasks.
- We explore multiple transfer learning strategies to leverage the large-scale Natural Instructions dataset and propose the following strategies.

  Multimodal Instruction Tuning ($\text{OFA}_{\text{MultiInstruct}}$) finetunes OFA on MultiInstruct.

  Mixed Instruction Tuning ($\text{OFA}_{\text{MixedInstruct}}$) finetunes OFA on the mixture of MultiInstruct and Natural Instructions.

  Sequential Instruction Tuning ($\text{OFA}_{\text{SeqInstruct}}$) first finetunes OFA on Natural Instructions and then on MultiInstruct.
- We also design a new evaluation metric – *Sensitivity*, to evaluate how sensitive the model is to the variety of instructions.

Effectiveness of Instruction Tuning on MultiInstruct

- Experimental results demonstrate strong zero-shot performance on various unseen multimodal tasks and the benefit of transfer learning from a text-only instruction dataset.
- Our results indicate that fine-tuning the model on a diverse set of tasks and instructions leads to a reduced sensitivity to variations in instructions for each task.

Model	Commonsense VQA				Visual Entailment		Visual Spatial Reasoning		NLVR	
	RougeL		ACC		ACC		ACC		ACC	
	Max	Avg ± Std	Max	Avg ± Std	Max	Avg± Std	Max	Avg± Std	Max	Avg± Std
OFA	17.93	14.97 ± 4.30	0.73	0.40 ±0.29	49.99	41.86 ± 10.99	54.99	35.29 ± 22.21	56.06	52.10 ± 3.35
$\text{OFA}_{\text{TaskName}}$	48.99	-	29.01	-	55.70	-	53.76	-	55.35	-
$\text{OFA}_{\text{MultiInstruct}}$	**52.01**	**50.60** ± 1.12	**33.01**	31.17 ± 1.59	**55.96**	**55.06** ±0.76	**55.81**	**53.90** ±1.38	56.97	56.18 ± 0.95
Transfer Learning from Natural Instructions										
$\text{OFA}_{\text{NaturalInstruct}}$	27.15	14.99 ± 9.12	7.35	2.04 ± 3.01	33.28	14.86 ± 16.68	51.44	36.44 ± 20.72	56.06	35.98 ± 21.64
$\text{OFA}_{\text{MixedInstruct}}$	50.40	49.34 ± 1.04	31.31	30.27 ± 0.94	54.63	53.74 ± 0.97	55.13	52.61 ± 1.64	56.67	55.96 ± 0.48
$\text{OFA}_{\text{SeqInstruct}}$	50.93	50.07 ± 1.07	32.28	**31.23** ± 1.09	53.66	52.98 ± 0.56	54.86	53.11 ± 1.45	**57.58**	**56.63** ± 0.66

Table 1. **Zero-shot Performance on Multimodal Commonsense Reasoning.** The best performance is in **bold**.

Model	Text VQA				Grounded VQA	
	RougeL		ACC		ACC	
	Max	Avg ± Std	Max	Avg ± Std	Max	Avg ± Std
OFA	15.21	9.30 ± 5.42	12.32	7.96 ± 4.20	0.02	0.00 ± 0.01
$\text{OFA}_{\text{TaskName}}$	23.80	-	20.02	-	0.00	-
$\text{OFA}_{\text{MultiInstruct}}$	**27.22**	26.46 ± 0.83	**22.70**	22.00 ± 0.70	**64.32**	47.22 ± 23.08
Transfer Learning from Natural Instructions						
$\text{OFA}_{\text{NaturalInstruct}}$	5.59	5.40 ± 0.24	4.18	3.78 ± 0.27	0.00	0.00 ± 0.00
$\text{OFA}_{\text{MixedInstruct}}$	24.15	23.67 ± 0.47	20.10	19.62 ± 0.46	63.79	**54.99** ± 18.16
$\text{OFA}_{\text{SeqInstruct}}$	27.03	**26.67** ± 0.47	22.64	**22.28** ± 0.46	64.19	54.46 ± 15.96

Table 2. **Zero-shot Performance on Question Answering datasets.** The best performance is in **bold**.

Impact of Increasing Multimodal Instruction Task Clusters

As we increase the number of task clusters, we observe an improvement in both the mean and maximum aggregated performance and a decrease in *sensitivity*.

