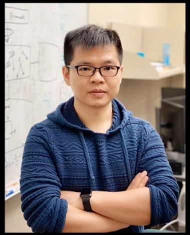
MULTIINSTRUCT: Improving Multi-Modal Zero-Shot Learning via Instruction Tuning

Zhiyang Xu*, Ying Shen*, Lifu Huang

Department of Computer Science, Virginia Tech





*Equal Contribution

Pre-trained Language Models for Downstream Tasks

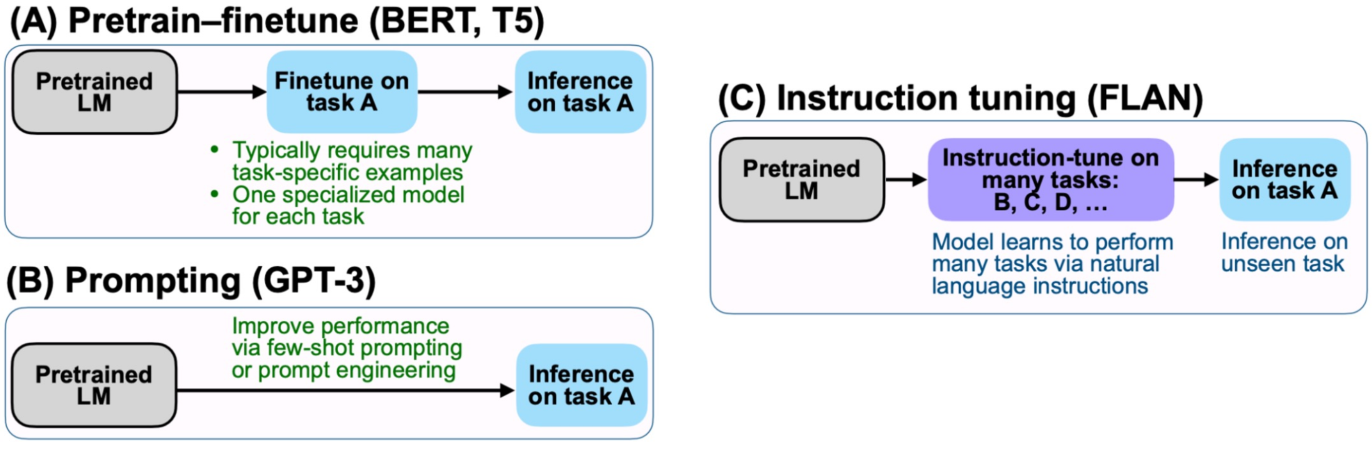


Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.

Image credit: Wei, Jason, et al. "Finetuned language models are zero-shot learners."

Language-only

Instruction Tuning on
Multimodal Pre-trained Models

Imbalance in Instructional Datasets between NLP and Multimodal

1600+ Language-only instruction tasks

NO large-scale, publicly-available multimodal instruction tasks

Wang, Yizhong, et al. "Benchmarking generalization via in-context instructions on 1,600+ language tasks." arXiv preprint arXiv:2204.07705 (2022).

