

Few-Shot Multimodal Named Entity Recognition based on Multimodal Causal Intervention Graph

Feihong Lu¹, Xiaocui Yang², Qian Li⁴, Qingyun Sun^{1*}, Ke Jiang¹,
Cheng Ji¹, Jianxin Li^{1,3}

¹School of Computer Science and Engineering, BDBC, Beihang University, Beijing, China

²School of Computer Science and Engineering, Northeastern University, Shenyang, China

³Zhongguancun Laboratory, Beijing, China

⁴Beijing University of Posts and Telecommunications, Beijing, China

{lufh,sunqy,jiangke22,jicheng,lijx}@act.buaa.edu.cn,yangxiaocui@stumail.neu.edu.cn

13240016260@163.com



北京航空航天大学
BEIHANG UNIVERSITY

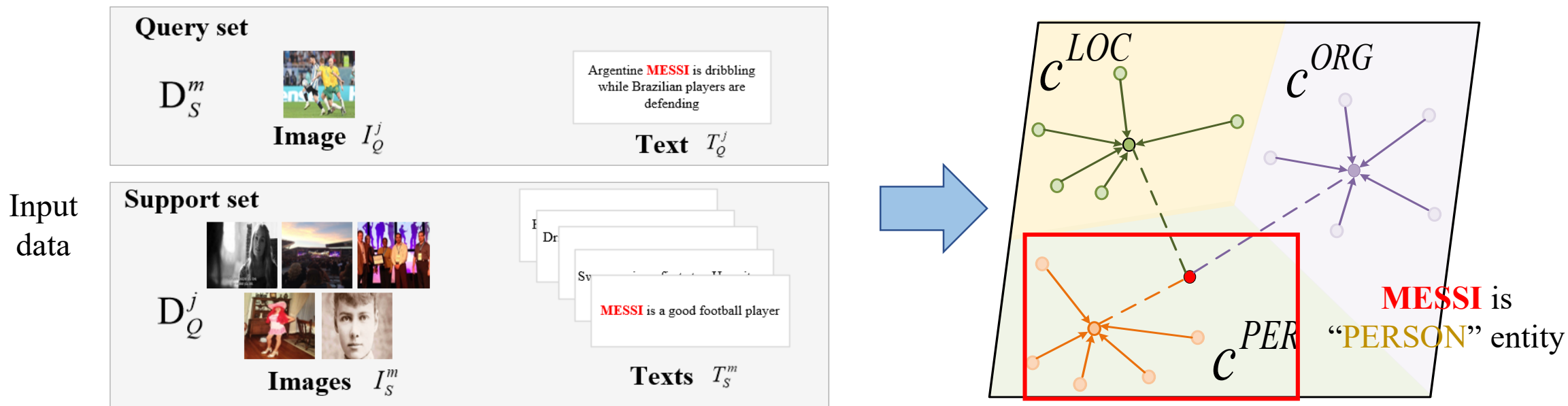


東北大學
Northeastern University



北京郵電大學
Beijing University of Posts and Telecommunications

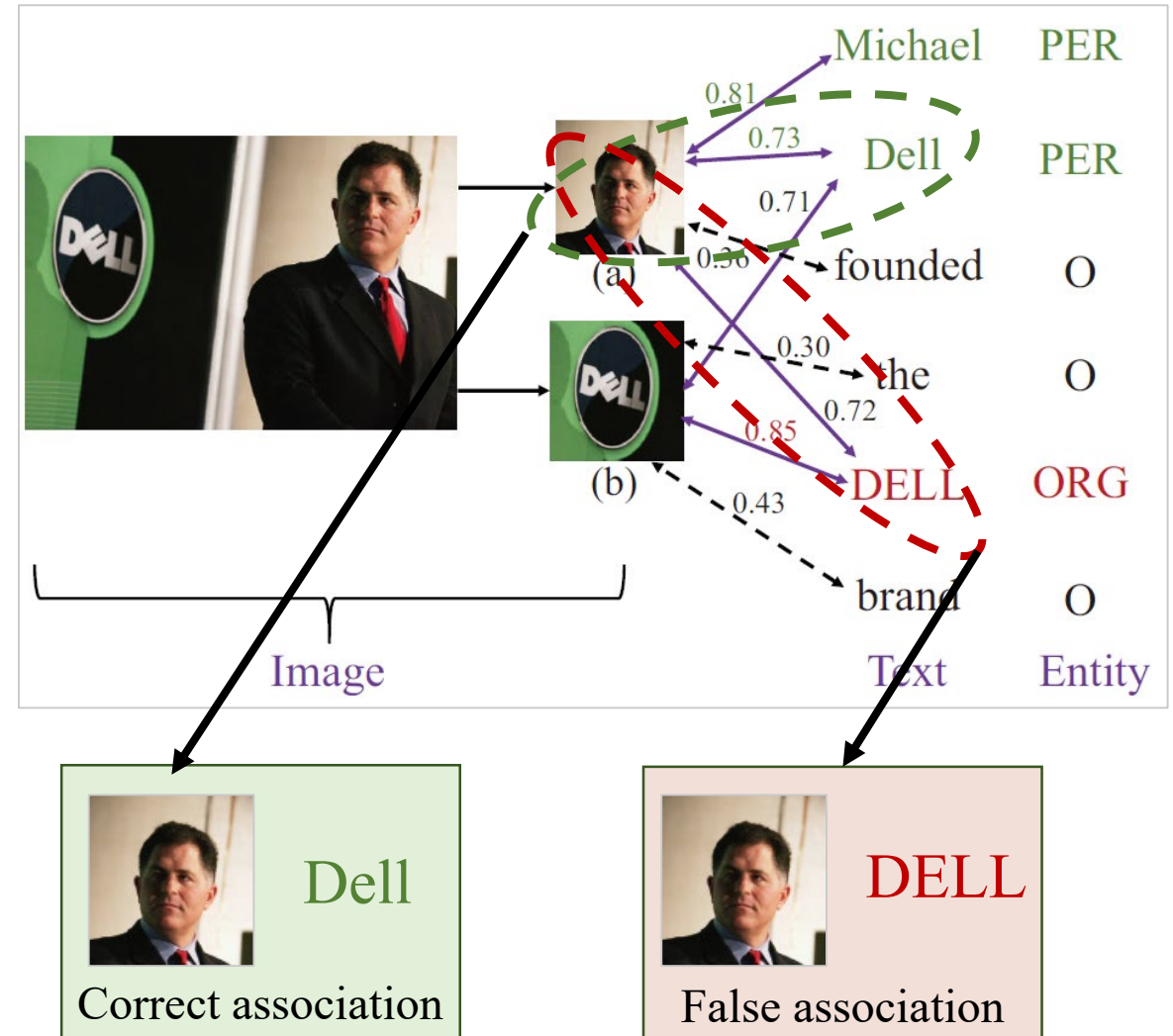