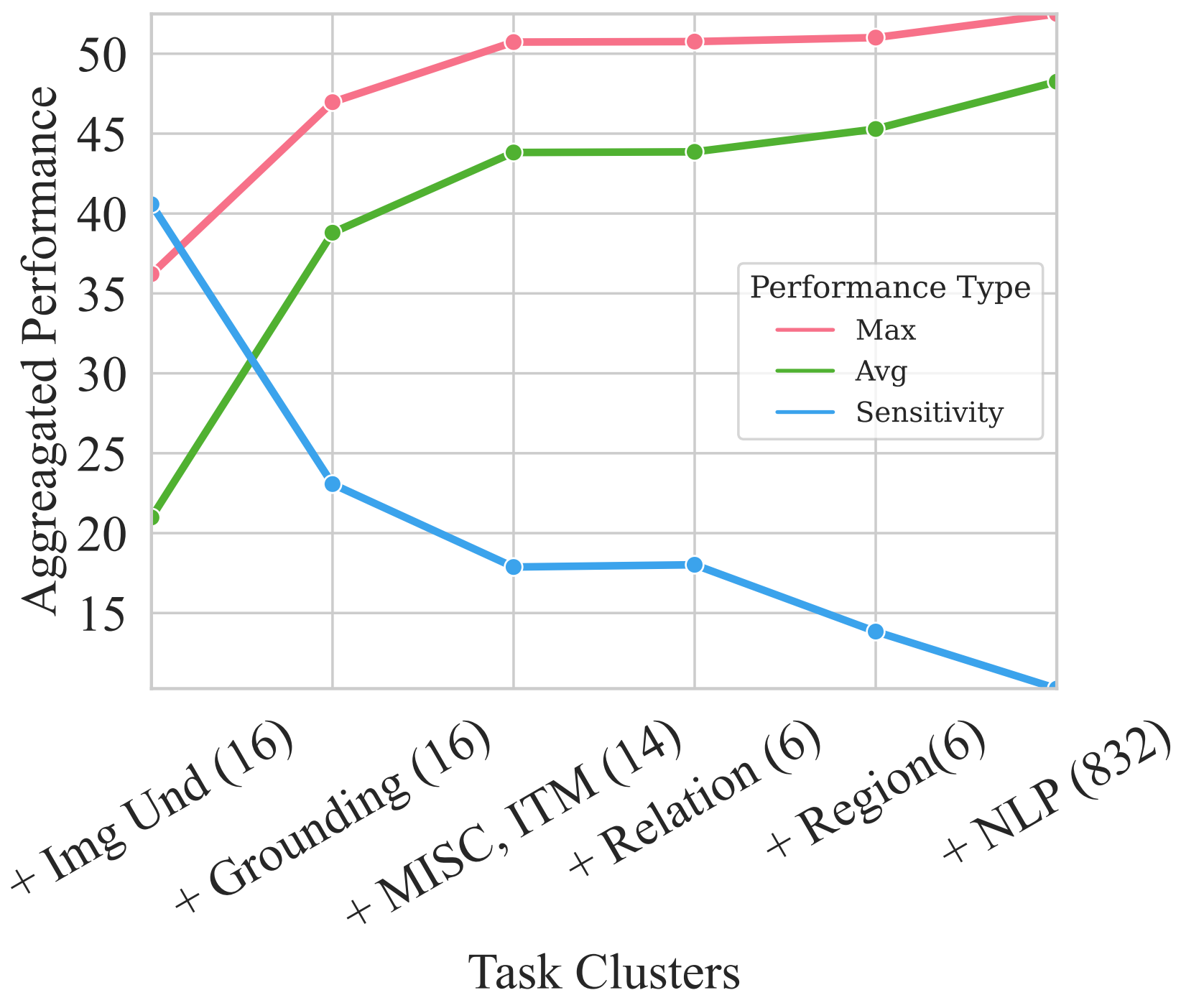


Figure 3. **Model Performance as the Number of Multimodal Instruction Task Clusters Increases.** The number in the parenthesis of each cluster denotes the number of tasks.

Effect of Diverse Instructions on Instruction Tuning

OFA finetuned on 5 instructions achieves much **higher aggregated performance** on all evaluation tasks and shows **lower sensitivity**.

# of Instructions	Aggregated Performance ↑	*Sensitivity* ↓
1 Instruction	42.81	24.62
5 Instructions	**47.82**	**10.45**

Table 3. **Effect of Different Number of Instructions.** Performance of $\text{OFA}_{\text{MultiInstruct}}$ finetuned on different numbers of instructions.

Effect of Fine-tuning Strategies on Model Sensitivity

- Instruction tuning on MultiInstruct can significantly reduce the sensitivity of OFA.
- Transfer learning from Natural Instructions dataset can further reduce the sensitivity of the model.