MultiInstruct

The ***first*** multimodal instruction tuning benchmark dataset

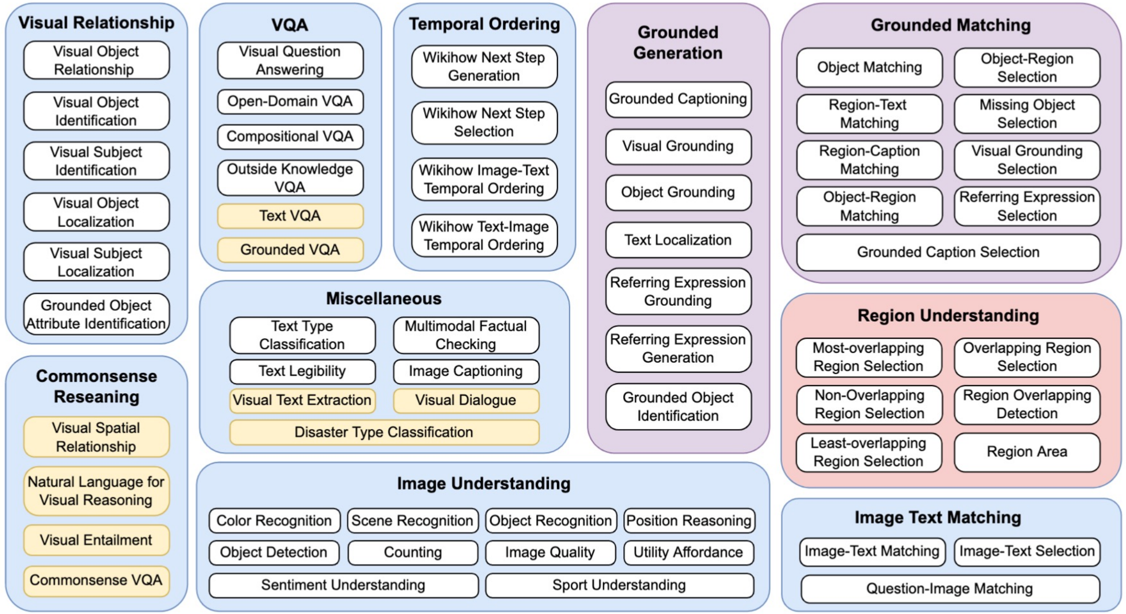
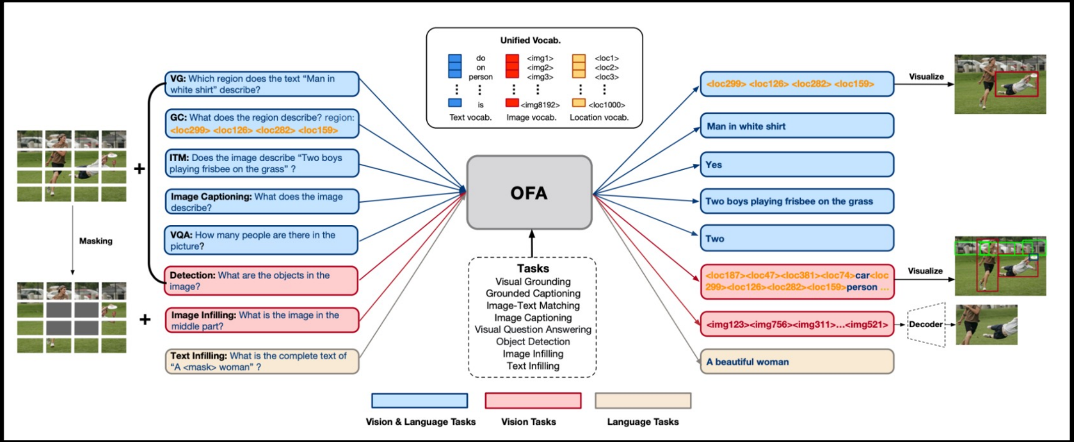


Figure 2: **Task Groups Included in MultiInstruct.** The yellow boxes represent tasks used for evaluation, while the white boxes indicate tasks used for training.

- 62 diverse multimodal tasks
- 10 broad groups
- 5 expert-written instructions

OFA (One For All)

- A unified multi-modal pre-trained model that is capable of performing both understanding and generation tasks with single or multiple modalities.
- OFA has a ***unified vocabulary*** for language, image tokens and the coordinates of a bounding box.



Wang, Peng, et al. "Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework."