Definition of FMNER Task



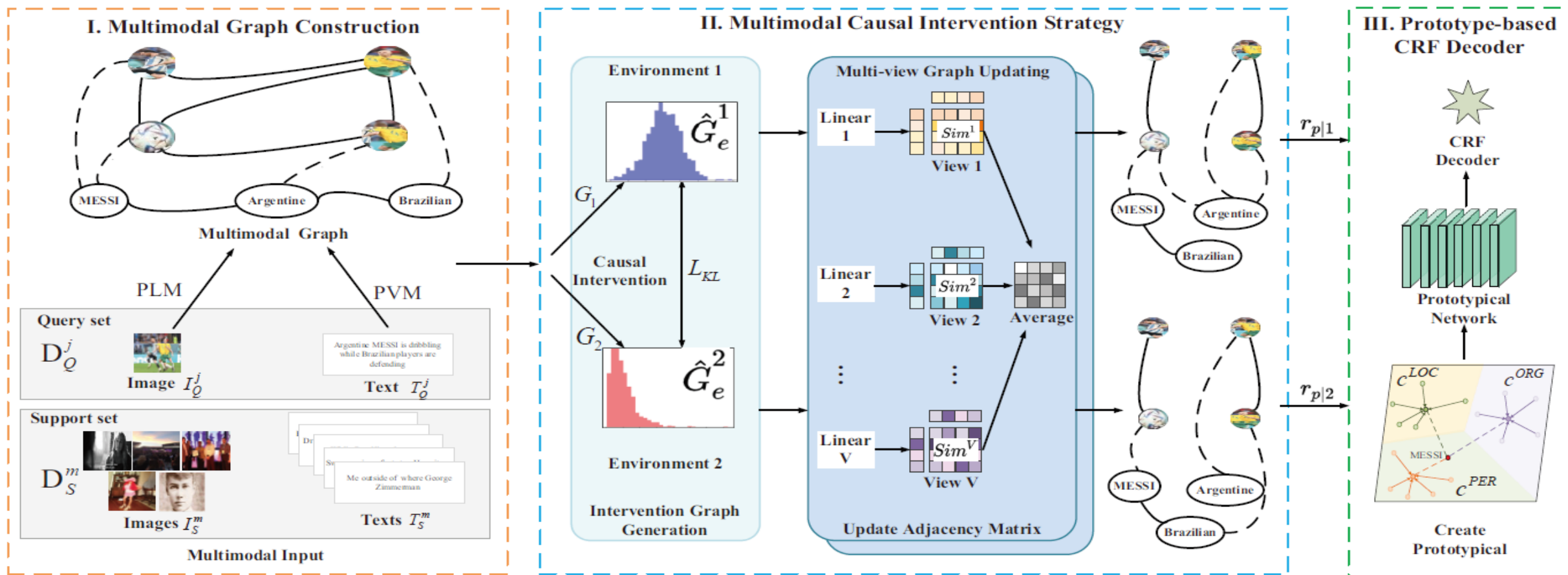
We define the FMNER setting where the model is trained on source domains with annotations $D_S^m = (\mathcal{T}_S^m, I_S^m)$ with source tag set C_S^m and then tested on target domains $D_Q^j = (\mathcal{T}_Q^j, I_Q^j)$ with target tag set C_Q^j by only providing a few labeled examples per entity type, where \mathcal{T} is the text modality, I is the image modality, m is the m -th entity type, j is j -th entity type, and $C_S^m \cap C_Q^j = \emptyset$. Formally, the setting of N-way K-shot is defined as follows: given K text-image pairs for each entity type from D_Q as input, $x = (t, i)_{k=1}^K \in (\mathcal{T}_Q, I_Q)$ and make the best tag sequence y , where $[C_Q] = N$.