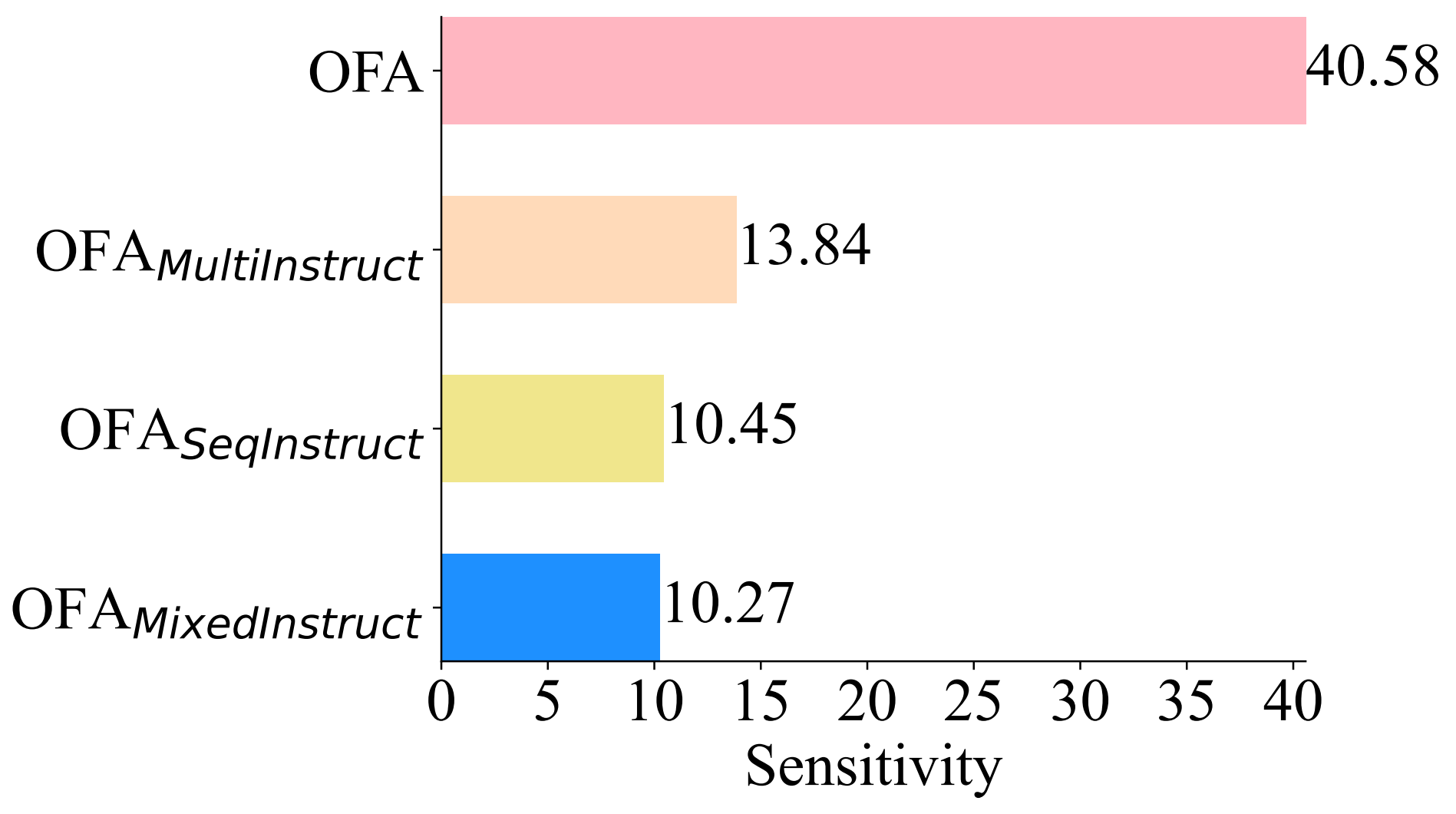


Figure 4. **Model *Sensitivity* on Unseen Evaluation Tasks.** Lower is better.

Zero-Shot Performance on NLP Tasks

- Instruction Tuning on MultiInstruct can improve zero-shot performance on unseen NLP tasks.
- The transfer learning strategy MixedInstruct can best preserve the zero-shot capability gained on Natural Instructions dataset.

Model	RougeL
OFA	2.25
$\text{OFA}_{\text{MultiInstruct}}$	12.18
Transfer Learning from Natural Instructions	
$\text{OFA}_{\text{NaturalInstruct}}$	**43.61**
$\text{OFA}_{\text{MixedInstruct}}$	43.32
$\text{OFA}_{\text{SeqInstruct}}$	30.79

Table 4. **Zero-shot Performance on NLP tasks.** The performance is reported in Rouge-L and the best performance is in **bold**.

Conclusion

- First large-scale multi-modal instruction tuning dataset. (Contains 62 multi-modal tasks from 10 broad categories.)
- Significantly improve the zero-shot capability of OFA via instruction tuning.
- Explore several transferring learning techniques and show their benefits. Design a new metric sensitivity.

Dataset, Code, and Model

The dataset, source code, and model checkpoints are publicly available at `https://github.com/VT-NLP/MultiInstruct`.

Zhiyang Xu and Ying Shen contributed equally to this work.