MultiInstruct

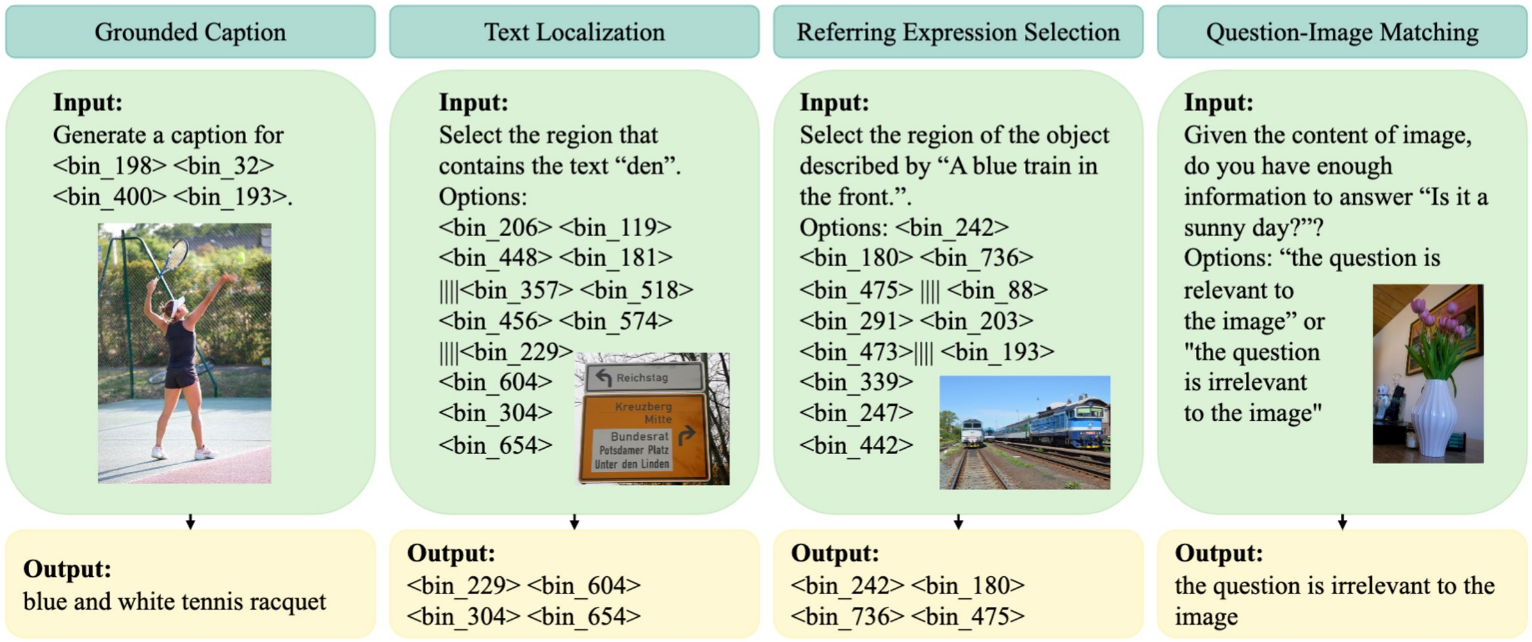


Figure 1: **Example Instances from MultiInstruct for Four Tasks.**

Multi-modal Instruction Tuning

Multi-Modal Instruction Turning

- Training Dataset Construction:
  - Use 53 tasks from 9 groups for training.
  - Sample 10,000 instances per task.
- Testing Dataset Construction:
  - Reserve the entire *Commonsense Reasoning* group for testing.
  - Select additional 5 tasks from VQA and Miscellaneous groups.
  - We use all the instances in the test split for each task.
  - Randomly sample 20 tasks from the test split of ***Natural Instructions*** dataset as unseen tasks for NLP.

Implementation Details

- Training details:
  - Pre-trained OFA-Large model (472M)
  - Mix all the instances for all tasks.
  - Each instance is randomly combined with one of its five instruction templates.

- Testing details:
  - For each task, we conduct a total of five experiments by evaluating the model using one of the five instructions in each experiment.
  - We report the mean and maximum performance and the standard deviation of the performance across all five experiments.

Evaluation Metrics

- For ***multi-modal classification tasks*** (Visual Entailment, Visual Spatial Reasoning, Natural Language Visual Reasoning, and Disaster Type Classification) we report the ***Accuracy***.
- For ***multi-modal generation tasks*** (Commonsense VQA, Text VQA, Grounded VQA, Visual Text Extraction, and Visual Dialogue) we report the ***Rouge-L***.
- For ***NLP tasks***, we report ***Rouge-L***.

- We also compute the ***aggregated performance*** for each model based on the mean of the model's performance on all multimodal and NLP unseen tasks. We use ***Rouge-L*** as the performance score for most tasks, and ***Accuracy*** for tasks that only have accuracy as a metric.

Sensitivity

How sensitive the model is towards to ***variety*** of instructions for the ***same task***:

- Ability to consistently produce the same results for the same task, regardless of slight variations in the wording of instructions.

$$\mathbb{E}_{t \in T}\left[\frac{\sigma_{i \in I^t}\left[\mathbb{E}_{(x,y) \in \mathcal{D}^t}\left[\mathcal{L}(f_\theta(i,x),y)\right]\right]}{\mu_{i \in I^t}\left[\mathbb{E}_{(x,y) \in \mathcal{D}^t}\left[\mathcal{L}(f_\theta(i,x),y)\right]\right]}\right]$$