Motivation

- **Motivation:** The *limited resource challenge*, where each sampling instance yields different content, resulting in data bias and alignment problems of multimodal units (image patches and words).
- **Lack of labeled data.** FMNER faces a significant challenge called data sampling bias, which creates spurious connections between multimodal data, thereby exacerbating the risk of overfitting due to incorrect projections between multimodal representations and entity types.
- **Differences in multimodal semantics.** Since global representations of different modalities cannot capture effective fine-grained semantic information, establishing accurate alignment of different modalities is difficult.

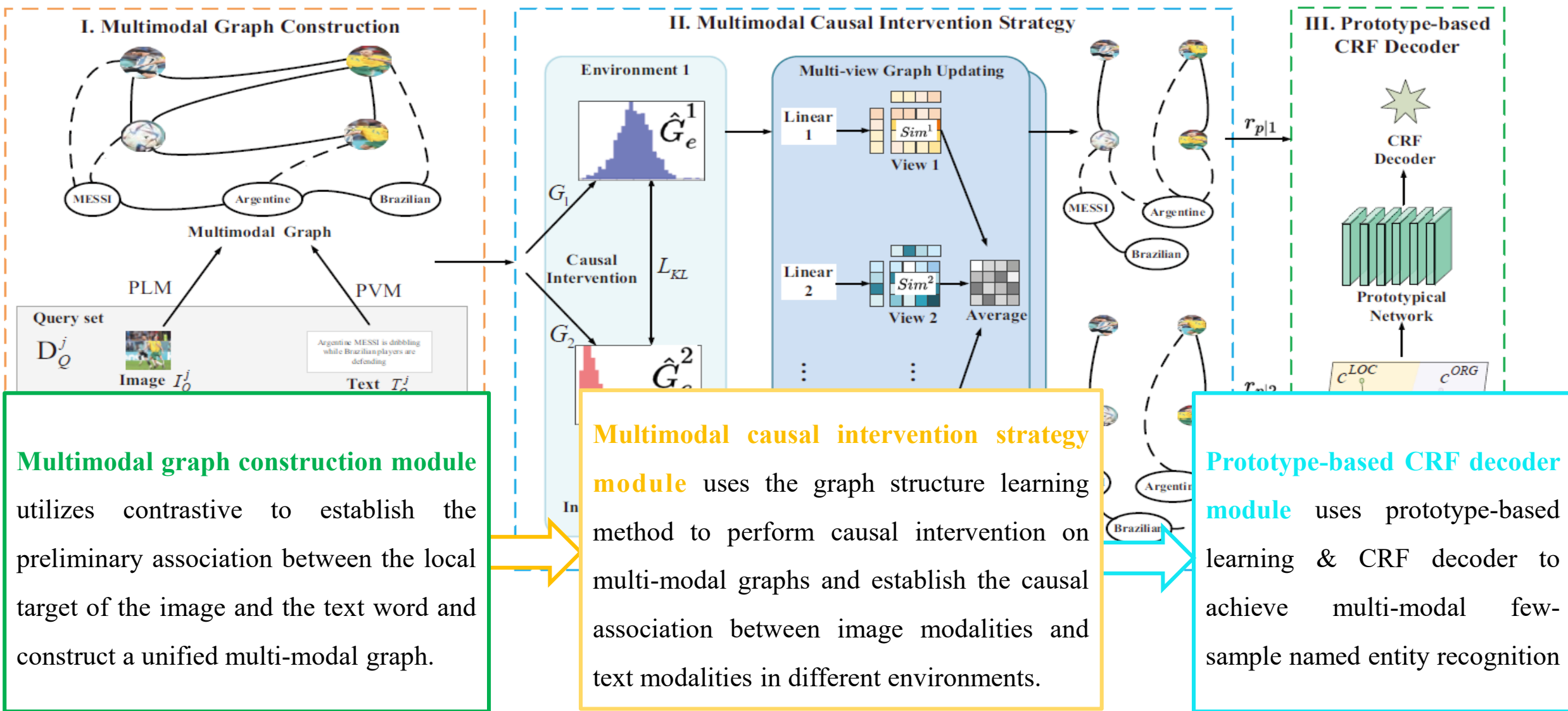


Our Framework



To overcome the above challenges, we propose a novel Multimodal causal Intervention Graphs (**MOUSING**) model for the FMNER task to more effectively capture the fine-grained alignment between different modalities. MOUSING including the Multimodal graph construction module, Multimodal causal intervention strategy module and Prototype-based CRF decoder module.

