

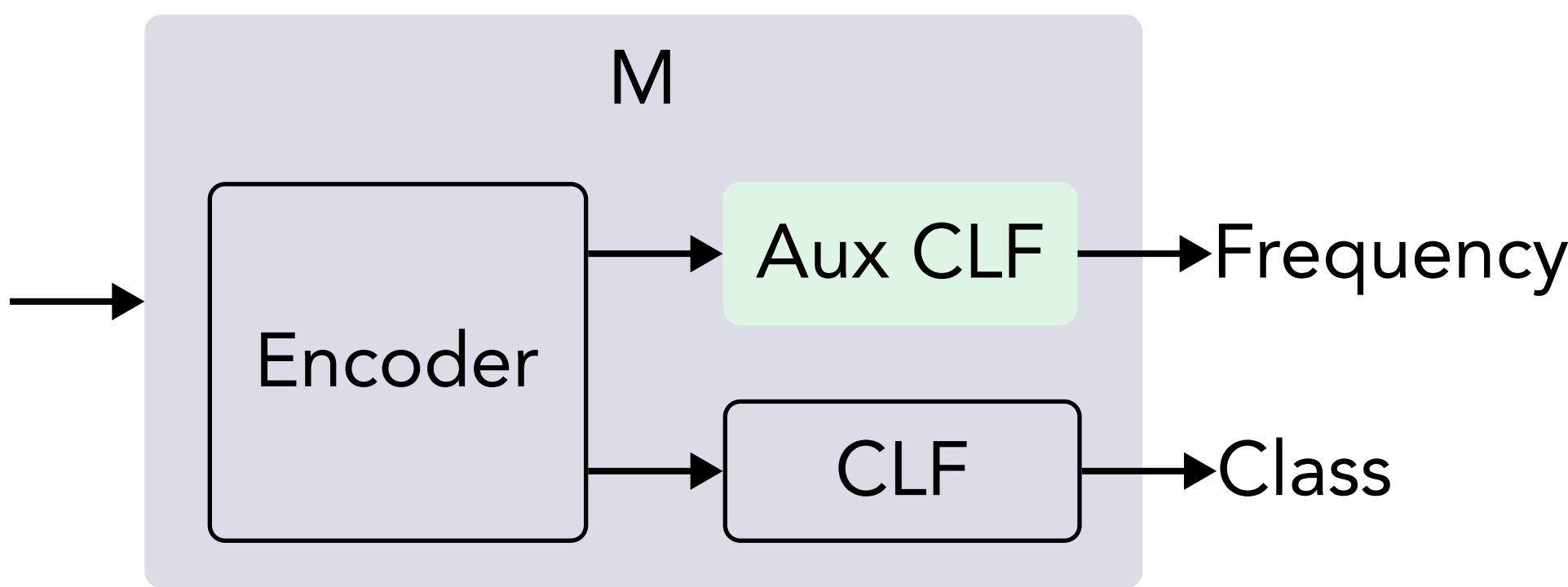
Motivation

- **Long-tailed distribution** dataset leads to biased model performance on different tasks (comm, freq, rare), we aim to teach the models how to predict the frequency of an input image.
- Facing **fine-grained image** classification task, we aim to use external information, i.e., text of the labels, for guiding our models.