Effectiveness of Instruction Tuning on MULTIINSTRUCT

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
$OFA_{TaskName}$	48.99	-	29.01	-	55.70	-	53.76	-	55.35	-
$OFA_{MultiInstruct}$	**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										
$OFA_{NaturalInstruct}$	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
$OFA_{MixedInstruct}$	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
$OFA_{SeqInstruct}$	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		Visual Text Extraction		Visual Dialogue		Disaster Type Classification	
	RougeL		RougeL		RougeL		RougeL		ACC	
	Max	Avg± Std	Max	Avg± Std	Max	Avg± Std	Max	Avg ± Std	Max	Avg ± Std
OFA	15.21	9.30 ± 5.42	0.02	0.00 ± 0.01	36.31	17.62 ± 16.82	45.46	28.71 ± 9.81	14.30	9.64 ± 4.34
$OFA_{TaskName}$	23.80	-	0.00	-	36.30	-	25.18	-	62.65	-
$OFA_{MultiInstruct}$	**27.22**	26.46 ± 0.83	**64.32**	47.22 ± 23.08	**74.35**	**62.43** ±11.56	**46.38**	32.91 ±7.59	64.88	56.00 ±12.96
Transfer Learning from NATURAL INSTRUCTIONS										
$OFA_{NaturalInstruct}$	5.59	5.40 ± 0.24	0.00	0.00 ± 0.00	5.65	1.24 ± 2.48	30.94	27.91 ± 2.16	56.64	38.21 ± 15.35
$OFA_{MixedInstruct}$	24.15	23.67 ± 0.47	63.79	**54.99** ± 18.16	62.43	46.56 ± 14.92	46.08	**38.02** ± 5.25	**68.31**	**64.31** ± 2.39
$OFA_{SeqInstruct}$	27.03	**26.67** ± 0.47	64.19	54.46 ± 15.96	71.63	60.62 ± 12.31	46.17	35.10 ± 6.92	64.46	57.89 ± 9.51

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

Impact of Increasing Multimodal Instruction Task Clusters

- Img Und
  - VQA + Image Understanding
- Grounding
  - Grounded Matching + Grounded Generation
- MISC, ITM
  - Temporal Ordering + Miscellaneous + Image Text Matching
- Relation
  - Visual Relationship
- Region
  - Region Understanding
- NLP
  - NLP tasks

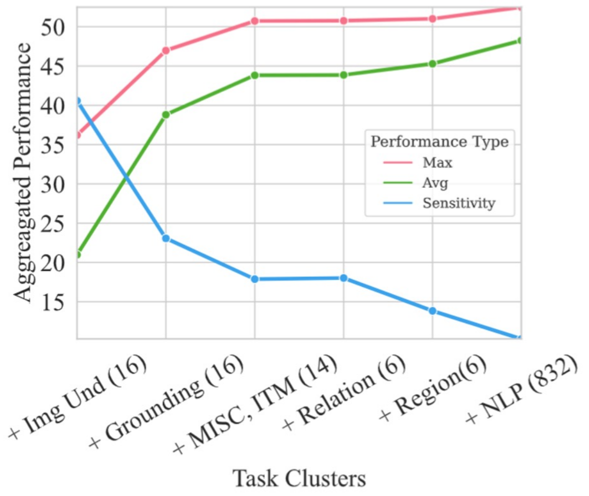


Figure 3: **Model Performance as the Number of Multimodal Instruction Task Clusters Increases.**

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.

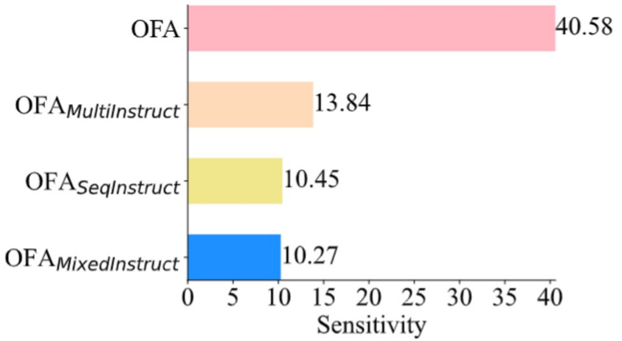


Figure 4: **Model *Sensitivity* on Unseen Evaluation Tasks.** Lower is better.

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*.

One More Thing!

We are collecting a much larger multimodal instruction tuning dataset with around 150 additional vision-language tasks and we will release them soon!