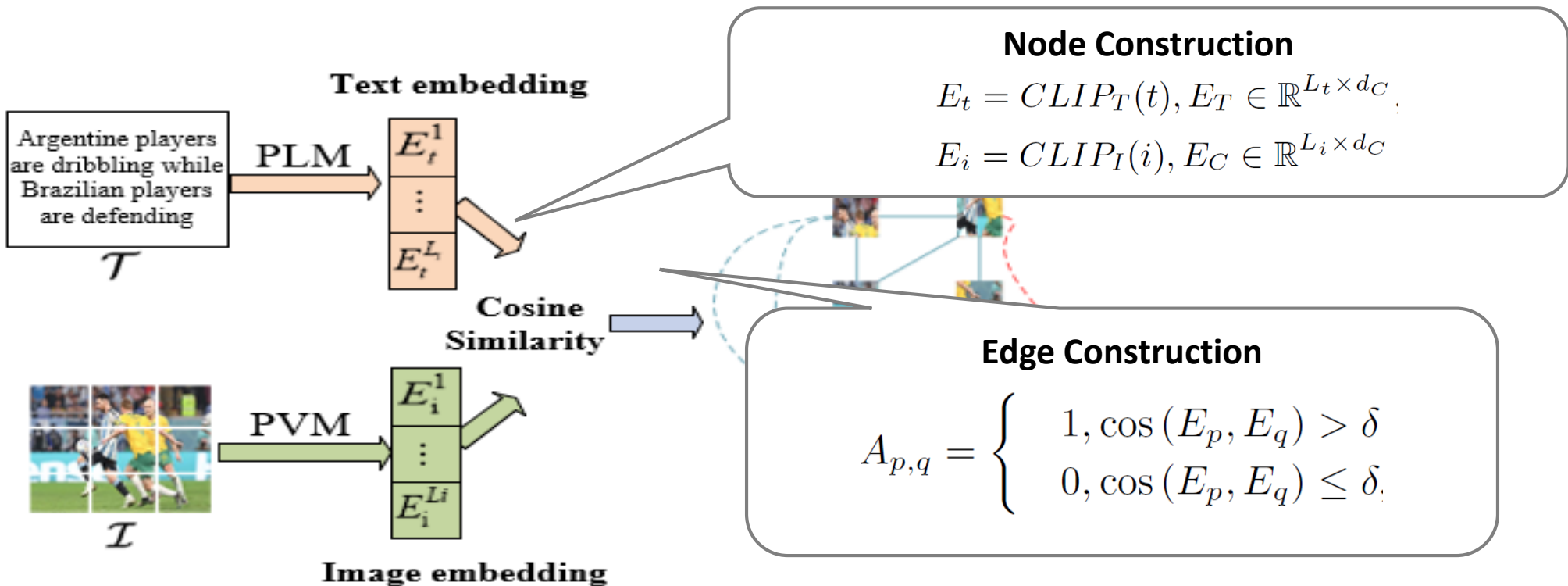
Our Framework



Our Framework

□ Multimodal graph construction

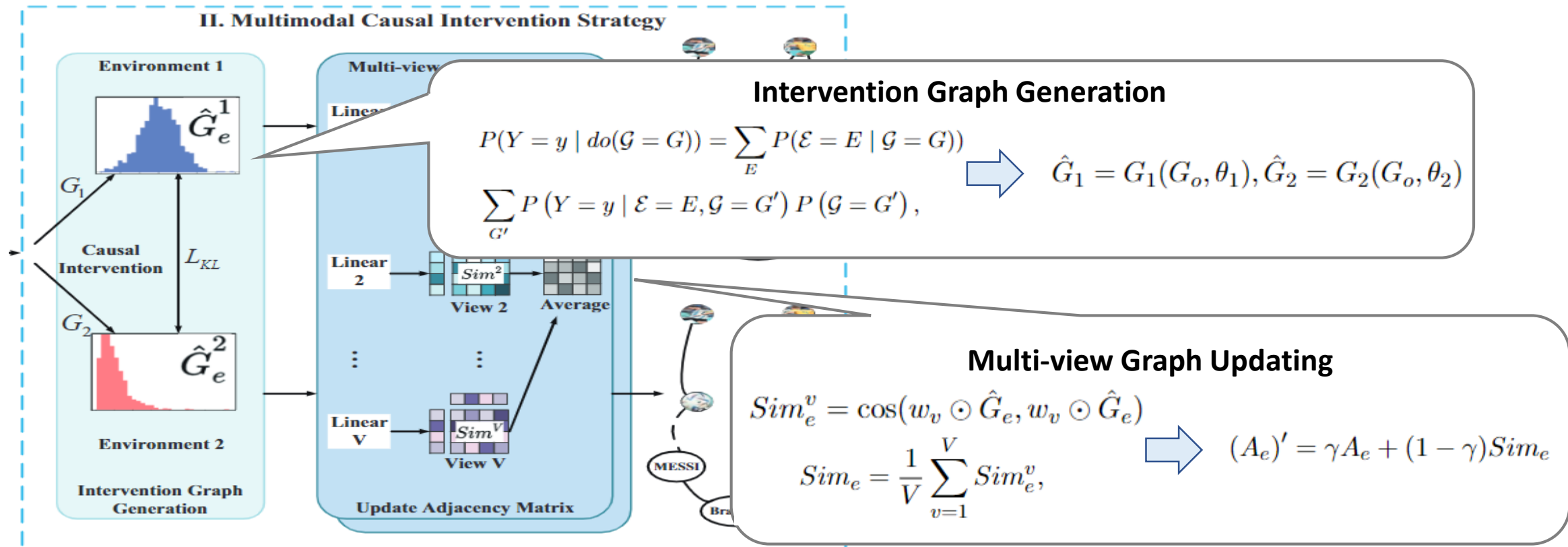
- To obtain a **unified representation of different modalities**, we use the CLIP model to extract text word representations and image patch representations, and use the encoding of each visual patch and text word as a node of the multimodal graph.
- In order to better **establish accurate associations between different nodes**, we calculate cosine similarity between different node features to capture valuable interactions between multi-modal semantic units (including intra-modal nodes and inter-modal nodes).



Our Framework

□ Multimodal causal intervention strategy

- We propose a novel multimodal causal intervention strategy (MCIS) to address the issue of unstable and biased correlations caused by data selection bias in the few-shot scenario. MCIS can adaptively learn stable associations between different modalities, reducing bias associations between different nodes while emphasizing the correct associations among them, which contains two steps: the **Intervention Graph Generation** and the **Multi-view Graph Updating**.