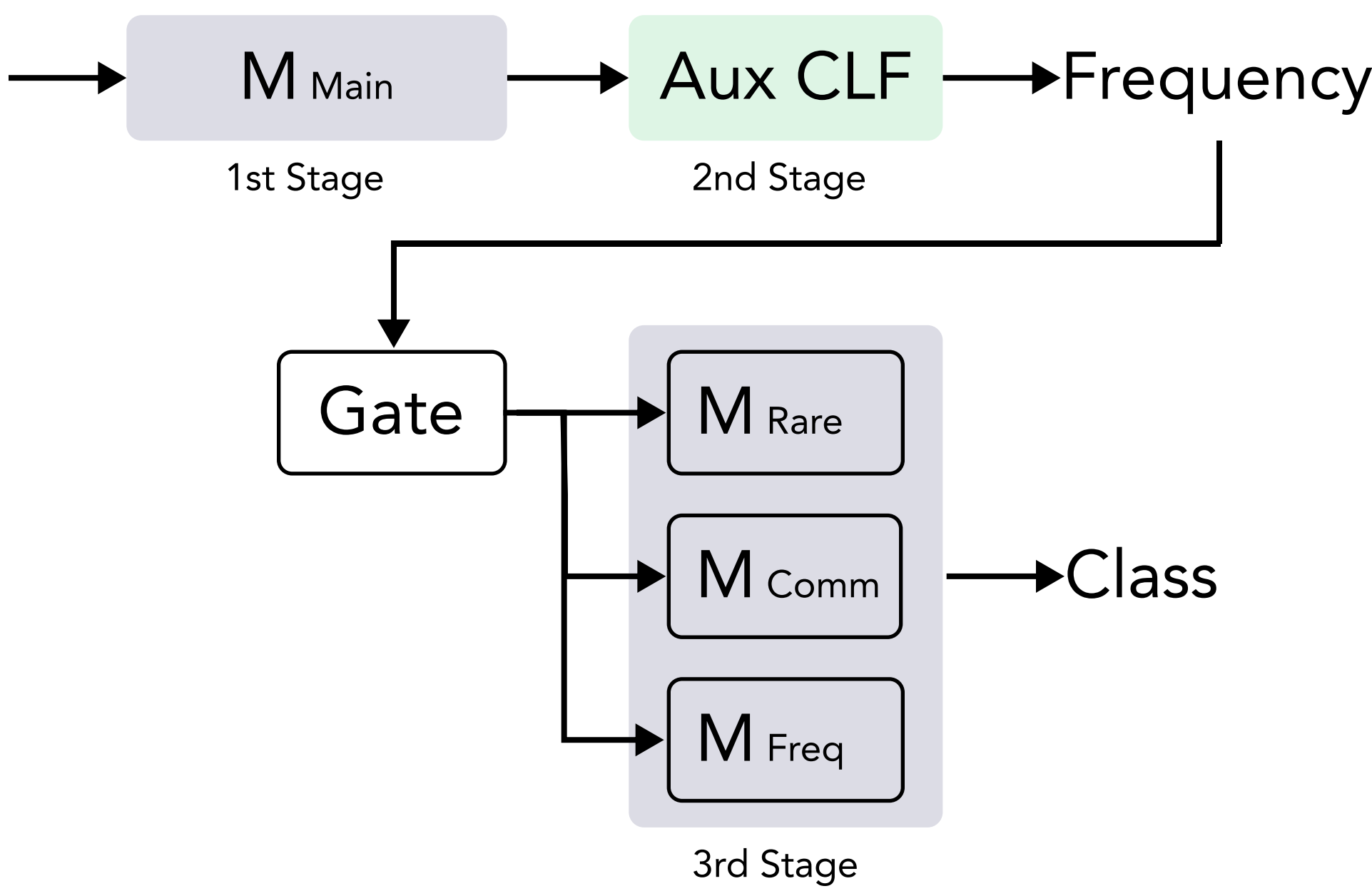
Methods

Auxiliary Classifier

- Problem: Long-tailed
- **Auxiliary classifier** in multi-task fashion



Auxiliary classifier as task separator

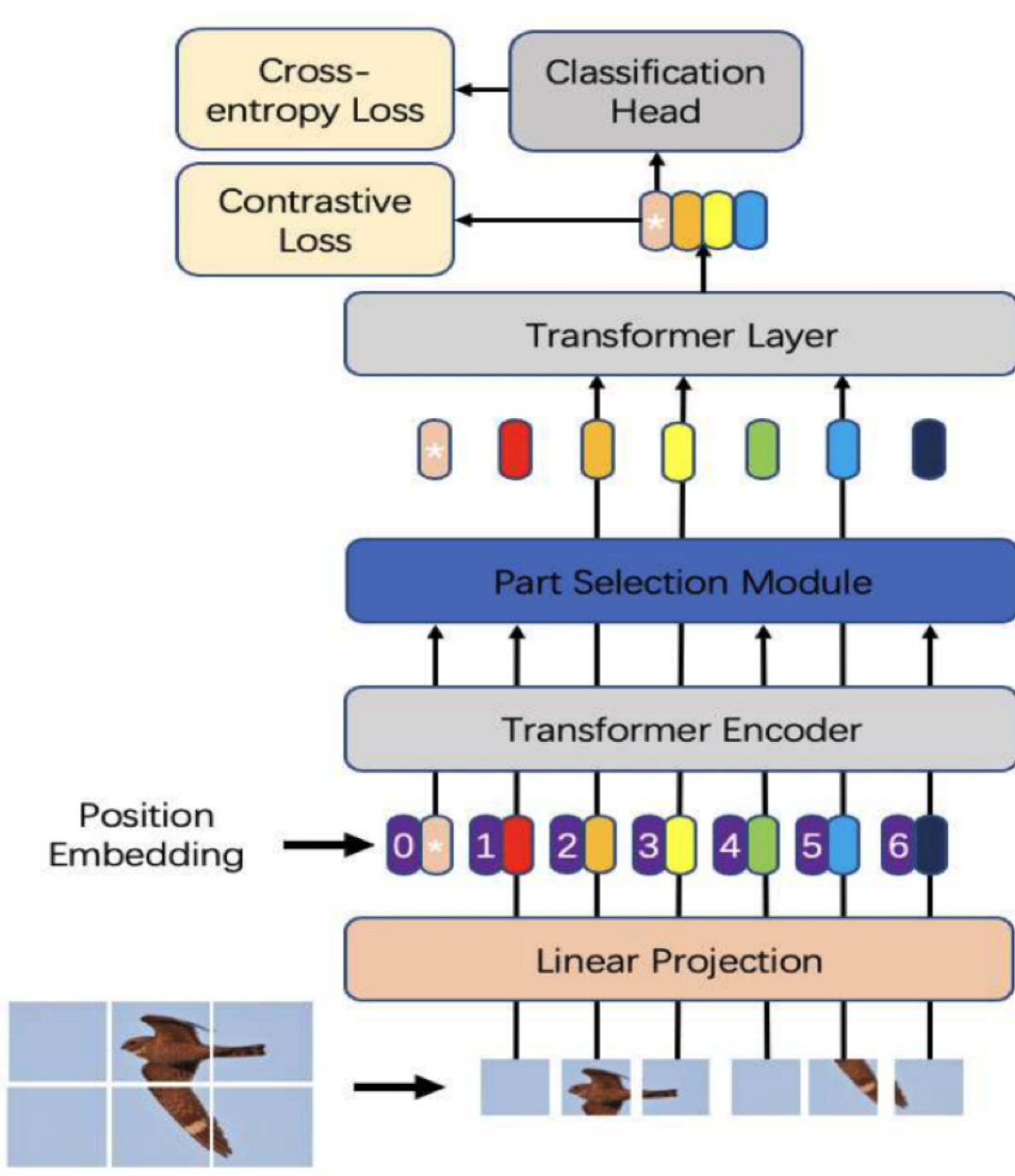


Results

Method	Freq.	Comm.	Rare	Main