Our Framework

□ Prototype-based CRF decoder

- **Build a prototype center.** We leverage Meta-Learning based on the prototypical network to handle FMNER. In each batch, we randomly sample a few instances as query set and other K instances as support set. Each prototype is the mean vector of the embedded support points belonging to its class:

$$c^m = \frac{1}{|S^m|} \sum_{(x_k, y_k) \in S^m} G_e(x_k),$$

- **Get the calculation results of the prototype.** The prototypical networks produce a distribution over classes for each word of $x_q \in Q$ based on a softmax over distances to the prototypes in the embedding space:

$$p(y_w = m | x^q, G_e) = \frac{\exp(-d(G_e(x_q), c^m))}{\sum_{m'} \exp(-d(G_e(x_q), c^{m'}))}$$

- **Optimize the model using the CRF loss function.** Finally, we employ Cross-Entropy loss, CE , to calculate classification loss.

$$L_{CRF} = CE(D(P|G_1), y) + CE(D(P|G_2), y).$$

Experiment

Modality	Model	Twitter-2015						
		Per+Loc	Per+Org	Per+Others	Loc+Org	Loc+Others	Org+Others	Avg.
TNER	BERT	17.99	13.33	15.43	14.85	13.71	17.61	15.49
	ProtoBERT	18.52	22.26	20.71	15.89	14.80	17.83	18.34
	RoBERTa	20.11	16.73	17.71	16.25	17.56	21.16	18.25
	ProtoRoBERTa	20.51	19.41	19.52	20.19	18.57	23.53	20.29
	NNshot	18.65	27.24	28.21	20.81	28.78	25.90	24.93
	Structshot	18.66	30.41	28.37	24.87	31.13	29.05	27.08
MNER	UMT	18.57	21.43	24.24	17.14	14.22	24.26	19.98
	UMT-CLIP*	28.57	22.23	31.25	27.27	12.46	22.50	24.05
	UMGF	20.00	17.77	24.94	17.20	16.23	25.71	20.31
	UMGF-CLIP*	23.30	21.73	32.21	22.50	20.54	22.31	23.76
	ProtoUMGF	23.33	25.39	27.14	25.10	18.65	26.34	24.33
	ProtoUMGF-CLIP*	23.79	20.30	31.88	23.45	20.85	22.46	23.78
	HVPNet	24.97	23.81	26.09	14.38	21.14	19.35	21.62
	ProtoHVPNet	19.35	29.13	37.20	21.43	17.03	16.74	23.48
	MOUSING	34.86	28.43	35.32	30.10	36.62	27.99	31.22

It shows performance comparisons on Twitter-2015 dataset with six different splits on few shot setting.

All the observations demonstrate the effectiveness of the MOUSING framework.

Experiment

Modality	Model	Twitter-2017						
		Per+Loc	Per+Org	Per+Others	Loc+Org	Loc+Others	Org+Others	Avg.
TNER	BERT	14.29	20.69	15.39	9.04	12.55	12.03	13.99
	ProtoBERT	12.60	23.26	14.03	21.95	18.67	15.87	17.73
	RoBERTa	20.59	15.39	22.73	11.76	16.44	13.14	16.68
	ProtoRoBERTa	21.21	24.10	23.14	23.36	17.51	19.67	21.50
	NNshot	24.27	29.82	25.29	23.23	23.82	22.78	24.86
	Structshot	25.41	29.39	22.89	24.45	24.20	27.22	25.59
MNER	UMT	24.00	16.76	12.50	20.59	17.54	17.39	18.13
	UMT-CLIP*	33.32	25.02	13.34	22.58	20.02	13.04	21.22
	UMGF	24.24	17.65	14.63	22.43	23.73	13.04	19.29
	UMGF-CLIP*	19.18	19.28	11.27	15.52	24.11	27.27	19.44
	ProtoUMGF	20.41	16.00	26.12	24.62	23.18	19.23	21.59
	ProtoUMGF-CLIP*	18.20	21.36	11.77	14.04	14.17	22.18	17.07
	HVPNet	32.50	16.28	16.67	18.23	14.07	25.41	20.53
	ProtoHVPNet	29.05	24.94	24.29	13.16	17.18	16.21	20.81
	MOUSING	34.12	28.10	27.14	24.12	24.85	25.79	27.35

It shows performance comparisons on Twitter-2017 dataset with six different splits on few shot setting.

All the observations demonstrate the effectiveness of the MOUSING framework.

Experiment

Variants	Precision	Recall	F1
w/o Image	26.59	26.96	26.72
w/o Intervention	24.94	25.09	25.00
w/o Multi-view	26.12	25.75	25.72
w/o MICS	19.36	19.11	19.06
w/ Random Intervention	29.99	30.69	30.24
w/ Gaussian Intervention	30.30	30.75	30.47
MOUSING (Ours)	32.89	33.70	33.18

To investigate the effectiveness of each module in MOUSING, we conduct variant experiments.

All the observations demonstrate the effectiveness of each component in our model.

Experiment

□ Hyperparameters Setting

- We conduct experiments for different hyperparameters, including V of Eq. (6), δ of Eq. (2), γ of Eq. (7), L of Eq. (9), and λ of Eq. 14. The experimental results are shown in Figure

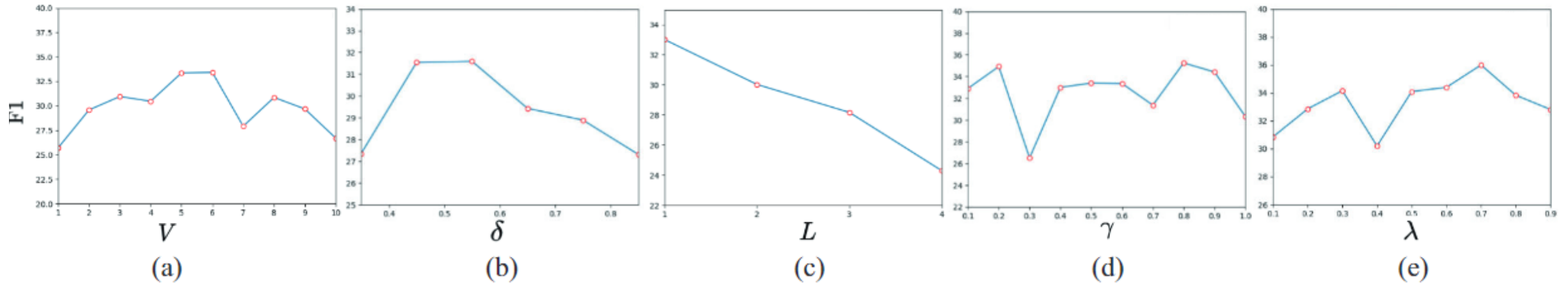


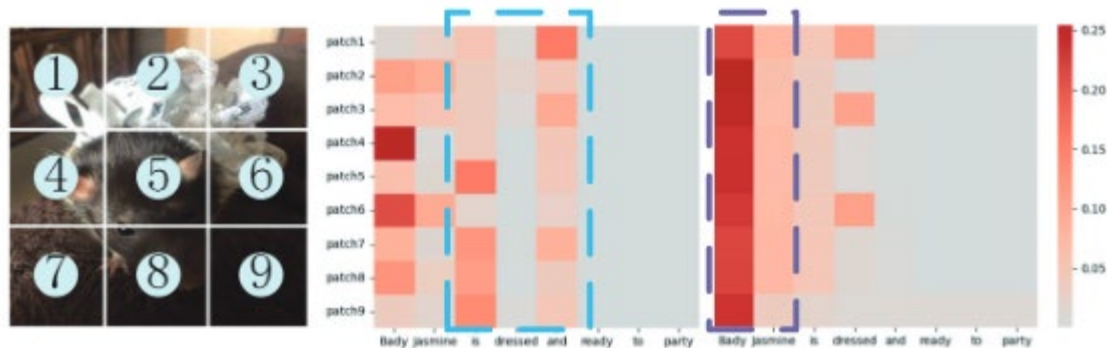
Figure 4: Hyperparameters experiments. **F1** comparisons of V for MOUSING.

The hyperparameters selected by our model are obtained through a large number of experiments and are reasonable and effective.

Experiment


□ Case Study

- The initial multimodal graph **establishes strong associations between ``is, and'' and all image patches**, so this association cannot assist in text prediction, as (b) shows. After updating our model, the new graph weakens invalid associations and **strengthens beneficial associations between patches and words, such as the association between image patches and ``Baby Jasmine''**, which will more effectively assist the model in making a detection.
- The comparative results for the case studies are shown in (d), where our model performs the best.