Baseline	0.894	0.715	0.284	0.730
Aux CLF / Multi-Task				
Aux CLF / 4-Stage				
Multi-Model / BERT				
Multi-Model / CLIP				

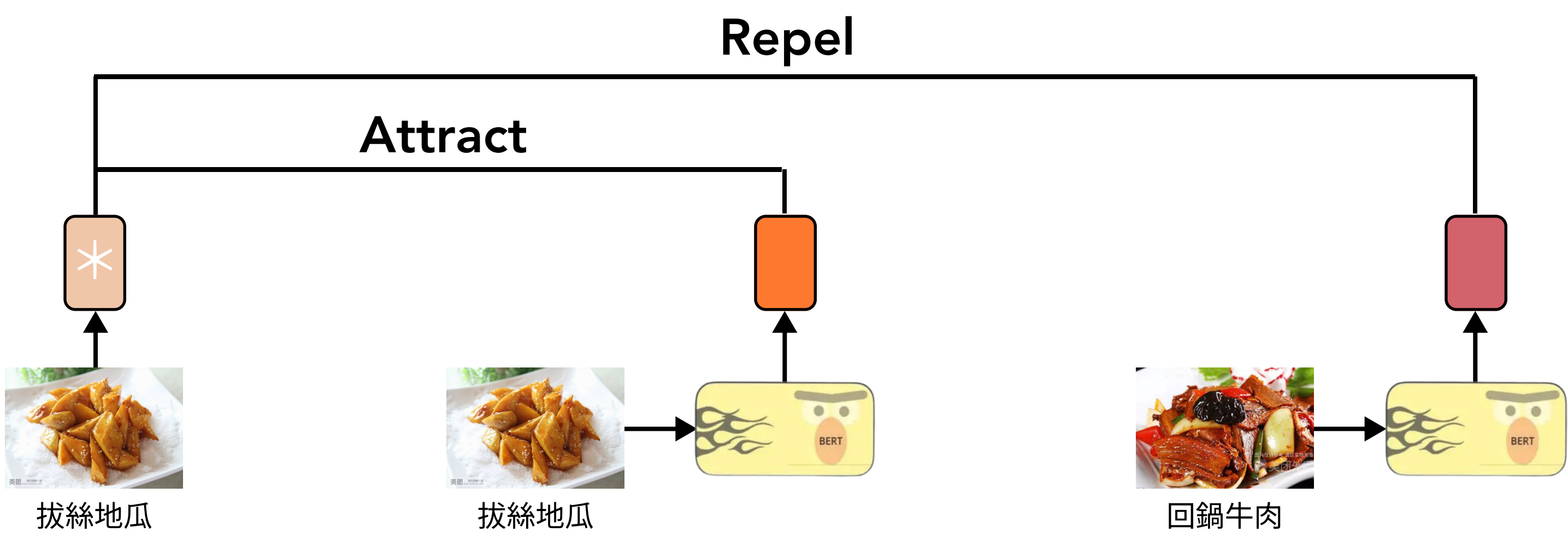
Baseline Approach

- TransFG
- Choose tokens with maximum attention for each head in the second last layer.