(a) Image data (b) Original graph (c) Updated graph

Text input: **Bady Jasmine** is dressed and ready to party

	Entity	Label	MOUSING	structshot	UMGF
	Bady	PER	✓	✗	✗
	Jasmine	PER	✓	✓	✓

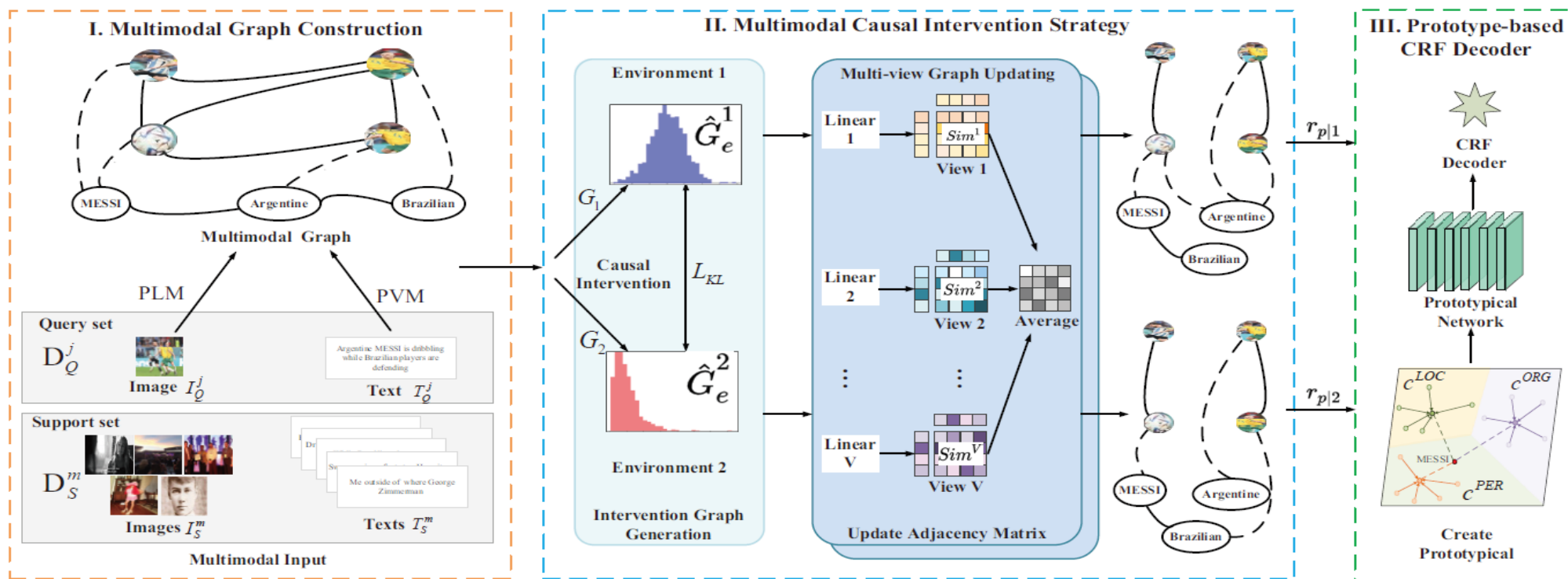
(d) Case example

It demonstrates how our model refines associations in a multimodal graph to enhance text prediction and detection, supported by visual examples and comparative case studies.

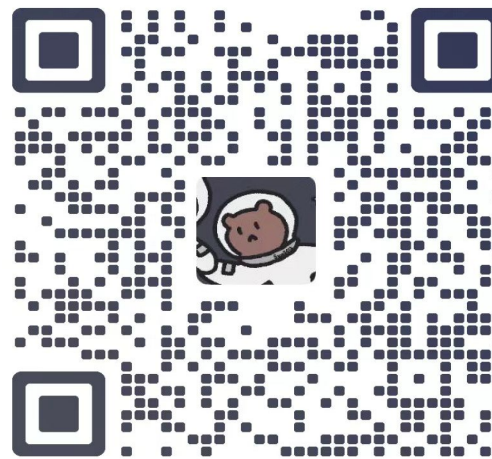
Conclusion

We propose a novel Multimodal causal Intervention Graphs (MOUSING) model for the FMNER task to more effectively capture the fine-grained alignment between different modalities.

- ① We propose a novel Multimodal causal intervention Graph (MOUSING) which builds deeper correlations among different modalities, to handle the Multimodal Named Entity Recognition task in a multimodal few-shot scenario.
- ② We first construct a multimodal graph to integrate fine-grained information. After that, Multimodal Causal Intervention Strategies (MCIS) are introduced to simulate multiple training environments to perform causal interventions, followed by a multi-view graph update method to improve fine-grained alignment across modalities.



WeChat



Thanks!
Questions and Advices?

Email: lufh@act.buaa.edu.cn