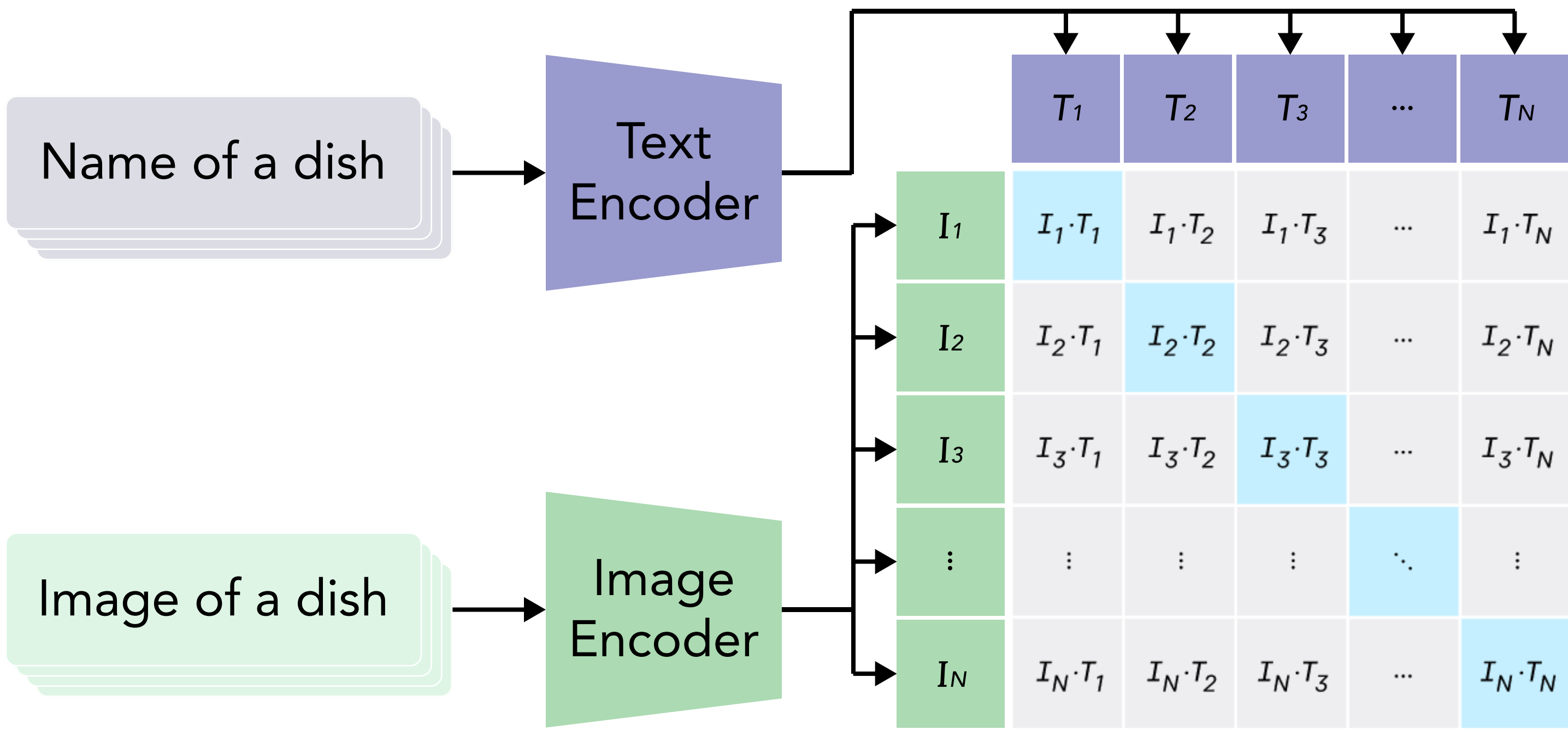


Multi-Modality

- Problem: Fine-grained
- LM embedding as cls token + contrastive learning



CLIP-zero shot vs fine-tune on CLIP



Conclusion

- We propose four methods, two of which based on auxiliary classifiers and the other two are multi-modality based.
- The approaches with auxiliary classifiers aim at the long-tailed distribution problem while the multi-modal approaches aim at the fine-grained classification problem.
- [Method] gives the best performance overall with the highest accuracy on [Task] track.